International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Improving Decision Tree Forest using Preprocessed Data

Archana R. Panhalkar, Dharmpal D. Doye

Abstract: ***Random forest is one of the best techniques in data mining for classification. It not only improves accuracy of classification but performing best for various data types. Data mining researchers concentrated on improving random tree forest by constructing trees by using various methods. In this paper, we are improving decision forest by applying various preprocessing techniques. Decision tree forest is created by using bootstrapped samples. Trees created using preprocessed data improves not only accuracy of classification but also improves time required to construct forest. Experiments are carried out on various UCI data sets to show better performance of our proposed system.***

Keywords: Random Forest, Decision Tree, C4.5, CART, Forest PA.

I. INTRODUCTION

In today's digital world, large amount of data get generated and stored. To process, analyze and mine useful information from stored data, effective techniques are addressed by researchers of data mining. Data mining plays very important role to perform various operations on data. To classify such a huge amount of data, various methods are proposed in literature and perform best. Decision tree is one of the promising classifier for accurate decisions from huge amount of data. Random forest is one of the decision tree classification techniques for promising decisions from data. Random forest is ensemble of decision trees. It produces diverse pool of tree classifiers for classification of data.

Decision trees are a type of model used for both classification and regression. Answer of decisions is taken by traversing tree to get accurate answer. The model acts with "if then else" conditions eventually yielding exact outcome. The powerful advantage of decision tree is that it is easy to interpret and gives straightforward visualization. A random forest is a collection of diverse decision trees whose results are cumulative into one final decision. Ability of random forest is to limit overfitting without significantly increasing error due to bias is why these contributes powerful models. Producing accurate and diverse decision trees is one of the important areas of research.

In the proposed, we are creating decision forest by constructing number of decision trees such that each tree is more accurate and diverse. Preprocessing is one of the best way to transfer data from information to knowledge. There are various techniques to preprocess data such as noise removal, reduction of data and normalizing data. We are creating decision tree forest by using ForestPA [1] method. We are preprocessing data in such a way that it produces accurate and diverse decision trees.

The remaining section of proposed paper is organized as follows. Section II gives brief idea about existing decision forest creation methods that outperforms best in literature. Section III explains our proposed approach. Section-IV shows evaluation of results and comparison with existing methods. In last section we had given some concluding remarks.

II. LITERATURE REVIEW

There are many decision tree forest methods are proposed in literature which gives best results for different data sets. Some of the following methods outperforms to classify data are discussed in brief.

Bagging: Bagging [4] is the method of generating bootstrap samples. Bootstrap sample is a new data set created by randomly choosing records from original dataset. These records are selected once or some records are selected multiple times. From each bootstrap sample, one decision tree gets created. Due to bagging we can create diverse decision trees. Dagging [5] and Wagging [6] are two variants of bagging to create samples.

Boosting: Generating weighted training samples is called as boosting[7]. In boosting, weights are assigned to training samples while building each decision tree and more penalty is given to misclassified records than correctly classified records. So accuracy and diversity is increased as compared to bagging. Adaboost is prominent example of boosting but not suitable for multiclass classification problems.

Random Subspace method: In bagging, records get divided to form new datasets but in random subspace [8]. Random subspace algorithm randomly draws subset of attributes from whole attributes of data. Subspace of attributes is selected at the node level or at the tree level. Drawback of random subspace is that each tree has only subset of attributes so may misclassify the data.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.

* Correspondence Author
Archana R. Panhalkar*, Department of Computer Science and Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nande, India. Email: archana10bhosale@rediffmail.com
Dharmpal D. Doye, Department of Electronics and Telecommunication, Shri Guru Gobind Singhji Institute of Engineering and, Technology, Nanded, India. Email: dddoye@yahoo.com

Random Forest: Random forest [9, 10] is combination of bagging and random subspace method. As compare to other decision tree forest creation methods, it produces accurate results. It is also used to give promising results on imbalanced data sets. An ensemble random forest model is inherently less interpretable than an individual *decision* tree. Training number of trees can have large computing costs (but can be computed parallel) and it uses a more area of memory.

Weighted Random Forest: In Random Forest only a random subset of attributes are considered for each node, but considers all of them. In weighted random forest[11] all nodes in a tree use the same set of randomly generated weight different weights than other tree. So, the significance given to the attributes will be diverse in each tree and that will make a distinction in creation.

Forest CERN: Forest by Continuously Excluding Root Node (Forest CERN) [12] is one of the best method to create diverse and accurate trees. It attempt to leave out attributes that take part in the root nodes of earlier trees. This is done by assigning unfavorable weights. So it creates more diverse trees. Drawback of Forest CERN is that it may create poor trees due to excluding high gained nodes.

ForestPA: Decision Tree Forest by penalizing attributes [1] is best and powerful method to create diverse and more accurate decision trees. It uses all attributes and creates 100 bootstrap samples of original training dataset. It assigns penalties by assigning more weights to those attributes that involved in the previous decision tree. ForestPA uses novel random weight generation method which is exponent based which gives larger penalties to the nodes at higher level and less penalties to the attributes at higher levels. So because of all reasons the individual trees produced are more accurate and diverse. The main limitation of this method is that it consumes more time in creating decision trees and data is not preprocessed.

Main problems in decision tree forest are computational overhead and take large memory. Generated trees not use any preprocessing methods. In proposed method we are trying to overcome the above drawbacks by preprocessing data and creating decision tree forest which is more accurate and requires less time in some datasets.

III. PROPOSED SYSTEM

In the proposed system we are creating 100 trees for each training datasets. We are preprocessing data such that there is improvement in decision tree forest. Standardization and normalization methods are used to preprocess data. Normalization is also called as Min-Max scaling method Normalization means to scale a variable's data to values between 0 and 1, while standardization is a method that transforms data to contain a mean of zero and a standard deviation of 1. This type of standardization is called a **z-score**, and data points can be standardized with the following formula (1):

$$X_i = \frac{X_i - \bar{X}}{S} \qquad (1)$$

Where:

- x_i is a data point (x_1, $x_2 \ldots x_n$).
- $\bar{x}$ is the sample mean.
- s is the sample standard deviation.

Z-scores are commonly used in statistics to evaluate diverse sets of data and to find likelihoods for sets of data using standardized tables (called z-tables). Various data sets alternatively use normalization and standardization. It not only increases accuracy but also decreases computation overhead. Here the proposed architecture of improved decision tree forest is shown in fig 1. We constructed trees using ForestPA [1] algorithm and to form random forest the concept of data preprocessing applied such as standardization or normalization based on data sets. Following are the steps used for improved decision tree forest proposed algorithm.

A. Take Input data

Read data set and divide data into training and testing datasets.

B. Preprocessing of Input data

Perform standardization and normalization on input training data set. The normalized Training data TrainDN and TrainDS is created for improving performance.

C. Create Bootstrap samples

Create 100 bootstrap samples of each training datasets TrainDN and TrainDS. Main motivation to create bootstrap samples is to increase diversity of the decision trees. More diverse trees get accurate results.

D. Generating decision trees from bootstrap sample

For each bootstrap sample i, create a weighted decision tree Treei by penalizing attributes in previous trees and attributes at lower level in trees using ForestPA algorithm [1]. To create decision tree from each sample we are using CART algorithm [13].

In CART, to select best attribute for splitting, GINI index is used but in this algorithm we are using merit value which is calculated by multiplying GINI index with selected attribute's weight.

E. Update weights and build next tree

We are using weight strategy of ForestPA algorithm. Gradual weight increment strategy is applied to the attributes in the latest tree. The weights are assigned not only to the attributes on the latest tree but also small weight increment is done to the attributes that does not appear latest tree.

F. Evaluating Decision trees

Decision forest is created for both preprocessed datasets $TrainD_N$ and $TrainD_S$. Then we are applying testing dataset and select best results from both dataset.

IV. EXPERIMENTAL RESULTS

The proposed method is applied on well known data sets that are freely available from the UCI Machine Learning Repository [14]. The diabetes, Balance scale, Flags color and Lung cancer data sets are used for experimentations.

In our experimentation, we implement Bagging (BG) [4], Random Subspace (RS) [8], Random Forest (RF) [9], and ForestPA algorithm [1]. We have taken datasets into .arff format. Ensemble Accuracy (EA) is used to calculate performance of decision tree forest algorithm [15, 16, 17, 18]. In Table I we present the EA result in percentage of Forest PA and proposed method for all data sets considered. We used 10-fold cross validation method to verify our results. To construct decision tree forest here it is used same normalized and standardized data set for training and testing. From Normalized and standardized datasets, for each fold we have created 100 trees. This proposed method not only preserves diversity but also it increases Ensemble accuracy. The increases Ensemble accuracy calculated w has been given in Table I.

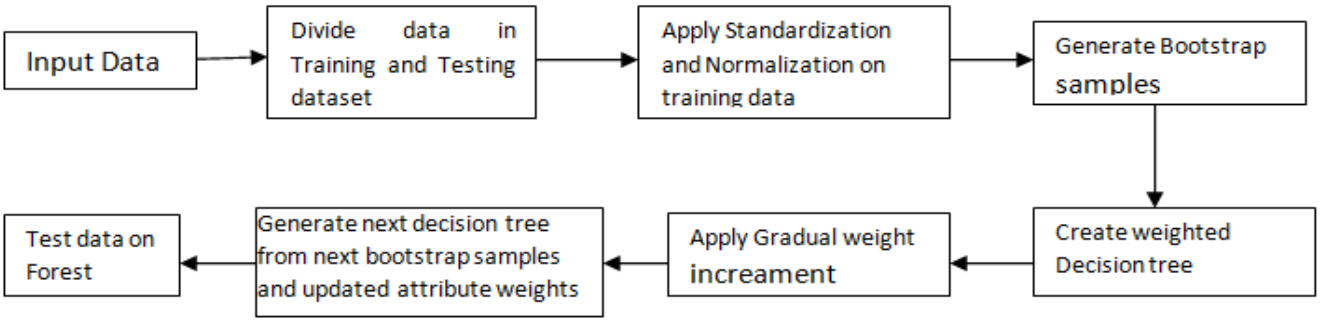

Fig. 1. Proposed architecture of preprocessed decision tree forest

Table I specifies that comparisons between the proposed system of Normalization and standardization and existing system of ForestPA has been shown. It is observed that slightly more ensemble accuracy is produced through our proposed system algorithms.

TABLE-I: Ensemble Accuracy Results

ENSEMBLE ACCURACY (%)			
DATASET	FORESTPA	PROPOSED SYSTEM	
		STANDARDIZATION	NORMALIZATION
Diabetes	75.26	74.86	75.92
Balance Scale	81.44	81.86	81.86
Flags	60.3	62.88	60.20
Lung Cancer	46.87	50	56.25

From Table-I, our proposed preprocessed dataset shows comparable better accuracy than ForestPA algorithm. One of the main drawbacks of ForestPA is that there is large computational overhead and it is due to use of unprocessed data ForestPA requires more computational time. This drawback is also removed using proposed system with preprocessing. Table-II shows improvement in time for the given datasets.

TABLE-II: Average Time in Millisecond

TIME(MILISECOND)			
DATASET	FORESTPA	PROPOSED SYSTEM	
		STANDARDIZATION	NORMALIZATION
Diabetes	4194	**4087**	4178
Balance Scale	3249	3155	**3110**
Flags	4404	**4235**	4367
Lung Cancer	1465	1397	**1359**

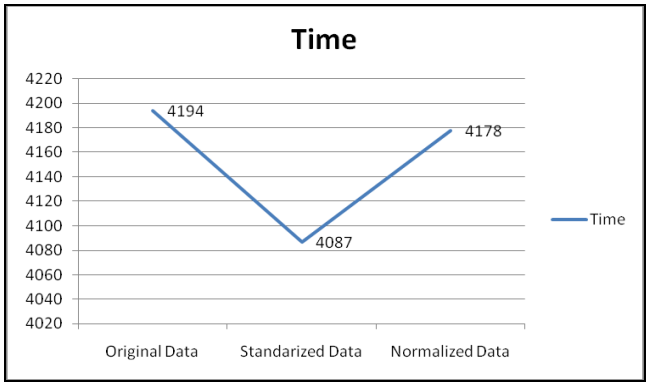

Fig. 2. Time Required for ForestPA, standardization and normalization (ms)

From Fig.2 We can say that there is considerable improvement in time to create Decision tree forest using ForestPA and proposed algorithm using normalization and standardization preprocessing methods. We carried out experiments on four datasets.

We can apply same method for other standard UCI datasets. From experiments, it is clear that standardized and normalized data always increases accuracy of decision tree forest and decreases time required for computation.

V. CONCLUSION

Decision tree Forest is one of the promising classifier which gives better accuracy than other classification methods. To construct decision tree forest, assignment of weights to attributes and use of bootstrap sampling increases accuracy of classifier. In the proposed method, we implemented weighted decision tree forest with preprocessed data. From Experimental results it is clear that preprocessing techniques always improves performance of classifiers. This method also reduces time required to create decision tree forest. There is considerable improvement in time is shown from experiments. So we can conclude that preprocessed data never decreases performance, on the other way it will either keep the performance same or increases performance of classifier.

REFERENCES

1. Adnan, Md Nasim, and Md Zahidul Islam. "Forest PA: Constructing a decision forest by penalizing attributes used in previous trees." Expert Systems with Applications 89 (2017): 389-403.
2. L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123–140.Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2006.
3. L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
4. L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123–140.
5. Ting, K., Witten., I.: Stacking bagged and dagged models. In: Fourteenth Int Conf on Machine Learning (ICML07). (1997) 367–375
6. Bauer, Eric, and Ron Kohavi. "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants." Machine learning 36.1-2 (1999): 105-139.
7. Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." In icml, vol. 96, pp. 148-156. 1996.
8. T. K. Ho, The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 832–844.
9. S. Bernard, S. Adam, L. Heutte, Dynamic random forests, Pattern Recognition Letters 33 (2012) 1580–1586.
10. S. Bernard, L. Heutte, S. Adam, Forest-RK: A new random forest induction method, Advanced Intelligent Computing Theories and applications, Lecture Notes in Computer Science 5227 (2008) 430–437.
11. J. Maudes, J. J. Rodriguez, C. G. Osorio, N. G. Pedrajas, Random feature weights for decision tree ensemble construction, Information Fusion 13 (2012) 20–30.
12. M. N. Adnan, M. Z. Islam, Forest CERN: A new decision forest building technique, in: Proceedings of the 20th Pacific Asia conference on Knowledge Discovery and Data Mining (PAKDD), 2016, pp. 304–315.
13. L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth International Group, CA, U.S.A., 1985.
14. M. Lichman, UCI machine learning repository, http://archive.ics.uci.edu/ml/datasets.html, last Accessed: 15/03/2016.
15. M. N. Adnan, On dynamic selection of subspace for random forest, in: Advanced Data Mining and Applications, Lecture Notes in Computer Science, Vol. 8933, 2014, pp. 370–379.
16. M. N. Adnan, M. Z. Islam, A comprehensive method for attribute space extension for random forest, in: Proceedings of 17th International Conference on Computer and Information Technology, 2014.
17. M. N. Adnan, M. Z. Islam, Complement random forest, in: Proceedings of the 13th Australasian Data Mining Conference (AusDM), 2015, pp.89–97.
18. V. Bhatnagar, M. Bhardwaj, S. Sharma, S. Haroon, Accuracy-diversity based pruning of classifier ensembles, Progresses in Artificial Intelligence2 (2014) 97–111.

AUTHORS PROFILE

Archana Panhalkar, Assistant Professor, Amrutvahini College of engineering, Sangamner, India. She has received the B.E., M.E., degrees from Baba Saheb Ambedkar University, Aurangabad, and SRTM University, Nanded, India, in 2003, 2008, respectively. Her main areas of research interest are Data Mining, Decision Tree and signal Processing, She is research scholar in Department of Computer Science and Engineering at Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India.

Dr. Dharmpal D. Doye Ph.D., Professor in Electronics and Telecommunication, His area of research is Digital Speech Processing and Recognition, Electronics Circuit Design, Linear Integrated Circuits, Operating Systems, Neural Networks and Fuzzy Systems, working in Department of Electronics and Telecommunication, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded, India.