Improved J48 Classification Algorithm for the Prediction of Diabetes

Gaganjot Kaur
Department of Computer Science and Engineering
GNDU, Amritsar (Pb.), India

Amit Chhabra
Department of Computer Science and Engineering
GNDU, Amritsar (Pb.), India

ABSTRACT
This research work deals with efficient data mining procedure for predicting the diabetes from medical records of patients. Diabetes is a very common disease these days in all populations and in all age groups. Diabetes contributes to heart disease, increases the risks of developing kidney disease, nerve damage, blood vessel damage and blindness. So mining the diabetes data in efficient manner is a critical issue. The Pima Indians Diabetes Data Set is used in this paper; which collects the information of patients with and without having diabetes. The modified J48 classifier is used to increase the accuracy rate of the data mining procedure. The data mining tool WEKA has been used as an API of MATLAB for generating the J-48 classifiers. Experimental results showed a significant improvement over the existing J-48 algorithm.

Keywords
J48 Decision Tree, MATLAB, Data Mining, Diabetes, WEKA.

1. INTRODUCTION
Data mining is a process to discover interesting knowledge, such as associations, patterns, anomalies, changes and significant structures from large amount of data stored in databases or other information repositories. In the procedure of data mining the former data is explained and future rules are calculated by data analysis. Data mining is a major advancement in the type of analytical tools. Data mining is a multi-disciplinary field which is a combination of machine learning, statistics, database technology and artificial intelligence. This technique includes a number of phases: Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. Data mining has proven to be very beneficial in the field of medical analysis as it increases diagnostic accuracy, to reduce costs of patient treatment and to save human resources [16] .There are various data mining techniques such as Association, Classification, Clustering, Neural Network and Regression.

1.1 Diabetes
Diabetes is an appropriate disease for data mining technology due to a number of reasons. In every age group this disease is common. It charges plenty of money and its effect is growing quickly. The body of a diabetic person does not produce or efficiently use insulin, the hormone that "unlocks" the cells of the body, allowing glucose to arrive and fuel them. A diabetic person has risk of having the other diseases as blood vessel harm, blindness, heart disease, nerve damage and kidney disease [3]. Diabetes is generally of 2 kinds: type 1(insulin dependent diabetes) and type 2(non-insulin-dependent diabetes).

Diabetes is a disease in which the blood glucose levels get increase which is due to the defects in secretion of insulin, or its action, or both. Diabetes is a prolonged medical disease. In diabetes, the cells of a person produce insufficient amount of insulin or defective insulin or may unable to use insulin properly and efficiently that further leads to hyperglycemia and type-2 diabetes. In type 1 diabetes there is absolute lack of insulin, usually secondary to a destructive process distressing the insulin-producing beta cells in the pancreas. There is excess decline of beta cells that enhances process of elevated blood sugars in type 2 diabetes. In current time it is one of the major public health problems. The International Diabetes Federation has claimed that presently 246 million people are suffering from diabetes worldwide and this number is expected to increase up to 380 million by 2025 [16].

2. J48 DECISION TREE
Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found [15]. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable [5].

J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

2.1 Basic Steps in the Algorithm: [15]
(i) In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class.

(ii) The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.

(iii) Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

2.2 Counting Gain

This process uses the "Entropy" which is a measure of the data disorder. The Entropy of $\vec{y}$ is calculated by

$$\boldsymbol{Entropy}(\vec{\boldsymbol{y}}) = -\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \frac{|\boldsymbol{y_i}|}{|\vec{\boldsymbol{y}}|} \mathbf{log}\left(\frac{|\boldsymbol{y_i}|}{|\vec{\boldsymbol{y}}|}\right)$$

$$\boldsymbol{Entropy}(\boldsymbol{j}|\vec{\boldsymbol{y}}) = \frac{|\boldsymbol{y_j}|}{|\overrightarrow{\boldsymbol{y}}|} \mathbf{log}\left(\frac{|\boldsymbol{y_j}|}{|\vec{\boldsymbol{y}}|}\right)$$

And Gain is

$$\boldsymbol{Gain}(\vec{\boldsymbol{y}}, \boldsymbol{j}) = \boldsymbol{Entropy}(\vec{\boldsymbol{y}} - \boldsymbol{Entropy}(\boldsymbol{j}|\vec{\boldsymbol{y}}))\ |$$

The objective is to maximize the Gain, dividing by overall entropy due to split argument $\vec{y}$ by value j.

2.3 Pruning

Because of the outliers this is a significant step to the result. Some instances are present in all data sets which are not well-defined and differ from the other instances on its neighbourhood.

The classification is performed on the instances of the training set and tree is formed. The pruning is performed for decreasing classification errors which are being produced by specialization in the training set. Pruning is performed for the generalisation of the tree.

2.4 Features of the Algorithm

(i) Both the discrete and continuous attributes are handled by this algorithm. A threshold value is decided by C4.5 for handling continuous attributes. This value divides the data list into those who have their attribute value below the threshold and those having more than or equal to it.

(ii) This algorithm also handles the missing values in the training data.

(iii) After the tree is fully constructed, this algorithm performs the pruning of the tree. C4.5 after its construction drives back through the tree and challenges to remove branches that are not helping in reaching the leaf nodes.

3. REATED WORK

The diabetes of the patients is calculated [1] by using the decision tree in two phases: data pre-processing in which the attributes are identified and second is diabetes prediction model constructed with the help of using the decision tree method. Both the phases are implemented using WEKA data mining tool. The performance comparison of Decision Tree Algorithms and Artificial Neural Network [2] on medical data was performed on the bases of parameters as kappa statistics, mean absolute error, relative squared error, time to model and mean-squared error. On the basis of results it has been examined that Decision Tree Algorithms performs better than the Artificial Neural Network. Hypertension has been predicted by generating [3] J-48 and Naive Bayesian classifiers in WEKA. The overall accuracy is around 83%. A slight improvement of ensemble five J-48 classifier was seen over pure Naive Bayesian and J-48 in sensitivity, accuracy and F-measure. Rough set tools were able to decrease the ensemble of five members to three but there is substantial growth of sensitivity. In [4] the diabetes disease analysis is performed by using the artificial meta plasticity on multilayer perceptron. The results attained by artificial meta plasticity on multilayer perceptron were compared with Bayesian classifier (BC), decision tree (DT) using same database. Decision tree performed the best classification on the basis of standard deviation. It [6] classifies the influence of various variables on life expectancy at birth. Cross Industry Standard Process and Sample, Explore, Modify, Model, Assess (SEMMA) are the two main methodologies that govern data mining. Decision tree method is chosen as the optimal for the problem as it has shown better results than the algorithms.

J48, Random Forest, Naive Bayes etc. algorithms [7] are used for disease diagnosis as they led to good accuracy. They were used to make predictions. The dynamic interface can also use the constructed models that mean the application worked properly in each considered case. The classification algorithms [8] Naive Bayes, decision tree (J48), Sequential Minimal Optimization (SMO), Instance Based for K-Nearest neighbour (IBK) and Multi-Layer Perception are compared by using matrix and classification accuracy. Three different breast cancer databases have been used and classification accuracy is presented on the bases of 10-fold cross validation method. A combination at classification level is accomplished between these classifiers to get the best multi-classifier approach and accuracy for each data set. Diabetes and cardiac diseases [10] are predicted using Decision Tree and Incremental Learning at the early stage. The i+Learning and i+LRA performs better than ID3 and other incremental learning algorithms on the bases of classification accuracy. These both algorithms can even handle the new attributes without affecting the learning performance. The main drawback of this method is that it adopts the binary tree rather than multi-branch tree.

For automatically detecting the disease detection in retinal image analysis an approach has been proposed [11]. The data mining techniques have been used to accurately categorize the Normal, Diabetic Retinopathy and Glaucoma affected retinal images. It has been proved that random tree and C4.5 classification techniques have achieved the maximum accuracy of 100% in classifying 45 images from the Gold Standard Database. [12] MLR and ADTree models are compared on the basis robustness against missing values. MLR is less robust against missing values than ADTree. At low boosting and ensemble number sufficient robustness is achieved and as these numbers increase it is compromised. Based on a hybrid method using unified Collaborative Filtering and multiple classifications, a Chronic Disease Diagnosis Recommender System approach [13] has been proposed. While recommending medical advices for patients the Unified CF method based on learning classification model using both historical binary ratings and external features will be used for attaining higher recommendation accuracy.

4. EXPERIMENTAL SETUP

4.1 WEKA

WEKA is an innovatory tool in the history of the data mining and machine learning research communities. By putting efforts since 1994 this tool was developed by WEKA team. WEKA contains many inbuilt algorithms for data mining and machine learning. It is open source and freely available platform-independent software. The people who are not having much knowledge of data mining can also use this software very easily as it provides flexible facilities for scripting experiments. As new algorithms appear in research

literature, these are updated in software. WEKA has also became one of the favourite tool for data mining research and helped to progress it by making many powerful features available to all.

The steps performed for data mining in WEKA are:

• Data pre-processing and visualization

• Attribute selection

• Classification (Decision trees)

• Prediction (Nearest neighbour)

• Model evaluation

• Clustering (Cobweb, K-means)

• Association rules

4.2 Improvement of J48

The proposed algorithm uses WEKA as API in MATLAB. WEKA is a comprehensive open source Machine Learning toolkit, written in Java. These functions provide a basic MATLAB interface to WEKA to allow the transformation of data back and forth and to use most of the features available in WEKA, such as training classifiers. By doing so the accuracy rate of the J48 algorithm has increased to large extent as compared to the accuracy of the same algorithm in WEKA. The proposed algorithm works as: The arff data file is loaded from WEKA into MATLAB. Then the refining of the dataset is done. Later the J48 classifier is applied. At the end the results are obtained that is the accuracy and error rate is calculated. Fig. 1 has shown the flow chart of the proposed algorithm.

5. PERFORMANCE EVALUATION

This section has shown the comparison of the different data mining algorithms. The formula to calculate accuracy is:

$$\text{(i)} \qquad TA = \frac{(TP+TN)}{TP+TN+FP+FN}$$

$$\text{(ii)} \qquad RA = \frac{(TP+FP)*(TN+FN)+(FN+TP)*(FP+TP)}{(Total*Total)}$$

In the equation (i) TA represents Total Accuracy, TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative. In equation (ii) RA represents Random Accuracy.

Fig 2 shows the tested negative and tested positive values of diabetes with respect to the different attributes. It shows that the diabetes function pedigree is the least significant and 2-hour serum insulin is the most significant attribute. If the value of the insulin is above 800 than most probably the person is diabetic.

Table 1 has shown that the results of proposed algorithm are quite significant as compared to other algorithms in data mining. All the algorithms other than proposed algorithm have accuracy rate below than 78% but the accuracy of the proposed algorithm is 99.87%.

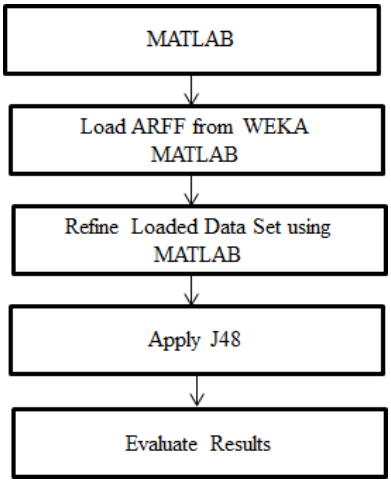

Fig 1 Flow Chart and Proposed Set-Up

Table 1: Performance comparison between different algorithms

Algorithm	**Accuracy**	**Error**
NaiveBayes	76.3021	23.6979
MultilayerPreception	73.3906	26.6094
MultiClassClassifier	77.2135	22.7865
RandomTree	68.099	31.901
REPTree	68.099	31.901
LADTree	74.0885	25.9115
BFTree	73.5677	26.4323
ADTree	72.9167	27.0833
RandomForest	73.9583	26.0417
J48	73.8281	26.1719
Proposed Algorithm	99.8700	0.1300

Table 1 shows the comparison graph of different algorithms on the base of accuracy and error. It clearly states that the proposed algorithm has large accuracy difference than other algorithms. It has accuracy rate of 99.87% rather than others that show maximum of 77.21%accuracy.

6. CONCLUSION & FUTURE WORK

This research work has proposed a new approach for efficiently predicting the diabetes from medical records of patients. The Pima Indians Diabetes Data Set has been used for experimental purpose. It has come up with the information of patients with and without having diabetes. The modified J48 classifier has been used to increase the accuracy rate of the data mining procedure. The data mining tool WEKA has been used as an API of MATLAB for generating the modified J-48 classifiers. Experimental results have shown a significant improvement over the existing J-48 algorithm. It has been proved that the proposed algorithm can achieve accuracy up to 99.87 %.

In near future we will use some more data sets to validate the proposed algorithm. Only 768 instances have been used in this research work presently; in future a large data set will also be considered.

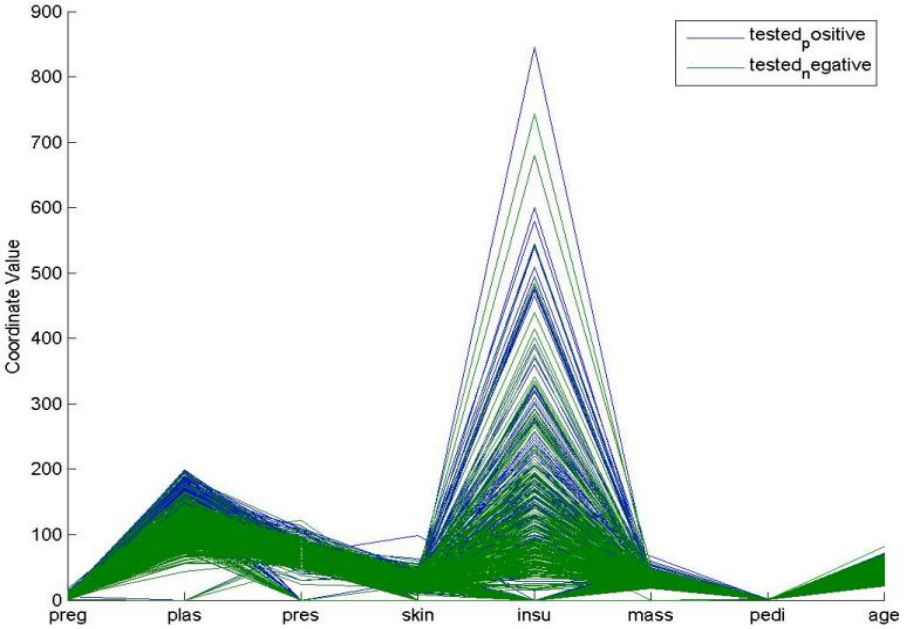

Fig 2: Evaluation of Diabetes

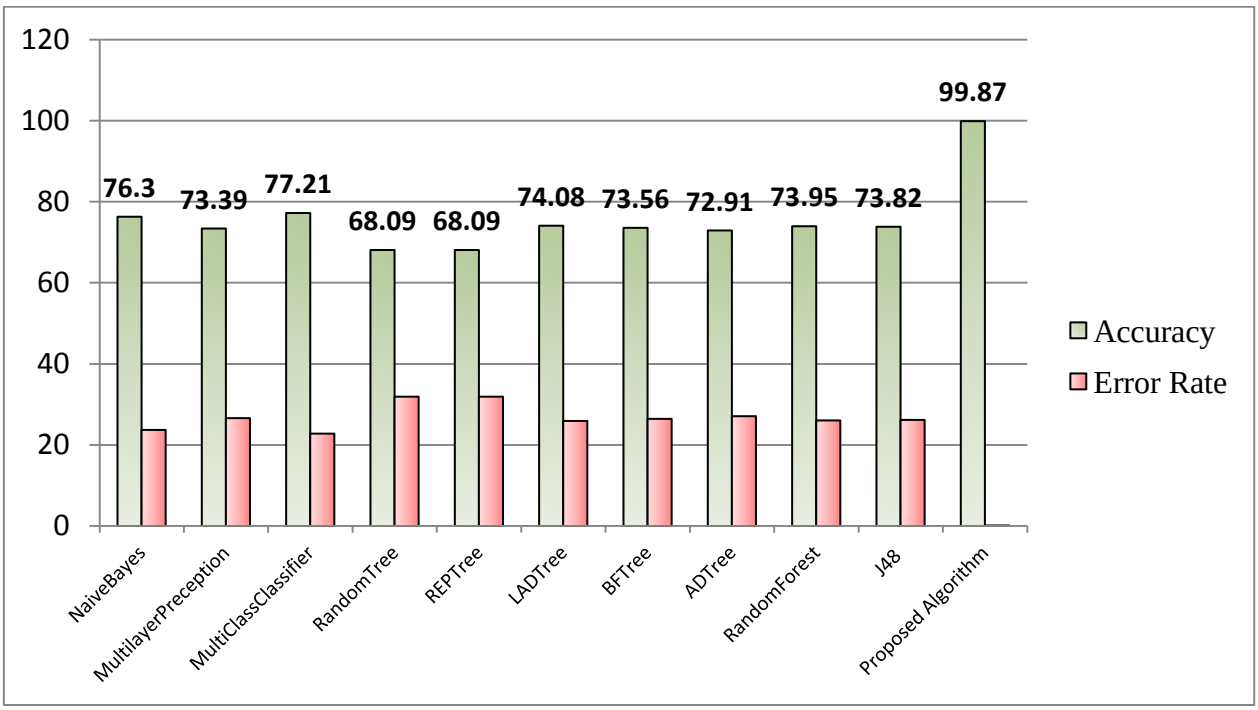

Fig. 3 Graph showing the accuracy and error rate of different algorithms

7. REFERENCES

[1] Al Jarullah, A.A., "Decision tree discovery for the diagnosis of type II diabetes," Innovations in Information Technology (IIT), 2011 International Conference on , vol., no., pp.303,307, 25-27 April 2011

[2] Folorunsho, Olaiya. "Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database." International Journal 3, no. 3 (2013).

[3] Huang, Feixiang; Wang, Shengyong; Chan, Chien-Chung, "Predicting disease by using data mining based on healthcare information system," Granular Computing (GrC), 2012 IEEE International Conference on , vol., no., pp.191,194, 11-13 Aug. 2012

[4] Marcano-Cedeno, Alexis; Andina, Diego, "Data mining for the diagnosis of type 2 diabetes," World Automation Congress (WAC), 2012 , vol., no., pp.1,6, 24-28 June 2012.

[5] Nadali, A; Kakhky, E.N.; Nosratabadi, H.E., "Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system," Electronics Computer Technology (ICECT), 2011 3rd International Conference on , vol.6, no., pp.161,165, 8-10 April 2011

[6] Nincevic, I.; Cukusic, M.; Garaca, Z., "Mining demographic data with decision trees," MIPRO, 2010 Proceedings of the 33rd International Convention , vol., no., pp.1288,1293, 24-28 May 2010

[7] Robu, R.; Hora, C., "Medical data mining with extended WEKA," Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on , vol., no., pp.347,350, 13-15 June 2012

[8] Salama, G.I.; Abdelhalim, M.B.; Zeid, M.A., "Experimental comparison of classifiers for breast cancer diagnosis," Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on , vol., no., pp.180,185, 27-29 Nov.,2012.

[9] S. Moertini Veronica ,"Towards The Use Of C4.5 Algorithm For Classifying Banking Dataset",Integeral Vol 8 No 2,October 2013.

[10] UM, Ashwinkumar, and Anandakumar KR. "Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques.",IEEE,pp:161-165,2011

[11] Geetha Ramani R, Lakshmi Balasubramanian, and Shomona Gracia Jacob. "Automatic prediction of Diabetic Retinopathy and Glaucoma through retinal image analysis and data mining techniques." In Machine Vision and Image Processing (MVIP), 2012 International Conference on, pp. 149-152. IEEE, 2012

[12] Sugimoto, Masahiro, Masahiro Takada and Masakazu Toi. "Comparison of robustness against missing values of alternative decision tree and multiple logistic regression for predicting clinical data in primary breast cancer." In Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, pp. 3054-3057. IEEE, 2013.

[13] Hussein Asmaa S,Wail M. Omar, Xue Li, and Modafar Ati. "Efficient Chronic Disease Diagnosis prediction and recommendation system." In Biomedical Engineering and Sciences (IECBES), 2012 IEEE EMBS Conference on, pp. 209-214. IEEE, 2012.

[14] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[15] Korting, Thales Sehn. "C4. 5 algorithm and Multivariate Decision Trees." Image Processing Division, National Institute for Space Research--INPE.

[16] Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." In Internet Technology And Secured Transactions, 2012 International Conferece For, pp. 471-472. IEEE, 2012.