Chapter 3
A Measure Optimized Cost-Sensitive Learning Framework for Imbalanced Data Classification

Peng Cao
Northeastern University, China & University of Alberta, Canada

Osmar Zaiane
University of Alberta, Canada

Dazhe Zhao
Northeastern University, China

ABSTRACT

Class imbalance is one of the challenging problems for machine-learning in many real-world applications. Many methods have been proposed to address and attempt to solve the problem, including sampling and cost-sensitive learning. The latter has attracted significant attention in recent years to solve the problem, but it is difficult to determine the precise misclassification costs in practice. There are also other factors that influence the performance of the classification including the input feature subset and the intrinsic parameters of the classifier. This chapter presents an effective wrapper framework incorporating the evaluation measure (AUC and G-mean) into the objective function of cost sensitive learning directly to improve the performance of classification by simultaneously optimizing the best pair of feature subset, intrinsic parameters, and misclassification cost parameter. The optimization is based on Particle Swarm Optimization (PSO). The authors use two different common methods, support vector machine and feed forward neural networks, to evaluate the proposed framework. Experimental results on various standard benchmark datasets with different ratios of imbalance and a real-world problem show that the proposed method is effective in comparison with commonly used sampling techniques.

DOI: 10.4018/978-1-4666-6078-6.ch003

INTRODUCTION

Recently, the class imbalance problem has been recognized as a crucial problem in machine learning and data mining (Chawla, Japkowicz &Kolcz, 2004; Kotsiantis, Kanellopoulos & Pintelas, 2006; He &Garcia, 2009; He & Ma, 2013). This issue of imbalanced data occurs when the training data is not evenly distributed among classes. This problem is also especially critical in many real applications, such as credit card fraud detection when fraudulent cases are rare or medical diagnoses where normal cases are the majority, and it is growing in importance and has been identified as one of the 10 main challenges of data mining (Yang, 2006). In these cases, standard classifiers generally perform poorly. classifiers usually tend to be overwhelmed by the majority class and ignore the minority class examples. Most classifiers assume an even distribution of examples among classes and assume an equal misclassification cost. Moreover, classifiers are typically designed to maximize accuracy, which is not a good metric to evaluate effectiveness in the case of imbalanced training data. Therefore, we need to improve traditional algorithms so as to handle imbalanced data and choose other metrics to measure performance instead of accuracy. We focus our study on imbalanced datasets with binary classes.

Much work has been done in addressing the class imbalance problem. These methods can be grouped in two categories: the data perspective and the algorithm perspective (He &Garcia 2009). The methods with the data perspective re-balance the class distribution by re-sampling the data space either randomly or deterministically (Chawla, Bowyer, Hall & Kegelmeyer, 2002; Chawla, Lazarevic, Hall & Bowyer, 2003; Chawla, Cieslak, Hall & Joshi, 2008; Barua, Monirul Islam, Yao & Murase, 2013; Galar, Fernández, Barrenechea & Herrera, 2013). The main disadvantage of re-sampling techniques are that they may cause loss of important information or the model overfitting, since that they change the original data distribution. In addition, the performance of sampling can vary significantly depending upon the data available.

Cost-sensitive learning is one of the most important topics in machine learning and data mining, and attracted high attention in recent years (Akbani, Kwek & Japkowicz, 2004; Ling & Sheng, 2008; Zhou & Liu, 2006). Cost-sensitive learning methods consider the costs associated with misclassifying examples, and try to learn more characteristics of samples with the minority class by setting a high cost to the misclassification of a minority class sample. It has been shown that the problem of learning from imbalanced datasets and the problem of learning when costs are unequal and unknown can be handled in the same manner even though these problems are not exactly the same (Maloof, 2003). Cost-sensitive learning does not modify the data distribution, and is generally more consistent in terms of performance than the sampling techniques (Chris, Taghi, Jason & Amri, 2008; Weiss, McCarthy & Zabar, 2007).

There are two challenges with respect to the training of cost sensitive classifier. The misclassification costs play a crucial role in the construction of a cost sensitive learning model for achieving expected classification results. However, in many contexts of imbalanced dataset, the misclassification costs cannot be determined. Beside the cost, the feature set and intrinsic parameters of some sophisticated classifiers also influence the classification performance. The imbalanced data distribution is often accompanied by high dimensionality in real-world data sets such as text classification and bioinformatics (Blagus, 2013; Van Hulse, Khoshgoftaar, Napolitano & Wald, 2009; Zheng, Wu & Srihari, 2004). Therefore, high-dimensionality poses additional challenges when dealing with class-imbalanced prediction. Optimal feature selection can concurrently achieve good accuracy and dimensionality reduction. The proper intrinsic parameter setting of classifiers, such as regularization cost parameter and