Impact of Instance Reduction Filters on Ensembled Decision Tree Classifier

G. Sujatha, K.Usha Rani

Abstract*: Data Mining is the process of discovering data from different perspectives and converts into useful information. Among various Data Mining techniques Classification is the most prominent supervised learning technique in Data Mining. Decision Trees are one among supervised learning techniques, plays vital role and widely used in medical domain to diagnose the problem of the patient. The performance of the Decision Tree classification increases by efficient Data Reduction Techniques. Instance Reduction is one among various techniques for Data Reduction. Supervised and Unsupervised Instance Filters are key approaches for Instance Reduction Filters. In this paper the experiment is conducted through Hybrid Instance Reduction Filters on Tumor Datasets.

Index Terms: Classification, Decision Tree, Instance Reduction Filters, Multiboot, Tumor Datasets.

I. INTRODUCTION

Data Mining is an essential step for extracting information from large data repositories [1], [2]. Classification and Prediction techniques are among the most suitable supervised learning techniques for extraction of hidden information from large volumes of data bases [3]. Classification technique in data mining, classify the data into a set of predefined classes or groups. Classification uses a diversity of algorithms and every algorithm is used to classify the records. Decision Tree (DT) algorithm is most predominant for data classification in Data Mining [4]-[7]. DT algorithm extracts the valuable rules and relationships among the large volumes of data stored in large database. The extracted rules and relationships are controlled in a formation which can be simply implicit by humans [8]. The decisions of the different decision tree learning algorithms combined as an ensemble method to achieve the multi objective function and for better prediction accuracy. Ensemble methods combined multiple learning algorithms to achieve better predictive performance than the individual learning algorithms. The main strategy of ensemble 1approach is to generate many classifiers and integrate the outputs of classifiers such that the combination of classifiers improves the performance of a single classifier [9]-[12]. In the literature survey an ensemble technique is more accurate than any of the individual classifiers. Some of the Ensemble techniques are Bagging [13], Boosting [14], [15], AdaBoost [16] and Multiboot [17] etc.

From our previous study, it was proved that Multiboot ensemble technique with C4.5 Decision Tree classifier (Hybrid Method) is best for Tumor Datasets [18]. Further, it was proved that committee size of 10 for Primary Tumor and committee size of 20 for Colon Tumor having higher accuracy by using the Hybrid Method [19]. The data in real world is highly inconsistent. The advancements in data gathering instruments a huge amount of data collection takes place within no time. Preprocessing plays vital role in any data mining technique's performance. The data reduction is one among various preprocessing techniques. Removal of unnecessary data from the learning classifier enhances the chances of getting more accurate results through the use of instance reduction come with filter methods. These filters are used for add, remove and transform all instances. Further it enhances the classification accuracy and reducing the building time [20]. For verification, in this study experiments are conducted on Instance Reduction Filter methods by using hybrid method on Primary Tumor (PT) and Colon Tumor (CT) Data sets. The performances of Instance Reduction Filters are compared on the basis of classification accuracy and time taken to build a model. The following figure shows the framework for instance reduction and model building after removal of redundant data with filter techniques.

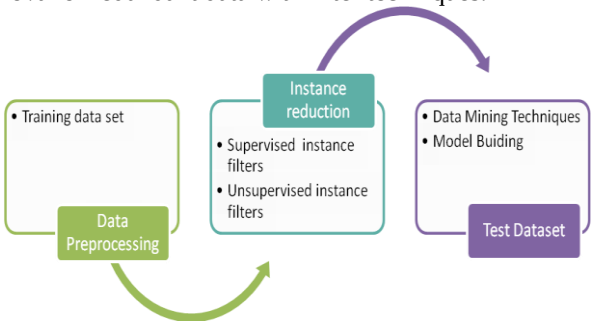

Fig 1: Framework on instance reduction application

The paper organized as follows: Section II contains literature survey on Instance Reduction Filter techniques on classifiers; Section III describes the details of Data Preprocessing and different Instance Reduction Filter techniques; Section IV deals with Experimental Results and section V presents the Conclusion.

Revised Manuscript Received on June 07, 2019.

G.Sujatha, Research scholar, Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati (AP), India .

Prof. K. Usha Rani, Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati (AP), India.

II. LITERATURE SURVEY

Literature survey on Instance reduction techniques with different classifiers are presented in this section.

Instance Reduction Techniques:

Pooja, Saroj Ratnoo [25] proved that the analysis on supervised and unsupervised instance filter methods with MultiBoostAB ensemble techniques on 14 datasets. Filter methods are more accurate and faster than wrapper method when applied on large datasets. In this, resample instance filter gave the best result in terms of classification accuracy and minimum mean absolute error.Bhavya [26] did experiments on the Pima Indian Diabetes Database (PIDD) with different filtering, attribute selection and classification methods by using MultiBoostAB ensemble technique. Among Resample and Stratified Remove Folds instance filters are performed equally on the PIDD.+Faiza Rahat, Syed Ahsan [27] focused on KDDCUP'99 which is a Benchmark dataset for intrusion detection and suffered highly from class imbalance problem. In this Stratified Remove folds and Resampling filter techniques with C4.5 classifier were applied to reduce dataset and to remove class imbalance problem.Pinar Yildirim [28] had done analysis on preprocessing techniques (Resample, Synthetic Minority Over Sampling Technique (SMOTE), Stratified Remove folds and Spread subsample) with classification algorithms on Aalbendazole dataset. The evaluation of results was performed based on accuracy measures and execution time. Among these algorithms, ID3 with resample has higher accuracy results on the dataset than other algorithms.M.Millan_Girlaldo, Vicente Garcia, Josep Salvador Sanchez [29] carried out experiments with 13 datasets taken from UCI repository. In this experiment stratified remove folds, SMOTE, Resample and Random under sampling filter techniques are applied on Fisher and 1-nn classifiers. By using these techniques accuracy of the datasets are improved.Lokesh S.Katore , J.S.Umale [30] did Experiments on Knowledge level dataset with four algorithms namely C4.5, Naïve Bayes, K Star and Simple Cart and compared according to their accuracies. Among these C4.5 algorithm gave better accuracy. Later this dataset was analyzed by using SMOTE and Resample Filters then the accuracy of this dataset was increased.JudickaelBah [31] had done experiments on Vertebral column dataset with Multi Layer Perception (MLP) technique. In this Resample filter technique was used. Results are compared with original and the reduced datasets. Resample technique was used for reducing the dataset and it works with high performance. The literature survey clearly shows instance reduction filter technique is the most promising research direction for ensemble decision tree classification algorithm design.

III. BACK GROUND

This section describes Data Preprocessing and Instance Reduction Techniques about different types of filter which are used in this study.

A. Data Preprocessing:

Data Preprocessing is the first and significant step in Data Mining process, since the performance of the data mining technique purely depends on the quality data. The Preprocessing techniques, once applied before performing Data mining techniques, will considerably improves the quality of the patterns mined and reduce the classifier build time. Data Preprocessing is a broad area which includes data cleaning, instance selection, feature extraction and transformation. Broadly data reduction task includes instance selection and feature selection in data preprocessing. Data Reduction is a process of reducing the volume of the data by removing unnecessary and repeated data but producing the same or similar analytical results.

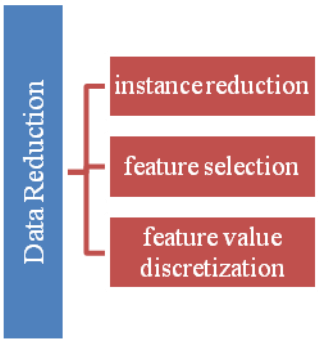

Fig 2: Data Reduction Methods

The removal of unnecessary instances from the original data set can avoid excessive storage, time complexity and also improve the classification accuracy [3, 20]. To verify this fact in this study Instance Reduction Techniques are experimented on hybrid classification method.

B. Instance Reduction Techniques:

Data Mining techniques apply on the enormous amount of data need to prepare the data with data reduction. Instance Reduction is most significant technique in data reduction methods. The removal of the missing and redundant data to achieve the tractable amount of data in Data Mining tasks especially in classification. Instance Reduction aims to reduce the size of the data set mainly to increase the classification accuracy by the removal of noise data and to increase the efficiency of the data set through the removal of redundant instances. The following shows the Data Reduction categories. The instance reduction process starts with input set ***S*** of all data and the output subset ***T*** of inputs consists of chose a subset of total available data to achieve classification accuracy. The ideal instance selection algorithm [21] is to produce the maximum reduction of redundant data from the large volume's datasets which input set is super set of output set means the superfluous data such as noise and redundant data removed to achieve the better accuracy [22,23]. Filter and wrapper methods are most significant methods among various Instance reduction methods. Wrapper methods apply on threshold function to select the features to reduce the redundant data. Filter method is independent of learning algorithm uses the characteristics of data to select and evaluate the features. Filter methods gives the best results if the datasets are very large and consists of large number of selected features and computationally efficient, further filter methods are accurate and faster than wrapper method [22]. So, Filter approach is considered for the experimentation purpose in this study. The below figure shows the most significant filter techniques applied in various classification algorithms. Class distribution is the idea of supervised instance filter technique. Unsupervised instance filter techniques are based on

random values, percentage and sampling methods.

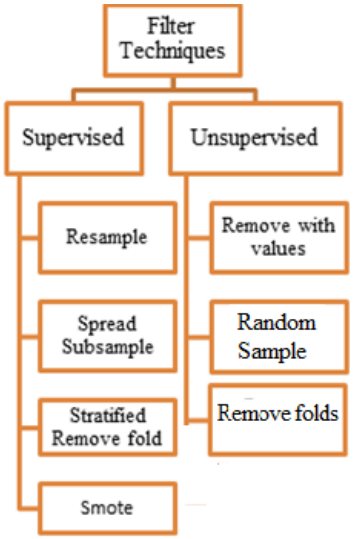

Fig 3: Filter Techniques for Classification

ues are experimented with the Ensemble method.

IV. EXPERIMENTAL RESULT

Primary Tumor (PT) is collected from UCI machine learning Repository [32] and Colon Tumor (CT) is collected from Bioinformatics Group Seville [33], which are openly available. It was proved that Hybrid Method is best for Tumor Datasets from our previous study [18]. Later, it was proved that committee size of 10 for PT and committee size of 20 for CT having higher accuracy by using the Hybrid Method [19]. The experiments are conducted by using by Waikato Environment for Knowledge Analysis (Weka 3.6.4) tool using 10-fold cross validation. The following table shows the characteristics of the Tumor Datasets.

TABLE I - Characteristics of Datasets

Data sets	Primary Tumor(PT)	Colon Tumor(CT)
Attributes	18	2001
Classes	2	2
Instances	339	62
Missing values	Yes	No

If the data set contains missing values or empty entries, those are pre processed. For the preprocessing Instance Filters are used.Instance Filters play a vital role to enhance the accuracy and performance. Hence, Supervised and Unsupervised filters with the Hybrid method are performed on Tumor data. The following TABLE II depicts the experimental results of the classification accuracy.

TABLE II-Accuracy (%) of C4.5 with MultiBoostAB (hybrid method) with and without Instance Reduction Filters

DATA SETS	With out filters	With filters			
		Supervised Instance Filter		Unsupervised Instance Filter	
		Stratified Remove Folds (SRF)	Resample (Res)	Random (Ran)	Remove folds (RemF)
PT	43.07	33.82	**62.53**	42.18	25.00
CT	82.26	84.62	**85.48**	80.61	84.62

The comparison of classifier accuracies on Two Tumor datasets are represented in Fig 4.

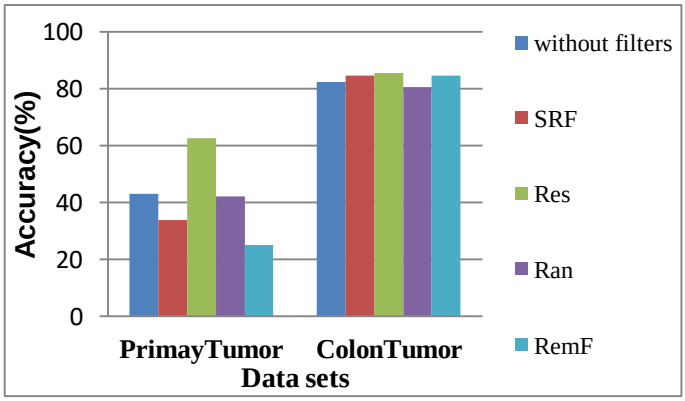

Fig 4: Accuracy (%) of hybrid method with and without Instance reduction filters

From the above figure and table it is clearly observed that Hybrid Method with Resample (Res) filter technique has higher accuracy than other instance reduction filter techniques on both the Tumor Datasets.The time taken by the classifier is also presented in TABLE III.

TABLE III: Execution time (Sec) of Hybrid Method with and without Instance Reduction Filters

Hence we did the comparative study on execution time to build a model with and without different instance reduction filters on Tumor data sets. Here conclude that the execution

DATA SETS	Hybrid Method				
	With out filters	With filters			
		Supervised Instance Filter		Unsupervised Instance Filter	
		Stratified Remove Folds (SRF)	Resample (Res)	Random (Ran)	Remove folds (RemF)
PT	0.27	**0.02**	0.22	0.19	0.03
CT	4.09	**0.02**	1.58	1.97	**0.02**

time with Hybrid method without using filters 0.27 Sec on PT, and 4.09 Sec on CT Datasets. The Stratified Remove folds filter yields lowest build time for both the Datasets.

V. CONCLUSION

Instance Reduction Filters are those in which we can add, remove and transform all instances. In this experiment Supervised and Unsupervised Filters are considered on the selected Tumor Datasets with the proved committee sizes by using Hybrid Method to evaluate the performance regarding Accuracy and Execution Time. From the results, conclude that Hybrid method with Resample Supervised Instance Filter technique gives better accuracy for Primary Tumor and Colon Tumor datasets. By applying the filters we can observe a significant reduction on execution time. Stratified Remove Folds is better for Primary Tumor and as well as for Colon Tumor datasets in the Execution time. But Accuracy is very important for taking decisions in the medical field than Execution Time. Hence Hybrid Method with Resample Instance Filter technique is better algorithm for Tumor Datasets for finding out whether the Tumor is a Benign or Malignant with 10 models for PT and CT with 20 models.

REFERENCES

1. J.Han and M.Kamber, “Data Mining; Concepts and Techniques”, Morgan Kaufmann Publishers, 2001.
2. T. M. Mitchell, “Machine learning and Data mining”, Communication. ACM, vol. 42, no, 11, 1999.
3. M.Venkatadri. C. Lokanatha Reddy, ”A Review on Data Mining from Past to the Future”, International Journal of Computer Applications (0975 -8887),Volume 15-No.7,February 2011,Pg.no:19-22
4. Zhou ZH, Jiang Y. “Medical diagnosis with C4.5 Rule proceeded by artificial neural network Ensemble”, IEEE Trans Inf Technol Biomed.2003 Mar; 7(1): Pg.no:37-42.
5. Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. “Artificial neural networks applied to survival prediction in breastcancer”., Oncology 1999; 57:281-6.
6. Delen D, Walker G, Kadam A. “Predicting breast cancer survivability: a comparison of three data mining methods”, Artificial intelligence in Medicine. 2005 Jun; 34(2):113-27.
7. M.Venkatadri, C.Lokanatha Reddy, “A Comparative Study on Decision tree Classification algorithms in Data Mining”, IJCAETS, ISSN: 0974-3596, April ’10 – Sept ’10, Volume 2: Issue 2, Pg.no:24-29.
8. Mitchell, T., Machine learning. McGraw-Hill, New York, 1997.
9. L. Rokach, “Ensemble-based classifiers,” Artificial Intelligence Review, vol. 33, no. 1-2, Pg.no: 1–39, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10462-009-9124-7.
10. R. Polikar, “Ensemble based systems in decision making,” IEEE Circuits and Systems Magazine, vol. 6, no.3, Pg.no: 21-45, 2006.
11. Venkatadri M, C.Lokanath Reddy, “A multi_objective genetic algorithm for feature selection in data mining,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 1, issue no. 5, Pg.no: 443–448, 2012.
12. Y. Ren, L. Zhang and P. N. Suganthan, "Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]," in IEEE Computational Intelligence Magazine, vol. 11, no. 1, Pg.no: 41-53, Feb. 2016
13. L. Breiman, “Bagging predictors, Machine Learning”, 26, 1996, Pg.no:123-140.
14. Y. Freund, R.E. Schapire” A short introduction to boosting” J. Jpn, Soc. Artificial Intelligence.14 (5):771, 1991.
15. Y. Freund, R.E. Schapire. “Experiments with a new boosting algorithm” in Proceedings Of the 13th International Conference on Machine Learning, Pg.no: 148-156, 1996.
16. Y. Freund, R. E. Schapire, "A decision theoretic generalization of on-line learning and an application to boosting," Proc. European Conference on Computational Learning Theory, Pg.no: 23-37, 1995.
17. Bauer, E. &Kohavi, R. (1999). “An empirical comparison of voting classification algorithms: Bagging, Boosting, and variants. Machine Learning”, 36, Pg.no:105–139.
18. G. Sujatha, K. Usha Rani, ” Advanced Ensemble Technique on Decision Tree Classifiers – An Experimental Study”, Special Issue on Computational Science, Mathematics and Biology, IJCSME-SCSMB-16-MAR-2016, Pg.no:264-268
19. G. Sujatha, K. Usha Rani,” Ensembled Decision Tree Classifier Performance with Varying Committee Sizes”. International Journal of Computer Engineering and Technology, 9(1), 2018, Pg.no: 96-101.
20. Ireneusz Czarnowski, Piotr Jedrzejowicz, “Instance reduction approach to machine learning and multi-database mining” AnnalesUMCS Informatica AI4, Pg.no: 60 -71, 2006.
21. Scheaffer, R.L., Mendenhall, W., and Ott, R.L., “Elementary Survey Sampling”, 5th Ed. New York: Duxbury Press 1996.
22. Jose Ramon Cano, Francisco Herrera, “A study on the combination of evolutionary algorithms and stratified strategies for training set selection in data mining”, Journal of Applied SoftComputing, vol.6, no. 3, Pg.no: 323-332, 2006.
23. Ifiok J. Udo and Babajide S. Afolabi, “Hybrid Data Reduction Technique for Classification of Transaction Data”, Journal of Computer Science and Engineering, vol. 6, no. 2, 2011.
24. A.K.Tanwani et.al.,”Guide lines to select machine learning scheme”, Evolutionary computation, Machine Learning and Data Mining in Bioinformatics”, Pg.no:128-138.
25. Pooja, Saroj Ratnoo,” A comparative study of instance reduction” ,Vol.3(3), Special issue, July ,2013, ICEREM , and Pg.no:7to13.
26. Bhavya et.al, “Comparative analysis of data reduction model for diabetes”, Vol.4,Issue 7,July 2015, IJCSMC, Pg.no:315-324
27. Faiza Rahat et.al, ”Comparative study of machine learning techniques for pre-processing network intrusion data”, international conference on Open source systems& Technologies, Published in IEEE explore , 2015, Pg.no:46-51,
28. Pinar yildrim et.al, “Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes “, Procedia Computer Science 83(2016), Pg.no:1013 – 1018 .
29. M.MillanGirlado, ”Instance selection methods and Resampling techniques for dissimilarity representation with imbalanced datasets”,Patternrecognition-applicationsandmethods”,AISC204, Springer-Verlag berlin 2013,Pg.no:149-160.
30. Lokesh S.Katore, ”Comparative Study of Recommendation Algorithms and Systems using WEKA “, International Journal of Computer Applications (0975 –8887),Volume 110 –No. 3, January 2015,Pg.no:14-17.
31. Judickael Bah,”Comparision of Data Reduction Algorithms for Biomedical Application”, online Scribed.
32. UCI Irvine Machine Learning Repository www.ics.uci.edu/~mlearn/ MLRepository.html
33. Bioinformatics Group Seville, http://www.upo.es/eps/bigs/datasets.html

AUTHORS PROFILE

G. Sujatha, research scholar in the Department of Computer Science at Sri Padmavati Mahila Visvavidyalayam, Tirupati, AP. She completed her M.Phil from Sri Padmavati Mahila Visvavidyalayam. She has more than 6 research publications in various international journals, conferences. Her Research areas include Data Mining and Big Data Analytics.

Prof K.Usha Rani, Professor in the Department of Computer Science at Sri Padmavati Mahila Visvavidyalayam, Tirupati, AP, Prof K.Usha Rani has more than 26 years of teaching experience and 19 years of research experience. She has more than 70 international and national publications in various reputed journals. She has attended more than 55 International and national conferences. She received a prestigious State Teacher Award by the Dept of Higher Education, Gov. of AP in 2018. Her Bio-data is placed in Asia Pacific Who's Who, International in 2018. Her research interest includes Data Mining, Soft Computing, Big Data Analytics, Neural Networks and Image Processing.