Int. J. of Computer and Communications
Vol. 1, No. 1, March 2011
Copyright © Mind Reader Publications
www.ijcc.yolasite.com

Predictive Modeling of clinical data using soft computing -Diabetes a Case Study

Ms. Madhavi Pradhan and G.R. Bamnote

HOD(Computer Engg), AISSMS college of Engg , Pune, Maharashtra, India.
e-mail: madhavipradhan@rediffmail.com
HOD(Computer Engg), PRMIT&R,Badnera- Amaravati , Maharashtra, India.
e-mail: grbamnote@rediffmail.com

Abstract

Several studies have reported that up to 50% of all patients with diabetes are undiagnosed. Although these patients are not aware of their disease, they are at risk for the micro and macro vascular complications. Early treatment of diabetes and associated cardiovascular risk factors may reduce the occurrence of these complications. Therefore early detection of diabetes will be of importance in reducing the burden of complications of diabetes. The early detection and prediction can give a warning at a stage where some medications and preventive action will help the patient to increase the span of his healthy life. In this paper we propose predictive model for diabetes using soft computing approach. The clinical data is generally redundant, incomplete, imprecise and inconsistent. The soft computing approach provides the flexible information processing capability for handling such real life data.

Keywords: *Classification, Data Preprocessing, Predictive accuracy, Predictive Model,, Soft computing .*

1 Introduction

Mankind have always aspired to predict the future based on the past and the nature of the unknown based on the qualities of known. It can be said that ‘knowledge is power’ only to the extent that it is useful in making such predictions as accurately as possible [1]. However, the large volume of past data exceeded our human ability for prediction without powerful tools. The predictive modeling exploits patterns found in historical and transactional data to predict the probability of an outcome. A predictive model is made up of a number of predictions, which are variable factors that are likely to influence future behavior or results. Classification and prediction are two forms of data mining functionalities that can be used to extract predictive models. There have been numerous comparisons of the different classification and prediction methods and the matter remains a research topic. No single method has been found to be superior over all others for all data sets [2]. The soft computing approach provides flexible information processing capability for handling real–life ambiguous situations in comparison with conventional hard computing.

2 REVIEW OF LITERATURE

Diabetes is a very complex disease which has affected millions of people of all ages, ethnicities, socio-economic backgrounds and is a growing problem not only in the India, but across the globe. For reasons of genetics and lifestyles, Indians form the world's largest diabetic population. Until the 1970s, it was widely believed that the prevalence of diabetes in India was low compared to the western world. But recent statistics now show that India has the world's largest diabetic population. In urban areas, 12 per cent of the adult population suffers from diabetes, compared to six per cent in the United States and the United Kingdom. Strangely, from India's vast rural population comprising 70 per cent of India's one billion people, a mere two to three per cent suffer from diabetes. Today, India has 35 million diabetic patients and the number is expected to rise to 57 million by 2025. Its incidence has been escalating in recent years and is a growing cause for alarm among health care because of its costs, morbidity and long term impact on the individual and the healthcare industry. Over 24 million people in the US alone have been diagnosed with diabetes and its treatment costs are currently estimated at $ 170 billion which keep growing at an alarming rate. Many causes have been attributed to the onset of diabetes including obesity, sedentary lifestyles, etc [3]. The predictive model for diabetes will help in early and accurate diagnosis thus can give a warning at a stage where some medications and preventive action will help the patient to increase the span of his healthy life.

Data mining (knowledge discovery from data) is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence are alternative names of data mining[2]. Classification and prediction are the data mining functionalities which can be used to extract predictive models describing important data classes or to predict future data trends.

Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty, partial truth and approximation. In effect, the role model for soft computing is the human mind. The guiding principle of soft computing is: “Exploit the tolerance for imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness and low solution cost”. The clinical data may consists of missing , incorrect and sometimes incomplete values set so using soft computing is the better alternative to handle such data.

The principal constituents of soft computing are fuzzy logic, neural computing, evolutionary computation and probabilistic reasoning. The principal constituent methodologies in soft computing are complementary rather than competitive. Fuzzy logic handles imprecision, neural computing deals with learning, evolutionary computation is for optimization and probabilistic reasoning handles uncertainty.

In this part, an attempt is made to compile the work done by various scientists and researchers on predictive models applied to clinical data;

[4] used recurrent artificial neural network to predict blood glucose sugar and then used neuro-fuzzy expert system to offer short-term therapeutic advice.

[5] used Fuzzy logic for representation of clinical guidelines for automated decision support.

In reference [6] presented a clinical decision support system based on On-Line Analytical Processing and data mining. It discovers hidden patterns in the data.

[7] used Bayesian analysis to make meaningful diagnostic and treatment decisions in an evidence-based validated framework.

In [8] the author has proposed a neural network based predictive control structure and applied to treat the problem of diabetes management. The control approach is to predict future plant behavior; hence it specifies accurate control actions necessary to stabilize slow process systems such as physiological systems.

[9] developed temporal abstraction using inference graph for knowledge based temporal abstraction framework. Inference graph are a form of knowledge base which consist of transition rules and inference rules.

[10] suggested a novel method of Thyroid disease diagnosis by using fuzzy cognitive map based decision support system. The system is structured by the combined knowledge from experts and generated fuzzy rules from data.

In [11] a clinical decision support system for metabolism synthesis is developed using a group of C language integrated production system.

A multi agent system for Clinical Decision Support is presented in [12]. It uses supervised learning techniques for supporting clinical decisions.

A modular clinical decision support system is discussed in [13].

[14] presented a data mining method on the study of medical information. They used improved apriori algorithm.

[15] presented a new expert system for diabetes disease diagnosis. They used modified spline smooth support vector machine.

[16] proposes intelligible support vector machines for diagnosis of diabetes mellitus.

Recently [17] proposed decision support system for breast cancer. They have used soft computing approach.

After reviewing the literature cited above it is identified that using the hybrid of soft computing and data mining techniques for preprocessing of clinical data and prediction can result in an effective prediction model.

3 PROPOSED SOLUTION

The objective of the proposed work is to design the predictive model for early and accurate detection of diabetes using hybrid of soft computing and data mining techniques.

Firstly the preprocessing of diabetes data is to be done using different soft computing techniques. The data preprocessing includes data cleaning and data transformation. Data cleaning involves handling noise, inconsistency and missing values in input data. Data transformation involves handling data format issues, discretization, normalization, generalization and feature selection.[1].

The second phase will apply the different classification algorithms to the preprocessed data. Classification algorithms aims to identify the characteristics that indicates the group to which each case belongs, This pattern can be used both to understand the existing data and to predict the behavior of the new instances. For any given problem, the nature of the data itself will affect the choice of models and algorithm chosen. Hence a variety of tools and techniques will be applied to find the best possible model.

Some of the techniques used to mine data are neural network, Decision tree, Multivariate adaptive regression splines (MARS), Rule induction, K-Nearest Neighbor , Memory based reasoning(MBR) , logistic regression, Discriminant, analysis, Generalized Additive Modela(GAM), Boosting and Genetic algorithms.

The leading tools for pure data mining are IBM Inteligent Miner, SGI MineSet, Oracle Darwin, SAS Enterprise Miner , SPSS Clementine and Weka.

The best model is often found after building models by using the different techniques and tools. Lastly the performance of selected best model is optimized.

4 PERFORMANCE ANALYSIS

The performance of the proposed predictive model is measured using the following criteria .

Accuracy: The accuracy of a predictive model is the probability of it correctly classifying records in the test dataset.

Accuracy = (TP+TN) / (TP+TN+FP+FN)

Where TP is the number of true positive
TN is the number of true negative
FP is the number of false positive
FN is the number of false negative

Area under receiver-operator characteristics curve(ROC): It characterize the relationship between sensitivity and (1-specificity). The sensitivity of a test is the probability(0-100%) that a test is positive for the patients with diabetes. The specificity is the probability that a test is negative for patients without diabetes. The curve that has a larger area under curve is better than the one that has a smaller area under curve.

5 Conclusion

The early detection and prediction of the diabetes can give a warning at a stage where some medications and preventive action will help the patient to increase the span of his healthy life. In long term, the on-set of diabetes may be delayed to such an extent that it is no longer a life threatening disease, but gets converted to a curable and controlled type of disease. The hybrid of the soft computing and data mining techniques for preprocessing of diabetes data and prediction can result in a effective prediction model.

References

[1] Vikram Pudi and P. Radha Krishna, "*Data Mining*", Oxford University Press 2009, ISBN 0-19-568628-4

[2] Jiawei Han and Micheline Kamber "*Data Mining: Concepts and Techniques*", 2nd ed. The Morgan Kaufmann series in Data Management Systems, March2006 ISBN 1-55860-901-6

[3]http://www.medicalnewstoday.com/articles/44967.php

[4] W.A.Sandham, D.J.Hamilton, A.Japp, K.Patterson., "Neural Network and Neuro-Fuzzy Systems For Improving Diabetes Therapy", *in the Proceedings ofthe 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* 1998, vol.3, pp.1438-1441, 29 Oct-1 Nov 1998.
doi: 10.1109/IEMBS.1998.747154

[5] J.Warren., G. Beliakov and Van der Zwaag. B., "Fuzzy Logic in Clinical Practice Decision Support Systems,", *in Proceedings of the 33rd Annual Hawai International Conference on System Science-2000.*, pp. 10, 4-7 Jan. 2000
doi: 10.1109/HICSS.2000.926789

[6] Chua Sook Ling and Sellappan Palaniappan,"Clinical Decision Support System Using OLAP Mining", *in the International Journal of computer Science and Network Security,* Vol.8 No.9, pp. 290-296, September 2008

[7] Howard Robin., John. S. Eberhardt III., W. D. Muller., R. Clark.,J. Kam.,"Classification of Pathology Data Using a Probabilistic (Bayesian) Model", *in proceedings of 18th International Conference on Systems Engineering,* 2005.,pp. 286- 291, 16-18 Aug.2005.
doi: 10.1109/ICSENG.2005.22

[8] M. F. Alamaireh, "A Predictive Neural Network Control Approach in Diabetes Management by Insulin Administration", *Information and Communication Technologies,*
2006. ICTTA'06, 2nd, vol.1, pp.1618-1623.
doi: 10.1109/ ICTTA.2006. 1684626

[9] Pham Van Chung and Duong Tuan Anh,"Applying Temporal Abstraction in Clinical Databases", *International Conference on Research,Innovation and Vision for the Future,* 2007 IEEE, pp.192-199, 5-9 March 2007.
doi: 10.1109/RIVF.2007.36915

[10] E.I. Papageorgiou, N.I.Papandrianos, D. Apostolopoulos and P Vassilakos, "Complementary use of Fuzzy Decision Trees Decision Making in Medical Informatics,", *International Conference on BioMedical Engineering and*

Informatics, 2008. BMEI 2008., vol.1, pp.888-892, 27-30 May 2008 ;
doi: 10.1109/BMEI.2008.275

[11] Qunyi Zhou, "A Clinical Decision Support System for Metabolism Synthesis."*International Conference on Computational Intelligence and Natural Computing* , 2009. CINC '09 . , vol.2, pp.323-325, 6-7 June 2009
doi: 10.1109/CINC.2009.89

[12] G. Czibula, I.G. Czibula, G.S. Cojocar and A.M. Guran, "IMASC - An Intelligent MultiAgent System for Clinical Decision Support", *First International Conference on Complexity and Intelligence of the Artificial and Natural Complex Systems , Medical Applications of the Complex Systems, Biomedical Computing, 2008*. CANS '08., pp.185-190, 8-10 Nov. 2008
doi: 10.1109/CANS.2008.28

[13**]** Fran Wu, Mitch Williams, Peter Kazanzides, Ken Brady and Jim Fackler,"A modular Clinical Decision Support System Clinical prototype extensible into multiple clinical settings," *3rd International Conference on Pervasive Computing Technologies for Healthcare*,2009. PervasiveHealth 2009., pp.1-4, 1-3 April 2009
doi:10.4108/ICST.PERVASIVEHEALTH2009.6078

[14] Xiaofeng Zhao, Liyan Jia, Linping An and Liyan Wang ,"A Data Mining Method on the Study of Medical Information",*2010 international Conference on Computer Application and SystemModeling (ICCASM)*, vol.9, pp. 183-186, 2224 Oct. 2010.
doi: 10.1109/ICCASM.2010.5623058

[15**]** Shanti Wulan Purnami, Jasni Mohamad and A. Emgbong, “A New Expert System for Diabetes Disease Diagnosis using Modified Spline Smooth Support Vector Machine”, *International Workshop on biomathematics ,bioinformatics and biostatistic,* 23-26 March 2010 Fukuoka, Japan

[16] N. Barakat.,A. P. Bradley Barakat and Mohamad. Nabil H., "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," *IEEE Transaction on Information Technology in Biomedicine* , vol.14, no.4, pp.1114-1120,July2010.
doi: 10.1109/TITB.2009.2039485

[17] R.R. Jangler, Anupama Shukla and Ritu Tiwari ," Intelligent Decision Support System for Breast Cancer”, *The International Conference on Swar Intelligence ICSI 2010,* Part II, LNCS 6146, pp. 351-358, 2010