To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling Techniques

Lida Abdi and Sattar Hashemi

Abstract—Class imbalance problem is quite pervasive in our nowadays human practice. This problem basically refers to the skewness in the data underlying distribution which, in turn, imposes many difficulties on typical machine learning algorithms. To deal with the emerging issues arising from multi-class skewed distributions, existing efforts are mainly divided into two categories: model-oriented solutions and data-oriented techniques. Focusing on the latter, this paper presents a new over-sampling technique which is inspired by Mahalanobis distance. The presented over-sampling technique, called MDO (Mahalanobis Distance-based Over-sampling technique), generates synthetic samples which have the same Mahalanobis distance from the considered class mean as other minority class examples. By preserving the covariance structure of the minority class instances and intelligently generating synthetic samples along the probability contours, new minority class instances are modelled better for learning algorithms. Moreover, MDO can reduce the risk of overlapping between different class regions which are considered as a serious challenge in multi-class problems. Our theoretical analyses and empirical observations across wide spectrum multi-class imbalanced benchmarks indicate that MDO is the method of choice by offering statistical superior MAUC and precision compared to the popular over-sampling techniques.

Index Terms—Multi-class imbalance problems, over-sampling techniques, Mahalanobis distance

1 INTRODUCTION

In recent years, with the accelerated developments in science and technology and availability of data, there is a need for more robust and accurate learning algorithms. Existence of imbalanced distributions among these data is very prevalent. In fact, a data set with unequal number of instances for different classes is called imbalanced data set. This skewness in the data underlying distribution causes many problems for typical machine learning algorithms. In particular, correctly classifying the minority class instances is a main issue in processing these data sets. Simply said, the key point of learning is to obtain a classifier which will provide high accuracy for the minority class without severely jeopardizing the accuracy of the majority class [31].

In many real world applications such as weld flaw [37] and protein fold [56] classifications, there is a presence of multi-class imbalanced data sets. These problems impose many new issues and challenges which have not been seen in two-class ones. Zhou and Liu [57] showed that cost sensitive learning with multi-class tasks is more difficult than two-class ones and a higher degree of class imbalance may increase the difficulty. They also revealed that almost all techniques are effective on two-class problems, while most are ineffective and even may cause negative effects on multi-class tasks.

Existing solutions in dealing with class imbalanced problems are at the data level and algorithmic level. Data level solutions are pre-process tasks which are applied to balance the skewed distributions directly. These solutions which can be used simply, are divided into over-sampling and under-sampling techniques. In over-sampling, the number of instances in minority classes increases to reach a desired level of balance. These synthetic samples, which are added to the original data set, may cause the algorithm to over-fit or over-generalize. On the other hand, under-sampling solutions eliminate some of the majority class instances, and in this way they can help ease the learning process. But these methods which remove some instances of the majority classes may cause lack of useful information and mislead the algorithm.

Algorithmic or model-based solutions such as cost sensitive methods, ensemble learning algorithms, and one-class learning (also known as novelty detection and recognition-based methodology) [8] are among the proposed ways of dealing with these problems. Cost sensitive methods consider higher misclassification costs for rare examples; however, in many cases these costs are not available. Ensemble learning algorithms showed great success in various learning problems. Bagging [5], Boosting [22], SMOTEBoost [9], and RareBoost [33] are popular in literature. One-class learning techniques address class imbalance problem by modifying the training mechanism with the more direct goal of better accuracy on the minority classes. Instead of differentiating examples of one class from the others, these methods learn a model by using mainly or only a single class of examples.

Many typical machine learning algorithms pose many difficulties dealing with uneven data distributions. Although over-sampling techniques are simple to be used as pre-process tasks, they indicated very great success in many applications. By generating artificial data for the minority classes and training a classifier on a balanced data distribution, the learning process improves significantly; consequently, over-sampling techniques have become an

• *The authors are with the Department of Computer Science and Engineering, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran. E-mail: l-abdi@cse.shirazu.ac.ir, s-hashemi@shirazu.ac.ir.*

Manuscript received 18 Oct. 2014; revised 1 July 2015; accepted 7 July 2015. Date of publication 21 July 2015; date of current version 3 Dec. 2015. Recommended for acceptance by X. He.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2458858