

ANKARA YILDIRIM BEYAZIT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES



**THE CLUSTERING OF PUBLIC HOSPITALS FOR THE
PRODUCTIVITY SCORECARD APPLICATION**

M.Sc. Thesis by

Ayşe KELEŞ

Department of Computer Engineering

July, 2017

ANKARA

**THE CLUSTERING OF PUBLIC HOSPITALS FOR THE
PRODUCTIVITY SCORECARD APPLICATION**

A Thesis Submitted to

The Graduate School of Natural and Applied Sciences of

Ankara Yıldırım Beyazıt University

**In Partial Fulfillment of the Requirements for the Degree of Master of
Science in Computer Engineering, Department of Computer Engineering**

by

Ayşe KELEŞ

July, 2017

ANKARA

M.Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**THE CLUSTERING OF PUBLIC HOSPITALS FOR THE PRODUCTIVITY SCORECARD APPLICATION**” completed by **Ayşe KELEŞ** under supervision of **Prof. Dr. Fatih V. ÇELEBİ** 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.

Prof. Dr. Fatih V. ÇELEBİ

Supervisor

Prof. Dr. Şahin EMRAH

Jury member

Assist. Prof. Dr. Hilal KAYA

Jury Member

Prof. Dr. Fatih V. ÇELEBİ

Director

Graduate School of Natural and Applied Sciences

ETHICAL DECLARATION

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Date:

Signature:.....

Name & Surname:.....



ACKNOWLEDGMENTS

First of all, I want to thank my supervisor Prof. Fatih V. ÇELEBİ for supporting me, not just about my thesis, about all other issues.

I would also like to thank my colleagues Dr. Hüdaver NURLU (General Surgeon) and Eng. Yusuf Yasin KURTLUK (Industrial Engineer). I benefit from their expert opinions in this work.

I would like to thank: Asst. Ayşe ARSLAN for her valuable and supporting contribution.

Finally, thanks to my husband and my lovely son and two daughters for their support, love and patience.

2017, 10 July

Ayşe KELEŞ

THE CLUSTERING OF PUBLIC HOSPITALS FOR THE PRODUCTIVITY SCORECARD APPLICATION

ABSTRACT

Identifying and Grouping the Roles of Hospitals on an Institutional Basis is one of the study of Ministry of Health. One of the consequences of the restructuring was the publication of Decree Law No. 663. In this context hospitals which are affiliated to Turkey Public Hospital Institution required to be evaluated in 6-month or annual period with this Legislative Decree. As required to this article for evaluation “The Productivity Scorecard Application” has started based on “The Balanced Institutive Scorecard Model”. For each indicator, which is involved in “The Productivity Scorecard Application”, acceptable values have been determined using different methods. In some indicator cards, acceptable value has been regarded as the average of the similar hospitals service classes.

In the study presented in this thesis; it is focused on the study of k-Means clustering algorithm of data mining techniques and hospitals which are affiliated to Turkey Public Hospital Institution and clustering of hospitals whose productivity scorecard will be estimated. Hospitals have been clustered in terms of their financial status, equipment capacity, staff capacity and produced medical services’ volume and variety.

As a result of clustering work carried out within the scope of this study; 597 hospitals to be assessed by The Productivity Scorecard Application were divided into 16 clusters. 149 attributes have been determined for the hospital data used as input in the clustering algorithm and data has been collected from the data of year 2016.

The validity of the clustering results was tested by taking expert opinions and also evaluated according to the distribution ratios of the hospital roles formed in the clusters. After these evaluations, it can be clearly stated that the clustering study has been found to be successful to substantially, taking into account the systematic gathering of data from the application resources and the detection of right outliers.

Keywords: Hospital clustering, k-means clustering algorithm, bisecting k-means, productivity scorecard application.

VERİMLİLİK KARNE UYGULAMASI İÇİN KAMU HASTANELERİNİN KÜMELENMESİ

ÖZ

Hastanelerin rollerinin belirlenmesi ve sınıflandırılması Sağlık Bakanlığı'nın temel kurumsal çalışmalarından biridir. Yeniden yapılanmanın sonuçlarından biri de 663 Sayılı Kanun Hükmünde Kararnamenin yayınlanmasıydı. Bu kararnamede Kamu Hastaneleri kurumuna bağlı hastanelerin 6 aylık ya da yıllık periyotlarda değerlendirilmesi gerektiğine dair madde yer almaktadır. Bu madde gereğince değerlendirme için “Dengeli Kurumsal Karne Modeli”nden yola çıkarak “Verimlilik Karne Uygulaması” başlamıştır. Verimlilik Karne Uygulamasında yer alan her bir gösterge için kabul edilebilir değerler farklı yöntemler kullanılarak belirlenmiştir. Bazı gösterge kartlarında kabul edilebilir değer, benzer hastanelerden oluşan hizmet sınıflarının ortalaması kabul edilmiştir.

Bu tezde sunulan çalışmada; veri madenciliği tekniklerinden olan K-means kümeleme algoritmasına ve Türkiye Kamu Hastaneleri Kurumu'na bağlı ve verimlilik karnesi hesaplanacak hastanelerin kümelenmesi çalışmasına odaklanılmıştır. Hastaneler; mali durum, cihaz kapasitesi, personel kapasitesi ve ürettiği tıbbi hizmetlerin hacmi ve çeşitliliği dikkate alınarak kümelenmiştir.

Bu çalışma kapsamında yapılan kümeleme çalışması sonucunda; verimlilik karne uygulaması tarafından değerlendirilecek 597 hastane 16 kümeye ayrılmıştır. Kümeleme algoritmasında girdi olarak kullanılan hastane veriseti için 149 nitelik belirlenmiş, veriler 2016 yılı verilerinden toplanmıştır.

Kümeleme sonuçların geçerliliği uzman görüşleri alınarak test edilmiş, ayrıca oluşan kümelerdeki hastane rollerinin dağılım oranlarına göre de değerlendirilme yapılmıştır. Yapılan bu değerlendirmelerden sonra açıkça söylenebilir ki kümeleme çalışması verilerin uygulama kaynaklarından sistematik olarak çekilmesi ve doğru sıradışı hastanelerin tespiti de dikkate alınarak büyük ölçüde başarılı bulunmuştur.

Anahtar Kelimeler: Hastanelerin kümelenmesi, k-means kümeleme algoritması, bisecting k-means, verimlilik karne uygulaması

CONTENTS

M.Sc. THESIS EXAMINATION RESULT FORM.....	ii
ETHICAL DECLARATION	iii
ACKNOWLEDGMENTS	iv
ABSTRACT.....	v
ÖZ	vi
NOMENCLATURE.....	ix
LIST OF TABLES	x
LIST OF FIGURES	xi
CHAPTER 1 - INTRODUCTION.....	1
1.1 The Productivity Scorecard Application.....	1
1.2 General Information about Data Mining.....	3
1.3 Related Works.....	4
1.4 The Aim of Work.....	6
1.5 Outline.....	7
CHAPTER 2 - MATERIAL AND METHODS.....	8
2.1 Data Mining Components	8
2.1.1 Database	8
2.1.2 Data Warehouse.....	8
2.2 Clustering	9
2.3 K-Means Algorithm	11
2.3.1 Limitations.....	13
2.4 Similarity Function	14
CHAPTER 3 - DATA PRE-PROCESSING	15
3.1 Data Selection	16
3.2 Data Cleaning.....	17
3.3 Data Integration.....	18
3.4 Data Transformation	19
3.5 Data Reduction.....	23
CHAPTER 4 - IMPLEMENTATION.....	26
4.1 Source Databases	27
4.2 The Data Warehouse Structure	28
4.2.1 Dimension Tables.....	28

4.2.2 Views and Staged Tables.....	30
4.3 K-Means Clustering	31
4.3.1 Hierarchical Clustering Algorithm	32
4.3.2 Standard K-Means Algorithm	32
4.3.3 Bisecting K-Means (a variant of K-Means)	33
CHAPTER 5 - RESULTS.....	36
CHAPTER 6 - DISCUSSION AND CONCLUSION.....	39
REFERENCES.....	41
APPENDICES	44
Appendix A - List of Attributes in the Hospital Dataset.....	45
Appendix B - Referral Rate Taken by Emergency Department Indicator Card	51
CURRICULUM VITAE.....	52

NOMENCLATURE

MoH	Ministry of Health
TPHI	Turkey Public Hospitals Institution (Türkiye Kamu Hastaneleri Kurumu)
NHS	National Health System
HCRS	Health Coding Reference Server
MRMS	Main Resources Management System
PRMS	Personnel Resources Management Systems
PPDS	Performance Payment Data System
DRMS	Device Resource Management System
ASS	Accounting Software System
BHSM	Basic Health Statistics Module
ICU	Intensive Care Unit
EMPI	Examinations and Medical Procedures and Interventions
PL/SQL	Procedural Language extension to Structured Query Language
OLTP	On-line Transaction Processing
SPSS	Statistical Package for the Social Science
SQL	Structured Query Language

LIST OF TABLES

Table 1.1 Indicator card list whose acceptable values are the average of the hospital clusters.....	2
Table 3.1 The list of hospitals affiliated to TPPI and their roles registered by MoH.	15
Table 3.2 The list of applications used in the MoH	18
Table 3.3 The details of the data integration processes.....	19
Table 3.4 Device Groups which has more than membership.....	21
Table 3.5 The Result of Correlation Analyze for Dimensionality Reduction.....	23
Table 4.1 Detailed technical information about database objects in the source databases.	28
Table 4.2 EMPI groups according to their score.....	30
Table 4.3 Information about CLSS_SURGEON_OPERATION table partitions.....	31
Table 4.4 For 10,15 and 20 k values, number of cases in each cluster	32
Table 4.6 Member count of final clusters.....	35
Table 5.1 Distribution ratio of all roles in clusters.....	37

LIST OF FIGURES

Figure 1.1 The steps of data mining and KDD[4].....	3
Figure 1.2 Architecture of a typical data mining system.	4
Figure 2.1 Architecture of a Data Warehouse [19].	9
Figure 2.2 The dendogram representation for hierarchical clustering of ten data objects	11
Figure 2.3 The K-Means partitioning algorithm (4).	12
Figure 2.4 Clustering of a set of objects based on the K-Means method. (The mean of each cluster is marked by a “+”)......	13
Figure 3.1 Aggregation and generalization about EMPI information on NHS data.	22
Figure 4.1 Flowchart of the implemented study	27
Figure 4.2 Hospital dimension table set up logic and implemented design	29
Figure 4.3 Schematic representation of views and staged tables created in data warehouse.....	30
Figure 4.4 Basic Bisecting K-means Algorithm for finding K clusters.	33
Figure 4.5 The formation steps of clusters with the bisecting k-means algorithm. ..	34
Figure 5.1 Maximum Ratio of hospitals has same Role in Cluster.....	36

CHAPTER 1

INTRODUCTION

One of the purposes of institutions that take data from many sources and store them in their databases is to convert the raw data to information. One of the main topics of this process called data mining is clustering. In data mining, clustering is the process of grouping and bringing together objects according to their similarities using certain properties of them.

Clustering is done in different areas using different methods for different purposes. The aim of this study is clustering of hospitals connected to TPHI using K-Means clustering algorithm to calculate acceptable value for productivity scorecards in “The Productivity Scorecard Application”. The study is also expected to assist the study of Ministry of Health about Identifying and Grouping the Roles of Hospitals on an Institutional Basis. In this chapter, it is explored basic concepts of data mining processes and also explain The Productivity Scorecard Application of TPHI.

1.1 The Productivity Scorecard Application

The Public Hospitals Association of Turkey was established with the "Restructuring", the second step of the "Health Transformation Program" which started in 2003. With the establishment of The Public Hospitals Institution, the integration of hospital management in the periphery, the formation of associations and the appointment of contracted administrators have been realized.

The "Hospitals; Medical and financial criteria and quality, patient and employee safety and education criteria according to the procedures and principles to be determined by the Institution for a period of six months or one year " clause is included in the decision of law no. 663 which caused the change of structure.

TPHI has begun to implement “The Productivity Scorecard Application” for evaluation by taking this law sanction to evaluate the productivity of resource using and service supplying. For evaluation “The Productivity Scorecard Application” has formed based on “The Balanced Scorecard Model” [1]. Based on this application, annual reports were prepared and shared with the collocutor between 2013 and 2016 [2].

The prepared indications are categorized in different dimensions. A separate indicator card has been prepared for each indicator in the hospital productivity scorecard in the Productivity Scorecard Application. There are 4 dimensions and 91 indicator cards in this scorecard application. Hospital scorecards dimensions are composed of I- Health Services Management, II- Financial Services Management, III- Administrative Services Management and IV- On-Site Assessment. Calculation methods and scoring were done for each indicator card.

For each indicator, acceptable values were determined using different methods. While for some indicators acceptable values have been regarded as the average of the service classes, in some cases the Ministry of Health and Institutional targets are taken as acceptable value. Most of these methods have been regarded as the average of the service classes. Indicator cards whose acceptable values are the average of the hospital clusters is listed in Table 1.1.

Table 1.1 Indicator card list whose acceptable values are the average of the hospital clusters

Department	Indicator Card
Emergency Services	Emergency service referral intensity
	Percentage of patients referred to the emergency department
	Emergency service mortality rate
	Emergency department rate of patients re-applied within 24 hours
Polyclinic Services	Clinician Number of patients per physician per day
	Patient admissions
	Patient's average waiting time for examination
Inpatient Services	Emergency Hospital Income Rate
	Bedding
	Inpatient Service Case Completion
	Inpatient Service Mortality Rate
Intensive Care Services	Brain Death Notification Rate
	Intensive care mortality rate
	Intensive Care Case Completion
Operating Room and Birth Services	Number of operations per surgery table
	Average number of days after surgery
Other services	Radiology Reasoning Ratio
	Certified Employee Status
	Proportion of Assistant Health Personnel Working in Clinic Care
	Medical Waste Produced per Bed

1.2 General Information about Data Mining

Data mining is simply the process by which a previously unknown, valid and applicable information is obtained from a heap of data through a dynamic process. Data mining is also a discipline that can be used in different domains to discover unknown significant knowledge [3].

Alternative as Data Mining actually is accepted as a part of the process of information discovery. The simplest definition of data analysis is the collection, organization, modeling and experimentation of data access. The steps of the discovery of information process are given in the Figure 1.1 and detailed below;

- Data Selection (Which is related to the analysis task retrieved from the databases)
- Data Cleaning (removing noisy and inconsistent data)
- Data Integration (combining many data sources)
- Data Transformation (To convert the data to useable data for Data Mining Technology)
- Data Mining (Apply intelligent data mining methods to catch data patterns)
- Pattern Evaluation (Defining interesting patterns represent information which obtained according to some measurements)
- Presentation (Presentation of information to the user) [4]

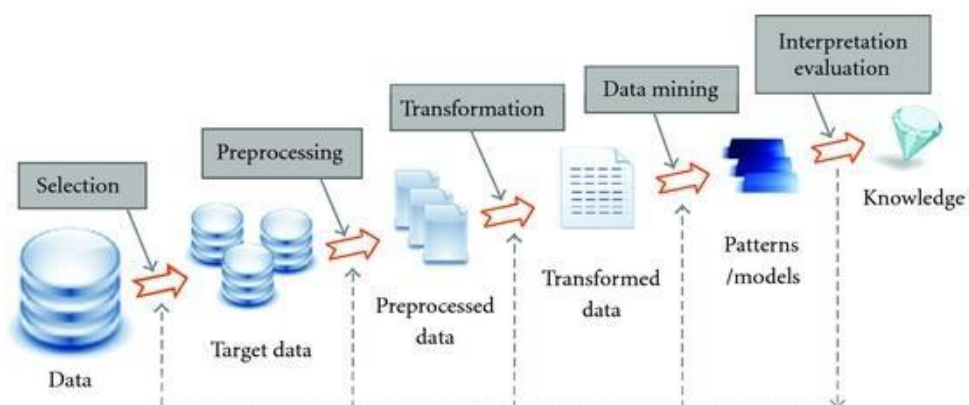


Figure 1.1 The steps of data mining and KDD[4]

Another point of view, data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. Based on this view, the architecture of a typical data mining system may have the Database, data warehouse, Worldwide Web, or other information repository, Database or data warehouse server, Knowledge base, Data mining engine, Pattern evaluation module and User interface(Figure 1.2) [4].

The input of data mining algorithms consists of a group of instances (rows, examples or observations). Every record is described by a number of attributes or columns which are considered to be either nominal or numeric and a class label which represents the outcomes of any observation. The set of observation is called a dataset [5].

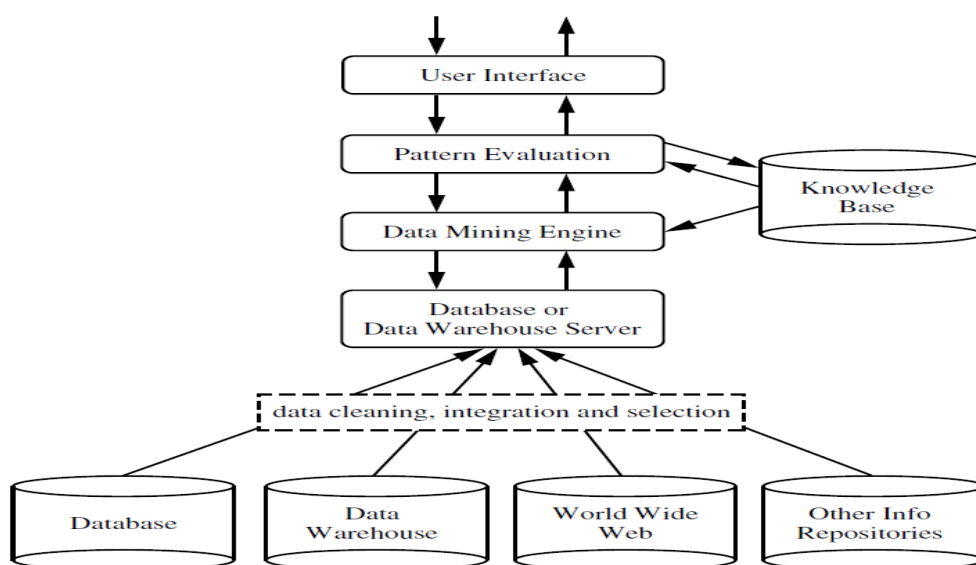


Figure 1.2 Architecture of a typical data mining system.

1.3 Related Works

The structure of the hospitals varies from a wide range, starting from architectural, laboratory services, operating room services, as well as hotel services, food and beverage and cleaning services, financial services and human resources management services. The grouping of hospitals in response to this diversity and complexity is generally aimed at a specific field and field success, such as clustering hospitals according to the organizational and financial performance indicators [6].

It was interesting to have a clustering work done on whole hospital specifications like the study method of this thesis [7]. The aim of the study which is different from this work was measuring evaluation index between the development status and the competitiveness of the hospitals. 567 hospitals were evaluated in this thesis, while the number of hospitals in the other study was 14. The data set formed in this study is much larger also.

Measuring the quality performance of hospitals in health care services improves healthcare outcomes of diagnostic procedures or specific treatment episodes. In addition, different modeling studies have been conducted for hospitals has large unobserved heterogeneous data [8]. Grouping according to the quality of health care [9] is another work from a clinical quality standpoint. The evaluation of hospitals was based not only on their own resources and outputs but also on patient hospitalized evaluations [10].

There are various studies on public hospitals in Turkey. It was aimed to examine the operational performance of general public hospitals of the MoH following the implementation of “The Health Transformation Program in Turkey” in one of these studies [11]. In another study, analyzing the effect of competition on technical efficiency for the hospital industry in Turkey, which are undertaken within the Health Transformation Program in Turkey are enhancing efficiency and increasing competition [12].

Many different methods have been used to improve the K-Means algorithm. The Bisecting K-Means algorithm [13], one of these studies, was implemented in this thesis. This algorithm starts with a single cluster containing all the objects. For each step, a cluster is selected to divide, 2 subsets according to the K-Means algorithm are found (Bisecting step). The previous step continues until the desired number of clusters is reached [13]. Due to member count constraint, a different way from the method applied in this paper was developed to choose which cluster is split. It has also been put forward that the Bisecting K-means algorithm performs better depending on the generation of relatively equal-sized clusters instead of clusters of widely varying sizes.

On each level divided by two in the Bisecting K-Means algorithm, there is no way of reassembling objects at different side of binary tree. “Cooperative Bisecting K-Means Clustering (CBKM)” method overcomes this drawback. It concurrently combines the results of the Bisecting K-Means and K-Means at each level of the binary hierarchical tree using cooperative and merging matrices [14].

In this study, a different way was implemented to determine the number of iterations in the Bisecting K-Means algorithm. In a different approach, which is proposed in the same subject, a limit is set on the number of iterations. In one of this study [15], it is presented that “limited-iteration Bisecting K-Means” with three iterations led to higher computing efficiency when compared with the Bisecting K-Means.

1.4 The Aim of Work

TPHI was established with the publication of the Decree Law No. 663 on "Health Transformation Program-Turkey", which started in 2003, and a clause indicating that the hospitals connected to Public Hospitals institution should be evaluated within 6 months or annual periods. In this context, the Balanced Scorecard Model for evaluation started with The Productivity Scorecard Application. A separate indicator card was prepared for each indicator in the hospital productivity score card. Acceptable value for some indicator cards is accepted as the average of the cluster with similar, while for some indicators the MoH and TPHI targets are taken to be calculated as acceptable value. The need for clustering of hospitals has arisen in order to be able to calculate indicator cards, which is the acceptable value is average of the hospitals in the same cluster.


Because of the variable structure of the hospitals, this clustering study should be repeatedly and systematically done every year. The hospital clustering work is recreated with the data from the previous year at the beginning of each year and included in The Productivity Scorecard Application. In this thesis, clustering of hospitals which are affiliated to TPHI was aimed using 2016 data.

In this study, 597 hospitals clustered and the hospital data set, which is the input of the cluster, consists of 109 qualifications. Classification of the hospitals was done according to the K-Means clustering algorithm and implemented is done with SPSS. As a result of the clustering, 597 hospitals in the Productivity Scorecard Application were divided into 16 clusters. The analysis of the results is described in the following sections.

1.5 Outline

The structure of this thesis consists of five main chapters. The first chapter gives general information about the subjects of the work done. Chapter two includes data pre-processing applied in this thesis is surveyed. At this stage, data integration, data cleaning, data conversion and data reduction are discussed. The next chapter, chapter 3, presents the experimental study implemented in the study; technical infrastructure and procedures established and evaluation of two kind K-Means algorithms applied.

Next to last chapter, chapter 4, the experimental results obtained are discussed. Finally, discussions, conclusions and some ideas for future work were held in chapter 5 and 6.



CHAPTER 2

MATERIAL AND METHODS

2.1 Data Mining Components

2.1.1 Database

Data Base Management Systems (DBMS) are programs developed to serve anyone who needs to manage very large amounts of data. The relational database approach designs data as tables within a given normalization rules, establishing a relationship between these tables through a primary key and a foreign key [16]. Within the scope of this thesis, the majority of data belong to hospitals which subject to classification were taken from relational databases. The data set generated in this study was prepared mostly from the operational databases of the applications operated in the Ministry of Health. Detailed information about mentioned databases is provided in the following sections.

2.1.2 Data Warehouse

Data Warehouse (DW) is a system used for reporting and data analysis, and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate sources [17].

From a technical point of view, a data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but can include data from other sources. In addition to a relational database, a data warehouse environment can include an extraction, transportation, transformation, and loading (ETL) solution, statistical analysis, reporting, data mining capabilities and other applications that manage the process of gathering data, transforming it into useful and delivering it to business users [18].

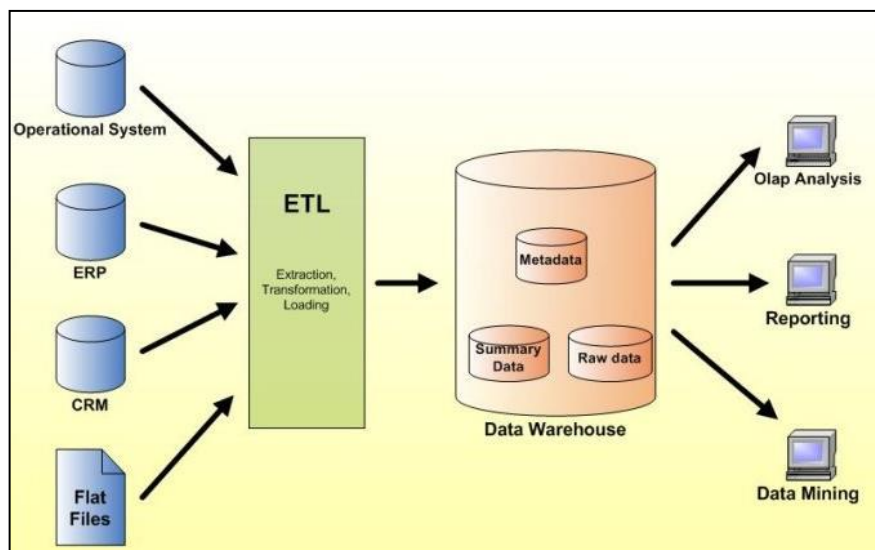


Figure 2.1 Architecture of a Data Warehouse [19].

From a data warehouse perspective, data mining can be viewed as an advanced stage of on-line analytical processing (OLAP). However, data mining goes far beyond the narrow scope of summarization-style analytical processing of data warehouse systems by incorporating more advanced techniques for data analysis.

2.2 Clustering

Data mining models as general acceptance are gathered under three main headings; classification, clustering and association rules, sequential patterns. Classification methods are covered under *supervised learning* heading in machine learning terminology. Cluster analysis is also a classification approach. However, there is no dependent variable value in cluster analysis, only classifications are made according to the attribute values of objects. The clustering approach is grouped as *unsupervised learning* in machine learning since there is no link between dependent and independent variables in the clustering process [20].

Cluster analysis or clustering is basically the decomposition of similar objects placed in the data array into the same groups. In cluster analysis, it is aimed that the objects placed in the same cluster are as homogeneous as possible and the objects placed in different clusters are as heterogeneous as possible [20].

Data clustering has been received considerable attention in many applications, such as data mining, document retrieval, image segmentation and pattern classification. The enlarging volumes of information emerging by the progress of technology, makes clustering of very large scale of data a challenging task. In order to deal with the problem, many researchers try to design efficient parallel clustering algorithms [21].

Clustering is named as *data segmentation* because clustering divides large data sets into groups according to similarities. Outliers (objects not entering any clusters) at the end of the clustering may be more interesting. As a data mining function, cluster analysis can be used to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. In this study, the hospitals that did not include any group in the clusters gave the experts some opinion that some exceptions should be observed in the evaluation of these hospitals.

Clustering methods have evolved with the contributions of all scientists from taxonomy, psychology, biology, statistics, social sciences and engineering. Cluster analysis methods were developed with two different approaches; *hierarchical* and *partitioning*. One of the most popular *partitioning clustering algorithms* is K-Means algorithm.

Hierarchical clustering algorithms successively produce nested clusters in two different ways; *agglomerative process* (starts with each object as individual clusters and successively merges the most similar or closest objects or groups until all of the groups are merged into one) and *divisive process* (starts with all of the objects in the same cluster and, at each step, a cluster is split up into smaller clusters, until a termination condition holds. The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendrogram. This tree graphically displays the merging process and the intermediate clusters. The dendrogram in Figure 2.2 shows how ten objects can be merged into a single cluster by agglomerative approach.

Agglomerative techniques are more common, and these are the techniques that implemented in this study for comparing with K-means and its variant.

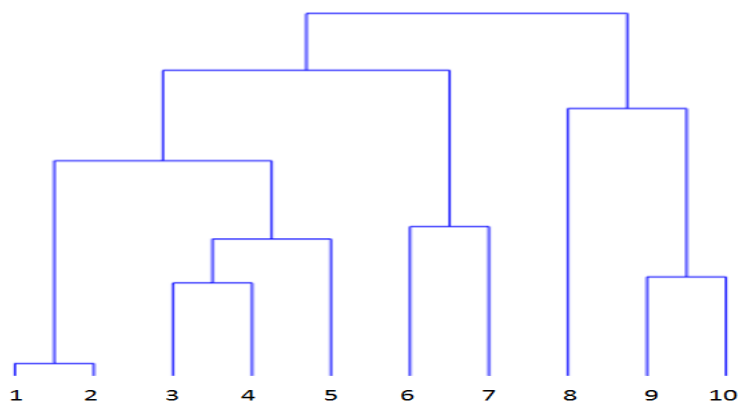


Figure 2.2 The dendrogram representation for hierarchical clustering of ten data objects

2.3 K-Means Algorithm

K-Means algorithm one of the most popular clustering algorithms in data mining, not just within clustering algorithms but within all data mining algorithms [22]. It was first published in 1955. This algorithm has been discovered by several researchers from different disciplines. The most important of these are Lloyd (1957, 1982) [23], Forgey (1965), Friedman and Rubin (1967) and McQueen (1967). A detailed history of K-Means with explanations of the various variations is given in [24]. It is still widely used, although it has been published 50 years ago and many clustering algorithms have been available. This also indicates the difficulty of designing a general purpose clustering algorithm [25].

The K-Means algorithm is dividing a given dataset into previously defined number of cluster(K), the input parameter, k , and a set of d -dimensional vectors, $D = \{x_i \mid i = 1, \dots, N\}$, where $x_i \in \mathcal{R}^d$ denotes the i th data point. Each of the resulting clusters in itself has high similarity and dissimilarity with the other clusters. Cluster similarity which can be viewed as cluster's centroid calculated by taking the average of the values of the objects in the cluster. The algorithm starts with by picking up k of the objects, each of which is also the start cluster center. For each of the remaining objects, then the algorithm iterates between two steps till convergence:

Step1: Data Assignment. Take an object and assign it to the nearest cluster, depending on the distance between the object and the cluster mean. Recalculate the mean of the cluster after each assignment.

Step2: Relocation of “means”. Recalculate the mean of the cluster after assignments. If there is no change about assignment (and hence the c_j values) end the algorithm, otherwise return to step 2.

Algorithm: k -means. The k -means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- k : the number of clusters,
- D : a data set containing n objects.

Output: A set of k clusters.

Method:

- (1) arbitrarily choose k objects from D as the initial cluster centers;
- (2) **repeat**
- (3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
- (4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
- (5) **until** no change;

Figure 2.3 The K-Means partitioning algorithm (4).

The algorithm execution is visually depicted in Figure 2.4. Each iteration needs $N \times k$ comparisons, which determines the time complexity of one iteration.

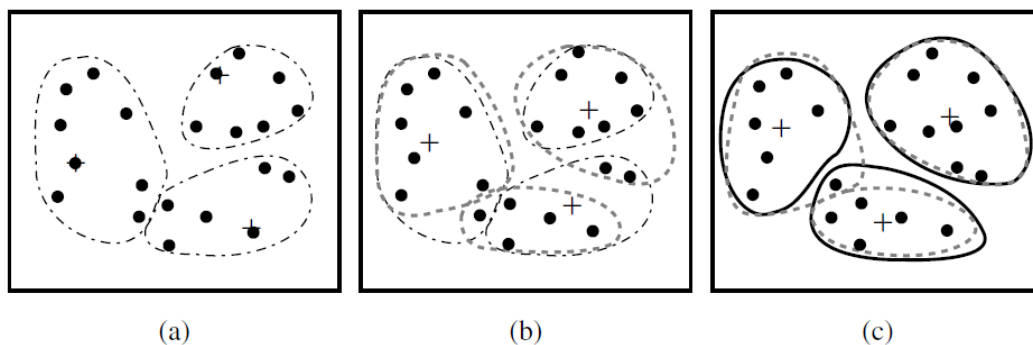


Figure 2.4 Clustering of a set of objects based on the K-Means method. (The mean of each cluster is marked by a “+”.)

2.3.1 Limitations

Cluster Model: The K-Means is not a suitable method for discovering data sets that are non-spherical, densities of which are different and which have clusters of different sizes. The concept is based on global clusters that may be separate for convergence towards the center of the mean value.

Difficult to predict K-Value: If the number of clusters (k) is not known in advance, the algorithm will be run for different k values and found the appropriate criterion to select one of the results. Some criteria were taken into account in determining the number of clusters in the study of the classification of hospitals, the maximum number of clusters and the minimum number of member for clusters. For this study, the maximum number of clusters was twenty-five and the minimum number of member for clusters was five.

Sensitive to noisy data: The algorithm is also sensitive to the presence of outliers, since “mean” is not a robust statistic. A preprocessing step to remove outliers can be helpful. Post-processing the results, for example to eliminate small clusters, or to merge close clusters into a large cluster, is also desirable.

Categorical values: Although the K-Means algorithm has a high efficiency in clustering large data sets, categorical data cannot be used directly because of it’s working with numerical data only [26]. The hospital dataset used in this study does not contain categorical data. The reason of why categorical data isn’t included is not

only by the limit of the algorithm but also to clustering hospitals according to the order of their size index. The hospital dataset used in this study does not contain categorical data. One aim is to classify not only by the limit of the algorithm but also by the order of the hospitals according to their size index.

2.4 Similarity Function

Similarities are a set of rules that are used to group or separate objects. Similarity or distance measures can be made in one dimension or in many dimensions. Every dimension is used to group objects. In this study, *Euclidean distance* measure was used for the calculation of distance between two objects in multi-dimensional space. This measurement is simply the geometric distance between two objects in space and calculated as;

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \quad (1.4.1)$$

Where d_{ij} : i. and j. distantness of the unit, x_{ik} : k variable of i. unit value, x_{jk} : k variable of j. unit value, $i = 1, \dots, n$; $j = 1, \dots, n$ ve $k = 1, \dots, p$. n is unit number and p is variable number.

CHAPTER 3

DATA PRE-PROCESSING

In this study, while beginning to create the data set, first, actual list of public hospitals to be clustered was prepared. Among these, which have been active more than six months in the related year, financially independent and general service hospitals (not branch) were chosen to be subjected for the application score card application. From the list of hospitals A1 branch, A2 branch and Dental Health Centers and Dental Health Hospitals were excluded. 597 general health facilities were subjected to this study. Table 3.1 includes the list of hospitals affiliated to TPHI and their roles registered by MoH.

Table 3.1 The list of hospitals affiliated to TPHI and their roles registered by MoH.

Hospital Role (Registered by Ministry of Health)	Clustered Hospitals	The Number of Hospitals
ADSH	NO	17
ADSM	NO	126
AI	YES	52
AI_SUBSPECIALTY	NO	24
AII	YES	63
AII_SUBSPECIALTY	NO	42
B	YES	124
C	YES	155
D	YES	121
EI	YES	82
	Total Number of Clustered Hospital	597

In the data pre-processing phases performed on the data that constitutes the dataset to be clustered, data cleaning (determination of incompatible data), data integration (creation of cross tables), data transformation (e.g. evaluation of personnel according to title coefficients, the sum of operations according to groups, etc.) studies were carried out.

3.1 Data Selection

The choice of features in the hospital data was determined especially by considering the cards of the Productivity Scorecard Application. For example, when the emergency services of hospitals were evaluated, qualifications related to the emergency departments were selected. One of the most important issues considered in the selection of qualities was whether this data was in the application databases and whether it provided the required conditions. A qualitative data which cannot be obtained systematically from the database was ignored because lack of possibility of manual obtaining from large number of hospitals covering the entire country. Some data were available in the database but it was excluded because the definition of the quality and the definition of the data in the application do not match. Qualifications were grouped under four titles; financial, equipment, personnel and service. The properties in the service group were also grouped as emergency, intensive care and other services. The groups and names of the selected attributes are listed in Appendix-A.

The data were obtained from four national applications of the ministry of health - National Health System (NHS), Main Resources Management System (MRMS), Accounting Software System (ASS) and Performance Payment Data System (PPDS).

The NSS is an electronic record system extending up to tele medical applications based on the sharing of a functional database of all citizens, accessible to each individual's own information, starting from the birth of the individual and consisting of health-related data for the whole lifetime, in a high-bandwidth within a communication framework. It is composed of a health management system module that includes an e-Pulse module and a National Clinical Data Storage built in MoH where the citizens of the country can access the self-health information (medical history) produced including examination, laboratory test results online from a single location. The E-pulse application data is managed in a no-sql database while the NHS data is stored in a relational data store. The data of services provided by the hospitals are obtained from the NHS data.

Main Resources Management System consists of; The Human Resources Management System (HRMS) controlling of personnel movements, the Material Resources Management System (MKYS) where all movements of materials in the Ministry of Health are actively monitored, the Investment Tracking System (YTS) The Private Health Institutions Management System (SKYS) where all the procedures of the private health institutions are followed and the Basic Health Statistics Module (TSIM) used to collect health statistics throughout the country. TSIM is a data collection application in which the data of the information determined by the Ministry of Health is recorded numerically by the authorized personnel of the health facility in monthly periods.

3.2 Data Cleaning

It brought problems with missing, inconsistent and noisy data because it is a national study involving a large number of hospitals and data taken from different applications. Data cleaning is filling missing values with appropriate data and correcting processes data noises and data inconsistencies. Sorting method was used to identify dirty and inconsistent data. According to their roles, the hospitals contrary to the order were identified and some of data was corrected in the application layer, while some of them filled with average values according to hospital role.

Many methods are chosen to complete missing values, ignoring the tuple, filling the data manually, filling with the average data, and so on. It was impossible for record to be ignored because the hospitals which are evaluated within scorecard application have be in a cluster. Because we are working with great data, manual filling of the data is not feasible because it takes a lot of time. Since the fact table is created automatically from the operational databases by the procedures written in the database layer, it updates all data when the procedure is executed after updating in the source code. So the update required rewriting the manually filled data, some of the work to complete the data was corrected at the application layer while some were filled in by taking averages according to the roles the hospitals were in. The most time consuming method of data cleaning was to correct the missing data to the application users.

3.3 Data Integration

The success of information discovery is proportional to the data harmony, especially when shortening time. In this study, data incompatibility was encountered because the data were taken from different applications. The data were taken from 6 application databases in the Ministry of Health. The list of these applications is given in the Table 3.2.

Table 3.2 The list of applications used in the MoH

1	NHS	National Health System
2	PRMS	Personnel Resources Management Systems
3	PPDS	Performance Payment Data System
4	DRMS	Device Resource Management System
5	ASS	Accounting Software System
6	BHSM	Basic Health Statistics Module

For data integration, Metadata defined in the Health Coding Reference Server (HCRS), which many applications in the MoH referenced for integration, is used at data warehouse. HCRS is a reference information system that shares open technologies (with XML Web Services). In order to provide common coding/classification systems that are available to all healthcare players, MoH Department of Information Processing developed HCRS which encapsulates all the international and national coding systems used in Turkey within a publicly accessible server. Some of the coding systems available from HCRS are ICD-10 [27], Drugs, ATC (Anatomic, Therapeutic, and Chemical Classification System), Associations, Clinics, Specialization, Careers, National Health Tariffs, Health Application Instructions, Supplies, Vaccines, Baby Monitoring Calendar, Pregnant Monitoring Calendar, Child Monitoring Schedule and Parameters [28].

The details of the processes during data integration in this study are summarized in the Table 3.3.

Table 3.3 The details of the data integration processes.

Source	Attribute	Integration Process	Description
All	Year	Type Transformation	Type of Year attributes(date, char, string) transformed to number
All	Hospital Codes	Type Transformation	Type of the Citizen Number attribute(char, varchar) transformed to number
PRMS	Personnel Title	Structure Transform	The titles, not determined coefficient of them, included "other" group by case statement
PRMS, PPDS	Citizen Number	Type Transformation	Type of the Hospital Codes(char, varchar) transformed to number
ASSD	Total Accrue	Cross Table	Cross table includes ASSD Hospital Code column and HCRS Hospital Code column
PPDS	Physician Numbers	Cross Table	Cross table includes PPDS Physician Title Codes column and PRMS Title Codes column

3.4 Data Transformation

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve *smoothing*, mentioned above to remove noise from the data, *aggregation*, *generalization of the data*, *normalization* and *attribute construction* (or feature construction) [4]. While preparing the hospital dataset, *aggregation*, *generalization* and *normalization* operations were performed for data transformation.

Aggregation: Where summary or aggregation operations are applied to the data. This issue is typically used in constructing a data cube for analysis of the data [4]. Because all the data in the application databases were more detailed than their intended purpose,

aggregation was done for all attribute of hospital dataset except “sege_endex” attribute. Due to safety concerns ASS application granted select permission for materialized view includes aggregated data.

Sum and *Avg* functions of aggregation was used for this study. In some cases, it was necessary to get a total, while in some cases the average was taken. For example, the number of beds of a hospital and the number of hospitalized patients per month were taken as the median number of beds and total number of hospitalized patients during transformation from the monthly data to the annual data.

Generalization: Generalization of the data, where low-level or “primitive” (raw) data are replaced by higher-level concepts through the use of concept hierarchies [4]. In this study, generalization process is made for time information, personnel information, EMPI [29]. (Surgery, venture, birth), emergency service triage information and medical devices. All time information of attributes was generalized to year. By the nature of the application, data are kept at the level of personnel in PRSM database. Since the personnel number attribute of the hospitals is based on the title, it is generalized according to the title of personnel information. EMPI data were generalized to groups under surgical, interventional, and birth headings. NHS application holds a record for each EMPI performed on the patient. Each record in the table includes hospital code, EMPI code, the time of EMPI, the patient's application information, the physician who performed the transaction, and the time to send the transaction to the system. The generalization process in NHS data was performed after the aggregation by EMPI code and date, shown in Figure 3.1. It was inevitable to make generalizations as the DRSM application devices because medical devices diversity was very high. For instance, 15 different rows of information in the database for each different types of operating tables in practice, as a result the "operating table" was taken as a top-level concept. Table 3.4 lists the number of different devices in each upper level device group which has more than one different device.

Table 3.4 Device Groups which has more than membership

Top-Level Device Group	The Number of Device Diversity	Top-Level Device Group	The Number of Device Diversity
Tomography	8		
Anesthesia Device	57	Heart-Lung Pump	3
Angiography Device	6	Chemotherapy Preparation	5
Arthroscopy	4	Colonoscopy-Sigmoidoscopy	12
Brachytherapy Device	2	Laparoscopy	4
Cryotherapy Device	2	Magnetic Resonance	15
Defibrillator	10	Nst/Cardio Tachograph	2
Dialysis / Renal Replacement	8	Audiometry - Tympanometry	6
Duodenoscopy	4	Puva	2
EEG (Electro Encephalography)	8	Radiographic Imaging	31
Eswt / Rswt / Shock Wave	3	Cystoscopy-Ureteroscopy	12
Phacoemulsification - Vitrectomy	5	Spect System	2
Phototherapy	22	Tomography	2

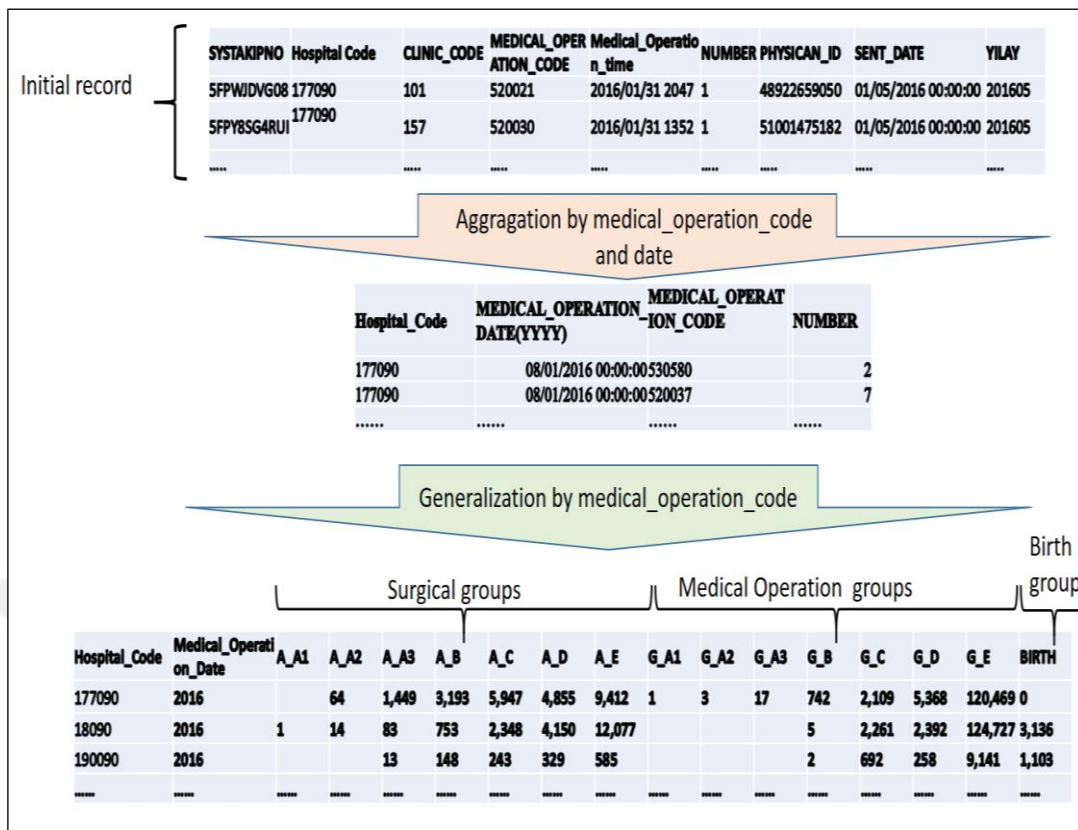


Figure 3.1 Aggregation and generalization about EMPI information on NHS data.

Normalization (z-score): Data normalization is required to remove dependency on the data unit and for equal weighting data. Normalization is particularly useful for classification algorithms involving neural networks, or distance measurements such as nearest-neighbor classification and clustering [4]. There are many methods for data normalization. In this study z-score normalization is dealt with. For our discussion, let A be a numeric attribute with n observed values, r_1, r_2, \dots, r_n . In z-score normalization (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean and standard deviation of A. A value, r_i , of A is normalized to $r_{iZScored}$ by computing

$$r_{iZScored} = \frac{r_i - \bar{A}}{\sigma_A} \tag{2.4.1}$$

Where \bar{A} and σ are the mean and standard deviation, respectively, of attribute A.

3.5 Data Reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results [4]. Data cube aggregation, Attribute subset selection, Dimensionality reduction, numerosity reduction, discretization and concept hierarchy generation techniques are used for data reduction. Aggregation and hierarchy generation are discussed above. In this section Dimensionality reduction, where encoding mechanisms are used to reduce the data set size, will be detailed.

A method of eliminating redundant attributes was applied for the dimension reduction. To discover redundant attributes in dataset “*Pearson correlation coefficient*” measure was used. According to Table 3.5 includes the result of correlation analyze, 40 attributes with high correlations (80% and over) were eliminated. The initial number of attributes, 149, was reduced to 109 after analyzing.

Table 3.5 The Result of Correlation Analyze for Dimensionality Reduction

Eliminated Attribute	Correlation Value	Attribute (Correlated with)
Inpatient	0.88	Total Accrue
Bed	0.89	Total Accrue
Pediatric 2. Step ICU Bed	0.93	Pediatric 2. Step ICU Inpatient#
Pediatric 3. Step ICU Bed	0.93	Pediatric 3. Step ICU Inpatient#
Infant 3. Step ICU Bed	0.96	Infant 3. Step ICU Inpatient#
Adult 3. Step ICU Bed	0.84	Total Accrue
Cobalt Therapy	1.00	Total Accrue
Radiographic Imaging	0.90	Total Accrue
Nst/Cardio Tachograph	0.81	Total Birth(C/S Included)
Anesthesia Device	0.93	Total Accrue

Table 3.5 (continues) The Result of Correlation Analyze for Dimensionality Reduction

Eliminated Attribute	Correlation Value	Attribute (Correlated with)
Defibrillator	0.91	Total Accrue
6_Examination#	0.81	Total Accrue
7_Examination#	0.84	Total Accrue
10_Examination#	0.80	Total Accrue
20_Examination#	0.81	Emergency Examination
22_Examination#	0.82	Total Accrue
22_Inpatient#	0.82	Total Accrue
24_Examination#	0.82	Total Accrue
24_Inpatient#	0.83	Total Accrue
27_Inpatient#	0.93	Total Birth(C/S Included)
29_Examination#	0.84	Total Accrue
29_Inpatient#	0.83	Total Accrue
31_Examination#	0.83	Total Accrue
32_Examination#	0.86	Total Accrue
26__Inpatient#	0.84	26_Examination#
30_Examination#	0.89	12_Examination#
30_Inpatient#	0.80	12_Examination#
32_Inpatient#	0.83	32_Examination#
14_Inpatient#	0.82	4_Inpatient#
A1 Group Operation	0.89	Total Accrue
A2 Group Operation	0.89	Total Accrue
A3 Group Operation	0.89	Total Accrue
B Group Operation	0.92	Total Accrue
C Group Operation	0.89	Total Accrue
Expert Physician	0.94	Total Accrue
Health Service Class	0.92	Total Accrue
Technical Service Class	0.92	Total Accrue

Table 3.5 (continues) The Result of Correlation Analyze for Dimensionality Reduction

Eliminated Attribute	Correlation Value	Attribute (Correlated with)
General Administrative Service Class	0.92	Total Accrue
Auxiliary Service Class	0.92	Total Accrue

CHAPTER 4

IMPLEMENTATION

At the beginning of the creation of the dataset that constitutes the subject of this study, a current list of the hospitals which the report card will be calculated has been set. After the attributes of each record in the dataset have been determined, the establishment of the technical infrastructure for the collection of the data from sources has begun.

In the implementation part of this study, a data warehouse is designed and implemented for pulling data out of the source systems and transformed them to desired format automatically. Data warehouse was installed in the oracle database software and the PL/SQL language was used for programming.

After creating of database schema, database links were established to the source databases in schema before the receiving to the data warehouse. Creating a database link is entailed a user account in the source database. After the security procedures have been fulfilled, database users were created at the related database and granted select privilege to the database objects specified by analysis results. Cross tables were created for which did not refer to HCRS whereas dimensions tables copied from HCRS database on the data warehouse for applications refer to it. Data transformation and discretization operation codes were written in view scripts. During populating table, partitioning on the table which the NHS data were stored in data warehouse was done because of great volumes of data. In fact, NHS tables should and should be partitions in the OLTP database as well. But the partition logic on the source side was different from our aggregation logic. For this reason, it cannot be aggregated and generalized directly from the source database and table partition was reconstructed according to the intended purpose in the data warehouse. The fact table is designed to generate the dataset file to be loaded into the SPSS application for running clustering algorithm. K-Means clustering algorithm was run on SPSS. *The bisecting K-Means algorithm* [13] was used because of the limitations of this study on the number of clusters and cluster members. Implemented study flowchart is shown in Figure 4.1.

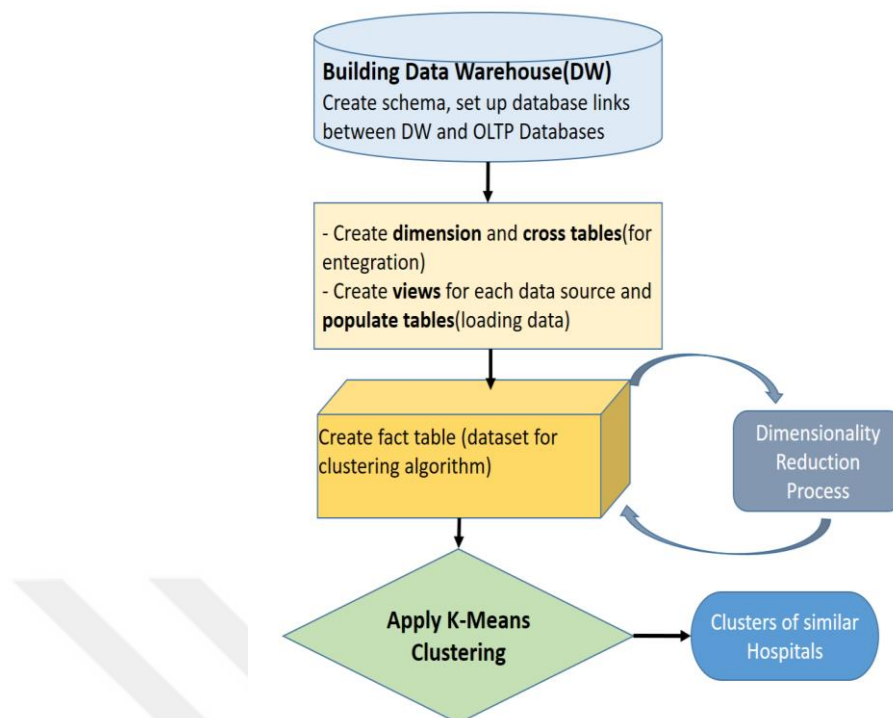


Figure 4.1 Flowchart of the implemented study

4.1 Source Databases

After the completion of the dataset formation task; the data was transferred from the operational databases to the created data warehouse. Data were obtained from four operational databases of applications which are running in the MoH; National Health System Database (NHSD), Main Resources Management Systems Database (MRMSD), Accounting Software System Database (ASSD) and Performance Payment Data System Database (PPDSD). Considering the security issues in databases; while some operational databases have been granted to select privilege to the schema level, partial tables in some databases and some databases have been opened for select privilege the desired database-specific views. Other technical information is given in the Table 4.1.

Table 4.1 Detailed technical information about database objects in the source databases.

Database	Accessed Database Object	Object Type	Partitioned	Size	Row count
NHSD	Patient Application Information	Table	Yes	250 GB	~ 1,5 billion
NHSD	EMPI Knowledge	Table	Yes	1,03 TB	~ 7 billion
NHSD	EMPI Score Knowledge	Table	Yes	292 GB	~ 4 billion
MRMSD	Device Knowledge	View	No	9 MB	~ 45 000
MRMSD	Personnel Knowledge	Materialized View	Yes	1,6 GB	~ 22 million
MRMSD	Services Knowledge	Table	Yes	1,3 GB	~ 120 million
ASSD	Accrue Knowledge	Materialized View	No	3 MB	~ 33 000
PPDSD	University Staff Knowledge	Table	No	392 MB	~ 2200

4.2 The Data Warehouse Structure

4.2.1 Dimension Tables

The most important dimension table in this study is the *hospitals dimension table* which contains the hospital codes corresponding to the records in the dataset used for clustering. The hospital dimensional table was formed in accordance with The Productivity Scorecard Application. For various reasons (economic, political, etc.), TPHI-affiliated hospitals may be reopened in a completely new status or connected to another hospital. Even though MRMS application includes association module which managed information of hospitals status, this dimension table could not be derived automatically from the application source tables, since the information chain for hospitals which are reopened at new status and linked to another could not be followed in the application. Due to this condition hospital dimension table records were manually filled by a team of specialists by examining the latest status of the hospitals.

The hospital dimension table was first constructed according to the current information in the application source. The "new hospital code" column was added to

the table to follow alteration information and the current hospital code field was copied to new column for all records. The "new hospital code" column was set for three different cases. Firstly, "new hospital code" column remained the same for hospitals which their status is not change during the year. Secondly, for hospitals to which the incorporate in another hospital, the "new hospital code" column was written as connected hospital code. In the last case, new recorded inserted to this dimension table for reopened hospitals in new status so that its code was changed. After completion of hospital dimension table, the source tables were joined with the "hospital code" column of the hospital dimension table during data transferring from source table to data ware house and data were grouped by the "new hospital code" column for aggregation functions.

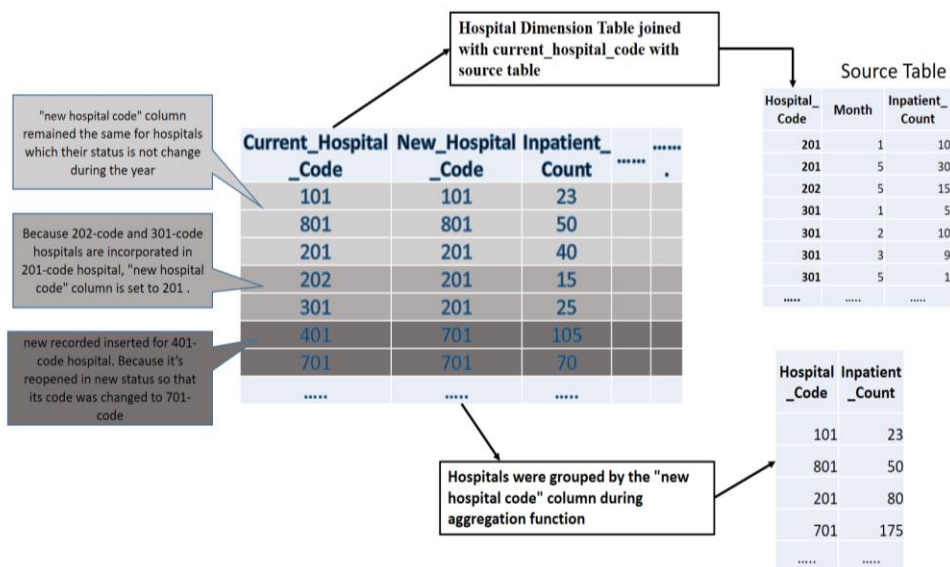


Figure 4.2 Hospital dimension table set up logic and implemented design

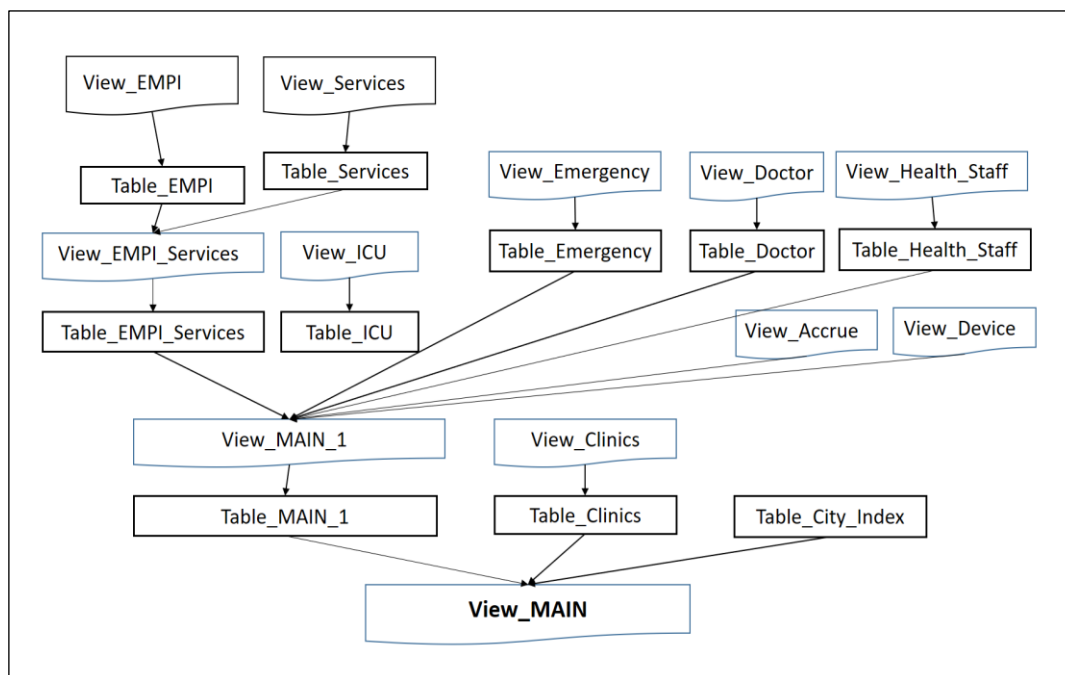
EMPI dimension table is copied from NKRS database which is the list of EMPI practiced in hospitals and other health institutions. There were only surgical groups in the source table, not other EMPI groups. EMPI group data are inserted according to their operation score (Table 4.2). Data discretization is implemented according to column which includes EMPI groups' information.

Table 4.2 EMPI groups according to their score

EMPI Group	EMPI Score
B Group	500-899
C Group	300-499
D Group	150-299
E Group	0-149

4.2.2 Views and Staged Tables

A database view is a logical object in a database. We can say “virtual tables” because it doesn't store data, instead describe information in the database. Since too many attributes are retrieved from different data sources, creating a view in a hierarchical structure has simplified coding. That's why the views were created for each data title. Before loading staging table data, creating a view makes it easy to format the data and keep data up to date. Most of the data pre-processing operations were done in view scripts (noisy data cleaning, integration, data discretization, etc.). After the views were created, data transfer is done to the stage tables. There is schematic representation of views and staged tables created in data warehouse in the Figure 4.3.

**Figure 4.3** Schematic representation of views and staged tables created in data warehouse

"Divide and Conquer" management has been implemented on the NHS database tables to simplify the management. Since the source table of the NHS is partitioned according to the "data sent time", V_CLSS_SURGEON_OPERATION view joined three tables for each partition, then inserted data into the partition stage table which partitioned by "EMPI Time". During transferring data, data was grouped by hospitals, EMPI code, month of EMPI time and physician and counted. After the data was transferred, aggregation was made on the EMPI column, firstly aggregated monthly and then yearly. After the data transfer, the partition information of the CLSS_SURGEON_OPERATION stage table is given in the followed Table 4.3.

Table 4.3 Information about CLSS_SURGEON_OPERATION table partitions

PARTITION_NAME	HIGH_VALUE	NUM_ROWS	BLOCKS
SYS_P176142	TO_DATE(' 2016-01-01 00:00:00')	2745333	6642
SYS_P176138	TO_DATE(' 2016-02-01 00:00:00')	2669512	6462
SYS_P176143	TO_DATE(' 2016-03-01 00:00:00')	2650426	6425
SYS_P176145	TO_DATE(' 2016-04-01 00:00:00')	2710421	6544
SYS_P176248	TO_DATE(' 2016-05-01 00:00:00')	2530867	6113
SYS_P176153	TO_DATE(' 2016-06-01 00:00:00')	2506696	6044
SYS_P176163	TO_DATE(' 2016-07-01 00:00:00')	2399820	5776
SYS_P176251	TO_DATE(' 2016-08-01 00:00:00')	2076554	5003
SYS_P176255	TO_DATE(' 2016-09-01 00:00:00')	2162102	5202
SYS_P176253	TO_DATE(' 2016-10-01 00:00:00')	1075577	2610
SYS_P176258	TO_DATE(' 2016-11-01 00:00:00')	1174286	2845
SYS_P176259	TO_DATE(' 2016-12-01 00:00:00')	1130025	2725
	* Block size: 4 MB		

4.3 K-Means Clustering

Agglomerative hierarchical and K-Means clustering algorithm is applied on the generated hospital dataset. Algorithm applying experiments in this thesis work are done in SPSS. The output will be clusters of hospitals to be averaged for the target value of the indicator cards. In this study, comparison with hospital roles (groups) determined by the MoH [30] and arguments of specialists are used to measure the performance of the clustering algorithm implemented. The algorithm is applied three

different ways discussed in the following subsections, it is examined according to this study-specific constraint.

4.3.1 Hierarchical Clustering Algorithm

While partitional clustering algorithms find all the clusters simultaneously as a partition of the data, hierarchical clustering algorithms recursively form clusters from the previous cluster. The most well-known hierarchical algorithms are single-link and complete-link. After implementation of complete-link algorithm with the hospital dataset, clusters of widely different sizes are produced, the distribution of the number of members in the formed clusters is unbalanced. For example, twelve clusters have 1 membership, six clusters have 2 memberships, one cluster has 3 memberships and one cluster has 570 memberships. Other problem is the membership count constraint. At the level of 15 clusters, only two clusters provisioned ‘at least five members in the cluster’ condition. This number is only one in the level of 20 clusters.

4.3.2 Standard K-Means Algorithm

Number of clusters k , cluster initialization and distance metric are user specified parameters of the K-Means algorithm. The most critical parameter is k . In this study, after running K-Means algorithm with k values between 10 and 20 (requested), it was observed that the distribution of number of members in the cluster did not satisfy the desired constraints. For each k values, minimum cluster member count was one which is lower than the constraint limit. For 10, 15 and 20 k values, number of cases in each cluster is shown in Table 4.4.

Table 4.4 For 10,15 and 20 k values, number of cases in each cluster

K value = 10		K value = 15		K Value = 20	
Cluster	Number of Case	Cluster	Number of Case	Cluster	Number of Case
1	2	1	1	1	3
2	1	2	14	2	1
3	11	3	1	3	421
4	133	4	129	4	23
5	2	5	1	5	7
6	1	6	1	6	1

Table 4.4 (continues) For 10,15 and 20 k values, number of cases in each cluster

K value = 10		K value = 15		K Value = 20	
7	1	7	373	7	1
8	32	8	2	8	1
9	413	9	53	9	1
10	1	10	5	10	2
		11	2	11	2
		12	7	12	3
		13	1	13	3
		14	1	14	3
		15	6	15	1
		16		16	1
		17		17	2
		18		18	118
		19		19	1
		20		20	2

4.3.3 Bisecting K-Means (a variant of K-Means)

To achieve the desired number of cluster member in a controlled way, it's chosen a very simple and efficient implementation of the K-Means algorithm, bisecting K-Means [13]. In this method, the number of clusters increased progressively, in conjunction with aimed number of cluster member. Bisecting K-Means achieves this by first putting all the data into a single cluster, and then recursively splitting the least compact cluster into two using 2-means.

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic K-means algorithm. (Bisecting step)
3. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

Figure 4.4 Basic Bisecting K-means Algorithm for finding K clusters.

To choose which cluster is split is another issue to handle. The criteria are the number of cluster and cluster member for this study. The clusters are chosen which has 10 or more than 10 members. If split cluster has less than 5 members, the process is undone and this cluster is set as result cluster. If the result cluster number at the end of the algorithm is less than the desired number, then the stopping criterion based on the number of cluster member is moved to the next level and the members of cluster which is not met criterion of member count in the previous level are accepted outliers. Since every hospital has to belong to a cluster, outliers are assigned to the appropriate cluster by specialists. Two hospitals were outlier according to algorithm result. The schematic representation of the way the algorithm works is shown in Figure 4.5. Result clusters and their member count are listed at Table 4.5.

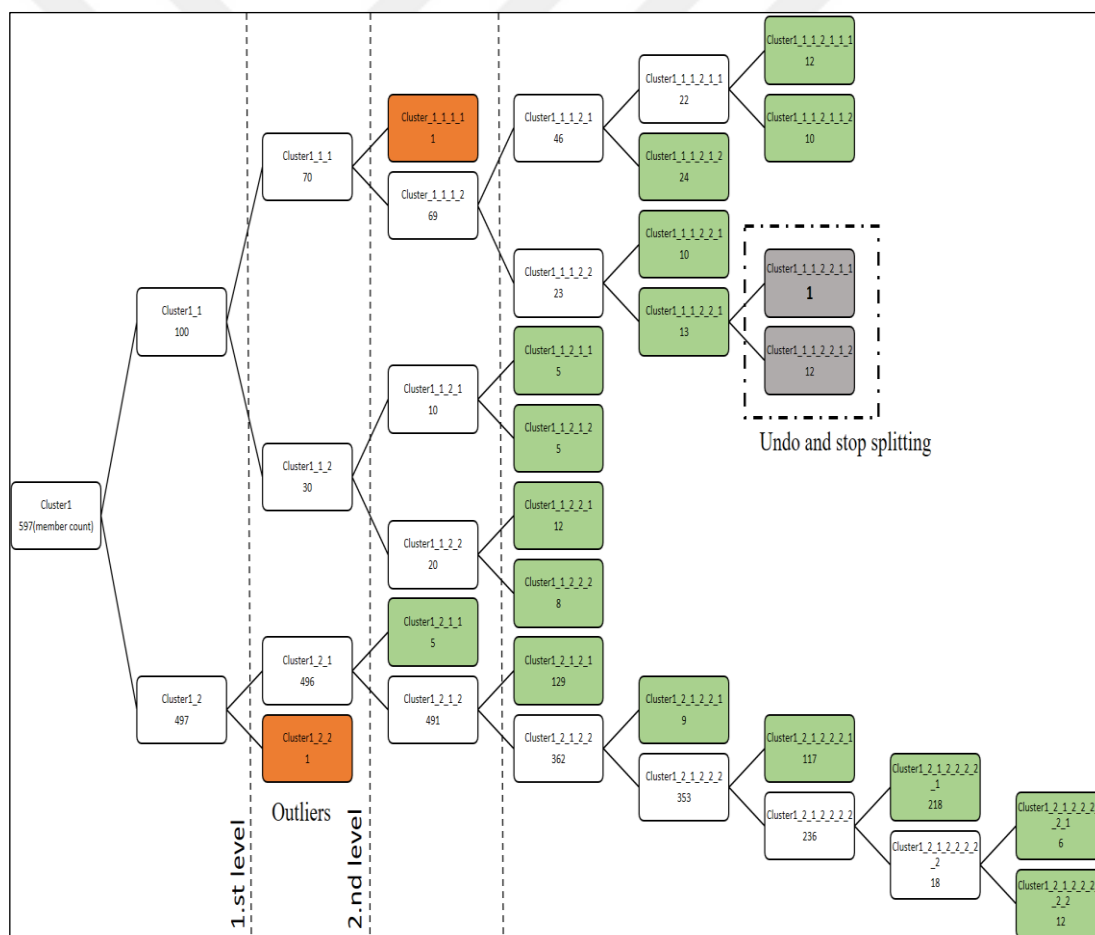


Figure 4.5 The formation steps of clusters with the bisecting k-means algorithm.

Table 4.5 Member count of final clusters

	Clusters															
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
Member counts	12	10	24	10	13	5	5	12	8	5	129	9	117	218	6	12

CHAPTER 5

RESULTS

As the major objective of this thesis study is clustering hospitals into groups where within-group hospitals are similar with respect to resource and output. K-Means clustering algorithm was applied for clustering relied largely on input characteristics. Measuring the quality of a clustering methods can be categorized into two groups *extrinsic methods*, which compare the clustering against the ideal clustering that is often built using human experts, and *intrinsic methods*, which evaluate the accuracy by considering separating how well the clusters.

The extrinsic method is used to measure clustering success which calculates the hospital role (category) determined by the MoH [30] distribution ratios in a cluster. General training hospitals studied on are classified by MoH, six different classes. The distribution ratio shows a homogeneity of the clusters, according to hospital roles.

The chart in the Figure 5.1 shows the highest percentage of roles in the cluster, while Table 5.1 lists the distribution of all the roles in the table. According to the chart, the number of members with the same role in just one cluster remained below 50%. Three clusters consisted of hospitals with the same role.

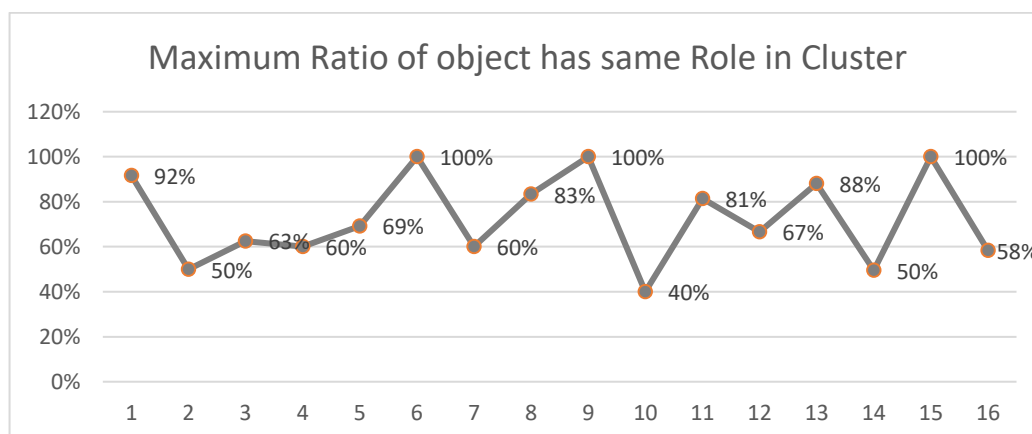


Figure 5.1 Maximum Ratio of hospitals has same Role in Cluster

Table 5.1 Distribution ratio of all roles in clusters

Distribution Ratio of Role Categories in Clusters						
Clusters	AI	AII	B	C	D	EI
1	8.33%	91.67%				
2	50.00%	50.00%				
3	29.17%	62.50%	8.33%			
4	60.00%	40.00%				
5	30.77%	69.23%				
6	100.00%					
7	60.00%	40.00%				
8	83.33%	16.67%				
9	100.00%					
10	20.00%		40.00%	20.00%	20.00%	
11		10.85%	81.40%	7.75%		
12			33.33%	66.67%		
13	0.85%		10.26%	88.03%	0.85%	
14				16.06%	49.54%	34.40%
15					100.00%	
16					41.67%	58.33%

The reasons about incompatibility between clustering method and role identification method are stated below;

- While the hospital roles were determined, the demographic information of the region where the hospital was located criteria has a significant importance, but any attribute concerning demographic information has not been added to the hospital dataset which is the clustering input.
- Although there is a change in the status of the hospitals (unification of hospitals, connect to another hospital financially, etc.) there is no change in the role. The reason for this may sometimes be procedural delays, but sometimes it may be due to political reasons. This situation cause inconsistency between

the clusters which are the result of the clustering done according to the current state of the hospitals and the hospital roles.

- The fact that the roles of some hospitals are not based on the current status but on the plans of the future period and some inconsistent political reasons can be considered as another reasons for inconsistencies between clusters and roles of hospitals.

The results can also be evaluated on *outlier objects*. The success of work can also be seen as the result of analysis of objects that cannot be merged with other objects. When the cases of two hospitals which are outliers according to clustering result were examined by specialists, it was concluded that these hospitals should be evaluated in completely different cases and should not be included in this grouping. The right *outlier detection* according to the experts can be regarded as another indicator of the success of this study.

CHAPTER 6

DISCUSSION AND CONCLUSION

In this study, clustering of hospitals affiliated to the TPHI was done in order to figure out target value for the indicator cards in the productivity scorecard application, which is the acceptable value is the service class average. It is planned to be the basis for the calculation of the 2017 the productivity scorecards according to the hospital clusters formed in this study. In addition, the study also supports the identification of hospital roles, one of the duties of the MoH. According to the hospital clusters formed by this study, the target values of the indicator cards in the productivity scorecards application came to be calculated.

As a result, the results of the applied study were evaluated as successful by the experts substantially. According to the results of this study, the automatic extraction of very large and scattered hospital data from the application databases and the data preprocessing according to the data mining techniques can reduce the error rate and speed up the process.

There were some difficulties in creating the dataset. Some information about hospital status changes could not be tracked. It was not possible to automatically retrieve all the qualifications that define hospitals automatically from data sources, such as demographic information. Therefore, the difficulties encountered in the development of an automated hospital grouping system in the institution suggest that some application needs to be developed and that some application needs to be regulated.

The structure of the hospitals varies from a wide range, starting from architectural, laboratory services, operating room services, as well as hotel services, food and beverage and cleaning services, financial services and human resources management services. No grouping study was found for the general similarity of the hospital according to all aspects in the literature review. This study may provide a more detailed review of the studies clustered hospitals according to a specific point of view in terms

of hospitals in different clusters according to the clusters formed in accordance with all aspects of hospitals.

NHDD is developed to enable the application to share the same meaning of data, and use them for the same purpose. The data whose definition and format determined within the NHDD establishes a reference for the information systems incorporated in MoH and used at health institutions. The challenges faced during data integration study in this work show the content interoperability among different applications importance. It is imperative that the applications in MoH to be developed be in rapport with to HCRS.

Future research and ideas to improve the performance of this work include:

- The data set used in this study is being developed for future studies. It will continue to be improved by adding not only numerical data but categorical data. In the future study, this study can be developed with more attributes of hospitals and different clustering techniques.
- Data set development for future studies can also be considered in terms of detailing of attributes. If some of the attribute data of hospital data set are decomposed, it will contribute to the success of the clustering. Disaggregation of physicians according to branches and the attributes related to emergency services according to the emergency levels is planned attribute detailing study in the future.
- Clustering is an unsupervised method for data analysis. However, there are often information about the problem domain as well as data instances. In the methodology of hospital role determination, it is often used that the experimenter possesses some background knowledge (about the domain or the data set). Domain knowledge is not used in the clustering algorithm implemented in this study. K-Means clustering algorithm can be improved by modifying it to use this information [31].

REFERENCES

- [1] Kaplan, R.S. and Norton, D., "The Balanced Scorecard - Measures That Drive Performance", Harward Business School, 1991.
- [2] TKHK, Verimlilik ve Kalite Yönetimi Daire Başkanlığı, Retrieved June 08, 2017, from <https://www.tkhk.gov.tr/DB/19/>, 2014
- [3] Carugo, O. and Eisenhaber, F., Data Mining Techniques for the Life Sciences (1st Ed.), Humana Press (Springer), 2010.
- [4] Han, J. and Kamber, M., Data Mining: Concepts and Techniques (2nd Ed.), Morgan Kaufmann Publishers, 2006.
- [5] Frank, E. and Witten, I.H., Data Mining, Morgan Kaufmann(3th Ed.), 2005.
- [6] Avcı, K., Sağlık Bakanlığı Hastaneleri'nin Örgütsel Ve Finansal Performans Göstergeleri Bakımından Kümelenmesi, Hacettepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 2015.
- [7] Li, L. and Xia, L.N., "Evaluation of Hospital Competitiveness in Jiangxi Province Based on the Cluster Analysis", Springer, 2013.
- [8] Berta, P., "Multilevel cluster-weighted models for the evaluation of hospitals", Metron-International Journal of Statistics, 2016.
- [9] Morton, D.J., "Statistical Methods for Classifying Hospital Quality", ProQuest LLC, 2015.
- [10] Iversen, H.H., Bjertnaes, O.A. and Skudal, K.E., "Patient evaluation of hospital outcomes: an analysis of open-ended comments from extreme clusters in a national survey", BMJ Publishing Group, 2014.
- [11] Şahin, İ., Özcan, Y.A. and Özgen, H., "Assessment of hospital efficiency under health transformation program in Turkey", Springer-Verlag, 2009.
- [12] Narcı, H.Ö., "An examination of competition and efficiency for hospital industry

in Turkey", Health Care Management Science, 2015.

- [13] Steinbach, M., Karypis, G., and Kumar, V., "A Comparison of Document Clustering Techniques", Department of Computer Science / Army HPC Research Center, University of Minnesota, 2000.
- [14] Kashef, R. and Kamel, M.S., "Enhanced bisecting k-means clustering using intermediate cooperation", Pattern Recognition, 2009.
- [15] Zhuang, Y., Yu, M. and Xin, C., "A Limited-Iteration Bisecting K-means for Fast Clustering Large Datasets", IEEE Trustcom BigDataSE ISPA, 2016.
- [16] Elmasri, R. and Navathe, S.B., Fundamentals of Database Systems(7th Ed.), Pearson Education, 2003.
- [17] Wikipedia, Data Warehouse, Retrieved July 01,2017 https://en.wikipedia.org/wiki/Data_warehouse, 2017.
- [18] Oracle Database Online Documentation-Data Warehousing and Business Intelligence, Retrieved June 01, 2017, from <https://docs.oracle.com/database/121/DWHSG/concept.htm>, 2017
- [19] datawarehouse4u.info, Data Warehouse, Retrieved June 01,2017, from <http://datawarehouse4u.info/>, 2017.
- [20] Akpınar, H., Data(1st Ed.), Papatya Yayıncılık, 2014.
- [21] Zhao, W., Ma, H. and He, Q., "Parallel K-Means Clustering Based on MapReduce", Springer-Verlag Berlin, 2009.
- [22] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou Z., Steinbach, M., Hand D.J., Steinberg D., "Top 10 algorithms in data mining". Springer, 2007.
- [23] Lloyd, SP., Least squares quantization in PCM, Unpublished Bell Lab. Tech. Note, 1957.

- [24] Jain, A.K. and Dubes, R.C., Algorithms for Clustering Data, Prentice Hall, 1988.
- [25] Jain, A.K., "Data clustering: 50 years beyond K-means", Elsevier Science, 2010.
- [26] Huang, Z., "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery , 1998.
- [27] WHO, International Classification of Diseases (ICD 10), Retrieved June 30, 2017 from <http://www.who.int/classifications/icd/en/>, 2017
- [28] SRDC, Turkey's National Health Information. Retrieved June 01,2017, from <http://www.srdc.com.tr/share/publications/2008/9.pdf>, 2005.
- [29] TKHK Ek Ödeme Daire Başkanlığı, Tıbbi İşlemler Yönergesi, Girişimsel İşlemler listesi. Retrieved July 01, 2017, from <https://www.tkhk.gov.tr/DB/20>, 2017.
- [30] Şencan, İ., Koç, O. and Bayraktar, O. G., Hastane rolleri, MoH, Tedavi Hizmetleri Genel Müdürlüğü, 2009.
- [31] Wagsta, K., "Constrained K-means Clustering with Background Knowledge", Proceedings of the Eighteenth International Conference on Machine Learning, 2001.

APPENDICES

Appendix A: The List of Attributes in the Hospital Dataset

Appendix B: An Indicator Card Example in the Productivity Scorecard Application



Appendix A - List of Attributes in the Hospital Dataset

Table A.1 The list of attributes of the hospital dataset

Group 1	Group 2	Attribute Name
Financial	Financial	Total Accrue
Staff	Physician	Trainer Physician
		Chief Assistant
		Specialist
		Physician
		Dentist
		Specialist Dentist
	Physician External Staff	Health Service Staff
		Technical Service Staff
		Managing Staff
		Support Service Staff
Services	Emergency Services	Number of Patient Admitted From Emergency Department
		Number of Emergency Examination
		Number of Total Examination
		Number of Yellow Area Examination
		Number of Red Area Examination
	Intensive Care Services	1. Step Adult ICU Inpatient#
		2. Step Adult ICU Inpatient#
		3. Step Adult ICU Inpatient#
		2. Step Pediatric ICU Inpatient#
		3. Step Pediatric ICU Inpatient#
		1. Step Neonatal ICU Inpatient#
		2. Step Neonatal ICU Inpatient#
		3. Step Neonatal ICU Inpatient#
		1. Step Adult ICU Bed#
	2. Step Adult ICU Bed#	

Table A.1 (continues) The list of attributes of the hospital dataset

Group 1	Group 2	Attribute Name
Services	Intensive Care Services	3. Step Adult ICU Bed#
		2. Step Pediatric ICU Bed#
		3. Step Pediatric ICU Bed#
		1. Step Infant ICU Bed#
		2. Step Neonatal ICU Bed#
		3. Step Infant ICU Bed#
	Other Services <i>(*Each of the attributes starting with the number corresponds to a clinic.)</i>	Number of A1 Group Operation
		Number of A2 Group Operation
		Number of A3 Group Operation
		Number of B Group Operation
		Number of C Group Operation
		Number of D Group Operation
		Number of E Group Operation
		Number of B Group Intervention
		Number of C Group Intervention
		Number of D Group Intervention
		Number of E Group Intervention
		Number of Inpatient
		Total Birth(C/S Included)
		Number of Examination
		Number of Bed
		2_Examination#
		3_Examination#
		3_Inpatient#
		4_Examination#
	4_Inpatient#	
5_Examination#		

Table A.1 (continues) The list of attributes of the hospital dataset

Group 1	Group 2	Attribute Name
Services	Other Services (*Each of the attributes starting with the number corresponds to a clinic.)	5_Inpatient#
		6_Examination#
		6_Inpatient#
		7_Examination#
		7_Inpatient#
		8_Examination#
		8_Inpatient#
		9_Examination#
		9_Inpatient#
		10_Examination#
		10_Inpatient#
		11_Examination#
		11_Inpatient#
		12_Examination#
		12_Inpatient#
		13_Examination#
		14_Examination#
		14_Inpatient#
		15_Examination#
		17_Examination#
		20_Examination#
		21_Examination#
		22_Examination#
		22_Inpatient#
		23_Examination#
23_Inpatient#		
24_Examination#		
24_Inpatient#		
25_Examination#		

Table A.1 (continues) The list of attributes of the hospital dataset

Group 1	Group 2	Attribute Name
Services	Other Services (*Each of the attributes starting with the number corresponds to a clinic.)	25_Inpatient#
		26_Examination#
		26_Inpatient#
		27_Examination#
		27_Inpatient#
		28_Examination#
		28_Inpatient#
		29_Examination#
		29_Inpatient#
		30_Examination#
		30_Inpatient#
		31_Examination#
		31_Inpatient#
		32_Examination#
		32_Inpatient#
		36_Examination#
		43_Examination#
		43_Inpatient#
45_Examination#		
46_Examination#		
46_Inpatient#		
Technology(Device)	Technology	Operation Table
		Anesthesia Device
		Angiography Device
		Arthroscopy
		(CT-Sim) Device
		Brachytherapy Device

Table A.1 (continues) The list of attributes of the hospital dataset

Group 1	Group 2	Attribute Name
Technology(Device)	Technology	Bronchoscopy Device
		Cryotherapy Device
		Cyber Knife
		Densitometer
		Defibrillator
		Dialysis / Renal Replacement
		Linear Accelerator
		Duodenoscopy
		EEG (Electro Encephalography)
		Effort
		ECT (Electro Convulsive Therapy)
		Emg/ Eng/ Enmg
		Endoscopic Robotic Surgery
		Eswt / Rswt / Shock Wave
		Phacoemulsification - Vitrectomy
		Phototherapy
		Fundus Camera
		Gamma Camera
		Gastroscopy
		Holter
		Iort
		Heart-Lung Pump
		Chemotherapy Preparation
		Cobalt Therapy
		Colonoscopy-Sigmoidoscopy
		Colposcopy
		Laparoscopy
		Lithotripter/ Stone Breaker
Mammography		

Table A.1 (continues) The list of attributes of the hospital dataset

GROUP 1	Group 2	Attribute Name
Technology(Device)	Technology	Magnetic Resonance
		Nst/Cardio Tachograph
		Audiometry - Tympanometry
		(Obt/Oct) System /Eye S
		Pet-Ct System
		Puva
		Radiographic Imaging
		Cystoscopy-Ureteroscopy
		Spect System
		Medical Laser
		Thyroid Uptake
		Tomography

Appendix B - Referral Rate Taken by Emergency Department Indicator Card

Figure A.1 “Referral rate taken by emergency department” indicator card

<i>SHY-ASH-06 Acil Servisin Almış Olduğu Sevk Oranı</i>											
Gösterge Kodu	SHY-ASH-06										
Gösterge Adı	Acil Servisin Almış Olduğu Sevk Oranı										
Boyut	Sağlık Hizmetleri Yönetimi										
Bölüm	Acil Servis Hizmetleri										
Amaç	112 ile diğer sağlık tesislerinden sevkli gelen hastaların, tedavi ve bakım hizmet süresinin yeterliliğini değerlendirilmek										
Hesaplama İçin Gerekli Veriler	Her bir hasta için; 112 ile Acil servise diğer sağlık tesislerinden sevk ile gelen hastanın taburcu saati (A) 112 ile Acil servise diğer sağlık tesislerinden sevk ile gelen hastanın kabul saati (B) 112 ile Acil servise diğer sağlık tesislerinden sevk ile gelen ilgili dönemdeki toplam hasta sayısı (C) Bir önceki dönem KED										
Sağlık Tesisi Değeri (STD)	$STD_1 = C$ $STD_2(A > B \text{ ise}) = (A - B)$ (Her hasta için ayrı ayrı hesaplanmaktadır.) $STD_3(B > A \text{ ise}) = ((24-B)+A)$ (Her hasta için ayrı ayrı hesaplanmaktadır.) $STD_3 = STD_2$ 'lerin aritmetik ortalaması										
Kabul Edilebilir Değer (KED)	KED = Sağlık tesisinin bulunduğu hizmet sınıfının 112 ile acil servise sevk ile gelen hasta sayılarının aritmetik ortalaması										
Gösterge Puanı (GP)	60										
Puan Katsayısı (k)	STD_1 / KED										
Puan Hesaplama	<table border="1"> <thead> <tr> <th>Puan Katsayısı (k)</th> <th>Sağlık Tesisi Puanı</th> </tr> </thead> <tbody> <tr> <td>$k \geq 1,50$</td> <td>GP</td> </tr> <tr> <td>$0,75 \leq k < 1,50$</td> <td>$GP - [GP \times (1,50 - k)^2]$</td> </tr> <tr> <td>$k < 0,75$</td> <td>$GP \times k^4$</td> </tr> <tr> <td>$STD_3 < 6 \text{ saat}$</td> <td>0</td> </tr> </tbody> </table> <p>Yukarıdaki tabloya uygun olarak dönem verileri ile puanın %50'si hesaplanır. Puanın diğer %50'si ise ;bu dönemdeki STD_1 ile bir önceki karne dönemindeki KED' in oranlanması sonucunda oluşan puan katsayısı (k) baz alınarak hesaplanır. (Bu hesaplamada STD_3 dikkate alınmaz.)</p>	Puan Katsayısı (k)	Sağlık Tesisi Puanı	$k \geq 1,50$	GP	$0,75 \leq k < 1,50$	$GP - [GP \times (1,50 - k)^2]$	$k < 0,75$	$GP \times k^4$	$STD_3 < 6 \text{ saat}$	0
Puan Katsayısı (k)	Sağlık Tesisi Puanı										
$k \geq 1,50$	GP										
$0,75 \leq k < 1,50$	$GP - [GP \times (1,50 - k)^2]$										
$k < 0,75$	$GP \times k^4$										
$STD_3 < 6 \text{ saat}$	0										
Açıklama	<ul style="list-style-type: none"> EK PUAN: Bu kriter 60 puan üzerinden değerlendirilmekle beraber acil servis hizmetlerinin bütününe doğrudan etki etmeden ek puan olarak kullanılacaktır. Yukarıdaki tabloda hesaplanan puan sağlık tesisinin acil servis hizmetlerinden aldığı toplam puana (tavan puanı aşmamak kaydı ile) eklenir. E1 rolündeki hastaneler, Meslek Hastalıkları Hastaneleri, Fizik Tedavi ve Rehabilitasyon Hastaneleri, Lepra, Deri ve Zührevi Hastalıklar Hastaneleri, Göz Hastaneleri göstergeden muafır. 										
Veri Kaynağı	Sağlık Net- Online / ASKOM										
Verinin Ait Olduğu Dönem	Altı aylık dönemlerde izlenir.										

CURRICULUM VITAE

PERSONAL INFORMATION

Name Surname: Ayşe KELEŞ

Date of Birth: 1976

E-mail: ayseinan@gmail.com



EDUCATION

Master Degree: 2017, Department of Computer Engineering, Yildirim Beyazıt University, Ankara, Turkey
GPA:3.57/4.00

Bachelor: 2005, Department of Computer Engineering, Baskent University, Ankara, Turkey
GPA:3.5/4.00

High School: 1995, Kahramanmaraş Sağlık Meslek Lisesi

WORK EXPERIENCE

Computer Engineer: Turkey Ministry of Health

TOPICS OF INTEREST

- Data Mining
- Database and Data warehouse
- Decision Support Systems