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.

- *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

3 Evaluation Criteria and 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.