Learning Structured Representations of Data

Etienne Barnard, Christiaan van der Walt, Marelie Davel, Charl van Heerden, Fred Senekal, Thegaran Naidoo

CSIR, Pretoria, South Africa
Department of Electrical, Electronic and Computer Engineering,
University of Pretoria, Pretoria, South Africa

{ebarnard,cvdwalt,mdavel,cvheerden,fsenekal,tnaidoo}@csir.co.za

Abstract

Bayesian networks have shown themselves to be useful tools for the analysis and modelling of large data sets. However, their complete generality leads to computational and modelling complexities that have limited their applicability. We propose an approach to simplify and constrain Bayesian networks that strikes a more useful compromise between generality and tractability. These constrained graphical will allow us to build computationally tractable models for large high-dimensional data sets.

We also describe examples of data sets drawn from image and speech processing on which we can (1) further explore this constrained set of graphical models, and (2) analyse their performance as a general-purpose statistical data analysis tool.

1. Introduction

Understanding the properties of large data sets has become a crucial activity in the Information Age: numerous tasks, ranging from the measurements of hundreds of distributed sensors, to the speech of thousands of speakers, each with distinct accents, voices and habits of speech, are undertaken continually in knowledge-intensive societies. Such data analysis plays a vital role in tasks as diverse as financial management, environmental monitoring and information technology (IT) service provision.

It is shown in [1] that the optimal choice of classifier for a classification task depends on the properties of the data set and the properties of the classifiers being considered. Significant expertise in the problem domain and data analysis is thus required to understand the properties of the classification data and significant insight into the properties of classifiers are also required to select the classifier that fits the properties of the classification task best. Mathematical modelling tools that are available for data analysis also remain highly specialized, requiring significant domain expertise to be useful in most applications. Furthermore, the dimensionalities of data sets have also increased significantly with the increased amount of available digital information, which increases the complexity of data analysis and limits the intuitive insight into the properties of the data. Understanding and modelling the properties of high-dimensional data has thus also become a crucial activity in the Information age.

It is clear that pattern recognition requires a unified framework for the analysis and modelling of arbitrary datasets. It is expected that the creation of a fully general framework will require one or more major breakthroughs [2]; however, even in the absence of such breakthroughs, much will be gained by developing representations and algorithms that treat diverse data sets in a more unified manner. The development of graphical models (also known as Bayesian networks) goes some way towards delivering such a description [3,4]; however, the excessive generality of these models has limited our intuitive insight into their properties (and has stymied attempts to create efficient learning algorithms for deriving their structure) [5].

We propose a constrained set of graphical models that strikes a more useful compromise between generality and tractability: more focused learning algorithms (e.g. the Expectation Maximization algorithm for mixture models or constrained optimisation algorithms for kernel models) tend to be ineffective at structural learning, whereas more general algorithms (e.g. Monte-Carlo based algorithms for graphical models) tend to be extremely expensive computationally. Our approach allows us to limit both of these risks, thus delivering algorithms that can efficiently describe the properties of complex real-world data sets with a limited amount of domain expertise and training data.

In Section 2 we give an overview of existing methods that are used for the analysis and modelling of data. We point out the advantages and disadvantages of these approaches, specifically in the context of high-dimensional data.

In Section 3 we formalize a new framework for the analysis of arbitrary data sets; we present a set of constrained graphical models that strikes a more useful compromise between generality and tractability than Bayesian networks. We also show, in Section 4, how these graphical models can represent the class-conditional probability density functions (pdf) learned by the Gaussian mixture model (GMM) classifier and how this framework can represent artificial data sets with known properties.

In Section 5 we discuss applications in image and speech processing with large datasets, on which our framework can be applied and evaluated, and we summarize our conclusions in Section 6.

2. Background

There are several existing approaches to analyse and model data; these techniques include (1) density estimation, (2) dimensionality reduction, (3) clustering, (4) bi-clustering, (5) topological models and (6) Bayesian networks.

There are two main approaches to density estimation, namely parametric and non-parametric density estimation [6]. Parametric approaches are fast and tractable but the

assumptions regarding the parametric shape of density functions are often too constraining, whereas non-parametric approaches make no assumptions about the form of the density functions; density functions are often estimated as e.g. a sum of kernels of sum of Gaussians. Non-parametric approaches are, however, computationally expensive (especially for high-dimensional data) and the final models give no transparent insight into the properties of the data.

Dimensionality reduction techniques attempt to project data to a lower-dimensional feature space, while still retaining as much of the information in the data as possible. A significant loss of information occurs, however, when the intrinsic dimensionality of the data (k) is higher than the dimensionality to which the data is projected. In order to visualize data, data is typically projected to 3-dimensions or lower. If k<=3, dimensionality reductions techniques can give insight into the properties of data by visualizing the projected data, if k>>3 (which is often the case for real-world data), dimensionality reduction is less useful.

Clustering techniques are very useful if the measure of similarity between objects in the feature space remains constant. A distance measure between objects is defined *a priori* for a clustering algorithm and remains the same throughout the clustering process. The properties of real-world data, however, tend to change throughout the feature space and clustering algorithms do not take this change in the relationship between variables into account.

Bi-clustering techniques try to account for this changing relationship between features throughout the feature space by clustering observations and features simultaneously. Bi-clustering techniques effectively divide the feature space into sub-feature spaces, where data points in sub-feature spaces have similar properties. Bi-clustering techniques are, however, computationally very expensive - it is known to be an NP-hard problem, for which efficient approximations have not been found.

Topological models attempt to describe the properties of data by describing how similar groups of data points are connected throughout the feature space [7]. They do not provide any information of the underlying structures of these groups of data points and do not provide probabilistic models that can be used to model the data. They are, however, useful to learn more from the data, when sufficient expertise is available for the analysis and interpretation of the data.

Bayesian networks are graphical models that can be used to learn and represent the joint pdf of data in a graphical framework. The fundamental insight motivating Bayesian networks is that multivariate probability distributions can be simplified if appropriate conditional independence relationships are recognized. If such relationships exist, an *n*-variable probability distributions can be represented in terms of conditional distributions for *n1, n2,..., nk* variables, where $n1 + n2 + ... + nk = n$. Since the complexity of estimating a multivariate probability distribution is exponential in the number of variables in the worst case [8], the savings implicit in this decomposition can be substantial. The graphical model is used to keep track of the independence relationships between groups of variables.

The application of such models requires the solution of two key problems:

- The learning problem: how does one estimate the structure and parameters of a graphical model (in practice, usually based on a number of training samples)
- The inference problem: given the values of some variables, how does one infer the most likely values of other, unknown variables, in order to compute a complete probability estimate.

Most current inference algorithms build on the message passing approach pioneered by Pearl [3]. Although such algorithms can have exponential worst-case behaviour, they are fairly efficient for appropriately-structured networks, and modern approaches have extended their applicability quite widely [8].

The learning problem usually factors into two parts, namely estimation of the appropriate structure, followed by parameter estimation for conditional distributions for all nodes within the structure [5]. For the latter problem, standard techniques from statistics are generally employed (see, for example, [6] for an overview). However, the success of these techniques depends on appropriate structure estimates, and these have not yielded well to the maximization approaches typically used in machine learning [9]. Hence, successful structure learning approaches currently require substantial domain-specific information [14].

In summary, graphical models offer an extremely attractive approach to the modelling of high-dimensional data sets. However, in full generality they require either significant domain expertise or large computational budgets, for both the inference and learning tasks. Their applicability would be greatly enhanced if an approach could be developed that removes some of these obstacles, even if their full generality is compromised in doing so.

3. Formal definition of graphical framework

3.1. Motivation for graphical models

It is shown in [10] that as the dimensionality of a feature space increases, the volume of a hyper-cube moves to the edges, whereas the volume of an ellipsoid moves to the outer shell. The volume of high-dimensional spaces thus tends to move to small regions of the feature space, which suggests that data tends to lie in manifolds (which make up small parts of the feature space) with high densities while the remaining part of the feature space is relatively empty. This phenomenon thus suggests that data, specifically in higher dimensional feature spaces, are generated from underlying manifolds.

If we consider an example of a body suit with N sensors capturing motion in 3 dimensions, we have a feature space of dimensionality 3N [11]. The exact position of a body can actually be specified by k angles between the joints of the body. The intrinsic dimensionality of the problem (k) is significantly smaller than the dimensionality of the feature space (3N). Mumford illustrated this same principle [12], by showing that high-dimensional natural images can be reduced to points on a 7-sphere. These examples suggest that high-dimensional data can be described by underlying manifolds with intrinsic dimensionalities (k) much lower than the dimensionality of the feature space (d).

In order to characterize a dataset, we thus need to identify the underlying manifolds from which the data points we observe have been generated, and we need to describe the

geometrical structure and intrinsic dimensionalities of these manifolds.

We propose a framework that consists of three components that are sufficient to describe the geometrical structure and intrinsic dimensionality of the underlying manifolds of data. This framework consists of (1) functional transformations, which are used to describe the geometrical structure of a manifold, (2) continuous pdfs which are used to (i) randomly select a point from a manifold and (ii) to model the intrinsic dimensionality of a manifold (i.e. the variation of a selected point from the manifold) and (3) discrete probability mass functions (pmf) which are used to switch between manifolds within a feature space.

Even in the case where the geometrical structure of a manifold cannot be described by a single parametric equation, we can approximate the geometrical structure with a combination of simplexes (simplicial complexes). Simplicial complexes can be used to approximate any arbitrary geometrical structure and have a point based representation which makes them easy to describe. In the next section we will illustrate how simplicial complexes can be used to generate artificial data and how they are represented in our graphical framework.

The proposed framework thus constrains our graphical networks to only three components to make practical learning algorithms more tractable, while still maintaining sufficient generality to describe datasets with underlying manifolds of any geometrical structure and intrinsic dimensionality.

In the following sub-section we propose the nomenclature of our proposed set of graphical models.

3.2. Nomenclature of graphical models

We denote random variables with bolded upper class Roman capital letters e.g. $\mathbf{X}$, a value drawn from a univariate random variable is denoted by a lower class Roman letter x and a vector drawn from a d-dimensional multivariate RV e.g. $X=[X_1,...,X_d]$ is denoted by a lower case bolded Roman letter $\mathbf{x}$. The i^{th} row vector of a matrix $\mathbf{X}$ is indicated by $\mathbf{x}_i$.

Vectors are indicated by lower case bolded Roman letters are column vectors, and a superscript capital T is used to indicate the transpose of a vector e.g. $\mathbf{x}^T$, which in this case will represent a row vector. Matrices are also indicated by bolded Roman capital letters $\mathbf{M}$, if random variables are used in the same context as matrices, the random variable will contain subscripts e.g. $\mathbf{X}_1$. $F_X(x)$ is used to indicate the cumulative distribution function (cdf) of a univariate random variable $\mathbf{X}$, and $f_X(x)$ is used to indicate the probability density function of a univariate random variable $\mathbf{X}$.

$\rho_X(x)$ is used to indicate the probability mass function (pmf) of a discrete univariate random variable. $p(\mathbf{y})$ refers to the probability that the vector $\mathbf{y}$ belongs to a specific class.

3.3. Components of graphical models

As discussed earlier, our graphical models will consist of three types of nodes: (1) continuous pdfs will be represented by circular nodes, (2) discrete pmfs will be represented by triangular nodes and (3) functional transformations will be represented by rectangular nodes.

The nodes in a graphical network will be connected by one-directional arrows, the directions of the arrows indicate the direction in which the function of each node is performed, and thus the sequence in which data is processed.

In the next section we give examples of how these graphical models can be applied.

4. Examples of graphical models

In this section we will illustrate how our graphical models can be used to (1) represent the class-conditional pdfs learned by a GMM classifier and (2) represent the underlying structure of artificially generated data.

4.1. Examples of class-conditional probability density funtions

A GMM classifier assumes that the class-condition pdf of each class consists of a mixture of Gaussian distributions; the class-conditional pdf is this the weighted sum of the pdfs of the mixtures. We can represent the class-conditional pdf learned by a GMM classifier in our graphical framework as illustrated in Figure 1.

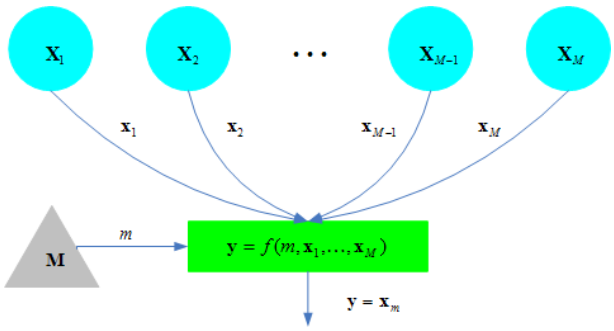

Figure 1: *Graphical representation of GMM classifier class-conditional pdf*

The output vector $\mathbf{y}$ can be expressed as

$$\mathbf{y} = \mathbf{x}_m, \tag{1}$$

where $\mathbf{x}_m$ is the sample drawn from mixture m.

As shown in Figure 1, the value of m is drawn from the discrete pmf $p_M(m)$. This pdf is characterised by the weight assigned to each mixture; the number of times the value m is drawn is proportional to the weight of mixture m, given by Π_i. The pdf of $\mathbf{y}$ can be expressed as

$$p(\mathbf{y}) = \sum_{i=1}^{M} \Pi_i p(\mathbf{x_i}) = \sum_{i=1}^{M} \Pi_i N(\boldsymbol{\mu}_\mathbf{i}, \boldsymbol{\Sigma}_\mathbf{i}), \qquad (2)$$

where M is the total number of mixtures per class and Π_i is the weight assigned to mixture i by the GMM classifier, $\boldsymbol{\mu}_i$ is the mean of mixture i and $\boldsymbol{\Sigma}_i$ is the covariance matrix of mixture i.

4.2. Examples of generating artificial datasets from geometrical structures

Artificial datasets can be generated from underlying manifolds with known geometrical structures. In this sub-section we will illustrate how these artificial datasets can be represented in the proposed graphical framework.

We firstly generate an artificial dataset by sampling data points from a 3rd order polynomial function $f_1(x) = 0.8x^3 - 0.2x^2 - 0.3x + 0.7$. Data points are uniformly sampled from the polynomial manifold by generating values for x from a uniform distribution, U(-1,1), and calculating the values of p(x). The data points sampled from this polynomial function can be regarded as data points lying on the same underlying manifold. We sample data from a second underlying manifold with the equation $f_2(x) = 0.5$ (a straight line parallel to the x-axis). Note that the manifold described by $f_1(x)$ has a dimensionality of 2 (data points vary in both dimensions) while the manifold described by $f_2(x)$ has a dimensionality of 1 (data points vary only in the x-direction). Figure 2 illustrates the dataset sampled from these two manifolds.

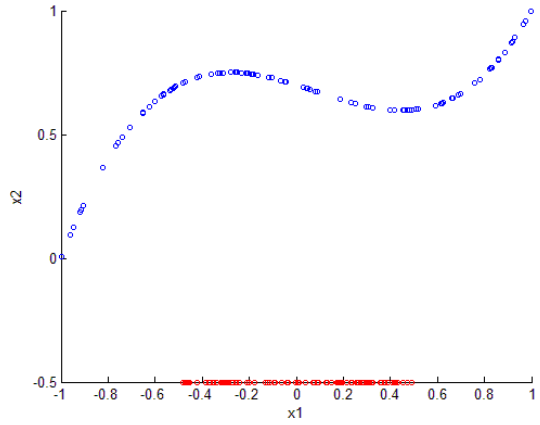

Figure 2: Dataset 1 (sampled from two polynomial functions

This dataset can be represented in our graphical framework as shown in Figure 3.

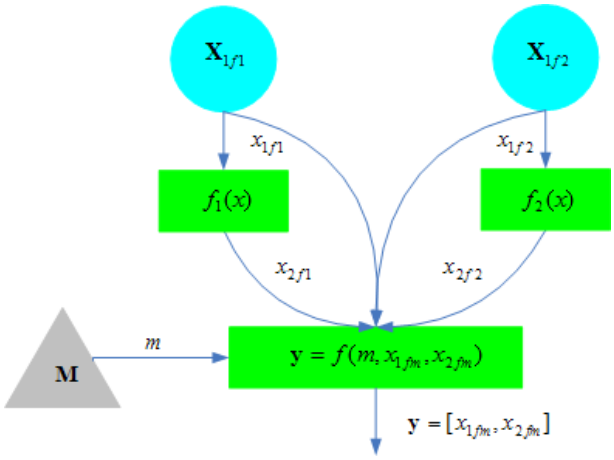

Figure 3: Graphical representation of dataset 1.

We can generate an equivalent dataset to dataset 1 by approximating the underlying structures of the two manifolds in our first example with simplexes. We can specify the anchor points of the simplexes as shown by the markers on the edges of each simplex.

We then make use of the Barycentric method to sample data points uniformly from the simplexes. The dataset sampled from the simplexes described by P, and their representations in our graphical framework are illustrated in Figures 4 and 5.

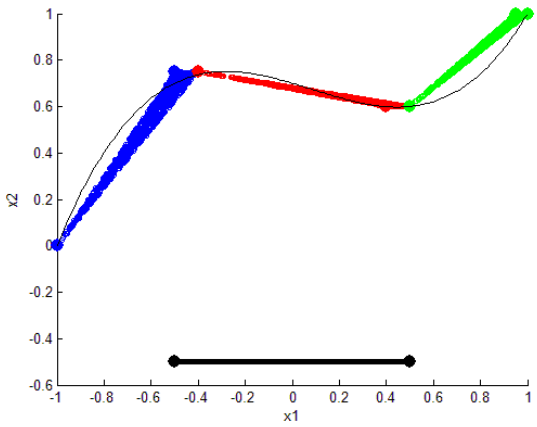

Figure 4: Dataset 2 (approximated structure of dataset 1).

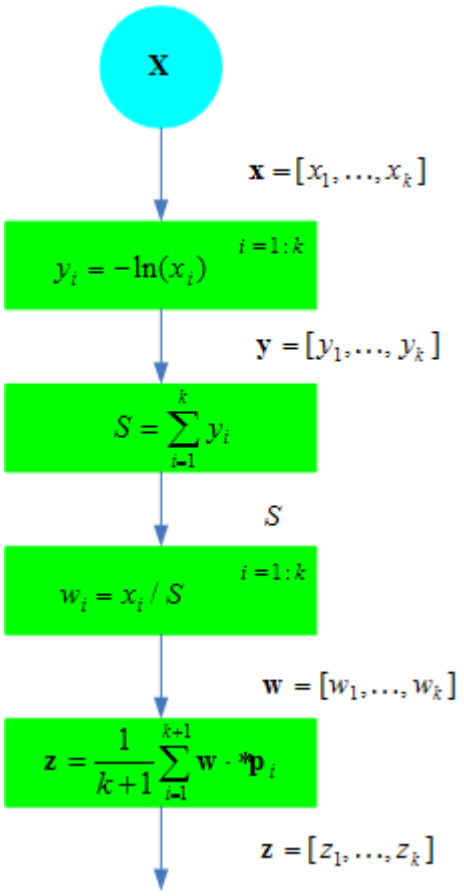

Figure 5: Graphical representation of dataset 2.

5. Application of framework to real-world projects

As part of a collaborative research project between the Meraka HLT Research Group, Material Science and Manufacturing and Modelling and Digital Science units at the CSIR, these graphical models are being applied and evaluated on three real-world projects, namely (1) age classification of speech data, (2) robotic perception and (3) computer vision for unmanned aerial vehicles. Each of these problems will be discussed in more detail in the following sub-sections.

5.1. Speaker age and gender classification

Estimating a speaker's age and gender from recordings of his or her voice is a task that has seen significant activity in the past five years [13]. On the one hand, this task is of psycho-acoustic interest: researchers wish to determine what the most important correlates of age and gender in a speaker's voice are. On the other hand, it also has significant practical importance - particularly for the design of spoken-dialogue systems that adapt to the characteristics of their users. Finally, insights gained from the study of this classification task will also be useful for the extraction of other meta-information (such as the speaker's cognitive load or physical exhaustion) from the speech signal; such meta-information is likely to become increasingly important as speech-based systems are deployed in, for example, automotive applications.

Researchers at Deutsche Telekom have developed a standard database for the age-and-gender classification problem, containing more than 50 000 recordings of speakers ranging from 7 to 72 years old. Each of these recordings is labelled as belonging to one of seven classes (children and 3 age ranges of males and of females). Mueller and Burkhardt [13] have defined a set of 22 features that can be used for this classification task, and in this sub-project we intend to optimize our graphical models using this feature set. Besides its relevance to a significant real-world task, this problem will allow us to refine our methods in a feature space of fairly low dimensionality (compared to the problems described below). State-of-the-art classification accuracy in this environment is around 50%, so there is substantial room for improvement.

5.2. Computer vision (Robotic perception)

In robotic systems, perception ability is often required to interpret external signals and build a conceptual model of the environment in order to interact with that environment. Examples of active research fields in robotic perception include speech recognition and computer vision.

Typical computer vision tasks, such as 3D reconstruction and object recognition, rely on the idea that certain features could be extracted from digital images and matched against other features. For example, in stereo vision, features are matched over the spatial domain (different images of the same scene captured at the same time instant) and geometrical calculations are performed, in structure from motion, features are matched over the temporal domain (images of a scene are captured at different time instants) and in object recognition, features are matched against previously extracted features, where features or sets of features are associated with a particular class label.

Clearly, the use of a feature-based approach is crucial to many computer vision tasks. However, the specific feature being used varies greatly from application to application. A variety of feature descriptors have been studied for different problems, for example SIFT, SURF, shape context, Harris corner and edge detectors, Haar filters, image histograms, steerable filters, differential invariants, etc. (for a review and comparison of different feature descriptors, see [14]). These techniques differ from one another at a fundamental level and although working well for different problems, there is no underlying unifying framework.

The approach of learning structured representation from data will be applied specifically to problems in object recognition. In our research, the objective is to determine the most appropriate features to use for a given classification problem, rather than trying to apply a range of existing feature descriptors (which may work well in other contexts but may not be suited to the problem). Given a set of labelled training images of natural scenes, we are extracting labelled image patches at different scales. Using this raw information, we aim to study the underlying manifolds and model the probability density functions associated with different classes. Such models should make it possible to approach classification problems in computer vision in a generic way.

5.3. Computer vision (Unmanned aerial vehicles)

This application focuses on the development of a vision-based positioning system that forms part of a rotary-winged aerial inspection platform. This sub-system is required by the inspection platform to ensure that the correct inspection data (images) are collected for offline processing.

The inspection system's onboard GPS information would be able to position it near the object that is to be inspected but it would most likely not be accurate enough for the purpose of inspection. It is therefore proposed to do the accurate positioning visually. This approach makes use of the rich information provided by the visual sensor to solve the positioning problem while benefiting from its relatively light weight to minimise the overall payload.

The first step in the proposed solution involves detecting and identifying the object of interest. The objects classification is then used to perform a model based 3D registration with the current view. This will allow the extraction of the 3D spatial transform required to move the platform into position (via the flight control system).

It is envisioned that the graphical modelling and analysis techniques being developed will be used to design effective feature extraction algorithms. This will be achieved by analysing the relationships between the features and their contribution at various stages of orientation and scale.

The feature vectors will be complex structures consisting of variable and invariant information whose influence will vary depending on the current task. The features for training and testing will be extracted from images of both simulated objects and scale-sized real world models.

6. Conclusions

In summary, graphical models offer an extremely attractive approach to the modelling of high-dimensional data sets. However, in full generality they require either significant domain expertise or large computational budgets, for both the inference and learning tasks.

We have formalised and presented a constrained set of graphical networks that will make practical learning algorithms more tractable than learning algorithms for Bayesian networks, while still maintaining sufficient generality to describe datasets with underlying manifolds of any geometrical structure and intrinsic dimensionality.

We are currently developing learning algorithms for these constrained networks; we will apply these learning algorithms to the real-world applications discussed in Section 5.

7. References

[1] C.M. van der Walt and E. Barnard, "Data characteristics that determine classifier performance", *SAIEE Africa Research Journal*, 98(3): 87-92, 2007

[2] E. Barnard, B. Palensky and P. Palensky, "Towards Learning 2.0", *in Proceedings of ICST IT-Revolutions*, 2008

[3] J. Pearl, "Fusion, propagation, and structuring in belief networks", *Artificial Intelligence*, 29(3):241 -288, 1986

[4] R.E. Neapolitan, "Learning Bayesian Networks", Prentice Hall, 2003

[5] D. Heckerman, "Learning Bayesian networks: The combination of knowledge and statistical data", *Machine Learning*, 20(3):197-243, 1995

[6] B.W. Silverman, "Density estimation for statistics and data analysis", London: Chapman and Hall, 1986

[7] P. Gaillard, M. Aupetit and G. Govaert, "Learning topology of a labeled data set with the supervised generative Gaussian graph", *Neurocomputing*, 71(7-9):1283-1299, 2008

[8] C. Andrieu, N. De Freitas, A. Doucet and M.I. Jordan, "An introduction to MCMC for Machine Learning", Machine Learning, 50(1-2):5-53, 2003

[9] S.J. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach", 2nd ed., Prentice Hall, 2003.

[10] L. Jimenez and D. Landgrebe, "Supervised classification in high dimensional space: geometrical, statistical and asymptotical properties of multivariate data", *IEEE Trans. on Systems, Man and Cybernetics*, 28(1):39-54, 1998.

[11] J. Bruske and G. Sommer, "Intrinsic dimensionality estimation with optimally topology preserving maps", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(5):572-575, 1998.

[12] A.B. Lee, K.S Pedersen and D. Mumford, "The nonlinear statistics of high-contrast patches in natural images", *International Journal of Computer Vision*, 54(1):83-103, 2003.

[13] C. Mueller and F. Burkhardt, "Combining Short-term Cepstral and Long-term Prosodic Features for Automatic Recognition of Speaker Age", in *Proc. Interspeech.*, Antwerp, Belgium, pp. 2277 – 2280, 2007

[14] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615-1630, 2005.