DOKUZ EYLUL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

MEDICAL DATA MINING BY USING MATLAB/SAS

by

İzzet ÇAVUŞLAR

August, 2008

İZMİR

MEDICAL DATA MINING BY USING MATLAB/SAS

A Thesis Submitted to the

Graduate School of Natural and Applied Sciences of Dokuz Eylül University

In Partial Fulfillment of the Requirements for the Master of Science in

Statistical, Statistics Program

by

İzzet ÇAVUŞLAR

August, 2008

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"MEDICAL DATA MINING BY USING MATLAB/SAS"** completed by **İZZET ÇAVUŞLAR** under supervision of **ASST. PROF. EMEL KURUOĞLU** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Emel KURUOĞLU
Supervisor

Prof. Dr. Efendi NASİBOĞLU
(Jury Member)

Prof. Dr. Şenay ÜÇDOĞRUK
(Jury Member)

Prof.Dr. Cahit HELVACI
Director Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

Special thanks to Asst. Prof. Emel Kuruoğlu who guided my life with the best way and to my family who has never ever stopped supporting me.

İzzet ÇAVUŞLAR

MEDICAL DATA MINING BY USING MATLAB/SAS

ABSTRACT

Nowadays each individual and organization - business, family or institution can access a large quantity of data and information about itself and its environment. Information is scattered within different archive systems that are not connected with one another, producing an inefficient organization of the data. Two developments could help to overcome these problems. First, software and hardware continually, offer more power at lower cost, allowing organizations to collect and organize data in structures that give easier access and transfer. Second, methodological research, particularly in the field of computing and statistics, has recently led to the development of flexible and scalable procedures that can be used to analyze large data stores. These two developments have meant that data mining is rapidly spreading through many businesses as an important intelligence tool for backing up decisions.

In this thesis data mining process has been expressed. Subjects obtained as how to process the data, which stages to migrate, which model would be available for the data. Statistical issues supporting data mining process has been analyzed. Then issue of how data mining in medical area which is one of the usage areas of data mining is obtained. Similarity between normal data mining process and medical data mining process has been analyzed. Problems faced in medical data mining process are obtained and methods to solve these problems are searched. Appliance is done on clustering and logistics regression, one of the statistical methods that support data mining. Clustering appliance is held on EEG data derived from the biophysics Program University of Dokuz Eylül. Appliance on logistics regressions has been made by data derived from patients of IVF department of the a private hospital. These appliances have been resulted as data mining is incredibly important in medical area and that is so beneficial in solution of the huge amounts.

Keywords: Medical data mining, clustering, logistic regression

MATLAB/SAS KULLANILARAK TIBBİ VERİ MADENCİLİĞİ

ÖZ

Bugünlerde her bir birey ve organizasyon – iş, aile, kurum – kendisi ve çevresi hakkında ciddi miktarda veriye ve bilgiye ulaşabilir. Bilgiler, birbirine bağlanamayan farklı arşiv sistemlerinde, dağınık şekilde yer almaktadır. Bu da verinin verimsiz şekilde biraya getirilmesine sebep olmaktadır. Bu problemlerin üstesinden gelebilecek iki gelişme vardır: Birincisi; yazılım ve donanımlar, sürekli olarak, daha düşük maliyetlerle, organizasyonların veri giriş ve transferini kolaylaştıran yapılar sunarak, veri toplanması ve organizasyonunu daha işlevsel hale getirmektedir. İkincisi; özellikle hesaplama ve istatistik alanlarındaki metodik araştırmalar, son dönemlerde, büyük veri depolarını analiz edebilen esnek ve ölçeklendirilebilir süreçlerin geliştirilmesine sebep olmuştur. Bu gelişmelerle birlikte tıbbi veri madenciliği yöntemleri kullanımı da yaygınlaşmıştır.

Bu tezde veri madenciliği kapsamında verilerin içerisindeki desenler, ilişkiler, değişimler ve istatistiksel olarak önemli olan yapılar incelenmiştir. Veri madenciliği sürecini destekleyen bazı istatistiksel yöntemler kullanılmıştır. İncelen konulardaki tıbbi verilerde karşılaşılan bazı problemler ele alınmış ve bu problemlerin çözüm yöntemleri araştırılmıştır. Veri madenciliği yöntemlerinden kümeleme ve lojistik regresyon yöntemleri uygulanmıştır. Kümeleme uygulaması Dokuz Eylül Üniversitesi Biyofizik Anabilim Dalı laboratuarından alınan Elektroensefalografi(EEG) verileri üzerinde yapılmıştır. Lojistik regresyon uygulaması için özel bir hastanenin IVF bölümündeki hastalardan alınan verilere bir model kurulmuştur. Yapılan ilk araştırma da EEG verileri farklı özellikleri ile kümelenmiştir. İkinci uygulamada kullanılan etken maddenin gebelik durumuna etkisi incelemiştir. Bunun sonucunda veri madenciliği yöntemleri ile tıbbi veriler bilgiye dönüştürülmüştür.

Anahtar Sözcükler: Veri madenciliği, kümeleme, lojistik regresyon

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

Nowadays each individual and organization - business, family or institution can access a large quantity of data and information about itself and its environment. This data has the potential to predict the evolution of interesting variables or trends in the outside environment, but so far that potential has not been fully exploited. There are two main problems. Information is scattered within different archive systems that are not connected with one another, producing an inefficient organization of the data. There is a lack of awareness about statistical tools and their potential for information elaboration. This interferes with the production of efficient and relevant data synthesis. Two developments could help to overcome these problems. First, software and hardware continually, offer more power at lower cost, allowing organizations to collect and organize data in structures that give easier access and transfer. Second, methodological research, particularly in the field of computing and statistics, has recently led to the development of flexible and scalable procedures that can be used to analyze large data stores. These two developments have meant that data mining is rapidly spreading through many businesses as an important intelligence tool for backing up decisions. (Giudici, 2003)

Data mining is being used in many scientific fields and also in medicine. Some of these fields are below:

The research that was done in 2000 by A. Kusiak, K.H. Kernstine and his friends is to find out if the tumor in the lung is a good or a bad one. According to the statistics, there are more than 160.000 lung cancer occurrences in America and 90% of them can't be saved. That's why the early diagnosis of the disease is important. By the help of the noninvasive tests, the correct diagnosis can be done with a percentage of 40% -60%. People prefer to have a biopsy to be certain that they are cancer or not. The invasive tests like the biopsy are risky and have a high cost. The data mining which was done on the invasive test data that were received in different laboratories and collected from different clinics gives the %100 correctness (Kusiak, 2000).

In 2000 Jurgen Paetz and his team did research on the people who lived between and 1993-1997 and lost their lives or stayed alive after a septic shock the target was to decrease the death toll by the help of the early warning system. For the data that was collected to find this out Neural Network data mining system was the appropriate one. The percentage of the dead due to the septic shock is 50%. To examine this case 140 different items were researched. Readings (temperature, blood pressure…), drags (dobutrex, dobutamin..) and therapy (diabetes, liver cirrhosis,..). There were 874 patients on the data base. So that the observation number on the data set is 122.600 in total. With the help of the neural network and correlation analysis methods that were performed on this data, the factors that effect the septic shocks were found out (Paetz, Hamker, Thöne, 2000).

In an article which was written in 2001 by Maria-Luiza, Antonie Osmar R. Za¨ıane and Alexandru Coman, the breast cancer was mentioned. In the last few years, the data mining methods named Neural Networks and Association were being used to diagnose and cure the breast cancer occurrences. By these methods the best diagnosis tool mammography and the mammogram rates were used on the statistical methods to record more scenes and to examine them. So that by supporting the mammography records with statistical methods the early diagnosis can be done and the possibility of missing the tumor since it's too little disappeared. (Luiza, 2000)

In the article that was written in 2001 by Ebubekir Emre Men, Özlem Yıldırımtürk and their team, the effectiveness of the medicine which is used to prevent the Atrial Fibrillation (AF) which occurs because of open heart surgery was examined. AF is the most common complication that is seen after an open heart surgery. To minimize the complication, the effectiveness of the medicine that should be given before and after the surgery was examined. In this research one of the data mining methods - the Regression – was used. 180 patients were examined. The medicine and the medicine rates were included. As a conclusion, the medicine which has the aritmic effect should be given during the 7 days before the surgery and 10 days after the surgery. So that the risk of AF could be reduced because of this the releasing of the patient became 7 days time instead of 10.

Cenk Şahin, S. Noyan Oğulata and their team did research in Çukurova Universtiy Medicine Faculty Balcalı Hospital Endocrinology department. They examined 502 patients who had tiroid disorder. The anatomical structure and the tiroid functions of the patients were examined in 5 different tests. The neural network method which is one of the data mining methods was used for their anatomical structure and to find out which of the 8 different kinds they belonged to. The work that was done by the help of the technological items that help doctors diagnose assists them in the artificial nerves method diagnosis. So that even if it is decided that there is tiroid disorder, it is also found out that the statistical method assistance leads doctors to the decision of which of the 8 different anatomies that the patient belong to.

The other study about hypertension that is done on the database which is prepared by The Korea Medical Insurence this study is done on the 127,886 records which belong to the year 1998.In the first stage it is studied on 9,103 record that has hypertension then on the same number of records that don't have hypertension. This example is done with the models training by dividing into learning that is formed of 13,689 records and the test sets that is formed of 4,588 records. Among the decision trees algorithms CHAD, C4.5, C5.0 are used in the learning algorithm. In the result of these studies the effective values about hypertension are BMI, urinary protein, blood glucose and cholesterol values. It is confirmed that living conditions (diet, intake salt, alcohol and tobacco amount) has no effect on none of these forecasts and it is also confirmed from the graphical values that only age has the effect. (Chae, Ho, Cho, Lee, Ji, 1998)

In 2002 Akat, A.Z., Doğanay, M., and their team research Laparoscopic cholecystectomy (LC) was to evaluate the factors affecting conversion rates, complication rates and operation times. Laparoscopic cholecystectomy (LC) has become the standard treatment method of cholelithiasis. This study was performed at Ankara Numune Education and Research Hospital, 4th Department of Surgery We have evaluated our first 1000 LC cases. The parameters included in the analyses were age, gender, presence of acute cholecystitis, previous abdominal surgery, concomitant diseases, liver function tests, experience of the surgeon, additional operations, findings on ultrasonography, and endoscopic retrograde

cholangiopancreatography. Consecutive univariate and multivariate regression analyses were applied to these parameters to evaluate their effect on conversion rates, complication rates and operation times. The conversion rate was 4.8%. The factors increasing the risk of conversion were male gender, previous abdominal surgery, acute cholecystitis, inexperienced surgeon, and increased gallbladder wall thickness on ultrasonography. Major operative complication rate was 3.1%. The most important risk factors for occurrence of complications were older age and acute cholecystitis. Mean operation time was 53.5 minutes. The independent factors increasing operation time were acute cholecystitis and inexperienced surgeon. Today, there is no absolute contraindication for LC. But when there is difficulty in dissection (especially acute cholecystitis), surgeon has to decide for conversion to open surgery at the right time regarding his experience, to minimize the complication risk (Akat, Doğanay, Koloğlu, Gözalan, Dağlar, Kama, 2002).

In the article that was written in 2001 by Çolak, C., Çolak, M.C., Orman, M.N., logistic regression model selection methods were compared for the prediction of coronary artery disease (CAD). Coronary artery disease data were taken from 237 consecutive people who had been applied to İnönü University Faculty of Medicine, Department of Cardiology. Logistic regression model selection methods were applied to CAD data containing continuous and discrete independent variables. Goodness of fit test was performed by Hosmer-Lemeshow statistic. Likelihood-ratio statistic was used to compare the estimated models. Each of the logistic regression model selection methods had sensitivity, specificity and accuracy rates greater than 91.9%. Hosmer-Lemeshow statistic showed that the model selection methods were successful in the description of CAD data. Related factors with CAD were identified and the results were evaluated. Logistic regression model selection methods were very successful in the prediction of CAD. Stepwise model selection methods were better than Enter method based on Likelihood-ratio statistic for the prediction of CAD. Age, diabetes mellitus, hypertension, family history, smoking, low-density lipoprotein, triglyceride, stress and obesity variables may be used for the prediction of CAD (Çolak, Orman, 2001).

In the article that was written in 2004 Okandan, M. & Kara, S. The acquired 18 normal sinus rhythm ECGs and 20 ECGs with atrial fibrillation (AF) are decomposed with db10 Daebauchies wavelets at level 6 and power spectral density was calculated for each decomposed signal with Welch method. Average power spectral density was calculated for six sub bands and normalized to be used as inputs to the neural network. Levenberg- Marquart Backpropagation feed forward neural network was built with logarithmic sigmoid transfer functions in three layer form. The trained network was tested on 6 normal and 7 AF ECGs. Performances of the classification were observed as 100 % accuracy, sensitivity and specify (Okandan, Kara, 2004).

In this thesis; the second part is about the data mining theory and the medical data mining, the third part is about the application of the clustering and logistic regression by using Matlab and SAS, and the fourth part is the conclusion.

CHAPTER TWO
MEDICAL DATA MINING

Data mining is to reach to large – scale data. In this chapter phases of data until analyzing, methods of analysis and the way it is used in medical data is mentioned. The first part is about the data mining and second part the medical data mining.

2.1 Data Mining

Data mining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. "Knowledge mining," a shorter term may not reflect the emphasis on mining from large amounts of data. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. Knowledge discovery as a process is depicted in Figure 2.1 and consists of an iterative sequence of the following steps: (Han, Kamber, 2001)

1. Data cleaning (to remove noise and inconsistent data)
2. Data Integration (where multiple data sources maybe combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

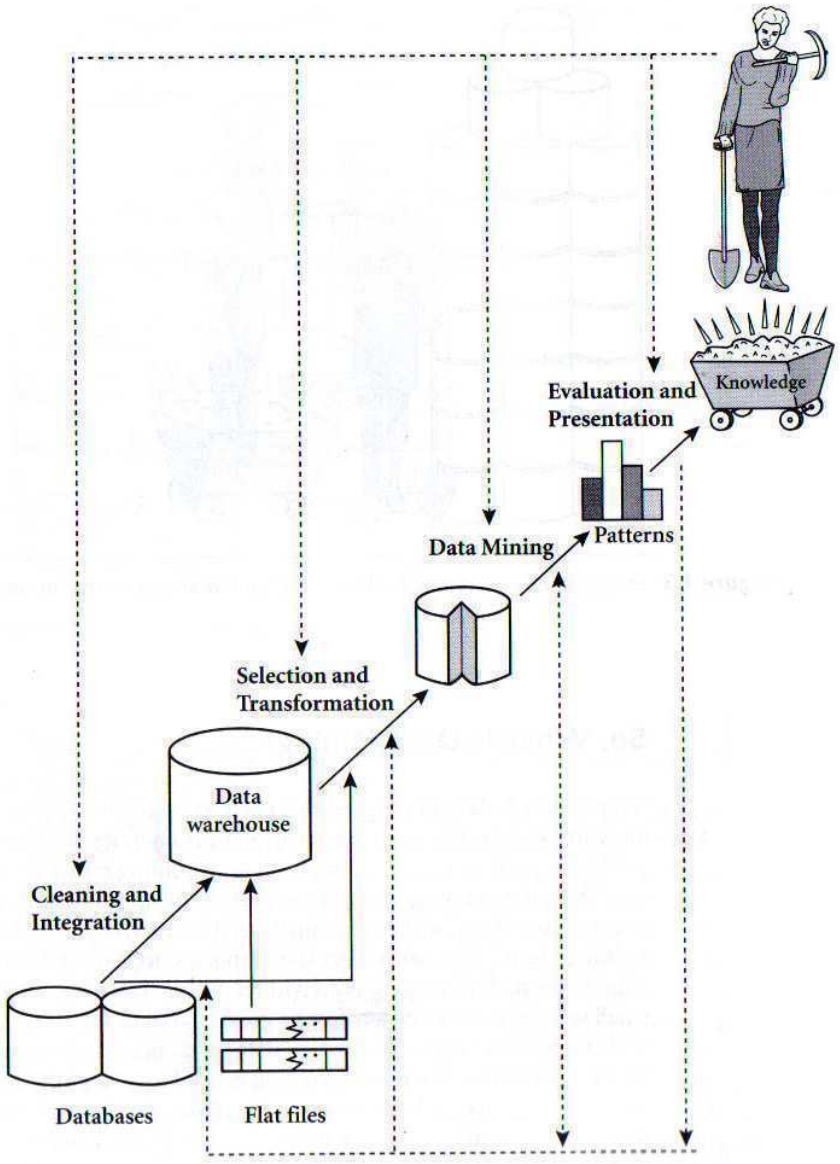


Figure 2.1 Data mining as a step in the process of knowledge discovery(Han, Kamber, 2001)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base. Note that according to this view, data mining is only one step in the entire process, albeit an essential one since it uncovers hidden patterns for evaluation.

Based on this view, the architecture of a typical data mining system may have the following major components (Figure2.1): (Han, Kamber, 2001)

1. Database, data warehouse, or other information repository
2. Database or data warehouse server
3. Knowledge base
4. Data mining engine
5. Pattern evaluation module
6. Graphical user interface.

2.1.1 Data Mining-On What Kind of Data?

We examine a number of different data stores on which mining can be performed. In principle, data mining should be applicable to any kind of information repository. This includes relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the Worldwide Web. Advanced database systems include object-oriented and object-relational databases, and specific application oriented such as spatial databases, time series databases, text databases, and multimedia databases. The challenges and techniques of mining may differ for each of the repository systems.

2.1.2 Relational Databases

A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as database, and a set of software programs to manage and access the data. The software programs involve mechanisms for the definition of database structures; for data storage; for concurrent, shared, or distributed data access; and for ensuring the consistency and security of the information stored, despite system crashes or attempts at unauthorized access.

A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (column or fields) and usually stores a

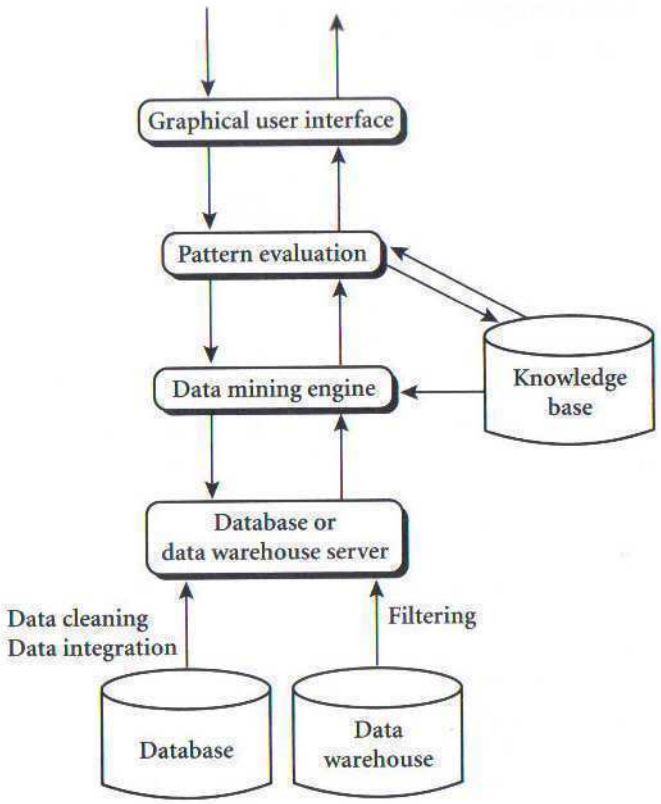


Figure 2.2 Architecture of a typical data mining system (Kamber, 2001)

large set of cells (records or rows). Each cell in a relational table represents an object identified by a unique *key* and described by a set of attribute values. A semantic data model, such as an entity-relationship (E.R) data model, which models the database as a set of entities and their relationships, is often constructed for relational databases.

2.1.3 Data Warehouses

Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. A large number of organizations have found that data warehouse systems are valuable tools in today's competitive, fast-evolving world. In the last several years, many firms have spent millions of dollars in building enterprise-wide data warehouses. Many people feel that with competition mounting in every industry, data warehousing is the latest must a way to keep customers by learning more about their needs.

According to W. H. Inmon, a leading architect in the construction of data warehouse systems, "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process". This short, but comprehensive definition presents the major features of a data warehouse. The four keywords, subject-oriented, integrated, time-variant, and nonvolatile, distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems. Let's take a closer look at each of these key features.

Subject-oriented; a data warehouse is organized around major subjects, such as customer, supplier, product, and sales. Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers. Hence, data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated; a data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

Time-variant; data are stored to provide information from a historical perspective (e.g., the past 5-10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

Nonvolatile; a data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

In sum, a data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information

on which an enterprise needs to make strategic decisions. A data warehouse is also often viewed as architecture, constructed by integrating data from multiple heterogeneous sources to support structured and/or ad hoc queries, analytical reporting, and decision making.

2.1.3.1 Differences between Operational Database Systems and Data Warehouses

The major task of on-line operational database systems is to perform on-line transaction and query processing. These systems are called on-line transaction processing (OLTP) systems. They cover most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting. Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are known as on-line analytical processing (OLAP) systems.

The major distinguishing features between OLTP and OLAP are summarized as follows.

2.1.3.2 Users and System Orientation

An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.

2.1.3.3 Data Contents

An OLTP system manages current data that, typically, are too detailed to be easily used for decision making. An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier to use in informed decision making.

2.1.3.4 Database Design

An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design. An OLAP system typically adopts either a star or snowflake model and a subject-oriented database design.

2.1.3.5 View

An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

2.1.3.6 Access Patterns

The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read only operations (since most data warehouses store historical rather than up-to-date information), although many could be complex queries.

Other features that distinguish between OLTP and OLAP systems include database size, frequency of operations, and performance metrics.

2.1.3.7 A Three-Tier Data Warehouse Architecture

Data warehouses often adopt a three-tier architecture, as a presented in Figure 2.3

1. The bottom tier is a warehouse database server that is almost always a relational database system. "How are the data extracted from this tier in order to create the data warehouse?" Data from operational databases and external sources (such as customer profile information provided by external consultants) are extracted using application program interfaces known as gateways. A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server. Examples of gateways include ODBC (Open Database Connection) and OLE-DB (Open Linking and Embedding for Databases), by Microsoft, and JDBC (Java Database Connection).

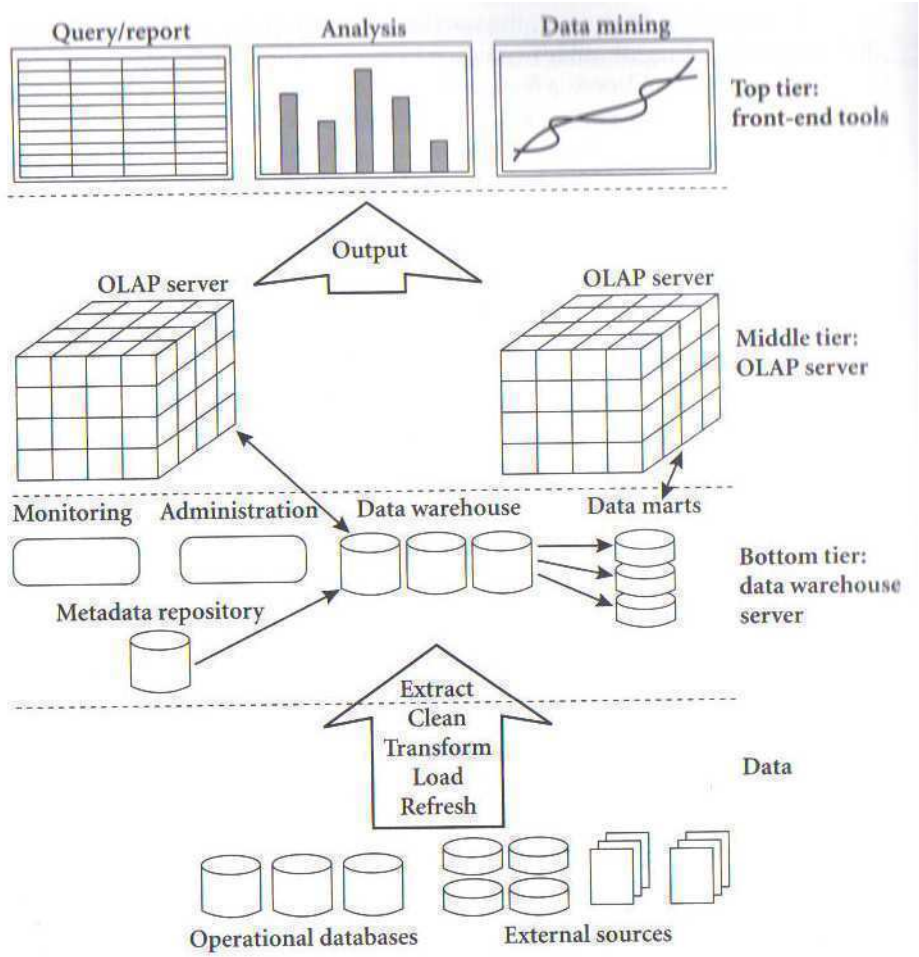


Figure 2.3 A three-tier data warehousing architecture (Han, Kamber, 2001)

2. The middle tier is an OLAP server that is typically implemented using either a relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or a multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

3. The top tier is a client, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

From the architecture point of view, there are three data warehouse models: the enterprise warehouse, the data mart, and the virtual warehouse.

Enterprise warehouse; an enterprise warehouse collects all of the information about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and

is cross-functional in scope. It typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond. An enterprise data warehouse may be implemented on traditional mainframes, UNIX super servers, or parallel architecture platforms. It requires extensive business modeling and may take years to design and build.

Data mart; a data mart contains a subset of corporate-wide data that is of value to a specific group of users. The scope is confined to specific selected subjects. For example, a marketing data mart may confine its subjects to customer, item, and sales. The data contained in data marts tend to be summarized.

Data marts are usually implemented on low-cost departmental servers that are UNIX or Windows/NT based. The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years. However, it may involve complex integration in the long run if its design and planning were not enterprise-wide.

Depending on the source of data, data marts can be categorized as independent or dependent. Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area. Dependent data marts are sourced directly from enterprise data warehouses.

Virtual warehouse; a virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized. A virtual warehouse is easy to build but requires excess capacity on operational database servers.

The top-down development of an enterprise warehouse serves as a systematic solution and minimizes integration problems. However, it is expensive, takes a long time to develop, and lacks flexibility due to the difficulty in achieving consistency and consensus for a common data model for the entire organization. The bottom-up approach to the design, development, and deployment of independent data marts provides flexibility, low cost, and rapid return of investment. It, however, can lead to problems when integrating various disparate data marts into a consistent enterprise data warehouse.

2.1.4 Data Preprocessing

There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube. Data transformations, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. These data processing techniques, when applied prior to mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining.

Figure 2.4 summarizes the data preprocessing steps described here. Note that the above categorization is not mutually exclusive. For example, the removal of redundant data may be seen as a form of data cleaning, as well as data reduction.

In summary, real-world-data tend to be dirty, incomplete, and inconsistent. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the knowledge discovery process, since quality decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge payoffs for decision making.

2.1.4.1 Data Cleaning

Real-world data tend to be incomplete, noisy, and inconsistent. *Data cleaning* routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Basic methods for data cleaning are at the following methods.

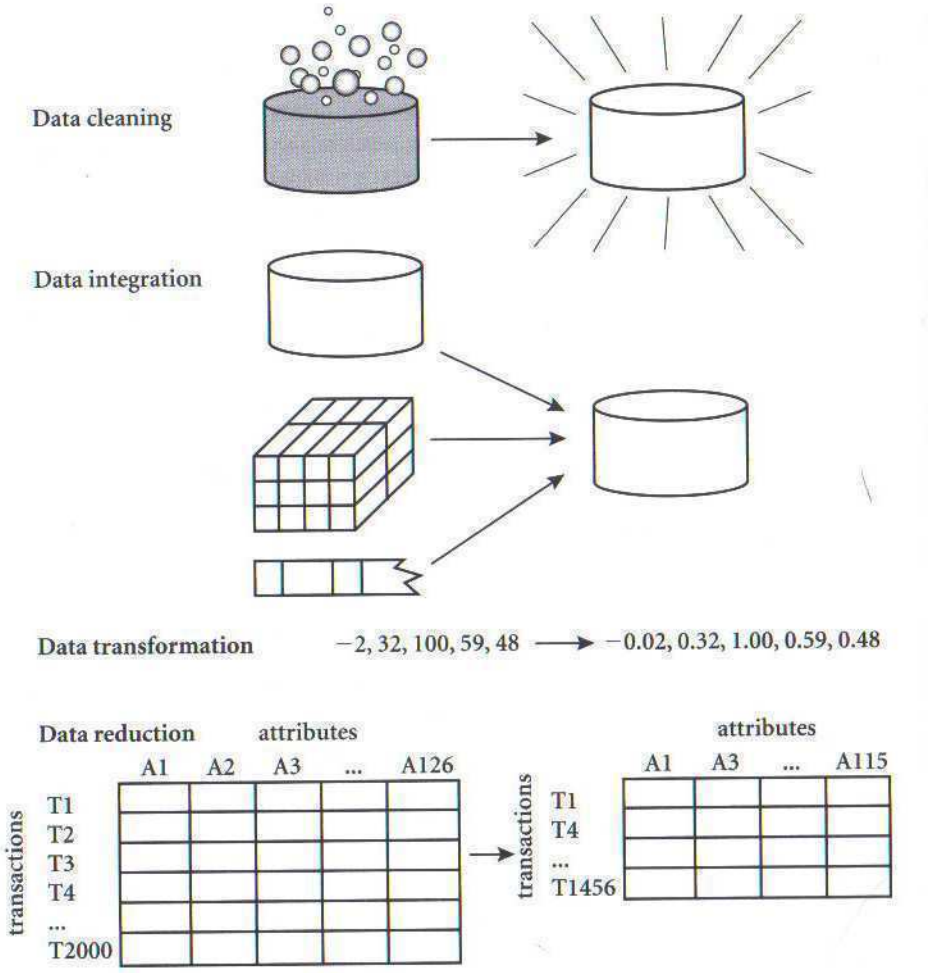


Figure 2.4 Forms of data preprocessing (Han,Kamber, 2001)

A. Missing Values.

1. Ignore the cell: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective, unless the cell contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
2. Fill in the missing value manually: In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.
3. Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like "Unknown" or ∞. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly

think that they form an interesting concept, since they all have a value in common—that of "Unknown." Hence, although this method is simple, it is not recommended.

4. Use the attribute mean to fill in the missing value
5. Use the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.

B. Noisy data. Noise is a random error or variance in a measured variable. Some attributes values might be invalid or incorrect. These values are often corrected before running data mining applications. (Dunham, 2002)

C. Inconsistent Data. There may be inconsistencies in the data recorded for some transactions. Some data inconsistencies may be corrected manually using external references. For example, errors made at data entry may be corrected by performing a paper trace. This may be coupled with routines designed to help correct the inconsistent use of codes. Knowledge engineering tools may also be used to detect the violation of known data constraints. For example, known functional dependencies between attributes can be used to find values contradicting the functional constraints.

There may also be inconsistencies due to data integration, where a given attribute can have different names in different databases. Redundancies may also exist.(Han, Kamber, 2001)

2.1.4.2 Data Integration and Transformation

Data mining often requires data integration. The data may also need to be transformed into forms appropriate for mining.

A. Data Integration. It is likely that your data analysis task will involve data integration, which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. There are a number of issues to consider during data integration. Schema integration can be tricky. How can equivalent real-world entities from multiple data sources be matched up? This is referred to as the entity identification problem.

For example, how can the data analyst or the computer be sure that customer-_id in one database and cust_number in another refer to the same entity? Databases and data warehouses typically have metadata—that is, data about the data. Such metadata can be used to help avoid errors in schema integration.

B. Data Transformation. Data from different sources must be converted into a common format for processing. Some data may be encoded or transformed into more usable formats. Data reduction may be used to reduce the number of possible data values being considered (Dunham, 2002).

2.1.5 Data Mining Functionalities

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform Inference on the current data in order to make predictions.

In some cases, users may have no idea which kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations or applications. Furthermore, data mining systems should be able to discover patterns at various granularities. Data mining systems should also allow users to specify hints to guide or focus the search for interesting patterns. Since some patterns may not hold for all of the data in the database, a measure of certainty or trustworthiness is usually associated with each discovered pattern.

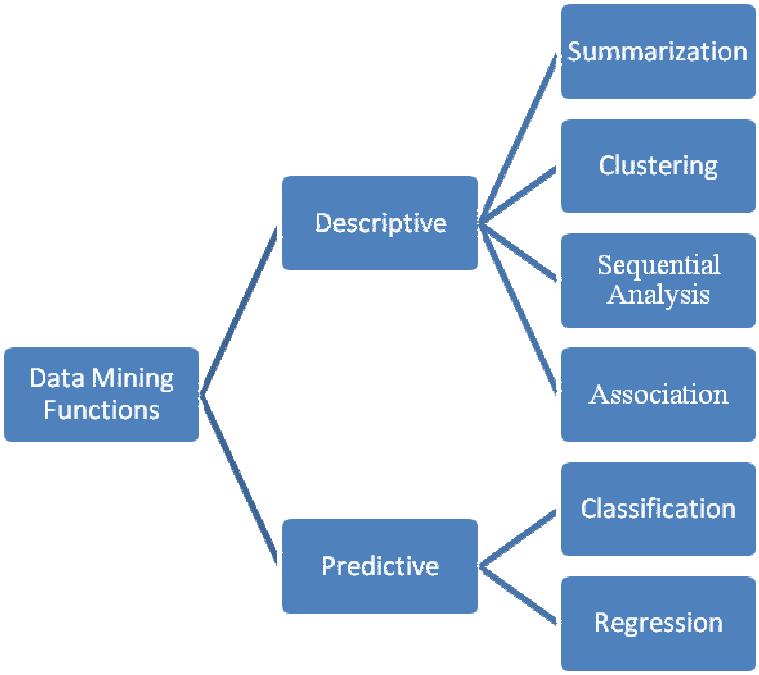


Figure 2.5 Data Mining Functions

2.1.5.1 Classification

This is the data analysis method that shows the important data classifications. The classification estimates the categorical data. The techniques that are used in the classification are these:

1. Decision Trees
2. Artificial Neural Networks (Han, 2001)

Decision trees are the most common ones in the data mining since their establishment do not cost much, their interpretation is easy, their data base system can be easily integrated and they have reliability.

The decision tree, as we can understand from the name, has an appearance of a tree and is an estimation technique.

With its tree structure it creates easily understood rules and can be integrated easily with the knowledge technology performances. It is the most popular classification technique.

Decision tree has decision knots, branches and leaves. The decision knot explains the test that is to be done. The result of this test makes the tree turn into branches without losing data if every know, test and turning into branches event occurs

consecutively and this procedure depends on the high level separations. Each branch of the tree is to complete the classification. If the classification does not occur on one edge of a branch, a decision knot appears. However, if a certain classification happens in the end of the procedure, there is a leaf on the edge of that branch. This leaf is one of the classes that are to be explained on the data. Decision tree procedure starts from the origin knot, and happens by following the consecutive knots from up to down.

For example, an education data is being examined and a pattern of estimating its class is being created.

The classification rule that makes this pattern

IF age = "41...50" AND income = high THEN credit situation = excellent

We can see that a person who is between and 41-50 and has a high income has an excellent credit situation.

After approving the correction of this pattern by a test data, the pattern can be carried out on a data whose class is uncertain and the classification of the new data can be explained excellently

2.1.5.2 Regression

The regression analysis is a technique which categorizes the relation between two (or more) variables which have a cause and effect relation with a mathematical regression pattern and is used to make estimations and predictions about that subject. After adapting the regression pattern, the inspection that explains whether the pattern is sufficient is the most important part of the regression analysis. To guarantee that the adapted pattern is close enough to the correct pattern and to be certain whether the regression analysis provides all the estimations is a must. If the regression pattern does not adapted sufficiently, it will give weak and incorrect results. It is not only the variations analysis to find out the pattern's sufficiency in the regression analysis but there are also more informative tests in addition to R^2, but not often used. (Şahinler, 2000)

In the situations that have two or more levels of the dependent variable, since the normality estimation breaks down, the logistic regression analysis is used as an alternative to the linear regression analysis. The logistic regression is an alternative

to the contingency table and discriminant analysis since it has a binary dependent variables (0,1) , does not have a normal distribution and a common covariance and it causes several estimations to break down.

Analyses since the normality estimation breaks down in the situations of dependent variables are two or more. It is estimated that the errors have a binom distribution in the logistic regression.

For the observation couple x, y which are independent from each other and are n number, the independent variable that is showed by X can be categorical or constant. $(i = 1,2,\ldots,n)$.

To add the categorizing and ordering independent variable to the pattern the design variables should be used.

The function that gives the linear relation between the dependent and independent variables is called the link function. In the linear regression the link function size is (I) but in the logistic regression it is logit or probit transformation. So, for the logistic regression pattern with only one variable, β_1 coefficient means the change that occurs in the logit when there is one unit increase in the X. The simple logistic regression pattern and $\pi(x_i)$ logit transformation is given in order in the (1) and (2) equalities.

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \frac{\exp(g(x_i))}{1 + \exp(g(x_i))} \tag{1}$$

$$g(x_i) = \ln\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] = \ln\left\{\frac{\frac{\exp(\beta_0 + \beta_1 x_i)}{(1 + \exp(\beta_0 + \beta_1 x_i))}}{\frac{1}{(1 + \exp(\beta_0 + \beta_1 x_i))}}\right\} = \ln\left(e^{\beta_0 + \beta_p x_i}\right) = \beta_0 + \beta_1 x_i \tag{2}$$

If the categorizing and ordering independent variable has a k category then k-1 design variable is used.

If the x which is the independent variable for the simple logistic regression pattern has a k category, k-1 design variable is D, and the coefficient is $u = 1,2,\ldots,k-1$. For the p independent variable pattern, the logit is in the third equality. ($i = 1,2,\ldots,p$)

$$g(x) = \beta_0 + \sum_{i=1}^{p} \sum_{u=1}^{k-1} \beta_{pu} D_{pu} \tag{3}$$

The prediction of the unknown parameters in the (1) equality logistic regression pattern can be made with the possibility method. This method gives the values of the unknown parameters which create the possibility of obtaining the observed data units maximum. For this, making up the function of the possibility is a must when the result variable is equal to 1, its contribution to the function of the possibility $\pi(x_i)$, when it's equal to 0, its contribution to the function of the possibility $1-\pi(x_i)$. In the observation couple (x_i, y_i), the independent possibility function's observation couples are explained by multiplicating and it is in the (4) equality.

$$L(\beta_0,\beta_1)=\prod_{i=1}^{n} \pi(x_i)^{y_i} \left(1-\pi(x_i)\right)^{1-y_i} \tag{4}$$

The target in the possibility method is to make the β make the (4) equality maximum. To find the β values which make this equality maximum -by having the derivative for $\ln L(\beta)$ according to β (β_0,β_1)- we can make these two possibility equalities equal to 0. The (β_0,β_1) values that are obtained from these equalities are named as the possibility estimation and showed as $(\hat{\beta}_0,\hat{\beta}_1)$. The estimation method of these variance and covariance of the coefficient are obtained from the matrix of the log-possibility functions' second level derivatives. However, it takes a long period of time so the package program are useful here.(Lemeshow, Hosmer, 1989).

After the estimation of the derivatives, the importance of the variables of the pattern is observed. The importance tests of the derivatives help create the best pattern with the least variables. The importance of the derivatives can be explained by these tests Likelihood Ratio Test, Wald test, Score test. The point is to find out whether the full model which has the variable to be observed has more knowledge than the model which does not include the variable. This point can be clear by comparing the result variable's observed values with the estimated values that are taken from the two models. If the estimated values of the variable model are better than the model which does not include the variable, then we can understand that the model we are examining is important. The possibility rate test is done by the G statistics.

$$G = -2\ln\left[\frac{\text{without the interaction term}}{\text{with an interaction term}}\right] \quad (5)$$

Under the hypothesis of $H_0 : \beta_1 = 0,\ H_1 : \beta_1 \neq 0$ n, G statistics has one degree of freedom level chi-square distribution. We reject the hypothesis if the G value is bigger than the chi-square value. In another way, we can check the log possibility value of the model which has and which does not have the variable that is observed. Adding the observed variable to the model causes an increase in the log possibility value. Log possibility rate test is equal to (-2) multiplication of the difference between the log possibility values of the two models. In the multi logistic regression model, the G statistics degree of freedom shows chi-square distribution under the hypothesis of the p education derivative is equal to 0. $\nu_2 = 1$ more than the variable number in all the model $\nu_1 = 1$ more than the reduced model's variable number $(\nu_2 - \nu_1)$: if the error possibility is more than 0.05, the reduced models is as good as the whole model.

The Wald test is done by comparing the estimation of the $\hat{\beta}_1$ parameter's possibility and its standard error's estimation. The rate that is gained shows a standard distribution under the (W), $H_0 : \beta_1 = 0$ hypothesis.
Wald statistics distributes chi-square and calculated by

$$W = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \quad (6)$$

In the logistic regression, interpretations of the derivatives are made by the odds and the odds' rate (OR). The differences are the logit's natural case which no logarithms' are taken. It is the $\ln(\pi(x)/(1-\pi(x)))$ logit value. The difference rate is the x=1 for the x=0 difference value. The natural logarithm of the difference rates are expressed as log odds rate or log odds and this is equal to the logit difference. According to this , the independent variable's being 2 – which means it is coded as 0 and 1- the difference rate is expressed as $OR = \exp(\beta_1)$ .The odds rate can take any value between 0 and ∞. The odds rate gives the 1 unit change of the x which has the possibility of being Y=1. If the odds rate is equal to 1, the calculated effect is equal to 0. If the odds rate is bigger than 1, like 1.3, then the 1 unit change in x

creates the possibility for Y to be 1 (Y=1) 0.3 times. In other words, if the odds rate is less than 1, like 0.7, then the effect of x on Y is negative. The one unit change in x decreases the possibility of Y to be equal to 1 1-0.7=0.3 times. (Şahin, 1999).

When the independent variable number is small it is easy to create a model. However, the more variables there are, the more difficult it is to choose. If there are more variables in the model, the estimated standard error increases and it becomes more dependent to the observed data unit. There are some methods in the logistic regression to choose the variable. These are Univariate Analysis (the test of the derivatives one by one), Stepwise Logistic Regression (step by step logistic regression) and Best Subsets. If the variable number is many, we can apply for the logistic regression analysis. This analysis is Forward Selection and Backward Elimination. Being included in the model or not is decided by Wald values and possibility rate values. The variable which has a big Wald value has a small p value and it's meaningful for the model. The example about the logistic regression is in chapter 3.

2.1.5.3 Association Rules

The association rules find the relation among the big data sets. Since the collected data is more and more every day, the companies want to use the association rules of the data base. Discovering the interesting association relations makes the companies' decision period more efficient. The most common example of the association rules is the market basket rule (or the trolley).This procedure analyses the shopping habits of the buyers (Han, 1999).

Discovering these kinds of associations shows us which products that the buyers buy and the market managers can improve better marketing strategies with this knowledge. For example is buyer buys milk, what is the percentage of the possibility for him to buy bread near milk? The market managers who design the shelves according to this knowledge can increase the sale rate. If the buyers' buying milk right after bread possibility is high, the market managers can increase the sale of bread by putting the two shelves together.

For example if the buyers who buy A product also attempt to buy B product, this can be showed by the association rule in (7).

A => B [support = %2, reliability = %60] (7)

These support and reliability expressions are the eccentricity measurement of the rules. They show the suitability of the rule. For the association rule in (7), 2% support shows that A and B products are sold at the same time in 2% of all the shopping the percentage of 60 shows that the 60% buyers who buy a product also buy the B product at the same shopping. The minimum support and reliability level is observed and the association rules that outdo these levels are taken into consideration (Zaki, 1999).

2.1.5.4 Sequence Analysis

Sequential analysis or sequence discovery is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. These patterns are similar to associations in that data (or events) are found to be related, but the relationship is based on time. Unlike a market basket analysis, which requires the items to be purchased at the same time, in sequence discovery the items are purchased over time in some order. Example illustrates the discovery of some simple patterns. A similar type of discovery can be seen in the sequence within which data are purchased. For example, most people who purchase CD players may be found to purchase CDs within one week. As we will see, temporal association rules really fall into this category.(Dunham, 2002)

For example; the Webmaster at the XYZ Corp. periodically analyzes the Web log data to determine how users of the XYZ's Web pages access them. He is interested in determining what sequences of pages are frequently accessed. He determines that 70 percent of the users of page *A* follow one of the following patterns of behavior: *(A, B, C)* or *(A, D, B, C)* or *(A, E, B, C)*. He then determines to add a link directly from page *A* to page *C*.

2.1.5.5 Summarization

Summarization maps data into subsets with associated simple descriptions. Summarization is also called characterization or generalization. It extracts or derives representative information about the database. This may be accomplished by actually retrieving portions of the data. Alternatively, summary type information (such as the mean of some numeric attribute) can be derived from the data. The summarization succinctly characterizes the contents of the database. Example illustrates this process.(Dunham, 2002)

For example; one of the many criteria used to compare universities by the U.S. News & World Report is the average SAT or ACT score [GM99]. This is a summarization used to estimate the type and intellectual level of the student body.

2.1.5.6 Clustering

Clustering is categorizing the data into classes. (Karypis, Han, ve Kumar, 1999). While the elements are similar to each other in one set, other sets' elements are different. In the clustering model there are no data classes which are in the classifying model. (Ramkumar, Swami, 1998). The data has no certain class. In the classifying model we can observe which class the data belongs to. However , in the clustering model there are data which has no certain classes. In some procedures the clustering model is a pre-period of the classifying model. (Ramkumar, Swami, 1998).

Discovering the buyers' classes in markets , classifying the animals and plants in biology , classifying the similar gens , classifying the buildings in a city planning according to their usage are typical clustering performances. Clustering can also be used on the web sites to classify the documents. (Seidman, 2001).Clustering the data has been improving. As the data increases on the data base, clustering analysis has had an important role in data mining research. There are a lot of clustering algorithm in sources. The clustering algorithm depends on the type of the data and the target. The main clustering methods are (Han, Kamber, 2001):

1. Partitioning methods (K-means method)
2. Hierarchical methods

In the partitioning methods n is the number of the items on the data base and k is the number of the sets that are to be made. The partitioning algorithm divides the n number item into k number sets. *(k _ n).* The items in one set are similar and different from the other sets' items. .(Han, Kamber, 2001)

K-means method chooses k number of items from the n number of items and all these items symbolizes one sets' center. The other items are divided into sets according to the center of the closest set. This means, one item goes to the set whose center is closest to it. Then an average is calculated and this average becomes the new center of that set. This procedure carries on until all the items go to their sets. (Han, Kamber, 2001).

Suppose that an item group is located in the space as in 2.6. When the user wants to divide these items into two sets, it is *k=2* (Han, Kamber, 2001)
2.6 show that First, the two items are chosen as the centers of the two sets and the other items are in the other sets -following the closest set center- With this procedure , the average of the items of the two sets is made and it becomes the new centers of these value sets. These new centers are showed in 2.6(b) with a cross. According to these new crossed centers, in every set one item becomes close to the other sets center 2.6(c).

The item that is coordinated (5,1) and the item that is coordinated (5,5) change their sets. With this addition the average and the centers of the sets change. (Han, Kamber, 2001).

The new calculated centers are in 2.6(d). K-means method is over now because there is no item left. And every item is the closest to its set's center.

K-means method can only be used when the average of the set is identified. The need to explain the number of the sets could sound like a disadvantage. But the more important disadvantage is the sensibility to the outliers. (Han, Kamber, 2001).

An item that has a big value can change the average of the set that it is in. This change could destroy the sensitivity of the set. The hierarchical methods depend on putting the data items in the set trees. The hierarchical methods can be classified from up to down or down to up (agglomerative and divisive hierarchical clustering). In agglomerative hierarchical clustering, as it is seen in 2.7 the hierarchical division is from down to up. First, every item makes its own set then these atomic sets gather and make bigger sets until every item is in one set.

In devise hierarchical clustering (2.7), the hierarchical division is from up to down. First, all the items are in one set then all the sets are divided into little pieces until every item is a set itself. (2.7)

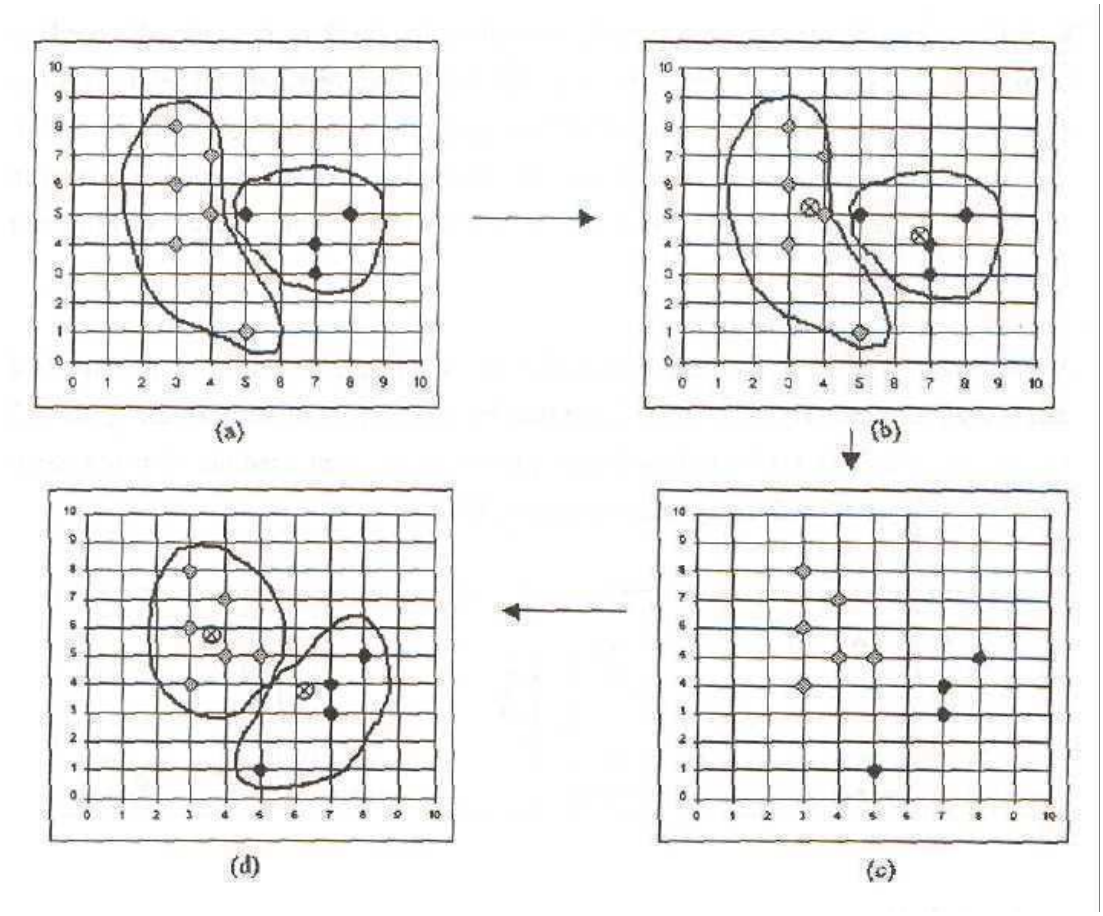


Figure 2.6 K-means

Agglomerative (AGNES)

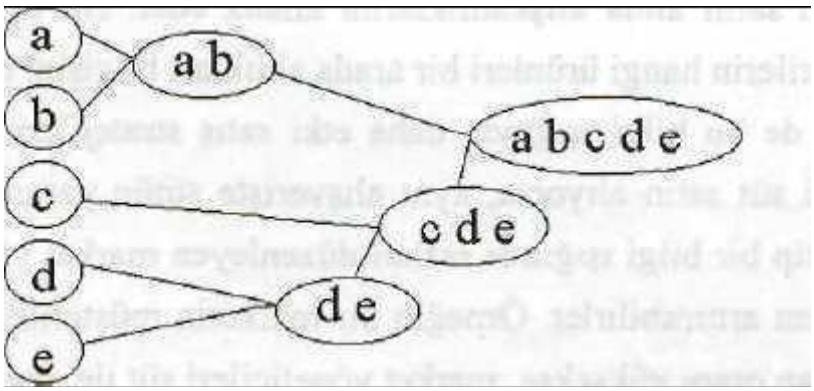


Figure 2.7 Hierarchical clustering

Divisive (DIANA)

In 2.7, you can see AGNES (AGlomerative NESting) and DIANA (Dlvise ANAlysis) methods. These methods are used on a five-item set (a, b, c, d, e). First, AGNES puts every item in a set. And the sets come together and attach. For example C1 and C2 set. If the certain distance is the same as any other items in any other sets, they can attach. This gathering procedure continues until every item is in one set. Han, Kamber, 2001). In DIANA, all the sets are divided into little pieces until every item is a set itself. The example about clustering is given in Chapter 3.

2.2 Medical Data Mining

Modern medicine generates huge amounts of data, but at the same time there is an acute and widening gap between data collection and data comprehension. Thus, there is growing pressure not only to find better methods of data analysis, but also need for automating them to facilitate the creation of knowledge that can be used for clinical decision-making. This is where data mining and knowledge discovery tools come into play, since they can help in achieving these goals.

2.2.1 Unique Features of Medical Data Mining

Data mining and knowledge discovery in medical databases are not substantially different from mining in other types of databases. There are some characteristic features, however, that are absent in non-medical data.

One feature is that more and more medical procedures employ imaging as a preferred diagnostic tool. Thus, there is a need to develop methods for efficient mining in databases of images, which is noy only different, but also more difficult, than mining in numerical databases. As an example, imaging techniques like SPECT, MRI, PET, and collection of ECG (Figure 2.8) or EEG signals, can generate gigabytes of data per day. A single cardiac SPECT procedure of one patient may contain dozens of 2D images. In addition, medical databases are always heterogeneous. For instance, an image of the patient's organ will almost always be accompanied by other clinical information, as well as the physician's interpretation (clinical impression; diagnosis). This heterogeneity requires high capacity data

storage devices, as well as new tools to analyze such data. It is obviously very difficult for an unaided human to process gigabytes of records, although dealing with images is relatively easier for humans, who are able to recognized patterns, grasp basic trends in data, and form rational decisions. The information becomes less useful as we are faced with difficulties of retrieving it, and making it available in an easily comprehensible format. Visualization techniques will play an increasing role in this setting, since images are the easiest for humans to comprehend, and they can provide a great deal of information in a single visualization of the results(Giudici, 2003).

A second feature is that the physician's interpretation of images, signals, or any other clinical data, is written in unstructured free-text English that is very difficult to

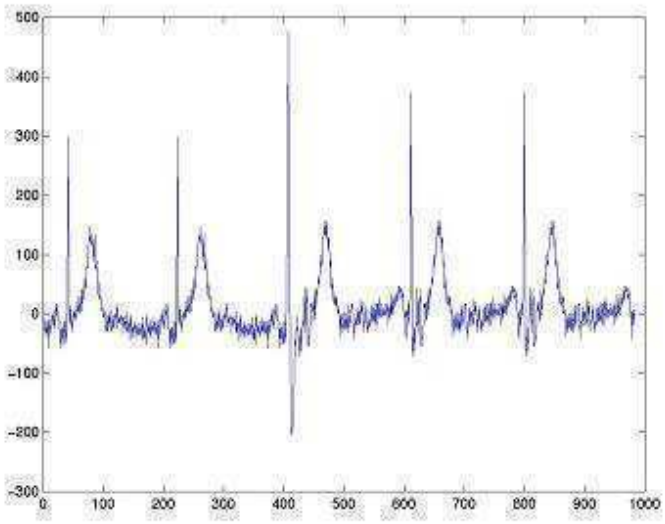


Figure 2.8 Electrocardiographic

Standardize and thus difficult to mine. Even specialists from the same discipline cannot agree on unambiguous terms to be used in describing a patient's condition. Not only do they use different names (synonyms) to describe the same disease, but they render the tas keven more daunting by using different grammatical constructions to describe relationships among medical entities. For example; (Cios, Moore, 2002)

Example: chest pain, radiates to left arm

CP ---- > L arm

Chest pressure with radiation to left arm

Chest pressure radiating to arm

A third unique feature of medical data mining is the question of data ownership. The corpus of medical data potentially available for mining is enormous. Some thousands of terabytes (quadrillions of bytes) are now generated annually in North America and Europe. However, these data are buried in heterogeneous databases, and scattered throughout the medical care establishment, without any common format or principles of organization. The questions of ownership of patient information is unsettled, and the object of recurrent, highly-publicized lawsuits and congressional inquiries. Do individual patients own data collected on themselves? Do their physicians own the data? Do their insurance providers own the data? Some HMOs now refuse to pay for patient participation in clinical treatment protocols that are deemed experimental. If insurance providers do not own their insurees' data, can they refuse to pay for the collection and storage of data? Some might argue that the ownership of human data, and therefore the ability to process and sell such data, is unseemly. If so, then how should the data managers, who organize and mine the data, be compensated? Or should this incredibly rich resource for the potential betterment of humankind be left unmined?

A fourth unique feature of medical data mining is a fear of lawsuits directed against physicians and other medical providers. Medical care in the U.S.A., for those who can afford it, is the best in world. However, U.S. medical care is some %30 more expensive than that elsewhere in North America and Europe, where quality is comparable; and U.S. medicine also has the most litigious malpractice climate in the world. Some have argued that the %30 surcharge on U.S. medical care, about one thousand dollars per capita annually, is mostly medico legal: either direct legal costs, or else the overhead of "defensive medicine", i.e., unnecessary tests ordered by physicians to cover themselves in potential future lawsuits. In this tense climate, physicians and other medical data-producers are understandably reluctant to hand over their data to data miners. Data miners could browse these data for untoward events. Apparent anomalies in the medical history of an individual patient might trigger an investigation. In many cases, the appearance of malpractice might be a data-omission or data-transcription error; and not all bad outcomes in medicine are necessarily the result of negligent provider behavior. However, an investigation inevitably consumes the time and emotional energy of medical providers. For exposing themselves to this risk, what reward to the providers receive in return?

A fifth feature is privacy band security concerns. For instance, pending U.S. Federal rules suggest that there should be unrestricted usage of medical data patients deceased for at least two years; but for live patients, all data must be encrypted, so that it impossible to identify a person, or to go back and decrypt the data. At stake is not only a potential breach of patient confidentiality, with the possibility of ensuing legal action; but also erosion of the physician-patient-relationship, in which the patient is extraordinarily candid with the physician in the expectation that such private information will never be made public. Thus, the encryption of data must be irreversible. A related privacy issue may apply if, for example, crucial diagnostic information were to be discovered on live-patient data and that a patient could be treated if we could only go back and inform the patient about the diagnosis and possible cure. According to the Health Insurance Portability and Accountability Act of 1996(HIPAA) legislation, this is unfortunately not possible, another issue is data security in data handling, and particularly in data transfer. Before the data are encrypted, only authorized persons should have access to the data. Since transferring the data electronically via the Internet is insecure, the data must be carefully encrypted even for transfers within one medical institution from one unit to another.

A sixth unique feature of medical data mining is that the underlying data structures of medicine are poorly characterized mathematically, as compared to many areas of the physical sciences. Physical scientists collect data which they can substitute into formulas, equations, and models that reasonably reflect the relationships among their data. On the other hand, the conceptual structure of medicine consists of word-descriptions and pictures, with very gew formal constraints on the vocabulary, the composition of images, or the allowable relationships among basic concepts. The fundamental entities of medicine, such as inflammation, ischemia, or neoplasia, are just as real to a physician as entities such as mass, length, or force are to a physical scientist; but medicine has no comparable formal structure into which a data miner can organize information, such as might be modeled by clustering, regression models, or sequence analysis. In its defense, medicine must contend with hundreds of distinct anatomic locations and thousands of diseases. Until now, the sheer magnitude of this concept space was insurmountable. Furthermore, there is some suggestion that be logic of medicine may

be fundamentally different from the logic of the physical sciences. However, it may now happen that fast computers ant the newer tools of data mining and knowledge discovery may overcome this prior obstacle.

Finally, human medicine is primarily a patient-care activity, and only secondarily a research resource. Generally the only justification for collecting data in medicine, or refusal to collect certain data, is to benefit the individual patient. Some patients might consent to be involved in research projects that do not benefit them directly, but such data collection is typically very small-scale, narrowly focused, and highly regulated by legal and ethical considerations. On the other hand, humans are the most closely-studied species in the world, and enormous quantities of data are generated as an incidental byproduct of patient care. In Europe and North America, much of this information is now primarily generated on electronic media. These data include observations that cannot be gained from animal studies, such as visual and auditory sensations, the perception of pain, and recollection of possibly relevant prior traumas and exposures. By contrast, most animal studies are short-term, and therefore cannot track long-term disease processes of medical interest, such as preneoplasia or atherosclerosis. Furth more, most animal studies are small-scale, and thus cannot detect or follow rare disease processes. And, there is no issue of having to extrapolate animal observations to the human species.

In summary, medical data are heavily image-based and text-based; the datasets are heterogeneous, replete with missing values; medicine lacks a formal, mathematical structure for organizing data; and there are significant issues of ownership, fairness, and privacy. Yet for all this, the unique promise of medical data mining is enormous and rich; and its potential for use in human betterment is unequalled. (Cios, 2001)

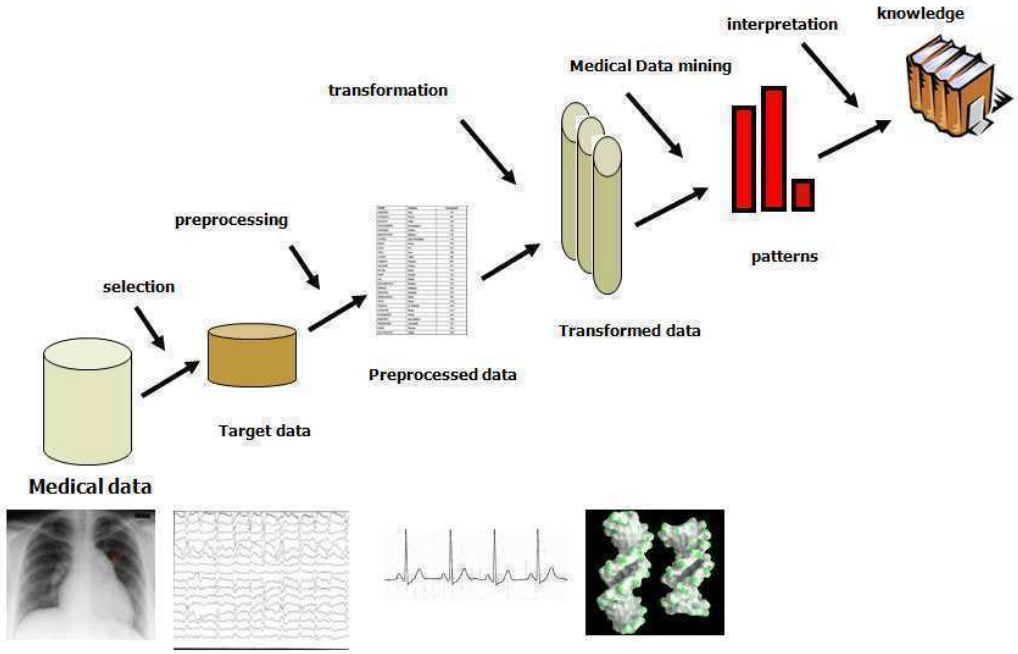


Figure 2.9 Medical knowledge discovery process

2.2.2 Medical knowledge Discovery Process

When we compose figure 2.1 which shows data mining process for medical data mining a process as in figure 2.9 is obtained.

STEP 1: For step one; our mission is to describe our goal and to define which database to use to get/receive the goal.

STEP 2: Let's suppose that we have millions of medical data. To be able to realize the necessary data steps we need to choose a subgroup of a database. To choose is to define enough of data to represent the database

STEP 3: Data preprocessing and data cleaning. In this step we try to eliminate noise that is present in the data. Noise can be defined as some form of error within the data. Some of the tools used here can be used for filling missing values and elimination of duplicates in the database

STEP 4: Transformation of data in this step can be defined as decreasing the dimensionality of the data that is sent for data mining. Usually there are cases where there are a high number of attributes in the database for a particular case. With the reduction of dimensionality we increase the efficiency of the data-mining step with respect to the accuracy and time utilization.

STEP 5: This is when the cleaned and preprocessed data is sent into the intelligent algorithms for classification, clustering, similarity search within the data, and so on. Here we chose the algorithms that are suitable for discovering patterns in the data. Some of the algorithms provide better accuracy in terms of knowledge discovery than others. Thus selecting the right algorithms can be crucial at this point.

STEP 6: Interpretation. In this step the mined data is presented to the end user in a Human - viewable format. This involves data visualization, which the user interprets and understands the discovered knowledge obtained by the algorithms.(Eapen, 2004)

CHAPTER THREE

APPLICATIONS OF MEDICAL DATA MINING

This chapter includes practices regarding clustering and logistic regression subjects that are one of the statistical methods of data mining mentioned in Chapter two.

3.1 Application of EEG Data

In the first work, the electroensefalography (EEG) data, which was obtained by Dokuz Eylül University Medicine Faculty Biophysics department laboratory, is examined. EEG is the method which measures the observation of the brain waves' activity with an electrical system. Since there is no electricity given to the patient, there is no pain or ache.

The record which is had by electroensefalography is named EEG. This technique is developed by Hans Berger in 1929. The recording in EEG is made by putting little electrots into the hairy head skin with a conductive element named "the cake". The electrical potential changes between the two electrots are recorded by the computer. The result is checked by the expert and the patient is informed. In the inspection of the record, many disorders of the brain can be diagnosed. (Epilepsy etc.)

Human nerve system includes 10 billion nerve cells. Most of them are in the brain and in the spine and the others parts of the body. Every brain cell is connected to 5.000-50.000 nerve cells. The nerve current is carried in the nerve fibers and creates waves in the brain. These waves can be tested on the head skin.

The target of this work is to solve the EEG structure by using the clustering and statistical methods. The data is gained by the work that was done by the Excel system in the Dokuz Eylül University Medicine Faculty Biophysics department laboratory computers.

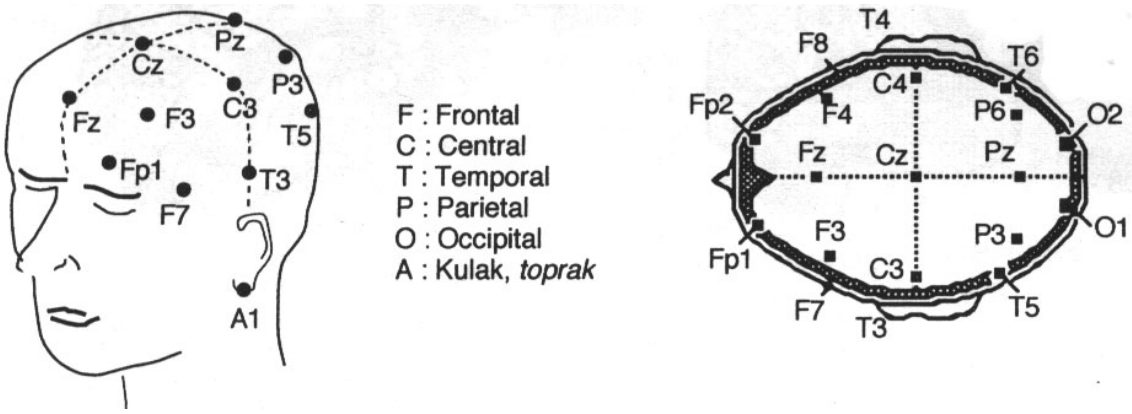


Figure 3.1 The electrot locations in EEG (Dasdag, n.d.)

Five persons' data structure is examined. Every person's data EEG record is made according to the 26 pre-stimulus and post-stimulus sweeps. These records are put into the computer from the Excel file. Then they are put into the MATLAB program to be organized and program. In this transformation the data is named. First the 5 person's names are coded: A, B, C, D, E. Every person has two data as pre-stimulus and post-stimulus. Pre-stimulus codes: A1, B1, C1, D1, E1. Post-stimulus codes: A1a, B1a, C1a, D1a, E1a. The measurement is repeated 26 times. Before giving a stimulus to one person 26 measurements are done. After giving the stimulus 26 measurement are recorded. Before and after every stimulus there is 500 ms-time. For example, for the A coded person the first 500 ms-time is recorded pre stimulus then by giving a stimulus sound there is 500 ms more time recorded. So the first part is completed. Then the same thing is done for the second part. There is no pausing between the parts. The record starts and continues until the end of the 500th ms of the 26th part. In table 1 you can see the 6 parts of the first 9 ms of the test of A1 pre stimulus. And also in table 2 you can see. The 6 parts of the first 9 ms of the test of A1 post-stimulus.

Table 3.1 6 parts of the first 9 ms of the test of A1 pre stimulus

	Sweep 1	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6
1. ms	-0,08679	18,65915	-10,5012	11,71621	-1,21501	0
2. ms	0,78108	19,17986	-10,3276	11,02191	-1,90931	-0,17357
3. ms	1,562161	19,70059	-10,154	10,24083	-2,69039	-0,34715
4. ms	2,343241	20,22131	-9,8069	9,546539	-3,47147	-0,52072
5. ms	3,037535	20,48167	-9,37297	8,852245	-4,16576	-0,69429
6. ms	3,645042	20,48167	-8,93903	8,244739	-4,86006	-0,69429
7. ms	4,165762	20,22131	-8,41831	7,810805	-5,46756	-0,60751
8. ms	4,512909	19,70059	-7,98438	7,550445	-5,98828	-0,43393
9. ms	4,599696	19,09308	-7,72402	7,550445	-6,509	-0,26036

Table 3.2 6 parts of the first 9 ms of the test of A1 post-stimulus

	Sweep 1	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6
1. ms	1,822521	-8,76546	15,44804	0	-1,47537	-0,43393
2. ms	1,735734	-9,19939	15,79518	0,347147	-1,64895	-0,95465
3. ms	1,648948	-9,63333	15,96876	0,694294	-1,56216	-1,64895
4. ms	1,735734	-9,98047	15,88197	1,041441	-1,38859	-2,34324
5. ms	1,822521	-10,0673	15,62161	1,388588	-1,04144	-3,12432
6. ms	1,822521	-9,98047	15,27446	1,648948	-0,60751	-3,9054
7. ms	1,735734	-9,63333	15,10089	1,822521	-0,17357	-4,68648
8. ms	1,562161	-9,19939	14,92732	1,909308	0,086787	-5,55435
9. ms	1,301801	-8,5051	14,66696	1,822521	0,347147	-6,509

Table 3.3 Min-max values

	Min	Max
A1	-25.1681	24.4739
A1a	-22.9117	28.3793
B1	-25.1671	27.9453
B1a	-22.0438	35.1486
C1	-16.8366	20.7420
C1a	-26.8171	26.6435
D1	-33.5865	41.0501
D1a	-43.0462	37.4051
E1	-19.6138	22.3042
E1a	-26.6435	21.7835

Since the pre-stimulus and post-stimulus results are examined in the research, the pre-stimulus data is collected and coded as “oncesi” .Then the post-stimulus results are gathered and coded as “sonrasi” and a 500*26 level matrix is created. To identify the data better EDA and the statistical definitions are used: table 3.3, 3.4, 3.5, 3.6, 3.7 and Figure 3.2 and 3.3.

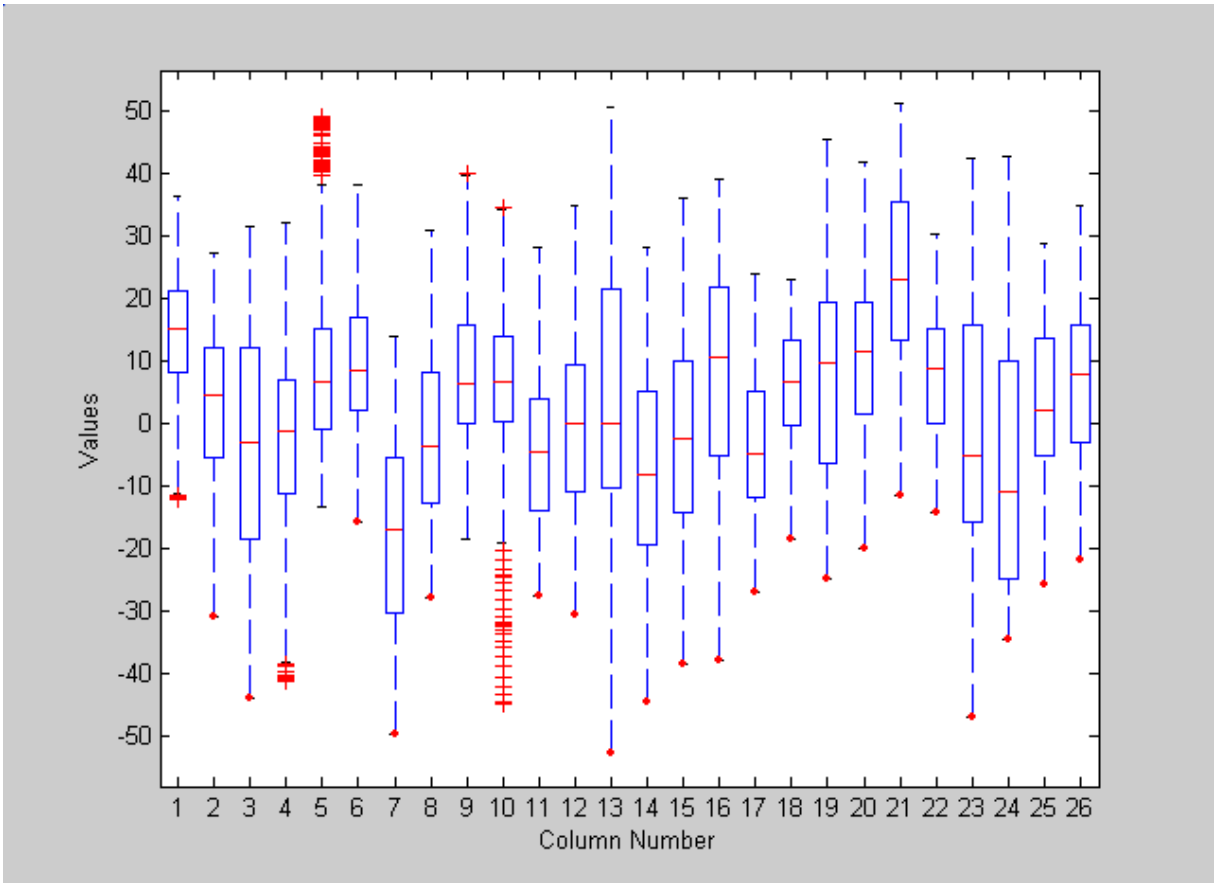


Figure 3.2 Boxplot graphs for 26 sweeps pre-stimulus

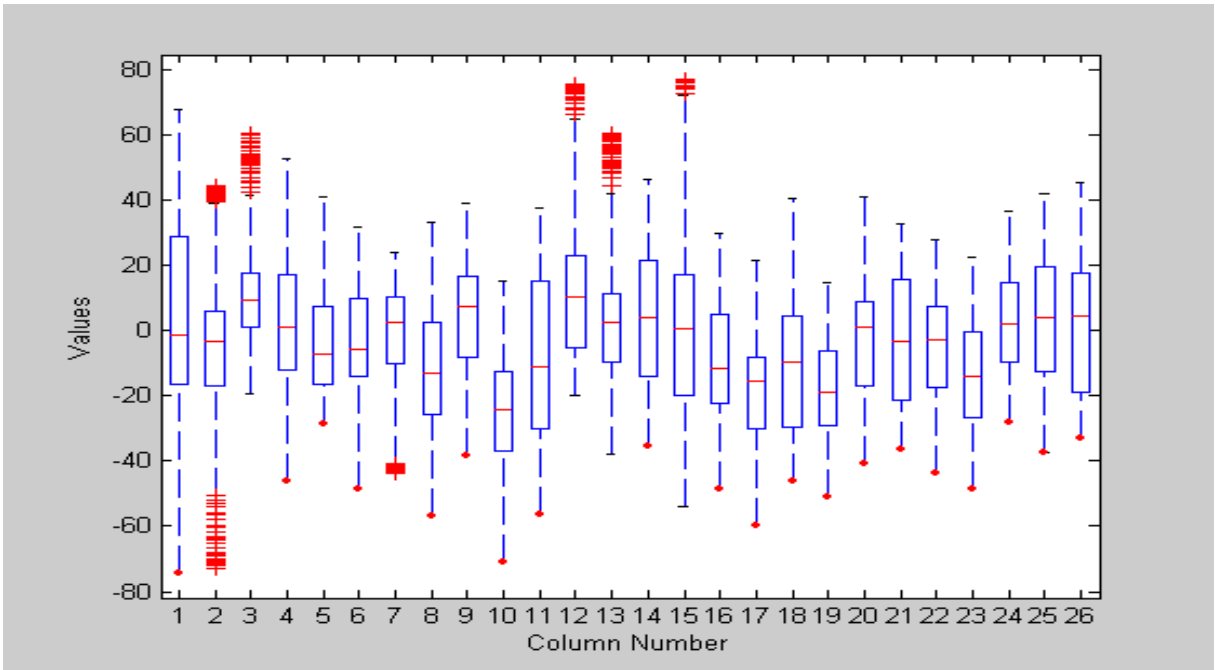


Figure 3.3 Boxplot graphs for 26 sweeps post-stimulus

When we look at the Figure 3.2 and 3.3 we can see the boxplot graphics .The extreme values post-stimulus are higher. Table 3.3, 3.4, 3.5, 3.6, show that the pre- and post-stimulus average and changes.

Table 3.4 ‘Oncesiort’

Sweep 1	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6	Sweep 7	Sweep 8
14.134	24.227	-34.758	-21.869	9.867	94.268	-17.697	-20.982
Sweep 9	Sweep 10	Sweep 11	Sweep 12	Sweep 13	Sweep 14	Sweep 15	Sweep 16
81.319	58.107	-4.064	-0.2034	45.179	-83.737	-13.521	75.204
Sweep 17	Sweep 18	Sweep 19	Sweep 20	Sweep 21	Sweep 22	Sweep 23	Sweep 24
-39.478	60.539	82.125	11.178	22.48	83.319	-17.786	-70.808
Sweep 25	Sweep 26						
27.428	6.713						

Table 3.5 ‘sonrasiort’

Sweep 1	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6	Sweep 7	Sweep 8
3,0974	-5,0178	11,051	1,7765	-2,128	-4,1349	-2,3288	-11,689
Sweep 9	Sweep 10	Sweep 11	Sweep 12	Sweep 13	Sweep 14	Sweep 15	Sweep 16
3,9198	-24,407	-8,8028	13,028	2,1261	4,7576	-0,0324	-9,2855
Sweep 17	Sweep 18	Sweep 19	Sweep 20	Sweep 21	Sweep 22	Sweep 23	Sweep 24
-18,701	-9,2537	-17,408	-1,4721	-3,6431	-5,1973	-13,848	2,1065
Sweep 25	Sweep26						
4,3744	2,3278						

Table 3.6 ‘stdoncesi’

Sweep 1	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6	Sweep 7	Sweep 8
10.511	13.256	19.524	15.732	15.471	12.074	15.423	13.697
Sweep 9	Sweep 10	Sweep 11	Sweep 12	Sweep 13	Sweep 14	Sweep 15	Sweep 16
13.221	12.793	13.098	14.971	23.355	17.897	16.941	18.176
Sweep 17	Sweep 18	Sweep 19	Sweep 20	Sweep 21	Sweep 22	Sweep 23	Sweep 24
12.043	91.376	16.388	13.155	15.584	10.818	20.759	20.885
Sweep 25	Sweep 26						
13.144	12.905						

Table 3.7 ‘stdsonrasi’

Sweep 1	Sweep 2	Sweep 3	Sweep 4	Sweep 5	Sweep 6	Sweep 7	Sweep 8
34.496	24.471	16.459	20.54	18.535	19.915	17.776	21.704
Sweep 9	Sweep 10	Sweep 11	Sweep 12	Sweep 13	Sweep 14	Sweep 15	Sweep 16
18.285	19.938	25.705	22.832	22.776	21.554	28.912	17.726
Sweep 17	Sweep 18	Sweep 19	Sweep 20	Sweep 21	Sweep 22	Sweep 23	Sweep 24
19.27	22.242	15.315	20.73	19.911	17.019	18.996	15.684
Sweep 25	Sweep 26						
20.066	21.269						

There are differences between the averages of pre and post-stimulus. When we look at the correlation coefficient numbers between the data of pre and post-stimulus;

$$\begin{bmatrix} 1.0000 & -0.0266 \\ -0.0266 & 1.0000 \end{bmatrix}$$

We can say that there is a weak correlation in the opposite way between the pre and post-stimulus data.

Correlation matrix(r);

$$\begin{bmatrix} oncesi * oncesi & oncesi * sonrasi \\ sonrasi * oncesi & sonrasi * sonrasi \end{bmatrix}_{52*52}$$

When we examine the correlation matrix, a 52*52 level matrix occurs.(Appendices 1-table 2). The first 26*26 part of the matrix occurred gives the correlation of the pre-stimulus data with itself. Also the 27*26 level parts gives the covariance matrix between pre and after. And 27*52 level part gives the post-stimulus data correlation with itself.

Table 3.8 Part of correlation matrix ‘Oncesi*oncesi’

	Sweep 1	Sweep 2	Sweep 3	Sweep 4	Sweep 5
Sweep 1	1	-0.23053	0.1714	-0.0626	-0.02505
Sweep 2	-0.23053	1	-0.27533	-0.44436	-0.15641
Sweep 3	0.1714	-0.27533	1	0.0853	0.10204
Sweep 4	-0.0626	-0.44436	0.0853	1	0.56407
Sweep5	-0.02505	-0.15641	0.10204	0.56407	1

Table 3.9 Part of correlation matrix 'Oncesi*sonrasi'

Post-stimulus Pre-stimulus	Sweep 1	Sweep 2	Sweep 3	Sweep 4	Sweep 5
Sweep 1	-0.43355	0.20392	0.078349	-0.12151	-0.057524
Sweep 2	-0.14591	-0.14972	0.11716	0.12733	-0.1389
Sweep 3	0.50771	0.1544	-0.1855	-0.10811	0.14067
Sweep 4	-0.0271	0.21986	-0.11577	-0.023316	0.19439
Sweep 5	-0.10268	0.52084	-0.16299	-0.30244	0.034008

Table 3.8 shows the correlation of the pre 'oncesi' data with itself. The first sweep and its correlation is one in the first 5*5 5.5 part. -0.23053 is the correlation coefficient between the 1st sweep and the 2nd sweep.

In table 3.9 we can see pre and after covariance matrix and an opposite way relation between the 1st sweep's pre and after covariance (-0.43355). The pre-stimulus 3rd sweep and post-stimulus first sweep has a strong relation. (0.50771).To identify the meaningful level of the 52*52 part of the correlation matrix, we have to identify the 70% part of it. So the values less than 1 are identified. As a result 20 values have a meaningful correlation.

The 2nd sweep show that the five person's pre and after data (figure 4). The order of the drawings in the 1st line and the data of the 5 persons' pre-stimulus case and the 2nd sweep post-stimulus data is in the figure 3.6.

The clustering analysis groups the data items. The target is to have the same or similar items in one group and to make the difference between the items in different groups.

The similarity or the difference rates show how good the clustering is done. That's why we use clustering for the EEG analysis. We have used the MATLAB program and k-means clustering method.

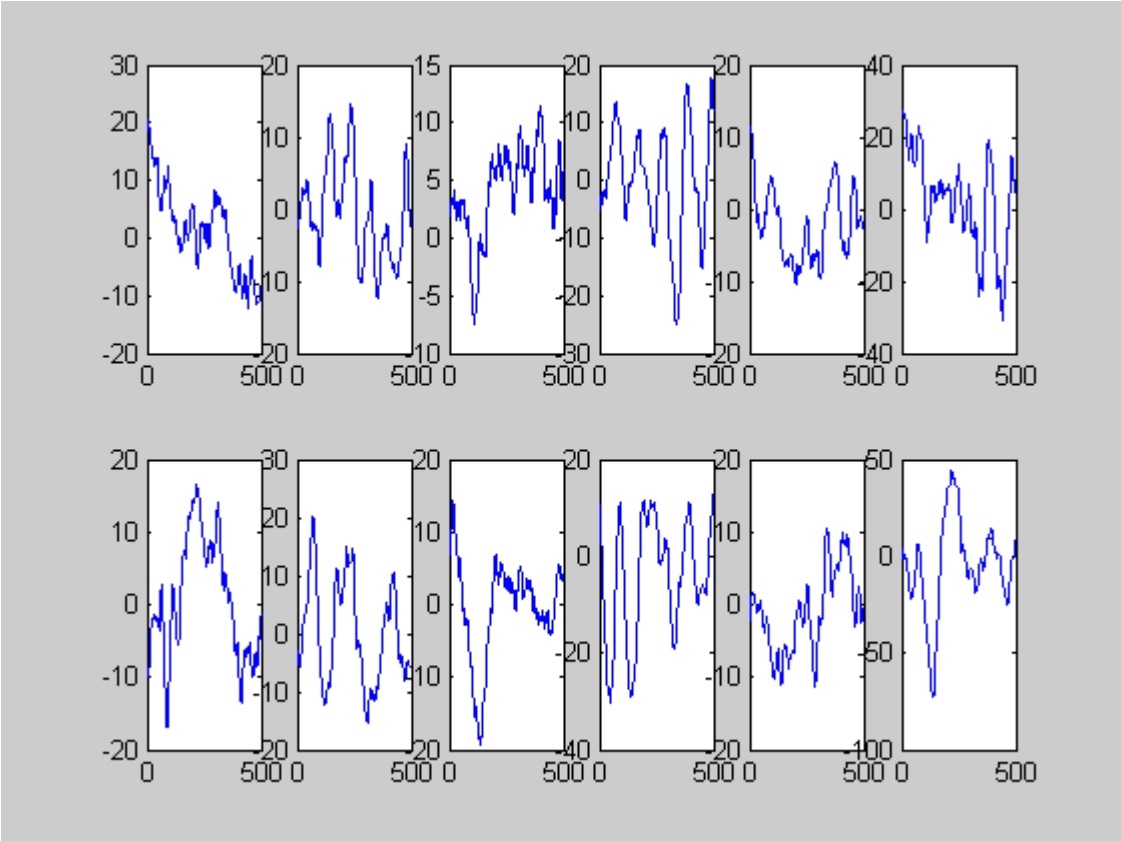


Figure 3.4 2.sweep subplot

For such data instead of k-means clustering method, hierarchical clustering method is used. Because the data's natural attachment is the point .The multiple set numbers is demanded. The measurement of the distance must be examined according to the hierarchical clustering method. The best method for measuring the distance is the Cosine. For this method, the cophenet coefficient which is in the MATLAB program is taken into consideration. Cophenet coefficient tells us which distance calculation is the best for the data. While the Euclidian method's coefficient is 0.6940, the cosine method's coefficient is 0.7121. After the hierarchical clustering method Dendrogram graphics are obtained.

In Figure 3.6 shows the sets of the EEG data and the similar sweeps in the same sets. According to the dendrogram graphics in figure 6 the clustering analysis that is done is the correlation matrix, there are 3 sets.

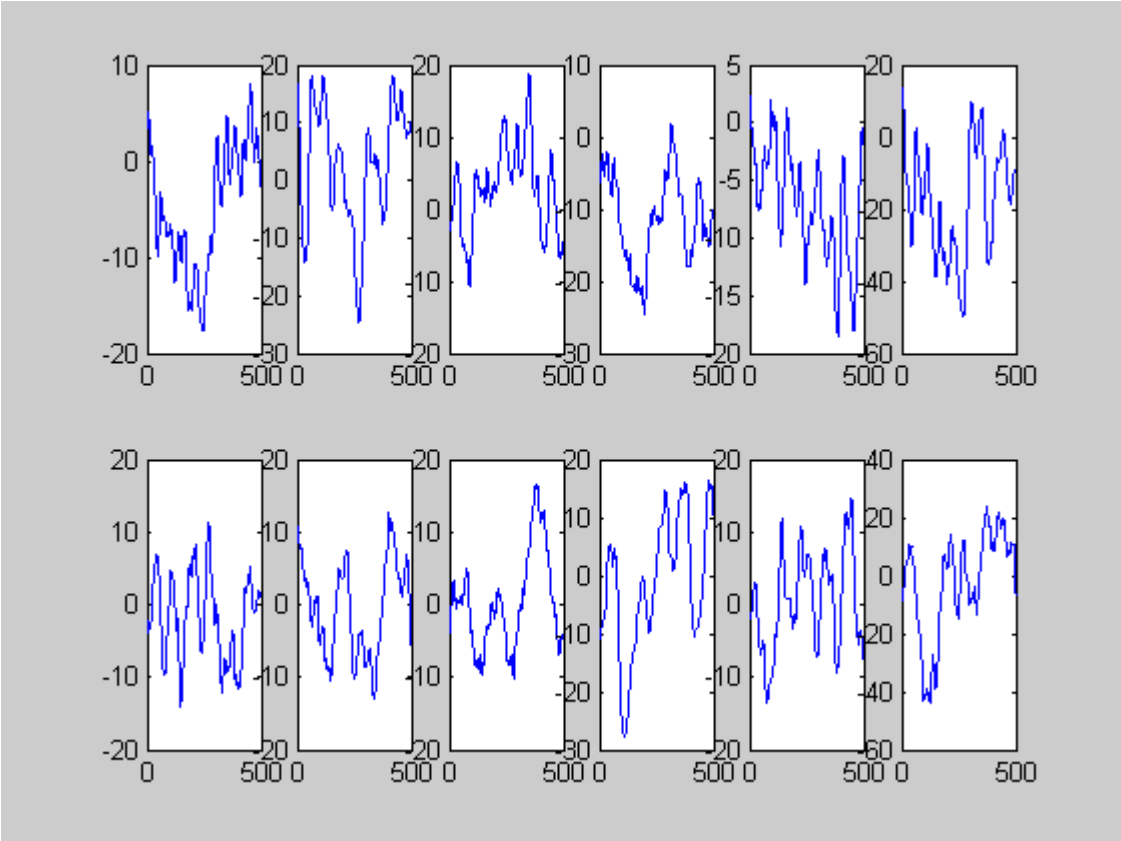


Figure 3.5: 7. Sweep subplot

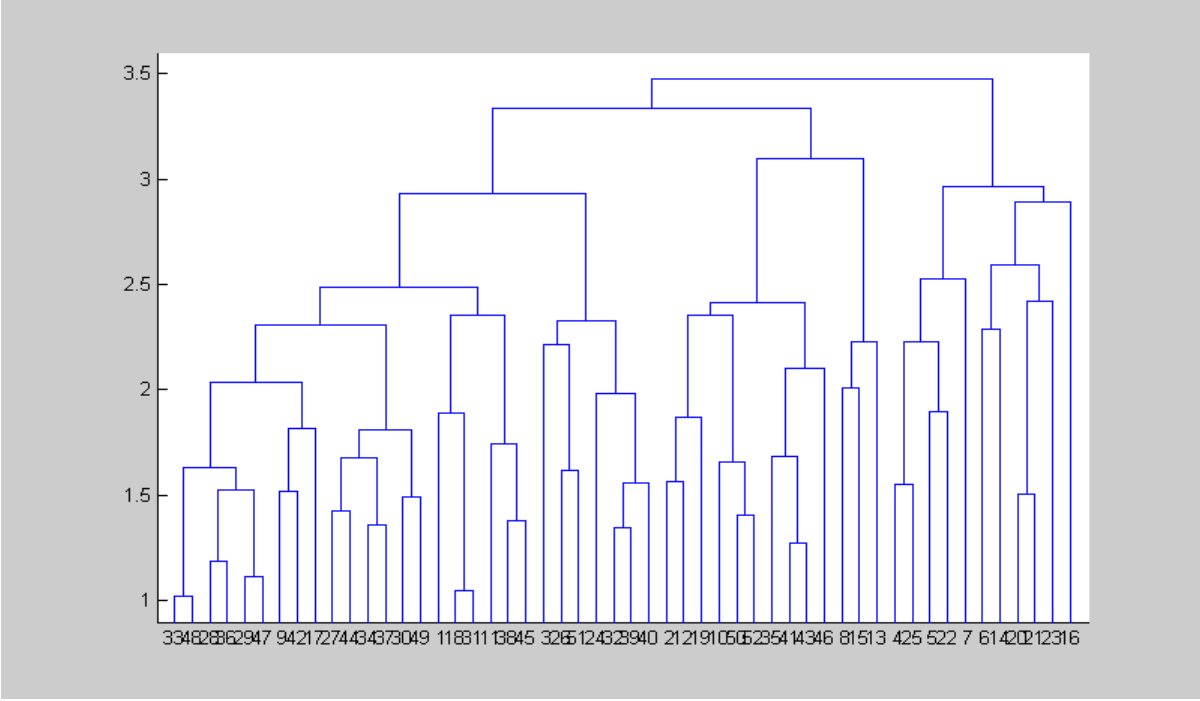


Figure 3.6 Dendrogram

In table 3.10 shows 52* 52 level matrix sets(appendices 4) according to their sweep numbers. In the examination the first 250 ms and the last 250 ms are grouped

differently and the correlation between them is inspected. And it is seen that the first 250 ms correlations are higher. This shows that the time between the sweeps should be shorter than 500 ms. (appendices 3)

Table 3.10 Clusters

1. cluster	33, 48, 28, 36, 29, 47, 9, 42, 17, 27, 44, 34, 37, 30, 49, 1, 18, 31, 11, 38, 45, 3, 26, 51, 24, 32, 39, 40
2. cluster	2, 12, 19, 10, 50, 52, 35, 41, 43, 46, 8, 15, 13
3. cluster	4, 25, 5, 22, 7, 6, 14, 20, 21, 23, 16

3.2 Application of IVF Data

In the second research that was done in a private in IVF unit, it was inspected whether the Active drug medicine that is used in tube baby has an effect on depending on some hormones FSH, LH, E2 and some variables like age, Oocyte, Endometrium, Cleavage. And this question was answered by the logistic regression usage.

There are a lot of medicines that are used in tube baby treatment. And also some hormones and variables are effective on this treatment. The point is to find out which of them cause the pregnancy.

The hormones and medicine's effecvetiveness in the tube baby treatment could not be identified easily because first we should identify what the hormones and the medicine introduce.

On the brain tissue hypothalamus part FSH is produced. If the FSH is less than 10 this means there are enough eggs and the pregnancy possibility is high. With the help of FSH the new eggs are produced in little bags named follicles. The E2 hormone that is produced in the follicles develops the womb. One of the follicles developed better than the others. The developing follicle becomes ready to open. If the estrogen hormone is less than 51, this causes the pregnancy possibility decrease.

On the other hand, the pituitary gland produces LH hormone. This hormone makes the follicle crack and releases the egg cell. The treatment can be started on the 2nd or 3rd day of the menstrual period after checking the blood values.

The point is to gain as many follicles as we can. Another observed part is the endometrium part because the pregnancy develops here when it is less than 7 mm the possibility of the pregnancy decreases.

In the tube baby method when the estrogen level is the maximum before the cracking of the follicle, the eggs are taken. The collected eggs and the sperm are put in one place and the fertilization is being observed. In the end of this fertilization division occurs (cleavage). After the division the most qualified one is chosen and put into the womb (ET=embryo transfer).Finally, with the blood test it is checked whether there is pregnancy.

Active drug includes progesterone. The monthly menstrual bleeding can be postponed by using active drug. It is the medicine for organizing the menstrual period.

In a Private Hospital IVF unit it has been inspected whether Active drug has an effect on getting pregnant by using the SAS package program and with the model of logistic regression.

Table 3.11 İndependent variables situation

Active drug Use	Discrete	1: use drug, 0: disuse drug
FSH Hormone	Discrete	1: 10 and under, 0: 11 and over
E2 Hormone	Discrete	1: 50 and under, 0: 51 and over
LH Hormone	Continuous	
Age	Continuous	
Cleavage	Continuous	

The research was done among the 175 patients in the last one year. But the 15 of these patients were omitted from this program because the cell division has never occurred on these 15 patients. The rest 160 patients' data are examined in great care in the IVF unit. The patients' names were never used. 6 different variables were seen on the data by using the SPSS Clementine feature selection. Here, 1 case refers to the pregnancy and the 0 case refers to non-pregnancy. The endometrium thickness and the oocytle are not in this table but they are used in the tube baby treatment.

In tables 3.11, 3.12, 3.13; shows the pregnancy case and the variables that have an effect. After giving the explanations about the dependent and independent variables we can start the establishment of a model. For the logistic regression model, choosing the variables must be the first step for this, the backward elimination. According to this method all the independent variables are taken into the model. Then the independent variables are selected by the Wald statistics and the p values.

Table 3.12 Frequency

		Frequency	Percent
Active drug	Use	59	36,9
	Disuse	101	63,1
	Total	160	100,0

Table 3.13 Descriptive statistics

	N	Minimum	Maksimum	average	Std. Deviation
LH	160	0,44	12,00	4,9136	2,25088
Age	160	20,00	45,00	34,8438	4,99754
Cleavage	160	1,00	16,00	4,7688	2,99312
N	160				

Table 3.14 shows that the codding of the categorical variables which are in the logistic regression model. According to this table the E2 and FSH variables' important point is 1.Because the factors that cause the pregnancy is in the 1 point. This means if the E2 variable is less than 50 and FSH variable is 10 or less, the pregnancy possibility increases. Active drug is thought to be effective in the pregnancy treatment. As a result of the logistic regression studies it is seen that the

effect of the certain variables are minimum. This result is decided by checking the Wald statistics and the p values. According to the results that are in table 3.16, in the first step all the variables are in the model. However, the LH variable which has a low level Wald value and a high level p value can't pass to the second step and is out of the model. 0.002 is its Wald value and 0.961 is the p value. With all the variables- except from the LH variable that was out of the model- a new logistic regression model is set. FSH is out of the model with a Wald value of 0.104 and p value of 0.748.In the 3rd step the Active drug variable and in the 4th step E2 variable is out of the model so that in the last step only 2 independent variables are effective on becoming pregnant. These are the division and the age variables.

Table 3.14 Categorical variables code

		Frequency	Parameter code
E2	0,00	27	1
	1,00	133	0
FSH	0,00	44	1
	1,00	116	0
Active drug	Use	59	1
	Disuse	101	0

The logistic regression model which was gained from the age and the division variables in the 5th step are in 95% reliability. We can't see the direct effect of the Active drug when we look at the logistic regression results of these variables. The logistic regression model is as below:

According to the OR values in table 5 there is no serious risk. But when we see the 80% reliable logistic regression models, on the 3rd step model we can see the effect of the active drug. As a result the variables in the model are the active drug, age, E2 and division variables. This logistic model can be expressed as below.(Table 3.15)

When we look at the OR value in table 5 we can see that young age has a positive effect on getting pregnant. Active drug is 1.5 times more effective and E2 hormone has 2 times more effect on getting pregnant.

$$g(x_i) = \text{Fixed} + \beta_0\ \text{Age} + \beta_1\ \text{Cleavage} \quad (1)$$

$$g(x_i) = 1{,}652 - 0{,}089\ X_{1i} + 0{,}179\ X_{2i} \quad (2)$$

$$\pi(x_i) = \frac{1}{1+\exp(-g(x_i))} = \frac{1}{1+\exp(-(1{,}652 - 0{,}089\ X_{1i} + 0{,}179\ X_{2i}))} \quad (3)$$

$$g(x_i) = \text{Fixed} + \beta_0\ \text{Age} + \beta_1\ \text{Cleavage} + \beta_3\ \text{Active drug} + \beta_4\ E2 \quad (4)$$

$$g(x_i) = 1{,}274 - 0{,}086\ X_{1i} + 0{,}177\ X_{2i} + 0{,}477\ X_{3i} + 0{,}683\ X_{4i} \quad (5)$$

$$\pi(x_i) = \frac{1}{1+\exp(-g(x_i))} = \frac{1}{1+\exp(-(1{,}274 - 0{,}086\ X_{1i} + 0{,}177\ X_{2i} + 0{,}477\ X_{3i} + 0{,}683\ X_{4i}))} \quad (6)$$

Table 3.15 3. and 5. Model comparison probability

Model	$g(x_i)$	$\pi(x_i)$
3. Model	Fixed + β_0 age + β_1 cleavage + β_3 Active drug+ β_4 E2 1,274 – 0,086 X 30 + 0,177 X 8 + 0,477 1 + 0,683 1	0,739
5. Model	Fixed + β_0 age + β_1 cleavage 1,652 – 0,089 X 30 + 0,179 X 8	0,602

Table3.16 First model

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	212.652	215.350
SC	215.727	270.703
-2 Log L	210.652	179.350

Table 3.17 Logistic regression analysis and model

		B	S.E.	Wald	Df	p-value	Exp(B)
1. Step	Drug(1)	,469	,367	1,634	1	,201	1,599
	Age	-,086	,036	5,579	1	,018	,918
	LH	-,004	,084	,002	1	,961	,996
	Cleavage	,182	,064	8,188	1	,004	1,200
	FSH(1)	,139	,439	,100	1	,751	1,149
	E2(1)	,689	,472	2,132	1	,144	1,993
	Fixed	1,221	1,363	,803	1	,370	3,392
2. Step	Drug(1)	,470	,367	1,637	1	,201	1,600
	Age	-,086	,036	5,635	1	,018	,918
	Cleavage	,182	,063	8,260	1	,004	1,199
	FSH(1)	,131	,407	,104	1	,748	1,140
	E2(1)	,688	,471	2,131	1	,144	1,990
	Fixed	1,210	1,344	,810	1	,368	3,354
3. Step	Drug(1)	,477	,367	1,693	1	,193	1,611
	Age	-,086	,036	5,661	1	,017	,917
	Cleavage	,177	,061	8,352	1	,004	1,193
	E2(1)	,683	,470	2,112	1	,146	1,981
	Fixed	1,274	1,329	,919	1	,338	3,577
4. Step	Age	-,089	,036	6,055	1	,014	,915
	Cleavage	,180	,061	8,732	1	,003	1,198
	E2(1)	,591	,460	1,647	1	,199	1,805
	Fixed	1,544	1,309	1,391	1	,238	4,685
5. Step	Age	-,089	,036	6,031	1	,014	,915
	Cleavage	,179	,060	8,853	1	,003	1,196
	Fixed	1,652	1,309	1,593	1	,207	5,218

Table 3.18 Fifth model

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	212.652	205.596
SC	215.727	211.746
-2 Log L	210.652	201.596

As seen in the table3.17 and 3.18 like hood interval is decreased in the most convenient model.

For example, estimation of probability of becoming pregnant for someone 30 years old with 8 cleavages, that takes active drug pills and whose E2 level is under 50 is as in the table below, for modes 3 and 5. According to this tablet he two models are effective on getting pregnant. But while the reliability of the 3rd model is 80%, the 5th models reliability is 95 %. So the people who are 30 and have 8 divisions can get pregnant with a possibility of 0.6.

At first, by backward elimination method, variables in step logistic regression model are designated. Concerning step logistic regression results obtained by the "Age" and "Division" variables, with 95% of conformity, usage of active drug is not directly effective for being pregnant. But if the model is designed with 80% of conformity rather than 95% than the model with "Age", "Division", "E2" and "active drug" is obtained. As always must be studied with high conformity, model 3 which does not include the pills active drug is chosen. As a result of study, in this date set, active drug is not efficient for pregnancy.

CHAPTER FOUR
CONCLUSIONS

Recently by the notable developments in the information Technologies the amounts of the stored data has increased. With each passing day computer systems are getting more and more powerful and cheap. The processors are getting faster and the capacities of the discs are increasing. Again they can store larger amounts of data and can process them in shorter time. Data can be collected and stored directly and numerically. As a result of this detailed and correct info can be achieved. Data is worthless when it's alone. It become as info in terms of our desire and aims. Information is data which is processed for a purpose. That's why the need for achieving to the beneficial information from large data masses has arose. To implement this data mining techniques has become daily research topics. By a simple definition the data mining is a job for achieving, mining information among large scaled dat. or in other means, searching the connections that help us form an estimate about future among large scaled data masses by using computer programmers. In this context in this study data mining process has been expressed. Subjects obtained as how to process the data, which stages to migrate, which model would be available for the data. Statistical issues supporting data mining process has been analyzed. These statistical methods are regression, classification, sequence analysis, association rules, summarization and clustering. Then issue of how data mining in medical area which is one of the usage areas of data mining is obtained. Modern medicine generates huge amounts of data, but at the same time there is an acute and widening gap between data collection and data comprehension. Thus, there is growing pressure not only to find better methods of data analysis, but also need for automating them to facilitate the creation of knowledge that can be used for clinical decision-making. This is where data mining and knowledge discovery tools come into play, since they can help in achieving these goals.

Data mining and knowledge discovery in medical databases are not substantially different from mining in other types of databases. Problems faced in medical data mining process are obtained and methods to solve these problems are searched. Appliance is done on clustering and logistic regression, one of the statistical methods that support data mining.

In the first work, the electroensefalography (EEG) data, which was gained by Dokuz Eylül University Medicine Faculty Biophysics department laboratory, is examined. EEG is the method which measures the observation of the brain waves' activity with an electrical system. The target of this work is to solve the EEG structure by using the clustering and statistical methods. The clustering analysis groups the data items. The target is to have the same or similar items in one group and to make the difference between the items in different groups. The similarity or the difference rates show how good the clustering is done. We have used the MATLAB program hierarchical and k-means clustering methods. For such data instead of k-means clustering method, hierarchical clustering method is used. Because the data's natural attachment is the point .The multiple set numbers is demanded. As a result 52* 52 dimensional matrix has been divided into 3 clusters. At the results of the correlation in data that are post- stimulus it is seen that data in the first 250ms has formed a group.

In the second research that was done in a private hospital in IVF unit, it was inspected whether the active drug medicine that is used in tube baby has an effect on depending on some hormones FSH, LH, E2 and some variables like age, Oocyte, Endometrium, Cleavage. And this question was answered by the logistic regression usage. There are a lot of medicines that are used in tube baby treatment. And also some hormones and variables are effective on this treatment. The point is to find out which of them cause the pregnancy. Active drug includes progesterone. The monthly menstural bleeding can be postponed by using active drug. It is the medicine ofor organizing the menstural period. In a hospital IVF unit it has been inspected whether active drug has an effect on getting pregnant by using the SAS package program and with the model of logistic regression. The research was done among the 175 patients in the last 1 year. But the 15 of these patients were omitted from this program because the cell division has never occurred on these 15 patients. The rest 160 patients'data are examined in great care in the IVF unit. The patients' names were never used. 6 different variables were seen on the data by using the SPSS Clementine feature selection. As a result; backward elimination method, variables in step logistic regression model are designated. Concerning step logistic regression results obtained by the "Age" and "Division" variables, with 95% of conformity, usage of Active drug is not directly effective for being pregnant.

As a result, data mining is the possibility of processing a huge amount of data what we choose. These methods are mostly preferred because their low cost and rapid process. That is why they are so valuable in medical areas. All data of patients could be stored in huge data bases which causes development in ways of cure.
We can say that, data mining is a “clever” robot that turns data into information.

REFERENCES

Akande, V.A., Keay, S.D., Hunt, L.P., Mathur, R.S., Jenkins, J.M., & Cahill, D.J. (2004). *The practical implications of a rarsed serum FSH and age on the risk of IVF treatment cancellation due tı a poor ovarian response*(7th ed.). Netherlands:Springerlink.

Akat, A.Z., Doğanay, M., Koloğlu, M., Gözalan, U., Dağlar, G., Kama, N.A.(2002). Evaluation Of 1000 Laparoscopic Cholecystectomiesperformed In One Institution. *Türkiye klinikleri tıp bilimleri dergisi,22*,2.

Antonie, M.L., Zaiane, O.R., Coman, A.(2001). Application of data mining techniques for medical image classification. *Proceedings of the second international workshop on multimedia data mining,* 2008, from http://www.cs.ualberta.ca/~luiza/mdmkdd01.pdf

Berry, M.J.A.(1999). *Mastering data mining the art and science of customer relationship management*. Lirnoff: John Wiley& Sons.

Chae, Y.M., Ho, S.H., Cho, K.W., Lee, D.H., Ji, S.H.(1998). Data mining approach to policy analysis in a healt insurance domain. . *Int J Med Inform 2001*, 62, 103-11.

Cios, K.J. & Moore, W.(2002). Uniqueness of medical data mining. *Artifical intellgence in Medical journal ,26,* 1-24,2007,PMID.

Cios, K.J.(2001). *Medical data mining and knowledge discovery*(5th ed.). NY:Physica-Verlag.

Çolak, C., Çolak, M.C., Orman, M.N.(2001). Koroner arter hastalığının tahmininde lojistik regresyon modeli seçim yöntemlerinin karşılaştırılması. *Anadolu kardiyoloji dergisi 2007,7,*6-11.

Daşdağ, S.(n.d.). *Elektroansefalografinin biyofizik temelleri,* 2007, http://www.dicle.edu.tr/~dasdag/EEG.ppt.

Dunham, M.H.(2002). *Data mining introductory and advanced topics*(10th ed.). New jursey: Prentice Hall.

Eanen, A.G.(2004).*Application of data mining in medical applications,* 2007, http://etd.uwaterloo.ca/etd/ageapen2004.pdf.

Everitt, B.(2003). *Medical statistics from A to Z*(2nd ed.). Edinburgh:Cambridge.

Giudici, P.(2003). *Applied data mining statistical methods for business and industry*(2nd ed.). London:John Wiley& Sons.

Han, J. & Kamber, M.(2001). *Data mining concepts and techniques*(5th ed.). San Diego: Academic pres.

Han, J., Fu, Y.(1999). *Mining multiple-level association rules in large databases* (5th ed.). USA: IEEE.

Karypis, G., Han, E., & Kumar, V.(1999). *Chameleon: Hierarchical clustering using dynamic modeling*(2nd ed.). USA:IEEE.

Kusiak, A., Kernstine, K.H., Kern, J.A., Mclaughlin, K.A., & Tseng, T.L.(May 21-23,2000). Data mining:medical and engineering case studies. *Proceedings of the industrial engineering research 2000 conferance,1*(1), 2008, from http://www.icaen.viowa.edu/~ankusiak/Res:in-prog/IIE-0.pdf.

Lemeshow, S., &Hosmer, D.(1989). *Applied logistic regression*(2nd ed.). Canada: John Wiley&Sons.

Martinez, W.L.& Martinez, A.R.(2002). *Computationa statistics handbook with matlab*(2nd ed.). Florida:chapman&Hall/CRC.

Men, E.E., Yıldırımtürk, Ö., Tuğcu, A., Aytekin, V.,&Aytekin, S.(2004). Açık kalp cerrahisi sonrası gelişen atriyal fibrilasyonu önlemek için kullanılan ilaçların etkinlik yönünden Karşılaştırılması. *Anadolu kardiyoloji dergisi, 8,* 206-213.

Okandan, M., Kara, S.(2004). Atrial fibrilasyonlu ekg'lerin yapay sinir ağları ve dalgacık dönüşümü kullanılarak sınıflandırılması. *Biyomedikal Mühendisliği Ulusal Toplantısı, BİYOMUT 2004,* 61-64.

Paetz, J., Hamker, F., &Thone, S.(2000). About the analysis of septic shock patient data. *Lecture note in computer science, 1,* 130-137.

Ramkumar, G.D.,& Swami, A.(1998). *Clustering data without distance functions*. USA:IEEE.

User reference for SAS ver. 9.1, 2005.

User reference for Matlab ver. 6.5, Mathworks Inc., 2003.

Seidman, C.(2000). *Data mining with micosoft SQL server*. Washington: Microsoft Press.

Şahin, C., Oğulata, S.N., Kırım, S., &Koçak, M.(2004). Troid bezi bozukluklarının yapay sinir Ağları ile teşhis. YA/EM'04, 24, 2008, http://yaem2004.cu.edu.tr/bildiriler/tammetin.pdf .

Şahinler, S.(2000).En küçük kareler yöntemi ile doğrusal regresyon modeli oluşturmanın temel prensileri. *MKÜ ziraat fakültesi dergisi,5*,57-73.

Şahinoğlu, M., Moralı, N., Kurt, S., &Ege, Ö.(2001). Biyoistatistik terimler sözlüğü. İzmir: Dokuz Eylül Yayınları. Tax, D.M.J.(n.d.).*Exploratory data analysis.* 2007, www.ph.tn.tudelft.nl/~csp/da/DA.pdf .

Zaki, M.J.(1999). *Parallel and distributed association mining*(5th ed.). USA:IEEE.

Appendix 1. Matlab Syntax

Mean: Average or mean value of arrays

Syntax:M = mean(A)

M = mean(A,dim)

Description:M = mean(A) returns the mean values of the elements along different dimensions of an array.

If A is a vector, mean(A) returns the mean value of A.

If A is a matrix, mean(A) treats the columns of A as vectors, returning a row vector of mean values.

If A is a multidimensional array, mean(A) treats the values along the first non-singleton dimension as vectors, returning an array of mean values.

M = mean(A,dim) returns the mean values for elements along the dimension of A specified by scalar dim.

Sum: of array elements

Syntax: B = sum(A)

B = sum(A,dim)

Description: B = sum(A) returns sums along different dimensions of an array.

If A is a vector, sum(A) returns the sum of the elements.

If A is a matrix, sum(A) treats the columns of A as vectors, returning a row vector of the sums of each column.

If A is a multidimensional array, sum(A) treats the values along the first non-singleton dimension as vectors, returning an array of row vectors.

B = sum(A,dim) sums along the dimension of A specified by scalar dim.

Std:Standard deviation of a sample

Syntax: y = std(X)

Description: y = std(X) computes the sample standard deviation of the data in X. For vectors, std(x) is the standard deviation of the elements in x. For matrices, std(X) is a row vector containing the standard deviation of each column of X.

std normalizes by n-1 where n is the sequence length. For normally distributed data, the square of the standard deviation is the minimum variance unbiased estimator of 2 (the second parameter). The standard deviation is

$$s = \left(\frac{1}{n-1} \sum \left(x_i - \bar{x} \right)^2 \right)^{\frac{1}{2}}$$

where the sample average is

$$\bar{x} = \frac{1}{n} \sum x_i$$

sqrt :Square root

Syntax: B = sqrt(X)

Description: B = sqrt(X) returns the square root of each element of the array X. For the elements of X that are negative or complex, sqrt(X) produces complex results.

corrcoef :Correlation coefficients

Syntax: R = corrcoef(X)

R = corrcoef(x,y)

[R,P]=corrcoef(...)

[R,P,RLO,RUP]=corrcoef(...)

[...]=corrcoef(...,'param1',val1,'param2',val2,...)

Description: R = corrcoef(X) returns a matrix R of correlation coefficients calculated from an input matrix X whose rows are observations and whose columns are variables. The matrix

R = corrcoef(X) is related to the covariance matrix C = cov(X) by

$$R(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}}$$

corrcoef(X) is the zeroth lag of the covariance function, that is, the zeroth lag of xcov(x,'coeff') packed into a square array.

R = corrcoef(x,y) where x and y are column vectors is the same as corrcoef([x y]). [R,P]=corrcoef(...) also returns P, a matrix of p-values for testing the hypothesis of no correlation. Each p-value is the probability of getting a correlation as large as the observed value by random chance, when the true correlation is zero. If P(i,j) is small, say less than 0.05, then the correlation R(i,j) is significant.

[R,P,RLO,RUP]=corrcoef(...) also returns matrices RLO and RUP, of the same size as R, containing lower and upper bounds for a 95% confidence interval for each coefficient. [...]=corrcoef(...,'param1',val1,'param2',val2,...) specifies additional parameters and their values. Valid parameters are the following.

'alpha' A number between 0 and 1 to specify a confidence level of 100*(1 - alpha)%. Default is 0.05 for 95% confidence intervals.

'rows' Either 'all' (default) to use all rows, 'complete' to use rows with no NaN values, or 'pairwise' to compute R(i,j) using rows with no NaN values in either column i or j.

The p-value is computed by transforming the correlation to create a t statistic having n-2 degrees of freedom, where n is the number of rows of X. The confidence bounds are based on an asymptotic normal distribution of 0.5*log((1+R)/(1-R)), with an approximate variance equal to 1/(n-3). These bounds are accurate for large samples when X has a multivariate normal distribution. The 'pairwise' option can produce an R matrix that is not positive definite.

Silhouette: Silhouette plot for clustered data

Syntax: silhouette(X,clust)

s = silhouette(X,clust)

[s,h] = silhouette(X,clust)

[...] = silhouette(X,clust,distance)

[...] = silhouette(X,clust,distfun,p1,p2,...)

Description: silhouette(X,clust) plots cluster silhouettes for the n-by-p data matrix X, with clusters defined by clust. Rows of X correspond to points, columns correspond to coordinates. clust can be a numeric vector containing a cluster index for each point, or a character matrix or cell array of strings containing a cluster name for each point. silhouette treats NaNs or empty strings in clust as missing values, and ignores the corresponding rows of X. By default, silhouette uses the squared Euclidean distance between points in X.

s = silhouette(X,clust) returns the silhouette values in the n-by-1 vector s, but does not plot the cluster silhouettes.

[s,h] = silhouette(X,clust) plots the silhouettes, and returns the silhouette values in the n-by-1 vector s, and the figure handle in h.

[...] = silhouette(X,clust,distance) plots the silhouettes using the inter-point distance measure specified in distance. Choices for distance are:

'Euclidean'Euclidean distance

'sqEuclidean'Squared Euclidean distance (default)

'cityblock'Sum of absolute differences, i.e., L1

'cosine'One minus the cosine of the included angle between points (treated as vectors)

'correlation'One minus the sample correlation between points (treated as sequences of values)

'Hamming'Percentage of coordinates that differ

'Jaccard'Percentage of non-zero coordinates that differ

Vector A numeric distance matrix in upper triangular vector form, such as is created by pdist. X is not used in this case, and can safely be set to [].

[...] = silhouette(X,clust,distfun,p1,p2, ...) accepts a distance function of the form d = distfun(X0,X,p1,p2,...)

where X0 is a 1-by-p point, X is an n-by-p matrix of points, and p1,p2,... are optional additional arguments. The function distfun returns an n-by-1 vector d of distances between X0 and each point (row) in X. The arguments p1, p2,... are passed directly to the function distfun.

Kmeans: K-means clustering

Syntax: IDX = kmeans(X,k)

[IDX,C] = kmeans(X,k)

[IDX,C,sumd] = kmeans(X,k)

[IDX,C,sumd,D] = kmeans(X,k)

[...] = kmeans(...,'param1',val1,'param2',val2,...)

Description: IDX = kmeans(X, k) partitions the points in the n-by-p data matrix X into k clusters. This iterative partitioning minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. Rows of X correspond to points, columns correspond to variables. kmeans returns an n-by-1 vector IDX containing the cluster indices of each point. By default, kmeans uses squared Euclidean distances. [IDX,C] = kmeans(X,k) returns the k cluster centroid locations in the k-by-p matrix C. [IDX,C,sumd] = kmeans(X,k) returns the within-cluster sums of point-to-centroid distances in the 1-by-k vector sumd. [IDX,C,sumd,D] = kmeans(X,k) returns distances from each point to every centroid in the n-by-k matrix D. [...] = kmeans(...,'param1',val1,'param2',val2,...) enables you to specify optional parameter name-value pairs to control the iterative algorithm used by kmeans.

hold Hold current graph in the figure

Syntax: hold on

hold off

hold

Description: The hold function determines whether new graphics objects are added to the graph or replace objects in the graph.

hold on retains the current plot and certain axes properties so that subsequent graphing commands add to the existing graph.

hold off resets axes properties to their defaults before drawing new plots. hold off is the default.

hold toggles the hold state between adding to the graph and replacing the graph.

Grid: Grid lines for two- and three-dimensional plots

Syntax: grid on

grid off

grid minor

grid

grid(axes_handle,...)

Description: The grid function turns the current axes' grid lines on and off.

grid on adds major grid lines to the current axes.

grid off removes major and minor grid lines from the current axes.

grid toggles the major grid visibility state.

grid(axes_handle,...) uses the axes specified by axes_handle instead of the current axes.

Pdist: Pairwise distance between observations

Syntax: Y = pdist(X)

Y = pdist(X,'metric')

Y = pdist(X,distfun,p1,p2,...)

Y = pdist(X,'minkowski',p)

Description: Y = pdist(X) computes the Euclidean distance between pairs of objects in m-by-n matrix X, which is treated as m vectors of size n. For a dataset made up of m objects, there are $(m-1).m/2$ pairs.

The output, Y, is a vector of length $(m-1).m/2$ containing the distance information. The distances are arranged in the order (1,2), (1,3), ..., (1,m), (2,3), ..., (2,m), ..., ..., (m-1,m). Y is also commonly known as a similarity matrix or dissimilarity matrix.

To save space and computation time, Y is formatted as a vector. However, you can convert this vector into a square matrix using the squareform function so that element i,j in the matrix, where i<j , corresponds to the distance between objects i and j in the original dataset.

Y = pdist(X,'metric') computes the distance between objects in the data matrix, X, using

the method specified by 'metric', where 'metric' can be any of the following character strings that identify ways to compute the distance.

'euclidean' Euclidean distance (default)

'seuclidean' Standardized Euclidean distance. Each coordinate in the sum of squares is inverse weighted by the sample variance of that coordinate.

'mahalanobis' Mahalanobis distance

'cityblock' City Block metric

'minkowski' Minkowski metric

'cosine' One minus the cosine of the included angle between points (treated as vectors)

'correlation' One minus the sample correlation between points (treated as sequences of values).

'hamming' Hamming distance, the percentage of coordinates that differ

'jaccard' One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ

Y = pdist(X,distfun,p1,p2,...) accepts a function handle to a distance function of the form

d = distfun(XI,XJ,p1,p2,...)

taking as arguments two q-by-n matrices XI and XJ each of which contains rows of X, plus zero or more additional arguments, and returning a q-by-1 vector of distances d, whose kth element is the distance between the observations XI(k,:) and XJ(k,:). The arguments p1,p2,... are passed directly to the function distfun.

Y = pdist(X,'minkowski',p) computes the distance between objects in the data matrix, X, using the Minkowski metric. p is the exponent used in the Minkowski computation which, by default, is 2.

Linkage: Create hierarchical cluster tree

Syntax:Z = linkage(Y)

Z = linkage(Y,'method')

Description: Z = linkage(Y) creates a hierarchical cluster tree, using the Single Linkage algorithm. The input matrix, Y, is a distance vector of length $\left((m-1).m/2\right)$-by-1, where m is the number of objects in the original dataset. You can generate such a vector with the pdist function. Y can also be a more general dissimilarity matrix conforming to the output format of pdist.

Z = linkage(Y,'method') computes a hierarchical cluster tree using the algorithm specified by 'method', where 'method' can be any of the following character strings that

identify ways to create the cluster hierarchy. Their definitions are explained in Mathematical Definitions. 'single' Shortest distance (default)

'complete' Largest distance

'average' Average distance

'centroid' Centroid distance. The output Z is meaningful only if Y contains Euclidean distances.

'ward' Incremental sum of squares

The output, Z, is an (m-1)-by-3 matrix containing cluster tree information. The leaf nodes in the cluster hierarchy are the objects in the original dataset, numbered from 1 to m. They are the singleton clusters from which all higher clusters are built. Each newly formed cluster, corresponding to row i in Z, is assigned the index m+i, where m is the total number of initial leaves.

Columns 1 and 2, Z(i,1:2), contain the indices of the objects that were linked in pairs to form a new cluster. This new cluster is assigned the index value m+i. There are m-1 higher clusters that correspond to the interior nodes of the hierarchical cluster tree.

Column 3, Z(i,3), contains the corresponding linkage distances between the objects paired in the clusters at each row i.

For example, consider a case with 30 initial nodes. If the tenth cluster formed by the linkage function combines object 5 and object 7 and their distance is 1.5, then row 10 of Z will contain the values (5, 7, 1.5). This newly formed cluster will have the index 10+30=40. If cluster 40 shows up in a later row, that means this newly formed cluster is being combined again into some bigger cluster.

Cophenet: Cophenetic correlation coefficient

Syntax: c = cophenet(Z,Y)

Description: c = cophenet(Z,Y) computes the cophenetic correlation coefficient which compares the distance information in Z, generated by linkage, and the distance information in Y, generated by pdist. Z is a matrix of size (m-1)-by-3, with distance information in the third column. Y is a vector of size $m.(m-1)/2$. For example, given a group of objects {1, 2, ..., m} with distances Y, the function linkage produces a hierarchical cluster tree. The cophenet function measures the distortion of this classification, indicating how readily the data fits into the structure suggested by the classification.

The output value, c, is the cophenetic correlation coefficient. The magnitude of this value should be very close to 1 for a high-quality solution. This measure can be used to

compare alternative cluster solutions obtained using different algorithms.

The cophenetic correlation between Z(:,3) and Y is defined as

$$c = \frac{\sum_{i<j}(Y_{ij} - y)(Z_{ij} - z)}{\sqrt{\sum_{i<j}(Y_{ij} - y)^2 \sum_{i<j}(Z_{ij} - z)^2}}$$

where:

- Yij is the distance between objects i and j in Y.
- Zij is the distance between objects i and j in Z(:,3).
- y and z are the average of Y and Z(:,3), respectively.

Dendrogram: Plot dendrogram graphs

Syntax: H = dendrogram(Z)

H = dendrogram(Z,p)

[H,T] = dendrogram(...)

[H,T,perm] = dendrogram(...)

[...] = dendrogram(...,'colorthreshold',t)

[...] = dendrogram(...,'orientation','orient')

Description: H = dendrogram(Z) generates a dendrogram plot of the hierarchical, binary cluster tree, Z. Z is an (m-1)-by-3 matrix, generated by the linkage function, where m is the number of objects in the original dataset.

A dendrogram consists of many U-shaped lines connecting objects in a hierarchical tree. Except for the Ward linkage (see linkage), the height of each U represents the distance between the two objects being connected. The output, H, is a vector of line handles.

H = dendrogram(Z,p) generates a dendrogram with only the top p nodes. By default, dendrogram uses 30 as the value of p. When there are more than 30 initial nodes, a dendrogram may look crowded. To display every node, set p = 0.

[H,T] = dendrogram(...) generates a dendrogram and returns T, a vector of length m that contains the leaf node number for each object in the original dataset. T is useful when p is less than the total number of objects, so some leaf nodes in the display correspond to multiple objects. For example, to find out which objects are contained in leaf node k of the dendrogram, use find(T==k). When there are fewer than p objects in the original data, all objects are displayed in the dendrogram. In this case, T is the identity map, i.e., T = (1:m)', where each node contains only itself.

When there are fewer than p objects in the original data, all objects are displayed in the dendrogram. In this case, T is the identity map, i.e., T = (1:m)', where each node contains only itself.

[H,T,perm] = dendrogram(...) generates a dendrogram and returns the permutation vector of the node labels of the leaves of the dendrogram. perm is ordered from left to right on a horizontal dendrogram and bottom to top for a vertical dendrogram.

[...] = dendrogram(...,'colorthreshold',t) assigns a unique color to each group of nodes in the dendrogram where the linkage is less than the threshold t. t is a value in the interval [0,max(Z(:,3))]. Setting t to the string 'default' is the same as t = .7(max(Z(:,3))). 0 is the same as not specifying 'colorthreshold'. The value max(Z(:,3)) treats the entire tree as one group and colors it all one color.

[...] = dendrogram(...,'orientation','orient') orients the dendrogram within the figure window. The options for 'orient' are

'top' Top to bottom (default)

'bottom' Bottom to top

'left' Left to right

'right' Right to left

Appendix 2. Matlab Code:

```
oncesi= A1 + B1 + C1 + E1 + D1;
 sonrasi=A1a + B1a + C1a + E1a + D1a;
oncesiort=mean(oncesi);
sonrasiort=mean(sonrasi) ;
stdoncesi=sqrt( 1/500 * sum((oncesi- ones(500,1)*oncesiort).^2));
stdsonrasi=sqrt( 1/500 * sum((sonrasi- ones(500,1)*sonrasiort).^2));
boxplot(oncesi);
boxplot(sonrasi);
corrcoef(oncesi,sonrasi)
oncesisonrasi=[oncesi sonrasi];
r=corrcoef(oncesisonrasi)
a=find(r>0.7)
d=find(r>=0.7 & r<1)
%K MEANS CLUSTERING
[c1,cmeans] = kmeans(r,2,'dist','cityblock');
[silh,h] = silhouette(r,c1,'cityblock');

ptsymb = {'bs','r^','md','go','c+'};
 for i=1:2
clust=find(c1==i);
plot3(r(clust,1),r(clust,2),r(clust,3),ptsymb{i});
hold on
end
plot3(cmeans(:,1),cmeans(:,2),cmeans(:,3),'ko');
plot3(cmeans(:,1),cmeans(:,2),cmeans(:,3),'kx');
 hold off
grid on
%HIERARCHICAL CLUSTERING
dist = pdist(r,'cosine');
 hclustering = linkage(dist,'average');
cophenet(hclustering,dist)
[h,n] = dendrogram(hclustering,0);
dist = pdist(r,'euclidean');
```

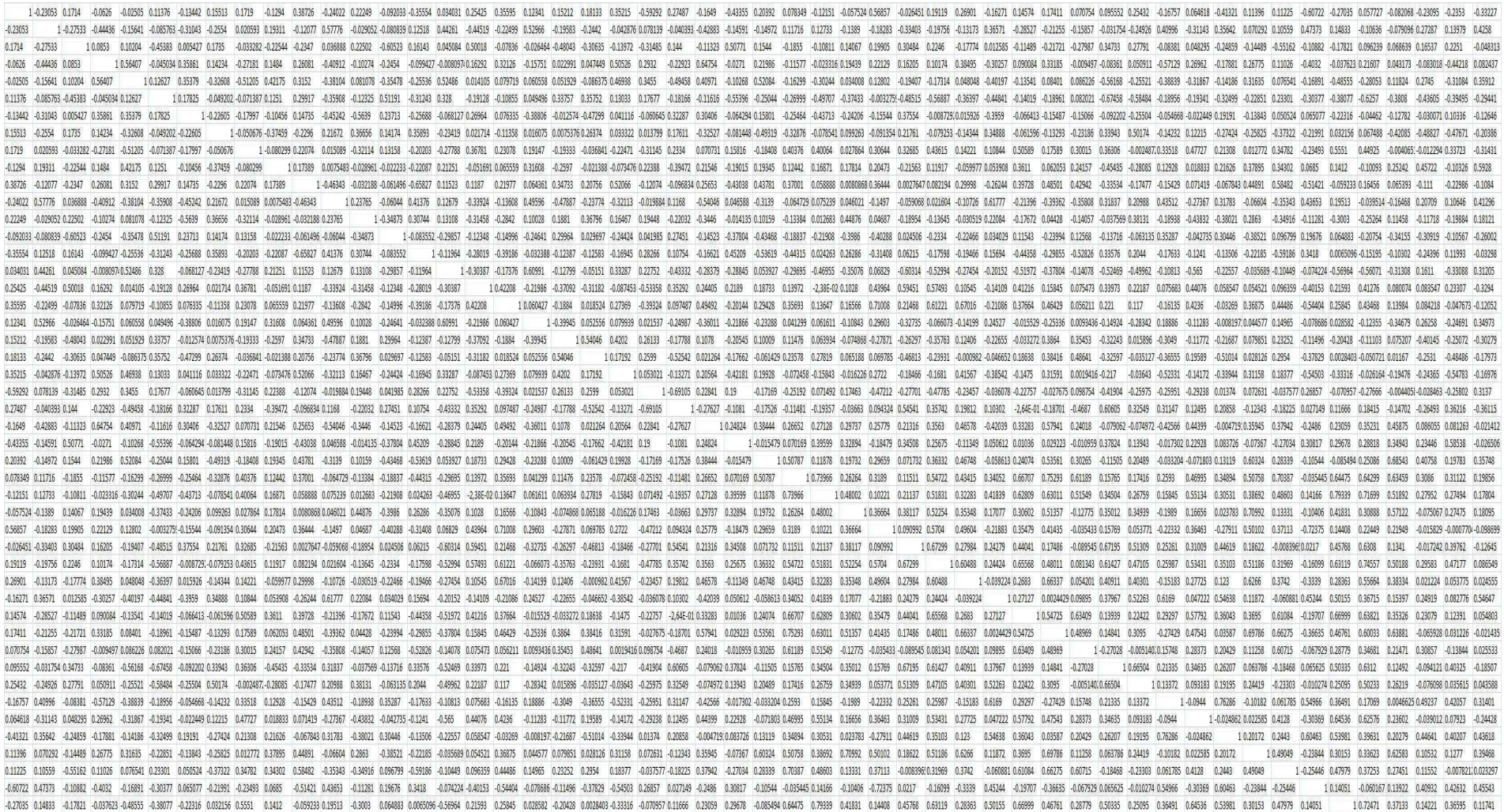
```
hclustering = linkage(dist,'average');
cophenet(hclustering,dist)
[h,n] = dendrogram(hclustering,0);
%subplot%2.sweep subplot çizimi
subplot(2,6,1);
plot(A1(:,2))
subplot(2,6,2);
plot(B1(:,2))
subplot(2,6,3);
plot(C1(:,2))
subplot(2,6,4);
plot(D1(:,2))
subplot(2,6,5);
plot(E1(:,2))
subplot(2,6,6);
plot(oncesi(:,2))
subplot(2,6,7);
plot(A1a(:,2))
subplot(2,6,8);
plot(B1a(:,2))
subplot(2,6,9);
plot(C1a(:,2))
subplot(2,6,10);
plot(D1a(:,2))
subplot(2,6,11);
plot(E1a(:,2))
subplot(2,6,12);
plot(sonrasi(:,2))

%7. Sweep subplot çizimi
 subplot(2,6,1);
 plot(A1(:,7))
 subplot(2,6,2);
 plot(B1(:,7))
 subplot(2,6,3);
```

```
plot(C1(:,7))
subplot(2,6,4);
plot(D1(:,7))
subplot(2,6,5);
plot(E1(:,7))
subplot(2,6,6);
plot(oncesi(:,7))
subplot(2,6,7);
plot(A1a(:,7))
subplot(2,6,8);
plot(B1a(:,7))
subplot(2,6,9);
plot(C1a(:,7))
subplot(2,6,10);
plot(D1a(:,7))
subplot(2,6,11);
plot(E1a(:,7))
subplot(2,6,12);
plot(sonrasi(:,7))
```

Appendix 3. Correlation Matrix of post-stimulus EEG Data

1	0.8759	0.53866	0.19057	0.46052	0.36127	0.68155	0.40975	0.40107	0.69147	0.58622	0.56693	0.246	0.19201	0.539	0.33613	0.64362	0.79223	0.69915	0.34123	0.633	0.56887	0.81009	0.55879	0.61188	0.34586
.8759	1	0.76549	0.20833	0.61645	0.30638	0.70808	0.61377	0.47864	0.85097	0.83432	0.79627	0.32731	0.37368	0.35234	0.67478	0.52136	0.78063	0.84835	0.10015	0.79055	0.808	0.76335	0.487	0.32627	0.29674
.53866	0.76549	1	0.64063	0.41347	0.50029	0.73488	0.62421	0.40338	0.78725	0.81437	0.72443	0.68222	0.5327	0.063356	0.74222	0.33935	0.80974	0.61432	0.1453	0.86129	0.8912	0.73533	0.48215	0.28577	0.27789
.19057	0.20833	0.64063	1	-0.011233	0.62885	0.43351	0.075443	0.1444	0.232	0.33483	0.10003	0.74428	0.2652	-0.14802	0.23838	0.14666	0.6293	-0.01535	0.43688	0.47591	0.42434	0.57391	0.36961	0.46484	0.23439
.46052	0.61645	0.41347	-0.011233	1	0.41997	0.67024	0.52304	0.55059	0.67509	0.43798	0.48208	0.44972	0.43439	0.46756	0.51963	0.64879	0.43369	0.59787	0.058851	0.67974	0.64299	0.27195	0.61471	-0.016311	0.49103
.36127	0.30638	0.50029	0.62885	0.41997	1	0.76061	0.034845	0.66527	0.50564	0.067937	0.16303	0.75863	0.40913	0.5336	0.12419	0.68629	0.69579	0.12089	0.58674	0.59482	0.53813	0.4094	0.8332	0.46879	0.785
.68155	0.70808	0.73488	0.43351	0.67024	0.76061	1	0.58664	0.56329	0.80318	0.47653	0.50863	0.69929	0.44217	0.56287	0.47827	0.82659	0.87183	0.58544	0.47168	0.88043	0.81032	0.67488	0.90605	0.5189	0.67385
.40975	0.61377	0.62421	0.075443	0.52304	0.034845	0.58664	1	0.0062087	0.56374	0.69704	0.58455	0.38681	0.21029	0.047982	0.72963	0.35898	0.47361	0.70595	-0.021786	0.70692	0.67575	0.48315	0.36755	0.1046	0.083874
.40107	0.47864	0.40338	0.1444	0.55059	0.66527	0.56329	0.0062087	1	0.66153	0.11312	0.41954	0.45298	0.5562	0.61879	0.24823	0.5637	0.45137	0.22383	0.12087	0.52338	0.58292	0.16553	0.62609	0.0078026	0.75294
.69147	0.85097	0.78725	0.232	0.67509	0.50564	0.80318	0.56374	0.66153	1	0.65123	0.87218	0.50011	0.7019	0.52491	0.51848	0.60254	0.71119	0.68212	0.086882	0.77466	0.83785	0.5505	0.64314	0.18856	0.39507
.58622	0.83432	0.81437	0.33483	0.43798	0.067937	0.47653	0.69704	0.11312	0.65123	1	0.77379	0.32544	0.35757	-0.14797	0.81694	0.1089	0.62773	0.80544	-0.17519	0.69822	0.76497	0.674	0.14964	0.090726	-0.10223
.56693	0.79627	0.72443	0.10003	0.48208	0.16303	0.50863	0.58455	0.41954	0.87218	0.77379	1	0.26781	0.69734	0.26001	0.53999	0.25087	0.49446	0.74808	-0.24414	0.56236	0.71489	0.4082	0.29911	-0.059265	-0.021639
.246	0.32731	0.68222	0.74428	0.44972	0.75863	0.69929	0.38681	0.45298	0.50011	0.32544	0.26781	1	0.45958	0.1857	0.42553	0.30434	0.62612	0.19898	0.38734	0.74452	0.85113	0.46915	0.67079	0.28169	0.54479
.19201	0.37368	0.5327	0.2652	0.43439	0.40913	0.44217	0.21029	0.5562	0.7019	0.35757	0.69734	0.45958	1	0.22023	0.19014	0.1648	0.3042	0.21587	-0.28368	0.41046	0.55302	0.035474	0.3704	-0.25111	0.09646
.539	0.35234	0.063356	-0.14802	0.46756	0.5336	0.56287	0.047982	0.61879	0.52491	-0.14797	0.26001	0.1857	0.22023	1	-0.18209	0.84976	0.3528	0.26346	0.51109	0.26009	0.193	0.2024	0.74131	0.41226	0.62705
.33613	0.67478	0.74222	0.23838	0.51963	0.12419	0.47827	0.72963	0.24823	0.51848	0.81694	0.53999	0.42553	0.19014	-0.18209	1	0.15963	0.51062	0.7073	-0.17009	0.76904	0.79211	0.48825	0.18459	-0.085161	0.19783
.64362	0.52136	0.33935	0.14666	0.64879	0.68629	0.82659	0.35898	0.5637	0.60254	0.1089	0.25087	0.30434	0.1648	0.84976	0.15963	1	0.63066	0.42391	0.67368	0.61923	0.45915	0.50049	0.90958	0.37909	0.77083
.79223	0.78063	0.80974	0.6293	0.43369	0.69579	0.87183	0.47361	0.45137	0.71119	0.62773	0.49446	0.62612	0.3042	0.3528	0.51062	0.63066	1	0.58136	0.47238	0.83619	0.79974	0.87781	0.71943	0.63006	0.52642
.69915	0.84835	0.61432	-0.01535	0.59787	0.12089	0.58544	0.70595	0.22383	0.68212	0.80544	0.74808	0.19898	0.21587	0.26346	0.7073	0.42391	0.58136	1	-0.015208	0.66644	0.68153	0.58922	0.35576	0.17107	0.11104
.34123	0.10015	0.1453	0.43688	0.058851	0.58674	0.47168	-0.021786	0.12087	0.086882	-0.17519	-0.24414	0.38734	-0.28368	0.51109	-0.17009	0.67368	0.47238	-0.015208	1	0.26522	0.03536	0.53343	0.60702	0.87072	0.56249
.633	0.79055	0.86129	0.47591	0.67974	0.59482	0.88043	0.70692	0.52338	0.77466	0.69822	0.56236	0.74452	0.41046	0.26009	0.76904	0.61923	0.83619	0.66644	0.26522	1	0.92703	0.72653	0.68775	0.33247	0.56945
.56887	0.808	0.8912	0.42434	0.64299	0.53813	0.81032	0.67575	0.58292	0.83785	0.76497	0.71489	0.85113	0.55302	0.193	0.79211	0.45915	0.79974	0.68153	0.03536	0.92703	1	0.61253	0.59477	0.14151	0.4634
.81009	0.76335	0.73533	0.57391	0.27195	0.4094	0.67488	0.48315	0.16553	0.5505	0.674	0.4082	0.46915	0.035474	0.2024	0.48825	0.50049	0.87781	0.58922	0.53343	0.72653	0.61253	1	0.47905	0.76145	0.2899
.55879	0.487	0.48215	0.36961	0.61471	0.8332	0.90605	0.36755	0.62609	0.64314	0.14964	0.29911	0.67079	0.3704	0.74131	0.18459	0.90958	0.71943	0.35576	0.60702	0.68775	0.59477	0.47905	1	0.52934	0.7891
.61188	0.32627	0.28577	0.46484	-0.016311	0.46879	0.5189	0.1046	0.0078026	0.18856	0.090726	-0.059265	0.28169	-0.25111	0.41226	-0.085161	0.37909	0.63006	0.17107	0.87072	0.33247	0.14151	0.76145	0.52934	1	0.38232
.34586	0.29674	0.27789	0.23439	0.49103	0.785	0.67385	0.083874	0.75294	0.39507	-0.10223	-0.021639	0.54479	0.09646	0.62705	0.19783	0.77083	0.52642	0.11104	0.56249	0.56945	0.4634	0.2899	0.7891	0.38232	1

Appendix 4. Correlation Matrix of EEG data



Appendix 5. SAS Code

```
mend _SASTASK_DROPDS;

%LET _EGCHARTWIDTH=0;
%LET _EGCHARTHEIGHT=0;

PROC SQL;
%_SASTASK_DROPDS(SASUSER.PRED1582);
%_SASTASK_DROPDS(WORK.SORT3238);
%_SASTASK_DROPDS(WORK.TEMP5953);
QUIT;

PROC SQL;
      CREATE VIEW WORK.SORT3238
            AS SELECT * FROM WORK.DATA1;
QUIT;
TITLE;
TITLE1 "Logistic Regression Results";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %SYSFUNC(DATE(), EURDFDE9.) at %SYSFUNC(TIME(), TIMEAMPM8.)";
PROC LOGISTIC DATA=WORK.SORT3238
      ;
      CLASS B     (PARAM=EFFECT) E  (PARAM=EFFECT) F  (PARAM=EFFECT)
A0    (PARAM=EFFECT);
      MODEL A=C A0 B E F            /
            SELECTION=BACKWARD
            SLS=0.05
            INCLUDE=0
            CORRB
            LINK=LOGIT
            CLPARM=BOTH
            CLODDS=BOTH
            ALPHA=0.05
      ;

      OUTPUT OUT=SASUSER.PRED1582(LABEL="Logistic regression
predictions and statistics for WORK.DATA1")
            PREDPROBS=INDIVIDUAL
            RESCHI=reschi_A
            RESDEV=resdev_A
            DIFCHISQ=difchisq_A
            DIFDEV=difdev_A ;

      OUTPUT OUT=WORK.TEMP5953 PREDPROBS=INDIVIDUAL
            PREDICTED=_predicted1 RESCHI=_reschi1 RESDEV=_resdev1
DIFCHISQ=_difchisq1 DIFDEV=_difdev1
            DFBETAS=_dfbetas0-_dfbetas1
            H=_h1 C=_c1;
      RUN;
DATA WORK.TEMP5953; SET WORK.TEMP5953; _index_ = _n_; RUN;
RUN;
QUIT;

TITLE;
TITLE1 "Regression Analysis Plots";
AXIS1 MAJOR=(NUMBER=5) WIDTH=1;
AXIS2 OFFSET=(10 PCT) WIDTH=1;
AXIS3 MAJOR=(NUMBER=5) OFFSET=(5 PCT) WIDTH=1;
```

```
GOPTIONS PUBLISH;
GOPTIONS xpixels=&_EGCHARTWIDTH ypixels=&_EGCHARTHEIGHT;
PROC GPLOT DATA=WORK.TEMP5953
      ;
      WHERE A IS NOT MISSING AND
            C IS NOT MISSING AND
            B IS NOT MISSING AND
            E IS NOT MISSING AND
            F IS NOT MISSING AND
            A0 IS NOT MISSING;


      TITLE4 "Deviance Residuals of A by Predicted A";
      SYMBOL1 C=BLUE V=DOT HEIGHT=2PCT INTERPOL=NONE L=1 W=1;
      LABEL _resdev1 = "Deviance Residual of A";
      LABEL _predicted1 = "Predicted A";
      WHERE _resdev1 IS NOT MISSING AND _predicted1 IS NOT MISSING;
      PLOT _resdev1 * _predicted1 /
            VAXIS=AXIS1 VMINOR=0 HAXIS=AXIS3 HMINOR=0 VREF=0
            DESCRIPTION = "Deviance Residuals of A by Predicted A"
      ;
      RUN;


      TITLE4 "Influence (DFBetas) of A by Predicted A";
      SYMBOL1 C=BLUE V=DOT HEIGHT=2PCT INTERPOL=NONE L=1 W=1;
      LABEL _predicted1 = "Predicted A";
      WHERE _dfbetas0 IS NOT MISSING AND _predicted1 IS NOT MISSING;
      PLOT _dfbetas0 * _predicted1 /
            VAXIS=AXIS1 VMINOR=0 HAXIS=AXIS3 HMINOR=0
            DESCRIPTION = "Influence (DFBetas) of A by Predicted A"
      ;
      RUN;

      TITLE4 "Influence (DFBetas) of A by Predicted A";
      SYMBOL1 C=BLUE V=DOT HEIGHT=2PCT INTERPOL=NONE L=1 W=1;
      LABEL _predicted1 = "Predicted A";
      WHERE _dfbetas1 IS NOT MISSING AND _predicted1 IS NOT MISSING;
      PLOT _dfbetas1 * _predicted1 /
            VAXIS=AXIS1 VMINOR=0 HAXIS=AXIS3 HMINOR=0
            DESCRIPTION = "Influence (DFBetas) of A by Predicted A"
      ;
      RUN;

SYMBOL;
QUIT;

TITLE;
TITLE1 "Regression Analysis Predictions";
PROC PRINT NOOBS DATA=SASUSER.PRED1582
;
RUN;

RUN; QUIT;
PROC SQL;
%_SASTASK_DROPDS(WORK.SORT3238);
%_SASTASK_DROPDS(WORK.TEMP5953);
QUIT;

TITLE; FOOTNOTE;
GOPTIONS RESET=ALL;RUN;
```