International Journal of Advanced Science and Technology
Vol. 29, No. 03, (2020), pp. 12977 - 12982

A Framework for Subset Pruning using REP Tree

[1]Dr.R.V.S.Lalitha, [2]Chaallapalli Sujana, [3]Dr.K.Kavitha, [4]Mylavarapu Kalyan Ram

[1]Professor, Department of C.S.E, Aditya College of Engineering & Technology, Surampalem,India,
[2]Assistant Professor, Department of C.S.E, Aditya College of Engineering & Technology, Surampalem,India,
[3]Associate Professor, Gokaraju Rangaraju Institute of Engineering & Technology, Hyderabad,India,
[4]Sr. Assistant Professor, Department of C.S.E, Aditya Engineering College (A), Surampalem,India.
Corresponding author:rvslalitha@gmail.com

Abstract: *Basically the count of confirmed cases depends upon number of tests so far carried out.The rate of confirmed cases relies on number of test carried out so far. Based on the figures arrived, the Government organization will impose precautionary measures. In case of patient data analysis for any pandemic disease decision making plays crucial role. If the test reports are inaccurate, underreporting takes place. The reports should be provided on time so as to safeguard the lives of patients. Identification of misclassified examples is a well known problem and is drawing significant attention in heath monitoring units. By and by exactness is the primary concern for evaluation and treatment of the individuals.In order to categorize correct classification labels, two typical algorithms are considered, ZeroR and RIPTree. The ZeroR classifies all the instances with majority of the labels without including any predictors. Since the RIPTree is less prone to error, it provides correct information about misclassified instances.The evaluation metrics can be streamlined with the adaptation of these two algorithms. The two algorithmic outcomes can be cross verified with Repeated Incremental Pruning to produce error reduction (RIPPER) algorithm which classifies true positives exactly. The application of the above algorithms assists in confirming the cases, with varied conditions and datasets.*

Keywords: *ZeroR, RIPTree, RIPPER, Classification, Pruning, Rules.*

I. Introduction

Corona virus is being spread globally since December. No proper diagnostics is available either for curing or for confirming the cases. As viral tests evaluate current infection of the individuals, the analysis of report for confirmed/unconfirmed is playingcrucial role. Finding outstanding classifier is a typical task, because the algorithm must support large datasets at par with accuracy. Researchers need to drill down their data exploration mechanisms by applying

ISSN: 2005-4238 IJAST

appropriate algorithms. Focusing on the analytical research, the study about correct classification of tests is experimented using ZeroR and RIPTree. Kaggle dataset has been taken for evaluation. This work explores how ZeroR is applied for corona database with each attribute. ZeroR is used for both classification and regression problems. Then using RIPTree learning algorithms rules are imposed for classifying based on impact of symptoms. The error rates are to be evaluated, because this algorithm works on large datasets. Finally, this analysis will again be evaluated using RIPPER algorithm which less error prone. By passing these three phases, we believe that the accuracy is maintained.

II. Related Work

Matej Petkovic proposed Multiple Target Regression (MTR) algorithm for multiple continuous dependent/target variables for simultaneous prediction. Feature ranking scores are calculated based on Symbolic, Genie3, Radom Forest and bagging and extra trees. Author concluded that Symbolic and Genie3 scores in connection with random forest yield best results[1].Haitao suggested Ensemble Multiboost based on RIPPER classifier for predicting imbalanced software. Initially, they used Principal Component Analysis(PCA) to filter key features for decision making. The classifier used NASA MDP public dataset and compared with existing algorithms[2].Dwi Normawati worked on coronary heart disease using Motivated Feature Selection(MTF)and rule based classifiers, VPRS and Repeaed Incremental Pruning Error Reduction(RIPPER)[3]. Heera compared various data mining tools like classification, WEKA, Pattern recognition tools and different classifiers. Datasets were considered from UCI repository and then accuracy is compared with different classifiers[4].Jyoti explored kernel based algorithms for noisy image datasets for non similarity measures for defining inter point similarities[5,9].Qichao Quo provided theoretical analysis of Radial Basis Function(RBF) networks which are basically for classification using supervised learning algorithms. Author worked on data dependent networks that are relied on kernel methods and k-means clustering also[6].The Mask R-CNN is the extension of faster R-CNN by adding prediction of object mask for bounding box recognition. Author explored the applications of CNN in object detection and feature extraction phenomina[7].Tong experimented with multi label learning with weekly labeled data for incomplete data instances. Author discussed about semi supervised multi label learning, weak label learning and extended weak label learning with relevant and irrelevant data[8-11].

III. Study of Algorithms

3.1. Data Analytics using Zero Rule(ZeroR)

ZeroR classifies data with respect to the labels that has maximum score without using any prediction model. The labels that are maximum in number are explored using WEKA tool as illustrated in Fig.1.

id	country	gender	...	symptom1	symptom	death
0	1	China	male	...	0.0	NaN	0
1	2	China	female	...	0.0	NaN	0
2	3	China	male	...	0.0	NaN	0
3	4	China	female	...	0.0	NaN	0
4	5	China	male	...	0.0	NaN	0

ISSN: 2005-4238 IJAST

```
..  ...      ...    ... ...     ...                  ...   ...
369  424    Japan  female  ...     1.0       cough, sore throat     0
370  425    Japan  female  ...     1.0                    fever     0
371  426    Japan    male  ...     1.0    fever, cough, headache     0
372  439 Malaysia  female  ...     1.0  fever, sore throat, cough     0
373  440 Malaysia  female  ...     1.0               mild fever     0
    [374 rows x 9 columns]
```

The Correlation coefficient is-0.1719, Mean absolute error is 0.1874, Root mean squared erroris 0.3067, Relative absolute error is 100% and Root relative squared erroris 100%.

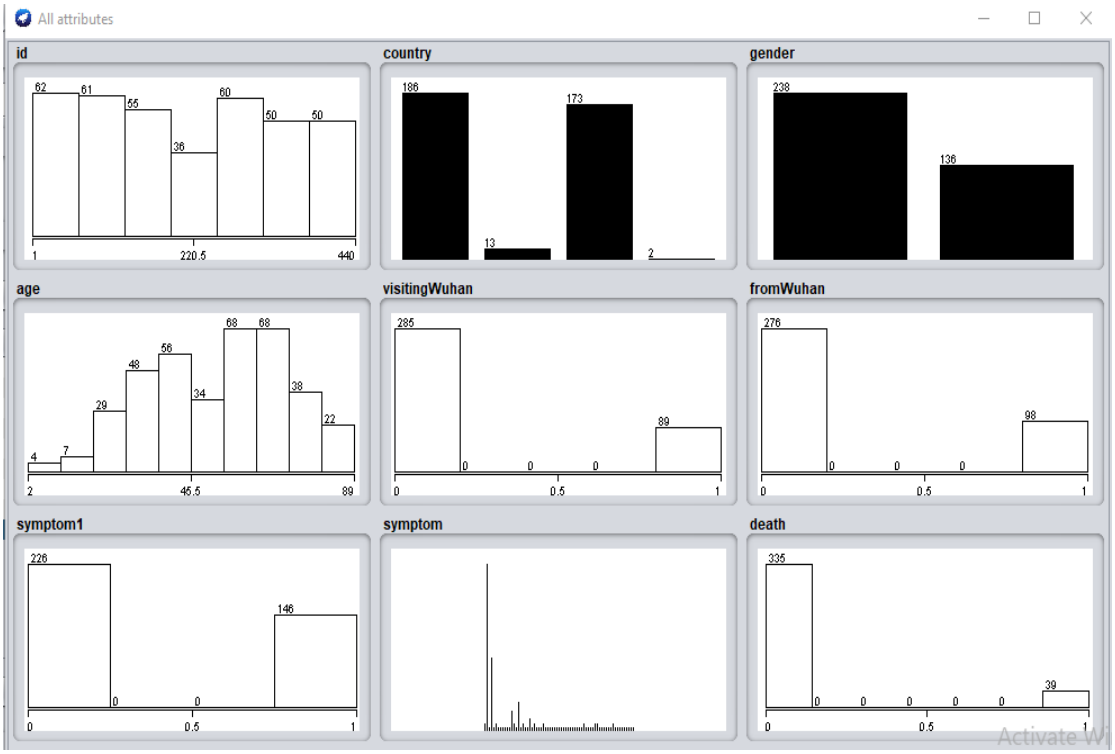

Fig.1. Analyzing each of the attributes using ZeroR for the chosen dataset.

3.2. Classification using REPTree

The REPTree is constructed using WEKA and the model is build in 0.02 sec. REPTree is based on C4.5 algorithm. It can produce classification based on discrete variables. It is also used for constructing regression trees by calculating gain or variance and prunes subsets. In the reduced pruning, rule set is first built and then pruning takes place. The output obtained using REPTree is as follows.

REPTree

```
id < 114.5
|   age < 57.5 : 0.06 (44/0.04) [21/0.09]
|   age >= 57.5
|   |   id < 45.5 : 0 (8/0) [2/0]
|   |   id >= 45.5 : 0.95 (23/0.04) [14/0.07]
```


ISSN: 2005-4238 IJAST

id >= 114.5 : 0 (174/0) [88/0]

The size of the tree deduced is 7. The Correlation coefficient is 0.8382. The Mean absolute error is 0.0399. The Root mean squared error is 0.1702. Relative absolute error is 21.2754 %. The Root relative squared error55.495 %. The decision tree obtained using REPTree is shown in Fig.2 and the comparison of error rate of the two algorithms ZeroR and REPTree is depicted in Fig.3.

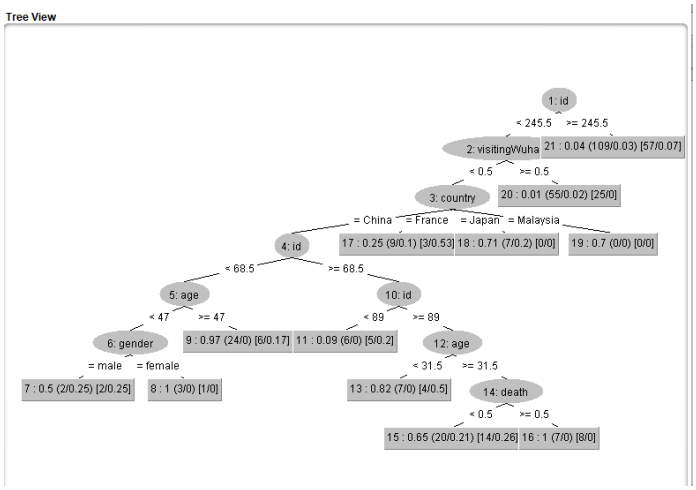

Fig.2.Decision Tree Using REPTree

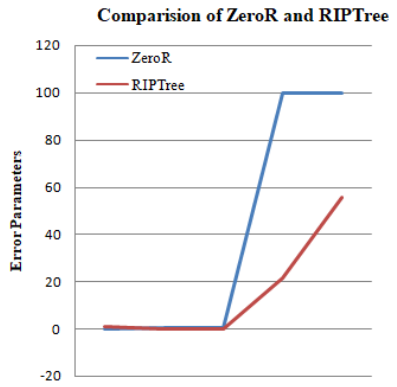

Fig.3.Comparison of error rates of ZeroR and REPTree.

IV. Classification using Repeated Incremental Pruning to produce error reduction(RIPPER)

RIPPER algorithm prunes dataset based on rules. This is an optimized version of Incremental Reduced Error Pruning (IREP). RIPPER tries to minimize the error, by identifying wrongly classified instances.RIPPER builds rules for two class problems. In incremental reduce pruning, each rule is pruned as soon as it is built. So validation of each rule is performed before expansion takes place. Initially all the class labels are sorted based on their popularity. The rule 1 is applied and subsequently all the instances that are covered in rule 1 are removed. This continues till all the rules are over. The significance and evaluation of these three algorithms are explored in Fig.4.

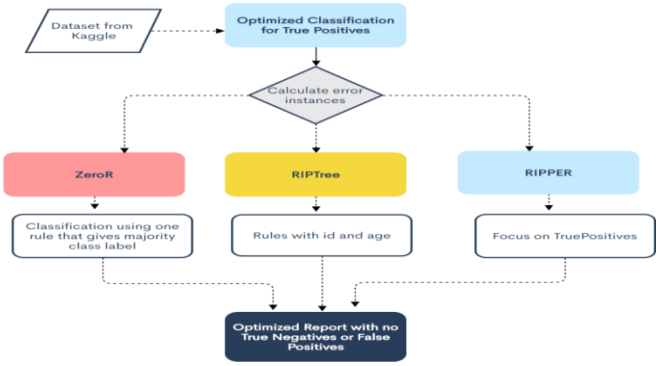

Fig.4. Stepwise analysis of ZeroR, RIPTree and Optimization using RIPPER

ISSN: 2005-4238 IJAST

Conclusion

The ZeroR is only one split decision tree classification. RIPPER and RIPTree are also decision tree classifiers and famous for engineering featured attributes into true positives and less error prone. It is shown that the advanced analysis shows that evaluation with the three algorithms gives true report in time. Also it is evident that error rate analysis is more in ZeoR and less in REPTree. The results of the study infers that REPTree and RIPPER are recommended learner classifiers for large datasets that computes in very less amount of time and with negligible error.

References

[1] Matej Petkovic, Dragi Kocey and Saso Dzeroski, Feature ranking for multi target regression, Machine Learning, 109, 1179-1204,2020.

[2] Haitao He, Xu Zhangi, Qian Wang, Jiadong Ren, Jiaxin Liu, Xiaolin Zhao, Yongqiang Cheng, Ensemble MultiBoost based on RIPPER classifier for prediction of imbalanced software defect data, Volume 7, pp 110333-110343, IEEE Access, August 2019.

[3] Dwi Normawati Sri Winarti, Feature selection with combination classifier use Rules based data mining for diagnosis of coronary heart disease, 2018 12th International Conference on Telecommunication Systems, Services and Applications(TSSA), 4-5, October 2018, 9 May 2019, Conference location Indonesia.

[4] Heera Begum Mirza, Varsha R Ratnaparkhe, Classifier Tools: A comparative study, 14-15, June 2018, Second International conference on Intelligent Computing and Control Systems (ICICCS), 11 March 2019, IEEE, Madurai, India.

[5] Jyoti Arora, Meena Tushir A new Kernel based probabilistic intuitionistic fuzzy c-means clustering, International Journal of Artificial Intelligence and Soft Computing, volume 6, no 4, 2018.

[6] Qichao Quo, Mikhail Belkin, Back to the Future: Radial Basis Function Network Revisited, August 2020, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp 1856-1867, volume 42.

[7] Kaiming He, Georgia Gkioxari, Mask R-CNN, IEEE Transactions on Pattern analysis and Machine Intelligence, Feb,2020, pp 386-397, vol 42.

[8] Tong Wei, Lan Zhe Guo, Yu Feng Li and Wei Gao, Learning safe multi label prediction for weekly labeled data, Machine learning, 107, 703-725, 2018.

[9] Yu Nishiyama, Motonobu, Arthur Gretton, Kenji Fukumizu, Model based kernel sum rule: kernel Bayesian inference with probalistic models, Jan 2020, Machine learning, 109, 939-972.

[10] Gregor H W G ebhardt, Andras Kupcsik, Gerhard Neumann, The Kernel Kalman rule, Efficient nonparametric inference by recursive least squares and subspace projections, Machine Learning, 108, 2113-2157, 2019.

[11] Raghavendran, C. V., Satish, G. N., Krishna, V. and Basha, S. M. Predicting Rise and Spread of COVID-19 Epidemic using Time Series Forecasting Models in Machine Learning. International Journal on Emerging Technologies, 11(4): 56–61,(2020).

ISSN: 2005-4238 IJAST

Dr.R.V.S. Lalitha received her Ph.D. from JNTUK, Kakinada in 2017. Her research includes Mobile Computing, Soft Computing and Machine Learning. She is working as Professor in the Department of CSE, at Aditya College of Engineering & Technology, Surampalem.

Ms.Challapalli Sujana received her M.Tech from JNTUK, Kakinada andworking as an Assistant Professor in the Department of CSE, at Aditya College of Engineering & Technology, Surampalem. Her research includes Blockchain technology, Machine learning and Data Mining.

Dr.K.Kavitha received her Ph.D. form BITS, Hyderabad in 2016 and is currently working as Associate Professor, in the Department of C.S.E, at Gokaraju Rangaraju Institute of Engineering & Technology, Hyderabad. Her research areas include Wireless Sensor Networks, Data Mining and Machine Learning.

Mr.M.Kalyan Ram, received M.Tech from JNTUK. He iscurrently pursuing Ph.D from K.L.University. He is working as Sr.AssistantProfessor, Department of Computer science engineering, AdityaEngineering College (A), Surampalem. Hisarea of Interest includes Blockchain Technology, Machine Learning, DataMining, Computer Networks and Cloud Computing.

ISSN: 2005-4238 IJAST