Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

BLENDING PROPENSITY SCORE MATCHING AND SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE FOR IMBALANCED CLASSIFICATION

William A. Rivera and Amit Goel and J. Peter Kincaid

Institute for Simulation and Training
University of Central Florida
Orlando, FL, USA
14wrivera@knights.ucf.edu, amit@goelresearch.com, pkincaid@ist.ucf.edu

ABSTRACT

Real world data sets often contain disproportionate sample sizes of observed groups making the task of prediction algorithms very difficult. One of the many ways to combat inherit bias from class imbalance data is to perform re-sampling. In this paper we discuss two popular re-sampling approaches proposed in literature, Synthetic Minority Over-sampling Technique (SMOTE) and Propensity Score Matching (PSM) as well as a novel approach referred to as Over-sampling Using Propensity Scores (OUPS). Using simulation we conduct experiments that result in statistical improvement in accuracy and sensitivity by using OUPS over both SMOTE and PSM

1 INTRODUCTION

Real world data sets often contain disproportionate sample sizes of observed groups making the task of prediction algorithms very difficult. Such examples of class imbalanced data frequently occur in different domains such as fraud detection, mammography of cancerous cells and post term births (Chawla, Bowyer, and Hall 2002). Re-sampling techniques offer simple solutions to dealing with class imbalanced data. In the next sections we discuss three re-sampling approaches followed by an experiment comparing them. This paper concludes with a discussion of the results from the experiment.

2 RE-SAMPLING TECHNIQUES

Synthetic Minority Over-sampling Technique (SMOTE) combines both over-sampling and under-sampling based on user defined thresholds.The under-sampling portion removes data randomly while the over-sampling portion randomly adds samples based on the difference of K-nearest neighbors. SMOTE combined with machine learning techniques such as Principal Component Analysis (PCA), Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) have been used to effectively increase predictive accuracy (Blagus and Lusa 2013, Farquad and Bose 2012, Xie and Qiu 2007, Tian, Gu, and Liu 2011, Naseriparsa and Kashani 2013).

Propensity score matching (PSM) is heavily used in the medical field to effectively compare observations between a control group and treatment group (Austin 2011). Reducing observations to similar pairs reduces bias allowing both groups to be equally represented and analyzed using statistical measures of significance to assess improvement for the treatment group. The propensity score is the conditional probability of group membership to the treatment group (usually the minority group) versus the control group (usually the majority group) based on its covariates: $e(x) = P(G = 1|x)$.

Over-sampling Using Propensity Scores (OUPS) blends both SMOTE and PSM by using the same approach that SMOTE uses for creating synthetic samples while using the propensity score for the match criteria (Rivera, Goel, and Kincaid 2014). New cases are created based on the over-sampling amount

needed to closely match the majority group amount. if an over-sampling rate of 400% is needed then 4 new cases per observation are created based on the propensity score match for each observation in the minority group.

3 EXPERIMENTAL DESCRIPTION

We extended previous work comparing the three approaches to include more iterative splitting samples and performing the Wilcoxon rank sum test to validate the statistical significance of improvement. The data was comprised of four data sets with varying degrees of imbalance and features from different industries with four different machine learning algorithms to measure the effectiveness of each sampling approach. A confusion matrix was then used to produce the following metrics: accuracy, sensitivity, specificity and precision.

4 RESULTS

Table 1 shows that the OUPS approach outperformed SMOTE and PSM for accuracy and sensitivity with statistical significance at the 0.05 level. SMOTE outperformed all other sampling techniques in specificity and in precision with statistical significant at the 0.05 level. OUPS produced fewer false negatives thus removing some of the bias caused by the imbalance and proving to be an effective approach for increasing accuracy and sensitivity rates. False negatives represent misclassification of a minority group example as belonging to the majority group. Thus OUPS did a better job removing some of the bias caused by the imbalance by approximately 5 - 20% when compared to the other sampling approaches.

Table 1: Results by Sample Technique

Sampling Technique	Performance Measure (n=1263)			
	Accuracy	Sensitivity	Specificity	Precision
SMOTE	76.77%	48.53%	**93.59% ***	**26.61% ***
PSM	60.93%	36.93%	83.33%	24.43%
OUPS	**79.82% ***	**53.41% ***	91.75%	23.42%
Note: * indicates a statistically significant improvement at 0.05 alpha over the entire group in that column				

REFERENCES

Austin, P. C. 2011, May. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". *Multivariate Behavioral Research* 46 (3): 399–424.

Blagus, R., and L. Lusa. 2013. "Open Access SMOTE for high-dimensional class-imbalanced data". *BMC Bioinformatics* 14 (106): 1–16.

Chawla, N. V., K. W. Bowyer, and L. O. Hall. 2002. "SMOTE : Synthetic Minority Over-sampling Technique". *Journal of Artificial Intelligence Research* 16:321–357.

Farquad, M. A. H., and I. Bose. 2012, April. "Preprocessing unbalanced data using support vector machine". *Decision Support Systems* 53 (1): 226–233.

Naseriparsa, M., and M. M. R. Kashani. 2013. "Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset". *International Journal of Computer Applications* 77 (3): 33–38.

Rivera, W. A., A. Goel, and P. J. Kincaid. 2014. "Resampling, a combined approach using SMOTE and Propensity Score Matching". *Accepted for publication in ICMLA 2014*.

Tian, J., H. Gu, and W. Liu. 2011, March. "Imbalanced classification using support vector machine ensemble". *Neural Computing and Applications* 20 (2): 203–209.

Xie, J., and Z. Qiu. 2007, February. "The effect of imbalanced data sets on LDA: A theoretical and empirical analysis". *Pattern Recognition* 40 (2): 557–562.