COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES USING DIFFERENT DATASETS

V. Vaithiyanathan[1], K. Rajeswari[2], Kapil Tajane[3], Rahul Pitale[3]

[1]Associate Dean Research, CTS Chair Professor, SASTRA University, Tanjore, India
[2]Associate Prof., PCCOE Pune & Ph.D Research Scholar, SASTRA Univ., Tanjore, India
[3]ME Student, Pimpri Chinchwad College of Engineering, Nigdi, Pune, India

ABSTRACT

In this paper different classification techniques of Data Mining are compared using diverse datasets from University of California, Irvine(UCI). Accuracy and time required for execution by each technique is observed. The Data Mining refers to extracting or mining knowledge from huge volume of data. Classification is an important data mining technique with broad applications. It classifies data of various kinds. Classification is used in every field of our life. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. This work has been carried out to make a performance evaluation of J48, MultilayerPerceptron, NaiveBayesUpdatable, and BayesNet classification algorithm. Naive Bayes algorithm is based on probability and j48 algorithm is based on decision tree. The paper sets out to make comparative evaluation of classifiers J48, MultilayerPerceptron, NaiveBayesUpdatable, and BayesNet in the context of Labour, Soyabean and Weather datasets. The experiments are carried out using weka 3.6 of Waikato University. The results in the paper demonstrate that the efficiency of j48 and Naive bayes is good.

KEYWORDS: *Data Mining, Classification, j48, Multilayer Perceptron, NaiveBayes Updatabl ,BayesNet.*

I. INTRODUCTION

Data mining involves the use of various sophisticated data analysis tools for discovering previously unknown, valid patterns and relationships in huge data set. These tools are nothing but the machine learning methods, statistical models and mathematical algorithm. Data mining consists of more than collection and managing the data, it also includes analysis and prediction. Classification technique in data mining is capable of processing a wider variety of data than regression and is growing in popularity. There are number of applicationsfor Machine Learning (ML), the most significant of which is data mining. The term Data Mining, also known as Knowledge Discovery in Databases (KDD) refers to the nontrivial extraction of implicit, potentially useful and previously unknown information from data in databases [1]. There are several data mining techniques are preprocessing, association, classification, pattern recognition and clustering, [7]. 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]. Classification methods includes Bayesian network,J48 Decision tree, Neural Network etc. Particularly this work is concerned with classification techniques.
The rest of the paper is organized as follows. Section II covers literature review. In Section III covers methodology. Finally in section IV, we summarize the comparative results.

II. LITERATURE REVIEW

J48: J48 a can be called as optimized implementation of the C4.5 or improved version of the C4.5. The output given by J48 is the Decision tree. A Decision tree is same as that of the tree structure

having different nodes, such as root node, intermediate nodes and leaf node. Each node in the tree contains a decision and that decision leads to our result as name is decision tree. Decision tree divide the input space of a data set into mutually exclusive areas, where each area having a label, a value or an action to describe or elaborate its data points. Splitting criterion is used in decision tree to calculate which attribute is the best to split that portion tree of the training datathat reaches a particular node.

MultilayerPerceptron:

Multi Layer Perceptron can be defined as Neural Network and Artificial intelligence without qualification. A Multi Layer perceptron (MLP) is a feedforward neural network with one or more layers between input and output layer. Basically there are three layers: input layer, hidden layer and output layer. Hidden layer may be more than one. Each neuron (node) in each layer is connected to every neuron (node) in the adjacent layers. The training or testing vectors are connected to the input layer, and further processed by the hidden and output layers. A detailed analysis and study of multi-layer perceptrons has been described by Żak[5] and by Hassoun [6].

BayesNet:

BayesNet classifier is based on the bayes theorm. So, in BayesNet classifier conditional probability on each node is calculated first and then a Bayesian Network get formed. Bayesian Network is nothing but a directed acyclic graph. The assumption made in BayesNet is, that all attributes are nominal and there are no missing values any such value replaced globally. Hill Climbing, Tabu Search, Simulated Annealing, Genetic Algorithm and K2 such a different types of algorithms are used to estimate conditional probability in BayesNet. In BayesNet, the output of can be visualized in terms of graph.

NaiveBayesUpdatable:

The name NaiveBayesUpdatable itself suggest that it is the updatable or improved version of NaiveBayes. A default precision used by this classifier when buildClassifier is called with zero training instances is of 0.1 for numeric attributes and hence it also known as incremental update.

III. METHODOLOGY

WEKA

The full form of WEKA: Waikato Environment for Knowledge Learning. Weka is a computer program that was developed by the student of the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains [2]. Data preprocessing, classification, clustering, association, regression and feature selection these standard data mining tasks are supported by Weka. It is an open source application which is freely available.
In Weka datasets should be formatted to the ARFF format. The Weka Explorer will use these automatically if it does not recognize a given file as an ARFF file. Classify tab in Weka Explorer is used for the classification purpose. A large different number of classifiers are used in weka such as bayes, function, tree etc.

Steps to apply classification techniques on data set and get result in Weka:

Step 1: Take the input dataset.
Step 2: Apply the classifier algorithm on the whole data set.
Step 3: Note the accuracy given by it and time required for execution.
Step 4: Repeat step 2 and 3 for different classification algorithms on different datasets.
Step 5: Compare the different accuracy provided by the dataset with different classification algorithms and identify the significant classification algorithm for particular dataset.
The experiments are conducted in a system with configuration Intel Pentium Processor P6100, 2 GB DDR3 Memory and 500 GB HDD. Experiments are conducted 3 times and an average accuracy and time is recorded.

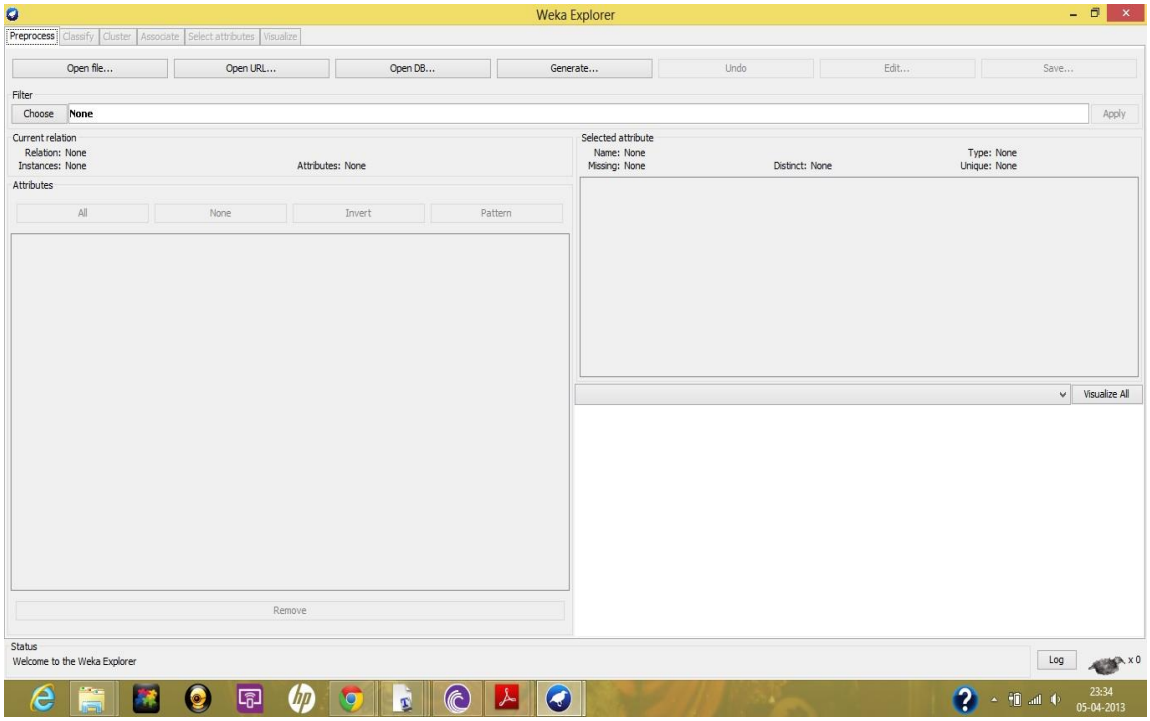

Figure:3.1 Weka Explorer**.**

IV. RESULTS AND DISCUSSION

A comparison of classifiers for different datasets based on the accuracy and time taken for execution is made. Accuracy is defined as the no of instances classified correctly. It is observed from table 4.1 that, NaiveBayesUpdatable performed well with Labour dataset, Multilayer Perceptron out performed with Soybean dataset and Weather dataset in terms of correctly classified instances.

Table:4.1 Comparison of Accuracy and Time taken for various classifiers

Name of the Classifier	Labour		Soybean		Weather	
	Accuracy	**Time**	**Accuracy**	**Time**	**Accuracy**	**Time**
J48	73.6842	0.02 sec	91.5	0.02sec	64.28	0 sec
Multilayer Perceptron	85.96	0.44 sec	93.41	55.88sec	78.57	0.02sec
BayesNet	87.71	0.02sec	93.26	0.02sec	68.28	0sec
NaiveBayesUpdatable	89.4	0sec	92.97	0sec	64.28	0sec

Table:4.2 Comparison of Accuracy of classifiers for various datasets.

Name of the Classifier	Labour	Soybean	Weather
J48	73.6842	91.5	64.28
Multilayer Perceptron	85.96	93.41	78.57
BayesNet	87.71	93.26	68.28
NaiveBayesUpdatable	89.4	92.97	64.28

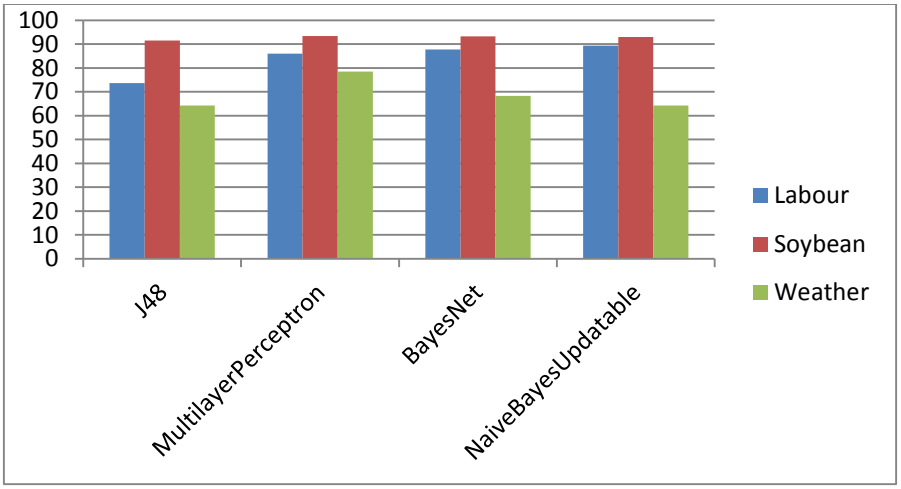

Figure:4.1 shows the graphical view of accuracy for various classifiers on different datasets**.**

V. CONCLUSION AND FUTURE WORK

Thus in this paper we have compared the performance of various classifiers. Three data sets from benchmark data set (UCI) is used for experimentation. It is found that the performance of classification techniques varies with different data sets. Factors that affect the classifier's performance are 1. Data set, 2. Number of tuples and attributes, 3. Type of attributes, 4. System configuration. Multilayer Perceptron outperformed with two datasets and NaivesBayesUpdatable has given good results with a data set.

Our future work will focus on improvement of Classification Technique thereby improving the efficiency of classification in a decreased time. Also a combination of classification techniques will be used to improve the performance.

REFERENCES

[1]. J. Han and M. Kamber, (2000) "Data Mining: Concepts and Techniques," Morgan Kaufmann.
[2]. Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/
[3]. Ian H.Witten and Elbe Frank, (2005) "Datamining Practical Machine Learning Tools and Techniques," Second Edition, Morgan Kaufmann, San Fransisco.
[4]. www.ics.uci.edu/~mlearn
[5]. Zak S.H., (2003), "Systems and Control" NY: Oxford Uninversity Press.
[6]. Hassoun M.H, (1999), "Fundamentals of Artificial Neural Networks", Cambridge, MA: MIT press.
[7]. Ritu Chauhan, Harleen Kaur, M.Afshar Alam, (2010) "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,(2012)" Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in DataMining using WEKA"
[9]. K. Rajeswari, Dr. V. Vaithiyannathan,(2012) "Mining Association Rules Using Hash Table", International Journal of Computer Applications(9132-3320).
[10]. Rajeswari, Dr. V. Vaithiyannathan,(2011), " 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.

AUTHORS BIOGRAPHY

V. Vaithiyanathan is a Professor, Associate Dean Research and CTS Chair Professor, SASTRA University. His areas of interest are Data Mining, Image Processing. He is associated with many funded projects. He has published many papers in reputed International journals and conferences.

K. Rajeswari has received her B.E and M.Tech in Computer Science & Engineering. She has published several papers in Data Mining. She is winner of Cambridge International Certification for Teachers with Distinction. She has about 15 years of Teaching experience. Currently she is working as an Associate Professor in –Computer Engg. in PCCOE, Pune. She is pursuing Ph.D in SASTRA University, Tanjore, India.

Kapil Tajane received Bachelor degree in Computer Science & Engg from Amravati University. Currently he is pursuing Master of Computer Engg from PCCOE, Pune University.

Rahul Pitale received Bachelor degree in Computer Science & Engg from Amravati University. Currently he is pursuing Master of Computer Engg from PCCOE, Pune University.