Pattern Recognition 49 (2016) 55–64

Contents lists available at ScienceDirect

Pattern Recognition

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

Smoothly approximated support vector domain description

Songfeng Zheng *

Department of Mathematics, Missouri State University, 901 S. National Avenue, Springfield, MO 65897, USA

ARTICLE INFO

Article history:
Received 14 February 2013
Received in revised form
21 November 2014
Accepted 4 July 2015
Available online 17 July 2015

Keywords:
Support vector domain description
Smooth approximation
Quadratic programming
conjugate gradient

ABSTRACT

Support vector domain description (SVDD) is a well-known tool for pattern analysis when only positive examples are reliable. The SVDD model is often fitted by solving a quadratic programming problem, which is time consuming. This paper attempts to fit SVDD in the primal form directly. However, the primal objective function of SVDD is not differentiable which prevents the well-behaved gradient based optimization methods from being applicable. As such, we propose to approximate the primal objective function of SVDD by a differentiable function, and a conjugate gradient method is applied to minimize the smoothly approximated objective function. Extensive experiments on pattern classification were conducted, and compared to the quadratic programming based SVDD, the proposed approach is much more computationally efficient and yields similar classification performance on these problems.

1. Introduction

There exist a range of pattern recognition problems, such as novelty detection, where the task is to discriminate the pattern of interest from imposters. In such cases, positive examples for training are relatively easier to obtain and more reliable. However, although negative examples are very abundant, it is usually difficult to sample enough useful negative examples to adequately model the imposters since they may belong to any class. In this situation, it is often reasonable to assume positive examples to cluster in a certain way. Under this assumption, the goal is to accurately describe the class of positive examples as opposed to the very wide range of negative examples, which are not of interest.

To this end, Tax and Duin [30–32] developed a support vector domain description (SVDD) method, which fits a tight hypersphere in the nonlinearly transformed feature space to enclose most of the positive examples. Thus, SVDD could be regarded as a description of the class of interest. Extensive experiments show that SVDD can correctly identify some negative examples even though it has not seen any negative example during the training phase [30–32].

SVDD is, like support vector machine (SVM) [34, chapter 10], a kernel method, thus inherits all the related advantages of SVM. Since it was proposed, SVDD has been applied to various application problems, including image classification [39], remote sensing image analysis [2,23,24], medical image analysis [29], machine diagnostics [33,38], and multi-class classification problems [18,37], among others. Furthermore, SVDD is a preliminary step for support vector clustering [3,19,20].

Similar to SVM, the formulation of SVDD leads us to a quadratic programming problem (see Section 2 for more details). Although the decomposition techniques [25,26] or sequential minimization methods [27] could be employed to solve the quadratic programming problem, the training of SVDD has time complexity about $O(n^3)$ (see the end of Section 2 for details), where n is the training set size. High training cost is undesirable, especially for model selection and some feature selection methods, where the training algorithm often needs to run multiple times. Therefore, it is highly appreciated to develop time-efficient yet accurate enough training algorithms for SVDD.

As an alternative, we can fit the SVDD model by directly optimizing the primal objective function, as the similar work for SVM [9]. However, the primal objective function of SVDD is not differentiable which prevents gradient based methods [4,36] from being applicable, although they are easy to implement, and converge fast to at least a local optimum. As such, we introduce a smooth approximation to the primal objective function of SVDD, which is an upper bound of the primal objective function and converges uniformly to the primal objective function as the controlling smoothing parameter increases. Then, conjugate gradient method is employed to minimize the proposed smoothly approximated objective function. We test the proposed approach on face detection and handwritten digit recognition problems, and detailed performance comparison on these problems demonstrates that the proposed smoothly approximated SVDD (SA-SVDD) often yields testing accuracy very close to that of the quadratic programming based SVDD (QP-SVDD). However, SA-SVDD is much more computationally efficient than QP-SVDD.

* Tel.: +1 417 836 6037; fax: +1 417 836 6966.
E-mail address: SongfengZheng@MissouriState.edu

http://dx.doi.org/10.1016/j.patcog.2015.07.003
0031-3203/

The rest of this paper is organized as follows: Section 2 briefly reviews the formulation of SVDD; Section 3 proposes the smoothly approximated SVDD model, and a conjugate gradient method is presented to minimize the smoothed objective function; a brief computational complexity analysis is also presented in Section 3; Section 4 compares the classification performances and training time of the proposed SA-SVDD algorithm to the original QP-SVDD on two publicly available real-world datasets; finally, Section 5 summarizes this paper and discusses some future research directions.

2. Support vector domain description

This section briefly reviews the general formulation of support vector domain description (SVDD) with only positive examples (Section 2.1) and with both positive and negative examples (Section 2.2). Refer to [30–32] for more detailed derivations.

2.1. SVDD with positive examples

Given training data $\{\mathbf{x}_i, i=1,\ldots,n\}$ with the feature vector $\mathbf{x}_i \in \mathbf{R}^p$, SVDD is looking for a hypersphere (in a high dimensional Hilbert feature space $\mathcal{H}$ where the examples have been mapped through a nonlinear transformation Φ) of radius $R>0$ and center $\mathbf{c}$ with a minimum volume containing most of the data. Therefore, we have to minimize R^2 with constraints $\|\Phi(\mathbf{x}_i)-\mathbf{c}\|^2 \le R^2$, for $i=1,\ldots,n$. In addition, since the training sample might contain outliers, we introduce a set of slack variables $\xi_i \ge 0$, as in the framework of support vector machine (SVM) [34, chapter 10]. The slack variable ξ_i measures how much the squared distance from the ith training example to the center exceeds the radius squared. Therefore, the slack variable could be understood as a measure of error. Taking the constraints into account, the problem becomes

$$\min_{R,\mathbf{c},\xi} \quad F(R,\mathbf{c},\boldsymbol{\xi}) = R^2 + C\sum_{i=1}^{n}\xi_i, \tag{1}$$

with constraints

$$\|\Phi(\mathbf{x}_i)-\mathbf{c}\|^2 \le R^2+\xi_i, \quad \xi_i \ge 0 \ \text{ for } i=1,\ldots,n, \tag{2}$$

where $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)'$ is the vector of slack variables, and the parameter $C>0$ controls the tradeoff between the volume of the hypersphere and the permitted errors.

The Lagrangian dual of the above optimization problem is

$$\min_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j) - \sum_{i=1}^{n}\alpha_i K(\mathbf{x}_i,\mathbf{x}_i), \tag{3}$$

with constraints

$$\sum_{i=1}^{n}\alpha_i = 1, \quad 0 \le \alpha_i \le C \ \text{ for } i=1,\ldots,n, \tag{4}$$

where $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_n)'$ with α_i being the Lagrangian multiplier for the ith constraint, and $K(\mathbf{x}_i,\mathbf{x}_j) = <\Phi(\mathbf{x}_i),\Phi(\mathbf{x}_j)>$ is the kernel function which satisfies Mercer's condition [34, chapter 10]. From the Karush–Kuhn–Tucker (KKT) conditions [4, chapter 3] [6, chapter 5], the center of the hypersphere in the high dimensional feature space $\mathcal{H}$ can be represented in terms of the Lagrangian multipliers as

$$\mathbf{c} = \sum_{i=1}^{n}\alpha_i\Phi(\mathbf{x}_i). \tag{5}$$

Once the parameters α_i's are obtained, the radius R can be computed from the set of support vectors.

In decision making stage, if the distance from a new example $\mathbf{x}$ is less than the radius R, it is classified as a positive example; otherwise, it is classified as a negative example. Thus, the decision rule is

$$f(\mathbf{x}) = \operatorname{sign}\left(R^2 - \|\Phi(\mathbf{x}) - \sum_{i=1}^{n}\alpha_i\Phi(\mathbf{x}_i)\|^2\right) = \operatorname{sign}\left(2\sum_{i=1}^{n}\alpha_i K(\mathbf{x}_i,\mathbf{x}) - K(\mathbf{x},\mathbf{x}) + b\right), \tag{6}$$

where $b = R^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\mathbf{x}_i,\mathbf{x}_j)$.

2.2. 2-Class SVDD

If negative examples are available, we could integrate this part of information to the formulation of SVDD. In this situation, we would prefer the hypersphere enclosing as many positive examples as possible and excluding as many negative examples as possible, and again, we want the volume of the hypersphere to be as small as possible. Let the training set be $\{(\mathbf{x}_i,y_i), i=1,2,\ldots,n\}$, where $y_i \in \{+1,-1\}$, with $y_i=+1$ for positive examples and $y_i=-1$ for negative examples. As in Section 2.1, we denote the radius of the hypersphere as R and denote its center as $\mathbf{c}$.

Suppose we impose different penalties for misclassifying positive and negative examples, then similar to Section 2.1, the optimization problem could be summarized as

$$\min_{R,\mathbf{c},\xi} \quad F(R,\mathbf{c},\boldsymbol{\xi}) = R^2 + C_{+1}\sum_{i:y_i=+1}\xi_i + C_{-1}\sum_{i:y_i=-1}\xi_i = R^2 + \sum_{i=1}^{n}C_{y_i}\xi_i, \tag{7}$$

where C_{+1} and C_{-1} are the penalties on mistakenly classifying a positive or negative example, respectively. As in Section 2.1, ξ_i is the slack variable on the ith example, and it should satisfy the constraints

$$\|\Phi(\mathbf{x}_i)-\mathbf{c}\|^2 \le R^2+\xi_i, \quad \xi_i \ge 0 \ \text{ for } y_i=+1, \tag{8}$$

and

$$\|\Phi(\mathbf{x}_i)-\mathbf{c}\|^2 \ge R^2-\xi_i, \quad \xi_i \ge 0 \ \text{ for } y_i=-1. \tag{9}$$

We compactly rewrite the constraints in one equation as

$$y_i\left(\|\Phi(\mathbf{x}_i)-\mathbf{c}\|^2 - R^2\right) \le \xi_i, \quad \xi_i \ge 0 \ \text{ for } i=1,\ldots,n. \tag{10}$$

By using the Lagrange multiplier method, we get the dual problem as

$$\min_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) - \sum_{i=1}^{n}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x}_i), \tag{11}$$

with constraints

$$\sum_{i=1}^{n}\alpha_i y_i = 1, \quad 0 \le \alpha_i \le C_{y_i} \ \text{ for } i=1,\ldots,n. \tag{12}$$

By the KKT condition, the center of the hypersphere can be represented as

$$\mathbf{c} = \sum_{i=1}^{n}\alpha_i y_i \Phi(\mathbf{x}_i). \tag{13}$$

Once the parameters α_i's are obtained, the radius R can be computed from the set of support vectors.

Given a new example $\mathbf{x}$, the decision rule is

$$f(\mathbf{x}) = \operatorname{sign}\left(R^2 - \|\Phi(\mathbf{x}) - \sum_{i=1}^{n}\alpha_i y_i\Phi(\mathbf{x}_i)\|^2\right) = \operatorname{sign}\left(2\sum_{i=1}^{n}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) - K(\mathbf{x},\mathbf{x}) + b\right), \tag{14}$$

where $b = R^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)$.

We should notice that minimizing the objective function in Eq. (7) does not imply strong generalization ability of the resultant