Prediction Of Diabetes Using Soft Computing Techniques- A Survey

M. Durairaj, G. Kalaiselvi

Abstract: 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.

Index Terms: Artificial Neural Network (ANN), C4.5 Classifier, Support Vector Machine (SVM), K-Nearest Neighbour (KNN).

1 Introduction

HUMAN body needs energy for activation. The carbohydrates are broken down to glucose, which is the important energy source for human body cells. Insulin is needed to transport the glucose into body cells. The blood glucose is supplied with insulin and glucagon hormones produced by pancreas. Insulin hormones produced by the beta cells of the islets of langerhans and glucagon hormones are produced by the alpha cells of the islets of langerhans in the pancreas. When the blood glucose increases, beta cells are stimulated and insulin given to the blood. Insulin enables blood glucose to get into the cells and this glucose is used for energy. So blood glucose is kept in a narrow range. There are two types of diabetes such as type 1 and type 2. The insulin deficiency is the outcome of diabetes. The ANN models have been widely used in predicting the data like time-series. Number of data mining algorithms has been proposed to classify, predict and diagnose diabetes. To apply Neural Network, data preprocessing should be done. It is a technique that involves transforming raw data into understandable format. This helps to fill the missing values in between the data. By analyzing the data using the values, it is possible for an expert to find values that are unexpected and erroneous. In this paper, the surveys of recent soft computing application on the prediction of diabetes are presented. This paper also aims to propose an effective technique for earlier detection of the disease diabetes. This paper organized as follows: Section 1 discusses various tools applied on the prediction and analysis. Section 2 presents the survey of soft computing application on predicting diabetes. Section 3 compares these techniques and discussed. Section 4 concludes with our findings.

- *M. Durairaj, Assistant Professor, School of Computer Science, Engineering and Applications, Bharathidasan University, India, Mobile No:+91 9487542202, (e-mail: durairaj.bdu@gmail.com)*
- *G. Kalaiselvi, Research Scholar, School of Computer Science, Engineering and Applications, Bharathidasan University, India, Mobile No: +917639565848, (e-mail: gkalaiselvik@gmail.com)*

2 Diabetes

Types of diabetes are discussed in this section. Type 1 diabetes can occur at any age. However, it is most often diagnosed in children, adolescents, or a young adult in this diabetes occurs when the body's immune system is attacked and the beta cells of pancreas are destroyed. This results in insulin deficiency. The only treatment for this is insulin. Type-2 diabetes occurs when the pancreas does not produce enough insulin to meet the body's needs. This is the most common type of diabetes developed at the age of 40. Recent studies have shown that 80% of type-2 diabetes complications can be prevented by earlier identification.

3 Tools Used For Diabetes Prediction

There are different soft computing techniques and tools are applied for the prediction and data analysis. In this section, some of the techniques are discussed.

3.1 Artificial Neural Network

The artificial neural network is much similar as natural neural network of a brain. Artificial Neural Network (ANN) basically has three layers, they are;

Input layer: Input neurons define all the input attribute values for the data mining model, and their probabilities.

Hidden layer: Hidden neurons receive inputs from input neurons and provide outputs to output neurons. The hidden layer is where the various probabilities of the inputs are assigned weights. A weight describes the relevance or importance of a particular input to the hidden neuron. The neuron with greater weight is assigned to an input. The value of that input is more important weights can be negative, which means that the input can inhibit, rather than favour, a specific result.

Output layer: Output neurons represent predictable attribute values for the data mining model.

3.2 C4.5 Algorithm

Accepted decision tree algorithms consist of C4.5, the equivalent time as the name imply. The performance of C4.5 is 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.

4 Review of Related Work on Diabetes Prediction

Data mining is the extraction of useful information from the large volume of data [13]. Data mining has been applied in various fields like medicine, marketing, banking, etc. In medicine, predictive data mining is used to diagnose the disease at the earlier stages itself and helps the physicians in treatment planning procedure."Asha Gowda Karegowda, et.al. [1***]"*** provided the application of hybrid GA and BPN. They experimented for classification of PIMA dataset. They concluded that the Back Propagation learns by making modifications in weight values by using gradient method starting at the output layer then moving backward through the hidden layers of the network and hence is prone to lead to troubles such as local minimum problem, slow convergence pace and convergence unsteadiness in its training procedure. ***"***Ravi Sanakal, et.al. [2]"presented a diagnostic FCM as well as SVM using SMO and decided which technique helps in diagnosis of Diabetes disease. The best result is obtained in a FCM with an accuracy of 94.3% and positive predictive value which is 88.57%. SVM has an accuracy of 59.5% which is quite low. These results are quite satisfactory, due to the fact that detecting the Diabetes is a very complex problem. Perhaps the most important result of this study was the understanding gained through the implementation and the results obtained here are also very encouraging."Rajesh, et.al.[3***]"*** presented a techniques of applying C4.5 algorithm for classification. The classification rate of 91% was obtained in C4.5algorithm. Future enhancement of this work includes improvisation of the C4.5 algorithms in order to improve the classification rate with greater accuracy. *"*Radha, et.al. [4]" demsonstrated the application of five classification techniques (C4.5, SVM, K-NN, PLR, and BLR) to predict the diabetes disease in patients. They pointed out that necessary to intend an automatic classification tool. In this study, these five techniques were chosen based on the computing time, in which BLR has the lowest computing time with 75% accuracy and error rate of 0.27. The second one with more accuracy rate is SVM while comparing with other techniques. The accuracy of BLR is 75% from the results obtained. The BLR algorithm plays a vital role in data mining techniques. "RajAnand, et.al. [5]" presented a novel approach to Pima Indian diabetes data diagnosis using PCA (principle component analysis) and HONN (Higher Order Neural Network). The HONN can perform diabetes classification with parsimonious representation of node architecture due to its generation of higher order terms. A lower mean square error and faster convergence is attained with PCA preprocessing. "Veena Vijayan, et.al.[6]" suggested for the expectation of maximization of algorithms, among K Nearest Neighbor algorithm, K-means algorithm, Amalgam KNN algorithm and Adaptive Neuro Fuzzy Inference System algorithm. From the observation EM (expectation maximization) possess the least classification accuracy. Amalgam KNN and ANFIS (Adaptive Neuro Fuzzy Inference system) provide the better classification accuracy results. Amalgam KNN comprises both the feature of KNN and K means. ANFIS in cooperates both the features. *"*Priya, et.al. [7]" proposed a method of applying Neural Networks for classification. The result produced by this model is higher than the other models since it performs classification using Neural networks in the Rapid miner tool. This has produced an improvement in the accuracy when compared to the other techniques. "Blanca S. Leon, et.al.[8]" demonstrated the application of Recurrent Neural Networks (RNN) for modeling and control of glucose–insulin dynamics in T1DM (type 1 diabetes mellitus) patients. The proposed RNN, used in these experiments, captures very well the complexity associated with blood glucose level for type 1 diabetes mellitus patients."Paul S. Heckerling, et.al. [9]" presented in their work, the predictor of variables derived from a neural network genetic algorithm accurately discriminated urinary tract infection from non infection in women with urinary complaints. Clinical variables are important in predicting infection differed depending on the uropathogen colony count used to define urinary infection. In addition, some variables predicted urine infection in unexpected ways, and interacted with other variables in making those predictions."Sebastian Polak, et.al. [10]" presented a study in which ANNs as well as other information technology tools are able to identify and analyze relationships in data even when some of the inputs are very complex and difficult to be defined. Therefore, the research carried-out in virtual space is gaining importance in medical applications as well. Superiority of ANNs is pronounced in their ability of automatic identification of complicated relationships.

5 Analysis of Reviewed Works

The accuracy of a learning system needs to be evaluated before it can become useful. Limited availability of data often makes estimating accuracy a difficult task. Choosing a good evaluation methodology is very important for machine learning systems development. There are several popular methods used for such evaluation. The reviews of these soft computing applications lead us to a conclusion that the applications of ANN show very encouraging results comparing with SVM, KNN, and C4.5 similar techniques [4]. Since, ANN is known to be an Universal Predictor, it predicts with relatively accuracy and efficiency.

6 Result and Discussion

In a reviewed work [1], the Pima Indian Diabetes dataset was used for clustering analysis. The dataset has 9 attributes and 768 instances. In this work, data mining was applied to explore the information in medical informatics. Various models have designed for each type of diabetic intervention using classification technique. In table 1, the comparison of different data mining algorithms with prediction accuracy is illustrated.

TABLE 1
PERFORMANCE OF DIFFERENT DATA MINING ALGORITHMS

Algorithm Used	Predicted Accuracy
SVM	74.8%
KNN	78%
C4.5	86%
ANN	89%

From the literature survey, it is observed that the ANN provides more accurate result than other classification techniques such as SVM, KNN and C4.5. The fig 1 shows the graphical presentation of the comparison results between different data mining algorithms.

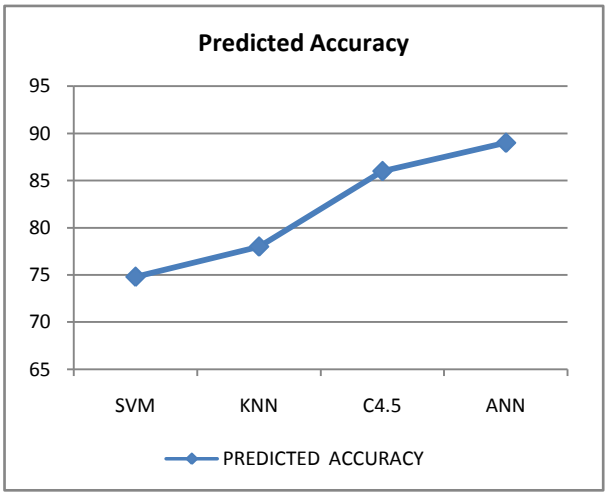

Fig. 1 *PREDICTED ACCURACY OF DIFFERENT DATA MINING ALGORITHMS*

Based on the reviewed literatures, the ANN is identified for applying effective prediction of diabetes disease and classification. This work will be implemented in our future course of experimentation in predicting the types of diabetes.

7 CONCLUSION

Different data mining classification techniques and its applications were reviewed in this paper. The applications of data mining techniques in various medical dataset were carefully studied. Comparison results of different data mining techniques shows that the ANN gives more accurate prediction result than other similar techniques. The data mining techniques compared in this works are namely: ANN, SVM, K-NN, and C4.5. Then one with the highest accuracy above 89% is ANN. As a universal predictor, ANN can play a vital role in various healthcare applications and predictions.

REFERENCES

[1] Asha Gowda Karegowda ,A.S. Manjunath , M.A. Jayaram,"Application Of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis Of Pima Indians Diabetes," International Journal on Soft Computing (IJSC), Vol.2, No.2, May 2011.

[2] Ravi Sanakal, Smt. T Jayakumari, "Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine," International Journal of Computer Trends and Technology (IJCTT) – volume 11 number 2 May 2014.

[3] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis," International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3 September 2012.

[4] P. Radha, Dr. B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques," IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.

[5] Raj Anand, Vishnu Pratap Singh Kirar, Kavita Burse, "K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA," International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[6] Veena Vijayan V. Aswathy Ravikumar, "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus,"International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, June 2014 .

[7] S.Priya R.R.Rajalaxmi," An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network," International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, 2012.

[8] Blanca S.Leona, AlmaY.Alanisb,n, EdgarN.Sancheza, Fernando Ornelas-Tellezc, EduardoRuiz-Velazquezb, "Inverse optimal neural control of blood glucose level for type1diabetes mellitus patients," Journal of the Franklin Institute 349 (2012) 1851–1870.

[9] Paul S. Heckerling, Gay J. Canaris, Stephen , Flach, Thomas G. Tape,Robert S. Wigton, Ben S. Gerber, "Predictors of urinary tract infection based on artificial neural networks and genetic algorithms," international journal of medical informatics 7 6, 2007.

[10] Sebastian Polak Aleksander Mendyk, "Artificial neural networks based Internet hypertension prediction tool development and validation," Applied Soft Computing 8 (2008) 734–739.

[11] Pankaj Srivastava, Neeraj Sharma,Richa Singh, "Soft Computing Diagnostic System for Diabetes," International Journal of Computer Applications (0975 – 888)Volume 47– No.18, June 2012.

[12] V.Karthikeyani, I.Parvin Begum, K.Tajudin, I.Shahina Begam, "Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction," International Journal on Computer Science and Engineering (IJCSE), December 2012.

[13] M. Durairaj, V. Ranjani, "Data Mining Applications In Healthcare Sector: A Study," international journal of scientific & technology research volume 2, issue 10, October 2013.