IJARIIT

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue3)

Available online at: www.Ijariit.com

Review of Diabetes Detection by Machine Learning and Data Mining

Preeti Verma[1], Inderpreet Kaur[2], Jaspreet Kaur[3]

Student[1], Assistant Professor[2], Assistant Professor[3]

Vermapreeti008@gmail.com[1], inderpreet.kaur029@gmail.com[2], jaspreetkaur.rb@gmail.com[3]

Dept. of C.S.E., Rayat Bahra Group of Institutes, Patiala, India

Abstract- The most common action in data mining is classification. It recognizes patterns that describe the group to which an item belongs. It does this by examining existing items that already have been classified and inferring a set of rules. Similar to classification is clustering. The major difference being that no groups have been predefined. Prediction is the construction and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of a given object is likely to have. The next application is forecasting

Keywords- data mining, diabetes, classifier, precision, machine learning.

I. INTRODUCTION

Data Mining refers to the process of extracting knowledge or other interesting patterns from large collection of data [1]. It involves iterative sequential steps such as Data Cleaning (removal of noise and inconsistent data), Data Integration (combining data from multiple heterogeneous sources), Data selection (selection of relevant data for analysis), Data Transformation (data is transformed into forms suitable for mining), Pattern Evaluation (identification of interesting patterns) and Knowledge Presentation (use of visualization and knowledge presentation techniques for presenting the mined knowledge to users). Data mining tasks can be classified into two categories: Descriptive and predictive data mining. Descriptive data mining provides information to understand what is happening inside the data without a predetermined idea. Predictive data mining allows the user to submit records with unknown field values, and the system will guess the unknown values based on previous patterns discovered form the database. Data mining models can be categorized according to the tasks they perform: Classification and Prediction, Clustering, Association Rules. Classification and prediction are predictive models but clustering and association rules are descriptive models.

The most common action in data mining is classification. It recognizes patterns that describe the group to which an item belongs. It does this by examining existing items that already have been classified and inferring a set of rules [2]. Similar to classification is clustering. The major difference being that no groups have been predefined. Prediction is the construction and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of a given object is likely to have. The next application is forecasting. This is different from predictions because it estimates the future value of continuous variables based on patterns within the data. Neural networks, depending on the architecture, provide associations, classifications, clusters, prediction and forecasting to the data mining industry. Neural networks can mine valuable information from a mass of history information and be efficiently used in financial areas, so the applications of neural networks to financial forecasting have been very popular over the last few years.

Soft computing is a field within computer science which is characterized by the use of inexact solutions to computationally-hard tasks such as the solution of NP complete problems, for which an exact solution cannot be derive polynomial time [3]. The advent of Soft Computing techniques dates back to early 1990s. Earlier computational approaches could model and precisely analyse only relatively simple systems. More complex systems arising in biology, medicine, the humanities, management sciences, computer science, engineering and similar fields often remained intractable to conventional mathematical and analytical methods. Major components of soft computing include Neural Networks, Fuzzy Systems, Evolutionary Computation (Genetic Algorithms, Differential Evolution), Meta-heuristic and Swarm Intelligent algorithms (Ant Colony Algorithm, Particle Swarm Optimization, Cuckoo Search, Harmony Search, Bees Algorithm)

Tools and technologies:

In this review use data mining and machine learning tools with mat lab environment. Different machine learning algorithms like tree, SVM, neural network use in classified the diabetes.

Diabetes Mellitus

Diabetes mellitus is the most common endocrine disease. The disease is characterized by metabolic abnormalities and by long-term complications involving the eyes, kidneys, nerves, and blood vessels. The diagnosis of symptomatic diabetes is not difficult. When a patient presents with signs and symptoms attributable to an osmotic dieresis and is found to have hyperglycemia essentially all physicians agree that diabetes is present. The two major types of diabetes are Type I diabetes and Type II diabetes. Type I diabetes is usually diagnosed in children and young adults, and was previously known as juvenile diabetes [4]. Type I diabetes mellitus (IDDM) patients do not produce insulin due to the destruction of the beta cells of the pancreas. Therefore, therapy consists of the exogenous administration of insulin. Type II diabetes is the most common form of diabetes. Type II diabetes mellitus (NIDDM) patients do produce insulin endogenously but the effect and secretion of this insulin are reduced compared to healthy subjects [5]. Currently cure does not exist for the diabetes, and then only option is to take care of the health of people affected, maintain their glucose levels in the blood to the nearest possible normal values.

II. Literature Review

Rian Budi Lukmanto etal. This paper has proposed the application of computational intelligence by using fuzzy logic that perform detection of diabetes mellitus (DM). This proposed method is based on the knowledge acquisition process. An accuracy of 87.46% is obtained by the method.

Cheng-Hsiung Weng et al.. In this paper Different types of neural network classifiers are used for disease prediction. First we compare the performance of single neural network with classifier with multiple neural network with authentic data set. Secondly use statistical testing to investigate the difference in performance among these classifiers. Multiple neural network should be better than single neural network.

Kamadi V.S.R.P. Varmaet al. This paper developing a decision tree model to predict the occurrence of diabetes disease. Much better decision rules can be identified from the data set with the use of the fuzzy decision boundaries. The modified Gini index-Gaussian fuzzy decision tree algorithm is proposed. This algorithm outperforms other decision tree algorithms.

Bum Ju Lee et al. This study aims to predict the fasting plasma glucose (FPG) status that is used in the diagnosis of diabetes. In this paper compared the 2 machine learning algorithms. That is logistic regression and naïve Bayes. Naive Bayes shows better result as compare to logistic regression.

Longfei HanPhyo Phyo et al. In this paper, present an ensemble learning approach for rule extraction from SVM, which uses Random Forest (RF) technique to develop an affordable and feasible rules for diagnosis of diabetes. The proposed method generates average precision 94.2% and average recall 93.9% for all classes.

Kandhasamy et al. The main aim of this study is to compare the performance of algorithms those are used to predict diabetes using data mining technique. In this paper we compare machine learning classifiers J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machines) to classify patients with diabetes mellitus.

Carpenter and Markuzon et.al. [12] implemented an instance counting algorithm ARTMAP-IC (Adaptive resonance theory match tracking algorithm) and obtained an accuracy of 81% to test set .Conventional (one training and one test) validation method has been used by them. ARTMAP-IC modifies the ARTMAP search algorithm to allow the network to encode inconsistent cases, and it combines both instance counting during training with distributed category representation during testing to obtain probabilistic predictions, even with fast learning and only one training epoch. The ARTMAP algorithm, control predictive errors. A new version facilitates prediction with sparse or inconsistent data. When original match tracking algorithm is compared with the new algorithm, it provides better approximates the real-time network differential equations and also reduces memory without any loss of performance. Predictive accuracy is estimated by simulations on four medical databases: Pima Indian diabetes, breast cancer, heart disease, and gall bladder removal. ARTMAP-IC results are similar to or better than those of logistic regression, K nearest neighbor (KNN), the ADAP perceptron, multisurface pattern separation, CLASSIT, instance-based (IBL), and C4.

Polat and Gunes et.al. [13] Discussing using principal component analysis (PCA) and adaptive neuro-fuzzy inference system (ANFIS). The aim of this study is to bring an improvement in the diagnostic accuracy of diabetes disease combining PCA and ANFIS. The proposed system has two stages. In the first stage, dimension of diabetes disease dataset that has 8 features, is minimized to 4 features with the usage of principal component analysis. In the second stage, diagnosis of diabetes disease is carried out via adaptive neuro-fuzzy inference system classifier. The dataset used in our study is taken from the UCI (from Department of Information and Computer Science, University of California) Machine Learning Database. The obtained classification accuracy of system was 89.47% and it was very promising as compared to the other classification applications.

Yue Huang [14] discussing feature selection technique, feature selection via supervised model construction (FSSMC), was used to identify the important attributes affecting diabetic control. After selection of suitable features, three complementary classification techniques (Naïve Bayes, IB1 and C4.5) were applied to the data in order to predict how well the condition of patient was controlled. FSSMC identified patients' 'age', 'diagnosis duration', the need for 'insulin treatment', 'random blood glucose' measurement and 'diet treatment' as the most important factors which influence blood glucose control. With the usage of this technique, the best predictive accuracy of 95% and sensitivity of 98% was achieved. The factors, like 'type of care' delivered, the use of 'home monitoring', and the importance of 'smoking 'are not so much important in diabetes control. The more important factors that are identified include: 'age of patients', 'diagnosis duration' and 'family history', which are beyond the control of physicians.

Kemal Polat et.al. [15] Discussing Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) for diagnosis of diabetes disease. Also, proposed a new cascade learning system based on Generalized Discriminant Analysis and Least Square Support Vector Machine. The proposed system includes two stages. The first stage, they used Generalized Discriminant Analysis to discriminant feature variables between healthy and patient (diabetes) data as pre-processing step. The second stage, they used LS-SVM for classification of diabetes dataset. While LS-SVM obtained 78.21% classification accuracy using 10-fold cross validation, the proposed system called GDA-LS-SVM obtained 82.05% classification accuracy using 10-fold cross validation and it is very promising compared to the previously reported classification techniques.

Hasan Temurtas [16] a multilayer neural network structure, trained by Levenberg–Marquardt (LM) algorithm and a probabilistic neural network structure were used. The results of the study were compared with the results of the previous studies that also focused on diabetes disease diagnosis and by using the same UCI machine learning database obtains 79.62% accuracy. The classification accuracy of MLNN with LM obtained by this study using correct training was comparatively better than those obtained by other studies except the classification accuracies by Polat and Gunes.

Santi Wulan et.al. [17] Implemented MKS-SSVM technique to improve accuracy of the result has been developed by many researchers. It is called Multiple Knot Spline SSVM (MKS-SSVM).Implement an experiment on Pima Indian diabetes dataset to evaluate the effectiveness of our method. The accuracy of previous results of this data is still below 80% using SSVM that is smooth support vector machine. Then, the proposed MKS-SSVM showed better performance in classifying diabetes disease diagnosis with accuracy of 93.2% which is better than previous reported results.

Hasan Temurtas [16] a multilayer neural network structure, trained by Levenberg–Marquardt (LM) algorithm and a probabilistic neural network structure were used. The results of the study were compared with the results of the previous studies that also focused on diabetes disease diagnosis and by using the same UCI machine learning database obtains 79.62% accuracy. The classification accuracy of MLNN with LM obtained by this study using correct training was comparatively better than those obtained by other studies except the classification accuracies by Polat and Gunes.

Santi Wulan [17] implemented MKS-SSVM technique to improve accuracy of the result has been developed by many researchers. It is called Multiple Knot Spline SSVM (MKS-SSVM).Implement an experiment on Pima Indian diabetes dataset to evaluate the effectiveness of our method. The accuracy of previous results of this data is still below 80% using SSVM that is smooth support vector machine. Then, the proposed MKS-SSVM showed better performance in classifying diabetes disease diagnosis with accuracy of 93.2% which is better than previous reported results.

Santi Wulan Purnami et.al. [18] Presents a novel method for diabetes disease diagnosis using modified spline smooth support vector machine (MS-SSVM) to obtain optimal accuracy results, firstly uniform Design method was used for selection of most relevant features. The performance of this method is evaluated using 10-fold cross validation accuracy, confusion matrix. The obtained classification accuracy using 10-fold cross validation is 96.58% in comparison with other spline SSVM technique. The results of this study showed that the modified spline SSVM was effective to detect diabetes disease diagnosis and this is very promising result compared to the previously reported results.

Muhammad Waqar Aslam et.al [19] implemented genetic programming (GP) and a variation of genetic programming called GP with comparative partner selection (CPS) for detection of diabetes. The proposed system includes two stages. In first stage, genetic programming is used to produce an individual from training data that converts the available features to a single feature such that it has different values for healthy and patient (diabetes) data. In the second stage, test data is used for testing of that individual features. The proposed system was able to achieve78.5±2.2% accuracy. The results showed that GP based classifier perform better in the diagnosis of diabetes disease.

Nahla H. Barakat et.al [20] support vector machines (SVMs) are proposed for the diagnosis of diabetes. In this, they used an explanation module, which is called as "black box" model of an SVM which we used for diagnostic (classification) decision. Results obtained on diabetes dataset by using "black box" shows that it is a promising tool that is provided by intelligible SVM's for the prediction of diabetes, with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94%. Future work is to conduct a prospective study to further refine the predictive results obtained by the proposed rules.

Problem statement:

Improve the accuracy of classification by Mache learning algorithms and analysis precision, recall, Reduce the feature set by feature selection and Extraction.

CONCLUSION AND FUTURE SCOPE

The performance of this method is evaluated using 10-fold cross validation accuracy, confusion matrix. The obtained classification accuracy using 10-fold cross validation is 96.58% in comparison with other spline SSVM technique. The results of this study showed that the modified spline SSVM was effective to detect diabetes disease diagnosis and this is very promising result compared to the previously reported results.

REFERENCES

[1] Rian Budi Lukmanto, Irwansyah E et all "The Early Detection of Diabetes Mellitus (DM)Kamadi V.S.R.P. Varma Using Fuzzy Hierarchical Model." ELSEVIER, Volume 59, 2015.

[2]Cheng-Hsiung Weng, T et all "Disease prediction with different types of neural network classifiers." ELSEVIER, Volume 33, 2014.

[3]Kamadi V.S.R.P. Varma, A et all "A computational intelligence approach for a better diagnosis of diabetic patients." ELSEVIER, Volume 40, 2014.

[4] Bum Ju Lee, Boncho N et all "Prediction of Fasting Plasma Glucose Status Using Anthropometric Measures for Diagnosing of Diabetes." IEEE, Volume 18, NO.2, 2014.

[5] Longfei Han, Beijing T et all "Rule Extraction from Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes." IEEE, Volume 19, 2014

[6]J.Pardeep Kandhasamy, S Balamurali "Perforamance analysis of Classifier Models to Predict Diabetes Mellitus." ELSEVIER, Volume 47, 2014.

[7] Ajith Abraham, Radha Thangaraj, Millie Pant, Pascal Bouvry, "Particle swarm optimization: Hybridization perspectives and experimental illustrations", Applied Mathematics and Computation, 2011

[8] Corne D, Dorigo M, Glover F, "New ideas in optimization", McGraw-Hill, USA, 1999

[9] D.Mishra, B.Sahu, "Feature selection for cancer classification: a signal-to-noise ratio approach", International Journal of Scientific and Engineering Research, vol.2, 2011.

[10] Expert Committee on the Diagnosis and Classification of Diabetes Mellitus, "Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus". Diabetes Care, vol 20, pp. 1183–1197, 1997

[11] http://en.wikipedia.org/wiki/Data_mining

[12] Carpenter, G. A., & Markuzon, "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases", Neural Networks, vol. 11, pp. 323–336, 1998.

[13] Kemal Polat and S. Gunes, "An expert system approach based on principal component analysis and adaptive NeuroFuzzy inference system to diagnosis of diabetes disease", Digital Signal Processing, vol. 17, pp. 702-710, Jul 2007

[14] Yue Huang, Paul McCullagh, Norman Black, Roy Harper, "Feature selection and classification model construction on type 2 diabetic patients", Volume 41 Issue 3, pp 251-262, Nov. 2007

[15] K. Polat, S. Gunes and A. Aslan, "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine", Expert Systems with Applications, vol. 34(1), pp. 214–221, 2008

[16] T. Hasan, Y. Nejat, T. Feyzullah, "A comparative study on diabetes disease diagnosis using neural networks", Expert Systems with Applications, vol. 36, pp. 8610- 8615, May 2009

[17] Santi Wulan Purnami, Abdullah Embong, Jasni Mohd Zain and S.P. Rahayu, "A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis", Journal of Computer Science vol.5 (12), pp.1006-1011, ISSN 1549-3636,2009

[18] Santi Wulan Purnami, Jasni Mohamad Zain and Abdullah Embong, "Data mining techniques for medical diagnosis using a new smooth SVM", Communications in Computer and Information Science, Vol 88, Part 1, pp.15-27,2010

[19] M.W. Aslam, A.K. Nandi, "Detection of diabetes using genetic programming", 18th European Signal Processing Conference, 2010
[20] Nahla H. Barakat, Andrew P. Bradley, Mohamed Nabil H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", IEEE transaction on Information technology in bioinformatics, Vol. 14, NO. 4, Jul 2010
[21]Jiawie Michelene Kamber Han, Data Mining: Concepts and Techniques, Morgan Kauffman Publishers, Second Edition
[22] http://en.wikipedia.org/wiki/Soft_computing
[23] Thair Nu Phyu, "Survey of Classification Techniques in Data", Proceedings of the International Multi Conference of Engineers and Computer Scientists, Vol I, Hong Kong, March 2009.
[24] Pinar Civicioglu, Erkan Besdok, "A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms", Springer Science and Business Media B.V. 2011

Authors

First Author Preeti verma M.Tech Student, Dept. of C.S.E., Rayat Bahra Group of Institutes, Patiala, India
Vermapreeti008@gmail.com

Second Author Inderpreet Kaur (B.TECH (CSE) SLIET, LANGOWAL) (M.TECH –CSE PUNJABI UNIVERSITY) Assistant Professor, Dept. of C.S.E., Rayat Bahra Group of Institutes, Patiala, India
inderpreet.kaur029@gmail.com

Third Author – Jaspreet Kaur (B.TECH –CSE, SVIT) (M.TECH-CSE- PUNJABI UNIVERSITY) Assistant Professor, Dept. of C.S.E., Rayat Bahra Group of Institutes, Patiala, India
jaspreetkaur.rb@gmail.com

.