1-1-2012

Medical Data Analysis Method For Epilepsy

Ameen Eetemadi
Wayne State University,

MEDICAL DATA ANALYSIS METHOD FOR EPILEPSY

by

AMEEN EETEMADI

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2012

MAJOR: COMPUTER SCIENCE

Approved by:

Advisor Date

DEDICATION

To my parents

and

my dear wife

ACKNOWLEGEMENTS

I have no ability but which comes from my Lord and it’s only through his power, grace, mercy and forgiveness which I was able to accomplish this work.

In addition, I would like to extend the deepest gratitude to my advisor, Professor Farshad Fotouhi, for supporting and mentoring me from my first day at Wayne State to date. My research, in general, and this thesis, in particular, would not have been possible without his relentless guidance and patience.

I am indebted to my family, especially my parents, for their continuous support and motivation throughout this effort. Words cannot express my appreciation for the selfless outlook and affectionate prayers of my parents and my wife. I thank them for always putting my interest ahead of their personal aspirations.

I would like to thank Dr. Siadat, Dr. Elisevich and Dr. Soltanian-Zadeh for being wonderful mentors. Dr. Siadat has inspired and helped me tremendously along the project. I should say, if I have any success here, it’s because of the foundation that has been put in place by him and others in HBIDS. Dr. Soltanian-Zadeh has been my advisor in Henry Ford Health Systems. He has given me the opportunity to work on this interesting subject and has supported me patiently all along. Dr. Elisevich as the medical expert on the project has given me great feedback and guidance on usefulness and direction which I should take throughout the project.

TABLE OF CONTENTS

Dedication ii

Acknowlegements iii

List of Figures vi

List Of Tables viii

Chapter 1 Introduction 1

Contributions 2

Medical Data Mining 2

Approaches in Medical Data Mining 2

Heterogeneous and volumetric 3

Ethical and legal issues 5

Introducing the Human Brain Image Database System (HBIDS) 6

Objective 6

State of the Existing System 6

Problem statement 8

Summary 10

Chapter 2 Background and Related Work 11

The Epilepsy Disease 11

Diagnosis methods 12

Treatment 13

Related work 14

Data Preprocessing 15

Data Mining 20

Visualization 26

Summary 33

Chapter 3 Contributions 35

Data pre-processing 35

Transformation 35

Filtering 41

Data mining 42

Similarity Measurement..42

Data visualization..44

Description of the method..45

Pros and Cons..48

Summary 50

Chapter 4 Experiments 52

Implementation 52

Data Analysis results 52

Similarity Measurement..52

Data Visualization 54

Classification 61

Summary 72

Chapter 5 Conclusion 73

References 75

Abstract...79

Autobiographical Statement...80

LIST OF FIGURES

Figure 1 - Components of the HBIDS system[7] 8

Figure 2 - Parallel Coordinate Plot Fisher's Iris Data [31] 27

Figure 3 - RADVIS visualization of DNA data[33] 28

Figure 4 - Chernoff faces in various 2D positions [34] 29

Figure 5 - Example of pixel-oriented visualization using spiral technique [36] 30

Figure 6 - Example of N-vision hierarchical visualization for 5 dimentions [37]........ 31

Figure 7 - A Graph-based visualization for microarray results [38] 32

Figure 8 - Dichotomies of database attributes 36

Figure 9 – Conventional procedure for constructing flat table for RDB 39

Figure 10 – Proposed new procedure forconstructing flat table for RDB* 41

Figure 11 – Patient Data Visualization with connected scatter plots 47

Figure 12 - Patients sorted based on the weighted similarity to a selected patient. 54

Figure 13 - Connected scatter plots using brushing technique that shows the relationship between three attributes 55

Figure 14 - Connected scatter plots that shows the relationship between five attributes and brushing the patients with value ‘I’ for the ‘Postoperative Engle Class’ attribute. 57

Figure 15 - Connected scatter plots that shows the relationship between five attributes and brushing the patients with right and left side brain surgery differently. 59

Figure 16 - Connected scatter plots showing the relationship between four attributes. Patients with non-successful surgery outcome are brushed with two colors according to the value of ‘Hippocampus Volume Asymmetry’. 60

Figure 17 - Connected scatter plots showing the relationship between four attributes. Patients with successful surgery outcome are brushed with two colors according to the value of ‘Hippocampus Volume Asymmetry’. 61

Figure 18 - J48 Classification Model to predict the surgery outcome. As an example, “No (92/42)” means the predicted surgery outcome is negative. There are 92 patients whom this rule applies to them and 42 patients which contradict this prediction. 62

Figure 19 – ADTree classification model to predict the surgery outcome 64

Figure 20 - DecisionStump classification model to predict surgery outcome 66

Figure 21 - AdaBoost first classifier to predict surgery outcome 68

Figure 22 - AdaBoost second classifier to predict surgery outcome 68

Figure 23 - AdaBoost third classifier to predict surgery outcome 68

Figure 24 - OneRule classification model to predict the surgery outcome 71

LIST OF TABLES

Table 1 - Data Transformation Example (Generalization) 19

Table 2 - Joining two tables with one-to-many relationship 40

Table 3 - Result of aggregating multiple results 40

Table 4 - Classification accuracy for the J48 decision tree 62

Table 5 - Confusion matrix for the J48 decision tree 62

Table 6- Classification accuracy for ADTree experiment 65

Table 7 - Confusion matrix for the ADTree experiment........ 65

Table 8- classification accuracy measures for each class in ADTree experiment 65

Table 9 - classification accuracy for DecisionStump experiment........ 66

Table 10 - confusion matrix for DecisionStump experiment 66

Table 11 - classification accuracy measures for each class in DecisionStump experiment 66

Table 12 - Classification accuracy for AdaBoost experiment........ 69

Table 13 - Confusion matrix for AdaBoost experiment 69

Table 14 - classification accuracy measures for each class in AdaBoost experiment........ 69

Table 15 - Classification accuracy for OneRule experiment 71

Table 16 - Confusion matrix for OneRule experiment 72

Table 17 - Classification accuracy measures for each class in OneRule experiment 72

Chapter 1 Introduction

The fundamental goal of a medical research project is to diagnose the patients more effectively, more efficiently and find better ways to treat patients. In this thesis, we focus on this goal from a database and data analysis perspective. The idea is to enhance the diagnosis based on historical data about patients, pretreatment measurements on those patients, applied treatments and outcomes. Designing such a system has two main steps which work in conjunction together:

1. Data gathering: This includes designing a database system that can contain complex patient records which may come from different systems. This step is crucial and requires careful data modeling as it gathers information from heterogeneous data sources. It should satisfy both integrity and accuracy of data.
2. Data analysis: This step has two challenges within itself.
    a. The first challenge is that like any other data analysis problem, it requires automatic and semi-automatic data analysis tools to include the domain knowledge of medical experts in the discovery process.

    b. The second challenge is how to deal with the complex database which contains un-structured data. Medical databases contain complex data types such as two and three dimensional medical images, signals, and free text as well as simple data types like numbers and categories.

The main focus of this thesis is on the second step, Data analysis. One may wonder why these two steps should work in conjunction together. At first glance, data analysis can be done independently as soon as the data gathering is done. Practically, the data gathering process never stops. It is also an expensive process especially when we gather information from multiple data sources. Therefore, it is essential to use the data analysis outputs and figure out what data is needed to generate new and accurate conclusions. We provide the following set of tools and methods to medical researchers:

- Data preparation for analysis
- Complex data visualization and exploration
- Hypothesis examination
- Pattern recognition and classification

Contributions

The following contributions distinguish this work from other approaches:

a) Developed a data pre-processing technique that transforms the HBIDS data into a format that is suitable for data mining.

b) Developed an interactive similarity measurement technique that allows the medical expert to find similar patients based on weighted attributes. This technique properly deals with missing data in the HBIDS system.

c) Developed a visual data exploration technique based on scatter-plots. This enables the medical expert to get a visual representation of the medical records in the system and the relationships between their attributes.

Medical Data Mining

Medical data mining (MDM) is the act of mining medical data. What distinguishes MDM from other types of data mining (DM)? It is the characteristics of the medical field itself and the types of data which are subjected to mining in this field. These characteristics need to be taken into account for a DM method to be effectively used against real data. The next section “Approaches in medical data mining” describes these characteristics in detail.

Approaches in Medical Data Mining

Among the very first things that one needs to be aware of when mining medical data, is its special characteristics and uniqueness. Human medical data are the most rewarding and difficult of all biological data to mine and analyze [1]. There is no animal that has its medical information recorded more than human. However, most of this data are recorded to benefit the diagnosis for the individual patient and not for a certain group study.

Different approaches that researchers take in mining medical data are due to various characteristics of medical data that make it unique. These characteristics can be categorized into the following two categories:

1. Heterogeneous and volumetric aspects.
2. Ethical and legal aspects.

Heterogeneous and volumetric

Medical data is usually collected from interviews with the patient, laboratory data, various images and physician’s interpretation and diagnosis. This constitutes massive and heterogeneous raw data that makes it very hard to analyze. Several approaches have been taken to mine this data.

Medical Image visualization and image processing approaches are used to analyze several types of images like MRI, SPECT, PET and EEG signals. Gigabytes of images are produced every day and it is extremely difficult for un-aided human to analyze such data. Automatic image segmentation methods are required to distinguish between different parts of the image. The segments can represent different tissues/organs or cells. Furthermore measurements of or extracted features from the segmented parts can be used for classification and clustering algorithms to find interesting patterns in data. Medical image processing is a well-established area of research.

Text mining approaches take on analyzing the medical text. A text based medical record may include laboratory reports, interview results and physician’s interpretation of the data. One of the main challenges here is that even the medical experts cannot agree on unambiguous terms to describe the patient condition [1]. One of the suggested approaches is to use computer translation to process physician’s interpretation[2].The underlying structures of medicine are poorly categorized mathematically compared to many of the other physical sciences[1]. There is very little formal constraint on the vocabulary and the relationship among the concepts. There is no inherent canonical form[1] available for medical terms. Data mining algorithms operate on highly structured data, so a unified medical language is a necessity. Unified Medical Language System (UMLS) is a language that contains 770,000 concepts and over 2 million synonyms in its 2002 edition [3]. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems; it may also be viewed as a comprehensive vocabluary and ontology of biomedical concepts. UMLS further provides facilities for natural language processing. It is intended to be used mainly by developers of systems in medical informatics [4]. XML based languages and formats are one of the best ways to represent semi-structured text data and are becoming more widespread due to its power to show relationships. TMA[5]is an example used in pathology. The major databases of today are equipped with XML processing and querying features that allow data mining methods to access and process information from XML documents. The XML format can be used during data preparation, storage, and retrieval. It can also be used to describe the domain knowledge. The XML format is already a de facto standard for data transportation and data storage [5].

Database oriented approaches use pattern recognition algorithms on massive amounts of data that are usually stored in a relational database. Classification, clustering and rule extraction are among the well-known approaches for data mining. These approaches work best when used on highly structured data. An ideal dataset that a classification algorithm can work on is a single table that consists of numerical or categorical attributes. The amount of missing attributes and noise in medical data is relatively high in contrast to other physical sciences. Missing a data attribute can be due to an exam that is not taken because it was not necessary,

[1] Canonical Form: In mathematics canonical form is a preferred notion that all other forms of the same concept can be reduced to using an algorithm [3]. For example 2/4, 3/6, 4/8 … can all be reduced to ½.

or the data was not entered in the patient’s medical record. Another issue with medical data is that complex data mining models may not be used because it is hard to explain them to experts. An example of such models is neural networks. A complex neural network might have high prediction accuracy but it does not provide straightforward rules, which is not understandable for a medical expert.

Furthermore, these highly structured datasets are not readily available in patient medical records, therefore, a medical specialist that understands patient medical records needs to carefully extract this valuable information so that they can be used in data mining algorithms. Text mining and XML based approaches to store medical records can make this process (semi) automatic and thus less costly.

Ethical and legal issues

When it comes to data about human subjects, enormous legal and ethical considerations need to be taken into account. The very first question is the **data ownership**. Does the patient or the medical institution own the data? This is still a disputed question [1]. Fear of lawsuits among healthcare providers and strict federal guidelines forcing the medical institutions to keep the patient data confidential. Therefore if a patient record is to be used in a study, there are federal guidelines that need to be followed. To protect the privacy of the patient when it is used in a research project, several approaches can be considered [1]:

- Acquiring data in an anonymous way from the very beginning of the data gathering process, so the patient identification is not recorded.
- Making data anonymous so the patient identification is removed after the data is acquired but before they are made available to a research project. In datasets with no patient identifier, it would not be clear whether two records are from the same patient or different patients and therefore the data mining model will be biased towards the patients with more records.

- De-Identifying the data by encrypting the patient identification. Such data can only be re-identified under certain conditions approved by the institution.
- Using the actual (identified) data under the condition that the patient has given written consent for data to be used in the research under federal law.

Some data mining approaches are proposed on how to make sure the patient information is not revealed as a result of data analysis or when the data are outsourced to a third party for analysis. Other approaches use incremental models that only use the data record once in the learning phase that creates the model and never stores the data[6].

Introducing the Human Brain Image Database System (HBIDS)

The Human Brain Image Database System is a database system developed in Henry Ford Health system in collaboration with Wayne State University.

Objective

The human Brain Image Database System is not only a system that stores and retrieves brain images. It is optimized for research on human brain diseases. It is imperative that brain images are one of the primary data used for such research. The data from other examinations such as EEG, WADA, physical examinations as well as data on medical status and family history of the patient come into play. The outcome of surgery and further observations before, during and after the surgery will also be used. Such heterogeneous and complex data is an asset but also a challenge for a human being to analyze. Therefore a software system is designed to gather and enter this data into the database. This system also provides tools to the medical experts to analyze the data examine hypotheses and reach conclusions.

State of the Existing System

HBIDS currently consists of a relational database equipped with a single patient view web interface that allows the medical experts to enter and see the data of an individual patient.

It consists of about 50 tables that are connected to each other. These tables together contain data that fit into the following categories:

- Patient's demographic and personal information (e.g., date of birth, sex)
- Medical history/risk factors (e.g., family history, meningitis, encephalitis)
- Disease description(e.g., seizure description) and given treatments(e.g., medication and surgery)
- image data and radiology
- EEG analysis
- outcomes (e.g. Engel classification, verbal learning test)
- Base tables (e.g., anti-epileptic drugs information)

Furthermore, several image analysis and segmentation algorithms have been developed and deployed on this system. The main goal of these algorithms is to extract imaging features from the brain; features such as volume, average intensity and texture of structures such as the hippocampus. The database empowers the researcher to correlate the imaging features with other attributes in the database to test a hypothesis and reach at a conclusion. Figure 1 shows how several components of this system are connected together.

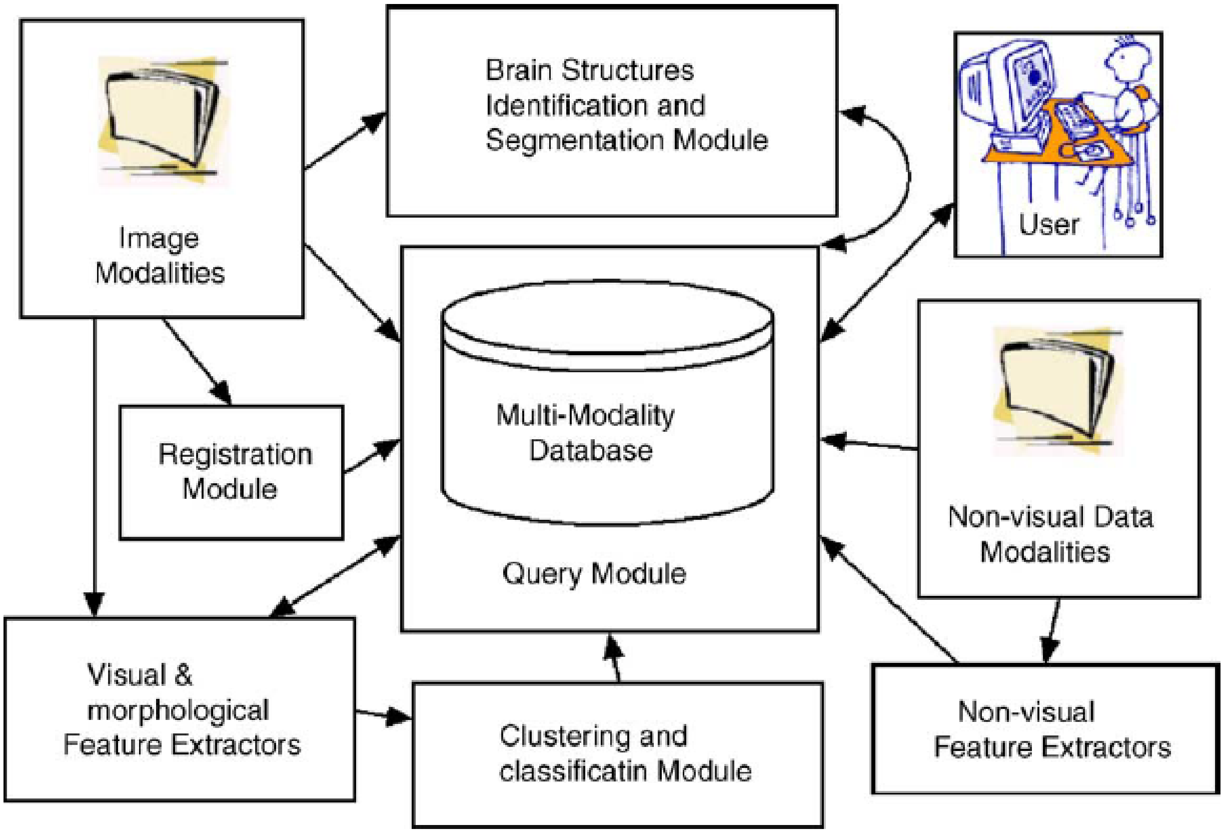


Figure 1 - Components of the HBIDS system[7]

Problem statement

The following are examples of questions that a medical expert may ask HBIDS:

a. What is the effectiveness of a surgery, e.g. temporal lobectomy, on temporal lobe epileptic patients? This can vary based on the patient’s condition. So the question is how the physician can predict the outcome based on pre-surgical information.

b. In which cases Phase II can be eliminated? EEG phase II includes implantation surgery to remove the scalp and place electrodes on the surface of the brain (subdural) and inside (depth).After implantation, patient will go through EEG monitoring for several days. Since this is a very invasive, risky and expensive process one would try to find ways to eliminate this phase and still reach similar or better performance.

c. How patient's quality of life would be affected? Brain performance such as IQ and Memory might change after the surgery and several parameters can participate in such events, e.g., surgery location and pre-surgical conditions.

One may ask "why is it difficult to answer the above questions? "As described in Introduction to HBIDS Section, the database is structured but yet it presents a complex system with lots of tables and attributes that do not necessarily and easily turn into a single flat table. Also, there are un-structured and loosely structured parts of data such as images and image features. Therefore, analyzing this data in itself is a challenge. Furthermore medical data is noisy and contains high levels of missing data in its nature. Thus, methods need to be developed to allow the medical experts analyze this data. In this study, we emphasize on the following steps:

1. Data preprocessing: It focuses on ways to automatically or semi-automatically structure un-structured data. This step is needed to make data and/or information ready for interactive data exploration and statistical analysis.
2. Data Analysis: This step includes designing and applying data mining techniques on medical data to extract interesting patterns. In a medical database it is hard to find similar patients especially when parameters of interest can be different based on the type of patient. This becomes even more difficult in a research project that the medical expert changes the set of parameters throughout the research.
3. Data Visualization: Human beings understand the data the best when presented in visualized formats such as diagrams and charts. Medical experts demand features such as ability to examine hypothesizes and visualize the results on the fly.

The objective of this study is to address the above for an existing human brain image database system that contain information about epileptic patients.

Summary

A database system for medical research project has two main parts: a) data gathering and b) data analysis. Due to the characteristics of the medical domain, each of these steps has its complexities and challenges. The focus of this thesis is on the data analysis part in which the special characteristics of medical data present some challenges. Medical data is often a) heterogeneous and volumetric & b) It has ethical and legal issues that need to be dealt with. Image processing, text mining and database oriented techniques are the main approaches which researchers take that are related to the heterogonous and volumetric characteristic of the data. This thesis describes database oriented techniques that are used in data mining for Human Brain Image Database System (HBIDS).

HBIDS stores the results of image processing techniques on patient's medical images along with other information on patient's medical record. This information gets stored a database which consists of 50 connected tables. These many tables each have 10-20 attributes in each table and have different types of relationships. This database model represents the complex nature and the patient's medical record. For a medical expert, finding information in this database becomes practically impossible. The developed techniques described in this thesis, make this data consumable for medical expert. They focus on data preprocessing, similarity measurement and data visualization.

Chapter 2 Background and Related Work

The Epilepsy Disease

Epilepsy is a brain disease that human race have been struggling with it during the history. It has been referred as “the sacred disease” in centuries regarding the mysterious behavior of the patient during seizure [8].

Definition: “When nerve cells in the brain fire electrical impulses at a rate of up to four times higher than normal, this causes a sort of electrical storm in the brain, known as a seizure. A pattern of repeated seizures is referred to as epilepsy [9].”

Importance: The percentage of people who experience this type of brain disorder is relatively high. More than 3 million people in the United States, about 1 in 100, have experienced an unprovoked seizure or been diagnosed with epilepsy[10]. This shows the importance of research on this type of brain disease.

Epilepsy is not caused by mental illness and it does not necessarily lead to mental illness or mental retardation. In contrast, many people with Epilepsy have normal or above-average intelligence [10]. There are some famous Nobel Prize winners, Olympic medalists and Philosophers who have had Epilepsy and were still performing well in their field of expertise.

There are two kinds of seizures:

1. Focal seizure (partial seizure), occur in only one part of the brain. About 60 percent of people with Epilepsy have focal seizure [10]. There are two types of focal seizures:
    a. *Simple focal seizure*: During the time of seizure the person is conscious but experiences unusual feelings and sensations. The feelings are sometimes unexplainable. They contain joy, anger, sadness or weird tastes and smells.
    b. *Complex focal seizure*: The patients lose their consciousness. They may show strange and repetitious behaviors. Blinks, muscle spasms, mouth movements, or

even walking in a circle when occur in a repetitive manner are signs of complex focal seizure [10].

Symptoms of focal seizures can be easily confused with other disorders and it may take many tests by the physician to identify it.

2. Generalized seizure is a result of abnormal neuronal activity in both sides of the brain. These seizures may cause loss of consciousness, falls, or massive muscle spasms [10].

It is not always easy to define the type of the seizure. Sometimes it starts as a focal seizure then spreads throughout the brain.

Diagnosis methods

There are number of tests to determine what type of epilepsy a patient has and where the origin of the epileptic activities in the brain is. Several methods might be used for a patient in the diagnosis process by the medical expert, some of which are listed below.

i. EEG Test

Electroencephalogram (EEG) test is designed to detect problems in the electrical activity of the brain [11].Electrodes are placed at several specific spots on the scalp to record brain electromagnetic signals. Electrodes can show, which part of the brain is engaged in abnormal electrical activity. Video monitoring is often used in conjunction with EEG to rule out other disorders that could be confused with epilepsy.

The output of this exam includes the location of each electrode with its recorded signal during the time of monitoring.

ii. Brain Scans

Brain scans are one of the most common ways of noninvasive or minimally invasive diagnosis in brain diseases. MRI (magnetic resonance imaging), CT (computed tomography), SPECT (single-photon emission computed tomography) and PET (positron emission tomography) are the most commonly used imaging technologies for Epilepsy. CT and MRI scans show the anatomy of the brain. They can reveal tumors, cysts and abnormalities in brain structures. PET, SPECT and functional MRI (fMRI) are used to show the activity of different parts of the brain. SPECT is a relatively new kind of imaging modality that can be used to locate seizure foci in the brain [10].

iii. MEG

Magnetoencephalogram is another type of brain scan. It detects magnetic signals generated by neurons over time. It can reveal signals that come from deeper parts of the brain. That is in contrast to EEG that detects signals only from the surface of the brain.

iv. Medical History

Taking the detailed medical history of the patient including symptoms and duration of the seizures is still one of the best methods available to determine the type of seizure [12].

v. Neurological and Behavioral Tests

Physicians often perform several behavioral and neurological tests to determine intelligence capacity of the patient. Such tests can also reveal the extent to which this disease might affect the person's life. The following is a list of some of the more important tests that measure the intelligence capacity of the brain from different aspects:

- IQ
- Boston Naming Test (BNT)
- Immediate and delayed verbal memory
- California Verbal Language Test (CVLT)

Treatment

The most common approach in treating epilepsy is to prescribe antiepileptic drugs. Surgery is the second choice of treatment that will be considered when medication is not effective. A counsel of doctors will consider different examinations to determine, which part of the brain is involved in the abnormal activities and what would be the consequences if the epileptogenic area(s) are resected. Surgeons usually avoid operating areas of the brain that are crucial in important activities such as hearing, speech or language. Tests such as Wada are to find areas in the brain that are responsible for language and memory. For cases with difficulties in localizing epileptogenic foci, the patient may undergo EEG phase II**.** A study published in 2000 has compared surgery treatment with an additional year of treatment with antiepileptic drugs in people with temporal lobe epilepsy. The results show that 64 percent of people have become seizure free after the surgery while only 8 percent of those who continued taking the drugs have become seizure free. Therefore, American Academy of Neurology (AAN) recommends surgery for temporal lobe epilepsy if the drugs are not effective [10].

The most common type of surgery is removal of seizure focus in the brain. Seizure focus is the portion of the brain where the seizure originates from. This surgical procedure is called lobectomy. The most common type of lobectomy is temporal lobe resection. For these patients, the surgery will result in significant reduction or complete elimination of seizures in 70-90 percent of the cases.

Related work

In this work, the medical data is first filtered and pre-processed into the appropriate format for data mining algorithms. Next we apply a number of data mining algorithms to extract relevant patient data. Finally, we provide a visual representation of the data, to help physicians explore the output of the data mining step. In this section, the related works for these three steps are reviewed:

1. Data Preprocessing
2. Data Mining
3. Visualization

Data Preprocessing

Data pre-processing is an often neglected but important step in the data mining process. The phrase "Garbage In, Garbage Out" is particularly applicable to data mining and machine learning projects. Data gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis[12]. Data preprocessing methods can be categorized into the following categories [13].

i. Descriptive data summarization
ii. Data cleaning
iii. Data Integration and Transformation
iv. Data reduction

i. Descriptive data summarization

Data summarization gives an overall picture of data. The techniques that are in this category try to identify typical characteristics of data and find out what parts of data are noise and outliers. Statistical measures such as mean and variance are for numerical data. Graphics based techniques such as histograms and scatterplots also give the data analyst an overall picture about the data. Having this big picture will help the data analyst to rule out outliers and suppress noise to increase the efficiency and accuracy in data mining and analysis.

ii. Data cleaning

Real world data often contains missing values, noise and inconsistency. Data cleaning methods are designed to estimate missing values, reduce noise and find outliers. Since in this project missing values is a major problem, we only focus on introducing methods for this purpose among other data cleaning methods.

Missing values: Although we should do our best to have a complete and accurate dataset but the real world data tends to be incomplete. The effect of missing values in a data set increases as the data dimensionality increases. In some data sets such as HBIDS data, there are very few data tuples - if any – that have no missing values and this makes a conventional data mining method useless. The proper way of treating a missing value depends on several factors:

1. Type of the attribute: A missing integer value can be treated by filling in the average of the attribute across the board while this cannot be done for a categorical type.
2. The data mining task that is intended to be used on the data set: If the goal is to apply a classification algorithm on data it would not make sense to fill out a missing class label.
3. Statistical distribution of the value among different tuples: Filling a missing value in a normal distribution with a small variance would be very different that the case of a uniform distribution. In the case of a normal distribution, the amount of information loss by using a representative average value for missing values is minimal compared to a uniform distribution.
4. Size of the data set and number of attributes: Ignoring a tuple in a big data set would have a low cost in contrast with a small data set.

Therefore there is not a single approach that works the best everywhere. Here is a list of ways to treat a missing value in the data set [13]:

a) Ignoring the tuple: This is the easiest way and can work very well for tuples that contain several attributes with missing value. This approach is not possible when most of the tuples have missing values.

b) Using a global constant for the missing value: The mining algorithm may consider such value as an interesting common pattern and result in incorrect output.

c) Using the attribute mean value to fill the missing values: This can be very useful especially when the attribute value has a normal distribution. Replacing the missing value of samples that belong to the same class with the mean value of non-missing values that belong to that class is used.

d) Using the most probable value: the can be determined using regression, decision tree induction, Bayesian classifier or inference-based tools.

The last three methods (b, c, d) can bias the data set and result in incorrect data mining results. The last method (d) is very popular [13], because there is a greater chance of preserving the relationship between attributes.

iii. Data Integration and Transformation:

Data integration is merging data from multiple data stores. Data may also need to be transformed into forms that are appropriate for mining.

- **Data Integration**: Analysis of heterogonous medical data often includes data integration task. The imaging data, image analysis data, medical information and patient history come from different sources and need to be integrated. For example a column named “surgery_date” in one data source might be named “date_of_surgery” in another.

  Redundancy is an important issue here. Some attributes might be derived from others. Such redundant attributes can cause inaccurate results or confusion in mining. **Correlation analysis** can be used to find some redundant attributes. This analysis can measure how strongly one attribute imply the other.

  Correlation coefficient is widely used in the literature for this measurement.

  Pearson Correlation Coefficient is one of the popular ones that measure the linear dependency of two numerical attributes:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

$r_{A,B}$, can vary from -1 to +1. If the attributes are highly correlated, $r_{A,B}$, becomes closer to +1. If they are not correlated the value closes to 0 and if they are negatively correlated the value will be negative.

By using correlation coefficients one can find candidate attributes to remove redundant attributes, in order to minimize the bias of data mining methods towards correlated attributes. Strong correlations can also be represented to the domain expert in order to assist forming hypothesizes about the data. This depends, highly on the domain knowledge and information about attributes.

- **Data Transformation:** Data transformation methods are used to transform data into forms appropriate for mining. The transformed data leads to better results when data mining methods are applied. The following is a list of different types of data transformation methods. One may use number of these methods all together.
  - **Smoothing,** removes noise from the data.
  - **Aggregation** is often used to construct data cubes at multiple granularities.
  - **Generalization,** replaces raw data with higher level concepts. For example it may replace the date of birth with the year of birth, or replaces the city with state. Other examples can be replacing verbal memory number with categories such as low, medium and high. This can be highly important when the number of categories comparing to the number of records is too high and therefore no meaningful pattern can be found based on those categories. See example in Table 1.

Table 1 - Data Transformation Example (Generalization)

#	City	Weather		#	State	Weather
1	Seattle	Rainy		1	Washington	Rainy
2	Redmond	Rainy		2	Washington	Rainy
3	Orlando	Sunny	=>	3	Florida	Sunny
4	Miami	Sunny		4	Florida	Sunny
5	Kirkland	Rainy		5	Washington	Rainy
6	Naples	Sunny		6	Florida	Sunny

In the left table no pattern can be found, but when you replace the cities by their higher categories (State), a new pattern can be found that, if the state in the record is Florida then the weather is sunny. Such pattern could not be discovered in the left table.

- **Normalization** specially is used on numerical data to scale the data in a common interval such as between 0 and 1. Without this process different records can be in different scales and will not be comparable with each other.

  Min-max normalization is a widely used linear transformation method that uses the minimum and maximum values of an attribute and maps them to a certain interval. In some classification methods such as neural networks this transformation is necessary to get proper results. The following is the min-max formula that maps any numerical value (v) to a new value (v'):

$$v' = \frac{v - min_A}{max_A - min_A}$$

- **Attribute construction** creates new attributes based on existing ones to help with the data mining process. An example of such attribute is the body mass index (BMI) that is calculated based on height and weight attributes in a data set.

iv. Data Reduction

Data reduction can be applied to a data set to obtain a smaller representation of the data that is smaller in volume but still maintains the integrity of the original data. Different methods of data reduction includes: Data cube aggregation, attribute subset selection, dimensionality reduction, numerosity reduction and discretization.

Data Mining

Data mining is the process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst [14]. Data mining problems are often solved using a mosaic of different approaches drawn from computer science, including multi-dimensional databases, machine learning, soft computing and data visualization, and from statistics, including hypothesis testing, clustering, classification and regression techniques[15].It is the appropriate choice and combination of these techniques which leads to solving a given problem effectively [15].

Statistics is an area very closely related to Data Mining. In the statistical approaches, the research hypothesis is often defined before the data is gathered. However data mining deals with secondary analysis. There, the data is not purposely collected to test some research hypothesis but is obtained from legacy databases or data warehouses where the volume of data may be much greater [15].

i. *Predictive Data Mining*

The data mining methods that construct models in order to predict future cases are in the category of predictive data mining. In general, the task of predictive data mining is to find the best fitting model that relates attributes to the outcome [15]. The goal of predictive data mining in clinical medicine is to derive models that can use patient specific information to predict the outcome of interest and to thereby support clinical decision-making [15]. This should help physicians improve their prognosis, diagnosis or treatment planning procedures.

When it comes to medical data, the nature of data is different than standard data mining data sets. Medical data sets may be smaller: typically, the number of instances is from several tens to several thousands. The number of attributes may widely range from several tens (classical problems from clinical medicine) to thousands (proteomics, genomics)[15].

ii. *Background Knowledge (Domain Knowledge)*

Literally, background knowledge is the 'information that is essential to understanding a situation or a problem' [16]. In the process of building a predictive model, using background knowledge means being capable of taking into account information which is already known and should not be rediscovered from the data. This issue may be particularly important in the analysis of medical data [17].

iii. *Predictive Data Mining Methods*

Data mining tasks can be classified into description and prediction. A description task tries to find human-interpretable patterns and associations. Examples of methods used in description tasks are data visualization and interactive data exploration. A prediction task seeks to foretell a response of interest by constructing a prediction model after considering the data. Therefore prediction tasks require data to have special response variable which is also called outcome variable. Examples of methods used in prediction tasks are neural networks and decision trees.

Methods used in prediction tasks can be generally categorized into classification and regression. Classification methods are used for cases where the response variable (outcome) is categorical while regression is used for numerical response variables. While the methods used in classification and regression is different, the data mining process is very similar. Among these two categories, only classification methods are used in this thesis. Here we list number of most popular predictive data mining methods according to recent rankings from KDNuggets[19]. These are categorized based on the resulting predictive model:

1. **Decision trees:** These methods recursively partition the records to construct a decision tree. Due to the heuristic nature of these methods, their complexity is low and they create sub-optimal models. Techniques such as pruning are used to trim the tree from leafs containing fewer instances. The reason why this is ranked first amongst other methods is largely due to the resulting model (decision tree) which is favorable and easy to understand by human. C4.5, See5 and CART are amongst these methods which exist in most data mining packages [18][19].
2. **Decision rules:** A decision rule is in the form of ‘IF condition-based-on-attribute-values THEN outcome-value’. These methods construct a set of rules either using decision trees as in 4.5rules [18] or directly from the data as in AQ and CN2 [20][21]. The complexity of these methods is generally higher than decision trees while sharing a similar performance. Decision rules can become abundant and pretty complex making it less readable for a human expert. On the other hand, it has a simpler representation which makes it favorable in some cases when combined with other techniques.
3. **Logistic regression:** It is a well-established method from statistics [22]. The model identifies the contribution of each variable with respect to a specific outcome. Therefore, when attribute values are defined, the probability of an outcome can be

estimated using the model. Missing value handling is not straight forward in this method and the model can be represented using a nomogram[2].

4. **Artificial neural networks:** They are amongst the methods with high predictive performance and therefore very popular in general. However they have their deficiencies [23]. They are very sensitive to parameters which define the architecture of the network. Also the resulting model is quite complex and cannot be easily understood by human medical expert. These deficiencies have made them less popular in the medical domain where understanding the predictive model by medical expert is crucial.
5. **Support vector machines (SVM):** They are probably the most powerful classification algorithm with regard to their predictive accuracy [24]. For a two-class data set, the resulting model of SVM is a hyper plane that separates the samples with different outcomes. The method tries to find the optimum hyper plane that has the maximum distance from the closest point of the two classes [25]. Since SVM is a linear classifier it is very robust. SVMs are based on strong mathematical and statistical foundation [26]. When predictive performance and robustness is the only factor for choosing a predictive data mining method, SVM is a strong candidate. However it is often not suitable for interpretation by human experts.
6. **Naïve Bayesian classifier:** It is considered as a baseline when evaluating the performance of a classifier. It has a well-established statistical foundation and creates a simple model. Its performance often is comparable to more sophisticated methods. In cases where other methods do better, it is usually due to non-linear relationship of the attributes in data. The resulting model is very suitable for interpretation by human experts.

[2] A graphic representation that consists of several lines marked off to scale and arrange in such a way that by using a straightedge to connect known values on two lines an unknown value can be read at the point of intersection with another line.

7. **Bayesian networks:** They are probabilistic graphical models which can express the joint probability distribution of variables. One of the most well-known ways to use them in medical applications is to consider the variables (attributes) as nodes. The nodes are connected by arcs. Each arc shows the conditional probability between the source node (variable) and the designation node. There are algorithms that can learn the Bayesian network based on given data. The output is very suitable and useful for human expert interpretation when the relationship between attributes is of interest. It's also used to solve classification problems. Classification labels are represented as nodes in the graph. Then based on well-known attribute values, the probability of different outcomes can be calculated.
8. **K-nearest neighbors:** This is very similar to what medical experts do when they take decisions based on similar cases that they have seen before [27]. For a given set of attributes of a new case, it looks at all the cases and finds the k most similar ones to the new case (according to their attributes). Then it looks at the outcome of those k most similar cases and picks the outcome class that is most popular among them as the predicted outcome.

iv. Ensemble Methods

In many cases a single method may not be able to provide the expected predictive performance. Ensemble methods introduce a way to combine multiple predictive data mining methods to get a better accuracy. In case of classifiers, each classifier can predict certain patterns better than others. Also using multiple classifiers often can result in more robustness as two classifiers can cancel the mistake done by one classifier doing a wrong prediction. Meta-Classifiers are the type of classifiers that can combine other classifiers to create a stronger classifier.

One of the most popular categories of Meta-Classifiers is Boosting. AdaBoost, BentleBoost, BrownBoost, RankBoost are among them.

v. Evaluating a classifier

There are several methods in the literature that experts use to evaluate a classifier given a data set. Below is the list of some of the widely used evaluation methods which are used in WEKA [30] and here in order to evaluate a classifier in Chapter 4:

Confusion matrix

In this matrix, each row shows the actual number of instances with the classification label mentioned in the left side of the row. Each column shows the number of instances that the classifier has predicted to have the classification label mentioned on top of the column. The importance of this evaluation method is that, it shows how the classifier has performed for each individual class. Given a classification label, one can see among the instances with that label, how many were predicted correctly. Also for the ones that were predicted incorrectly, which labels they have been predicted to have. You can see an example of a confusion matrix for a binary classifier in Table 5 - Confusion matrix for the J48 decision tree

. Here the expert can see if the classifier is biased towards specific classification labels.

Classification accuracy measures for each class

- True positive (TP) rate: The percentage of instances that are correctly classified in this class.
- False positive (FP) rate: The percentage of instances that are incorrectly classified in this class.
- False negative (FN) rate: The percentage of instances that are incorrectly classified to other classes (which should have been classified in this class).

- Precision $= \frac{Number\ of\ true\ positives}{Number\ of\ true\ positives\ + Number\ of\ false\ positives}$. Precision for a certain class label shows the accuracy of a class label predicted by the classifier.
- Recall $= \frac{Number\ of\ true\ positives}{Number\ of\ true\ positives\ + Number\ of\ false\ negatives}$. Recall for a certain class label shows the percentage of times when the classifier correctly predicts a classification label for a certain class.
- $F - Measure = 2 \times \frac{Precision\ \times Recall}{Precision + Recall}$. F-Measure is the harmonic mean of precision and recall. This is a single measurement that combines precision and recall.

Visualization

Since us as humans can understand visual representations much better than numerical representations, visual data exploration techniques are essential for a data analyst. These techniques for multivariate and multidimensional data can be categorized into five classes [31]: geometric projection, icon-based, pixel-based, hierarchical, graph based.

i. Geometric projection

Geometric data projection techniques work based on projecting values to geometrical positions on the screen. The techniques listed below are examples of this technique.

a) Scatter plot matrices is one of the widely used ones. A scatter plot matrix contains one scatter plot for each pair wise combination of two attributes in the data set. Scatter plot is a very well-known and therefore one of the easiest to understand forms of data visualization. However they only can show the 2-dimensional relationships in a data set and therefore don’t provide the ability to find patterns that include multiple dimensions.

b) Parallel coordinates project the data into several equally spaced parallel axes in a two dimensional space. Each axis is associated with an attribute in the data set. A data point in this projection is identified by connecting the points on all the axes which the attribute value of that data point is associated with. This technique can be very effective in revealing multi-dimensional patterns. However the downside of using it is that as the number of data point increase the lines overlap each other and it becomes extremely hard to see the pattern. Figure 2 visualizes Fisher's Iris data using parallel coordinates.

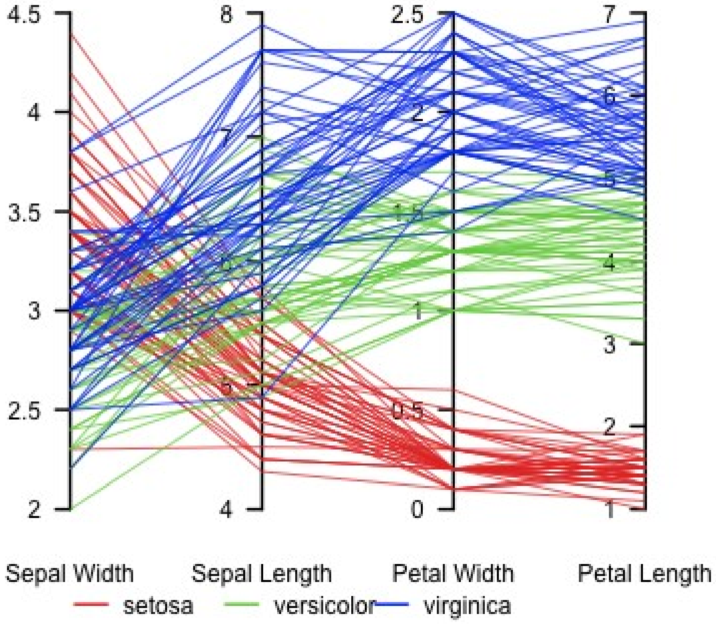


Figure 2 - Parallel Coordinate Plot Fisher's Iris Data [31]

c) Radial coordinate visualization (RADVIS). Suppose there is an object in a circle. Several springs attach the object to the surface of the circle in fixed locations. A data point can be represented with this where the attribute values of the data point are the spring constants (the tightness of the spring). If the object is released with only the springs attached to it, it finds its balance finally and that would be the position of the data point in a circle. This method can be used for a multi-dimensional visualization. One drawback of RADVIS is that it is possible to get the same data projection with different attribute values [29]. Therefore, one should be very careful with interpreting this result. Figure 3 is visualizing a 12 dimensional DNA data using this method.

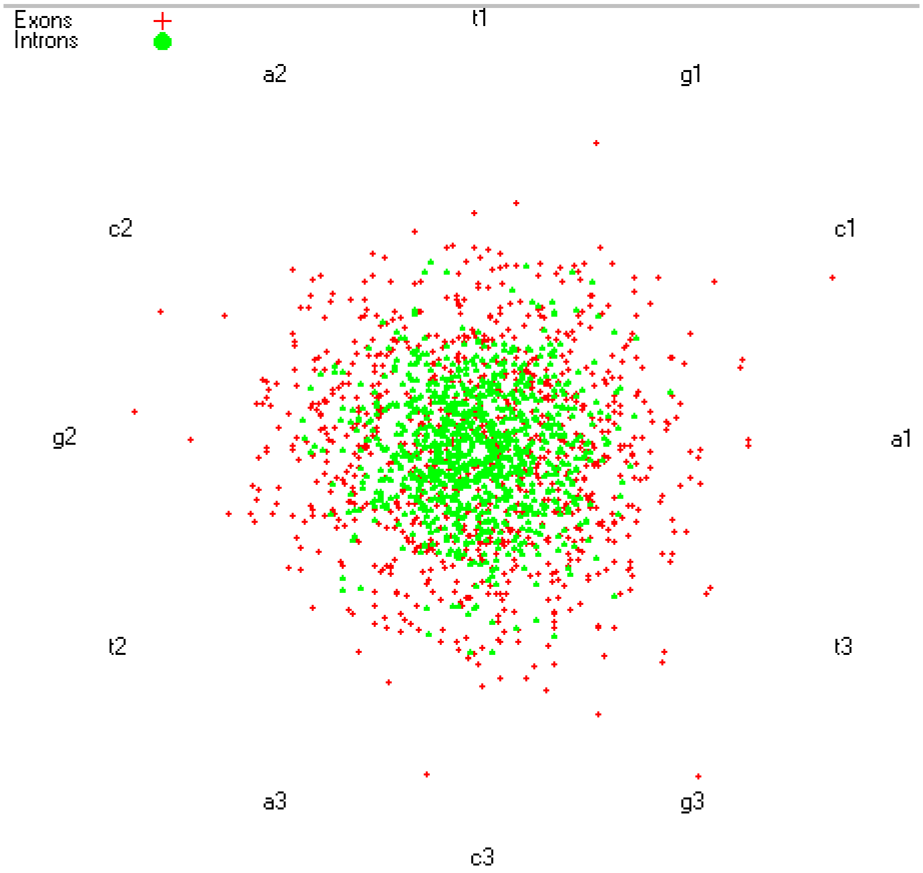


Figure 3 - RADVIS visualization of DNA data[33]

ii. *Icon-based*

Icon based techniques map each multidimensional data point into an icon. The visual features of the icon vary for each feature but all of the features together create an icon. One of the well-known icons used for this type of visualization is face as in Figure 4. Different features of the face represent different values for a specific attribute. The drawback of using face as the icon is the visual recognition system of human pays more attention to certain features such as eye than the others and can hardly compare values of two features together. Another example is a “stick figure” with several arms with the direction of an arm being the value of an attribute.

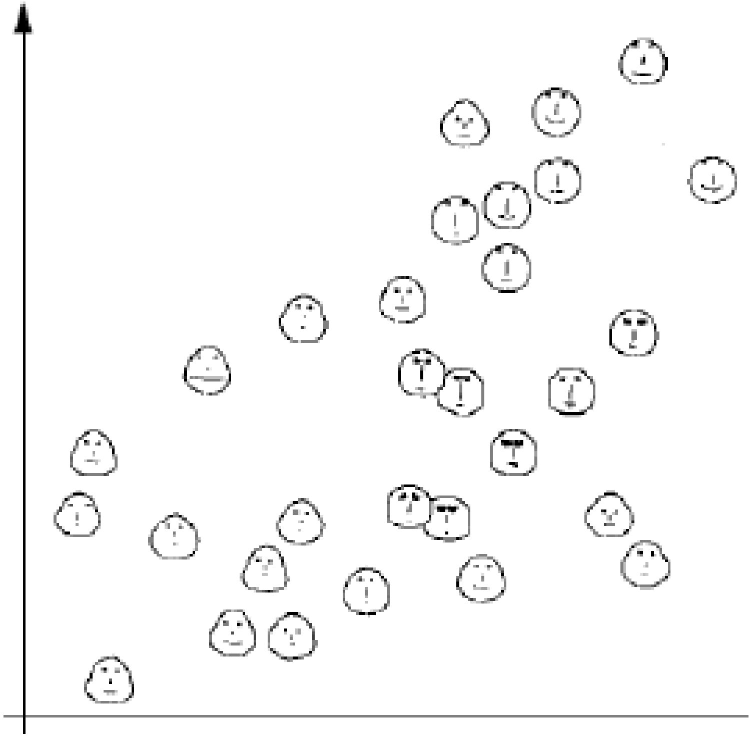


Figure 4 - Chernoff faces in various 2D positions [34]

iii. Pixel-based

Pixel-based techniques show the data in several sub windows, each sub window being a different dimension. Each pixel in a sub window represents the attribute value of a specific data point with its color. Positioning of data points in the window vary in different methods. One of the methods is spiral positioning of data points to a selected data point. This method allows the most number of data points to be seen on the screen among all the visualization methods. Figure 5 shows an example of this visualization technique for 8 attributes. The top left window shows the overall distance of the data points to a query positioned in the center. In each of the other 8 windows, the spiral distance is based on the value of the data points to the query regarding one attribute. However the color of each data point is the same as the one assigned to it in the top left window.

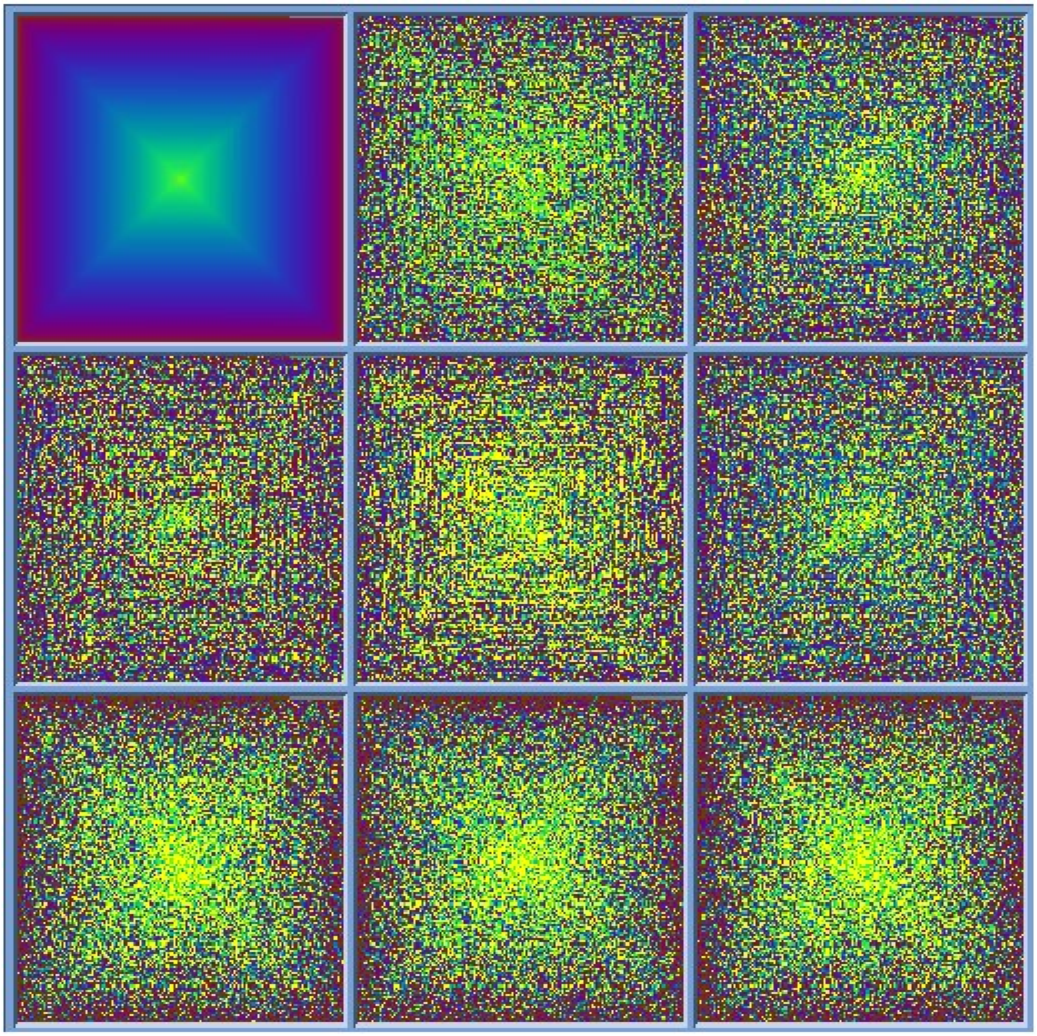


Figure 5 - Example of pixel-oriented visualization using spiral technique [36]

iv. Hierarchical

Hierarchical techniques select two to three dimensions in the beginning and the data about other dimensions can be shown when the value of the initial dimension is specified. N-vision is an example of this technique where it shows the data in a 3D scatterplot. Then by selecting a point or a slice by the user, it shows the other dimension. This can specially be useful in cases where the data has hierarchical structures. Figure 6 is an example of hierarchical visualization for 5 dimensions. The height of each point in the surface (the inner world) is defined by its distance to a constant point in x3-x4-x5 (the outer world) using it's x3, x4, x5 attributes. Therefore by moving the x1-x2 coordinate relative to x3-x4-x5 the surface changes according to the distance of each data point to the previously defined constant point in x1-x2-x3 coordinate.

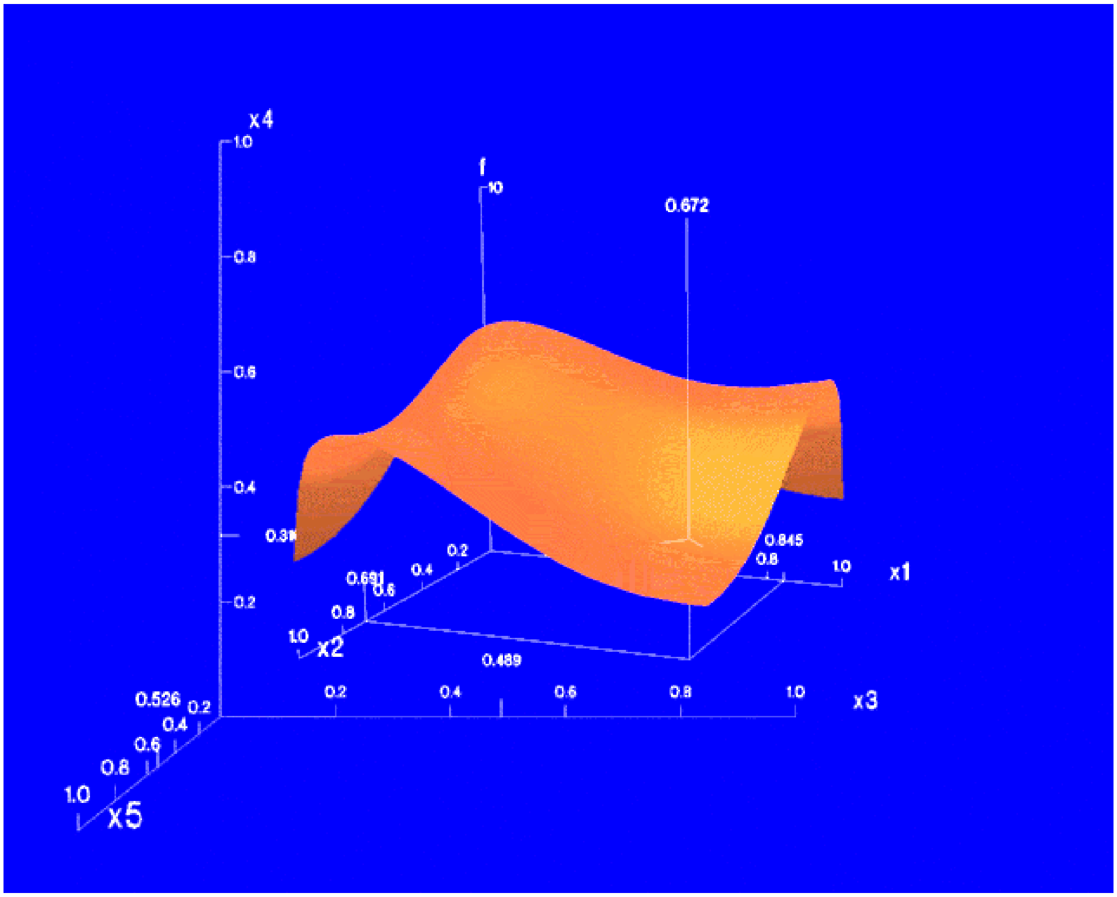


Figure 6 - Example of N-vision hierarchical visualization for 5 dimentions [37]

v. *Graph-based*

Graph-based techniques can be used in relational databases to reveal the object dependencies and relationships in the database. An Entity Relationship Diagram (ERD) shows the relationships between objects in an abstract way. Graph-based visualization shows the relationships between existing entities in the database. These graphs can become very complex in a medium sized database due to the amount of arcs that cross over each other. Graph exploration techniques are designed to assist the users better see the relationships in 2D/3D layouts(Graph based visualization). Furthermore, query visualization techniques allow the expert to design queries in a visual and interactive manner. Figure 7 is a graph-based visualization example microarray results in the context of hundreds of thousands of concepts integrated from several heterogeneous databases.

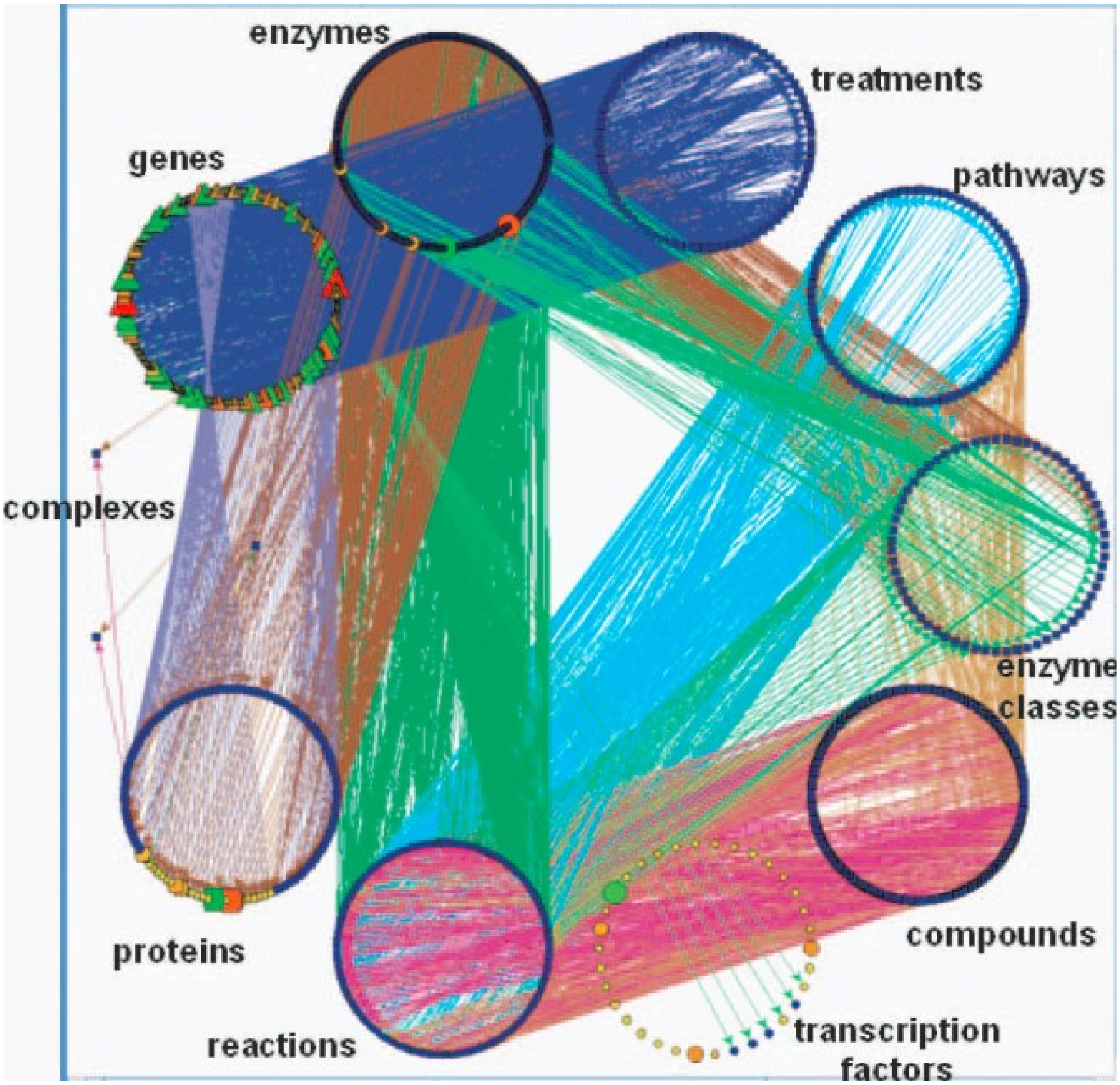


Figure 7 - A Graph-based visualization for microarray results [38]

Summary

Epilepsy is a brain disease that often shows itself when patient frequently has seizure. The cause of the seizure is a malfunctioning tissue[s] of the brain which generates abnormal electrical signals. One of the most common treatments for epileptic patients is to respect the abnormal tissue. Therefore the experts use techniques such as MRI, SPEC imaging and EEG tests to find such tissues. The patient’s medical history and laboratory results from other tests are also used in order to make decisions on whether to operate on the patient’s brain and also how to conduct the surgery to get the best results with minimal damage to the patient’s brain. The HBIDS system stores all the data which the medical expert uses in the diagnosis and treatment process.

One of the goals of the HBIDS system is to use the gathered data for analysis. However, data analysis methods do not produce meaningful results if they use the raw medical data. In several cases, the method is not able to consume the raw data. Therefore data preprocessing is needed to prepare the data for analysis. The type of techniques used for data preprocessing are: I) Descriptive data summarization, II) Data cleaning, III) Data integration and transformation and IV) Data reduction. They use different algorithms in order to: a) Remove noise, b) Reduce the amount of missing values, c) Remove redundant parts of the data, d) Transform the data to make it more suitable for analysis by normalization, aggregation, etc. After data is prepared, data mining and data visualization techniques can be very effective analyzing the data.

Predictive data mining is referred to a category of data mining methods which construct models in order to predict future cases. Incorporating the domain knowledge in the data mining process is extremely valuable in medical data. Therefore data mining methods which produce models that can be easily understood are highly preferable by the domain expert. Decision trees, Decision rules, Bayesian classifier, Baysian networks and K-nearest neighbor are among the methods which produce human readable models. Artificial neural networks and Support vector machines are prediction methods which produce models with high prediction accuracy. However their models are often not readable for medical experts. In order to compare the predictive power of different methods they need to be evaluated. Confusion matrix and Classification accuracy measures are commonly used for this purpose.

One of the most powerful ways to extract knowledge from data is to present it to the domain expert using visualization techniques. Geometric projection, icon-based, pixel based, hierarchical and graph based visualization techniques have been introduced in the literature. Each of these techniques has pros and cons. Therefore depending on the a) amount of data b) number of dimensions and c) the type of pattern which the domain expert is looking for, each of these visualization techniques can be used.

This chapter explained what epilepsy is, and what methods are in the literature for data mining and data visualization. By looking at the next chapter you will see what are the newly developed methods for data analysis in HBIDS and how do they correlate to existing methods in the literature which this chapter described.

Chapter 3 Contributions

This chapter describes the work that is done in pre-processing, similarity measurement and visualization categories

Data pre-processing

The HBIDS consists of a relational database. It stores each type of data differently in order to have normalized tables. Conventional data-mining algorithms can only deal with single tables. Therefore, upon selection of attributes, data needs to be transformed from this database to a single table format to be used by data-mining applications such as Weka. Furthermore, the filtering needs to be performed to address issues such as errors in data entry.

Transformation

The data model in HBIDS contains different types of data. Not all of this data is appropriate for medical data mining. Therefore, we need to differentiate between different types of data in the database to be able to query them systematically. We dichotomize data about data (metadata) in different way. One possible dichotomy is: *content-dependent metadata* and *content-independent metadata*:

- *Content-independent metadata*: The metadata that is independent of the content of the unstructured data. For instance, for an MRI image, the pixel size is not derived from the content of the image (existence of a tumor, size of the brain structures, etc.). These types of metadata have been represented in a conventional database modeling paradigm as attributes: they are actual columns in tables. A_i is the notation used for these attributes.
- *Content-dependent metadata* [31]: The metadata that is dependent on the content of the unstructured data. They are usually computed using an unstructured data. An example is the volume of the hippocampus computed using a T1-weighted MR image. Since there are infinite ways to describe an unstructured data (e.g. image) the number of such

metadata can grow indefinitely. Therefore, we cannot deal with these types of metadata in a conventional way (conventional data modeling). A general-purpose table has been proposed in this paper to keep track of this kind of metadata. $A_i^*$is the notation used for these attributes.

Metadata: Data about information-bearing entities to help in identification, discovery, assessment, and management of the described entities [32].

Unstructured data: We consider any data that cannot be *directly* used in a basic SQL statement as unstructured, e.g., audio and video clips, body of an email, human brain images, and segmented models of brain structures.

Using this concept of content-dependent and content-independent metadata, we can categorize the type of attributes in the database as in Figure 8.

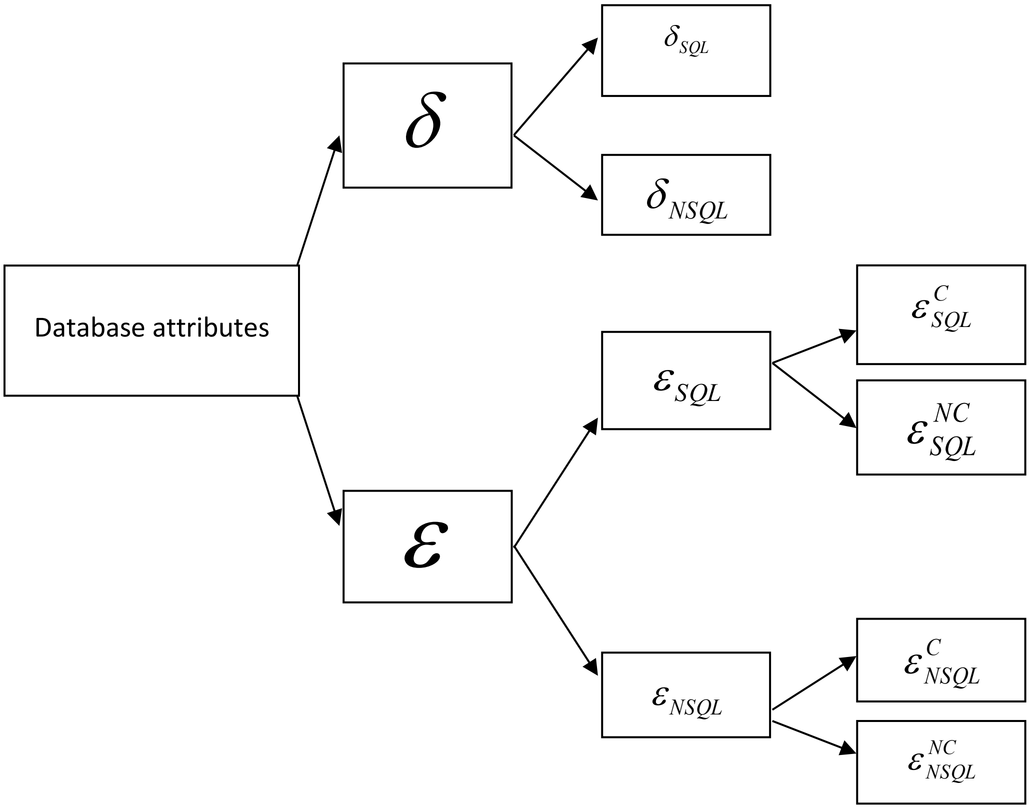


Figure 8 - Dichotomies of database attributes

δ : Set of all epilepsy-related data.

ε : Set of all metadata that describes δ .

δ_{SQL}: Set of structured data; examples are: gender, race, verbal memory and surgery side.

δ_{NSQL}: Set of non-structured data; examples are: MRI and SPECT images.

ε^{C} : Content-dependent metadata; is a set of metadata that describes the contents of the raw data.

ε^{NC} : Content-independent metadata; treats the raw data as black boxes and describes them regardless of their contents.

ε_{SQL}: Set of structured metadata;

ε_{NSQL}: Set of unstructured metadata

ε^{C}_{SQL}: Set of structured content-dependent metadata; examples are: hippocampus volume, average intensity and hippocampus texture feature.

ε^{NC}_{SQL}: Set of structured content-independent metadata; examples are: pixel size, date of imaging, image thickness.

ε^{C}_{NSQL}: Set of unstructured content-dependent metadata; examples are: hippocampus model and registration information

ε_{NSQL}^{NC}: Set of unstructured content-independent metadata; we don't have an example of this type of metadata in the system.

Definitions

D₁. $A_i \in \varepsilon_{SQL}$ is an attribute of a conventional relational data base (RDB).

D₂. $a_{\mathrm{i,j}}^{\mathrm{p}}$ is the jth value of A_i for patient p.

D₃. A_i^* is an all-purpose attribute (APA) if its jth value for patient $p(a_{i,j}^{*p})$ is the value of feature A_k where $A_k \in \varepsilon_{SQL}^{C}$

D₄. $a_{\mathrm{k,j}}^{*\mathrm{p}}$ is the jth value for patient p where $\mathrm{A}_{\mathrm{i}}^{*} = 'A_k'$ and $A_k \in \varepsilon_{SQL}^{C}$

D₅. RDB is the part of the database that contains A_i attributes.

D₆. RDB* is the part of the database that contain A_i^* attributes.

D₇. RDB$^+$ is the part of the database that contain A_k^* attributes.

For example in HBIDS, there is a single table which constructs RDB*.One of the columns is the attribute name and it's part of the key in this table. All the Ai* attributes and their values are stored in this table. There is a column for attribute value which is of type string. Each value of this column represents $a_{\mathrm{i,j}}^{\mathrm{p}}$.The following theorems explains how this data-preprocessing phase transforms RDB to RDB+.

Theorem 1.

D₄ transforms RDB to RDB$^+$. To show this we just need to consider the following two observations. This theorem constitutes the way that an RDB$^+$ can be mapped to a flat table which is an essential part of our work for similarity measure computation. Figure 9 and Figure 10 show the conventional and proposed way of transforming RDB and RDB$^+$ to their corresponding flat table, respectively.

Observation 1.

If A$_i$ is not an APA then $a^{p}_{k,j}$ is always mapped to the same location in RDB$^+$

Observation 2.

If A^{*}_{i} is an APA then $a^{*p}_{k,j}$ will not be mapped to the same location in RDB$^+$.

Implementation of this method and dealing with missing values

The pre-processing software is implemented as part of the work for this thesis. As shown in Figure 9, the user goes through three phases using this software:

1. *Attribute selection*
2. *Automatic joint*
3. *Data Aggregation*

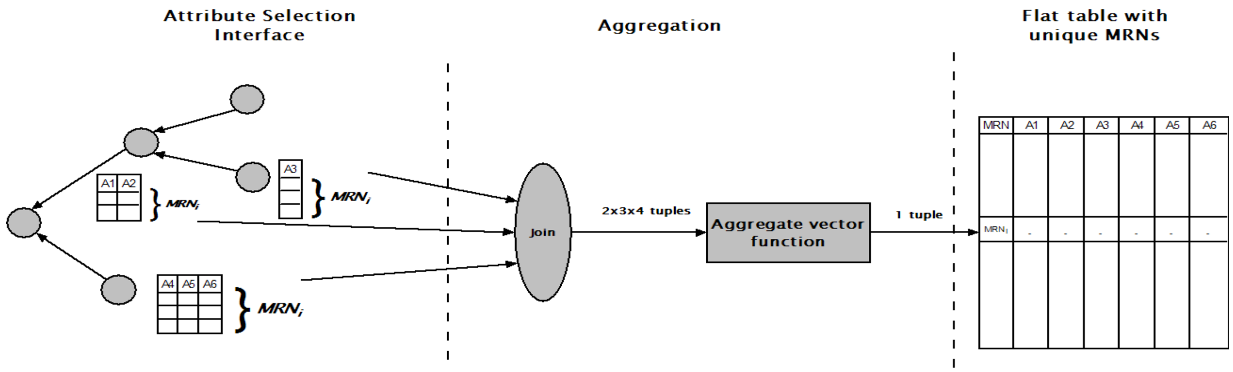


Figure 9 – Conventional procedure for constructing flat table for RDB

Attribute selection: The user selects the desired attributes from different tables.

Automatic join: As the attributes can be selected from different tables, they have to be joined together in order to construct a single flat table. Making a flat table has some advantages and disadvantages. The main advantage is that it makes the data analysis very straightforward and each patient will turn into one point within feature space. Since the flat table representation of an RDB discards the one-to-many relationships we lose some information that has been presented using the relation database architecture through the foreign keys. For example, when two tables are joined, the aggregated information by the group-by statement will not be available any more.

Aggregation: In the result of an automatic join, there may be more than one value per attribute for each patient that makes the analysis inaccurate and biased to the number of values for a specific attribute [33]. Table 2 shows an example of joining two tables with one-to-many relationship that have resulted in multiple records for a single patient. In the aggregation process for each patient one tuple will be produced (flat table). Data aggregation will help in the issue of missing values as well. In the example of Table 3, average has been used as the aggregate function for numerical attributes. Other alternatives are median, max and min. Max and min are not good representative values as they can be outliers. Median may select the outlier when selecting from two values. Therefore average is a fairly good representative value for an attribute. As it can be seen in Tables 2 and 3, multiple records per patients has been aggregated into one and number of missing values has been reduced from two to zero.

Table 2 - Joining two tables with one-to-many relationship

MRN	BNT	DelVmem	ImVmem
1	40	20	
1		24	25
2	20	30	21

Table 3 - Result of aggregating multiple results

MRN	BNT	DelVmem	ImVmem
1	40	22	25
2	20	30	21

Figure 10, shows the need for a *mapping filter* for RDB* databases. A mapping filter is needed for A* attributes. The mapping is done considering definition D_4. A base table that defines A_k attributes and the mapping filter maps A_k using the base table to a virtual attribute in its parent table. After this mapping, the rest of data preparation process becomes like a RDB database.

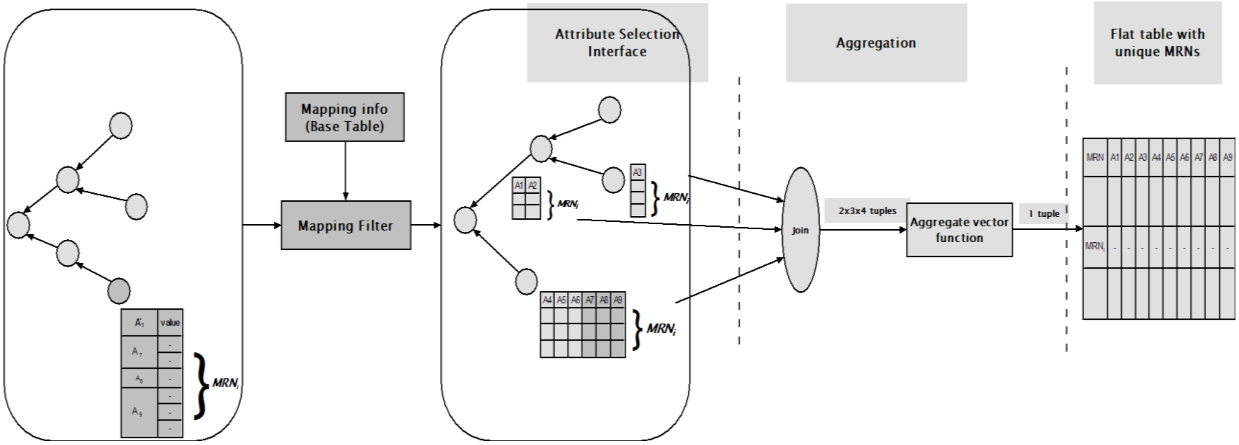


Figure 10 – Proposed new procedure forconstructing flat table for RDB*

Filtering

The table generated from the transformation phase needs extra filtering for the following reasons:

1. The classification modules require the data to have a specific format otherwise they do not produce results. They treat nominal, numerical and string values differently. Namely the string values for the selected attributes in the database need to be converted into nominal values.
2. There are noise and inconsistencies in the data. For example, sometimes the data entry has used capital 'Y' and lower case 'y' representing a positive answer to a question. While they have the same meaning, the classification module will treat them differently. Filtering needs to be done to fixe these values.
3. Data has too many classification labels compared to the number of patients in each class. This will result in poor classification accuracy and results. Custom

filtering rules need to be defined to decrease the number of distinct labels by grouping them.

A filtering phase is added before performing any data-mining. Standard and custom filters from the Weka programming library are used here after the Transformation phase.

Data mining

There are only a small portion of data mining techniques that can be used directly in diagnosis. They help in the medical expert's decision based on the medical information about past patients and their responses to diagnosis. Classical data mining methods (e.g. Clustering, Classification, and Association Rule Mining) often cannot be used directly in these tasks because they are not designed to be interactive. Also they need lots of manual preprocessing and post processing. For this project an interactive data mining method called **Similarity measurement** is designed and implemented, which is described in the following section.

Similarity Measurement

Finding similar retrospective patients to a prospective one can help in diagnosis as well as prognosis. This facilitates evidence-based decisions, which are statistically more significant. The challenge would be how to define similarity measure, how to deal with missing values and how to prioritize and weigh different patients' features.

A method is developed that allows the medical expert to select the attributes that are important when trying to find a similar patient. After querying the database, a table is presented to the user. Each row stands for one patient. The user selects the patient of interest that all the other patients need to be compared with. A range of 5 different weights can be assigned to attributes to define how important the attribute is in the similarity measurement. Then the patients will be sorted based on their similarity to the selected patient. This allows the expert to see the similar patients and use their history and diagnosis as a factor when diagnosing the new patient.

Method

There are several similarity measures in the literature. We use the well-known Euclidean distance followed by feature normalization and feature weight assignment.

Normalization: Normalize the feature values and map into [0 1] so that the range of values does not contribute to the distance.

$$Normalized(x_i) = x_i / (\max(X) - \min(X))$$

Linear Combination: Weights can be assigned to features based on expert's domain knowledge.

The Weighted Euclidean distance is defined as following:

$$D(x, y) = \sqrt{\sum_{i=1}^{l} w_i (x_i - y_i)^2}$$

Missing value treatment

The problem here in a data set with large portion of missing values is how to find the Euclidean distance between two vectors when some of the attributes are not available. Furthermore, how to compare two distances and come to the conclusion that the smaller distance stands for a higher similarity when the number of attributes that the distance is computed based on, is not the same.

There are several ways to deal with missing values [34]:

1. To ignore all features with missing values. This approach can be used when the rate of missing values is very low and losing them will not have a major effect. In our case, if we apply this approach we end up having no feature according to the large number of missing values.

2. To consider the mean or median value of all the existing values for the feature-patient. This approach would bias the distance measure in the medical case that we have a lot of missing values toward the mean or median.
3. To consider a penalty ratio for the number of missing values when calculating the distance between two patients. This approach outperforms the other two when the medical expert is looking for the most similar patients –who are diagnosed in the past - to a new one under diagnosis.

$$b_i = \begin{cases} 0, & \text{if both } x_i \text{ and } y_i \text{ are available} \\ 1, & \text{otherwise} \end{cases}$$

$$D(x,y) = \frac{l}{l - \sum_{i=1}^{l} b_i} \sqrt{\sum_{\text{all } i: b_i = 0} w_i (x_i - y_i)^2}$$

where l is the number of values.

This approach allows us to use all the records and still have a fair distance measure with respect to missing values, thus we adopt it in this project. Compared to the previous two methods it has the following benefits:

a. It doesn’t ignore all the similar patients with missing values.
b. It doesn’t have any bias towards the mean or median
c. It can bring the highly similar patients which have missing values to the attention of the medical expert.

Data visualization

Looking at the data exploration methods not all of them are suitable for every data set and sometimes a combination of them need to be used. Choosing the appropriate technique can be based on:

a. Structure of the data set: Examples of data set structures are single table, time series, and relational database.

b. Attribute types: Examples are numerical, string, nominal/categorical.

c. The User: Examples of different types of users are statistical analyst, domain expert, business manager.

d. Number of data points: Different visualization techniques perform differently depending on the size of the dataset. For example pixel-based techniques are among the best when the number of data points is in the order of 10,000 but they are not the best for 100 data points.

e. The type of pattern we are looking for. For example, if pattern is two dimensional, there is probably no need to use more complex visualization techniques and a single scatterplot would be enough.

f. The amount of interaction needed. For example highly interactive techniques are not useful when the visualization output needs to be printed.

g. The amount of missing data.

Considering the above criteria an interactive “scatter plot matrix” technique is designed for data exploration in HBIDS. The pros and cons section evaluates this technique based on the above parameters.

Description of the method

After selecting the candidate attributes for visualization, a scatter plot matrix is drawn on the screen. Each cell in this matrix contains a 2d scatter plot. For example cell (2, 3) is the projection of the data set on the A2xA3 plane - A2, A3 being the second and third attributes -. One of the problems with scatter plot matrixes is that, not all of the attribute combinations are interesting to the user. Therefore they may even become a distraction that prevents the user from focusing on the more important combinations. For this reason, the ability to remove some combinations is considered.

The brushing technique is used in the scatter plots. Using the brushing technique one can select the projected data points on a scatter plot and assign a color to them (brush). These data points will appear with their new color on all other scatter plots. This empowers the user with a multi-dimensional understanding of data using the interactive brushing of scatter-plots. Medical experts can apply their hypothesis on the data by assigning different colors to groups of patients of their choice in one diagram and see how these groups differentiate in other diagrams. Figure 11shows an example where five attributes of a patient are selected and only three combinations are selected for visualization. All the points are initially red. The user defines two groups of patients in the bottom scatter-plot in green and blue colors and all the data points in the scatter plots are colored accordingly.

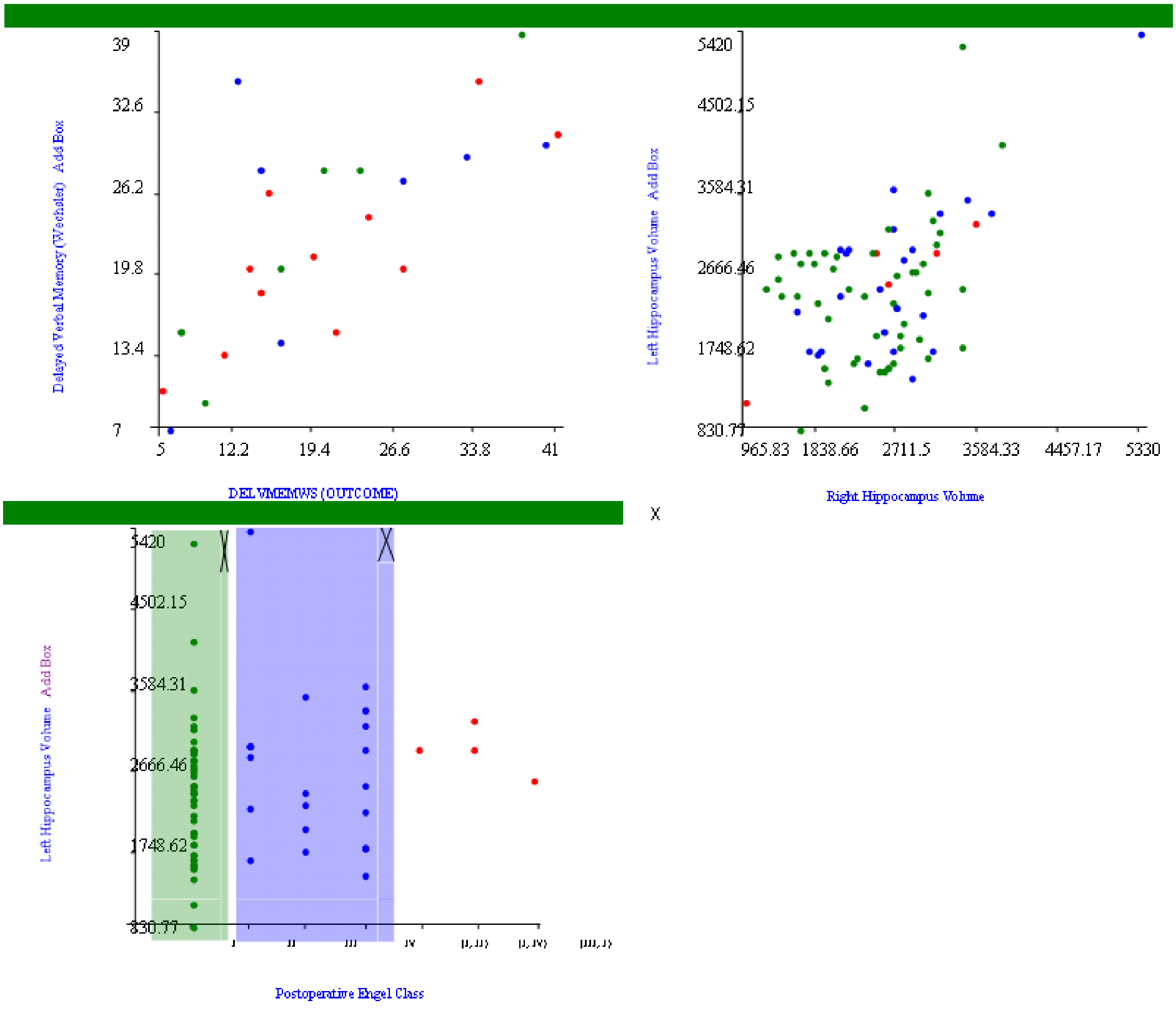


Figure 11 – Patient Data Visualization with connected scatter plots

There are some points that need to be discussed about the developed method:

1. **Missing Data**: This is a very powerful technique to visualize data sets that contain missing data. To show a data point in a 2d scatter plot, it doesn't matter if the value of other attributes are available. One need to note that, the points in a scatter plot can stay un-changed even if all of the data points in another scatter plot are colored. Therefore if a data point is not colored in a diagram it can have two meanings:

    a. The data point was not selected.

   b. The data point had missing value in at least one of the attributes of the diagram which it was selected.

2. **Overlapping groups:** If two groups overlap, the color of the group which was selected most recently will be considered.
3. **Categorical (Nominal) attributes:** For the nominal attributes the axis is equally divided by the number of categories but the order of the categories doesn't have any specific meaning. The benefit of this approach is that all the data types can be shown in diagrams similar to numerical data types. The downside of this approach appears in scatter plots that both of the attributes are categorical and therefore all the data points that have the same attribute values appear as a single point. Consequently, it is not clear how many data points have that value. There are ways to overcome this shortcoming. For example by drawing a larger circle when two data points overlap each other.
4. **Axis ranges (min-max):** The scale and range of values on the axis changes the way a user interpret a pattern in the diagram. The minimum and maximum value of a numerical attribute is not always available in the database therefore there is no way other than using the existing minimum and maximum values of the attribute in the database. This approach does not perform well when the number of data points is limited because the existing minimum value could be far above the possible minimum value. Therefore since the axes have the same length, they have different scales which make the comparison harder.

Pros and Cons

We use the parameters in the previous section to talk about the pros and cons and if this technique is among the best to explore the data in HBIDS.

- *Structure of the data set*: This technique works the best for cases where each data point is represented as a row in a table and columns represent the attributes. This is the preferred structure in most of the conventional data exploration techniques, which is not the case in HBIDS. However, the data aggregation and preprocessing discussed in the previous chapter converts the HBIDS data to this structure, i.e., flat table. Therefore, this technique works on HBIDS data set very well.

- *Attribute types*: Since HBIDS contains both numerical and nominal data types, this technique which deals with both data types is very appropriate.
- *The User*: Scatter plot is widely known and used in all disciplines. Since this technique is based on a known technique, it is amongst the best for being user friendly and does not need extra knowledge or training.
- *Number of data points*: Number of data points is not a limitation in this approach. The radius of a circle representing a data point can be changed according to the number of data points. It can be as small as a pixel like the pixel-based visualization approaches. Another way to mitigate this is to add noise to differentiate between same data-points in a diagram.
- *The type of pattern we are looking for*: Scatter plots empowered by the brushing technique do not help the user to see patterns which relate to more than 4 attributes. By brushing one 2-d scatter plot, a pattern in a separate 2-d scatter plot can be observed. If a third scatter-plot is colored, it erases the effect of coloring of the first scatter plot and therefore it cannot show patterns which relate to more than 4 dimensions.
- *The amount of interaction needed*: Interactive techniques are usually more attractive than non-interactive techniques since the users can experience their curiosities in exploring the data and they can try different ways of grouping data points. In the same time, the result of the data exploration can be easily printed to show the extracted pattern.
- *The amount of missing data*: As discussed earlier, missing data in one attribute of a data point does not limit the technique in showing the data point in combinations that do not

contain that attribute. Therefore this technique is very appropriate in data sets with missing data.

Summary

This chapter describes the developed data pre-processing, similarity measurement and data visualization techniques for data analysis in HBIDS.

In order to prepare the data in HBIDS for analysis, a set of pre-processing techniques have been developed. The goal is to have a single attribute value for every patient. To achieve this goal, a) transformation, b) attribute selection, c) aggregation and d) filtering are applied to the data.

A single attributed called "all-purpose" is used to allow the researchers add new content-dependent metadata without the need to introduce new database columns. Transformation is needed to present such attribute values like all structured database attributes for attribute selection. Once a set of attribute is selected, multiple database tables are queried. At this point, a single table is ready. The only problem is that, each patient will have multiple rows. This can be due to multiple examinations for the patient. After this, data aggregation is done to consolidate multiple values for an attribute given a patient id. For numerical attributes we use average. For categorical attributes the assigned attribute value is the set of all distinct values regarding the patient.

When the medical expert wants to make a decision for a new patient, he or she wants to look at the similar patients which have been diagnosed in the past. It is very hard to do this task using a relational database. A tool is developed to find the similar patients to a new patient given the output of the developed pre-processing phase. The medical expert selects the attributes that he is interested in and assigns weights to them. Then the distances from all patients are calculated regarding the assigned attribute weights. Since there is a lot of missing values in the data set, the distance formula takes into account the missing values and assigns a penalty to them during this calculation. This way, past patients with missing attribute values can still come to the medical expert's attention if the value of their existing attributes is close to the selected patient.

A data visualization technique is developed which gives the medical expert the ability to see multi-dimensional patterns. After the medical expert selects the attributes that he or she is interested in, the data is shown in a set of connected scatter-plots. He can further examine hypothesizes by coloring different sets of patients in one scatter-plot to see their value in other scatter-plots. This technique is very suitable for data with missing values because to show a patient data point in one scatter-plot, only two attribute values needs to be present.

The valuable data in HBIDS couldn't be consumed directly by the medical expert without the developed methods described here. These techniques handle the complexity of this system and present the data in simple interactive forms. The aggregation phase in data pre-processing step is a key to make the data visualization and similarity measurement steps possible. This step creates a summary of the patient record to make the next steps possible.

Chapter 4 Experiments

Implementation

As part of this thesis, an interactive web application is developed. It lets the medical expert to select the attributes of interest in the web browser. The application then performs the pre-processing method mentioned in Chapter 2 using SQL statements and Java code. This data is then shown to the user as a single table. The medical expert can use both the developed similarity measurement and data visualization techniques.

Both similarity measurement and visualization techniques work based on XML data and use XSLT to separate data from presentation. JavaScript is used to make the application interactive by enabling the user to assign weights to attributes and sort patients based on the measured similarity to a patient of interest. Vector Markup Language (VML) is used to draw the connected scatter-plots for visualization. Enabling the user to connect scatter plots by brushing is also done using JavaScript and VML.

Data Analysis results

After applying the data preprocessing techniques described in Chapter 3, data mining techniques discussed in Chapter 3 have been applied. Below are the results of this experiment on HBIDS data for each technique:

Similarity Measurement

A web based application is designed to assist the medical experts to find the most similar patient to the one under treatment. Below are the steps that the medical expert takes:

1- The expert selects the relevant attributes for the study and queries the database. The application returns a flat table. There is one row per patient and columns represent attributes.

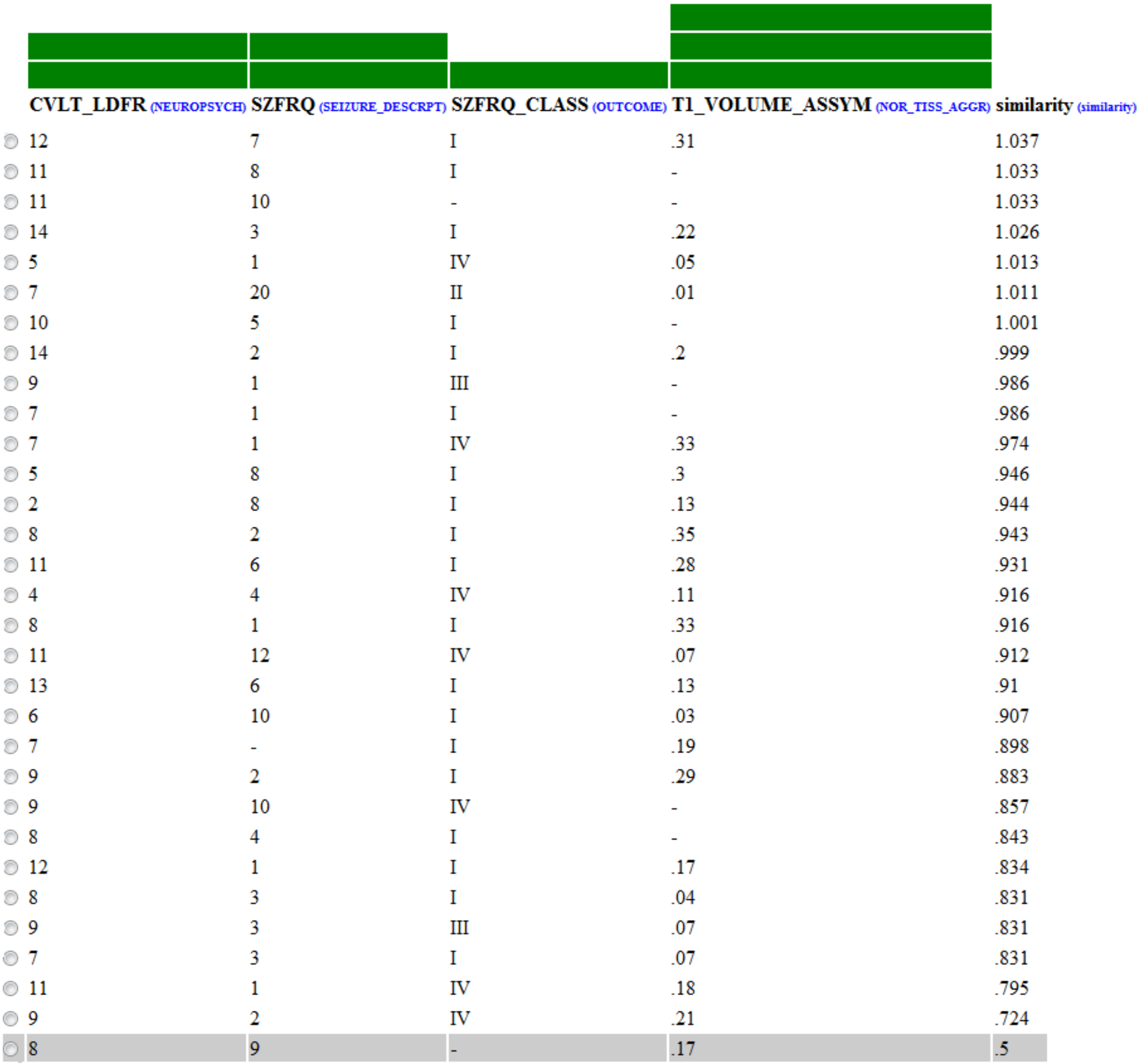
2- The expert then selects a patient.

3- Then the expert assigns a weight to each attribute based on their relevance.

4- The application calculates the similarity of each patient to the selected one and sort them based on such similarity measure.

5- At this point the physician can use this ordered list to make decisions based on similar patients to the patient in hand.

Figure 12, shows the sorted list of patients based on a new patient. The medical expert has selected three attributes and assigned a higher weight to Volume asymmetry (the column name in the figure is: T1_VOLUME_ASSYM).



	CVLT_LDFR (NEUROPSYCH)	SZFRQ (SEIZURE_DESCRPT)	SZFRQ_CLASS (OUTCOME)	T1_VOLUME_ASSYM (NOR_TISS_AGGR)	similarity (similarity)
○	12	7	I	.31	1.037
○	11	8	I	-	1.033
○	11	10	-	-	1.033
○	14	3	I	.22	1.026
○	5	1	IV	.05	1.013
○	7	20	II	.01	1.011
○	10	5	I	-	1.001
○	14	2	I	.2	.999
○	9	1	III	-	.986
○	7	1	I	-	.986
○	7	1	IV	.33	.974
○	5	8	I	.3	.946
○	2	8	I	.13	.944
○	8	2	I	.35	.943
○	11	6	I	.28	.931
○	4	4	IV	.11	.916
○	8	1	I	.33	.916
○	11	12	IV	.07	.912
○	13	6	I	.13	.91
○	6	10	I	.03	.907
○	7	-	I	.19	.898
○	9	2	I	.29	.883
○	9	10	IV	-	.857
○	8	4	I	-	.843
○	12	1	I	.17	.834
○	8	3	I	.04	.831
○	9	3	III	.07	.831
○	7	3	I	.07	.831
○	11	1	IV	.18	.795
○	9	2	IV	.21	.724
○	8	9	-	.17	.5

Figure 12 - Patients sorted based on the weighted similarity to a selected patient.

Data Visualization

In this section we show few examples on how the developed visualization technique assists the medical expert in understanding the data.

i. Example 1

Figure 13 is constructed by querying the data for all the patients regarding three attributes –using the web interface- from the HBIDS system. Here are the steps that the medical expert follows given the connected scatter plot:

1. In the first look, the distribution of attribute values can be seen in each of the 2-d scatter plots. All of the patient data points are colored as red.
2. Medical expert wants to understand if side of the brain which the surgery has been performed on, has any significant correlation with the examination results of the WADA test. Using the right diagram, he brushes all the patients with right side surgery using blue and the ones with left side surgery using green.
3. The result on the left diagram which shows the WADA test results reveals a pattern. A straight line can be drawn that separates the blue points from the red points (with few exceptions).

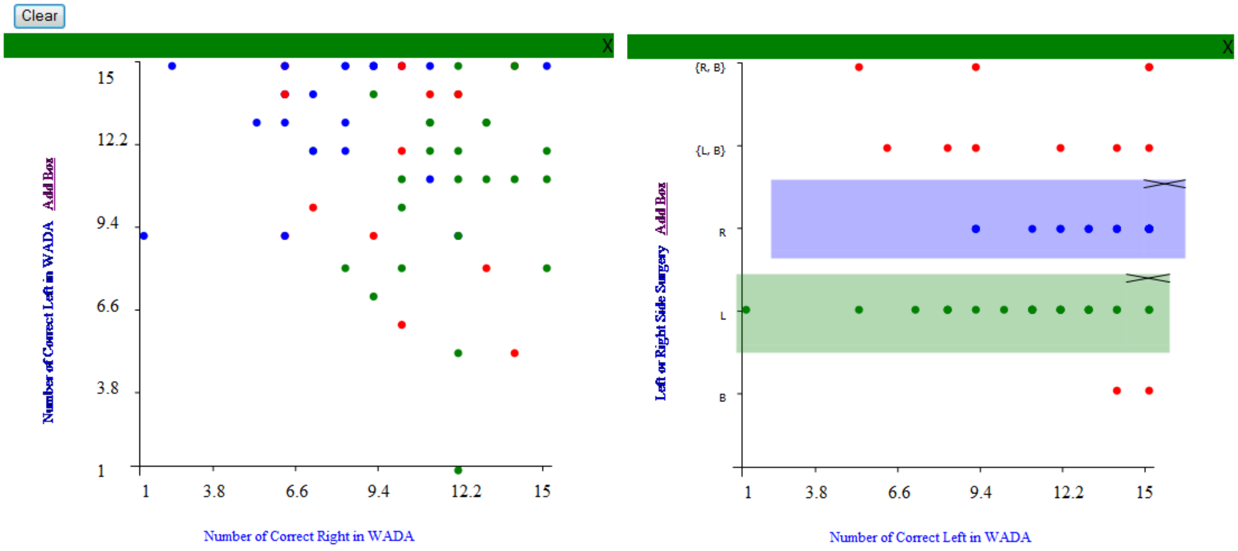

Clear
X
Number of Correct Left in WADA
Add Box
15
12.2
9.4
6.6
3.8
1
1
3.8
6.6
9.4
12.2
15
Number of Correct Right in WADA
X
Left or Right Side Surgery
Add Box
{R, B}
{L, B}
R
L
B
1
3.8
6.6
9.4
12.2
15
Number of Correct Left in WADA

Figure 13 - Connected scatter plots using brushing technique that shows the relationship between three attributes

ii. Example 2

Figure 14, is constructed by selecting five attributes from the HBIDS system. The following steps show how the brushing technique helps the medical expert to see patterns in this data. The medical expert's goal in this scenario is to see the difference in data distribution for the selected attributes regarding two groups of patients: a) The patients which have had a successful surgery and therefore the value for 'Postoperative Engel Class' attribute in them is 'I'. b) The patients which didn't have a successful surgery and therefore have a value different than 'I' regarding the 'Postoperative Engel Class'.

1. The targeted group is selected in the top-right diagram and colored as black and therefore all such patients get colored as black throughout the scatterplots. Then the expert focuses on each of the scatter-plots to see if interesting patterns can be observed.
2. At the first glance on the top-left scatterplot, an interesting pattern can be seen. There are more red points on the diagonal line compared to the blue points. The blue points are pretty much evenly distributed all over. From the medical point of view, this would mean that, patients with similar WADA scores on the right and left side of the brain have a lower chance of success in the surgery compared to others.
3. The two bottom scatterplots show the relationship between, right and left WADA scores with 'Hippocampus volume asymmetry'. These diagrams don't seem to reveal any interesting patterns.

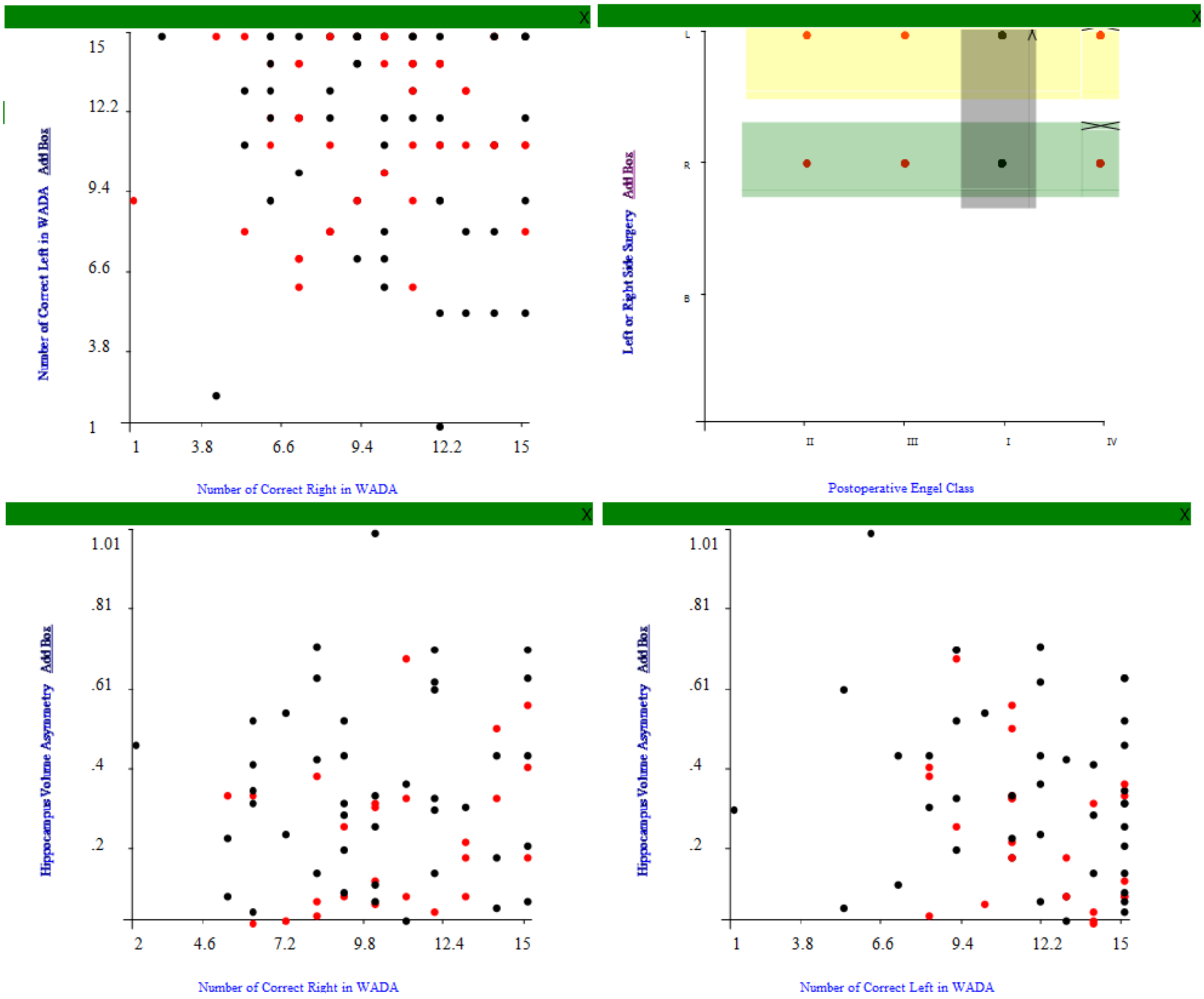


Figure 14 - Connected scatter plots that shows the relationship between five attributes and brushing the patients with value 'I' for the 'Postoperative Engle Class' attribute.

iii. Example 3

The medical expert now uses the same attributes used in Example 2 but does the brushing differently as shown in. The goal here is to see if interesting patterns can be observed in two groups of patients regarding the selected attributes: a) Right side of the brain has been operated on b) Left side of the brain has been operated on. Medical expert does the following steps:

1. Select the targeted group using the top-right scatterplot and color the patients with right side brain surgery as green and patients with left side brain surgery as yellow.
2. The data on the top-left scatterplot has already been observed in Example 1 by brushing the same group of patients.
3. Looking at the bottom-left scatterplot, there are only green points above the diagonal line. Similarly in the bottom-right scatterplot, there are only yellow points above the diagonal line. It seems that there is a potential here to design a rule to help in identifying the side of the brain that should be subject to surgery. This needs to be investigated and examined further by the medical expert.

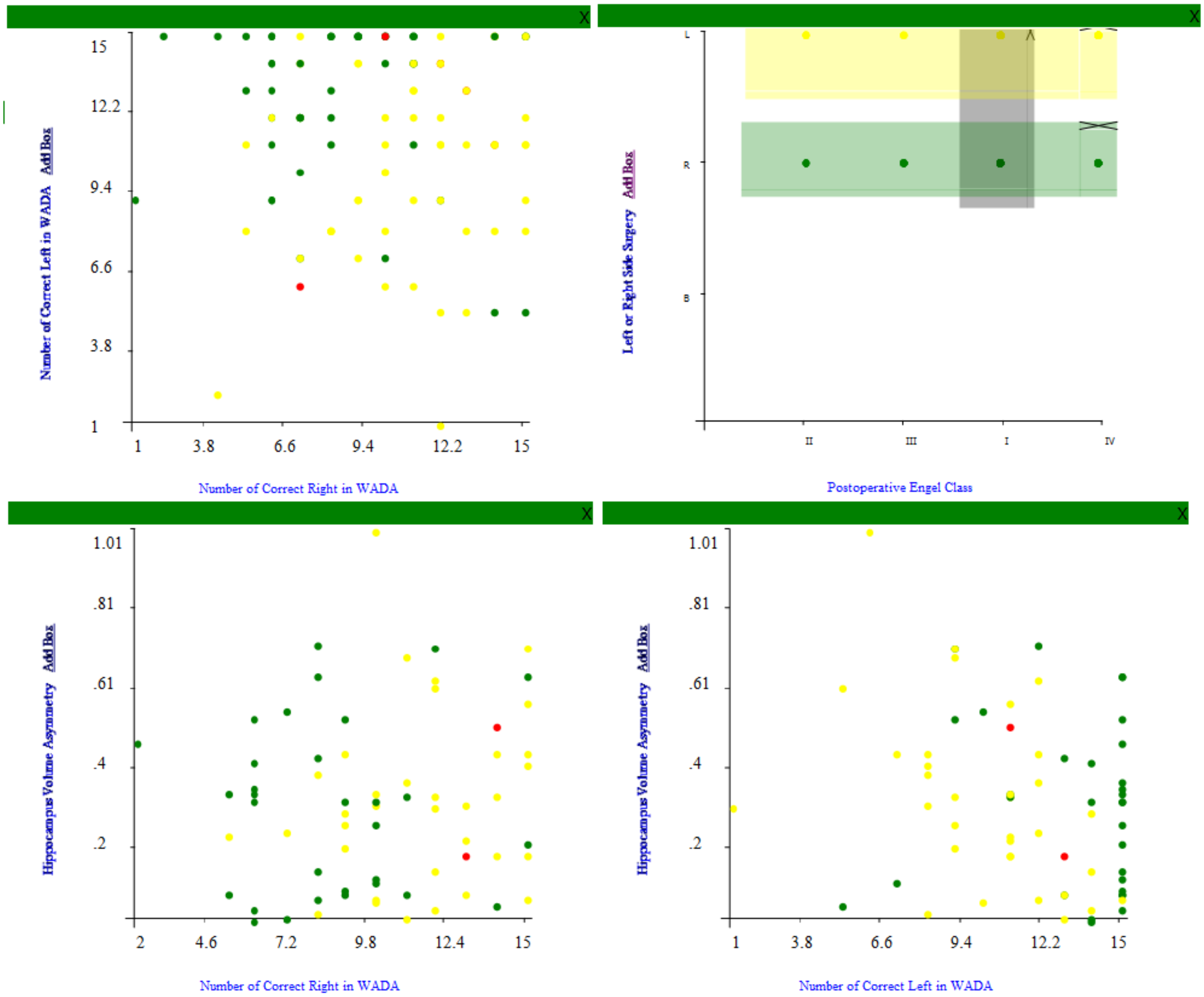


Figure 15 - Connected scatter plots that shows the relationship between five attributes and brushing the patients with right and left side brain surgery differently.

iv. Example 4

In the above examples the medical expert has only done the brushing based on categorical attributes. Figure 16 and Figure 17show an example of the brushing technique for a combination of categorical and numerical values. The goal of the medical expert is to see how the attribute values are distributed among the three pre-surgical numerical assessments with regards to the success of the surgery. The medical expert goes through the following steps:

1. Select the three pre-surgical numerical attributes and surgery outcome and identify a threshold based on domain knowledge for Hippocampus volume asymmetry

(patients with values above 0.2 are colored green and the ones with values bellow 0.2 are yellow).

2. In the first figure (Figure 16), patients with poor surgery outcome are only colored (the rest are red which is the default color).In the second figure (Figure 17), patients with good surgery outcome are only colored.
3. Now the medical expert needs to compare the plot on the left in both find if any patterns can be found.

In this example, there is no interesting pattern that looks apparent from the first look. However this shows how this visualization technique can help in revealing multi-dimensional patterns.

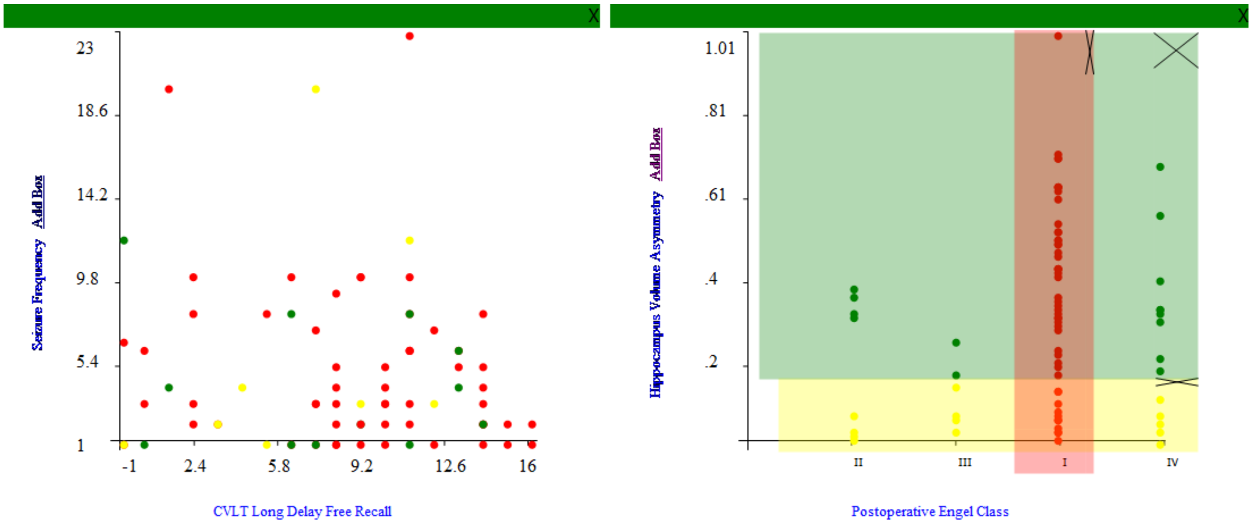


Figure 16 - Connected scatter plots showing the relationship between four attributes. Patients with non-successful surgery outcome are brushed with two colors according to the value of 'Hippocampus Volume Asymmetry'.

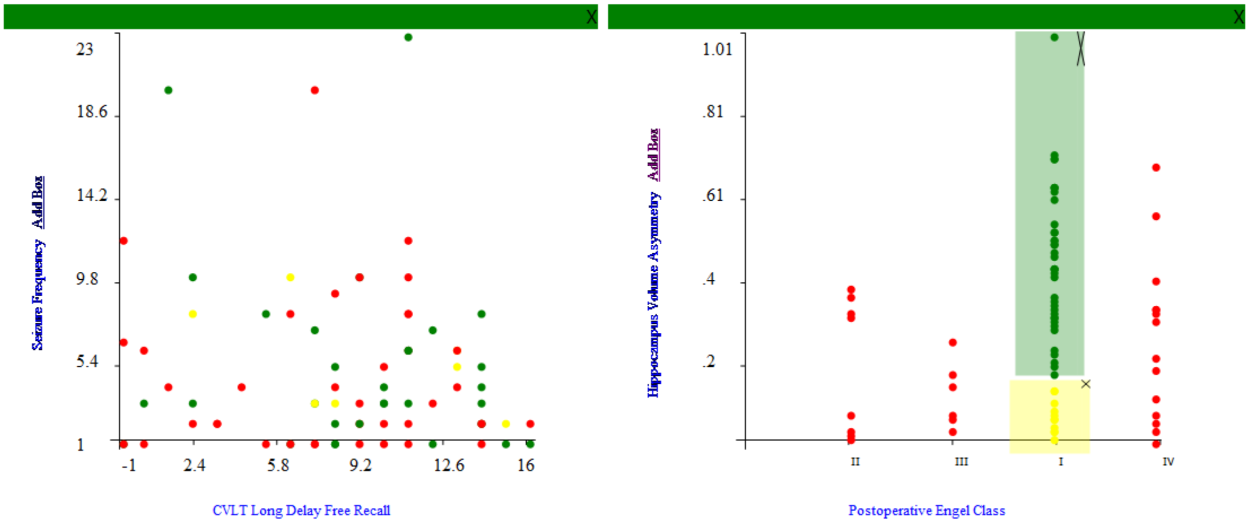


Figure 17 - Connected scatter plots showing the relationship between four attributes. Patients with successful surgery outcome are brushed with two colors according to the value of 'Hippocampus Volume Asymmetry'.

Classification

Predicting the surgery outcome based on medical assessments prior to the surgery is one of the highly interesting subjects to the medical experts. In this section we intend to use the data in HBIDS for this purpose. After data pre-processing, the data is imported in the Weka application and different classification algorithms are executed. Notice that the data contains all the attributes of the patient and the classifiers try to select the best set of attributes for classification. Following is the results of the classification experiments performed on the data. The classification label is binary and the values are either "Yes" or "No" ("Yes" identifies a successful surgery and "No" identifies an unsuccessful surgery).

i. J48

J48 is a decision tree based algorithm based on the famous C4.5. The following parameters are set for this classifier prior to the experiment:

a) Minimum number of instances per leaf: This has been set to 15. This is because we are interested only in rules with enough evidence.

b) Sub-tree raising: This is to consider raising the sub tree after pruning.

Classification Result

The generated decision tree is shown in Figure 18. Table 4andTable 5 show the classification accuracy and confusion matrix for this experiment.

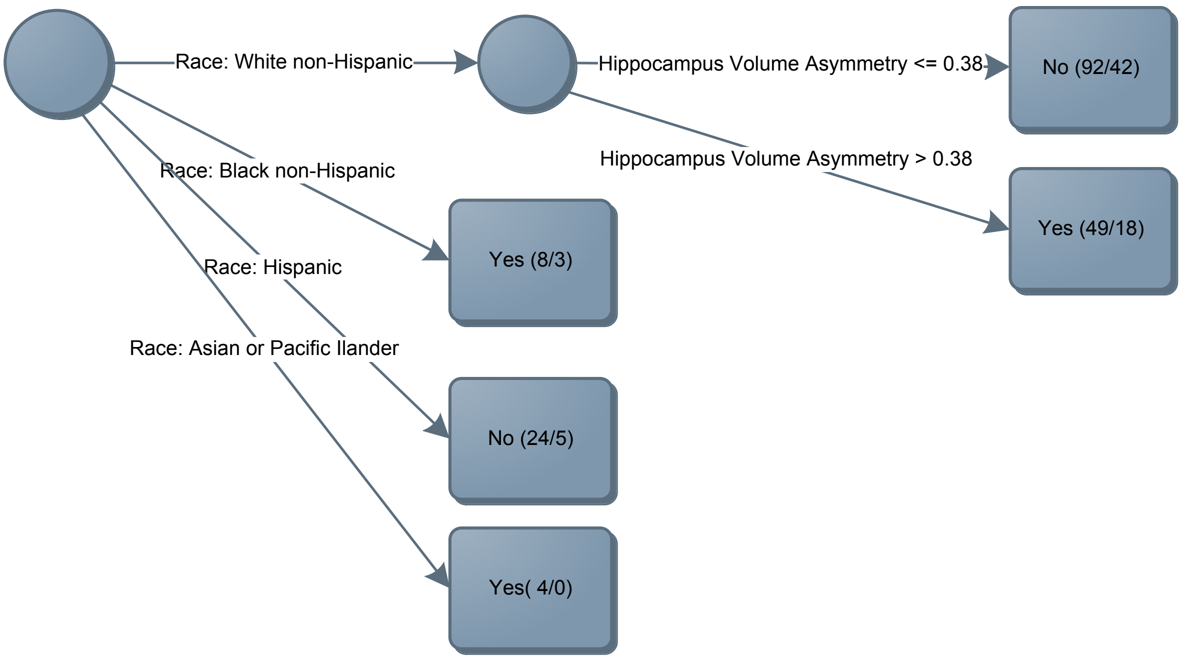


Figure 18 - J48 Classification Model to predict the surgery outcome. As an example, “No (92/42)” means the predicted surgery outcome is negative. There are 92 patients whom this rule applies to them and 42 patients which contradict this prediction.

	Count	Percentage
Correctly Classified Instances	91	51.12%
Incorrectly Classified Instances	87	48.88%

Table 4 - Classification accuracy for the J48 decision tree

		Predicted class	
		Yes	No
Actual Class	Yes	60	27
	No	60	32

Table 5 - Confusion matrix for the J48 decision tree

Analysis

The classification accuracy clearly shows that J48 has performed poorly on this dataset with 51% correctly classified instances. Looking at the confusion matrix, the results are proportional to the number of instances in each class. It means this classifier is similar to a random classifier that generates classification labels in a rate proportional to the percentage of each class in the training data. The generated decision table has one interesting data. The Hispanic race shows a very low success rate on surgery (only 5 out of 29 cases end up with good results). This can be a candidate for further analysis.

ii. Alternating Decision Tree (ADTree)

ADTree algorithm generates an alternating decision tree. The output of an alternating decision tree to a given data point is a predicted class label with an associated confidence measure. Each classification label has a number assigned to it (e.g. Yes: 1, No: -1). Smaller distance to an associated number of a class label means a higher confidence in the prediction.

In an alternating decision tree, there are two types of nodes: a) Decision nodes which identify the child node that needs to be traversed based on the attribute values. b) Prediction nodes, which are leaf nodes in the tree and assign a number to a single traversal of the tree (this number resembles the confidence measure according to the traversed path in the tree). In order to calculate the confidence measure for a given data point we only need to find all the possible Prediction nodes (the leaf nodes) which can be reached from the root through the Decision nodes and calculate the summation.[39]

The following parameter values are set in the Weka ADTree classifier prior to the experiment:

Parameters

a) Number of boosting iterations: 5

b) Search path: expand all paths

Classification Result

The generated classification tree is shown in Figure 19. A positive confidence measure indicates successful surgery and a negative one indicates non-successful surgery. Table 6, Table 7 and Table 8 show the classification accuracy, confusion matrix and the classification accuracy measures for each class in this experiment.

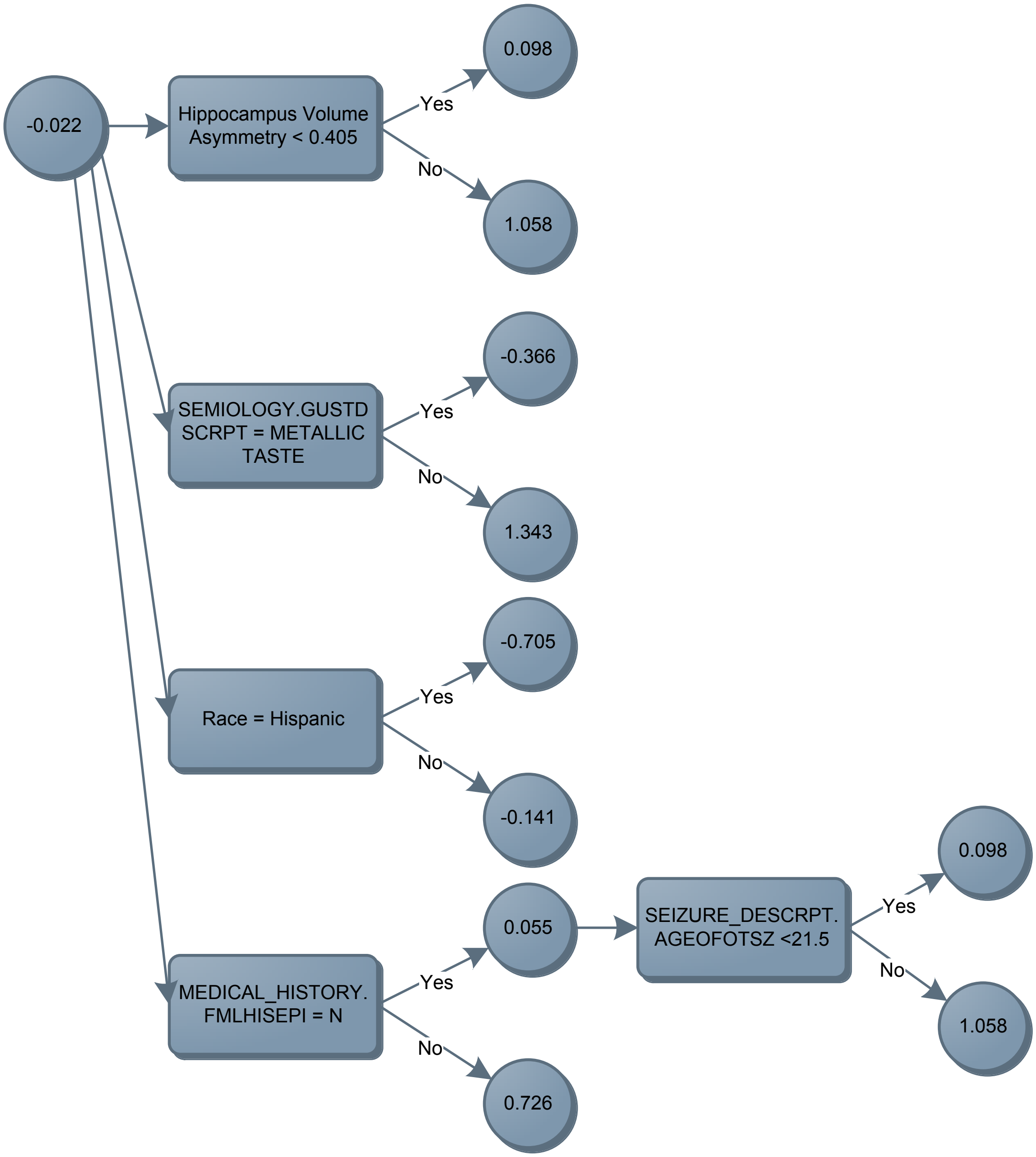

-0.022
Hippocampus Volume Asymmetry < 0.405
Yes
0.098
No
1.058
SEMIOLOGY.GUSTD SCRPT = METALLIC TASTE
Yes
-0.366
No
1.343
Race = Hispanic
Yes
-0.705
No
-0.141
MEDICAL_HISTORY. FMLHISEPI = N
Yes
0.055
No
0.726
SEIZURE_DESCRPT. AGEOFOTSZ <21.5
Yes
0.098
No
1.058

Figure 19 – ADTree classification model to predict the surgery outcome

	Count	Percentage
Correctly Classified Instances	115	64.60%
Incorrectly Classified Instances	63	35.40%

Table 6- Classification accuracy for ADTree experiment

		Predicted class	
		Yes	No
Actual Class	Yes	55	32
	No	31	60

Table 7 - Confusion matrix for the ADTree experiment

TP Rate	FP Rate	Precision	Recall	F-Measure	Predicted Class
0.659	0.368	0.652	0.659	0.656	No
0.632	0.341	0.64	0.632	0.636	Yes

Table 8- classification accuracy measures for each class in ADTree experiment

Analysis

ADTree not only provides a decision tree, but the numbers associated with each branch provides a confidence factor. That indicates how much an attribute weighs in the decision making. This classifier has better accuracy (64%) compared to J48. However, 64% cannot be considered a good accuracy. One interesting pattern here is that, the "Hippocampus volume asymmetry" attribute weighs differently for values less than 0.405 compared to values greater than 0.405. The weight is 1.058 for higher numbers compared to .098 for lower numbers. This means if volume asymmetry is greater than .405 the classifier has much higher confidence in predicting a good outcome.

iii. DecisionStump

About this algorithm: It's a classification method consisting of a one-level decision tree.

Parameters: This classifier has no parameters

Classification Result

This classifier attempts to find the single best attribute which can be used for classification. The generated classification tree is shown in Figure 20. Table 9, Table 10 and Table 11 show the classification accuracy, confusion matrix and classification accuracy measures for each class.

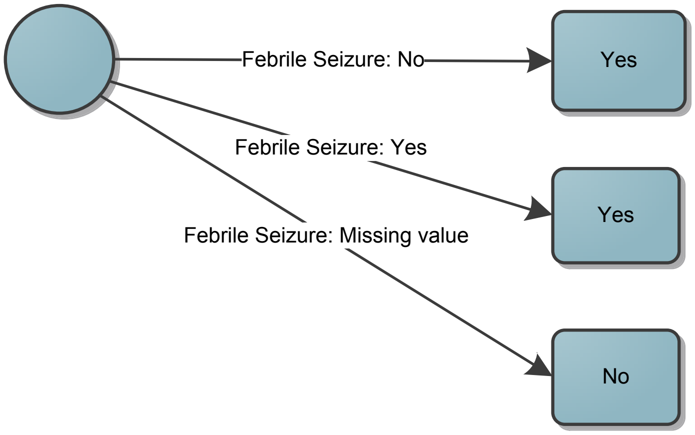


Figure 20 - DecisionStump classification model to predict surgery outcome

	Count	Percentage
Correctly Classified Instances	125	70.22%
Incorrectly Classified Instances	53	29.78%

Table 9 - classification accuracy for DecisionStump experiment

		Predicted class	
		Yes	No
Actual Class	Yes	85	2
	No	51	40

Table 10 - confusion matrix for DecisionStump experiment

TP Rate	FP Rate	Precision	Recall	F-Measure	Predicted Class
0.44	0.023	0.952	0	1	No
0.977	0.56	0.625	1	1	Yes

Table 11 - classification accuracy measures for each class in DecisionStump experiment

Analysis

This classifier so far had the best classification accuracy (70%), however this is not yet a good enough accuracy. The number of instances with missing value in the “Febrile Seizure” attribute which have value for the “Surgery Outcome” attribute is very low (3 instances) and therefore the related branch in the classification model is not useful.

iv. *AdaBoost M1*

About this algorithm: This is a Meta-classifier for boosting a nominal class classifier using the Adaboost M1 method. Only nominal class problems can be tackled by this method. It often dramatically improves performance, but sometimes it over-fits.

Parameters

c) Classifier: DecisionStump is used as the base classifier.

d) Number of iterations: 3.This means it creates three decisionStump classifiers based on the data and assigns different weights to each of them.

Classification Result

The created model is shown in Figure 21, Figure 22 and Figure 23. It has created three classifiers bellow based on decision stump and assigned different weights to them. The left-most circle in each figure shows the weight that is associated with the classifier by the meta-classifier. Table 12, Table 13 and Table 14 show the classification accuracy, confusion matrix and the classification accuracy measures for each class.

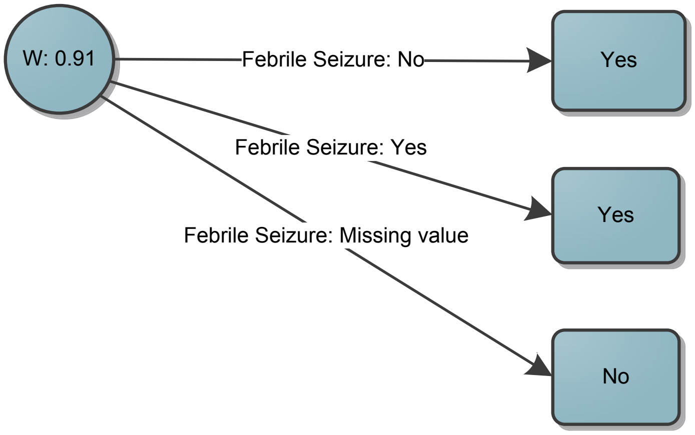


Figure 21 - AdaBoost first classifier to predict surgery outcome

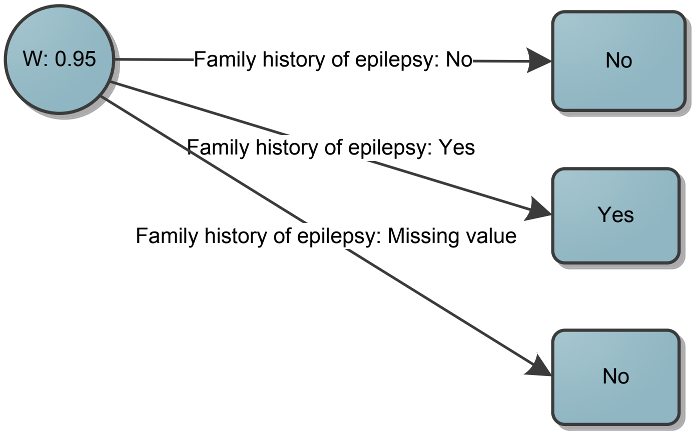

W: 0.95
Family history of epilepsy: No
No
Family history of epilepsy: Yes
Yes
Family history of epilepsy: Missing value
No

Figure 22 - AdaBoost second classifier to predict surgery outcome

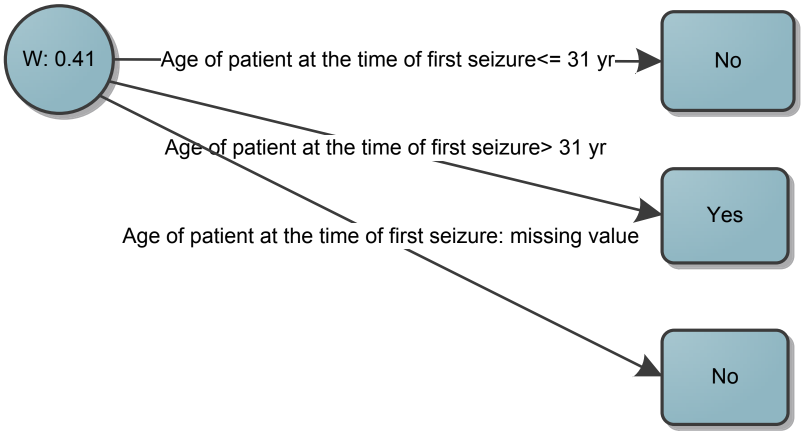

W: 0.41
Age of patient at the time of first seizure<= 31 yr
No
Age of patient at the time of first seizure> 31 yr
Yes
Age of patient at the time of first seizure: missing value
No

Figure 23 - AdaBoost third classifier to predict surgery outcome

	Count	Percentage
Correctly Classified Instances	124	69.66%
Incorrectly Classified Instances	54	30.34%

Table 12 - Classification accuracy for AdaBoost experiment

		Predicted class	
		Yes	No
Actual Class	Yes	85	2
	No	52	39

Table 13 - Confusion matrix for AdaBoost experiment

TP Rate	FP Rate	Precision	Recall	F-Measure	Predicted Class
0.429	0.023	0.951	0	1	No
0.977	0.571	0.62	1	1	Yes

Table 14 - classification accuracy measures for each class in AdaBoost experiment

Analysis

The classification accuracies are very similar to that of the DecisionStump. Comparing the weights of each classifier is an interesting subject for further investigation. For example the first two classifiers use “Family history of epilepsy “and “Febrile seizure “and have very similar weights. In contrast, the third classifier which uses “Age of patient at the time of first seizure” has half of the weight of the former attributes. Here, the missing values again have played a role in increasing the classification accuracy which needs to be investigated further. One important point to consider is that using multiple classifiers did not increase the classification accuracy. This is another area for investigation on why Boosting have not helped with accuracy. Boosting is considered to improve classification accuracy substantially.

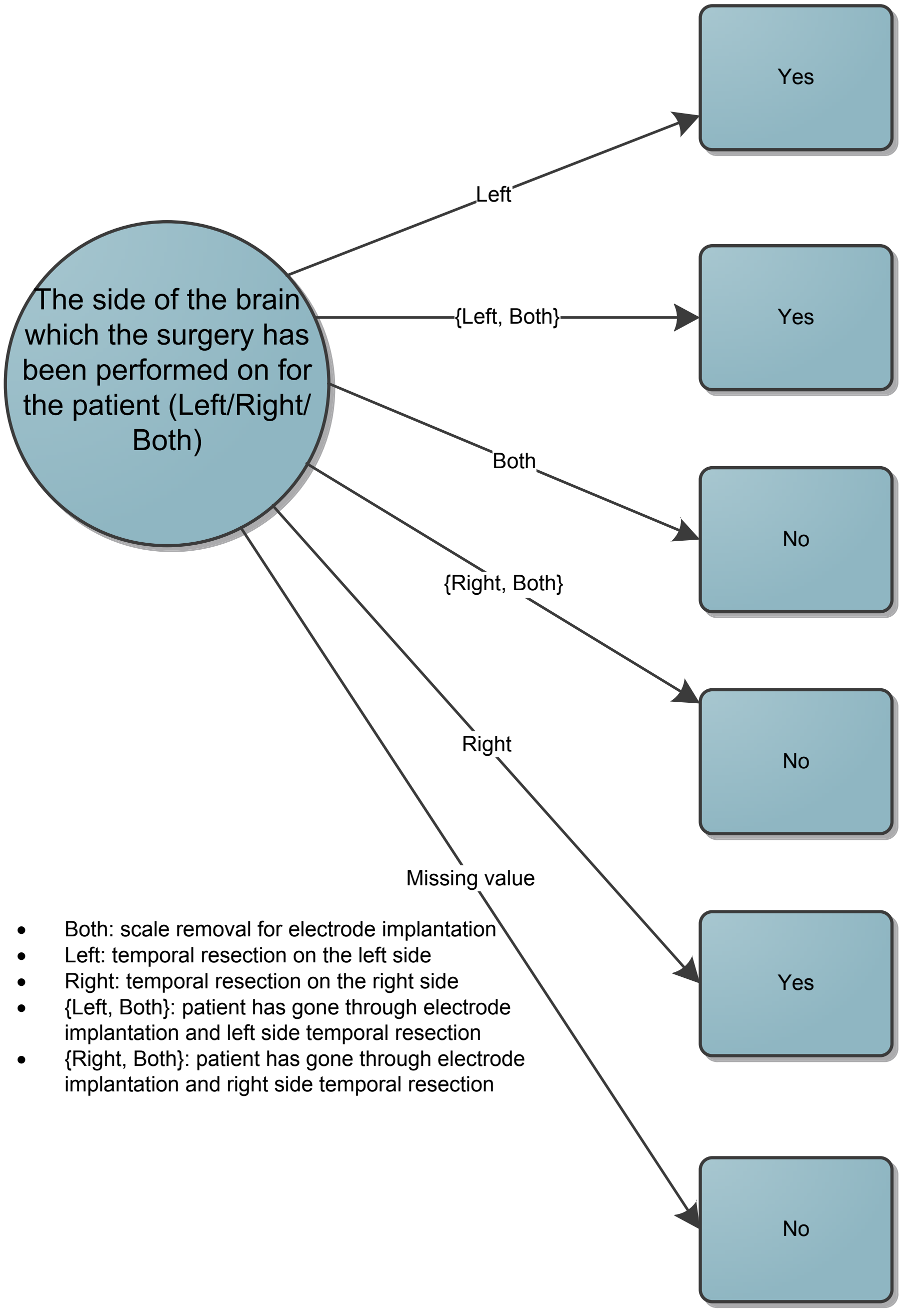
v. *One Rule*

About this algorithm: This classifier uses the minimum-error attribute for prediction, discretizing numeric attributes.

Parameter: Minimum number of instances that needs to be in each numeric bucket for discretizing: 10.

Classification Result

The OneRule classifier's goal is to find the best attribute that can predict the class label and extracts one rule using that attribute. Figure 24 is the trained classification model. Table 15, Table 16, Table 17 show the classification accuracy, confusion matrix and the classification accuracy measures for each class.


The side of the brain which the surgery has been performed on for the patient (Left/Right/ Both)
Left
{Left, Both}
Both
{Right, Both}
Right
Missing value
Yes
Yes
No
No
Yes
No

- Both: scale removal for electrode implantation
- Left: temporal resection on the left side
- Right: temporal resection on the right side
- {Left, Both}: patient has gone through electrode implantation and left side temporal resection
- {Right, Both}: patient has gone through electrode implantation and right side temporal resection

Figure 24 - OneRule classification model to predict the surgery outcome

	Count	Percentage
Correctly Classified Instances	126	70.78%
Incorrectly Classified Instances	52	29.21%

Table 15 - Classification accuracy for OneRule experiment

		Predicted class	
		Yes	No
Actual Class	Yes	74	13
	No	39	52

Table 16 - Confusion matrix for OneRule experiment

TP Rate	FP Rate	Precision	Recall	F-Measure	Predicted Class
0.571	0.149	0.8	0.571	0.667	No
0.851	0.429	0.655	0.851	0.74	Yes

Table 17 - Classification accuracy measures for each class in OneRule experiment

Analysis

The classification accuracy of the OneRule classifier is very similar to that of the DecisionStump. One missing result here compared to DecisionStump, is the number of instances that match each rule. This can be very beneficial for the domain experts to understand the rule in the context of patients which match with a leaf node. A future work here is to have information on what is the accuracy of each leaf node.

Summary

This chapter briefly describes how the data visualization and similarity measurement were implemented. It also shows few examples of the usage of the developed methods.

Several classification algorithms are applied to the data in order to create a model for predicting the resection surgery outcome for patients with temporal lobe epilepsy. J48, ADTree, DecisionStump, AdaBoost and OneRule are the classifiers that are used. There are two factors involved in selecting these classifiers for this experiment. a) They are amongst the most popular classifiers and b) They can deal with both numerical and categorical attributes. The highest classification accuracies are from One-Rule and Decision-Stump and it’s around %70 which are the simplest classifiers among the ones which are used here.

Chapter 5 Conclusion

In this project, we showed how a pallet of techniques is used to extract new knowledge from a Human Brain Image Database System (HBIDS). This includes: data preparation, filtering, similarity measurement, interactive visual data mining and classification techniques. Amongst them, data preparation is essential to do any of the rest because of the nature of data in this system which is spread in multiple tables. First, a method was designed that creates flat tables out of the HBIDS's relational database. Some filtering was done on top of that due to inconsistency issues in the data entry.

The designed interactive visual data mining tool lets the medical expert to browse through this complex database to find interesting patterns and study different hypotheses in connected scatter plots. This tool is integrated with a similarity measurement tools that lets the expert find similar diagnosed patients to a new patient.

Different classification methods are applied to the data using Weka. The classification results did not provide good accuracies (Max: 70%) when used to predict outcome.

Future Work:

The following are several areas which are important for future work:

a) Converting a relational database into a flat table causes data loss. Methods should be designed to a) minimize this data loss by adding constrains when several values of a specific field need to be aggregated. b) Design smart/interactive aggregators which can take domain knowledge to find better categorical values for an attribute.

b) A feedback loop should be included in the data mining systems to improve data quality. Adding this module to the system will ensure increasingly higher data quality. Classification methods sometimes find patterns which reveal issues either in the data entry software or the data gathering mechanism. While the medical expert is investigating a hypothesis using the interactive data mining application he/she figures out such issues with data. Feedback loop should be used here to fix the data or order gathering some data.

c) The similarity measurement technique only considers numerical attributes. Therefore method needs to be developed to take calculate the similarity measure based on both numerical and categorical attributes.

d) As mentioned in Chapter 4, the classification methods used did not produce models with high accuracies. The low classification accuracy needs to be further investigated. Classic classification methods may not represent the optimal way for extracting knowledge on how to predict a surgery outcome. These methods create classification models for predicting the class label for **any** instance. As a result the models have very low accuracy for specific types of patients. Some modifications to these methods or filtering their results can help in creating more accurate models.

REFERENCES

[1] K. J. Cios and W. Moore, "Uniqueness of Medical Data Mining," vol. 26, 2002.

[2] W. Ceusters, "Medical Natural Language Understanding as a Supporting Technology for Data Mining in Healthcare," in *Medical Data Mining and Knowledge Discovery*, Springer, 2001.

[3] U. N. L. o. Medicine, "Unified Medical Language System® (UMLS®)," 29 July 2009. [Online]. Available: http://www.nlm.nih.gov/research/umls/index.html. [Accessed 15 November 2009].

[4] "Unified Medical Language System," 12 November 2009. [Online]. Available: http://en.wikipedia.org/wiki/Unified_Medical_Language_System. [Accessed 15 November 2009].

[5] J. J. Berman, M. E. Edgerton and B. A. Friedman, "The tissue microarray data exchange specification: A community-based, open source tool for sharing tissue microarray data," vol. 3, 2003.

[6] J.-W. Byun, T. Li, E. Bertino, N. Li and Y. Sohn, "Privacy-preserving incremental data dissemination," *Journal of Computer Security,* vol. 17, no. 1, pp. 43-68, 2009.

[7] M.-R. Siadata, H. Soltanian-Zadehb, F. Fotouhid, A. Eetemadib and K. Eliseviche, "Content-based image database system for epileps," *Computer Methods and Programs in Biomedicine,* vol. 79, 2005.

[8] E. Reynolds and J. W. Kinnier, "Psychoses of epilepsy in Babylon: the oldest account of the disorder.," vol. 49, no. 9, 2008.

[9] 26 3 1998. [Online]. Available: http://www.medterms.com. [Accessed 24 12 2008].

[10] "MedicineNet.com," [Online]. Available: http://www.medicinenet.com/seizure/page2.htm.

[11] P. M. Daniel B. Hoch, "National Library of Medicine - National Institutes of Health," 29 5 2008.

[Online]. Available: http://www.nlm.nih.gov/medlineplus/ency/article/003931.htm.

[12] C. W. Bazil, "Medical History," Epilepsy.com, 8 March 2004. [Online]. Available: http://www.epilepsy.com/EPILEPSY/testing_medhistory. [Accessed 15 February 2012].

[13] D. Pyle, Data Preparation for Data Mining (the Morgan Kaufmann Series in Data Management Systems), Los Altos, CA: Morgan Kaufmann, 1999.

[14] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2/e, Elsevier Inc., 2006.

[15] P. Giudici, Applied Data Mining Statistical Methods for Business and Industry, Wiley & Sons, 2003.

[16] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," vol. International Journal Of Medical Informatics, no. 77, 2008.

[17] C. Fellbaum, WordNet An Electronic Lexical Database, MIT Press, 1998.

[18] B. Zupan, J. Holmes and R. Bellazzi, "Knowledge-based data analysis and interpretation," vol. Artifitial Intelligence in Medicine, no. 37, 2006.

[19] "Poll: Data Mining Methods," KDnuggets, April 2006. [Online]. Available: http://www.kdnugets.com/polls/2006/data_mining_methods.htm.

[20] J. Quinlan, C4.5: Programs for Machine Learning, San Mateo: Morgan Kaufmann Publishers, 1993.

[21] L. Breiman, Classification and Regression Trees, Chapman and Hall/CRC, 1984.

[22] P. Clark and T. Niblett, "The CN2 Induction Algorithm," *Machine Learning,* p. 261–283, 1989.

[23] R. S. Michalski and K. A. Kaufman, "Learning Patterns in Noisy Data: The AQ Approach," *Machine Learning and Its Applications.*

[24] D. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed., New York: Wiley, 2000.

[25] G. Schwarzer, W. Vach and M. Schumacher, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Statistics in Medicine,* pp. 541-61, 2000.

[26] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines: and other kernel-based learning methods, reprint ed., Cambridge University Press, 2006.

[27] R. Bellazzi and B. Zupanb, "Predictive data mining in clinical medicine: Current issues and guidelines," *international journal of medical informatics,* vol. 77, pp. 81-97, 2008.

[28] V. Vapnik, Statistical Learning Theory, New York: Wiley.

[29] T. Hastie, R. Tibshirani and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, New York: Springer, 2001.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations,* vol. 11, no. 1, 2009.

[31] D. Keim and H.-P. Kriegel, "Visualization Techniques for Mining Large Databases: A Comparison," *IEEE Transactions on Knowledge and Data Engineering,* vol. 8, no. 6, 1996.

[32] Wikipedia Foundation, "Parallel coordinates," Wikipedia, 19 April 2012. [Online]. Available: http://en.wikipedia.org/wiki/Parallel_coordinates. [Accessed 19 April 2012].

[33] C. Brunsdon, A. S. Fotheringham and M. E. Charlton, "An Investigation of Methods for Visualising Highly Multivariate Datasets," 1998.

[34] E. Keough, "Institute for Visualization and Perception Research," University of Massachusetts Lowell, [Online]. Available: http://web.uml.edu/gallery/index.php/IVPRPhotos/. [Accessed 2012 April 2012].

[35] H. Chernoff, "The Use of Faces to Represent Points in k-Dimensional Space Graphically," *Journal American Statistical Association,* vol. 68, pp. 361-368, 1973.

[36] D. A. Keim, "Pixel-oriented Visualization Techniques for Exploring Very Large Databases," *Journal of Computational and Graphical Statistics,* vol. 5, pp. 58-77, 1996.

[37] K. S. Feiner, "Worlds Within Worlds:Metaphors for Exploring N-Dimensional Virtual Worlds," in *UIST '90 Proceedings of the 3rd annual ACM SIGGRAPH symposium on User interface software and technology*, New York, 1990.

[38] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier and S. Philippi, "Graph-based analysis and visualization of experimental results," *Bioinformatics,* vol. 22, no. 11, pp. 1383-1390, 2006.

[39] S. Mohammad-Reza, S.-Z. Hamid, F. Farshad and E. Kost, "Content-based image database system for epilepsy," *Computer Methods and Programs in Biomedicine,* March 2005.

[40] "American Library Association, Task Force on Metadata Summary Report," Chicago, 1999.

[41] X. Yin, J. Han, J. Yang and P. S. Yu, "Efficient Classification across Multiple Database Relations: A CrossMine Approach," *IEEE Trans. On Knowledge and Data Eng, IEEE computer society,,* vol. 18, no. 6, June 2006.

[42] S. Theodoridis and K. Koutroumas, Pattern Recognition, Greece: Academic Press, 1998.

[43] Y. Freund and L. Mason, "The Alternating Decision Tree Algorithm," in *Proceedings of the 16th International Conference on Machine Learning*, Bled, Slovenia, 1999.

[44] B. AG, B. M, M. MD, B. R and A. SS, "Data mining and XML: current and future issues. In: Proceedings of the First International Conference on Web Information Systems," Hong Kong, 2000.

[45] S. Tsumoto, "Problems with Mining Medical Data," Washington, DC, 2000.

ABSTRACT

MEDICAL DATA ANALYSIS METHOD FOR EPILEPSY

by

AMEEN EETEMADI

December 2012

Advisor: Dr. FarshadFotouhi

Major: Computer Science

Degree: Master of Science

Applying data mining techniques on medical databases which contain un-structured and semi-structured data is a challenging task. It is not only due to the complexity of such databases but also due to the characteristics of the medical domain. This thesis describes how multiple layers of data mining techniques have been applied to a Human Brain Image Database system. It starts with data preparation which paves the way for conventional data analysis techniques to be applied to the data. A similarity based patient retrieval tool has been designed and developed to assist in treatment planning and outcome estimation for epileptic patients. Finally connected scatter-plot visualization tool has been designed and implemented in order to assist the medical experts to see the relationship between attributes and visually compare a new patient's similarity scores against patients that have been previously operated on in the hospital.

AUTOBIOGRAPHICAL STATEMENT

After receiving his bachelor’s degree from Sharif University of Technology, majoring in Software Engineering, he was admitted to the graduate program at the Department of Computer Science, Wayne State University. While a graduate student, he collaborated with the Radiology Image Analysis Lab at Henry Ford Health System, working on the Human Brain Image Database System (HBIDS), during which he has developed an interest in Medical Data Mining and Image Processing. His master’s thesis presented here was performed under supervision of Dr. Fotouhi at Wayne State University and Dr. Soltanian-Zadeh at Henry Ford Health System. Contributions of his thesis can be described within the framework of the existing HBIDS developed originally by Dr. Siadat, Oakland University. After finishing the coursework, he had the opportunity to work on DNA sequencing algorithms at Microsoft Research as a research intern. He is now working with Microsoft Workflow team as a Software Engineer in Test.