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.

• *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

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.