An effective noise reduction technique for class imbalance classification

[1]Dr. P Ratna Babu, [2]P. Lokaiah

Abstract--*The paper presents a unique approach to handle noisy instances in the data sources using the novel technique of priority instance picking for weak range feature subsets. The technique used in the proposed approach quickly identifies the noisy instances in the data source than the benchmark C4.5 algorithm. The C4.5 algorithm also removes the noisy instances from the formed decision tree but in the final stage by applying the pruning technique. The results conducted on 12 UCI datasets suggest that the proposed approach performs better than the benchmark algorithm.*

Keywords--*Data Mining, Knowledge Discovery, Feature subset, priority instance picking*.

I INTRODUCTION

The quality of predictive model developed is depended on the training data provided for classification model building. The training data provided is of erroneous in nature such as with noisy, missing and outlier instances then the predictive model developed will be insufficient for unseen instance prediction. The training data can also be in the form of class imbalance nature. The existing approaches are not capable of effectively handling all real world datasets [1]. Exclusively, the datasets of noisy, class imbalance in nature have decreased the overall classification performance [2-7]. The binary dataset have two sub classes named majority and minority where an effective classification results can be seen in majority class with good number of instances and very ineffective classification results can be seen for minority class with less number of instances [8].

Class imbalance ratio (IR) for different real world datasets can be range up to 1:99 % for different real world datasets where the intrinsic datasets properties projected in a exclusive way[9]. The main areas of applicability of data sources are widely from secure data transfer [10-11], clinical analysis [12], geographical image analysis [13], credit loss detection [14] and computer aided biological analysis [15]. In this paper, we propose a new technique for effective classification for noisy class imbalance datasets.

The remaining paper is organized as follows: section 2 presents the distinct related literature in the research area. The proposed methodology with brief description is presented in the section 3. Section 4 describes the complete experimental set up with all the datasets used in the study. Section 5 presents the results of the experimental simulation with explanation for the behaviour of specific datasets. The conclusion and future work of the study are presented in section 6.

[1] Professor, Department of CSE, RISE Krishna Sai Prakasham Group of Institutions, Ongole, AP, India, Email:ratnajoyal@gmail.com

[2] Research Scholar, Department of CSE, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal, MP, India, Email:lokaiah75@gmail.com

Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

II RELATED WORK

In data mining approach, the key concept is to build efficient model from the existing data. One of the drawbacks that arise in the model building is about the encounter of noisy instances in the data source.

Noisy data removal is one of the key issues for better knowledge discovery from the data sources. K. V. Uma [16] has investigated on the concept of noisy data rectification using feature selection technique. They have also pointed out various drawbacks with the noisy data such as excessive storage space, adverse branch formation for improper model building etc.One of the main issues occurs due to the noisy instances data set is known as “over fitting”. In over fitting the tree build is of oversize leading to noisy and unnecessary branches. Nittaya Kerdprasop et al.,[17] have worked on the noisy instance removal using the clustering and visualization technique prior to tree building process.

The existence of noisy instances in the medical data sources also severely affects the predictive outcome of the classification. Dragan Gamberger et al.,[18] have investigated on the issue of noisy removal for predicting coronary artery disease diagnosis.Ensemble classifier learning is also one of the effective technique for counter balancing the noisy instance in the data set. THOMAS G. DIETTERICH [19] have done an empirical comparative study for forming different ensemble of decision tree approaches.

Uncertainty in the dataset is one of the reasons to analyse when dealing with scalable dataset. Cèsar Ferri et al., [20] have experimented with different level of uncertainty conditions for different threshold levels.Ensemble technologies are one of the effective techniques for improving the performance of the classifiers. Ludmila I et al., [21] have proposed a method for classifier combination using different ensemble methods to generate efficient classification for a vast benchmark datasets.Shaghayegh Gharghabi et al.,[22] have worked in the area of data streams for effective classification using the dimensionality reduction technique. The above literature indicates the need of an improved approach for handling datasets with class imbalance nature and misleading instances.

III PROPOSED APPROACH

In this subs section, we present the proposed approach for improving noisy datasets.

In the initial phase, the dataset is applied for preprocessing stage, where missing values are removed by using different techniques. The techniques used for removing missing values are like average the reaming existing values and replacing the average value in the missing spaces. Another technique used for removing missing values is by default values in the missing spaces.

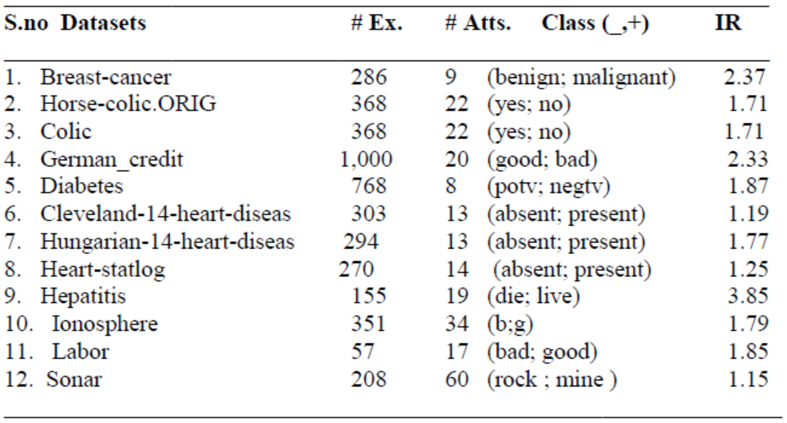
In the next stage, the processed data is applied for a attribute selection algorithm which can pick the best attributes in the dataset and the reaming irrelevant or un important attributes are removed for the existing approaches. But, in the novel proposed approach the so called irrelevant attribute are not completely removed, unless the weak ranges of the instances in these attribute are identified and removed.

In the third stage, a specific visualization technique is used to project the instances on a two dimensional plane and the far away instances from a form origin are identified and removed. In the final stage, the improved dataset is applied to a base algorithm for generating the evaluation metrics. In this case C4.5 algorithm is used as a base algorithm. Indeed, the point to be noted is that in C4.5 decision tree post pruning technique is applied for removal of unnecessary branches of the formed tree.

IV EXPERIMENTAL SETUP

The experimental study is conducted on well-known 12 binary class imbalance datasets from UCI [23] machine learning repository. Table 1 presents the properties of datasets such as S.no, name of the datasets, number of examples, number of features, majority and minority class and class imbalance ratio.

Table 1: Summary of benchmark imbalanced datasets



S.no	Datasets	# Ex.	# Atts.	Class (_,+)	IR
1.	Breast-cancer	286	9	(benign; malignant)	2.37
2.	Horse-colic.ORIG	368	22	(yes; no)	1.71
3.	Colic	368	22	(yes; no)	1.71
4.	German_credit	1,000	20	(good; bad)	2.33
5.	Diabetes	768	8	(potv; negtv)	1.87
6.	Cleveland-14-heart-diseas	303	13	(absent; present)	1.19
7.	Hungarian-14-heart-diseas	294	13	(absent; present)	1.77
8.	Heart-statlog	270	14	(absent; present)	1.25
9.	Hepatitis	155	19	(die; live)	3.85
10.	Ionosphere	351	34	(b;g)	1.79
11.	Labor	57	17	(bad; good)	1.85
12.	Sonar	208	60	(rock ; mine)	1.15

The evaluation metrics used for the experimentation are accuracy, root mean square error and tree size.The experimental methodology used is 10 fold cross validation for 10 runs. All the performance metrics are measures with an average of 10 runs. The algorithms compared are simulated using the open source tool kit WEKA [24] for standard configurations of system.

V RESULTS

The experimental results of the proposed approach are presented in this section. The results are compared on the evaluation metrics such as accuracy, root mean square error and tree size.The accuracy results are presented in the table 2, with both mean and standard deviation values. Out of the 12 datasets used for comparison, the proposed algorithm has improved its value on 6 datasets, decreased its value on 5 datasets and remains unchanged on 1 datasets.The summaries of results for root mean square error are presented in the table 3. The results indicate that the proposed algorithm have performed well on 10 out of 12 datasets. The results of tree size are presented in the table 4. These observations from the results present that the proposed algorithm have produced better values on 10 out of 12 datasets. The proposed algorithm has not performed well on two datasets of horse colic. The reason may be due to less amount of noise present in these datasets.

Table 2: Summary of tenfold cross validation performance for Accuracy on all the datasets

Datasets	C4.5	Proposed

Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

Breast-cancer	74.28± 6.05●	75.26±5.04
Horse-colic.ORIG	66.31± 1.23●	77.20± 5.63
Horse-colic	85.16± 5.91	85.43± 5.95
German_credit	71.25± 3.17○	70.00± 0.00
Pima_diabetes	74.49± 5.27○	65.11± 0.34
Cleveland-14-heart-diseas	76.94± 6.59●	77.23± 6.47
Hungarian-14-heart-diseas	80.22± 7.95●	81.05± 7.39
Heart-statlog	78.15± 7.42○	77.37± 6.34
Hepatitis	79.22± 9.57●	82.93± 8.43
Ionosphere	89.74± 4.38○	74.93± 4.91
Labor	78.60±16.58●	84.97±14.24
Sonar	73.61± 9.34○	53.38± 1.63

Table 3: Summary of tenfold cross validation performance for Root_mean_squared_error on all the datasets

Datasets	**C4.5**	**Proposed**
Breast-cancer	0.444±0.037●	0.436±0.034
Horse-colic.ORIG	0.473±0.004●	0.398±0.050
Horse-colic	0.352±0.060●	0.351±0.063
German_credit	0.476±0.028●	0.458±0.000
Pima_diabetes	0.439±0.042●	0.477±0.001
Cleveland-14-heart-diseas	0.281±0.039●	0.260±0.037
Hungarian-14-heart-diseas	0.252±0.043●	0.247±0.038
Heart-statlog	0.429±0.077●	0.417±0.053
Hepatitis	0.404±0.096●	0.367±0.070
Ionosphere	0.299±0.081○	0.424±0.030
Labor	0.401±0.170●	0.304±0.167
Sonar	0.491±0.093○	0.499±0.001

Table 4: Summary of tenfold cross validation performance for Tree Size on all the datasets

Datasets	**C4.5**	**Proposed**
Breast-cancer	12.78●	7.93
Horse-colic.ORIG	1.00○	56.38
Horse-colic	8.80○	10.37

German_credit	126.85●	1.00
Pima_diabetes	43.40●	1.00
Cleveland-14-heart-diseas	42.52●	27.34
Hungarian-14-heart-diseas	10.53●	4.27
Heart-statlog	34.64●	21.43
Hepatitis	17.66●	4.08
Ionosphere	26.74●	3.00
Labor	6.92●	5.92
Sonar	27.90●	1.00

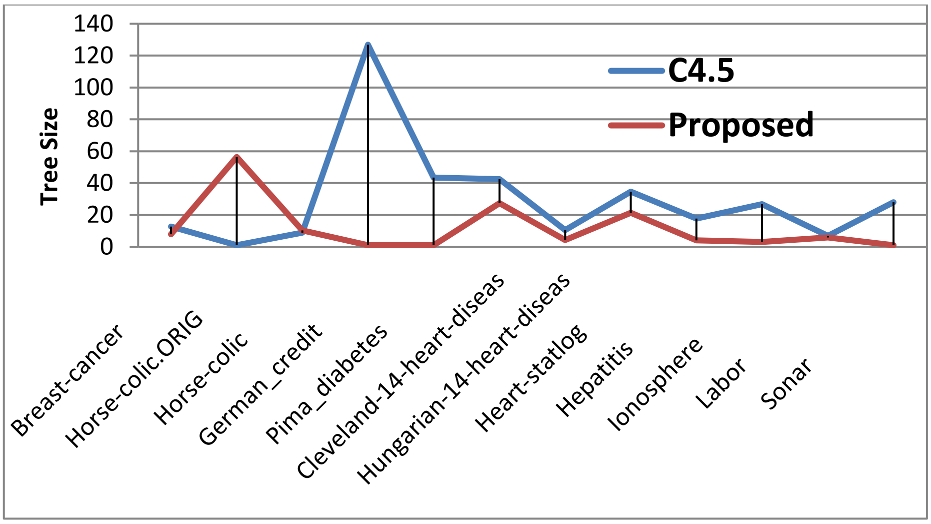


Figure 1: Trends in Tree Size for Proposed approach versus C4.5 on 12 datasets

VI CONCLUSION

The technique used in the proposed approach quickly identifies the noisy instances in the data source than the benchmark C4.5 algorithm. The C4.5 algorithm also removes the noisy instances from the build model by applying the pruning technique. The experimental results observed from 12 UCI datasets indicated that the proposed approach is a completive one when compared with existing techniques. In the future direction, we apply the proposed technique to uncertain and incomplete data sources.

REFERENCES

1. H. He and E. A. Garcia, “Learning fromimbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
2. A. Estabrooks, T. Jo, and N. Japkowicz, “Amultiple resampling method for learning fromimbalanced data sets,” *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

3. H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning,” in *Advances in Intelligent Computing*, pp. 878–887, Springer, 2005.

4. N. V. Chawla, N. Japkowicz, and A. Kotcz, “Editorial: special issue on learning from imbalanced data sets,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

5. U. Bhowan, M. Johnston, and M. Zhang, “Developing new fitness functions in genetic programming for classification with unbalanced data,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42,no. 2, pp.406–421, 2012.

6. J.-H. Xue and P. Hall, “Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis?” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1109–1112, 2015.

7. R. Batuwita and V. Palade, “Class imbalance learning methods for support vector machines,” in *Imbalanced Learning: Foundations, Algorithms, and Applications*, pp. 83–99, John Wiley & Sons, Berlin, Germany, 2013.

8. V. L´opez, A. Fern´andez, S. Garc´ıa, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics,” *Information Sciences*, vol. 250, pp. 113–141, 2013.

9. F. Provost, “Machine learning from imbalanced data sets 101,” in *Proceedings of the AAAI’2000Workshop on Imbalanced Data Sets*, pp. 1–3, 2000.

10. L. Pelayo and S. Dick, “Applying novel resampling strategies to software defect prediction,” in *Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS ’07)*, pp. 69–72, June 2007.

11. J. Long, J.-P. Yin, E. Zhu, and W.-T. Zhao, “A novel active cost sensitive learning method for intrusion detection,” in *Proceedings of the 7th International Conference on Machine Learning and Cybernetics (ICMLC ’08)*, pp. 1099–1104, IEEE, Kunming, China, July 2008.

12. K. Zahirnia, M. Teimouri, R. Rahmani, and A. Salaq, “Diagnosis of type 2 diabetes using cost-sensitive learning,” in *Proceedings of the 5th International Conference on Computer and Knowledge Engineering (ICCKE ’15)*, pp. 158–163, October 2015.

13. M. Kubat, R. C.Holte, and S.Matwin, “Machine learning for the detection of oil spills in satellite radar images,” *Machine Learning*, vol. 30, no. 2-3, pp. 195–215, 1998.

14. T. Fawcett and F. Provost, “Adaptive fraud detection,” *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 291–316, 1997.

15. I. Triguero, S. del R´ıo, V. L´opez, J. Bacardit, J. M. Ben´ıtez, and F. Herrera, “ROSEFW-RF: the winner algorithm for the ECBDL’14 big data competition: an extremely imbalanced big data bioinformatics problem,” *Knowledge-Based Systems*, vol. 87, pp. 69–79, 2015.

16. K. V. Uma,” Improving the Classification accuracy of Noisy Dataset by Effective Data Preprocessing”, *International Journal of Computer Applications (0975 – 8887) Volume 180 – No.36, April 2018*

17. Nittaya Kerdprasop and Kittisak Kerdprasop,” A Heuristic-Based Decision Tree Induction Methodfor Noisy Data”, T.-h. Kim et al. (Eds.): DTA/BSBT 2011, CCIS 258, pp. 1–10, 2011.

18. Dragan Gamberger, Nada Lavrac," FILTERING NOISY INSTANCES AND OUTLIERS", H. Liu et al. (eds.), *Instance Selection and Construction for Data Mining,* © Springer Science+Business Media Dordrecht 2001

19. THOMAS G. DIETTERICH," An Experimental Comparison of Three Methodsfor Constructing Ensembles of Decision Trees:Bagging, Boosting, and Randomization", Machine Learning, 40, 139–157, 2000, Kluwer Academic Publishers. Manufactured in The Netherlands.

20. Cèsar Ferri, José Hernández-Orallo, Peter Flach," Setting decision thresholds when operating conditions are uncertain", Data Mining and Knowledge Discovery (2019) 33:805–847, https://doi.org/10.1007/s10618-019-00613-7

21. Ludmila I. Kuncheva · Juan J. Rodríguez," A weighted voting framework for classifiers ensembles", Knowl Inf Syst (2014) 38:259–275, DOI 10.1007/s10115-012-0586-6

22. Shaghayegh Gharghabi · Chin-Chia Michael Yeh · Yifei Ding, Wei Ding · Paul Hibbing Samuel LaMunion · Andrew Kaplan, Scott E. Crouter · Eamonn Keogh," Domain agnostic online semantic segmentation formulti-dimensional time series",Data Mining and Knowledge Discovery (2019) 33:96–130, https://doi.org/10.1007/s10618-018-0589-3

23. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

24. A Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.