3rd International Conference on Science,Technology and Management
India International Center, New Delhi
17 January 2016, www.conferenceworld.in
(ICSTM-16)
978-81-932074-0-6

ANALYSING PERFORMANCE MONITORING OF DIFFERENT CLASSIFICATION TECHNIQUES FOR DIEBETIC DATA

Ms.M.C.S.Geetha[1], Dr.I.Elizabeth Shanthi[2]

[1]Assistant Professor, Department of Computer Applications, Kumaraguru College of Technology, (India)

[2]Associate Professor, Computer Science,

Avinashilingam Institution for Home Science and Higher Education for Women,

Avinashilingam University, (India)

ABSTRACT

Data mining is the current research area to solve various problems and classification is one of the main problem in the field of data mining. It allows users to analyze data from different dimensions or angles, categorize it, and summarize the relationships identified. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. In this paper we present algorithmic discussion of J48, Random tree, J48 Graft, LAD and REP. Here the performance is compared for computing time, correctly classified instances, kappa statistics, RMSE, MAE, RRSE, RAE and to find the error rate measurement for different classifiers using weka tool. We have taken the classification of data for diabetic patients data set is developed by collecting data from hospital repository which has 1865 instances with different attributes. The dataset instances consist of two categories of blood tests and urine tests. Weka tool classifies the data and is evaluated using 10 fold cross validation and the results are then measured. It discovered that J48 performs better in most of the cases.

Keywords: Data Mining, Classification, J48, Random tree

I. INTRODUCTION

The goal of this paper is to correctly classify the datasets, therefore a doctor can safely and cost effectively select the most excellent datasets for the diagnosis of the disease. The main motivation is that diabetes affects a large number of people and it is a hard to diagnose the disease. A diagnosis is a constant process in which a doctor collects information from a patient and other sources, like friends and family, and from substantial datasets of the patient. The practice of creating a diagnosis begins with the identification of the patient's symptoms.

The symptoms will be the source of the hypothesis from which the doctor will start investigate the patient. This is the main concern to optimize the duty of appropriately selecting the set of medical tests that a patient must carry out to have the finest, the less expensive and time consuming analysis possible. A result will not only help doctors in making assessment, and make all this process livelier, it will also lessen health care costs and waiting times for the patients. This paper will focus on the study of data from a data set called Diabetes data set.

The paper is organized as follows: Chapter 2 discusses the related work. Chapter 3 discusses the materials and methods. Chapter 4 discusses the methodology and Chapter 5 discusses the conclusion.

II. RELATED WORK

A lot of research is carried out in this medical field. Past research in dealing with this problem can be explained with the subsequent approaches: (a) Discover all rules initially and then permit the user to query and recover those he/she is interested in. The approach is that of templates [3]. This approach lets the user to specify what rules he/she is involved as templates. The system then uses the templates to recover the rules that match the templates from the set of exposed rules. (b) Use constraints to restrict the mining process to generate only related rules. [4] Report their knowledge in trying to repeatedly acquire medical knowledge from medical databases. They have tested three medical databases and the rules encouraged are used to evaluate against a set of predefined medical rules. [12] Suggest an algorithm that can take item constraint specified by the user in the association rule mining processor that only those rules that assure the user specified item constraints are produced.

The study helps in forecasting the status of diabetes i.e., whether it is in an early stage or in a highly developed stage based on the typical results and also helps in guessing the maximum number of patients suffering from diabetes with particular characteristics. Thus patients can be given helpful treatment by efficiently diagnosing the characteristics.

The major use of the technique is to have a vigorous working model of this technology. The process of designing a model helps to classify the different blood groups with existing Hospital Classification techniques for analysis of Blood group data sets. The capability to classify regular diabetic patients will enable to plan scientifically for organizing in an effective manner. Expansion of data mining technologies to forecast treatment errors in populations of patients characterize a major advance in patient security research.

III. MATERIALS AND METHODS

Weka is a popular suite of machine learning software developed in Java at the University of Waikato, New Zealand. The Weka suite includes a collection of visualization tools and algorithms for data investigation and predictive modeling, together with graphical user interfaces for easy access to this functionality. The data that is used should be in the ARFF (Attribute Relation file format) format and the file should have the extension (.arff). WEKA is a collection of machine learning algorithms for explaining real world data mining problems. Weka supports numerous standard data mining tasks, data preprocessing, classification, clustering, regression, visualization and feature selection.

3.1 Data Preprocessing

The data preprocessing is the first step in the data mining process. One of the challenges that is faced in the knowledge discovery process of medical database is reduced data quality. So the data is prepared carefully to obtain accurate and correct results. First we decide the most related attributes to our mining task.

3.2 Data Mining Stages

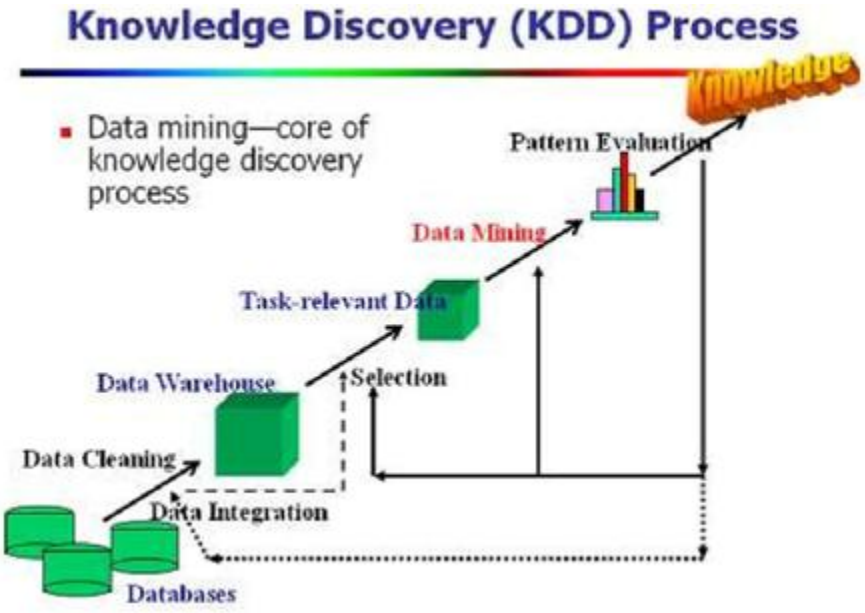

Fig 1: Data Mining Process

The data mining stage was divided into different phases. At each phase all the algorithms were used to analyze the medical datasets. The testing method assumed for this research was parentage split that train on a fraction of the dataset, cross validate on it and test on the remaining fraction. Sixty six percent (66%) of the health dataset which were arbitrarily selected was used to instruct the dataset using all the classifiers. The validation was agreed using ten folds of the training sets. The representation was now applied to unseen or new dataset which was prepared with thirty four percent (34%) of randomly selected records of the datasets.

3.3 Pattern Evaluation

This is the stage where interesting patterns instead of knowledge are identified based on given metrics.

3.4 Evaluation Matrics

In selecting the suitable algorithms and parameters that model the diabetes forecasting variable, the following performance metrics were used:

3.4.1. **Time:** This is referred as the time required to complete training or modeling of a dataset. It is characterized in seconds.

3.4.2. Kappa Statistic: Evaluating the degree of nonrandom harmony between observers or measurements of the same uncompromising variable.

3.4.3. Root relative squared error: Root relative squared error is the total squared error comparative to the error would have been if the prediction is the average of the absolute value.

3.4.4. Relative Absolute Error: Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values.

3.4.5. Mean Absolute Error: Mean absolute error is the average of the dissimilarity between predicted and the real value in all test cases; it is the average prediction error.

3.4.6. Mean Squared Error: It is one of the most frequently used actions for numeric prediction. This value is calculated by taking the average of the squared difference between each computed value and its equivalent correct value. The mean-squared error is just the square root of the mean-squared-error. The mean-squared error provides the error value the same dimensionality as the actual and predicted values.

IV. METHODOLOGY

4.1 Classification

Classification is a process of finding a model that describes data classes or concepts. It is Based on a set of training data.

4.2 J48 Pruned Tree

J48 is a module for generating a pruned or unpruned C4.5 decision tree. When J48 is applied onto refreshed data, the results got are shown below on Fig 2.

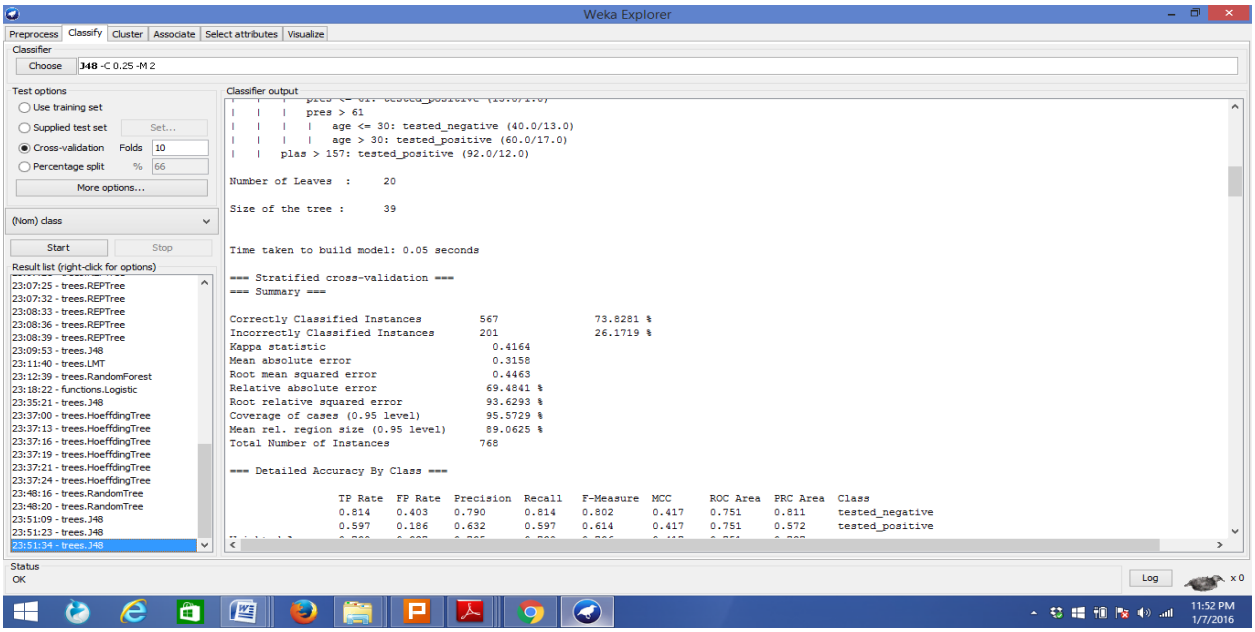

Fig 2 : J48

4.3 Randomtree

Perhaps C4.5 algorithm which was developed by Quinlan [13] is the most popular tree classifier till today. When Randomtree is applied the following results are got in Fig 3.

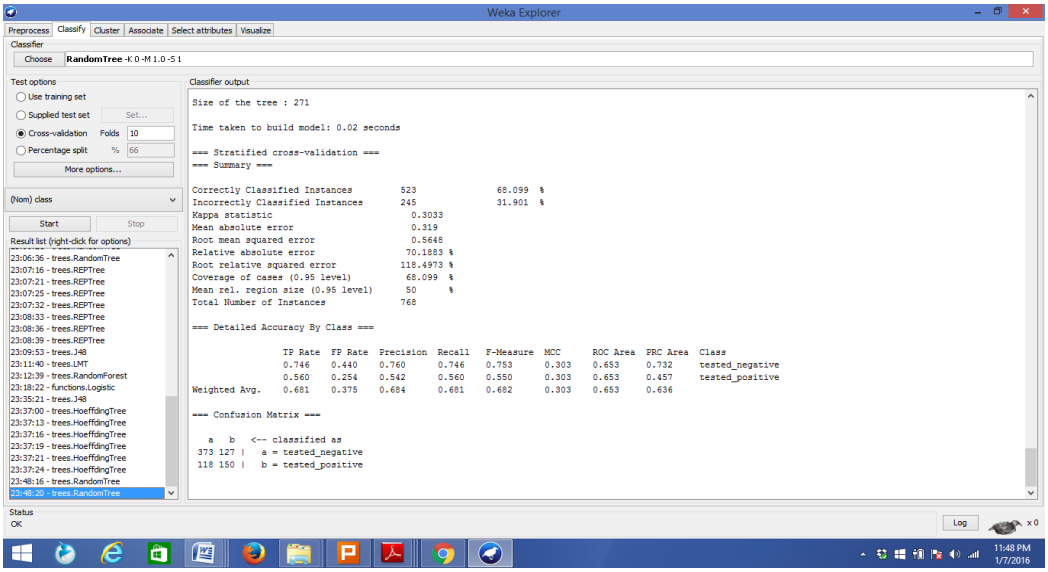

Fig 3: Random tree

4.4 LAD tree

LADTree is a class for generating a multiclass alternating decision tree using logistics strategy. LADTree produces a multi- class LADTree. It has the capability to have more than two class inputs. It performs additive logistic regression using the Logistics Strategy.

4.5 REP Tree

Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).

V. RESULT AND DISCUSSION

J48 algorithm was selected for the prediction because out of the five classifiers used to train the data, it has the best performance measures.

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: py1

Instances: 2804

Attributes: 11

NAME

GENDER

AGE

HEIGHT

BLOOD GROUP

BLOOD SUGAR(F)
BLOOD SUGAR (PP)
BLOOD SUGAR (R)
URINE SUGAR(F)
URINE SUGAR(PP)
URINE SUGAR (R)
Test mode: evaluate on training data
=== Classifier model (full training set) ===
J48 pruned tree

J48 pruned tree

AGE <= 46
| AGE <= 35
| | GENDER = Male
| | | AGE <= 26: B positive (2.0/1.0)
| | | AGE > 26: A positive (3.0/1.0)
| | GENDER = Female
| | | AGE <= 34: O negative (2.0)
| | | AGE > 34: A positive (2.0/1.0)
| AGE > 35: B positive (7.0/4.0)
AGE > 46
| GENDER = Male
| | AGE <= 60: O positive (5.0/3.0)
| | AGE > 60: AB positive (4.0/2.0)
| GENDER = Female
| | AGE <= 63
| | | AGE <= 55: AB positive (2.0/1.0)
| | | AGE > 55: A1B positive (4.0/2.0)
| | AGE > 63: A negative (2.0/1.0)
Number of Leaves : 10
Size of the tree : 19
Time taken to build model: 0.29 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 1865

70.5905%

Incorrectly Classified Instances 777

29.4095%

Kappa statistic 0.6703

Mean absolute error 0.0489

Root mean squared error 0.1564

Relative absolute error 35.5333 %

Root relative squared error 59.6144%

Total Number of Instances 2642

Ignored Class Unknown Instances 162

Table-1: Different Performance Metrices Running In Weka

Classifier	Correctly classified Instances	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
J48	1865 (70.5%)	0.706	0.036	0.727	0.706	0.702	0.981
J48 GRAFT	1524 (57.6%)	0.607	0.024	0.678	0.520	0.600	0.781
LAD TREE	553 (20.9%)	0.05	0.116	0.038	0.05	0.043	0.464
RANDOM TREE	350 (13.2%)	0.111	0.122	0.098	0.111	0.07	0.464
REP TREE	348 (0.13%)	0.132	0.132	0.017	0.132	0.31	0.548

In this study, we examine the performance of different classification methods that generates accuracy and some error. According to Table 1, we can say that the highest accuracy is 70.5% in J48 and lowest accuracy is 0.13% in REP.

Table- 2: Errors Measurement For Different Classifiers In Weka

	J48	J48GRAFT	RANDOM TREE	REP	LAD
TIME	0.29	0.42	0.02	0.05	1.85
CORRECTLY CLASSIFIED INSTANCES	1865 (70.5%)	1524 (57.6%)	350 (13.2%)	348 (0.13%)	553 (20.9%)
KAPPA STATISTIC	0.011	0.6700	0.011	0.012	0.0654
MAE	0.0123	0.0480	0.1798	0.1377	0.1821
RMSE	0.1154	0.1560	0.3199	0.2624	0.3171
RAE%	12.53%	35.50%	100.24%	99.98%	101.55%
RRSE%	22.61%	58.63%	106.82%	100%	105.87%

Based on table 2, errors are evaluated among different classifiers in WEKA and found J48 is the best. The blood groups in both positive and negative are shown in Table-1. Overall blood group A was the commonest (24.03 %), followed by B (18.77%), AB (19.11%), O (23.65) and A1B

Table-3: Overall blood group

Blood group spectrum	Nos (%)	+ve(%)	–ve(%)
A	635 (24.03)	348 (13.17)	287 (10.85)
B	496 (18.77)	289 (10.93)	207 (7.83)
AB	505 (19.11)	196 (7.41)	309 (11.69)
A1B	453 (17.14)	300 (11.35)	153 (5.79)
O	625 (23.65)	345 (10.59)	280 (13.05)

In the present blood group-A was the largest (24.03%) while A1B was the least common (17.14%).

VI. CONCLUSION AND FUTURE WORK

The purpose of this study is to estimate and examine 5 selected classification algorithms on WEKA. The performance reported by J48 classifier is better with an accuracy of 70.59% that takes 0.29 seconds for training. They are used in different healthcare units all over the world. In this paper, WEKA is used to identify the diabetic patient's behavior using the classification algorithms in data mining. The analysis has been carried out using a diabetes data set and J48 decision tree algorithm implemented in WEKA. This paper classifies the diabetic patient based on the age, gender, weight, height, blood group, blood sugar(PP), blood sugar(F), urine sugar(PP), urine sugar(F). The J48 model provided a good classification accuracy..

The future work will be focused on using the other classification algorithms of data mining. The performance of an algorithm is reliant on the domain and the kind of the data set. Therefore, the practice of other classification algorithms like machine learning can be explored in future. Diagnosing cancer patients found on blood cells or predicting the cancer types on the blood groups, blood pressure, personality traits and medical diseases.

REFERENCES

[1] Mats Jontell, Oral medicine, Sahlgrenska Academy, Göteborg University (1998) "*A Computerised Teaching Aid in Oral Medicine and Oral Pathology.* " Olof Torgersson, department of Computing Science, Chalmers University of Technology, Göteborg.

[2] T. Mitchell, "*Decision Tree Learning*", in T. Mitchell, Machine Learning (1997) the McGraw- Hill Companies, Inc., pp. 52-78.

[3] Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I (1994) "*Finding interesting rules from large sets of discovered association rules*," CIKM.

[4] Tsumoto S., (1997), "*Automated Discovery of Plausible Rules Based on Rough Sets and Rough Inclusion*," Proceedings of the Third Pacific-Asia Conference (PAKDD), Beijing, China, pp 210-219.

[5] Liu B., Hsu W., (1996) "*Post-analysis of learned rules*," AAAI, pp. 828-834.

[6] Liu B., Hsu W., and Chen S., (1997) "*Using general impressions to analyze discovered classification rules*," Proceedings of the Third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[7] Stutz J., P. Cheeseman. (1996) Bayesian classification (autoclass): Theory and results. In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press

[8] Witten Ian H., E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Ch. 8, © 2000 Morgan Kaufmann Publishers

[9] http://www.cs.waikato.ac.nz/ml/weka/, accessed 06/05/21.

[10] http://grb.mnsu.edu/grbts/doc/manual/J48_Decision_T rees.html, accessed

[11] Wikipedia, ID3-algorithm (accessed 2007/12/09) (URL: http://en.wikipedia.org/wiki/ID3_algorithm)

[12] Srikant,R.,Vu,Q.andAgrawal,R.,(1997), *"Mining association rules with item constraints,"* Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, USA, pp 67-73.