# Analysis of Different Classifiers for Medical Dataset using Various Measures

Payal Dhakate
ME Student,
Pimpri Chinchwad College of Engineering, Nigdi,
Pune, India.

K. Rajeswari
Associate Professor
Pimpri Chinchwad College of Engineering, Nigdi,
Pune,India

Deepa Abin
Assistant Professor,
Pimpri Chinchwad College of Engineering, Nigdi,
Pune, India

## ABSTRACT

The process of extracting information from a dataset and transforming it into an understandable structure for further use is called as data mining. A number of important techniques such as preprocessing, classification, clustering are performed in data mining using WEKA tool. In medical diagnoses the role of data mining approaches is being increased. Particularly Classification algorithms are very helpful in classifying the data, which is important for decision making process for medical practitioners. To increase the accuracy in the short time ensemble is used. The ensemble is formed by combination of two or more classifiers. For experimentation of ensembles, different types of base classifiers such as Bagging and Adaboost in combination with classifiers and classifiers such as C4.5, J48, and AD tree are used in the medical data set. The experiment is carried out in the WEKA tool on the UCI machine repository. Experimental results for ensemble with bagging classifier shows good accuracy for FT Tree in less time. Also arrthmia dataset shows the highest average accuracy.

## Keywords

AD Tree; J48; Random Tree; REP Tree; Simple cart; WEKA;

## 1. INTRODUCTION

Data mining is the process of automatic classification based on data patterns obtained from a dataset. It is the extraction or mining of knowledge from large amounts of the data, also called as Knowledge mining, knowledge discovery, knowledge extraction[1] in databases. Different types of algorithms have been developed and implemented for extracting information and discovering knowledge patterns which are useful for decision support.

### 1.1 Weka

WEKA is open source software written in Java, introduced by Waikato University. It contains implementations of algorithms for classification and association rule mining, along with graphical user interfaces and visualization utilities for data exploration and algorithm evaluation. It is used in the machine learning and data mining community as an educational tool for teaching both applications and the technical internals of machine learning algorithms, a research tool for developing and comparing new techniques. It is applied increasingly widely in other academic fields, and in commercial settings. It is free and open source software is the secret of WEKA's success. However, there are several other factors such as portability, graphical user interface, extensibility documentation and support [33]. Figure 1 shows the WEKA interface. We can perform preprocessing and classification in WEKA using different types of classifiers.

Classification is the process of finding a model or function which describes and distinguishes data classes or concepts, for the intention of using the model to predict the class of objects whose class label is unknown [1]. Different types of classifiers are used for classification such as naïve Bayes, J48, C4.5 and decision tree etc.

Waikato Environment for Knowledge Analysis
Version 3.6.2
(c) 1999 - 2010
The University of Waikato
Hamilton, New Zealand

**Figure1. WEKA Interface**

### 1.2 Ensemble

An ensemble is a supervised learning algorithm, because it can be trained and then used to make predictions. Ensembles are grouped two or more classifiers. These ensemble systems contain redundant members those if removed, may further increase group diversity and produces better results. The ensembles are smaller in size relaxes the memory and storage requirements, reducing system's run-time overhead along with improving overall efficiency. The trained ensemble represents a single hypothesis. Ensembles can be shown to have more flexibility in the functions they can represent. Figure 2. Shows the process creating of an ensemble. The term ensemble is usually reserved for methods that generate multiple hypotheses using the same base learner. The prediction of an ensemble typically requires more computation than to predict a single model, so ensembles may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation. Faster algorithms such as decision tree are commonly used with ensembles, although slower algorithms can benefit from ensemble techniques as well.

The problem is a comparative study of classification technique such as Random Forest, FT tree, REP Tree, Simple cart and J48 using base classifiers called as ensembles using various parameters using different data sets. Here we are using base classifier such as Bagging, Adaboost etc. Bagging algorithms used to improve model stability and accuracy. Bagging works well for unstable base models and can reduce variance in predictions[5].

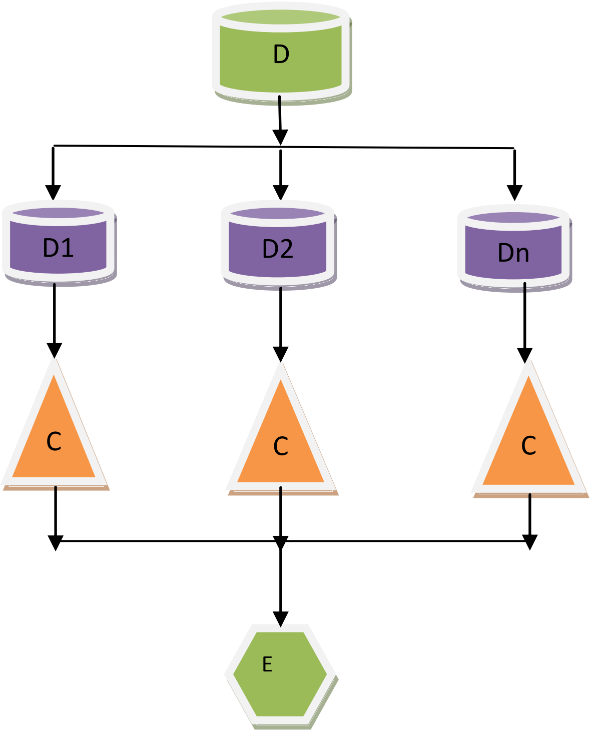

**Figure 2. Creation of ensemble**

# 2. LITERATURE SURVEY

## 2.1 Classifiers

### *2.1.1 FT tree*

FT tree classifier used for building functional trees, which are classification trees that could have logistic regression functions at the inner nodes and leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values.REP Tree is a fast decision tree learner, it builds a decision or regression tree using information gain or variance and prunes it using reduced-error pruning. It only sort values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces.

### *2.1.2 J48 tree*

It builds the decision tree from labeled training data set using information gain and it examines the same that results from choosing an attribute for splitting the data. To make the decision the attribute with highest normalized information gain is used. Then the algorithm recurs on smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class[3]. It is also based on Hunt's algorithm. J48 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, J48 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. J48 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biases of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum[4].

### *2.1.3 Random Tree*

It is a class for constructing a tree that considers K randomly chosen attributes at each node, it performs no pruning. Also has an option to allow estimation of class probabilities based on a holdout set [4]

### *2.1.4 Naive Bayes:*

Naive Bayes classifiers have worked well in many complex real-world situations. Naive Bayes or Bayes Rule is the basis for many machine-learning and data mining methods. The rule is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the evidence by calculating the correlation between the target and other variables. By theory, this classifier has minimum error rate, but it may not be the case always. However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. Observations show that Naïve Bayes has performed consistently before and after reduction of a number of attributes. Naïve bayes is based on probability theory to find the most likely possible classifications [5].

### *2.1.5 J48 Decision Tree*

It is a popular classifier which is simple and easy to implement. J48 Decision Tree with reduced error. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more useful for Feature Selection and knowledge discovery. The performance of decision trees can be enhanced with suitable attribute selection.

### *2.1.6 Bagging*

Bagging is an ensemble method used to classify the data with good accuracy. It is also called as Bootstrap Aggregation. Here first the decision trees are derived by building the base classifiers c1, c2,…, cn on the bootstrap samples D1, D2, .., Dn with replacement from the data set D. Later the final model or decision tree is derived as a combination of all base classifiers c1, c2,…, cn with the majority votes. It can be applied on any classifier such as REP Tree, random forest, C4.5 and J48 etc. Bagging plays an important role in the field of medical diagnosis.

### *2.1.7 AdaBoost*

It is the most famous boosting algorithm. It uses the same training set over and over again also combine an arbitrary number of base learners. AdaBoost is sensitive to noisy data and outliers. It generates and calls a new weak classifier in each of a series of rounds t=1,….,T For each call, a distribution of weights Dt is updated that indicates the importance of examples in the data set for the classification[ 6].

Here in Section 2 proposed method is discussed experimental results and performance evaluation are discussed in Section 3 and in Section 4 conclusion is written.

# 3. PROPOSED METHOD

Classification is the process of finding a model that describes data classes or concepts, for the purpose of predicting the class of objects whose class label is unknown [3]. It is a technique which is used to predict group membership for data instances. Classification is having two steps, first builds a model using training data for that class label must be known and in second the model tested by assigning class labels to data objects in a test data set. The implementation of the ensembles are done in WEKA 3-6-6 and experimented on to standard medical datasets, they are from UCI Data repository. The datasets such as diabetes, Arrhythmia, Wine and breast cancer considered because nowadays the percentage of diabetes patients are growing very fast[7].

Heart disease and Heart attack are one of the major diseases. Heart disease was the major cause of deaths in the different countries including the India. Heart disease kills one person every 34 seconds in the United States. And the cost is about 393.5 billion dollars. Coronary heart disease, and Cardiovascular disease are some categories of heart diseases[8]. The dataset for Heart Cleveland contains 14 attributes and number of instances are 303. Another problem observed in females is breast cancer. It contains total 10 attributes including class attribute and 286 instances. Diabetes dataset contains 9 attributes and 768[7]. India continues to be the "diabetes capital" of the world, and by 2030, nearly 9 per cent of the country's population is likely to be affected with the disease It is estimated that every fifth person with diabetes will be an Indian. This means that India has the highest number of diabetes in any one of the country in the world.

WEKA having facility to convert the data sets from arff format into csv format. 10 fold cross validation is used for the evaluation. For constructing the ensemble we are considering base classifiers such as bagging and adaboost in combinations with classifiers such as J48, C4.5, REP tree. Accuracy and time is very important in the field of medical domain, the performance measure accuracy of classification is considered in this study.

## 4. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

### 4.1 Measures for Performance Evaluation:

To measure the performance accuracy and time are used for the evaluation of any ensembles. As accuracy and time are the important factors for calculating the results. Table I shows the accuracy of different classifiers applied on medical dataset and Figures 3, 4, 5 shows it graphically for the medical datasets. Whereas Table II shows the time required for the construction of the ensemble and figures 6, 7, 8 and 9 shows it graphically.

### *4.1.1 Accuracy:*

It is a ratio of number of correctly classified instances to the total number. of instances and it can be defined as [2].

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + False\ Negative + True\ Negative} \quad (1)$$

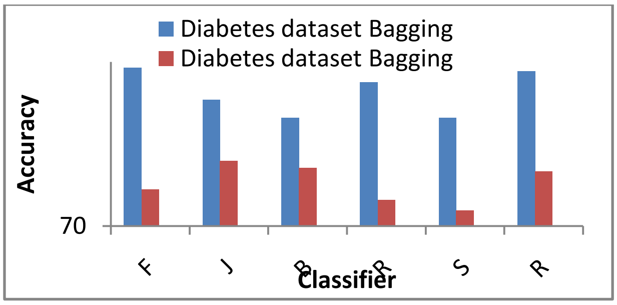

**Figure 3. Accuracy verses classifier graph for diabetes dataset.**

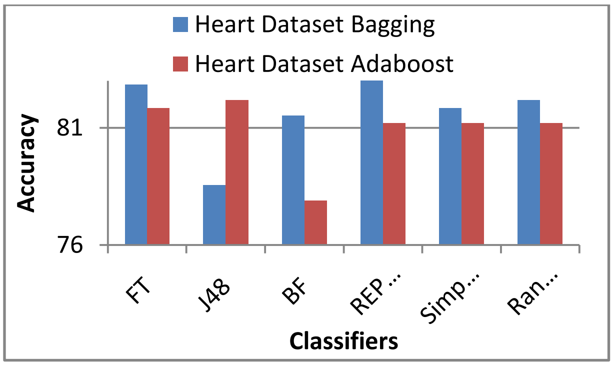

**Figure 4. Accuracy verses classifier graph for heart disease dataset.**

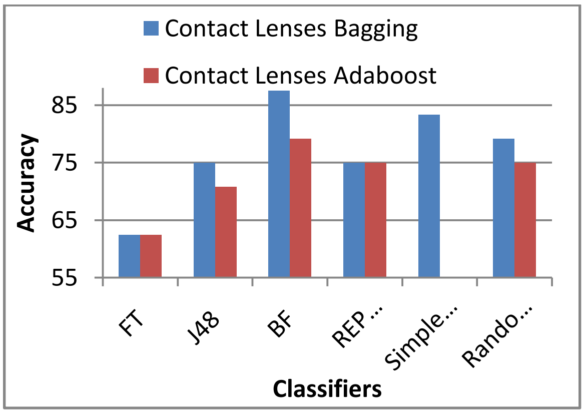

**Figure 5. Accuracy verses classifier graph for contact lense dataset**

**TABLE I Different classifiers applied to medical dataset**

| Classifier | Arrhythmia dataset | | Diabetes dataset | | Heart Dataset | | Contact Lenses | |
|---|---|---|---|---|---|---|---|---|
| | Bagging | Adaboost | Bagging | Adaboost | Bagging | Adaboost | Bagging | Adaboost |
| **FT** | 86.6 | 84.52 | 75.7813 | 71.3542 | 82.8383 | 81.8482 | 62.5 | 62.5 |
| **J48** | 84.52 | 70.57 | 74.6094 | 72.3958 | 78.5479 | 82.1782 | 75 | 70.8333 |
| **BF** | 83.63 | 72.78 | 73.9583 | 72.1354 | 81.51 | 77.8878 | 87.5 | 79.1667 |
| **REP Tree** | 84.22 | 67.92 | 75.2604 | 70.9635 | 83.4983 | 81.1881 | 75 | 75 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Simple cart** | 84.22 | 74.33 | 73.9583 | 70.5729 | 81.8482 | 81.1881 | 83.3333 | .. |
| **Random forest** | 86.011 | 66.59 | 75.651 | 72.0052 | 82.1782 | 81.1881 | 79.1667 | 75 |
| **Average** | 92.011 | 72.785 | 74.869 | 71.571 | 81.736 | 80.913 | 77.083 | 60.41 |

**TABLE II Different classifiers applied on medical dataset for timing**

| **Arrthmia** | **Arrthmia** | | **Diabetes** | | **heartc** | | **contact lenses** | |
|---|---|---|---|---|---|---|---|---|
| | **Bagging** | **Adaboost** | **Bagging** | **Adaboost** | **Bagging** | **Adaboost** | **Bagging** | Adaboost |
| FT | 0.81 | 3.51 | 0.61 | 3.45 | 1.64 | 2.01 | 0.02 | 0 |
| J48 | 0.19 | 3.49 | 0.11 | 0.13 | 0.06 | 0.05 | 0 | 0 |
| BF | 0.28 | 9.26 | 0.34 | 0.44 | 1.15 | 0.28 | 0.02 | 0.02 |
| REP Tree | 0.13 | 0.8 | 0.05 | 0.06 | 0.05 | 0.03 | 0 | 0 |
| Simple cart | 0.3 | 7.17 | 0.33 | 0.39 | 1 | 0.3 | 0..02 | |
| Random forest | 0.22 | 0.55 | 0.31 | 0.5 | 0.22 | 0.08 | 0..02 | 0 |

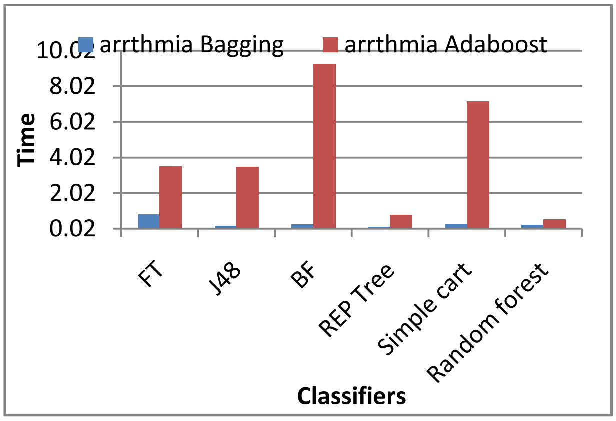

**Figure 6. Accuracy verses classifier graph for diabetes dataset**

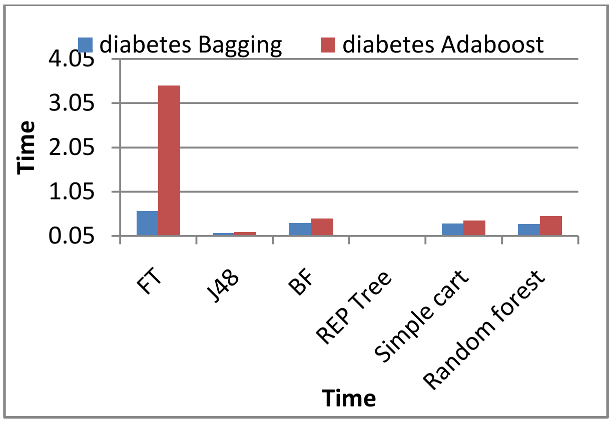

**Figure 7. Accuracy verses classifier graph for diabetes dataset**

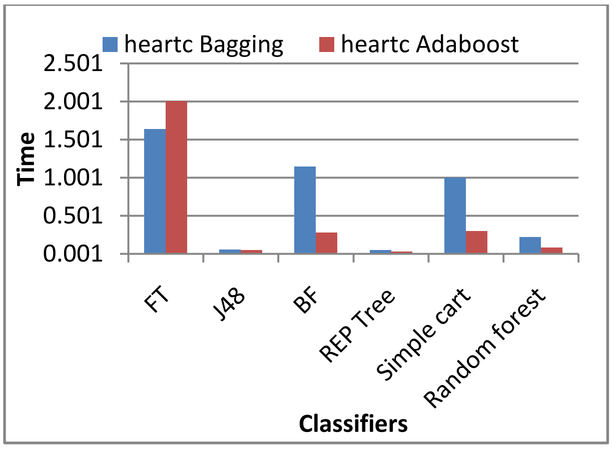

**Figure 8 .Accuracy verses classifier graph for diabetes dataset**

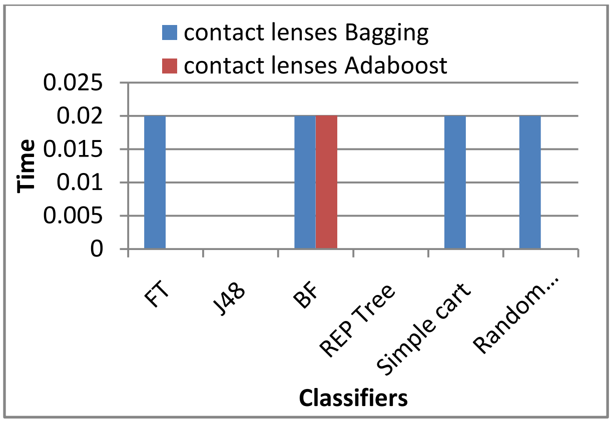

**Figure 9. Accuracy verses classifier graph for diabetes dataset.**

## 5. CONCLUSION

The paper discussed about data mining, and different classification techniques applied on medical database using WEKA tool. For medical diagnosis various data mining techniques are available. In the proposed technique Bagging ensembles and Adaboost ensembles are constructed in WEKA using 10 fold cross validation. The results for bagging show that FT Tree shows good results.

In all if considering average accuracy of all datasets, Arrhythmia dataset shows better accuracy for bagging. Whereas adaboost showsgood accuracy for heart dataset. In future we apply feature selection on classifier before forming the ensemble so that the noisy, irrelevant data should be removed.

## 6. REFERENCES

[1] Payal Dhakate , Suvarna Patil , K. Rajeswari , Dr. V.Vaithiyananthan , Deepa Abin, "Preprocessing and Classification in WEKA using different classifiers", Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 8( Version 1), August 2014, pp.

[2] Remco R. Bouckaert, Eibe Frank, Mark A. HallGeoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, "WEKA—Experiences with a Java Open-Source Project", Journal of Machine Learning Research, November 2010.

[3] Trilok Chand Sharma, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013.

[4] P.Yasodha, M. Kannan, "Analysis of a Population of Diabetic PatientsDatabases in Weka Tool", Research Vol 2, Issue 5, May-2011.

[5] Vikas Chaurasia, Saurabh Pal, "Data Mining Approach to Detect Heart Dieses", International Journal of Advanced Computer Science and Information Technology Vol. 2.

[6] D.Lavanya and Dr.K.Usha Rani, "Ensemble decision tree classifier for breast cancer data"International Journal of Information Technology Convergence and Services, Vol.2, No.1. February 2011.

[7] Prof.K.Rajeswari , Dr.V.Vaithiyanathan and Shailaja V.Pede, "Feature Selection for Classification in Medical Data Mining",International journal of emerging treands and technology in computer science.Vol 2, Issue 2, March – April 2013.

[8] Ren Diao, Fei Chao, Member, IEEE, Taoxin Peng, Neal Snooke, and Qiang Shen, "Feature Selection Inspired Classifier Ensemble Reduction", IEEE TRANSACTIONS ON CYBERNETICS, Vol. 44, NO. 8, AUGUST 2014.

[9] Remco R. Bouckaert, Eibe Frank, Mark A. Hall, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H.Witten," WEKA—Experiences with a Java Open-Source Project" Journal of Machine Learning Research 11 (2010)