Application of Clustering as a Data Mining Tool in Bp systolic diastolic

Assist. Proffer Dr. Zeki S. Tywofik
Department of Computer, Dijlah University College (DUC),Baghdad, Iraq.
Assist. Lecture. Ali T. YASEEN.
Department of Computer, Dijlah University College (DUC),Baghdad, Iraq.

Abstract:

This work demonstrates the application of clustering , a data mining tool, in the health care system. Hemoglobin A1c is the most parameters for the monitoring of metabolic control of patients with diabetes mellitus[1]. The aim of this study is to determine the reference rang of glycosylated hemoglobin (Hb A1c%) in an Iraqi population (males and females) effect and predict Bp systolic diastolic(Blood pressure systolic diastolic) by using demonstrates the application of clustering, as data mining tool, in the health care system. Data mining has the capability for clustering, prediction, estimation, and pattern recognition by using health databases.

Blood samples were collected from 100 healthy subjects (50 females and 50 males) are ranged between (20-75) years old as dataset. The reference value of HbA1c% was (5.34 + 0.67)% in female and (5.67 + 0.73)% in males. The present clustering and found a strong relation between HbA1c% and systolic diastolic blood pressure in males whereas the relation in females no significant.

Keywords: Data mining, Cluster Analysis, K-Means Algorithm.

1. INTRODUCTION

Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns.

Data mining is not new has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling[2].

In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Several factors have motivated the use of data mining applications in healthcare. The existence of medical insurance fraud and abuse, for example, has led many healthcare insurers attempt to reduce their losses by using data mining tools to help them find and track offenders. Fraud detection using data mining applications is prevalent in the commercial world, for example, in the detection of fraudulent credit card transactions. Recently, there have been reports of successful data mining applications in healthcare fraud and abuse detection. As a possible solution to this problem, the concept of *Knowledge Discovery in Databases – KDD* has emerged [3]. The process of the formation of significant models and assessment within KDD is referred to as data mining [2,4]. Data mining is used to uncover hidden or unknown information that is not apparent, but potentially useful [3].

Cluster analysis is a technique used in data mining. Cluster analysis involves the process of grouping objects with similar characteristics [6], and each group is referred to as a cluster. Cluster analysis is used in various fields, such as marketing, image processing, geographical information systems, biology, and Genetic[4].

In this paper Blood samples collect from healthy subjects of 100 patients to study the relation between Bp systolic diastolic and HbA1c, patients Age, Weight by forming clusters also found reference rang effect between female and male. The clustering process was carried out using a Kmeans algorithm.

2. Cluster Analysis

Cluster analysis is a multivariate analysis technique where individuals with similar characteristics are determined and classified (grouped) accordingly [2,7]. Through cluster analysis, dense and sparse region can be determined in the distribution, and different distribution patterns may be achieved. The concepts of similarities and differences are used in cluster analysis. Different measures may be used in determining similarities and differences. This study utilizes the Euclidian distance measure[5].

2.1. Euclidian Distance Measure

The Euclidian distance measure is frequently used as a distance measure, and is easy to use in two dimensional planes. As the number of dimensions increases, the calcite salability time also increases [2].

$$d(i, j) = \sqrt{(\left| X_{11} - X_{11} \right|^2 + \left| X_{i2} - X_{j2} \right|^2 + \ldots + \left| X_{ip} - X_{jp} \right|^2)} \qquad (2\text{-}1)$$

The formula defines data objects i and j with a number of dimension equal to p. The distance between the two data objects d(i,j) is expressed as given in formula (2.1.). X_{ip}: is the measurement of object i in dimension[6].

3. The Basic algorithm of K-mean

The K-means algorithm is a cluster analysis algorithm used as a partitioning method, and was developed by MacQueen in 1967 [8]. K-means is the most widely used used and studied clustering algorithm. Given a set of n data points in real d-dimensional space, Rd, and an integer k, the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center[7].

The K-means algorithm defines a random cluster centroid according to the initial parameters [8]. Each consecutive case is added to the cluster according to the proximity between the mean value of the case and the cluster centroid. The clusters are then re-analysed to determine the new centroid point. This procedure is repeated for each data object.

The algorithm is composed of the following steps:[8,9]

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.
5. Until the centroids dos not change.

4. APPLICATION

The data gathered from were collected from 100 healthy subjects (50 females and 50 females) are ranged between (20-75) years old as dataset.

4.1. Database

The database management system used in the study was the Microsoft SQL Server 2000. This system was used for two reasons; the software used in analysis

was compatible and efficient to use with the database management system, and the data to be analyzed was maintained in the database prior to the study.

Data Set the gathered from were collected from 100 healthy subjects 2010, and included records of 100 patients.

4.2. Application Software

The programming environment for the application was Matlab. The Matlab software application was suitable for the development of the application, and compatible with the SQL Server 2000 in which the data was maintained. The K-means algorithm used in the application was defined in the Matlab software as a function.

4.3. The Data Process

The data exploration and presentation process consisted of various steps. These steps were data preparation, data selection and transformation, data mining and presentation.

4.3.1. Data Preparation and Discuses table1.

In these steps, the data that was maintained in different tables was joined in a single table. The database of female(F) and male(M) are collected from peoples and select the data that more effect Bp systolic diastolic such as Age Year are from 21-66 for females and 25-57 for male, weight in different rang kilograms, Height in centimeters relation between them and Hb A1C put it in range after taken samples from female and male. The other attributes are less effect. The process errors in the data were corrected because some data in collection out of rang in some attribute the data arranged and put in table 1 blow.

Table 1 Dataset training

Sex	Age Yrs	WE. Kgm	Higfht Cm.	Hb A1c	Bp systolic diastolic (class)
F	21-32	55-66	170-162	4.0-4.8%	110/70
F	32-39	50-64	170-162	4.8-5.1%	120/70
F	34-40	64-74	162-175	5.1-5.5%	120/80
F	47-64	60-85	153-165	5.6-6.2 %	130/80
F	36-67	70-85	152-171	5.8-6.6%	130/90
F	66	79	165	63%	150/90
M	25-38	64-75	183-176	4.0-5.3%	120/80
M	**34-35**	**70-74**	**170-173**	**4.3-5.4%**	**130/80**
M	**25-32**	**109-129**	**174-180**	**5.6-6.1%**	**120/90**
M	**36-37**	**30-37**	**167-181**	**5.0-6.9%**	**130/90**
M	**35-57**	**84-95**	**174-174**	**5.6-5.9%**	**150/90**

4.3.2. Data Selection and Transformation

In Data mining process of extracting interesting pattern from database. After the database preparation, transformation to process was performed. In this step the database used in the study were determined and transformed if necessary in range then database are regularly collected and analyzed by applications of program using data mining.

4.4.3. Data Mining

he prepared data was then put through the data mining process. The K-means algorithm was used in this step. The number of clusters was determined as an external parameter. Different cluster numbers were tried, and a successful partitioning was achieved with 6 clusters. The cluster centurions are given in table 2.

Table 2.

Cluster	sex	Bp systolic diastolic range	Percentage %
1	**Female**	**110/70 – 120/80**	**80%**
2	**Female**	**130/80 – 130/90**	**16%**
3	**Female**	**140/90**	**4%**
4	**Male**	**120/80 – 120/90**	**46%**
5	**Male**	**130/80 – 130/90**	**42%**
6	**Male**	**150/90**	**12%**

4.4.4. Discussion

The results of the data mining step are presented in this step.

1. The effect of Hb A1c in Bp systolic diastolic for female less from table
2. The effect of Hb A1c in Bp systolic diastolic for female large the female from table 2.
3. There strong relation between Age Yrs, WE. Kgm, Higfht Cm. and Hb A1c from 1 and table 2.
4. The result of clustering as shown in table 2.
5. The resulting clusters are shown in figure 2.

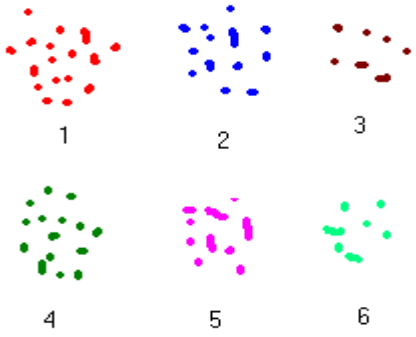

Figure 2. Each number represents different cluster

5. CONCLUSION

This study utilizes data mining in the field of clustering the data. Cluster analysis and K-means analysis were used as data mining techniques. The steps of the data mining process were carried out and explained in detail. The area of application was predict effect of Age Year, weight in kilograms, Height in centimeters witch relation between them and Hb A1C on Bp systolic diastolic for patients. The use of the data mining technique in clustering may provide us with more varied and significant findings.

REFERENCES

[1]. Salwa H.N., Emad Abdul-Rehman (2010) "Determining the reference Rang Values of Glycosylated Hemoglobin (HbA1c) by Immunoiturbid Assayin Iraqi Population" *Journal of* College of science.

[2] Erdoğan, Ş. Z., "Veri Madenciliği ve Veri Madenciliğinde Kullanılan K-Means Algoritmasının Öğrenci Veri Tabanında Uygulanması", Yüksek Lisans Tezi, İstanbul Üniversitesi, 2004.

[3]. Milley, A. (2000). Healthcare and data mining. *Health Management Technology*, 21(8), 44-47.

[4] Yuqing, P., Xiangdan, H., Shang, L., "The KMeans Clustering Algorithm Based On Density and Ant Colony", IEEE Int. Conf. Neural Networks & Signal Processing Nanjing, China, 457-460, December 14-17, 2003.

[5] Thearling, K., "An Introduction to Data Mining",http://thearling.com/text/dmwhite/dm white.htm, 01 December 2003.

[6] Han, J., Kamber, W., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA, 5-10, 2001.

[7] Fayyad, U.M., Piatesky-Shapiro, G., Smyth, P., Uthurusamy, R., "Advances in data mining and knowledge discovery", MIT Pres, USA, 1994.

[8] Kanungo, T., Mount, D., S. Netanyahu, N., Piatko D. C., Silverman, R., Wu, A.Y., "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002.

[9] Luke, B. T., "K-Means Clustering", http://fconyx.ncifcrf.gov/~lukeb/kmeans.html, 20 October 2004.