International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 5
May 2018

Technique to Improve the Accuracy of a Decision Tree Induction by using Data Mining Extort Boosting Algorithm

Mrs. K. Laxminarayanamma
Assoc. Professor
Department of Information Technology,
Institute of Aeronautical Engineering
Dundigal Vill., Medchal Dist.,
Hyderabad,Telangana,India

Mr. Yannam Apparao
Assoc. Professor
Computer Science and Engineering Dept.,
Marri Laxman Reddy Institute of Technology and Management,
Dundigal Vill., Medchal Dist.,
Hyderabad,Telangana, India

Abstract: *In this paper, we introduce a technique to improve the accuracy of a decision tree by using new data mining extort boosting algorithm. Which a large database has huge amount of records is analyzed and predicted to retrieve useful information and to make decisions, we analyze the data classify it based on our requirement.*

Extract Boosting is one of the ways to improve the accuracy of a decision tree induction. In this proposal initially weights are assigned to each of the training tuples. After the classifiers are learned, the weights are updated such that the subsequent classifier gives more attention towards the tuples which were previously missed out. The final classifier is the combination of votes of each individual classifier.

Keywords: Decision Tree Induction, Boosting, extort Boosting, Adaboost, Misclassification error of tuple. Data Partition,

I. INTRODUCTION

Data mining is a process of entraining knowledge from massive volume of data, It refers to a way of finding significant and useful information from an organization's Database.

The following are the reasons for using data mining [1].

1. Knowledge discovery
2. Data visualization
3. Data correction

Knowledge discovery: The objective of knowledge discovery process is to identify the invisible correlation, patterns, and trends available in the database [1].

Data visualization: The objective of data visualization is to "harmonize" large volume of data so as to find a sensible way of displaying data [1].

Data Correction: Process is used to identify and correct incomplete, erroneous, inconsistent data [1].

Classification and prediction can be viewed as two kinds of data analysis that is used to.

i) Retrieve models that describe important data classes.
ii) Ii) To predict future data trends.

Classification is analyzed and predicted to retrieve useful information and to make decisions. Classification is one of the methods used for data analysis.

II. EVALUATING THE ACCURACY OF A CLASSIFIER

The techniques used for estimating classifier accuracy are.

a) Holdout technique
b) K-fold cross validation technique
c) Bootstrapping technique

Holdout Technique: The holdout technique is one of the commonly used techniques for estimating classifiers accuracy, In this technique, the given data are randomly fragmented into two independent sets called the training set and the test set[2]. The data fragmentation is carried out in such a way that the training set contains two-thirds of the entire training data and the test set contains the remaining one-third of the data. The data in training set is used to derive

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 5
May 2018

the classifier, whereas the data in the test set is used to estimate the accuracy of the classifier. As the classifier is derived using only a portion of initial data, a pessimistic estimate is obtained[2].

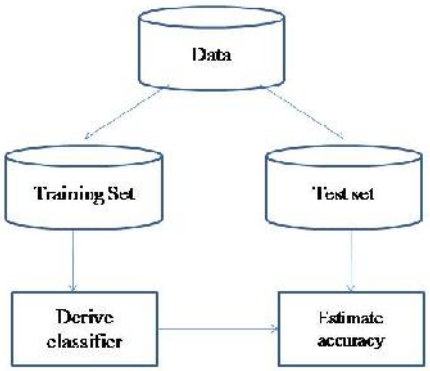

Figure 1: Holdout Technique

The Estimation obtained using holdout method can be made more certain by repeating the holdout method 'n' times and computing the average of the n accuracies to obtain the overall accuracy estimate. This modified hold-out technique is called Random sub sampling technique [2].

K-fold Cross Validation Technique:The K-fold cross validation technique is one of the commonly used techniques for estimating classifier accuracy. In this technique, the given data are randomly fragmented into **K** equal sized mutually exclusive folds or subsets $\mathbf{S_1, S_2, S_3,\ldots, S_K}$. The classifier is trained using K-1 subsets and is tested using a subsets $\mathbf{S_i}$ in the $\boldsymbol{i^{th}}$ iteration. For example, in the first iteration, the classifier C_1 is trained using the subsets $\mathbf{S_2,S_3,\ldots,S_K}$ and is tested on the subset S_1. Similarly in the second iteration, the classifier C_2 is trained using the subsets $\mathbf{S_1, S3, S4,\ldots, S_K}$ and is tested on the subsets $\mathbf{S_2}$.Thus ,in this technique, **K** training and $\mathbf{L_K}$ testing are performed ,resulting in **K** classifications,$\mathbf{C_1,C_2,\ldots,C_k}$. The accuracy is estimated by computing the ratio of the total number of correct classifications obtained from **K** iterations, and the total number of samples in the given data.

$$\text{Accuracy estimate} = \frac{\text{Number of correct classification obtained from K iteration}}{\text{Number of samples in the given data}}$$

Generally, a 10-fold cross validation technique is used to estimate classifiers accuracy. Alternatively, the stratified cross-validation technique can also be used, which is a modified K-fold cross validation technique. In this technique, the subsets $\mathbf{S_1, S_2, S_3,\ldots,S_K}$. are stratified(or arranged) in such a way the class distribution of the initial data, and the samples in each subset are approximately the same.

Bootstrapping Technique: In the boot strapping technique, the given training tuples are sampled uniformly with replacement.

That is, a training instance selected once is repeatedly chosen and added to the training set. It also applied leave-one-out (i.e., a K-fold cross – validation with sets K to S) technique to the initial samples.

III. Bagging And Boosting

Bagging is one of the methods used for increasing the accuracy of the classifier, In this method, the training set $\mathbf{DP_j}$ sampled by replacing certain tuples in data partition with original tuples. Hence the $\mathbf{DP_j}$ may not have all tuples as in original data partition DP and may also contain repetitive tuples. For this new data partition, a new model M_j is generated.

To classify a tuple, we use a newly generated model M_j. each model returns its own prediction. The prediction which is made by majority of the model is considered as appropriate class for the tuple.

IV. EXTORT BOOSTING ALGORITHM

Boosting is one of the ways to improve the accuracy of a decision tree induction.

Initially weights are assigned to each of the training tuples. After the classifiers are learned, the weights are updated such that the subsequent classifier gives more attention towards the tuples which were previously missed out. The final classifier is the combination of votes of each individual classifier.

"Adaboost" is the, most commonly used algorithm. It initially assigns equal weights to the data tuples in DPA training set Dp_j is later generated by sampling the tuples in the original data set DP. Sampling is done by replacement which depends on the weight of the data tuples. A model Mj is created from the training set Dp_j.

The error is then calculated. The weights of training tuples are adjusted based on the calculated error. If the tuple is wrongly classified, then its weight is increased and if it is correctly classified its weight is decreased .In the next pass, the tuples with the higher weight are concentrated .The error rate is calculated using.

$$E(M_j)= \sum_j^d wjMS(xj)$$

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 5
May 2018

Where MS(x_j) is the Misclassification error of tuple, x_J. If it is misclassified its value is 1 else it's 0.If the error rate of a model. We generate another new model Mj.

If a tuple is correctly classified then its weight is multiplied by E(M_j)|(1-E(M_j)). After this, the weights are normalized by multiplying it with the sum of old weights, divided by the sum of new weights, In this way, he weights of misclassified tuples are increased and the weights of correctly classified tuples are decreased.

In Beginning we assign equal votes to all classifiers, but in booting votes depend on the error rate,

$$\text{Log}\ \frac{1-E(M_j)}{E(M_j)}$$

IV. ALGORITHM

Input

- Do, Data partition with certain data tuples
- N, number of models
- A learning scheme

Output

An accurate composite model.

Method

1. Initize the weight of each tuple to 1/d
2. For j=1 to n
3. Sample Dp with replacement for each tuple and derive a boosting sample DP_j.
4. Derive a model M_j from DP_j
5. $E(M_j)= \sum_j^d wjMS(xj)$
6. If E(Mj)>0.5
7. Discard M_i and reinitialize the weights to 1/d
8. and goto setp 3.
9. end if
10. for all correctly classified tuples do.
11. Multiple the weights by

E (M_j) | (1-E(M_j))

12. Normalize the weights
13. End for

To classify an Unknown Tuple, x

1. Initialize all weights to 0.
2. For j=1 to n

$$\text{Log}\ \frac{1-E(M_j)}{E(M_j)}$$

3. e=M_j(x)
4. Add weight W_j to the weight for class 'e'.
5. end for
6. Return class with maximum weight.

For example: if we had 3 training instances with the weights 0.01, 0.5 and 0.2. The predicted values were -1, -1 and -1, and the actual output variables in the instances were -1, 1 and -1, then the terrors would be 0, 1, and 0. The misclassification rate would be calculated as:

error = (0.01*0 + 0.5*1 + 0.2*0) / (0.01 + 0.5 + 0.2)

or

error = 0.704

A stage value is calculated for the trained model which provides a weighting for any predictions that the model makes. The stage value for a trained model is calculated as follows:

stage = ln((1-error) / error)

Where stage is the stage value used to weight predictions from the model, ln() is the natural logarithm and error is the misclassification error for the model. The effect of the stage weight is that more accurate models have more weight or contribution to the final prediction.

Conclusion:

In this paper, we introduce a technique to improve the accuracy of a decision tree by using new data mining extort boosting algorithm. By using EXTORT BOOSTING ALGORITHM we can retrieve the records from large amount of data is analyzed ,predicted and and to make decisions, we analyze the data classify it based on our requirement.

References:

[1] Mrs. Bharati M. Ramageri, "Data Mining Techniques And Applications," Indian Journal of Computer Science and Engineering, Vol. 1 No. 4, pp. 301-305, Available : http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf

[2]. Evaluating Data Mining Models: A Pattern Language Jerffeson Souza∗ Stan Matwin Nathalie Japkowicz School of Information Technology and Engineering

University of Ottawa K1N 6N5, Canada {jsouza,stan,nat}@site.uottawa.ca.

[3]. A Study Of Bagging And Boosting Approaches To Develop Meta-Classifier G.T. Prasanna Kumari Associate Professor, Dept of Computer Science and Engineering, Gokula Krishna College of Engg, Sullurpet-524121, AP, INDIA, Email:tabi_prasanna@yahoo.com

[4]. Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning, 40, 139–158.

[5] Alan Fern and Robert Givan, "Online Ensemble Learning: An Empirical Study," The Seventeenth International Conference on Machine Learning, Stanford, CA, pp. 279-286, July 2000.

[6] Herbert K.H. Lee and Merlise A. Clyde, "Lossless Online Bayesian Bagging," Journal of Machine Learning Research, Vol. 5, pp. 143-151, 2004.

[7] Nikunj C. Oza and Stuart Russell, "Online Bagging and Boosting," in Artificial Intelligence and Statistics 2001, Key West, FL, USA, pp. 105-112. January 2001.