Neurocomputing 238 (2017) 328–339

Contents lists available at ScienceDirect

Neurocomputing

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

Ellipsoidal data description

Kunzhe Wang, Huaitie Xiao*

Science and Technology on Automatic Target Recognition Laboratory, National University of Defense Technology, Changsha 410073, China

ARTICLE INFO

Article history:
Received 21 May 2016
Revised 5 December 2016
Accepted 28 January 2017
Available online 9 February 2017

Communicated by Prof. Zidong Wang

Keywords:
Novelty detection
Support vector machines
Kernel methods
Minimum volume enclosing ellipsoid
Mahalanobis distance
Rademacher complexity

ABSTRACT

Support vector data description (SVDD) is a leading classification method for novelty detection, which minimizes the volume of a spherically shaped decision boundary around the normal class. While SVDD has achieved promising performance, it will lead to a loose boundary for multivariate datasets of which the input dimensions are usually correlated. Inspired by the relationship between kernel principal component analysis (kernel PCA) and the best-fit ellipsoid for a dataset, this study proposes the ellipsoidal data description (ELPDD) which considers feature variance of each dimension adaptively. A minimum volume enclosing ellipsoid (MVEE) is constructed around the target data in the kernel PCA subspace which can be learned via a SVM-like objective function with log-determinant penalty. We also provide the Rademacher complexity bound for our model. Some relating problems are investigated in detail. Experiments on artificial and real-world datasets validate the effectiveness of our method.

1. Introduction

Support vector machine (SVM) [1,2] has received considerable interest in the community of machine learning and gained great success in many applied domains in the past two decades. As a derivative, support vector data description (SVDD) [3,4] was designed to the problem of novelty detection, which encloses most or all of the target data within a spherical boundary. It has been shown that SVDD is equivalent to one-class SVM (OCSVM) which separates the data from the origin with maximal margin when the Gaussian kernel is used [5]. All support vector algorithms can be generalized to their kernel counterparts in the Hilbert space using kernel trick. In practice there are some issues in the optimization procure of SVDD, which will lead to infeasible dual problem or local optima. Chang et al. [6] conducted a comprehensive analysis of SVDD in the view of convex optimization and derived the dual problem of SVDD with strong duality. They also considered $L2$ loss in the formulation of SVDD as a novel extension.

Though powerful description ability has been justified for SVDD, it will collapse with polynomial kernel as samples with large norms in feature space will overwhelm others, resulting in a superfluous region which does not contain target training objects [4]. In essence, the problem lies in the fact that a spherical boundary only characterizes the data by its center and radius. Even for the Gaussian kernel, SVDD can lead to large empty areas around the normal class in the input space in cases that the input dimensions are not isotropic or independent, which are usually true for real datasets [7]. In addition, a hypersphere in the feature space will significantly overestimate the support of the distribution in directions perpendicular to the subspace where the data lies in.

As mentioned above, an ellipsoid is thus preferred to model the heterogeneous dataset. It takes into account the differences in variance for each dimension as well as covariance between them. Ellipsoid estimation is a rather general problem in statistics and finds its application in many areas. For novelty detection, it turns out to be the problem of identifying a unique ellipsoid with minimum volume covering most or all of the normal class, which is usually referred to as minimum volume enclosing ellipsoid (MVEE). The calculation of MVEE has been demonstrated to be the problem of D-optimality [8] which seeks to maximize the determinant of the covariance matrix of the design. Previous researches mainly focused on developing algorithms for the computation of MVEE [9,10]. However, a little work has attempted to generalize it to the feature space.

There are certain difficulties in kernelization of the algorithm. Firstly, kernel trick cannot be applied directly as in SVDD. The objective function of MVEE is expressed as outer products between samples, and thus the extension is not trivial. Secondly, the M matrix will be singular or positive semi-definite in high dimensional feature space, which gives an ill-posed problem. Existing literatures tried to tackle the above problems in several ways. A hyper-ellipse was built to detect intrusion in computer networks and further

* Corresponding author.
E-mail addresses: htxiao@126.com, wangkunzhe@163.com (H. Xiao).

http://dx.doi.org/10.1016/j.neucom.2017.01.070

extended to high dimensional space through approximation of the kernel function by its Taylor expansion [11]. A kernel whitening algorithm was proposed as a preprocessing method for SVDD, which is identical to fitting an ellipsoid in the original kernel space [7]. Similar work can also be found in [12]. Another technique is to convert the D-optimal experimental design into an iterative algorithm in terms of a Mahalanobis distance defined in the kernel space [13]. Although these approaches outperform SVDD, their kernelizations are all not conducted naturally. As for the singularity of the covariance matrix, a regularized MVEE was proposed to make it applicable in high dimensional feature space [14].

This study proposes a generalized version of SVDD, referred to as ellipsoidal data description (ELPDD), which considers feature variance of each dimension adaptively. An ellipsoid with soft margin is constructed in the kernel PCA subspace without kernel feature extraction and the complexity of the proposed model is measured using log-determinant. Our thought comes from the fact that PCA can be thought of as fitting an ellipsoid to the data, while each axis of the ellipsoid corresponds to a principal component. Following similar line of derivation in kernel PCA [15], the resulting model can be learned via a SVM-like objective function with log-determinant penalty instead of $L2$-norm penalty. A Mahalanobis distance for novelty detection is obtained based on the projection method. Time complexity of ELPDD is discussed through simulations. We also provide the Rademacher complexity bound for our model. Some properties on model selection, including the kernel type, hyperparameters as well as error estimate are investigated in detail. The experimental results demonstrate that ELPDD gets better performances than other traditional classifiers and SVDD in terms of F_1 score and AUC.

The remainder of this paper is organized as follows. We review the standard SVDD in Section 2. In Section 3, the formulation of ELPDD is derived in detail both in the input space and the feature space. The implementation of the algorithm is also provided there. Discriminant for novelty detection is given in Section 4. Section 5 compares the computational complexity of ELPDD with that of SVDD through simulations. A stability analysis for novelty detection is provided in Sections 6 and 7 investigates the problem of model selection for ELPDD. Experimental results can be found in Section 8. We conclude with a discussion in Section 9.

2. Support vector data description (SVDD)

Given a class of objects $\{\mathbf{x}_i\}$ of size m in the input space $\mathbb{R}^n$, SVDD intends to find a sphere with minimum volume containing all or most of this class. By introducing the slack variables ξ_i, it turns out to be finding the smallest sphere with center $\mathbf{c}$ and radius R that solves the optimization problem

$$\begin{aligned}&\min_{\mathbf{c},R,\boldsymbol{\xi}}\ R^2+\tfrac{1}{\nu m}\sum_i \xi_i\\&\text{subject to } \|\mathbf{x}_i-\mathbf{c}\|^2\le R^2+\xi_i\end{aligned}\tag{1}$$

The Lagrangian dual of the above problem reduces to be

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\ \sum_i \alpha_i \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i-\sum_{i,j}\alpha_i\alpha_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\\&\text{subject to } 0\le\alpha_i\le\tfrac{1}{\nu m},\quad \sum_i\alpha_i=1\end{aligned}\tag{2}$$

and the solution $\mathbf{c}=\sum_i\alpha_i\mathbf{x}_i$. $\nu\in(0,1]$ is an upper bound on the fraction of outliers and lower bound on the fraction of support vectors (SV's), which is referred to as ν-property [5]. The radius R is the distance from the center to any SV with $0<\alpha_i<1/\nu m$. The corresponding decision function for a test point is of the form

$$f(\mathbf{x})=\operatorname{sgn}\left(R^2-\sum_{i,j}\alpha_i\alpha_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j+2\sum_i\alpha_i\mathbf{x}^{\mathrm{T}}\mathbf{x}_i-\mathbf{x}^{\mathrm{T}}\mathbf{x}\right)\tag{3}$$

The above derivation can be kernelized by simply replacing the inner product by some more general kernel function. Thus more flexible data descriptions can be obtained.

3. Ellipsoidal data description (ELPDD)

As has been pointed out in Section 1, MVEE provides better description ability for a dataset. An enclosing ellipsoid in $\mathbb{R}^n$ is characterized by center $\boldsymbol{\mu}$ and scatter $\mathbf{Q}$, and is defined by

$$E_{\mathbf{Q},\mu}=\left\{\mathbf{x}\in\mathbb{R}^n:(\mathbf{x}-\mu)^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{x}-\mu)\le 1\right\}\tag{4}$$

where $\mathbf{Q}\in\mathbb{R}^{n\times n}$ is a symmetric positive-definite matrix of which the eigenvectors define the axes of the ellipsoid and the corresponding eigenvalues λ are the squares of the semi-axes. We minimize the volume of the ellipsoid by minimizing $\det\mathbf{Q}=\Pi_{i=1}^n\lambda_i$. It has been demonstrated that the problem of MVEE is strictly convex [16,17]. There is, however, a practical objection when $\mathbf{Q}$ is singular which is actually inevitable in high dimensional space. In this section, we first define the ELPDD in the input space, which gives a closed boundary around target data. Then a kernel version of this algorithm is designed for nonlinear separable case between targets and outliers. At the end of this section we will show that the solution of ELPDD can be obtained by standard SDP toolbox.

3.1. ELPDD in the input space

To define the ELPDD in the input space, we solve the following convex program

$$\begin{aligned}&\min_{\mathbf{Q}^{-1},\boldsymbol{\mu},R,\boldsymbol{\xi}}-\ln\det\left(\mathbf{Q}^{-1}\right)+R^2+\tfrac{1}{\nu m}\sum_i\xi_i\\&\text{subject to } (\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\le R^2+\xi_i\end{aligned}\tag{5}$$

where there is an implicit constraint $\mathbf{Q}\succ 0$. The above formulation allows for mislabeled examples by introducing the slack variables $\xi_i\ge 0$. The constant ν controls the tradeoff between volume of the ellipsoid and error on the target. R is the Mahalanobis distance from the ellipsoid center.

Following the standard optimization technique, we construct the Lagrange functional for the above problem

$$\begin{aligned}L(\mathbf{Q}^{-1},\boldsymbol{\mu},R,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=&-\ln\det\left(\mathbf{Q}^{-1}\right)+R^2+\tfrac{1}{\nu m}\sum_i\xi_i\\&+\sum_i\alpha_i\left[(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})-R^2-\xi_i\right]-\sum_i\beta_i\xi_i\end{aligned}\tag{6}$$

where α_i, β_i are nonnegative Lagrange multipliers. The Lagrangian has to be minimized with respect to $\mathbf{Q}^{-1}$, $\boldsymbol{\mu}$, R and $\boldsymbol{\xi}$ while maximized with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

At the saddle point, the solutions $\mathbf{Q}^{-1}$, $\boldsymbol{\mu}$, R and $\boldsymbol{\xi}$ should satisfy the conditions

$$\frac{\partial L}{\partial\mathbf{Q}^{-1}}=-\mathbf{Q}+\sum_i\alpha_i(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}=0\tag{7}$$

$$\frac{\partial L}{\partial\boldsymbol{\mu}}=-2\sum_i\alpha_i\mathbf{Q}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})=0\tag{8}$$

$$\frac{\partial L}{\partial R}=2R-2R\sum_i\alpha_i=0\tag{9}$$

$$\frac{\partial L}{\partial\xi_i}=\frac{1}{\nu m}-\alpha_i-\beta_i=0\tag{10}$$

From the equalities (7)–(10), one obtains

$$\mathbf{Q}=\sum_i\alpha_i(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\tag{11}$$

$$\boldsymbol{\mu}=\sum_i\alpha_i\mathbf{x}_i\tag{12}$$