International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 11 • Issue 2 June - Dec 2019 pp. 109-114 www.csjournals.com

Comparative Analysis of Different Classification Methods for Data Mining

Niharika, Naresh Kumar
Department of Computer Sc. and Engg, UIET, Kurukshetra University, Kurukshetra, INDIA

Abstract: There are many state of the art classification models that have already been defined in the field of data mining. With whopping amount of data being generated in the modern era, there is a continuous need to analyze data and discover the interesting patterns. This paper gives the comparative analysis of different classifier methods based on the correlation coefficient, mean absolute error, root mean squared error, relative absolute error, root relative squared error and time taken by that method which reveals that reduced error pruning takes the least amount of time while the ensemble method 'random forest' has the highest score when the analysis is done on the titanic dataset.
Keywords: Data mining, classification, WEKA, decision tree, reduced error pruning, linear regression, random forest.

I. INTRODUCTION

DATA mining means to discover the hidden patterns and it has a great potential to be an efficient means to discover something is interesting as well as knowledgeable [3]. The data mining process includes steps such as preprocessing, association, classification and clustering of data. This paper gives the comparative analysis of different classifiers in data mining.

A. Classification:

Classification is the step to predict the class labels of the data. In classification step, the dataset itself has two properties i.e. attributes and class labels. Classification is a process which is comprised of two steps namely, learning step and classification step. The construction of classification model is done in the learning step while the prediction of various class label is done in the classification step. The various classification methods are as follows:

1. Decision tree:

A decision tree is a classification model which has a tree like structure and contain the following parts:

i) Decision node:

The decision nodes are the internal nodes of the tree and are rectangular in shape. These nodes represents the test on the attributes.

ii) Leaf nodes:

The leaf nodes are the terminal nodes of the decision tree and are oval in shape. These nodes represents the outcome or the class labels for the defined attribute.

iii) Arc/Edges:

An arc on node is the line that denotes the split for the particular node.

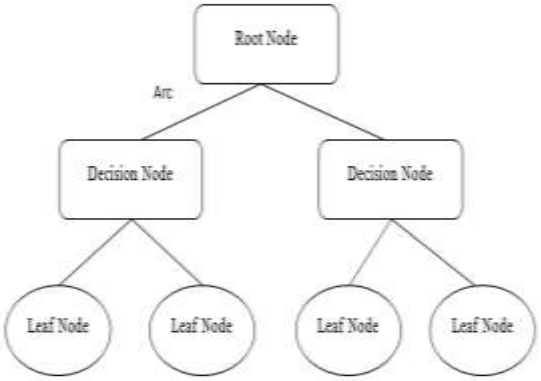

Fig.1. Decision tree representation

To construct the decision tree, it has two phases [1]:

1) Tree building phase:

In the tree building phase, the tree is constructed and include cross validation. The cross validation is the technique where the data-set is partitioned into k-number of mutually exclusive sets where, k-1 sets are the training sets and one set is the test set.

2) Pruning phase:

Sometimes, while modeling the decision tree , noise or irrelevant information is also considered and included in the construction and is called over-fitting. To alleviate this conundrum , the solution is pruning. Pruning is also of two types:

i) Post pruning:

Post-pruning is done after the completion of the decision tree modeling.

ii) Pre-pruning:

Pre-pruning is also called early pruning. Pre-pruning is done before the completion of the modeling of the decision tree.

Reduce error pruning:

The technique of reduced error pruning is implemented in bottom-up fashion where the error rate of child node is compared with the error rate of the parent node and then it is decided whether to prune the decision tree or not [5].

It is denoted as:

If (parent_node(error_rate)<child_node(error_rate)):
  Then prune the decision tree
else:
  Do not prune the decision tree

B. Linear Regression:

Linear regression is the classification technique where the classification is denoted by a line in the x-y plot where x is the predictor variable and y is the response variable [4].

C. Random forest:

Random forest is one of the ensemble classification method which is a collection of several decision trees. Random forest method is used to improve the overall efficiency of the method [2].

II. RELATED WORK

The comparison of various classification methods for mining the medical dataset is done in [7] which describes the accuracy for various classification methods such as Naïve Bayes, Decision tree and ANN. The result depicted that the decision tree method has the highest accuracy of 89% followed by Naïve Bayes method with accuracy 86.53%.

III. PROPOSED WORK

This paper analyzes the titanic dataset [8] in WEKA (Waikato Environment for Knowledge Analysis) is a data mining tool [6,10]. . The attributes of the titanic dataset are as depicted in the table.

Table.1. Input attributes of titanic dataset

Input attributes
1. Survived
2. Sex
3. Age
4. Fare
5. Pclass_1
6. Pclass_2
7. Pclass_3
8. Embarked_0
9. Embarked_1
10. Embarked_2
11. Family Size

Weka's main interface for all of its user is the Explorer [11], but the knowledge flow explorer and command line can provide the same functionalities to the user as the explorer. The Explorer interface includes various tabs to provide the access to the main components of the weka workbench[10]:

1) Preprocess:
   The Preprocess tab has the feature to enable the user to import data from a arff file, a database, a CSV file, etc..

2) Classify
   The Classify tab enables the user to select and apply the classification algorithms to the dataset selected in the preprocess tab, to estimate the accuracy of the resulting predictions, and to visualize erroneous predictions .

3) Cluster
   The Cluster tab gives the users access to the clustering, e.g., the simple k-means algorithm.
4) Associate:
   The Associate tab provides access to association rule that attempt to identify all important interrelationships between attributes in the data.

5) Select:
   The Select attributes tab includes the algorithms for identifying attributes that are most predictive in the dataset.

6) Visualize:
   The Visualize tab shows the visualization of the data-set such as scatter plot matrix.

The comparison of various classifiers based on the several factors such as correlation coefficient, mean absolute error, root mean squared error, relative absolute error, root relative squared error and time taken by that method by using 10 cross validation to make a prediction for the dataset by using WEKA .

i) Time:
   Time taken by that algorithm
ii) Score:
   Accuracy score for the algorithm
iii) Correlation coefficient:
   Correlation coefficient denoted the strength of relationship between two variables. The value of correlation coefficient lies between -1.0 and +1.0.
iv) Mean absolute error (MAE) :
   Mean absolute error (MAE) is the average of values of errors in the set of prediction [10].
v) Root mean squared error (RMSE):
   Root mean squared error (RMSE) is the standard deviation of the error in the set of prediction[10].
vi) Relative absolute error (RAE):
   Relative absolute error is the absolute error divided by The corresponding error of the classifier predicting the prior probabilities of the class [10].
vii) Root relative squared error:
   Root relative squared error is the root mean squared error (RMSE) divided by the root mean prior squared error (RMPSE) [10].

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 11 • Issue 2 June - Dec 2019 pp. 109-114 www.csjournals.com

IV. RESULTS

The results obtained are depicted in Table 2.

Table.2. Comparative analysis of different classifiers

Algorithm used	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Time (s)
Linear Regression	0.4003	0.9999	1.4805	91.64%	91.77%	0.22
Random forest	0.8994	0.4107	0.7063	37.64%	43.78%	0.63
Decision tree	0.8303	0.5165	0.906	47.33%	56.16%	0.22
Reduced error pruning	0.6736	0.4296	0.7897	39.38%	48.95%	0.05

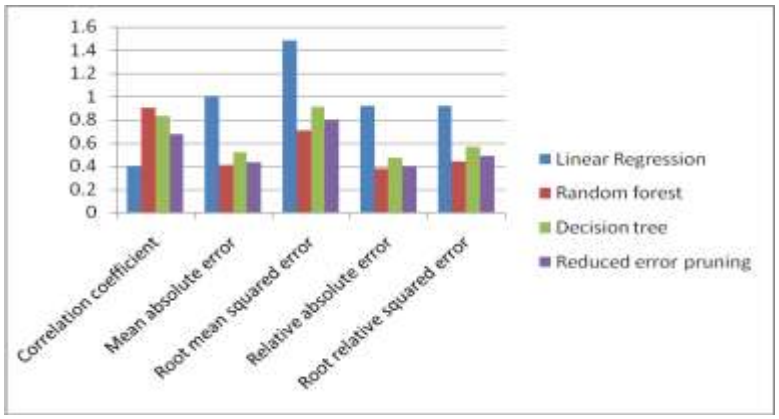

Fig.2. Bar graph for the error comparison

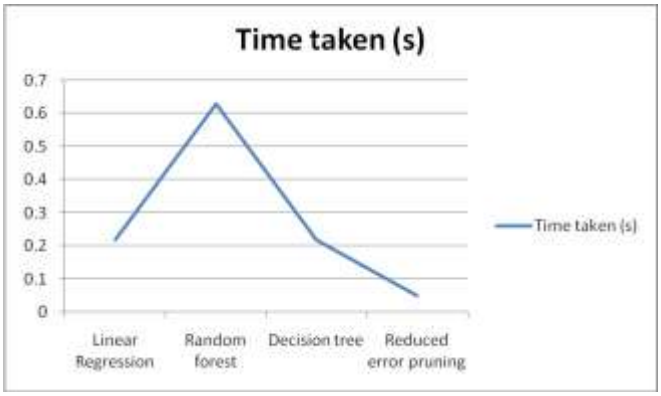

Fig.3. Line graph for the comparison

It can be analyzed that the reduced error pruning method takes the minimum time of 0.05 second while the random forest method has the minimum mean absolute error, root mean squared error, relative absolute error, root relative squared error. The linear regression method has the lowest correlation coefficient while random forest has the highest correlation coefficient. Decision tree and linear regression ,both the methods , took equal time of 0.22 second.

From the bar graph, it is obtained that random forest has the highest accuracy_score among other classification method which is followed by linear regression.

The comparison analysis of accuracy score of the classification methods is as follows:

Table.3. Comparative analysis of Accuracy_score

Algorithm Used	Accuracy_score
Linear regression	80.22
Random Forest	81.34
Decision Tree	76.49

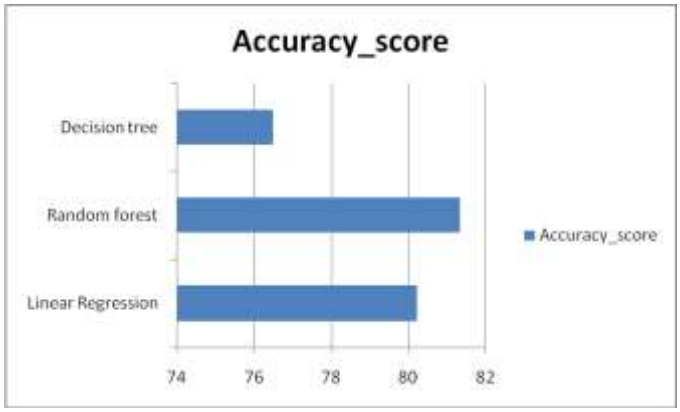

Fig.4. Bar graph for the Accuracy_score

V. CONCLUSION

The selection of the classification technique is dependent on the number of factors such as correlation coefficient, mean absolute error, root mean squared error, relative absolute error, root relative squared error and time taken by that method. Based on the comparison analysis of above classification methods, corresponding classification technique can be chosen such as if time is main consideration then the reduced error pruning technique should be opted for.

VI. ACKNOWLEDGMENT

First and foremost, praises and thanks to the God, the Almighty, for his showers of blessings throughout my research work to complete the research successfully. I would like to express my deep and sincere gratitude to my research supervisor, Dr. Naresh Kumar, Assistant Professor, Department of Computer Science and Engineering , University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, Haryana for giving me the opportunity to do research and providing invaluable guidance throughout this research. His dynamism, vision, sincerity and motivation have deeply inspired me. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honor to work and study under his guidance. I am extremely grateful for what he has offered me. I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future.

REFERENCES

[1] Mansour Y. "Pessimistic Decision Tree Pruning Based on Tree Size " Press of Proc. 14th International Conference on Machine Learning pp.195- 201, 1997

[2] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling ", J. Chem. Inf. Comput.Sci. Vol. 43(6) , pp. 1947-1958 ,November 2003

[3] Polaka, Inese & Tom, Igar & Borisov, Arkady (2010), "Decision Tree Classifiers in Bioinformatics " Scientific journal of J. Riga Technical
University, Vol. – 42(1) pp- 118-123., January 2010

[4] I. Naseem, R. Togneri and M. Bennamoun , "Linear Regression for Face Recognition," IEEE Transactions on Pattern Analysis and Machine
Intelligence, vol. 32(11), pp. 2106-2112, 2010

[5] Mohamed, W.N.H.W., Salleh M.N.M., Omar, A.H.(2012). "A Comparative Study of Reduced ERROR Pruning Method in DecisionTree Algorithms " IEEE International Conference On Control system Computing and Engineering, pp.393-397, November 2012
[6] WEKA at [http://www.cs.waikato.ac.nz/~ml/weka].
[7] Soni, J., Ansari, U., Sharma, D., & Soni, S." Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction."
International Journal of Computer Applications, vol .- 17(8), pp. 43– 48, March 2011
[8] Dataset at [https://www.kaggle.com/c/titanic]
[9 Nasa, Chitra. "Evaluation of Different Classification Techniques for WEB Data." International Journal of Computer Applications , vol.-52(9), pp.
34-40,August 2012.
[10] Jayakameswaraiah, M." Design and Development of Data Mining System to Estimate Cars Promotion using Improved ID3 Algorithm."
International Journal of Advanced Research in Computer and Communication Engineering Vol. 3(9) , pp. 8052-8061 ,September 2014
[11] http://en.wikipedia.org/wiki/Weka_%28machine_learnig %29

AUTHORS

Dr. Naresh Kumar is faculty in Computer Sc. and Engineering branch at University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra (India). He has B.E, M.Tech and Ph.D in the discipline of Computer Sc. and Engineering. (Mobile : +91-94670-12567; E-mail : naresh_duhan@rediffmail.com)

Niharika is M.Tech Computer Engg. student at University Institute of Engg. and Technology, Kurukshetra University, Kurukshetra (India) and holds B.Tech in Computer Sc. and Engg. from Geeta Inst. of Engg. and Tech., Kurukshetra University, Kurukshetra (India) . Her research interests include machine learning methods, classification methods, supervised learning , unsupervised learning and big data analytics.