Information Sciences 288 (2014) 347–373

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A new approach for imbalanced data classification based on data gravitation

Lizhi Peng [a,b], Hongli Zhang [a,*], Bo Yang [b], Yuehui Chen [b]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150002, PR China
[b] Shandong Provincial Key Laboratory for Network Based Intelligent Computing, University of Jinan, Jinan 250022, PR China

ARTICLE INFO

Article history:
Received 6 March 2013
Received in revised form 15 April 2014
Accepted 27 April 2014
Available online 5 August 2014

Keywords:
Data gravitation
Classification
Imbalanced data
Machine learning

ABSTRACT

Imbalanced classification is an important machine learning research topic that troubles most general classification models because of the imbalanced class distribution. A newly developed physical-inspired classification method, i.e., the data gravitation-based classification (DGC) model, performs well in many general classification problems. However, like other general classifiers, the performance of DGC suffers in imbalanced tasks. In this study, we develop a specific DGC model namely Imbalanced DGC (IDGC) model for imbalanced problems. The amplified gravitation coefficient (AGC) is introduced for gravitation computing. AGC is a type of coefficient that contains class imbalance information, which can strengthen and weaken the gravitational field of the minority and majority classes. We also design a fitness evaluation function in the weight optimization procedure of the data distribution to ensure that the model parameters adapt to the imbalanced class distributions. A total of 44 binary class data sets and 15 multiclass imbalanced data sets are used to test the performance of the proposed method. Experimental results show that the adapted DGC model is effective for imbalanced problems.

1. Introduction

Imbalanced classification problems have attracted considerable research interests in recent years. Unlike standard classification problems, an imbalanced task involves a data set that has an "imbalanced" class distribution, i.e., the number of instances in one class (majority class) is outnumbered by the number of instances in another class (minority class). This phenomenon is mainly attributed to the limited instances of the minority class. For example, in Internet traffic classification problems, Web browsing traffic is a dominant type of traffic that occurs in the Internet at each moment [70,50]. However, capturing enough malicious traffic samples, such as attacks and virus traffic, for training is difficult. Many real-world classification tasks, such as medical diagnosis [44], fraud detection [14], finance risk management [7], network intrusion detection [9], stream classification [26], and bioinformatics [67], have similar diagnosis characteristics. In these imbalanced tasks, the minority class is usually more important than the majority class [18,68]. Thus, to maximize the recognition rate of the minority class on the premise of considering a good tradeoff for both of the minority and majority classes is the main goal of an imbalanced task.

* Corresponding author.
E-mail addresses: plz@ujn.edu.cn (L. Peng), zhanghongli@hit.edu.cn (H. Zhang), yangbo@ujn.edu.cn (B. Yang), yhchen@ujn.edu.cn (Y. Chen).

http://dx.doi.org/10.1016/j.ins.2014.04.046
0020-0255/

Most standard classification models are not suitable for imbalanced tasks because such models seek high classification accuracies across the entire data set. Owing to class imbalance, a standard classifier usually classifies a minority class instance incorrectly as the model is over-trained by the majority class instances. Therefore, although a standard classifier can achieve high accuracy in an imbalanced task, the actual performance is poor because its identifying rate of the minority class is low. However, a class imbalance is not the only factor that influences classification performance [34]. Other data distribution characteristics, such as small disjuncts [63], class overlaps [25], noise [53], and borderline samples [45], also hinder classification performance.

The data gravitation-based classification (DGC) [49] model is a new classification model that is based on Newton's law of universal gravitation. The DGC model refers to a data instance in the data space as a data "particle" and considers the type of "gravitation" between any two data particles in the computation. This gravitation is directly proportional to the product of the "masses" of two data particles and is inversely proportion to the square of the distance between the data particles. By comparing the gravitation from different data classes in the training set, DGC can effectively classify a testing data instance in a simple manner. Simić et al. [56] combined DGC with case-based reasoning to create a hybrid intelligent tool for financial forecasting. Similar classification models inspired by gravitation, such as GBC [48] and CGM [64], have also been presented in recent years.

However, DGC suffers from imbalanced data sets [49], which also affect other general classifiers. In a highly imbalanced training set, the gravitational fields of the minority and majority classes are usually weak and extremely strong, respectively. The strong gravitational field of the majority class attracts most of the minority class samples to the majority class, thus resulting in a low TPR. Cano et al. have done an excellent work to adapt DGC model for imbalanced tasks [8], and the improved model is called DGC+. They construct a class-independent attribute-class weights matrix instead of the weight vector in the basic DGC model. The attribute-class weights matrix is able to carry the accurate distribution information of different classes. By using the attribute-class weights matrix, DGC+ modify the computation method of gravitations. The modified computation method of gravitation is able to effectively stress the minority class, and weaken the majority class at the same time. And they use the covariance matrix adaption evolution strategy (CMA-ES) to search the optimum class-independent attribute-class weights matrix. Empirical studies have shown that DGC+ is effective for standard, noisy, and imbalanced data [8]. However, DGC+ is time-consuming because its model parameters in the attribute-class weights matrix are far more than the parameters of the basic DGC model. And to get higher classification accuracies, DGC+ defines each instance as a single data particle, which also increases the computational complexity.

The objective of this study is to design an imbalanced classification model based on the basic DGC model to improve its behavior for imbalanced problems, especially for high imbalanced tasks. Additionally, the study is to obtain a robust approach with respect to the state-of-the-art on this topic. We introduce a new factor namely amplified gravitation coefficient (AGC) to carry the imbalanced class distribution information. AGC is used to modify the gravitation computation method. The modified gravitation computation method strengthens and weakens the gravitational field of the minority and majority classes, respectively. Unlike DGC+, we do not increase the number of model parameters, i.e. we use the feature weight vector in the basic DGC model. We also apply AGC for the weight optimization procedure to ensure that the model parameters adapt to the imbalanced class distributions.

This paper is organized as follows. Section 2 introduces the imbalanced classification problem. Section 3 discusses the basic DGC model including its theoretical basis, classification principles, feature weighting method, and data particle-creating method. Section 4 presents our proposals. Section 5 depicts our experimental setup. Sections 6 and 7 present the results and analysis of the binary class data sets and multiclass data sets, respectively. And then we summarize some lessons learned in the study in Section 8. Finally, Section 9 concludes and outlines the recommendations for the future research.

2. Imbalanced classification problem

2.1. Imbalanced data sets

Many real-world classification tasks face a challenging problem: the instances of one class outnumbers the instances of another class. These classification data sets are called imbalanced data sets, and such classification tasks are called imbalanced classification problems [30,57]. The dominant class in an imbalanced data set is called the majority class, and the subordinate class is called the minority class. In a typical binary classification task, the majority and minority classes are considered negative and positive, respectively. The imbalance ratio (IR) is the basic measure in evaluating the degree of imbalance of an imbalanced data set. It is defined as the difference between the number of instances of the majority and minority classes [46,31].

$$IR = \frac{n_{maj}}{n_{min}} \tag{1}$$

Imbalanced data sets pose a significant challenge to traditional classification techniques. Minority class instances are easily "forgotten" by general classifiers because the classifiers are over-trained by numerous majority class instances. Thus, general classifiers usually misclassify minority class instances. The most important target for an imbalanced classification task is the accurate identification of the minority class instances. Therefore, traditional classification techniques tend to be ineffective when confronting imbalanced tasks.