SMOTE: Synthetic Minority Over-sampling Technique

Siddhant Prakash, *1211092724* and Kunal Bansal, *1211XXXXXX*

Abstract—The abstract goes here.

Index Terms—Computer Society, IEEE, IEEEtran, journal, LaTeX, paper, template.

1 Introduction

THE seminal work, "SMOTE: synthetic minority over-sampling technique" [1] deals with the fundamental problem of imbalanced class data-sets in machine learning, where, data points of one class are in majority over data points of another class. A number of times the data points belonging to the minority class are more important than the majority class. This majority-minority skew in the data-set leads to classification in favor of majority class samples when one uses standard classification techniques like Naive Bayes or Decision Trees. In several cases, the penalty for mis-classifying these minority classes are much higher than mis-classifying the majority class. For example, the problem of classifying the images of mammograms for cancerous cell detection is very sensitive and mis-classification may lead to disastrous outcomes. Still, the number of positive class samples is way outnumbered by the number of negative class samples with the majority (negative) class sample constituting 98% of total samples.

The authors in [1] propose a new algorithm to augment the minority samples in a data-set by creating synthetic data points for minority class to even the data distribution between majority and minority classes. They compare the result obtained by this new technique with the results of plain under-sampling the majority samples, as well as over-sampling the minority samples as previously done in other works dealing with the problem. The results show that their method leads to learning of better classifiers. Further, they show that the classifiers improve with near equal representation from all class in training data. The authors use Receiver Operating Characteristics (ROC) curves as their performance measure. ROC curves provide trade-off between true positive (TP) vs. false positive (FP) which is much more suited to the class imbalance problem than the error rate (accuracy) metric. Area Under the Curve (AUC) and convex hull of ROC curve are used for the experiments as they provide good comparison of the classifiers' performance in class imbalance scenarios.

In our project, we learn to deal with the challenges of class imbalanced data-sets and how to overcome them. We replicate the work done in [1] and understand the problem by going over the previous works cited in this paper. Additionally, we summarize the various techniques researchers came up with regards to dealing with this issue. We overview the performance metrics used in this paper and explain why the authors chose to use these metrics, as well as how are they relevant to this particular problem. We then move on to the implementation of the algorithm SMOTE [1] proposed in the paper, along with two more approaches which was used to emphasize the novelty of this technique. We replicate the experiments section on 3 of the 9 datasets listed in the paper, using Decision Trees, Nearest Neighbours and Nave Bayes classifiers to learn classification models. We also use the prediction results from these models to plot ROC curves, and taking AUC and convex hull of the curve as evaluation metric, compare our graphs with the graphs obtained in the paper.

• *S. Prakash and K. Bansal are with School of Computing, Informatics and Design Systems Engineering, Arizona State University, Arizona, AZ, 85281.*
E-mail: {sprakas9, kbansal3}@asu.edu

2 Previous Works

Most of the cases of imbalanced dataset is dealt in two ways, viz. under-sampling the majority class samples or over-sampling the minority class samples. Different domain requires different techniques for the same based on their requirement. When it comes to under-sampling the majority class, Kubat et al (1997, 1998) experimented with the same. In Kubat and Matwin (1997), the majority class is selectively under sampled while the minority class sampling remains fixed. The performance metric used for the classifier is geometric mean which is not as expressive as a ROC curve and corresponds to just one point on it. Related to the above, the SHRINK system of Kubat et al (1998) classified the overlapping reasons of both majority and minority classes as positive, leading to a "best positive region" classification.

Another study on under-sampling of dataset was performed by Japkowicz (2000). In her study, she explored different sampling techniques on artificial 1D data for better evaluation of concept complexity. Her exploration involved under-sampling as well as resampling of data. Both strategies involved two different methods, viz. random and focused. Random resampling used samples from minor class to be sampled randomly until they matched major class samples, while focused resampling used only the boundary

points between minor and major class. In random under-sampling, the samples from majority class were removed randomly to match the minority class samples, in contrast to focused under-sampling which under sampled majority class samples lying further away. Her study revealed the efficacy of both the sampling techniques but did not provide any clear advantage in the domain considered.

While under sampling approach works, other works uses under-sampling of majority class samples along with over-sampling of minority class samples for learning a better classifier. Ling and Li et al (1998) uses lift analysis to measure classifier's performance in the domain of marketing analysis problem. They ranked the test examples by confidence measures and used lift as the evaluation criteria. In one of the experiments they performed, they under sampled the majority class and observed that the best lift index is obtained when there is equal representation of the classes. In another experiment, they over-sampled the minority samples with replacement to match the negative samples but could not prove the same as significant. The work present in this paper is similar in strategy, but the over-sampling techniques is different. Another work which uses the idea of under-sampling as well as over-sampling of data to overcome class imbalance problem is Solberg and Solberg et al (1996). They use SAR imagery dataset obtained for classification of oil slicks which is heavily biased towards look-alike data compared to oil slicks (98%-2%). They created a new dataset by over-sampling the oil slicks data randomly and under-sampling the look-alike data to create equal class distribution. As a result, on learning a classification tree on the balanced dataset they obtained better error rates on both classes compared to training on imbalanced dataset. Domingos et al (1999) also take the same approach to deal with class imbalance by introducing a "metacost" term to under-sampling as well as over-sampling. The work shows the metacost improves over either, and proves that under-sampling of majority class does better than over-sampling of minority class.

Other researchers (DeRouin et al 1991) tried to use the same on feed-forward neural networks which is not able to learn to discriminate between classes sufficiently due to the same class imbalance problem. The learning rate of the neural network was adapted according to the class distribution in the data set. Experimenting over artificial as well as real-world training data with multi-class problem provided better classification accuracy for minority class. In information retrieval domain, document classification is one of the challenging problems which is affected by this class imbalance. Creating a simple bag-of-words model results in interesting words samples as a minority due to very limited instances of such words in the document. Thus, in IR domain, the performance metric is replaced from error rates and instead, precision and recall terms are used for performance measurements.

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

We will see what we mean by the terms True Positive (TP), False Positive (FP) and False Negative (FN) in the next section. In the same domain, Mladenic and Grobelnik (1999) proposed a feature subset selection approach to deal with class imbalance. They found out that using odd ratio along with Naive Bayes classifier performs best in the domain. Odds ratio incorporates target class information giving better result over information gain which is computed per word for each class. In 1997, Provost and Fawcett introduced the ROC convex hull method for performance evaluation of the classifier in which ROC space is used to separate classification performance from class and cost distribution.

3 Evaluation Criteria and Performance Metrics

Confusion matrix, as shown in Table 1, is one of the most common method in machine learning used to evaluate performance of a (2-class) classification problem.

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

TABLE 1
Confusion Matrix

As shown in the table, the columns are predicted class and the rows are actual class. Over a dataset of finite samples, the count of correctly classified negative samples is termed as True Negative (TN), while the count of incorrectly classified negative samples is termed as False Positive (FP). Similarly, the count of incorrectly classified positive sample is termed as False Negative (FN), while the count of correctly classified positive sample is termed as True Positive (TP).

For any classification task, predictive accuracy is defined as the total number of correctly classified samples over total number of samples. Mathematically, it is given by,

$$Accuracy = \frac{TP + TN}{TN + FP + FN + TP}$$

In machine learning, we evaluate the performance of a classifier by its error rate which is given by,

$$ErrorRate = 1 - Accuracy$$

This performance measure works well for balanced class data sets, but for imbalanced datasets a much wider used metric is the Receiver Operating Characteristics (ROC) curves.

A typically ROC curve is a plot of percentage True Positive vs percentage False Positive. One such curve is shown in Figure **XX**. We have %ge FP on the X-axis given by,

$$\%ge\ FP = \frac{FP}{TN + FP}$$

and %ge TP on the Y-axis given by,

$$\%ge\ TP = \frac{TP}{TP + FN}$$

As evident from the definition, the ideal point on the curve will be (0,100), signifying the all positive examples are classified correctly while no negative samples are classified wrongly as positive. The Area Under the Curve (AUC) can

be taken as a good metric for comparing different classifiers, but these can be suboptimal for some specific cost and class distributions. Thus, the convex hull of ROC curve, being potentially optimal, is also taken as one of the performance metrics.

4 Implementation

In this section we provide the details of the various approaches we implemented from the paper. Using the "mammography" dataset [3], we show the effectiveness of SMOTE in this class imbalanced scenario. The dataset consist of 11,183 samples of which 10,923 are negative samples while we have only 260 positive samples. The dataset has 6 attributes, while we learn the classifier on 2 attributes, by decomposing the dataset using PCA, for better visualization. In Figure 1, we show the decision boundaries in the feature space, created by classifying 10% of the total data containing **18** positive samples and **1,100** negative samples. We have used the classifiers implemented in Scikit Learn library [2] with codes written in Python 2.7.1. The codes and dataset are provided in the supplementary material.

4.1 Over-sampling with replacement

Learning a classifier on the imabalanced class data gives us a biased classification towards the majority class samples (red circles). From Figure 1a we observe that the classification decision boundary using the original data is not very accurate. Many minority class samples (blue circles) are classified under majority class (red region) resulting in decrease in true positive rate. The decision boundary is more general and spread away from minority (positive) class samples too.

One of the earliest ways suggested by researchers to overcome this issue is over-sampling the minority class samples, in order to reduce the bias towards the majority class. We implement the over-sampling approach by duplicating the minority class samples. The degree of over-sampling in our implementation is 500%, i.e. if the dataset had 18 positive class samples and 1,100 negative class samples, as is the case in Figure 1, we upsample the positive class sample to 90. As a result, we see the decision boundary for minority class samples (blue region) shrink to enclose each individual positive sample. In other words, the data starts overfitting towards the positive class samples as more and more leaf nodes are added to the decision tree leading to very specific decision boundaries in the feature space. We see the same happening in Figure 1b with the 6 positive samples in the region around (0, 5).

4.2 SMOTE

4.3 SMOTE with under-sampling

5 Experiments

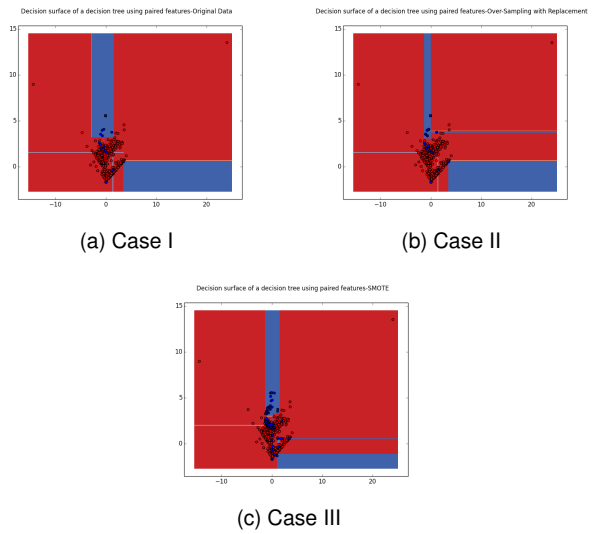
5.1 Datasets

5.2 Classifiers

6 Discussions

7 Conclusion

The conclusion and future scope goes here.



(a) Case I (b) Case II

(c) Case III

Fig. 1. (a) Decision boundary on raw original data gives a biased classification towards negative (red) class samples.

Acknowledgments

The authors would like to thank...

References

[1] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[2]

[3]