**Effective Techniques for Handling Imbalanced Datasets in Machine Learning**

**Sunny Singh**

AI-Tech Systems

Rewari, India

[Snnsingh625@gmail.com](mailto:Snnsingh625@gmail.com)

***Abstract* – Data is powerful and important. All the machine learning algorithm will be useless without data. Without proper data, even the best of algorithms cannot deliver good results. On the other hand, even a simple algorithm can deliver brilliant results with proper datasets that have a substantial number of data points. The data which is collected online is raw and unstructured. In other words, one or more attributes can be imbalanced in such a way that obtaining any information from the data becomes quite difficult and visualization equally misleading. This is a very common problem that is encountered while training machine learning models called imbalanced data. This imbalanced dataset gives inaccurate results even when best of the models are used to process that data. When the data is biased, the results will also be biased, which is the last thing that any of us will want from a machine learning algorithm. This paper talks about, different techniques used to balance the data in hand. This paper also talks about the results of each of these techniques. This paper also focuses on the most and least effective techniques used. This paper talks about some of the most common techniques used like, under-sampling, over-sampling, feature selection, evaluation metrics, modified learning algorithmic solutions.**

***Keywords - Data, Imbalanced data, Under-sampling, over-sampling, feature selection, evaluation metrics, modified learning algorithmic solutions.***

1. INTRODUCTION

In recent years, machine learning is one filed in which a lot of research is going on. In machine learning, this problem of imbalance data has been an active area of research. In machine learning the algorithms build models based on train data and they work best when the number of samples in each class are about equal. This is because the ML algorithms are designed in such a way that it tends to maximize accuracy and reduce error. In real life problems, usually, datasets are unbalanced and this can cause seriously negative effect on built model. Imbalanced data set are the datasets which contains majority of the samples from one class, rather than rest of the classes. In unbalanced data sets at least one of the class is represented by only a small number of training examples (called the minority class) while other classes make up the majority class.

Machine Learning is being used in various fields, such as medical diagnosis, text categorization, fraud detection, oil-spills detection in satellite images, toxicology, cultural modeling, detecting network intrusions, managing risk, predicting failures of technical equipment and many more. The unbalanced data set problem appears in almost every sphere, in such problems most of the machine learning algorithms are biased towards the majority classes, and therefore, they show very poor classification rates on minority classes. Sometimes, it is also possible that the algorithms predicts everything as majority class and ignores the minority class. And this can have large scale implications. For example, when a medical diagnosis is done of a certain cancer, if the patient with cancer is a positive class, and non-cancer (healthy patient) a negative class, then missing a cancer patient (positive class) is much more serious. This could be fatal to the patient and he/she can lose his/her life because of the delay in the correct diagnosis and treatment. Therefore it becomes imperative to deal with data imbalance problems. In this paper I will talk about different techniques like, sampling, using algorithms, cost sensitive learning, used to solve this issue and classify them on the basis of their effectiveness.

1. WORK DONE IN DATA IMBALANCE

Basically there are two main methods in which this problem of data imbalance can be solved. The first method is called the “external” approach in which transformation or sampling is done from original unbalanced data set to create a balanced data set. This is called external approach because the external training data is balanced while the algorithms are unchanged. The second approach is the “internal” approach in which weights are adjusted within learning algorithms to use the original imbalanced data set in training.

3. DIFFERENT STRATEGIES USED

Sampling, bagging, boosting etc, are popular approaches which are used by the researchers to handle data imbalance problem. The details of these methods are as described.

3.1. SAMPLING STRATEGIES

Sampling is used widely to solve the problem of imbalance. This method is divided into two categories: Under sampling and Oversampling.

3.1.1. UNDER SAMPLING

Under sampling is a method which removes examples from the majority class to balance the data set. This is a method which is good for large scale applications, in which the number of majority class examples is very large. Therefore reducing the training samples reduces the training time and storage. The most important drawback of this method is that it removes potentially useful information that could be important for algorithms. Under sampling is divided into random and informative. In Random Under sampling it randomly removes the examples from the majority class until the data set gets balanced. In Informative Under sampling method, it selects the examples from majority class based on a selection criterion to make the data set balanced.

3.1.2. OVER SAMPLING

Oversampling is a method which is used to balance the data set by replicating the examples of minority class. In this method there is no loss of data as in under sampling, which is an advantage of this method. One main disadvantage of this method is that it leads to over-fitting and can introduce an additional computational cost if the data set is already fairly large but imbalanced. Just like under sampling, oversampling is also divided into two types. Random Oversampling and Informative Oversampling. Random Oversampling is the method which balances the data set by replicating the randomly chosen minority class examples. Informative Oversampling method synthetically generates minority class examples based on a pre-specified criterion. There are number of Oversampling methods available in the literature like SMOTE, Borderline SMOTE, OSSLDDD-SMOTE etc.

3.1.3 SMOTE

Over-sampling which replicates samples can lead to similar but more specific regions in the feature space as the decision region for the minority class. This can also lead to over fitting due to multiple copies of minority class examples. To overcome the problem of over fitting and broaden the decision region of minority class examples, this method of SMOTE is used. To illustrate how this technique works, let’s consider some training data which has s samples, and f features in the feature space of the data. For simplicity, let’s consider these example as continuous. As an example, consider a dataset of birds for classification. 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.

Many modifications and extensions have been made to the SMOTE method ever since its proposal.

3.2. BAGGING

In this method the original training data set is divided into N subsets of the same size. Each subset created from the training data is used to create one classifier (classifier learned from those subsets). A compound classifier is created as the aggregation of particular classifiers. This is a technique which can be used with many classification methods and also, it applies regression methods to reduce the variance associated with prediction which improves the prediction process. Prediction method is applied on each of the bootstrap sample and then the results are combined by averaging for regression and simple voting procedure for classification to obtain the overall prediction. Tests on real life and simulated data sets using classification and regression trees and subset selection in linear regression shows that this method can be used to get substantial gains in accuracy.

3.3. BOOSTING

Boosting is a technique used in machine learning which is based on the observation that finding many rough rules is easier than finding highly accurate prediction rule.

“Kerns et al.” introduced a method called **Boosting,** which weak learners into strong learners. The base machine learning algorithm is called repeatedly by the boosting algorithm, every time feeding it with a different subset of training examples. Every time it is called, the base learning algorithm generates a new prediction rule, and after many rounds, the boosting algorithm combines these rules into a single prediction rule that will be much more accurate than any one of the single rules. Variant of this algorithm is as described.

“W. Lee” also presented with a boosting algorithm to solve the problem of unbalanced data. In this method, a number of classifiers are trained using smaller and usually balanced subsets of the original data, which are combined in final classification step in ensemble process. These subsets which are created from the original data usually contain all minority instances and the same number of randomly selected majority instances. It focuses on those instances which are not already accurately learned using weights to decide the values of probability of selection.

“Yoav et al.” presenteda new method called **AdaBoost** algorithm also called the adaptive boosting to solve the problem of data imbalance. It main focus in on difficult data points. Difficult data points are generally the data points that are mostly misclassified by the previous weak classifier. AdaBoost works in such a way that it combines these weak classifiers into a comprehensive prediction by an optimally weighted majority vote of weak classifier. AdaBoost method is fast, simple and easy to program. Hyper parameter tuning is not required in AdaBoost. AdaBoost does not need prior knowledge about weak learner. It’s an effective method but it is vulnerable to uniform noise. AdaBoost uses weak classifiers which lead to low margins and over fitting.

3.4. ADAPTIVE SAMPLING METHODS AND SYNTHETIC DATA

GENERATION

The main objective here is to provide a balanced distribution of data from over-sampling and/or under-sampling techniques to improve prediction accuracy. When we look at synthetic sampling, SMOTE created synthetic data in minority class by selecting few of the nearest minority neighbors of a minority data and generating synthetic minority data along with the lines between the minority data and the nearest minority neighbors on the other hand adaptive sampling methods are proposed to generate synthetic data. The idea of Borderline-SMOTE technique was to find out the borderline minority samples. Then, synthetic samples were generated along the line between the borderline samples and their nearest neighbors of the same class.

3.5. COST SENSITIVE LEARNING

There are some more methods at the algorithmic level, in this the data imbalance problem can be tackled by adjusting the costs of the various classes. In case of decision trees adjusting the probabilistic estimate at the tree, adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning.

**Cost Sensitive Learning (CSL)** is a very common method used to handle the classification problem of imbalanced data sets. Cost sensitive learning apply the miss classification cost to incorrectly classified examples. In case of correct classifications, there is no penalty. The cost of FN will be more than the cost of FP and the costs of TP and TN is zero.

There are three different ways in which cost sensitive learning is used, the first technique apply misclassification costs to the data set as a form of data space weighting, the second technique applies cost-minimizing techniques to the combination schemes of ensemble methods, and the third and the last technique incorporates cost sensitive features directly into classification paradigms to essentially fit the cost sensitive framework into these classifiers.

3.6. RECOGNITION BASED METHODS

This is a type of method in which the classifier learns only on the minority class samples (target class). This method thus improves the performance of the classifier on unseen data and recognizes only those data points that belong to the majority class. This method which is also called one class learning can be a robust technique when dealing with unbalanced data and highly dimensional noisy feature space. This method of recognition based or one-class learning can perform better under certain conditions such as high dimensional data, however, many classifiers such as decision trees and Naive Bayes cannot be built by one class learning.

3.7. ALGORITHM BASED METHODS

3.7.1 CLASS CONFIDENCE PROPORTION DECISION TREE (CCPDT)

Wei Liu, Sanjay Chawla, et al., proposed a new decision tree algorithm, the Class Confidence Proportion Decision Tree (CCPDT), which is robust and doesn’t take into consideration the size of classes and generates rules which are statistically significant. To generate these specific rules which are statistically significant they design a novel and efficient top-down and bottom-up approach which uses Fisher’s exact test to prune branches of the tree which are not statistically considerable. Together these two changes defer a classifier that performs statistically better than not only traditional decision trees but also trees learned from data that has been balanced by well known sampling techniques.

3.7.2 SEMI SUPERVISED CLUSTERING

Mingwei Leng, et al., proposed an active semi supervised clustering algorithm that uses an energetic method for data selection to minimize the amount of labeled information, and it operates multi threshold to enlarge labeled datasets on multi density and imbalanced datasets. In this, three standard datasets and one synthetic dataset are used to demonstrate the algorithm, and the tentative outcomes show that the semi supervised clustering algorithm has a higher accuracy and a more stable performance in comparison to other clustering and semi supervised clustering algorithms, particularly when the datasets are multi density and imbalanced

3.7.3 SMOTE-BOOST

SMOTEBoost algorithm combines SMOTE and the standard boosting procedure (Chawla et al., 2003b). We want to utilize SMOTE for improving the accuracy over the minority classes, and we want to utilize boosting to maintain accuracy over the entire data set. The major goal is to better model the minority class in the data set, by providing the learner not only with the minority class instances that were misclassified in previous boosting iterations, but also with a broader representation of those instances.

The standard boosting procedure gives equal weights to all misclassified examples. Since boosting samples from a pool of data that predominantly consists of the majority class, subsequent samplings of the training set may still be skewed towards the majority class. Although boosting reduces the variance and the bias in the final ensemble, it might not hold for datasets with skewed class distributions. There is a very strong learning bias towards the majority class cases in a skewed data set, and subsequent iterations of boosting can lead to a broader sampling from the majority class. Boosting (Adaboost) treats both kinds of errors (FP and FN) in a similar fashion. Our goal is to reduce the bias inherent in the learning procedure due to the class imbalance, and increase the sampling weights for the minority class. Introducing SMOTE in each round of boosting will enable each learner to be able to sample more of the minority class cases, and also learn better and broader decision regions for the minority class. SMOTE-Boost approach outperformed boosting.

4. PERFORMANCE EVALUATION

UNDER-SAMPLING

ADVANTAGES

1. Independent on underlying classifier

2. Can be easily implemented

DISADVANTAGES

1. May remove significant patterns and cause loss of useful information

OVER-SAMPLING

ADVANTAGES

1. Independent on underlying classifier

2. Can be easily implemented

DISADVANTAGES

1. Time consuming: Introduce additional computational cost

2. May lead to over-fitting

SMOTE

ADVANTAGES

1. Alleviates over fitting caused by random oversampling as synthetic examples are generated rather than replication of instances.

2. It is simple to implement and interpret.

3. No loss of information.

DISADVANTAGES

1. SMOTE is not very practical for high dimensional data.

2. While generating synthetic examples, SMOTE does not take into consideration neighboring examples can be from other classes. This can increase the overlapping of classes and can introduce additional noise.

BAGGING

ADVANTAGES

1. Improves the stability and accuracy of machine learning algorithms used in statistical classification and regression.

2. It reduces variance and helps to avoid over fitting.

DISADVANTAGES

1. It can mildly degrade the performance of stable methods such as K-nearest neighbors.

BOOSTING

ADVANTAGES

1. Supports different loss function.

2. Works well with interactions.

DISADVANTAGES

1. Prone to over-fitting.

2. Requires careful tuning of different hyper-parameters.

ADAPTIVE SAMPLING METHODS AND SYNTHETIC DATA

GENERATION

ADVANTAGES

1. It reduces the bias introduced by the class imbalance.

2. This algorithm can be extended for integration with ensemble based learning algorithms.

3. It adaptively shifting the classification decision boundary toward the difficult examples.

COST SENSITIVE LEARNING

ADVANTAGES

1. Minimize the cost of misclassification (by biasing the classifier toward the minority class)

DISADVANTAGES

1. The misclassification costs (the actual cost of errors) often are unknown.

RECOGNITION BASED

ADVANTAGES

1. Have better performance especially on high dimensional data.

DISADVANTAGES

1. Many classifiers such as decision trees and Naive Bayes cannot be built by one class learning.

ENSEMBLE METHODS

ADVANTAGES

1. Better classification performance than individual classifiers.

2. More resilience to noise.

DISADVANTAGES

1. Time consuming

2. Over fitting

5. LITERATURE SURVEY

* A Robust Decision Tree Algorithm for Imbalanced Data Sets by( Wei Liu and Sanjay Chawla, David A. Cieslak and Nitesh V. Chawla )
* ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning by ( Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li)
* On the Classification of Imbalanced Datasets 1C.V. KrishnaVeni, 2T. Sobha Rani (ISSN : 0976-8491(Online) | ISSN : 2229-4333(Print) IJCST Vol. 2, SP 1, December 2011)
* Learning from imbalanced data: open challenges and future directions by ( Bartosz Krawczyk1 )
* DATA MINING FOR IMBALANCED DATASETS: AN OVERVIEW by Nitesh V. Chawla
* International Journal of Computing and Business Research (IJCBR) ISSN (Online) : 2229-6166 Volume 5 Issue 4 July 2014 (IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW)
* International Journal of Computer Applications (0975 – 8887) Volume 127 – No.15, October 2015 37 - A Survey on Methods for Solving Data Imbalance Problem for Classification by (Arpit Singh and Anuradha Purohit).
* A Survey on Methods to Handle Imbalance Dataset by Apurva Sonak and R.A.Patankar – at International Journal of Computer Science and Mobile Computing (ISSN 2320–088X IJCSMC, Vol. 4, Issue. 11, November 2015, pg.338 – 343)
* CLASS IMBALANCE AND ACTIVE LEARNING by (Josh Attenberg and S¸ eyda Ertekin).
* Applying Support Vector Machines to Imbalanced Datasets by (Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz).
* Class-Boundary Alignment for Imbalanced Dataset Learning by (Gang Wu and Edward Y. Chang).
* Mixture of Expert Agents for Handling Imbalanced Data Sets by (S. B. Kotsiantis and P. E. Pintelas).
* Journal of Information Engineering and Applications - ISSN 2224-5782 (print) ISSN 2225-0506 (online) Vol.3, No.10, 2013
* Journal of Network and Innovative Computing
* ISSN 2160-2174, Volume 1 (2013) pp. 332-340 - A Review of Class Imbalance Problem by (Shaza M. Abd Elrahman and Ajith Abraham).
* The Class Imbalance Problem: Significance and Strategies by (Nathalie Japkowicz).

6. CONCLUSION

Class imbalance is a hot topic being investigated recently by machine learning and data mining researchers. The researchers for solving the imbalance problem have proposed various approaches. However, there is no general approach proper for all imbalance data sets and there is no unification framework. This paper summarizes various solutions for dealing with class imbalance problems.

7. REFERENCES

[1] <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.724.1946&rep=rep1&type=pdf>

[2] ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning by (Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li).

[3] <https://pdfs.semanticscholar.org/6ee8/41531de7ebef4675c580b71241339aa4a0b3.pdf>

[4] <https://link.springer.com/content/pdf/10.1007%2Fs13748-016-0094-0.pdf>

[5] <https://www3.nd.edu/~dial/publications/chawla2005data.pdf>

[6] <http://www.researchmanuscripts.com/July2014/2.pdf>

[7] <https://pdfs.semanticscholar.org/10bd/f08bddcf6067d62152edd7c2953ccebe5c0d.pdf>

[8] <https://ijcsmc.com/docs/papers/November2015/V4I11201573.pdf>

[9] <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8642396>

[10] <http://web.mit.edu/seyda/www/Papers/IMBBook2013_AL.pdf>

[11] <http://www.site.uottawa.ca/~nat/Workshop2003/Wu-final.pdf>

[12] <https://pdfs.semanticscholar.org/fb76/4c7bdaa19550e520f7f0eeb8003c11b1b0fb.pdf>

[13]<https://s3.amazonaws.com/academia.edu.documents/32182007/Evaluation_Measures_for_Models_Assessment_over_Imbalanced_Data_Sets>

[14] <http://ias04.softcomputing.net/jnic2.pdf>

[15] <https://www.jair.org/index.php/jair/article/view/10302/24590>

[16] <http://web.mit.edu/seyda/www/Papers/SIGIR07_ALTextCateg.pdf>

[17] <https://pdfs.semanticscholar.org/907b/02c6322d0e7dff6b0201b03e3d2c6bc1d38f.pdf>