Chapter 40
Advances in Algorithms for Re-Sampling Class-Imbalanced Educational Data Sets

William Rivera
Institute for Simulation and Training, University of Central Florida, USA

Amit Goel
Institute for Simulation and Training, University of Central Florida, USA

J Peter Kincaid
Institute for Simulation and Training, University of Central Florida, USA

ABSTRACT

Real world data sets often contain disproportionate sample sizes of observed groups making it difficult for predictive analytics algorithms. One of the many ways to combat inherent bias from class imbalance data is to perform re-sampling. In this book chapter we discuss popular re-sampling methods proposed in research literature, such as Synthetic Minority Over-sampling Technique (SMOTE) and Propensity Score Matching (PSM). We provide an insight into recent advances and our own novel algorithms under the umbrella term of Over-sampling Using Propensity Scores (OUPS). Using simulation we conduct experiments that result in statistical improvement in accuracy and sensitivity by using these new algorithmic approaches.

INTRODUCTION

With any real world data there is often difficulty in creating prediction models that are highly accurate. In classification of outcomes there is typically a large disparity between the amount of observations collected from equally represented groups or classes. This makes the task of accurately predicting group membership on new data difficult. The problem of disparity between groups is called class imbalance.

Class imbalance is a common property of real world data sets but the issue with class imbalance is that the classifier tends to classify new observations as belonging to the over represented group or

DOI: 10.4018/978-1-5225-1759-7.ch040

majority group because of the inherit bias. The problem is intensified with larger levels of imbalance most commonly found in observational studies. Extreme cases of class imbalance are commonly found in fraud detection, mammography of cancerous cells and post term births. Reported cases of imbalance have been as extreme as 100,000 to (Chawla, 2005; D'Agostino, 1998; Mendes-Moreira & Soares, 2012; Tian, Gu, & Liu, 2010).

Another inherent problem in class imbalance classification is that the classifier will usually contain high prediction accuracy because the underrepresented group is so small thus nullifying the misclassification cost of those observations since the impact is not noticeable. In most cases the target of interest is prediction of the underrepresented group which results in poor predictability.

The first major study to evaluate class imbalance was conducted in 2000. Japkowicz performed experiments on 125 randomly (using uniform distribution) synthesized data sets with varying degrees in complexity, training set size and imbalance in order to search for factors that impact class imbalance data. Using multilayer perceptron networks they identified that domains that contained linearly separable data sets did not suffer misclassification from imbalance. Second, the degree of complexity increases with the level of imbalance and lastly that the error rate is subject to the proportion of imbalance.

Further studies followed suit in highlighting additional reasons why classifiers perform poorly. These include inappropriate metrics for highly class imbalanced data, lack of generalization of classification rules for minority examples and the view minority examples as noise. Data intrinsic properties that perpetuate the class imbalance problem include the degree of class imbalance, complexity of the target concept and the classifier involved (Fernández, García, & Herrera, 2011; X. Guo, Yin, Dong, Yang, & Zhou, 2008; Japkowicz, 2000; López, Fernández, García, Palade, & Herrera, 2013; R. Prati, Batista, & Monard, 2004; R. C. Prati, Batista, & Monard, 2004; Weiss, 2010a, 2010b). The next few sections provide a further overview of these characteristics.

Disjunction

In typical data the minority group will be represented as small disjuncts overwhelmingly surrounded by majority cases. Disjuncts represent clusters spread throughout the data. The size of a disjunct are represented by the amount of observations that it correctly classifies and small disjuncts represent a small region where only a few training examples predict correctly. Small disjuncts have been empirically shown to have higher errors rates compared to large disjuncts which also tend to contribute significantly to the total test error (Weiss, 2010a).

Small disjuncts may appear as noise and most classifiers use induction to represent small disjuncts as a rule. In general there is a lack of information to provide for good generalizations. For instance a classification tree will typically represent each disjunct as a leaf or a given decision at a certain path. The smaller the disjunct the more error prone the classier tends to be (López et al., 2013; Quinlan, 1991; Weiss, 2010a, 2010b). This inherit problem has been of much study and part of what's been considered the data difficulty factor in dealing with imbalanced data sets (Stefanowski, 2014).

Weiss argues that small disjuncts produce higher error rates compared to large disjuncts which is strongly associated with class imbalance (Weiss, 2010a). However the true relationship between them remain uncertain and most research related to disjuncts have typically used decision tree classification algorithms with pruning to provide broader generalization coverage. These algorithms remain highly subject to this problem while other classification algorithms are considered less prone to the issue.