Vol 8. No. 3(2), October 2015 ISSN 2006-1781
African Journal of Computing & ICT
© 2016 Afr J Comp & ICT – All Rights Reserved
www.ajocict.net

Comparative Analysis of Selected Supervised Classification Algorithms

M.A. Mabayoje, A.O. Balogun, S. Salihu & K.R. Oladipupo
Department of Computer Science,
University of Ilorin
Ilorin, Nigeria
E-mails: mmabayoje@gmail.com , balogun.ao1@unilorin.edu.ng, shaksoft@yahoo.com,
Phones: 08063185885, , 08120282939, , 08033974515

ABSTRACT

Information is not packaged in a standard easy-to-retrieve format. It is an underlying and usually subtle and misleading concept buried in massive amounts of raw data. From the beginning of time it has been man's common goal to make his life easier. The prevailing notion in society is that wealth brings comfort and luxury, so it is not surprising that there has been so much work done on ways to sort large volume of data. Over the year, there are various data mining techniques and used to sort large volume of data. This paper considers Classification which is a supervised learning technique. Therefore the need to come up with the most efficient way to deal with voluminous data with very little time frame has been one of the biggest challenges to the AI community. Hence, this paper presents a comparative analysis of three classification algorithms namely; Decision Tree (J-48), Random Forest and Naïve Bayes. A 10-fold cross validation technique is used for the performance evaluation of the classifiers on KDD''99, VOTE and CREDIT datasets using WEKA (Waikato Environment for Knowledge Analysis) tool. The experiment shows that the type of dataset determines which classifier is suitable.

Keywords*:* Classification, Decision Tree (DT J-48), Random Forest (RF), Naïve Bayes (NB).

African Journal of Computing & ICT Reference Format:
M.A. Mabayoje, A.O. Balogun, S. Salihu & K.R. Oladipupo (2015): Comparative Analysis of Selected Supervised Classification Algorithms.
Afri J Comp & ICTs Vol 8, No.3 Issue 2 Pp 47-52

1. INTRODUCTION

Knowledge discovery in databases (KDD) is the process of sorting through large amounts of data and picking out relevant information. It is the automated extraction of hidden predictive information form large databases [4], hence it is useful for collecting and interpreting data from huge database [5]. Data mining in relation to Enterprise Resource Planning is the statistical and logical analysis of large sets of transaction data, looking for patterns that can aid decision making. Now, statisticians view data mining as the construction of a statistical model, that is, an underlying distribution from which the visible data is drawn [9]. There are some who regard data mining as synonymous with machine learning. There is no question that some data mining appropriately uses algorithms from machine learning. Machine types used by machine-learning practitioners, such as Bayes nets, Support Vector Machines, decision trees, hidden Markov models, and many others.

Classification is the process of finding the hidden pattern in data. Classification is one of data mining functionalities. It finds a model or function that separates classes or data concepts in order to predict the classes of an unknown object. The data analysis task is classification, where a model or classifier is constructed to predict class (categorical) labels, such as "safe" or "risky" for the loan application data. These categories can be represented by discrete values, where the ordering among values has no meaning. Because the class labels of training data is already known, it is also called supervised learning.

Classification consist two processes: (1) training and (2) testing. The first process, training, builds a classification model by analyzing training data containing class labels. While the second process, testing, examines a classifier (using testing data) for accuracy (in which case the test data contains the class labels) or its ability to classify unknown objects (records) for prediction [3].

This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way. Classification algorithms are laid under classification techniques such as Decision Tree based Methods, Rule-based Method, Memory – based Reasoning, Neural Networks, Naïve Bayes and Bayesian Belief Networks, Support Vector Machines and so on.
The rest of the paper is organized as follows: Section 2,briefs about classification algorithm such as Decision Tree (DT-J48), Random Forest (RF) and Naïve Bayes (NB). Section 3 explain briefly about experimental analysis and results. Section 4 presents a conclusion for this paper.

© 2016 Afr J Comp & ICT – All Rights Reserved

2. CLASSIFICATION ALGORITHMS

A. DECISION TREE

Decision tree is a predictive modeling technique most often used for classification in data mining [10]. The Classification algorithm is inductively learned to construct a model from the pre-classified data set. An advantage of using decision tree algorithms is that its construction does not require any domain knowledge. Hence a data mining expert with little knowledge of networking can help build accurate decision tree models and decision trees can handle high dimensional data. Each data item is defined by values of the attributes and classification may be viewed as mapping from a set of attributes to a particular class. Each non-terminal node in the decision tree represents a test or decision on the considered data item. Choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached [11]. In decision tree, Each internal node tests an attribute, Each branch corresponds to attribute value, Each leaf node assigns a classification and When DT is used instances are describable by attribute. Target function is discrete valued, Disjunctive hypothesis may be required very useful when there is possibly noisy training data.

A. Random Forest

Random Forest is an ensemble of trees specifically decision trees, which has been ensemble using different methods such as bagging, boosting ,random split selection. Random forests, a meta-learner comprised of many individual trees, was designed to operate quickly over large datasets and more importantly to be diverse by using random samples to build each tree in the forest. Randomly sample with replacement (bootstrap) the training set and select 2/3 of data to be used for tree construction, choose a random number of attributes from the in Bag data and select the one with the most information gain to comprise each node and continue to work down the tree until no more nodes can be created due to information loss (). Diversity is obtained by randomly choosing attributes at each node of the tree and then using the attribute that provides the highest level of learning. Performance of the random forests algorithm is linked to the level of correlation between any two trees in the forest. The more the correlation increases, the lower the overall performance of the entire forest of trees.

The way to vary the level of correlation between trees is by adjusting the number of random attributes to be selected when creating a split in each tree. Increasing this variable (m) will both increase the correlation of each tree and the strength of each tree. At some point the tree correlation and tree strength will complement each other providing the highest performance. In addition, increasing the number of trees will provide a more intelligent learner just as having a large diverse group will make intelligent decisions. A random forest is a classifier consisting of a collection of tree structured classifiers {h(x,Qk), k=1, ...} where the {Qk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [2].

B. Naïve Bayes

The Bayesian classification represents a supervised learning method as well as a statistical method for classification Assuming an underlying probabilistic model, it allows to capture an certainty about the model in a principled way by determining probabilities of the outcomes [1]. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Naïve Bayesian classifiers simplify the computations and exhibit high accuracy and speed when applied to large databases. A disadvantage of using Bayesian networks is that their results are similar to those derived from threshold-based systems, while considerably higher computational effort is required [11]. Another disadvantage is that in naïve bayes approach it is assumed that the data attributes are conditionally independent [12] which is not always so (it should be noted however that despite this, Bayesian classifiers give satisfactory results because focus is on identifying the classes for the instances, not the exact probabilities). Naive Bayes (NB): Handles continuous attributes three ways: model them as a single normal, model them with kernel estimation, or discretize them using supervised discretization. For each trial we use 4000 cases to train the different models, 1000 cases to calibrate the models and select the best parameters, and then report performance on the large final test set. We would like to run more trials, but this is a very expensive set of experiments. Fortunately, even with only five trials we are able to discern interesting differences between methods [13].

The naive Bayesian classifier works thus: Each data sample is represented by an n dimensional feature vector, X = (x1, x2.... xn) Suppose that there are m classes H1, H2.... Hm .Given an unknown data sample, X, the classifier will predict that X belongs to the class having the higher posterior probability, conditioned on X. That is, the naive Bayesian classifier assigns an unknown sample X to the class Hi if and only if: P (Hi / X) > P (Hj / X) for $1 \leq j \leq m$. this posterior probabilities are computed using Bayes theorem. In other words an unknown sample X is assigned to the class Hi for which the P (Hi/X) is the maximum.

© 2016 Afr J Comp & ICT – All Rights Reserved

3. EXPERIMENTAL RESULTS

This section presents the result of experimental studies using both crisp-valued and real-valued data sets. We evaluate algorithms on KDD''99 and on datasets, which are available in the WEKA tool. In our experiment, DT(J-48), Random Forest and Naïve Bayes were compared using **Weka**. A short experimental evaluation for benchmark datasets is presented. The information of the data sets contains names of dataset, number of instances and number of attributes which are given in Table 1.

Table 1: Experimental datasets

Index	Dataset	Instances	Attributes
1	KDD''99	487,271	42
2	VOTE	435	17
3	CREDIT	1,000	21

A. Weka Classification

The Waikato Environment for Knowledge Analysis (Weka) is a comprehensive suite of Java class libraries that implement many state-of-the-art machine learning and data mining algorithms. Weka is freely available on the World-Wide Web and accompanies a new text on data mining [7] which documents and fully explains all the algorithms it contains. Applications written using the Weka class libraries can be run on any computer with a Web browsing capability; this allows users to apply machine learning techniques to their own data regardless of computer platform.

Tools are provided for pre-processing data, feeding it into a variety of learning schemes, and analyzing the resulting classifiers and their performance [8].

An important resource for navigating through Weka is its on-line documentation, which is automatically generated from the source. The primary learning methods in Weka are ―classifiers‖, and they induce a rule set or decision tree that models the data. Weka also includes algorithms for learning association rules and clustering data.

The core package contains classes that are accessed from almost every other class in Weka. The most important classes in it are *Attribute*, *Instance*, and *Instances*. An object of class Attribute represents an attribute—it contains the attribute's name, its type, and, in case of a nominal attribute, it's possible values. An object of class Instance contains the attribute values of a particular instance; and an object of class Instances contains an ordered set of instances—in other words, a dataset.

In this paper we have taken the classifiers such as Decision Table, Random Forest and Naive Bayes. The datasets that are used are KDD''99, VOTE and CREDIT (both of WEKA tool) are classified using the above referred classifiers. Table 2, 3, 4 shows the correctly and incorrectly classified instances and classification time of mentioned classification algorithms respectively.

TABLE 2: Classification Accuracy and Time For KDD''99

Algorithms	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**Classification Time (Seconds)**
DECISION-TREE	99.9598	0.0402	130.98
RANDOM FOREST	99.9733	0.0267	142.71
NAÏVE BAYES	99.6661	1627	32.79

© 2016 Afr J Comp & ICT – All Rights Reserved

Figure 1, depicts the performance of the discussed classification algorithms on KDD''99 dataset. Random Forest exhibit highest classification accuracy and is the best supervised classification algorithm for KDD''99 data set.

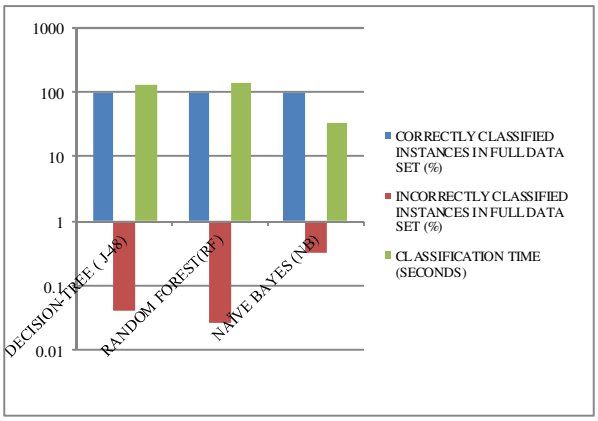

Figure 1: Classification Accuracy and Time for KDD''99 dataset

TABLE 3: Classification Accuracy And Time For Vote Dataset

Algorithms	Correctly Classified Instances	Incorrectly Classified Instances	Classification Time (Seconds)
DECISION-TREE	96.3218	3.6782	0.06
RANDOM FOREST	95.4023	4.5977	0.28
NAÏVE BAYES	90.119	9.881	0

Figure 2, depicts the performance of the discussed classification algorithms on VOTE dataset. Decision Tree exhibit highest classification accuracy and is the best supervised classification algorithm for VOTE data set.

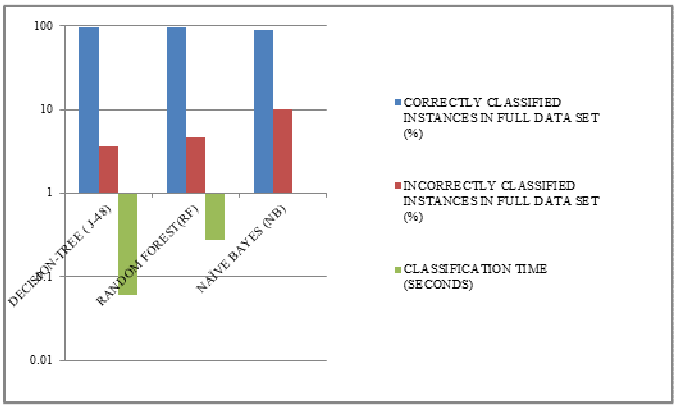

Figure 2: Classification Accuracy and Time for VOTE data set

© 2016 Afr J Comp & ICT – All Rights Reserved

Table 4: Classification Accuracy And Time For Credit Data Set

Algorithms	Correctly Classified Instances	Incorrectly Classified Instances	Classification Time (Seconds)
DECISION-TREE	70.5	29.5	0.22
RANDOM FOREST	73.6	26.4	0.31
NAÏVE BAYES	75.4	24.6	0.03

Figure 3, depicts the performance of the discussed classification algorithms on CREDIT dataset. Naïve Bayes exhibit highest classification accuracy and is the best supervised classification algorithm for CREDIT data set.

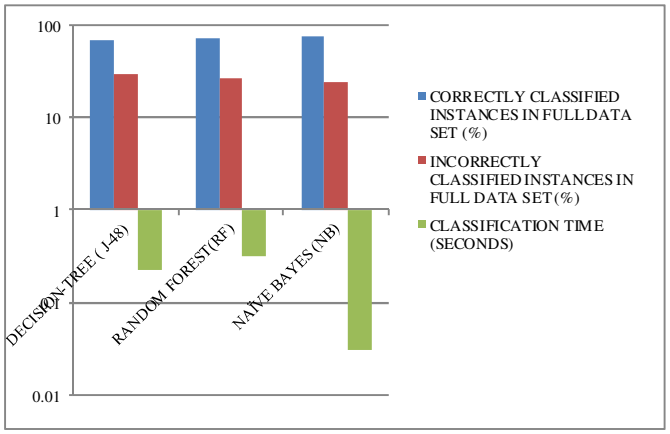

Figure 3: Classification Accuracy and Time for CREDIT data set

4. CONCLUSION

Inarguably, various algorithms have been used for many researches; it is of high importance to note that each of the algorithms has its own advantages and disadvantages. Figure 1, Figure 2 and Figure 3 above show the performance of some selected algorithms in classifying connection records (KDD Cup '99 data set, VOTE and CREDIT (WEKA) datasets). Despite the fact that algorithms gave different detection rate and one is better than the others albeit on different dataset, none is actually said to be best. It is pertinent to note that different classifiers have different knowledge regarding the problem and they approach the problems differently. The type of dataset determines which is best.

REFERENCES

[1] D.sheela, Jeyarani, R. Rajeswari, A. Pethakikshmi. (2013) "Comparative study of Decision Tree and Naive Bayesian", International Journal Computer Applications.

[2] Breiman L. (2001),Classification and Regression by Random **Forest 2001.**

[3] Alex, Stephen, & Kurt, —Building Data Mining application for CRM, USA 1999.

[4] Elena Zhelera, (2009) "Intelligent Technique for Warehousing and Mining Sensor Network" Data, pp. 159, 2009. ISBN 1605663298.

[5] C. Velayutham and K. Thangavel, (2011) —" Unsupervised Quick Reduct Algorithm Using Rough Set Theory", ‖nternational Journal Of Electronic Science And Technology, Vol.9 (3).

[6] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham, —Weka: Practical Machine Learning Tools and Techniques with Java Implementations.

[7] Jiewei Han Micheline Kamber and Jian Pei, (2011), "Data Mining: Concept and Techniques", 3rd edition Morgan Kaufmann Publishers.

© 2016 Afr J Comp & ICT – All Rights Reserved

[8] Carbone, P. L. (1997). "Data mining or knowledge discovery in databases: An overview", In Data Management Handbook, New York: Auerbach Publications.
[9] E.Kesavulu Reddy, Member IAENG, V.Naveen Reddy, P.Govinda Rajulu, (2011) "A study of Intrusion Detection in Data Mining". WCE 2011, July 6 -8, 2011.
[10] Barbara, D., Wu, N. and Jajodia, S. [2001]. "Detecting Novel Network Intrusions Using Bayes Estimators", Proceedings Of the First SIAM Int. Conference on Data Mining, (SDM 2001), Chicago, IL.
[11] Rich Caruana and Alexandru Niculescu-Mizil, (2006) "An Empirical Comparison of Supervised Learning Algorithms" (2006).

Author's Biography

MABAYOJE, Modinat is a Lecturer of Computer Science at the Department of Computer Science, University of Ilorin, Nigeria. She obtained her BSC Computer Science at the University of Ilorin, Iloirn, Nigeria in 2003, a Master of Science Degree in Computer Science at the University of Ilorin, Ilorin in 2009 and a PhD Degree in Computer Science from the University of Ilorin, Ilorin, Nigeria in 2015. Her research interests include Information Retrieval, Data Mining, Machine Learning and Information system. A distinguished member of Computer Professionals (Registration) Council of Nigeria, Computer Science and Information Technology Community (CSITC). She can be reached by phone on +23480635885 and throughE-mail mabayoje.ma@unilorin.edu.ng mmabayoje@gmail.com

BALOGUN, Abdullateef is a Lecturer of Computer Science at the Department of Computer Science, University of Ilorin, Nigeria. He obtained his B.Sc. and M.Sc. degrees in Computer Science at the University of Ilorin, Ilorin, Nigeria in 2012 and 2015 respectively. His research interests include Data Mining, Machine Learning, Information Security, and Software engineering. He can be reached by phone on +234-812-028-2939 and through E-mail balogun.ao1@unilorin.edu.ng bharlow058@gmail.com

Salihu Shakirat A. is a lecturer at the department of Compter Science, University of Ilorin, Kwara State, Nigeria. She obtained B.Sc and M.Sc degrees in Computer Science at University of Ilorin in 2006 and 2011 respectively. Her research works has been based on implications of ICT tools in cashless economy, classroom activities and Good Governance. Other areas of interests includes Knowledge Management and Information Retrieval. She can be reached by phone on +2348033974515 and E-mail shaksoft@yahoo.com