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

THIS paper deals with the problem of imbalanced datasets.
Often times the dataset that are used for classification have data points of "interesting" class as a minority.
This skews the classification in favor of majority class samples.
In many cases, the penalty for mis-classifying these minority classes are much higher than mis-classifying the majority "normal" class.
Examples: pictures of mammograms for cancerous cell detection (98%-2%).
The paper proposes a new algorithm to augment the datasets by creating synthetic data points for minority class to even the data distribution.
Thus, leading to creation of better classifiers with near equal representation from all class in training data.
Evaluation criteria of Receiver Operating Characteristics (ROC) provides trade-off between true positive (TP) vs. false positive (FP).
Area Under the Curve (AUC) of ROC curve is an accepted metric for classification performance.

Many previous work tries to tackle the problem of imbalanced datasets in broadly two ways.
First, to assign distinct penalty for training data and
Second, to change the dataset by either under-sampling the majority class or over-sampling the minority class.
The authors approach mixes the two and uses a unique algorithm to over-sample the minority class.
They show there performance using the AUC of ROC curve and ROC convex hull method.
They compare there classification for C4.5 Decision Tree, Ripper and Naive Bayes classifiers.

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

• *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

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

4.1 SMOTE

4.2 Others

5 Results

5.1 Datasets Used

6 Discussions

7 Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

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.