

Review of Prediction of Diabetes Diagnosis patient Using Classification Based Data Mining Techniques

Chandan Azad , Dr. Dinesh Kumar Sahu , Dr. Varsha Namdeo
1, 2, 3SRK University, Bhopal M.P., India

ABSTRACT

Data mining is a process of extracting information from a dataset and transform it into understandable structure for further use, also it discovers patterns in large data sets . Data mining has number of important techniques such as preprocessing, classification. Classification is one such technique which is based on supervised learning.. diabetic is a life threatening disease prevalent in several developed as well as developing countries like India. the data classification is diabetic patients data set is developed by collecting data from hospital repository consists of 1865 instances with different attributes. The instances in the dataset are two categories of blood tests, urine tests. In this paper we discuss various algorithm approaches of data mining that have been utilized for diabetic disease prediction. Data mining is a well known technique used by health organizations for classification of diseases such as diabetes and cancer in bioinformatics research. In the proposed approach we have used WEKA with 10 cross validation to evaluate data and compare results. Weka has an extensive collection of different machine learning and data mining algorithms.

Keywords

Weka, Data mining, Classification, Diabetic Disease Prediction.

1. INTRODUCTION

The main focus of this paper is the classification of different types of datasets that can be performed to determine if a person is diabetic. The solution for this problem will also include the cost of the different types of datasets. For this reason, the goal of this paper is classifier in order to correctly classify the datasets, so that a doctor can safely and cost effectively select the best datasets for the diagnosis of the disease. The major motivation for this work is that diabetes affects a large number of the world population and it's a hard disease to diagnose. A diagnosis is a continuous process in which a doctor gathers information from a patient and other sources, like family and friends, and from physical datasets of the patient. The process of making a diagnosis begins with the identification of the patient's symptoms. The symptoms will be the basis of the hypothesis from which the doctor will start analyzing the patient. This is our main concern, to optimize the task of correctly selecting the set of medical tests that a patient must perform to have the best, the less expensive and time consuming diagnosis possible. A solution like this one, will not only assist doctors in making decisions, and make all this process more agile, it will also reduce health care costs and waiting times for the patients.

The major contributions of this paper are:

- (1) To extract useful classified accuracy for prediction of Diabetes diseases.
- (2) Comparison of different data mining algorithms on Diabetes dataset.
- (3) Identify the best performance algorithm for prediction of diseases.

This paper will focus on the analysis of data from a data set called diabetes data set

2. RELATED WORK

The few medical data mining applications as compared to other domains. Reported their experience in trying to automatically acquire medical knowledge from clinical databases. They did some experiments on three medical databases and the rules induced are used to compare against a set of predefined clinical rules. Past research in dealing with this problem can be described with the following approaches:

- (a) Discover all rules first and then allow the user to query and retrieve those he/she is interested in. The representative approach is that of templates . This approach lets the user to specify what rules he/she is

interested as templates. The system then uses the templates to retrieve the rules that match the templates from the set of discovered rules.

(b) Use constraints to constrain the mining process to generate only relevant rules. Proposes an algorithm that can take item constraints specified by the user in the association rule mining processor that only those rules that satisfy the user specified item constraints are generated. The study helps in predicting the state of diabetes i.e., whether it is in an initial stage or in an advanced stage based on the characteristic results and also helps in estimating the maximum number of women suffering from diabetes with specific characteristics. Thus patients can be given effective treatment by effectively diagnosing the characteristics. Our research work based on the concept from Data Mining is the knowledge of finding out of data and producing it in a form that is easily understandable and comprehensible to humans in general. These further extended in this to make an easier use of the data's available with us in the field of Medicine.

The main use of this technique is to have a robust working model of this technology. The process of designing a model helps to identify the different blood groups with available Hospital Classification techniques for analysis of Blood group data sets. The ability to identify regular diabetic patients will enable to plan systematically for organizing in an effective manner. Development of data mining technologies to predict treatment errors in populations of patients represents a major advance in patient safety research.

3. OVERVIEW OF PROPOSED APPROACH

3.1 WEKA

In order to carry out experimentations and implementations Weka was used as the data mining tool. Weka (Waikato Environment for Knowledge Analysis) is a data mining tool written in java developed at Waikato. WEKA is a very good data mining tool for the users to classify the accuracy on the basis of datasets by applying different algorithmic approaches and compared in the field of bioinformatics. Explorer, Experimenter and Knowledge flow are the interface available in WEKA that has been used by us. In this paper we have used these data mining techniques to predict the survivability of Diabetes disease through classification of different algorithms accuracy .

3.2 Classification

In data mining tools classification deals with identifying the problem by observing characteristics of diseases amongst patients and diagnose or predict which algorithm shows best performance on the basis of WEKA's statistical output .

Three techniques have been adopted in this paper, the first technique uses explorer interface and depends on algorithms like Naïve Bayes, SMO, J48, REP Tree and RANDOM Tree, used in areas to represent, utilize and learn the statistical knowledge and significant results have been achieved.

The second technique uses Experimenter interface. This study allows one to design experiments for running algorithms such as Naïve Bayes, J48, REP Tree and RANDOM Tree on datasets. These algorithms can be run on experimenter and analyze the results. It configures the test option to use cross validation 10 folds. This interface provides provision for running all the algorithms together and thus a comparative result was obtained.

The third technique uses Knowledge Flow. In this study we classified the accuracy of different algorithms Naïve Bayes, SMO, J48, REP Tree and random Tree on different data sets and compared the results to know which algorithm shows best performance. In order to predict Diabetes Disease for survivability by user one can select this weka component from toolbar, place them in a layout like manner and connect its different components together in order to form a knowledge flow web for preprocessing and analyzing data.

All the algorithms used by us were applied to a Diabetes data set explained in detail in section 4. In order to obtain better accuracy 10 fold cross validation was performed. For each classification we selected training and testing sample randomly from the base set to train the model and then test it in order to estimate the classification and accuracy measure for each classifier. The thrust classifications and accuracy used by us are:

3.2.1. Correctly Classified Accuracy

It shows the accuracy percentage of test that is correctly classified.

3.2.2. Incorrectly Classified Accuracy

It shows the accuracy percentage of test that is incorrectly classified.

3.3.3. Mean Absolute Error

It shows the number of errors to analyze algorithm classification accuracy.

3.3.4. Time

It shows how much time is required to build model in order to predict disease.

3.3.5. ROC Area

Receiver Operating Characteristic¹⁹ represent test performance guide for classifications accuracy of diagnostic test

4. CONCLUSION / FUTURE WORK

The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. The aim of data mining is to extract knowledge from information stored in database and generate clear and understandable description of patterns. The main aim of this paper is to predict diabetes disease using WEKA data mining tool. It has four interfaces. Out of these four we have used three interfaces: Explorer, Experimenter and knowledge flow. Each interface has its own classifier algorithms. We have used five algorithms i.e. Naïve Bayes, J48, SMO, REP Tree and Random tree for our experimentation. Then these algorithms were implemented using WEKA data mining technique to analyze algorithm accuracy which was obtained after running these algorithms in the output window.

5. REFERENCES

- [1] S , Liver Disease Prediction Using Bayesian Classification , Special Issues , 4th National Conference on Advance Computing , Application Technologies, May 2014
- [2] Solanki A.V., Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 5857-5860,2014.
- [3] Joshi J, Rinal D, Patel J, Diagnosis And Prognosis of Breast Cancer Using Classification Rules, International Journal of Engineering Research and General Science,2(6):315-323, October 2014.
- [4] David S. K., Saeb A. T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38,2013.
- [5] Vijayarani, S., Sudha, S., Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, International Journal of Innovative Research in Computer and Communication Engineering, 1(3): 735-741, 2013.
- [6] Kumar M. N., Alternating Decision trees for early diagnosis of dengue fever .arXiv preprint arXiv:1305.7331,2013.
- [7] Durairaj M, Ranjani V, Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.
- [8] Sugandhi C , Ysodha P , Kannan M , Analysis of a Population of Cataract Patient Database in WEKA Tool , International Journal of Scientific and Engineering Research ,2(10) ,October ,2011.
- [9] Yasodha P, Kannan M, Analysis of Population of Diabetic Patient Database in WEKA Tool, International

Journal of Science and Engineering Research, 2 (5), May 2011.

- [10] Bin Othman M. F , Yau, T. M. S., Comparison of different classification techniques using WEKA for breast cancer, In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, Springer Berlin Heidelberg, 520-523, January 2007.
- [11] Wikipedia, http://en.m.wikipedia.org/wiki/Dengue_fever, accessed in January 2015.
- [12] Wikipedia, [http://en.m.wikipedia.org/wiki/weka_\(machine_learning\)](http://en.m.wikipedia.org/wiki/weka_(machine_learning)), accessed in January 2015.
- [13] Waikato, <http://www.cs.waikato.ac.nz/ml/weka>, accessed in January 2015.
- [14] Wikipedia, en.m.wikipedia.org/wiki/Data_set, accessed in January 2015.
- [15] KirkbyR, Frank E, WEKA Explorer User Guide for version 3-4-3, November 2004.
- [16] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000.
- [17] Varun Kumar and Nisha Rathee, "Knowledge discovery from database Using an integration of clustering and classification", (IJACSA) International Journal of Advanced Computer Science and Applications, 2011.
- [18] Swasti Singhal, Monika Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering(IJITEE), 2013
- [19] Arodz, M. Kurdziel, E. O. D. Sevre, and D. A. Yuen, "Pattern recognition techniques for automatic detection of suspicious-looking anomalies in mammograms," *Comput. Methods Programs Biomed.*, vol. 79, pp. 135–149, 2005.
- [20] L. Ramirez, N. G. Durdle, V. J. Raso, and D. L. Hill, "A support vector machines classifier to assess the severity of idiopathic scoliosis from surface topology," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 84–91, Jan. 2006.
- [21] A. Swets, R. M. Dawes, and J. Monahan. "Better decisions through science", *Scientific American*, 283:82–87, October 2000.
- [22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 1988.
- [23] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
- [24] I. H. Witten and E. Frank. *Data Mining - Practical Machine Learning Tools and Techniques with JAVA Implementations*. Morgan Kaufmann Publishers, 2000.
- [25] Chen, Y.-W., & Lin, C.-J. (2005). Combining SVMs with various feature selection strategies. Available from <http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf>.
- [26] Cheng-Lung Huang, Hung-Chang Liao b, Mu-Chen Chen c, "Prediction model building and feature selection with support vector machines in breast cancer diagnosis ", *Expert Systems with Applications*, 2008, 578-587 doi:10.1016/j.eswa.2006.09.041