Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.389 – 393

RESEARCH ARTICLE

COMPARISON OF DIFFERENT DATASETS USING VARIOUS CLASSIFICATION TECHNIQUES WITH WEKA

Deepali Kharche[1], K. Rajeswari[2], Deepa Abin[3]

[1]ME Student, Pimpri Chinchwad College of Engineering, Nigdi, Pune, India

[2]Associate Professor, PCCOE, Pune & Ph.D Research Scholar, SASTRA University, Tanjore, India

[3]Assistant Professor, PCCOE, Pune

1 Kharche.deepali@gmail.com; [2] raji.pccoe@gmail.com; [3] deepaabin@gmail.com

Abstract-- Data Mining refers to mining or extracting knowledge from huge volume of data. Classification is used to classify each item in set of data into one of the predefined set of classes. In data mining, an important technique is classification, generally used in broad applications, which classifies various kinds of data. In this paper, different datasets from University of California, Irvine (UCI) are compared with different classification techniques. Each technique has been evaluated with respect to accuracy and execution time and performance evaluation has been carried out with J48, Simple CART (Classification and Regression Testing), and BayesNet and NaiveBayesUpdatable Classification algorithm.

Keywords: Data Mining; Classification; J48; Simple CART; BayesNet; NaiveBayesUpdatable; WEKA

I. INTRODUCTION

Data Mining is knowledge mining from data, knowledge extraction, and data analysis. Data Mining involves the various data analysis tools for identifying previously unknown, valid patterns and relationships in huge data set. The term Data Mining, also known as Knowledge Discovery in Databases (KDD) is the process of discovering interesting patterns and knowledge from large amount of data. The data sources can include databases, data warehouses, web etc.[1].There are number of applications for machine Learning (ML), the most significant of which is data mining. There are different data mining techniques like classification, association, preprocessing, transformation, clustering, and pattern evaluation. Classification and Association are the popular techniques used to predict user interest and relationship between those data items which has been used by users [9, 10].

II. LITERATURE REVIEW

2.1 J48:

J48 [3] is an optimized implementation and improved version of C 4.5. J48 builds decision trees from a set training data using the Information entropy. J48 examines the normalized to information gain that results from choosing an attribute for splitting the data. It uses the facts that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. To make a decision, the attribute with the highest normalized information gain is used. The output of the J48 is in the form of Decision tree. Decision tree is has a tree structure with different nodes, like root node, intermediate nodes and leaf nodes. Each node in the tree contains the decision that forms a Decision Tree. Splitting criteria has been used in decision Tree, to calculate which attribute is the best split for the portion tree of training data.

2.2 CART (Classification And Regression Tree):

Classification and regression trees (CART) decision tree [3] is a learning technique, which gives the results as either classification or regression trees, depending on categorical or numeric data set. It also describes the generation of binary decision tree because CART generates only two branches at each node.

2.3 BayesNet:

BayesNet [8] Classifier depends on the Bayes theorem. In BayesNet classifier conditional probability of each node is calculated first and then Bayesian Network is formed. Bayesian Networks is a directed acyclic graph. One of the assumption made in BayesNet is, that all attributes are nominal and there are no missing values, if any such values are present then they replaced globally.

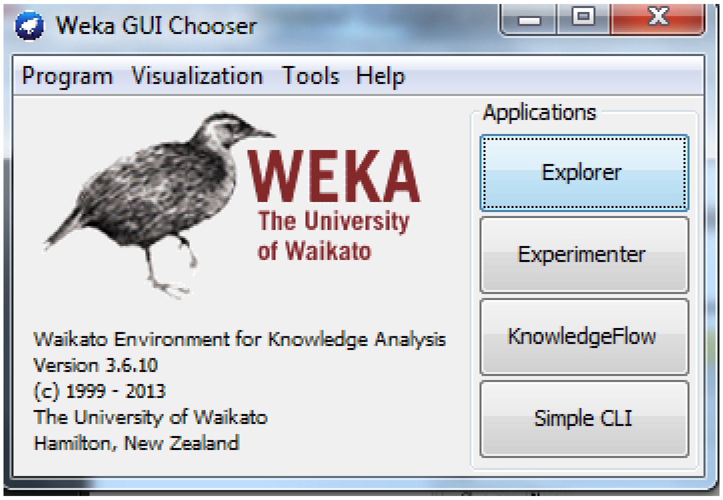
2.4 NaiveBayesUpdatable:

The NaiveBayesUpdatable [3] is the improved version of NaiveBayes. The default precision is used by this classifier when build Classifier is called with zero training instances of 0.1, for numeric attributes, and hence it called as incremental update.

III. METHODOLOGY

The open Source Application – WEKA [2] (Waikato Environment for knowledge learning) is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. A large different number of classifiers are used in WEKA such as Bayes, function, tree etc. The Classify tab in WEKA explorer as shown in figure 3.1 can be used to perform classification of the given dataset. Data preprocessing, classification, clustering, association, regression, feature selection, etc such techniques in data mining that are supported by WEKA.

Steps to apply different classification techniques on data set and obtain result in Weka:
Step 1: Accept the input dataset and preprocessed.
Step 2: Apply the classifier algorithm on the whole data set in Classification.
Step 3: Note the accuracy given by it and time required for execution. Also check the confusion matrix.
Step 4: For comparison of different classification algorithms on different datasets repeat step 2nd and 3rd with respect to accuracy and execution time.
Step 5: Compare the different accuracy results provided by the dataset with different classification algorithms and identify the significant classification algorithm for particular dataset.



Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Fig: 3.1. WEKA Startup Window

IV. RESULTS AND DISCUSSION

In our work comparison of different datasets using different classifiers, based on the accuracy and execution time required for analyzed and discussed. Accuracy is defined as the number of instances classified correctly. The Accuracy of a classifier on a given set is the percentage of test set tuples that are correctly classified by the classifier.

$$\text{Accuracy} = \frac{TP + TN}{P + N} \qquad \text{Eq. (4.1)}$$

Eq.(4.1) gives the formula for calculating the Accuracy, where TP is the True Positive Rate; TN is the True Negative Rate ,P and N are the Positive and Negative values respectively.

It is observed from Table 4.1 that, BayesNet Performed well with Soyabean Dataset, Simple CART performed well with Nursery Dataset (UCI Dataset) but takes some more time than other datasets used to build the model, J48 as well as Simple CART also performed well with Mushroom Dataset (UCI Dataset), J48 and NaiveBayesUpdatable performed well with Iris Dataset and NaiveBayesUpdatable performed fine with Labour dataset (UCI Dataset).
Similarly, Table 4.2 shows that, for Labour and Iris dataset J48 and NaiveBayesUpdatable performs well while Iris dataset also performs well with BayesNet method. Mushroom and Nursery dataset works well with NaiveBayesUpdatable. Soyabean dataset works well with J48 classification method in respect to execution time.

	Soyabean	**Nursery**	**Mushroom**	**Iris**	**Labour**
Name of the classifier	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**
J48	91.5081	97.0525	100	96	78.6842
Simple CART	91.0688	99.5756	99.9385	95.3333	78.9424
BayesNet	93.2651	90.3318	96.2211	92.6667	87.7193
NaiveBayesUpdatable	92.9722	90.3241	95.8272	96	89.4731

Table: 4.1 Comparison of Accuracy for various datasets.

Name of the classifier	**Soyabean**	**Nursery**	**Mushroom**	**Iris**	**Labour**
	Time	**Time**	**Time**	**Time**	**Time**
J48	0.05 Sec	0.09 Sec	0.08 Sec	0.00 Sec	0.00 Sec
Simple CART	2.04 Sec	8.11 Sec	8.08 Sec	0.02 Sec	0.06 Sec
BayesNet	0.09 Sec	0.06 Sec	0.19 Sec	0.00 Sec	0.02 Sec
NaiveBayesUpdatable	0.30 Sec	0.02 Sec	0.03 Sec	0.00 Sec	0.00 Sec

Table:4.2 Comparison of Time for various classifiers

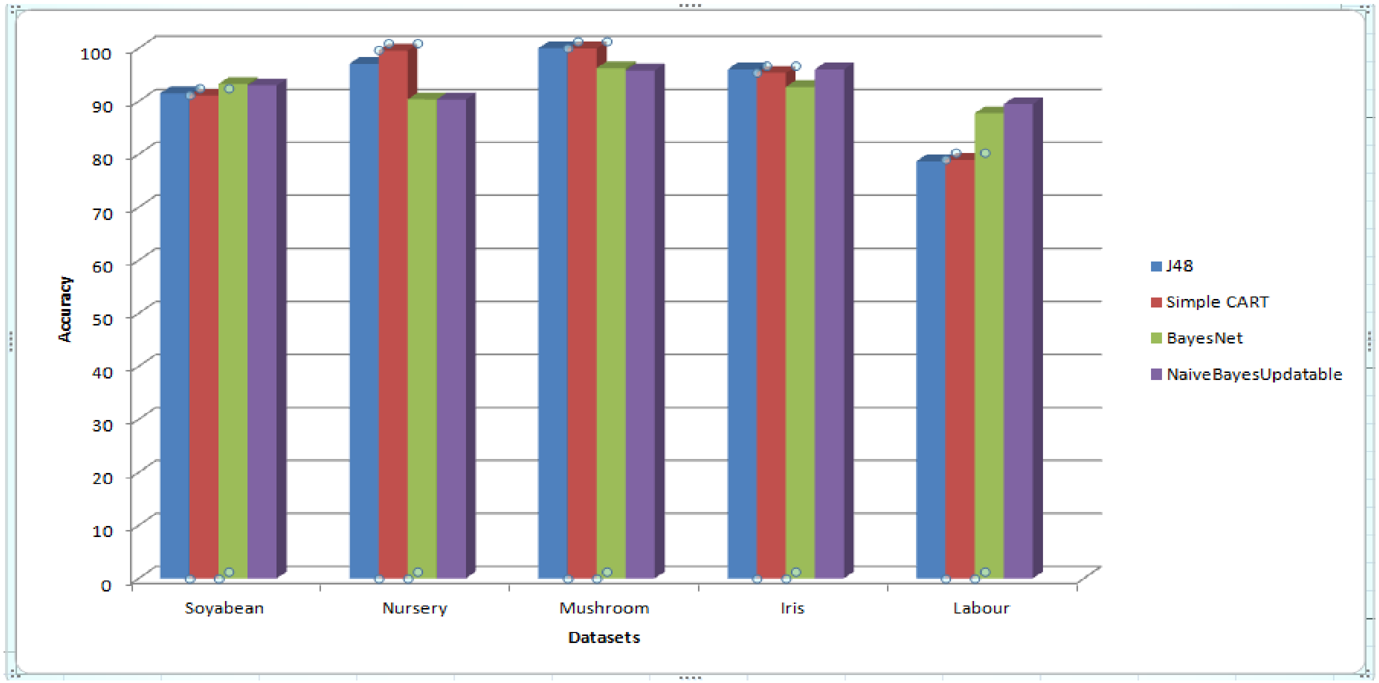


Fig. 4.1 Graphical Comparison of Different dataset using different classification techniques.

Figure 4.1 shows the graphical comparison of accuracy with different datasets using different classification techniques.

V. CONCLUSION AND FUTURE WORK

In this paper we have compared the performance of the datasets using the different Classification techniques with evaluation criteria as accuracy and execution time. It is observed that performance of classification techniques varies with different datasets. Factors that affect the classifier's performance are 1. Dataset 2. Number of instances and attributes, 3. Type of attributes. J48 and NaiveBayesUpdatable have gives better results with other data sets used in comparison.

Our future work will focus on the combination of classification techniques that can be used to improve the performance.

ACKNOWLEDGMENT

This work was supported by Computer Engineering Department in Pimpri Chinchwad College of Engineering, Nigdi, Pune

REFERENCES

[1] J. Han and M. Kamber, (2000) "*Data Mining: Concepts and Techniques*"

[2] Weka: Data Mining Software in Java *http://www.cs.waikato.ac.nz/ml/weka/*

[3] Ian H.Witten and Elbe Frank, (2005) *"Data mining Practical Machine Learning Tools and Techniques,"* Second Edition, San Fransisco.

[4] www.ics.uci.edu/~mlearn

[5] Zak S.H., *"Systems and Control"* NY: Oxford Uniniversity Press.

[6] Hassoun M.H, *"Fundamentals of Artificial Neural Networks"*, Cambridge, MA: MIT press.

[7] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, *"Data Clustering Method for Discovering Clusters in Spatial Cancer Databases"*, International Journal of Computer Applications (0975 – 8887)Volume 10– No.6.

[8] Yugal kumar and G. Sahoo *"Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in DataMining using WEKA"*

[9] K. Rajeswari, Dr. V. Vaithiyannathan *"Mining Association Rules Using Hash Table"*, International Journal of Computer Applications (9132-3320).

[10] K. Rajeswari, Dr. V. Vaithiyannathan *"Heart Disease Diagnosis: An Efficient Decision Support System Based on Fuzzy Logic and Genetic Algorithm"*, International Journal of Decision Sciences, Risk and Management by Inderscience Publications. ISSN: 1753-7169.