Knowledge-Based Systems 95 (2016) 75–85

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A maximum margin and minimum volume hyper-spheres machine with pinball loss for imbalanced data classification

Yitian Xu*, Zhiji Yang, Yuqun Zhang, Xianli Pan, Laisheng Wang

College of Science, China Agricultural University, Beijing 100083, China

ARTICLE INFO

Article history:
Received 2 August 2015
Revised 5 December 2015
Accepted 14 December 2015
Available online 21 December 2015

Keywords:
Pinball loss
Maximum margin
Minimum volume
Imbalanced data classification
Hyper-sphere

ABSTRACT

The twin hyper-sphere support vector machine (THSVM) classifies two classes of samples via two hyper-spheres instead of a pair of nonparallel hyper-planes as in the conversional twin support vector machine (TSVM). Moreover THSVM avoids the matrix inverse operation when solving two dual quadratic programming problems (QPPs). However it cannot yield a desirable result when dealing with the imbalanced data classification. To improve the generalization performance, we propose a maximum margin and minimum volume hyper-spheres machine with pinball loss (Pin-M^{3}HM) for the imbalanced data classification in this paper. The basic idea is to construct two hyper-spheres with different centers and radiuses in a sequential order. The first one contains as many examples in majority class as possible, and the second one covers minority class of examples as possible. Moreover the margin between two hyper-spheres is as large as possible. Besides, the pinball loss function is introduced into it to avoid the noise disturbance. Experimental results on 24 imbalanced datasets from the repositories of UCI and KEEL, and a real spectral dataset of Chinese grape wines indicate that our proposed Pin-M^{3}HM yields a good generalization performance for the imbalanced data classification.

1. Introduction

The support vector machines (SVMs) are amongst the most widely used machine learning techniques today. They were motivated by the celebrated work of Vapnik and his colleagues on generalization and the complexity of learning [1,2]. The basic idea of the classical SVM is to maximize the margin between two classes by minimizing the regularization term and minimize classification errors with hinge loss function. Although SVMs have a long history of literatures, there are still great efforts on improvements. It has evolved into a multitude of diverse formulations with different properties, e.g. ν-SVM [3], Pin-SVM [4], and SSLM [5].

A main challenge for the standard SVM is the high computational complexity [6,7]. Many improved versions have been proposed, e.g. Chunking [1], SMO [8], SVMLight [9] , LSSVM [10,11], Libsvm [12], Liblinear [13], PSVM [14], and TSVMs [15–18]. All of these methods classify the samples via hyper-plane. In addition, some variants based on hyper-sphere, such as twin hyper-sphere support vector machine (THSVM) [19,20], have emerged in recent years. The THSVM determines two hyper-spheres by solving two related SVM-type problems. Each of them is smaller than the classical SVM, which makes the THSVM more efficient than the classical SVM. Moreover, the THSVM avoids the matrix inverse operation [21] when solving two dual quadratic programming problems (QPPs) compared with the TSVM.

The methods mentioned above for data engineering have shown great success in many real-world applications. However, they may yield an undesirable result when dealing with imbalanced data classification. Actually, most standard algorithms assume or expect balanced class distributions or equal misclassification cost. Therefore, when presented with complex imbalanced data sets, these algorithms fail to properly represent the distributive characteristics of the data and resultantly provide unfavorable result. The problem of learning from imbalanced data is a relatively new challenge that has attracted growing attention from both academia and industry.

Researchers have studied the performance of SVM for the imbalanced data [22]. Many modifications of the classical SVM to fit the imbalanced classification also have been proposed, including sampling, cost-sensitive, ensemble and kernel modification methods, such as the combination approach of SMOTE and biased-SVM for imbalanced datasets [23,24], the efficient weighted Lagrangian twin support vector machine (WLTSVM) [25], the k-category proximal support vector machine (PSVM) [26], the weighted least squares SVMs (WLSSVMs) [27,28], and the total margin-based adaptive fuzzy SVM kernel method (TAF-SVM) [29,30].

In this paper, we propose the maximum margin and minimum volume hyper-spheres machine with pinball loss (Pin-M^{3}HM) for

* Corresponding author. Tel.: +86 1062737077.
E-mail address: xytshuxue@126.com (Y. Xu).

http://dx.doi.org/10.1016/j.knosys.2015.12.005
0950-7051/

two-class imbalanced data. For the between-class imbalance [31], our Pin-M^{3}HM aims at constructing two hyper-spheres with different centres and radiuses. The first one contains majority class of examples and the second one covers minority class of examples. Two hyper-spheres are generated in sequential order. Furthermore, by employing the pinball loss [4,32–36], our algorithm is equipped with noise insensitivity which has a significant advantage in the imbalanced data classification. Moreover, the Pin-M^{3}HM has a faster learning speed as the following three reasons: Firstly, it has a simpler form to solve the minority class hyper-sphere since it replaces a time-consuming term with a given majority class center a_1 from the previous step; Besides, it also solves two smaller-sized QPPs instead of a large QPP as in the classical SVM; Thirdly, it successfully avoids the matrix inverse operations when solving the dual QPPs.

The rest of this paper is organized as follows: Section 2 outlines the related works. The Pin-M^{3}HM is proposed in Section 3. In Section 4, we theoretically compare the proposed Pin-M^{3}HM with the THSVM. Experiments on twenty-four imbalanced datasets from the repository of UCI and KEEL, and a real spectral datasets of Chinese grape wines are conducted to verify the validity of our Pin-M^{3}HM in Section 5. The last section gives the conclusion.

2. Related works

In this section, we provide a review on the classification methods for imbalanced problems and some innovative works based on SVM.

2.1. Classification methods for imbalanced problems

With the great influx of attention devoted to the imbalanced learning problem, numerous techniques have been developed to deal with this issue. This section is a sort summary of the classification methods for imbalanced datasets [31,37–39].

– **Data level approaches**
  Typically, the use of data level approaches (sampling methods) in imbalanced learning applications consists of the modification of an imbalanced data set by some mechanisms in order to provide a balanced distribution.
  • Random oversampling and undersampling. In the case of undersampling, removing examples from the majority class may cause the classifier to miss important concepts pertaining to the majority class; in regards to oversampling, since it simply appends replicated data to the original data set, multiple instances of certain examples become “tied”, leading to overfitting.
  • Informed undersampling. Some examples are presented as the *EasyEnsemble* and *BalanceCascade, NearMiss* and the one-sided selection (OSS) method.
  • Synthetic sampling with data generation. The synthetic minority oversampling technique (SMOTE) is a powerful method that has shown a great deal of success in various applications.
  • Adaptive synthetic sampling. Some representative work includes the borderline-SMOTE and adaptive synthetic sampling (ADASYN) algorithms.
  • Other sampling method, such as data cleaning techniques, cluster-based methods and integration of sampling and boosting [40,41].
– **Algorithm level approaches**
  This category contains variants of existing classifier learning algorithms biased towards learning more accurately the minority class.
  • Cost-sensitive methods. Cost-sensitive learning targets the imbalanced learning problem by using different cost matrices that describe the costs for misclassifying any particular data example, such as cost-sensitive dataspace weighting with adaptive boosting (e.g. *AdaC1, AdaC2*, and *AdaC3*) , cost-sensitive decision trees, cost-sensitive neural networks (e.g. the popular multilayer perceptron (MLP) model), cost-sensitive Bayesian classifiers and some works exist which integrate cost functions with SVMs [27,28,42].
  • Kernel-based methods and active learning methods. Some examples include the SMOTE with different costs (SDCs) method and the ensembles of over/undersampled SVMs.
  • Additional methods for imbalanced learning. These approaches aim to recognize instances of a concept by using mainly, or only, a single class of examples rather than differentiating between instances of both positive and negative classes as in the conventional learning approaches, such as the one-class learning or novelty detection methods.

Generally speaking, various empirical studies have shown that in some application domains, including certain specific imbalanced learning domains [43–45], the algorithm level approaches are superior to sampling methods. Our proposed method in this paper belongs to the algorithm level approaches.

2.2. Works based on support vector machine

Traditional SVM facilitates learning by using specific examples near concept boundaries (support vectors) to maximize the separation margin between the support vectors and the hypothesized concept boundary (hyperplane) [46], meanwhile minimizing the total classification error.

2.2.1. Loss function

A loss function describes the cost of the discrepancy between the prediction $f(x)$ and the observation y at the point x. Let $(X, \mathcal{A})$ be a measurable space and $Y \subset R$ be a closed subset. Then function $L: X \times Y \times R \to [0, \infty)$ is called a loss function, or simply a loss, if it is measurable.

Loss functions can be used to formalize many learning goals, including classification, regression, and density level detection problems. They have a long history in mathematical statistics and machine learning. For example, the least squares loss for regression was already used by Legendre, Gauss, and Adrain in the early 19th century [36] and the classification loss function dates back to the beginning of machine learning.

Since the classification loss typically leads to computationally hard optimization problems, margin-based surrogates were considered in many machine learning algorithms, such as hinge loss, squared hinge loss, logistic loss, and so on.

2.2.2. Support vector machine with hinge loss

For SVM, the first surrogate for the classification loss was the hinge loss, i.e.,

$$L_{hinge}(x, y, f(x)) = max\{1 - yf(x), 0\}, y = \pm 1. \tag{1}$$

We consider a binary classification problem with a training set $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\}$, where $x_i \in R^n$ is the input and $y_i \in \{1, -1\}$ is the corresponding output. The SVM searches for a separating hyper-plane $f(x) = w^T x + b = 0$, where $w \in R^n$ and $b \in R$. On one hand, maximizing the margin between two classes is equivalent to minimization of the regularization term $\frac{1}{2}\|w\|^2$. On the other hand, minimizing the training error $L(x, y, f(x))$ can be represented as the term $\sum_{i=1}^{l} L(x_i, y, f(x_i))$. Finally, we have the following formulation:

$$\min_{w,b} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} L(x_i, y, f(x_i)), \tag{2}$$