SIGNAL PROCESSING METHODS FOR MENTAL FATIGUE MEASUREMENT AND MONITORING USING EEG

SHEN KAIQUAN

(B. Sci., USTC)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2008

Acknowledgments

I am deeply indebted to my supervisors Prof. Li Xiaoping, Prof. Einar P. V. Wilder-Smith and Assoc. Prof Ong Chong-Jin. Without their wide spectrum of expertise, this interdisciplinary doctoral research would not be possible. Prof. Li, the director of our research laboratories, has a very strong bioengineering background, steering the research with his insightful envisions; Prof Einar, as an experienced neurologist, flavors this research with a strong neurophysiology-driven appetite; Assoc. Prof Ong has given freely of his precious time and expertise to contribute on signal processing methodologies and many signal processing ideas in this research stemmed from enlightening discussions with him.

I also wish to record my deep gratitude to my friends and colleagues in Neurosensors Laboratories for their valuable suggestion, support and encouragement. The life with them is memorable and inspiring.

Last but by no means least, I am most grateful to my parents and brothers for their loves, encouragements and moral supports. Special thanks to my wife, Karen, and my daughter, Amanda. Their loves made me strong to adventure ahead.

Table of Contents

Summary

In recent years, there have been increasing interests in mental-fatigue tracking technologies with the widespread hope that they will be invaluable in the prevention of fatigue-related accidents. This thesis is concerned with developing novel signal-processing methods that enable automatic mental-fatigue measuring and monitoring in human individuals from their electroencephalogram (EEG) recordings. New methods for automatic EEG artifact removal, feature selection and multi-class classification are proposed and tested in the present work.

EEG is easily contaminated by physiological artifacts from electrocardiograph (ECG), electrooculogram (EOG) and electromyogram (EMG). These artifacts typically have much higher amplitude than cerebral signals and thus impose great difficulties in EEG interpretation. In this study, a novel independent-component-analysis (ICA) based automatic EEG artifact-removal method is proposed, in which a weighted support vector machine (SVM) together with an error-correction algorithm is used for automatic identification of artifactual independent components in EEG. This combination of weighted SVM and error-correction mechanism is motivated by the special structural information of the learning problem at hand, with the former dealing with the inherent unbalancing of data and the latter exploiting some useful constraints readily available from empirical studies. Our experiments show that a significant performance advance has been obtained by the proposed method, comparing with several existing methods in the literature.

Feature selection plays an important role for the performance of a mental-fatigue measuring and monitoring system. When the underlying important features are known and irrelevant / redundant features are removed, the learning problem can be greatly simplified, resulting in an improved generalization capability and enhanced system interpretability. The work proposes new feature-selection methods. They use a novel feature-ranking criterion based on the sensitivity analysis of posterior probabilities. In loose terms, this criterion evaluates the importance of a specific feature by computing the aggregate value, over the feature space, of the absolute difference of the probabilistic outputs of the learning method with and without the feature. The proposed methods are competitive with, if not better than, some popular feature-selection methods in the literature, based on the datasets that we have tested.

For reliably classifying mental fatigue into different levels, a multi-class classification system is established using a recently-developed probabilistic support vector machine (PSVM) method. The numerical results show that it does not only give superior classification accuracy but also provides a valuable estimate of confidence in the prediction of mental fatigue levels in a given 3-second EEG epoch.

The thesis is organized as followed. Chapter 1 provides the motivation and objectives of the present work. The background knowledge needed for the subsequent chapters is given in Chapter 2. Chapter 3 gives an overview of the approach taken in this work and the detailed description of the collection and labeling of mental fatigue EEG used in the present work. The next four Chapters provide the detailed account of the proposed automatic EEG artifact removal method (Chapter 4), feature selection method (Chapters 5-6) and multi-class classification method (Chapter 7). It is worth noting that Chapter 7 also presents the prototype of the developed automatic mental-fatigue measuring and monitoring system and includes a comprehensive performance evaluation of the developed system. Conclusions are drawn in Chapter 8.

List of Tables

List of Figures

List of Symbols

Φ	the mapping function
$\|\|\cdot\|\|$	the Euclidean norm
$(\cdot)$	the inner product (or dot product) operator
α_i	the Lagrangian multiplier for the i^{th} sample
$\boldsymbol{\alpha}$	the vector of Lagrangian multipliers
$\boldsymbol{\xi}$	the vector of slack parameters
γ	the kernel parameter for the Gaussian kernel used in the support-vector-machines
ω_i	the i^{th} class
ξ_i	the slack parameter for the i^{th} sample
A	the sigmoid parameter used in Platt's probabilistic outputs
$\mathbf{A}$	the mixing matrix
A_{ij}	the sigmoid parameter used in Platt's probabilistic outputs for class i and class j
a_{ij}	the mixing coefficient for the j^{th} source to the i^{th} channel
B	the sigmoid parameter used in Platt's probabilistic outputs
b	the bias term of the hyperplane
B_{ij}	the sigmoid parameter used in Platt's probabilistic outputs for class i and class j
C	the regularization parameter used in the support-vector-machines
c	the number of classes in a c-class classification problem
C_+	the generalization parameter for the positive class
C_-	the generalization parameter for the negative class

$\mathbf{D}$	a diagonal matrix
D	the dimensionality
$d(\cdot)$	the decision function of a classifier
$\mathbb{D}$	dataset
$\mathbb{D}_{ij}$	the subset of $\mathbb{D}$ formed by samples from class i and class j
E	mathematical expectation operator
$\mathbb{H}$	the Hilbert space
$\mathbf{I}$	a unit matrix
$K(\mathbf{x}_1 \cdot \mathbf{x}_2)$	the kernel function used in the support-vector-machines
m	the number of independent components resulting from an EEG epoch
N	the total number of samples
n	the index of sample or the number of EEG channels
N_+	the number of training samples from positive class ($y = +1$) in a two-class classification problem
N_-	the number of training samples from negative class ($y = -1$) in a two-class classification problem
N_i	the number of training samples from the i^{th} class
$p_i(\mathbf{x})$	the probability of belonging to class i given $\mathbf{x}$, i.e. $P(\omega_i\|\mathbf{x})$
p_i	equivalent to $p_i(\mathbf{x})$
$p_{ij}(\mathbf{x})$	the pairwise probability of belonging to class i knowing that $\mathbf{x}$ is from class i or class j, i.e. $P(\omega_i\|\mathbf{x}, \mathbf{x} \in \omega_i \cup \omega_j)$
p_{ij}	equivalent to $p_{ij}(\mathbf{x})$
$\hat{p}_{ij}(\mathbf{x})$	the estimate of $p_{ij}(\mathbf{x})$
$\hat{p}_{ij}$	equivalent to $\hat{p}_{ij}(\mathbf{x})$
$\hat{p}_i(\mathbf{x})$	the estimate of $p_i(\mathbf{x})$
$\hat{p}_i$	equivalent to $\hat{p}_i(\mathbf{x})$
$\mathbb{R}$	the real space
$\mathbb{R}^d$	d-dimensional real space

$\mathbf{S}$	the matrix denoting the source signals corresponding to the mixture epoch $\mathbf{Z}$
s_i	the i^{th} source signal at time instance t (t omitted)
$\mathbf{s}_i$	the time series of the i^{th} source signal
^{T}	matrix transpose
t	the time instance
$\mathbf{v}$	the virtual scaling vector
$\mathbf{w}$	the normal to the hyperplane
$\mathbf{W}$	the inverse of mixing matrix $\mathbf{A}$
$\mathbf{x}$	the feature vector
$\mathbf{x}_i$	the i^{th} sample
y	the class label
y_i	the class label for the the i^{th} sample
$\mathbf{Z}$	the matrix denoting an epoch of EEG (or mixture signals)
z_i	the EEG signal (or mixture signals) recorded from the i^{th} channel at time instance t (t omitted)
$\mathbf{z}_i$	the EEG time series (or mixture time series) recorded from the i^{th} channel

Acronyms

ANN	artificial neural networks
AR	autoregressive
ARMA	autoregressive moving average
AWVT	auditory working-memory vigilance task
DSTA	Defence Science and Technology Agency
ECG	electrocardiograph
EEG	electroencephalogram
EMG	electromyogram
EOG	electrooculogram
ESS	Epworth Sleepiness Scale
fMRI	functional magnetic resonance imaging
FastICA	fixed-point ICA algorithm using gradient descent searching approach
FIR	finite-impulse-response
GMM	Gaussian mixture models
IC	independent component
ICA	independent-component-analysis
IIR	infinite-impulse-response
KNN	*k*-nearest neibor algorithm
KSS	Karolinska Sleepiness Scale
LDA	linear discriminant analysis
LDF	linear discriminant function

MEG	magnetoencephalogram
MRI	magnetic resonance imaging
MSLT	Multiple Sleep Latency Test
MWT	Maintenance of Wakefulness Test
NREM	non-rapid-eye-movement sleep
NTSB	National Transportation Safety Board
OVA-SVM	the “one-versus-all” SVM
OVO-SVM	the “one-versus-one” SVM
PCA	principal-component-analysis
PERCLOS	PERcentage CLOSure of eyelids
PSVM	probabilistic support vector machine
PVT	Psychomotor Vigilance Task
PWC-PSVM	the probabilistic SVM method using the pairwise coupling strategy
qEEG	quantitative electroencephalogram
REM	rapid-eye-movement sleep
RF	random forests
RHS	right hand side of equation
RP	Random Permutation
SFS	Situational Fatigue Scale
SSS	Stanford Sleepiness Scale
SVM	support vector machine
VAS	Visual Analogue Scale

Chapter 1

Introduction

Mental fatigue, defined by Grandjean (1980) as “a state of reduced mental alertness that impairs performance”, has become one of the most significant causes of accidents throughout modern society (see Dinges, 1995; Idogawa, 1991; Lal and Craig, 2001a; Mitler et al., 1988). In recent years, there have been increasing interests in electroencephalogram (EEG) based automatic mental-fatigue measurement and monitoring system (Artaud et al., 1994; Dinges and Mallis, 1998; Gevins et al., 1995; Lal et al., 2003), with the widespread hope that such system will become invaluable in the prevention of mental-fatigue related accidents.

This thesis is concerned with developing novel signal processing methods that enable automatically measuring and monitoring mental fatigue in human individuals from their EEG recordings. Various methods tackling the problems related to EEG signal processing, such as artifact removal, feature selection and multi-class pattern classification, are proposed and tested.

As an introduction, this chapter examines the role of mental fatigue in increasing the occurrences of various accidents throughout our modern society and provides an overview of the past related work on mental-fatigue detection using EEG (detailed literature review deferred to Chapter 2). The contributions of the current work are then outlined, followed by the organization of the thesis given at the end of this chapter.

1.1 Motivation

Typical symptoms of mental fatigue include decreased physiological arousal, slowed functioning of sensorimotor and impaired capability of information processing in the brain (Mascord and Heath, 1992). Such adverse physiological changes can seriously deteriorate operator's ability to respond effectively to emergency situations and numerous evidence has shown that mental fatigue has become one of the most significant causes of accidents throughout our society.

Mental fatigue is receiving increasing attention in the field of road safety. According to the early work by Idogawa (1991), mental fatigue accounts for 35% to 45% of all vehicle accidents on the road. A recent estimation (Stutts et al., 1999) made by the National Highway Traffic Safety Administration in the United States has also announced that, each year in United States alone, there are approximately 100,000 road accidents reported due to mental-fatigue related drowsy driving, claiming over 1,500 lives.

Another important area that calls for further research on mental fatigue is airline industry (both commercial and military). The National Transportation Safety Board (NTSB) in the United States cited pilot fatigue as either the cause or a contributing factor in 69 plane accidents from 1983 to 1986 (Kaplan, 1996; Stanford Sleep Disorders Clinic and Research Center, 1991). According to a recent report (Ryan and Heath, 2007), the NTSB has linked pilot fatigue to at least 10 commercial aviation accidents since 1993. While these reported accidents represent only a small percentage of the more than 40 million airline flights during the period, these crashes killed over 260 people .

Mental fatigue is critical not only in transportation industries, but also in other occupations, for instance, factory operators and health care professional, where sustained attention is required. The consequence of the potential incidents caused by mental fatigue in these occupations may not be fatal, but the accumulated costs for health care, lost productivity and damage to machinery and property can easily amount to billions of dollars.

Mental fatigue is believed to be a nonlinear, temporally dynamic, and complex process which results from various factors (Dinges, 1995). Typical factors causing mental fatigue include sleep restriction or deprivation and circadian rhythm (see Cajochen et al., 2004; Hartley et al., 1994; Pearson, 2004; Philip et al., 2005), irrelevant work schedules (see Åerstedt et al., 2000; Brictson, 1966; Horne and Reyner, 1995), length of journey and monotonous driving environment (see Horne and Reyner, 1995), and demanding delivery schedule (see Hartley et al., 1994).

Among other causes of mental fatigue, sleep deprivation and circadian rhythm are generally considered the most significant cause for the increasing occurrences of mental-fatigue related accidents. Nowadays, it is becoming increasingly common for us to stretch our limits to squeeze more time for work or for play. That extra time is usually taken by reducing the time period for which we sleep. This is true not only for students preparing for exams or office workers, but also for industrial workers, health care-professionals, drivers and pilots. Though it seems as an easy concession to make, but slowly and surely this lack of sleep catches up with us and makes ourselves prone to the impairment of mental fatigue. The sleep loss is a “sleep debt” that is cumulative. A modest loss of sleep on each single night may end up with a serious sleep debt over several nights. The more sleep debt we accumulate, the greater impairment does mental fatigue have. Moreover, the impairment due to mental fatigue can also be amplified by the bi-modal circadian rhythm. Some evidence of this can be seen by examining the temporal patterns of mental-fatigue related accidents. It has been documented (Miller,

2001) that there are two surges in the occurrences of mental-fatigue related accidents which match nicely with our circadian rhythm: one surge in the early morning and another surge in the mid afternoon

The nature of mental fatigue may also partly explain why there are increasing occurrences of mental-fatigue related accidents. Mental fatigue is *ubiquitous*, *pervasive* and *insidious* in nature (Miller, 2001). By *ubiquitous*, we mean that mental fatigue affects everybody. Although the individual difference does exist, we however often feel, without basis, that we are more resistant to mental fatigue than others. By *pervasive*, we mean that mental fatigue affects everything we do, physically, emotionally and cognitively. However, the impairment of mental fatigue is often under-estimated. By *insidious*, we mean that often when we are fatigued, we are quite unaware of how badly we are performing. In fact, several studies (Arnedt et al., 2001; Dawson and Reid, 1997; Lamond and Dawson, 1999) have provided strong basis of the equivalency of mental fatigue to alcohol in terms of impairment of our brain functioning. Moreover, we often do not recognize that we are too fatigued to be safe and may deny the impairment induced by mental fatigue, in the same manner as a drunk person does.

Another contributing factor to the increasing occurrences of mental-fatigue related accidents is the increasing level of automation (Okogbaa et al., 1994). Although automation has provided tremendous benefits, it also makes operators more susceptible to mental fatigue because automation significantly suppresses the stimulating influences by reducing the need of active operation.

If an automatic system could be developed to measure and monitor mental fatigue, a considerable number of accidents can be prevented and many lives could be saved. This is exactly the reason why mental fatigue tracking technology has been a perennial priority in the list of NTSB's “most wanted” safety improvements. In Singapore, Defence Science and Technology Agency (DSTA) is also greatly interested in a “mental-fatigue screening system". Specifically, this screening system is required to detect the extreme mental fatigue of the pilots and to raise the alarm, before their reaching a state in which they are incapable of fulfilling their cruise duties. The current doctorial research has been partly motivated by this local relevance.

To this end, abundant efforts have been devoted to develop an objective, non-intrusive and automatic mental-fatigue measurement and monitoring method. Some pilot studies have correlated mental fatigue with different physiological measures such as electrocardiograph (ECG), electrooculogram (EOG) and EEG. A good review of these methods can be found in the thesis by Mallis (1999) and a review by Lal and Craig (2001a). Among the numerous physiological indicators which have been linked to mental fatigue in the literature, EEG has been shown to be one of the most predictive and reliable techniques for detecting subtle changes in the brain due to mental fatigue (Artaud et al., 1994; Dinges and Mallis, 1998; Gevins et al., 1995; Horne and Reyner, 1995; Lal and Craig, 2001a; Lal et al., 2003; Lal and Craig, 2002; Makeig and Jung, 1995).

More recently, several studies have also reported the feasibility of measuring mental fatigue indexed by subject's task performance, based on EEG data in attention-sustained experiments using auditory or visual stimuli (Duta et al., 2004; Jones, 2006; Jung et al., 1997; Lal et al., 2003; Makeig et al., 2000; Peiris et al., 2004; Sommer et al., 2002; Vuckovic et al., 2002). Most of these pilot studies have focused on the detection of performance lapses in the specific tasks that they studied (i.e. prediction of a mistake in a specific task) without measuring subjects' mental-fatigue levels directly. Moreover, most of these pilot studies used fairly simple linear or nonlinear regression or neural networks, and the recent advance in the signal processing methods, like automatic artifact removal, feature selection and multi-category pattern classification, have been overlooked. More importantly, very little evidence exists on the efficacy of incorporating EEG into a practically-usable automatic mental-fatigue measurement and monitoring system, and the literature continues to produce varying and even conflicting results. This is likely due to the challenge of developing effective mathematical framework, signal processing methods and learning algorithms for the analysis of EEG signals in relationship to mental fatigue.

To measure and monitor mental fatigue in (near) real-time fashion, at least three challenges remain in developing or adapting powerful signal processing methods (running on fast enough computer or processing chip which were not available before) to extract the relevant information from the EEG.

First, the technical challenge of automatic removal of the pervasive EEG artifacts has rarely been addressed. These EEG artifacts typically have much higher amplitude than cerebral signals and thus impose great difficulties in EEG interpretation. This, coupled with the fact that mental fatigue produces much less distinguishable changes in terms of EEG waveforms than other brain states like sleep (Kecklund and Åerstedt, 1993), makes it imperative to have an effective automatic EEG artifact removal module in a workable EEG-based mental fatigue monitoring system.

Second, it remains unclear what EEG features are important for measuring and monitoring mental fatigue. Past studies have computed features on one or more spectral bands from a priori defined one or more EEG channels, rather than computing full-spectrum of each of the EEG channel in full mapping EEG recordings; Features that have been selected to relate to mental fatigue were often limited to powers of some specific standard frequency bands (often without giving the justification), rather than considering combination of multiple types of features; Moreover, the recent advance in feature selection in the domain of machine learning have been largely overlooked, despite the apparent multi-fold benefits of adapting such data mining technique: when the underlying important EEG features are known and irrelevant / redundant EEG features are removed, the learning problem can be greatly simplified, resulting in improved accuracy and enhanced system interpretability.

Third, a comprehensive pattern recognition system is required for continuous measuring and monitoring mental fatigue using EEG. It is not only complicated but also challenging to predict the subject's mental-fatigue level given an EEG epoch of few seconds.

1.2 Objectives

This thesis is concerned with developing novel signal processing methods that enable automatically measuring and monitoring mental fatigue in human individuals from their EEG recordings.

The approach taken in this work is to first identify important features in the EEG signals that correlate with mental fatigue in an individual from an collected mental-fatigue EEG dataset. Then, these key features are used to construct an intelligent system that tracks the state of mental fatigue of an individual.

1.3 Organization of the Thesis

Chapter 1 serves as an introduction. It examines the role of mental fatigue in the increasing occurrences of various accidents throughout our modern society and provides an overview of the past related work on mental-fatigue detection, followed by the description of the objectives of the present work.

Chapter 2 provides the relevant background information on EEG, standard EEG signal processing methods, and the detailed review of the past related work on EEG-based mental fatigue monitoring. Some formulations of the relevant signal processing methods from the literature needed for subsequent chapters are also given in the chapter.

Chapter 3 gives an overview of the approach taken in this work and the detailed description of the collection and labeling of mental fatigue EEG used in the present work.

Chapter 4 is devoted to the proposed automatic artifact removal method and the report of its performance in comparison with some existing methods in the literature.

Chapter 5 and **Chapter 6** describe the proposed new feature-selection methods and the related numerical experiments. For the ease of presentation, feature selection methods for two-class classification are first discussed in **Chapter 5**, followed by its non-trivial extension to multi-class feature-selection methods described in **Chapter 6**. Although the proposed feature-selection methods are proposed for EEG signal processing, they in fact represent novel approaches that are generally useful in the domain of machine learning.

Chapter 7 gives the details of our method for automatic classification of multi-level mental fatigue EEG using a probabilistic multi-class support vector machine (SVM). This chapter also presents the prototype of the developed automatic mental-fatigue measurement and monitoring system. The comprehensive performance evaluation of such a system is also reported.

It is worth noting that, in organizing the thesis, **Chapter 4** to **Chapter 7** are presented to be as self-contained as possible because each of these chapters deals with different aspects of EEG signal processing. Accordingly, each method presented in **Chapter 4** to **Chapter 7** is also tested separately on well-known publicly-available benchmark datasets whenever possible. The performance evaluation of those methods using mental-fatigue EEG is deferred to **Chapter 7**. An additional benefit of doing so is that the validity of the proposed signal processing methods can be evaluated broadly in the domain of machine learning before they are used in the specific application for mental-fatigue measurement and monitoring.

Chapter 8 concludes the thesis with a discussion on the significance of current research, its limitations and future directions.

Chapter 2

Literature Review

This chapter serves to familiarize the readers with the relevant background information on EEG, such as EEG electrode placement, montage (an EEG jargon for differential referencing), commonly-referenced standard EEG frequencies and their use in the study of sleep patterns. This chapter also gives a detailed literature review on the past work pertaining to the detection of mental fatigue using EEG, followed by a review of the EEG signal processing methods with an emphasis on those needed for the subsequent chapters.

2.1 EEG: Physiological Basis

The electroencephalogram is a recording of electrical activities in the brain as recorded from electrodes placed on the scalp. The first EEG recordings on human were made by Berger (1929), although similar measurements on animals had been carried out as early as 1875 by Caton (1875). Soon after the invention of EEG, it has been one of the major tools to investigate brain functionality.

The EEG measures mainly summated potential field generated by post-synaptic currents (Speckmann and Elger, 1999). The synapse, a tiny interface between the terminal bouton of a neuron and the membrane of another neuron or non-neuronal cell (such as glandular cell), is the site where one neuron communicates with another cell. The number of synapses in the human brain is about 10^{15} to 5×10^{15} (1-5 quadrillion). They allow neurons to form interconnected circuits within the central nervous system and thus are crucial to all cognitive functions of our brain. They are also the major source of the EEG signals. An action potential in a pre-synaptic axon causes the release of neurotransmitter into the synapse. The neurotransmitter diffuses across the synaptic cleft and binds to receptor in a post-synaptic dendrite, triggering a flow of ions into or out of the dendrite. This results in compensatory currents in the extracellular space. It is these extracellular currents that are responsible for the generation of the EEG signals.

It is generally believed that it is not possible to measure the potential field generated by a single post-synaptic activation using the scalp EEG. Rather, the scalp EEG represents the summation of the synchronous activities of thousands of neurons that have similar spatial orientation. The synchronous activation of such neuron cluster is commonly modeled by using a dipole source activation. The relationship between the EEG and a dipole source activation can be illustrated by Fig. 2.1. This schematic drawing treats the brain as a volume conductor that is roughly spherical. As shown in the figure, what the EEG measures is the potential difference between two locations on the scalp.

Besides the electrical field, the dipole source activation also generates a magnetic field as shown in Fig. 2.1. This magnetic field can also be measured and the resulting measurement is called magnetoencephalogram (MEG). Basically, EEG and MEG are just different manifestations of brain activities, but MEG has some remarkable advantages over EEG. For example, the skull insulation distorts the EEG but it is transparent to magnetic fields that the MEG measures. However, the MEG generated by the brain is very weak (50 to 100 femto-teslas, about one-billionth the strength of the Earth's mag-

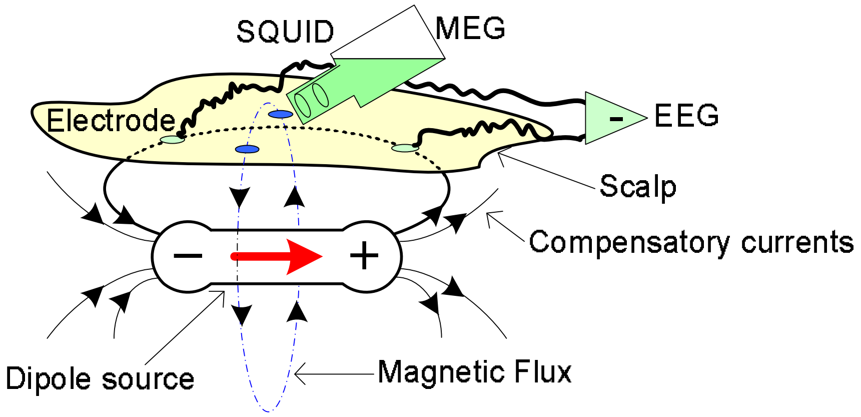


Figure 2.1: Schematic drawing of the bio-electrical field and bio-magnetic field generated by a dipole source activation

netic field) and is easily overwhelmed by environmental magnetic noises. Therefore, the measurement of MEG has to be carried out in a magnetically shielded room using sophisticated equipment called Super-conducting QUantum Interference Device (SQUID) and thus is inappropriate at present for non-clinical use.

2.2 EEG: Technological Basis

In Berger's time (Berger, 1929), the EEG recording systems were very cumbersome and could only be used in research laboratory or in a hospital. With the recent development of electronics, there are more portable and powerful mini-systems for EEG recording. This section provides the technological basis of EEG recording. Both hardware aspects (electrode and filtering) and procedural aspects (the standard electrode placement, the setting of differential referencing) are discussed in this section.

2.2.1 Electrode

The electrical contact between the input of the EEG recording system and the brain from which the electrical signals originate is made by means of electrodes. Various types of EEG electrodes can be found in (Spehlmann, 1981). Currently, the most commonly used electrodes for scalp EEG are surface electrodes that are affixed to the skin with conductive jelly. Indirect contact is established by an electrolyte bridge formed by the conductive jelly applied between the electrode and the skin (see Kamp and da Silva, 1999, page 110).

It is worth noting that the recent development of dry EEG electrodes (Fonseca et al., 2007; Griss et al., 2002; Taheri et al., 1994) equipped with the wireless transmission technology may largely benefit the use of EEG beyond clinics in the near future, for example, the use of EEG for mental fatigue measurement and monitoring in working environment.

2.2.2 The International 10-20 System

The international 10-20 system of electrode placement (Jasper, 1958) has become the standard electrode placement method in the context of EEG measurement. It ensures accurate placement of electrodes on same subject in repeated measurements and allows comparison of the EEG signals between subjects.

As shown in 2.2, two bony landmarks are used for the essential positioning of the EEG electrodes: first, the nasion which is the point between the forehead and the nose; second, the inion which is the lowest point of the skull from the back of the head and is normally indicated by a prominent bump. The “10” and “20” in the name of the international 10-20 system refer to the fact that the surface distances between adjacent

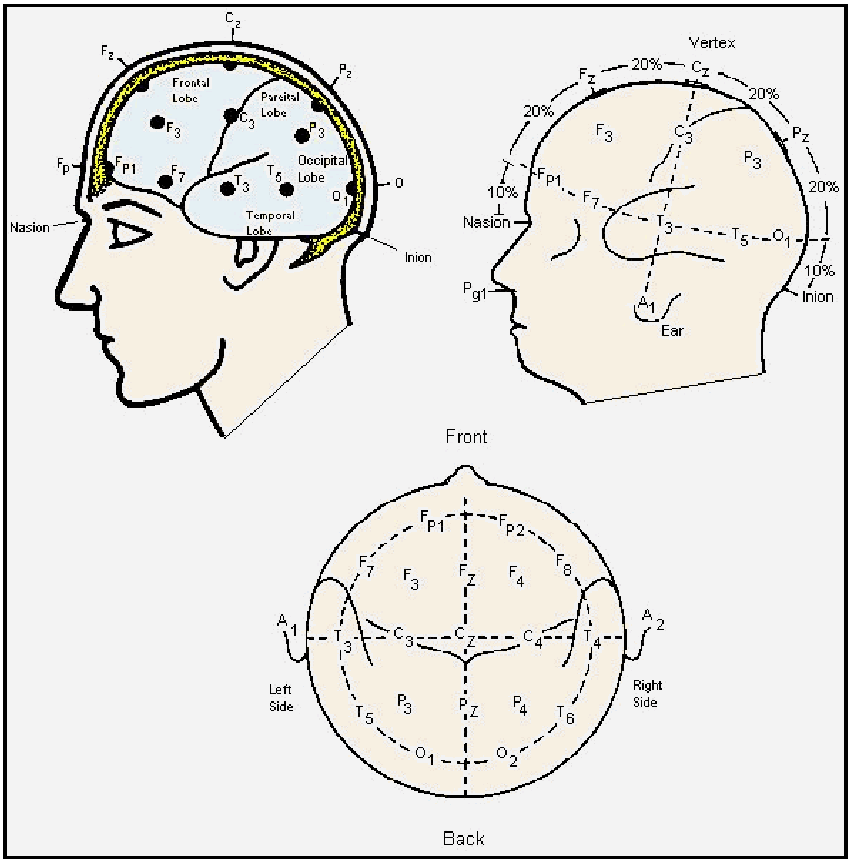


Figure 2.2: The international 10-20 system of electrode placement (Aguiar et al., 2000)

electrodes are either 10% or 20% of the total front-back or right-left surface distance of the skull.

Each site has a letter to identify the underlying brain functional lobe and a number to identify the hemisphere location. The letters F, T, C, P and O stand for Frontal, Temporal, Central, Parietal and Occipital respectively. Note that there is no central lobe in brain anatomy, the "C" letter is used for identification purposes only. Even numbers (2,4,6,8) refer to electrode positions on the right hemisphere, whereas odd numbers (1,3,5,7) refer to those on the left hemisphere. For electrodes on the midline between left and right hemisphere, a "z" letter is used in place of a number.

The international 10-20 system involves 21 electrodes, but it can be modified to accommodate extra electrodes when necessary. For example, in the modified combinatorial nomenclature (MCN) system used for 32-channel EEG recording, extra electrodes are placed in-between the existing 10-20 system. However, the naming system used by MCN is more complicated and the new letters introduced to name the extra electrodes do not necessarily refer to the underlying cerebral cortex.

2.2.3 Montage

Since a reading of EEG, as shown in 2.1, represents a voltage difference between two electrodes or two locations, the display of the EEG may be set up in one of following ways, depending on the choice of differential referencing. Such differential referencing method for displaying EEG is termed a montage.

Bipolar montage: Each channel (i.e., waveform) represents the voltage difference between two adjacent electrodes. The entire montage consists of a series of such pairs of electrodes and it typically includes chains running anteroposteriorly or transversely, using the same linkage over both hemispheres. For example, in the commonly-referenced “double banana” montage, the channel “Fp1-F3” represents the voltage difference between the Fp1 electrode and the F3 electrode. Next, the channel “F3-C3” represents the voltage difference between F3 and C3, and so on through the entire array of electrodes anteroposteriorly.

Referential montage or unipolar montage: Each channel represents the voltage difference between an active electrode and an designated “inactive” one, known as the reference. Ideally, the reference should be completely silent, having a zero potential. In practice, however, all locations on the scalp are active to some degree. Therefore, the choice of reference electrode is mainly determined by the available domain knowledge.

For example, midline positions (the middle of Fz and Cz or Cz and Pz) are often used as the reference because they do not amplify the signals from one hemisphere vs. the other. In the literature, such reference is called cephalic reference since reference electrode is put on scalp.

More often, a non-cephalic reference (reference electrode near clavicle) is used. It is hard to say whether a non-cephalic reference is superior to a cephalic reference. On the one hand, a non-cephalic reference can be used to address the problem of cerebral contamination caused by an cephalic reference. On the other hand, a non-cephalic reference is subject to the contamination of electrocardiograph (ECG) artifact and measures must be taken to remove the resulting large amplitude ECG artifact introduced. Nevertheless, a non-cephalic reference becomes more popular among EEG signal processing community with merging techniques for minimizing ECG artifact.

Average reference montage: The outputs of all of the amplifiers are summed and averaged, and this averaged signal is used as the common reference for each channel.

Laplacian montage: Each channel represents the voltage difference between an electrode and a weighted average of the surrounding electrodes.

When analog (paper) EEG are used, the EEG technician switches between montages during the recording in order to highlight or better characterize certain features of the EEG. With digital EEG, handling of montage becomes much easier. Typical, all EEG are digitized and stored in unipolar montage. This is simply because any other montage, if it is desired, can be constructed mathematically from the stored EEG.

2.2.4 Filtering

In theory, the greater the recorded frequency band, the greater the fidelity of reproduction of the actual cerebral activity. In practice, however, recording a larger frequency band increases the amount of outside interference or noise in the EEG signals. Filters are used to make a compromise between reduction of extraneous noise and preservation of cerebral signals (see Reilly, 1999, page 132).

Nowadays, the EEG is usually sampled at a frequency of about 256 Hz, which is more than sufficient to cover the most commonly-referenced frequency bands (Niedermeyer, 1999) as shown in Table 2.1. Correspondingly, a routine EEG system typically comes with an integrated low-pass filter (cut-off frequency at about 35 Hz) and a high-pass filter (cut-off frequency about 0.1 Hz) as well as a 50 Hz or 60 Hz notch filter depending on the frequency of local power system.

Table 2.1: Standard EEG frequency bands

Frequency Band	Range
Delta	0.5–4 Hz
Theta	4–8 Hz
Alpha	8–13 Hz
Beta	13–30 Hz

It is worth noting that individual work in the literature may use slightly different lower and upper limit for each frequency band than those in Table 2.1. Moreover, higher frequencies are also considered in the literature. For example, “gamma band” was used to designate frequencies above 30 Hz as early as 1938 (Jasper and Andrews, 1938). This term was then abandoned and “gamma” frequencies became a part of “beta” frequencies. However, the use of the term “gamma band” has made an impressive comeback during the 1990s (Başar, 1992; Bullock, 1992; Eckhorn et al., 1992; Gray et al., 1992; Kaplan, 1996). The “gamma” frequencies are conceived mainly as induced rather than as spontaneous rhythms and they are therefore usually not included in the list of standard frequency bands for spontaneous EEG (Niedermeyer, 1999).

2.3 EEG: Characteristics

The scalp EEG is the secondary measure of brain functions. As discussed in Section 2.1, the scalp EEG is a presentation of synchronized post-synaptic activations. A considerable change in EEG indicates that there is some brain activity occurring in terms of millions of cells acting together, in a synchronized fashion. In this sense, the measured scalp EEG should be thought as “epiphenomena”, which is the manifestation or byproduct of brain functions. The brain does not communicate or perform a function using the EEG. Rather, it is a secondary measure, much like the vibration from an working engine or the temperature of an active circuit. Understanding this characteristic of EEG is important because it defines what the EEG can tell and what it can not. For example, the brain does not, for example, produce alpha waves for any purpose. The existence of the alpha waves is simply a result of certain brain function or brain activity. Alpha waves can however be utilized for the investigator’s advantage, by investigating what they represent and what they imply when they are changed (such as increase / decrease in their amplitudes or shift in their frequencies).

Another characteristic of the scalp EEG is complexity. The EEG complexity originates in the intricate neural system in the brain. Moreover, both internal and external noise factors also largely increase the complexity in the interpretation of EEG. On the one hand, EEG is subject to many modifiers including brain anatomy (for example, the skull insulation distorts the EEG signals), neuron alignment and even metabolism in the brain. On the other hand, it is nearly always contaminated by other non-cerebral signals called artifacts. The most common types of artifacts include EOG artifacts, ECG artifacts and electromyogram (EMG) artifacts. In addition to internal artifacts, there are other noises which originate from outside of the subject, for example, the power line noise of 50 or 60 Hz, depending on the frequency of local power system. Poor contact of EEG electrode to scalp may also distort the EEG signals due the momentary change in the impedance.

From a signal processing point of view, EEG has the following characteristics: (i) EEG is noisy. It is often contaminated by EOG, ECG and EMG artifacts and thus effective artifact removal method is needed in order to improve the reliability of EEG interpretation. (ii) EEG is nonstationary. It varies with physiological and psychological states of the brain. In practice, the EEG is often treated as a stationary process over a relatively short duration (about 3 seconds). (iii) EEG is nonlinear. Although the traditional linear methods show to be very useful in EEG analysis, the EEG is a highly nonlinear process.

2.4 EEG: A Major Tool to Study Brain

EEG has been one of the major tools to investigate brain functionality since Berger (1929). In fact, before the brain-imaging techniques, such as computerized tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) and, more recently, functional magnetic resonance imaging (fMRI), EEG was the main, if not the only, tool for study of the brain. The rest of this section gives the reader the flavour of the diversity of EEG applications.

(a) **Study of physiological or psychological brain states:** EEG has been used in study of physiological and psychological brain states since Berger (1929) as documented by gloor (1969). The fascinating aspect of Berger's work is that many of the ideas that he proposed are still relevant today (see Shaw, 2003, page 9). Following Berger, many research efforts have been devoted to the use of EEG in the study of various physiological and psychological brain states, such as sleep (e.g. Anderer et al., 1999; Erwin et al., 1984; Penzel and Conradt, 2000; Rechtschaffen and Kales, 1968), arousal (e.g. Bonnet and Arand, 2001; Kok and Zeef, 1991), fatigue (e.g. Artaud et al., 1994; Dinges and Mallis, 1998; Gevins et al., 1995; Lal et al., 2003), attention (e.g. Arruda et al., 2007; Dockree et al., 2007; Oken et al., 2006; White et al., 2005), anxiety (e.g. Gordeev, 2007; Hogan et al., 2007; Lee et al., 1997; Schiff et al., 1949; Shagass, 1955; Warbrick et al., 2006; Weinstein, 1995), anesthesia (e.g. Davidson, 2006; Esmaeili et al., 2007; Feinberg and Campbell, 1997; Herregods et al., 1989; Jospin et al., 2007; Koskinen et al., 2005; Maksimow et al., 2006; McEwen et al., 1975; Mi et al., 2003; Modena et al., 1969; Suttmann et al., 1989; Zhang et al., 2001) and pain (e.g. Bromm et al., 1992; De Benedittis and De Gonda, 1985; Diers et al., 2007; Dowman et al., 2008; Gucer et al., 1978; Huber et al., 2006; Le Pera et al., 2000; Lutzenberger et al., 1997; Sarnthein et al., 2006).

(b) **Study of neural diseases:** EEG has also been shown useful in study of various neural diseases, such as epilepsy (e.g. Barkley and Baumgartner, 2003; Binnie et al., 1981; Collura et al., 1990; Ebersole, 1991; Foldvary et al., 2001; Gigli and Valente, 2000; Goodin et al., 1990; Kershman et al., 1951; Kuhl and Lund, 1967; Legg et al., 1973; Matsuoka et al., 2000; Narayanan et al., 2008; Wray and Hablitz, 1978), brain tumor (e.g. Bassett et al., 1967; Deboer et al., 2002; Kubota et al., 2001; Murphy, 1957; Ochi and Sakata, 1955; Silverman et al., 1961), Parkinson's disease (e.g. Ban and Hojo, 1971; Delval et al., 2006; Kuhn et al., 2005; Lalo et al., 2008; Novak et al., 2007; Vardi et al., 1978; Visser and Postma, 1971), ADD/ADHA (e.g. Alexander et al., 2008; Becker et al., 2004; Diamond, 1997; Kuperman et al., 1996; Laporte et al., 2002; Murias et al., 2007; Snyder et al., 2008; Trudeau et al., 1999), depression (e.g. Fingelkurts et al., 2006; Hongo et al., 1963; Kerkhofs et al., 1988; Kupfer et al., 1976; Li et al., 2008; Roschke et al., 1994), Alzheimer's disease (e.g. Besthorn et al., 1994; Ehle and Johnson, 1977; Ihl et al., 1996; Jelles et al., 1999; Kowalski et al., 2001; Nobili et al., 1999;

Ponomareva et al., 2008; Soininen et al., 1988; Strik et al., 1997), Wilson's disease (e.g. Chu et al., 1991; Giagheddu et al., 2001; Hansotia et al., 1969; Nevsimalova et al., 1986).

(c) **Study of neural injuries:** Study of neural injuries using EEG is another important area of research. The relationships between EEG and stroke (e.g. Finnigan et al., 2006, 2008; Platz et al., 2000; van Putten and Tavy, 2004; Vock et al., 2002; Wood et al., 1984), trauma (e.g. Goransson et al., 1988; Khanna et al., 1991; Naquet et al., 1968; Tezer et al., 2004; Thatcher et al., 1989) and coma (e.g. Brenner, 2005; Calhoun and Ettinger, 1966; Fenwick et al., 1969; Kassab et al., 2007; Young, 2000) have been extensively studied.

(d) **Study of brain conditioning:** The typical example of the use of EEG in brain conditioning is biofeedback techniques (Duffy, 2000; Miller, 1969a,b; Thatcher, 2000). The biofeedback techniques have been shown to be able to reinforce, or to reduce, any rhythms or combination of rhythms, or for more complex configurations such as training different brain locations to be synchronized, or desynchronized. Impressive efforts have been made to correlate such different kinds of training with the behavioral or cognitive enhancement in subjects. However, many critiques opposing such approach have also been documented in the literature (Shaw, 2003; Steiner and Dince, 1981).

(e) **The EEG-MRI Combo:** It is well-known that EEG has high temporal resolution and low spatial resolution. Recently, it has became popular to combine EEG with high spatial-resolution brain-imaging techniques, such as MRI, providing a power tool with both high temporal and spatial resolutions (Alper, 1993; Mirsattari et al., 2004).

2.5 EEG and Sleep

Among all the afore-mentioned EEG applications, the use of EEG in sleep study is probably most influencing. The EEG methods used in the sleep study, involving the definition, naming, quantification of typical frequency bands as shown in Table 2.1, are still widely followed in many EEG research areas, including in the study of mental fatigue. Moreover, sleep deprivation or sleep loss is the most significant cause of mental fatigue and there are also some similarities between sleep and mental fatigue (the discussion of this is deferred to the later chapters). Therefore, this section is devoted to a brief discussion of the relationship between EEG and sleep.

The average adult needs about 8 hours of sleep per night, but sleep need is approximately normally distributed. That means, few people may need as little as 6 hours, while other few may need as much as 10 hours. On an individual basis, the amount of sleep that an individual requires is the amount necessary to achieve full alertness and effortless functioning during the waking hours, even when sitting quietly and being bored. When an individual feels that they have to keep moving in order to stay alert, that is a strong sign of too little sleep. One may assess the relative severity of this problem by using a reliable subjective rating scale called the Epworth Sleepiness Scale or Stanford Sleepiness Scale.

Sleep is a complex and active physiological process. Sleep progresses in a cyclic fashion between two types of sleep: non-rapid-eye-movement sleep (NREM) and rapid-eye-movement sleep (REM). Within NREM sleep, sleep are often further classified into four stages: stage 1, stage 2, stage 3 and stage 4. The cycling of sleep, or commonly known as “sleep architecture”, can be easily monitored by EEG. Fig. 2.3 shows the various sleep stages, with EEG signals shown on the left and the corresponding schematic of sleep stages as a function of time of night shown on the right.

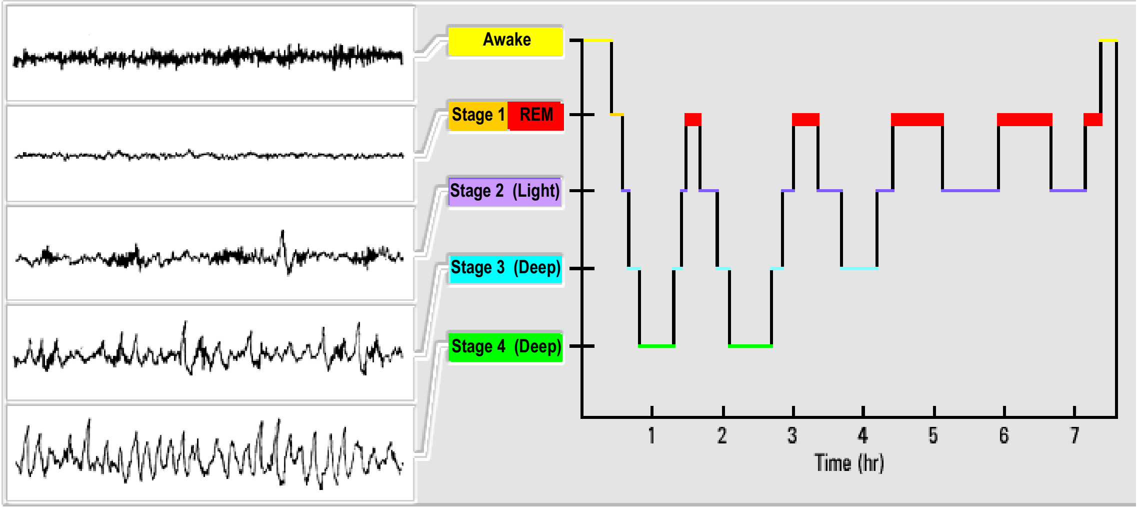


Figure 2.3: Brain electrical activity (on left) illustrates the stages of sleep(on right). Note that sleep progresses in a cyclic fashion through the sleep period. Morning awakening often occurs from the stage REM. (McCallum et al., 2003)

As shown in Fig. 2.3, the five stages of sleep, including their repetition, occur cyclically. The first cycle, which ends after the completion of the first REM stage, usually lasts for 100 minutes. Each subsequent cycle lasts longer, as its respective REM stage extends. So a person may complete five cycles in a typical night's sleep.

As an individual drifts off to sleep at night, he/she enters Stage 1. This is followed by a slowing of the heart rate, decrease of body temperature and relaxing of muscle tension as Stage 2 is entered. Stages 3 and 4 are known as slow-wave or delta sleep (because the energy is dominated in the delta band) and these slow-wave brain activities are known to be associated with very deep and restorative levels of sleep. During these stages, it is particularly difficult to wake the individual. REM sleep occurs periodically throughout, but the longer periods of REM sleep during normal, nocturnal sleep are most likely to occur during the pre-dawn hours. REM sleep shows a brain electrical pattern similar to Stage 1 or waking. Most dreaming occurs during REM sleep, and many normal morning awakenings occur from REM sleep. With respect to an individual's principal sleep period for each 24-hour period, it is important that the entire cyclic process of sleep be completed to receive the restful effects of a sleep period. Anything that interferes with sleep, such as noise disruptions, medication, alcohol, or simply insufficient duration, will change the physiological structure of the sleep cycles and impair alertness the next day.

The use of EEG for sleep scoring, i.e. classifying sleep stages, shows one of the most prominent approaches to EEG interpretation. Such an approach is to identify patterns that associate with specific physiological or pathological brain states and it is involved in a great deal of EEG history.

2.6 Mental-Fatigue Basics

Everyone knows how it feels to get too little sleep. Many individuals refer to this feeling as “mental fatigue” or “sleepiness”. Sleep loss is the primary causes of mental fatigue and humans have specific physiological, psychological and environmental requirements for getting adequate sleep (McCallum et al., 2003). The information in this section talks about the basis of mental fatigue which includes the definition, effects and physiological measurement methods of mental fatigue. The emphasis is put on the mental fatigue induced by sleep deprivation because it is the most important cause of mental-fatigue related accidents.

2.6.1 Mental Fatigue: Definition

“Mental fatigue”, is easier felt, than defined. There is no common definition accepted by the scientific community. The most basic definition of mental fatigue can be feeling tired, sleepy or exhausted, while a more complex definition can be a state of an central nervous system, in which prior physical activity and/or mental processing and/or wastefulness, in the absence of sufficient rest, result in insufficient cellular capacity or system-wide energy to maintain the original level of alertness and/or processing by using normal resources (Staal, 2004). Many definitions of mental fatigue in the literature focus on the functional manifestations of mental fatigue. For example, Grandjean (1980) defined the mental fatigue as “a state of reduced mental alertness that impairs performance”. This is simply due to the fact that only very limited knowledge about the physiological mechanism of mental fatigue is available while the mental-fatigue related impairment to the brain functions has become one of the most significant causes of accidents throughout the modern society (see Dinges, 1995; Idogawa, 1991; Lal and Craig, 2001a).

Before we continue to discuss mental fatigue, it is important to clarify the terms denoting mental fatigue. Human fatigue can be divided into two categories: physical fatigue and mental fatigue. Physical fatigue refers to the reduction of performance of the muscular system, while mental fatigue is in general related to the brain or central nervous system in a state of reduced mental alertness and with noticeable functional impairment. With the increasing number of knowledge/information-based occupations and the rapid development of machinery/automation, the complex phenomenon of mental fatigue becomes more and more important while physical fatigue is decreasing. The present work concerns with mental fatigue only. Hence, in the rest of the thesis, “fatigue” may also be used to refer without ambiguity to mental fatigue.

It is also not uncommon that the terms mental fatigue, sleepiness, drowsiness and to a less extend, alertness/vigilance were used interchangeably in the literature (Broadbent, 1979; Dinges, 1995; Grandjean, 1979, 1980; Lal and Craig, 2001b; Torsvall and Åkerstedt, 1987). Mental fatigue and sleepiness were often used synonymously to refer to mental fatigue resulting from the neurobiological processes that regulate the circadian rhythm and the need to sleep (Dinges, 1995). Although the term sleepiness has a more precise definition than mental fatigue (hence the latter is not preferred by many sleep specialists.), the term mental fatigue is widely used to indicate the influence of long working periods, sleep debt, circadian rhythm, and being unable to sustain a certain level of vigilance or task performance (Dinges, 1995; Lal and Craig, 2002). The present work concerns mainly on the mental fatigue due to sleep deprivation and circadian rhythm which is the most common cause of fatigue-related accidents. Therefore, though the popular term mental fatigue is used throughout the thesis, the terms mental fatigue and sleepiness can actually be used interchangeably in this thesis without ambiguity. The terms alertness and vigilance were also used to denote the phenomenon of mental fatigue in some sustained attention studies where the effect of mental fatigue on attention capability was emphasized, but alertness and vigilance differ significantly by definition from mental fatigue.

Mental fatigue is believed to be a nonlinear, temporally dynamic, and complex process which results from various factors (Dinges, 1995). Typical factors causing mental fatigue include sleep restriction or deprivation and circadian rhythm (see Cajochen et al., 2004; Hartley et al., 1994; Pearson, 2004; Philip et al., 2005), irrelevant work schedules (see Åerstedt et al., 2000; Brictson, 1966; Horne and Reyner, 1995), length of journey and monotonous driving environment (see Horne and Reyner, 1995), and demanding delivery schedule (see Hartley et al., 1994). Other secondary factors may also contribute to mental fatigue psychologically, such as mood, motivation, noise, heat (Rhodes and Gil, 2002; Rogers et al., 2003).

Among all causes of mental fatigue, sleep deprivation and circadian rhythm are generally considered the most significant cause for increasing occurrences of mental-fatigue related accidents. Nowadays, it is becoming increasingly common for us to stretch our limits to squeeze more time for work or for play. That extra time is usually taken by reducing the time period for which we sleep. This is true not only for students preparing for exams or office workers, but also for industrial workers, health care-professionals, drivers and pilots. Though it seems as an easy concession to make, but slowly and surely this lack of sleep catches up with us and makes ourselves prone to the impairment of mental fatigue. The sleep loss is a “sleep debt” that is cumulative. A modest loss of sleep on each single night may end up with a serious sleep debt over several nights. The more sleep debt we accumulate, the greater impairment does mental fatigue have. Moreover, the impairment due to mental fatigue can also be amplified by the bi-modal circadian rhythm. Some evidence of this can be seen by examining the temporal patterns of mental-fatigue related accidents. It has been documented (see Miller, 2001) that there are two surges in the occurrences of mental-fatigue related accidents which match nicely with our circadian rhythm: one surge in the early morning and another surge in the mid afternoon

As discussed in Section 1.1, the nature of mental fatigue may also partly explain why there are increasing occurrences of mental-fatigue related accidents. Mental fatigue is *ubiquitous*, *pervasive* and *insidious* in nature (see Miller, 2001) and we often do not recognize that we are too fatigued to be safe. Another contributing factor to the increasing occurrences of mental-fatigue related accidents is the increasing level of automation (see Okogbaa et al., 1994). Although automation has provided tremendous benefits, it also makes operators more susceptible to mental fatigue because automation significantly suppresses the stimulating influences by reducing the need of active operation.

2.6.2 Mental Fatigue: Effects

Mental fatigue has a variety of effects on the functionality of the brain. The most extensively-studied effect is that on sustained attention or vigilance. It has been shown that the reaction time in a vigilance task like the Psychomotor Vigilance Task (PVT), is directly proportional to mental fatigue (Dinges and Powell, 1985; Pack et al., 2006; Rogers et al., 2003). Other faculties like working memory, judgment and decision making, and mood are also affected as mental fatigue progresses (Cajochen et al., 2004; Staal, 2004). It has even been shown that after a certain level of sleep deprivation, the performance deterioration is similar to that caused by alcohol (Arnedt et al., 2001; Dawson and Reid, 1997; Lamond and Dawson, 1999): a striking fact showing the danger of mental fatigue.

According to a report on a study of Canadian Marine pilots, the tasks rated by pilots to be most affected by fatigue were decision-making, attention, remaining awake and reaction time (Rhodes and Gil, 2002). In addition the study also pointed out that mental fatigue led to decreased performance on memory tasks. Other studies too have pointed out these same deficits. Ferguson et al. (2005) also asserted that slowed reaction time, impaired decision making, memory difficulties and vigilance decrements are primary impairments caused by mental fatigue.

Various studies have also shown that performance of working memory deteriorates due to sleep deprivation (Caldwell et al., 2004; Chee and Choo, 2004; Ferguson et al., 2005; Murphy and Delanty, 2007; Smith et al., 2002). Similarly, executive functioning and decision-making are also degraded due to mental fatigue (Bruck and Pisani, 1999; Killgore et al., 2006; Neri et al., 1992; Nilsson et al., 2005; Raaijmakers, 1990).

In summary, below are the main cognitive impairments for mental fatigue due to sleep deprivation:

(a) Impaired alertness or sustained attention or vigilance,

(b) Memory difficulties,

(c) Poor decision making,

(d) Reasoning abilities become slower,

(e) Language and verbal skills are compromised,

(f) Mathematical skill deteriorate,

(g) Psychological effects like low mood, loss of motivation to continue working,

(h) Microsleeps — few moments of falling asleep while doing a task, followed by regaining the conscious control of the task. These few moments can be most dangerous specially for driver, pilots and industrial workers.

(i) Attention tunneling — as mental fatigue progresses, the ability of the person to analyze a large amount of factors while making a decision deteriorates, making him/her focus on one or two factors which seem important to him/her. This can lead to dangerous results in situations where a person is required to assess data from different sources for making a right decision - like pilots, air traffic controllers and military commanders.

2.6.3 Mental Fatigue: Measurements

Indicators of mental fatigue can be generally categorized into four types: (a) subjective feelings of tiredness, sleepiness, loss of motivation, low mood, impatience, frustration, confusion, (b) performance decrements on cognitive or psychomotor tasks, such as sustained-attention tasks, working-memory tasks, decision-making tasks, mathematical tasks, verbal and language tasks, (c) behavioral changes - being lethargic or irritable, episodes of microsleeps, (d) changes in EEG, EMG and other physiological measures.

Accordingly, the measurement methods for mental fatigue can also be broadly divided into four categories: (a) subjective self-report measures, (b) objective performance measures, (c) behavioral measures, (d) physiological measures.

This subsection is devoted to all the above-mentioned four types of mental-fatigue measurement methods, except that EEG-based methods, as the closely-related past work to this doctoral study, will be discussed in the next section. It is also worth noting that some of these non-EEG mental-fatigue measurement methods are often used as the benchmark methods in collecting mental-fatigue EEG database used for EEG-based mental-fatigue measurement methods. This will be clearer in the discussion of EEG-based mental-fatigue measurement methods in the next section.

2.6.3.1 Subjective Self-Report Measures

The simplest measure of mental fatigue can be a subjective self-report measure, such as Visual Analogue Scale (VAS), Stanford Sleepiness Scale (SSS), Epworth Sleepiness Scale (ESS), Karolinska Sleepiness Scale (KSS) and, more recently, Situational Fatigue Scale (SFS) (Yang and Wu, 2005). Subjective self-report mental-fatigue measurement methods require subjects to rate their level of mental fatigue either indirectly (see Piper et al., 1998; Zachrisson et al., 2002) or directly (see Schapire, 1992). In these methods, the subjects are requested to rate their current states about their own assessments of mental fatigue, typically through a questionnaire. The level of mental fatigue is estimated by scoring their responses on the questionnaire.

Some of these scales like SSS, VAS and KSS are quite simple, involving only a few questions about how the subjects are feeling at the moment. While other scales like ESS and SFS are very detailed and require the subjects to estimate their level of mental fatigue if they were in specific situations. The SFS has very detailed scenarios in which the subjects have to imagine and they have to estimate how fatigued they would be if they were in those scenarios, like watching TV for 2 hours, jogging for 20 minutes, or reading for 1 hour, etc. Though these latter scales are claimed to have good results in assessing mental fatigue, it raises the question whether such situations as “watching TV” or “reading”, would have the same effect on all people, without considering the nature of what they were watching on TV or what book they were reading. These measures are purely based on a psychological estimate by the subjects, which may lead to less accurate estimate of mental fatigue.

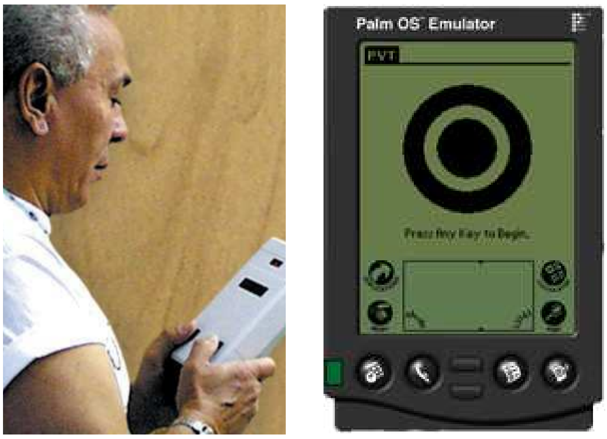
Despite some known issues about the possible subjective bias (see Frey et al., 2004), the subjective self-report measures are easy to administer and are generally believed to have good reliability and good validity, especially in the setting of clinical assessment when subjects are likely cooperative and faithful in their self-rating. However, they cannot be used in some domains, such as transportation industry, where an objective and non-intrusive mental-fatigue measurement method is required.

2.6.3.2 Objective Performance Measures

It is widely known that the mental fatigue due to sleep deprivation causes decrements in the functioning of the brain, so mental fatigue can be measured objectively by the performance of the subjects in performing various mental tasks.

There have been many mental tasks that were designed to assess the functioning of the brain (see Bonnet and Arand, 1999; Griffin and Koonce, 1996; Williamson et al., 2001; Wilson, 2002; Wilson et al., 2007). Some tasks estimate vigilance of the subject by measuring reaction times, while some tasks measure memory or/and decision-making functions. Other objective measures of mental fatigue may not necessarily involve active mental tasks, rather they measure the mental fatigue by finding the propensity to fall asleep by Multiple Sleep Latency Test (MSLT) or Maintenance of Wakefulness Test (MWT) (Bonnet and Arand, 1999).

One of the most commonly used objective measures of mental fatigue is the Psychomotor Vigilance Task (PVT), developed by Dinges and Powell (1985). In this task, as shown in Fig. 2.4a, a visual stimulus is given on the device screen and the subject has to press a response button as soon as possible after perceiving the stimulus. Subtracting the time-stamp of the stimulus from that of the subject's response gives the simple reaction time for that response. The reaction time, averaged over a certain period (usually 10 minutes) with multiple stimuli, is used as the measure of mental fatigue. Many stud-



(a) (b)

Figure 2.4: Objective performance measures of mental fatigue:(a) PVT-192 and (b) PalmPVT (Source: www.ambulatory-monitoring.com)

ies have shown that the reaction time increases as the mental fatigue progresses. The Walter Reed Army Institute of Research has developed a PDA-based PVT, known as PalmPVT , which has also shown close correspondence with the original PVT in terms of results when used in sleep deprivation studies (Ferguson et al., 2005; Lamond et al., 2005; Thorne et al., 2005). The PalmPVT is shown in Fig. 2.4b.

Benefit from its simplicity and portability, the PVT is often treated as the gold standard of performance measures and it has been used extensively in sleep-deprivation studies. However, since the PVT requires the subjects to actively perform mental tasks, such objective mental-fatigue measurement method again cannot be used in some domains, such as transportation industry, where a non-intrusive mental-fatigue measurement method is required. Like other objective performance measures, another disadvantage of PVT is that adminstration of PVT is relatively time-consuming (usually 10 mins).

Apart from general cognitive tasks used for measuring mental fatigue, there are some specific tasks like driving simulator for drivers, multitasking for pilots and some complex decision-making tasks for military purposes (like the Warship Commander Task). These tasks are customized to cater for the specific needs. For example, driving sim-

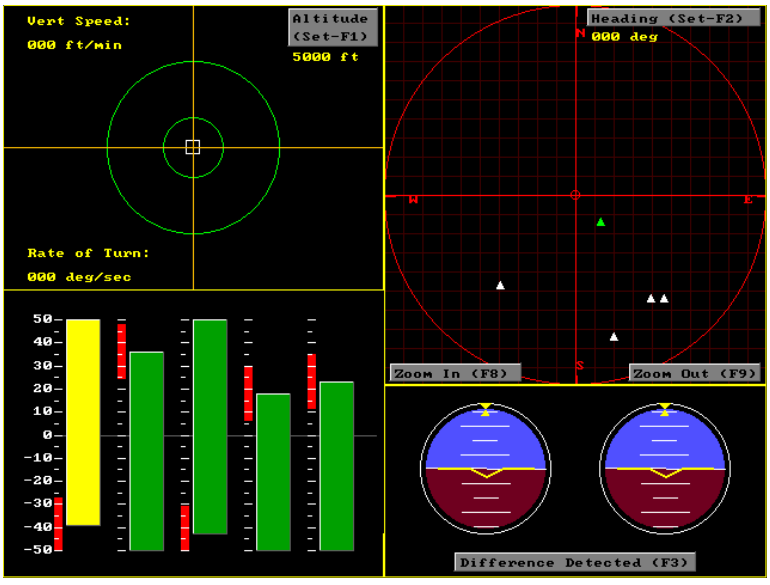


Figure 2.5: The multitasking for pilots includes a visual-motor tracking task, a display of way points over which the pilot has to "fly", a display of two attitude indicators, which sometimes differ, and a series of histograms, the length of which changed from time to time. Another two complex tasks that are directly interacted. (Weinberg et al., 1998)

ulators use driving-specific performance indicators, such as lane deviation, number of crashes and number of times speed limit was crossed over a period of time, to measure mental fatigue. Similarly, multitasking for pilots, as shown in Fig. 2.5, uses tasks similar to what a pilot is supposed to do when on a flight. It shows four tasks on different quadrants of the computer screen to measure metrics like vigilance, tracking etc. Such tasks are very specific for certain domains (drivers, pilots or commanders) and can not be used for other domains. Moreover, they requires complex hardware and software, which greatly limits their use.

2.6.3.3 Behavioral Measures

The behavioral measures are not very commonly used to measure mental fatigue. They are mainly used in field studies to assess the sleep cycles of subjects. One such device is the Actigraph^{TM}, which is watch shaped device containing accelerometers to detect wrist movements in subjects. They are to be worn all the time on the wrist and they record the wrist movement data. The wrist movement is less during sleep, while it is more during waking hours. Analysis of the data collected from an Actigraph$^{\text{TM}}$ device can reveal the sleep cycle of an individual.

2.6.3.4 Physiological Measures

Recently, physiological measures of mental fatigue has sparkled a lot of interests. Besides the EEG-based physiological measures which will be discussed in detail in the next two sections, other physiological measures are mainly based on monitoring the face (especially the eyes) of the subject.

For example, PERCLOS is a physiological measure based on the PERcentage CLOSure of eyelids (PERCLOS). PERCLOS is calculated by processing the subject's face image data coming from an infra red camera. It appears especially promising for driver safety where the subject, i.e. the driver, is supposed to sit in the same location facing the relatively bulky equipment. PERCLOS assumes that the drooping of the eyelids increases as the person becomes more fatigued. Such system can raise alarm to the driver once the percentage of eye closure crosses a certain threshold, thus allowing to taking some necessary counter measures (Dinges, 1998; Vaca, 2005). Besides PERCLOS, other physiological measures of mental fatigue includes eye-blink frequency, nodding frequency, face position and fixed gaze (see Bergasa et al.).

These non-EEG physiological mental-fatigue measurement methods appear to perform well for driving scenarios, but they are also subject to some pitfalls. For example, they may perform badly when there is considerable movements on the part of the driver and they are not suitable for situations where the subject requires to be mobile due to job demand.

As seen in this section, there are many methods being used to measure the level of mental fatigue. Every method has some benefits and some limitations depending on the location and subjects for which it is intended to use. While the self-report questionnaires are easy to administer and can even be computerized or programmed into a portable PDA, they are subject to individual bias and they are obtrusive methods which are not suitable for continuous monitoring of mental fatigue. Objective performance measures provide objective and often more-reliable estimate of mental fatigue, but they are also obtrusive methods that require subjects to perform actively in some mental tasks. Some non-EEG based physiological measures, such as PERCLOS, give good results for driving scenarios, but may fail if there is too much movement on the part of the driver, and also they are not suitable for situations where the job demand of the subjects requires them to be mobile.

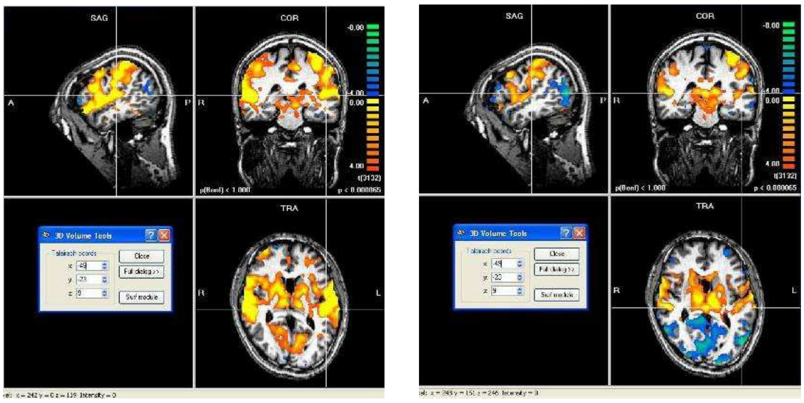
2.7 Neurophysiological Basis of EEG-based Mental-Fatigue Measurement

Among the numerous physiological indicators which have been linked to mental fatigue in the literature, EEG has been shown to be one of the most predictive and reliable one (Horne and Reyner, 1995; Lal and Craig, 2001b). This is directly motivated by the neurophysiological manifestation of mental fatigue, that is, mental fatigue can be with regard to the cortical deactivation in the brain. This cortical deactivation causes the miscommunications between the cortical regions, resulting in various cognitive impairments on alertness, attention and decision making, etc.

The postulation of such cortical deactivation occurring during mental fatigue can be traced back to early 1990s (Brookhuis and de Waard, 1993; de Waard and Brookhuis, 1991; Kecklund and Åerstedt, 1993). In recent years, there has been accumulating evidence for this cortical deactivation pattern in relation to mental fatigue. It is particularly interesting to see a similar deactivation/activation pattern changes occurred in sleep process which have been documented in recent positron emission tomography (PET) studies (Kecklund and Åerstedt, 1993; Maquet et al., 1996; Nofzinger et al., 1997). Similarly, the cortical deactivation due to mental fatigue has also been recently verified in our previous fMRI study (Li et al., 2005). For reference purposes, the major results of this fMRI study are briefly reported in the next. The fMRI study shows: 1) Cortical deactivation in the brain generally mirrored the task performance and hence the mental fatigue. The circadian fatigue caused general decreased activity of the brain which was coherent with the decreased performance. 2) Temporal cortex exhibited consistent activation decreasing trend with the task performance, indicating a direct involvement of the temporal cortex in mediating task performance. 3) As the circadian fatigue progresses, thalamus also showed a similar activation deactivation trend with the time. Thalamus has been found to be involved in mediating attention (Portas et al., 1998). The decreased activation in thalamus could be the indicator of attention loses due to mental fatigue. 4) The medial frontal cortex showed a consistent pattern of activation change: higher in session 1; drops in session 2 and maintains thereafter.

The striking different activation patterns, between a fresh brain in the morning after one night sleep and the fatigued brain after one-night sleep deprivation, can be seen from Figs. 2.6a and 2.6b. Compared with the fresh brain, the fatigued brain shows significant lower activation in various functional lopes.

The fMRI was particularly useful in revealing the neurophysiology of mental fatigue, it is however not practical to use fMRI for mental fatigue measurement due to its huge cost, requirement of shielding room, prohibition of head movement and its low temporal resolution. One of the possible approaches for catching the changes of neuronal activation in the brain can be using EEG. EEG is the recording of the electric activity in the brain, direct capture of the deactivation/activation pattern related to mental fatigue.



(a) (b)

Figure 2.6: The activation patterns shown in fMRI scans for (a) a fresh brain after one night sleep; (b) the fatigued brain after one night sleep deprivation.

In addition, with the recent development of dry EEG sensors (Griss et al., 2002; Taheri et al., 1994), the preparation time for EEG data acquisition is largely reduced, which makes EEG well suited for a mental fatigue tracking device in operational settings.

2.8 Past Work on EEG-based Mental-Fatigue Measurement and Monitoring System

In recent years, the EEG-based mental-fatigue tracking technology has been a focal point of research. As discussed in Section 2.4, EEG is a common physiological indicator that have been successfully used to study physiological or psychological brain states (like wakefulness, sleep cycles), neural diseases and neural injuries. Among the numerous physiological indicators which have been linked to mental fatigue in the literature, EEG has been shown to be one of the most predictive and reliable one (Horne and Reyner, 1995; Lal and Craig, 2001b). Moreover, using EEG for mental-fatigue measurement has many desirable properties: it is an objective, non-obtrusive and efficient mental-fatigue measurement method which is well-suited for traffic safety and other domains where online measurement and monitoring of mental fatigue is crucial.

As early as in the 1970s, researchers have documented considerable evidence of a strong correlation between EEG waveform and mental fatigue (see Grandjean, 1970; H. Fruhstorfer, 1977; Howitt et al., 1978; Kanamori, 1985; Kecklund and Åerstedt, 1993). However, mental fatigue produces much less distinguishable changes in terms of EEG waveforms than other brain states like sleep (Kecklund and Åerstedt, 1993). There have been little, if any, evidence showing the efficacy of measuring mental fatigue by characterizing the EEG waveforms. In recent decades, such EEG waveform approach has given rise to the more powerful quantitative electroencephalogram (qEEG) methods that are equipped with digital signal-processing algorithms for the study of mental fatigue.

Although mental fatigue differs significantly from sleep cycles, many of past EEG-based mental-fatigue studies follow the same approaches used in sleep studies. Using spectrum analysis (a typical qEEG method which was extensively used in sleep studies), researchers have quantitatively shown the associations between EEG delta, theta, alpha, beta activities and mental fatigue during driving (Artaud et al., 1994; Beatty et al., 1974; Dinges and Mallis, 1998; Gevins et al., 1995; H. Fruhstorfer, 1977; Horne and Reyner, 1995; Lal and Craig, 2001a; Lal et al., 2003; Lal and Craig, 2002; Makeig and Inlow, 1993; Makeig and Jung, 1995; Ogilvie et al., 1991; O'Hanlon and Beatty, 1977; O'Hanlon and Kelley, 1977; Torsvall and Åkerstedt, 1987, 1988). Though it requires careful inspection, it is not difficult to see, from the rest of this Section, the common conclusion between most of spectrum analyses—mental fatigue is generally associated with a decrease in the frequencies of the predominant energy bands. However, most previously published studies on EEG changes during mental fatigue have found varying results that could be due to methodological differences and limitations (Lal and Craig, 2002).

For example, Beatty et al. (1974) reported that occipital theta activity in EEG is the most reliable spectral indicator of mental fatigue. It contradicts with the study of O'Hanlon and Beatty (1977) which reported that alert individuals with eyes open show a predominance of beta activity in EEG and mental fatigue can be characterized by a shifting of energy to the alpha band. Though the subsequent studies of H. Fruhstorfer (1977) and O'Hanlon and Kelley (1977) appear partially in favor of O'Hanlon and Beatty (1977) by documenting that increases in both theta and alpha are significant for drowsy individuals with eyes open, somewhat surprising results were obtained in a study by Ogilvie et al. (1991) who claimed that increases in power were found across all standard frequency bands at sleep onset (a state of extreme mental fatigue). This is apparently not the end of the variable literature on mental fatigue. A recent work by Makeig and Inlow (1993) again produced different results to the above.

In the 1980s and 1990s, in contrast to most previous studies that are based on laboratory tests or simulations, a group of Swedish researchers, led by Torsvall and Åkerstedt, carried out a series of field studies to examine mental fatigue in shift workers using ambulatory EEG, EOG and ECG (Åerstedt, 1988; Åkerstedt and Gillberg, 1990; Åkerstedt et al., 1991; Torsvall and Åkerstedt, 1987, 1988; Torsvall et al., 1989). Three-shift workers, such as train drivers and paper-mill workers, were studied during day, afternoon and night shifts. In those studies, only selected spectral bands from a pre-selected single-channel EEG were studied. A 4-channel Medilog tape-recorder was used to record the single-channel EEG (Cz-Oz or O2-P4), the EOG (oblique) and the ECG. The EEG was sampled at 68 Hz and the hourly-averaged spectra of delta (0.5–3.9 Hz), theta (4–7.9 Hz) and alpha (8–11.9 Hz) were sequently obtained by using a special purpose spectrum analyzer. Epochs containing artifacts or sleep ($\geq$4 min of consecutive stage 1 or higher) were removed after visual inspection. The single-channel ECG was also recorded, although it was found that this physiological measurement did not correlate with mental fatigue. The EOG was mainly used for visual identification of slow rolling eye movements which was often considered as the reliable indicator for sleep onset. Throughout the work period, subjective self-report measure of sleepiness was also recorded on a seven-point scale (1–very, very alert; 7–very, very sleepy).

The results from these field studies showed that subjective sleepiness increased significantly during the night shift. The hourly-averaged alpha power density increased marginally at the same time, but no significant changes were found in the hourly-average theta and delta power densities. Interestingly, these studies suggested that the detection of sleepiness may be more clearly determined from burstlike transients in the spectral content of the EEG, rather than from the usual approach of looking at spectral averages over a long period (usually an hour). More specifically, they reported that the very short duration of increased power in the alpha, theta and to a lesser extent, delta bands were the indicators of increases sleepiness, although such burstlike alpha, theta and delta indicators of sleepiness might be engulfed in the much longer time periods without such activity (Åkerstedt and Gillberg, 1990; Åkerstedt et al., 1991; Torsvall and Åkerstedt, 1987; Torsvall et al., 1989).

In another study (Åkerstedt and Gillberg, 1990) where eight subjects were kept awake and active overnight in a sleep lab, it has been showed nicely that whether the eyes are open or closed can make a substantial difference in the EEG spectrum of sleepiness. The results showed that intrusions of slow rolling eye movements and of alpha and theta power density during waking, open-eyed activity strongly differentiated between high and low self-rated sleepiness, while the differentiation was poorer for subjects with eyes closed. They also noticed that slow rolling eye movements might be one of the major reasons for the increased alpha activity during extreme sleepiness and it was much more difficult to differentiate between sleepiness and alert states with eyes closed. This suggests that EEG-based mental fatigue measurement and monitoring method should use the EEG recorded under the setting that the subject's eyes are open. This finding is not trivial because it is not uncommon that EEG with eyes closed were unwisely used in some past work on mental fatigue for the sake of less EOG artifacts. It is worth to point out that, if it is really desirable to keep subjects' eyes closed, it would be valuable to document eye rolling separately by EOG with extra EOG channels.

Since the 1990s, study on the relationship between performance degradation and the EEG spectrum has made an impressive comeback. As early as in the 1970s, Beatty et al. (1974) reported that the performance deterioration in monotonous visual monitoring tasks could be predicted by increased theta band activity in the occipital brain region. Further evidence was reported by (Townsend and Johnson, 1979) who studied the pre-stimulus EEG spectrum to reaction time in some monitoring task for both well-rested subjects and sleep-deprived subjects, although the authors also warned there could be considerable variation in EEG activity which was unrelated to performance. Since the early 1990s, a series of studies have re-affirmed the feasibility of measuring mental fatigue or drowsiness indexed by subject's task performance, based on EEG data in attention-sustained experiments using auditory or visual stimuli (Duta et al., 2004; Jones, 2006; Jung et al., 1997; Makeig and Inlow, 1993; Makeig et al., 2000; Peiris et al., 2004; Sommer et al., 2002; Vuckovic et al., 2002). Most of these studies have focused on the detection of performance lapses in the specific tasks that they studied (i.e. the prediction of a mistake in a specific task) without measuring the subjects' mental-fatigue levels directly. Moreover, most of these pilot studies used fairly simple linear, nonlinear regression or neural networks.

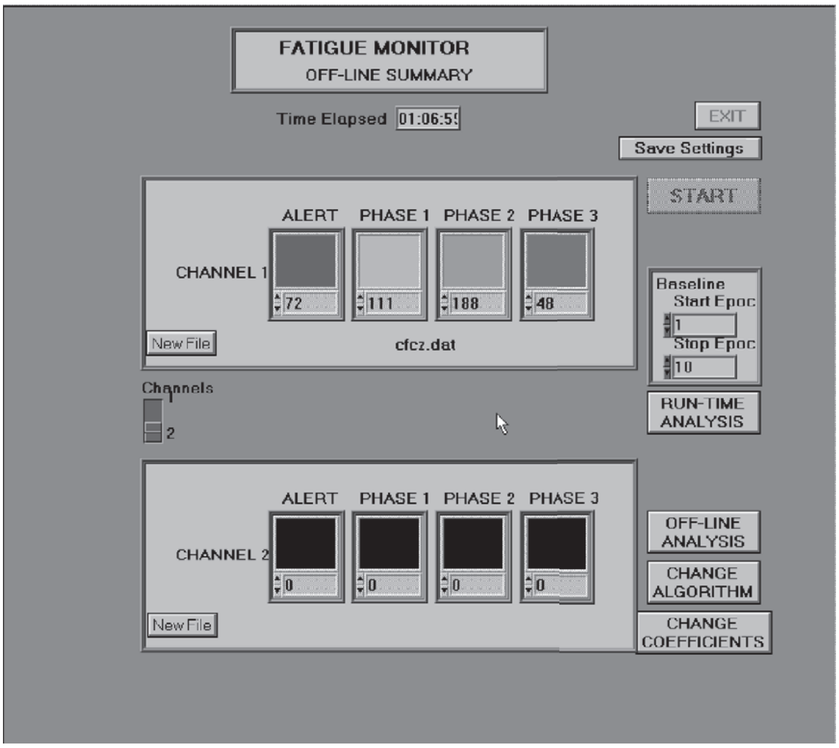
For example, the early work by Makeig and Inlow (1993) showed nicely the coherence of fluctuations in performance and EEG spectrum. In the study, the spontaneous EEG were recorded from thirteen subjects when they were performing an simulated passive-sonar-target detection task with eyes closed. No automatic artifact removal was done, except simple rejection of large eye movements via visual inspection on the recording from an extra periocular channel. The fluctuations in performance was measured by local error rate which was derived by computing the fraction of undetected targets within a time window with a constant width of about 33s, while the power time series in a given frequency was estimated by using a 2.46s moving window. The results of coherence analysis showed that changes in performance were accompanied by nearly simultaneous shifts in EEG spectral power which included drops in alpha, increases in low theta and delta, and moderate increases in sigma band power (13 Hz). The work also demonstrated the possibility of the use of EEG for prediction of performance lapses via a multiple linear regression method.

Sleep onset detection is another relevant research area (see Gennaro et al., 2001; Ogilvie et al., 1991; Virkkala et al., 2007; Yeo et al., 2007). However, detecting the sleep onset in an individual, be it a pilot or be it a driver, may not always mitigate a potentially dangerous consequence: if the individual is already in the sleep stage, even if awakened, there may be insufficient time to avoid an impending accident (Kaplan, 1996). It is more important to measure and monitor mental fatigue in real-time (or close to real-time) before sleep onset so that effective counter-measures or preventions can be put into place at an appropriate stage.

Although the relationship between mental fatigue and EEG has been extensively studied, there have been only a handful of attempts in the literature that were directly aimed to develop an EEG-based mental-fatigue measurement and monitoring system. The automated drowsiness-detection system developed by Gevins et al. (1977) using four referential-channels (C3-P3, P3-01, C4-P4 and P4-02) is probably the first such attempt. In their system, simple decision-heuristics, based on increased ratios of both delta-band to alpha-band and theta-band to alpha-band spectral intensity as compared with thresholds determined for each subject from a randomly-chosen alert EEG baseline, were used. Though their system were only able to differentiate between alert EEG and drowsy EEG (in other words, mental fatigue was measured at only two levels—drowsy v.s. alert) and the signal processing method used might appear pretty humble compared with the state-of-the-art pattern-recognition techniques, the approach of using the spectrum features of ongoing EEG for automatic measuring and monitoring mental-fatigue remains relevant today.

Another case in point is the work by Ninomija et al. (1993), in which a system was developed to detect sleepy states of drivers using a single ongoing EEG feature, i.e. the grouped EEG alpha waves, so as to warn them of the dangerous states. They reported a type II error of 25%–35%. In order to improve the reliability of the system, they suggested to monitor simultaneously the ECG (change of R-R intervals) during driving. As further substantiated by the researchers (Fukuda et al., 1994), the detection of grouped alpha waves was based on moving regression coefficients. The apparent disadvantage in this system is the use of extra electrodes to monitor two separate physiological signals, making it more cumbersome. The use of ECG for mental-fatigue detection also appears controversial. For example, several studies have claimed that ECG did not correlate with mental fatigue (Åerstedt, 1988; Åkerstedt and Gillberg, 1990; Åkerstedt et al., 1991; Torsvall and Åkerstedt, 1987, 1988; Torsvall et al., 1989). Moreover, the literature does not favor the use of alpha activity for mental-fatigue detection. For example, Lal et al. (2003) pointed out that even though alpha marginally increases during drowsiness, the magnitude of change in the delta and theta waves are larger and easier to detect.

The EEG-based driver-fatigue countermeasure system as presented in a series of recent papers by a group of Australian researchers (Lal and Craig, 2001a,b; Lal et al., 2003; Lal and Craig, 2002) are probably the latest effort in the literature to develop an EEG-based mental-fatigue measurement and monitoring system. They used solely the spectrum features of ongoing EEG to differentiate mental fatigue at 4 levels (Fig. 2.7). Specifically, for each EEG channel, the following values were calculated: Dm, Dsd, Tm, Tsd, Am, Asd, Bm, and Bsd, where D, T, A, and B represent the spectral magnitudes in the delta, theta, alpha, and beta bands, respectively, and m and sd represent the mean and standard deviation of those magnitudes. Thresholds were then defined for each frequency band in each channel. The classification of mental fatigue at 4 levels was again determined by the instantaneous spectral magnitudes in each frequency band of a given channel (without artifact removal) and the relation of those magnitudes to the thresholds by us-



Time Elapsed 01:06:5

CHANNEL 1: ALERT 72, PHASE 1 111, PHASE 2 188, PHASE 3 48

CHANNEL 2: ALERT 0, PHASE 1 0, PHASE 2 0, PHASE 3 0

Baseline Start Epoc 1

Stop Epoc 10

Figure 2.7: The display panel of the EEG-based driver-fatigue countermeasure system developed by Lal et al. (2003). Each 30s epoch was allocated to mental fatigue at 4 levels: alert, Phase 1 (transition to fatigue), Phase 2 (transitional–posttransitional phase), and Phase 3 (post-transitional phase). An example of mental-fatigue detection shown in one channel only, i.e. detection from one site on the brain, in this instance the Cz.

ing algorithmic Boolean logic: the approach that is not very different from the one used more than 20 years ago by Gevins et al. (1977). No quantitative performance evaluation was reported in these initial trials of the first prototype of the system.

It needs to be pointed out that, in many EEG-based mental-fatigue studies, mental fatigue was classified into discrete levels. For example, in the above-mentioned studies (Lal and Craig, 2001a,b; Lal et al., 2003; Lal and Craig, 2002), mental fatigue was classified into four phases/levels: early, medium, extreme fatigue phase, and an arousal phase. It is arguable whether mental fatigue should be measured continuously or discretely, but it is reasonable to believe that the progression of mental fatigue may not be entirely smooth or continuous. On the contrary, mental fatigue could be very much like sleep stages where only quasi-categorical sleep stages can be defined. Evidence of such quasi-categorical mental-fatigue stages has also been shown in a recent EEG study by Trejo et al. (2007).

In summary, the literature shows substantial evidence of changes in EEG, such as simultaneous changes in slow-wave activity (e.g. delta and theta activity) as well as alpha activity during mental fatigue. However, most previously published studies on EEG changes during mental fatigue have found varying results and very little evidence exists on the efficacy of incorporating EEG signal detection and analysis into an effective mental-fatigue measurement and monitoring system. This is likely due to methodological limitations. To measure and monitor mental fatigue in (near) real-time fashion, the challenge remains in developing or adapting powerful signal processing methods (running on fast enough computer or processing chip which were not available before) to extract the relevant information from the EEG. As shown in the above review, most studies have computed measures on one or more spectral bands from a priori defined one or more EEG channels, rather than computing full-spectrum of each of the EEG channel in full mapping EEG recordings; Features that have been selected to relate to mental fatigue were often limited to powers of some specific standard frequency bands (often without giving the justification), rather than considering combination of multiple types of features; The technical challenge of automatic removing the pervasive EEG artifacts has rarely been addressed; Moreover, the recent advance in the signal processing methods in the domain of machine learning, like feature selection and multi-category pattern classification, have not been applied in this field.

2.9 EEG Signal Processing

EEG signals are the signatures of brain activities. EEG implementation is all about evaluating and quantifying the EEG signals. Generally, the goal is to relate certain physiological or psychological brain states to particular patterns present in the EEG via appropriate EEG signal processing methods. This section briefly reviews the signal processing methods commonly used in the EEG analysis.

The reader should be aware that most of the mathematics are omitted in this brief review and only the mathematics needed for the subsequent chapters are collected at the end of this chapter in Section 2.10. Further details on EEG signal processing methods can be found in the review paper by Thakor and Tong (2004) or the review book by Sanei and Chambers (2007).

2.9.1 Waveform Inspection

In the beginning of EEG history, clinical researchers relied heavily on visual inspection of EEG waveforms. This conventional visual analysis method of observing the EEG waveform is thought to be subjective and laborious (Thakor and Tong, 2004). In past decades, various qEEG methods (using digital signal processing techniques) have been extensively studied and convincing evidence has been shown that they are capable of capturing EEG patterns that may be difficult, if not impossible, to be captured by manual waveform inspection. The qEEG methods mainly include various methods for EEG signal modeling, filtering and denoising, signal transform, blind signal separation and pattern classification. In the following sections, these qEEG methods will be discussed.

2.9.2 Filtering and Denoising

As discussed in Section 2.3, the raw EEG signals are usually contaminated with various sources of noises and artifacts, such as EOG artifact, ECG artifact, EMG artifacts and 50 Hz or 60 Hz power line noise depending on the local power supply.

Artifacts in EEG are commonly handled by discarding the affected segments of EEG. The simplest approach is to discard a fixed length segment (usually 1-3 seconds) from the time an artifact is detected. The recognition of some artifacts, like eye blink artifacts, are generally effected by detecting a voltage exceeding a threshold (usually 100 μV) in separate EOG channel. Other artifacts are generally ignored or manually marked by a EEG practitioner and then manually discarded. Discarding segments of EEG data with artifacts can greatly decrease the amount of data available for analysis.

Automatic removing/suppressing artifacts and noises is certainly very useful, especially for EEG applications where (near) real-time processing of EEG signals is required. Some noise and artifacts are easy to recognize and can usually be removed by filtering (for example the power line noise). Nevertheless, most artifacts, such as EOG, ECG and EMG artifacts, are present consistently and are difficult to remove. The removal of EOG and ECG artifacts is important because they overlap in amplitude and spectrum of EEG and easily interfere with EEG interpretation.

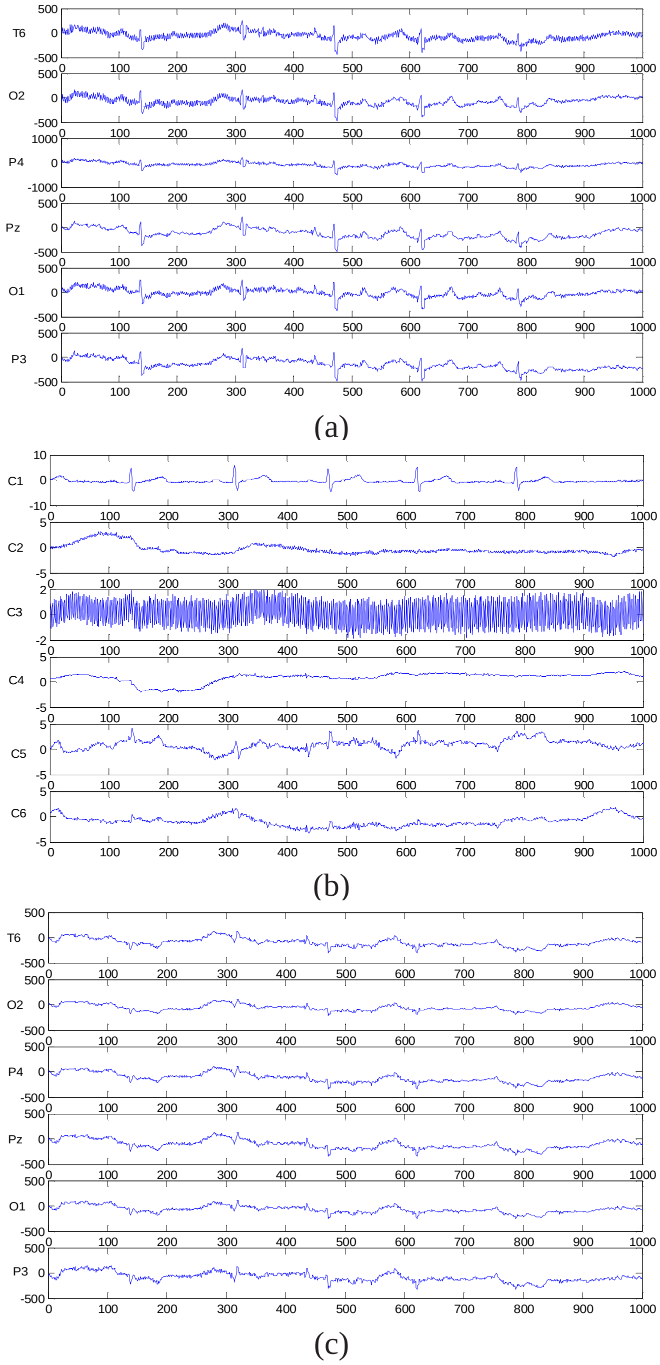
For example, mental fatigue produces much less distinguishable changes in EEG waveforms than other brain states like sleep (Kecklund and Åerstedt, 1993). Meanwhile, it has also been shown in Section 2.8 that the increases in low frequency activities, such as delta and theta band, are important to detect mental fatigue. The normal ECG rhythm of a human is approximately 1.0-1.5 Hz and its second-order harmonics (2.0-3.0 Hz) are within the delta band, whereas the EMG artifacts typically span the whole frequency band. In such case, the influence of ECG and EMG artifacts cannot be ignored and some automatic artifact removal methods are certainly needed.

Regression using the separate EOG/EMG channel (placed near the artifact sources) are the most common type of artifact removal in the literature. The need of the extra channels is apparently one of the drawbacks. Moreover, since the EOG/EMG channel may also contain EEG signals, the regression approach has the undesired effect of removing part of EEG signals. More detailed discussion on this approach can be found in (Croft and Barry, 2000a,b).

In recent years, there has been increasing interest in applying independent-component-analysis (ICA) to separate and remove artifact in EEG (Castellanos and Makarov, 2006; Jung et al., 1998, 2000b; Urrestarazu et al., 2004; Vigário et al., 2000; Vigário, 1997; Wallstrom et al., 2004). For the ease of presentation, the detailed description of ICA algorithm is deferred to Section 2.10. The use of ICA for artifact removal is mainly motivated by the fact that ICA is effective in decomposing raw EEG recordings into artifactual and non-artifactual independent components. Non-artifactual components represent signals from brain activations while artifactual components represent electrical signals originating from non-cerebral artifacts (see Fig. 2.8).

As shown by the example as in Fig. 2.8, the ICA appears impressively promising for EEG artifact removal. Conventionally, artifactual independent components are manually identified (usually by visual inspection) and then removed. This process is very time-consuming and not suitable for real-time applications. However, the automatic identification of artifactual independent components from non-artifactual independent components remains a challenge. Recent effort towards automatic artifact removal using ICA includes (Nicolaou and Nasuto, 2004; Shoker et al., 2005) where a standard support vector machine (SVM) classifier, trained on equal number of artifactual and non-artifactual samples, was used for automatic identification of artifactual independent components. Such a combination of ICA and SVM offers a promising approach for automatic artifact removal. Unfortunately, it will be seen in the subsequent chapters that unique properties of the problem at hand have not been taken into consideration and further research is needed.

Several filtering methods have also been used to examine particular frequency bands of interest. Particularly, lowpass, highpass and bandpass filters are routinely used in pre-



(a)

(b)

(c)

Figure 2.8: An example of ICA-based artifact removal: (a)One segment of real EEG data– the ECG artifact is prominent in all channels and the 50 Hz power line noise is significant in T6,O2; (b)The resulting independent components separated by the ICA– the component c1 is ECG artifact source while the c3 is 50 Hz power line noise source; (c)The reconstructed EEG segment after discarding ECG artifact and 50 Hz power line noise (i.e. the components c1 and c3).

processing of EEG signals. During the data collection stage, embedded analog infinite-impulse-response (IIR) filters are normally used for anti-aliasing and removing the high frequency activities of no concern. During the post-processing stage, both IIR filters (usually a digital implementation) and finite-impulse-response (FIR) filters are commonly used to further enhance the EEG signals.

2.9.3 EEG Signal Modelling

After removing or depressing the artifacts and noise, the question of EEG signal modelling naturally arises: can we have a model that regulates the EEG signals? The linear modeling and nonlinear modelling methods described in the next few subsections are the most common methods used to extract a parametric description of the EEG signals.

2.9.3.1 Linear Modelling

The main objective of linear modelling is to find a set of model parameters that best describe the EEG signal generation system for each EEG channel. The most common method is the autoregressive (AR) modelling. For the i^{th} channel, we have

$$z_i(n) = -\sum_{k=1}^{p} a_k z_i(n-k) + \varepsilon_i(n), \tag{2.1}$$

where $z_i(n)$ denotes the EEG time series of the i^{th} channel, a_k, $k = 1, \cdots, p$, are the model parameters, n denotes the discrete samples (time interval normalized to unity), and $\varepsilon_i(n)$ is the noise input. The following variant of AR modelling, i.e. the autoregressive moving average (ARMA) modelling, has also been used:

$$z_i(n) = -\sum_{k=1}^{p} a_k z_i(n-k) + \sum_{k=0}^{q} b_k \varepsilon_i(n-k), \tag{2.2}$$

where $b_k, k = 1, \cdots, q$, are the additional model parameters. It differs from the AR modelling in that each sample is predicted not only by the previous samples but also by the previous noise inputs.

In contrast to the AR and ARMA modelling methods which treat each channel independently, multivariate AR approach has also been considered to model multi-channel EEG signals as a whole. In such multivariate approach, each sample is predicted by both its previous samples from the same channel and the previous samples from the other channels. Hence, for the i^{th} channel,

$$z_i(n) = -\sum_{k=1}^{p} a_{ik} z_i(n-k) - \sum_{j=1, j \neq i}^{M} \sum_{k=1}^{p} a_{jk} z_j(n-k) + \varepsilon_i(n), \tag{2.3}$$

where M represents the number of channels, and $z_i(n)$, $\varepsilon_i(n)$ represents the output sample and input noise for the i^{th} channel respectively.

In general, the model parameters in the above linear modelling methods are estimated either directly (such as through maximum likelihood estimation) or by employing some iterative optimization schemes (Sanei and Chambers, 2007).

In recent years, it has been argued that there are advantages of AR modelling over the classical Fourier transform. However, there are significant known issues on the AR modelling method, such as the difficulty in choosing the appropriate model order and the challenge of selecting an appropriate length of EEG segment, which largely limits its use in EEG signal modelling (Sanei and Chambers, 2007; Thakor and Tong, 2004).

2.9.3.2 Nonlinear Modelling

There are also nonlinear methods being considered for the purpose of EEG signal modelling, in which the output EEG samples are nonlinearly related to its previous samples.

In the generalized autoregressive conditional heteroscedasticity (GARCH) method, each sample are related to its previous samples through a sum of nonlinear functions. This model was originally introduced for time-varying volatility (Nobel Prize in Economic Sciences in 2003). The study of such approach in EEG signal modelling is still in its infancy.

2.9.4 Non-stationarity and Signal Segmentation

A time series signal can be deemed stationary if there is no considerable variation in its statistics. The EEG signals are typically non-stationary and they are considered stationary only within short intervals (usually about 3 s for human resting EEG).

Since the EEG signals are non-stationary (or quasi-stationary within a short interval), it is often necessary to segment the EEG signals into epochs of similar characteristics. Such EEG segmentation is not only meaningful to clinicians in EEG diagnosis, but also to many EEG signal processing methods that assume the stationarity of the signals (for example, the AR modelling method).

To address this need, various adaptive parametric segmentation methods have been proposed for automatic EEG segmentation. In loose terms, the procedure of adaptive parametric segmentation is based on the estimation of the similarity index of an initial fixed interval of EEG with an EEG interval of the same duration viewed through the moving window running along the EEG recording. The similarity index will presumably drop sharply when the window runs over a segment boundary, signaling a transition to the following segment. For example, the afore-mentioned autoregressive modelling method, which predicts the EEG sample at a given moment by its previous samples, can be useful for automatic EEG segmentation. The discordance between predicted and real EEG samples could be a sufficient indication of a local non-stationarity (Jansen, 1991).

Studies on automatic EEG segmentation in turn reveal the non-stationarity of the EEG signals. From the reported results of the EEG adaptive segmentation based on the autoregressive models, the quasi-stationary segment, in general, spans from 2–20 seconds (Barlow et al., 1981; Creutzfeldt et al., 1985; Jansen, 1991). Use of multiple regression analysis for EEG adaptive segmentation also reveals a similar duration of 2–10 seconds for the majority of quasi-stationary EEG segments (Inouye et al., 1995). It will be seen in a later section that this piece-wise stationary (or quasi-stationary) structure should be taken into consideration when the EEG signals are processed piece-wisely using a moving window.

Although the adaptive parametric segmentation methods appears to be effective for segmentation of EEG signals, there is inherent contradiction in the parametric segmentation (Kaplan and Shishkin, 2000). In principle, the parametric methods of adaptive segmentation makes it possible to describe adequately the piece-wise stationary structure of the EEG signals. However, all these methods designed for the analysis of non-stationary processes are based on a procedure (usually based on autoregressive model) which may be applicable only to stationary processes. It is evident that accurate fitting of a model can be achieved only on a stationary interval. The longer the interval, the finer characteristics of the process can be represented by the model. But the longer the analyzed interval of the real EEG, the more probable the incidence of heterogeneities within it. If the model is constructed on a very short interval, it will be very rough and the results of segmentation based on the parameters of this model cannot be expected to be of high quality (Brodsky et al., 1999).

Thus, parametric segmentation methods were rarely used to detect brain activity or pathological abnormality. Better systems (usually a pattern recognition system) should be considered for such purposes (Sanei and Chambers, 2007). Section 2.9.7 in this chapter will review such approach.

2.9.5 Signal Transforms

Since EEG signals can be deemed piece-wise stationary, it is straightforward to characterize them in either the time or frequency domain (after applying fast Fourier transforms). The Wavelet transform also offers another alternative for a time-frequency analysis of EEG signals.

2.9.5.1 Fast Fourier transform

The frequency-domain representation of a finite-length signal can be obtained by using the (discrete) fast Fourier transform. The spectrum analysis using Fast Fourier transform are routinely used to relate the standard frequency bands (as in Table 2.1) to specific physiological or pathological brain states.

Since the EEG signals are non-stationary, to track the temporal dynamics in the frequency contents of the signal, the EEG signals are usually segmented into epochs (via a fixed-length moving window running along the EEG time series) and consecutive transforms are then performed on each epoch. It needs to be pointed out that the choice of window length, overlapping between adjacent epochs, window type in performing the fast Fourier transform are critical to successful capture of subtle frequency shifts in the EEG signals (Gevins, 1987; Thakor and Tong, 2004).

It is worth noting that parametric spectrum estimation methods such as those based on AR or ARMA modelling can potentially outperform the (discrete) fast Fourier transform in presenting the frequency contents of the EEG signals, but they may also underperform due to poor estimation of the model parameters (mainly because of the nonstationarity of the EEG signals). The selection of model order in AR or ARMA models is another problem that has not been fully addressed. A high model order may artificially split a true peak in spectrum, whilst a low model order may lead to aliasing between nearby peaks in spectrum.

The (discrete) fast Fourier transform has fixed time and frequency resolutions. Higher time and frequency resolution can be obtained through joint time-frequency analysis such as the time-frequency presentation obtained through Wigner-Ville distribution (see Thakor and Tong, 2004). Nevertheless, such joint time-frequency analysis has notable limitations: cross-term calculations may give rise to negative energy and the aliasing effect may distort the spectrum.

2.9.5.2 Wavelet Transform

The wavelet transform is another alternative for a time-frequency analysis. The unique property of the wavelet transform is that it provides adaptive time-frequency resolutions by using scalable time-frequency kernel functions instead of the fixed-scale window function. It is significant because one usually needs more time accuracy in locating the transient waves while being more interested in the frequency resolution for the EEG signals dominated by slow waves. By using a variable-scaling which is shorter for higher frequency and longer for lower frequency, the wavelet transform method can potentially better localize the signal components in time-frequency domain. The details are readily available from the literature (e.g. Murenzi et al., 1988) and hence omitted.

2.9.6 Nonlinearity

Researchers have also been looking for nonlinear characteristics in the EEG signals such as chaotic measures (Pritchard and Duke, 1992, 1995; Sarbadhikari and Chakrabarty, 2001). These nonlinear chaotic measures are generally borrowed from chaos theory or nonlinear dynamics which has been a rapidly developing area in physics since the 1980s.

The most commonly used nonlinear measures include dimension estimation (see Lee et al., 2001), Kolmogorov entropy (see Pritchard and Duke, 1992, 1995; Sarbadhikari and Chakrabarty, 2001) and Lyapunov exponent spectrum (see Aftanas et al., 1997; Fell et al., 1993; Iasemidis et al., 1990; Kim et al., 2000). These nonlinear measures offer a new class of features that can be generally useful for the EEG pattern classification systems as described in the next.

2.9.7 Patten Classification

The objective of classification is to draw a discriminant boundary between two or more classes and to be able to label a new sample to an appropriate class based on its measured features. In the context of EEG signal processing, the classification of the data in feature space is often preferred. In fact, formulating an EEG signal processing problem into a classification problem is involved in a great deal of recent EEG history. For example, the EEG-based automatic classification of sleep stages has proved to be a big success. Researchers have also been studying the classification of mental tasks by using EEG, such as classification of left and right finger movements which has been nicely demonstrated in the rapidly-growing EEG research area of brain-computer-interface. Also, as discussed in Section 2.9.2, automatic artifact removal can be boiled down to a classification problem after the EEG signals have been decomposed into both artifactual and non-artifactual source signals.

Many classification methods have been developed in the domain of machine learning (Duda et al., 2000; Vapnik, 1995). Among them, linear discriminant analysis (LDA), *k*-nearest neibor algorithm (KNN), artificial neural networks (ANN) and, more recently, support vector machine (SVM) has been widely used in many real-world pattern classification problems . However, many classification methods do not necessarily perform well on classification of EEG patterns due to the inherent challenges of the learning problems at hand.

First, the number of features involved in EEG pattern classification is usually large. This is mainly due to the following two reasons: (i) the inherent redundancy of the EEG data (i.e. the big number of channels, say 19 channels according to the international 10-20 electrode placement system); (ii) the limited knowledge of the neural circuitry in the brain (because lack of domain knowledge may force us to include as many types of features as possible, as long as these features show some correlation with the targets). However, direct classification using all possible features is apparently undesirable, since irrelevant and redundant features have adverse effect on the overall classification performance and generalization ability of the system. Therefore, a data-driven approach of selecting the key features is generally useful.

Second, the classification of EEG patterns typically involves unbalanced data where minority class(es) can be very much under-represented in the data (with relatively few samples). It will be seen in the later sections of this chapter that handling such unbalanced data is a challenging problem that is still an ongoing research topic.

Third, the classification of EEG patterns can be somewhat ill-posed where there are a large number of features but with a relatively small number of samples. This, coupling with the fact that the classification generally involves more than two classes, makes it difficult to relate input features to output targets.

Last, there is generally a nonlinear-mapping between the input features and the output targets in EEG classification tasks. Fitting such unknown nonlinear-mapping function is by itself a challenging task.

2.10 Mathematical Background

This section collects the mathematics from the literature and establishes the necessary notations needed for the subsequent chapters. They serve as the mathematical background of this doctoral study.

2.10.1 Independent-Component-Analysis

Independent-component-analysis is a recently-developed algorithm for blind source separation (Common, 1994; Hyvarinen, 2000; Jutten and Herault, 1991), in which case the original independent sources are assumed to be unknown, and yet to be separated from their weighted mixtures.

2.10.1.1 The Concept

The ICA is best explained by the cocktail party problem as shown in Fig. 2.9. Imagine that you are in a room where two people (denoted by two speakers in Fig. 2.9) are speaking simultaneously. There are two microphones which are placed in different locations in the room. The microphones give you two recorded speech signals, denoted by z_1 and z_2 at the time instance t (the time index t is omitted in the expression for simplicity). Each of these recorded signals is a weighted sum of the speech signals emitted by the two speakers, denoted by s_1 and s_2 at the time instance t. Let's disregard any time delay in the speech transmission for simplicity. The relationship can be expressed by the following linear equations:

$$\begin{aligned} z_1 &= a_{11}s_1 + a_{12}s_2; \\ z_2 &= a_{21}s_1 + a_{22}s_2, \end{aligned} \tag{2.4}$$

where a_{11}, a_{12}, a_{21}, a_{22}, are so-called mixing coefficients that depend on the distances of the microphones from the speakers. It would be very useful if you could now recover the two original speech signals s_1 and s_2, using only the recorded mixed signals z_1 and z_2. Actually, if we knew the mixing coefficients a_{ij}, we could solve the linear equations in (2.4) easily. The point is, however, that if we do not know the a_{ij}, it becomes a much more difficult problem, i.e. the well-known *cocktail-party problem*.

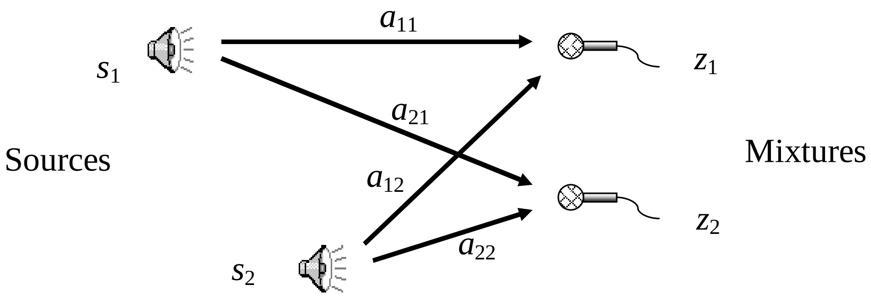


Figure 2.9: Cocktail party problem

The ICA was originally developed to deal with problems that are closely related to the cocktail-party problem (Common, 1994; Hyvarinen, 2000; Jutten and Herault, 1991). It uses some information on the statistical properties of the signals s_i to estimate the a_{ij}. Actually, and perhaps surprisingly, it turns out that it is enough to solve the above-mentioned cocktail-party problem by assuming that s_1 and s_2, at each time instant t, are statistically independent. It needs to be pointed out that the independence assumption is not an unrealistic assumption in many cases and that the assumption needs not be exactly true in practice. That means, even if s_1 and s_2 are loosely dependent to each other, the ICA can still give very good estimates of them.

Fig. 2.10 gives a simple but impressive experiment on ICA. Fig. 2.10a shows the original source signals which include a sinusoid wave s_1, a funny wave s_2, a saw-tooth wave s_3 and an impulsive noise wave s_4, while Fig. 2.10b shows the linearly-mixed signals, z_1, z_2, z_3 and z_4, generated from the source signals using a (unknown) randomly-generated coefficients a_{ij}, $i, j = 1, \cdots, 4$. The ICA is used to estimate the a_{ij} by using only the

z_i, $i = 1, \cdots, 4$. As can be seen from Fig. 2.10c, ICA outputs the estimated independent components (ICs) that are very close to the original source signals. Their signs and appearing order may be different, but these generally have no significance for signal processing.

The ICA appears especially useful for EEG signal processing. The EEG data consist of recordings of electrical potentials in many different locations on the scalp. These potentials are presumably generated by mixing the underlying components of both brain activities and artifacts. This scenario is very similar to the cocktail-party problem: it is very useful to recover the original components (of either brain activities or artifacts), but we can only observe/record mixtures of these underlying components. The ICA has shown great potential in EEG signal processing by giving access to its independent components.

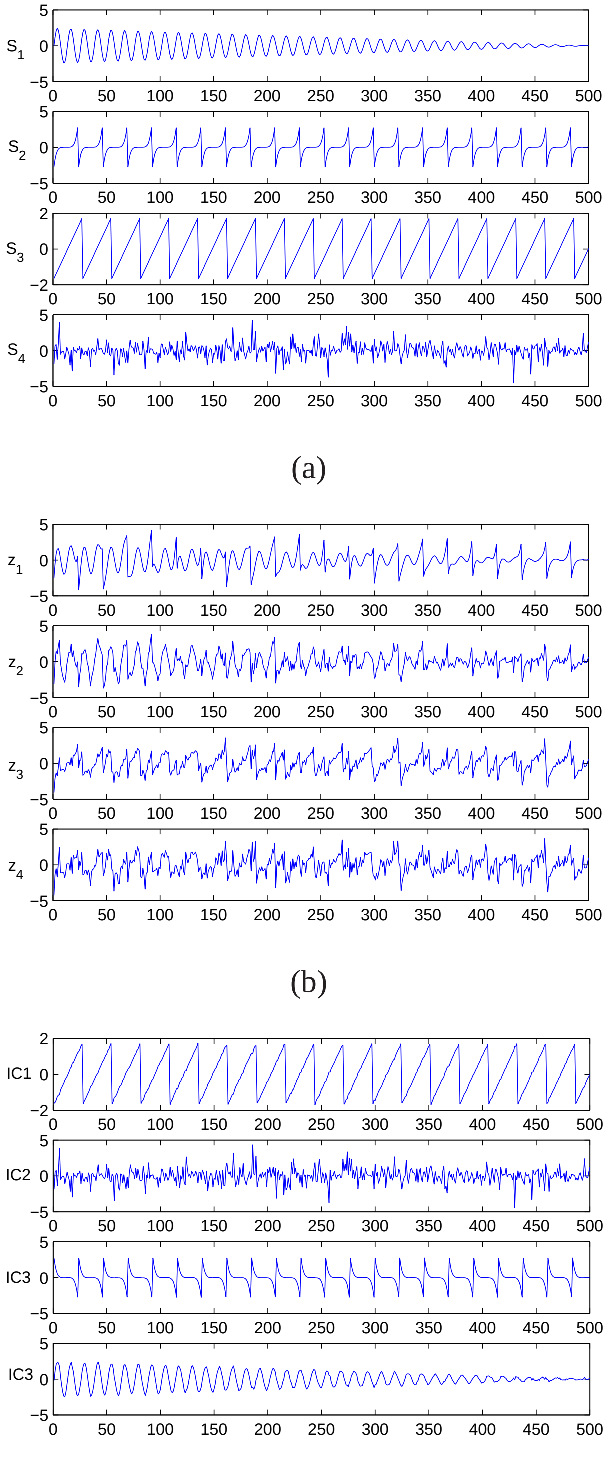
2.10.1.2 The Model

The basic data model used in defining ICA assumes that the observed n-dimensional data vector at time instant t, $\mathbf{z} = [z_1, \cdots, z_n]^T$, is given by

$$\mathbf{z} = \sum_{i=1}^{m} \mathbf{a}_i \mathbf{s}_i = \mathbf{A}\mathbf{s}, \tag{2.5}$$

where $\mathbf{s} = [s_1, \cdots, s_m]^T$ is m independent source signals with zero mean (which can be guaranteed by explicitly extracting the mean of each z_i without loss of generality), and $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_m]$ is a constant mixing matrix which is a function of the location of the sources, the channel positioning in the EEG recording, the shape and the conductivity distribution of the brain as a volume conductor (Vigário, 1997).

As in a general blind signal separation problem, $\mathbf{A}$ is assumed to be an $n \times m$ matrix of full rank (there are at least as many mixtures as the number of independent sources,



(a)

(b)

(c)

Figure 2.10: An experiment on ICA using artificial signals: (a) original source signals; (b) mixed signals using a randomly-generated mixing coefficients; (c) the recovered source signals by ICA using only the mixed signals.

i.e. $n > m$). In addition, although $\mathbf{A}$ is unknown, we assume it to be constant, or semi-constant (preserving local constancy) in order to perform ICA.

If let $\mathbf{W}^T$ denote the pseudo-inverse of $\mathbf{A}$, the problem of solving Equation (2.5) can be redefined equivalently as to find the separating matrix $\mathbf{W}$ that satisfies

$$\mathbf{s} = \mathbf{W}^T\mathbf{z}. \tag{2.6}$$

2.10.1.3 The ICA Algorithm

It has been documented that the preprocessing of the input data (mixtures) by whitening can significantly ease the separation of the source signals (Karhunen et al., 1997). Therefore, standard principal-component-analysis (PCA) for whitening $\mathbf{z}$ is implemented in preprocessing. It can be shown in the compact form (noting that we have dropped the time index t):

$$\mathbf{v} = \mathbf{V}\mathbf{z}, \tag{2.7}$$

where $E\{\mathbf{v}\mathbf{v}^T\} = I$ with I denoting the $n \times n$ unit matrix. The whitening matrix $\mathbf{V}$ is given by

$$\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^T, \tag{2.8}$$

where $\mathbf{D} = diag[1, \cdots, m]$ is a diagonal matrix comprising the eigenvalues of covariance matrix $E\{\mathbf{z}\mathbf{z}^T\}$ as its diagonal elements, and $\mathbf{E}$ is a matrix with the corresponding eigenvectors as its columns.

The starting point for ICA is the very simple assumption that the components s_i are statistically independent. There are several ICA algorithms proposed in the literature, each using a different measure of independence. The most popular ICA algorithms are based on mutual information or non-Gaussianity. It has been shown (Hyvarinen, 1999, 2000) that mutual information (MI) is a measure of independence and that maximizing the non-Gaussianity of the source signals is equivalent to minimizing the mutual information between them.

To illustrate the idea of ICA algorithm, let's focus on the ICA algorithm that uses the non-Guassianity as the measure of independence in which the classical fourth-order cumulant or kurtosis is used to quantify the non-Guassianity of a signal. Let's consider a projection $u = \mathbf{w}^T\mathbf{v}$ and the kurtosis as defined by

$$kurt(y) = E\{u^4\} - 3[E\{u^2\}]^2, \tag{2.9}$$

where the operator E denotes the mathematical expectation. In this context, finding an independent source signal is essentially to find a projection $\mathbf{w}$ of the recorded mixtures $\mathbf{z}$ that maximizes the norm of the kurtosis in (2.9).

Then, a fixed-point ICA algorithm using gradient descent searching approach (FastICA) algorithm (Hyvarinen, 1999, 2000) is used to search the expectation maximization of (2.9). As a result, rows of the separating matrix $\mathbf{W}$ and corresponding independent sources are identified one by one, up to a maximum of m. The basic steps of this ICA algorithm are as follows.

Algorithm 2.1: Main steps of FastICA algorithm.

```
Choose initial vector $\mathbf{w}_0$ randomly and let the iteration step $k = 0$;
while Convergence/Stop Criterion is not met do
    Let $\mathbf{w}_k = E\{\mathbf{v}(\mathbf{w}_{k-1}^T\mathbf{v})^3\} - 3\mathbf{w}_{k-1}$;
    Let $\mathbf{w}_k = \mathbf{w}_k/||\mathbf{w}_k||$;
end
```

2.10.2 Support Vector Machine

The support vector machine (SVM) is a supervised learning method used for classification and regression. It was originally designed for two-class classification. Unlike other statistical learning methods (such as neural networks and decision trees) which usually aim only to minimize the empirical classification error, SVM simultaneously minimizes the empirical classification error and maximizes the geometric margin in classification; hence it is also known as maximum margin classifier (Boser et al., 1992; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Vapnik, 1995, 1998).

The SVM is a powerful supervised learning method and it has a firm mathematic foundation in the framework of statistical learning theory (Vapnik, 1995). The literature has documented its superior performance on a variety of applications (Boser et al., 1992; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Vapnik, 1995, 1998).

2.10.2.1 Two-Class SVM

To understand the concept of the SVM, let's start from a two-class classification problem with a two-dimensional linearly-separable training dataset. It will be seen that there will be no change in SVM formulation for the multi-dimensional cases. For such linearly-separable samples as shown in Fig. 2.11, a discriminant plane (or hyperplane) is sufficient to separate the samples from the two classes. Apparently, there is an infinite number of such possible planes that could correctly classify the training data without any error. However, there is only one plane that is optimal. It is straight-forward that the optimal plane is the one that separates the samples without error and, in the meantime, its distances to the closest samples from both classes is maximal. This optimal plane can be similar to the one shown in Fig. 2.11, which can presumably generalize best (i.e. classifying the unseen test data with the lowest error) since it gives the maximum geometric margin in the classification.

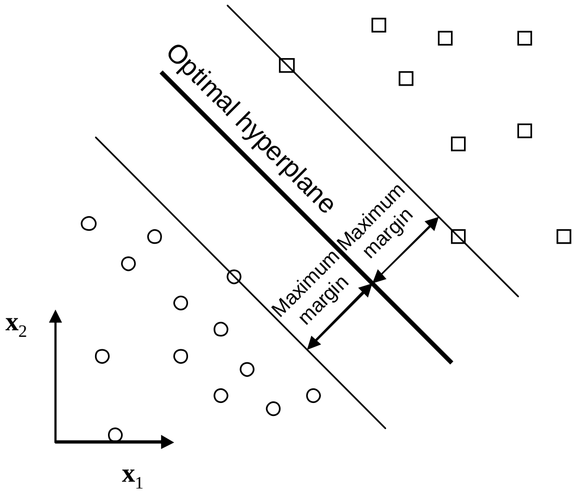


Figure 2.11: Optimal separating hyperplane

One way to find the optimal separating plane in a linearly-separable case is through constructing the so-called *convex hulls* of samples from each class as shown in Fig. 2.12. The enclosed regions are the convex hulls for the respective class. By examining the hulls, it is possible, albeit not automatically, to determine the closest two points lying on the hulls of each classes (note that these do not necessarily coincide with actual data points as in the case of Fig. 2.12. The optimal separating plane is then determined as the perpendicular and equidistant plane to these points as shown in Fig. 2.12.

The SVM actually stems from the idea of formulating the seeking of the optimal separating plane that separates samples from two classes with maximum margin. To formulate the SVM, let's again start with the simplest case, i.e. the linear machine for linearly-separable samples. It will be seen latter that the SVM for the general case, nonlinear machine for non-linearly-separable data results in a very similar mathematical formulation.

Suppose that a dataset $\mathbb{D}$ for the linear-separable case is given in the form of $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ where $\mathbf{x}_i \in \mathbb{R}^D$ is the i^{th} sample, $y_i \in \{1, -1\}$ is the corresponding class label. Here, N is

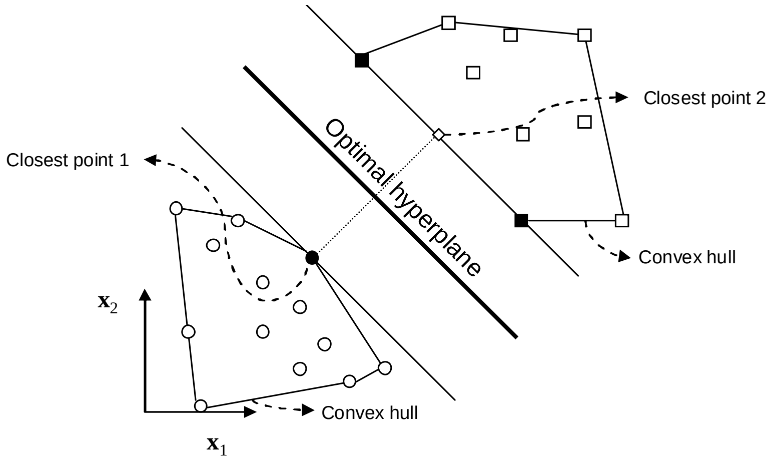


Figure 2.12: Determination of the optimal separating hyperplane using the concept of convex hulls

the number of samples and D is the dimensionality. The objective is to find the optimal separating hyperplane with maximum margin.

Consider the hyperplane is given as

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \tag{2.10}$$

where $(\cdot)$ refers to inner product (dot product) operator, $\mathbf{w}$ is normal to the hyperplane, $b/||\mathbf{w}||$ is the perpendicular distance from the hyperplane to the origin, and $||\mathbf{w}||$ is the Euclidean norm of $\mathbf{w}$ as shown in Fig. 2.13. The SVM simply looks for the separating hyperplane that provides the maximum margin between the two classes. It will be shown that the approach is to reduce the search of the optimal hyperplane to a convex optimization problem by minimizing a quadratic function under some linear inequality constraints. It should be noted that the hyperplane as in Equation (2.10) can be scaled arbitrarily. This allows to transform the problem of maximizing the margin to a problem of minimizing the norm of the weight vector, by setting the functional margins to be equal to unity, i.e. setting $|\mathbf{w} \cdot \mathbf{x} + b| = 1$ for the closest samples to the hyperplane, which can always be guaranteed by appropriate scaling of Equation (2.10) (The hyperplane

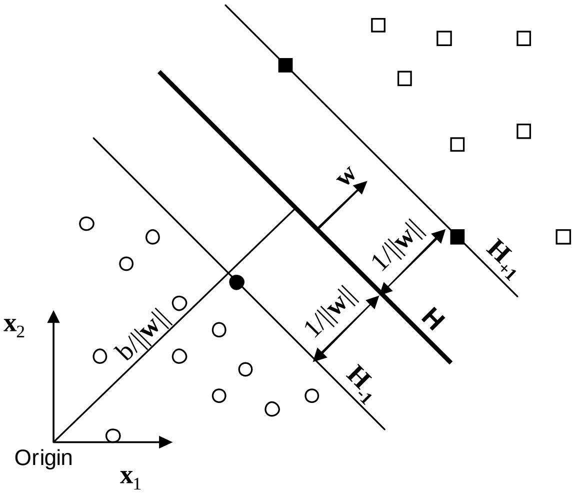


Figure 2.13: Training of the linear SVM, for a linearly-separable case, is to find the optimal hyperplane (thick line) which separates the samples from two classes (circles vs. squares) with maximum margin. The support vectors are shown as solid circles or squares.

with a functional margin of unity is sometimes referred to as canonical hyperplane). As the result, the optimal hyperplane can be determined by the following optimization problem:

$$\min L(\mathbf{w}, b) = \min \frac{1}{2}||\mathbf{w}||^2, \tag{2.11}$$

$$\text{subject to} \quad \mathbf{w}\cdot\mathbf{x}_i + b \geq +1, \text{ for } y_i = +1; \tag{2.12}$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1, \text{ for } y_i = -1. \tag{2.13}$$

It is not difficult to see that the closest samples to the optimal hyperplane are those that satisfy the equality in Equations (2.12) and (2.13). They are so-called *support vectors* which has a unity distance to the optimal hyperplane and, in other words, which define the supporting hyperplane (H_{-1} and H_{+1} in Fig. 2.13). The margin between these supporting planes can be shown to be $2/||\mathbf{w}||$ as shown in Fig. 2.13. In the figure, the support vectors are also highlighted using solid circles or solid squares.

The inequality constraints as in Equations (2.12) and (2.13) can be combined into one set of inequality constraints as $y_i(\mathbf{w}\cdot\mathbf{x}_i+b)-1\geq 0,\ \forall i$. Therefore, the optimization problem that determines the optimal separating hyperplane can be simplified to the following *primal form*:

$$\min J(\mathbf{w},b)=\min\frac{1}{2}||\mathbf{w}||^2, \tag{2.14}$$

$$\text{subject to}\quad y_i(\mathbf{w}\cdot\mathbf{x}_i+b)-1\geq 0,\ i=1,\cdots,N. \tag{2.15}$$

The above primal form is usually solved via its equivalent *dual form*. The following dual form is obtained through introducing the Lagrangian multipliers:

$$J(\mathbf{w},b,\boldsymbol{\alpha})=\frac{1}{2}||\mathbf{w}||^2-\sum_{i=1}^{N}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i+b)-1], \tag{2.16}$$

where $\alpha_i\geq 0$ is the non-negative Lagrangian multiplier for the i^{th} equality in Equation (2.15). Please note that $J(\mathbf{w},b,\boldsymbol{\alpha})$ has to be minimized with respect to $\mathbf{w}$, b and maximized with respect to α_i. Hence,

$$\frac{\partial J(\mathbf{w},b,\boldsymbol{\alpha})}{\partial\mathbf{w}}=\mathbf{w}-\sum_{i=1}^{N}y_i\alpha_i\mathbf{x}_i=0, \tag{2.17}$$

i.e.

$$\mathbf{w}=\sum_{i=1}^{N}y_i\alpha_i\mathbf{x}_i=0, \tag{2.18}$$

and

$$\frac{\partial J(\mathbf{w},b,\boldsymbol{\alpha})}{\partial b}=\sum_{i=1}^{N}y_i\alpha_i=0. \tag{2.19}$$

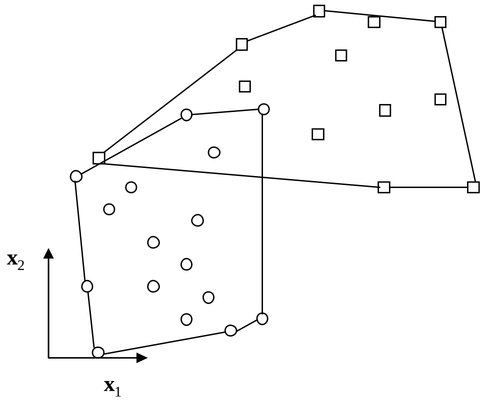


Figure 2.14: Overlapping convex hulls for the non-linearly-separable case

By substituting these into Equation (2.16), the dual form is obtained as

$$\begin{aligned}
&\max J(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}}[\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j],\\
&\text{subject to} \quad \sum_{i=1}^{N} y_i\alpha_i = 0; \qquad (2.20)\\
&\qquad\qquad\quad \alpha_i \geq 0 \quad \forall i.
\end{aligned}$$

It is a well-known quadratic programming problem, for which many numerical solutions are available (Hsu et al., 2004; Joachims, 1999; Platt, 1999).

So far, the SVM formulation for determination of the optimal separating hyperplane for linearly-separable cases has been described. However, many practical classification problems deal with non-separable data as shown in Fig. 2.14 (they have overlaps in their convex hulls for the two classes). Obviously, it may be possible to define a complicated nonlinear hyperplane to separate the data perfectly but it is well-known in machine learning community that it causes the overfitting problem which adversely affects the generalization ability of the classifier.

Assuming that a discriminant hyperplane is still desirable for these non-separable cases, the SVM handles such cases via the concept of so-called *soft margin classifier* (see Fig.

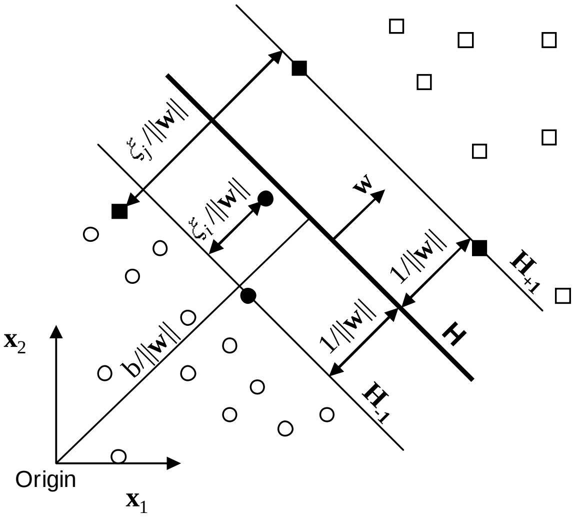


Figure 2.15: The concepts of the soft margin and the slack parameter used for the linear SVM for the non-separable case.

2.15). The term “soft margin” means that the margin constraints are relaxed to allow for some violation of some samples (the violation is denoted by a non-negative slack variable ξ_i as shown below) and that this violation is meanwhile given proportionate influence on the location of the hyperplane. Thus the primal form as in Equations (2.14) and (2.15) is changed to

$$\min J(\mathbf{w}, b, \boldsymbol{\xi}) = \min \left[\frac{1}{2}||\mathbf{w}||^2 + C \sum_{i=1}^{N} \xi_i \right], \tag{2.21}$$

$$\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i \geq 0, \; i = 1, \cdots, N, \text{ and} \tag{2.22}$$

$$\xi_i \geq 0, i = 1, \cdots, N, \tag{2.23}$$

where C is the generalization parameter that offers a trade-off between accuracy of data fit and regularization.

As shown in Fig. 2.15, only those samples that violate the supporting hyperplane (H_{-1} and H_{+1} for class -1 and class $+1$ respectively) have a positive slack parameter ξ and their distance to the respective supporting hyperplane is $\xi/||\mathbf{w}||$. Therefore, for an error to occur, the corresponding ξ must exceed unity. Hence, $\sum_i \xi_i$ is an upper bound on the number of training errors. Equation (2.21) clearly shows the motivation of the soft margin concept used in the SVM for non-separable cases: it offers a trade-off between the empirical classification accuracy (the training error) and regularization capability (the geographic margin). A small value of C significantly limits the influence of outliers, whereas a large value of C give heavy penalty on the errors made by the hyperplane which may lead to a discriminant hyperplane biased to the outliers. Therefore, appropriate selection of C is of great importance and it is still an ongoing research topic. Typically, the parameter C is selected using the cross-validation procedure although other methods have also been discussed (Chapelle et al., 2002; Keerthi, 2002; Lee and Lin, 2000).

Using the similar strategy of introducing Lagrandian multipliers to Equations (2.21) and (2.22), the dual form of soft margin SVM can be obtained as

$$\begin{aligned} &\max J(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}}[\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j], \\ &\text{subject to} \quad \sum_{i=1}^{N} y_i \alpha_i = 0; \\ &\qquad\qquad\quad 0 \leq \alpha_i \leq C \quad \forall i. \end{aligned} \tag{2.24}$$

It is again a quadratic programming problem. Comparing this with Equation (2.20), it is clear that the only difference is the new constraint of $0 \leq \alpha_i \leq C$ (replacing the previous one of $\alpha_i \geq 0$). The change of the constraints has no significant implication on the method that solves the quadratic programming problem.

While the SVM stemmed from of the idea of (soft) margin classifier as described above, it was the idea of *kernel trick* that popularized the SVM. The idea of kernel trick offers an alternative solution to approximate any nonlinear discriminant function in original feature space (the input space) by nonlinearly projecting the data into a high (possibly infinite) dimensional feature space, using an appropriate nonlinear mapping function.

The key success of kernel trick lies in the special types of mapping functions that obey Mercer's theorem (also called reproducing kernel Hillbert spaces). These mapping functions offer an implicit mapping that maps the original feature vector $\mathbf{x} \in \mathbb{R}^D$ into a high (possibly infinite) dimensional Hillbert feature space, $\mathbb{H}$, using a nonlinear mapping function Φ:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2), \qquad \mathbb{R}^D \xrightarrow{\Phi} \mathbb{H}. \tag{2.25}$$

This means the explicit mapping needs not be known or calculated; rather the cheap inner product is sufficient to provide the mapping. Further, this means that the input feature inner product can simply be substituted with the appropriate kernel function to obtain the nonlinear SVM formulation that is capable of approximating any complicated nonlinear discriminant functions in the input space, without involving any other changes. In this way, all the benefits of the original linear SVM method are maintained, yet the use of kernel trick transforms a simple linear classifier into a powerful nonlinear classifier.

Hence, the primal form of the nonlinear SVM is

$$\min J(\mathbf{w}, b, \boldsymbol{\xi}) = \min \left[\frac{1}{2} ||\mathbf{w}||^2 + C \sum_{i=1}^{N} \xi_i \right], \tag{2.26}$$

$$\text{subject to} \quad y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) - 1 + \xi_i \geq 0, \ i = 1, \cdots, N, \text{ and} \tag{2.27}$$

$$\xi_i \geq 0, i = 1, \cdots, N, \tag{2.28}$$

and the corresponding dual form is

$$\begin{aligned} \max J(\boldsymbol{\alpha}) &= \max_{\boldsymbol{\alpha}} [\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)], \\ \text{subject to} \quad & \sum_{i=1}^{N} y_i \alpha_i = 0; \\ & 0 \leq \alpha_i \leq C \quad \forall i. \end{aligned} \tag{2.29}$$

Popular kernel functions used in SVM include

$$\text{Polynomial:} \qquad K(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 \cdot \mathbf{x}_2 + a)^b, \text{ for } a = 0 \text{ or } 1, b > 1; \tag{2.30}$$

$$\text{Gaussian:} \qquad K(\mathbf{x}_1, \mathbf{x}_2) = exp\left[-\gamma(-||\mathbf{x}_1 - \mathbf{x}_2||^2)\right] \text{ for } \gamma > 0; \tag{2.31}$$

$$\text{Sigmoid:} \qquad K(\mathbf{x}_1, \mathbf{x}_2) = \tanh(a\mathbf{x}_1 \cdot \mathbf{x}_2 + b), \text{ for some } a > 0 \text{ and } b > 0. \tag{2.32}$$

After the machine training, for a given unseen feature vector (representing a test pattern), the trained SVM outputs its predicted class label (-1 or +1) based on the half space (defined by the hyperplane) into which that feature vector falls, by the following output function

$$f(\mathbf{x}) = \sum_{i=1}^{N} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b, \tag{2.33}$$

and the decision function

$$d(\mathbf{x}) = sgn\left[\sum_{i=1}^{N} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right]. \tag{2.34}$$

The hyper-parameters of SVM, C and others (for example, the γ in the Gaussian kernel), are typically selected through a grid-search via a cross-validation procedure. The optimal values for these hyper-parameters that produce the highest cross-validation accuracy are used in the final training of the SVM. Such model selection also provides significant immunity to the overfitting problem.

2.10.2.2 Platt's Probabilistic Outputs for SVM

Standard SVM classifies a sample $\mathbf{x}$ depending on the sign of $f(\mathbf{x})$, or the half space in $\mathbb{H}$ into which $\Phi(\mathbf{x})$ falls. Such an approach, however, ignores the relative confidence in the classification, or the distance $\Phi(\mathbf{x})$ is from the separating hyperplane. Platt (2000) addressed this shortcoming through the use of a sigmoid function and mapped $f(\mathbf{x})$ into $p(\omega|\mathbf{x})$ (i.e. the probability of belonging to the class ω given $\mathbf{x}$), providing probabilistic information from standard SVM output. Here, $\omega = +1$ or $\omega = -1$ for class $+1$ and class -1, respectively. The benefit of $p(\omega|\mathbf{x})$ over $f(\mathbf{x})$ in improving classification accuracy has been demonstrated on several numerical experiments in the domain of machine learning (Duan and Keerthi, 2005; Platt, 2000), but it has rarely been studied in real-life applications.

Suppose N_+ and N_- are the numbers of positive ($y = +1$) and negative ($y = -1$) samples respectively in the dataset $\mathbb{D}$. The Platts probability output is

$$\hat{p}(\omega|\mathbf{x}) = \frac{1}{1+\exp[(Af(\mathbf{x})+B)]}, \tag{2.35}$$

where $f(\mathbf{x})$ is the SVM output given by Equation (2.33) and the parameters A and B are obtained from minimizing the negative log likelihood (or the cross-entropy error function) of $\mathbb{D}$ in the form of

$$\min F(A,B) = \min\{-\sum_i [t_k \log \hat{p}(\omega|\mathbf{x}) + (1-t_k)\log(1-\hat{p}(\omega|\mathbf{x}))]\}, \tag{2.36}$$

where $t_k = (N_+ + 1)/(N_+ + 2)$ if $y_k = +1$ and $t_k = 1/(N_- + 2)$ if $y_k = -1$. It needs to be noted that Lin et al. (2003) have suggested some modifications for numerical stability in obtaining the above Platt's probabilistic output. Hereafter, $\hat{p}(\omega|\mathbf{x})$ refers to the estimated posterior probability of belonging class ω given $\mathbf{x}$ obtained from Equations (2.36) and (2.36), while $p(\omega|\mathbf{x})$ refers to the true but typically unknown posterior probability of belonging to class ω given $\mathbf{x}$. They will be used extensively in the subsequent chapters.

2.10.2.3 Multi-Class SVM

The SVM was originally designed for two-class classification. Several researchers have studied the extension of the two-class SVM to multi-class classification, although it has rarely been discussed and used in biomedical signal processing.

Consider a prototypical multi-class classification problem having c classes (ω_1, ω_2, $\cdots$, ω_c) and a given dataset $\mathbb{D}$ in the form of $\{\mathbf{x}_i, y_i\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^D$ is the i^{th} sample, $y_i \in \{1, \cdots, c\}$ is the corresponding class label. Hence, $y_i = k$ if and only if $\mathbf{x}_i \in \omega_k$. Let N_k be the number of samples that belong to class ω_k, $N := N_1 + \cdots + N_c$ be the total number of samples in $\mathbb{D}$ and $\mathbb{D}_{ij} := \{\mathbf{x}_k, y_k\}_{\mathbf{x}_k \in \omega_i \cup \omega_j}$ be the subset of $\mathbb{D}$ formed by samples from classes ω_i and ω_j. In addition, let $p_i(\mathbf{x}) \equiv P(\omega_i|\mathbf{x})$ denote the posterior probability of belonging to class ω_i given $\mathbf{x}$ and let $\hat{p}_i(\mathbf{x})$ denote its estimate. Similarly, $p_{ij}(\mathbf{x}) \equiv P(\omega_i|\mathbf{x}, \mathbf{x} \in \omega_i \text{ or } \omega_j)$ refers to the pairwise probability of belonging to class ω_i knowing that $\mathbf{x}$ is from class ω_i or class ω_j and $\hat{p}_{ij}(\mathbf{x})$ is its estimate.

A standard multi-class SVM (for multi-class classification problems) is usually implemented by combining several two-class SVMs. The most popular standard multi-class SVM is the “one-versus-one” SVM (OVO-SVM). The final classification is based on voting by all the pair-wise two-class SVMs. Specifically, for a given test feature vector, count the times that each class wins in all these pair-wise classifications and choose the class that wins most as the class for that test feature vector. Besides OVO-SVM, other forms of standard multi-class SVM also exist, such as the “one-versus-all” SVM (OVA-SVM) and various error-correction schemes. They follow similar principle as OVO-SVM and perform similarly (Hsu and Lin, 2002; Rifkin and Klautau, 2004).

2.10.2.4 Probabilistic Multi-Class SVM

Duan and Keerthi (2005) have reported an interesting study, showing a probabilistic-based multi-class SVM performs significantly better than the standard multi-class SVMs (such as OVO-SVM and OVA-SVM).

The probabilistic-based multi-class SVM is based on the estimations of posterior probabilities using SVM and it has the decision function in the form of

$$d(\mathbf{x}) = \arg\max_i \{p_i(\mathbf{x})\}. \tag{2.37}$$

Typically, $p_i(\mathbf{x})$ is estimated by $\hat{p}_i(\mathbf{x})$, obtained from solving the following pairwise-coupling (PWC) optimization problem (Hastie and Tibshirani, 1998; Wu et al., 2004):

$$\min_{\hat{p}_i(\mathbf{x})} \sum_{i=1}^{c} \sum_{j:j\neq i} \left[\hat{p}_{ji}(\mathbf{x})\hat{p}_i(\mathbf{x}) - \hat{p}_{ij}(\mathbf{x})\hat{p}_j(\mathbf{x})\right]^2, \text{ subject to} \sum_{i=1}^{c} \hat{p}_i(\mathbf{x}) = 1. \tag{2.38}$$

where $\hat{p}_{ij}(\mathbf{x}), \hat{p}_{ji}(\mathbf{x})$ are known Platt's probabilistic outputs of the two-class SVM classifiers (Platt, 2000) as discussed in Section 2.10.2.2. Specifically, suppose the standard output of the two-class SVM trained using $\mathbb{D}_{ij}$ is

$$f_{ij}(\mathbf{x}) = \sum_{\mathbf{x}_k \in \mathbb{D}_{ij}} y_k \alpha_k K(\mathbf{x}_k, \mathbf{x}). \tag{2.39}$$

The probabilistic SVM output, $\hat{p}_{ij}(\mathbf{x})$, is

$$\hat{p}_{ij}(\mathbf{x}) = \frac{1}{1 + \exp\left(A_{ij} f_{ij}(\mathbf{x}) + B_{ij}\right)}, \tag{2.40}$$

where the parameters A_{ij} and B_{ij} are determined from minimizing the negative log likelihood (or the cross-entropy error function) function, or

$$\min F(A_{ij}, B_{ij}) = \min\{-\sum_{\mathbf{x}_k \in \mathbb{D}_{ij}} [t_k \log \hat{p}_{ij}(\mathbf{x}_k) + (1-t_k)\log(1-\hat{p}_{ij}(\mathbf{x}_k))]\}, \tag{2.41}$$

where $t_k = (N_i+1)/(N_i+2)$ if $y_k = i$ and $t_k = 1/(N_j+2)$ if $y_k = j$. It is worth noting that a cross-validation process was suggested by (Platt, 2000) to remove the requirement of keeping a hold-out validation dataset for fitting the parameters A_{ij} and B_{ij}, which is especially useful when the number of training samples is small.

The above procedure of obtaining $\hat{p}_i(\mathbf{x})$ from $\hat{p}_{ij}(\mathbf{x})$ is hereafter referred as PWC-PSVM. Both quantities, $\hat{p}_i(\mathbf{x})$ from Equation (6.2) and $\hat{p}_{ij}(\mathbf{x})$ from Equation (6.4) are used extensively in the subsequent chapters.

2.10.2.5 The Weighted SVM for Unbalanced Problem

Unbalanced problem refers to the scenario where one class is very much under-represented in the data (with relatively few samples). This is a common and challenging problem that the machine-learning researchers have to tackle in real-world applications, especially in many pattern-recognition applications using bio-medical signals. For example, consider the problem of automatic detection of a certain disease where cases of that disease in a very large population are perhaps less than 1%. In such circumstance, if we build a model in the usual way where the aim is to minimize the error rate, this may easily lead to a biased classifier to say that there is no cases of the disease. The accuracy of classifier is up to 99%, but of little use.

Classification of a unbalanced dataset often involves adjustments to the modelling in some way. One conventional approach is to down-sample the majority class to even up the classes. Alternatively, one might over-sample the rare class and as such increase the weight of the minority. Such conventional approaches can work, but it is not always clear whether they will and how much they can help. Under-sampling can lead to a loss of information, whilst over-sampling may lead to over-fitting.

The modified SVM, so-called the weighted SVM (Osuna et al., 1997), provides a promising alternative to deal with the unbalanced problem. Conceptually, the weighted SVM is to impose higher penalty on the classification errors made on the samples from the minority class. This can be better explained by the following primal and dual formulations of the weighted SVM:

$$\begin{aligned}
\min J(\mathbf{w}, b, \boldsymbol{\xi}) &= \min \left[\frac{1}{2}||\mathbf{w}||^2 + C_- \sum_{y_i=-1} \xi_i + C_+ \sum_{y_i=+1} \xi_i \right], \\
\text{subject to} \quad & y_i(\mathbf{w}\cdot\Phi(\mathbf{x}_i)+b)-1+\xi_i \geq 0,\ i=1,\cdots,N,\ \text{and} \\
& \xi_i \geq 0, i=1,\cdots,N,
\end{aligned} \tag{2.42}$$

and the corresponding dual form is

$$\begin{aligned}
\max J(\boldsymbol{\alpha}) &= \max_{\boldsymbol{\alpha}} [\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)], \\
\text{subject to} \quad & \sum_{i=1}^{N} y_i \alpha_i = 0; \\
& 0 \leq \alpha_i \leq C_- \text{ if } y_i = -1, \quad 0 \leq \alpha_i \leq C_+ \text{ otherwise,}
\end{aligned} \tag{2.43}$$

where C_- and C_+ are different regularization parameters for the negative class and positive class, respectively. A useful choice (Eitrich and Lang, 2006; Osuna et al., 1997) is to let

$$\frac{C_-}{C_+} = \frac{N_+}{N_-}. \tag{2.44}$$

In this way, the weighted SVM deals with the unbalanced problem within the well-defined framework, the misclassification cost. In this setting, by imposing a very high cost on misclassification on the minority class, the aim of the weighted SVM is to minimize the misclassification cost, rather than to minimize the error rate for the case of the conventional SVM.

Chapter 3

Proposed Research Approach and Data Collection

This doctoral research work is concerned with developing novel signal processing methods that enable automatically measuring and monitoring mental fatigue in individuals from their EEG recordings. The work is of great interest in traffic safety and other domains where measurement and monitoring of mental fatigue is crucial. Let's begin with the overview of the approach taken in this work.

3.1 Rationale

Despite its clear importance, there is no gold method for mental-fatigue measurement. The conventional mental-fatigue measurement methods can be classified into two categories: subjective and objective measurements. Subjective mental-fatigue measurement methods require subjects to rate their level of mental fatigue either indirectly (e.g. Piper et al., 1998; Zachrisson et al., 2002) or directly (e.g. Shapiro et al., 2002), whereas objective methods assess mental fatigue via quantifying subjects' performance on a specific task (e.g. Dinges and Powell, 1985; Thorne et al., 2005). There is a general agreement that these conventional measurement methods can have good reliability and good validity. However, they cannot be used in some domains, such as transportation industry, where an objective and non-intrusive mental-fatigue measurement method is required.

In attempts to develop an objective and non-intrusive mental fatigue measurement method, some pilot studies have correlated mental fatigue with physiological measures such as electrocardiogram (ECG), electrooculogram (EOG) and EEG. A good review of these approaches can be found in the thesis by Mallis (1999) and a review by Lal and Craig (2001a). More recently, several studies have reported the feasibility of measuring mental fatigue or drowsiness indexed by subject's task performance, based on EEG data in attention-sustained experiments using auditory or visual stimuli (e.g. Duta et al., 2004; Jones, 2006; Jung et al., 1997; Lal et al., 2003; Makeig et al., 2000; Peiris et al., 2004; Sommer et al., 2002; Vuckovic et al., 2002).

Most of these pilot studies have focused on the detection of performance lapses in the specific tasks that they studied (i.e. the prediction of a mistake in a specific task) without measuring subjects' mental-fatigue levels directly. Also, most of these pilot studies used fairly simple linear or nonlinear regression or neural networks and recent advance in the signal processing methods, like automatic artifact removal, feature selection and multi-category pattern classification, have been largely overlooked. More importantly, the literature continues to produce varying and even conflicting results and very little evidence exists on the efficacy of incorporating EEG into a practically-usable automatic mental-fatigue measurement and monitoring system. This is likely due to the challenge of developing effective mathematical framework, signal processing methods and learning algorithms for the analysis of EEG signals in relationship to mental fatigue.

3.2 Approach Taken In This Work

Mental fatigue produces much less distinguishable changes in terms of EEG waveforms than other brain states like sleep (e.g. Kecklund and Åerstedt, 1993). Therefore, the design of an EEG-based mental fatigue measurement and monitoring system, as shown in Fig. 3.1, requires some advanced signal processing methodology to address the interference of artifacts. In addition, good answers are also required for the problem of selecting useful features in relationship to mental fatigue as well as for the problem of classifying these features in relationship to mental fatigue. A stringent experimental design ensuring that the developed system measures a meaningful fatigue-induced EEG change is also crucial.

Fig. 3.1 shows the flowchart of the data-driven approach taken in this work. The approach is (i) to (automatically) remove the artifacts that very much undermine the reliability of the EEG as a physiological indicator of mental fatigue; then (ii) to (automatically) identify the key features in the EEG signals that correlate with mental fatigue in an individual; and finally (iii) to construct an intelligent system that tracks the state of mental fatigue of an individual based on these identified key features.

The present study may serve as a key step towards an EEG-based mental-fatigue monitoring device or an EEG-based mental-fatigue screening system. The EEG-based mental-fatigue monitoring device is particularly useful for preventing fatigue-related driving accidents, where the driver's mental fatigue needs to be continuously monitored, so as to trigger necessary countermeasures when the driver becomes too fatigued to be safe. The EEG-based mental-fatigue screening system is widely demanded in defence where checking the mental fatigue level is one of the most important parameters in the routinely-performed fitness-for-duty screening on those military personnel who perform critical missions (such as cruise duty on a strategic bomber).

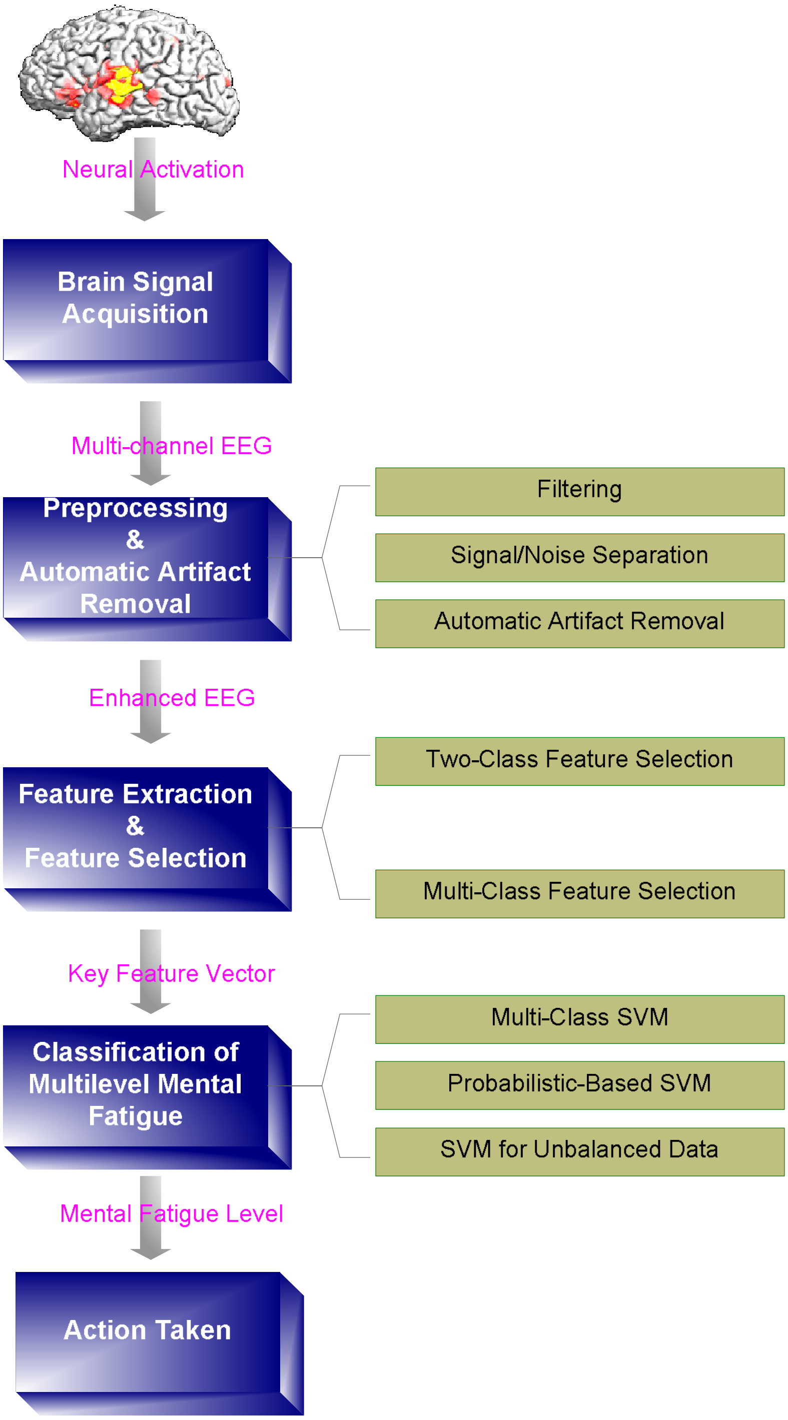


Figure 3.1: Flowchart of the proposed EEG-based mental-fatigue measurement and monitoring system.

The rest of this chapter will discuss the experimental design for collecting the mental-fatigue EEG database that will be used in subsequent chapters to train the proposed EEG-based mental-fatigue measurement and monitoring system in a supervised regime.

3.3 Experimental Design and Data Collection

Scientific validity and reliability should first be established in a controlled laboratory experiment. The present work used a rigid, albeit costly, controlled laboratory studies, involving a 25-hour sleep deprivation on volunteers and sampling over all circadian phases (or mental-fatigue levels), to ensure that the developed system measures a meaningful fatigue-induced EEG change. This section gives the detailed account of the experimental design used in this work. The resulting mental-fatigue EEG database (each EEG segment with a reliable label of mental-fatigue level) was used, in the subsequent chapters, to train the proposed EEG-based mental-fatigue measurement and monitoring system in a supervised regime.

3.3.1 Mental-Fatigue EEG Experiments

Let's begin with the mental-fatigue EEG experiments involving a 25-hour sleep deprivation on each subject. Mental-fatigue EEG data, sampling over all circadian phases (or mental-fatigue levels) throughout the 25-hour sleep deprivation experiment, were recorded hourly from the subjects. The procedure is as follows.

3.3.1.1 Hardware and software environment

Monopolar EEG data were acquired at a sampling frequency of 167 Hz using a Medtronic PL-Winsor 2.35 EEG system together with a 19-channel electrode cap, according to the international 10-20 system (Jasper, 1958). The EEG data were pre-filtered by the EEG system through its integrated low-pass filter (cut-off frequency at 35 Hz) and high-pass filter (cut-off frequency at 0.1 Hz) as well as a 50 Hz notch filter.

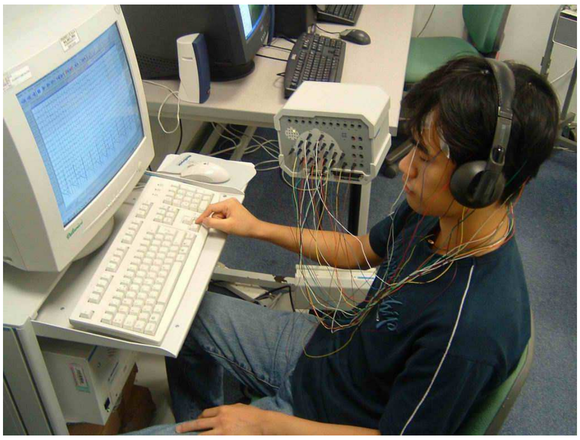


Figure 3.2: The experiment set-up for mental-fatigue EEG database collection.

3.3.1.2 Subjects

In total, 22 subjects were selected from right-handed volunteers of local tertiary institutions who fulfilled the inclusion criteria of not being on any medication, no history of sleep disorders and with regular sleep hygiene as evidenced by a one-week sleep diary prior to the experiment. The recruitment of human subjects for this study was approved by the Institutional Review Board of the National University of Singapore. Informed consents were obtained and nominal monetary incentives sufficient to cover transportation costs were given for their participation.

3.3.1.3 Procedure

Each subject underwent a 25-h sleep-deprivation experiment in a temperature-controlled laboratory (23–25 °C) from 8:30 am to 9:30 am next day. Caffeine, tea, smoking were prohibited for about two days (from one-day before the experiment till the end of experiment). Subjects were required to perform an auditory working-memory vigilance task (AWVT) session once an hour throughout the experiment (with eyes open) and they were allowed to engage in non-strenuous activities in non-AWVT-session period. EEG data were recorded simultaneously during every AWVT session and they were labeled to 5-level mental fatigue according to the subject's performance in AWVT. The details of such labeling of the mental-fatigue EEG data using AWVT is given in the next section.

3.3.2 Labeling of Mental-Fatigue EEG

In order to train the proposed system in a supervised regime, an AWVT was proposed to label the recorded mental-fatigue EEG resulting from the above 25-hour mental-fatigue EEG experiments. Specifically, according to a subject's performance on AWVT, the EEG data, collected hourly throughout his/her 25-hour sleep-deprivation experiment, were manually classified into mental fatigue at 5 levels.

3.3.2.1 Why AWVT?

The simplest measure of mental fatigue can be a subjective self-report measure, such as a the Visual Analogue Scale, Stanford Sleepiness Scale, Epworth Sleepiness Scale, Karolinska Sleepiness Scale and the recent Situational Fatigue Scale (Yang and Wu, 2005). Here a person is made to rate his current state about his own assessment of mental fatigue. These scales though easy to administer, have shown that many times a person is not the best judge of how much his functioning capacity is compromised due to mental fatigue (Frey et al., 2004). Some of these tests are very detailed and require the subject to estimate their level of fatigue if they were in specific situations (like watching TV, shopping, etc), for a certain period. Though these scales claim to have good results in assessing fatigue, but it raises the question whether such situations as ”watching TV”, would have the same effect on all people, without depending on the nature of what they were watching on TV. These measures could have many psychological factors affecting the estimate, and thus they may not be so accurate. This kind of subjective self-report measures have given rise to performance measures which are objective and mostly free from drawbacks of self-report measures.

Subjective mental fatigue measurement methods require subjects to rate their level of mental fatigue either indirectly (e.g. Piper et al., 1998; Zachrisson et al., 2002) or directly (e.g. Shapiro et al., 2002), whereas objective methods assess mental fatigue via quantifying subjects performance on a specific task (e.g. Dinges and Powell, 1985; Thorne et al., 2005). One of the most commonly used objective measure of mental fatigue is the Psychomotor Vigilance Task (PVT), developed by Dinges and Powell (1985). In this task, a visual stimulus is given and the subject has to press a response button as soon as possible. Many studies have shown that the performance on this task shows an increase in reaction time as the mental fatigue increases. The popularity of the PVT is mainly because it is a simple task, which could be easily performed by anyone, not depending on the aptitude and education of the person. The Walter Reed Army Institute of Research has developed a PDA based PVT, known as PalmPVT, which has also shown close correspondence with the original PVT in terms of results when used in sleep deprivation studies (Ferguson et al., 2005; Lamond et al., 2005; Thorne et al., 2005).

A task like PVT is useful to detect arousal of a person, and how it changes with the progress of mental fatigue. Arousal is defined as the general readiness of the nervous system to respond to a novel stimulus. It is seen that arousal mainly affects the reaction time in a vigilance task, while the accuracy does not only depend on arousal (Tassi et al., 2003). This kind of a task can be an accurate measure in situations where people are required to do monotonous tasks like electronic component assembling. On the other hand, it may not be such a good measure in situations where the job demands are more complex and thus the higher faculty of brain functions are required.

Mental fatigue has a variety of effects on the functioning of the brain. The most extensively studied effect is that on sustained attention or vigilance. It is seen that the reaction time in a vigilance task like the PVT, is directly proportional to mental fatigue (Dinges and Powell, 1985; Pack et al., 2006; Rogers et al., 2003). Other faculties like working memory, judgment and decision making are also affected as fatigue progresses (Cajochen et al., 2004; Staal, 2004). It has even been shown that after a certain level of sleep deprivation, the performance deterioration is similar to that caused by raised levels of alcohol in the body (Arnedt et al., 2001; Dawson and Reid, 1997; Lamond and Dawson, 1999).

Though the PVT is a simple task that gives a fairly accurate measure of sleepiness in individuals, one can wonder about how many tasks in real life situations require a reaction time type response from individuals. Is a driver just required to press a button or pedal when he sees an obstacle in front, or is he also required to make a decision on which pedal to press, or which side to steer the car to avoid the pending danger? In practical situations such kind of decision-making goes hand in hand with a timely response, and thus the outcome in such a situation not only depends on how fast a person responds, but how adequately he decides to take corrective action. Similarly working memory is used to keep relevant information in the mind, like speed limits while driving. The burden on working memory is more for pilots, who have to take many variables in consideration from various instruments, to make their decisions. These basic functions of working memory and decision making thus form an important foundation on which the functioning of a person depends in real life conditions. It has been seen in studies that these functions also deteriorate due to sleep deprivation, apart from the decrease in vigilance capacity.

As pointed out by Rogers et al. (2003), working memory and attention are important factors that may influence the ability to perform neurobehavioral tasks and determine the efficiency of neurocognitive functioning. Similarly, according to a report on a study of Canadian Marine pilots, the tasks rated by pilots to be most affected by fatigue were decision-making, attention, remaining awake and reaction time (Rhodes and Gil, 2002). In addition, the study also pointed out that mental fatigue led to decreased performance on memory tasks. Other studies too have pointed out these same deficits. Ferguson et al. (2005) asserted that slowed reaction time, impaired decision making, memory difficulties and vigilance decrements are caused by mental fatigue. Various studies on the changes in working memory due to sleep deprivation have shown that the performance deteriorates (Caldwell et al., 2004; Chee and Choo, 2004; Murphy and Delanty, 2007; Smith et al., 2002). Similarly executive functioning and decision-making are also degraded due to mental fatigue (Bruck and Pisani, 1999; Killgore et al., 2006; Neri et al., 1992; Nilsson et al., 2005; Raaijmakers, 1990).

There are also many different tasks that assess different functions of the brain (Bonnet and Arand, 1999; Griffin and Koonce, 1996; Wilson, 2002; Wilson et al., 2007). These tasks usually take a long time, or are required to be done by professionals or operators. Objective ways to find the propensity to fall asleep as measured by Multiple Sleep Latency Test and a measure of alertness like Maintenance of Wakefulness Test are also lab based (Bonnet and Arand, 1999).

The need exists for such tasks that can be used independently by subjects in real work environments, with minimum hindrance to their daily work routine, without the need for operators or observers. Such a task, which also incorporates other higher faculties of brain functions in addition to reaction time, will be a better test of a person's capabilities to perform in a real world situation. Thus, a reaction time task is modified to introduce the elements of working memory and decision making, so that a more realistic way of measuring the performance decrements due to mental fatigue can be developed.

3.3.2.2 Characteristics of An Ideal Objective Performance Task

Considering the wide range of functions that deteriorate due to sleep deprivation, there is a need for a comprehensive task, able to obtain a more realistic measure of a person's ability to perform his duties safely. An ideal task, for wide use should have the following characteristics:

(a) It should accurately measure mental fatigue.

(b) The amount of skill involved in doing the task should be absent or minimal. This will make the task suitable for use without depending on the aptitude or education of a person.

(c) It should have minimum dependence on motivation and other psychological factors, so that it can be more reliable for repeated use.

(d) The task should be comprehensive - measuring a range of human faculties (like working memory, decision making, attention), so that it can be a more realistic test of the ability of a person to do his job safely.

(e) The task should be sensitive enough to test the functioning capability of a person in a short time, so that it causes minimal disruption of the normal routine of a person.

(f) The task should measure those functions that are used in most real life scenarios, so that it can be used in different situations and have wide application.

(g) The task should be reliable and consistent.

3.3.2.3 The AWVT

The auditory working-memory vigilance task is our attempt to measure mental fatigue in a more realistic way, in which the task performance involves higher mental abilities of a person, like working memory and decision making, in addition to vigilance.

Description: The AWVT can be seen as a variation of the Serial Choice Vigilance Task (SVT), using auditory stimulus. In the usual SVT, there are multiple stimulus types, and multiple response buttons corresponding to those stimuli. At a particular time, one of the stimuli is given, and the person has to decide which response button to press, keeping in mind, which button corresponds to that particular stimulus. This appearance of stimuli is random and continues for a fixed number of trials or a fixed period. This kind of task introduces the decision-making element to the simple reaction time task, like PVT.

In AWVT, stimulus-delivering software was used on a PC. Programming was done to deliver a set of four direction commands (left, right, up, down), in random order, every 5 seconds. Each direction command, within a set, was given at 500 ms interval. Subjects were required to constantly concentrate and, after the command set is given through headphones connected to the PC, to press the pre-specified buttons, within 1.5 s after each complete command set, in the order of commands that they heard. This gave the subject enough time to press the response keys, but still not giving too much time to relax. The period between the stimuli set was fixed, so there was no variation in foreperiod, like in the PVT and PalmPVT. The software creates text files showing the sequence of events (both stimulus command and response) and their time stamps, with millisecond resolution. There is no feedback system in the program to show the reaction time after a trial. The reaction times and errors were calculated separately later from the collected text files. An error is one where the response sequence is not the same as the given command set.

Mental-Fatigue Scoring Using AWVT: For every AWVT session, an AWVT score was calculated in terms of percentage of correct responses. EEG data recorded simultaneously during the AWVT session were labeled to 5-level mental fatigue according to the subject's AWVT performance. Specifically, for each subject, his/her individual performance span (the highest AWVT score to the lowest AWVT score) was evenly divided into five segments corresponding to fatigue level $1, \cdots, 5$, respectively. The label (i.e. mental fatigue at 5 levels) of the EEG data for an AVT session was determined by which segment the corresponding AWVT performance score fell into.

The AWVT differs from other simple reaction time tasks like PVT in that it introduces the decision-making element so as to get a more realistic measure of a person's cognitive performance impairment due to mental fatigue. Beside this unique property, standard measures, such as the learning curve, test-retest reliability and within-subject consistency, also show that the AWVT offers as good classification of mental fatigue, if not better, than the other simple reaction time tasks like PVT:

(a) The AWVT appears to have a similar learning curve to the PVT. The learning curve for a task shows how many trials it will take for the performance on the task of a subject to become saturated (defined as less than ten percent change in consecutive trials). It captures the subject's learning effect on the task performance. A shorter learning curve is usually preferred due to the concern of potential compounding of the learning effect (extraneous factor, i.e. the noise) and the underlying brain functional impairment due to mental fatigue (factor to be measured, i.e. the signal). The study of the learning curve of AWVT on randomly-chosen 5 subjects in our pilot study shows that the AWVT performance score saturated after 1-3 trials, indicating a similar learning curve to PVT.

(b) The AWVT appears to have a better test-retest reliability than PVT. The test-retest reliability is to measure the reproducibility of the circadian rhythms throughout the 25-hour sleep deprivation of the same subject on different experiment days. Table 3.1 shows the test-retest Pearson's correlation of the randomly-chosen 5 subjects. It is worth noting that the higher correlation between different experiment days on the same subject indicates the higher test-retest reliability.

(c) The AWVT also demonstrated better within-subject consistency than PVT. Both AWVT and PVT offers a relative measure of mental fatigue, benchmarked against the maximum and minimum values over a 25-hour sleep-deprivation experiment. This requires that the highest and lowest performance scores for a particular subject should not vary much, when the task is done on different days. Thus, the consistency in maxima and minima of the task performance score over different experiment days is a good indicator of the within-subject consistency of that task. The test-retest studies on the randomly-chosen 5 subject show that, between original and repeat studies, for each particular subject, the percentage change in the maxima and minima values of AWVT scores ranged from 0.8% to 8.3% (Mean 3.96%; SD 2.67%), while the percentage change for PVT lapses ranged from 22.2% to 36.3% (Mean 27.1%; SD 8%): a strong indicator of the better within-subject consistency of AWVT.

Table 3.1: Pearsons correlation values between initial and repeat trials on five subjects for AWVT performance score and PVT lapses. The higher correlation indicates the higher test-retest reliability.

	sub1	sub2	sub3	sub4	sub5
AWVT	0.69	0.88	0.69	0.76	0.71
PalmPVT Lapses	0.60	0.77	0.29	0.51	0.60

3.4 Concluding Remarks

As a result of the 25-hour sleep deprivation experiments, a relative large database of mental fatigue EEG (with reliable labels of mental fatigue levels), collected from 22 subjects (each underwent a 25-hour sleep deprivation), was established. As it will be seen in the subsequent chapters, part of the resulting mental-fatigue EEG database (with corresponding labels of mental fatigue level given by auditory working-memory vigilance task (AWVT)) were used to train the proposed EEG-based mental-fatigue measurement and monitoring system as shown in Fig. 3.1. The trained system was then tested on unlabeled EEG data and subsequently checked for concordance with the manual classification by AWVT.

Chapter 4

Weighted SVM with Error Correction for Automatic EEG Artifact Removal

The presence of artifacts, such as eye blinks, in electroencephalographic (EEG) recordings obscures the underlying signals and makes EEG analysis difficult. Large amounts of EEG data must often be discarded because of contamination by eye blinks, muscle activity, line noise, and pulse signals. In this chapter, a novel automatic EEG artifact removal method based on independent-component-analysis (ICA) is presented. In this method, no EEG data are discarded. In stead, artifacts are automatically identified and subsequently removed after they are decomposed from raw EEG data by ICA. Comparing with past methods, the proposed method has two unique features: 1) a weighted version of Support Vector Machine (SVM) formulation that handles the inherent unbalanced nature of component classification; 2) ability to accommodate structural information typically found in component classification. The advantages of the proposed method are demonstrated on real-life EEG recordings with comparisons made with several benchmark methods in the literature. Results show that the proposed method is preferable to the other methods in the context of artifact removal by achieving a better tradeoff between removing artifacts and preserving inherent brain activities. Qualitative evaluation of the reconstructed EEG epochs also demonstrates that after artifact removal inherent brain activities are largely preserved.

4.1 Introduction

Electroencephalogram recordings are known to be contaminated by physiological artifacts from various sources, such as blinking and movements of the eyes, beating of the heart and movements of other muscle groups (e.g. Jung et al., 2000a). These artifacts are mixed together with the brain signals, making interpretation of the EEG signals difficult (Urrestarazu et al., 2004).

Artifacts in EEG are commonly handled by discarding the affected segments of EEG. The simplest approach is to discard a fixed length segment (usually 1-3 seconds) from the time an artifact is detected. The recognition of some artifacts, like eye blink artifacts, are generally effected by detecting a voltage exceeding a threshold (usually 100 μV) in separate EOG channel. Other artifacts are generally ignored or manually marked by a EEG practitioner and then manually discarded. Discarding segments of EEG data with artifacts can greatly decrease the amount of data available for analysis.

In recent years, there has been increasing interest in applying independent-component-analysis (ICA) (Common, 1994; Hyvarinen, 2000; Jutten and Herault, 1991) to artifact removal in EEG (Castellanos and Makarov, 2006; Jung et al., 1998, 2000a,b; Makeig et al., 1996; Urrestarazu et al., 2004; Vigário et al., 2000; Vigário, 1997; Wallstrom et al., 2004). This is mainly motivated by the fact that ICA is effective in decomposing raw EEG recordings into artifactual and non-artifactual independent components (ICs) (e.g. Castellanos and Makarov, 2006; Jung et al., 1998; Vigário et al., 2000). Non-artifactual ICs represent signals from brain activations while artifactual ICs represent electrical signals originating from non-cerebral artifacts. In conventional ICA-based artifact removal methods (e.g. Makeig et al., 1996; Urrestarazu et al., 2004; Vigário et al., 2000), artifactual ICs are manually identified (usually by visual inspection) and removed. This process is very time-consuming and, hence, not suitable for real-time applications. Recent effort towards automatic artifact removal includes the pilot work by Nicolaou and Nasuto (2004) and Shoker et al. (2005) where a standard SVM (Boser et al., 1992; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Vapnik, 1995, 1998) trained on equal number of artifactual and non-artifactual samples, is used for automatic identification of artifactual ICs.

Such a combination of ICA and SVM offers a promising approach for automatic artifact removal. Unfortunately, unique properties of the problem at hand have not been taken into consideration. First, the real data is extremely unbalanced-only a few of the ICs are artifactual ICs and the majority is non-artifactual ICs (e.g. Castellanos and Makarov, 2006; Joyce et al., 2004; Jung et al., 2001; Onton et al., 2006; Romero et al., 2003). It is well known in the machine learning community that the performance of a standard SVM, trained on balanced dataset, may perform poorly when the real data is unbalanced. Second, the number of artifactual ICs responsible for each type of artifact, decomposed from a given EEG epoch, is often small. This structural information of the underlying data can be very useful for improving the accuracy of automatic artifact identification. To the best of our knowledge, such structural information has however not been exploited in past literature.

This chapter shows a formulation that exploits the above-mentioned unique properties by: 1) using weighted SVM (Osuna et al., 1997) to handle the unbalanced data, and 2) imposing constraints on the number of artifactual ICs through a novel error correction algorithm. It is worth noting that the proposed formulation is conceptually different from past ICA-based artifact removal methods. It considers all the ICs derived from a given EEG epoch collectively while past methods treat each IC independently. In a

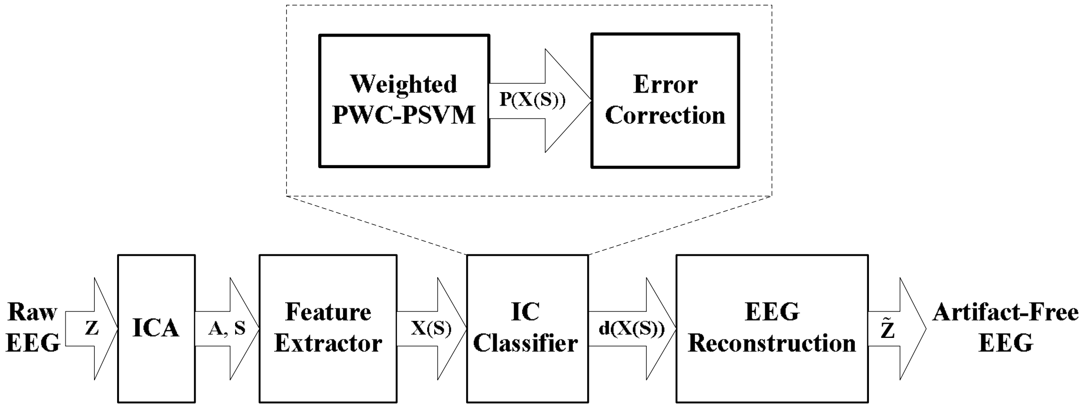


Figure 4.1: Block diagram of the proposed ICA-based automatic artifact removal system. The system consists of four main modules: ICA, feature extractor, IC classifier and EEG reconstruction module. The novelty of the proposed IC classifier is explicitly shown. It has two sub-modules: a modified probabilistic multi-class SVM to address the unbalance nature of the data and an error correction block to handle the unique structural information of the data.

carefully controlled experiment using real-life EEG data, the proposed method shows significant performance advance as compared with a number of past methods in the literature.

4.2 Overview of the Proposed Artifact Removal System

This section provides an overview of the proposed automatic artifact removal system and establishes the necessary notations needed for subsequent sections. Like other ICA-based artifact removal systems in the literature, the proposed system (see Fig. 4.1) consists of four main modules: ICA, feature extractor, IC classifier and EEG reconstruction. The contribution of the present work is mainly on the new method used in the IC classifier, though the feature extractor also includes some new features.

The continuous raw EEG recording is first segmented into epochs with a fixed length. The resulting EEG epochs are then fed, epoch by epoch, into the artifact removal system. Given a raw EEG epoch as the input, the output of the system is the reconstructed artifact-free EEG epoch.

Consider a given n-channel raw EEG epoch, $\mathbf{Z} = [\mathbf{z}_1\ \mathbf{z}_2\ \cdots\ \mathbf{z}_n]^T$ where $\mathbf{z}_i \in \mathbb{R}^l,\ \forall i$, is the time series for the i^{th} EEG channel with a fixed length, l. The ICA module decomposes $\mathbf{Z}$ into m ($\leq n$) ICs, each representing an independent source. Let $\mathbf{S} = [\mathbf{s}_1\ \mathbf{s}_2\ \cdots\ \mathbf{s}_m]^T$ denote the resulting ICs where $\mathbf{s}_i \in \mathbb{R}^l,\ i = 1, \cdots, m$, is the i^{th} IC and $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_m]$ denote the mixing matrix with $\mathbf{a}_i \in \mathbb{R}^n$ containing the scalp distribution coefficients of $\mathbf{s}_i$. While many implementations of ICA are available, the popular FastICA package (Gävert et al., 2005) is used in the present work.

The feature extractor generates a set of feature vectors from each of the $\mathbf{s}_i$'s. Suppose D features are extracted from $\mathbf{s}_i$. Then, $\mathbf{x}(\mathbf{s}_i) = [g_1(\mathbf{s}_i)\ g_2(\mathbf{s}_i)\ \cdots\ g_D(\mathbf{s}_i)]^T$ denotes the feature vector extracted from $\mathbf{s}_i$ and $\mathbf{X}(\mathbf{S}) = [\mathbf{x}(\mathbf{s}_1)\ \mathbf{x}(\mathbf{s}_2)\ \cdots\ \mathbf{x}(\mathbf{s}_m)]^T$ denotes the set of feature vectors obtained from $\mathbf{S}$.

Suppose the $\mathbf{s}_i$'s are attributed to c different classes $\{\omega_1, \cdots, \omega_c\}$ with ω_1 referring to the class of brain sources and the rest $c-1$ classes, i.e. $\omega_2, \cdots, \omega_c$, referring to different artifactual sources. Standard IC classifier used in the literature (Nicolaou and Nasuto, 2004; Shoker et al., 2005) classifies $\mathbf{s}_i$ into one of c classes, or the decision function $d(\mathbf{x}(\mathbf{s}_i))$ maps $\mathbf{x}(\mathbf{s}_i)$ into $\{\omega_1, \cdots, \omega_c\}$. Such a setup considers $\mathbf{s}_i,\ i = 1, \cdots, m$, independently and is the framework used in most work in the literature. However, it is difficult for such a setup to account for the unique structure of the underlying data. The proposed classifier, as shown in Fig. 4.1, considers the $\mathbf{s}_i$'s collectively and yields all m predicted class labels via the decision function $d(\mathbf{X}(\mathbf{S}))$. Such a setup aims to incorporate structural information of the dataset and address the unbalanced nature of the data.

The proposed $d(\mathbf{X}(\mathbf{S}))$ is based on a modified version of probabilistic multi-class SVM. The choice of SVM stems from its superior performance on many learning problems. Justification to this choice is verified by experimental results compared with other classification approaches, like KNN, Gaussian mixture models (GMM) and linear discriminant function (LDF). It is worth noting that the proposed modification to the probabilistic multi-class SVM to address the unbalanced nature of the data is also a novel contribution of this work. The detailed account of the proposed IC classifier will be given in the next section.

The EEG reconstruction module reconstructs artifact-free EEG epoch by zeroing the contribution of all artifactual sources from raw EEG epoch. Suppose $\tilde{\mathbf{S}}$ is obtained from $\mathbf{S}$ by zeroing all the identified artifactual ICs. The reconstructed artifact-free EEG epoch, denoted by $\tilde{\mathbf{Z}}$, can be obtained as follows: $\tilde{\mathbf{Z}} = \mathbf{A}\tilde{\mathbf{S}}$.

4.3 The Proposed Approach

The proposed IC classifier is a combination of a modified probabilistic multi-class SVM and a novel error correction algorithm. It is our attempt to address the unique properties of the problem. Given m ICs decomposed from an EEG epoch, let m_{ω_i} be the number of ICs corresponding to class ω_i. The unique properties of the problem can be effected in terms of the following constraints:

$$m_{\omega_1} \gg m_{\omega_2}, \quad m_{\omega_1} \gg m_{\omega_3}, \quad \cdots, \quad m_{\omega_1} \gg m_{\omega_c}, \quad and \tag{4.1}$$

$$l_{\omega_2} \leq m_{\omega_2} \leq u_{\omega_2}, \quad l_{\omega_3} \leq m_{\omega_3} \leq u_{\omega_3}, \quad \cdots, \quad l_{\omega_c} \leq m_{\omega_c} \leq u_{\omega_c}. \tag{4.2}$$

The constraints in Equations (4.1) represent the inherent unbalanced nature of the data, while those as in Equations (4.2) are the unique structural information that define the upper and lower bounds, denoted by u_{ω_i} and l_{ω_i}, respectively, with regards to the number of ICs belonging to each type of artifactual source. Optimal values of u_{ω_i} and l_{ω_i} depend on the bioelectrical nature of that artifact (e.g., electrocardiogram (ECG) and electrooculogram (EOG) artifacts generally have three spatial components each) and the protocol under which the EEG data are collected (e.g. the number of EEG channels used). Typically, they can be tuned by a data-driven approach. The details will be given in the description of the numerical experiments in this chapter.

4.3.1 The Modified Probabilistic Multi-Class SVM

A modified probabilistic multi-class SVM is proposed to address the unbalanced nature of the learning problem (as shown in Equations (4.1)). It is modified from a recently-developed probabilistic multi-class SVM proposed by Hastie and Tibshirani (1998), by replacing all the standard binary SVMs with weighted SVMs (Osuna et al., 1997). This modified probabilistic multi-class SVM is hereafter denoted as the weighted PWC-PSVM method. The implementation of the weighted PWC-PSVM involves three major steps as follows.

Let's begin with the general notations needed. Consider a nominal c-class unbalanced classification problem with dataset $\mathbb{D}$ in the form of $\{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$ is the i^{th} sample and $y_i \in \{\omega_1, \cdots, \omega_c\}$ is the corresponding class label and N is the total number of training samples. Let N_i denote the number of training samples belonging to class ω_i, and $\mathbb{D}_{ij} \equiv \{\mathbf{x}_k, y_k\}_{\mathbf{x}_k \in \omega_i \cup \omega_j}$ be the subset of $\mathbb{D}$ formed by the samples from class ω_i and ω_j.

Construction of Weighted Binary SVMs: In total, $c(c-1)/2$ weighted binary SVMs are constructed, each classifying a pair of classes. The weighted binary SVM classifying class ω_i and class ω_j is trained using $\mathbb{D}_{ij}$ by solving the following optimization problem

(Osuna et al., 1997):

$$
\begin{aligned}
\min J(\mathbf{w}, b, \boldsymbol{\xi}) = \min & \left[\frac{1}{2} ||\mathbf{w}||^2 + C_{ij} \sum_{\mathbf{x}_k \in \omega_i} \xi_k + C_{ji} \sum_{\mathbf{x}_k \in \omega_j} \xi_k \right], \\
\text{subject to} \quad & \mathbf{w} \cdot \Phi(\mathbf{x}_i) + b \geq -1 - \xi_k, \quad \text{if } \mathbf{x}_k \in \omega_i, \\
& \mathbf{w} \cdot \Phi(\mathbf{x}_i) + b \leq -1 + \xi_k, \quad \text{if } \mathbf{x}_k \in \omega_j, \text{ and} \\
& \xi_k \geq 0, \forall k,
\end{aligned}
\tag{4.3}
$$

where $\Phi(\cdot)$ is a nonlinear mapping function that maps feature vectors into a high (possibly infinite) dimensional Euclidean space H; $\mathbf{w} \in H$, $b \in \mathbb{R}$ are the parameters that determine the optimal separating hyperplane: $\mathbf{w} \cdot \Phi(\mathbf{x}) + b = 0$; $\xi_k \in \mathbb{R}$ is the non-negative slack variable. Different regularization parameters, C_{ij} and C_{ji}, are introduced for the classes ω_i and ω_j, respectively. A useful choice (Eitrich and Lang, 2006) is to let

$$
\frac{C_{ij}}{C_{ji}} = \frac{N_j}{N_i}. \tag{4.4}
$$

Conceptually, this is to impose higher penalty on the classification errors made on the samples from the minority class.

Practically, the optimization problem of Equation (4.3) is solved using its dual formulation (Boser et al., 1992; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Vapnik, 1995, 1998):

$$
\begin{aligned}
\max J(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} & [\sum_{\mathbf{x}_k \in \mathbb{D}_{ij}} \alpha_k - \frac{1}{2} \sum_{\mathbf{x}_p \in \mathbb{D}_{ij}} \sum_{\mathbf{x}_q \in \mathbb{D}_{ij}} y_p y_q \alpha_p \alpha_q K(\mathbf{x}_p, \mathbf{x}_q)], \\
\text{subject to} \quad & \sum_{\mathbf{x}_k \in \mathbb{D}_{ij}} y_k \alpha_k = 0; \\
& 0 \leq \alpha_k \leq C_{ij} \text{ if } \mathbf{x}_k \in \omega_i, \quad 0 \leq \alpha_k \leq C_{ji} \text{ otherwise,}
\end{aligned}
\tag{4.5}
$$

where α_k is the non-negative Lagrangian multiplier for the k^{th} sample and $K(\mathbf{x}_p, \mathbf{x}_q) =$

$\Phi(\mathbf{x}_p) \cdot \Phi(\mathbf{x}_q)$ is the kernel function. Let's denote the output function of this weighted binary SVM by

$$f_{ij}(\mathbf{x}) = \sum_{\mathbf{x}_k \in \mathbb{D}_{ij}} \alpha_k y_k K(\mathbf{x}_k, \mathbf{x}) + b. \tag{4.6}$$

The choice of kernel function $K(\cdot,\cdot)$ in the above equation is general and our study is done with the popular Gaussian kernel, $K(\mathbf{x}_p, \mathbf{x}_q) = \exp(-\gamma \|\mathbf{x}_p - \mathbf{x}_q\|^2)$ where γ is the kernel parameter.

Generating Pairwise Class Probabilities: Standard SVM classifies a sample $\mathbf{x}$ depending on the sign of $f(\mathbf{x})$, or the half space in H into which $f(\mathbf{x})$ falls. Such an approach, however, ignores the relative confidence in the classification, or the distance that $\mathbf{x}$ is from the separating hyperplane. Platt (2000) proposes the use of the following sigmoid function to map $f_{ij}(\mathbf{x})$ into $p_{ij}(\mathbf{x}) \equiv P(\omega_i | \mathbf{x}, \mathbf{x} \in \omega_i \cup \omega_j)$, the pairwise probability of belonging to class ω_i knowing that $\mathbf{x}$ is from class ω_i or ω_j:

$$p_{ij}(\mathbf{x}) = 1 - p_{ji}(\mathbf{x}) = \frac{1}{1 + \exp[A_{ij} f_{ij}(\mathbf{x}) + B_{ij}]}, \tag{4.7}$$

where the parameters A_{ij} and B_{ij} are determined from minimizing the negative log likelihood function (see Section 2.10.2.2 for details).

It is worth noting that a 5-fold cross-validation procedure is implicitly used in fitting the parameters A_{ij} and B_{ij}, as suggested by Platt (2000). This cross-validation process removes the requirement of keeping a hold-out validation dataset for fitting the parameters A_{ij} and B_{ij}, which is especially useful when the number of training samples is small. Our implementation also includes the modifications suggested by Lin et al. (2003) for numerical stability.

Estimating Multi-class Posterior Probability: Given $p_{ij}(\mathbf{x}), \forall i \neq j$, the multi-class posterior probabilities of belonging to ω_i given $\mathbf{x}$, denoted by $p_i(\mathbf{x}) \equiv P(\omega_i|\mathbf{x}), \forall i$, can be estimated by solving the following optimization problem (Hastie and Tibshirani, 1998; Wu et al., 2004):

$$\min_{p_i(\mathbf{x})} \sum_{i=1}^{c} \sum_{j:j\neq i} \left[p_{ji}(\mathbf{x})p_i(\mathbf{x}) - p_{ij}(\mathbf{x})p_j(\mathbf{x})\right]^2, \text{ subject to} \sum_{i=1}^{c} p_i(\mathbf{x}) = 1. \tag{4.8}$$

Let $\mathbf{p}(\mathbf{x}(\mathbf{s}_i)) = [p_1(\mathbf{x}(\mathbf{s}_i))p_2(\mathbf{x}(\mathbf{s}_i))\cdots p_c(\mathbf{x}(\mathbf{s}_i))]^T$, representing the vector of multi-class posterior probabilities as given by Equation (4.8) for the $\mathbf{s}_i$, the i^{th} IC derived from a given EEG epoch. It will be used in the proposed error correction algorithm as given in the next.

4.3.2 Error Correction

Consider the classification of m feature vectors corresponding to m ICs from a given EEG epoch, $\mathbf{x}(\mathbf{s}_i),\ i = 1,\cdots,m$. Instead of simply using $d(\mathbf{s}_i) = \arg\max_k\{p_k(\mathbf{x}(\mathbf{s}_i))\}$ to classify each IC independently, the proposed IC classifier includes a novel error correction algorithm, $\mathbf{d}(\mathbf{S})$, which aims to incorporate the structural information as given in Equations (4.2) by considering all the $\mathbf{s}_i,\ i = 1,\cdots,m$, collectively and yielding all m predicted class labels simultaneously.

In loose terms, the proposed error correction algorithm is to find the m predicted class labels that satisfy the constraints as in Equations (4.2) and, at the same time, match the $\mathbf{P}(\mathbf{X}(\mathbf{S})) = [\mathbf{p}(\mathbf{x}(\mathbf{s}_1))\ \mathbf{p}(\mathbf{x}(\mathbf{s}_2))\ \cdots\ \mathbf{p}(\mathbf{x}(\mathbf{s}_m))]$ as much as possible.

Let $\mathbf{q}_i \in \mathbb{R}^c$ be the code vector representing the predicted class label for $\mathbf{s}_i$. This implies that, if the predicted class label of $\mathbf{s}_i$ is k, the k^{th} element q_{ik} is equal to one and all the other elements in $\mathbf{q}_i$ are zeros. Then, the proposed error correction algorithm can be formulated into the following mixed integer quadratic problem:

$$\begin{aligned}
&\min_{\mathbf{Q}} \sum_{i=1}^{m} \|\mathbf{q}_i - \mathbf{p}(\mathbf{x}(\mathbf{s}_i))\|, \\
&\text{subject to} \quad q_{ij} = 0 \text{ or } 1, \quad \text{for } i = 1, \cdots, m, \text{ and } j = 1, \cdots, c, \\
&\qquad\qquad\quad \sum_{j=1}^{c} q_{ij} = 1, \\
&\qquad\qquad\quad l_{\omega_2} \le \sum_{i=1}^{m} q_{i2} \le u_{\omega_2}, \quad \cdots, \quad l_{\omega_c} \le \sum_{i=1}^{m} q_{ic} \le u_{\omega_c},
\end{aligned} \tag{4.9}$$

where the optimization is over $Q = [\mathbf{q}_1\ \mathbf{q}_2\ \cdots\ \mathbf{q}_m]^T$.

While various efficient solvers of the above optimization problem are available, the present study uses the solver developed by Bemporad and Mignone (2001). With the solution, $\mathbf{Q}$, the $\mathbf{s}_i$'s ($i = 1, \cdots, m$) are simultaneously classified by

$$\mathbf{d}(\mathbf{X}(\mathbf{S})) = [\arg\max_k \{q_{1k}\} \quad \cdots \quad \arg\max_k \{q_{mk}\}]^T. \tag{4.10}$$

4.4 Numerical Experiments

In numerical experiments, we limited ourselves to the problem of automatic removal of ECG artifact and EOG artifact in real-life EEG. The proposed IC classifier was compared quantitatively with several benchmark methods in a stringent subject-wise cross-validation procedure. In addition, the reconstructed EEG epochs were reviewed by an independent EEG expert to qualitatively evaluate the performance of the proposed artifact removal method.

4.4.1 Data Preparation

Ten right-handed volunteers from local tertiary institutions were selected for EEG measurements. These subjects fulfilled the inclusion criteria of no history of cardio-vascular disease, normal eye sight and with regular eye blinks. Informed consent was obtained and nominal monetary incentive was given for their participation. Multi-channel unipolar EEG data were recorded at about 167 Hz (or 6ms sampling interval) from 17 electrodes (excluding Fp1, Fp2) placed on the scalp according to the International 10-20 system (Jasper, 1958) using the PL-EEG Wavepoint System (Medtronic, Inc. Denmark). Five minutes of EEG data were recorded from each subject with their eyes open and in resting state. The EEG data were bandpass-filtered with cutoff frequencies of 0.02 Hz and 35 Hz, using a customized bandpass filter implemented in LabView (version 6.1, National Instruments, USA).

These EEG recordings were firstly segmented into 12-second epochs ($l = 2000$). Each EEG epoch was then decomposed into ICs by ICA. The ICs were presented to an EEG expert for manual classification independently and in a random order. The EEG expert labeled each IC as EEG IC (class ω_1), ECG IC (class ω_2) or EOG IC (class ω_3). These labels were regarded as “true” labels, against which the performance of IC classifiers was benchmarked.

Six features ($D = 6$) were extracted from each IC and they were used as the chief information source, in place of the IC, for classification. Four features were directly adopted from the literature (Shoker et al., 2005) for characterizing EOG artifacts and two new features were proposed in the present study for characterizing ECG artifacts. The detailed definitions of these features are given in Appendix A. Combining the resulting feature vectors with the “true” labels given by manual classification, a subject-wise data subset of 425 samples (25 epochs × 17 ICs), $\mathbb{D}_k := \{\mathbf{x}_i, y_i\}_{i=1}^{425}$, $k = 1, \cdots, 10$, was obtained for each subject. This relatively large dataset, available separately for each subject, were used to evaluate the proposed IC classifier in a stringent subject-wise cross-validation procedure (see Section).

4.4.2 Parameter Selection

For the proposed IC classifier, two groups of parameters need to be tuned: a) the hyper-parameters for each weighted SVM, i.e. the regularization parameters, C_{ij}, C_{ij} and the kernel parameters, γ_{ij}; b) the lower and upper bounds for each type of artifactual ICs, i.e. u_{ω_j} and l_{ω_j}.

Tuning of hyper-parameters: Since C_{ij} and C_{ji} are connected through Equation (4.4), only one of them needs to be tuned. In the experiments, (C_{ij}, γ_{ij}) were jointly tuned by a 5-fold cross-validation (Muller et al., 2001) using the model selection tool in the LIB-SVM package (Hsu et al., 2004) on the following grids: $[2^{-5}, \cdots, 2^{10}] \times [2^{-10}, \cdots, 2^{3}]$ with a step size of $2^{0.5}$.

Tuning of u_{ω_i} and l_{ω_i}: As mentioned earlier, for ECG and EOG artifacts, they generally have three spatial components each (Schlögl et al., 2007). ICA may output three artifactual ICs corresponding to the three spatial components if high-density EEG recordings (such as 64-channel EEG recording) are used. However, the EEG data used in the present study were recorded from 17 locations in the standard 10-20 system and ICA tended to output less than three artifactual ICs for both ECG and EOG artifacts. In the present experiment, a grid-search, with both u_{ω_i} and l_{ω_i} ranging from 0 to 3 and a search step size of 1, was performed for ECG ICs and EOG ICs respectively to obtain optimal values for u_{ω_i} and l_{ω_i} that gave the highest balanced accuracy.

4.4.3 Quantitative Performance Evaluation

Subject-Wise Cross Validation: To evaluate the performance of the proposed system, a subject-wise resampling scheme was used. Among the data subsets $\{\mathbb{D}_i\}_{i=1}^{10}$ collected from 10 subjects, samples from 9 subjects were used to form a training set $\mathbb{D}_{tra}$, and the samples from the left-out subject were used to form a testing set $\mathbb{D}_{tes}$. Practically, this resampling procedure results in 10 pairs of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$ in total. In the numerical experiments, for each pair of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$, $\mathbb{D}_{tra}$ was used for tuning of parameters and training of SVM. The trained classifier was then tested on left-out dataset $\mathbb{D}_{tes}$. The major advantages of such subject-wise cross-validation procedure include that: a) each testing set is independent of the training set and thus the test error simulates the classifier's generalization performance on other unseen subjects; b) classifier performance obtained on multiple testing sets can be used for evaluating the statistical significance in the performance difference between the two classifiers.

Performance Measures: The following popular measures were used for evaluating the performance of the proposed IC classifier: balanced accuracy, relative classifier information (RCI) (Sindhwani et al., 2001), Cohen's kappa (Cohen, 1960), overall agreement and specific agreement on each class (Hripcsak and Heitjan, 2002). For a given testing set with c classes, let n_{ij} be the number of samples from ω_i (true label) being classified to ω_j by the classifier (predicted label).

a) *Balanced Accuracy:* It is the average accuracy on all classes, i.e. $BA = 1/2\sum_{i=1}^{c}(n_{ii}/\sum_{j=1}^{c} n_{ij}) \times 100\%$.

b) *RCI:* It measures the amount of uncertainty about the class label of an input reduced by a classifier, i.e. $RCI = (H_I - H_O)/H_I \times 100\%$, where H_I and H_O denote the prior and posterior uncertainty about the class of an unseen input respectively. RCI has been shown to be a useful performance measure that captures a detailed picture of classifier performance while being immune to the effect of prior class probabilities. More details about computation of *RCI* can be found in Sindhwani et al. (2001).

c) *Overall Agreement:* It measures the portion of cases that two classifiers agree upon (without distinguishing between agreements on different classes), i.e. $OA = \sum_{i=1}^{c} n_{ii} / \sum_{i=1}^{c} \sum_{j=1}^{c} n_{ij}$.

d) It measures the degree of agreement on each class. Specific agreement on class ω_k is calculated as $SA_{\omega_k} = 2n_{kk} / (\sum_{i=1}^{c} n_{ik} + \sum_{j=1}^{c} n_{kj})$.

e) *Cohen's kappa:* Cohen's kappa is probably the most popular metric used in the literature, despite its known issues related to effect of prevalence in the cases of unbalanced classification problems (Hripcsak and Heitjan, 2002). It is calculated as $kappa = (OA - EA)/(1 - EA)$, where *OA* refers to the overall agreement and $EA \equiv \sum_{k=1}^{c} (\sum_{i=1}^{c} n_{ik} \sum_{j=1}^{c} n_{kj}) \ / \ (\sum_{i=1}^{c} \sum_{j=1}^{c} n_{ij})^2$ is the agreement expected by chance.

Other Methods for Comparison: The proposed IC classifier (i.e. weighted PWC-PSVM + error correction) was compared with the following five benchmark methods: (i) the weighted PWC-PSVM without error correction, (ii) the standard SVM trained on under-sampled balanced dataset as used in the work by Nicolaou and Nasuto (2004); Shoker et al. (2005), (iii) GMM (with class conditional probability densities estimated by using the software package developed by Bouman (1997), (iv) KNN (K from 1 to 9 were tested and the best results obtained with K=5 were reported), and (v) LDF using the minimum-squared-error solution.

4.4.4 Qualitative Performance Evaluation by Reviewing Reconstructed EEG

An independent EEG expert was invited to qualitatively evaluate the performance of the proposed artifact removal system by examining each of the 250 raw EEG epochs and its corresponding reconstructed EEG epoch simultaneously. The evaluation of each epoch was based on three aspects: the removal of ECG artifact, the removal of eye-blinking artifact and the preservation of brain activities. The EEG expert was required to give detailed judgment on each of these three aspects. For the evaluation of the removal of ECG or EOG artifact, “No improvement” was used to indicate that almost no change was observed in the amount of artifacts before and after artifact removal; “minor improvement” indicated that artifacts were partially removed but still observed in the reconstructed EEG; a score of “almost removed” was given when almost no considered artifact was observed in the reconstructed EEG. For the evaluation on the preservation of brain activities, “major attenuation” was used to indicate that typical brain activities were significantly attenuated; a score of “minor attenuation” was given when the amplitude of typical brain activities was reduced but still visible; “well preserved” indicated that almost no change in brain activities was observed before and after artifact removal.

4.4.5 Experimental Results

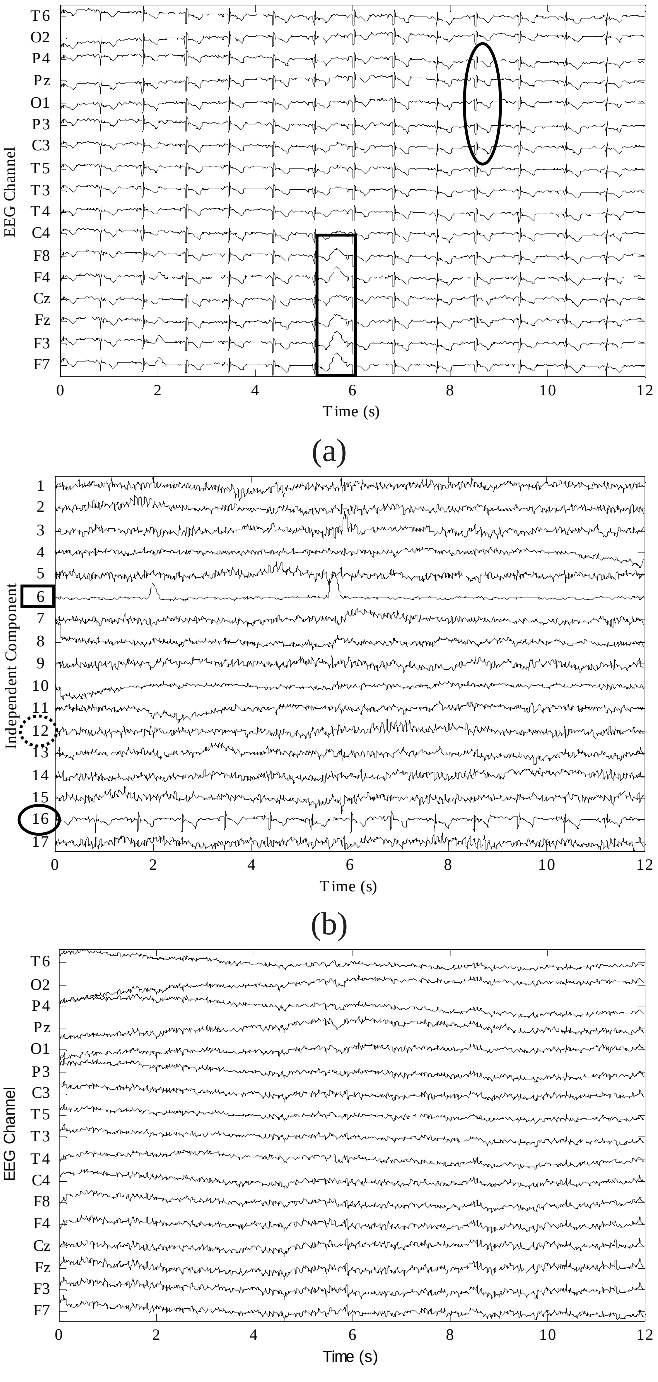
4.4.5.1 Validation of the Unique Properties of the Learning Problem

The collected data as described in Section 4.4.1 showed that among the 17 ICs separated from each 12-second EEG epoch, there were only one ECG IC and no more than two EOG ICs. In total, 250 ECG ICs, 292 EOG ICs and 3,708 EEG ICs were separated from 250 EEG epochs from the ten subjects. Fig. 4.2a, Fig. 4.2b and Fig. 4.2c show a typical 12-second EEG epoch, the resulting ICs and the reconstructed EEG after artifact removal, respectively. As can be seen, only one EOG IC (marked by a square) and one ECG IC (marked by a circle) were separated from the EEG epoch. This verified the unique properties of the learning problem at hand: the uneven class distributions and the underlying structural information as given in Equations (4.1) and (4.2). The evidence of such unique properties of the learning problem can also be seen from the optimal values of inequality constraints for EOG ICs (ω_2) and ECG ICs (ω_3) determined by the afore-mentioned grid-search, i.e. $u_\omega = 2$, $l_{\omega_2} = 1$, $u_{\omega_3} = l_{\omega_3} = 1$.

4.4.5.2 Quantitative Comparison

Detailed classification results and performance measures of the proposed method and the benchmark methods are summarized in Table 4.1. The numbers shown are the averages over 10 test datasets corresponding to 10 pairs of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$. The P-values (given in parentheses) were obtained in the paired t-test between the proposed method and each of the benchmark methods. Based on the results in Table , the proposed method appears to be superior to all the benchmark methods. Details are as follows.

a) *Comparison between the proposed method and the modeling approaches:* As shown in Table 4.1, the proposed method achieved significantly higher balanced accuracy and RCI than the modeling approaches. It performed well on both the majority class (i.e. EEG ICs) and the minority classes (EOG ICs and ECG ICs). In comparison, all the modeling approaches showed very good performance on EEG ICs, which is evidenced by the significant higher specific agreement on EEG ICs given by KNN and GMM and the significant higher overall agreement given by GMM; however, their performance on the minority classes was not satisfying. In the context of artifact removal, the proposed method which achieved a good tradeoff among classification performance on each class is preferable.



(a)

(b)

(c)

Figure 4.2: A typical example of (a) a 12-second EEG epoch (the waveforms marked with ellipse and rectangular are typical ECG and eye-blinking artifacts), (b) the resultant ICs (The IC marked by a rectangular was “true” EOG IC and the one marked by a ellipse was “true” ECG IC, as labeled by the EEG expert. The IC marked by an dashed ellipse which was a “true” EEG IC was misidentified as an ECG IC by the weighted PWC-PSVM. This misidentification was subsequently corrected by the proposed error correction algorithm, (c) the corresponding reconstructed EEG epoch after artifact removal by the proposed method.

b) *Comparison between the proposed method and the standard SVM:* As can be seen from Table 4.1, the standard SVM showed significant performance deterioration on all performance measures as compared with the proposed method. The results given by the confusion matrices suggest that the better performance of the proposed method is mainly due to its higher accuracy on EEG ICs as compared with the standard SVM (3520/3708 v.s. 3424/3708). One plausible reason is that the standard SVM, as used in past work by Nicolaou and Nasuto (2004); Shoker et al. (2005), was trained on down-sampled balanced training data (with large portion of samples of EEG ICs being discarded). Such down-sampling causes a significant loss of information and thus leads to suboptimal performance on the majority class.

c) *Comparison between the weighted PWC-PSVM with error correction and the weighted PWC-PSVM without error correction:* As shown in Table 4.1, almost all performance measures show that the weighted PWC-PSVM with error correction significantly outperformed the weighted PWC-PSVM method without error correction. The confusion matrices show that the incorporation of error correction resulted in an large increase in the number of correctly classified EEG ICs (3540 v.s. 3474) at a tiny cost of the number of correctly classified ECG ICs (246 v.s. 248). It is a strong indication of the goodness of the proposed error correction algorithm. Consider all the ICs resulting from a given EEG epoch. On the one hand, the error correction algorithm prevents the classifier from attributing more ICs to artifacts than it should, i.e. avoids exceeding the corresponding upper limits as in Equations (4.2). On the other hand, if the classifier fails in picking up the minimum number of artifactual ICs, the error correction algorithm enforces assigning certain number of most probable ICs (with largest posterior probability of belonging to the classes of ECG/EOG) to artifactual ICs. Fig. 4.2 shows a typical example when the weighted PWC-PSVM classified two ICs (marked with a circle and an ellipse respectively) as ECG ICs but the IC marked with the ellipse was actually an EEG IC. The proposed error correction algorithm corrected this error by incorporating the constraint on ECG ICs: there is only one ECG IC decomposed from a given EEG epoch.

Table 4.1: Performance comparison between the proposed method (i.e. weighted PWC-PSVM + ER) and five benchmark methods (weighted PWC-PSVM, standard SVM, GMM, KNN and LDF). The numbers shown are averages over 10 test datasets corresponding to 10 pairs of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$. The number in parenthesis is the P-value obtained in the paired t-test between each of the benchmark methods and the proposed method. The symbols '$^+$' and '$^-$' indicate statistically significant wins or losses over the proposed method (P-value < 0.05).

Classifier	Confusion Matrix				BA	RCI	Kappa	OA	SA		
	T/P	ω_1	ω_2	ω_3	(%)	(%)			ω_1	ω_2	ω_3
Weighted PWC-PSVM + EC	ω_1	3540	20	4	95.67	76.76	0.82	0.95	0.97	0.75	0.98
	ω_2	164	272	0							
	ω_3	4	0	246							
Weighted PWC-PSVM	ω_1	3474	20	2	95.35	70.33	0.78	0.94	0.95	0.74	0.88
	ω_2	171	272	0	(0.05)	(0.00^{-})	(0.01^{-})	(0.01^{-})	(0.00^{-})	(0.32)	(0.01^{-})
	ω_3	63	0	248							
Standard SVM	ω_1	3424	20	4	94.63	67.15	0.74	0.93	0.96	0.7	0.87
	ω_2	212	272	0	(0.02^{-})	(0.00^{-})	(0.00^{-})	(0.00^{-})	(0.00^{-})	(0.03^{-})	(0.00^{-})
	ω_3	72	0	246							
GMM	ω_1	3653	71	20	88.73	68.34	0.85	0.97	0.98	0.79	0.95
	ω_2	49	221	0	(0.00^{-})	(0.00^{-})	(0.23)	(0.02^{+})	(0.00^{+})	(0.07)	(0.00^{-})
	ω_3	6	0	230							
KNN (K=5)	ω_1	3661	69	27	87.87	66.46	0.85	0.97	0.98	0.82	0.9
	ω_2	28	221	0	(0.00^{-})	(0.00^{-})	(0.45)	(0.07)	(0.00^{+})	(0.23)	(0.02^{-})
	ω_3	19	2	223							
LDF	ω_1	3691	129	197	58.71	29.08	0.53	0.92	0.96	0.7	0.343
	ω_2	12	162	0	(0.00^{-})	(0.00^{-})	(0.00^{-})	(0.00^{-})	(0.03^{-})	(0.01^{-})	(0.00^{-})
	ω_3	5	1	53							

4.4.5.3 Review of Reconstructed EEG

The qualitative evaluation of the proposed artifact removal system by the independent EEG expert is given in Table 4.2. Artifacts were largely reduced, without attenuating brain activities, in most of the reconstructed EEG epochs. The amount of ECG artifacts was reduced in 98.4% of the EEG epochs, with 98.0% indicated as "almost removed" and 0.4% indicated as "minor improvement". Eye-blinking artifacts were removed in 96.8% of the reconstructed EEG epochs, with 92.0% indicated as "almost removed" and 4.8% indicated as "minor improvement". In 88.4% of the epochs, brain activities

Table 4.2: Qualitative evaluation of the proposed method on the removal of ECG, eye-blinking artifact and the preservation of brain activities by an independent EEG expert

ECG Removal	*No improvement* 1.60%	*Minor improvement* 0.40%	*Mostly removed* 98.00%
Eye-blinking Removal	*No improvement* 3.20%	*Minor improvement* 4.80%	*Mostly removed* 92.00%
EEG Preservation	*Major attenuation* 0.80%	*Minor attenuation* 10.80%	*Well preserved* 88.40%

were well preserved. Only 0.8% of the epochs suffered from major attenuation in brain activities and 10.8% of the epochs were slightly attenuated in brain activities.

4.5 Discussion

A novel IC classifier which combines a modified probabilistic multi-class SVM and an error correction algorithm has been proposed in the present study. The proposed approach has been compared with several benchmark methods: the modified probabilistic multi-class SVM without error correction, the standard SVM, GMM, KNN and LDF. In a stringent subject-wise cross-validation procedure, numerical experiments have shown that the proposed IC classifier achieved significantly better performance than the benchmark methods. Moreover, a qualitative evaluation of the reconstructed EEG by an independent expert has demonstrated that the proposed artifact removal method effectively removes artifacts while fairly well preserving brain activities in EEG. The superiority of the proposed approach can be attributed to the following reasons.

Firstly, the unbalanced nature of the underlying data is properly addressed by using the modified probabilistic multi-class SVM. This multi-class SVM is modified from the probabilistic multi-class SVM proposed by Hastie and Tibshirani (Hastie and Tibshirani, 1998), by replacing all standard binary SVMs with weighted SVMs. It uses real unbalanced data for training and penalizes more on the classification errors made on the samples from the minority class. As shown by experimental results, in comparison with the modeling approaches (GMM, KNN and LDF), the modified multi-class SVM appears more effective in compensating the bias of prior class probabilities. The proposed method is also superior to the standard SVM used in the past work by Nicolaou and Nasuto (2004) and Shoker et al. (2005) which was trained on a balanced training set formed by down-sampling of the majority class. The down-sampling inevitably causes loss of information and thus leads to the suboptimal performance of the standard SVM on the majority class.

Secondly, useful structural information of the underlying data is incorporated in decision making through the error correction algorithm. The structural information in the present study is the constraints on the number of ICs responsible for each type of artifact resulting from a given EEG epoch, as described in Equations (4.2). It is worth noting that this structural information is different from class priors: class priors can be directly included by many modeling method (e.g. KNN, GMM); however, the constraints as of Equations (4.2) exist among the batch of ICs resulting from the same EEG epoch and thus can only be exploited by considering the batch of ICs collectively (as the proposed error correction algorithm does). Conventional classifiers, such as KNN, GMM and LDF, which consider each sample independently, are unable to incorporate such structural information. As shown by experimental results, a better tradeoff among the classification accuracy on each class is achieved by incorporating this structural information through the error correction algorithm. The proposed error correction algorithm may be significant in both theoretical and practical aspects. It appears generally useful for classification problems where similar structural information is contained in the test samples and thus simultaneous classification of several test samples is necessary.

Moreover, the use of a probabilistic SVM may also contribute to the superior performance of the proposed method. Given a test sample, $\mathbf{x}$, the decision of conventional SVM is based on the sign of SVM outputs, $f_{ij}(\mathbf{x})$. Such an approach ignores the relative confidence in classification, or the distance that **x** is from the separating hyperplane. In contrast, the probabilistic SVM is based on the calibrated confidence measures, i.e. the estimates of posterior probabilities. The superiority of probabilistic SVM over standard SVM has been recently demonstrated in a few studies in the domain of machine learning (see Duan and Keerthi, 2005; Hastie and Tibshirani, 1998; Platt, 2000), while its application in EEG signal processing remains rare.

In the present study, we limited ourselves to the removal of ECG and eye-blinking artifacts. However, the idea of the proposed method can be generally extended to the removal of other types of artifacts that can be isolated to one or more ICs by ICA (for example, artifact due to muscle tension), provided that suitable features are available. It should be acknowledged that, like all the ICA-based ICA removal methods, the proposed method may produce discontinuities at the beginning and end of each reconstructed EEG epoch, although it appears minimal in our experiments. As a precaution, the segmentation should be retained in the review/use of the reconstructed EEG to prevent the potential discontinuities from interfering EEG interpretation.

4.6 Concluding Remarks

This chapter presents an advanced and comprehensive solution to the difficult problem for almost all EEG implementations: the problem of automatic EEG artifact removal. The proposed method takes into account the unique properties of the learning problem at hand by (i) using weighted probabilistic SVM to handle the unbalanced data, and (ii) implementing an error correction algorithm to accommodate useful structural information of the underlying data. Quantitative comparisons between the proposed method and several benchmark methods on real-life EEG data showed that the proposed method significantly outperforms the other methods. A qualitative review on the reconstructed

EEG also revealed that artifacts were largely attenuated while brain activities were well preserved in most of the EEG epochs. The proposed method appears to be well suited for automatic EEG artifact removal.

Chapter 5

Feature Selection via Sensitivity Analysis of SVM Probabilistic Outputs

Designing effective feature selection method which allows the identification of critical EEG features to mental fatigue is an important part of our effort in the development of the EEG-based mental-fatigue measurement and monitoring system. It is in general an important aspect of solving data-mining and machine-learning problems. This chapter proposes a feature-selection method for the SVM learning for two-class classification problems. Like most feature-selection methods, the proposed method ranks all features in decreasing order of importance so that more relevant features can be identified. It uses a novel criterion based on the probabilistic outputs of SVM. This criterion, termed Feature-based Sensitivity of Posterior Probabilities (FSPP), evaluates the importance of a specific feature by computing the aggregate value, over the feature space, of the absolute difference of the probabilistic outputs of SVM with and without the feature. The exact form of this criterion is not easily computable and approximation is needed. Four approximations, FSPP1-FSPP4, are proposed for this purpose. The first two approximations evaluate the criterion by randomly permuting the values of the feature among samples of the training data. They differ in their choices of the mapping function from standard SVM output to its probabilistic output: FSPP1 uses a simple threshold function while FSPP2 uses a sigmoid function. The second two directly approximate the criterion but differ in the smoothness assumptions of criterion with respect to the features. The performance of these approximations, used in an overall feature-selection scheme, is then evaluated on various artificial problems and real-world problems, including datasets from the recent Neural Information Processing Systems (NIPS) feature selection competition. FSPP1-3 show good performance consistently with FSPP2 being the best overall by a slight margin. The performance of FSPP2 is competitive with some of the best performing feature-selection methods in the literature on the datasets that we have tested. Its associated computations are modest and hence it is suitable as a feature-selection method for SVM applications.

5.1 Introduction

Feature selection is an important issue in machine-learning problems. When the underlying important features are known and irrelevant/redundant features are removed, learning problems can be greatly simplified resulting in improved generalization capabilities. Feature selection can also help reduce online computational costs, enhance system interpretability (Boser et al., 1992; Cortes and Vapnik, 1995; Vapnik, 1995, 1998) and improve performance of the learning problems (see Günter and Bunke, 2004; Guyon and Elisseef, 2003; Guyon et al., 2002; Saon and Padmanabhan, 2001; Weston et al., 2001). Several feature-selection methods have been proposed in recent years and a good review of them can be found in the recent book by (Guyon et al., 2006b). In general, feature-selection methods can be classified into three categories: filter-based, wrapper-based and embedded-based (Guyon and Elisseef, 2003; Kohavi and John, 1997; Neumann et al., 2005). Filter-based methods are independent of the underlying learning algorithm while wrapper-based methods use the underlying learning algorithm to measure the quality of the features but without exploiting the structure of the learning algorithm. In contrast, embedded-based methods exploit the knowledge of the specific structure of the learning algorithm (Guyon and Elisseef, 2003; Lal et al., 2006) and cannot be separated from it. Generally, embedded-based methods are superior in performance relative to filter-based or wrapper-based methods but carry with them a heavier computational load (Guyon et al., 2006a).

This chapter develops a new embedded-based feature-selection method specifically for support vector machine (SVM) learning. The focus on SVM stems from the interests in it as a learning method following its encouraging results on a variety of applications (Boser et al., 1992; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Vapnik, 1995, 1998). Unlike past feature-selection methods for SVM, this chapter proposes the use of the probabilistic outputs of SVM as a more accurate measure of feature importance. For the prototypical two-class (ω_{+} and ω_{-1}) classification problem, probabilistic output of SVM for a given sample, $\mathbf{x}$, can be interpreted (Hastie & Tibshirani, 1998; Platt, 2000) as the posterior probability of $\mathbf{x}$ belonging to class ω , $p(\omega|\mathbf{x})$. Here, class ω can be either ω_{+1} or ω_{-1}. Such an interpretation under the Bayesian framework has also been established (Williams & Rasmussan, 1996; Chu et al., 2003, 2004). This work proposes a criterion based on the sensitivity of probabilistic outputs of SVM to each feature as a measure of importance of that feature, and is termed Feature-based Sensitivity of Posterior Probabilities (FSPP). In loose terms, this criterion is the aggregate value, over the feature space, of the absolute difference of the probabilistic outputs of SVM with and without the feature.

The evaluation of this criterion is investigated using four approximations, termed FSPP1-FSPP4 respectively. These approximations are then combined with the recursive feature-elimination approach (Guyon et al., 2002) and other heuristic feature-ranking approaches to yield an overall feature-selection scheme. The first two approximations are motivated by the random forests feature-selection method (Breiman, 2001) where the idea of Random Permutation (RP) of the values of a feature is used to eliminate the contribution of that feature. They differ from each other in that FSPP1 uses a simple threshold function to obtain the probabilistic output of SVM while FSPP2 uses a sigmoid function. The second two are direct approximations of the criterion. FSPP3 assumes mild dependence of the criterion with respect to the features while FSPP4 assumes that criterion is differentiable with respect to the features. The proposed methods are tested on several learning problems, including the MONK's problems, breast cancer and heart disease problems from the UCI Repository (Newman et al., 1998), the nonlinear synthetic problem of Weston et al. (2001) and another two challenging problems, ARCENE and MADELON, from the NIPS 2003 feature selection competition (Guyon, 2003). Numerical comparisons with two well-known existing SVM feature-selection methods (SVM-RFE by Guyon et al. (2002) and the margin method by Rakotomamonjy (2003)) are also presented. The results show that FSPP2 performs consistently well on these datasets and compares favorably with the best methods available in the literature.

The rest of the chapter is organized as follows. Past related results from the literature needed for the subsequent sections are collected in Section 5.2. Section 5.3 provides the basis of the proposed criterion and the descriptions of the four approximations of the criterion. Section 5.4 outlines the overall feature-selection schemes using the proposed criterion. Extensive experimental results are reported in Section 5.5, followed by discussion in Section 5.6. The concluding remarks are drawn in Section 5.7.

5.2 Background

The section reviews closely-related past work on SVM feature-selection methods. The intention is to set the notations for the remainder of this chapter and to make the chapter as self-contained as possible. We begin with the general notations used. This work considers the typical two-class classification problem with dataset $\mathbb{D}$ in the form of $\{\mathbf{x}_j, y_j\}_{j=1}^N$ where $\mathbf{x}_j \in \mathbb{R}^D$ is the j^{th} sample, $y_j \in \{-1,+1\}$ is the corresponding class label, and D is the dimensionality of $\mathbf{x}_j$. Also, $\mathbf{x}^i$ denotes the i^{th} feature of vector $\mathbf{x}$, hence, $\mathbf{x}_j^i$ is the i^{th} feature of the j^{th} sample and $\mathbf{x}_{-i} \in \mathbb{R}^{D-1}$ is the vector obtained from $\mathbf{x}$ with the i^{th} feature removed. Double subscripted variable $\mathbf{x}_{-i,j}$ is also used and it refers to the j^{th} sample of variable $\mathbf{x}_{-i}$.

5.2.1 Probabilistic SVM

This chapter assumes the availability of the solution of the probabilistic SVM as proposed by Platt (2000). The full details of such probabilistic SVM formulation have been given in Sections 2.10.2.12.10.2.2 in Chapter 2. For the convenience of the reader, the key equations needed for the rest of this chapter are summarized as follows.

The SVM decision boundary of the two-class problems takes the form of an optimal separating hyperplane, $\mathbf{w}.\Phi(\mathbf{x})+b=0$, in Hillbert feature space H, obtained by solving the convex optimization problem

$$\min J(\mathbf{w}, b, \boldsymbol{\xi}) = \min \left[\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N} \xi_i \right], \tag{5.1}$$

$$\text{subject to} \quad y_i(\mathbf{w}\cdot\Phi(\mathbf{x}_i)+b) - 1 + \xi_i \geq 0,\ i = 1,\cdots,N, \text{ and} \tag{5.2}$$

$$\xi_i \geq 0, i = 1,\cdots,N, \tag{5.3}$$

over $\mathbf{w}\in H$, $b \in \mathbb{R}$ and the non-negative slack variable $\boldsymbol{\xi} \in \mathbb{R}^N$. The $\Phi(\cdot)$ defines the kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x}_j)$. In the above, C is a parameter that balances the size of $\mathbf{w}$ and the sum of ξ_i.

The solution of the above convex optimization problem gives the expression of the separating hyperplane, i.e.

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i K(x_i, x) + b, \tag{5.4}$$

serving as the decision function for all unseen samples **x** in that the predicted class is +1 if $f(\mathbf{x}) > 0$ and -1 otherwise. The normal of this hyperplane in H is

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \Phi(x_i). \tag{5.5}$$

For ease of presentation, the exposition hereafter uses, without loss of generality, the popular Gaussian kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2), \tag{5.6}$$

where γ is the kernel parameter. For accurate prediction of unseen samples, proper values of the parameters C and γ should be used. Typically, these parameters are obtained using the cross-validation procedure although other methods have also been discussed (Chapelle et al., 2002; Keerthi, 2002; Lee and Lin, 2000).

Standard SVM output classifies a sample **x** depending on the sign of $f(\mathbf{x})$, or the half space in H into which $\Phi(\mathbf{x})$ falls. Such an approach, however, ignores the relative confidence in the classification, or the distance $\Phi(\mathbf{x})$ is from the separating hyperplane. Platt (2000) addressed this shortcoming through the use of a sigmoid function and mapped $f(\mathbf{x})$ into $p(\omega|\mathbf{x})$, providing probabilistic information from standard SVM output. The benefit of $p(\omega|\mathbf{x})$ over $f(\mathbf{x})$ in improving classification accuracy has been demonstrated on several numerical experiments (Duan and Keerthi, 2005; Platt, 2000).

Suppose N_+ and N_- are the numbers of positive (y=+1) and negative (y=-1) samples respectively in dataset $\mathbb{D}$. The Platt's probability output is

$$\hat{p}(\omega|x) = \frac{1}{1+\exp(Af(x)+B)}, \tag{5.7}$$

where $f(\mathbf{x})$ is the SVM output given by Equation 5.4 and the parameters A and B are obtained from minimizing the negative log likelihood (or the cross-entropy error function) of $\mathbb{D}$ in the form of

$$\min F(A,B) = \min \left\{ -\sum_i [t_i \log \hat{p}(\omega|x_i) + (1-t_i)\log(1-\hat{p}(\omega|x_i))] \right\}, \tag{5.8}$$

with $t_i = (N_+ + 1)/(N_+ + 2)$ if $y_i = +1$ and $t_i = 1/(N_- + 2)$ if $y_i = -1$. Our implementation of the above includes the modifications suggested by Lin et al. (2003) for numerical stability. Hereafter, $\hat{p}(\omega|\mathbf{x})$ refers to the estimated posterior probability of belonging class +1 given $\mathbf{x}$ obtained from Equations (5.7)-(5.8), while $p(\omega|\mathbf{x})$ refers to the true but typically unknown posterior probability of belonging to class ω given $\mathbf{x}$. The quantity $\hat{p}(\omega|\mathbf{x})$ is used extensively in the approximations of the proposed feature-ranking criterion.

5.2.2 Past Work in SVM Feature Selection

Several feature-selection methods for SVM have been proposed in the literature (Barkley and Baumgartner, 2003; Guyon et al., 2002; Rakotomamonjy, 2003; Weston et al., 2001). In most of these methods, the feature-ranking criterion relies on the sensitivity of some suitable index of performance, or its estimate, with respect to the feature. Features with low sensitivity are deemed less important while those with high sensitivity are more.

Index of performance is typically linked to generalization ability of SVM and several estimates of this ability have been used in the literature. Guyon et al. (2002) used the cost function of (5.2) and proposed a feature-ranking criterion based on the sensitivity of this cost function with respect to a feature. In loose terms, this criterion measures the importance of a feature by the difference in the sizes of the margin with and without the feature. For notational convenience, this criterion is denoted by $\Delta||\mathbf{w}||^2$ hereafter. Using this criterion as a basis, less important features are dropped successively, resulting in a feature-selection method known as SVM Recursive Feature-Elimination (SVM-RFE). Similarly, Weston et al. (2001) used, as the performance index, the SVM radius/margin bound (Vapnik, 1998)

$$R^2 \left\|\mathbf{w}\right\|^2, \tag{5.9}$$

where R is the radius of the smallest sphere, centered at the origin, that contains all $\Phi(\mathbf{x}_i)$, $i = 1, \cdots, N$. The sensitivity of this index with respect to a feature was obtained through the use of a virtual scaling factor. As suggested by Weston et al. (2001), the idea could also be extended to the span estimate (Vapnik and Chapelle, 2000) which is a tighter upper bound on the expected generalization error. Rakotomamonjy (2003) extended SVM-RFE algorithm using radius/margin bound and span estimate and proposed feature-selection methods based on their zero-order and first-order sensitivity with respect to the features. As reported (Rakotomamonjy, 2003) to be the best among the considered methods, the first-order sensitivity, denoted by $\nabla \left\|\mathbf{w}\right\|^2$, is included in our numerical experiments for comparison.

5.3 The Ranking Criterion Based On Posterior Probabilities

The proposed ranking criterion for the i^{th} feature is

$$C_t(i) = \int |p(\omega|\mathbf{x}) - p(\omega|\mathbf{x}_{-i})| p(\mathbf{x}) d\mathbf{x}, \tag{5.10}$$

where $\mathbf{x}_{-i} \in \mathbb{R}^{D-1}$ is the vector derived from $\mathbf{x}$ with the i^{th} feature removed. The motivation of the above criterion is clear: the greater the absolute difference between $p(\omega|\mathbf{x})$ and $p(\omega|\mathbf{x}_{-i})$ over the space of $\mathbf{x}$, the more important is the i^{th} feature. As the true values of $p(\omega|\mathbf{x})$ and $p(\omega|\mathbf{x}_{-i})$ are usually unknown, they are approximated by $\hat{p}(\omega|\mathbf{x})$ and $\hat{p}(\omega|\mathbf{x}_{-i})$ respectively, obtained via Equations (5.7)-(5.8). The value of $\hat{p}(\omega|\mathbf{x}_{-i})$ corresponds to the probabilistic output of a SVM trained with data $\{\mathbf{x}_{-i,j}, y_j\}_{j=1}^{N}$ instead of $\{\mathbf{x}_j, y_j\}_{j=1}^{N}$. Since $\mathbf{x}$ has D features, this means that training of the SVM has to be done D times so that a ranked list of $\{C_t(i),\ i = 1, \cdots, D\}$ is obtained showing the relative importance of all features in $\mathbb{D}$. This is a computationally expensive process since each SVM training is expensive, having a known complexity (Joachims, 1999; Platt, 1999) of at least $O(N^2)$ and that D can be large. The remainder of this section shows four approximations (FSPP1-FSPP4) of (5.10) that avoid the retraining process.

Motivated by the random forests (RF) method (Breiman, 2001), the first two approximations involve a process of Random Permutation (RP) that randomly permutes the values of a feature. Specifically, the values of the i^{th} feature of $\mathbf{x}$ are randomly permuted over the N examples. All other features of $\mathbf{x}$, except $\mathbf{x}^i$, remain unchanged. Suppose $\zeta_1, \cdots, \zeta_{N-1}$ is a set of uniformly distributed random numbers from (0,1) and $\lfloor \zeta \rfloor$ is the largest integer that is less than ζ. The random permutation process is executed as follows (Page, 1967): For each k starting from 1 to $N-1$, compute $j = N \times \lfloor \zeta \rfloor + 1$ and swap the values of $\mathbf{x}_k^i$ and $\mathbf{x}_j^i$. At the end of this process, the values of $\mathbf{x}_i$ will have been randomly permuted.

We now state a general theorem relating the posterior probability and the RP process and it serves as the theoretical basis for FSPP1 and FSPP2. To state this theorem precisely, let $\mathbf{x}_{(i)} \in \mathbb{R}^D$ be the vector derived from $\mathbf{x}$ with the i^{th} feature randomly permuted.

Theorem 5.1 $p(\omega|\mathbf{x}_{(i)}) = p(\omega|\mathbf{x}_{-i})$.

Proof: As the uniform distribution is used in the RP process, the distribution of $p(\mathbf{x}^i)$ is unchanged, or

$$p(\mathbf{x}^i_{(i)}) = p(\mathbf{x}^i). \tag{5.11}$$

Hence, we have

$$p(\mathbf{x}_{(i)}) = p(\mathbf{x}^i_{(i)}, \mathbf{x}_{-i}) = p(\mathbf{x}^i_{(i)})p(\mathbf{x}_{-i}) = p(\mathbf{x}^i)p(\mathbf{x}_{-i}), \tag{5.12}$$

where the second equality follows from the fact that the distribution of the $p(\mathbf{x}^i_{(i)})$ is independent from $p(\mathbf{x}_{-i})$ following the RP process. Using similar argument, we have

$$p(\mathbf{x}_{(i)}, \omega) = p(\mathbf{x}^i_{(i)})p(\mathbf{x}_{-i}, \omega) = p(\mathbf{x}^i)p(\mathbf{x}_{-i}, \omega). \tag{5.13}$$

Hence,

$$p(\omega|\mathbf{x}_{(i)}) = \frac{p(\omega, \mathbf{x}_{(i)})}{p(\mathbf{x}_{(i)})} = \frac{p(\mathbf{x}^i)p(\mathbf{x}_{-i}, \omega)}{p(\mathbf{x}^i)p(\mathbf{x}_{-i})} = p(\omega|\mathbf{x}_{-i}). \tag{5.14}$$

■

A corollary of Theorem 5.1 is the mutual information equality of $I(\omega, \mathbf{x}_{(i)}) = I(\omega, \mathbf{x}_{-i})$.

This result follows from

$$\begin{aligned}
I(\omega,\mathbf{x}_{(i)}) &= \sum_{\omega}\int_{\mathbf{x}_{(i)}} p(\omega,\mathbf{x}_{(i)})\log\frac{p(\omega,\mathbf{x}_{(i)})}{P(\omega)p(\mathbf{x}_{(i)})}\mathrm{d}\mathbf{x}_{(i)} \\
&= \sum_{\omega}\int_{\mathbf{x}_{-i}}\int_{\mathbf{x}^{i}_{(i)}} p(\mathbf{x}^{i}_{(i)})p(\omega,\mathbf{x}_{-i})\log\frac{p(\omega,\mathbf{x}_{-i})}{P(\omega)p(\mathbf{x}_{-i})}\mathrm{d}\mathbf{x}^{i}_{(i)}\mathrm{d}\mathbf{x}_{-i} \qquad (5.15)\\
&= \sum_{\omega}\int_{\mathbf{x}_{-i}} p(\omega,\mathbf{x}_{-i})\log\frac{p(\omega,\mathbf{x}_{-i})}{P(\omega)p(\mathbf{x}_{-i})}\mathrm{d}\mathbf{x}_{-i}\int_{\mathbf{x}^{i}_{(i)}} p(\mathbf{x}^{i}_{(i)})\mathrm{d}\mathbf{x}^{i}_{(i)} \\
&= I(\omega,\mathbf{x}_{-i}),
\end{aligned}$$

where Equations (5.12) and (5.13) are invoked.

Theorem 5.1 and its corollary show that the RP process has the same effect as removing the contribution of that feature for classification. Using this fact, criterion as of Equation (5.10) can be equivalently stated as

$$C_t(i) = \int |p(\omega|\mathbf{x}) - p(\omega|\mathbf{x}_{(i)})|p(\mathbf{x})\mathrm{d}\mathbf{x}. \qquad (5.16)$$

With the above equivalent form of the the proposed ranking criterion, we are now in a position to state its first two approximations.

Method 1 (FSPP1): *Approximation using threshold function*

The first approximation uses a threshold function for the approximation of Equation (5.16) in the form of

$$p(\omega|\mathbf{x}) \approx \varphi(f(\mathbf{x})) \qquad (5.17)$$

and

$$p(\omega|\mathbf{x}_{(i)}) \approx \varphi(f(\mathbf{x}_{(i)})), \tag{5.18}$$

where $\varphi(\cdot)$ is the threshold function given by

$$\varphi(f) = \begin{cases} 1 \ \text{ if } f \geq 0 \\ 0 \ \text{ if } f < 0 \end{cases} . \tag{5.19}$$

It is worthy to note that $p(\omega|\mathbf{x}_{(i)})$ uses the same f function as given by Equation (5.4) and does not involve the retraining of the SVM. Further approximation of the integration over $\mathbf{x}$ in Equation (5.16) yields

$$\text{FSPP1}(i) = \frac{1}{N}\sum_{j=1}^{N} |\varphi(f(\mathbf{x}_j) - \varphi(f(\mathbf{x}_{(i),j}))|, \tag{5.20}$$

where $\mathbf{x}_{(i),j}$ refers to the j^{th} example of the input data where the i^{th} feature has been randomly permuted.

Method 2 (FSPP2): *Approximation using SVM probabilistic outputs*

Motivated by the good results reported by Platt (2000) and Duan and Keerthi (2005), FSPP2 approximates $p(c|\mathbf{x})$ by the Platt's probabilistic output, $\hat{p}(\omega|\mathbf{x})$, in Equation (5.16). Obviously, other methods that obtain probabilistic outputs from SVM can also be used (Hastie and Tibshirani, 1998; Vapnik, 1998). Similarly, $p(\omega|\mathbf{x}_{(i)})$ in Equation (5.16) is approximated by $\hat{p}(\omega|\mathbf{x}(i))$ using the same trained SVM and the same trained sigmoid for $\hat{p}(\omega|\mathbf{x})$. Hence,

$$\text{FSPP2}(i) = \frac{1}{N}\sum_{j=1}^{N} |\hat{p}(\omega|\mathbf{x}_j) - \hat{p}(\omega|\mathbf{x}_{(i),j})|. \tag{5.21}$$

Method 3 (FSPP3): *Approximation via virtual vector* $\mathbf{v}$

Unlike the previous, the next two methods (FSPP3 and FSPP4) approximate Equation (5.10) via an additional virtual scaling factor. The use of an additional virtual vector $\mathbf{v} \in \mathbb{R}^D$ for the purpose of feature selection has been attempted in the literature (Rakotomamonjy, 2003; Weston et al., 2001) and it simplifies the computation of Equation (5.10). Specifically, this approach uses one $\mathbf{v}_i$, having a nominal value of 1, for each feature and replaces every $\mathbf{x}^i$ by $\mathbf{v}^i\mathbf{x}^i$. Let $\mathbf{vx} = [\mathbf{v}^1\mathbf{x}^1 \ \cdots \ \mathbf{v}^D\mathbf{x}^D]$ and $\mathbf{v}_{-i}\mathbf{x}$ refers to $\mathbf{vx}$ with $\mathbf{v}^i = 0$. In this setting, the criterion as of Equation (5.10) can be approximated by

$$\mathrm{C}_t(i) = \int |p(\omega|\mathbf{vx}) - p(\omega|\mathbf{v}_{-i}\mathbf{x})|p(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{5.22}$$

Using standard approximation, the above becomes

$$\mathrm{FSPP3}(i) = \frac{1}{N}\sum_{j=1}^{N} |\hat{p}(\omega|\mathbf{vx}_j) - \hat{p}(\omega|\mathbf{v}_{-i}\mathbf{x}_j)|, \tag{5.23}$$

where $\hat{p}(\omega|\mathbf{vx}_j)$ refers to the Platt's posterior probability of the j^{th} example and $\hat{p}(\omega|\mathbf{v}_{-i}\mathbf{x}_j)$ = $[1 + \exp(Af(\mathbf{v}_{-i})\mathbf{x}) + B)]^{-1}$ as given by Equation (5.7) and $f(\cdot)$ is the SVM output expression (5.4) obtained from the training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$.

Method 4 (FSPP4): *Approximation via derivative of* $p(\omega|\mathbf{vx})$ *with respect to* $\mathbf{v}$

The criteria as of Equation (5.22) can also be represented, under the assumption that $p(\omega|\mathbf{vx})$ is differentiable with respect to $\mathbf{v}$, by

$$C_t(i) = \int \left| \int_{\mathbf{v}^i=1}^{\mathbf{v}^i=0} \frac{\partial p(\omega|\mathbf{vx})}{\partial \mathbf{v}^i} \mathrm{d}\mathbf{v}^i \right| p(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{5.24}$$

Instead of the integral over $\mathbf{v}^i$ from 1 to 0, FSPP4 uses the sensitivity with respect to $\mathbf{v}^i$

evaluated at $\mathbf{v}^i = 1$ and the above is approximated by

$$C_t(i) = \int \left| \frac{\partial p(\omega|v\mathbf{x})}{\partial \mathbf{v}^i} \Delta \mathbf{v}^i |_{\mathbf{v}^i=1} \right| p(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \left| \frac{\partial p(\omega|v\mathbf{x})}{\partial \mathbf{v}^i} |_{\mathbf{v}^i=1} \right| p(\mathbf{x}) \mathrm{d}\mathbf{x}, \qquad (5.25)$$

where $\Delta\mathbf{v}^i = -1$. It is important to note that, when $p(c|\mathbf{x})$ is approximated by $\hat{p}(\omega|\mathbf{x})$ via Equation (5.7), $\partial \hat{p}(\omega|\mathbf{vx})/\partial \mathbf{v}^i$ admits a closed-from expression using the results of Equation (5.4) and Equation (5.7). For the ease of presentation, its expression and derivation are given in Appendix B. Hence, the fourth method is

$$\text{FSPP4}(i) = \frac{1}{N} \sum_{j=1}^{N} \left| \frac{\partial \hat{p}(\omega|\mathbf{vx}_j)}{\partial \mathbf{v}^i} |_{\mathbf{v}^i=1} \right|. \qquad (5.26)$$

The above shows four possible approximations to the ranking criterion as of (5.10). The use of these four methods, in an overall scheme for the purpose of feature selection, is shown in the next section.

5.4 Feature-Selection Methods

This section presents two overall feature-selection schemes by combining FSPP1-FSPP4 with either the initial feature-ranking (INIT) approach (FSPP-INIT) or the recursive feature-elimination (RFE) approach (FSPP-RFE). Both INIT and RFE approaches are commonly used for feature selection, with INIT being closer to the filter-based method and the RFE being closer to the embedded method (Guyon et al., 2006b; Guyon and Elisseef, 2003).

For both of the proposed feature-selection schemes (FSPP-INIT and FSPP-RFE), it is assumed that an SVM output function $f(\mathbf{x})$ is available and that all hyper parameters, C, γ or others, have been determined through a proper model selection process. For the cases where FSPP2-FSPP4 are involved, it is also assumed that the posterior probabilities are available according to Equations (5.7) and (5.8).

The FSPP-INIT scheme has as its inputs dataset $\mathbb{D}$, the index set $I = \{1,\ 2,\ \cdots,\ D\}$ containing indices of features to be considered and the choice of the approximation method $m \in \{1,\ \cdots,\ 4\}$. The output of FSPP-INIT is a ranked list of the features in the form of an index set $J_r = \{j_1, \cdots, j_D\}$ with $j_k \in I$ and FSPP$m(j_k) \geq$ FSPP$m(j_{k+1})$ for $k = 1, \cdots, D-1$. The major steps involved are shown in Algorithm 2.

Algorithm 5.1: Main steps of FSPP-INIT feature-selection scheme.

```
Input: $\mathbb{D}$, $I$, $m$
Output: Ranked list $J_r$
Train the SVM and obtain the posterior probabilities via Equations (5.7) and
(5.8) using the dataset $\mathbb{D}$; For each $i \in I$, compute FSPP$m(i)$;
Output ranked list $J_r = \{j_1, \cdots, j_D\}$ with $j_k \in I$ and FSPP$m(j_k) \geq$
FSPP$m(j_{k+1})$ for $k = 1, \cdots, D-1$.
```

The FSPP-RFE scheme is similar to the one given by Guyon et al. (2002) but with the FSPPm used as the ranking criterion. The steps involved in this approach are summarized in Algorithm 3. The inputs are the dataset $\mathbb{D}$, I and m, with the output being the ranked list of features J_R.

Algorithm 5.2: Main steps of FSPP-RFE feature selection scheme.

```
Input: $\mathbb{D}$, $I$, $m$
Output: Ranked list $J_R$
while $I \neq \emptyset$ do
    set $l = \text{size}(I)$;
    if $l > 1$ then
        Invoke FSPP-INIT($\mathbb{D}$, $I$, $m$)and obtain the output $J_r$;
        Let the last element of $J_r$ be $k^*$;
        Assign $k^*$ to the $l^{th}$ element of $J_R$;
        else
            Assign the only element in $I$ to the $l^{th}$ element of $J_R$;
        end
    end
    Let $I = I \setminus k^*$, remove feature $k^*$ from every sample in $\mathbb{D}$ and clear $J_r$;
end
```

As the FSPP-INIT scheme computes the ranked list only once, it is closer in spirit to a filter-based feature-selection scheme although the SVM algorithm is used. On the other hand, the FSPP-RFE scheme uses FSPP-INIT as an inner-loop and invokes it $D-1$ times, each time with a smaller index set I. The FSPP-RFE(D, I, m) scheme removes one feature (the one with the lowest FSPPm score) from the dataset at a time. Obviously, more than one feature can be removed at one time with slight modifications to Steps 5, 6 and 11 in Algorithm 3. The current description of FSPP-RFE does not involve the determination of parameters C and γ for each of the inner loop. Such a process is possible albeit with even higher costs. For notational convenience, FSPPm-INIT and FSPPm-RFE are used to specify the feature selection scheme using FSPPm as the choice of the approximation method.

5.5 Experiments

Extensive experiments on both artificial and real-world benchmark problems were carried out using the proposed methods. Like others, the artificial problems, i.e. MONK's problems from UCI Repository (Newman et al., 1998) and Weston's nonlinear synthetic problem (Weston et al., 2001), were used because the key features are known and are suitable for comparative study of the four FSPPs. Two real-world problems, i.e. breast cancer and heart disease problems from UCI Repository (Newman et al., 1998; Rätsch, 2005), were chosen as they have been used by other feature-selection methods (Guyon et al., 2002; Rakotomamonjy, 2003) and serve as a common reference for comparison. Finally, the proposed methods were tested on ARCENE and MADELON problems used in the NIPS 2003 feature selection competition (Guyon, 2003), a well-known set of challenging feature-selection problems.

In general, our method requires, for each problem, three subsets of data in the form of

$\mathbb{D}_{tra}$, $\mathbb{D}_{val}$ and $\mathbb{D}_{tes}$ for training, validation and testing purposes. In cases where only $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$ were available, $\mathbb{D}_{tra}$ was further split randomly into a new $\mathbb{D}_{tra}$ and $\mathbb{D}_{val}$ in the ratio of 70% to 30%. The subset $\mathbb{D}_{tra}$ was normalized to zero mean and unit standard deviation. Its normalizing parameters were also used to normalize $\mathbb{D}_{val}$ and $\mathbb{D}_{tes}$. The subset $\mathbb{D}_{tra}$ was meant for the training of the SVM including the determination of the optimal C and γ using 5-fold cross-validation procedure. The subset $\mathbb{D}_{val}$ was needed for the determination of parameters A and B in Equation (5.7). The $\mathbb{D}_{tes}$ subset was used for obtaining an unbiased testing accuracy of the underlying method. In cases where there were 100 realizations of a given dataset, the procedure by Rätsch et al. (2001) was followed: parameters C and γ were chosen as the median of the five sets of (C, γ) of the first five realizations. Here each set of (C, γ) was obtained by standard 5-fold cross-validations for one realization.

5.5.1 Artificial Problems

MONK's problems: These problems (MONK-1 to 3) are available in UCI Repository of machine learning databases (Newman et al., 1998). As the provided data do not have $\mathbb{D}_{val}$ and the size of $\mathbb{D}_{tra}$ is relatively small, our experiments used part of the test set to form $\mathbb{D}_{val}$ and $\mathbb{D}_{tra}$. The exact data split and the descriptions of the dataset are given in Table .

The results for MONK-1 experiment using the optimal parameters (C = 32 and γ = 0.125) are shown in Fig. 5.1. Fig. 5.1a shows the FSPPm scores for the four methods using the INIT approach. It is easy to see that all four methods were effective in determining the key features. Figure 5.1b shows the test error rates of SVM using only the top-ranked features obtained via the RFE approach. The monotonic decrease in the testing error rates with increasing top-ranked features is a clear indication of the effectiveness of the feature-selection procedure. The results for MONK-2 and MONK-3

Table 5.1: Description of MONK's datasets (Five discrete features: $\mathbf{x}^1, \mathbf{x}^2, x^4 \in \{1,2,3\}$; x^3, $x^6 \in \{1,2\}$; $x^5 \in \{1,2,3,4\}$)

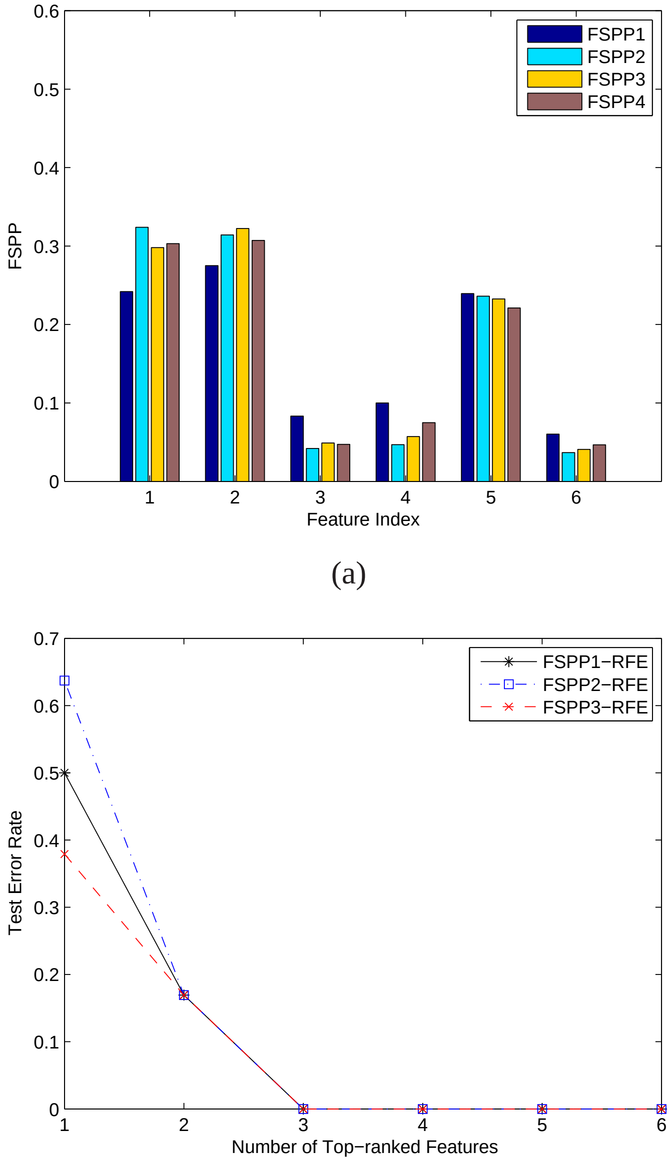
	$\mathbb{D}_{tra}$	$\mathbb{D}_{val}$	$\mathbb{D}_{tes}$	*Target Concept*
MONK-1	216	216	124	($\mathbf{x}^1$ = $\mathbf{x}^2$) or ($\mathbf{x}^5$ = 1) for Class 1, otherwise Class -1
MONK-2	216	216	169	Exactly two of $\{$ $\mathbf{x}^1$ = 1, $\mathbf{x}^2$ = 1, $\mathbf{x}^3$ = 1, $\mathbf{x}^4$ = 1, $\mathbf{x}^5$ = 1, $\mathbf{x}^6$ = 1$\}$ for Class 1, otherwise Class -1
MONK-3	216	216	122	($\mathbf{x}^5$ = 3 and $\mathbf{x}^4$ = 1) or ($\mathbf{x}^5 \neq 4$ and $\mathbf{x}^2 \neq 3$) for Class 1, otherwise Class -1

show similar trends to Fig. 5.1 and are hence not shown.

The test error rates for FSPP4-RFE are not shown in Fig. 5.1b as the computation of Equation (B.5) failed. This problem arose due to the existence of multiple identical examples in the training data, resulting in the matrix in Equation (B.5) being singular. While less likely to occur in real-life datasets, such situations can be handled using pseudo inverses and/or Singular Value Decomposition (SVD) of the matrix in Equation (B.5). However, they were not pursued because the performance of FSPP4 for other examples is not promising, as shown in the next few examples.

Weston's nonlinear synthetic problem: We followed the procedure given in (Weston et al., 2001) and generated 10,000 samples of 10 features each. Only the first two features ($\mathbf{x}^1$, $\mathbf{x}^2$) are relevant while the remaining features are random noise, each taken from a normal distribution $N(0,\ 20)$. The output $y \in \{-1,\ +1\}$ and the number of samples with y = +1 is equal to that with y=-1. If $y = -1$, ($\mathbf{x}^1$, $\mathbf{x}^2$) were drawn from $N(\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma})$ or $N(\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma})$ with equal probability, with $\boldsymbol{\mu}_1 = (-3/4,\ -3)$, $\boldsymbol{\mu}_2 = (3/4,\ 3)$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$. If y = +1, ($\mathbf{x}^1$, $\mathbf{x}^2$) were drawn again from two normal distributions with equal probability, with $\boldsymbol{\mu}_1 = (3,\ -3)$, $\boldsymbol{\mu}_2 = (-3,\ 3)$ and the same $\boldsymbol{\Sigma}$. $\mathbb{D}_{tra}$ and $\mathbb{D}_{val}$ contained 100 random samples each and the rest were included in $\mathbb{D}_{tes}$ for one realization of the dataset.

Average feature-selection performance over 100 realizations is shown in Fig. 5.2. with



(a)

(b)

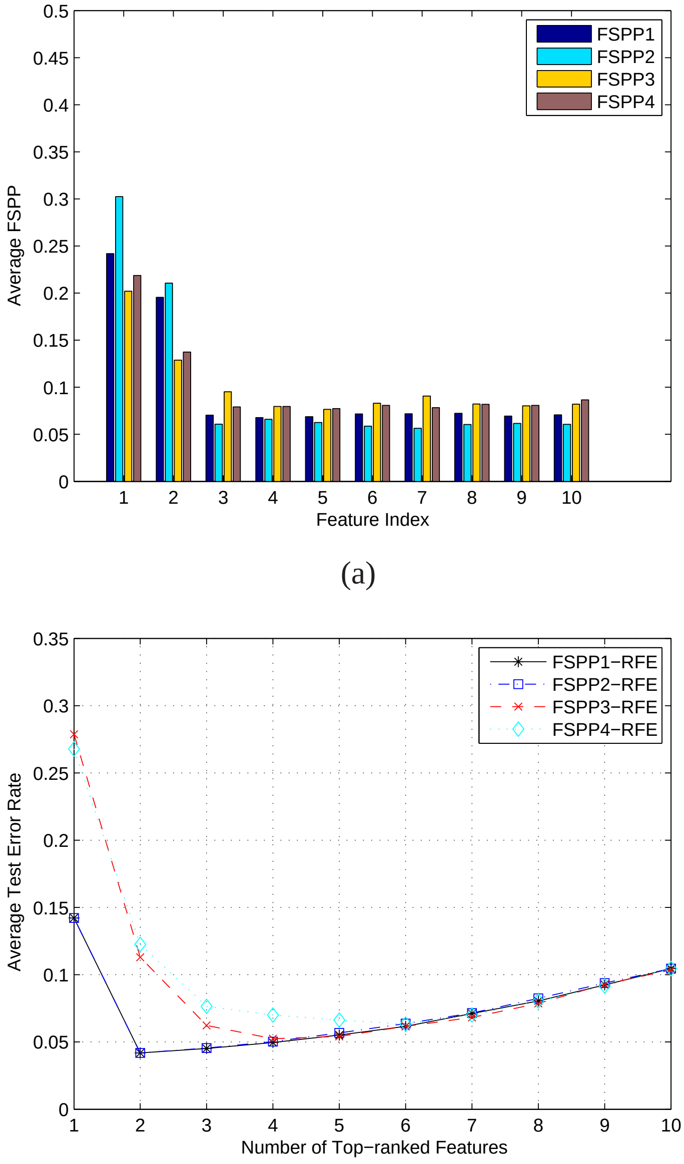
Figure 5.1: Performance of proposed methods on MONK-1 problem: (a) values of FSPP*m*, *m* = 1, 2, 3, 4 using FSPP*m*-INIT; (b) test error rates against top-ranked features identified by FSPP*m*-RFE.

the parameters set at $C = 32.0$, $\gamma = 0.03125$. Similar to the MONK's problems, Fig. 5.2a and 5.2b were obtained from the use of FSPPm-INIT and FSPPm-RFE respectively. Fig. 5.2a shows the correct identification of the first two features having FSPPm scores that are significantly larger (P-value < 0.01 based on paired t-test over the 100 realizations) than the FSPPm scores of a redundant feature. Fig. 5.2b shows that FSPP1-RFE and FSPP2-RFE correctly identified the two key features as the test error rates were the lowest with only two surviving features. However, FSPP3-RFE and FSPP4-RFE produced less appealing results. Additional experiments were conducted to verify the statistical significance of the advantage of FSPP1 and FSPP2 over FSPP3 and FSPP4 under the RFE approach. Four paired t-tests on the test error rates were conducted: FSPP1 vs FSPP3, FSPP1 vs FSPP4, FSPP2 vs FSPP3 and FSPP2 vs FSPP4. Each of these t-tests was further repeated with only 1, 2, 3 or 4 surviving features. For all of these paired t-tests, the P-values obtained were less than 0.03.

The difference between the performance of FSPP2 and FSPP3 is interesting and deserves attention. Both criteria use the same $\hat{p}$ expression obtained from Equations (5.7) and (5.8) but differ in that $\hat{p}(\omega|\mathbf{x}_{(i),j})$ is used in FSPP2 and $\hat{p}(\omega|\mathbf{v}_{-i}\mathbf{x}_j)$ in FSPP3. The sample $\mathbf{x}_{(i),j}$ has the i^{th} feature taking value that is randomly permuted while $\mathbf{v}_{-i}\mathbf{x}_j$ has the i^{th} feature set to 0. The better performance of FSPP2 over FSPP3 appears to suggest that the distribution $\hat{p}(\omega|\mathbf{v}_{-i}\mathbf{x})$ differs more from $p(\omega|\mathbf{x}_{-i})$ than $\hat{p}(\omega|\mathbf{x}_{(i)})$.

5.5.2 Real-World Benchmark Problems

The real-world benchmark problems are the breast cancer and heart disease datasets obtained from Rätsch (2005), used also by Mika et al. (1999); Rakotomamonjy (2003); Rätsch et al. (2001) in their experiments. Sizes of feature/$\mathbb{D}_{tra}$/$\mathbb{D}_{val}$/$\mathbb{D}_{tes}$ are 9/140/60/77 and 13/119/51/100 respectively and each problem has 100 realizations. For comparison purposes, the format of presentation of results by Rakotomamonjy (2003) was adopted.



(a)

(b)

Figure 5.2: Performance of the proposed methods on Weston's nonlinear dataset: (a) values of FSPPm, m = 1,2,3,4 using FSPPm-INIT; (b) test error rates against top-ranked features identified by FSPPm-RFE. Note that the stated FSPPm values and test error rates are the averages over 100 realizations.

Plots of the mean test error rates of SVM are provided with decreasing number of top-ranked features. Each plot is the mean over 100 realizations using either FSPP-RFE or FSPP-INIT feature-selection scheme.

For comparison purposes, performance of two feature-ranking criteria, the $\Delta||\mathbf{w}||^2$ method by Guyon et al. (2002) and the $\nabla\,||w||^2$ method by Rakotomamonjy (2003), is also included. They were chosen because they appear to have performed well (Rakotomamonjy, 2003; Weston et al., 2001). Their performance was reproduced together with those using FSPP1-4 in Figs. 5.3 and 5.4 for the two problems. While Fig. 5.3 is for breast cancer dataset and Fig. 5.4 is for the heart disease dataset, Figs. 5.3a and 5.4a report on the results based on the INIT approach while Figs. 5.3b and 5.4b are results of the RFE approach. These results were obtained for the optimal parameters: (C = 2.83, γ = 0.05632) for the breast cancer dataset and (C = 2.38, γ = 0.00657) for the heart disease dataset.

Under the INIT approach, Fig. 5.3a shows that all the methods considered (except FSPP4) produced similar test error rates for the breast cancer dataset. This is confirmed by the P-values ($>$0.05) obtained from paired t-tests for the 100 realizations, except for FSPP4 which gave P-values of less than 0.01 when compared with other methods. This was, however, not observed for the heart disease dataset. Fig. 5.4a shows that the FSPP1-4 are significantly better than the$\Delta||\mathbf{w}||^2$ and the$\nabla||\mathbf{w}||^2$ methods with P-values being less than 0.01 in the paired t-tests for FSPPm vs $\Delta||\mathbf{w}||^2$ and FSPPm vs $\nabla||\mathbf{w}||^2$. The performance of FSPP4 is not appealing for the breast cancer data. One possible reason is that the function $\hat{p}(\omega|\mathbf{vx})$ as a function of $\mathbf{v}^i$ is highly nonlinear and not well approximated by $\partial\hat{p}(\omega|\mathbf{vx})/\partial\mathbf{v}^i$ evaluated at $\mathbf{v}^i$= 1 as in Equation (5.26).

For the RFE approach, Fig. 5.3b shows that FSPP1 and FSPP2 again yielded significantly lower average test error rates than FSPP3, $\Delta||\mathbf{w}||^2$ and $\nabla||\mathbf{w}||^2$. This is confirmed by the paired t-tests with P-values $<$ 0.05 when only the top 2 or 3 features were used.

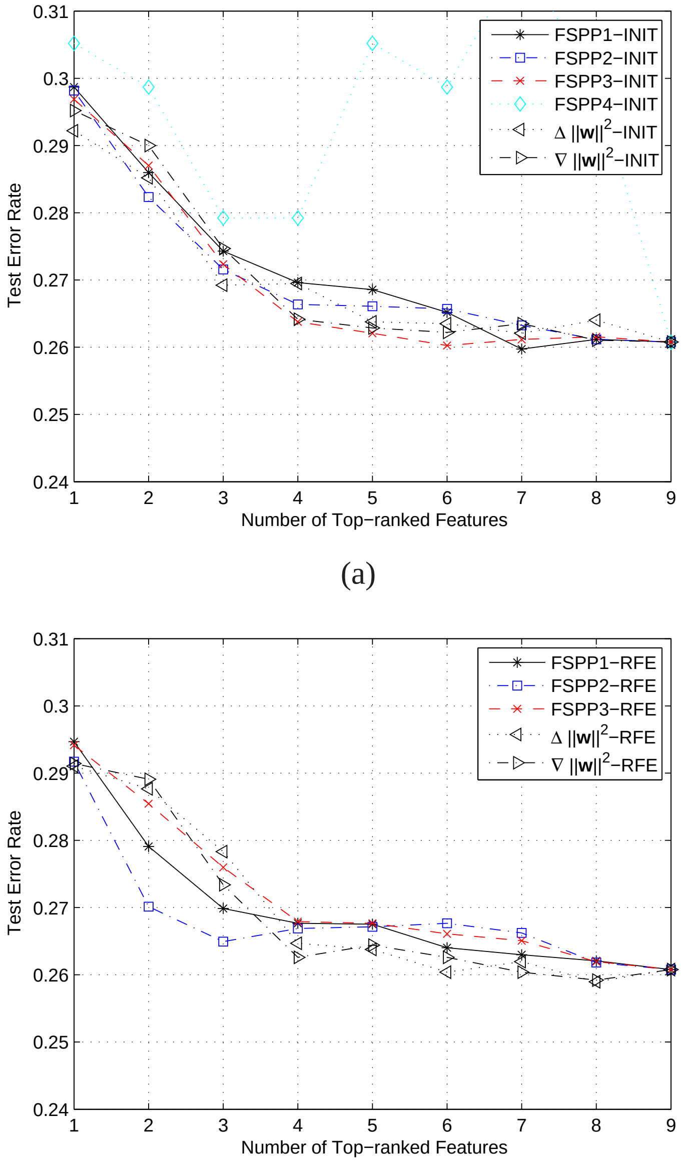
Table 5.2: Description of ARCENE and MADELON datasets

Dataset	Features	$\mathbb{D}_{tra}$	$\mathbb{D}_{val}$	$\mathbb{D}_{tes}$
MADELON	500	2000	600	1800
ARCENE	10000	100	100	700

Fig. 5.3b further shows that FSPP2 had a slight edge over FSPP1 and produced lower average test error rates when only the top 2 or 3 features were used (P-values$<$0.05), suggesting that FSPP2 could be the best performing method. In Fig. 5.4b, the advantage of the FSPP*m* over the other two methods is obvious. The paired t-tests between FSPP*m* versus either of the two methods yielded P-values of less than 0.03. The variation in performance among FSPP1-3 are, however, not significant as the P-values were greater than 0.05. Also, FSPP4-RFE is not shown in Fig. 5.3b or Fig. 5.4b as the computation of Equation (B.5) failed during the recursive feature elimination process.

5.5.3 NIPS Challenge Problems

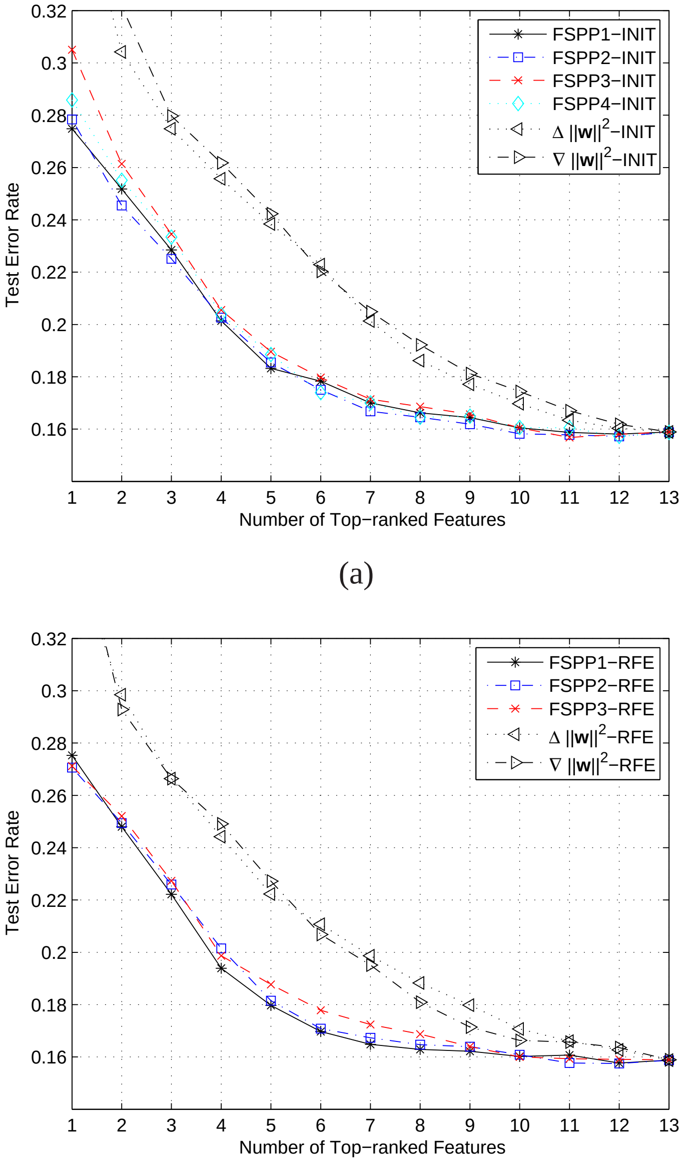
A well-known set of challenging feature-selection problems is that given in the NIPS challenge problems (Guyon, 2003). These problems are known to be difficult and are designed to test various feature-selection methods using an unbiased testing procedure without revealing the labels of the test set. The problem sets ARCENE and MADELON were chosen to evaluate our proposed method. In view of time and space constraints, only the results of FSPP2-RFE are reported. The details of the ARCENE and MADELON datasets are given in Table 5.2. ARCENE is probably the most challenging among all the datasets from the NIPS competition as it is a sparse problem with the smallest examples-to-features ratio (num-of-training-examples/num-of-features=100/10000), while MADELON is a relatively easier problem with a bigger examples-to-features ratio (2000/500). They were chosen to show effectiveness of the proposed methods for both sparse and non-sparse problems.



(a)

(b)

Figure 5.3: Test error rates against top-ranked features on breast cancer dataset where the top-ranked features were chosen based on (a) FSPPm-INIT (b) FSPPm-RFE, m=1,2,3,4. Results of two other methods, $\Delta||\mathbf{w}||^2$ and $\nabla||\mathbf{w}||^2$, were also included. The test error rates shown are the averages over 100 realizations.



(a)

(b)

Figure 5.4: Test error rates against top-ranked features on heart disease dataset where the top-ranked features were chosen based on (a) FSPPm-INIT (b) FSPPm-RFE, m =1, 2, 3, 4. Results of two other methods, $\Delta||\mathbf{w}||^2$ and $\nabla||\mathbf{w}||^2$, were also included. The test error rates shown are the average over 100 realizations.

Table 5.3: Results on NIPS 2003 challenge datasets as of February 01, 2006. (note: BER is the balanced error rate on Dtes, while AUC refers to area under the ROC curve.)

Dataset	**Our best entry by FSPP2-RFE**					**Top entry by other researchers**				
	Rank	BER	AUC	Feat. No.	Probe (%)	Rank	BER	AUC	Feat. No.	Probe (%)
MADELON	1	0.0622	0.9378	12	0	2	0.0622	0.9807	500	96
ARCENE	2	0.106	0.894	5000	27.82	1	0.072	0.9811	100	0

Based on the results of the earlier experiments, FSPP2-RFE was chosen for these two datasets. Our version of FSPP2-RFE used a three-tier removal of features for MADELON: 100 features at each recursion until 100 features were left followed by 20 features at each recursion until 20 features were left and finally one feature at each recursion. A more aggressive removal scheme was used for ARCENE: 1000 features were deleted at each recursion. For each dataset, our result of FSPP2-RFE having the best validation accuracy was chosen. Our entries were respectively ranked 1st and 2nd (as of February 01, 2006) in the MADELON and ARCENE group of entries. A comparison between our results and the best entries by other participants of the challenge (see Guyon, 2003) is given in Table 5.3 (as of February 01, 2006).

5.6 Discussion

In summary, FSPP1-3 performed well for all the artificial datasets. This is to be expected of any good feature-selection method. For the real-world datasets, FSPP1-2 had better performance than FSPP3 with the edge going to FSPP2, especially when small numbers of top-ranked features were used. The excellent performance of FSPP2 in the two NIPS challenge problems reaffirmed its suitability for real-world datasets.

FSPP2-RFE appears to do well on sparse datasets (datasets with large number of features but small training samples), as seen in the experiment associated with the ARCENE problem. The reason for its good performance is not exactly clear, but one possible reason is that the FSPP2 is based on the ensemble of all training examples of $|p(\omega|\mathbf{x}) -$

$p(\omega|\mathbf{x}_{-i})|$ over the feature space, as seen in Equation (5.10). This ensemble over all $\mathbf{x}_j$ is likely to be more accurate in measuring the contribution of a feature and is more robust against decreasing training examples. This is different from other methods that rely on bounds of index of performance where many of these bounds are known to be loose (Rakotomamonjy, 2003; Vapnik, 1998) and its effect could be more severe when the ratio of samples-to-features is low.

One significant advantage of FSPP2 is the modest computations needed for its evaluation. Suppose the SVM output $f(\mathbf{x}_i)$ is available for all $\mathbf{x}_i$ in the training data. The evaluation of $\hat{p}(\omega|\mathbf{x})$ requires a one-time determination of variables A and B from the optimization problem as of Equation (5.8). Since Equation (5.8) is an unconstrained convex optimization problem in two variables, its numerical determination is straight forward (Lin et al., 2003). The random permutation of every feature over the training data is required and it is a simple $O(DN)$ operation which can be done efficiently. Hence, the FSPP2 scales linearly with respect to the number of features or training samples and is suitable for large problems in high dimensions.

The proposed idea of using sensitivity of posterior probabilities for feature selection appears general and should be extendable to other machine learning algorithms where probabilistic outputs are also available.

The idea of using sensitivity of posterior probabilities for feature selection has been demonstrated in the context of two-class classification problem. Possible extensions of the current work could also include the adaptation of the criterion to regression problems and multi-class classification problems where feature selection methods remain rare in the literature. The next chapter of this dissertation will cover such an attempt.

5.7 Concluding Remarks

This chapter introduces a new feature-ranking criterion based on the posterior probability of the SVM output. It is motivated from the advantage gained in using posterior probability as a decision function for classification instead of the direct SVM output function. Four approximations are proposed for the evaluations of the criterion. These approximations are used in two overall feature-selection approaches, recursive feature-elimination approach and initial feature-ranking approach.

The experimental results on various datasets show that three of the four approximations (FSPP1, 2 and 3) yield good overall performance under the recursive feature-elimination approach. Among them, FSPP2 has the overall edge in terms of accuracy and shows performance that is comparable with some of the best methods in the literature. In addition, FSPP2 has modest computation and hence, is suitable for large problems in high dimensional feature space. In addition, it appears to perform well for datasets with low samples-to-features ratios. Consequently, this method is a good candidate for feature selection for SVM applications.

As discussed in Section 2.8, mental fatigue is usually classified into more than two discrete levels in the context of EEG-based mental-fatigue measurement and monitoring system. In this case, a feature-selection method for multi-class classification is required. The next chapter will focus on the endeavor to extend the proposed feature-selection method for two-class classification problems, as presented in this chapter, to a feature selection method for multi-class classification problems.

Chapter 6

Sensitivity of Posterior Probability as a Measure of Feature Importance for Multi-Class Classification Problems

EEG-based mental-fatigue measurement and monitoring is usually formulated into a multi-class classification problem of differentiating mental fatigue at several discrete levels. Therefore, a feature-selection method for multi-class classification problems is required. This chapter proposes two feature-selection criteria for multi-class classification problems based on the sensitivity of the posterior probability of the classifier with respect to the feature. They are extensions of a two-class feature-selection method presented in the previous chapter and are based on two new criteria. In loose terms, each of the two criteria measures the importance of a feature by computing the aggregate value, over the feature space, of the absolute difference of the posterior probabilities of the classifier with and without the feature. In their original form, the evaluations of the criteria are computationally expensive and three approximations are proposed. Using the support vector machine (SVM) multi-class classifier as the working example, performances of the three approximations are tested on several artificial and real-world benchmark datasets. Comparisons are also made with two other popular feature-selection methods from the literature. Numerical results show that the proposed approximations, when used in an overall feature-selection method, generally outperform the two popular feature-selection methods for the datasets considered. In addition, one of the three approximations performs slightly better than the other two.

6.1 Introduction

Relatively little attention has been given to feature selection in a multi-class classification setting in the literature. Most existing feature-selection methods are intended for two-class classification problems. This is not surprising since a multi-class classifier can be implemented by appropriately combining several two-class classifiers and hence feature selection can be done separately for each of the two-class classifiers. However, multi-class feature-selection methods are important and deserve attention for at least two reasons. First, some classifiers, such as neural networks (see Haykin, 1999), logistic regression (see Hosmer and Lemeshow, 1989) and random forests (Breiman, 2001), are naturally multi-class classifiers and thus feature selection has to be performed in the context of multi-class classification. Second, two-class feature-selection methods may not be applicable for some multi-class classifiers built up from two-class classifiers. For example, consider the multi-class classifier built up from several "one-versus-all" (OVA) two-class support vector machine (SVM) classifiers (see Rifkin and Klautau, 2004). Let $f_i(\mathbf{x})$ denote the output of the i^{th} two-class SVM classifier classifying class ω_i against the rest of classes for a given test sample $\mathbf{x}$. The decision rule for OVA multi-class classifier is $\arg\max_i f_i(\mathbf{x})$. This approach implicitly assumes that all two-class classifiers use the same set of features. If different feature subsets are used for the different two-class classifiers, the afore-mentioned decision rule can be wrong since the output of each two-class classifier has its own bias level. This same difficulty exists for multi-class SVM classifier constructed using pair-wise coupling strategy (PWC) (Hastie and Tibshirani, 1998).

New feature-selection methods for multi-class classification problems are considered in this chapter. They are extensions of a two-class feature-selection method as presented in the previous chapter and are based on two new criteria: one when $p(\omega_i|\mathbf{x})$, the posterior probability of belonging to class ω_i given $\mathbf{x}$, is available and the other when only the pairwise posterior probability, $p(\omega_i|\mathbf{x}, \mathbf{x} \in \omega_i \text{ or } \omega_j)$, is available. Both criteria use the sensitivity of the posterior probability with respect to a feature as a measure of feature importance, and are collectively termed Multi-class Feature-based Sensitivity of Posterior Probabilities (MFSPP). In loose terms, both criteria correspond to the aggregate value over the feature space of the absolute difference of the posterior probabilities with and without the feature. Conceptually, these criteria are very different from those considered in the existing feature-selection literature. For example, the typical criteria used for two-class SVM are the sensitivity of the cost function (Guyon and Elisseef, 2003; Guyon et al., 2002), or the sensitivity of bounds on generalization error (Rakotomamonjy, 2003; Weston et al., 2001).

In their original form, the evaluation of the proposed two criteria are expensive and approximations are needed. Three approximations, MFSPP1-3, are used and their performances are tested on various benchmark datasets. Comparisons are also made with several existing multi-class feature-selection methods in the literature. The results show that all three approximations perform consistently well and compare favorably with the other methods considered, with a slight edge going to MFSPP1.

The proposed feature-selection methods require the use of probabilistic classifiers. For the ease of presentation, this work considers only probabilistic classifiers obtained from the SVM methods as they are known to have superior performance (Duan and Keerthi, 2005; Hastie and Tibshirani, 1998; Platt, 2000). It is well known (Platt, 2000) that a probabilistic classifier can be obtained from the output function of a standard two-class SVM classifier. Hastie and Tibshirani (1998) also show a pairwise coupling strategy to combine several probabilistic two-class classifiers to form a single probabilistic multi-class SVM classifier.

The rest of this chapter is organized as follows. Results from the literature needed for the subsequent sections are collected and reviewed in Section **??**. Section 6.3 states the proposed feature-ranking criteria and the descriptions of the three approximations. Section 6.4 outlines the overall feature-selection scheme incorporating the criteria. Extensive experimental results are reported and discussed in Section 6.5, followed by the conclusions in Section 6.6.

6.2 Review of Past Work

The section provides a review of probabilistic multi-class SVM classifier and other closely-related past work on multi-class feature-selection methods. We begin with the general notations used. Consider a prototypical multi-class classification problem having c classes $(\omega_1, \omega_2, \cdots, \omega_c)$ and a given dataset $\mathbb{D}$ in the form of $\{\mathbf{x}_i, y_i\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^D$ is the i^{th} sample, $y_i \in \{1, \cdots, c\}$ is the corresponding class label. Hence, $y_i = k$ if and only if $\mathbf{x}_i \in \omega_k$. Let n_k be the number of samples that belong to class ω_k, $N := N_1 + \cdots + N_c$ be the total number of samples in $\mathbb{D}$ and $\mathbb{D}_{ij} := \{\mathbf{x}_k, y_k\}_{\mathbf{x}_k \in \omega_i \cup \omega_j}$ be the subset of $\mathbb{D}$ formed by samples from classes ω_i and ω_j. Also, let $\mathbf{x}^i$ denote the i^{th} feature of vector $\mathbf{x}$. Hence, $\mathbf{x}_j^i$ refers to the i^{th} feature of the j^{th} sample and $\mathbf{x}_{-i} \in \mathbb{R}^{D-1}$ is the vector obtained from $\mathbf{x}$ with the i^{th} feature removed. Double subscripted variable $\mathbf{x}_{-i,j}$ is also used and it refers to the j^{th} sample of vector $\mathbf{x}_{-i}$. In addition, $p_i(\mathbf{x}) \equiv P(\omega_i|\mathbf{x})$ refers to the posterior probability of belonging to class ω_i given $\mathbf{x}$ and $\hat{p}_i(\mathbf{x})$ is used to denote its estimate. Similarly, $p_{ij}(\mathbf{x}) \equiv P(\omega_i|\mathbf{x}, \mathbf{x} \in \omega_i \text{ or } \omega_j)$ refers to the pairwise probability of belonging to class ω_i knowing that $\mathbf{x}$ is from class ω_i or class ω_j and $\hat{p}_{ij}(\mathbf{x})$ is its estimate.

6.2.1 Probabilistic Multi-Class SVM

The description of the probabilistic multi-class SVM has been given in Section 2.10.2.4 in Chapter 2. Here, only key equations are given below for easy reference.

If $p_i(\mathbf{x})$ of a probabilistic multi-class SVM classifier is available, the decision function is

$$d(\mathbf{x}) = \arg\max_i \{p_i(\mathbf{x})\}. \tag{6.1}$$

Typically, $p_i(\mathbf{x})$ is estimated by $\hat{p}_i(\mathbf{x})$, obtained from solving the following pairwise-coupling (PWC) optimization problem (Hastie and Tibshirani, 1998; Wu et al., 2004):

$$\min_{\hat{p}_i(\mathbf{x})} \sum_{i=1}^{c} \sum_{j:j\neq i} \left[\hat{p}_{ji}(\mathbf{x})\hat{p}_i(\mathbf{x}) - \hat{p}_{ij}(\mathbf{x})\hat{p}_j(\mathbf{x})\right]^2, \text{ subject to} \sum_{i=1}^{c} \hat{p}_i(\mathbf{x}) = 1. \tag{6.2}$$

where $\hat{p}_{ij}(\mathbf{x}), \hat{p}_{ji}(\mathbf{x})$ are known probabilistic outputs of the two-class SVM classifiers (Platt, 2000). Specifically, suppose the standard output of the two-class SVM trained using $\mathbb{D}_{ij}$ is

$$f_{ij}(\mathbf{x}) = \sum_{\mathbf{x}_k \in \mathbb{D}_{ij}} y_k \alpha_k K(\mathbf{x}_k, \mathbf{x}), \tag{6.3}$$

where $K(\cdot,\cdot)$ is the kernel function (Boser et al., 1992; Cortes and Vapnik, 1995; Vapnik, 1995). The probabilistic SVM output, $\hat{p}_{ij}(\mathbf{x})$, is

$$\hat{p}_{ij}(\mathbf{x}) = \frac{1}{1 + \exp\left(A_{ij} f_{ij}(\mathbf{x}) + B_{ij}\right)}, \tag{6.4}$$

where the parameters A_{ij} and B_{ij} are determined from minimizing the negative log likelihood (or the cross-entropy error function) function, or

$$\min F(A_{ij}, B_{ij}) = \min\{-\sum_{\mathbf{x}_k \in \mathbb{D}_{ij}} [t_k \log \hat{p}_{ij}(\mathbf{x}_k) + (1-t_k)\log(1-\hat{p}_{ij}(\mathbf{x}_k))]\}, \qquad (6.5)$$

where $t_k = (N_i+1)/(N_i+2)$ if $y_k = i$ and $t_k = 1/(N_j+2)$ if $y_k = j$. It is worth noting that a 5-fold cross-validation process is implicitly used in Equation (6.5) as suggested by (Platt, 2000). This cross-validation process removes the requirement of keeping a hold-out validation dataset for fitting the parameters A_{ij} and B_{ij}, which is especially useful when the number of training samples is small. Our implementation of the above includes the modifications suggested by Lin et al. (Lin et al., 2003) for numerical stability. Obviously, the choice of kernel function $K(\cdot,\cdot)$ in Equation (6.3) is general and our study is done with the popular Gaussian kernel, $K(\mathbf{x}_p, \mathbf{x}_q) = \exp(-\gamma\|\mathbf{x}_p - \mathbf{x}_q\|^2)$ where γ is the kernel parameter.

The above procedure of obtaining $\hat{p}_i(\mathbf{x})$ from $\hat{p}_{ij}(\mathbf{x})$ is hereafter referred as PWC-PSVM. Both quantities, $\hat{p}_i(\mathbf{x})$ from Equation (6.2) and $\hat{p}_{ij}(\mathbf{x})$ from Equation (6.4) are used extensively in the approximations of the proposed feature-ranking criterion hereafter.

6.2.2 Other Feature-Selection Methods for SVM

For comparison purposes, three other feature-selection methods for multi-class classification problems are reviewed below.

6.2.2.1 Multi-Class Version of Fisher's Score

The multi-class version of Fisher's score (F-Score) assesses the importance of each feature independent of other features by the ratio of between-class variances and within-class variances. The F-score for the k^{th} feature is

$$F(k) = \frac{\sum_{i=1}^{c} N_i(\boldsymbol{\mu}_i^k - \boldsymbol{\mu}^k)^2}{\sum_{i=1}^{c} \sum_{\mathbf{x}_j \in \omega_i} (\mathbf{x}_j^k - \boldsymbol{\mu}_i^k)^2}, \tag{6.6}$$

where $\boldsymbol{\mu}_i^k := \sum_{\mathbf{x}_j \in \omega_i} \mathbf{x}_j^k$ is the i^{th} class mean of the k^{th} feature and $\boldsymbol{\mu}^k := \sum_{i=1}^{c} (N_i \boldsymbol{\mu}_i^k)/n$ is the mean over all class means.

6.2.2.2 Multi-Class Versions of SVM-RFE algorithm

The SVM Recursive-Feature-Elimination algorithm (SVM-RFE) is a well-known feature-selection method for two-class SVM classifiers (Guyon et al., 2002). It uses the sensitivity of the cost function of two-class SVM classifier with respect to each feature as the feature-ranking criterion and the least important feature is recursively removed from the training data between successive training of the classifier. Such a procedure, combined with multi-class SVM under the OVA setting (SVM-OVA-RFE), was suggested by Weston et al. (2003) based on its feature-ranking criterion

$$D(k) = \sum_{i=1}^{c} |J_i - J_{-k,i}|, \tag{6.7}$$

where $J_i := \frac{1}{2}\|\mathbf{w}_i\|^2 + C\sum_j \xi_j$ is the cost function of the two-class SVM classifier between class ω_i and the rest of classes, and $J_{-k,i}$ is the corresponding cost function after the k^{th} feature is removed. Assuming no change in solution of the SVM, Equation (6.7) is shown (Weston et al., 2003) to be equivalent to

$$\hat{D}(k) = \sum_{i=1}^{c} \left| \|\mathbf{w}_i\|^2 - \|\mathbf{w}_{-k,i}\|^2 \right|, \tag{6.8}$$

where $\mathbf{w}_i$ and $\mathbf{w}_{-k,i}$ are the respective normals to the separating hyperplanes of the SVM classifiers.

Based on Equation (6.7) or Equation (6.8), it is straight-forward to extend the SVM-RFE algorithm to "one-versus-one" (OVO) multi-class SVM (SVM-OVO-RFE) by using the following feature-ranking criterion:

$$\hat{D}^p(k) = \sum_{i=1}^{c} \sum_{j:j\neq i} \left| \|\mathbf{w}_{ij}\|^2 - \|\mathbf{w}_{-k,ij}\|^2 \right|, \tag{6.9}$$

where $\mathbf{w}_{ij}$ is the normal to the separating hyperplane of the two-class SVM classifier classifying class ω_i against class ω_j and $\mathbf{w}_{-k,ij}$ is the corresponding vector obtained after the k^{th} feature is removed.

6.3 The Proposed Criteria

Consider the prototypical c-class classification problem with posterior probability $p_i(\mathbf{x})$ and $p_{ij}(\mathbf{x})$. Let $\mathbf{x}_{-k} \in \mathbb{R}^{D-1}$ be the vector derived from $\mathbf{x}$ with the k^{th} feature removed. The proposed ranking-criterion for the k^{th} feature is

$$S(k) = \sum_{i=1}^{c} \int \lambda_i |p_i(\mathbf{x}) - p_i(\mathbf{x}_{-k})| p(\mathbf{x}) d\mathbf{x}, \tag{6.10}$$

where

$$\lambda_i \geq 0, \ \ i = 1, \cdots, c \text{ and } \sum_{i=1}^{c} \lambda_i = 1. \tag{6.11}$$

The motivation of above criterion is clear: the greater the weighted absolute difference between $p_i(\mathbf{x})$ and $p_i(\mathbf{x}_{-k})$ over the space of $\mathbf{x}$, the more important is the k^{th} feature. The λ_i are introduced to account for the different emphases placed on the sensitivity of posterior probabilities for the various classes. One useful choice is to let

$$\lambda_i = \frac{\beta}{N_i} \text{ with } \beta = (\sum_{i=1}^{c} \frac{1}{N_i})^{-1} \tag{6.12}$$

so as to avoid the key features of a majority class from dominating the ranking in a multi-class feature-selection setting.

The evaluation of $S(k)$ requires the availability of $p_i(\mathbf{x})$. If $p_i(\mathbf{x})$ are not available but only $p_{ij}(\mathbf{x})$ are, a second proposed feature-ranking criterion is

$$S^p(k) = \sum_{i=1}^{c} \sum_{j=1, j\neq i}^{c} \int \lambda_{ij} |p_{ij}(\mathbf{x}) - p_{ij}(\mathbf{x}_{-k})| p(\mathbf{x}|\mathbf{x} \in \omega_i \text{ or } \omega_j) d\mathbf{x} \tag{6.13}$$

where λ_{ij} play the same role as λ_i in Equation (6.10) and are subject to similar constraints:

$$\lambda_{ij} \geq 0 \text{ for all } (i, j) \text{ with } i \neq j, \ \lambda_{ij} = \lambda_{ji}, \ \sum_{i\neq j} \lambda_{ij} = 1. \tag{6.14}$$

As the true values of $p_i(\mathbf{x})$ and $p_{ij}(\mathbf{x})$ in Equation (6.10) and Equation (6.13) are unknown, they are approximated by $\hat{p}_i(\mathbf{x})$ and $\hat{p}_{ij}(\mathbf{x})$ obtained by solving Equations (6.2) and (6.5) respectively. The values of $p_i(\mathbf{x}_{-k})$ in Equation (6.10) and $p_{ij}(\mathbf{x}_{-k})$ in Equation (6.13) correspond to the probabilistic outputs of PWC-PSVM trained using data $\{\mathbf{x}_{-k,i}, y_i\}_{i=1}^n$ in place of $\{\mathbf{x}_i, y_i\}_{i=1}^n$. Since $\mathbf{x}$ has D features, to obtain $p_i(\mathbf{x}_{-k})$ for $k = 1, \cdots, D$ means that training of PWC-PSVM has to be performed D times, each time with one feature removed from the training set. This is a computationally expensive process. The remainder of this section shows three approximations (MFSPP1–MFSPP3) of Equation (6.10) or Equation (6.13) which avoid the retraining process.

The approximations of $p_i(\mathbf{x}_{-k})$ and $p_{ij}(\mathbf{x}_{-k})$ in Equation (6.10) and Equation (6.13) involve a process of random permutation (RP) of the values of a feature as discussed in the previous chapter. Specifically, the values of the k^{th} feature of $\mathbf{x}$ are randomly permuted over the n samples of a dataset while all other features of $\mathbf{x}$, except $\mathbf{x}^k$, remain unchanged. Suppose $\{\zeta_1, \cdots, \zeta_{N-1}\}$ is a set of uniformly distributed random numbers from $(0,1)$ and $\lfloor \zeta \rfloor$ is the largest integer that is less than ζ. The random permutation of the values of the k^{th} feature is executed as follows (Page, 1967): for each i starting from 1 to $n-1$, compute $j = \lfloor N \times \zeta_i \rfloor + 1$ and swap the values of $\mathbf{x}_i^k$ and $\mathbf{x}_j^k$.

Let $\mathbf{x}_{(k)} \in \mathbb{R}^D$ be the vector derived from $\mathbf{x}$ with the k^{th} feature randomly permuted by the RP process. The next theorem states a result on $p_i(\mathbf{x})$ and $p_{ij}(\mathbf{x})$ following the RP process and serves as the theoretical basis for the approximations. The proof of which has been given in the proof of Theorem 5.1 in the context of a two-class classification problem but is included in Appendix C. for easy reference.

Theorem 6.1 *Suppose $p_i(\mathbf{x})$ is the posterior probabilities of $\mathbf{x}$ belonging to class ω_i and $p_{ij}(\mathbf{x})$ is the posterior probabilities of belonging to class ω_i given that $\mathbf{x} \in \omega_i \cup \omega_j$. Then,*

$$p_i(\mathbf{x}_{-k}) = p_i(\mathbf{x}_{(k)}) \tag{6.15}$$

and

$$p_{ij}(\mathbf{x}_{-k}) = p_{ij}(\mathbf{x}_{(k)}). \tag{6.16}$$

Theorems 6.1 shows that the RP process has the same effect as removing the contribution of that feature for classification. Using this fact, the criterion of Equation (6.10) can be equivalently stated as

$$S(k) = \sum_{i=1}^{c} \int \lambda_i |p_i(\mathbf{x}) - p_i(\mathbf{x}_{(k)})| p(\mathbf{x}) d\mathbf{x}, \tag{6.17}$$

while Equation (6.13) is equivalent to

$$S^p(k) = \sum_{i=1}^{c} \sum_{j=1:j\neq i}^{c} \int \lambda_{ij} |p_{ij}(\mathbf{x}) - p_{ij}(\mathbf{x}_{(k)})| p(\mathbf{x}|\mathbf{x} \in \omega_i \text{ or } \omega_j) d\mathbf{x}. \tag{6.18}$$

Based on Equation (6.17) and Equation (6.18), the three approximations are stated next.

MFSPP1: As mentioned earlier, $p_i(\mathbf{x})$ is not known exactly and is approximated by $\hat{p}_i(\mathbf{x})$ of Equation (6.2), trained using dataset $\mathbb{D}$. As for $p_i(\mathbf{x}_{(k)})$, it is approximated by the same $\hat{p}_i$ expression of Equation (6.2) but with $\mathbf{x}$ replaced by $\mathbf{x}_{(k)}$. This means that no retraining of the classifier is involved in the approximation of $p_i(\mathbf{x}_{(k)})$ by $\hat{p}_i(\mathbf{x}_{(k)})$. Further approximation of the integration over the $\mathbf{x}$ space in Equation (6.17) yields

$$\hat{S}_1(k) = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \lambda_i |\hat{p}_i(\mathbf{x}_j) - \hat{p}_i(\mathbf{x}_{(k),j})|, \tag{6.19}$$

where $\mathbf{x}_{(k),j}$ refers to the j^{th} sample of the training data where the k^{th} feature has been randomly permuted.

MFSPP2: In a manner similar to MSFPP1, Equation (6.18) becomes

$$\hat{S}^p(k) = \sum_{i=1}^{c} \sum_{j=1:j\neq i}^{c} \sum_{\mathbf{x}_p \in \mathbb{D}_{ij}} \frac{1}{N_i + N_j} \lambda_{ij} |\hat{p}_{ij}(\mathbf{x}_p) - \hat{p}_{ij}(\mathbf{x}_{(k),p})|. \tag{6.20}$$

The criteria of Equation (6.17) and Equation (6.18) assess the importance of a feature by the sensitivity of $p_i(\mathbf{x})$ or $p_{ij}(\mathbf{x})$ with respect to that feature. Obviously, this idea can be applied to other multi-class classification methods so long as probabilistic outputs are available. For example, one can obtain another approximation of Equation (6.17) for the OVA multi-class SVM. In OVA multi-class SVM, c two-class SVM classifiers are constructed where the i^{th} two-class classifier is for separating class ω_i from the rest of the classes. One suggested (Duan et al., 2003) simple estimate of $p_i(\mathbf{x})$ is:

$$\bar{p_i}(\mathbf{x}) = \tilde{p}_i(\mathbf{x}) / \sum_{j=1}^{c} \tilde{p}_j(\mathbf{x}) \tag{6.21}$$

where

$$\tilde{p}_j(\mathbf{x}) = 1/[1 + \exp\left(A_j f_j(\mathbf{x}) + B_j\right)] \tag{6.22}$$

is the probabilistic output obtained from $f_j(\mathbf{x})$, the standard output of the i^{th} two-class classifier, in a similar manner to Equation (6.4) and Equation (6.5).

MFSPP3: Using Equation (6.21) in Equation (6.17), a feature-ranking criterion for OVA multi-class SVM is

$$\hat{S}_2(k) = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \lambda_i |\bar{p_i}(\mathbf{x}_j) - \bar{p_i}(\mathbf{x}_{(k),j})|. \tag{6.23}$$

Clearly, other more sophisticated schemes (see Duan et al., 2003; Roth, 2001) for obtaining posterior probabilities from standard outputs of OVA multi-class SVM can also be used.

6.4 Feature Selection Method

This section describes the overall scheme of multi-class feature selection using MFSPP1-MFSPP3. The overall scheme follows the well-known recursive-feature-elimination (RFE) approach (Guyon et al., 2002), in which the least important feature, as ranked by MFSPP1-MFSPP3, is removed from successive SVM training. Accordingly, the overall scheme is referred to as MFSPP1-RFE to MFSPP3-RFE. It is assumed that estimates of $p_i(\mathbf{x})$ and $p_{ij}(\mathbf{x})$ are available under the formulation of PWC-PSVM or other probabilistic SVMs. This also implies that all the hyper-parameters of SVMs, (C,γ), have been determined through a proper model selection process, followed by a subsequent determination of (A_{ij},B_{ij}) for Equation (6.4) or (A_i,B_i) for Equation (6.22).

As a review, the steps involved in the MFSPP1-RFE scheme are summarized in Algorithm 4. This scheme has its inputs the dataset $\mathbb{D}$ and the index set $I = \{1,2,\cdots,D\}$ containing all the indices of features. The output is a ranked list of features in the form of an index set $J_R = \{j_1, j_2, \cdots, j_D\}$ where $j_k \in I$ for each k. The scheme starts with the full feature set I. The while loop is invoked, which trains the PWC-PSVM classifier and gets a ranked list of the features J_l containing all elements in I. Next, the last element of J_l (corresponding to the feature having the smallest $\hat{S}_1$) is removed from I and stored in the rightmost position of the ranked list J_R. The while loop is then invoked on the reduced set I. This process continues, each time removing the least important feature from I and storing it in the rightmost free position of J_R, until I is empty.

With a slight modification to steps 5 and 6 in Algorithm 4, one can easily get the steps involved in MFSPP2-RFE and MFSPP3-RFE. It is also worth noting that more than one feature can be removed at one time with a slight modification to steps 5 and 6 of Algorithm 4 and that the current description of feature-selection scheme does not involve the determination of hyper-parameters (C,γ)'s in the step 4 in the while loop of Algorithm 4. Such a process is possible albeit with higher costs.

Algorithm 6.1: Main steps of MFSPP1-RFE feature selection scheme.

```
Input: $\mathbb{D}$, $I$
Output: Ranked list $J_R$
while $I \neq \emptyset$ do
    set $l = \text{size}(I)$;
    if $l > 1$ then
        Train PWC-PSVM using $\mathbb{D}$;
        For each $i \in I$, compute $\hat{S}_1(i)$ via Equation (6.19);
        Put each $j_k \in I$ into a rank list $J_l = \{j_1, j_2, \cdots, j_l\}$ satisfying
        $\hat{S}_1(j_k) \geq \hat{S}_1(j_{k+1})$ for $k = 1, \cdots, l-1$;
        Let the last element of $J_l$ be $k^\star$;
        Assign $k^\star$ to the $l^{th}$ element of $J_R$;
        else
            Asign the only element in $I$ to the $l^{th}$ element of $J_R$;
        end
    end
    Let $I = I \setminus k^\star$ and remove feature $k^\star$ from every sample in $\mathbb{D}$.
end
```

Table 6.1: Basic information of the four real-world benchmark problems used in the present study

Problem	#class	#realization	#feature	#training	#testing
wine	3	100	13	125	53
lung cancer	3	100	56	22	10
waveform	3	100	40	3500	1500
DNA	3	1	180	2000	1186

6.5 Experiments and Discussions

Extensive experiments on both artificial and real-world benchmark problems are carried out using the proposed methods. Like other studies, artificial problems are used because the key features are known, making comparative study easy. Four real-world benchmark problems from UCI repository of machine learning datasets (Newman et al., 1998) and the Statlog collection (Michie et al., 1994) are chosen to serve as references for comparison. Descriptions of these are given in Table 6.1.

It is important to note that the purpose of the experiments is not to compare the performances of different multi-class SVMs (excellent work on this can be found in Duan and Keerthi, 2005; Hsu and Lin, 2002; Rifkin and Klautau, 2004), but to measure the effectiveness of proposed multi-class feature-selection methods. To do so, feature sets from different feature-selection methods are used and their performances are evaluated using one consistent multi-class classifier: PWC-PSVM.

The experiment for each dataset uses two data subsets, $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$, for training and testing purposes. $\mathbb{D}_{tra}$ is normalized to be zero mean and unit standard deviation and $\mathbb{D}_{tes}$ is correspondingly adjusted using the normalization parameters of $\mathbb{D}_{tra}$. The normalized $\mathbb{D}_{tra}$ is used for training of each binary SVM classifier including the determination of the optimal hyperparameters (C,γ) and the fitting of sigmoid functions for its probabilistic outputs. The parameters (C,γ) for each binary SVM are selected by 5-fold cross-validation over the following grid: $[2^{-7},\cdots,2^{7}]\times[2^{-10},\cdots,2^{3}]$ and is done separately for each binary SVM using LibSVM (Hsu et al., 2004). Obviously, $\mathbb{D}_{tes}$ is used only to obtain the unbiased test accuracy rate of the underlying method.

For each datasets in Table 6.1 except the DNA, the 100 realizations are generated by random (stratified) splitting of the total samples into $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$ in the ratio of 70% to 30%. The choice of one realization for the DNA dataset is due to the high computational cost involved and missing entries in the lung-cancer dataset are filled by the mode values of their respective classes. The parameters (C,γ) for each binary SVM classifier are chosen as the median of five sets of (C,γ) of the first five realizations. Here, each set of (C,γ) is obtained by 5-fold cross-validation for one realization (Rätsch et al., 2001). In all experiments, λ_i follows that given by Equation (6.12) and $\lambda_{ij}=\lambda_{ji}=\frac{\beta}{N_i+N_j}$ with $\beta=(\sum_{i,j:i\neq j}\frac{1}{N_i+N_j})^{-1}$.

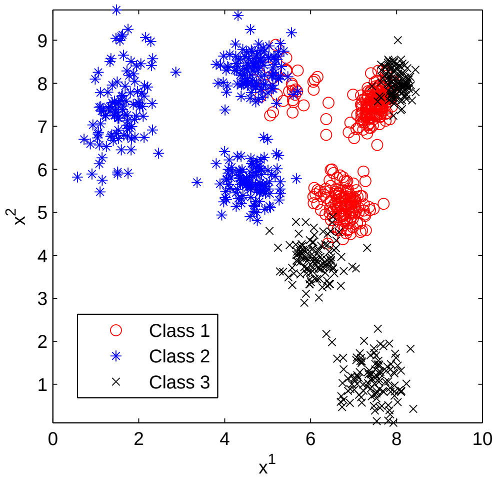


Figure 6.1: The distribution of first two features in the three-class nonlinear synthetic problem, with the data in each class generated from a mixture of Gaussians.

6.5.1 Artificial Problem

The first experiment involves a three-class version of Weston's nonlinear problem (Weston et al., 2001) where the samples for each class are generated from a mixture of Guassians. Following the procedure by Weston et al. (2001), 10,000 samples are generated with 10 features each for the three classes. Only the first two features, $(\mathbf{x}^1, \mathbf{x}^2)$, are relevant while the rest are random noise, each taken from a normal distribution $N(0,20)$. Figure 6.1 shows the distribution of the first two features. Note that a similar dataset has been used in (Hastie and Tibshirani, 1998).

The experiment aims to study the effect of sparsity of the training set on the performance of various feature-selection methods. Four sizes (30, 50, 70 and 100) of $\mathbb{D}_{tra}$ are chosen. For each size, 100 realizations of $\mathbb{D}_{tra}$ are obtained by random selection from the 10,000 samples. When one realization of $\mathbb{D}_{tra}$ is selected, the rest of the 10,000 samples are used for $\mathbb{D}_{tes}$.

Two sets of the SVM parameters, (C,γ), are used. The first (set I) is chosen by the aforementioned method using the first five realizations of $\mathbb{D}_{tra}$. The second (set II) is tuned

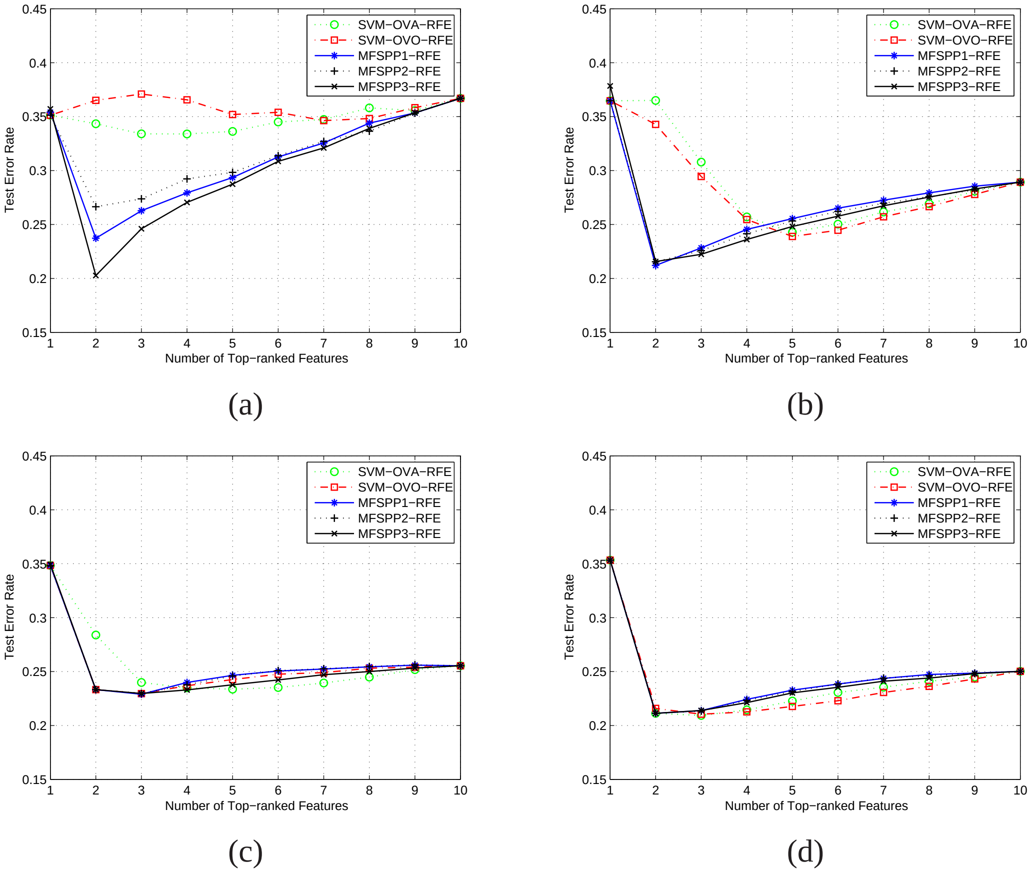
Table 6.2: Mean and standard deviation of test errors on the three-class version of Weston's nonlinear problem using different feature-selection methods and different training set sizes. The numbers in brackets are the percentage of runs that $(\mathbf{x}^1, \mathbf{x}^2)$ are successfully identified as the first two most-important features by each feature-selection method over 100 realizations. Two settings of parameters (C, γ) are considered: (I)the median of five sets of (C, γ) resulting from a 5-fold cross-validation process on each of the first five realizations of $\mathbb{D}_{tra}$; (II) a 5-fold cross-validation process on the randomly-selected 3,000 samples.

		Training Set Size			
	Method	30	50	70	100
(I)	SVM-OVA-RFE	0.34±0.07 (12%)	0.37±0.03 (1%)	0.28±0.06 (58%)	**0.21±0.02 (100%)**
	SVM-OVO-RFE	0.37±0.02 (0%)	0.34±0.07 (19%)	**0.23±0.02 (100%)**	0.22±0.03 (97%)
	MFSPP1-RFE	**0.20±0.08 (83%)**	0.22±0.04 (96%)	**0.23±0.02 (100%)**	**0.21±0.02 (100%)**
	MFSPP2-RFE	0.27±0.11 (53%)	**0.21±0.03 (98%)**	**0.23±0.02 (100%)**	**0.21±0.02 (100%)**
	MFSPP3-RFE	0.24±0.10 (67%)	0.22±0.04 (96%)	**0.23±0.02 (100%)**	**0.21±0.02 (100%)**
(II)	SVM-OVA-RFE	0.27±0.13 (69%)	0.21±0.07 (93%)	0.17±0.04 (99%)	**0.16±0.02 (100%)**
	SVM-OVO-RFE	0.32±0.07 (24%)	0.26±0.09 (60%)	0.21±0.08 (84%)	0.18±0.06 (98%)
	MFSPP1-RFE	0.26±0.08 (60%)	**0.20±0.04 (98%)**	**0.17±0.02 (100%)**	**0.16±0.02 (100%)**
	MFSPP2-RFE	0.24±0.08 (75%)	**0.20±0.04 (98%)**	**0.17±0.02 (100%)**	**0.16±0.02 (100%)**
	MFSPP3-RFE	**0.24±0.08 (79%)**	**0.20±0.04 (98%)**	**0.17±0.02 (100%)**	**0.16±0.02 (100%)**

using randomly chosen 3,000 samples (1,000 for each class) from the 10,000 training samples and this set is used for all four sizes of the training set. This second set is taken to be the optimal parameter values and serves as a reference to decouple the effect of a wrong choice of (C, γ) from the effect of different feature-selection methods.

Table 6.2 shows the means and the standard deviations of the test errors over the 100 realizations when only the two highest-ranked features are used. In addition, Figure 6.2 and Figure 6.3 show plots of the mean of the test-error rates versus the number of top-ranked features used. The two figures differ in the choice of (C, γ) used: Set I for Figure 6.2 while Set II for Figure 6.3. Besides MFSPP1-3, results of SVM-OVA-RFE and SVM-OVO-RFE (existing wrapper methods reviewed in Section II.B) are also included for comparison purposes. The F-score measure is a filter method and is not included for comparison for artificial dataset as it is not very meaningful.

From Table 6.2, Figs 6.2 and 6.3, it is easy to see that MFSPP1-RFE, MFSPP2-RFE and MSPP3-RFE perform at least as good, if not better, than SVM-OVA-RFE and



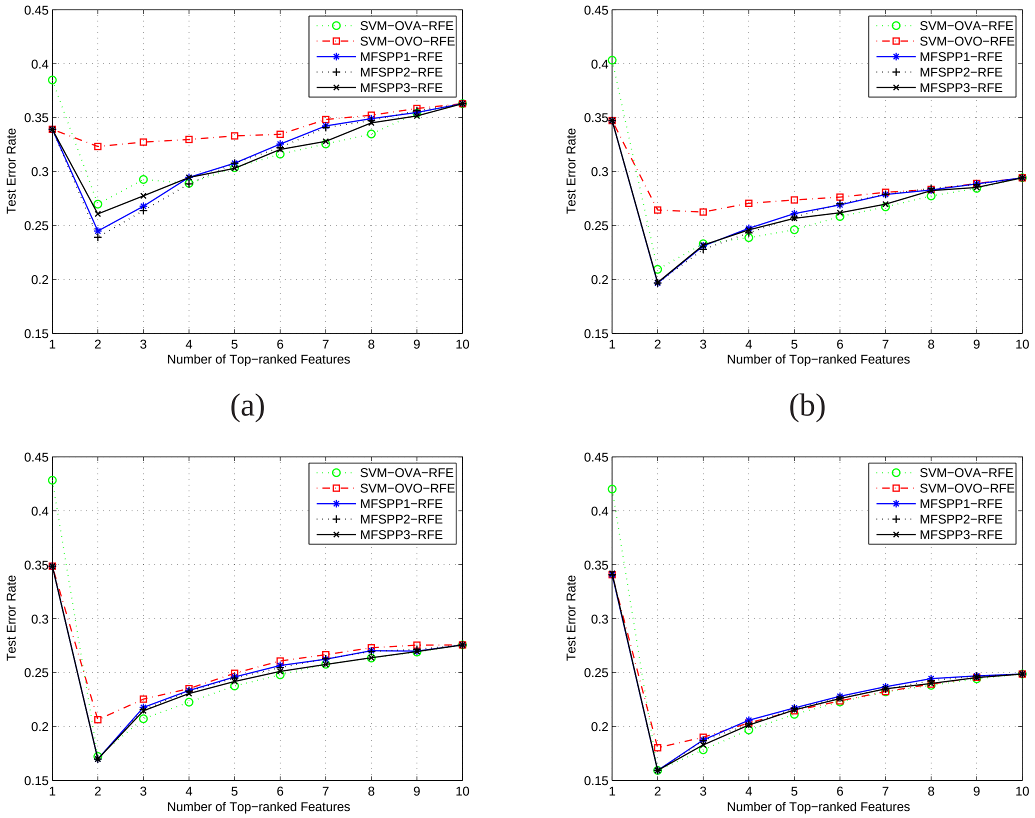
(a)

(b)

(c)

(d)

Figure 6.2: Average test-error rates against top-ranked features over 100 realizations of the three-class version of Weston's nonlinear problem for four training set sizes: (a) 30 samples; (b) 50 samples; (c) 70 samples; (d) 100 samples. The set I of parameters (C, γ) are used and they are chosen as the median of five sets of (C, γ) resulting from a 5-fold cross-validation process on each of the first five realizations of $\mathbb{D}_{tra}$.



(a) (b)

(c) (d)

Figure 6.3: Average test-error rates against top-ranked features over 100 realizations of the three-class version of Weston's nonlinear problem for four training set sizes: (a) 30 samples; (b) 50 samples; (c) 70 samples; (d) 100 samples. The set II of parameters (C, γ) are used and they are chosen chosen by a 5-fold cross-validation process on the randomly-selected 3,000 samples.

SVM-OVO-RFE. The advantage of MFSPP1-RFE, MFSPP2-RFE and MSPP3-RFE over SVM-OVA-RFE and SVM-OVO-RFE is evident when the feature-selection problem becomes more challenging (as the size of training set gets smaller). The statistical significance of the difference is verified by additional paired t-tests. Six paired t-tests on the test-error rates (over 100 realizations when only the first two top-ranked features are provided to the predictor) between the proposed methods (MFSPP1-RFE, MFSPP2-RFE and MSPP3-RFE) and the benchmark methods (SVM-OVA-RFE and SVM-OVO-RFE) are conducted. Most of the resulting P-values are less than 0.05 (especially when the training set size is small, i.e. 30 or 50)—a clear indication of the statistical significance on the advantage of the proposed methods over the benchmark methods.

On the other hand, the performances of the three proposed ranking criteria (MFSPP1-RFE, MFSPP2-RFE and MSPP3-RFE) are very similar. Additional paired t-tests on the test-error rates (over 100 realizations when only the two highest-ranked features are used) confirm that no significant difference exists in the performances of these three methods on this problem.

6.5.2 Real-World Benchmark Problems

The real-world benchmark problems and their respective realizations used in the experiments are given in Table 6.1. Figures 6.4, 6.5 and 6.6 show the average test-error rates for the proposed methods (MFSPP1-RFE, MFSPP2-RFE and MFSPP3-RFE) and the benchmark methods (F-Score, SVM-OVA-RFE and SVM-OVO-RFE) on all real-world benchmark datasets except DNA. It is evident that, at almost all values of top-ranked feature used, the performances of the proposed methods are better than those of the benchmark. Also, the best-performing method appears to be MFSPP1-RFE, it gives the lowest test-error rate using the smallest number of top-ranked features.

The statistical significance of the above-mentioned performance difference is also confirmed by additional paired t-tests for datasets having 100 realizations. Tables 6.3, 6.4 and 6.5 show the P-values for comparisons between the best-performing method, MFSPP1-RFE, and the other methods on all real-world benchmark datasets except DNA. Additional six sets of paired t-tests on the test-error rates between the proposed methods (MFSPP1-RFE, MFSPP2-RFE and MSPP3-RFE) and the benchmark methods (SVM-OVA-RFE and SVM-OVO-RFE) are also conducted and they show that all the three proposed methods perform at least as good, if not better, than the benchmark methods.

The result for the DNA dataset is shown in Figure 6.7. As shown, the proposed set of methods generally give lower test-error rates than the benchmark methods. Among all the methods, MFSPP1-RFE again produces the lowest error rate using the smallest number of top-ranked features.

6.5.3 Discussion

The difference in performances of the proposed methods (MFSPP1-RFE, MFSPP2-RFE and MSPP3-RFE) to the other methods (SVM-OVA-RFE and SVM-OVO-RFE) is interesting and deserves attention. Both groups of methods use the RFE approach but differ in the choice of ranking criteria used: conceptually the former uses the sensitivity of the posterior probability with respect to a feature, $\frac{\partial p_i(\mathbf{x})}{\partial \mathbf{x}^k}$ while the latter uses the sensitivity of the SVM cost function with respect to a feature, $\frac{\partial J(\mathbf{x})}{\partial \mathbf{x}^k}$ or equivalently, $\frac{\partial \|\mathbf{w}\|^2}{\partial \mathbf{x}^k}$. The better performances of the proposed methods appear to suggest that $\frac{\partial p_i(\mathbf{x})}{\partial \mathbf{x}^k}$ is a better measure of importance of the k^{th} feature than $\frac{\partial \|\mathbf{w}\|^2}{\partial \mathbf{x}^k}$. The exact reasons for this are not entirely clear but several possibilities exist.

One possibility relates to the measure of performance. Testing accuracy is used to determine the superiority of one feature selection method over another and this accuracy

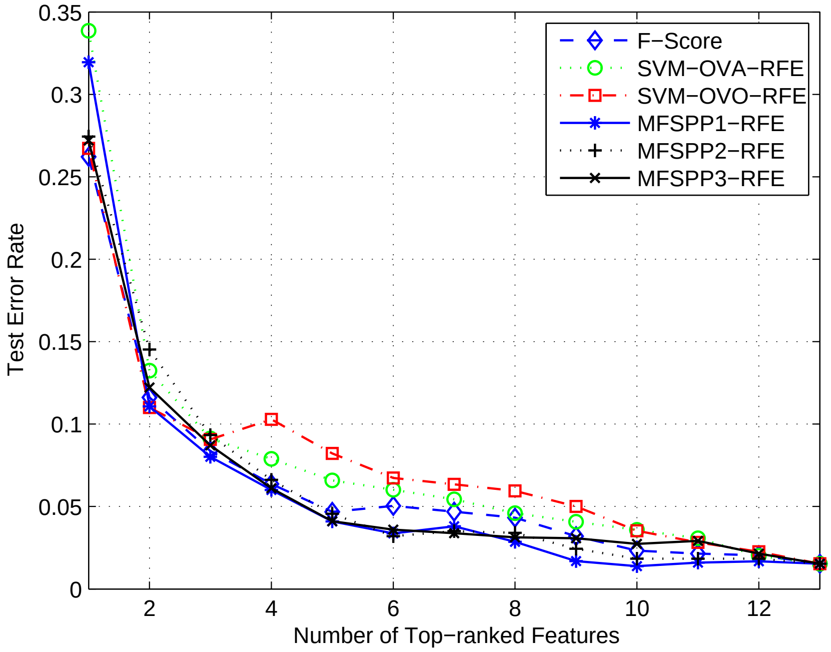


Figure 6.4: Average test-error rates against top-ranked features over 100 realizations of the wine dataset.

Table 6.3: Performance comparison between the best-performing method (i.e. MFSPP1-RFE) and the other methods (F-Score, SVM-OVA-RFE, SVM-OVO-RFE, MFSPP1-RFE, MFSPP2-RFE) on the wine dataset. The *P*-value is obtained in the paired *t*-test between each method to the best-performing method MFSPP1-RFE. The symbols “$^{+}$” and “$^{-}$” indicate statistically significant wins or losses over MFSPP1-RFE (*P*-value $<$ 0.05).

N_{top}	MFSPP1 -RFE	F-Score		SVM-OVA -RFE		SVM-OVO -RFE		MFSPP2 -RFE		MFSPP3 -RFE	
	Mean TER (%)	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value
1	31.9	26.2	0.34	33.9	0.06	26.7	0.51	27.4	0.37	27.2	0.63
2	11.1	11.6	0.16	13.2	0.00^{-}	11.0	0.90	14.5	0.00^{-}	12.2	0.01^{-}
3	8.0	8.4	0.28	9.1	0.00^{-}	9.1	0.01^{-}	9.3	0.00^{-}	8.7	0.02^{-}
4	6.0	6.4	0.19	7.9	0.00^{-}	10.3	0.00^{-}	6.6	0.01^{-}	6.1	0.64
5	4.1	4.7	0.00^{-}	6.6	0.00^{-}	8.2	0.00^{-}	4.6	0.00^{-}	4.1	0.83
6	3.4	5.0	0.00^{-}	3.6	0.00^{-}	6.7	0.00^{-}	3.2	0.18	3.6	0.16
7	3.8	4.7	0.00^{-}	5.4	0.00^{-}	6.3	0.00^{-}	3.5	0.17	3.4	0.07
8	2.9	4.3	0.00^{-}	4.6	0.00^{-}	5.9	0.00^{-}	3.4	0.02^{-}	3.1	0.28
9	1.7	3.2	0.00^{-}	4.1	0.00^{-}	5.0	0.00^{-}	2.4	0.00^{-}	3.1	0.00^{-}
10	1.4	2.3	0.00^{-}	3.6	0.00^{-}	3.5	0.00^{-}	1.8	0.00^{-}	2.7	0.00^{-}
11	1.6	2.1	0.00^{-}	3.1	0.00^{-}	2.8	0.00^{-}	1.8	0.02^{-}	2.9	0.00^{-}
12	1.7	2.0	0.00^{-}	2.0	0.00^{-}	2.3	0.00^{-}	1.8	0.03^{-}	2.1	0.00^{-}
13	1.5	1.5	1.00	1.5	1.00	1.5	1.00	1.5	1.00	1.5	1.00

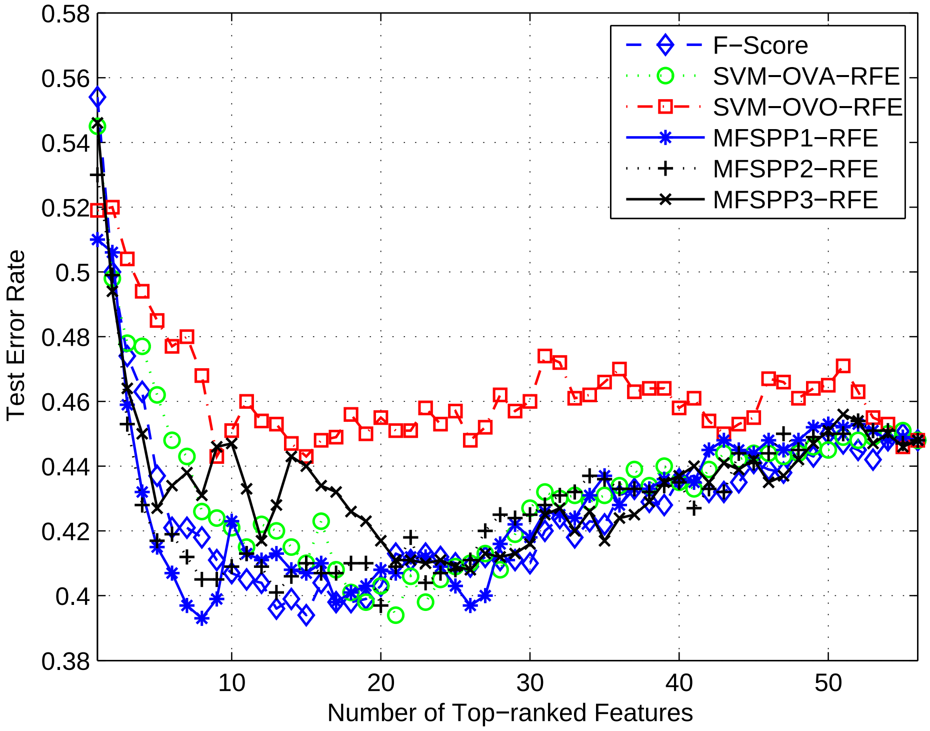


Figure 6.5: Average test-error rates against top-ranked features over 100 realizations of the lung-cancer dataset.

Table 6.4: Performance comparison between the best-performing method (i.e. MFSPP1-RFE) and the other methods (F-Score, SVM-OVA-RFE, SVM-OVO-RFE, MFSPP1-RFE, MFSPP2-RFE) on the lung-cancer dataset. The *P*-value is obtained in the paired *t*-test between each method to the best-performing method MFSPP1-RFE. The symbols "+" and "−" indicate statistically significant wins or losses over MFSPP1-RFE (*P*-value < 0.05).

N_{top}	MFSPP1 -RFE	F-Score		SVM-OVA -RFE		SVM-OVO -RFE		MFSPP2 -RFE		MFSPP3 -RFE	
	Mean TER (%)	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value
1	51.0	55.4	0.00^-	54.5	0.00^-	51.9	0.45	53.0	0.06	53.0	0.00^-
4	43.2	46.3	0.03^-	47.7	0.01^-	49.4	0.00^-	42.8	0.72	42.8	0.20
8	39.3	41.8	0.06	42.6	0.01^-	46.8	0.00^-	40.5	0.20	40.5	0.02^-
10	42.3	40.7	0.25	42.1	0.89	45.1	0.06	40.9	0.23	40.9	0.09
20	40.8	40.3	0.68	40.3	0.71	45.5	0.02^-	39.7	0.25	39.7	0.36
30	41.8	41.0	0.44	42.7	0.42	46.0	0.01^-	42.5	0.51	42.5	0.82
40	43.6	43.6	1.00	43.5	0.91	45.8	0.03^-	43.5	0.88	43.5	0.88
50	45.3	45.0	0.73	44.5	0.13	46.5	0.16	45.0	0.47	45.0	0.75
56	44.8	44.8	1.00	44.8	1.00	44.8	1.00	44.8	1.00	44.8	1.00

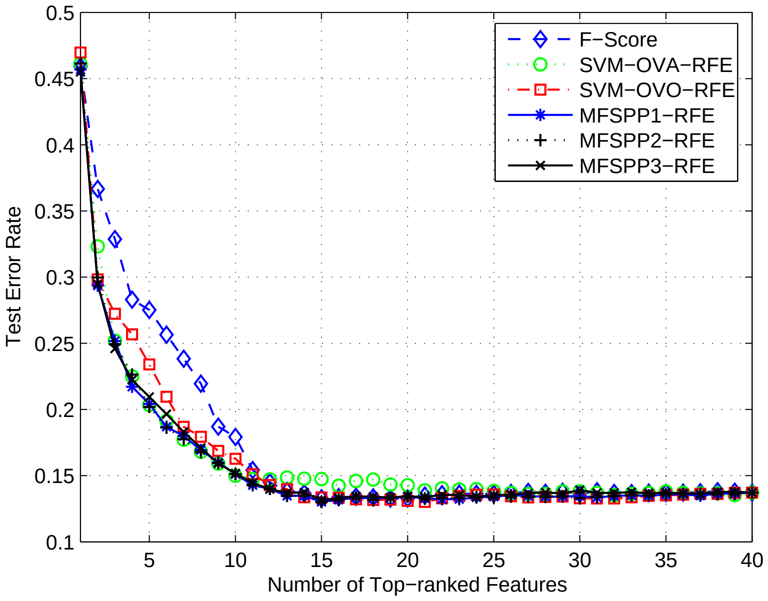


Figure 6.6: Average test-error rates against top-ranked features over 100 realizations of the waveform dataset.

Table 6.5: Performance comparison between the best-performing method (i.e. MFSPP1-RFE) and the other methods (F-Score, SVM-OVA-RFE, SVM-OVO-RFE, MFSPP1-RFE, MFSPP2-RFE) on the waveform dataset. The *P*-value is obtained in the paired *t*-test between each method to the best-performing method MFSPP1-RFE. The symbols "$+$" and "$-$" indicate statistically significant wins or losses over MFSPP1-RFE (*P*-value < 0.05).

N_{top}	MFSPP1 -RFE	F-Score		SVM-OVA -RFE		SVM-OVO -RFE		MFSPP2 -RFE		MFSPP3 -RFE	
	Mean TER (%)	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value	Mean TER (%)	*P*-value
1	45.7	45.9	0.27	46.1	0.16	47.0	0.01^{-}	46.1	0.07	45.5	0.34
5	20.4	27.5	0.00^{-}	20.3	0.17	23.4	0.00^{-}	20.2	0.15	20.9	0.04^{-}
10	15.1	17.9	0.00^{-}	15.0	0.36	16.3	0.01^{-}	15.1	0.87	15.2	0.26
15	13.0	13.3	0.14	14.8	0.01^{-}	13.4	0.04^{-}	13.0	0.65	13.2	0.27
20	13.4	13.3	0.19	14.3	0.00^{-}	13.1	0.12	13.4	0.81	13.4	0.74
25	13.4	13.6	0.24	13.8	0.04^{-}	13.6	0.31	13.3	0.42	13.5	0.37
30	13.5	13.6	0.76	13.7	0.71	13.3	0.80	13.3	0.85	13.8	0.27
35	13.6	13.7	0.90	13.7	0.75	13.6	0.76	13.6	0.97	13.7	0.80
40	13.7	13.7	1.00	13.7	1.00	13.7	1.00	13.7	1.00	13.7	1.00

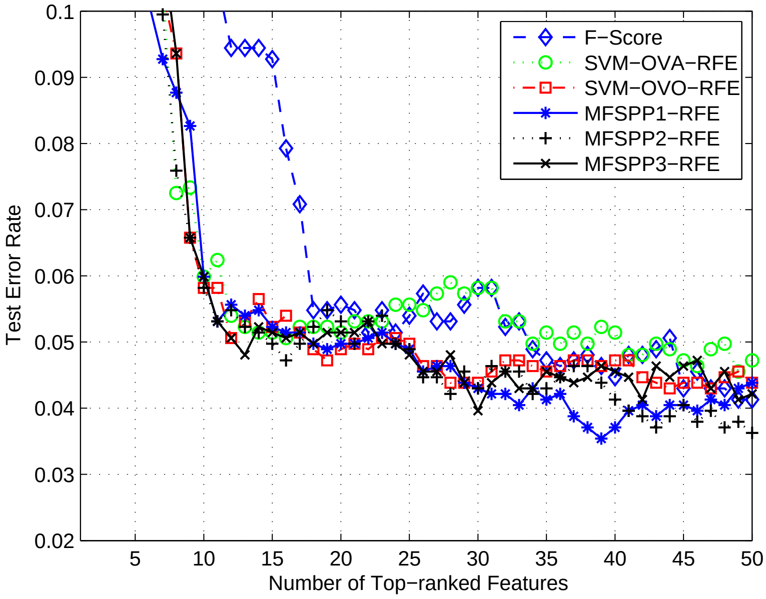


Figure 6.7: Test-error rates against top-ranked features on the DNA dataset.

is largely dependent on the choice of the decision function, $d(\mathbf{x})$. One possible reason for the better performance of $\frac{\partial p_i(\mathbf{x})}{\partial \mathbf{x}^k}$ is that $p_i(\mathbf{x})$ is ”closer” to the decision function, $d(\mathbf{x})$, of the classifier ($d(x) = \arg\max_i\ p_i(\mathbf{x})$) than $\|\mathbf{w}\|^2$ since $\frac{\partial d(\mathbf{x})}{\partial \mathbf{x}^k} = \arg\max_i\ \frac{\partial p_i(\mathbf{x})}{\partial \mathbf{x}^k}$. While the relation between $\|\mathbf{w}\|$ (or $J(\mathbf{x})$) and $d(\mathbf{x})$ is connected via the expressions of $p_i(\mathbf{x}) = (1 + \exp(A_i f(\mathbf{x}) + B_i)^{-1}$ and $f(\mathbf{x}) = \sum \mathbf{w}^T \Phi(\mathbf{x}) + b$. As the decision function directly affects testing accuracy, measures based on $\frac{\partial p_i(\mathbf{x})}{\partial \mathbf{x}^i}$ is a better choice for feature selection.

Another possible reason is that, $\|\mathbf{w}\|^2$ which is inversely proportionate to the SVM margin, could be more sensitive to the effect of a wrong choice of (C, γ) in the presence of sparse training data. Some evidence of this can be seen in our numerical experiments on the Weston’s nonlinear datasets. By comparing the results obtained using Set I and II in Table 6.2, Figures 6.2 and 6.3, it is evident that, for small training set sizes of 30 and 50, the test errors of SVM-OVA-RFE and SVM-OVO-RFE improve significantly from Set I to Set II, while the changes are much less significant for MFSPP1-RFE, MFSPP2-RFE and MSPP3-RFE.

6.6 Concluding Remarks

This chapter proposes the use of two feature-ranking criteria for feature selection in multi-class classification systems. It is based on posterior probabilities of multi-class SVM and is motivated by the advantage gained in using posterior probability as a decision function for classification instead of the direct SVM output function. The three approximations used for the two criteria are tested on various artificial and real-world benchmark problems in an overall feature-selection scheme using the popular recursive feature-elimination approach. The experimental results show that all the three approximations yield good overall feature-selection performance in the datasets considered. Among them, one of the approximation that uses the probabilistic outputs of the multi-class SVM proposed by Hastie and Tibshirani (Hastie and Tibshirani, 1998) has an overall edge and gives consistently better performance than the other feature-selection methods considered. In addition, it also performs best (among the other methods considered) on sparse datasets with low samples-to-features ratios. It is especially of interest to biomedical applications, such as the EEG application in this dissertation, which usually involve such sparse datasets.

Chapter 7

Continuous Measurement and Monitoring of Mental Fatigue: A Comprehensive Pattern Recognition System

This chapter presents an EEG-based mental-fatigue measurement and monitoring system using a probabilistic-based multi-class support vector machine (SVM) method. This pattern-recognition system uses the mental-fatigue EEG database established in Chapter 3, and it also includes the novel functions of automatic artifact removal algorithm and the automatic feature selection algorithm developed in Chapter 4 and Chapter 6 respectively. The experiments that follow provide evidence that this pattern-recognition system not only gives superior accuracy in predicting the subjects' mental-fatigue level but also provides a valuable estimate of confidence in the prediction that it makes in a given 3-second EEG epoch.

7.1 Introduction

In attempts to develop an objective and non-intrusive mental fatigue measurement method, some pilot studies have correlated mental fatigue with physiological measures such as electrocardiogram (ECG), electrooculogram (EOG) and EEG. A good review of these approaches can be found in the thesis by Mallis (1999) and a review by Lal and Craig (2001a). More recently, several studies have reported the feasibility of measuring mental fatigue or drowsiness indexed by subjects task performance, based on EEG data in attention-sustained experiments using auditory or visual stimuli (see Duta et al., 2004; Jones, 2006; Jung et al., 1997; Lal et al., 2003; Makeig et al., 2000; Peiris et al., 2004; Sommer et al., 2002; Vuckovic et al., 2002).

Most of these pilot studies have focused on the detection of performance lapses in the specific tasks that they studied (i.e. the prediction of a mistake in a specific task) without measuring subjects mental-fatigue levels directly. Also, most of these pilot studies used fairly simple linear or nonlinear regression or neural networks and recent advance in the signal processing methods, like automatic artifact removal, feature selection and multi-category pattern classification, have been overlooked. More importantly, the literature continues to produce varying and even conflicting results and very little evidence exists on the efficacy of incorporating EEG into a practically-usable automatic mental-fatigue measurement and monitoring system.

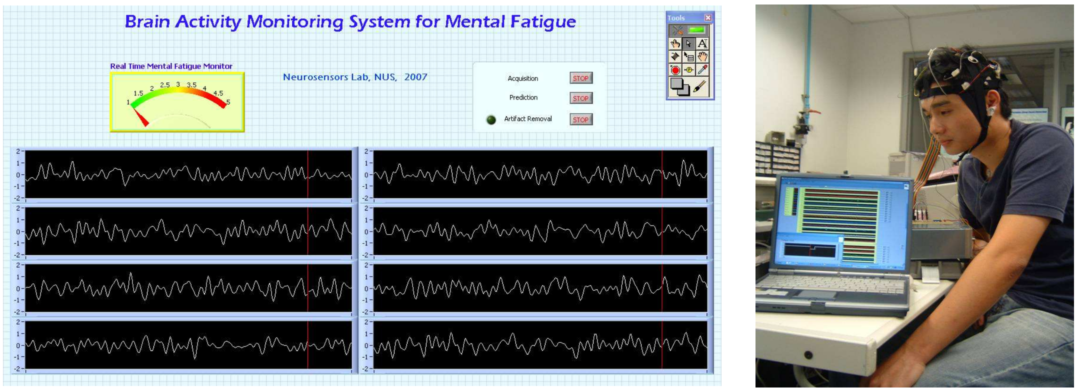
This chapter investigates whether a recently established technology similar to neural networks, probabilistic-based multi-class SVM, together with the novel the automatic artifact removal algorithm and the automatic feature selection algorithm developed in Chapter 4 and Chapter 6, can be used to automate the measurement and monitoring of subjects mental fatigue at different levels. Unlike standard multi-class SVM and other statistical learning methods which only give the bare classification, the probabilistic-based multi-class SVM provides not only superior classification accuracy but also useful estimates of confidence in the classification decision (Duan and Keerthi, 2005). This chapter also tests, through rigid performance evaluation, whether the probabilistic-based multi-class SVM together with the EEG signal processing methods developed in the previous Chapters can be used to establish a robust EEG-based mental fatigue measurement and monitoring system that is potentially of use in automated fatigue detection systems.

7.2 The Demonstration System

Fig. 7.1 shows the demonstration system of the EEG-based mental-fatigue measurement and monitoring system that was developed in this study. Monopolar EEG data were acquired at a sampling frequency of 167 Hz together with an electrode cap, according to the international 10-20 system (Jasper, 1958). The EEG data were pre-filtered by the EEG system through its integrated low-pass filter (cut-off frequency at 35 Hz) and high-pass filter (cut-off frequency at 0.1 Hz) as well as a 50 Hz notch filter. The EEG data were piped to a laptop through a data acquisition card (DAQCard- 6036E, National Instruments, USA) and then processed by a customized LabView software system running on a laptop for automatic measurement of subjects mental fatigue at different levels. The predicted mental fatigue levels were shown by a curve varying with the time (or by a virtual meter) on the laptop monitor, together with plots of real-time EEG data (after automatic artifact removal). The developed system has real-time capacity, but the present study focuses on the offline evaluation of its accuracy.

7.3 Data Preparation and Artifact Removal

As discussed in Chapter 3, 22 subjects were selected from right-handed volunteers of local tertiary institutions who fulfilled the inclusion criteria of not being on any medi-


Brain Activity Monitoring System for Mental Fatigue
Real Time Mental Fatigue Monitor
1
1.5
2
2.5
3
3.5
4
4.5
5
Neurosensors Lab, NUS, 2007
Acquisition
STOP
Prediction
STOP
Artifact Removal
STOP
Tools

(a) (b)

Figure 7.1: The developed demonstration system: (a) the display panel of the system, (b) the set-up of the demonstration system.

cation, no history of sleep disorders and with regular sleep hygiene as evidenced by a one-week sleep diary prior to the experiment. The recruitment of human subjects for this study was approved by the National University of Singapore (NUS) ethical committee. Informed consents were obtained and nominal monetary incentives sufficient to cover transportation costs were given for their participation. In order to train the system in a supervised regime, an auditory working-memory vigilance task (AWVT) was used as a validation measurement of mental fatigue. The detailed account of AWVT has been given previously in Chapter 3.

Each subject underwent a 25-h sleep deprivation experiment in a temperature-controlled laboratory (23–25 °C) from 8:30 am to 9:30 am next day. Caffeine, tea, smoking were prohibited for about two days (from one-day before the experiment till the end of experiment). Subjects were required to perform AWVT session once an hour throughout the experiment (with eyes open) and they were allowed to engage in non-strenuous activities in non-AWVT-session period. EEG data were recorded simultaneously during every AWVT session and they were labeled to 5-level mental fatigue according to the AWVT performance score. Specifically, for each subject, his/her individual performance span (the highest AWVT score to the lowest AWVT score) was evenly divided into five segments corresponding to fatigue level 1 to level 5, respectively. The label (i.e. mental fatigue at 5 levels) of the EEG data for an AWVT session was determined by which segment the corresponding AWVT performance score fell into.

As the result, a relative large database of mental fatigue EEG (with reliable labels of mental fatigue levels), collected from 22 subjects (each underwent a 25-hour sleep deprivation), was available separately for each subject. As it will be seen in the subsequent sections, about half of the data (from 12 subjects) were used for identifying the key EEG features that are relevant to mental fatigue. The rest of the data (from 10 subjects) were used to evaluate the performance of the proposed system in a stringent subject-wise cross-validation procedure (see Section 7.6.4).

The collected EEG data were enhanced by a customized FIR bandpass filter with a pass band of 0.1–25 Hz implemented in Matlab (version 6.5, MathWorks, USA). The electroencephalogram artifacts and electrocardiogram artifacts were automatically removed by using the artifact removal method described in Chapter 4.

7.4 Feature Extraction

The purpose of feature extraction was to extract a set of features that optimally distinguish mental fatigue at 5 levels. They were used as the chief information source, in replace of the EEG data, for classification.

Specifically, the multi-channel EEG data were segmented into 3-second-long EEG epochs by passing through feature extraction windows (length of 3 seconds or 500 samples). There was two-second lag (or 334 samples) between two adjacent segmentations. Fast Fourier transform (FFT) with Hann window (length of 256 samples and 50% of overlap between adjacent segments) (Oppenheim and Schafer, 1989) was performed on each of these 3-second EEG epochs. The resulting power spectrum density function (normalized by its total power) for each channel was divided into four segments according to the four standard EEG frequency bands (Niedermeyer and Silva, 1999): delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), and beta (13-20 Hz). For each channel, four frequency features were defined for each standard EEG frequency band based on the EEG power spectrum $P(f_i)$ in that considered frequency band, capturing both spatial and temporal information that were useful for distinguishing mental fatigue at 5 levels. These features are defined as follows.

(a) *Dominant Frequency (DF):* For every peak in a considered frequency band, two frequencies in relation to a peak were defined - one was on the rising slope and the other was on the falling slope having the power equal to (or closest to) half the power of the peak. These two frequencies defined a frequency zone called full width half maximum band of the peak (Weisstein, 2007). Among all peaks in the considered frequency band, the peak with the largest average power in its full width half maximum band was called the dominant peak, while its corresponding frequency was called dominant frequency. In loose terms, this feature was to capture the dominant peak with the most significant bandwidth within a considered frequency band.

(b) *Average Power of Dominant Peak (APDP):* This was defined as the average power on the full width half maximum band of a dominant peak. It represented the significance/importance of that dominant peak.

(c) *Center of Gravity Frequency (CGF)*: It was defined as

$$CGF = [\sum_i P(f_i) \cdot f_i]/[\sum_i P(f_i)], \tag{7.1}$$

where f_i is discretized frequency and $P(f_i)$ is the estimated power spectral density. This feature is significantly different from the first feature (dominant frequency), which can be illustrated by an example: if the spectrum for a considered frequency band is dominated by two narrow peaks (one larger than the other), it is not difficult to see from Equation (7.1) that the center of gravity frequency will fall in between these two peaks, whereas the dominant frequency will be the frequency of the largest peak.

(d) *Frequency Variability (FV)*: It is defined as

$$FV = \frac{\sum_i P(f_i) \cdot f_i^2 - [\sum_i P(f_i) \cdot f_i]^2 / \sum_i P(f_i)}{\sum_i P(f_i)}. \tag{7.2}$$

Considering $P(f_i)$ as the probability distribution of frequency, this feature is in fact the variance of the frequency in the defined frequency band.

As a result of feature extraction, the mental-fatigue EEG data recorded from each subject throughout the 25-hour sleep-deprivation experiment, were first segmented into 2,100 epochs. Each 3-second EEG epoch was then converted into a 304×1 vector of quantitative EEG features (4 kinds of features × 19 channels × 4 frequency bands). It is acknowledged that small portion of the EEG data (less than 20-second EEG data) in the beginning of each recording period was discarded for two reasons: 1) to minimize the bias of warming-up effect on the subject at the beginning of each session of auditory working-memory vigilance task; 2)to ensure the well-balanced samples for all the five levels of mental fatigue.

7.5 Feature Selection

Feature selection concerned with the identification of a minimum set of key EEG features necessary for accurate classification of mental fatigue at 5 levels (out of abovementioned 304 features). It is important for at least two reasons: 1) from ergonomics point of view, fewer features and thus fewer EEG channels are desirable for user's comfort; 2) from machine-learning point of view, when the underlying important features are known and redundant features are removed, the classification problem can be greatly simplified, resulting in improved classification accuracy.

The MFSPP-RFE, the feature-selection method for multi-class classification problems as presented in Chapter 6, was used to select the key features for multi-level mental-fatigue EEG classification. In view of time constraints, only the results of MFSPP1-RFE was obtained. The choice of MFSPP1-RFE was due to its encouraging performance in the rigid numerical experiments on various artificial and real-world benchmark problems as reported in Chapter 6.

To ensure that the performance evaluation which will be reported in later sections in this chapter is not biased, only part of the database (12 out of 22 subjects) was used for identifying the key features. A subject-wise cross-validation procedure was used to form the data subset $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$ as required by the MFSPP1-RFE. Specifically, 2100×11 samples from 11 subjects were used to form a $\mathbb{D}_{tra}$, and the samples from the left-out subject were used to form a testing set $\mathbb{D}_{tes}$. Practically, this subject-wise cross-validation procedure results in 12 pairs of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$. For each pair of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$, $\mathbb{D}_{tra}$ was used by feature-selection approach, MFSPP1-RFE (as described in Algorithm 4), producing a ranked feature list J_R showing the all the features in decreasing order of importance. To estimate the predictive performance of the selected features, the PWC-PSVM, i.e. the probabilistic multi-class SVM, was iteratively fit on $\mathbb{D}_{tra}$, at each iteration retraining a new PWC-PSVM after discarding the least important features according to J_R. To estimate the predictive performance of the selected features, the test errors were obtained by testing on $\mathbb{D}_{tes}$ on the iteratively trained PWC-PSVM and calculating the percentage of misclassifications on $\mathbb{D}_{tes}$. In order to save the computational time, a three-tier feature-removal scheme was used, in which 20 least-important features were removed at each recursion until 44 features left, and then 5 features were removed at each recursion until 24 features were left, followed by two features removed at each

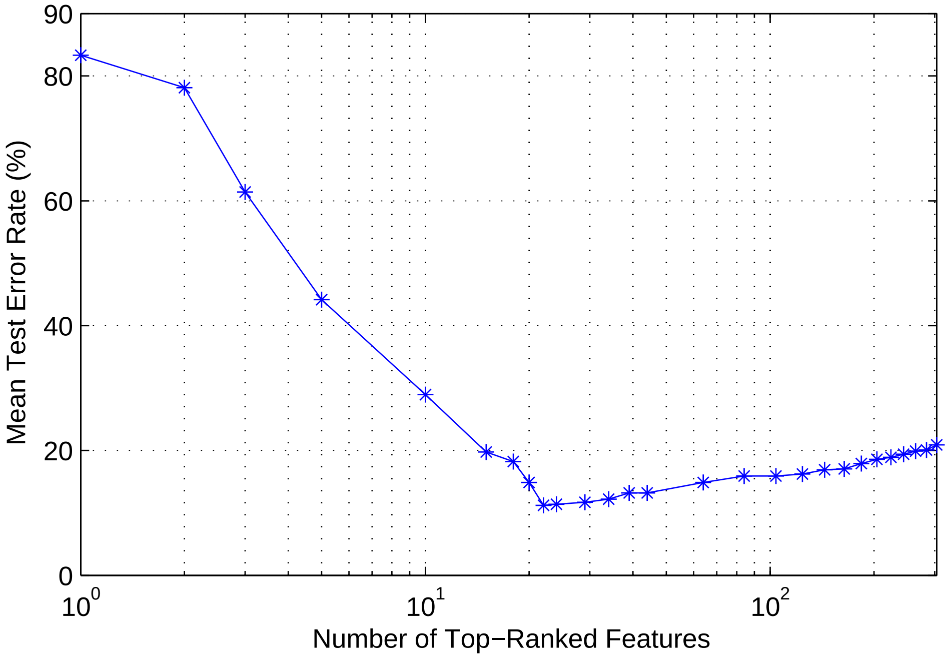


Figure 7.2: Mean test error rate against the number of top-ranked features where the top-ranked features were selected by MFSPP1-RFE. The test error rates were obtained by averaging 12 test error rates on all resampled subsets $\mathbb{D}_{tes}$'s.

recursion.

Fig. 7.2 shows the mean test error rates of PWC-PSVM on unseen testing sets $\mathbb{D}_{tes}$'s with the decreasing number of top-ranked features, where the top-ranked features were selected the MFSPP1-RFE approach. The mean test error rates shown were the average values over 12 test error rates corresponding to 12 pairs of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$. Error bars of standard deviation have not been plotted for the sake of clarity. The results show that the standard deviation is rather stable with regards to the number of the top-ranked features used for classification (between 4% and 7%).

Like most EEG-based automatic diagnostic systems, it is imperative to use only key features. The full feature set constitutes a high dimensional vector (304 for the present study) that contains key features pertinent to the classification of mental fatigue, irrelevant features for other cognitive states or artifacts, as well as redundant features which can be replaced by key features. Direct classification using the full feature set is apparently undesirable, since irrelevant and redundant features have adverse effect on the overall classification performance and generalization ability of the system. On the other hand, if some of key features were further removed after removal of irrelevant and redundant features, the classification accuracy would drop dramatically. The results as shown in Fig. 7.2 matches exactly the said scenario. Using a multi-class classifier, the lowest mean test error rate (approximately 12%) was obtained using only about 22 features, compared with a mean test error rate of about 21% using the full feature set (304 features) and a mean test error rate of about 83% using only the most important feature. The classification performance in differentiating mental fatigue at 5 levels could be greatly improved by using only the key features pertinent to the classification via feature selection.

The determination of the number of top-ranked features to be retained can be tricky. Based on the results as shown in Fig. 7.2, it may be a good idea to examine the first 22 top features as ranked by the MFSPP1-RFE using each of 12 pairs of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$. It is reasonable to believe that, if a feature is repeatedly selected as a key feature in the experiments on 12 pairs of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$ following the above-mentioned cross-validation procedure, the feature is indeed important. There were 18 features which were repeatedly ranked within the first 22 top features in the experiments on at least 6 pairs of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$ (out of 12 pairs). These features are shown in Fig. 7.3 and Table 7.1. They were used in the proposed EEG-based mental-fatigue measurement and monitoring system.

It is worth noting that, although the key features were selected by the automatic feature selection method using a data-driven approach, the selection generally makes sense in terms of neurophysiology. For example, features in delta, theta and alpha bands from frontal region of the brain (i.e. F3, F4, Fz, F7 and F8 as shown in Fig. 7.3) were identified important for the classification of mental fatigue at different levels. It corresponds with various studies in the literature (see Broughton et al., 1994; Cajochen et al., 1999;

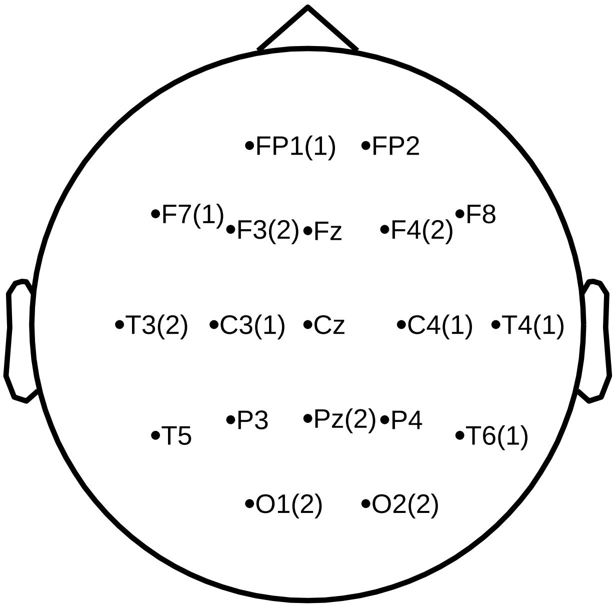


Figure 7.3: Distribution of the 18 key features which were repeatedly ranked within the first 22 top features in the experiments on 12 pairs of $\mathbb{D}_{tra}$ and $\mathbb{D}_{tes}$ by MFSPP1-RFE. The number in bracket following the channel name is the number of key features deriving from that channel.

Jung et al., 1997) which have shown nicely relative increase in delta, theta and alpha activity after sleep deprivation. Also, occipital sites (i.e. O1 and O2) were selected as shown in Fig. 7.3, which is consistent with many other studies on neurophysiology of mental fatigue (see Alloway et al., 1997; Cajochen et al., 1995; Cantero and Atienza, 2000; Stampi et al., 1995). Besides these features with clear neurophysiological interpretation, the key features selected by the automatic feature selection method also include some new features which have not been captured from a purely neurophysiologic angle, such as key features from T3 and T6. The automatic feature selection procedure identifies features that provide improvement to the classification accuracy, and hence, can usually discover a larger set of features. Besides new key features, features that work only in the presence of other features may also be discovered. Interested readers may refer to the review paper (Guyon and Elisseef, 2003) for related discussions. We see the new features identified by the procedures like ours which serve as candidates for

Table 7.1: List of the selected 18 key features

Feature Rank	Channel Name	Frequency Band	Feature Name
1	T4	Theta	APDP
2	T6	Theta	APDP
3	F3	Alpha	APDP
4	C3	Theta	APDP
5	C4	Theta	APDP
6	Pz	Theta	APDP
7	T3	alpha	APDP
8	T3	Beta	FV
9	F3	Theta	APDP
10	O2	Theta	APDP
11	F4	Beta	APDP
12	Fp1	Delta	APDP
13	O1	Delta	FV
14	O2	Beta	APDP
15	F7	Delta	CGF
16	Pz	Beta	APDP
17	F4	Alpha	APDP
18	O1	Beta	APDP

subsequent neurophysiologic investigation.

As a result of afore-mentioned feature extraction and feature selection, the collected mental fatigue EEG data were transformed into subject-wise datasets $\mathbb{D}_k$, $k = 1, \cdots, 10$ (for 10 subjects), in the form of $\{\mathbf{x}(n), y(n)\}_{n=1}^{2100} \in \mathbb{R}^{18} \times \{1, 2, 3, 4, 5\}$, where $\mathbf{x}(n)$ is the 18-dimensional feature vector derived from the n^{th} EEG epoch of the k^{th} subject and $y(n)$ is the corresponding mental-fatigue level determined by the manual classification.

7.6 Automatic Measurement of Mental Fatigue Using Probabilistic-Based SVM

A probabilistic-based multi-class SVM was used in the proposed system for automatic measurement and monitoring of mental fatigue. This is part of our attempt to achieve higher accuracy in predicting the subjects' mental fatigue at 5 levels and to obtain a useful confidence estimate telling how confident/reliable each prediction is.

7.6.1 Two-class SVM

SVM is a supervised learning method used for classification and regression. It was originally designed for two-class classification. Unlike other statistical learning methods (such as neural networks and decision trees) which usually aim only to minimize the empirical classification error, SVM simultaneously minimizes the empirical classification error and maximizes the geometric margin in classification; hence it is also known as maximum margin classifier (Boser et al., 1992; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Vapnik, 1995, 1998).

The training of SVM is essentially seeking an optimal separating hyperplane that separates samples from two classes with maximum margin, but the trick is to find the hyperplane in a high (possibly infinite) dimensional space obtained by transforming the original feature space using an appropriate nonlinear mapping function, rather than in the original feature space (Duda et al., 2000). Fig. 7.4 shows an illustrative example of a hyperplane that SVM constructs. The support vectors of SVM are the training samples that define the optimal separating hyperplane and are the most difficult patterns to classify. The SVM identifies these support vectors and simultaneously constructs the optimal separating hyperplane which optimally separates them with maximum margin; that is the machine training for SVM. After the machine training, for a given unseen feature vector (representing a test pattern), the trained SVM outputs its predicted class label (-1 or +1) based on the half space (defined by the hyperplane) into which that feature vector falls.

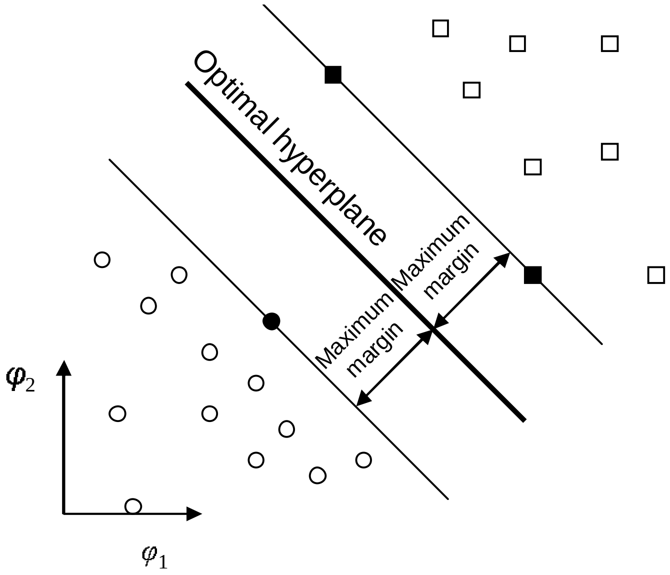


Figure 7.4: Training of SVM is to find the optimal hyperplane (thick line) which separates the samples from two classes (circles vs. squares) with maximum margin. The support vectors are shown as solid circles or squares. The figure shows the projection view of the hyperplane in two dimensions (φ_1 and φ_2) in transformed space.

7.6.2 Standard Multi-Class SVM

SVM was originally designed for two-class classification. A multi-class SVM (for multi-class classification problems) is usually implemented by combining several two-class SVMs. The most popular standard multi-class SVM is the 'one-versus-one' SVM (OVO-SVM). The final classification is based on voting by all the pair-wise two-class SVMs. Specifically, for a given test feature vector, count the times that each class wins in all these pair-wise classifications and choose the class that win most as the class for that test feature vector. Besides OVO-SVM, other forms of standard multi-class SVM also exist, such as 'one-versus-all' SVM (OVA-SVM) and various error-correction schemes. They follow similar principle as OVO-SVM and perform similarly (Hsu and Lin, 2002; Rifkin and Klautau, 2004).

7.6.3 Probabilistic-Based Multi-Class SVM

Two-class SVM classifies a sample depending on the half space (defined by separating hyperplane) into which that sample falls and standard multi-class SVM is simply a brute-force combination of two-class SVMs. Such an approach, however, ignores the relative confidence in the classification, or the distance that the sample is from the separating hyperplane. Standard multi-class SVM, such as OVO-SVM, only gives a bare classification, i.e. prediction of the class that the test sample belongs to. No extra information is provided to show how confident the classification is.

Unlike standard multi-class SVM, probabilistic-based multi-class SVM gives not only a bare classification but also a useful confidence estimate showing how confident the classification is. A discussion of the basic principle and mathematics of probabilistic-based multi-class SVM has been shown in Chapter 2. For simplicity, this probabilistic-based multi-class SVM is hereafter denoted by PWC-PSVM.

In this study, PWC-PSVM was used for the purpose of automatic measurement of mental fatigue at 5 levels. For a new feature vector $\mathbf{x}(n)$ derived from the n^{th} EEG epoch, the decision rule of PWC-PSVM is to assign $\mathbf{x}(n)$ to the most probable mental fatigue level as follows:

$$d(\mathbf{x}(n)) = \arg\max_{i}\{P(\omega_i|\mathbf{x}(n)),\ i = 0\ ,\cdots,\ 5\}, \tag{7.3}$$

where ω_i denotes the i^{th} mental fatigue level and $P(\omega_i|\mathbf{x}(n))$ is the confidence estimate of assigning the n^{th} EEG epoch to the i^{th} mental fatigue level. The PWC-PSVM method has the advantage that it provides the output of multi-class SVM a new interpretation in the form of posterior probability or confidence estimate of assigning the new sample to that class. If the posterior probability for one class is significantly higher than the other classes, then the strength of the prediction is sufficiently high. On the contrary, if there are multiple or no classes which claim the test sample with relatively high posterior probability, then the strength of the prediction is low. It is more accurate to make predictions with high strength than those with low strength. In the present work, a bar plot of these confidence estimates associated with a test sample could also be shown together with decision by Equation (7.3). The bar plot tells the user how sure (in a qualitative sense) it is of that decision.

The usefulness of confidence estimates was also studied in an attempt to achieve high accuracy for the proposed mental fatigue measurement and monitoring system. Instead of using Equation (7.3) to predict mental fatigue level on single EEG epoch, the prediction can be made through aggregation of confidence estimates on multiple EEG epochs, say τ epochs. After a number of EEG epochs, the current mental fatigue level of $\mathbf{x}(n)$ can be determined by the most probable class as follows:

$$\hat{d}(\mathbf{x}(n)) = \arg\max_i \{P(\omega_i|\mathbf{x}(n-k)),\, k = 0\,,\ldots,\, \tau - 1 \text{ and } i = 0\,,\cdots,\, 5\}, \tag{7.4}$$

where $P(\omega_i|\mathbf{x}(n-k))$ again denotes the confidence estimate of assigning the $(n-k)^{th}$ epoch to the i^{th} mental fatigue level as given by PWC-PSVM. The classification rule makes the prediction of the mental fatigue level by the associated class label with the largest confidence estimate within τ consecutive EEG epochs. In the present study, the effectiveness of Equation (7.4) was compared with those of other two commonly used methods for aggregating multiple predictions:

$$\bar{d}(\mathbf{x}(n)) = \arg\max_i \{\sum_{k=0}^{N-1} P(\omega_i|\mathbf{x}(n-k))\}; \tag{7.5}$$

$$\tilde{d}(\mathbf{x}(n)) = \arg\max_i \{\sum_{k=0}^{N-1} \log P(\omega_i|\mathbf{x}(n-k))\}. \tag{7.6}$$

These two methods as in Equations (7.5) and (7.6) are similar except that the latter is expected to have better scaling than the former for confidence estimates given by PWC-PSVM ($0 < P(\omega_i|\mathbf{x}(n-k)) < 1$).

7.6.4 Subject-Wise Cross-Validation for Performance Evaluation

To evaluate the generalization performance of the proposed EEG-based mental-fatigue measurement system, a blocking re-sampling scheme "leave-one-proband-out" (Lahiri, 2003) was used. The data from 10 subjects (that were different from the 12 subjects used in feature selection) were divided subject-wise so that samples used for training and for testing were not from same subjects. Specifically, samples of one subject were used for testing, and samples of the rest subjects were used to form a serial of nested training datasets, starting from a smallest training dataset comprising samples of one subject to a biggest training dataset comprising samples of nine subjects by progressively incorporating more and more subjects' samples. As a result, for each hold-out subject for testing, a serial of 9 nested training datasets were formed. It is fair to note that this "leave-one-proband-out" re-sampling scheme forms an extremely stringent evaluation of subject-independent performance of the proposed system.

7.7 Results

7.7.1 Mental-fatigue classification accuracy

Fig. 7.5 showed 10 curves of the testing accuracy for the PWC-PSVM method, each curve showing the testing accuracies on a hold-out subject varying with the number of subjects used in training. The testing accuracy was calculated in terms of the percentage of correct classifications. For comparison, the average testing accuracies for both of

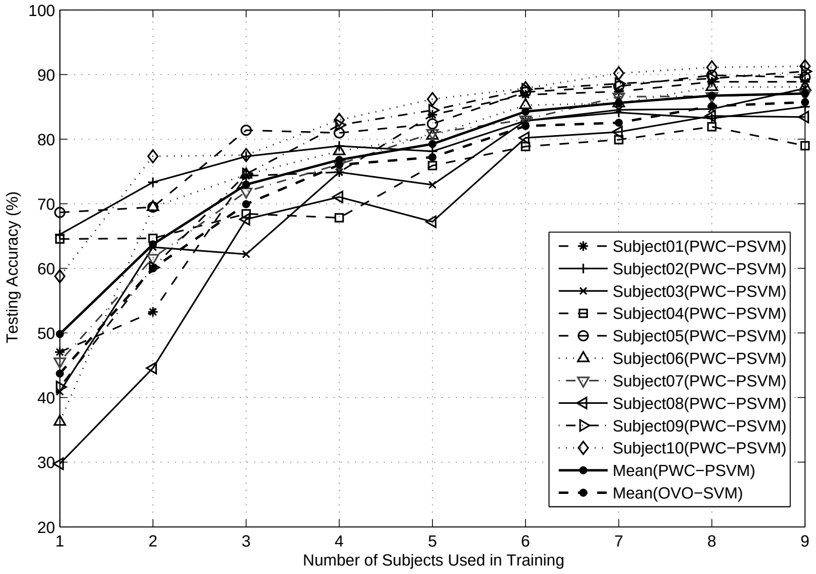


Figure 7.5: The testing accuracy varying with number of subjects for training in single-trial classification using the PWC-PSVM method. The testing accuracy was evaluated on a hold-out subject. Each curves in the figure corresponded to a hold-out subject, with the thick solid line showing the mean. For comparison, the mean of testing accuracies using OVO-SVM method was also shown by the thick dashed line.

PWC-PSVM (thick solid line) and OVO-SVM method (thick dashed line) were shown in Fig. 7.5. The PWC-PSVM method consistently obtained higher mean accuracy than OVO-SVM. In fact, paired *t*-test showed that PWC-PSVM significantly outperformed OVO-SVM (p-value < 0.05). The paired *t*-test's were done, for each number of subjects used in training, on the 10 paired accuracies resulting from 10-fold subject-wise cross-validation. This suggests the goodness of the probabilistic multi-class SVM over standard non-probabilistic multi-class SVM, which is consistent with the empirical study of Duan and Keerthi (2005).

Fig. 7.5 represents the generalization performance of the proposed mental fatigue monitoring system on new subjects. As shown in Fig. 7.5, the highest average testing accuracy of 87.5% for 5-level mental fatigue classification was obtained with PWC-PSVM trained on 9 subjects, compared with the highest average testing accuracy of 85.1% for

Table 7.2: Mean confusion matrix resulting from subject-wise 10-fold cross-validation

Actual \ Predicted	Level 1	Level 2	Level 3	Level 4	Level 5
Level 1	389 (92.6%)	18 (4.3%)	10 (2.4%)	1 (0.2%)	2 (0.5%)
Level 2	20 (4.8%)	373 (88.9%)	12 (2.9%)	8 (1.9%)	7 (1.7%)
Level 3	16 (3.8%)	30 (7.1%)	342 (81.4%)	15 (3.6%)	17 (4.0%)
Level 4	5 (1.2%)	20 (4.8%)	25 (6.0%)	355 (84.5%)	15 (3.6%)
Level 5	8 (1.9%)	10 (2.4%)	7 (1.7%)	19 (4.5%)	376 (89.5%)

OVO-SVM. They are in fact 10-fold cross-validation accuracies with all the available data being split into 10 folds subject-wise.

Fig. 7.5 is also useful for deducing the minimum number of subjects required for training. The proposed mental-fatigue monitoring system was aimed to be applicable to different subjects and hence inter-subject dependence of mental-fatigue pattern is of concern. As shown in Fig. 7.5, the mean testing accuracy using PWC-PSVM (thick solid line) monotonically increased with the number of subjects for training, suggesting a mild inter-subject dependence of mental fatigue pattern. The testing accuracy was about 84.3% (±3.6%) with the classifier PWC-PSVM trained on 6 subjects. It then slowly increased to about 87.5% (±3.2%) with the classifier trained on 9 subjects where it almost saturated.

It is worth noting that equal costs for any misclassification were assumed in the present study. Table 7.2 showed the average confusion matrix, which was the mean of 10 confusion matrices resulting from the afore-mentioned subject-wise 10-fold cross-validation procedure. It showed that gross errors (such as mental fatigue level 1 being misclassified to mental fatigue level 5) did not often occur.

7.7.2 Relating classification confidence estimate to classification accuracy

Fig. 7.5 and Table 7.2 showed the hard accuracy for the PWC-PSVM method: the accuracy in terms of the percentage of correct classifications using the single-trial decision rule of Equation (7.3). However, the strength of the classification, which was readily provided by the PWC-PSVM method, has not been taken into consideration. Since the PWC-PSVM method gave the output of multi-class SVM a more subtle interpretation as a classification confidence estimate, it provides us a way to evaluate the classification results by comprehensively examining these confidence values. Table 7.3 listed the rank of confidence estimate for the correct class, counting on all the hold-out validating samples (resulting from subject-wise 10-fold cross-validation) when the PWC-PSVM method achieved the highest mean accuracy of 87.5%. Following the procedure of getting the mean confusion matrix in Table 7.2, the counts in Table 7.3 were divided by 10 (the number of folds) and rounded to the nearest integers. The corresponding percentage was shown in parentheses. Given a test sample, if the confidence estimate for the correct class is ranked first by the classifier, the single-trial classification using Equation (7.3) is correct. Otherwise, an error occurs. From Table 7.3, it is evident that most errors were due to the reason that the confidence estimate for the correct class was ranked second or third by the classifier PWC-PSVM. Nearly half of the errors occurred because the confidence estimate for the correct class was ranked second, indicating that errors tend to occur in the overlapping regions of multiple classes where there are multiple classes claiming the test sample with high confidence estimate.

Table 7.3 explored the possibility of further improving the multi-class classification accuracy by aggregating confidence estimates on multiple epochs. Subject-wise 10-fold cross-validation accuracies were obtained, on different number of epochs used (τ = 1 to 5) for each aggregation method as in Equations (7.4), (7.5) and (7.6). As shown in

Table 7.3: Categorization of the single-trial decision results based on the ranking of confidence estimate (percentages are shown in parentheses following the corresponding counts)

Correct Class Label	**Rank of Confidence Estimate for Correct Class**				
	First	Second	Third	Fourth	Fifth
Level 1	389 (92.6%)	16 (3.8%)	6 (1.4%)	6 (1.4%)	3 (0.7%)
Level 2	373 (88.9%)	25 (6.0%)	8 (1.9%)	6 (1.4%)	8 (1.9%)
Level 3	342 (81.4%)	33 (7.9%)	20 (4.8%)	13 (3.1%)	12 (2.9%)
Level 4	355 (84.5%)	30 (7.1%)	14 (3.3%)	10 (2.4%)	11 (2.6%)
Level 5	376 (89.5%)	25 (6.0%)	10 (2.4%)	6 (1.4%)	3 (0.7%)

Table 7.4: Comparison of different aggregation methods on different numbers of epochs used for aggregation

Aggregation Rule	Number of Epochs for Aggregation				
	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
$\hat{\underline{d}}(\mathbf{x}(n))$	**87.50%**	**89.30%**	**90.10%**	**90.60%**	**90.60%**
$d(\mathbf{x}(n))$	**87.50%**	**89.30%**	**90.10%**	90.10%	90.11%
$\tilde{d}(\mathbf{x}(n))$	**87.50%**	88.90%	89.90%	89.91%	89.91%

Table 7.3, when τ is set to 1, they were equivalent to single-trial decision rule as in Equation (7.3). As the number of epochs for aggregation increased from 1 to 5, the cross-validation accuracy increased for all methods, with the edge going to the aggregation method as in Equation (7.4). It gave the best cross-validation accuracy of 90.6%, aggregating on 4 or 5 epochs.

As expected, all the three aggregation methods as in Equations (7.4), (7.5) and (7.6) improved the testing accuracy. However, the slight edge of method as in Equation (7.4) over those as in Equations (7.5) and (7.6) is most interesting. The aggregation method as in Equation (7.4) made the prediction of the mental fatigue level for the current test sample $\mathbf{x}(n)$ by the associated class label with the largest confidence estimate within τ consecutive EEG epochs $\mathbf{x}(n-\tau+1), \cdots, \mathbf{x}(n)$. It is different from Equations (7.5) and (7.6) which gave aggregate prediction by simply averaging the confidence estimates over multiple epochs. One of plausible reasons for the slightly better performance of Equation (7.4) is that it increased the accuracy most by effectively increasing the chance that the final prediction of mental fatigue level was made on the single trial prediction with the highest strength.

It is also interesting to see from Table 7.3, that the accuracy saturates at about $\tau = 5$. A plausible explanation could be that reliable EEG pattern change due to change in mental fatigue or brain functional state spans 15 seconds to 1 minute (Lal et al., 2003; Torsvall and Åkerstedt, 1987).

7.8 Discussion

An EEG based mental fatigue measurement and monitoring system using a multi-class SVM with confidence estimate has been presented. Three aggregate prediction methods have also been proposed and compared in an attempt on further improving classification accuracy. The results show that a multi-class SVM with confidence estimate outperformed standard multi-class SVM method. Moreover, the classification accuracy was further increased to about 90% using one of the proposed aggregate prediction methods.

The developed system may serve as a key step towards an EEG based mental-fatigue monitoring device. In the literature, several studies have reported the feasibility of detecting operator drowsiness based on EEG data in attention-sustained experiments (see Duta et al., 2004; Jones, 2006; Jung et al., 1997; Lal et al., 2003; Makeig et al., 2000; Peiris et al., 2004; Sommer et al., 2002; Vuckovic et al., 2002). Most of these pilot studies used a fairly simple linear or nonlinear regression or neural networks, as opposed to the more sophisticated multi-class SVM with confidence estimate use in the present study. Another shortcoming of these pilot studies was the lack of subject-wise cross-validation in their performance evaluation (see Lal et al., 2003). The present study used a relative large data (22 subjects, each for a 25-hour duration), applied a stringent “leave-one-proband-out” scheme in the evaluation of subject-independent performance, and showed a high accuracy (about 90%) in classifying mental fatigue EEG.

A remarkable property of the probabilistic-based multi-class SVM used in the present study is that it provides not only superior classification accuracy but also useful estimates of confidence in the classification that it makes. Its benefits have recently been studied in the domain of machine learning (Duan and Keerthi, 2005), while its application in biomedical engineering remains rare. The present study has provided additional evidence by demonstrating the use of resulting posterior probabilities for in-depth evaluation of classification results (via comprehensive examination of these posterior probabilities) and for further performance boosting (via aggregation of these posterior probabilities). The highest accuracy of 90.6% achieved in this study is also one of the significant contributions of the present study.

The present study has focused on circadian mental fatigue caused by sleep deprivation. Inefficient functioning due to sleep deprivation and working at the time of circadian dips has been a major cause of accidents in shipping, aviation, industrial and military scenarios. There has also been a general agreement that sleep deprivation causes a degradation of many of human abilities like vigilance, sustained attention, working memory, judgment and executive decision making. Consequently, the present study used the Auditory Vigilance Task in indexing the mental fatigue level caused by sleep deprivation. This task, in comparison with the popular PVT (Dinges and Powell, 1985; Thorne et al., 2005) which only measures vigilance, measures not only vigilance but also working memory, decision making and sustained attention - the higher faculties of the human brain which are used for complex tasks in real life. Nevertheless, this task tests more functions while still being simple and having minimal variability (due to aptitude or education of a person). A comprehensive comparison between AWVT and other mental fatigue measurement methods could be worthy of future investigations.

As in many other mental fatigue or vigilance studies using EEG (e.g. Lal et al., 2003), mental fatigue was classified into different levels (5 levels in the present work). It is arguable whether mental fatigue should be measured continuously or discretely, but it is reasonable to believe that progression of mental fatigue may not be entirely smooth or continuous. On the contrary, mental fatigue could be very much like sleep staging where only quasi-categorical sleep stages can be defined. In the study done by Lal et al. (2003), mental fatigue was similarly classified into 4 phases: early, medium, extreme fatigue phases, and an arousal phase. Evidence of such quasi-categorical mental fatigue states has also been shown in a recent EEG study by Trejo et al. (2007).

In the present study, a relative measure of mental fatigue (with the full range of individual performance divided into five) instead of a universal measure for all subjects is used because the task performance of AWVT (same as other performance tasks like PVT) is subject-dependent. It is reasonable to believe that all mental fatigue levels in a full cycle of circadian fatigue were sampled by the 25-hour sleep deprivation experiment used in the present study. Therefore, the maximum (or minimum) task performance corresponds to the lowest (or highest) mental fatigue level, i.e. mental fatigue level 1 (or level 5). However, the minimum and maximum performance might be rather noisy, thus possibly distorting the manual classification of intermediate mental fatigue levels (level 2-4). The classification results by the proposed probabilistic-based multi-class SVM (as shown in Table 7.2) also imply that mental fatigue levels 2-3 as defined by this relative measure were less distinguishable as SVM gave lower classification accuracies on these two classes than to other classes.

7.9 Concluding Remarks

A pattern-recognition system for automatic classification of subjects' mental fatigue at 5 levels has been presented in this chapter. This chapter has also described a workable demonstration system of an EEG-based mental-fatigue measurement and monitoring device, through the use of all the EEG signal processing methods developed in previous chapters. The performance evaluation of the system via a stringent “leave-one-proband-out” demonstrates the feasibility of an automatic EEG method for assessing and monitoring of mental fatigue at a time scale of 3s EEG epoch.

Chapter 8

Conclusions and Recommendations

8.1 Conclusions

The present work represents a new data-driven approach to automatic mental-fatigue measurement and monitoring. The developed signal processing software has resulted in a demonstration prototype that shows promising performance in the prediction of mental fatigue levels given a 3-second EEG data. It may serve as a key step towards an EEG-based mental-fatigue monitoring device.

This research has resulted in a novel method for automatic EEG artifact removal. EEG artifacts, like from ECG, EOG and EMG, typically have much higher amplitude than cerebral signals and thus impose a great difficulty in EEG interpretation. Comparing with some existing methods, the results of our numerical experiment show that a significant performance advance has been achieved in automatic EEG artifact removal using the proposed method.

The study also resulted in the invention of a serial of feature-selection methods based on a new feature-ranking criterion that is conceptually different from those used in the literature. In loose terms, this criterion evaluates the importance of a specific feature by computing the aggregate value, over the feature space, of the absolute difference of the probabilistic outputs of the learning method with and without the feature. As a result, all features can be ranked in a decreasing order of importance so that more relevant features can be identified. These new feature-selection methods are significant not only in theoretical aspect but also in application aspect. Using only important features in a mental-fatigue measurement and monitoring system can result in a higher accuracy and improved system interpretability with a simpler architecture.

In addition to the development of new methods for automatic artifact removal and feature selection, this work has investigated the use of a probabilistic multi-class SVM for measuring and monitoring mental fatigue using EEG. The numerical results show that it not only gives superior classification accuracy but also provides a valuable estimate of confidence in the prediction of mental fatigue level in a given 3-second EEG epoch.

Finally, the experiments conducted as a part of the research have also shed light on mental-fatigue assessment methods that are critical for setting up mental-fatigue EEG dataset. A new auditory working-memory vigilance task has been proposed as a critical improvement to conventional vigilance task, providing a more realistic measure of a person's mental fatigue based on more comprehensive measurable cognitive performance impairment. Moreover, new EEG features are investigated for characterizing mental fatigue and serve as good candidates for subsequent neurophysiologic investigation.

8.2 Recommendations

It should be noted that although this study has produced encouraging results on automatic measurement of mental fatigue by EEG, there are a number of challenges that need to be addressed in future investigation. The first is to consider widening the selection criteria of subjects. The subjects used in the present study were restricted to young healthy tertiary students. This might largely minimize the effect of individual differences in EEG. Future studies should include a wider spectrum of subjects to investigate a possible effect of other variables such as age, race and even some pathological conditions (e.g. chronic fatigue syndrome). The second is with regards to hardware implementation. The hardware implementation of the proposed EEG-based mental-fatigue measurement and monitoring system is also vital in the application of such system in working environment. This study focused on the signal processing methodologies, and did not attempt to find electronic and mechanical textiles for the proposed system.

The feature selection methods proposed in this study are useful for classification of mental fatigue in specific and for machine learning in general. The proposed idea of using sensitivity of posterior probabilities for feature selection appears general and should be extendable to other machine learning algorithms where probabilistic outputs are also available. One of such possibility is the softmax-based probabilistic multi-layer perceptrons neural networks, combined with which the idea of using sensitivity of posterior probabilities for feature selection can lead to a new feature selection method for MLP neural networks.

Journal Publications Related To This Thesis

Kai-Quan Shen, Chong-Jin Ong, Xiao-Ping Li, and Einar P V Wilder-Smith. Sensitivity of posterior probability as a measure of feature importance for multi-class classification problems. *Journal of Machine Learning Research*, 2008. Submitted for Publicatioin.

Kai-Quan Shen, Chong-Jin Ong, Xiao-Ping Li, and Einar P V Wilder-Smith. Feature selection via sensitivity analysis of SVM probabilistic outputs. *Machine Learning*, 70: 1–20, 2008.

K. Q. Shen, X. P. Li, C. J. Ong, S. Y. Shao, and E. P. Wilder-Smith. EEG-based mental fatigue measurement using multi-class support vector machines with confidence estimate. *Clin Neurophysiol*, 119(7):1524–33, 2008a.

Shi-Yun Shao, Kai-Quan Shen, Chong-Jin Ong, Einar P. V. Wilder-Smith, and Xiao-Ping Li. Automatic EEG artifact removal: a weighted support-vector-machine approach with error correction. *IEEE Trans Biomed Eng*, 2008. In Press.

Mervyn V. M. Yeo, Xiao-Ping Li, Kai-Quan Shen, and Einar P. V. Wilder-Smith. Can SVM be used for automatic EEG detection of drowsiness during car driving? *Safety Science*, 2008. In Press.

Rohit Tyagi, Kai-Quan Shen, Shi-Yun Shao, and Xiao-Ping Li. A novel auditory working-memory vigilance task for mental fatigue assessment. *Safety Science*, 2008. In Press.

Jian-Bo Yang, Kai-Quan Shen, Chong-Jin Ong, and Xiao-Ping Li. Feature selection via sensitivity analysis of MLP probabilistic outputs. *IEEE Trans Neural Networks*, 2008. Submitted for Publication.

K. Q. Shen, C. J. Ong, X. P. Li, Z. Hui, and E. P. Wilder-Smith. A feature selection method for multilevel mental fatigue EEG classification. *IEEE Trans Biomed Eng*, 54 (7):1231–7, 2007.

Kai-Quan Shen, Chong-Jin Ong, Xiao-Ping Li, Hui Zheng, and Einar P V Wilder-Smith. Feature selection using SVM probabilistic outputs. *LNCS*, 4232:782–791, 2006.

Lian-Yi Zhang, Chong-Xun Zheng, Xiao-Ping Li, and Kai-Quan Shen. Measuring kolmogorov entropy of EEG for studying the state of mental fatigue. *Chinese Journal of Biomedical Engineering*, 26(2):170–6, 2007.

Lian-Yi Zhang, Chong-Xun Zheng, Xiao-Ping Li, and Kai-Quan Shen. Feasibility study of mental fatigue grade based on kolmogorov entropy. *Space Medicine & Medical Engineering*, 18(5):375–380, 2005.

Bibliography

T. Åerstedt. Sleepiness as a consequence of shift work. *Sleep*, 11(1):17–34, 1988.

T. Åerstedt, G. Kecklund, M. Gillberg, and A. Lowden. Days of recovery. In L. Hartley, editor, *Proceedings of the Fourth International Conference on Fatigue and Transportation*, Perth, Australia, 2000.

L. I. Aftanas, N. V. Lotova, V. I. Koshkarov, V. L. Pokrovskaja, S. A. Popov, and V. P. Makhnev. Non-linear analysis of emotion EEG: calculation of kolmogorov entropy and the principal Lyapunov exponent. *Neurosci Lett*, 226(1):13–16, 1997.

Paulo Aguiar, André David, Sandra Paulo, and Agostinho Rosa. EEGSOLVER—brain activity and genetic algorithms. In *The 15th Sympoium on Applied Computing (SAC 2000)*, Como, Italy, 2000.

T. Åkerstedt and M. Gillberg. Subjective and objective sleepiness in the active individual. *Int J Neurosci*, 52(1-2):29–37, 1990.

T. Åkerstedt, G. Kecklund, and A. Knutsson. Manifest sleepiness and the spectral content of the EEG during shift work. *Sleep*, 14(3):221–225, 1991.

D. M. Alexander, D. F. Hermens, H. A. Keage, C. R. Clark, L. M. Williams, M. R. Kohn, S. D. Clarke, C. Lamb, and E. Gordon. Event-related wave activity in the EEG provides new marker of ADHD. *Clin Neurophysiol*, 119(1):163–79, 2008.

C. E. Alloway, R. D. Ogilvie, and C. M. Shapiro. The alpha attenuation test: assessing excessive daytime sleepiness in narcolepsy-cataplexy. *Sleep*, 20(4):258–66, 1997.

J. Alper. EEG + MRI: a sum greater than the parts. *Science*, 261(5121):559, 1993.

P. Anderer, S. Roberts, A. Schlogl, G. Gruber, G. Klosch, W. Herrmann, P. Rappelsberger, O Filz, M.J. Barbanoj, G Dorffner, and B. Saletu. Artifact processing in computerized analysis of sleep EEG – A review. *Neuopsychobiology*, 40(3):150–157, 1999.

J. Todd Arnedt, Gerald J. S. Wilde, Peter W. Munt, and Alistair W. MacLean. How do prolonged wakefulness and alcohol compare in the decrements they produce on a simulated driving task? *Accident Analysis and Prevention*, 33(3):337–344, 2001.

James Arruda, R. Amoss, Kerry Coburn, and Heather McGee. A quantitative electroencephalographic correlate of sustained attention processing. *Applied Psychophysiology and Biofeedback*, 32(1):11–17, 2007.

P. Artaud, S. Planque, C. Lavergne, H. Cara, P. de Lepine, C. Tarriere, and B. Gueguen. An on-board system for detecting lapses of alertness in car driving. In *The 14th International Technical Conference on Enhanced Safety of Vehicles*, volume 1, Munich, Germany, 1994.

E. Başar. Brain natural frequencies are causal factors for resonances and induced rhythms. In E. Başar and T. H. Bulock, editors, *Induced rhythms in the brain*, pages 425–467. Birkhäuser, Boston, 1992.

T. Ban and M. Hojo. A comparative study of the effects of anti-Parkinson drugs on the oxotremorine-induced EEG and muscular activities. *Psychopharmacologia*, 19(1): 1–15, 1971.

G. L. Barkley and C. Baumgartner. MEG and EEG in epilepsy. *J Clin Neurophysiol*, 20 (3):163–78, 2003.

J. S. Barlow, O. D. Creutzfeldt, D. Michael, J. Houchin, and H. Epelbaum. Automatic adaptive segmentation of clinical EEGs. *Electroencephalogr Clin Neurophysiol*, 51 (5):512–525, 1981.

R. C. Bassett, J. H. Murphy, H. P. Velten, and B. K. Bagchi. Correlative EEG and mercury scan findings in 80 cases of brain tumor suspect. *Electroencephalogr Clin Neurophysiol*, 23(6):591–2, 1967.

J. Beatty, A. Greenberg, W. P. Deibler, and J. F. O'Hanlon. Operant control of occipital theta rhythm affects performance in a radar monitoring task. *Science*, 183(127):871–873, 1974.

K. Becker, J. K. Sinzig, and M. Holtmann. Attention deficits and subclinical epileptiform discharges: are EEG diagnostics in ADHD optional or essential? *Dev Med Child Neurol*, 46(7):501–2, 2004.

A. Bemporad and D. Mignone. A matlab function for solving mixed integer quadratic programs (version 1.06). Online, 9 May 2001. URL `:http://control.ethz.ch/~hybrid/miqp/`.

L.M. Bergasa, L.M. Bergasa, J. Nuevo, M.A. Sotelo, R. Barea, and M.E. Lopez. Real-time system for monitoring driver vigilance. *IEEE Transactions on Intelligent Transportation Systems*.

Hans Berger. Über das elektrenkephalogramm des menschen (on the EEG in humans). *Arch. Psychiatr. Nervenkr.*, 87:527–570, 1929.

C. Besthorn, H. Forstl, C. Geiger-Kabisch, H. Sattel, T. Gasser, and U. Schreiter-Gasser. EEG coherence in Alzheimer disease. *Electroencephalogr Clin Neurophysiol*, 90(3): 242–5, 1994.

C. D. Binnie, A. J. Rowan, J. Overweg, H. Meinardi, T. Wisman, A. Kamp, and F. Lopes da Silva. Telemetric EEG and video monitoring in epilepsy. *Neurology*, 31(3):298–303, 1981.

M. H. Bonnet and D. L. Arand. Impact of activity and arousal upon spectral EEG parameters. *Physiol Behav*, 74(3):291–8, 2001.

Michael H. Bonnet and Donna L. Arand. Level of arousal and the ability to maintain wakefulness. *Journal of Sleep Research*, 8(4):247–254, 1999.

Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152, 1992.

C. A. Bouman. Cluster: An unsupervised algorithm for modeling gaussian mixtures. Online, April 1997. URL `http://cobweb.ecn.purdue.edu/~bouman/software/cluster/`.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

R. P. Brenner. The interpretation of the EEG in stupor and coma. *Neurologist*, 11(5): 271–84, 2005.

C. A. Brictson. Measures of pilot performance: Comparative analysis of day and night carrier recoveries. Report, compiled and distributed by the NTIS, U.S. Department of Commerce, 1966.

D. E. Broadbent. Is a fatigue test now possible? *Ergonomics*, 22(12):1277–1290, Dec 1979.

B. E. Brodsky, B. S. Darkhovsky, A. Y. Kaplan, and S. L. Shishkin. A nonparametric method for the segmentation of the EEG. *Comput Methods Programs Biomed*, 60(2): 93–106, 1999.

B. Bromm, W. Forth, E. Richter, and E. Scharein. Effects of acetaminophen and antipyrine on non-inflammatory pain and EEG activity. *Pain*, 50(2):213–21, 1992.

K. A. Brookhuis and D. de Waard. The use of psychophysiology to assess driver status. *Ergonomics*, 36(9):1099–1110, 1993.

R. J. Broughton, J. Hasan, and W. Dunham. Anterior slow alpha of drowsiness: topography, source dipole analysis. *Sleep Research*, 23:5, 1994.

Dorothy Bruck and Danielle L. Pisani. The effects of sleep inertia on decision-making performance. *Journal of Sleep Research*, 8(2):95–103, 1999.

T.H. Bullock. Introduction to induced rhythms: a widespread, heterogeneous class of oscillations. In E. Başar and T. H. Bulock, editors, *Induced rhythms in the brain*, pages 1–26. Birkhäuser, Boston, 1992.

C. Cajochen, D. P. Brunner, K. Krauchi, P. Graw, and A. Wirz-Justice. Power density in theta/alpha frequencies of the waking EEG progressively increases during sustained wakefulness. *Sleep*, 18(10):890–4, 1995.

C. Cajochen, R. Foy, and D. J. Dijk. Frontal predominance of a relative increase in sleep delta and theta EEG activity after sleep loss in humans. *Sleep Res Online*, 2(3):65–9, 1999.

C. Cajochen, K. Blatter, and D. Wallach. Circadian and sleep-wake dependent impact on neurobehavioral function. *Psychologica Belgica*, 44(1-2):59–80, 2004.

Jr John A. Caldwell, J. Lynn Caldwell, David L. Brown, and Jennifer K. Smith. The effects of 37 hours of continuous wakefulness on the physiological arousal, cognitive performance, self-reported mood, and simulator flight performance of F-117A pilots. *Military Psychology*, 16(3):163–181, 2004.

C. L. Calhoun and M. G. Ettinger. Unusual EEG in coma after cardiac arrest. *Electroencephalogr Clin Neurophysiol*, 21(4):385–8, 1966.

Jose Luis Cantero and Mercedes Atienza. Spectral and topographic microstructure of brain alpha activity during drowsiness at sleep onset and REM sleep. *Journal of Psychophysiology*, 14(3):151–158, 2000.

Nazareth P Castellanos and Valeri A Makarov. Recovering EEG brain signals: artifact suppression with wavelet enhanced independent component analysis. *J Neurosci Methods*, 158(2):300–312, 2006.

Richard Caton. The electric currents of the brain. *British Medical Journal*, 2:278, 1875.

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46:131159, 2002.

Michael W. L. Chee and Wei Chieh Choo. Functional imaging of working memory after 24 hr of total sleep deprivation. *The Journal of Neuroscience*, 24(19):4560–4567, 2004.

N. S. Chu, C. C. Chu, S. C. Tu, and C. C. Huang. EEG spectral analysis and topographic mapping in Wilson's disease. *J Neurol Sci*, 106(1):1–9, 1991.

J. Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20: 37–46, 1960.

T. F. Collura, H. Luders, and R. C. Burgess. EEG mapping for surgery of epilepsy. *Brain Topogr*, 3(1):65–77, 1990.

P. Common. Independent component analysis – A new concept? *Signal processing*, 36: 287–314, 1994.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995.

O. D. Creutzfeldt, G. Bodenstein, and J. S. Barlow. Computerized EEG pattern classification by adaptive segmentation and probability density function classification. clinical evaluation. *Electroencephalogr Clin Neurophysiol*, 60(5):373–393, 1985.

N. Cristianini and J. Shawe-Taylor. *Introduction to support vector machines*. Cambridge University Press, Cambridge, 2000.

R. J. Croft and R. J. Barry. Removal of ocular artifact from the EEG: a review. *Neurophysiol Clin*, 30(1):5–19, 2000a.

R. J. Croft and R. J. Barry. EOG correction: which regression should we use? *Psychophysiology*, 37(1):123–125, 2000b.

A. J. Davidson. Measuring anesthesia in children using the EEG. *Paediatr Anaesth*, 16 (4):374–87, 2006.

D. Dawson and K. Reid. Fatigue, alcohol and performance impairment. *Nature*, 388 (6639):235, 1997.

G. De Benedittis and F. De Gonda. Hemispheric specialization and the perception of pain: a task-related EEG power spectrum analysis in chronic pain patients. *Pain*, 22 (4):375–84, 1985.

D. de Waard and K. A. Brookhuis. Assessing driver status: a demonstration experiment on the road. *Accid Anal Prev*, 23(4):297–307, 1991.

T. Deboer, A. Fontana, and I. Tobler. Tumor necrosis factor (TNF) ligand and TNF receptor deficiency affects sleep and the sleep EEG. *J Neurophysiol*, 88(2):839–46, 2002.

A. Delval, L. Defebvre, E. Labyt, X. Douay, J. L. Bourriez, N. Waucquiez, P. Derambure, and A. Destee. Movement-related cortical activation in familial Parkinson disease. *Neurology*, 67(6):1086–7, 2006.

J. M. Diamond. ADHD and EEG. *J Am Acad Child Adolesc Psychiatry*, 36(5):575–7, 1997.

M. Diers, C. Koeppe, E. Diesch, A. M. Stolle, R. Holzl, M. Schiltenwolf, K. van Ackern, and H. Flor. Central processing of acute muscle pain in chronic low back pain patients: an EEG mapping study. *J Clin Neurophysiol*, 24(1):76–83, 2007.

D. Dinges and M. Mallis. Managing fatigue by drowsiness detection: Can technological promises be realized? In L. R. Hartley, editor, *Managing Fatigue in Transportation: Selected Papers from the 3rd Fatigue in Transportation Conference*, pages 209–229. Elsevier, Oxford, U.K, 1998.

D. F. Dinges. An overview of sleepiness and accidents. *J Sleep Res*, 4(S2):4–14, 1995.

D. F. Dinges and J. W. Powell. Microcomputer analysis of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments and Computers*, 17:652–655, 1985.

D.F. Dinges. Perclos: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance. Technical report, Office of Motor Carrier Research and Standards, Oct. 1998.

Paul M. Dockree, Simon P. Kelly, John J. Foxe, Richard B. Reilly, and Ian H. Robertson. Optimal sustained attention is linked to the spectral content of background EEG activity: greater ongoing tonic alpha ($\sim$ 10 hz) power supports successful phasic goal activation. *European Journal of Neuroscience*, 25(3):900–907, 2007.

R. Dowman, D. Rissacher, and S. Schuckers. EEG indices of tonic pain-related activity in the somatosensory cortices. *Clin Neurophysiol*, 119(5):1201–12, 2008.

K. B. Duan and S. S. Keerthi. Which is the best multiclass SVM method? an empirical study. *Lecture Notes in Computer Science*, 3541/2005:278–285, 2005.

Kaibo Duan, S. Sathiya Keerthi, Wei Chu, Shirish Krishnaj Shevade, and Aun Neow Poo. Multi-category classification by soft-max combination of binary classifiers. *Lecture Notes in Computer Science*, 2709:125–134, 2003.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. Wiley, New York, 2nd edition, 2000.

F. H. Duffy. The state of EEG biofeedback therapy (EEG operant conditioning) in 2000: an editor's opinion. *Clin Electroencephalogr*, 31(1):V–VII, 2000.

M. Duta, C. Alford, S. Wilson, and L. Tarassenko. Neural network analysis of the mastoid EEG for the assessment of vigilance. *International Journal of Human-Computer Interaction*, 17(2):171–199, 2004.

J. S. Ebersole. EEG dipole modeling in complex partial epilepsy. *Brain Topogr*, 4(2): 113–23, 1991.

R. Eckhorn, T. Schanze, M. brosch, W. Salem, and R. Bauer. Stimulus-specific synchronizations in cat visual cortex: Multiple miscroelectrode and correction studies from several cortical areas. In E. Başar and T. H. Bulock, editors, *Induced rhythms in the brain*, pages 47–80. Birkhäuser, Boston, 1992.

A. L. Ehle and P. C. Johnson. Rapidly evolving EEG changes in a case of Alzheimer disease. *Ann Neurol*, 1(6):593–5, 1977.

Tatjana Eitrich and Bruno Lang. Efficient optimization of support vector machine learning parameters for unbalanced datasets. *Journal of Computational and Applied Mathematics*, 196(2):425–436, 2006.

C. W. Erwin, E. R. Somerville, and R. A. Radtke. A review of electroencephalographic features of normal sleep. *J Clin Neurophysiol*, 1(3):253–74, 1984.

V. Esmaeili, M. B. Shamsollahi, N. M. Arefian, and A. Assareh. Classifying depth of anesthesia using EEG features: a comparison. *Conf Proc IEEE Eng Med Biol Soc*, 2007:4106–9, 2007.

I. Feinberg and I. G. Campbell. Coadministered pentobarbital anesthesia postpones but does not block the motor and sleep EEG responses to MK-801. *Life Sci*, 60(12):PL 217–22, 1997.

J. Fell, J. Röschke, and P. Beckmann. Deterministic chaos and the first positive lyapunov exponent: a nonlinear analysis of the human electroencephalogram during sleep. *Biol Cybern*, 69(2):139–146, 1993.

P. B. Fenwick, J. B. Dollimore, and S. Walker. Digital analysis of the EEG in coma. *Electroencephalogr Clin Neurophysiol*, 26(5):535, 1969.

Sally Ferguson, Nicole Lamond, and Drew Dawson. Great barrier reef coastal pilots fatigue study. Final report for AMSA, Centre for Sleep Research, University of South Australia, 2005.

A. A. Fingelkurts, H. Rytsala, K. Suominen, E. Isometsa, and S. Kahkonen. Composition of brain oscillations in ongoing EEG during major depression disorder. *Neurosci Res*, 56(2):133–44, 2006.

S. P. Finnigan, S. E. Rose, and J. B. Chalk. Rapid EEG changes indicate reperfusion after tissue plasminogen activator injection in acute ischaemic stroke. *Clin Neurophysiol*, 117(10):2338–9, 2006.

S. P. Finnigan, S. E. Rose, and J. B. Chalk. Contralateral hemisphere delta EEG in acute stroke precedes worsening of symptoms and death. *Clin Neurophysiol*, 2008.

N. Foldvary, G. Klem, J. Hammel, W. Bingaman, I. Najm, and H. Luders. The localizing value of ictal EEG in focal epilepsy. *Neurology*, 57(11):2022–8, 2001.

C. Fonseca, J. P. Silva Cunha, R. E. Martins, V. M. Ferreira, J. P. Marques de Sa, M. A. Barbosa, and A. Martins da Silva. A novel dry active electrode for EEG recording. *IEEE Trans Biomed Eng*, 54(1):162–5, 2007.

Danielle J. Frey, Pietro Badia, and Kenneth P. Wright. Inter- and intra-individual variability in performance near the circadian nadir during sleep deprivation, 2004.

C. Fukuda, M.F. Funada, S.P. Ninomija, Y. Yazu, N. Daimon, S. Suzuki, and H. Ide. Evaluating dynamic changes of driver's awakening level by grouped &alpha; waves. In *Proc. 16th Annual International Conference of the IEEE Engineering Advances: New Opportunities for Biomedical Engineers Engineering in Medicine and Biology Society*, volume 2, pages 1318–1319, 1994.

H. Gävert, J. Hurri, J. Särelä, and A. Hyvärinen. The FastICA package for MATLAB (version 2.5). Online, Oct 2005. URL `http://www.cis.hut.fi/projects/ica/fastica/`.

L. De Gennaro, M. Ferrara, and M. Bertini. The boundary between wakefulness and sleep: quantitative electroencephalographic changes during the sleep onset period. *Neuroscience*, 107(1):1–11, 2001.

A. Gevins, H. Leong, R. Du, M. E. Smith, J. Le, D. DuRousseau, J. Zhang, and J. Libove. Towards measurement of brain function in operational environments. *Biol Psychol*, 40(1-2):169–86, 1995.

A. S. Gevins. *Overview of computer analysis*, pages 131–147. Amsterdam:Elsevier, 1987.

A. S. Gevins, G. M. Zeitlin, S. Ancoli, and C. L. Yeager. Computer rejection of EEG artifact. ii. contamination by drowsiness. *Electroencephalogr Clin Neurophysiol*, 43 (1):31–42, 1977.

M. Giagheddu, G. Tamburini, M. Piga, P. Tacconi, A. Giagheddu, A. Serra, P. Siotto, L. Satta, L. Demelia, and F. Marrosu. Comparison of MRI, EEG, EPs and ECD-SPECT in Wilson's disease. *Acta Neurol Scand*, 103(2):71–81, 2001.

G. L. Gigli and M. Valente. Sleep and EEG interictal epileptiform abnormalities in partial epilepsy. *Clin Neurophysiol*, 111 Suppl 2:S60–4, 2000.

P. gloor. *Hans Berger on the Electroencephalogram of Man*. Amsterdam:Elsevier, 1969.

D. S. Goodin, M. J. Aminoff, and K. D. Laxer. Detection of epileptiform activity by different noninvasive EEG methods in complex partial epilepsy. *Ann Neurol*, 27(3): 330–4, 1990.

A. M. Goransson, D. H. Ingvar, and F. Kutyna. Remote cerebral effects on EEG in high-energy missile trauma. *J Trauma*, 28(1 Suppl):S204–5, 1988.

S. A. Gordeev. Brain bioelectrical activity in humans with high anxiety level. *Fiziol Cheloveka*, 33(4):11–7, 2007.

E. Grandjean. Fatigue in industry. *Br J Ind Med*, 36(3):175–186, 1979.

E. Grandjean. *Fitting the task to the man : an ergonomic approach*. Taylor & Francis, London, 3rd edition, 1980.

E. P. Grandjean. Fatigue. *Am Ind Hyg Assoc J*, 31(4):401–11, 1970.

C.M. Gray, A.K. Engel, P. Konig, and W. Singer. Mechanisms underlying the generation of neuronal oscillations in cat visual cortex. In E. Başar and T. H. Bulock, editors, *Induced rhythms in the brain*, pages 29–45. Birkhäuser, Boston, 1992.

Glenn R. Griffin and Jefferson M. Koonce. Review of psychomotor skills in pilot selection research of the U. S. military services. *International Journal of Aviation Psychology*, 6(2):125–147, 1996.

P. Griss, H. K. Tolvanen-Laakso, P. Merilainen, and G. Stemme. Characterization of micromachined spiked biopotential electrodes. *IEEE Trans Biomed Eng*, 49(6):597–604, 2002.

G. Gucer, E. Niedermeyer, and D. M. Long. Thalamic EEG recordings in patients with chronic pain. *J Neurol*, 219(1):47–61, 1978.

S. Günter and H. Bunke. Feature selection algorithms for the generation of multiple classifier systems and their application to handwritten word recognition. *Pattern Recognition Letters*, 25(11):1323–1336, 2004.

I. Guyon. NIPS 2003 feature selection competition. Online, 2003. URL `http://www.nipsfsc.ecs.soton.ac.uk/`.

I. Guyon, S. Gunn, A. B. Hur, and Dror G. Design and analysis of the NIPS2003 challenge. In *Feature Extraction, Foundations and Applications*, chapter 9. Physica-Verlag:Springer, 2006a.

I. Guyon, S. Gunn, M. Nikravesh, and Zadeh L. A. *Feature Extraction, Foundations and Applications*. Physica-Verlag:Springer, 2006b.

Isabelle Guyon and André Elisseef. An introduction to variable and featue selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3): 389–422, 2002.

K. meinzer J. H. Peter U. Pfaff H. Fruhstorfer, P. Langanke. Neurophysiological vigilance indicators and operational analysis of a train vigilance monitoring device: A laboratory and field study. In R. R. Mackie, editor, *Vigilance*, pages 147–162. Plenum Press, 1977.

P. Hansotia, R. Harris, and J. Kennedy. EEG changes in Wilson's disease. *Electroencephalogr Clin Neurophysiol*, 27(5):523–8, 1969.

L. R. Hartley, P. K. Arnold, G. Smythe, and J. Hansen. Indicators of fatigue in truck drivers. *Applied Ergonomics*, 25(3):143–156, 1994.

Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

Simon Haykin. *Neural Networks*. Prentice Hall, second edition, 1999.

L. Herregods, G. Rolly, E. Mortier, M. Bogaert, and C. Mergaert. EEG and semg monitoring during induction and maintenance of anesthesia with propofol. *Int J Clin Monit Comput*, 6(2):67–73, 1989.

A. M. Hogan, E. L. Butterfield, L. Phillips, and J. A. Hadwin. Brain response to unexpected novel noises in children with low and high trait anxiety. *J Cogn Neurosci*, 19 (1):25–31, 2007.

T. Hongo, K. Kubota, and H. Shimazu. EEG spindle and depression of gamma motor activity. *J Neurophysiol*, 26:568–80, 1963.

J.A. Horne and L.A. Reyner. Driver sleepiness. *Journey of sleep research*, 4:23–29, 1995.

David W. Hosmer and Stanley Lemeshow. *Applied logistic regression*. Wiley series in probability and mathematical statistics . Applied probability and statistics. Wiley, New York, 1989.

J. S. Howitt, A. E. Hay, G. R. Shergold, and H. M. Ferres. Workload and fatigue–in-flight EEG changes. *Aviat Space Environ Med*, 49(10):1197–202, 1978.

George Hripcsak and Daniel F Heitjan. Measuring agreement in medical informatics reliability studies. *J Biomed Inform*, 35(2):99–110, 2002.

Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.

C.W. Hsu, C.C. Chang, and C.J. Lin. LibSVM: a library for support vector machines, 2004.

M. T. Huber, J. Bartling, D. Pachur, S. Woikowsky-Biedau, and S. Lautenbacher. EEG responses to tonic heat pain. *Exp Brain Res*, 173(1):14–24, 2006.

A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on neural networks*, 10(3):626–634, 1999.

A. Hyvarinen. Independent component analysis: algorithms and applications. *neural networks*, 13(4-5):411–430, 2000.

L. D. Iasemidis, J. C. Sackellares, H. P. Zaveri, and W. J. Williams. Phase space topography and the Lyapunov exponent of electrocorticograms in partial seizures. *Brain Topogr*, 2(3):187–201, 1990.

Kyoko Idogawa. On the brain wave activity of professional drivers during monotonous work. *Behaviormetrika*, 18(30):23–34, 1991.

R. Ihl, T. Dierks, E. M. Martin, L. Froolich, and K. Maurer. Topography of the maximum of the amplitude of EEG frequency bands in dementia of the alzheimer type. *Biol Psychiatry*, 39(5):319–25, 1996.

T. Inouye, S. Toi, and Y. Matsumoto. A new segmentation method of electroencephalograms by use of Akaike's information criterion. *Brain Res Cogn Brain Res*, 3(1): 33–40, 1995.

B. H. Jansen. Quantitative analysis of electroencephalograms: is there chaos in the future? *Int J Biomed Comput*, 27(2):95–123, 1991.

H. Jasper. Report of committee on methods of clinical exam in EEG. *Electroencephalography and Clinical Neurophysiology*, 10:370–375, 1958.

H.H. Jasper and H.L. Andrews. Electroencephalograph. III: noraml differentiation of occipital and precentral regions in man. *Archives of Neurology and Psychiatry*, 39: 96–115, 1938.

B. Jelles, J. H. van Birgelen, J. P. Slaets, R. E. Hekster, E. J. Jonkman, and C. J. Stam. Decrease of non-linear structure in the EEG of alzheimer patients compared to healthy controls. *Clin Neurophysiol*, 110(7):1159–67, 1999.

T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in kernel methods: support vector learning*. 1999.

R.D. Jones. Measurement of sensory-motor control performance capacities: tracking tasks. In J.D. Bronzino, editor, *The Biomedical Engineering Handbook – Biomedical Engineering Fundamentals*, pages 77:1–77:25. CRC Press, Boca Raton, Florida, 3rd edition, 2006.

M. Jospin, P. Caminal, E. W. Jensen, H. Litvan, M. Vallverdu, M. M. Struys, H. E. Vereecke, and D. T. Kaplan. Detrended fluctuation analysis of EEG as a measure of depth of anesthesia. *IEEE Trans Biomed Eng*, 54(5):840–6, 2007.

Carrie A Joyce, Irina F Gorodnitsky, and Marta Kutas. Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. *Psychophysiology*, 41(2):313–325, 2004.

T. P. Jung, S. Makeig, M. Stensmo, and T. J. Sejnowski. Estimating alertness from the EEG power spectrum. *IEEE Trans Biomed Eng*, 44(1):60–9, 1997.

T.-P. Jung, C. Humphries, T.-W. Lee, S. Makeig, M.J. McKeown, V. Iragui, and T.J. Sejnowski. Removing electroencephalographic artifacts: comparison between ICA and PCA. In *Proc. IEEE Signal Processing Society Workshop Neural Networks for Signal Processing VIII*, pages 63–72, 1998. doi: 10.1109/NNSP.1998.710633.

T. P. Jung, S. Makeig, C. Humphries, T. W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, 2000a.

T. P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T. J. Sejnowski. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clin Neurophysiol*, 111(10):1745–1758, 2000b.

T. P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T. J. Sejnowski. Analysis and visualization of single-trial event-related potentials. *Hum Brain Mapp*, 14(3):166–185, 2001.

C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24:1–10, 1991.

Anton Kamp and Fernando Lopes da Silva. Technological basis of EEG recording. In Ernst Niedermeyer and Fernando Lopes Da Silva, editors, *Electroencephalography:*

Basic Principles, Clinical Applications, and Related Fields, pages 110–121. Lippincott Williams & Wilkins, Hong Kong, 4th edition, 1999.

N. Kanamori. A spindle-like wave in the cat hippocampus: a novel vigilance level-dependent electrical activity. *Brain Res*, 334(1):180–2, 1985.

A. Ya. Kaplan and S. L. Shishkin. Application of the change-point analysis to the investigation of the brain's electrical activity. In B. E. Brodsky and B. S. Darkhovsky, editors, *Non-Parametric Statistical Diagnosis: Problems and Methods (Mathematics and Its Applications)*, pages 333–388. Kluwer Academic Publishers, 2000.

Richard Frederic Kaplan. *An innovative EEG based approach to drowsiness detection*. PhD thesis, Case westen reserve university, 1996.

J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo. A class of neural networks for independent component analysis. *IEEE transactions on neural networks*, 8(486-504), 1997.

M. Y. Kassab, M. U. Farooq, R. Diaz-Arrastia, and P. C. Van Ness. The clinical significance of EEG cyclic alternating pattern during coma. *J Clin Neurophysiol*, 24(6): 425–8, 2007.

G. Kecklund and T. Åerstedt. Sleepiness in long distance truck driving: an ambulatory EEG study of night driving. *ergonomics*, 36(9):1007 – 1017, 1993.

S. S. Keerthi. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms. *IEEE Transactions on Neural Networks*, 13(5):12251229, 2002.

M. Kerkhofs, C. Kempenaers, P. Linkowski, V. de Maertelaer, and J. Mendlewicz. Multivariate study of sleep EEG in depression. *Acta Psychiatr Scand*, 77(4):463–8, 1988.

J. Kershman, J. Vasquez, and S. Golstein. The incidence of focal and non-focal EEG abnormalities in clinical epilepsy. *Electroencephalogr Clin Neurophysiol*, 3(1):15–24, 1951.

R. Khanna, S. H. Nizamie, and A. Das. Electrical trauma, nonictal EEG changes, and mania: a case report. *J Clin Psychiatry*, 52(6):280, 1991.

William D. S. Killgore, Thomas J. Balkin, and Nancy J. Wesensten. Impaired decision making following 49 hr of sleep deprivation. *Journal of Sleep Research*, 15(1):7–13, 2006.

D. J. Kim, J. Jeong, J. H. Chae, S. Park, S. Yong Kim, H. Jin Go, I. H. Paik, K. S. Kim, and B. Choi. An estimation of the first positive lyapunov exponent of the EEG in patients with schizophrenia. *Psychiatry Res*, 98(3):177–189, 2000.

R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

A. Kok and E. J. Zeef. Arousal and effort: a review and theoretical synthesis of studies of age-related changes in event-related potentials. *Electroencephalogr Clin Neurophysiol Suppl*, 42:324–41, 1991.

M. Koskinen, S. Mustola, and T. Seppanen. Relation of EEG spectrum progression to loss of responsiveness during induction of anesthesia with propofol. *Clin Neurophysiol*, 116(9):2069–76, 2005.

J. W. Kowalski, M. Gawel, A. Pfeffer, and M. Barcikowska. The diagnostic value of EEG in alzheimer disease: correlation with the severity of mental impairment. *J Clin Neurophysiol*, 18(6):570–5, 2001.

T. Kubota, J. Fang, Z. Guan, R. A. Brown, and J. M. Krueger. Vagotomy attenuates tumor necrosis factor-alpha-induced sleep and EEG delta-activity in rats. *Am J Physiol Regul Integr Comp Physiol*, 280(4):R1213–20, 2001.

V. Kuhl and M. Lund. The prognosis of epilepsy with special regard to the course of EEG. *Electroencephalogr Clin Neurophysiol*, 23(4):394, 1967.

A. A. Kuhn, M. I. Hariz, P. Silberstein, S. Tisch, A. Kupsch, G. H. Schneider, P. Limousin-Dowsey, K. Yarrow, and P. Brown. Activation of the subthalamic region during emotional processing in Parkinson disease. *Neurology*, 65(5):707–13, 2005.

S. Kuperman, B. Johnson, S. Arndt, S. Lindgren, and M. Wolraich. Quantitative EEG differences in a nonclinical sample of children with ADHD and undifferentiated ADD. *J Am Acad Child Adolesc Psychiatry*, 35(8):1009–17, 1996.

D. J. Kupfer, F. G. Foster, L. Reich, S. K. Thompson, and B. Weiss. EEG sleep changes as predictors in depression. *Am J Psychiatry*, 133(6):622–6, 1976.

S. N. Lahiri. *Resampling methods for dependent data*. Springer series in statistics. Springer, New York, 2003.

S. K. Lal and A. Craig. A critical review of the psychophysiology of driver fatigue. *Biol Psychol*, 55(3):173–94, 2001a.

Saroj K. L. Lal and Ashley Craig. Electroencephalography activity associated with driver fatigue: Implications for a fatigue countermeasure device. *Journal of Psychophysiology*, 15(3):183–189, 2001b.

Saroj K. L. Lal, Ashley Craig, Peter Boord, Les Kirkup, and Hung Nguyen. Development of an algorithm for an EEG-based driver fatigue countermeasure. *Journal of Safety Research*, 34(3):321–328, 2003.

Saroj K.L. Lal and Ashley Craig. Driver fatigue: Electroencephalography and psychological assessment. *Psychophsiology*, 39:313–321, 2002.

T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff. Embedded methods. In I. Guyon, S. Gunn, M. Nikravesh, and Zadeh L. A., editors, *Feature Extraction, Foundations and Applications*, chapter 5. Physica-Verlag:Springer, 2006.

E. Lalo, S. Thobois, A. Sharott, G. Polo, P. Mertens, A. Pogosyan, and P. Brown. Patterns of bidirectional communication between cortex and basal ganglia during movement in patients with Parkinson disease. *J Neurosci*, 28(12):3008–16, 2008.

N. Lamond and D. Dawson. Quantifying the performance impairment associated with fatigue. *J Sleep Res*, 8(4):255–62, 1999.

Nicole Lamond, Drew Dawson, and Gregory D. Roach. Fatigue assessment in the field: Validation of a hand-held electronic psychomotor vigilance task. *Aviation, Space, and Environmental Medicine*, 76:486–489, 2005.

N. Laporte, G. Sebire, Y. Gillerot, R. Guerrini, and S. Ghariani. Cognitive epilepsy: ADHD related to focal EEG discharges. *Pediatr Neurol*, 27(4):307–11, 2002.

D. Le Pera, P. Svensson, M. Valeriani, I. Watanabe, L. Arendt-Nielsen, and A. C. Chen. Long-lasting effect evoked by tonic muscle pain on parietal EEG activity in humans. *Clin Neurophysiol*, 111(12):2130–7, 2000.

J. H. Lee and C. J. Lin. Automatic model selection for support vector machines. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2000. URL `Onlineathttp://www.csie.ntu.edu.tw/cjlin/papers/modelselect.ps.gz`.

M. S. Lee, B. H. Bae, H. Ryu, J. H. Sohn, S. Y. Kim, and H. T. Chung. Changes in alpha wave and state anxiety during chundosunbup qi-training in trainees with open eyes. *Am J Chin Med*, 25(3-4):289–99, 1997.

Y. J. Lee, Y. S. Zhu, Y. H. Xu, M. F. Shen, S. B. Tong, and N. V. Thakor. The nonlinear dynamical analysis of the EEG in schizophrenia with temporal and spatial embedding dimension. *J Med Eng Technol*, 25(2):79–83, 2001.

N. J. Legg, P. C. Gupta, and D. F. Scott. Epilepsy following cerebral abscess. a clinical and EEG study of 70 patients. *Brain*, 96(2):259–68, 1973.

X.P. Li, W. Zhou, W.W. Lee, L. Zhan, H. Zheng, K.Q. Shen, F.S. Sheu, E.P.V. Wilder-Smith, S. Graham, and C.S. Soon. Study of human brain mental fatigue by fMRI and EEG. In *12th ICBME*, Singapore, 2005.

Y. Li, S. Tong, D. Liu, Y. Gai, X. Wang, J. Wang, Y. Qiu, and Y. Zhu. Abnormal EEG complexity in patients with schizophrenia and depression. *Clin Neurophysiol*, 119(6): 1232–41, 2008.

H. T. Lin, C. J. Lin, and R. C. Weng. A note on platt's probabilistic outputs for support vector machines. Technical report, Technical Report, Department of Computer Science, National Taiwan University, 2003.

W. Lutzenberger, H. Flor, and N. Birbaumer. Enhanced dimensional complexity of the EEG during memory for personal pain in chronic pain patients. *Neurosci Lett*, 226 (3):167–70, 1997.

S. Makeig and M. Inlow. Lapses in alertness: coherence of fluctuations in performance and EEG spectrum. *Electroencephalography And Clinical Neurophysiology*, 86:23–35, 1993.

S. Makeig and T. Jung. Changes in alertness are a principal component of variance in the EEG spectrum. *NeuroReport*, 7:312–216, 1995.

S. Makeig, A. J. Bell, Jung Tzyy-Ping, and T. J. Sejnowski. Independent component analysis of electroencephalographic data. Advances in Neural Information Processing 8. Proceedings of the 1995 Conference, pages 145–51, Denver, CO, USA, 1996. MIT Press.

Scott Makeig, Tzyy-Ping Jung, and Terrence J. Sejnowski. Awareness during drowsiness: Dynamics and electrophysiological correlates. *Canadian Journal of Experimental Psychology*, 54(4):266–273, 2000.

A. Maksimow, M. Sarkela, J. W. Langsjo, E. Salmi, K. K. Kaisti, A. Yli-Hankala, S. Hinkka-Yli-Salomaki, H. Scheinin, and S. K. Jaaskelainen. Increase in high frequency EEG activity explains the poor performance of EEG spectral entropy monitor during s-ketamine anesthesia. *Clin Neurophysiol*, 117(8):1660–8, 2006.

Melissa Mercedes Mallis. *Evaluation of techniques for drowsiness detection: Experiment on performance-based validation of fatigue-tracking technologies*. Ph.d., Drexel University, 1999.

P. Maquet, J. Pters, J. Aerts, G. Delfiore, C. Degueldre, A. Luxen, and G. Franck. Functional neuroanatomy of human rapid-eye-movement sleep and dreaming. *Nature*, 383 (6596):163–166, 1996.

D. J. Mascord and R. A. Heath. Behavioral and physiological indices of fatigue in a visual tracking task. *Journal of Safety Research*, 23(1):19–25, 1992.

H. Matsuoka, T. Takahashi, M. Sasaki, K. Matsumoto, S. Yoshida, Y. Numachi, H. Saito, T. Ueno, and M. Sato. Neuropsychological EEG activation in patients with epilepsy. *Brain*, 123 (Pt 2):318–30, 2000.

M McCallum, T Sanquist, M Mitler, and G Krueger. Commercial transportation operator fatigue management reference, 2003.

J. A. McEwen, G. B. Anderson, M. D. Low, and L. C. Jenkins. Monitoring the level of anesthesia by automatic analysis of spontaneous EEG activity. *IEEE Trans Biomed Eng*, 22(4):299–305, 1975.

W. Mi, T. Sakai, T. Kudo, M. Kudo, and A. Matsuki. The interaction between fentanyl and propofol during emergence from anesthesia: monitoring with the EEG-bispectral index. *J Clin Anesth*, 15(2):103–7, 2003.

D. Michie, D.J. Spiegelhalter, and C.C. Taylor (Eds). Machine learning, neural and statistical classification, 1994. URL `http://www.amsta.leeds.ac.uk/~charles/statlog/`.

S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.R. Mullers. Fisher discriminant analysis with kernels. In *Proc. IEEE Signal Processing Society Workshop Neural Networks for Signal Processing IX*, pages 41–48, 1999.

James C. Miller. *Controlling Pilot Error: Fatigue*. McGraw-Hill, 2001.

N. E. Miller. Learning of visceral and glandular responses. *Science*, 163(866):434–45, 1969a.

N. E. Miller. Visceral learning and other additional facts potentially applicable to psychotherapy. *Int Psychiatry Clin*, 6(1):294–312, 1969b.

S. M. Mirsattari, D. H. Lee, D. Jones, F. Bihari, and J. R. Ives. MRI compatible EEG electrode system for routine use in the epilepsy monitoring unit and intensive care unit. *Clin Neurophysiol*, 115(9):2175–80, 2004.

M. M. Mitler, M. A. Carskadon, C. A. Czeisler, W. C. Dement, D. F. Dinges, and R. C. Graeber. Catastrophes, sleep, and public policy: consensus report. *Sleep*, 11(1): 100–109, 1988.

G. Modena, M. Angiolillo, R. Bertelli, and M. Bini. EEG activity in newborns delivered from mothers with and without full anesthesia. *Electroencephalogr Clin Neurophysiol*, 27(7):701, 1969.

K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. 12(2):181–201, 2001. ISSN 1045-9227. doi: 10.1109/72.914517.

R. Murenzi, J. M. Combes, A. Crossman, and P. Tchmitchian. *Wavelets*. Springer-Verlag, 1988.

M. Murias, J. M. Swanson, and R. Srinivasan. Functional connectivity of frontal cortex in healthy and ADHD children reflected in EEG coherence. *Cereb Cortex*, 17(8): 1788–99, 2007.

J. P. Murphy. The role of the EEG in the differential diagnosis of brain tumor. *South Med J*, 50(8):1013–7, 1957.

Kevin Murphy and Norman Delanty. Sleep deprivation: A clinical perspective. *Sleep and Biological Rhythms*, 5(1):2–14, 2007.

R. Naquet, R. P. Vigouroux, M. Choux, C. Baurand, and P. Chamand. The EEG of recent cranial trauma in a reanimation service. *Electroencephalogr Clin Neurophysiol*, 25 (1):88, 1968.

J. T. Narayanan, D. R. Labar, and N. Schaul. Latency to first spike in the EEG of epilepsy patients. *Seizure*, 17(1):34–41, 2008.

David F. Neri, Scott A. Shappell, and Charles A. DeJohn. Simulated sustained flight operations and performance, part 1: Effects of fatigue. *Military Psychology*, 4(3): 137–155, 1992.

J. Neumann, C Schnörr, and G. Steidl. Combined SVM-based feature selection and classification. *Machine Learning*, 61(1-3):129–150, 2005.

S. Nevsimalova, Z. Marecek, and B. Roth. An EEG study of Wilson's disease. findings in patients and heterozygous relatives. *Electroencephalogr Clin Neurophysiol*, 64(3): 191–8, 1986.

D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

N. Nicolaou and S. J. Nasuto. Temporal independent component analysis for automatic artefact removal from EEG. In *In Proceedings of the 2nd International Conference on Advances in Medical Signal and Informaton Processing*, pages 5–8, 2004.

Ernst Niedermeyer. The normal EEG of waking adult. In Ernst Niedermeyer and Fernando Lopes Da Silva, editors, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, pages 149–173. Lippincott Williams & Wilkins, Hong Kong, 4th edition, 1999.

Ernst Niedermeyer and Fernando Lopes Da Silva. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. 1999.

Jens P. Nilsson, Marie Soderstrom, Andreas U. Karlsson, Mats Lekander, Torbjorn Akerstedt, Nina Erixon Lindroth, and John Axelsson. Less effective executive functioning after one night's sleep deprivation. *Journal of Sleep Research*, 14(1):1–6, 2005.

S.P. Ninomija, M.F. Funada, Y. Yazu, H. Ide, and N. Daimon. Possibility of ecgs to improve reliability of detection system of inclining sleep stages by grouped a waves. In *Proc. 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1410–1411, 1993.

F. Nobili, F. Copello, P. Vitali, T. Prastaro, S. Carozzo, G. Perego, and G. Rodriguez. Timing of disease progression by quantitative EEG in alzheimer' s patients. *J Clin Neurophysiol*, 16(6):566–73, 1999.

E. A. Nofzinger, M. A. Mintun, M. Wiseman, D. J. Kupfer, and R. Y. Moore. Forebrain activation in REM sleep: an FDG PET study. *Brain Res*, 770(1-2):192–201, 1997.

P. Novak, S. Daniluk, S. A. Ellias, and J. M. Nazzaro. Detection of the subthalamic nucleus in microelectrographic recordings in Parkinson disease using the high-frequency ($>$ 500 hz) neuronal background: Technical note. *J Neurosurg*, 106(1):175–9, 2007.

Y. Ochi and K. Sakata. Some experiences with the photic stimulation and photometrazol activation in EEG examination in cases of brain tumor. *Folia Psychiatr Neurol Jpn*, 9 (3):243–52, 1955.

R. D. Ogilvie, I. A. Simons, R. H. Kuderian, T. MacDonald, and J. Rustenburg. Behavioral, event-related potential, and EEG/FFT changes at sleep onset. *Psychophysiology*, 28(1):54–64, 1991.

J. F. O'Hanlon and J. Beatty. Concurrence of electroencephalographic and performance changes during a simulated radar watch and some implications for the arousal theory of vigilance. In R. R. Mackie, editor, *Vigilance*, pages 189–202. Plenum Press, New York, 1977.

J. F. O'Hanlon and G. R. Kelley. Comparison of performance and physiological changes between drivers who perform well and poorly during prolonged vehicular operation. In R. R. Mackie, editor, *Vigilance*, pages 87–110. Plenum Press, New York, 1977.

B. S. Oken, M. C. Salinsky, and S. M. Elsas. Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical Neurophysiology*, 117(9):1885–1901, 2006.

O. Geoffrey Okogbaa, Richard L. Shell, and Davorka Filipusic. On the investigation of the neurophysiological correlates of knowledge worker mental fatigue using the EEG signal. *Applied Ergonomics*, 25(6):355–365, 1994.

Julie Onton, Scott Makeig, Neuper Christa, and Klimesch Wolfgang. Information-based modeling of event-related brain dynamics. In *Progress in Brain Research*, volume Volume 159, pages 99–120. Elsevier, 2006.

Alan V. Oppenheim and Ronald W. Schafer. *Discrete-time signal processing*. Prentice-Hall signal processing series. Prentice Hall, Englewood Cliffs, N.J., 1989.

Edgar E. Osuna, Robert Freund, and Federico Girosi. Support vector machines: Training and applications. Technical report ai memo 1602, MIT, 1997.

Allan I. Pack, Greg Maislin, Bethany Staley, Frances M. Pack, William C. Rogers, Charles F. P. George, and David F. Dinges. Impaired performance in commercial drivers: Role of sleep apnea and short sleep duration. *American Journal of Respiratory and Critical Care Medicine*, 174(4):446–454, 2006.

E. S. Page. A note on generating random permutations. *Applied Statistics*, 16(3):273–274, 1967.

K. A. Pearson. *Circadian Rhythms, Fatigue, and Manpower Scheduling*. Master's thesis, Naval Postgraduate School, Monterey, CA, 2004.

M. T. R. Peiris, R. D. Jones, G. J. Carroll, and P. J. Bones. Investigation of lapses of consciousness using a tracking task: Preliminary results. In *26th Annual International Conference of IEEE Engineering in Medicine and Biology Society*, pages 4721–4724, San Francisco, CA, 2004.

T. Penzel and R. Conradt. Computer based sleep recording and analysis. *Sleep Med Rev*, 4(2):131–148, 2000.

Pierre Philip, Patricia Sagaspe, Nicholas Moore, Jacques Taillard, Andre Charles, Christian Guilleminault, and Bernard Bioulac. Fatigue, sleep restriction and driving performance. *Accident Analysis and Prevention*, 37(3):473–478, 2005.

B. F. Piper, S. L. Dibble, M. J. Dodd, M. C. Weiss, R. E. Slaughter, and S. M. Paul. The revised piper fatigue scale: psychometric evaluation in women with breast cancer. *Oncol Nurs Forum*, 25(4):677–84, 1998.

J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in kernel methods: support vector learning*, chapter 12, pages 185–208. MIT Press, 1999.

J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A.J. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in Large Margin Classiers*, pages 61–74. MIT Press, Cambridge, MA:MIT, 2000.

T. Platz, I. H. Kim, H. Pintschovius, T. Winter, A. Kieselbach, K. Villringer, R. Kurth, and K. H. Mauritz. Multimodal EEG analysis in man suggests impairment-specific changes in movement-related electric brain activity after stroke. *Brain*, 123 Pt 12: 2475–90, 2000.

N. V. Ponomareva, G. I. Korovaitseva, and E. I. Rogaev. EEG alterations in non-demented individuals related to apolipoprotein E genotype and to risk of Alzheimer disease. *Neurobiol Aging*, 29(6):819–27, 2008.

C. M. Portas, G. Rees, A. M. Howseman, O. Josephs, R. Turner, and C. D. Frith. A specific role for the thalamus in mediating the interaction of attention and arousal in humans. *J Neurosci*, 18(21):8979–8989, 1998.

W. S. Pritchard and D. W. Duke. Measuring chaos in the brain: a tutorial review of nonlinear dynamical EEG analysis. *Int J Neurosci*, 67(1-4):31–80, 1992.

W. S. Pritchard and D. W. Duke. Measuring "chaos" in the brain: a tutorial review of EEG dimension estimation. *Brain Cogn*, 27(3):353–397, 1995.

J.G.W. Raaijmakers. Decision making under mental and physical stress. Technical report, TNO Institute for Perception, 1990.

Alain Rakotomamonjy. Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.

G Rätsch. Benchmark repository. Online, 2005. URL `http://ida.first.fhg.de/projects/bench/benchmarks.htm`.

Gunnar Rätsch, Takashi Onoda, and Klaus-R. Müler. Soft margins for adaboost. *Machine Learning*, 42(3):287–320, 2001.

A Rechtschaffen and A Kales. *Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. UCLA Brain Information Services/Brain Research Institute, Los Angeles, 1968.

Edward L. Reilly. EEG recording and operation of the apparatus. In Ernst Niedermeyer and Fernando Lopes Da Silva, editors, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, pages 122–142. Lippincott Williams & Wilkins, Hong Kong, 4th edition, 1999.

W. Rhodes and V. Gil. Development of a fatigue management program for canadian marine pilots. Technical report, Transportation Development Centre, Canada, 2002.

R. Rifkin and A. Klautau. In defence of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

NL Rogers, J Dorrian, and DF. Dinges. Sleep, waking and neurobehavioral performance. *Frontiers in Bioscience*, 8:1056–1067, 2003.

S. Romero, M. A. Mananas, S. Clos, S. Gimenez, and M. J. Barbanoj. Reduction of EEG artifacts by ICA in different sleep stages. volume 3 of *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, pages 2675–2678, Cancun, Mexico, 2003. Institute of Electrical and Electronics Engineers Inc.

J. Roschke, K. Mann, and J. Fell. Nonlinear EEG dynamics during sleep in depression and schizophrenia. *Int J Neurosci*, 75(3-4):271–84, 1994.

Volker Roth. Probabilistic discriminative kernel classifier for multi-class problems. *Lecture Notes in Computer Science*, 2191:246, 2001.

Anne Ryan and Brad Heath. Fatigue key to mistakes among pilots, 2007. URL `http://www.usatoday.com/news/nation/2007-11-07-airfatigue$\_$N.htm`.

Saeid Sanei and Jonathon Chambers. *EEG signal processing*. John Wiley & Sons, Chichester ; Hoboken, NJ, 2007.

G. Saon and M. Padmanabhan. Minimum Bayes error feature selection for continuous speech recognition. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 800–806. 2001.

S. N. Sarbadhikari and K. Chakrabarty. Chaos in the brain: a short review alluding to epilepsy, depression, exercise and lateralization. *Med Eng Phys*, 23(7):445–455, 2001.

J. Sarnthein, J. Stern, C. Aufenberg, V. Rousson, and D. Jeanmonod. Increased EEG power and slowed dominant frequency in patients with neurogenic pain. *Brain*, 129 (Pt 1):55–64, 2006.

R.E. Schapire. *The design and analysis of efficient learning algorithms*. 1992.

E. Schiff, C. Dougan, and L. Welch. The conditioned PGR and the EEG as indicators of anxiety. *J Abnorm Psychol*, 44(4):549–52, 1949.

A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, and G. Pfurtscheller. A fully automated correction method of EOG artifacts in EEG recordings. *Clin Neurophysiol*, 118(1):98–104, 2007.

C. Shagass. Anxiety, depression, and the photically driven electroencephalogram. *AMA Arch Neurol Psychiatry*, 74(1):3–10, 1955.

C. M. Shapiro, M. Flanigan, J. A. Fleming, R. Morehouse, A. Moscovitch, J. Plamondon, L. Reinish, and G. M. Devins. Development of an adjective checklist to measure five faces of fatigue and sleepiness. data from a national survey of insomniacs. *J Psychosom Res*, 52(6):467–73, 2002.

John Crossley Shaw. *The Brain's Alpha Rhythms and the Mind*. Elsevier, 2003.

L. Shoker, S. Sanei, and J. Chambers. Artifact removal from electroencephalograms using a hybrid BSS-SVM algorithm. 12(10):721–724, 2005.

D. Silverman, S. Parandian, and H. Shenkin. Effect of intravenous urea on the EEG of brain tumor patients. *Electroencephalogr Clin Neurophysiol*, 13:587–90, 1961.

Vikas Sindhwani, Pushpak Bhattacharya, and Subrata Rakshit. Information theoretic feature crediting in multiclass support vector machines. In *In Proceesing of 1st SIAM Int. Conf. on Data Mining (SDM 2001)*, Chicago, IL, USA, 5-7 April 2001.

M. E. Smith, L. K. McEvoy, and A. Gevins. The impact of moderate sleep loss on neurophysiologic signals during working-memory task performance. *Sleep*, 25(7): 784–94, 2002.

S. M. Snyder, H. Quintana, S. B. Sexson, P. Knott, A. F. Haque, and D. A. Reynolds. Blinded, multi-center validation of EEG and rating scales in identifying ADHD within a clinical sample. *Psychiatry Res*, 2008.

H. Soininen, P. J. Riekkinen, J. Partanen, E. L. Helkala, V. Laulumaa, J. Jolkkonen, and K. Reinikainen. Cerebrospinal fluid somatostatin correlates with spectral EEG variables and with parietotemporal cognitive dysfunction in alzheimer patients. *Neurosci Lett*, 85(1):131–6, 1988.

D. Sommer, T. Hink, and M. Golz. Application of learning vector quantization to detect driver dozing-off. In *European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (EUNITE 2002)*, pages 119–123, 2002.

Erwin-Josef Speckmann and Christian E. Elger. Introduction to the neurophysiological basis of the EEG and DC potentials. In Ernst Niedermeyer and Fernando Lopes Da Silva, editors, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, pages 15–27. Lippincott Williams & Wilkins, Hong Kong, 4th edition, 1999.

Rainer Spehlmann. *EEG primer*. Elsevier/North-Holland Biomedical, New York, N.Y., 1981.

Mark A. Staal. Stress, cognition, and human performance: A literature review and conceptual framework. NaSA technical memorandum 2004-212824, Ames Research Center, Aug 2004 2004.

Claudio Stampi, Polly Stone, and Akihiro Michimori. A new quantitative method for assessing sleepiness: The alpha attenuation test. *Work & Stress*, 9(2):368 – 376, 1995.

Stanford Sleep Disorders Clinic and Research Center. Why should we care about sleep? the toll of daytime sleepiness and sleep disorders on society. 1991.

S. S. Steiner and W. M. Dince. Biofeedback efficacy studies: a critique of critiques. *Biofeedback Self Regul*, 6(3):275–88, 1981.

W. K. Strik, R. Chiaramonti, G. C. Muscas, M. Paganini, T. J. Mueller, A. J. Fallgatter, A. Versari, and R. Zappoli. Decreased EEG microstate duration and anteriorisation of the brain electrical fields in mild and moderate dementia of the alzheimer type. *Psychiatry Res*, 75(3):183–91, 1997.

Jane C. Stutts, Jean W. Wilkins, and Bradley V. Vaughn. Why do people have drowsy driving crashes? input from drivers who just did, 1999.

H. Suttmann, G. Juhl, B. Baur, W. Morgenstern, and A. Doenicke. Visual EEG analysis in controlling intravenous anesthesia using propofol. *Anaesthesist*, 38(4):180–8, 1989.

B. A. Taheri, R. T. Knight, and R. L. Smith. A dry electrode for EEG recording. *Electroencephalogr Clin Neurophysiol*, 90(5):376–83, 1994.

Patricia Tassi, Anne Bonnefond, Alain Hoeft, Roland Eschenlauer, and Alain Muzet. Arousal and vigilance: Do they differ? study in a sleep inertia paradigm. *Sleep Research Online*, 5(3):83–87, 2003.

F. I. Tezer, N. Dericioglu, and S. Saygi. Generalized spike-wave discharges with focal onset in a patient with head trauma and diffuse cerebral lesions: a case report with EEG and cranial mri findings. *Clin EEG Neurosci*, 35(3):151–7, 2004.

N.V. Thakor and S Tong. Advances in quantitative electroencephalogram analysis methods. *Annual Review of Biomedical Engineering*, 6(9):1–43, 2004.

R. W. Thatcher. EEG operant conditioning (biofeedback) and traumatic brain injury. *Clin Electroencephalogr*, 31(1):38–44, 2000.

R. W. Thatcher, R. A. Walker, I. Gerson, and F. H. Geisler. EEG discriminant analyses of mild head trauma. *Electroencephalogr Clin Neurophysiol*, 73(2):94–106, 1989.

D. R. Thorne, D. E. Johnson, D. P. Redmond, H. C. Sing, and J. M. Shapiro. The walter reed palm-held psychomotor vigilance test. *Behavior Research Methods, Instruments and Computers*, 37(1):111–118, 2005.

L. Torsvall and T. Åkerstedt. Sleepiness on the job: continuously measured EEG changes in train drivers. *Electroencephalography and Clinical Neurophysiology*, 66: 709–719, 1987.

L. Torsvall and T. Åkerstedt. Extreme sleepiness: quantification of EOG and spectral EEG parameters. *Int J Neurosci*, 38(3-4):435–441, 1988.

Lars Torsvall, Törbjourn Åkerstedt, Katja Gillander, and Anders Knutsson. Sleep on the night shift: 24-hour EEG monitoring of spontaneous sleep/wake behavior. *Psychophysiology*, 26(3):352–358, 1989.

R. E. Townsend and L. C. Johnson. Relation of frequency-analyzed EEG to monitoring behavior. *Electroencephalogr Clin Neurophysiol*, 47(3):272–279, 1979.

L. J. Trejo, K. Knuth, R. Prado, R. Rosipal, K. Kubitz, R. Kochavi, B. Matthews, and Y. Zhang. EEG-based estimation of mental fatigue: Convergent evidence for a three-state model. *Lecture Notes in Computer Science*, 4565/2007:201–211, 2007.

D. L. Trudeau, P. Thuras, and H. Stockley. Quantitative EEG findings associated with chronic stimulant and cannabis abuse and adhd in an adult male substance use disorder population. *Clin Electroencephalogr*, 30(4):165–74, 1999.

Elena Urrestarazu, Jorge Iriarte, Manuel Alegre, Miguel Valencia, César Viteri, and Julio Artieda. Independent component analysis removing artifacts in ictal recordings. *Epilepsia*, 45(9):1071–1078, 2004.

F. Vaca. Get safe...get sleep! *Annals of Emergency Medicine*, 45:434–436, 2005.

M. J. van Putten and D. L. Tavy. Continuous quantitative EEG monitoring in hemispheric stroke patients using the brain symmetry index. *Stroke*, 35(11):2489–92, 2004.

V. N. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9):2013–2036, 2000.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer-Verlag, 1995.

Vladimir Naumovich Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York, 1998. Vladimir N. Vapnik. ill.

J. Vardi, H. Glaubman, J. M. Rabey, and M. Streifler. Myoclonic attacks induced by L-dopa and bromocryptin in Parkinson patients: a sleep EEG study. *J Neurol*, 218(1): 35–42, 1978.

R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Trans Biomed Eng*, 47 (5):589–593, 2000.

R.N. Vigário. Extraction of ocular artefacts from EEG using independent component analysis. *Electroencephalography And Clinical Neurophysiology*, 103(3):395–404, 1997.

J. Virkkala, J. Hasan, A. Värri, S.-L. Himanen, and M. Härmä. The use of two-channel electro-oculography in automatic detection of unintentional sleep onset. *J Neurosci Methods*, 163(1):137–144, 2007.

S. L. Visser and J. U. Postma. Influence of L-dopa on the EEG and EMG in Parkinson patients. *Psychiatr Neurol Neurochir*, 74(4):315–21, 1971.

J. Vock, P. Achermann, M. Bischof, M. Milanova, C. Muller, A. Nirkko, C. Roth, and C. L. Bassetti. Evolution of sleep and sleep EEG after hemispheric stroke. *J Sleep Res*, 11(4):331–8, 2002.

A. Vuckovic, V. Radivojevic, A. C. Chen, and D. Popovic. Automatic recognition of alertness and drowsiness from EEG by an artificial neural network. *Med Eng Phys*, 24(5):349–60, 2002.

Garrick L Wallstrom, Robert E Kass, Anita Miller, Jeffrey F Cohn, and Nathan A Fox. Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based methods. *Int J Psychophysiol*, 53(2):105–119, 2004.

T. Warbrick, D. Sheffield, and A. Nouwen. Effects of pain-related anxiety on components of the pain event-related potential. *Psychophysiology*, 43(5):481–5, 2006.

H. Weinberg, J.J. Jantzen, D. Cheyne, Paul Carson, and Alex Vincent. Measurement and monitoring of the effects of work schedule and jet lag on the information processing capacity of individual pilots. Technical Report TP 13193E, Report for Transportation Development Centre and Civil Aviation Directorate of Safety and Security Group, Transport Canada, March 1998 1998.

A. M. Weinstein. Visual erps evidence for enhanced processing of threatening information in anxious university students. *Biol Psychiatry*, 37(12):847–58, 1995.

Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. Feature selection for SVMs. In *NIPS*, volume 13, pages 668–674, 2001.

Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003.

J. White, Teresa Hutchens, and Joel Lubar. Quantitative EEG assessment during neuropsychological task performance in adults with attention deficit hyperactivity disorder. *Journal of Adult Development*, 12(2):113–121, 2005.

A. M. Williamson, A. M. Feyer, R. P. Mattick, R. Friswell, and S. Finlay-Brown. Developing measures of fatigue using an alcohol comparison to validate the effects of fatigue on performance. *Accid Anal Prev*, 33(3):313–26, 2001.

Glenn F. Wilson. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *International Journal of Aviation Psychology*, 12(1): 3–18, 2002.

Glenn F. Wilson, John A. Caldwell, and Christopher A. Russell. Performance and psychophysiological measures of fatigue effects on aviation related tasks of varying difficulty. *International Journal of Aviation Psychology*, 17(2):219–247, 2007.

J. H. Wood, K. S. Polyzoidis, C. M. Epstein, G. L. Gibby, and G. T. Tindall. Quantitative EEG alterations after isovolemic-hemodilutional augmentation of cerebral perfusion in stroke patients. *Neurology*, 34(6):764–8, 1984.

D. V. Wray and J. J. Hablitz. Selective amplitude histograms: a statistical approach to EEG-single unit relationships in generalized epilepsy. *Brain Res*, 154(2):317–29, 1978.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.

C. M. Yang and C. H. Wu. The situational fatigue scale: a different approach to measuring fatigue. *Qual Life Res*, 14(5):1357–62, 2005.

Mervyn V M Yeo, Xiaoping Li, and Einar P V Wilder-Smith. Characteristic EEG differences between voluntary recumbent sleep onset in bed and involuntary sleep onset in a driving simulator. *Clin Neurophysiol*, 118(6):1315–1323, 2007.

G. B. Young. The EEG in coma. *J Clin Neurophysiol*, 17(5):473–85, 2000.

O. Zachrisson, B. Regland, M. Jahreskog, M. Kron, and C. G. Gottfries. A rating scale for fibromyalgia and chronic fatigue syndrome (the fibrofatigue scale). *J Psychosom Res*, 52(6):501–9, 2002.

X. S. Zhang, R. J. Roy, and E. W. Jensen. EEG complexity as a measure of depth of anesthesia for patients. *IEEE Trans Biomed Eng*, 48(12):1424–33, 2001.

Appendix A

Definition of the Six Features Used in the Automatic Artifact Removal System

Given an IC, $\mathbf{s}_i$, the six features extracted from the IC were defined as follows.

Feature 1: It is defined (Shoker et al., 2005) as the ratio between the peak amplitude and the variance of the IC:

$$g_1(\mathbf{s}_i) = \frac{\max|\mathbf{s}_i|}{\sigma^2_{\mathbf{s}_i}}, \tag{A.1}$$

where $\sigma_{\mathbf{s}_i}$ is the standard deviation of time series $\mathbf{s}_i$.

Feature 2: It is essentially the normalized skewness of $\mathbf{s}_i$ as follows (Shoker et al., 2005).

$$g_2(\mathbf{s}_i) = \frac{E\{\mathbf{s}_i^3\}}{\sigma^3_{\mathbf{s}_i}}, \tag{A.2}$$

where the operator E denotes the mathematical expectation. (A-2) *Feature 3:* This feature measures the cross-correlation between $\mathbf{s}_i$ and reference EEG signals collected from eye-blinking dominated EEG channels, i.e. Fp1, Fp2, F3, F4, O1, O2. The reference EEG signals are chosen from an EEG database distinct from the database used for training and testing of the artifact removal system (see Shoker et al. (2005) for details). It is given by

$$g_3(\mathbf{s}_i) = \frac{1}{6}\sum_{j=1} 6 \max_{\tau} |E\{\mathbf{z}_j^0(t)\mathbf{s}_i(t+\tau)\}|. \tag{A.3}$$

Feature 4: This feature is the Kullback-Leibler (KL) distance between the probability density function (PDF) of $\mathbf{s}_i$ and that of a reference EOG IC which is decomposed from an EEG epoch distinct from those used for training and testing (Shoker et al., 2005). It is given by

$$\begin{aligned} g_4(\mathbf{s}_i) &= D_{KL}(\mathbf{P}(\mathbf{s}_i) \parallel \mathbf{P}(\mathbf{s}_{eog}^0)) \\ &= \int \mathbf{P}(\mathbf{s}_i) \ln \frac{\mathbf{P}(\mathbf{s}_i)}{\mathbf{P}(\mathbf{s}_{eog}^0)} \mathrm{d}\mathbf{s}_{\mathrm{i}}, \end{aligned} \tag{A.4}$$

where $\mathbf{P}(\mathbf{s}_i)$ and $\mathbf{P}(\mathbf{s}_{eog}^0)$ are the PDF of $\mathbf{s}_i$ and the reference EOG IC, $\mathbf{s}_{eog}^0$, respectively.

Feature 5: The fifth feature is the variance of scalp distribution of $\mathbf{s}_i$, given by

$$g_5(\mathbf{s}_i) = \mathrm{var}(\frac{\mathbf{a}_{\mathrm{i}}}{||\mathbf{a}_{\mathrm{i}}||}), \tag{A.5}$$

where $\mathbf{a}_i$ refers to the scalp distribution coefficients in mixing matrix corresponding to $\mathbf{s}_i$. This feature is specially proposed for ECG ICs because empirical evidences have shown that their unique scalp distribution gives smaller variance than other types of ICs.

Feature 6: This feature is similar to the feature 4 and it computes the KL distance between the PDF of $\mathbf{s}_i$ and that of a reference ECG IC, $\mathbf{s}^0_{ecg}$ via the following equation:

$$\begin{aligned} g_4(\mathbf{s}_i) &= D_{KL}(\mathbf{P}(\mathbf{s}_i) \parallel \mathbf{P}(\mathbf{s}^0_{ecg})) \\ &= \int \mathbf{P}(\mathbf{s}_i) \ln \frac{\mathbf{P}(\mathbf{s}_i)}{\mathbf{P}(\mathbf{s}^0_{ecg})} \mathrm{d}\mathbf{s}_\mathrm{i}. \end{aligned} \tag{A.6}$$

This feature is proposed to capture the distinct PDF of ECG ICs due to their unique composition of P wave, QRS complex and T wave.

It is worth noting that features 3, 4 and 6 require reference signals obtained from distinct EEG epochs that are not part of training and testing datasets. They do not require additional reference EEG channels which are generally required in many non-ICA based artifact removal methods.

Appendix B

Derivation of FSPP4 in Chapter 5

This appendix shows the derivation of $\partial \hat{p}(\omega|\mathbf{vx}_j)/\partial \mathbf{v}^i$ used in Equation (5.25) of FSPP4. Let $\hat{p}_j$ and f_j denote $\hat{p}(\omega|\mathbf{vx}_j)$ and $f(\mathbf{vx}_j)$ respectively. Suppose there are m support vectors after the training/tuning of SVM. Let $I_1 = \{k|0 < \alpha_k < C\}$ and $I_1 = \{k|\alpha_k = C\}$ with cardinalities m_1 and m_2 respectively with $m_1 + m_2 = m$. From Equations (5.4), (5.6) and (5.7), it is easy to see that

$$\left.\frac{\partial \hat{p}_j}{\partial \mathbf{v}^i}\right|_{\mathbf{v}^i=1} = -\frac{\exp(Af_j+B)}{[1+\exp(Af_j+B)]^2}\left.\left[A\frac{\partial f_j}{\partial \mathbf{v}^i} + f_j\frac{\partial A}{\partial \mathbf{v}^i} + \frac{\partial B}{\partial \mathbf{v}^i}\right]\right|_{\mathbf{v}^i=1}, \tag{B.1}$$

with

$$\begin{aligned}\frac{\partial f_j}{\partial \mathbf{v}^i} = \sum_{k=1}^{m} \big[&(-2\gamma)\alpha_k y_k(\mathbf{x}_{k,i} - \mathbf{x}_{j,i})^2 K(\mathbf{vx}_k, \mathbf{vx}_j) + \\ &y_k K(\mathbf{vx}_k, \mathbf{vx}_j)\partial\alpha_k/\partial v^i\big] + \partial b/\mathbf{v}^i.\end{aligned} \tag{B.2}$$

Expression of the 1^{st} term in the RHS of Equation (B.1) involves the evaluations of $\partial\alpha_k/\partial\mathbf{v}^i$ for $k \in I_1$ and $\partial b/\partial\mathbf{v}^i$ as shown in Equation (B.2), where the mild assumption of $\partial\alpha_k/\partial v^i = 0$ for $k \in I_2$ is used. Using the Karush-Kuhn-Tucker (KKT) conditions

(Cristianini and Shawe-Taylor, 2000) of the SVM solutions, it is not difficult to show that

$$\left\{\begin{array}{c} \sum_{k\in I_1}\alpha_k y_k K(\mathbf{v}\mathbf{x}_k,\mathbf{v}\mathbf{x}_p)+\sum_{k\in I_2}\alpha_k y_k K(\mathbf{v}\mathbf{x}_k,\mathbf{v}\mathbf{x}_p)+b=y_p,\forall\, p\in I_1 \\ \sum_{k\in I_1}\alpha_k y_k+\sum_{k\in I_2}\alpha_k y_k=0 \end{array}\right. , \tag{B.3}$$

or

$$\begin{bmatrix} \mathbf{A} & \mathbf{e} \\ \tilde{\mathbf{y}}^T & 0 \end{bmatrix}\begin{bmatrix} \tilde{\boldsymbol{\alpha}} \\ b \end{bmatrix}+\begin{bmatrix} \boldsymbol{\beta} \\ \beta_0 \end{bmatrix}=\begin{bmatrix} \tilde{\mathbf{y}} \\ 0 \end{bmatrix}, \tag{B.4}$$

where $\mathbf{A}_{pk}=y_k K(\mathbf{v}\mathbf{x}_k,\mathbf{v}\mathbf{x}_p)$, $\tilde{\mathbf{y}}$ is the vector of y_i $(i\in I_1)$, $\mathbf{e}$ is $m_1\times 1$ vector of all 1, $\tilde{\boldsymbol{\alpha}}$ is the vector of α_i $(i\in I_1)$, $\beta_0=\sum_{k\in I_2}\alpha_k y_k$ and $\boldsymbol{\beta}_p=\sum_{k\in I_2}\alpha_k y_k K(\mathbf{v}\mathbf{x}_k,\mathbf{v}\mathbf{x}_p)$. Differentiate Equation (B.4) with respect to $\mathbf{v}^i$ yields

$$\begin{bmatrix} \frac{\partial\tilde{\boldsymbol{\alpha}}}{\partial\mathbf{v}^i} \\ \frac{\partial b}{\partial\mathbf{v}^i} \end{bmatrix}=-\begin{bmatrix} \mathbf{A} & \mathbf{e} \\ \tilde{\mathbf{y}}^T & 0 \end{bmatrix}^{-1}\left\{\begin{bmatrix} \frac{\partial\boldsymbol{\beta}}{\partial\mathbf{v}^i} \\ 0 \end{bmatrix}+\begin{bmatrix} \frac{\partial\mathbf{A}}{\partial\mathbf{v}^i} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}\begin{bmatrix} \tilde{\boldsymbol{\alpha}} \\ b \end{bmatrix}\right\}. \tag{B.5}$$

The 2nd and 3rd terms in the RHS of Equation (B.1) involve differentiations of A and B. From Equation (5.8), the solutions for A and B have to satisfy

$$\frac{\partial F(A,B)}{\partial A}=-\sum_j\left(\frac{t_j}{\hat{p}_j}+\frac{1-t_j}{1-\hat{p}_j}\right)\frac{\partial\hat{p}_j}{\partial A}=0; \tag{B.6}$$

$$\frac{\partial F(A,B)}{\partial B}=-\sum_j\left(\frac{t_j}{\hat{p}_j}+\frac{1-t_j}{1-\hat{p}_j}\right)\frac{\partial\hat{p}_j}{\partial B}=0. \tag{B.7}$$

Differentiate both sides of Equations (B.6) and (B.7) with respect to $\mathbf{v}^i$, we have

$$\begin{aligned}\frac{\partial^2 F(A,B)}{\partial v^i \partial A} =& \sum_j \left(\frac{t_j}{\hat{p}_j^2} - \frac{1-t_j}{(1-\hat{p}_j)^2}\right)\frac{\partial p_j}{\partial A}\frac{\partial \hat{p}_j}{\partial v^i} - \\ & \sum_j \left(\frac{t_j}{\hat{p}_j} + \frac{1-t_j}{1-\hat{p}_j}\right)\left(\frac{\partial^2 \hat{p}_j}{\partial^2 A}\frac{\partial A}{\partial v^i} + \frac{\partial^2 \hat{p}_j}{\partial B \partial A}\frac{\partial B}{\partial v^i} + \frac{\partial^2 \hat{p}_j}{\partial f_j \partial A}\frac{\partial f_j}{\partial v^i}\right) \\ =& 0;\end{aligned} \tag{B.8}$$

$$\begin{aligned}\frac{\partial^2 F(A,B)}{\partial v^i \partial B} =& \sum_j \left(\frac{t_j}{\hat{p}_j^2} - \frac{1-t_j}{(1-\hat{p}_j)^2}\right)\frac{\partial p_j}{\partial B}\frac{\partial \hat{p}_j}{\partial v^i} - \\ & \sum_j \left(\frac{t_j}{\hat{p}_j} + \frac{1-t_j}{1-\hat{p}_j}\right)\left(\frac{\partial^2 \hat{p}_j}{\partial^2 B}\frac{\partial B}{\partial v^i} + \frac{\partial^2 \hat{p}_j}{\partial A \partial B}\frac{\partial A}{\partial v^i} + \frac{\partial^2 \hat{p}_j}{\partial f_j \partial B}\frac{\partial f_j}{\partial v^i}\right) \\ =& 0.\end{aligned} \tag{B.9}$$

Note that $\partial \hat{p}_j / \partial \mathbf{v}^i$ of Equations (B.8) and (B.9) are further expressed in terms of $\partial A/\partial \mathbf{v}^i$ and $\partial B/\partial \mathbf{v}^i$ using Equation (B.1), while $\partial f_j/\partial \mathbf{v}^i$ is known from Equations (B.2), (B.5). Hence, $\partial A/\partial \mathbf{v}^i$and $\partial B/\partial \mathbf{v}^i$ can be solved from this expanded set of equations derived from Equations (B.8) and (B.9).

The evaluation of $\partial \hat{p}_j/\partial \mathbf{v}^i$ involves the full set of training samples and is often computationally expensive. Fortunately, numerical evidence shows that the magnitudes of the 2^{nd} and 3^{rd} terms in the RHS of Equation (B.1) are typically several orders smaller than the 1^{st} term. Hence, an approximate value of $\partial \hat{p}_j/\partial \mathbf{v}^i$ can be found by making the assumption that $\partial A/\partial \mathbf{v}^i = 0$ and $\partial B/\partial \mathbf{v}^i = 0$. Under this assumption, $\partial \hat{p}_j/\partial \mathbf{v}^i$ reduces to the evaluation of the 1^{st} term in the RHS of Equation (B.2), which can be obtained by Equations (B.2) and (B.5). Our numerical experiments in Chapter 5 use this approximation.

Appendix C

Proof of Theorem 6.1 in Chapter 6

Since $\mathbf{x}_{(k)}$ is derived from $\mathbf{x}$ with the values of the k^{th} feature uniformly randomly permuted by the RP process, the probability distribution of feature $\mathbf{x}^k$, $p(\mathbf{x}^k)$, is unchanged by the RP process, i.e.,

$$p(\mathbf{x}^k_{(k)}) = p(\mathbf{x}^k). \tag{C.1}$$

The vector $\mathbf{x}_{(k)}$ is that obtained from $\mathbf{x}$ with its k feature randomly perturbed. Then, its distribution

$$p(\mathbf{x}_{(k)}) = p(\mathbf{x}^k_{(k)}, \mathbf{x}_{-k}) = p(\mathbf{x}^k_{(k)})p(\mathbf{x}_{-k}) = p(\mathbf{x}^k)p(\mathbf{x}_{-k}), \tag{C.2}$$

where the second equality follows from the fact that the distribution of $p(\mathbf{x}^k_{(k)})$ is independent from $p(\mathbf{x}_{-k})$ following the RP process. Using same argument, the joint distribution

$$p(\mathbf{x}_{(k)}, \omega_i) = p(\mathbf{x}^k_{(k)})p(\mathbf{x}_{-k}, \omega_i) = p(\mathbf{x}^k)p(\mathbf{x}_{-k}, \omega_i). \tag{C.3}$$

Hence,

$$p_i(\mathbf{x}_{(k)}) = \frac{p(\omega_i, \mathbf{x}_{(k)})}{p(\mathbf{x}_{(k)})} = \frac{p(\mathbf{x}^k)p(\mathbf{x}_{-k}, \omega_i)}{p(\mathbf{x}^k)p(\mathbf{x}_{-k})} = p_i(\mathbf{x}_{-k}). \tag{C.4}$$

Using similar argument, it is not difficult to prove $p_{ij}(\mathbf{x}_{-k}) = p_{ij}(\mathbf{x}_{(k)})$.