A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique

Miss. Mayuri S. Shelke¹, Dr. Prashant R. Deshmukh², Prof. Vijaya K. Shandilya
¹Student of Master of Engineering in (CSE), Sipna College of Engineering and Technology, Amravati, India
²Professor, Amravati, Maharashtra, India
³Associate professor, Sipna College of Engineering and Technology, Amravati, India

Abstract: In today’s era of internet the amount of data generation is growing on increasing. Some of the data related to medical, e-commerce, social networking, etc. are of great importance. But many of these datasets are imbalanced that is some records belonging to same category are in much large number and some are very rare. For extracting useful date from such large dataset different data mining or machine learning techniques are used. But these imbalanced nature of the datasets affects the performance of a classifier very greatly. To deal with this it is necessary to understand the problem of imbalanced learning. There are various Undersampling and oversampling techniques available which try to resolve imbalanced learning problem. This paper, performs the study of this imbalance nature of the datasets and different techniques of oversampling and Undersampling that are used to balance the datasets.

Keywords: Imbalanced learning, Machine Learning, Oversampling, Undersampling.

I. INTRODUCTION

Oversampling and undersampling in data analysis are techniques used to adjust the class distribution of a data set (i.e. the ratio between the different classes/categories represented). Oversampling and undersampling are opposite and roughly equivalent techniques. They both involve using a bias to select more samples from one class than from another [1]. In today’s era of machine learning and data mining, many real world applications work on datasets mainly for performing analysis and generating recommendations and predictions. For performing these calculations the dataset should be properly balance. But sometimes it is seen that these datasets are imbalance in nature. Leading to the problem of imbalanced data. The data which has an unequal distribution of samples among classes is known as imbalanced data. The class having more samples is generally a majority class and a class which contains very scarce samples is a minority class. Such type of data sets pose a great challenge to the classifier as it becomes problematic to classify the minority samples precisely because of their fewer amount. Standard classification algorithms also fail to classify such form of imbalanced data accurately with least misclassification error.

There are major methods available to resolve the imbalanced learning problems which are nothing but a sampling, active learning, cost sensitive learning and kernel based methods. Sampling based methods provide the solution at data level by balancing the number of samples among classes. Undersampling and oversampling are two main methods of sampling in which samples are either reduced from majority class or samples are added in the minority class. Both techniques have their own advantages as well as drawbacks [2]. Active learning approaches focus mainly on acquiring labels to the unlabeled data. Another method is cost based method which provides solution to an imbalanced dataset at the algorithmic level. It uses cost matrix which represents costs associated with each representation. Besides of these methods, kernel based methods also work well in handling imbalanced datasets.
Most of the machine learning algorithms perform better when data sets are almost balanced. But problem arises when given data sets are very much imbalanced in nature. Classification of these imbalanced datasets is a very crucial task for the classifier as classifier may tend to favor the majority class samples. As a result of unequal distribution of data, majority class significantly dominates the minority class. To deal with such imbalanced learning problem many undersampling as well as oversampling techniques available. Random undersampling removes the majority instances randomly, which may lose some important information. But these methods ignore the distribution information on the training data set. Many oversampling techniques such as SMOTE, Borderline-SMOTE, ADASYN, RAMOBoost tries to solve class imbalance problem efficiently, but there are some limitations to this method. To overcome these inabilities, it is desirable to develop a solution which will integrate benefits of both undersampling and oversampling and handles the imbalanced data in the medical diagnostic field efficiently.

II. IMPORTANT TERMS

1) Imbalanced learning:
Imbalanced learning occurs whenever some types of data distribution significantly dominate the instance space compared to other data distributions. There are two types of imbalance learning problem [11]. i) Between class imbalances: Imbalance that exists between the samples of two classes is usually known as between class imbalance. ii) Within class imbalances: If the samples of the majority and minority classes have more than one concept in which some concepts are rarer than others then it is called within class imbalance.

2) Undersampling:
Undersampling methods work by reducing the majority class samples. This reduction can be done randomly in which case it is called random undersampling or it can be done by using some statistical knowledge in which case it is called informed undersampling. Some informed undersampling methods and iteration methods also apply data cleaning techniques to further refine the majority class samples.

3) Oversampling:
In oversampling method, new samples are added to the minority class in order to balance the data set. These methods can be categorized into random oversampling and synthetic oversampling. In random oversampling method, existing minority samples are replicated in order to increase the size of a minority class. In synthetic oversampling technique, artificial samples are generated for the minority class samples. These new samples add the essential information to the minority class and prevents its instances from the misclassification.

III. RELATED WORK
The authors in this paper [3] says that a combination of method of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance (in ROC space) than only under-sampling the majority class. Author also shows that a combination of our method of over-sampling the minority class and under-sampling the majority class can achieve better classifier performance (in ROC space) than varying the loss ratios in Ripper or class priors in Naive Bayes. Our method of over-sampling the minority class involves creating synthetic minority class examples. Experiments are performed using C4.5, Ripper and a Naive Bayes classifier. The method is evaluated using the area under the Receiver Operating Characteristic curve (AUC) and the ROC convex hull strategy.

The authors in paper [4] said that, traditional classification algorithms perform not very well on imbalanced data sets and small sample size. To deal with the problem, a novel method is proposed to
change the class distribution through adding virtual samples, which are generated by the windowed regression over-sampling (WRO) method [2]. The proposed method WRO not only reflects the additive effects but also reflects the multiplicative effect between samples. A comparative study between the proposed method and other over-sampling methods such as synthetic minority over-sampling technique (SMOTE) and borderline over-sampling (BOS) on UCI datasets and Fourier transform infrared spectroscopy (FTIR) data set is provided. Experimental results show that the WRO method can achieve better performance than other methods. But at the same time, the method of solving regression coefficients is in the local window, so the efficiency is not high.

The author in this paper [5], author gives way to handle data mining problems where class prior probabilities and/or misclassification costs between classes are highly unequal is to resample the data until a new, desired class distribution in the training data is achieved. The author here introduced the generative oversampling algorithm, a resampling algorithm that creates artificial data points from a probability distribution learned from the minority class. Empirically, author have shown that generative oversampling works well for a range of text classification datasets using linear SVMs. Moreover, if the best minority class training prior is unknown, generative oversampling has the added benefit of producing results which are robust to changes in minority class training prior. Generative oversampling is also simple to implement and can be used on a variety of different data types by selecting appropriate generative models. Therefore, it is a viable and flexible alternative whenever resampling methods are used.

The paper [6] introduces the importance of imbalanced data sets and their broad application domains in data mining, and then summarizes the evaluation metrics and the existing methods to evaluate and solve the imbalance problem. Synthetic minority oversampling technique (SMOTE) is one of the over-sampling methods addressing this problem. Based on SMOTE method, this paper [6] presents two new minority over-sampling methods, borderline-SMOTE1 and borderline-SMOTE2, in which only the minority examples near the borderline are over-sampled. The borderline examples of the minority class are more easily misclassified than those ones far from the borderline. Thus our methods only over-sample the borderline examples of the minority class, while SMOTE and random over-sampling augment the minority class through adding virtual samples, which are generated by the windowed regression over-sampling (WRO) method [2]. Experimental results show that these approaches achieve better TP rate and F-value than SMOTE and random over-sampling methods.

The author in this paper [7] suggested, the accuracy rate of the predictive model is not an appropriate measure when there is imbalanced problem due to the fact that it will be biased towards the majority class. Thus, the performance of the classifier is measured using sensitivity and precision. Oversampling and undersampling are found to work well in improving the classification. Here the author suggested two new method for balancing the datasets called as bagging and boosting. The bagging method is a bootstrap ensemble method that can be applied to enhance model stability. In the Bagging approach, all instances in the training dataset have equal probability to be selected. All samples were replicates based on bootstrap approach. The replicates are samples drawn with replacement and with the same size as the training sample. For each bootstrap set, one model is fitted. The final predictions of the cases are produced using the voting approach. Meanwhile, after calculations boosting and bagging did not improve the Decision Tree performance.

The author in this paper [8] said that Down-sizing the majority class results in a loss of information that may result in overly general rules. In order to overcome this drawback of the under-sampling approach Yen and Lee (2009) proposed an unsupervised learning technique for supervised learning called cluster based under-sampling. Their approach is to first cluster all the training samples into K clusters then chose appropriate training samples from the derived clusters. The main idea is that there
are different clusters in a dataset, and each cluster seems to have distinct characteristics. If a cluster has more majority class samples and less minority class samples, it will behave like a majority class sample. On the other hand, if a cluster has more minority class samples and less majority class samples, it does not hold the characteristics of the majority class samples and behaves more like the minority class samples. Therefore, their approach selects a suitable number of majority class samples from each cluster by considering the ratio of the number of majority class samples to the number of minority class samples in the derived cluster [9]. They first cluster the full data to K clusters. A suitable number (M) of majority class samples from each cluster are then selected by considering the ratio. This approach may be suitable for datasets where class labels are confidently defined and truly reflect the property of the labeled class. But as mentioned earlier that in some cases, especially for medical datasets, there is no guarantee that the given class labels are truly reflect the actual class of that record.

The author in this paper [10] said that many machine learning classification algorithms assume that the target classes share similar prior probabilities and misclassification costs. The problem of classification when one class has a much lower prior probability in the training set is called the imbalanced dataset problem. One popular approach to solving the imbalanced dataset problem is to resample the training set. However, few studies in the past have considered resampling algorithms on data sets with high dimensionality. Here, author examine the imbalanced dataset problem in the realm of text classification. Text has the added problems of both sparsity and high dimensionality. They first describe the resampling techniques we use in this thesis, including several resampling techniques we are introducing. After resampling, we classify the data using multinomial naïve Bayes, k nearest neighbor, and SVMs. Finally, compare the results of our experiments and find that, while the best resampling technique to use is often dataset dependent, certain resampling techniques tend to perform consistently when coupled with certain classifiers.

IV. SAMPLING TECHNIQUES

A. Undersampling

1. MLPUS
MLP-based undersampling technique (MLPUS) which will preserve the distribution of information while doing undersampling [1]. The MLPUS involves three key mechanisms: a) clustering of majority class samples b) selection of important samples using SM evaluation c) training of MLP using selected samples in SM evaluation. To select the representative samples from the majority class, it is again clustered using k-means and here, k is equal to the number of samples in the minority class. Here, Multilayer Perceptron (MLP) is used for the classification. Performance of MLP is depends mainly on values of learning rate and epoch.

2. EasyEnsemble
In EasyEnsemble method, majority class is divided into several subsets and the size of each subset is equal to the size of a minority class. Then for each subset, it develops a classifier using whole minority class and majority class subset. Results generated from all the classifiers are combined to get the final decision. To develop a classifier Adaboost is used. EasyEnsemble approach has been shown in Figure 3. As EasyEnsemble uses independent random sampling with replacement, it can be considered as an unsupervised learning algorithm [12].

3. BalanceCascade
This method follows the supervised learning approach [12]. BalanceCascade method works as follows: Subset of majority class is formed which contains a number of samples equal to the number of minority class sample. When C1 classifier is trained using the majority class subset and whole minority class, the samples from a majority subset which are correctly classified are removed. This
new generated sampled set of majority class is given as an input to $C_2$. The same procedure is iterated until final classifier is reached. At every classifier, the size of the majority subset gets reduced. In BalanceCascade there is a sequential dependency between classifiers. BalanceCascade differ from EasyEnsemble as it removes true majority samples in order to reduce redundancy.

### B. Oversampling:

#### 1. SMOTE

There are a number methods available to oversample a dataset used in a typical classification problem (using a classification algorithm to classify a set of images, given a labelled training set of images). The most common technique is known as SMOTE: Synthetic Minority Over-sampling Technique [2]. To illustrate how this technique works consider some training data which has $s$ samples, and $f$ features in the feature space of the data. Note that these features, for simplicity, are continuous. As an example, consider a dataset of birds for clarification. The feature space for the minority class for which we want to oversample could be beak length, wingspan, and weight (all continuous). To then oversample, take a sample from the dataset, and consider its $k$ nearest neighbors (in feature space). To create a synthetic data point, take the vector between one of those $k$ neighbors, and the current data point. Multiply this vector by a random number $x$ which lies between 0, and 1. Add this to the current data point to create the new, synthetic data point.

#### 2. RAMOBoost

Ranked Minority Oversampling in Boosting (RAMOBoost) is a technique which systematically generates synthetic samples depending on sampling weights. It adjusts these weights of minority samples according to their distribution. This method works in two stages. In first stage decision boundary is shifted towards the samples which are difficult to learn from both majority and minority classes. In the second stage to generate synthetic samples a ranked sampling probability distribution is used. If RAMOBoost adopts techniques used in SMOTE-N method, then it can handle datasets having nominal features.

#### 3. Borderline-SMOTE

As SMOTE generates synthetic samples for each minority sample it may lead to over generalization [11]. The main objective of Borderline-SMOTE is to identify minority samples located near decision boundary. Then these samples are used further for oversampling. This method focuses on borderline samples because classifier may misclassify them. In [6] two methods borderline-SMOTE1 and borderline-SMOTE2 has been proposed. Both methods give better results on TP rate and F-value as compared to SMOTE.

#### 4. ADASYN

Haibo He, E.A. Garcia, proposed a novel approach adaptive synthetic sampling to handle imbalanced data set. In synthetic sample generation process, there is no need to consider all minority samples as there may be problem of overlapping. ADASYN uses the weighted distribution of minority samples. It assigns weight to minority sample depending on importance of minority sample. Samples which are difficult to classify got higher weight than others. More samples are generated for the sample having a higher weight. ADASYN

#### 5. MWMOTE

Existing synthetic oversampling methods may have some insufficiencies and inappropriateness in many scenarios [13]. In order to overcome these problems, a new method has been proposed, namely, Majority Weighted Minority Oversampling Technique (MWMOTE). The objective of MWMOTE is twofold: to improve the sample selection process and to improve the synthetic sample generation process. MWMOTE involves three key phases.
• In the first phase, MWMOTE identifies hard-to-learn and the most important minority class samples from the original minority set, Smin and construct a set, Simin by the identified samples.
• In the second phase, each member of Simin is given a selection weight, Sw, according to its importance in the data.
• In the third phase, using the clustering approach, MWMOTE generates the synthetic samples from Simin using Sws and produces the output set, by adding the synthetic samples to Smin.

V. CONCLUSION
Currently, imbalanced learning becomes a challenging and active research topic in the area of machine learning. This paper gives a brief review on various methods for learning and solving imbalanced learning problem. It also provides a brief description of all methods under consideration. To handle imbalanced learning problem, there are several research directions for undersampling as MLPUS, EasyEnsemble, BalanceCascade and oversampling techniques such as ADASYN, RAMOBoost and MWMOTE. All these techniques can be generalized to solve multiclass imbalanced learning problem. Furthermore, these methods can also be modified to facilitate for incremental learning applications and balancing the datasets.

REFERENCES