Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 48 (1) 185-188 (2016)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Performance Analysis of Dissimilar Classification Methods using RapidMiner

M. A. ABRO*, H. NAWAZ[++], W. A. ABRO**

Department of Computer Science, SZABIST, 90 Clifton, Karachi, Pakistan

Received 15th March 2015 and Revised 08th January 2016

Abstract: The centre of attention of this research paper is to assess the performance of dissimilar functioning methods of classification technique on data set of Urban Land Cover using RapidMiner software and propose one that gives good performance on mentioned data set of Urban Land Cover. This is an attractive dataset for classifying the high resolution image of urban land cover; may be utilized for several purposes with the tree planning etc. The performance of classification methods C4.5 Decision Tree, Random forest, K-Nearest Neighbor and Naïve Bayes is examined on basis of their accuracy cost of error and kappa values. The class precision and recall is derived from confusion matrix.
.

Keywords: Classification, Naïve Bayes, K-NN, Decision Tree, Random Forest, RapidMiner.

1. INTRODUCTION

The classification is one of the imperative techniques of Data mining that uses the training data sets and test data sets and forecast the results. Srivastava and et.al describes the classification as noteworthy technique of Data mining that utilizes the amount of examples consisted of attributes which may be categorical or continuous. The examples are part of training set. The classification is performed on target class attribute. The continuous attributes are ordered while categorical attributes are unordered. (Lee *et.al* 1999) explained that a Classification method functions to locate the classifier to predict the target class label. This study looks at the accuracy, precision, recall, Kappa and cost of classification error of various methods of classification after the assessment and comparison. Therefore, this research study is valuable for the researchers, scientists and other concerned personals in order to choose the classification method for prediction with high accuracy and with minimum classification rate error.

1.1 Rapidminor

There is list of data mining and knowledge discovery tools which process the all types of datasets and give the summarized results to take proper decision. RapidMiner is one of them. This tool is open source and provides integrated atmosphere, solutions, services and working platform for Machine learning analysis. This tool first developed in 2001 and it was named YALE at the University of Dortmund. It was renamed as RapidMiner-1 in year 2007. A number of industries, researchers, academicians and etc are using the RapidMiner for the purposes of analysis, research and education etc.

1.2 Data Set Information

Urbanization is the change of the society that is shifting of people from rural areas to city areas. The recent and exact layouts of cities or urban areas are important sources for the economic, census development as well as cultural and public protection. This data set of Urban Land Cover is classification dataset. In this dataset information is retrieved from a high resolution image taken through satellite. There is targeted class: Land Cover class (nominal) following by Class attributes which are consisted of Concrete, Trees, Grass, Soil, Building, Cars, Pools, Shadows and Asphalt. There is no missing value in this data set. It is multivariate data set. The training dataset is consisted of 168 numbers of instances and 148 attributes. Using this data set, two studies are carried out; by (Johnson *et.al* 2013) and second study done by (Johnson and Brian 2013) in same year, both studies are performed by means of classification technique of data mining to help sustainable metropolitan planning efforts.

Table-1.2.1 shows the detail and status dataset and target Class.

Land Cover Data set and Class analysis	
Class Name	**Land Cover Class**
Class Type:	Nominal
Attributes	148
Instances	168
Missing Values	0

[++]Corresponding Author: Haque Nawaz Lashari, E-mail: hnlashari@smiu.edu.pk , Cell# +92-333-7588408
* Department of Computer Science, Shaheed Benazir Bhutto University, Liyari, Karachi, Pakistan.
[++]** Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan

Class elements:		9
Sr. No	**Label**	**No. Of frequencies in Training set**
1	Concrete	23
2	Trees	17
3	Grass	29
4	Soil	14
5	Building	25
6	Cars	15
7	Pools	15
8	Shadows	16

2. RELATEDWORK

Several studies are carried out on different data sets to solve the problems of real world but still there is need of more research to find out the solutions of the existing problems. The center of attention of this research study is to weigh up the performance of divergent functioning algorithms of classification on data set of Urban Land Cover using RapidMiner software and decide on a classification method which performs well on the data set of Urban Land Cover. Different studies show importance of classification methods regarding their high accuracy and low cost of classification error. (Bhargavi *et.al* 2009) did their study on agriculture land soils using different classification algorithms including Naive Bayes. In this study, Naive Bayes is recognized as efficient and accurate classification method. (Delen, *et.al*, 2005) mention decision tree classification algorithm competent and proficient algorithm in predicting the accuracy than the artificial neural network and logic regression algorithms. (Kuramochi, and Karypis,, 2005) in their study on Gene Classification explains the functionality of K-NN approach that this method performs very good in classification of Genes than the SVM induced and K-NN induced (Rodriguez-Galiano *et.al* 2012) discuss in their paper that Random Forest method is recognized as one of the note worthy method of classification for categorizing the land cover.

Classifying the land cover, (Johnson *et.al*, 2013) used a multi scale approach. In his research study on classification of urban-land-cover classification accuracy is weigh against classification accuracy of image segmentations. It is suggested in this study to compare classification accuracy further by using the Random forest with other classifiers. Viewing this suggestion, in this research study, we use the Random forest along with other classifiers. Viewing this suggestion, in this research study, we use the Random forest along with other classification methods for the classification of urban land cover using RapidMiner tools.

3. MATERIALS AND METHODS

This research study starts from problem statement and data collection. The data set is downloaded from UCI Machine Learning Repository and prepared in RapidMiner tool. All the missing values are set and dataset prepared as recognizable and normal. Data preparation measures are required to make the data set clean and usable for the classification. All the values are processed and make them normalized. The class attribute is labeled as class to classify the urban land. The data set is converted intoCSV format. The example set is consisted of 168 examples / instances, one special attribute that is the class attribute and 147 regular attributes. Following are the process model figures of each technique.

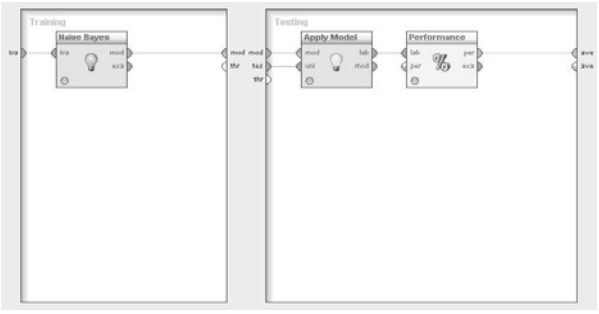
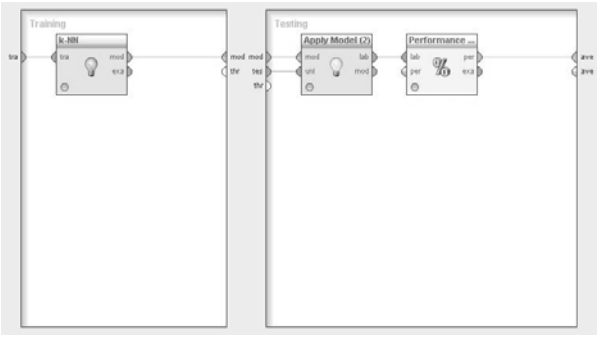


Fig. 3.1 Describes Process Model of Naïve Bayes



Fig, 3.2 Describes Process Model of K-NN

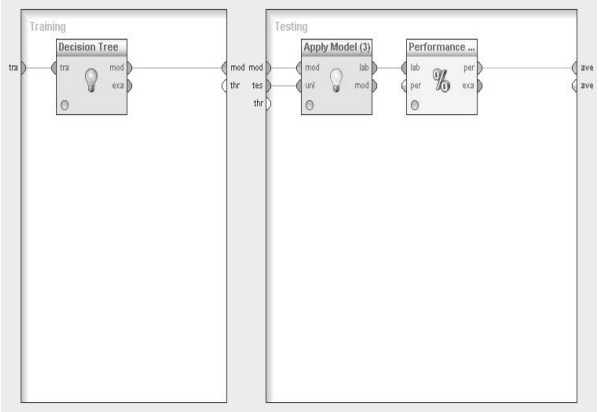


Fig.3.3 Descri Process Model of C4.5 Decision Tree

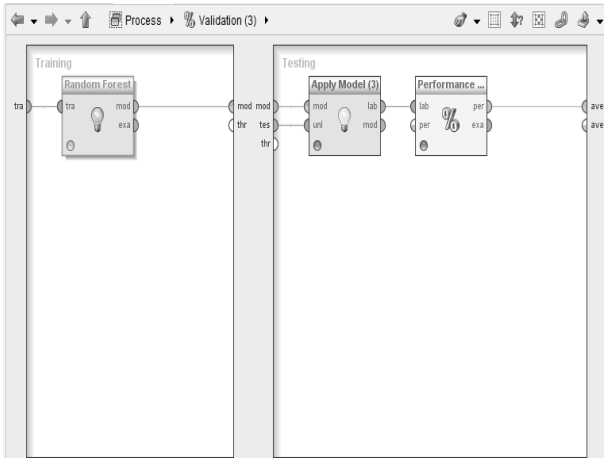


Fig. 3.4 shows the process Model of Random Forest

4. RESULTS EVALUATION AND ANALYSIS

The performance of all four classification methods is evaluated and analysed on the basis of accuracy, classification error, Kappa, class precision and Recall.

The purpose behind this research study is to evaluate the performance of dissimilar classification methods using the RapidMiner data mining tool. All the mentioned methods are processed on dataset of Land Cover. The Cross validation parameters are set same for all classification methods.

The Chart 3.1.1explains the performance of classification methods processed in RapidMiner. There is divergent performance of each method on mentioned dataset. Johnson and et.al (2013) suggested in his study on Land cover to compare classification accuracy further with accuracy of Random forest.

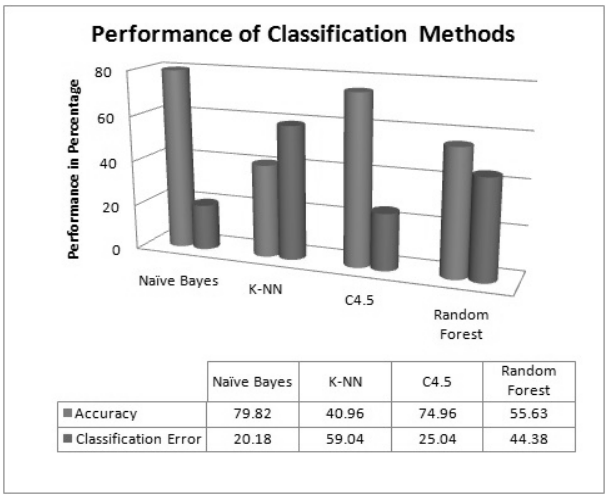


Chart 3.1.1 portray performance of different classification methods

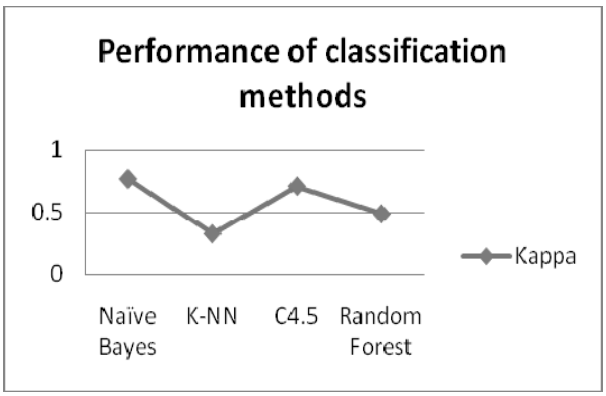


Table 3.1.1 gives details of Class Precision of classification methods. There is significant difference among the class precision of all four methods derived from target class of urban land cover data set.

Class Precision in %				
Predicted Class Elements	**Naïve Bayes**	**K-NN**	**C4.5**	**Random Forest**
car	86.67	75	93.33	100
concrete	70.83	36.11	48.48	41.67
tree	71.43	40	100	87.5
building	75	45.83	72.73	40.54
asphalt	100	10	75	71.43
grass	83.33%	34.62	77.78	42.59
shadow	88.24	16.67	87.5	83.33
soil	60	14.29	33.33	0
pool	93.33	69.23	100	90.91

Table 3.1.2 explains the class Recall differences obtained from the functionality of classification methods used in this study. The good results of class recall are acquired from Naive Bayes and C4.5.

Class Recall in %				
Predicted Class Elements	**Naïve Bayes**	**K-NN**	**C4.5**	**Random Forest**
car	86.67	80	93.33	53.33
concrete	73.91	56.52	69.57	43.48
tree	88.24	70.59	58.82	41.18
building	72	44	64	60
asphalt	92.86	7.14	85.71	71.43
grass	68.97	31.03	96.55	79.31
shadow	93.75	6.25	87.5	62.5
soil	64.29	7.14	14.29	0
pool	93.33	60	93.33	66.67

5. RESULTS DISCUSSION

The performance of Random forest in shape of Accuracy is little with high rate of classification error and low values of Kappa. The performance of C4.5is high than the Random forest. The K-NN does not give the suitable results. The accuracy and kappa values of K-NN are lowest in this study performed on urban land cover. The Naive Bayes gives good performance with high accuracy, Kappa and low cost of classification error. Here Naive Bayes gives 79.82% accuracy with 0.769 Kappa values and C4.5 gives 74.96% accuracy with 0.711 values. While Random Forest gives 55.63% accuracy with 0.487 Kappa values and K-NN gives 40.96% with 0.325 Kappa Values.

The Class Precision and Recall resultsare not same of all four classification methods. However, Naive Bayes and C4.5 methods of classification performed well and give high ratio of Class precision and Recall in contrast of Random forest and K-NN

6. CONCLUSION

This paper looks at the performance of classification methods after the assessment and comparison. Therefore, this research study is important for the researchers, scientists and other concerned personals in order to decide the classification method for prediction with high accuracy and with minimum classification rate error. Four classification methods which are Naive Bayes, C4.5, Random Forest and K-NN have used in this research study Whole study is performed on urban land dataset. Urbanization is the change of the society that is shifting of people from rural areas to city areas. This data set of Urban Land Cover is classification dataset. In this research study, Naive Bayes performed very well with highest accuracy and Kappa values and lowest rate of classification error while C4.5 decision tree remains on second number, Random Forest on third level and K-NN on last level. Therefore, Naive Bayes is suggested for further studies on these types of datasets.

REFERENCES:

Bhargavi, P., and S. Jyothi, (2009). Applying naive Bayes data mining technique for classification of agricultural land soils. International journal of computer science and network security, 9(8), 117-122.

Delen, D., G. Walker, and A. Kadam, (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.

https://archive.ics.uci.edu/ml/datasets/Urban+Land+Cover#

Johnson, B., and X. Zhixiao. (2013) "Classifying a high resolution image of an urban area using super-object information." ISPRS Journal of Photogrammetry and Remote Sensing 83, 40-49

Kuramochi, M., and G. Karypis, (2005). Gene classification using expression profiles: a feasibility study. International Journal on Artificial Intelligence Tools, 14(04), 641-660.

Lee, W., S. J., Stolfo, and K. W. Mok, (1999). A data mining framework for building intrusion detection models. In Security and Privacy. Proceedings of the IEEE Symposium on 120-132.

Prekopcsak, Z., G. Makrai, T. Henk, and C. Gaspar-Papanek, (2011). Radoop: Analyzing big data with rapidminer and hadoop. In Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM) 865-874.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. (2012) "An assessment of the effectiveness of a random forest classifier for land-cover classification." ISPRS Journal of Photogrammetry and Remote Sensing 67, 93-104.

Srivastava, A., E. H. Han, V. Kumar, and V. Singh, (2002). Parallel formulations of decision-tree classification algorithms. 237-261. Springer US.