Comparison of various classification algorithms on iris datasets using WEKA

Kanu Patel[1], Jay Vala[2], Jaymit Pandya[3]

[1]*Assist. Prof, I.T Department, BVM Engineering College, V.V.Nagar kanu.patel@bvmengineering.ac.in*

[2]*Assist. Prof., I.T Department, GCET Engineering College, V.V.Nagar, jayvala@gcet.ac.in*

[3]*Assist. Prof, I.T Department, GCET Engineering College, V.V.Nagar, jaymitpandya@gcet.ac.in*

Abstract: Classification is one of the most important task of data mining. Main task of data mining is data analysis. For study purpose various algorithm available for classification like decision tree, Navie Bayes, Back propagation, Neural Network, Artificial Neural, Multi-layer perception, Multi class classification, Support vector Machine, k-nearest neighbor etc. In this paper we introduce four algorithms from them. Study purpose we take iris.arff dataset. Implement this all algorithm in iris dataset and compare TP-rate, Fp-rate, Precision, Recall and ROC Curve parameter. Weka is inbuilt tools for data mining. So we used weka for implementation.

Keyword: classification, K-nn, ROC, FP-rate, Decision tree, WEKA

I.   INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining algorithms which carry out the assigning of objects into related classes are called classifiers. Classification algorithms include two main phases; in the first phase they try to find a model for the class attribute as a function of other variables of the datasets, and in the second phase, they apply previously designed model on the new and unseen datasets for determining the related class of each record [1][3]. There are different methods for data classification such as Decision Trees (DT), Rule Based Methods, Logistic Regression (LogR), Linear Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Artificial Neural Networks (ANN), Linear Classifier (LC) and so forth. The comparison of the classifiers and using the most predictive classifier is very important. Each of the classification methods shows different efficacy and accuracy based on the kind of dataset [2]. In addition, there are various evaluation metrics for comparing the classification methods that each of them could be useful depending on the kind of the problem. Among the other criteria for comparing the classification methods, one could mention; precision, recall, error rate, confusion matrix.

In this article, using a new method, five usual data classification methods (Decision tree, Multi-layer perception, Naïve Bayes, C4.5, SVM) have been compared based on the AUC criterion. These mentioned methods have been applied on the random generated datasets which are

independent from a special problem. This comparison is based on the effect of the numbers of existing discrete and continuous attributes and the size of the dataset on the AUC [2].

The rest of the paper is organized as follows: In section 2, preliminary work or previous works related to this area and the motivations of performing the new work have been presented. Section 3 provides an explanation about dataset for classification methods. Reporting the results of the applying classification methods on the datasets is presented in section 4. Section 5 evaluates the results and investigates the efficacy of the classifiers. Finally, section 6 concludes the paper and describes future works.

II.   PRELIMINARY

As our intention is to choose the best algorithms for iris datasets which can be integrated in our weka tool, we have to search among those that can support categorical and numeric data, handle incomplete data, offer a natural interpretation to instructors and be accurate working with small samples. Therefore, we analyses five of the most common machine learning techniques, namely Decision Trees, Bayesian classifiers and Multi-layer perception,C4.5, SVM. We rule out the use of Support Vector Machine techniques because of their lack of a comprehensive visual representation was used to perform the analysis [5].

a.   Classification:

Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy". Popular classification techniques include decision trees and neural networks.

b.   Decision tree

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. This page deals with decision trees in data.

There are many specific decision-tree algorithms. Notable ones include:

*   ID3[6][7] (Iterative Dichotomiser 3)
*   C4.5 (successor of ID3)
*   CART (Classification And Regression Tree)
*   CHAID (CHi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.
*   MARS: extends decision trees to better handle numerical data.

c.   Multi-layer perceptron

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each

node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network.[1][2] MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable

d. SVM

In this section, we study Support Vector Machines, a promising new method for the classification of both linear and nonlinear data. In a nutshell, a support vector machine (or SVM) is an algorithm that works as follows. It uses a nonlinear mapping to transform the original training data into a higher dimension [8]. Within this new dimension, it searches for the linear optimal separating hyper plane (that is, a "decision boundary "separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane. The SVM finds this hyper plane using support vectors ("essential" training tuples) and margins (defined by the support vectors).We will delve more into these new concepts further below.

e. Bayesian classifiers

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem, described below. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given classis independent of the values of the other attributes [8]. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naïve." Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes. Bayesian belief networks can also be used for classification.

f. C4.5

C4.5 is one of the decision tree based algorithm. It's a higher version of ID3. We implemented this all five algorithm in to Weka and compare their results.

g. WEKA Tools

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state of- the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules; It also includes visualization tools. The new machine learning schemes can also be developed with this package. WEKA is open source software issued under General Public License [4].

The data file normally used by Weka is in ARFF file format, which consists of special tags to indicate different things in the data file (foremost: attribute names, attribute types, and attribute values and the data). The main interface in Weka is the Explorer. It has a set of panels, each of which can be used to perform a certain task. Once a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis

III.   DATASET

In this paper for comparison of all classification we used "iris.arff" dataset the basic information of iris is given below.

Relevant Information: This is perhaps the best known database to be found in the pattern recognition literature

--- Predicted attribute: class of iris plant.

--- This is an exceedingly simple domain.

Number of Instances: 150 (50 in each of three classes)

Number of Attributes: 4 numeric, predictive attributes and the class

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
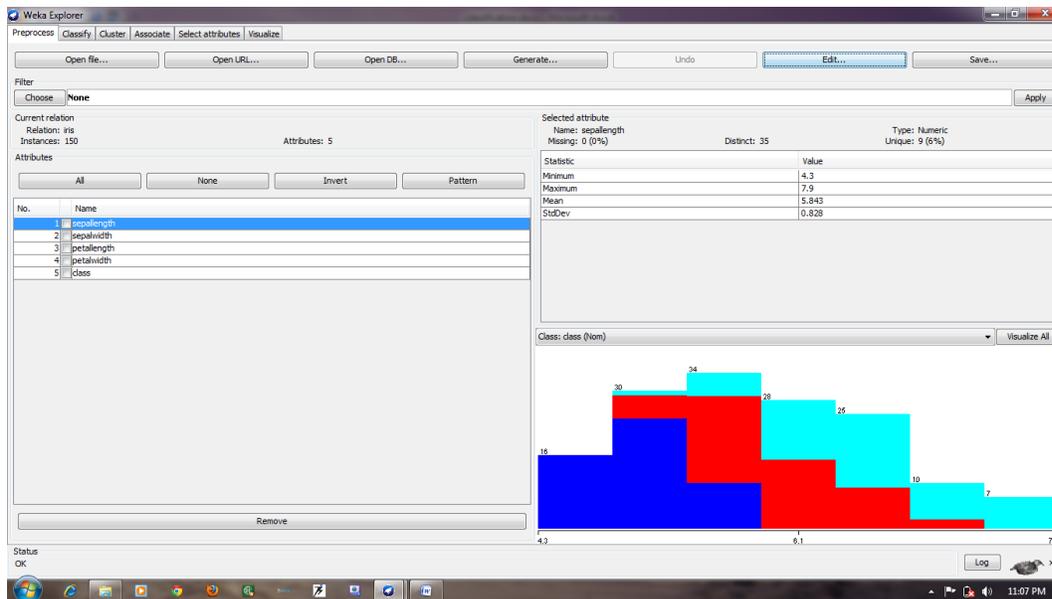4. petal width in cm
5. class:
   -- Iris Setosa
   -- Iris Versicolour
   -- Iris Virginica

Missing Attribute Values: None


IV.   IMPLEMENTATION

For implementation, we used weka tools. First of all open weka tools and select dataset in this paper we used iris dataset and it's inbuilt of weka. First look of weka look like.

Above figure show the basic information of dataset. like list of attribute and class variable. A data set contains value of all attribute. We can change that value using weka.
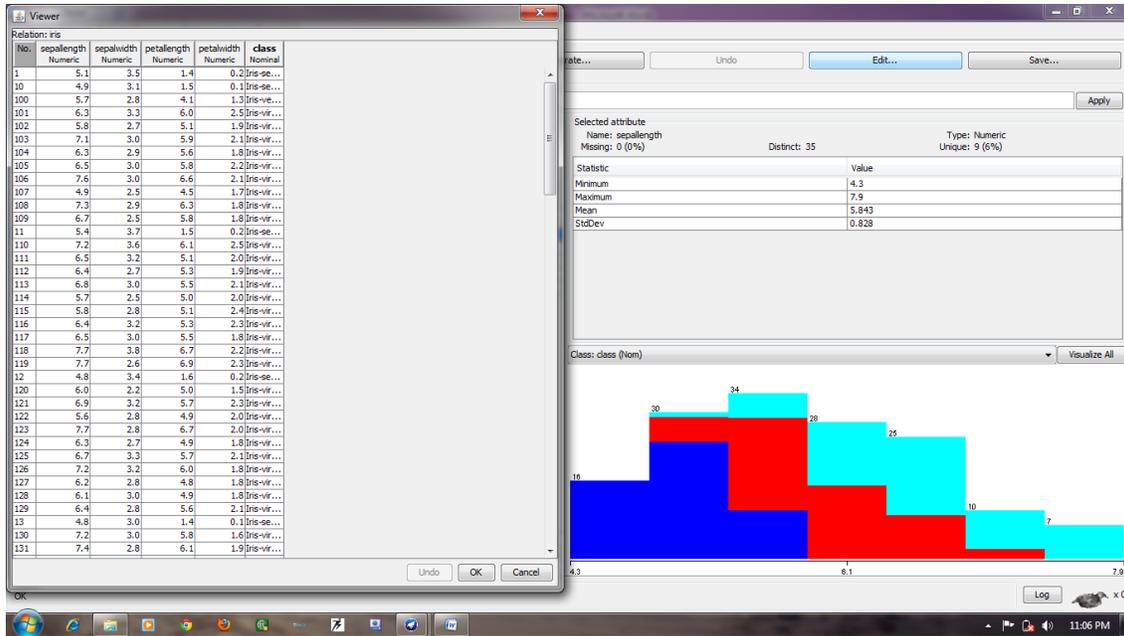
Above figure gives the value of iris dataset of every attribute. After preprocessing we directly go for classification select classifier from classification algorithm and click on start button. weka run that selected algorithm on iris dataset. its display output regarding all parameter like error rate and confusion matrix. In this paper we implemented four algorithm Decision tree, multiclass, naive byes and multi layer perceptron.
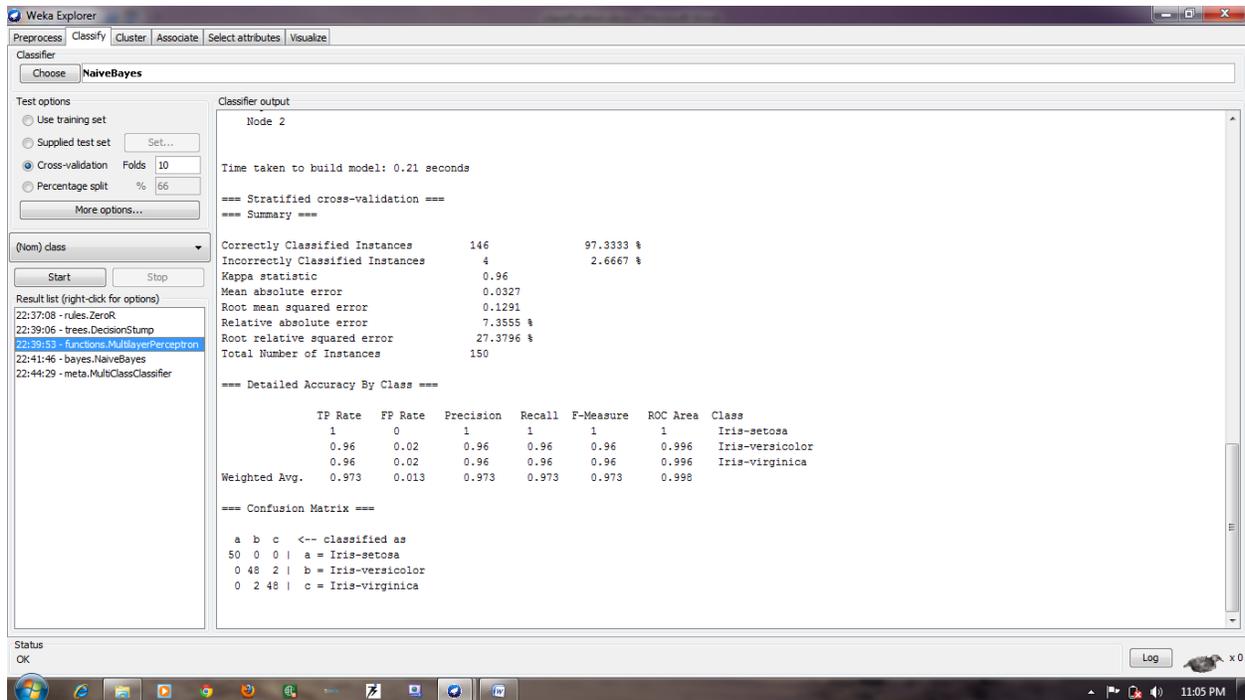
At last confusion matrix of multi layer perceptron algorithm is.

```
=== Confusion Matrix ===
a    b    c        <-- classified as
50   0    0        | a = Iris-setosa
0    48   2        | b = Iris-versicolor
0    2    48       | c = Iris-virginica
```

V.   EXPERIMENTAL RESULT

We tested the four aforementioned algorithms using different parameter settings and different numbers of folds for cross validation, in order to discover whether they have a great effecton the result. Finally, we set the algorithms with default parameters and used 10-foldcross validation we show the classification accuracy and rates obtained with the four algorithms for the iris dataset.

	TP-rate	FP-Rate	Precision	Recall	ROC Area
DecisionStump	0.667	0.167	0.5	0.667	0.833
MultilayerPerceptron	0.973	0.013	0.973	0.973	0.998
NaiveBayes	0.96	0.02	0.96	0.96	0.994
MultiClassClassifier	0.96	0.02	0.96	0.96	0.977

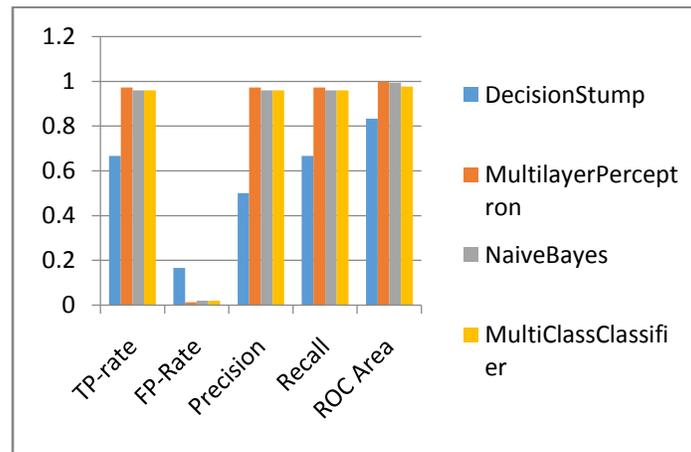*Table 1 Comparison of all four algorithms*

Fig. 1 Comparison of all four algorithms in chart form

VI.   CONCLUSION

In this paper, We compare four algorithm on iris dataset with some parameter. In iris dataset contain simple and class attribute. Decision tree will implement on iris dataset then its less efficient than all other like multilayer perceptron, Naive Bayes and Multiclass classifier. In all algorithm multilayer perceptron is more accurate and efficient in all parameter like TP-rate, FP-rate, Precision, Recall and ROC area.

VII.   REFERENCES

[1].P-N. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining," Addison-Wesley Publishing, 2006.

[2] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms," John Wiley & Sons Publishing, 2003.

[3] I.H. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann Publishing, Second Edition, 2005.

[4] Y.S. Kim, "Comparision of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size," Journal of Expert Systems with Application, Elsevier, 2008, pp. 1227-1234.

[4]. WEKA at http://www.cs.waikato.ac.nz/~ml/weka.

[6] J.S R Jang (1993). ANFIS Adaptive Network Based Fuzzy inference System. IEEE transaction on Systems, Man and Cybernetics. Vol. 23, no3, pp 665-685

[7] Mansour Y (1997). Pessimistic decision tree pruning based on tree size. *In Press of Proc. 14th International Conference on Machine Learning.* Pp.195-201.

[8] Basic of Classification in data mining source from: Wikipedia