ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology
Vol. 3, Special Issue 20, April 2016

An Efficient Study on Usage of Data Mining Techniques for Predicting Diabetes

R. Manimaran[#1] and Dr. M.Vanitha [*2]
 # Research Scholar, PG and Research Department of Computer Science, J.J.College of Arts and Science (Autonomous), Pudukkottai,Tamilnadu,India.
Email: rmanimaran_jjc@rediffmail.com
**Assistant professor, Department of Computer Science and Engineering, Alagappa University,Karaikudi, India.*
Email: mvanitharavi@gmail.com

Abstract: **Data mining extracts the knowledge from a large amount of data which stores in multiple heterogeneous data base. Knowledge is conveying the message through direct or indirect. This paper provides a survey of various data mining techniques. These techniques include Clustering and Classification. This research paper also conducts a formal review of the application of data mining such as the Diabetes. This paper discusses the topic based on past research paper and also studies on the data mining techniques.**

***Keywords: Clustering*, and Classification.**

1. Introduction

Data mining is a powerful concept for data analysis and process of discovery interesting pattern from the huge amount of data, data stored in various databases such as data warehouse, World Wide Web, external sources. Interesting pattern that is easy to understand, unknown, valid, potential useful [1]. Data mining is a type of sorting technique which is actually used to extract hidden patterns from large databases. The goals of data mining are fast retrieval of data or information, knowledge Discovery from the databases, to identify hidden patterns and those patterns which are previously not explored, to reduce the level of complexity, time saving, etc[2]. Data mining refers extracting knowledge and mining from large amount of data. Sometimes data mining treated as knowledge discovery in database (KDD)[3] . KDD is an iterative process, consist a following steps[4]

Selection: select data from various resources where operation to be performed.

- **Preprocessing:** also known as data cleaning in which remove the unwanted data.
- **Transformation:** transform into a new format for processing.
- **Data mining:** identify the desire result.
- **Interpretation:** interpret the result/query to give meaningful report/information

1.1 Diabetes

Diabetes is an appropriate disease for data mining technology due to a number of reasons. In every age group this disease is common. It charges plenty of money and its effect is growing quickly. The body of a diabetic person does not produce or efficiently use insulin, the hormone that "unlocks" the cells of the body, allowing glucose to arrive and fuel them. A diabetic person has risk of having the other diseases as blood vessel harm, blindness, heart disease, nerve damage and kidney disease [5]. Diabetes is generally of 2 kinds: type 1(insulin dependent diabetes) and type 2(non-insulin-dependent diabetes).

Diabetes is a disease in which the blood glucose levels get increase which is due to the defects in secretion of insulin, or its action, or both. Diabetes is a prolonged medical disease. In diabetes, the cells of a person produce insufficient amount of insulin or defective insulin or may unable to use insulin properly and efficiently that further leads to hyperglycemia and type-2 diabetes. In type 1 diabetes there is absolute lack of insulin, usually secondary to a destructive process distressing the insulin-producing beta cells in the pancreas. There is excess decline of beta cells that enhances process of elevated blood sugars in type 2 diabetes. In current time it is one of the major public health problems. The International Diabetes Federation has claimed that presently 246 million people are suffering from diabetes worldwide and this number is expected to increase up to 380 million by 2025 [6].

2. Data Mining Techniques

Data mining means collecting relevant information from unstructured data. So it is able to help achieve specific objectives. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model .A descriptive model presents, in concise form, the main characteristics of the data set. The purpose of a predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable; the target variable [7]. The goal of predictive

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology
Vol. 3, Special Issue 20, April 2016

and descriptive model can be achieved using a variety of data mining techniques [8].

2.1 Classification: Classification based on categorical (i.e. discrete, unordered) . This technique based on the supervised learning (i.e. desired output for a given input is known) .It can be classifying the data based on the training set and values (class label). These goals are achieve using a decision tree, neural network and classification rule (IF- Then).for example we can apply the classification rule on the past record of the student who left for university and evaluate them. Using these techniques we can easily identify the performance of the student.

2.2 Regression: Regression is used to map a data item to a real valued prediction variable [8]. In other words, regression can be adapted for prediction. In the regression techniques target value are known. For example, you can predict the child behavior based on family history.

2.3 Time Series Analysis: Time series analysis is the process of using statistical techniques to model and explain a time-dependent series of data points. Time series forecasting is a method of using a model to generate predictions (forecasts) for future events based on known past events [9]. For example stock market.

2.4 Prediction: It is one of a data mining techniques that discover the relationship between independent variables and the relationship between dependent and independent variables [4].Prediction model based on continuous or ordered value.

2.5 Clustering: Clustering is a collection of similar data object. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic. This technique based on the unsupervised learning (i.e. desired output for a given input is not known). For example, image processing, pattern recognition, city planning.

2.6 Summarization: Summarization is abstraction of data. It is set of relevant task and gives an overview of data. For example, long distance race can be summarized total minutes, seconds and height.

2.7 Association Rule: Association is the most popular data mining techniques and fined most frequent item set. Association strives to discover patterns in data which are based upon relationships between items in the same transaction. Because of its nature, association is sometimes referred to as "relation technique". This method of data mining is utilized within the market based analysis in order to identify a set, or sets of products that consumers often purchase at the same time [6].

2.8 Sequence Discovery: Uncovers relationships among data [8]. It is set of object each associated with its own timeline of events. For example, scientific experiment, natural disaster and analysis of DNA sequence.

3. RELATED WORKS

P. Thangaraju and B.Deepa [10], 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 nonrelevant 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 [11], 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 [12], analyzed the comparison of three major clustering algorithms are KMeans, 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 [13], presented a survey paper, a review of clustering and its different techniques in data mining is done. Kawsar Ahmed, et.al[14], 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 predictionsystem.

Dr.N. Rajalingam, K. Ranjini [15], 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 [16], 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

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology
Vol. 3, Special Issue 20, April 2016

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 [20], presented a survey is to provide a comprehensive review of different clustering techniques in data mining.

4. METHODOLOGY

4.1 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 [16]. The set of clusters resulting from a cluster analysis can be referred to as a Clustering [21].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 [22]. The different types of clustering techniques are available. But here we are discussing K-means clustering techniques.

4.1.1 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 [13]. 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 [10]. In K-means algorithm, the numerical attributes are works competently. K-means clustering tool is widely used in industrial and scientific applications [13].

4.2 Classification

Classification is the process of finding a set of models that describe and distinguish data classes and concepts, for the purpose of being able to use the model to predict the class whose label is unknown.

4.2.1. C4.5 Algorithm:

Decision trees are controlling categorization algorithms. Accepted decision tree algorithms consist of C4.5. At the equivalent time as the name imply, this performance recursively separate inspection in branches to build tree for the purpose of improving the calculation accuracy. Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. C4.5 generates classifiers expressed as decision trees, but it can also construct Classifiers in more comprehensible rule set form [22].

4.2.2. The K-Nearest Neighbour Algorithm:

The *k*-nearest neighbour's algorithm is a technique for classifying objects based on the next training data in the feature space. It is among simplest of all mechanism learning algorithms. Then the algorithm iterates between two steps till junction:

Step 1: Data Assignment each data point is assign to its adjoining centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of "means". Each group representative is relocating to the center (mean) of all data points assign to it. If the data points come with a possibility measure (Weights), then the relocation is to the expectations (weighted mean) of the data partitions [23].

4.2.3. Naive Bayes Algorithm:

Naïve Bayes Classifier is a term dealing with simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. It assumes that the presence or absence of particular feature of a class is unrelated to the presence or absence of any other feature [23].

The Naive Bayes algorithm is based on conditional probabilities. It uses Bayes' theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows.

Prob (B given A) = Prob (A and B) /Prob (A)

To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone [24].

4.2.4. Support Vector Machines (SVM):

Support vector machines are a moderately new-fangled type of learning algorithm, originally introduced. Naturally, SVM aim at pointed for the hyper plane that most excellent separates the classes of data. SVMs have confirmed the capability not only to accurately separate entities into correct classes, but also to identify instance whose establish classification is not supported by data. Although SVM are comparatively insensitive define distribution of training examples of each class. SVM can be simply extended to perform numerical calculations. Two such extension, the first is to extend SVM to execute regression analysis, where the goal is to produce a linear function that can fairly accurate that target function. An extra extension is to learn to rank elements rather than producing a classification for individual

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology
Vol. 3, Special Issue 20, April 2016

elements. Ranking can be reduced to comparing pairs of instance and producing a +1 estimate if the pair is in the correct ranking order in addition to −1 otherwise [25].

4.2.5. The Apriori Algorithm:

One of the most popular data mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum sup-port) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candi-date generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order[26].

5.RESEARCH FINDINGS

5.1. DATASOURCE:

Table 1. Attributes of diabetes dataset [27] [28].

NO	NAME	DESCRIPTION
1	Pregnancy	Number of times pregnant
2	Plasma	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Pres	Diastolic blood pressure (mm Hg)
4	Skin	Triceps skin fold thickness (mm)
5	Insulin	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)

5.2. PERFORMENCE OF ANALYSING ALGORITHM

Table 2. Performance of Analyzing Algorithm [28] [29].

Algorithm used	Accuracy %
C4.5	86%
SVM	74.8%
KNN	78%
Naïve Bayes	75%
Apirori	75%
K-Mean algorithm	97%

From Table 2 [28],[29], it is been proved that K-Mean algorithm provides an accurate result when compared to Classification techniques C4.5 SVM, KNN, Naïve Bayes, and Apriori.

6. CONCLUSION

There are different data mining classification and clustering techniques can be used for the identification and prevention of diabetes disease among patients. This paper describes some classification and clustering techniques in data mining to predict diabetes disease in patients. Namely: C4.5, SVM, K-NN, Naive Bayes, Apriori and K-Mean algorithm. These techniques are compared by disease among patients using six algorithms accuracy. Then one has highest accuracy above 97%. The K-mean algorithm is best among sixes. They are used in various healthcare units all over the world. In future work we review various classifications and clustering algorithm and its significance's

REFERENCES:

[1] Yongjian Fu " data mining: task, techniques and application"

[2] Er. Rimmy Chuchra "Use of Data Mining Techniques for the Evaluation of Student Performance:A Case Study" International Journal of Computer Science and Management Research Vol 1 Issue 3 October 2012

[3] J. Han and M. Kamber. "Data Mining, Concepts and Techniques", Morgan Kaufmann, 2000.

[4] Aakanksha Bhatnagar, Shweta P. Jadye, Madan Mohan Nagar" Data Mining Techniques & Distinct Applications: A Literature Review" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9, November- 2012

[5] Huang, Feixiang; Wang, Shengyong; Chan, Chien-Chung, "Predicting disease by using data mining based on healthcare information system," Granular Computing (GrC), 2012 IEEE International Conference on , vol., no., pp.191,194, 11-13 Aug. 2012 .

[6] Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." In Internet Technology And Secured Transactions, 2012 International Conference For, pp. 471-472. IEEE, 2012.

[7] Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013,

[8] Dr. M.H.Dunham, "Data Mining, Introductory and Advanced Topics", Prentice Hall, 2002.

[9] Time Series Analysis and Forecasting with Weka , http://wiki.pentaho.com/display/datamining

[10] 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

[11] 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

[12] 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.

[13] Amandeep Kaur Mann and Navneet Kaur, "Survey Paper on Clustering Techniques", International Journal of Science, Engineering and Technology Research (IJSETR) Volume2, Issue4, April 2013.

[14] 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.

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

[16] Khaled Hammouda, Prof. Fakhreddine Karray, "A Comparative Study of Data Clustering Techniques", University of Waterloo,Ontario, Canada

[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] Pradeep Rai and Shubha Singh , "A Survey of Clustering Techniques", International Journal of Computer Applications,Volume 7-No. 12, October 2010.

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

[22] . N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik."Data Mining Machine Learning Approaches and Medical Diagnose Systems" A Survey. International journal of computer applications, Vol. 2, No. 2, 2009.

[23]. Nitin Bhatia,Vandana,| Survey of Nearest Neighbor Techniques| (IJCSIS) Vol. 8, No. 2, 2010, ISSN 1947-5500.

[24]. Vidhya.K.A, G.Aghila| A Survey of Naïve Bayes Machine Learning approach in Text Document Classification| (IJCSIS) Vol. 7, No.2, 2010.

[25]. J. Smola, B. Scholkopf, |A tutorial on support vector regression|, Stat Comput 14 (2004) 199–222.

[26]. Charanjeet Kaur, —Association Rule Mining using Apriori Algorithm: A Survey|, IJARCET Volume 2, Issue 6,June 2013.

[27]. G.Visalatchi *et al*, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014, pg. 100-105 **ISSN: 2321-8363**

[28]. " Data Mining based Prediction of Medical data " Dr. Bushra M. Hussan Computer Science Department - College of Science - Basrah University.