International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Review on Diagnosis of Diabetes in Data Mining

Sukhjinder Singh[1], Kamaljit Kaur[2]

[1]M.Tech Student, Department of Computer Science & EngineeringSGGSWU, Fatehgarh Sahib

[2]Assistant Professor, Department of Computer Science & Engineering SGGSWU, Fatehgarh Sahib

Abstract: ***Data Mining is used for various purposes in many applications like industries, medical etc. This is used for extracting the useful information from the huge amount of data set. Health monitoring is also used the data mining concept for predict the diagnosis of the diseases. In health monitoring diabetes is the common health problem nowadays, which affects peoples. There are various data mining techniques and algorithm is used for finding the diabetes. Neural Network, Artificial neural fuzzy interference system, K-Nearest-Neighbor (KNN), Genetic Algorithm, Back Propagation algorithm etc. These techniques and the algorithms provide the better result to the people and the doctors regarding the diagnosis of the diabetes. From these results the people can predict he is affected with the diabetes or non-diabetes.***

Keywords: Data Mining, Artificial neural fuzzy interference system, K-Nearest-Neighbor (KNN), Machine Learning (ML), Principal Component Analysis (PCA)

1. Introduction

Knowledge discovery in databases is well-defined process consisting of several distinct steps. In Fig: 1 shows the architecture of Knowledge Discovery in Database. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows: ―Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Anticipating patient's future behavior on the given history is one of the important applications of data mining techniques that can be used in health care management. Health care organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care[5].

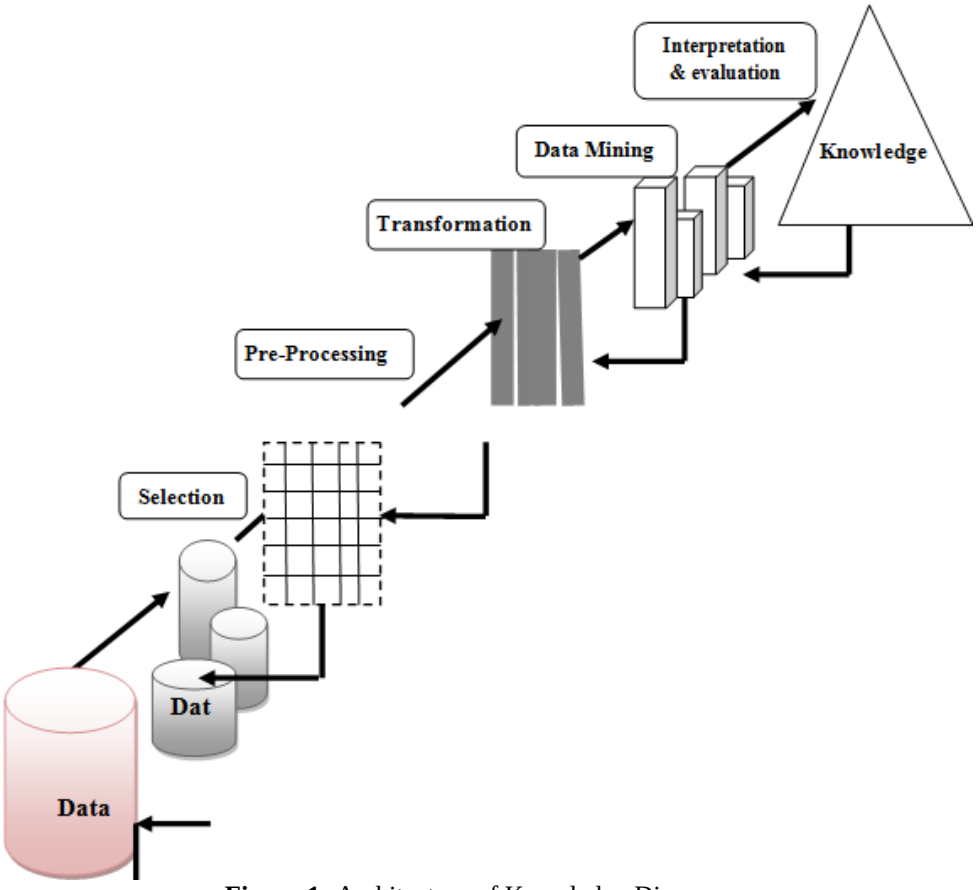

Figure 1: Architecture of Knowledge Discovery

2. Literature Survey

Chaudhari et al[3] Disease diagnosis is one of the most important applications of such system as it is one of the leading causes of deaths all over the world. Predict the human use the inputs from complex tests conducted in labs and also predict the disease based on risk factors such as tobacco smoking, alcohol intake, age, family history, diabetes, hypertension, high cholesterol, physical inactivity, obesity. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. K-Nearest-Neighbor (KNN) is one of the successful data mining techniques used in classification problems. Recently, researchers are showing that combining different classifiers through voting is outperforming other single classifiers.This paper investigates applying KNN to help healthcare professionals in the diagnosis of disease specially heart disease. It also investigates if integrating voting with KNN can enhance its accuracy in the diagnosis of heart disease patients. The results show that applying KNN could achieve higher accuracy than neural network ensemble in the diagnosis of heart disease patients. The results also show that applying voting could not enhance the KNN accuracy in the diagnosis of heart disease.

Prof.Mythili et al[2]Diabetes mellitus, in simple terms called as diabetes, is a metabolic disease, where a person is affected with high blood glucose level. Diabetes is a metabolic disorder caused due to the failure of body to produce insulin or to properly utilize insulin. This condition arises when the body does not produce enough insulin, or because the cells do not respond to the insulin that is produced. Blood glucose test is the crucial method for diagnosing diabetes. Also, there have been numerous computerized methods proposed for diagnosis of diabetes. All these methods have some input values which would be the result of different tests that should be carried out in hospitals. This paper proposes a methodology that aims to ease the patients undergoing various medical tests, which most of them consider as a tedious task and time consuming. The parameters identified for diagnosing diabetes have been designed in such a way that, the user can predict if he is affected with diabetes himself. Back Propagation algorithm is used for diagnosis.

Ahmed et al [5] Heart disease is a major cause of morbidity and mortality in modern society. Medical diagnosis is extremely important but complicated task that should be performed accurately and efficiently. The powerful data analysis tools are used to extract useful knowledge from the huge amount of medical data. There is a huge data available within the healthcare systems. However, there is a task of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found many applications in business and scientific domain. One of the applications is disease diagnosis where data mining tools are proving successful results. This research paper proposed to find out the heart diseases through data mining, Support Vector Machine (SVM), Genetic Algorithm, rough set theory, association rules and Neural Networks. In this study, we briefly examined that out of the above techniques Decision tree and SVM is most effective for the heart disease. So it is observed that, the data mining could help in the identification or the prediction of high or low risk heart diseases.

Thangaraju et al [6] 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 andNeural Network are some of the most significant techniques in data mining. In Health care industries, Data miningplays 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.

Durairaj et al [7]Neural Networks are one of the soft computing techniques that can be used to make predictions on medical data. Neural Networks are known as the Universal predictors. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The Artificial Neural Networks (ANNs) based system can effectively applied for high blood pressure risk prediction. This improved model separates the dataset into either one of the two groups. The earlier detection using soft computing techniques help the physicians to reduce the probability of getting severe of the disease. The data set chosen for classification and experimental simulation is based on Pima Indian Diabetic Set from (UCI) Repository of Machine Learning databases**.** In this paper, a detailed survey is conducted on the application of different soft computing techniques for the prediction of diabetes. This survey is aimed to identify and propose an effective technique for earlier prediction of the disease.

3. Techniques

Diabetes Mellitus has become a common health problem nowadays, which would affect people and lead to various complications like visual impairment, cardio vascular disease, leg amputation and renal failure if diagnosis is not done in the right time [2]. In this discussed the two classifier techniques with principal component analysis component analysis are implemented for the forecasting of Diabetes and concluded with best forecasting techniques which has a maximum accuracy [1]. These are given below:

1. Neural Network
2. Principal Component Analysis (PCA) with Neural Network
3. Artificial neural fuzzy interference system (ANFIS)
4. PCA with ANFIS

Data

For our work the Pima Indians Diabetic database is used from UCI Repository of Machine Learning Databases. . The data set was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney

Diseases. All patients in this database are Pima-Indian women at least 21 years old and living near Phoenix, Arizona, USA[1]. The binary value 0 and 1 is used for the negative and positive test. Binary value 0 is used for negative test and the value 1 is used for the positive test[1].

Principal Component Analysis

Principal component analysis (PCA) is a standard tool in modern data analysis. It is a simple non parametric method for extracting relevant information from confusing data sets. Principal components analysis method is used for achieving the simplification and generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data. The procedure can be followed in many ways i.e. a) Using singular value decomposition method (SVD) b) using the covariance matrix method. In this work we have used MATLAB software for deriving the principal components [1].

Neural Network Classifier

This phase involves the classification of the diabetic dataset utilizing Neural Networks as a classifier. Here the complete dataset is utilized for the training and testing purpose. The network is trained using the Back propagation algorithm and the testing is done through Post Regression analysis [8][9].

PCA with Neural Network Classifier

In this the dataset was also first reduced to a lower dimension and the reduced dataset is used for training through neural networks.The results of all the classifier techniques are tabulated and compared at the end in order to find the best classifier technique out of the four proposed ones [1].

ANFIS Classifier

In this phase the training and testing of the dataset will be done using only ANFIS as the classifier. First the dataset containing 768 samples and 8 features is selected for the training through ANFIS using MATLAB. In ANFIS the two different models is used. Result obtain with best model is maximum71% and quite similar to work done in past, hence this method is not suitable for the forecasting purpose[1].

PCA with ANFIS Classifier

In this section making the modification in the data set and evaluate the result at the end using the classifier ANFIS. Firstly the dataset is reduced to a lower dimension using Principal Component Analysis. After this the reduced dataset is trained usingthe ANFIS classifier and then the testing of the ANFIS model is carried out using the Cross validation approach.The accuracy of classification using only ANFIS classifier was nearly 71% so in order to improve the accuracy some kind of modification was required to be done in the classification technique. For that purpose Principal Component Analysis (PCA) was used along with ANFIS [1].

4. Conclusions

In this paper, the various techniques are discussed for predict the diagnosis of diabetes. Using the data mining technique the health care management predicts the disease and diagnosis of the diabetes and then the health care management can alert the human being regarding diabetes based upon this prediction. The Principal Component Analysis (PCA) is also the technique used for the analysis. The PCA is the feature extraction technique has more act upon on the accuracy of classification techniques. But when the PCA combined with the Neural Networks for classification achieved the best classification accuracy and the PCA performs better for non-diabetic samples than the diabetic samples when combined with Neural Networks. Classification speed of ANFIS is not better than the Neural Networks.

References

[1] Rakesh Motka, Viral Parmar, Balbindra Kumar, A. R. Verma, " Diabetes Mellitus Forecast Using Different Data Mining Techniques", International conference on computer and Communication Technology
[2] Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, "Diagnosis of Diabetes Mellitus based on Risk Factors", International Journal of Computer Applications, Vol.10, Issue No.4, November.2010
[3] Anand A. Chaudhari, Prof.S.P.Akarte, " Fuzzy and Data Mining based Disease Predection using K-NN Algorithm", International Journal of Innovations in Engineering and Technology, Vol. 3, Issue No. 4, April 2014
[4] Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar, " Diagnosis of Diabetes Mellitus based on Risk Factors", International Journal of Computer Applications, Vol. 10, Issue No. 4, November 2010
[5] Aqueel Ahmed, Shaikh Abdul Hannan, " Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering, Vol. 1, Issue No. 4, September 2012
[6] P. Thangaraju, B.Deepa, T.Karthikeyan, "Comparison of Data mining Techniques for Forecasting Diabetes Mellitus", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue No. 8, August 2014
[7] M. Durairaj, G. Kalaiselvi, " Prediction Of Diabetes Using Soft Computing Techniques- A Survey", International Journal of Scientific & Technology Research, Vol. 4, Issue No.3, March 2015
[8] S.F.B, Jaafar and Darmawaty Mohd Ali. "Diabetes Mellitus Forecast using Artificial Neural Network (ANN), Asian conference on sensors and the internationalconference on new techniques in pharmaceutical andmedical research proceedings (IEEE), Kuala Lumpur,Malaysia, 5-7 September 2005, pp 135-139.
[9] Dey R, Bajlai V and Gandhi G, et al. " Application of Artificial neural network technique for thediagnosing diabetes mellitus", IEEE Third International Conferenceon Industrial and Information System, Kharagpur, India , Page 1-4,2008.