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. 3, March 2014, pg.562 – 568

RESEARCH ARTICLE

A Comparative Study on Performance Evalution of Eager versus Lazy Learning Methods

Solomon Getahun Fentie[1] **Abebe Demessie Alemu**[2] **Bhabani Shankar D. M.**[3]

[1]Department of Computer Science, Bahir Dar University, Ethiopia

[2]Department of Computer Science, Bahir Dar University, Ethiopia

[3]Department of Computer Science, Bahir Dar University, Ethiopia

[1] get.masters@gmail.com, [2] abebe_demessie@yahoo.com, [3] dm_bhabani@yahoo.co.in

Abstract: - Classification is one of the fundamental tasks in data mining and has also been studied extensively in statistics, machine learning, neural networks and expert systems over decades. Naïve Bayes, k-nearest, and decision tree are the most commonly known classification algorithms ever used in different researches. In this study, the performance evaluation of eager (naïve Bayes, ADTree) and lazy (IBk, KStar) classification algorithms are experimented. Our findings show that based on the evaluation metrics precision, recall, F-measure and accuracy, eager learners are slow in training but faster at classification than lazy classification algorithms because they constructs a generalization model before receiving any new tuples to classify. Moreover, based on our investigation eager learners outperform the lazy learners in their accuracy.

Keywords: Classification; Eager learner; Lazy learner; ADTree; Naïve Bayes

I. INTRODUCTION

In the recent past few years, the fields of machine learning and data mining have been studied to a great extent and applied in various fields of studies. It is now realized among the research communities that the contribution of machine learning has become immense for the development of science and technology. Classification, which is one of the supervised machine learning methodologies, is related to one of the fundamental tasks in data mining and has also been studied extensively in statistics, neural networks and expert systems over decades [5]. Classification of instances or tuples from a given large data sets requires techniques of data mining to apply.

© 2014, IJCSMC All Rights Reserved

Classification involves two phases-construction of a model for classification/prediction and testing & usage of it for determining the class labels/ prediction. In this paper, performance evaluation of eager (naïve Bayes, ADTree) and lazy (IBk or K-NN, KStar) learning algorithms are experimented based on the standard UCI Bank Marketing data set donated on 2012-02-14. The data is related to direct marketing campaigns through phone calls of a banking institution with the classification goal of predicting if the client will subscribe a term deposit or not.

The learning algorithms are passed through four different training data sets. The rest of paper is organized as: learning algorithms for classification, data preparation, experiment, result & discussion and finally, the conclusions that have been drawn from the experiment.

II. LEARNING ALGORITHMS FOR CLASSIFICATION

Naïve Bayes: - Naïve Bayes classification algorithm is one of eager learning algorithms that promotes the class conditional independence and predicts the class label in the fastest manner. Naïve Bayesian classifiers assume that there are no dependencies amongst attributes. This assumption is called class conditional independence [1] .As in [2], the performance of the naive Bayes classifier is surprisingly good even if the independence assumption between attributes is unrealistic in most of the data sets.

ADTree: - In classification, there are a number of tree algorithms that are used to classify a number of data instances as accurate as possible. AdaBoost generates rules that have majority votes over the simpler so called “weak” rules. By representing alternating trees (AD Tree) as such votes, we can make it easy to use AdaBoost to learn alternating trees from data [3]. This representation generalizes both voted stumps and decision trees in a natural way. One of the nice features of ADTrees is that they give a measure of confidence which we call the classification margin in addition to a classification.

IBk:-The lazy IBk (commonly known as K- nearest neighbor) is one of classification algorithms that uses distance weighting measures with capability of various attributes like Date attributes, Numeric attributes, Unary attributes, Nominal attributes, Missing values, Binary attributes and Empty nominal attributes.

KStar:- KStar (K*) is an Instance-based learners (IBL) that classifies an instance by comparing it to a database of pre-classified examples [4]. The fundamental assumption is that similar instances will have similar classifications. The question lies in how to define “similar instance” and “similar classification”. The corresponding components of an instance-based learner are the distance functions which determine how similar two instances are, and the classification functions which specify how instance similarities yield a final classification for the new instance. It is one of the lazy learners that take long time to classify an instance of test data sets.

III. DATA PREPARATION

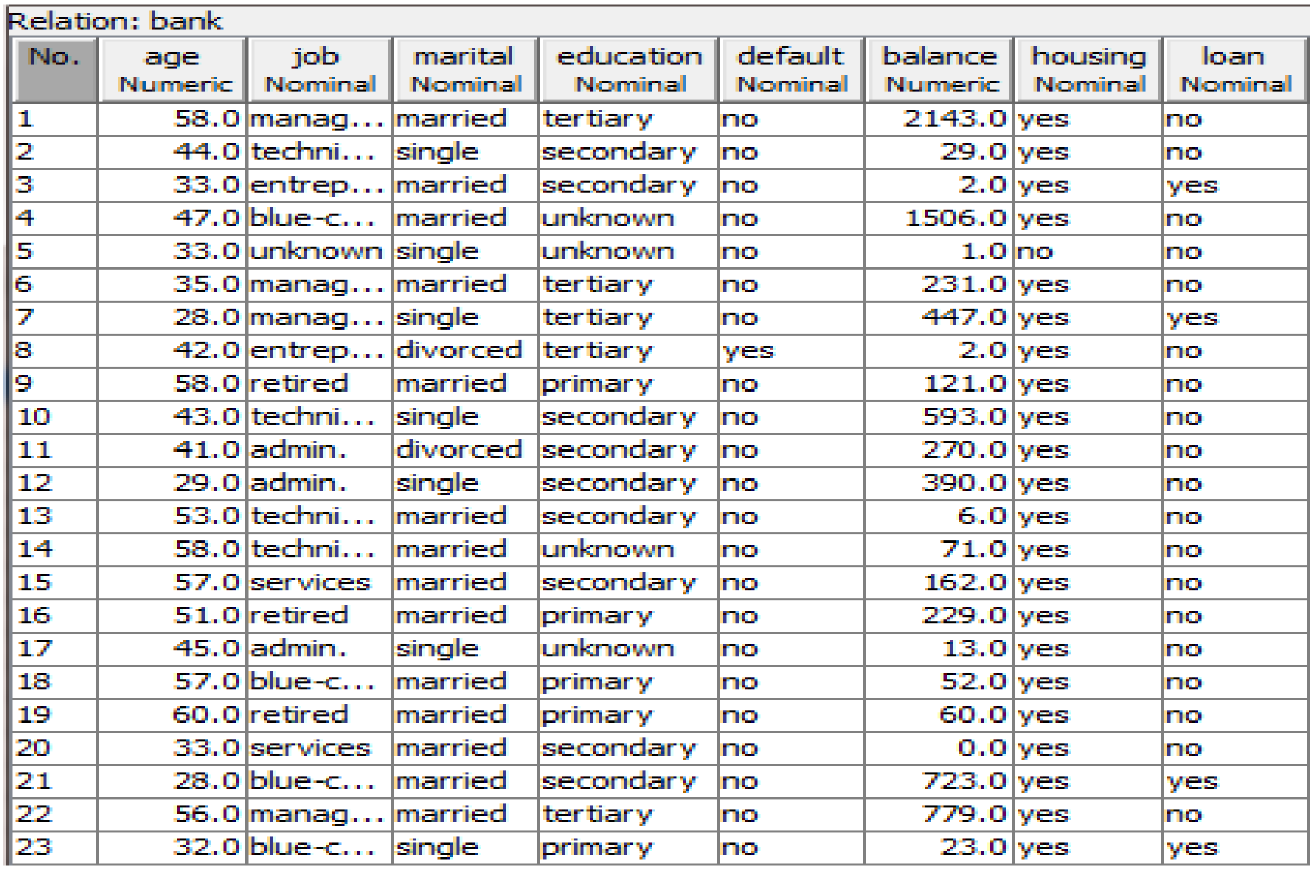
In the classification algorithms, data sets are transformed into training sets and test sets in order to build a model and use it for the classification purpose respectively. The training set involves the various attributes having one as classifying attribute. On the other hand the test set includes the same attributes with the unseen tuples of data that the model is going to classify the instances.

The data under study is related to direct marketing campaigns based on the phone calls. Often, more than one contact to the same client was required, in order to access if the bank client deposit was (or not) subscribed.

In our case we used four different training sets with 17 attributes of each. Each training sets contain 1000, 2000, 3000 and 4000 instances.

The attributes of our training sets are age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, pout and classify.

On these selected 4-training sets, 4-types of learning algorithms have been employed. The experiment is done using one of the best known data mining software called Weka. To show the performance of the classifiers or learning algorithms, we used the same experiment procedure of Weka as 66% of the data set as training set and the remaining as test set.



Relation: bank

No.	age Numeric	job Nominal	marital Nominal	education Nominal	default Nominal	balance Numeric	housing Nominal	loan Nominal
1	58.0	manag...	married	tertiary	no	2143.0	yes	no
2	44.0	techni...	single	secondary	no	29.0	yes	no
3	33.0	entrep...	married	secondary	no	2.0	yes	yes
4	47.0	blue-c...	married	unknown	no	1506.0	yes	no
5	33.0	unknown	single	unknown	no	1.0	no	no
6	35.0	manag...	married	tertiary	no	231.0	yes	no
7	28.0	manag...	single	tertiary	no	447.0	yes	yes
8	42.0	entrep...	divorced	tertiary	yes	2.0	yes	no
9	58.0	retired	married	primary	no	121.0	yes	no
10	43.0	techni...	single	secondary	no	593.0	yes	no
11	41.0	admin.	divorced	secondary	no	270.0	yes	no
12	29.0	admin.	single	secondary	no	390.0	yes	no
13	53.0	techni...	married	secondary	no	6.0	yes	no
14	58.0	techni...	married	unknown	no	71.0	yes	no
15	57.0	services	married	secondary	no	162.0	yes	no
16	51.0	retired	married	primary	no	229.0	yes	no
17	45.0	admin.	single	unknown	no	13.0	yes	no
18	57.0	blue-c...	married	primary	no	52.0	yes	no
19	60.0	retired	married	primary	no	60.0	yes	no
20	33.0	services	married	secondary	no	0.0	yes	no
21	28.0	blue-c...	married	secondary	no	723.0	yes	yes
22	56.0	manag...	married	tertiary	no	779.0	yes	no
23	32.0	blue-c...	single	primary	no	23.0	yes	yes

Fig. 1: Example of Data view

IV. EXPERIMENT

To conduct our experiment, we used four classifiers, namely Eager learners (Naïve Bayes, Alternating Decision Tree), and Lazy learners (IBK, Kstar) on the UCI Bank Marketing data set. The experiment was conducted on the four training data sets based on test mode of k-fold cross-validation with supplied test sets (where k=10). For each classifier, we considered a classification time and their respective performance value. For the performance issue of classifiers, we focused on the evaluation parameters: precision, recall, accuracy and F-measure.

© 2014, IJCSMC All Rights Reserved

TABLE I: Data Description

	Training Data	**Testing Data**

Number of instances	Train-data-1=1000
	Train-data-2=2000
	Train-data-3=3000
	Train-data-4=4000
		Test-data-1=340
		Test-data-2=680
		Test-data-3=1020
		Test-data-4=1360

Number of attributes	17 (7-Numeric, 10-Nominal)	17 (7-Numeric, 10-Nominal)
Missing data	none	none

A. *Experiment Procedures*

The experiment was carried out as follows:

Preparing working data: The data for the purpose of manipulation was taken from the UCI Bank Marketing data set which consists of more than 45,000 instances. The training and test data sets are taken from these data tuples independently.

Setting classes of the training set: The last attribute (classify) represents our result set (class label) which has values of "Yes" or "No". Every qualifying data values of the attributes of our dataset will result "yes" and every inappropriate attribute values will result "no".

Supplying the chosen values of data: The data values of all 17 attributes are entered to the WEKA data mining software as training and test sets.

Classification: In this step, the processed data has been taken one-by-one and has been submitted to all classifiers, eager (Naïve Bayes, Decision Tree) and lazy (IBK, K*) against the training and test sets to determine the result set.

V. RESULT AND DISCUSSION

Our experiment demonstrates that the eager learning algorithm (ADTree) outperforms better in its accuracy in all training sets as shown in Fig-2. On the other hand the precision and recall are compromised as the number of instances increases in the training sets.

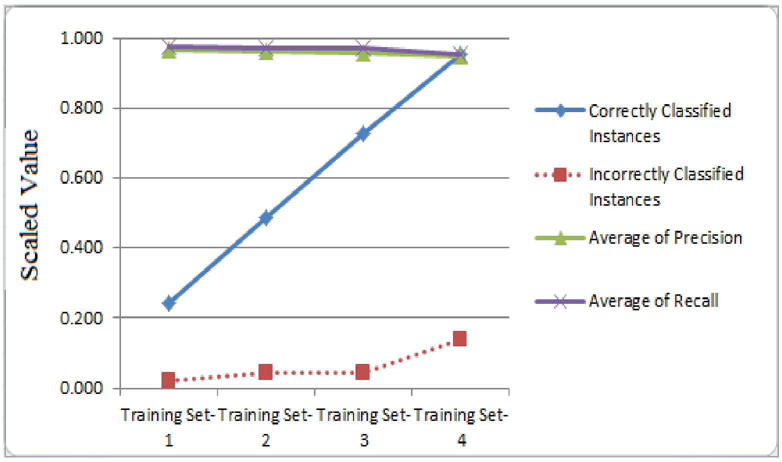


Fig. 2: Performance of eager (ADTree) classifier

© 2014, IJCSMC All Rights Reserved

The average weighted precisions for ADTree outperforms better than Naïve Bayes, IBK and Kstars respectively as shown in the Figure-3 in all test sets.

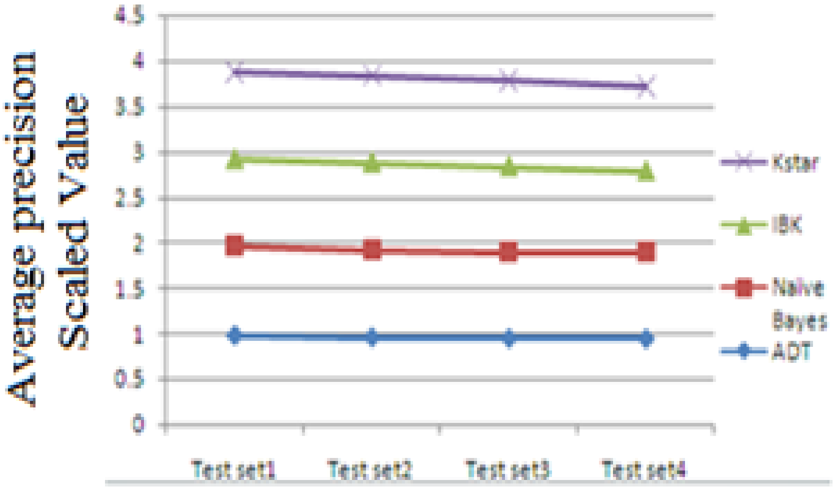


Fig. 3. Average precession of all classifiers

F–measure is the harmonic mean of recall and precession, and hence to know which learner is the best in terms of classifiers both precision and recall, we can infer the F-measure values.

As we can infer from the Figure-4, the average F-measure value of ADTree is the largest among all the other classifiers followed byNaïve Bayes, IBk and KStar respectively for all the test sets indicating that in terms of precession and recall (F-measure), the eager learners outperform than the lazy classifiers.

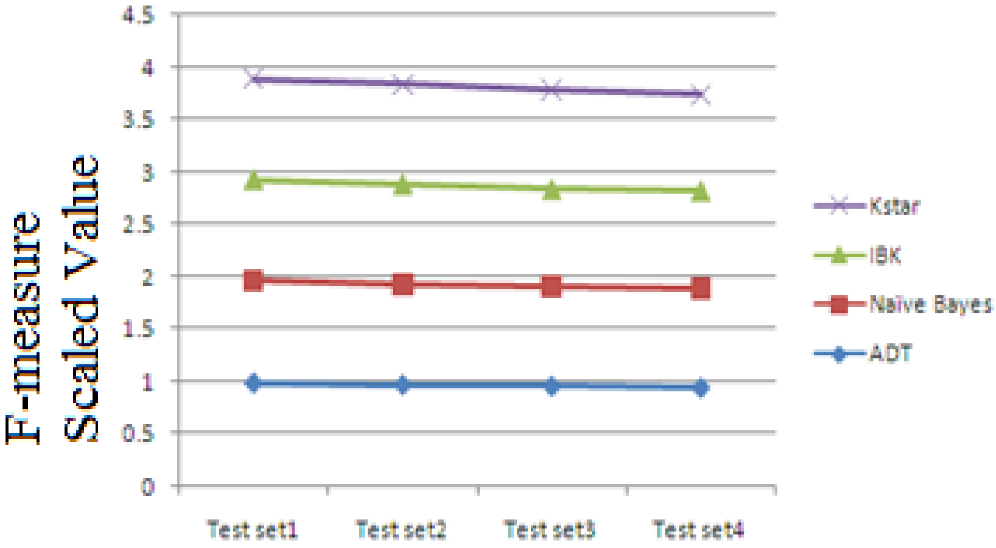


Fig. 4: F-measure of all classifiers

As shown in the Table 2, the ADTree outperforms in its accuracy than Naïve Bayes, IBk and KStar respectively.

© 2014, IJCSMC All Rights Reserved

TABLE 2: Accuracy comparison of different classifiers

Classifiers	Correctly classified Instances in Test 1 (%)	Incorrectly classified Instances in Test 1 (%)
Naïve Bayes	97.9412	2.0588
ADTree	98.5294	1.4706
IBk-(KNN)	97.3529	2.6471
KStar	97.2059	2.7941

On the other hand, in the Table-3, eager learning algorithms (Naïve Bayes, ADTree) require more time to spend in constructing a model while learning whereas the lazy learning algorithms (IBk, KStar) do not require more time during model construction.

TABLE 3: The time taken to build a model and classify

Learning Algorithms		Time taken to build a Model on Training set-4	Time taken to Classify Test set-4
Eager Learning Algorithms	Naïve Bayes	0.03 seconds	0.00 seconds
	ADTree	0.66 seconds	0.00 seconds
Lazy Learning algorithms	IBk	0.00 seconds	0.20 seconds
	KStar	0.00 seconds	0.50 seconds

VI. CONCLUSION

We consider the precision, recall, F-measure and accuracy as our evaluation matrices for conducting our experiment on the performance evaluation of different classifiers. On the basis of these evaluation metrics, we conclude that eager learners are slow in training but faster and accurate at classification than lazy classification algorithms because they construct a generalization model before receiving any new tuples for classification. Moreover, in overall parameters, ADTree (eager) classification algorithm outperforms all the other classifiers followed by Naïve Bayes (eager learner), IBk (lazy learner) and KStar (lazy learner) respectively. However, as the number of instances in the training set and test set increases for the large data sets, eager learning algorithms seek much time to construct and train the model which is an open issue for the researchers to minimize the tradeoff between model construction and classification time.

REFERENCES

[1] Ahmad Ashari,ImanParyudi, A Min Tjoa:"*Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alterntive Design in an Energy Simulation Tool*", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013.

[2] Franz,Pernkopf: "*Bayesian network classifiers versus selective k-NN classifier*", Pattern Recognition 38 (2005) 1 – 10.

[3] YoavFreund ,Llew Mason: "*The alternating decision tree learning algorithm*".

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

[5] D.Lavanya ,Dr. K.Usha Rani : "*Performance Evaluation of Decision Tree Classifiers on Medical Datasets*", International Journal of Computer Applications (0975 – 8887)Volume 26– No.4, July 2011.

[6] S.L. Ting, W.H. Ip, Albert H.C. Tsang: "*Is Naïve Bayes a Good Classifier for Document Classification?*, International Journal of Software Engineering and Its Applications Vol. 5, No. 3, July, 2011.

[7] D.Lavanya and Dr.K.Usha Rani: "*ENSEMBLE DECISION TREE CLASSIFIER FORBREAST CANCER DATA*", International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012.

[8] Reza Entezari-Maleki, Arash Rezaei, and Behrouz Minaei-Bidgoli :"*Comparison of Classification Methods Based on the Type of Attributes and Sample Size*".

[9] Wei-Hao Lin and Alexander Hauptmann : "*Meta-classification: Combining Multimodal Classifiers*", O.R. Zaïane et al. (Eds.): Mining Multimedia and Complex Data, LNAI 2797, pp. 217–231, 2003.