Neurocomputing 138 (2014) 248–259

Contents lists available at ScienceDirect

Neurocomputing

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

PDFOS: PDF estimation based over-sampling for imbalanced two-class problems ☆

Ming Gao [a], Xia Hong [a], Sheng Chen [b,c,*], Chris J. Harris [b], Emad Khalaf [c]

[a] School of Systems Engineering, University of Reading, Reading RG6 6AY, UK
[b] Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
[c] Electrical and Computer Engineering Department, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 14 October 2013
Received in revised form
8 January 2014
Accepted 1 February 2014
Communicated by K. Li
Available online 16 February 2014

Keywords:
Imbalanced classification
Probability density function based over-sampling
Radial basis function classifier
Orthogonal forward selection
Particle swarm optimisation

ABSTRACT

This contribution proposes a novel probability density function (PDF) estimation based over-sampling (PDFOS) approach for two-class imbalanced classification problems. The classical Parzen-window kernel function is adopted to estimate the PDF of the positive class. Then according to the estimated PDF, synthetic instances are generated as the additional training data. The essential concept is to re-balance the class distribution of the original imbalanced data set under the principle that synthetic data sample follows the same statistical properties. Based on the over-sampled training data, the radial basis function (RBF) classifier is constructed by applying the orthogonal forward selection procedure, in which the classifier's structure and the parameters of RBF kernels are determined using a particle swarm optimisation algorithm based on the criterion of minimising the leave-one-out misclassification rate. The effectiveness of the proposed PDFOS approach is demonstrated by the empirical study on several imbalanced data sets.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

In a typical two-class imbalanced classification problem, the instances in one class outnumber the instances of the other class. The majority class is usually referred to as the negative class, while the minority one as the positive class. Machine learning based on imbalanced data, whereby the imbalance in class distribution renders the positive class instances to be submerged in the negative class, is of great interest. The problem typically arises in life threatening or safety critical applications, such as mammography for breast cancer detection [1], mobile phone fraud detection [2], and detection of oil spills in satellite radar images [3]. In addition, many engineering applications, including information retrieval and filtering [4], direct marketing [5], risk management [6], and so on, are inherently imbalanced. In these applications, the primary objectives are often to target and explore the rare cases/classes which are less probable yet highly risky/costly. The imbalance between two classes is problematic for many standard classification algorithms [7–11]. The performances of these algorithms deteriorate as class imbalance degree increases, or as the data samples of positive class become sparser [9]. For example, the kernel-based methods, which are regarded as robust classifiers [12], construct a decision hyperplane separating two classes. Without special countermeasure, the resultant hyperplane will tend to be placed in favour of the classification performance for the negative class, but the classification performance for the target class becomes unsatisfactory. There exist a large amount of works to deal with the imbalanced learning, and the reader is referred to the excellent survey paper [12] for more information. Typical techniques of tackling the imbalanced problem can be categorised into two categories: resampling methods, also known as external methods, and imbalanced learning algorithms, often referred to as internal methods.

Imbalanced learning algorithms are obtained by modifying some existing learning algorithms internally so that they can deal with imbalanced problems effectively, without 'artificially' altering or re-balancing the original imbalanced data set. For example, the kernel classifier construction or model selection procedure can be modified, in order to cope with the imbalanced distribution during the

☆This work was supported by UK EPSRC.

* Corresponding author at: Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK.
E-mail addresses: ming.gao@pgr.reading.ac.uk (M. Gao), x.hong@reading.ac.uk (X. Hong), sqc@ecs.soton.ac.uk (S. Chen), cjh@ecs.soton.ac.uk (C.J. Harris), ekhalaf@kau.edu.sa (E. Khalaf).

http://dx.doi.org/10.1016/j.neucom.2014.02.006
0925-2312 © 2014 Elsevier B.V. All rights reserved.

classifier construction process [11,13]. A well-known radial basis function (RBF) modelling approach is the two-stage procedure [14], in which the RBF centres are first determined using the κ-means clustering [15] and the RBF weights are then obtained using the least squares estimate (LSE). To cope with imbalanced data sets, a natural extension of [14] is to modify the latter stage as the weighted LSE (WLSE), where the same weighted cost function of [13] is used. This κ-means+WLSE algorithm provides a viable technique for this category of imbalanced learning.

The resampling methods are external as they operate on original imbalanced data set, aiming to provide a re-balanced input to train a conventional classifier. One scheme is to assign different weights to the samples of the data set in accordance with their misclassification costs [16,17]. There have been a large number of studies focusing on this simple yet effective methodology to combine with the conventional classifiers for the rebalanced data set. Clearly the ultimate classification performance will be dependent on the adopted resampling strategy as well as the choice of classifier. In terms of classifier development, recently, the particle swarm optimisation (PSO) algorithm [18] has been applied to minimise the leave-one-out (LOO) misclassification rate in the orthogonal forward selection (OFS) construction of tunable RBF classifier [19,20]. PSO [18] is an efficient population-based stochastic optimisation technique inspired by social behaviour of bird flocks or fish schools, and it has been successfully applied to wide-ranging optimisation applications [21–28]. Owing to the efficiency of PSO, the tunable RBF modelling approach advocated in [19,20] offers significant advantages over many existing kernel or RBF classifier construction algorithms, in terms of better generalisation performance and smaller classifier size as well as lower complexity in learning process. With regarding to the choice of resampling strategy, we note that various resampling methods can be divided into the two basic categories, according to whether they re-balance the class distribution by under-sampling or over-sampling.

Random under-sampling is the non-heuristic method aiming to re-balance class distribution by randomly eliminating instances in the negative class [29]. Despite its simplicity, random under-sampling is considered to be one of the most effective re-sampling methods [30]. A major drawback of this technique is that it may discard data potentially important for building the classifier. Thus, many studies focus on heuristic selection techniques [31–40] to eliminate negative class instances. The method presented in [35] selectively under-samples the negative class, while keeping all the samples of the positive class. Specifically, the negative class instances are divided into the four categories: class-label noise instances A that overlap the positive class decision region; borderline instances B that are unreliable and can easily cause misclassification; redundant instances C that do not harm classification accuracy but increase classification costs; and safe instances D that are worthy of being kept for classification process. The categories A and B are detected by the use of Tomek links concept [41], as the instances complying with Tomek links are either borderline or noisy samples. Also a SHRINK [3] system attributes the overlapping regions of both the negative and positive classes as the positive class, and searches for the best positive-class region. Alternatively, Wilson's edited nearest neighbour (ENN) rule [42] is introduced to eliminate noisy instances in the negative class [43]. The ENN rule removes any instance whose class label differs from the class label of at least two of its three nearest neighbours, and a neighbourhood cleaning rule (NCL) [44] modifies the ENN by removing any negative-class instance whose class label differs from that of its 3-nearest neighbours. In order to find a consistent subset, the categories C and D are identified by involving Hart's condensed nearest neighbor (CNN) rule [45].

Under-sampling tends to be an ideal option when the imbalance degree is not very severe. However, as pointed out in [46], the use of over-sampling is necessary when the imbalance degree is high. Random over-sampling is the non-heuristic method aiming to re-balance class distribution by randomly replicating instances in the positive class. Studies [9,29] highlight that this method is simple yet very competitive to more complex over-sampling methods. However, over-fitting is a recognised serious problem for random over-sampling, because the exact copies of the instances in the positive class are made. In the study of imbalanced data sets in marketing analysis, over-sampling the positive instances with replacement is applied to match the number of negative instances [5]. The study [47] proposed a synthetic minority over-sampling technique (SMOTE), which aims to enhance the significance of some specific regions in the feature space by over-sampling the positive class. Instead of mere data oriented duplicating, SMOTE generates synthetic instances in the feature space formed by random samples along the line linking the instance and its k-nearest neighbours (k-NN). Although SMOTE is well acknowledged by the academic community, it still has some drawbacks, including over generalisation and large variance [48]. Thus, SMOTEBoost [49], borderline-SMOTE [50] and adaptive synthetic sampling (ADASYN) [51] were proposed to alleviate its limitations. Despite the empirical evidences that the foregoing methods have been effective in improving the classification performance for the target class, the reason behind the success of the oversampling approaches, such as SMOTE, is not fully understood. In fact, there exist little theoretical studies to justify most of the oversampling methods. This raises the fundamental questions as how to measure the quality of synthetic instances and why these can be used as training samples.

Against this background, we propose a novel oversampling approach based on the kernel density estimation from positive-class data samples. The estimation of the probability density function (PDF) from observed data samples is a fundamental problem in many machine learning and pattern recognition applications [52–54]. The Parzen window (PW) estimate is a simple yet remarkably accurate nonparametric density estimation technique [53–55]. According to the estimated PDF, synthetic instances are generated as the additional training data. The RBF classifier proposed in [20] is then applied to the rebalanced data set, to complete the classification process. In the generic density estimation application, the PW estimator has a well-known drawback, owing to the fact that it employs the full data sample set in defining the density estimate for a subsequent observation and, therefore its computational cost for testing directly scales with the sample size. Note that we apply the PW estimator for estimating the distribution of the minority class, which by nature consists of a small number of data samples. Therefore the potential disadvantage of the PW estimate does not exist in our application. In fact, if the sample size of the positive class is large, there will be no need to oversample it by introducing artificial samples, and the imbalance of the two classes can be better dealt with by removing some samples from the majority class, in other words, by undersampling the negative class.

The significance of our PDFOS+PSO-OFS method is twofold. Firstly, in comparison to the existing oversampling techniques, our PDFOS based oversampling approach has much stronger theoretical justification. This is because an ideal or “optimal” oversampling technique should generate synthetic data according to the same probability distribution which produces the observed positive-class data samples. By using the estimated PDF of the minority class to generate synthetic samples, the generated synthetic data follow the same statistical properties as the observed positive-class data samples. Therefore, the proposed PDFOS technique generates synthetic instances with better quality

Download English Version:

https://daneshyari.com/en/article/406530

Download Persian Version:

https://daneshyari.com/article/406530

Daneshyari.com