Performance Test on Classification Algorithms

Jeevitha Sampath, Sunitha N V, Arpana Shetty

ABSTRACT--- Nowadays, a huge amount of data is generated due to the growth in the technologies. There are different tools used to view this massive amount of data, and these tools contain different data mining techniques which can be applied for the obtained data sets. Classification is required to extract useful information or to predict the result from these enormous amounts of data. For this purpose, there are different classification algorithms. In this paper, we have compared Naive Bayes, K*, and random forest classification algorithm using Weka tool. To analyze the performance of these three algorithms we have considered three data sets. They are diabetes, supermarket and weather data set. In this work, an analysis is made based on the confusion matrix and different performance measures like RMSE, MAE, ROC, etc.

Keywords — Naive Bayes, K*, Random Forest, Root Mean Squared Error (RMSE), Mean Absolute Error(MAE), Receiver Operating Characteristic(ROC), Weka

I. INTRODUCTION

Data Mining is a technique where valid information or knowledge is obtained by analyzing the hidden patterns in the data. Various tools are available where we can run different classification algorithms to gain the required information. The classification algorithms can determine the category to which the new observation belongs. In this paper, we are comparing the accuracy of three well-known classification algorithms: Naïve Bayes, K* and Random forest. The algorithms are run on three data sets: Diabetics, Supermarket and Weather.

A. Naïve Bayes: It is a simple algorithm used for building classifiers. These classifiers assume that for each class variable the value of a particular feature is independent of the amount of any other feature.
B. K*: This algorithm belongs to the family of "Lazy Learners." In this algorithm, entropy is used as the distance measure.
C. Random forest: It is a classification algorithm which constructs multiple decision trees and combines them to get a better prediction result.

II. LITERATURE SURVEY

In Paper [1] random forest is described as tree predictors where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The study consists of randomly selected input or a combination of the data at each node to grow each tree. Based on the characteristics such as accuracy, error, strength, and correlation the result was compared with other algorithms. It is shown that for large data set low error rate is possible, but the improvement was less on the small data set.

In paper [2] random forest classification algorithm is used to predict the risk of diabetes. Here electronic health record used to store the data of each patient during his admission to the hospital. They have used map reducing programming where map function maps the data set values with similarity variables to predict the risk on early stage. Most of the data in this record are unstructured, and the reduce function reduces the variables. The accuracy obtained by using this technique is 0.87.

In paper [3], random forest algorithm is used to predict the behavior of the customer. A customer plays an important role when it comes to buying the products from any store. Here the survey was conducted by giving the different questions to the customer based the product, their fascinating thing, a place they want to visit, the price of the product, priority on choosing the product and many more. Based on this data and random forest algorithm the prediction is made and the accuracy obtained is 94 %.

In paper [4], Naive Bayes algorithm is used to classify the SMS received by the Rescue Agencies during disasters. Initially, using NLP a bag of words per item is created. Then, Naïve Bayes Classifier is developed using the training set. Finally, the Naïve Bayes Text Classification is done. It classifies the messages into five different classes (Spam, Invalid, Alert1, Alert2, Alert3) based on the pre-classified information that is used as the learned classifier. Up to 89% accuracy is achieved by using this technique. Further improvement can be achieved by increasing the number of entries in the learned classifier.

In paper [5], the Naive Bayes algorithm is used to classify the large text document into the specified domain. Animal and plant domain article in Wikipedia is the data set used for the analysis. The accuracy provided by this method is about 98.8%. Here entire work is divided into two phases. In the first phase, content is extracted from the web page, tokenized, stop word is removed and stemming is done. Stemming is the process of mapping different morphological variant of the word into base word. To help the classifier to learn, or the classifier to classify the document, feature extraction is done in the second phase.

In paper [6], K* algorithm is used to overcome the curse dimensionality problem by using different characteristics.

Revised Manuscript Received on July 10,2019.

Jeevitha Sampath, Department of CSE BIT, Mangalore. (E-mail: jeevs2010@gmail.com)
Sunitha N V, Department of CSE BIT, Mangalore. (E-mail: sunithanv6720@gmail.com)
Arpana Shetty, Department of CSE BIT, Mangalore. (E-mail: arpana.shetty1@gmail.com)

Here author used Weka tool for the analysis where *the globalBlend* parameter was enabled, and it is found that the accuracy of classified results is low by increasing this parameter for the given data set.

In paper [7], the study of K* algorithm is done by comparing its performance with other algorithms. Here analysis is done based on missing values, imbalanced attributes and mixed values for the data set taken from Keel Repository. The tool used to obtain the classifier algorithm is Weka. According to the analysis done by the author based on different parameters, it is proved that the performance of K* algorithm is equivalent to Naïve Bayes and Random Forest. It deals best in noisy and imbalanced attributes.

In paper[8], it is shown how the different classifiers are used for the classification purpose using the tool Weka. Here they have considered predicting the disease in an early stage. They founded the accuracy of each classifier and used other parameters to detect the disease in early stage.

III. RESULT ANALYSIS

In this experiment, the WEKA tool is used. It was developed by Machine Learning Group developed it at the University of Waikato, which contains classification, clustering and association rules. By using this tool, the data set can be exported from another database or any other file.

Weka tool is used to run the classification algorithms, and comparison of the algorithms is made based on its performance measurements.

A. Data sets used in Experiment

I. In the first scenario data set taken for the experiment is diabetes, where it contains 768 instances and nine attributes. In the test mode, 66% is considered as a training set, and the remaining 34% is regarded as a test set. This data set is analyzed with three different algorithms such as Naïve Bayes, K* and Random Forest.

II. In the next scenario data set taken for the experiment is the Supermarket, where it contains 4627 instances and 217 attributes which is more extensive data set when compared to the diabetes data set. In the test mode, 66% is considered as training set and the remaining 34% is regarded as test set. This data set is analyzed with three different algorithms such as Naïve Bayes, K* and Random Forest.

III. In the last scenario data set taken for the experiment is the weather, where it contains 14 instances and five attributes. In the test mode, 66% is considered as training set and the remaining 34% is regarded as test set. This data set is analyzed with three different algorithms such as Naïve Bayes, K* and Random Forest.

B. Performance Measures

Accuracy:

The accuracy of the algorithm can be calculated as follows:

Accuracy = (True Positive + True Negative) / (Number of Instances)

Kappa statistic:

The Kappa statistic is a metric that compares Accuracy which is observed with an Accuracy that is expected.

Kappa= (Observed Accuracy-Expected Accuracy)/(1-Expected Accuracy)

Mean Absolute Error: MAE refers to the mean of the absolute values of the individual prediction error for all instances of the data set which is used for the test.

Root Mean Squared Error /Deviation: RMSD is a measure of the differences between the model predicted values and the observed values.

ROC (Receiver Operating Characteristic) *Area:* The ROC curve is a graphical plot obtained by plotting the true positive rate against the false positive rate for different threshold settings.

TTB (Time tobuildmodel): Time is taken to create the model. This is measured in seconds.

Correctly Classified Instances (CCI): Number of the correct classified instance.

Incorrectly Classified Instances (ICI): Number of the incorrect classified instance from the total number of instance.

C. Classification Results

Three classifiers considered for the analysis are Naive Bayes, Random forest and K* using WEKA.

Different performance measures obtained for the classifiers are tabulated below.

a) Diabetes Data Set

Table 3.1: Confusion Matrix using Naïve Bayes Algorithm

Number of Instances =261	Predicted		
Actual		Not Found	Found
	Not Found	150	28
	Found	32	51

The accuracy of the Naïve Bayes algorithm obtained by using the data mentioned in Table 3.1 is 0.77

Table 3.2: Confusion Matrix for K* Algorithm

Number of Instances =261	Predicted		
Actual		Not Found	Found
	Not Found	147	31
	Found	45	38

The accuracy of the K* algorithm obtained by using the data mentioned in Table 3.2 is 0.71

Table 3.3: Confusion Matrix for the Random Forest Algorithm

Number of Instances =261	Predicted	
Actual	Not Found	Found
Not Found	155	23
Found	33	50

The accuracy of the Random Forest algorithm obtained by using the data mentioned in Table 3.3 is 0.785

Table 3.4: Classification Result

Algorithm	CCI %	ICI %	ROC	Kappa	RMSE	MAE	TTB (sec)
Naive Bayes	77.0115	22.9885	0.854	0.4631	0.3822	0.266	0.02
K*	70.8812	29.1188	0.736	0.297	0.4789	0.3128	0.94
Random forest	78.5441	21.4559	0.838	0.4889	0.3879	0.3046	0.02

Three classifiers Algorithms are compared using different performance measures which are shown in table 3.4. According to the result, ROC and Kappa are better in Naïve Bayes, and Random Forest algorithm whereas RMS and MAE are better in K * in comparison with other two algorithms.

a) Weather Data Set

Table 3.5: Confusion Matrix using Naïve Bayes Algorithm

Number of Instances =5	Predicted		
Actual		Yes	No
	Yes	3	0
	No	2	0

The accuracy of the Naïve Bayes algorithm obtained by using the data mentioned in Table 3.5 is 0.60

Table 3.6: Confusion Matrix for K* Algorithm

Number of Instances =5	Predicted		
Actual		Yes	No
	Yes	2	1
	No	2	0

The accuracy of the K* algorithm obtained by using the data mentioned in Table 3.6 is 0.40

Table 3.7: Confusion Matrix for the Random Forest Algorithm

Number of Instances =5	Predicted		
Actual		Yes	No
	Yes	2	1
	No	2	0

The accuracy of the Random Forest algorithm obtained by using the data mentioned in Table 3.7 is 0.4

Table 3.8: Classification Result

Algorithm	CCI %	ICI %	ROC	Kappa	RMSE	MAE	TTB (sec)
Naive Bayes	60	40	0.333	0	0.5706	0.5129	0
K*	40	60	0.000	-0.3636	0.7445	0.6489	0
Random forest	40	60	0.333	-0.3636	0.7746	0.6	0

Three classifiers Algorithms are compared using different performance measures which are shown in table 3.8. According to the result, ROC is better in Naïve Bayes and Random Forest algorithm whereas Kappa, RMS and MAE are better in K * and Random Forest compares to Naïve Bayes algorithms.

b) Super Market Data Set

Table 3.9: Confusion Matrix using Naïve Bayes Algorithm

Number of Instances =1573	Predicted	
Actual	Low Transaction	High Transaction
Low Transaction	986	0
High Transaction	587	0

The accuracy of the Naïve Bayes algorithm obtained by using the data mentioned in Table 3.9 is 0.63

Table 3.10: Confusion Matrix for K* Algorithm

Number of Instances =1573	Predicted		
Actual		Low Transaction	High Transaction
	Low Transaction	986	0
	High Transaction	587	0

The accuracy of the K* algorithm obtained by using the data mentioned in Table 3.10 is 0.63

Table 3.11: Confusion Matrix for the Random Forest Algorithm

Number of Instances =1573	Predicted		
Actual		Low Transaction	High Transaction
	Low Transaction	986	0
	High Transaction	587	0

The accuracy of the Random Forest algorithm obtained by using the data mentioned in Table 3.11 is 0.63

Table 3.12: Classification Result

Algorithm	CCI %	ICI %	ROC	Kappa	RMSE	MAE	TTB (sec)
Naive Bayes	62.6828	37.3172	0.500	0	0.4839	0.4639	0.03
K*	62.6828	37.3172	0.536	0	0.4839	0.4639	28.88
Random forest	62.6828	37.3172	0.500	0	0.4839	0.4639	0.05

Three classifiers Algorithms are compared using different performance measures which are shown in table 3.12. Since the data set used contains more set of data all the parameters in three algorithms almost remains the same except the ROC value of K * which is a bit higher than the other two algorithms.

IV. CONCLUSION

In this paper, the analysis of three different classification algorithms when applied on three different data sets is made. The size and type of dataset used for the analysis vary. The behavior of these algorithms differs for each of these datasets.

After performing the analysis on data set we found that for Diabetes data set Random Forest algorithm gives higher accuracy, for Super Market data set all three algorithms give the same accuracy and for weather data set Naïve Bayes algorithm gives high accuracy. However, the performance of the algorithm varies depending on the type of data set provided.

REFERENCES

1. Leo Breiman,"RANDOM FORESTS," on Statistics Department University of California Berkeley, CA 94720,2001
2. SS Rallapalli and T. Suryakanthi, "Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm," in International Conference on Advances in Computing and Communication Engineering (ICACCE), Durban, 2016, pp. 281-284.
3. Harsh Valecha, AparnaVarma, IshitaKhare, AakashSachdeva, MuktaGoyal*, "Prediction of Consumer Behaviour using Random Forest Algorithm," In 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2018.
4. A. Ordonez, R. E. Paje and R. Naz, "SMS Classification Method for Disaster Response Using Naïve Bayes Algorithm," 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 233-236.
5. J. Santoso, E. M. Yuniarno and M. Hariadi, "Large Scale Text Classification Using Map Reduce and Naive Bayes Algorithm for Domain Specified Ontology Building," 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, 2015, pp. 428-432.
6. RafetDuriqiViganRaca, BetimCico, "Comparative Analysis of Classification Algorithms on Three Different Datasets using WEKA," in 5th Mediterranean Conference on Embedded Computing, MECO,2016.
7. Dayana C. Tejera Hernandez, "An Experimental Study of K* Algorithm," in I.J.Information Engineering and Electronic Business, March 2015
8. Narander Kumar,Sabita Khatri "Implementing WEKA for medical data classification and Early disease Prediction," in 3rd IEEE International Conference on "Computational Intelligence and Communication Technology," 2017.