Soft Comput
DOI 10.1007/s00500-014-1291-z

FOCUS

To combat multi-class imbalanced problems by means of over-sampling and boosting techniques

Lida Abdi · Sattar Hashemi

Abstract Imbalanced problems are quite pervasive in many real-world applications. In imbalanced distributions, a class or some classes of data, called minority class(es), is/are under-represented compared to other classes. This skewness in the data underlying distribution causes many difficulties for typical machine learning algorithms. The notion becomes even more complicated when machine learning algorithms are to combat multi-class imbalanced problems. The presented solutions for tackling the issues arising from imbalanced distributions, generally fall into two main categories: data-oriented methods and model-based algorithms. Focusing on the latter, this paper suggests an elegant blend of boosting and over-sampling paradigms, which is called MDOBoost, to bring considerable benefits to the learning ability of multi-class imbalanced data sets. The over-sampling technique introduced and adopted in this paper, Mahalanobis distance-based over-sampling technique (MDO in short), is delicately incorporated into boosting algorithm. In fact, the minority classes are over-sampled via MDO technique in such a way that they almost preserve the original minority class characteristics. MDO, in comparison with the popular method in this field, SMOTE, generates more similar minority class examples to original class samples. Moreover, the broader representation of minority class examples is provided via MDO, and this, in turn, causes the classifier to build larger decision regions. MDOBoost increases the generalization ability of a classifier, since it indicates better results with pruned version of C4.5 classifier; unlike other over-sampling/boosting procedures, which have difficulties with pruned version of C4.5. MDOBoost is applied to real-world multi-class imbalanced benchmarks and its performance is then compared with several data-level and model-based algorithms. The empirical results and theoretical analyses reveal that MDOBoost offers superior advantages compared to popular class decomposition and over-sampling techniques in terms of MAUC, G-mean, and minority class recall.

Keywords Multi-class imbalance · Over-sampling · Mahalanobis distance · Boosting algorithm · Class decomposition techniques

1 Introduction

In recent years, with the explosive growth of available data in many scientific fields, there is a need for more robust and efficient learning techniques. Among these huge data, the imbalanced distributions are very pervasive. A data set with uneven number of instances for different classes is called an imbalance data set. This skewness in data underlying distribution poses many issues to typical machine learning algorithms. Correctly classifying the rarest or minority class instances is of great importance in imbalanced learning. Consequently, obtaining a classifier with high accuracy for the minority class without severely jeopardizing the accuracy of the majority class is a key point (He and Garcia 2009).

In many real-world applications such as protein fold and weld flaw classifications (Zhao et al. 2008; Liao 2008), there is a presence of multi-class imbalanced distributions. Processing these problems imposes many new challenges and issues which have not been seen in two-class cases.

Communicated by E. Lughofer.

L. Abdi · S. Hashemi (✉)
CSE and IT Department, Shiraz University, Engineering Campus Number 2, Mollasadra Ave., Shiraz, Iran
e-mail: s-hashemi@shirazu.ac.ir

L. Abdi
e-mail: l-abdi@cse.shirazu.ac.ir

Published online: 30 April 2014