RESEARCH ARTICLE OPEN ACCESS

Comparison of Classification Algorithms in Diabetic Dataset

J.Anitha [1], Dr.A.Pethalakshmi [2]
M.Phil Scholar [1], Associate professor and Head [2]
Department of Computer Science,
M.V.Muthiah Government Arts College for Women, Dindigul.
Tamil Nadu - India

ABSTRACT
Data mining Techniques has proved for early prediction of disease with higher accuracy in order to save human life. Diabetes is one of the most common and rapidly increasing diseases in the world. Diabetes has affected over 246 million people worldwide with the majority of them being women. World Health Organization report (WHO), this number is expected to rise over 380 million by 2025. In this paper two classification algorithms, namely Naive Bayes and J48 are studied and applied on the diabetic dataset. The so- called algorithms are tested using WEKA tool for comparing its accuracy rate, time and error rate.

Keywords:- Data mining, Diabetes,Dataset, Naive Bayes, J48.

I. INTRODUCTION

Data mining:Data mining is the process of discovering interesting patterns and knowledge from large amount of data [3].It is a self-knowledge discovery and a process for the analysis of large dataset providing unknown, hidden, meaningful patterns automatically obtained from large-scale databases [9].A physician has to analyze lot of factors before diagnosing the diabetes which makes physician's job difficult.

Recently, there are many methods and algorithms used mine bio-medical dataset for hidden information including Neural networks (NNs), Decision Tree(DT), Fuzzy logic systems, Naive Bayes, SVM and so on. These algorithms decrease the time spent for processing symptoms and producing diagnoses, making them more precise at the same time.

Diabetes:Diabetes is a major health problem in most of the countries. Among all countries, India is in 3rd place according to this .It is a condition in which your body is unable to produce the required amount of insulin needed to regulate the amount of sugar in the body. Insulin is the principle hormone that regulates uptake of glucose from the blood into most cells (muscle, fat cells).If the amount of insulin available is in-sufficient, and then glucose will not have its usual effect so that glucose will not be absorbed by the body cells that require it. WHO reports state that almost one-third of the women who suffer from diabetes have no knowledge about it [1].

The common symptom, for the diabetic patients are frequent urination, increased thirst, weight loss, slow-healing in wound, giddiness, increased hunger etc.

Types of Diabetes

- ***Type I:***It is called Insulin--dependent diabetes, it usually appears before age of 30, due to lack (or) deficiency of insulin. Majority of these diabetes causes were in children. Persons with type I diabetes, the beta cells of the pancreas, (which are responsible for insulin production), are destroyed due to autoimmune system.

- ***Type II:***It is called non-Insulin dependent diabetes. It is usually occurs over 40 years of age. The causes of type II diabetes are overweight, obesity, lack of physical activity, poor diet and family history.

- ***Gestational Diabetes:***It is the 3rd main form and occurs when pregnant women without a previous history of diabetes develop a high blood glucose level.[7]

Diabetes affects human organs such as kidney, eye, heart, nerves, foot, etc…Type I, Type II diabetes can't be cured, they can be controlled and treated by special diets, regular exercise and insulin injection.

The paper is organized as follows: Section II describes the related works. Section III deals with the methodology of two algorithms. Section IV discusses about the results of two algorithms and Section V concludes the paper.

II. RELATED WORKS

AiswaryaIyer, et al. [1] have employed Decision tree (J48), Naïve Bayes algorithms for predicting diabetes. They used Pima Indian Diabetes dataset; it was implemented using WEKA tool. They found Naive Bayes algorithm gave 79.56% accuracy than another for predicting diabetes.V.AnujaKumari, R.Chitra ,[2] used SVM with Radial Basis Function Kernal for classification of diabetes

disease. They used MATLAB, R2010a for implementation. They found the accuracy rate as 78%.

N. Sarma, et al,[4] used Bayesian net classifier and decision tree for Predicting Diabetes Type 2. They used PIMA indian diabetic dataset.They used WEKA tool for their implementation in that they found bayes net classifier gives the accuracy level of 71-74% depending upon the number of cross validation applied on the dataset when performing the test.and decision tree gives the accuracy level of 78-80% Which is the best accuracy without implementing any neural network structure.

P.Padmaja et al.[5] used clustering concepts for character evaluation of diabetes. They evaluated 5 different clusters by using 4 algorithms, namely 1) K-means, 2) Partitioning Around Medoids(PAM), 3) Minimum spanning tree (MST), 4) Nearest Neighbours used to identify good quality clusters. The result they found was, PAM provides cluster of good quality.

G.Parthiban, S.K.Srivatsa [6] used Naïve Bayes, SVM Techniques for Diagnosing Heart Disease for Diabetic Patients. They used WEKA tool and got the result as 94.6% of accuracy for SVM.Dr. M. Renuka Devi andJ. Maria Shyla[7] explored various Data mining techniques such as Naïve Bayes, MLP, Bayesian Network, C4.5 , ANN, Modified J48, etc…They used MATLAB and WEKA tool. In that paper, Modified J48 classifier gave 99.87% of highest accuracy. RupaBagdi et al. [8] compared ID3 and C4.5 decision tree algorithm results. Finally they found C4.5 was more precise than ID3.

Sadri sa'di et al.[9] used Naive Bayes, RBF Network and J48 datamining algorithms for diagnosing type II diabetes. They used WEKA tool. Finally they found Naive Bayes, having the accuracy rate of 76.96% than other algorithms.Sankaranarayanan.S et al. [10] intended to discover the hidden knowledge from a particular dataset to improve the quality of health care for diabetic patients.

Satheeskumar.B, Gayathri.P,[11] used Data mining Classification Algorithms such as CART,J48,NBTree forAnalysis of Adult-Onset Diabetes. They used WEKA tool for implementing these algorithms. They found the accuracy rate as 80% for J48 algorithm when compared to other algorithms.

TahaniDaghistani and RiyadAlshammari, [13] used MNGHA, saudhi Arabia dataset to predict diabetic patients using 18 risk factors. They found RandomForest achieved the best performance when compared to other data mining classifiers. V. Kumar and L. Velide,[14] used Data mining Approach for Prediction and Treatment Of diabetes Disease. The techniques they used as Naïve Bayes, JRip, J48 (4.5), DT, NN .They used WEKA tool for implementation. They got 68.5% of accuracy level for J48 algorithm.

Venkatesan, P., and S. Anitha.[15] they made to study the applicability of a general purpose, supervised feed forward neural network with one hidden layer, namely. Radial Basis Function (RBF) neural network. It uses relatively smaller number of locally tuned units and is adaptive in nature. RBFs are suitable for pattern recognition and classification. Performance of the RBF neural network was also compared with the most commonly used multilayer perceptron network model and the classical logistic regression. Diabetes database was used for empirical comparisons and the results show that RBF network performs better than other models .

III. METHODOLOGY

A. Naïve Bayesian classifier:

Bayesian classification represents a supervised learning method as well as statistical method for classification. It is simple probabilistic classifier based on Bayesian theorem with strong independence assumption. It is particularly suited when the dimensionality of input is high. They can predict the probability that a given tuple belongs to a particular class. This classification is named after Thomas Bayes (1702-1761) who proposed the bayes theorem.

Bayesian formula can be written as :

$$P(H \mid E) = [P(E \mid H) * P(H)] / P(E)$$

The basic idea of Bayes's rule is that the outcome of a hypothesis or an event (H) can be predicted based on some evidences (E) that can be observed from the Bayes's rule.[12]

This algorithm provides a prediction model in relation to the likelihood of certain outcomes. Naive Bayes algorithm measures patterns or relationships among data by counting the number of observations. The algorithm then creates a model that reflects the patterns and their relationships. After creating this model, it can be used as a prediction of several objectives.

B. J48

Decision Tree learning is one of the most widely used andpractical methods for inductive inference over superviseddata.

- It represents a procedure for classifyingcategorical data based on their attributes.It is also efficient for processing large amount of data, itis often used in data mining application.
- The construction of decision tree does not require anydomain knowledge or parameter setting, and thereforeappropriate for exploratory knowledge discovery.Their representation of acquired

knowledge in tree formis intuitive and easy to assimilate by humans.

J48 is a decision tree that uses the concept of entropy with a training dataset. The decision tree is a method to display a series of rules, leading to a class or value. In J48 algorithm, every feature of the data is used to make a decision by splitting into smaller subsets. J48 uses a statistical value called the Information Gain to determine how much a property can separate the training data according to their classification. The information interest of a feature is the amount of entropy reduction that can be achieved by separating data through this feature.

IV. RESULTS AND DISCUSSION

The Diabetic 130-US hospital dataset in the year 1999-2008 was taken [16].It has 55 attributes, 101766 instances .In that we have taken 8 clinical attributes and 101766 Instances. They are : Patient_nbr, gender, age, number_diagnoses, max_glu_serum, A1Cresult, insulin, DiabetesMad. Naïve Bayes, J48 algorithmsresults are shown in Table1, Table 2.

TABLE 1. PERFORMANCE EVALUATION OF TWO DIFFERENT ALGORITHMS USING WEKA TOOL

Algorithms	Naive Bayes	J48
Correctly classified Instances	77.2173 %	79.6857 %
Incorrectly classified Instances	22.7827 %	20.3143 %
Time taken to test model on training data	0.94 sec	0.27 sec
TP Rate	0.772	0.797
FP Rate	0.363	0.380
Precision	0.785	0.795
Recall	0.772	0.797
F-measure	0.778	0.796
MCC	0.391	0.421
ROC Area	0.854	0.874
PRC Area	0.858	0.874
Kappa Statistics	0.3889	0.4212

In table 1 shows Naïve bayesproduces 77.2% of correctly classified instances and J48 produces 79.6% of correctly classified instances.Time taken to test the model for Naïve bayes is 0.94 sec and J48 is 0.27sec. The precision value for Naïve bayes is 0.785 and for J48 is 0.795. The recall value for Naïve bayes is 0.772 and for J48 is 0.797.

TABLE 2. ERROR RATE OF TWO DIFFERENT ALGORITHMS USING WEKA TOOL

Algorithms	Naive Bayes	J48
Mean absolute error	0.2308	0.2261
Root mean squared error	0.3415	0.3362
Relative absolute error	65.1692 %	63.84 %
Root relative squared error	81.1543 %	79.9002

In table 2 shows Mean absolute error rate for Naïve bayes is 0.2308 and J48 is 0.2261.Root mean squared error is 0.3415 and J48 is0.3362. Relative absolute error for Naïve Bayes is 65.1692% and for J48 is 63.84%.Root relative squared error for Naïve bayes is 81.15% and for J48 is 79.9%

V. CONCLUSION

The automatic diagnosis of diabetes is an important real-world medical problem. Detection of diabetes in its early stageis the key for treatment. In this work, we have compared two classification algorithms. Among this, J48 algorithm outperformed when compared to Naïve Bayes algorithm. In future study, the work can be extended and improves for the automation of diabetes analysis.

REFERENCES

[1] A. Iyer, J. Jeyalatha, and R. Sumbaly, "Diagnosis of Diabetes using Classification mining Techniques," *IJDKP*, vol. 5, no. 1, Jan. 2015.

[2] V. A. Kumari and R. Chitra, "Classification of Diabetes Disease using Support Vector Machine," *IJERA*, vol. 3, no. 2, pp. 1797–1801, Apr. 2013.

[3] Jiawei Han and MichelineKamber "Data Mining Concepts and Techniques", second edition. Morgon Kaufmann Publishers, 2007.ISBN: 978-81-312-0535-8.

[4] N. Sarma, S. Kumar, and A. Kr. Saini, "A Comparative Study On Decision Tree And Bayes Net Classifier For Predicting Deabetes Type 2," *IJSRET*, 2014.

[5] P. Padmaja, S. Viikkurty, N. I. Siddiqui, P. Dasari, B. Ambica, V. B. V. . VenkataRao, M.ValiShaik, and V. J. P. R. Rudraraju, "Characteristic evaluation of Diabetes data using Clustering Techniques," *IJCSNS*, vol. 8, no. 11, Nov. 2008.

[6] G. Parthiban and S. K. Srivatsa, "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients," *IJAIS*, vol. 3, no. 7, 2012.

[7] Dr. M. RenukaDevi,J. Maria Shyla,"Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730.

[8] R. Bagdi and P. P. Patil, "Diagnosis of Diabetes using OLAP and Data mining Integration," *IJCSCN*, vol. 2, no. 3, pp. 314–322.

[9] S. sa'di, A. Maleki, R. Hashemi, Z. Panbechi, and K. Chalabi, "Comparison of Datamining Algorithms in the Diagnosis of Type II Diabetes," *IJCSA*, vol. 5, no. 5, Oct. 2015.

[10] Sankaranarayanan.S and DrPramanandaPerumal.T, "Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies", World Congress on Computing and Communication Technologies, 2014, pp. 231-233.

[11] S. Kumar B and G. P, "Analysis of Adult-Onset Diabetes using Data mining Classification Algorithms," *IJMCS*, vol. 2, no. 3, Jun. 2014.

[12] Sunita Joshi, BhuwaneshwariPandeyNitin Joshi,"Comparative analysis of Naive Bayes and J48 Classification Algorithms",IJARCSSE,vol. 5,no .12, Dec 2015.

[13] T. Daghistani and R. Alshamimar, "Diagnosis of Diabetes by Applying Data Mining Classification Techniques," *IJACSA*, vol. 7, no. 7, 2016.

[14] V. Kumar and L. Velide, "A Data mining Approach for Prediction and Treatment Ofdiabetes Disease," *IJSIT*, vol. 3, no. 1, pp. 073–079, 2014.

[15] Venkatesan, P., and S. Anitha. "Application of a radial basis function neural network for diagnosis of diabetes mellitus."Current Science 91, no. 9, pp. 1195-1199, 2006.

[16] https://archive.ics.**uci**.edu/.../**datasets**/**Diabetes**+130-US+hospitals+for+years+1999-200