ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Forecasting Diabetes Mellitus Using Different Clustering Algorithms in Data Mining

Pathan Jawwad Khan
M.Tech Student,
Computer Science Department,
Everest Educational Society's Group of Institutions
College of Engineering & Technology, Aurangabad.

Auti R A, M.Tech
Assistant Professor,
Computer Science Department,
Everest Educational Society's Group of Institutions
College of Engineering & Technology, Aurangabad.

Abstract:

Data Mining is the process of extracting information from large data sets through algorithms and techniques drawn from the field of Statistics, machine Learning and Data Base Management Systems. There are different kinds of data mining techniques are available. Clustering, Classification, Association Rule and Neural Network are some of the most major techniques in data mining.

In Health care trades, Data mining plays an important role. Most frequently the data mining is used in health care industries for the process of forecasting diseases. Diabetes is a lingering condition.

This means that is lasts for a long time, often for someone's whole life [1]. This research paper explains the use of data mining and the comparison of diabetes to predict the future decisions using clustering techniques.

Here we are using three different kinds of clustering techniques named as Hierarchical clustering; Density based clustering, and Simple K- Means clustering. Weka (Waikato Environment for Knowledge Analysis) is used as data mining tool.

Keywords:

Data mining, Diabetes, Forecast, Clustering, K-means, Hierarchical clustering, Density based clustering, Weka.

I.INTRODUCTION :

Diabetes, often referred to by doctors as diabetes mellitus, describes a group of metabolic diseases in which the person has high blood glucose (blood sugar), either because insulin production is inadequate, or because the body's cells do not respond properly to insulin, or both. Patients with high blood sugar will typically experience polyuria (frequent urination), they will become increasingly thirsty (polydispia) and hungry (ploy hagia).

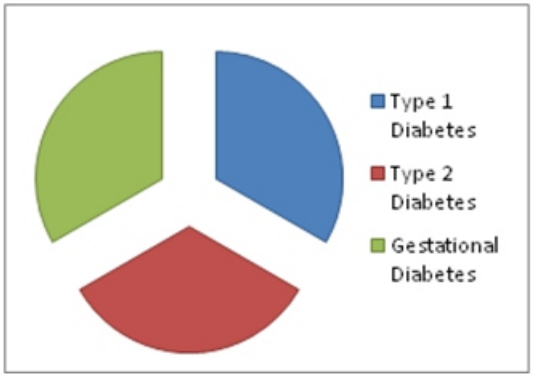


Figure 1: Diabetes Mellitus Types.

Diabetes Mellitus are separated into three types.

Type I Diabetes Mellitus:

It results from the body's failure to produce enough insulin. This form was referred to as "Insulin-dependent diabetes mellitus" or "Juvenile diabetes". The cause is unknown.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Type II Diabetes Mellitus:

It begins with insulin resistance, a condition in which cells fail to respond to insulin properly. As the disease progresses a lack of insulin may also develop. This form was previously referred to as "Non-insulin-dependent diabetes mellitus" or "adult-onset diabetes". The primary cause is excessive body weight and not enough exercise.

Gestational Diabetes:

This is the third main form and occurs when pregnant women without a previous history of diabetes develop a high blood glucose level. Anticipation and management of diabetes mellitus involves a healthy diet, physical exercise, avoid using tobacco and being a normal body weight. Proper foot care and blood pressure control are also important for people with the disease [5].

II.RELATED WORKS:

Mohd Fauzi bin Othman and Thoms Moh Shan Yau [6], presented his paper is o examine the performance of different classification and clustering methods for a set of bulky data. The algorithms or methods tested are Bayes Network, Redial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithms.

Dr. N. Rajalingam, K. Ranjini [16], presented a comparative study of implementation of hierarchical clustering algorithm- agglomerative and divisive clustering for various attributes. The Visual Programming Language is used for implementation of these algorithms. The result of this paper study is the performance of divisive algorithm works as twice as fast as the agglomerative algorithm.

Khaled Hammouda, Prof. Fakhreddine Karray [14], presented the reviews of four off-line clustering algorithms are K-means clustering, Fuzzy C-means clustering, Mountain clustering , and Subtractive clustering. The algorithms are implemented and tested against a medical problem of heart disease diagnosis. The accuracy and performance are compared. Aastha Joshi, Rajneet Kaur [17], proposed a brief review of six different types of clustering techniques are K-means clustering, Hierarchical clustering, DNSCAN clustering, OPTICS and STING. Manish Verma, et.al [18], proposed as analysis of six type of clustering techniques are K-Means, Hierarchical, DBSCAN, OPTICS, Density Based and EM clustering Algorithm. WEKA tool is used for implemented and analyzed.Bharat Chaudharil, Menan Parikh [7], analyzed the comparison of three major clustering algorithms are K-Means, Hierarchical and Density Based clustering algorithm. In this paper, the performances of these three algorithms are compared based on the feature of correctly class wise clustering. The performance of these three clustering algorithms is compared using a Data Mining tool WEKA.

P. Thangaraju and B.Deepa [3], proposed a survey on preclusion and discovery of skin melanoma risk using clustering techniques. The skin melanoma patient's data are gathered from different diagnostic center which contains both cancer and non-cancer patient's information. The gathered data are pre-processed and then clustered using K-means algorithm for separating relevant and non-relevant data to skin melanoma. Then significant frequent patterns are discovered using MAFIA algorithm.

Sharddha K.Popat, et.al [19], focused on survey of different clustering techniques. They are partitional algorithms, Hierarchical algorithms, Density Based algorithms. The result of this survey was hierarchical clustering can be perform better than the other techniques. Pradeep Rai and Shubha Singh [23], presented a survey is to provide a comprehensive review of different clustering techniques in data mining.

III.BACKGROUND:

A Decision Support System (DSS) is a computer based information system that support organizational decision-making activities. DSS's serve the management, operations, and planning levels of an organization and help to make decision, which may be rapidly changing and not easily specified in advance. DSS can be either fully computerized of both.DSS's include knowledge-based system. A properly designed DSS is an interactive software-based system intended to help decision makers compile useful information from a combination of raw data, documents, and personal knowledge, to identify and solve problems and make decision. Now a day's data mining helps the users of any DDS by reducing the time required for decision making activities.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

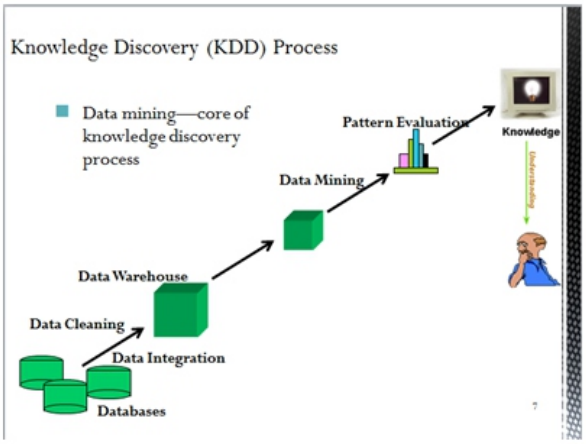


Fig: 3.1 Data Mining

Data mining is about finding new information in a lot of data. The information obtained from data mining is both new and useful. In many cases, data is stored so it can be used later. Now a day's data mining may take important role in several areas like Medical Science, Railway etc. Clustering is one of the techniques which are widely using forecasting diabetes mellitus. Many data mining tool are available in the market such as WEKA etc. for applying the functions of data mining.

IV.METHODOLOGY:

4.1.WEKA:

WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. Weka is open source software which consists of a collection of machine learning algorithms for data mining tasks [6]. WEKA is freely available and it is also platform-independent [7].

The WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation. This means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.

4.2.CLUSTERING:

Cluster analysis or Clustering is the process of partitioning a group of data objects into subsets. The main aim is that the objects in a group will be similar to one another and different from the objects in other groups. The objects are similarity within a group is larger than that among groups [14]. The set of clusters resulting from a cluster analysis can be referred to as a Clustering [2].

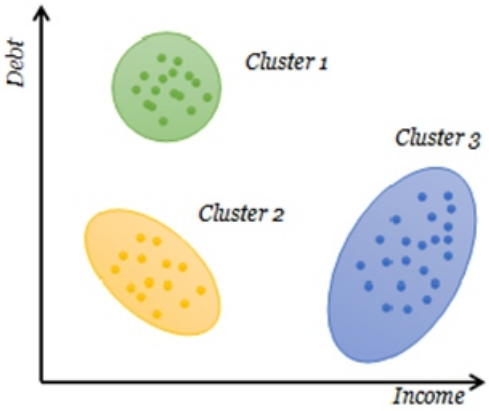


Fig: 4.2.1 Clustering Elements

4.3.Simple k-means clustering:

K-means is an unsupervised learning and iterative clustering algorithm in which objects are moved among sets of clusters until the desired set is reached. Within a cluster, a centroid denotes a cluster, which is a mean point within cluster [10].The main goal of the K-means clustering is to subset n observations into K clusters in which each observation belongs to the cluster with the nearest mean [3]. In K- means algorithm, the numerical attributes are works competently. K-means clustering tool is widely used in industrial and scientific applications [10].

4.4.Hierarchical clustering:

Hierarchical clustering is a method of cluster analysis which hunts to a construct a hierarchy of cluster. A tree data structure called a Dendrogram, can be used to illustrate the hierarchical clustering technique and the sets of different clusters. The root in a dendrogram contains one cluster where all elements are together the leaves in a dendrogram each consist of a single element cluster.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Internal nodes represent new cluster formed by merging the cluster that appears as its children. Each level is associated with the distance measure that was used to the cluster.

Hierarchical Clustering are two types. They are, Agglomerative- Each observation stars in its own cluster and pairs of cluster are merged as one move up the hierarchy. Agglomerative clustering is a "Bottom Up" approach. Divisive – All observations starts in one cluster and splits are performed recursively as one moves down the hierarchy. Divisive clustering is a "Top Down" approach.

4.5.Density based clustering:

Density based clustering algorithm try to seek clusters based on higher density of data points in a region [13]. For each instance of a cluster, the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances [7]. One of the most density based clustering algorithm is the DBSCAN.

DBSCAN data points are separated into three classes. They are:

» Core Points: Core points are the interior of cluster.

» Border Points: A border point is not a core point, but it falls within the neighborhood of a core point.

» Noise Points: It is not a core point or a border point. A noise point is any point.

Another density based clustering algorithm is OPTICS. It is an iterative clustering algorithm. It performs by creating an ordering of the data set representing its density based clustering structure.

V. DATASET DESCRIPTION:

In our work we have used Pima Indian Diabetes Dataset[12] for comparing the clustering algorithm for forecasting Diabetes Mellitus. The data set consist of 8 attributes that are used to forecast the Diabetes Mellitus. The details descriptions of the attributes are given as in the table 5.1.

No	Name of the Attributes	Description
1	Plas	Plasma glucose concentration a 2 hours in a oral glucose tolerance test
2	Pres	Diastolic blood pressure (mm Hg)
3	Skin	Triceps skin fold thickness (mm)
4	Insu	2-Hour serum insulin (mu U/ ml)
5	Mass	Body mass index (weight in kg/ (height in m)^2)
6	Pedi	Diabetes pedigree function
7	Age	Age (years)
8	Class	Class Variable (0 or 1)

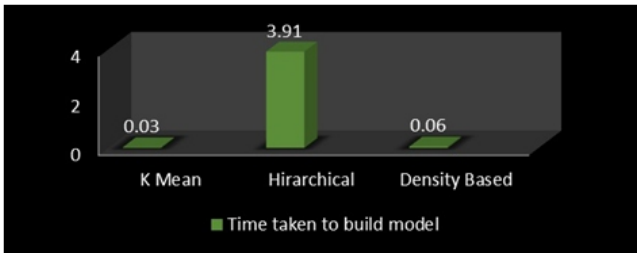
Table 5.1: Attributes for Diabetes Mellitus:

The attributes are given based on data types. The data set is based on the numeric and nominal data type.

VI.EXPERIMENTAL RESULTS:

The given three types of algorithms as K-means clustering, Hierarchical clustering and Density based clustering are applied on the Diabetes Mellitus data set in WEKA and the performance of the algorithm are given based various factors. The performance can be obtained based on the time taken to build model and correctly clustered instances. Table 6.1 describes that the time taken to a model by each algorithms. Figure 6.1 represents that the graph of time taken by the algorithms to build a model.

Name of the Algorithm	Time Taken to build model
K-means Clustering	0.03 seconds
Hierarchical clustering	3.91 seconds
Density based clustering	0.06 seconds



ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Table 6.1 Time taken by the algorithms:

X-Axis: Clustering Algorithms, Y-Axis: Time Range Figure 6.1 Performance of the Algorithms based on the time taken the dataset consists of 768 instances and they are applied as a test case in the clustering algorithms. The performance of the algorithms can be known from the instances that are clustered. The instances which are clustered using the WEKA tool can be given as table 6.2.

Name of the Algorithm	No. of clusters	Clustered instances	No. of Iterations	Within clusters sum of squared error	Log likelihood	Unclustered Instances
K-means	2	0: 500 (65%) 1: 268 (35%)	4	149.517766 4581119	–	0
Hierarchical	2	0: 268 (35%) 1: 500 (65%)	–	–	–	0
Density based clustering	2	0: 495 (64%) 1: 273 (36%)	–	–	-30.21166	0

Table 6.2 Comparison result of algorithms using WEKA tool:

VII.CONCLUSION:

Data mining plays an important role in extracting the hidden information in the medical database. The data pre-processing is used in order to improve the quality of the data. The experiment has been successfully performed with several data mining clustering techniques and it is found that the K-means algorithm gives a better performance over the supplied data set with the time taken of 0.03%. It is believed that the data mining can significantly help in the Diabetes Mellitus research and finally improve the quality of health care of Diabetes Mellitus patients.

In this research paper we have taken the time taken to build a model by the algorithms as a parameter for clustering data. The future work of this paper will be taken the quality of the clustered data as a parameter for clustering the dataset. It can also be implemented using several clustering techniques.

REFERENCE:

[1]http:// www.diabetesaustralia.com.au/Understanding-Diabetes/What is Diabetes.

[2]Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, second Edition (2006).

[3]P.Thangaraju and B.Deepa. "A Case study on Perclusion and Discovery of Skin Melanoma Risk using clustering techniques", International Journal of Advance Research in Electronics and Communication Engineering (IJARECE), Volume 3, Issu e 7, July 2014.

[4]http://www.medicalnewstoday.com/info/diabetes.

[5]http:// en.wikipedia.org/wiki/Diabetes_mellitus.

[6]Mohd Fauzi bin Othama, Thomas Moh Shan Yau, "Comparison of Different Classification Techniques using WEKA for Breast Cancer", F.Ibrahim, N.A. Abu Osman, J. Usman and N.A. Kadir (Eds.): Biomed 06, IFMBE Processing 15, pp.520-523,2007.

[7]Bharat Chaudharil, Manan Parikh, "A Comparative Study of Clustering Algorithms using WEKA tool", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 1, Issue 2, Oct 2012 ISSN 2319-4847.

[8]http:// en.wikipedia.org/wiki/Hierarchical_clustering.

[9]Data mining Introductory and Advance Topics, Margaret H.Dunham.

[10]Amandeep Kaur Mann and Navneet Kaur, "Survey Paper on clustering Techniques", International Journal of Science Engineering and Technology Research (IJSETR) Volume 2 Issue 4, April 2013.

[11]Kawsar Ahmed, Abdullah-Al-Emran, Tasnuba Jesmin, Roushney Fatima Mukti,, Md Zamilur Rahman and Farzana Ahmed, "Early Detection of Lung Cancer Risk Using Data Mining", Asian Pacific Journal of Cancer Prevention, Volume 14, 2013.

[12]UCImachine learning repository and archive. ics.uci.edu/ml/datasets/Diabetes+130US+hospitals+for+years+1999 2008.

[13]http://en.wikipedia.org/wiki/Cluster_analysis$Density-based_clustring.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

[14]Khaled Hammouda, Prof.Fakhreddine Karrey, "A Comparative Study of Data clustering techniques", University of Waterloo, Ontario, Canada.

[15]Han J. and Kamber M.:"Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000.

[16]Dr. N. Rajalingam, K. Ranjini, "Hierarchical Clustering Algorithm – A Comparative Study", International Journal of Computer Applications (0975-8887), Volume 19-No 3, April 2011.

[17]Aastha Joshi Rajneet Kaur, " A Review : Comparative Study Of Various Clustering Technique in Data mining", International Journal Of Advance Research in Computer Science and Software Engineering, Volume 3 Issue 3, March 2013.

[18]Manish Verma, et.al, "A Comparative study of Various Clustering Algorithms In Data Mining" International Journal of Engineering Research and Application (IJERA), Vol.2 Issue 3, May-Jun 2012, pp.1379-1384.

[19]Sharddha K. Popat, et.al, "Review and Comparative Study of Clustering Techniques" International Journal of Computer Science and Information Technology, Volume 5(1), 805-812, 2014.

[20]K. RuthRamya, et. al, "A Class Based Approach for Medical Classification of Chest Pain", International Journal of Engineering Trends and Technology, Vol. 3, Issue.2, pp.89-93, 2012.

[21]Rui Xu, "Survey Of Clustering Algorithms" IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 16, no. 3, May 2005.

[22]Prof. Pier Luca Lanzi, "Density-Based, Grid-Based and Model-Based Clustering", Data Mining and Text Mining (583@Politecnico di Milano).

[23]Pradeep Rai and Shubha Sing, "A Survey Of Clustering Techniques", International Journal of Computer Application, Volume 7-No.12, Oct 2010.