International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

Analysis of Different SMOTE based Algorithms on Imbalanced Datasets

Akangkshya Pathak

CSE Undergraduate student, Jorhat Engineering College, Assam, India

---

Abstract - *Imbalanced datasets pose a problem when we use them to make machine learning models, as oftentimes the machine learning algorithms give poor, inaccurate results on the minority class. Synthetic Minority Overlapping techniques (SMOTE) is a method to manage the problem of class imbalance in datasets, and it has become a very popular method. There are many SMOTE based algorithms. This paper seeks to compare 7 different SMOTE based algorithms on the basis of different parameters by applying it to different datasets to find out how each algorithm performs.*

Key Words: Dataset, Oversampling, SMOTE, Imbalanced Dataset

1. INTRODUCTION

A dataset can be referred to as a collection of data. A dataset is said to be imbalanced if the classes composing the dataset are not approximately equally represented. In fraud detection, imbalance of the order of 100 to 1 is prevalent and in some other applications the imbalance may be up to 100,000 to 1(1).

In her paper Japkowicz(2) discussed the effect of imbalance in a dataset. Three different strategies were used and evaluated for this. These were: resampling, under-sampling and a recognition-based induction scheme. so as to simply measure and construct concept complexity she experimented on artificial 1D data. She considered two resampling methods. The first one is random resampling, which is a method which consists of resampling the smaller class indiscriminately until it consists of as many samples as the majority class. The second method was focused resampling, which is another sampling method which consists of resampling only those minority examples that occurs on the boundary between the minority and majority classes. She also considered random under-sampling, and it involved under-sampling the bulk class samples willy-nilly until their numbers matched the quantity of minority class samples. Focused under-sampling involved under-sampling the bulk class samples lying further away. Her observations noted that both the sampling approaches were effective, which using the subtle sampling techniques didn't give any clear advantage within the domain considered.

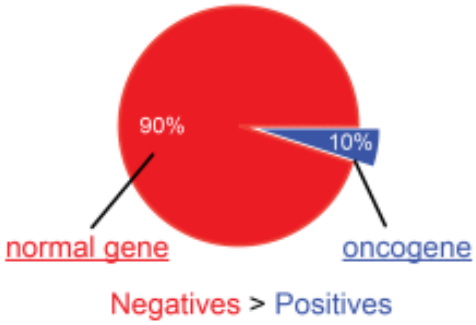


Fig -1: Imbalanced data

Imbalanced datasets pose a problem when we use machine learning algorithms on them. Predictive accuracy is typically used for evaluating the performance of machine learning algorithms. But when the data used is imbalanced and the costs of different errors vary markedly, predictive accuracy is not appropriate.

There are two ways in which the machine learning community has addressed the issue of class imbalance in datasets. Assigning distinct costs to training examples is one way. The other is to re-sample the original dataset, either by oversampling the minority class or under-sampling the majority class. Synthetic Minority Oversampling Technique is a technique proposed to solve this problem. This is a technique which blends under-sampling of the majority class with a special form of over-sampling the minority class (1).

2. SMOTE

SMOTE is an oversampling approach in which the minority class is over-sampled by creating synthetic examples rather than by over-sampling with replacement. The synthetic examples are generated by operating in feature space rather than data space, thus making it a less application specific manner.

We take each minority class sample and over-sample the minority class and then we introduce synthetic examples along the line segments which join any or all of the k minority class nearest neighbours. We randomly chose neighbours from the k nearest neighbours, depending upon the amount of over-sampling required. We generate the synthetic samples through the following steps: The first step is to take the difference between the feature sample which is under consideration and its nearest neighbor. We then multiply this difference by a random number which is between 0 and 1, and add the resulting value to the feature vector under consideration. The selection of a random point along the line segment between two specific features is caused by this. The

decision region of the minority class is effectively forced to become more general by this approach. Rather than smaller and more specific regions, the synthetic examples cause the classifier to create larger and less specific decision regions. Rather than being subsumed by the majority class samples around them, more general regions are now learned for the minority class samples (3).

The SMOTE algorithm carries out an oversampling approach in order to rebalance the original training set. The key idea of SMOTE is to introduce synthetic examples, instead of applying a simple replication of the minority class instances. This new data is generated by interpolation between several minority class instances that are within a defined neighborhood. The procedure is said to be focused on the "feature space" rather than on the "data space" for this reason, in other words, the algorithm is based on the values of the features and their relationship, in lieu of considering the data points as a whole.

In the framework of learning from imbalanced data, the Synthetic Minority Oversampling Technique is considered as an effective preprocessing standard. This is due to two reasons. The first one is its simplicity in the design of the procedure, and the second one is its robustness when applied to different type of problems. The SMOTE algorithm performs an oversampling approach, in order to rebalance the original training set. The main idea of SMOTE is to introduce synthetic examples within a defined neighborhood, rather than using simple replication of the minority class instances.

In the research community, the SMOTE preprocessing method has become a pioneer in imbalanced classification. Several extensions and options have been proposed since its launch to enhance its efficiency in various situations. Additionally, it is regarded as the most significant data preprocessing algorithm in machine learning.

3. VARIANTS OF SMOTE USED

Six different variants of SMOTE based algorithms were used to test out their performance in imbalanced datasets. These variants are:

3.1 SMOTE

SMOTE is an oversampling technique within which the synthetic samples are created for the minority class by oversampling the minority class. This algorithm helps to beat the over fitting problem that is caused by random oversampling. It focuses on the feature space to come up with new instances with the assistance of interpolation between the positive instances that lie together.

3.2 ADASYN

ADASYN refers to Adaptive Synthetic Minority Oversampling Technique and is a generalized variety of the SMOTE algorithm. This method is comparable to SMOTE but it differs in the fact that it generates different number of samples depending on an estimate of the local distribution of the category to be oversampled. This algorithm also aims to oversample the minority class by creating synthetic instances for it (4).

3.3 Borderline SMOTE

Borderline-SMOTE is another variation of the SMOTE algorithm. Borderline-SMOTE only generates synthetic data along the decision boundary between the two classes, which is unlike the SMOTE algorithm because in SMOTE the synthetic data are created randomly between the two data. If we know that the wrong classification happens near the boundary decision, then the best algorithm to use in this case is Borderline-SMOTE (5).

3.4 Safe Level SMOTE

Safe Level SMOTE is a variation of the Synthetic Minority Oversampling Technique. In this technique, each positive instance is assigned its safe level and the synthetic instances are generated after that. In order to ensure that all synthetic instances are generated only in safe regions, each synthetic instance is positioned closer to the largest safe level. If the safe level of an instance is close to 0, then the instance is almost entirely noise. The instance is considered safe if it is close to k. The safe level ratio is used for selecting the safe positions to generate synthetic instances (6).

3.5 Cluster SMOTE

Cluster SMOTE is another variation of SMOTE in which K-means algorithm is used. In this technique, the k-means algorithm is used to cluster minority samples, which is followed by finding the minority sub clusters and then the SMOTE algorithm is applied. The optimal size of the sub clusters is not determined by this algorithm, and similarly, the sample size generated by each sub-cluster is also not calculated by it.

3.6 SN SMOTE

SN SMOTE is a variant of SMOTE in which the SMOTE algorithm is extended by using a different formula for neighborhood which is called as surrounding neighborhood. One of the key features of this neighborhood type is the fact that the neighbors of a sample are considered in terms of both spatial distribution and proximity with respect to the sample. This shows some practical advantages over the conventional neighborhood, which is only based on the minimum distance. The use of the surrounding neighborhood for over-sampling the minority class generates new synthetic examples which will be homogeneously distributed throughout the original positive instances, contributing to spread the influence region of the minority class (7).

3.7 SVM Smote

SVM-SMOTE is a variant of SMOTE algorithm in which an SVM algorithm is used to detect sample which is to be used for generating new synthetic samples. In this algorithm, SVM classifiers are trained on the original training set after which

the borderline area is approximated by the support vectors. The synthetic data is created in a random manner along the lines that join each minority class support vector with a number of its nearest neighbours. In this algorithm, more data is synthesized away from the region of class overlap. This algorithm focuses more on where the data is separated.

4. PARAMETERS USED TO MEASURE ACCURACY

Four different parameters have been used to measure the efficiency of the different algorithms on a given dataset. They are:-

4.1 Accuracy Score

In classification problems, accuracy can be defined as the number of correct predictions made by the model compared to the total number of predictions made. It can be calculated by taking the correct predictions in the numerator, which consists of true positives and true negatives, and taking the total different predictions made in the denominator, which consists of correct as well as incorrect predictions, and then calculating their ratio. The ratio gives us the accuracy.

4.2 Precision Score

Precision is a measure, which is about being precise. In diagnostic binary classification, it is also known as positive predictive value. From the total data that we diagnosed as being correct, precision is a measure which tells us what portion of that data was actually correct. It is a good measure to determine the accuracy of an algorithm in the case when the cost of False Positive is high.

4.3 Recall Score

Recall is a measure which is used to calculate the proportion of data which was actually correct, that the algorithm diagnosed as being correct. It calculates the total number of Actual Positives that our model captures via labelling it as Positive. And because of this, recall is the model metric that we use to select our best model when there is a high cost associated with False Negative. In diagnostic binary classification, recall is also known as sensitivity.

4.4 F1 Score

F1 score is a measure to find out the accuracy of a test. It is calculated by using two other measures of accuracy, which are precision and recall of a test. The harmonic mean of the precision and recall gives us the F1 score. If the F1 score has the highest value which is 1.0, it indicates a perfect precision and recall. On the other hand, if the F1 score has the lowest value which is 0, it indicates that either the precision or the recall is zero.

5. DATASETS USED

Two different datasets were used to measure the accuracy parameters. These datasets have different imbalance ratios. They are:

5.1 Ecoli dataset

The features of this dataset are:-

Table -1: Ecoli dataset

General Information			
Type	Imbalanced	Origin	Real world
Features	7	(Real / Integer / Nominal)	(7 / 0 / 0)
Instances	336	IR	8.6
% Positive instances	10.42	% Negative instances	89.58
Missing values?	No		

5.2 PageBlock dataset

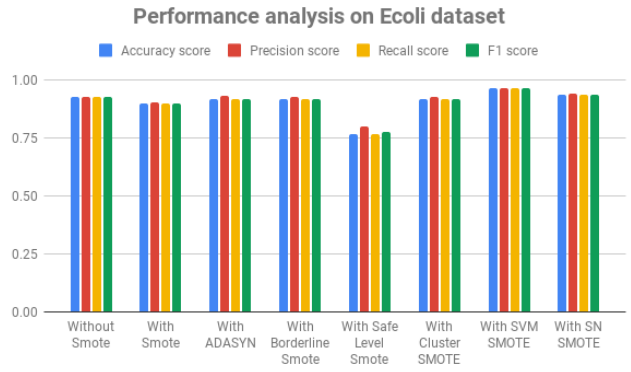
The features of this dataset are:-

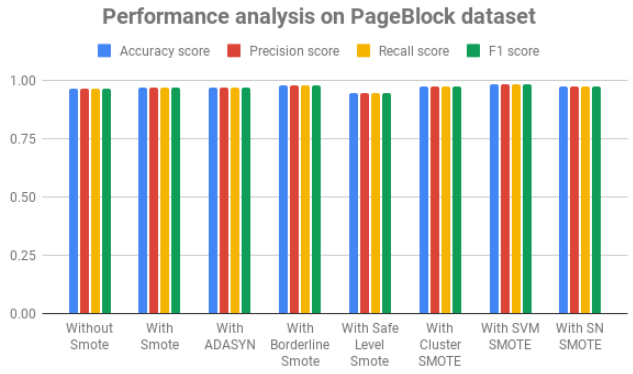
Table -1: PageBlock dataset

General Information			
Type	Imbalanced	Origin	Real world
Features	10	(Real / Integer / Nominal)	(4/ 6 / 0)
Instances	5472	IR	8.79
% Positive instances	10.21	% Negative instances	89.79
Missing values?	No		

6. RESULTS

On applying the different SMOTE based algorithms on imbalanced dataset we got different precision scores for each algorithm. These results have been summarized in the graph below:-





From these results we can see that different algorithms give different performances when used in a particular dataset. In the Ecoli dataset, we can clearly demarcate the difference in the performance of the different algorithms used and we can see that SVM Smote gives the best performance. But in the PageBlock Dataset we see that the algorithms give comparable levels of performance and the best performing algorithm does not show a very distinct leap in performance from the other algorithms.

We can also see a distinct difference in the performance of an algorithm, based on the different parameters used to measure the accuracy.

7. CONCLUSION

SMOTE based algorithms have become very useful when creating highly accurate machine learning models that make use of imbalanced datasets. On testing these algorithms on different datasets, we have come to the conclusion that different algorithms give different levels of accuracy when used on datasets having different class imbalance ratios. It can be concluded that all of these algorithms are very essential in order to get good accuracy scores, though different models give the best precision score depending on the imbalance ratio of the datasets used.

REFERENCES

[1] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research 16 (2002) 321–357

[2] Nathalie Japkowicz, "Learning from Imbalanced Data Sets: A Comparison of Various Strategies", AAAI Technical Report WS-00-05. 2000, AAAI.

[3] Alberto Fern´andez, Salvador Garc´ıa, Francisco Herrera and Nitesh V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary", Journal of Artificial Intelligence Research 61 (2018) 863-905.

[4] Hazel A. Gameng, Bobby B. Gerardo and Ruji P. Medina, "Modified Adaptive Synthetic SMOTE to Improve Classification Performance in Imbalanced Datasets," 2019 6th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS).

[5] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part I, LNCS 3644, pp. 878 – 887, 2005. Springer-Verlag Berlin Heidelberg 2005.

[6] Chumphol Bunkhumpornpat, Krung Sinapiromsaran and Chidchanok Lursinsap, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem", T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 475–482, 2009. © Springer-Verlag Berlin Heidelberg 2009.

[7] V. García, J. S. Sánchez, R. Martín-Félez and R. A. Mollineda, "Surrounding neighborhood-based SMOTE for learning from imbalanced data sets", Springer-Verlag Berlin Heidelberg 2012.