Applied Mathematics and Computation 283 (2016) 141–152

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Distance-based margin support vector machine for classification

Yan-Cheng Chen[a], Chao-Ton Su[b,*]

[a] National Chung-Shan Institute of Science and Technology, Taoyuan City, Taiwan
[b] Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Room 820, Engineering Building I, 101, Sec. 2, Kuang Fu Road, Hsinchu 30013, Taiwan

ARTICLE INFO

Keywords:
Support vector machine
Class imbalance
Class overlapping
Classification

ABSTRACT

Recently, the development of machine-learning techniques has provided an effective analysis tool for classification problems. Support vector machine (SVM) is one of the most popular supervised learning techniques. However, SVM may not effectively detect the instance of the minority class and obtain a lower classification performance in the overlap region when learning from complicated data sets. Complicated data sets with imbalanced and overlapping class distributions are common in most practical applications. Moreover, they negatively affect the classification performances of the SVM. The present study proposes the use of modified slack variables within the SVM (MS-SVM) to solve complex data problems, including class imbalance and overlapping. Artificial and UCI data sets are provided to evaluate the effectiveness of the MS-SVM model. Experimental results indicate that the MS-SVM performed better than the other methods in terms of accuracy, sensitivity, and specificity. In addition, the proposed MS-SVM is a robust approach for solving different levels of complex data sets.

1. Introduction

The SVM, which has become one of most powerful classification techniques in the data mining field, was proposed by Vapnik [1]. SVM has a strong statistical learning theory and has shown excellent classification results in many applications, from handwritten digit recognition [2] to text categorization [3]. SVM also works very well with high-dimension data sets and avoids the dimensionality problem [4]. The aim of the SVM is to minimize classification errors by maximizing the margin between the separating hyperplane and the data sets. A special property of the SVM is that it simultaneously minimizes the empirical classification error and maximizes the geometric margin.

However, the SVM may not effectively detect the instance of the minority class when learning from complex data sets. Complicated data sets with class imbalance and overlapping distributions are common in most practical applications, and affect the classification performances of the SVM. In predicting the rarest objects, the hyperplane or decision boundary generated by the SVM can be severely skewed toward the majority class, especially on the class imbalance if the training instances of the majority class outnumber that of the minority class. For class overlapping problem, the clear interval of separation of the two classes was almost non-existent. On the complex overlapping regions, the decision boundary without flexible property is difficult to separate from the corrected classes because the regions of the different classes usually have

* Corresponding author. Tel.: +886 3 5742936; fax: +886 3 5722204.
E-mail address: ctsu@mx.nthu.edu.tw (C.-T. Su).

http://dx.doi.org/10.1016/j.amc.2016.02.024
0096-3003/

the same ranges in the single attribute or multi-attributes. SVM usually produces high predictive values over a majority class and poor values over a minority class when learning from complex data sets. This condition results in a serious bias in the minority class, especially in medical disease and credit scoring problems. For example, if 1% of the patients have medical disease, then the model that predicts every patient as healthy has a 99% accuracy, although it fails to detect any of the disease conditions.

Recent studies have suggested that four main directions cope with data complexity problems. Several researchers have examined various methods using sampling methods, the modified kernel boundary, trade-off cost matrices, and adaptive margins. Preprocessing the data via down-sampling the majority class or over-sampling the minority class is adopted as the first approach. Several researchers have studied over-sampling and down-sampling techniques for skewed or imbalanced data sets [5,6]. The first approach focuses on the concept of sampling methods. The performance of the classification methods changes from imbalanced to newly balanced data sets. Synthetic minority over-sampling technique (SMOTE) is a well-known algorithm to fight unbalanced classification problems [5]. The general idea of this method is to generate new examples of the minority class artificially using the nearest neighbors. Furthermore, the majority class examples are also under-sampled, leading to a more balanced data set. The lack of rigorous and systematic treatment for imbalanced data is a common problem in sampling methods [7]. For example, down-sampling training data sets can potentially remove certain important information, whereas over-sampling may introduce noise. The goal of the second approach is to place the learned hyperplane farther from the positive class. Wu and Chang [8] proposed the adjustment of the class boundary by modifying the kernel matrix according to the class distributions to avoid the hyperplane from getting closer to the positive instances. The original kernel matrix and conformal transformation kernel matrix will produce bias during the iterative procedure. In the third approach, Veropoulos et al. [9] suggested the control of the trade-off between the false positives (FPs) and false negatives (FNs) by the data mining technique to overcome class imbalance problems. For the fourth approach, Trafalis and Gilbert [10,11] and Song et al. [12] proposed to replace the original elements by modifying the slack variables and augmented data points. Replacing the slack variables with modified slack variables makes the adaptive margin and decision function less affected by the abnormal example. Using the composed mean and standard deviation of each data point replaces the original data point to overcome data perturbations [10,11] based on the linear and non-linear data sets. Augmented terms designed by large margins in geometry have been proposed by considering the original elements of SVMs in a convex quadratic programming model.

The present study aims to develop the modified slack variables within the SVM to enhance its classification performance and adjust the suitable distance between the decision boundary and the margin. The main idea of the proposed method follows that in the previous studies [12,13], which use the novel modified margin to reformulate the model. Novel modified margins use distance metrics to replace the original slack variables and enlarge the distance between the rarest objects and the normal groups when learning from complex data sets. The original slack variables in the SVM model are not sensitive to the rarest objects in the learning stage. Considering the advantage of distance metrics, the proposed method can achieve a good performance by eliminating the influence of class distributions and separating the majority and minority classes effectively. Slack variables based on distance metrics are included in the optimal function of the dual problem. The selected parameters of which are determined by the gradient descent method. The proposed method, with the aligned slack variables, will be not affected easily by the complex data sets.

The proposed method performs similar functions as that of the standard SVM in terms of synthetic and UCI data sets. Data complexity metrics are used to evaluate the performance of the proposed method in various scenarios. The experimental results of these metrics provide a qualitative description of the training data characteristics. Similarly, they could help explain the good performance of the proposed method under different data set scenarios.

The present paper is structured as follows. Section 2 presents the proposed modified SVM. Section 3 consists of the systematic experiments on the synthetic and UCI data sets based on the measurement of the performance metrics. Section 4 summarizes the experimental results. Finally, the main remarks and future works are presented in Section 5.

2. Proposed modified slack variables within the SVM (MS-SVM)

2.1. Original SVM

Vapnik [1] first proposed the SVM, a set of related machine-learning techniques, to solve classification problems. In a simple pattern classification problem, a hyperplane separates two classes of patterns based on given examples $\boldsymbol{x_i}$, $\boldsymbol{y_i}$, for $i = 1, \ldots, n$, where $\boldsymbol{x_i}$ is a vector in input space $\mathcal{S} \in \mathbb{R}^n$ and denotes the class index, taking the value $+1$ or -1. Kernel trick transforms data $\boldsymbol{x_i}$ from $\mathcal{S}$ into the feature space $\mathcal{F} \in \mathbb{R}^N$ (N may be infinite) using nonlinear mapping $\phi(\mathbf{x})$. Kernel trick uses a classifier algorithm to solve a nonlinear problem by mapping the original nonlinear observations into a higher dimensional $\mathcal{F}$ in machine learning. It then searches for a linear decision function

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b \tag{2.1}$$

in the feature space. Patterns are classified by the sign in Eq. (2.1). If no hyperplane splits the positive and negative instances, the soft margin method selects a hyperplane that splits the instances as cleanly as possible. At the same time, the distance to the nearest cleanly split instances is maximized. This method can be accomplished by introducing positive-valued slack