International Journal of Research in Engineering, Science and Management
Volume-3, Issue-2, February-2020
www.ijresm.com | ISSN (Online): 2581-5792

A Survey of Data Balancing Technique for Multi-Class Imbalanced Problem

Deore Mrunalee Chhotu[1], J. R. Mankar[2]
[1]*M.E. Student, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, India*
[2]*Professor, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Nashik, India*

***Abstract*: The imbalanced dataset contains skewed distribution of data. Such data distribution generates difficulties for machine learning algorithms. It may lead to generate false results in case of data imbalance. Various techniques proposed in a literature to balance a dataset using oversampling or under sampling methods. The study of these techniques is done independently. A little work has been done with the combined study of these two techniques. Some techniques have constraint on dataset type i.e. some techniques only works on numerical dataset. This paper aims to study different techniques for data balancing its advantages and limitations.**

***Keywords*: Oversampling, Under sampling, Hybrid dataset, Mahalanobis distance, Cluster based under sampling, Imbalance data, Classification.**

1. Introduction

A Lot of applications generates supervised data. i.e. data is divided in number of classes depending on its feature values. The number of records (instances) varies from class to class. If there is a large difference between numbers of instances for each class in a dataset, then such dataset is called as imbalanced dataset. The class contains less number of instances is called as minority class whereas class having greater number of instances as compared to other classes is called as majority class. If the dataset has two classes and there is large difference in number of instances in each class then one class is called as majority class and the other one is called as minority class. In multiclass dataset, more than one class can be labeled as minority class. Variety of applications generates such imbalanced data. The application includes disease diagnosis, activity recognition, fraud detection, protein fold etc.

The performance of machine learning algorithm degrades in case of imbalanced data. The main aim of the study of imbalanced data handling is to improve the accuracy of machine learning algorithm for minority class without hampering the accuracy of majority class.

The imbalanced data handling solution is classified in two main categories:

A. Data level

The data level solution is the preprocessing task. This is applied before machine learning algorithms. Initially data is balanced and then it is given to the machine learning algorithm. The data level solution has two techniques for data balancing:

B. Oversampling

In oversampling the number of instances of minority class is increases by adding few dummy instances. The dummy instances are created by analyzing the existing instances in that class.

C. Under sampling

In under sampling the majority class instance count is reduced with respect to the minority class instance count.

The oversampling technique may lead to overfitting and overgeneralization problem whereas under sampling may face problem of information loss and may mislead the classification technique.

1) Algorithmic level

In the algorithmic level data balancing is virtually done in machine learning algorithms. In learning phase the balancing can be done using adjusting the threshold value or adjusting the probabilistic estimate. It can also be done using one class learning and ensemble learning.

The binary class imbalanced problem can be handled using resampling or under sampling using data level solution or the classifier threshold can be shifted towards minority class using algorithmic level. But the same solution cannot be directly applied for multiclass imbalanced problem because: the relation among classes is not obvious and the classes boundaries may overlap. Following section elaborates the work done in the domain of data balancing. In the section 3, analysis of existing systems and problem formulation is elaborated. At the end conclusion is stated.

2. Related Work

The data reduction methods are classified in two categories based on the technique used in it. The wrapper methods take the help of classifier for reduction purpose. Hart at. El. [2] proposes a wrapper based reduction algorithm.

The multiclass imbalanced problem is converted in to two

class data imbalance problem by converting the problem as multiple two class sub problems. The widely used two class imbalance handling strategies are: one-versus-one(OVO) and one –versus-all(OVA). Z.-L. Zhang [3] proposes an ensemble learning based on OVA and OVO.

Simply duplicating the instances in minority class may create overfitting problem whereas randomly deleing the samples of majority class may lose some important information. Chuanxia Jian, et. al. [4] proposes a new method called as different contribution sampling method (DCS). It works on contributions of the support vectors (SVs) and non-support vectors in classification technique. This technique uses biased support vector machine (B-SVM) method to find SVs and the NSVs from imbalanced data.

SMOTE is widely used technique to balance the class distribution. It is similar to interpolation. In this technique initially K nearest neighbors are selected from minority class and then calculate the difference between them. The new samples are generated within the difference range. For this the difference value is multiplied by the random value between [0,1] and the generated value is added to the original minority class sample. SMOTE technique fails to preserve the class covariance structure. It increases the overlapping between various classes and disturbs the class boundaries [5].

Das et al. [6] proposes an oversampling technique based on joint probability distribution of data attributes. It uses Gibbs sampling for minority class sample generation process. Unlike SMOTE this strategy focuses on individual class properties and mutual relation among multiple classes. This increases the classification accuracy and the resultant dataset preserves the class covariance structure.

Lin et al. [7] proposes a neural network based oversampling technique. A dynamic sampling procedure DyS technique uses train multilayer perceptron's (MLP). It iteratively selects instances from dataset based on the probability and generate final training set for MLP for multi-class imbalanced classification.

Abdi and Hashemi [2] proposes a Mahalanobis distance based oversampling technique. This is a distance based oversampling technique. In this technique one class with highest samples is treated as majority class and all other classes are treated as minority class. The new samples are generated for each minority class equal to the number of instances in majority class. This is a good sampling technique for multiclass imbalanced problem with overlapped class structure. It preserves the class co-variance structure. This MDO technique is applicable for only numeric dataset.

Xuebing Yang, et. al [1] proposes an Adaptive Mahalanobis Distance-based Over-sampling AMDO technique. This is an extended version of MDO. This AMDO works on hybrid dataset. This technique balances the dataset based on Imbalanced Ratio(IR). For hybrid dataset learning it uses Heterogeneous Value Difference Metric distance (HVDM) and Generalized Singular Value Decomposition(GSVD). Using HVDM it calculates the K2 nearest neighbor and using GSVD it transforms the minority class samples to the PC space. The new samples are generated using Mahalanobis distance. The technique preserves the class covariance structure.

Unlike oversampling under sampling is also used to balance the dataset. Wei-Chao Lin, et., al. [9] proposes a clustering based under sampling technique. In this technique majority class instances are clustered and form each cluster only a single representative instance is selected. The representative instance is preserved and all other instances in a cluster are deleted. The representative of cluster is selected using 2 strategies: 1 is the cluster centroid and 2: nearest neighbor of cluster center. The c4.5 classifiers is used to evaluate the classification accuracy.

Yen et. al. [10] proposes a study on Cluster-based under-sampling approaches. This technique selects the representative data from majority class to improve the classification accuracy of minority class. The experimental results shows that cluster based under sampling generate better results than the other under sampling technique.

Li H et.al.[11] proposes a combined approach of oversampling and under sampling technique to balance a dataset. Improved SMOTE (ISMOTE) as over-sampling technique with distance-based under-sampling (DUS) technique are combined together. The proposed solution generates better results than the individual oversampling or under sampling technique.

Junsomboon et. al. [12] also proposes a combined approach of oversampling and under sampling. This technique combines the Neighbor Cleaning Rule (NCL) and Synthetic Minority Over-Sampling Technique (SMOTE) techniques. NCL technique removes the outliers in majority class. SMOTE is used to increase the samples of minority class. The results show that recall rate is improved as compared to the individual techniques.

3. Analysis and Problem Formulation

There are various techniques for data balancing these techniques are mainly classified as: data level and algorithmic level solutions. The data level solutions are applied as a preprocessing step. It is mainly categorized in 2 sections: oversampling and under sampling. These techniques are mostly studied independently. Very few approaches are proposed to have combined oversampling and under sampling solution.

There is need to develop an ensemble approach for multiclass hybrid data balancing which combines the oversampling and under sampling strategies.

A. Algorithm

Algorithm: AMDO: Adaptive Mahalanobis Distance-based Over-sampling
Input: S: Imbalanced dataset,
A: Metadata of all attributes
K1, K2: System constants
Output: S': updated dataset

Processing:

1. Initialize: c: set of classes, p1: set of numeric attributes, p2: set of nominal attributes, m: distinct values of nominal attribute, D: Class distribution, nmaj: number of samples in majority class
2. Calculate Orate for all minority Cn-1 classes using algorithm
3. For x =1 to cn-1
4. Sx = Read samples of class x
5. For each sample i in Sx
6. Find K2 nearest neighbors using HVDM matrix
7. Num(i) = number of nearest neighbors of selected sample i
8. Weight(i) = num(i)/ K2
9. If Num(i)<K1
10. Remove i from Sx
11. End
12. End
13. Obtains Xsup = transform nominal attributes to m distinct attributes
14. Calculate μ_s: Mean and σs : standard deviation of Xsup
15. Xsup = $\frac{Xsup-\mu_s}{\sigma s}$, Normalize sample set
16. Generate matrix N and X from Xsup
17. Update Xsup = N1/2XsupM1/2
18. Compute V via GSVD of Xsup using N and M
19. Obtain diagonal vector of coefficients of matrix V
20. If Orate if class >0 then
21. For i=0 to Orate
22. Choose random sample from Xsup
23. Compute α using Mahalanobis distance function
24. Generate p11 + m positive random numbers
25. r1,..rp11+m and make them sum to one
26. for k= 1: p11 + m
27. calculate Xk = $\frac{Xk^2}{\alpha Vk}$
28. end
29. Add Xk to Xsnew
30. End
31. End
32. Xsnew = σs(N-1/2XsnewVTM(-1/2) + μ)
33. End
34. Update S' = S+Xsnew
35. Return S'

[13]

4. Conclusion

In this paper various data balancing techniques are studied. In under sampling, cluster based solution is widely used technique. For oversampling AMDO technique is better solution for multiclass hybrid imbalance dataset. The combined approach generates better results than the individual oversampling or under sampling techniques. There is need to develop a system that proposes a combined approach for multiclass imbalance data handling problem that should be able to deal with hybrid dataset. A combined approach using Cluster based under sampling and AMDO oversampling technique may generate better results than the individual approaches.

References

[1] Xuebing Yang, Qiuming Kuang, Wensheng Zhang, and Guoping Zhang, "AMDO: An Over-Sampling Technique for Multi-Class Imbalanced Problems", IEEE Trans. Knowl. Data Eng., vol.30, no. 9, pp. 1672 - 1685 Sept. 2018

[2] L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," IEEE Trans. Knowl. Data Eng., vol. 28, no. 1, pp. 238-251, Jan. 2016.

[3] Z. L. Zhang, B. Krawczyk, S. Garcia, A. Rosales-Perez, and F. Herrera, "Empowering ono-vs-one decomposition with ensemble learning for multi-class imbalanced data," Knowl.-Based Syst., vol. 106, no. C, pp. 251–263, Aug. 2016.

[4] C. X. Jian, J. Gao, and Y.-H. Ao, "A new sampling method for classifying imbalanced data based on support vector machine ensemble," Neurocomputing, vol. 193, no. C, pp. 115–122, Jun. 2016.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol 16, no. 1, pp. 321–357, Jan. 2002.

[6] B. Das, N. C. Krishnan, and D. J. Cook, "RACOG and wRACOG: Two probabilistic oversampling techniques," IEEE Trans. Knowl. Data Eng., vol. 27, no. 1, pp. 222–234, Jan. 2015.

[7] M.-L. Lin, K. Tang, and X. Yao, "Dynamic sampling approach to training neural networks for multiclass imbalance classification," IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 4, pp. 647–660, Apr. 2013.

[8] A. Orriols-Puig and E. Bernado-Mansilla, "Evolutionary rulebased systems for imbalanced data sets," Soft. Comput., vol. 13, no. 3, pp. 213–225, Oct. 2008.

[9] Lin, Wei-Chao & Tsai, Chih-Fong & Hu, Ya-Han & Jhang, Jing-Shang. (2017). Clustering-based undersampling in class-imbalanced data. Information Sciences, 2017.

[10] Yen, Show-Jane & Lee, Yue-Shi., "Cluster-based Under-Sampling Approaches for Imbalanced Data Distributions", Expert Systems with Applications, vol. 36, 2008.

[11] Li H., Zou P., Wang X. and Xia R., "A New Combination Sampling Method for Imbalanced Data", In: Sun Z., Deng Z. (eds) Proceedings of 2013 Chinese Intelligent Automation Conference. Lecture Notes in Electrical Engineering, vol 256. Springer, Berlin, Heidelberg

[12] Junsomboon, N. and Phienthrakul, T. Combining over-sampling and under-sampling techniques for imbalance dataset. In Proceedings of the 9th International Conference on Machine Learning and Computing, 243–247 (ACM, 2017).