International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 2539-2546

An Empirical Analysis on Data Preprocessing Over Two-Class versus Multi-Class Imbalance Learning

K. Santhi[1*], A. Rama Mohan Reddy[2]
[1] *Research Scholar, CSE Department, SV University College of Engineering*
[2] *Professor, CSE Department, SV University College of Engineering*
[1]*santhi@svcolleges.edu.in*, [2]*ramamohansvu@gmail.com*

Abstract

Traditional artificial learning methodologies consider that the number of samples in each class is approximately same in size. But coming to real-time situations, instances distribution is uneven because some of class samples appear more frequent compare to others. This causes difficulty to learning algorithms which give favor to the majority class which has large no. of samples. This paper addresses useful methods related to data preprocessing on class imbalance problems and further this paper presents empirical analysis on data preprocessing techniques on binary class as well as multi class classification problems using evaluation metrics like Accuracy, AUC, G-Mean.

Keywords: *Imbalance learning, binary class imbalanced, multi-class imbalanced, data preprocessing.*

1. Introduction

Class Imbalance issue [14] alludes to characterization issue in which a few classes contain a larger number of occasions than different classes. In twofold class issue, one class contains progressively number of tests (major class), where as another with less number of samples (minority class). Imbalanced informational collections happen, all things considered, issues like clinical finding, fraudulant location, content mining, video handling, picture retrieval, etc. As increasingly number of cases have a place with the larger part class, conventional AI arrangement calculations present kindness to the greater part class in the learning strategy which cause tendency achieves the awful demonstrating rather to minority class. This occurs in circumstances where minor class is given more significant than major class, for example, in the analysis of uncommon maladies.

In multi-class scenario, some classes contain more number of samples compare to others. Perhaps the greatest test is unevenness learning in managing multi class characterization issues. More lagging in research on multi-class imbalance problems over two-class imbalance learning as few papers discussed multi-class classification problems. Multi class imbalance learning is insignificant troublesome when contrasted with double class on the grounds that the connection between classes is not, at this point clear and one class is treated as larger part when contrast with one class, simultaneously it might be treated as minority class when it contrasted with others. So handling multi class imbalance datasets is considered as the current hot research topics.

To solve the imbalanced data problems, a spread of approaches and methods are proposed in literature, which may be separated into three types: data-level methods, algorithmic-level methods, cost-sensitivity methods and ensemble based approaches [1, 2]. Data-level methods are pre-processing of knowledge before the training process or constructing any classifier, during which resampling of knowledge are performed externally, to balance the ratio of instances in minority and majority class. Algorithmic-level methods are the creation or modification of algorithms in which minority class is considered. These methods reinforce the learner towards minority class, not allowing to bias for majority class [3]. In cost-sensitivity methods, the value of misclassification are reduced also as total cost of errors is minimized [4].

Pre-processing methods mainly focuses in this paper are preprocessing can be performed independently without considering any classifier [5]. preprocessing techniques are more versatile

ISSN: 2233-7857 IJFGCN

and can be applied globally. There is plenty of work performed by research community on preprocessing of data to overcome the imbalanced class issues because it enhances the learnability of data by any classifier. Various techniques and methods which address these problems have been proposed for their solutions.

The article is listed like this : Section 2 mention the preprocessing strategies on twofold class grouping. Information preprocessing systems for awkwardness learning on multi-class arrangement are introduced in Section 3. Observational examination close by appraisal estimations are coordinated in Section 4. End notes and future works are presented in area 5.

2. Preprocessing on Binary Class Imbalance Learning

Preprocessing of data by resampling techniques, in order to balance the ratio of instances present in majority and minority classes, can be divided in to three categories: oversampling, undersampling, and hybrid solutions (Fig. 1) [4].

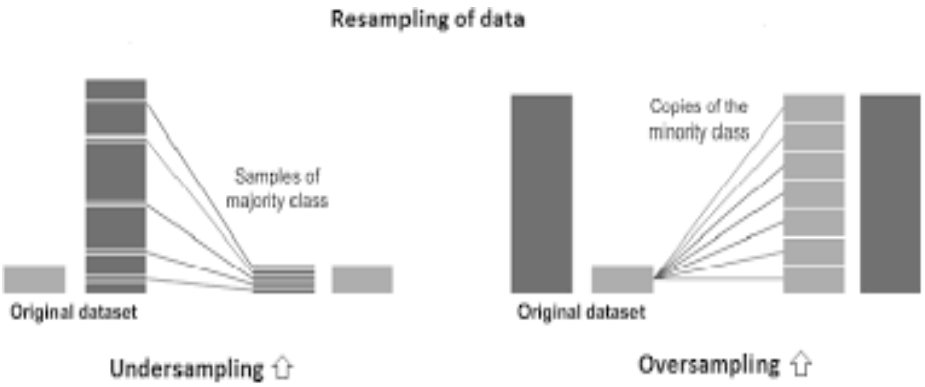

Fig. 1 Preprocessing on Binary class Imbalance learning

2.1 Oversampling Methods

The simplest technique used for oversampling is random oversampling (Fig. 2). It selects the samples randomly and produces new samples in minority class. Although, it increases the number of samples, but new samples are often quite similar to the original samples which may result in overfitting as the generated samples are exact replication of samples [5].

Fig. 2.a. Imbalanced dataset **Fig.2.b. Oversampling of dataset**

Synthetic Minority Oversampling Technique (SMOTE) is proposed by Chawla et al. in 2002 to overcome the overfitting problem [6]. In this technique new samples are generated by linear interpolation of an inferior sample with their randomly selected k-Nearest Neighbors (kNN). This technique generates new samples without examining the majority class samples, which may induce overlapping between majority and minority samples, causing over-generalization along with amplifying the noise. Though there are drawbacks, researchers widely adopts SMOTE due to its simplicity [10]. Various extensions and modification of this method have been proposed to eliminate its weaknesses. Some of the used filtering-based methods for avoiding the noise are SMOTE-TL and SMOTE-EL [11].

ISSN: 2233-7857 IJFGCN

Over-generalization occurs as the minority class instances distribute into majority class region creates noise and overlapping when generated new samples. Borderline-Smote [8] presents a solution by identifying borderline between major and minor classes. It then considers the minority samples at border line and increase them. Safe-level SMOTE [7] addresses this issue as generating the synthetic samples into safest level. For each minority sample, it calculates a value which is defined as number of minority samples between kNN and the new samples generated.

2.2 Undersampling Methods

Random undersampling(RUS) is the simplest method for resampling of an imbalanced dataset, during which the samples of majority class are randomly eliminated from the class to balance the distribution of classes for learning process (Fig. 3).

Fig. 3.a. Imbalanced data **Fig.3.b. Undersampling**

This method is simple and comparatively less complex than other methods or oversampling of data. However, due to its significant drawback that it losses potential samples which will be useful in learning process, or removing the samples randomly also removes useful data with it. Therefore, the researchers and practitioners usually prefer oversampling techniques to undersampling. RUS is simple to use and its weakness are overcome when it is used with other methods [7], [8].

A fast clustering-based undersampling method is proposed to find the problem of uneven data distribution [9]. It has many characteristics, time complexity of this method is confined to the number of samples of minority class. Moreover, each cluster is trained by a specific classifier. It takes into consideration the distinguished problem of undersampling which is loss of information in majority class samples.

2.3 Hybrid Methods

The hybrid method is combining oversampling and undersampling as well as integration of resampling techniques with ensemble classifiers appears to be effective and enhance the performance significantly. A new method, Cluster Based Instance Selection (CBIS) [15] uses undersampling approach during which clustering analysis groups the majority class instances into the subclasses in dataset and instance selection remove the unrepresented data instances from each subclass. Another application of undersampling adapted with one class SVM are recently proposed for data overlapping and imbalanced problem. Tomek-link undersampling is employed to eliminate overlapped, redundant and borderline instances from the majority class and overcome the imbalances and overlapped cases. Using the bagging method along with an oversampling technique, a replacement ensemble method Bagging of Extrapolation Borderline-SMOTE SVM [16] is proposed to integrate borderline information so as to affect imbalanced data problem.

3. Multi-Class Imbalance Data Preparation

Data Preprocessing uses outfit based approachs which use dynamic decision methods for multi-class ungainliness learning. Dynamic Ensemble Selection framework which combines

preprocessing strategy maintained balancing the data randomly and a dynamic selection plan that consigns a major capacity to classifiers which precisely name minor class tests inside the local region where the request test was found. Our proposed strategy uses Random Under-Sampling , Irregular Over-Sampling and SMOTE for getting balanced sets for setting up the base classifiers from available group.

The clustering based system isolates the part space into regions and apportions different burdens to the base classifiers in every zone. The result of framework is the weighted response of local gathering consigned to the area where the inquiry test is found. The classifiers' heaps and thus the gatherings' zones are gotten using a formative arrangement with an inclination genuine upgrade standard, with a view of decreasing the class tendency inside the responses of the described neighborhood social affairs.

An important feature about dynamic selection is that prediction of abilities of classifiers solid with every sample. Normally the expectation of classifiers aptitudes is predicated on social affair of named tests, called the dynamic assurance dataset (DSEL). DS execution is exceptionally unstable to the movement of tests in DSEL. In case the appointment of DSEL gets balancedless, by there will be a deep probability that the zone of capacity for a test event will get lopsided. So that, Dynamic Selection counts may end up uneven towards selecting base classifiers that are experts for the larger part class. Considering this, we use data preprocessing techniques for setting up a group of classifiers also as modifying class appointment in DSEL for strategies of Dynamic Selection.

Changing the transport of the planning data to find a good pace with bad representation of minority class is a fruitful response for issues of imbalance, and lot of systems are available during regards. Branco et al. disconnected such methodologies to classes such as, to be explicit, stratified testing, incorporating new data and mixes of the two past strategies.

3.1 Stratifed Sampling Methods

One critical characterization is under-trying, which removes events from the mass class to alter the movement. Random Under Sampling(RUS) is one such procedure. RUS has been including boosting (RUSBoost) and with Bagging. a veritable drawback of RUS is that it can discard significant data which may be a drag while using DS moves close.

3.2 Synthesizing New Data Methods

Consolidating new cases has a couple of benefits, and more proposals are available for constructing fabricated models. At the present time, well known technique that uses prologue to make new models is SMOTE[12]. Crushed over-models the minority class by making new fabricated data. Different procedures have been made subject to standard of SMOTE, for instance, Borderline-Synthetic Minority Oversampling Technique [12], AdaptiveSYN [13], Ranked Minority Over-Sampling and Random Balance.

3.3 Ensemble Methods

The RB (Random Balance) procedure relies upon the measure of under-looking at and over-testing that is issue unequivocal which remembers an essential impact for the introduction of the classifier concerned. RB keeps up segments of the dataset, yet changes degrees of the bigger part and minority classes, using a self-assertive extent. This consolidates the circumstance where the

minority class is over-addressed and subsequently the anomaly extent is turned around. As needs be, repeated employments of RB produce datasets having an outsized anomaly extent change, which propels fair assortment.

Algorithm RB procedure

Input: T,MP,MN
Input of the algorithm- A
Output of the algorithm- Result
Step 1. No.of.samples←|T|
Step 2. majsamples←|MP|
Step 3. minsamples←|MN|
Step 4. newMajsamples ←random(2,No.of.samples−2)
Step 5. newMinsamples←No.of.samples−newMajsamples
Step 6. Result←{}
Step 7. if newMajsamples < majsamples then
Step 8. Result ← MP
Step 9. Result ← T ∪RUS(MN,newMajsamples)
Step 10. Result ← Result∪SMOTE(MP,(newMinsamples−minsamples)/|MP|,A)
Step 11. Else
Step 12. TR← MN
Step 13. Result← Result ∪RUS(MP,newMinsamples)
Step 14. Result ← Result ∪SMOTE(MN,(newMajsamples−majsamples)/|MN|,A)
Step 15. End of if
Step 16. write Result

4. Empirical Analysis and Evaluation Metrics

Preprocessing of the data has its own significance in every field of artificial intelligence especially in data mining; such as, many realworld applications endure imbalanced data problems. Researchers focuses more on resampling of data in preprocessing although they face many challenges in this area. Researchers have overcome many weaknesses in techniques and approaches used for re-sampling. However, still some drawbacks remain uncovered.

4.1 Discussion about Results on Preprocessing

Pre-processing strategies give yield, in light of issue explicit especially. Pre-processing methods are progressively flexible and can be applied freely to any classifier. Pre-processing procedures are non-viable of run and preliminary overheads for the cost assessment draws near.

Altogether, five imbalanced multiclass datasets are chosen from storage facilitycalled KEEL. The disproportion extent is handled as the degree of amount of larger part class advisers for the amount of minority class models. At the present time, class with most outrageous number of models is the larger part class, and class with base number of models is named minority one. Exploratory assessment finished on a great deal of arranged multi-class imbalanced benchmarks allowed to get understanding into the show of oversampling and outfit method using sporadic evening out. The preprocessing frameworks, ROS, SMOTE uses customer showed parameters. For RB and SMOTE, five nearest neighbors are considered . RUS,ROS,RB approaches on imbalanced data showed up in Fig. 4, Fig. 5, Fig. 6 individually.

The investigation shows that thinking about the accuracy, G-mean and area under curve(AUC), the outcome acquired utilizing preprocessing strategies is in every case measurably better when contrasted with not utilizing preprocessing. Accuracy of 0.95, AUC of 0.83,G-mean

of 0.70 is obtained utilizing preprocessing by arbitrary equalization and dynamic gathering determination. Thinking about different classes and model sorts can prompt diminished issue difficulty and expanded classification execution.

4.2 Evaluation Metrics

Performance of re-sampling techniques and approaches are measured by some common matrices, but in case when distribution of the class is not uniform, all the metrics are not suitable.

Precision (1), Recall (2), Accuracy (3), G-mean (4), and AUC are calculated. By using Precision and Recall, we can calculate F-measure, in confusion matrix, majority samples are treated as negative (N), and minority samples as positive (P). Precision evaluates the classifier's exactness, which means the total number of samples that labelled correctly as positive (minority) which are positive actually. Classifier's completeness is evaluated by Recall in a way that the number of positive samples are classified as positive correctly. Consequently, if the classifier favoured the negative samples and ignores the positive samples, then low G-mean will be obtained by the classifier.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$\text{G-mean} = \sqrt{\left(\frac{TP}{TP+FN} * \frac{TN}{TN+FP}\right)} \quad (4)$$

AUC is suitable for performance evaluation for the class imbalance problem since it is not dedicated to the distribution of two classes in any dataset. AUC is acquired by scheming the ratio of FPR to TPR (5), where number of negative (majority) instances are denoted by NN and number of positive (minority) instances are referred by NP.

FPR is considered as False positive rate and TPR is true positive rate.

$$\text{TPR} = \frac{TP}{NP}, \quad \text{FPR} = \frac{FP}{NN} \quad (5)$$

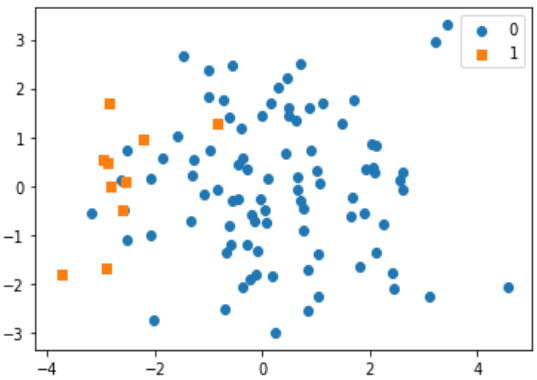

Fig. 4.a. Imbalanced Dataset

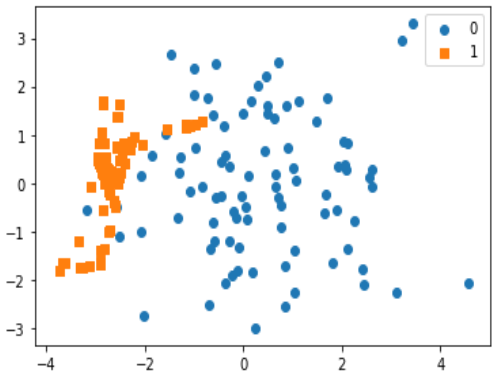

Fig. 4.b. Balancing by Oversampling

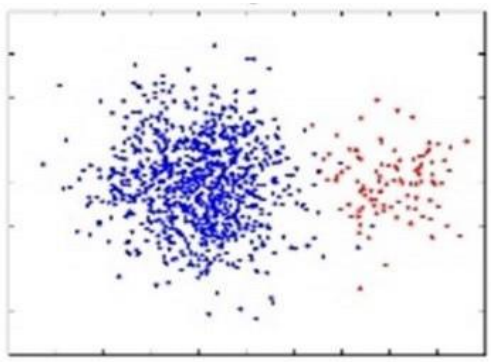

Fig.5.a.Imbalanced Data

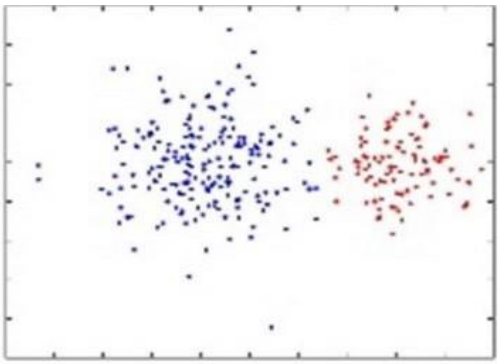

Fig.5.b.Balancing by RUS

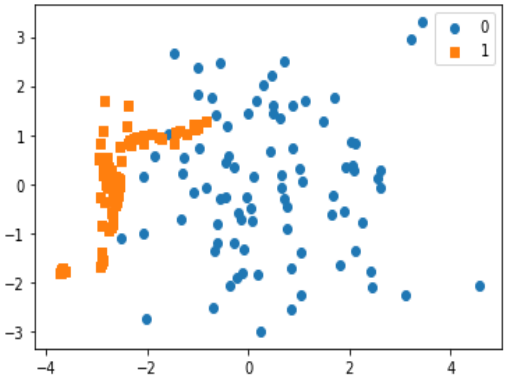

Fig. 6.a. Imbalanced Dataset

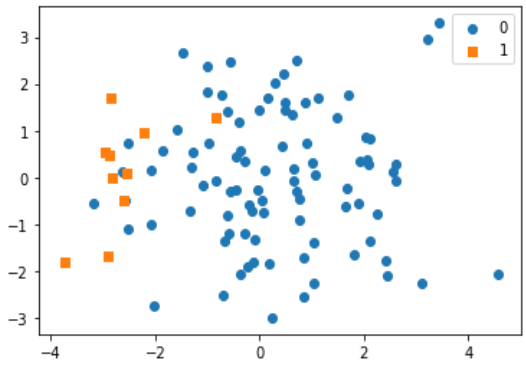

Fig.6.b. Balancing by Hybrid Method

5. Conclusion

we have inspected the issue of picking up from multi-class uneven datasets in this paper. Such circumstances speak to a significantly more imperative test than parallel classes. One needs to consider the lopsidedness extent, yet also the associations among classes as we have distinctive majority and minority gatherings.

The proposed examination allowed to expand a progressively significant comprehension into the possibility of multi-class imbalanced issues and model sorts present inside. General closures came to right currently be used in extended work to design new preprocessing learning estimations that may combine this establishment data about issue structure. Closures gave right currently structure an explanation behind various suggestion related to multi-class imbalanced learning.

References

[1] Beyan, Cigdem, and Robert B. Fisher, "Classifying Imbalanced Data Sets Using Similarity Based Hierarchical Decomposition," Pattern Recognition, vol. 48, no. 5, pp. 1653–1672, 2015. https://doi.org/10.1016/j.patcog.2014.10.032.

[2] Galar, Mikel, et al., "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," Systems Man and Cybernetics, vol. 42, no. 4, pp. 463–484, 2012. https://doi.org/10.1109/TSMCC.2011.2161285.

[3] Joshi, Mahesh V., et al., "Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements," Proceedings 2001 IEEE International Conference on Data Mining, 257–264, 2001.

[4] Ling, Charles X., and Victor S. Sheng, Cost-Sensitive Learning and the Class Imbalance Problem, University of Western Ontario, 2008.

[5] Chawla, N., et al., "Special issues on learning from imbalanced data sets," ACM SigKDD Explorations Newsletter, vol. 6, no. 1, pp. 1–6, 2004. https://doi.org/10.1145/1007730.1007733.

[6] Chawla, Nitesh V., et al., "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, no. 1, pp. 321–357, 2002. https://doi.org/10.1613/jair.953.

[7] Bunkhumpornpat, Chumphol, et al., "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem", PAKDD '09 Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pp. 475– 482, 2009. https://doi.org/10.1007/978-3-642-01307-2_43.

[8] Han, Hui, et al., "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," International Conference on Intelligent Computing, pp. 878–887, 2005. https://doi.org/10.1007/11538059_91.

[9] N. Ofek, L. Rokach, R. Stern, and A. Shabtai, "Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem," Neurocomputing, vol. 243, pp. 88–102, 2017. https://doi.org/10.1016/j.neucom.2017.03.011.

[10] He, Haibo, and Edwardo A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009. https://doi.org/10.1109/TKDE.2008.239.

[11] Błaszczyński, Jerzy, and Jerzy Stefanowski, "Neighbourhood Sampling in Bagging for Imbalanced Data," Neurocomputing, vol. 150, pp. 529–542, 2015. https://doi.org/10.1016/j.neucom.2014.07.064.

[12] H. Han, W.-Y. Wang and B.-H. Mao, Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, in Proc. Int. Conf. Advances in Intelligent Computing (ICIC) (2005), pp. 878–887.

[13] H. He, Y. Bai, E. Garcia and S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in Proc. Int. Joint Conf. Neural Networks (2008), pp. 1322–1328.

[14] H. He and E. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21(9) (2009) 1263–1284.

[15] D. Devi, S. K. Biswas, and B. Purkayastha, "Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique," Conn. Sci., vol. 31, no. 2, pp. 105–142, 2019. https://doi.org/10.1080/09540091.2018.1560394.

[16] F. Rayhan, S. Ahmed, A. Mahbub, R. Jani, S. Shatabda, and D. M. Farid, "CUSBoost: Cluster-Based Under-Sampling with Boosting for Imbalanced Classification," 2nd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2017, pp. 1–5, 2018. https://doi.org/10.1109/CSITSS.2017.8447534.