ISSN No. 0976-5697

Volume 5, No. 4, April 2014 (Special Issue)

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Performance Investigation of different Classification Algorithms Using Waikato Environment for Knowledge Analysis

G.Victor Daniel[1]
[1]Dept.Of Information Technology, GITAM University, Hyderabad, India
victordaniel.gera@gitam.edu

G.Sunny Demoli[2]
[2]Dept.of CSE, EITW, Hyderabad, India.
blessy.sunny@gmail.com

P.Prathyusha[3]
[3]Dept of ECE, Vignan Bharati Instite of Technology, Ghatkesar, Hyderabad, India
prathyusha.pedapalli@@gmail.com.

K.Suresh Kumar[4]
[4]Dept.Of Information Technology, GITAM University, Hyderabad, India

Abstract : Data mining refers to the task of extracting knowledge or hidden interesting patterns from the large volumes of data. Classification is one of the data mining functionalities which refer to the process of finding a model to predict the class label of objects whose class is unknown. This paper analyze the four major classification algorithms such as Naïve Bayes classifier, ADTree classifier, PART Rule based classifier and Kstar classifier using Waikato Environment for Knowledge Analysis or in short, WEKA. The aim of this paper is to investigate the performance of the classification algorithms on the aspect of correctly classified instances. The data 'vote data' with a total data of 7395 and a dimension of 435 rows and 17 columns are used to experiment and to rationalize different classification algorithms. The performance of these four algorithms are presented and compared.

Keywords: Clsssification Algorithms; Weka; Performance Evaluation Of Machine Learning Algorithms

I. INTRODUCTION

Knowledge discovery is a process which consists of an iterative sequence of steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining functionalities are concept/class description, mining frequent patterns, associations and correlations, classification and prediction, cluster analysis, outlier analysis and evolution analysis [1]. Considerable data mining techniques are classification and clustering, regression. Classification predicts categorical (discrete, unordered) labels. Classification is considered as two-step process. The first step is, learning step (or training phase), in which a classifier or a model is built based on the training data. This model is used to predict the class of objects whose class label is unknown. In the second step, accuracy of this classifier is predicted. In order to avoid data 'overfit' problem, training data would not be used as testing data, instead a separate set of test tuples along with their associated class labels are used. Considerable amount of research is being done on classification algorithms, with much importance given to accuracy of a classifier. The accuracy of a classifier can be measured in terms of number of correctly classified instances.

II. METHODS

A. *Naïve Bayesian Classification:*

Bayesian classifiers predict the probability that a given tuple belongs to a specific class[7]. Let there be set of tuples along with their labels. Assume that there are 'i' classes, i>=1.The task of the classifier is to predict the class of the given tuple. The classifier predicts that the given tuple belongs to the class with maximum posterior probability.

Simple Bayesian classifier works based on bayes' theorem, given by the equation

$$P(K_i|X) = \frac{P(X|K_i)\,P(K_i)}{P(X)}$$

Let $P(K_i)$ represent the class with highest posterior probability (Ki) may be estimated by $|K_{i,D}| / |D|$,
$K_{i,D}$ = number of training tuples of class ki in D.
$P(X|k_i)$ is estimated by formula:

$$P(x_1|K_i) * P(x_2|K_i) *\ldots\ldots\ldots *P(X_n|K_i),$$

Where $P(x_1|K_i) * P(x_2|K_i) *\ldots\ldots\ldots *P(X_n|K_i)$ can be estimated from the training tuples. Bayes Network learning using various searches algorithms and quality measures. A Bayesian network is a probabilistic model which represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG).

B. *Alternating Decision Tree Classifier:*

Decision tree induction builds decision trees from the training data set. A decision tree consists of internal nodes and external nodes. Internal node is a test conducting unit and external node is corresponding to the class label. For a given a tuple, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class label for that tuple. In order to find the associated class label of a given tuple, field values of the tuple are verified against the nodes of the decision tree, which leads to the class label at the leaf node. Tree pruning tries to recognize and eliminate the branches of the decision tree which are outliers. ADTree generates an alternating decision tree [3].

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014

C. Rule Based Classification:

Rule-based classifiers, represents the classifier in terms of rules. It uses a set of rules for classification. A rule can be estimated by coverage and accuracy. A rule's coverage is the number of tuples covered by the rule, where as a rule's accuracy, is percentage of tuples that the rule can classify correctly. Rules are arranged according to their priority list. Rules generated by the rule based classifier are easy to understand when compared to decision tree especially when the tree generated is very large. A rule is generated corresponding to every path from root to leaf node in the tree, where leaf node is labeled with the class label. Rules accuracy can be estimated by comparing them with the tuples in the training set. Rules that do not improve the accuracy can be removed. PART rule classifier uses separate-and-conquer. Builds partial decision tree at every step and considers the best leaf as the rule [4].

D. Lazy Learner: Kstar:

Classification methods can be categorized as Eager learners and Lazy learners. Eager learners build the classifier as soon as they receive training tuples. They don't wait for any test tuples, where as Lazy learners wait simply preserving the given train tuples, till they receive tuples to classify. K-Nearest-Neighbor classifier is an example of lazy learners, which uses efficient storage structures. Nearest-Neighbor classifier learns by analogy, by comparing testing tuples with training tuples that are analogous to it. Kstar is an instance-based classifier, class label of the test tuples instance is determined based on the similarity of the class label of the training instance and it uses entropy base distance function[5].

III. THE DATA

The data used in this investigation is the voter data. It has a total of 7395 data and a dimension of 435 rows and 17 columns. Percentage split of training and testing data is taken as 66% and 44%. Test data is used to test the accuracy of the classifier.

IV. WEKA

University of Waikato in New Zealand developed a data mining tool called WEKA which is a short form for Waikato Environment for Knowledge Analysis [6]. Several data mining algorithms are implemented in WEKA using java language. WEKA provides a facility to apply data mining techniques to the real-world data. This tool makes machine learning tools readily available, also new machine learning algorithms can be developed using WEKA. This package is publicly available and offers collection of algorithms. It uses ARFF (Attribute-Relation File Format) or XRFF (Xml attribute Relation File Format) Data Formats. Initially, data is to be loaded into the tool .Once the data is loaded, then preprocessing techniques and also classification techniques can be applied on it[6].

V. RESULTS

The following algorithms namely Naïve Bayesian Classification, Alternating Decision tree, PART rule learner and Kstar classifiers are compared and investigated in terms of accuracy. Details such as time taken to build the model, correctly classified instances, incorrectly classified instances, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error, confusion matrices are discussed.

Table 1: Simulation Results of Algorithms

Algorithm (Total Number of Instances :148)	**% of Correctly classified Instances (No of Tuples)**	**% of Incorrectly Classified Instances (No of Tuples)**	**Time taken to build model (in seconds)**
NaiveBayes	91.2162 (135)	8.7838 (13)	0.09
ADTree	97.2973 (144)	2.7027 (4)	0.06
PART	96.6216 (143)	3.3784 (5)	0.06
Kstar	93.2432 (138)	6.7568 (10)	0

Table. 2 Training and Simulation Errors

Algorithm (Total Number of Instances :148)	**Mean absolute error**	**Root mean squared error**	**Relative absolute error**	**Root relative squared error**
NaiveBayes	0.0912	0.2858	19.04%	57.6557%
ADTree	0.0565	0.1401	11.79%	28.25%
PART	0.0686	0.1811	14.31%	36.54%
Kstar	0.0722	0.2053	15.08%	41.41%

Table.3 Confusion Matrix for Naivebayes

a	b	←Classified as
75	11	a
2	60	b

Table.4 Confusion Matrix for ADTree

a	**b**	**←Classified as**
84	2	**a**
2	60	**b**

Table.5 Confusion Matrix for ADTree

a	**b**	**←Classified as**
83	3	**a**
2	60	**b**

Table.6 Confusion Matrix for Kstar

a	**b**	**←Classified as**
77	9	**a**
1	61	**b**

The confusion matrix is a useful tool which is used to analyze classifier's ability to recognize tuples of different classes. Confusion matrix can be represented with the following matrix.

Table.7 Confusion Matrix

	C1	**C2**
C1	True positives	False negatives
C2	False positives	True negatives

True positives denote the positive tuples that are correctly classified by the classifier; True negatives are the negative tuples that are correctly denoted by the classifier. False positives are the negative tuples that are incorrectly labeled .false negatives are the positive tuples that are incorrectly identified.

Loss functions like absolute error, squared error measure the error between the actual value and predicted value. Let p_i and p_i' represents actual value and predicted value then absolute error can be calculated by formula $|p_i-p_i'|$, similarly squared error can be calculated by formula $(p_i-p_i')^2$.

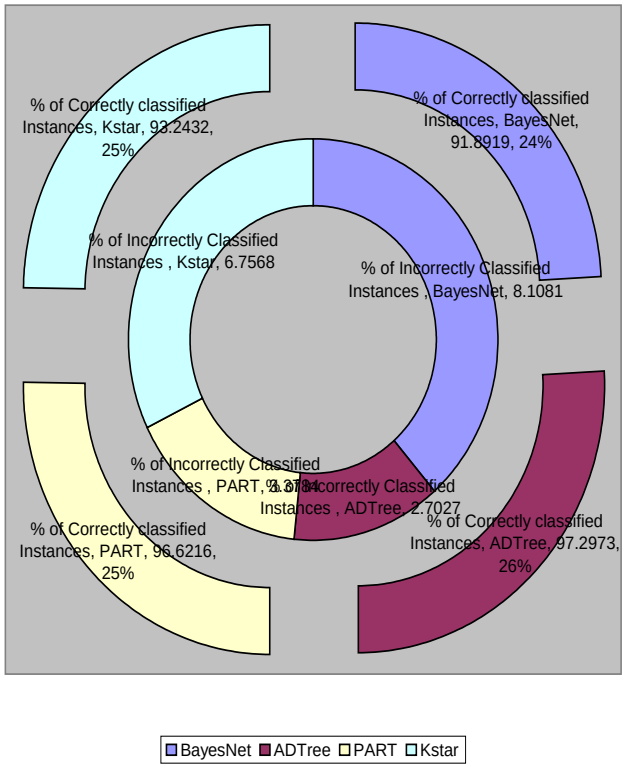

Figure. 1 Results

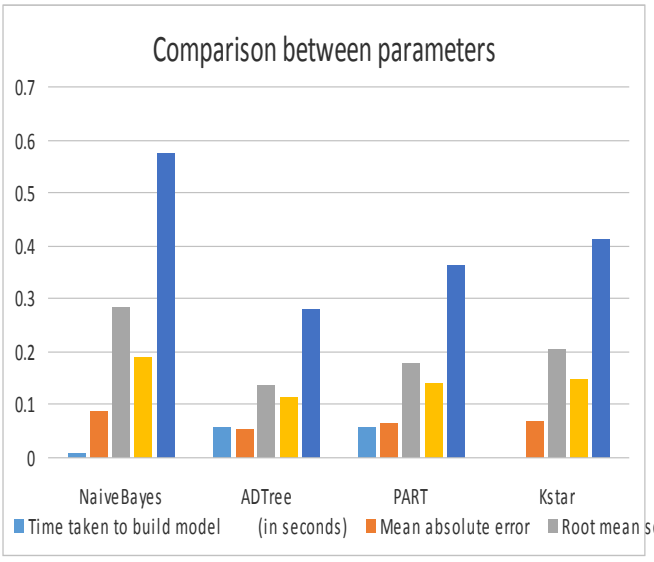

Figure.2 Comparison between parameters

VI. DISCUSSIONS

From the above data tables and Figures, we can observe that the highest accuracy is 97.2973% and the lowest is 91.8919 %. Classifiers Naïve Bayes, ADTree, PART and KStar give accuracy of 91.2162%, 97.2973%, 96.6216%, and 93.2432% respectively. From the results, ADTree classifier gives highest accuracy for the given data tuples. Out of 148 instances, ADTree can correctly classify 144 tuples, PART classifier can classify 143 tuples correctly, KStar can classify 138 tuples correctly, and Naïve Bayes can classify 135 tuples correctly. In terms of the total time taken to build the model, KStar classifier is found to be better algorithm, then ADTree & PART classifiers, then Naïve Bayes classifier respectively.

According to the simulation results, mean absolute errors of the algorithms NaïveBayes, ADTree, PART, and Ktar classifiers are observed as 0.0912, 0.0565, 0.0686, and 0.0722 respectively. Root mean square error vales are 0.2858, 0.1401, 0.1811, and 0.2053 respectively for classifiers NaiveBayes, ADTree, PART, and Kstar It is easy to calculate relative absolute error and relative absolute squared error from these values. High error rate is found in NaiveBayes classifier. An algorithm with lower error rate is preferred because of its high classification ability.

VII. CONCLUSION

In this study and dissertation we investigated Naïve Bayes classifier, Alternating Decision Tree, PART rule based classifier and Kstar lazy learner. Simulation results have been presented in terms of accuracy and error rate. The confusion matrices are also presented for better understanding of the results. ADTree algorithm has shown high accuracy with 97.2973 % and the Time taken to build the model is 0.06 sec. The same algorithm has also shown lowest average error rate of 0.0983. These results put forward the consideration of the fact that ADTree has higher potential of correctly classifying the given data tuples under consideration.

VIII. REFERENCES

[1] Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Pe-ter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[3] Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proceeding of the Sixteenth International Conference on Machine Learning, Bled, Slovenia.

[4] Eibe Frank, Ian H. Witten: Generating Accurate Rule Sets Without Global Optimization.

[5] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Pe-ter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[7] George H. John, Pat Langley: Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence.

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014