© 2019 IJSRSET | Volume 6 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET196225

A Review on Data Level Approaches for Managing Imbalanced Classification Problem

Ayushi Chaplot, Naveen Choudhary, Kalpana Jain
Department of CSE, College of Technology and Engineering, Udaipur, Rajasthan, India

ABSTRACT

In real world, the distribution of dataset is not in symmetric form. It can vary from application to application and distribution of data in that application. The un-symmetric form of this distribution is called imbalanced class distribution or skewed class distribution. So, the classification of data with skewed distribution of class can lead to the poor performance of the classifier. To solve the problem of imbalanced dataset in which the instances of one class is more than the instances of other class, there are different data level approaches for handling imbalanced classes. So, in this paper we will discuss about different data level approaches and have comparative study among them.

Keywords : Imbalanced data, Oversampling, Undersampling, Multiclass Classification.

I. INTRODUCTION

In recent years, class imbalance issue has received huge consideration in different fields like Machine Learning and Data Mining. Class imbalance problem occurs when the margin between the sizes of different classes in the dataset is very high. The proportion of classes present in dataset is not same.

The class which has less number of instances i.e. it has low proportion is known as minority class/positive class while the class with high proportion is known as majority class/negative class.

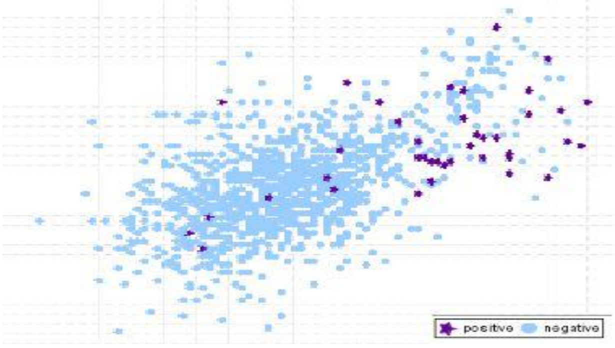


Fig.1 Imbalanced Dataset

In Fig.1 [1] purple stars represent minority class and blue circles represent majority class.

When the proportion of classes present in the dataset is not same then while doing classification the classifier will be biased towards the class which has more number of instances i.e., the majority class and this will lead to misclassification and poor performance of classifier. The class imbalance problem is important mainly in the applications where the misclassification of examples from the positive class (minority class) is expensive.

Let's take an example of the diagnosis of rare disease like tumour. If the tumour is considered as the positive class and the non-tumour is considered as the negative class then diagnosis of the patient who has tumour is very important as the sort of misclassification can lead to the death of the patient. Some other real world applications include rare disease diagnosis, detection of fraud in electricity service, detection of credit card fraud etc.

So to solve class imbalance problem there are many data level approaches. These data level approaches includes different form of resampling technique as shown in Fig.2. Resampling technique consist of series of methods which is used to reconstruct the dataset by either increasing the instances in the minority class to improve its weight or decreasing the instances from majority class to reduce the biasness of classifier towards the majority class.

Resampling includes: Oversampling and undersampling.

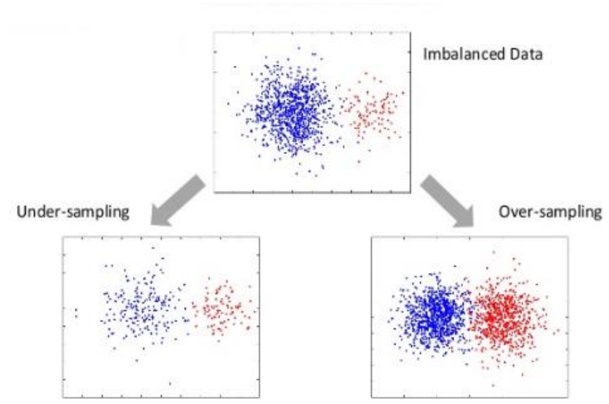


Fig.2 Resampling Techniques

In oversampling technique, increase frequency of minority class by either replicating the samples from the dataset or by creating new synthetic samples. In undersampling technique, the main objective is to decrease the instances of the minority class by deleting instances from the dataset or by data cleaning methods by removing inconsistent instances like noisy data, borderline data, outliers and redundant data.

As classification is data dependent so the quality and property of data is very important for the accuracy of the classification. So presence of inconsistent data may lead to the poor classification of dataset.

Data quality issues are:

Noisy Data: Noisy data affects the data quality to a great extent. Noisy data is data which is present in the area of another class or we can say that instances of minority and instances of majority are overlapping each other. So while doing classification it creates worst impact on classifier.

Borderline Data: Borderline data are present near the class boundaries of majority and minority classes. It is located at the overlap area of classes. So if any amount of perturbation added to instance of any class then it will lead to misclassification because any amount of noise added can lead to classification of instance of minority class as the instance of majority class or vice versa.

Outliers: An outlier is an extreme point that is far away from the normal distribution of data. The point that is outlier neither belongs to minority class nor to the majority class. Data outliers can diminish and deceive the classification process and it will provide less accurate results.

Redundant Data: Data redundancy is repetition of data means same data points are repeated at multiple locations. So removal of these redundant instances are important otherwise there will be problem of overfitting which will lead to improper classification of classes.

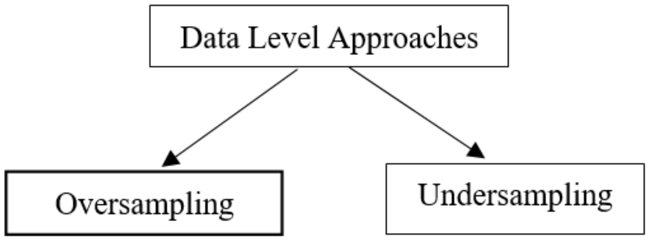
So to overcome the imbalance class problem and data quality issue there are resampling techniques which includes oversampling techniques and undersampling techniques which are further discussed in next section.

II. DATA LEVEL APPROACHES

Managing imbalanced datasets requires strategies such as enhancing classification algorithms or balancing classes in the dataset. This step is called data processing. Data processing is an important step before providing data to any classification algorithm.

The main objective is either to increase instances of minority class or decrease instances of majority class.

This is done to obtain approximately same number of instances in both majority and minority classes. Increasing frequency of minority class is known as oversampling and decreasing frequency of majority class is known as undersampling.



A. Oversampling

Oversampling is the part of data level approach in which we step-up the instances of minority class either by replicating the instances present in minority class or by creating synthetic samples with the help of instances present in minority class to increase frequency of minority class. So it will reduce imbalance ratio among classes and also misclassification lead by classifier.

Some oversampling techniques include:

1) Random Oversampling Technique (ROS): ROS is an oversampling technique in which the number of instances of minority class are increased. In this we take instances from minority class and replicate that minority instances to some random time according to the imbalance ratio. In this number of instances of minority class will be increased but not the variety of instances in class. So it will lead to the problem of overfitting.

2) Synthetic Minority Oversampling Technique (SMOTE): To solve the problem of overfitting in Random Oversampling Technique, new technique is introduced by Chawla et al. [2] that is SMOTE. So in SMOTE synthetic samples are created instead of replicating same instances. So in SMOTE we work on feature space rather than data space.

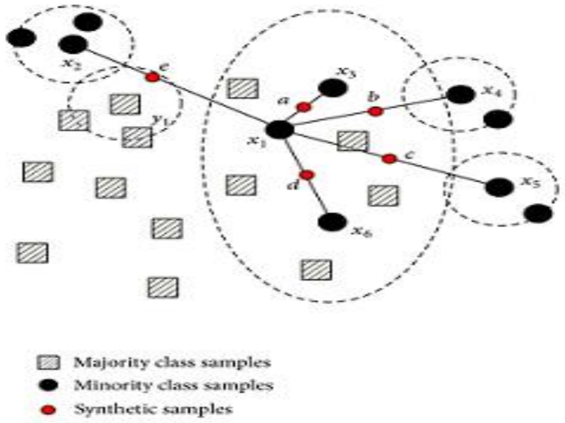


Fig.3 SMOTE

The working rule of the SMOTE technique is based on k-nearest neighbour method by looking for the value of nearest neighbour which is adjacent for each instance in the minority class. After that synthetic data point will be created as many as to reduce imbalance ratio between minority class instances and k-nearest neighbour which is selected at random. The main objective is to increase weight of minority class.

$$X \text{ new} = X + \text{Rand } (0, 1) * (X' - X)$$

The stages of the SMOTE method are:

(1) X represents every minority class on the dataset.

(2) Then, looking for k-nearest neighbour by selecting one of the closest k's represented as X'.

(3) The next stage is using linear interpolation between X and X' to build a new minority class sample.

So SMOTE will increase weight as well as the variety of minority class.

3) Adaptive Synthetic (ADASYN) Oversampling Technique: He et al. [3] proposed ADASYN is same as SMOTE just with slight change in it. In the SMOTE it creates synthetic samples in same way for all the minority samples. So all the synthetic samples are

linearly correlated with their parents. While in ADASYN after creating synthetic samples, a small value is added to all the synthetic samples so that it can have more variance in the minority class. And also all the samples are more scattered with ADASYN than samples created by SMOTE.

4) Borderline Synthetic Minority Oversampling Technique(B-SMOTE): Han et al.[4] proposed B-SMOTE is a variant of SMOTE in which instead of creating synthetic instances from all the minority data point, we will only consider borderline data points of minority class to create synthetic samples. As borderline instances are very prone to misclassification so we will divide minority instances into safe, noise and borderline instances and only use borderline instances to increase weight and variety of minority class.

B. Undersampling

Undersampling is the part of data level approach in which we step-down the instances of majority class either by eliminating the instances present in majority class randomly or by deleting inconsistent instances present in majority class to decrease frequency of majority class. So it will reduce imbalance ratio among classes and also misclassification lead by classifier.

Some undersampling techniques include:

1) Cluster Centroid: Cluster Centroid is an undersampling technique in which firstly we will cluster majority samples into k-clusters and for undersampling we will replace majority samples with the centroids of different clusters.

2) Random Undersampling (RUS): RUS is an undersampling technique in which the number of instances of majority class are decreased .In this we randomly select some instances from majority class to represent themselves as the majority class and other all instances are removed. So number of instances that are removed from majority class may contain some important information, so while doing classification it may lead to poor performance of classifier.

3) NearMiss: Nearmiss is a catechistic approach which used K-nearest neighbour technique. In these it uses different rules to select instances from majority class. Nearmiss1 selects samples from majority class for which the average distance of the minority samples which are k-nearest is smallest. Nearmiss2 selects samples from majority class for which average distance of farthest samples from minority class is smallest.

4) Tomek Links: Random undersampling has disadvantage that it can lead to removal of important information from majority class. So to avoid this Tomek [5] proposed Tomek links in which instead of removing important information it will lead to the removal of inconsistent instances such as noisy and borderline data. Tomek Links discards unwanted overlap between instances of different classes. It will remove the instance of the majority class if its nearest neighbours are not of majority class and of different class. Tomek links are based on the closeness of two instances belonging to different classes. Given two instances Ij and Ik belonging to different classes, and the distance between them is $dist(Ij , Ik)$, if there is no other instances Il such that $dist(Ij , Il) < dist(Ij , Ik)$ then the pair (Ij , Ik) is called a Tomek link. Thus if two instances form a Tomek Link then either an instance is noisy data or borderline data.

5) Edited Nearest Neighbour: Wilson [6] proposed Edited nearest neighbour works on principle of nearest neighbour rule and it “edit” the dataset meaning removing instances of majority class which do not follow neighbourhood principle. For each data point to be undersampled its nearest neighbour are considered and if all the neighbours are of different class than the instances then that instance will be removed.

Another extended version of Edited nearest neighbour is **Repeated Edited Nearest Neighbours** in which algorithm will be repeated will be multiple times. Thus it lead to more removal of data than Edited Nearest Neighbour.

6) Condensed Nearest Neighbour (CNN): Hart [7] proposed Condensed Nearest Neighbour is an undersampling cleaning methods which removes inconsistent instance like redundant data. CNN works on principle of 1-NN rule as put all minority class instance in set S, take one majority class instance and put in set S and all other majority class instances in set C. Take each sample from set C one by one and classify each sample from set C using 1-NN rule. If that sample is classified than discard that sample and if not classified properly than put in set S and reiterate on set C until all the instances are checked by 1-NN rule.

But problem with 1-NN rule is that it is very sensitive to noise and it will noisy samples to the dataset which can lead to misclassification.

7) One Sided Selection: To overcome the problem of condensed nearest neighbour, Kubat and Matwin [8] proposed One Sided Selection which is a hybrid approach obtained from combination of Tomek link and Condensed Nearest Neighbour. In this firstly Tomek link is used to remove noisy and borderline data. Then CNN is used to remove redundant data. But in these firstly noisy data is removed so CNN will not be very sensitive and will not produce any noisy samples. But as in One sided Selection, CNN is used and it follows 1-NN rule, so whole undersampling will be on the basis of one single instances selected from majority class. So if in any dataset if there are two subset of majority class than there will be undersampling of only one subset as only one instance is selected from majority class and other subset will be untouched. So it will lead to overfitting because of undersampling of only one subset.

8) Neighbourhood Cleaning Rule (NCL): Laurikkala [9] proposed Neighbourhood Cleaning Rule is undersampling cleaning method which is used to clean inconsistent instances. NCL's main idea is that in this K- nearest neighbour are considered for a sample in majority class. If the label/class of an instance of majority class is different from half of its nearest neighbour then that sample will be removed considering that sample as noisy or borderline sample. Thus in NCL more concentration is on data cleaning than just data reduction.

C. Hybrid Approaches

Hybrid approaches is the combination of undersampling and oversampling. In this oversampling is used for increasing frequency of minority class and then undersampling is used for data cleaning where inconsistent instances will be removed.

SMOTE + Tomek Links [10] and SMOTE + ENN [11]: In minority class synthetic samples are created to step-up frequency of instances present using SMOTE technique. But SMOTE technique while creating synthetic samples can also create noisy samples. So for the removal of noisy samples Edited Nearest Neighbour technique or Tomek link can be used.

III. CLASSIFICATION OF DATASET WITH MULTIPLE CLASSES

While doing balancing in different datasets different number of classes are present. Proportion of classes present in different datasets are not same.

If a dataset is having two classes then while balancing we will either oversample the minority class or under sample the majority class or we can apply hybrid approach by applying undersampling on majority class and oversampling on minority class. This type of distribution is known as Binary Class Distribution.

If more than two classes are present in any dataset then this type of distribution is known as multiclass distribution. But there is a problem with multiple class [12] because solutions that are applicable for balancing binary class may not be directly applied to multiple classes. If we apply data level approaches for balancing multiclass dataset then there will be increased search space problem. So for the conversion of multiclass dataset into binary class dataset there are binarization techniques such as:

A. One Vs. One (OVO): One vs. One strategy is applied on a dataset to solve multiclass classification problem. In this we will built binary classifier for each pair of classes from multiclass dataset. We will built classifier for each class with every other class. Like if we take N classes in a dataset then one vs. one strategy will build N(N-1)/2 binary classifiers.

B. One Vs. Rest (OVR) / One Vs. All (OVA): In one vs. all approach we will build individual classifier for all the classes by keeping instances of current class as the positive and rest instances of all other classes as negative.

At the time of classification each model F1, . . . ,FC created will be checked such that particular instance belong to which of the model by checking degree of the membership of the instance from all the models.

IV. CONCLUSION

Thus this paper represents an overview of the problems faced while classifying imbalanced dataset and also data quality issues which lead to misclassification of classes. So to solve imbalanced class problem and data quality issues we discussed data level approaches in this paper. This data level approaches includes oversampling and undersampling. Oversampling is used to balance the dataset by increasing instances of minority class. Also Undersampling is used for balancing the dataset by decreasing instances of majority class and also it is used for data cleaning purpose. Also there are hybrid approaches which is combination of oversampling for balancing instances of classes and undersampling is used for data cleaning purpose. In this paper we also discuss binarization techniques for balancing multiple class problems. We can't apply data level approaches directly on multiclass data so we convert it into binary class using binarization techniques.

Thus this paper will be helpful for researchers who wants to have knowledge about class imbalance problems and data level approaches for solving class imbalance problems.

V. REFERENCES

1. V. Lopez, A. Fernandez, S. Garcia, V. Palade and F. Herrera, "An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics," Elsevier Journal of Information Sciences, vol. 250 pp. 113-141, November 2013.
2. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16 pp. 321-357, June 2002.
3. H. He, Y. Bai, E.A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," IEEE International Joint Conference on Neural Networks, pp. 1322-1328, June 2008.
4. H. Han, W.Y. Wang, and B.H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," International Conference on Intelligent Computing, vol. 3644 pp. 878-887, August 2005.
5. I. Tomek, "Two modifications of CNN" IEEE Transactions on System Man and Cybernetics, vol. 6: pp.769-772, November 1976.

6. D. Wilson, “Asymptotic Properties of Nearest Neighbour Rules Using Edited Data” IEEE Transactions on Systems, Man, and Cybernetrics, vol. 2 pp. 408-421, July 1972.
7. P. Hart, “The condensed nearest neighbour rule,” Information Theory, IEEE Transactions on, vol. 14 pp. 515-516, May 1968.
8. M. Kubat, S. Matwin, “Addressing the curse of imbalanced training sets: one-sided selection,” the Fourteenth International Conference on Machine Learning, vol. 97 pp. 179-186, July 1997.
9. J. Laurikkala, “Improving identification of difficult small classes by balancing class distribution,” Springer Berlin Heidelberg, vol. 2101 pp.63-66 June 2001.
10. G. Batista, R. C. Prati, M. C. Monard. “A study of the behavior of several methods for balancing machine learning training data,” ACM Sigkdd Explorations Newsletter, vol. 6 pp. 20-29, June 2004.
11. G. Batista, B. Bazzan, M. Monard, “Balancing Training Data for Automated Annotation of Keywords:a Case Study,” The Second Brazilian Workshop on Bioinformatics, pp. 35-43, December 2003.
12. R. Barandela., J.L.Sánchez, V. García and E. Rangel, “Strategies for learning in class imbalance problems,” Pattern Recognition, vol.36 pp. 849-851, September 2003.

Cite this article as :

Ayushi Chaplot, Naveen Choudhary, Kalpana Jain, "A Review on Data Level Approaches for Managing Imbalanced Classification Problem", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), ISSN : 2456-3307, Volume 6 Issue 2, pp. 91-97, March-April 2019.
Available at doi :
https://doi.org/10.32628/IJSRSET196225
Journal URL : http://ijsrset.com/IJSRSET196225