Chapter 5
Learning From Class Imbalance: A Fireworks-Based Resampling for Weighted Pattern Matching Classifier (PMC+)

Sreeja N. K.
PSG College of Technology, India

ABSTRACT

Learning a classifier from imbalanced data is one of the most challenging research problems. Data imbalance occurs when the number of instances belonging to one class is much less than the number of instances belonging to the other class. A standard classifier is biased towards the majority class and therefore misclassifies the minority class instances. Minority class instances may be regarded as rare events or unusual patterns that could potentially have a negative impact on the society. Therefore, detection of such events is considered significant. This chapter proposes a FireWorks-based Hybrid ReSampling (FWHRS) algorithm to resample imbalance data. It is used with Weighted Pattern Matching based classifier (PMC+) for classification. FWHRS-PMC+ was evaluated on 44 imbalanced binary datasets. Experiments reveal FWHRS-PMC+ is effective in classification of imbalanced data. Empirical results were validated using non-parametric statistical tests.

INTRODUCTION

Despite intense research in the field of Machine learning, learning from imbalanced data still remains a challenging problem. Class imbalance is a phenomenon in which the number of instances belonging to one class (majority class) is much more than the instances belonging to the other class (minority class). Minority class instances may be regarded as rare events or unusual patterns in daily life which are difficult to detect. Although rare, they are considered important and require immediate response.

Developments in learning from imbalanced data have been motivated by numerous real-life applications involving such rare events. Data for forecasting natural disasters (Hong et al.,2013), fraudulent credit card transactions (Panigrahi et al., 2009), target identification from satellite radar images (Kubat et

DOI: 10.4018/978-1-7998-1659-1.ch005

al., 1998), classifying biological anomalies (Choe et al.,1998), computer-assisted medical diagnosis and treatment (Mazurowski, et al., 2008) exhibit class imbalance. Conventional classifiers are designed to work with balanced class distributions and therefore they are biased towards the majority class distribution. Therefore, conventional classifiers do not predict rare events. The extent of imbalance in a dataset is determined by the imbalance ratio. It is defined as the ratio of the number of instances belonging to the majority class to the number of instances belonging to the minority class.

The problem of class imbalance may be addressed in three ways: (i) Data level approaches that aim to rebalance the class distributions using over-sampling, under-sampling or hybrid techniques, (ii) Cost-sensitive approaches that introduces penalty for every misclassification and (iii) Ensemble approaches that model multiple classifiers for better classification.

Resampling techniques are data level approaches that aim at restoration of the balance in the dataset. Cost sensitive approaches considers different numeric costs for different misclassification types during model building. However, this method requires a cost matrix to be defined. Ensemble classifiers combine several classifiers thereby improving the classifier performance. Ensemble classifiers are time consuming as they build several base classifiers. Among these, the most common and an effective technique to tackle class imbalance is to rebalance the data using data level approaches. Many times, resampling methods are combined with a base classifier for better performance (Alberto et al., 2013).

Classifiers like Weighted Data Gravitation based classification (DGC+), Weighted Pattern Matching based Classification (PMC+) and Imbalanced DGC (IDGC) were also proposed for classification of imbalanced data. These classifiers do not use resampling or cost sensitive method for tackling the class imbalance.

This chapter proposes a FireWorks-based Hybrid ReSampling (FWHRS) algorithm to restore the balance in the dataset. The proposal is used with Weighted Pattern Matching based classifier (PMC+) for classification. Experiments have been performed on 44 imbalanced binary datasets collected from the KEEL repository (Alcalá-Fdez et al., 2011). The experiments consider various problem domains, number of instances and imbalanced ratio. Experiments indicate the competitive nature of the proposed FWHRS-PMC+ algorithm obtaining significantly better results in terms of Cohen's kappa rate and Area Under the Curve (AUC). Statistical analysis like Iman and Davenport test and Bonferroni-Dunn post hoc test was performed to evaluate whether there are significant differences in the results of the classifiers.

RELATED WORK

Class imbalance is a common scenario encountered in the field of machine learning and pattern recognition. Methods that handle class imbalance are categorized into Data level approaches, Cost sensitive approaches and ensemble approaches. Resampling is a data level approach that artificially inflates the minority class instances or reduces the number of majority class instances. Cost sensitive approaches introduces penalty for each misclassification. Ensemble classifiers constructs several classifiers and combine them to obtain a new classifier that outperforms the individual ones.

Data Level Approaches: Resampling

Data level approaches preprocesses the original imbalanced dataset to balance the class distribution. Thus, the base classifier need not be modified. It is empirically proved that these methods are a use-