ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

Comparison of Data mining Techniques for Forecasting Diabetes Mellitus

P. Thangaraju[1], B.Deepa[2], T.Karthikeyan[3]

Assistant Professor, Department of Computer Applications, Bishop Heber College (Autonomous), Tiruchirappalli, India[1]

M.Phil, Scholar, Department of Computer Applications, Bishop Heber College (Autonomous), Tiruchirappalli, India[2]

Associate Professor, PSG College of Arts & Science, Coimbatore, India[3]

Abstract: Data mining is the practice of examining large pre- existing databases in order to generate new information. There are different kinds of data mining techniques are available. Classification, Clustering, Association Rule and Neural Network are some of the most significant techniques in data mining. In Health care industries, Data mining plays a significant role. Most frequently the data mining is used in health care industries for the process of forecasting diseases. Diabetes is a chronic condition. This means that is lasts for a long time, often for someone's whole life [1]. This paper studies the comparison of diabetes forecasting approaches 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 is used as a tool.

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

I. INTRODUCTION

Diabetes Mellitus also known as simply Diabetes is a group of metabolic diseases in which the person has high blood glucose, either because insulin production is inadequate, or because the body's cells do not respond properly to insulin, or both [4]. The causes of diabetes vary depending on their genetic makeup, family, ethnicity, health and environmental factors. There is no common diabetes cause that fits every type of diabetes. The reason there is no defined diabetes cause is because the causes of diabetes vary depending on the individual and the type [4]. Diabetes Mellitus are separated into three types [5]:

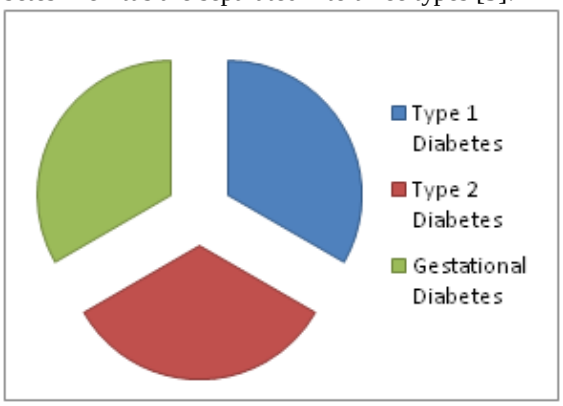


Figure 1: Diabetes Mellitus Types.

Type 1 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.

Type 2 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

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 centre 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. Finally implement a system using c#.net to predict skin melanoma risk level with suggestions which is easier, cost reducible and time savable.

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

Bharat Chaudharil, Manan Parikh [7], analyzed the comparison of three major clustering algorithms are K-Means, Hierarchical clustering 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.

Amandeep Kaur Mann and Navneet Kaur [10], presented a survey paper, a review of clustering and its different techniques in data mining is done. Kawsar Ahmed, et.al [11], in his paper, they proposed a system to detect the Lung cancer risk. Their proposed system was easy, cost effective and time saving. The data are collected from different diagnosis centres. The collected data are preprocessed and clustered using K-means algorithm. Then AprioriTid and Decision tree algorithm are used to find significant frequent pattern. Then they developed a significant frequent pattern tool for lung cancer prediction system.

Dr.N. Rajalingam, K. Ranjini [16], presented a comparative study of implementation of hierarchical clustering algorithms- 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, DBSCAN clustering, OPTICS, and STING. Manish Verma, et. al [18], proposed a analysis of six types of clustering techniques are k-Means Clustering, Hierarchical Clustering, DBSCAN clustering, Density Based Clustering, Optics and EM Algorithm. WEKA tool is used for implemented and analyzed.

Shraddha K.Popat, et.al [19], focused on survey of different clustering techniques. They are Partitional algorithms, Hierarchical algorithms, Density based clustering algorithm. 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

Data mining is the practice of examining large pre-existing databases in order to generate new information. Data Mining involves the following performance such as extract, transform, and load transaction data onto the data warehouse system, Store and manage the data in the multidimensional database system, Provide data access to business analysts and information technology professionals, Analyze the data by application software, and Present the data in a useful format [3].

The different steps in Data Mining are Selection, Preprocessing, Transformation, Data Mining and Pattern evaluation [3]. Data mining has some techniques to examine the data. They are classification, clustering, correlations, association rule etc and it has been used intensively and widely by many organizations [11]. Nowadays data mining may take important role in several areas like Medical, Science, Railway etc. Clustering is one of the techniques which is widely using for forecasting diabetes mellitus.

IV. METHODOLOGY

4.1 WEKA Tool

WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, NewZealand. 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].

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].

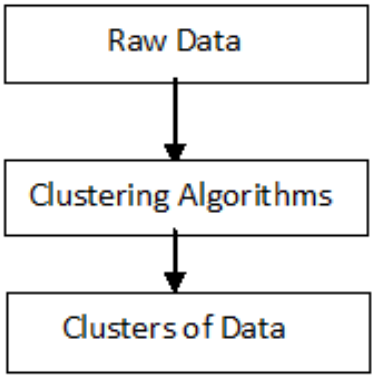


Figure 2: Stages of clustering [10].

Clustering analysis has been widely used in many applications such as business intelligence, image pattern recognition, web search, biology, and security. Clustering is known as unsupervised learning [2]. The different types of clustering techniques are available. But here we are using K-means clustering, Hierarchical clustering and Density based clustering techniques.

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

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

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].

4.3 Hierarchical clustering:

Hierarchical clustering is a method of cluster analysis which hunts to construct a hierarchy of clusters. 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. Internal nodes represent new clusters formed by merging the clusters that appear as its children. Each level is associated with the distance measure that was used to merge the clusters.

Hierarchical clustering are two types. They are,

Agglomerative- Each observation starts in its own cluster and pairs of clusters are merged as one move up the hierarchy. Agglomerative clustering is a "Bottom up" approach.

Divisive- All observations start in one cluster and splits are performed recursively as one moves down the hierarchy. Divisive clustering is a "Top down" approach.

4.4 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 (Minpts) [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 at the interior of a 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 algorithms for forecasting Diabetes Mellitus. The data set consists of 9 attributes that are used to forecast the Diabetes Mellitus. The detail descriptions of the attributes are given as in the table 5.1.

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

Table 5.1: Attributes for Diabetes Mellitus

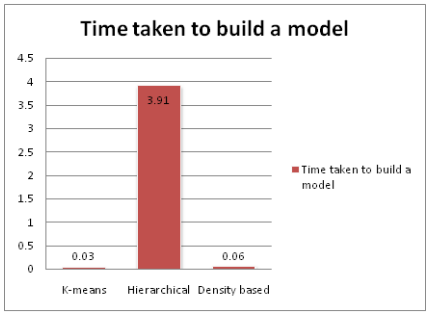
The attributes are given based on data types. The data set is based on the numeric and nominal data type.

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

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

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 8, August 2014

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.

Table 6.2 Comparison result of algorithms using WEKA tool

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

VII. CONCLUSION

Data mining plays a major role in extracting the hidden information in the medical data base. The data pre-processing is used in order to improve the quality of the data. This model is built based as a test case on the UCI repository dataset. 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 ultimately improve the quality of health care of Diabetes Mellitus patients. It can also be implemented using several clustering techniques.

In this paper we have taken the time taken to build a model by the algorithms as a parameter for clustering the dataset. 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 Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 3, Issue 7, July 2014
[4] http://www.medicalnewstoday.com/info/diabetes
[5] http://en.wikipedia.org/wiki/Diabetes_mellitus
[6] Mohd Fauzi bin Othman, 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. Kadri (Eds.): Biomed 06, IFMBE Proceedings 15, pp.520-523, 2007
[7] Bharat Chaudharil, Manan Parikh, "A Comparative Study of clustering algorithms using weka tools", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 1, Issue 2, October 2012 ISSN 2319-4847.
[8] http://en.wikipedia.org/wiki/Hierarchical_clustering
[9] Data mining Introductory and Advanced 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] UCI machine learning repository and archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008.
[13] http://en.wikipedia.org/wiki/Cluster_analysis#Density-based_clustering
[14] Khaled Hammouda, Prof. Fakhreddine Karray, "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 Techniques in Data Mining", International journal of Advanced 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 Applications (IJERA), Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.
[19] Shraddha K. Popat, et. al, "Review and Comparative Study of Clustering Techniques" International Journal of Computer Science and Information Technologies, 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 (UIC 583 @ Politecnico di Milano)
[23] Pradeep Rai and Shubha Singh , "A Survey of Clustering Techniques", International Journal of Computer Applications, Volume 7-No. 12, October 2010.