Pattern Recognition Letters 30 (2009) 1236–1240

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Least squares one-class support vector machine

Young-Sik Choi *

Department of Computer Engineering at Korea Aerospace University, Goyang City, Gyeonggi Province 412-791, Republic of Korea

ARTICLE INFO

Article history:
Received 29 September 2008
Received in revised form 13 April 2009
Available online 18 May 2009

Communicated by R.P.W. Duin

Keywords:
LS (least squares) one-class SVM
Proximity measure
Relevance ranking
One-class SVM (support vector machine)

ABSTRACT

In this paper, we reformulate a standard one-class SVM (support vector machine) and derive a least squares version of the method, which we call LS (least squares) one-class SVM. The LS one-class SVM extracts a hyperplane as an optimal description of training objects in a regularized least squares sense. One can use the distance to the hyperplane as a proximity measure to determine which objects resemble training objects better than others. This differs from the standard one-class SVMs that detect which objects resemble training objects. We demonstrate the performance of the LS one-class SVM on relevance ranking with positive examples, and also present the comparison with traditional methods including the standard one-class SVM. The experimental results indicate the efficacy of the LS one-class SVM.

1. Introduction

One-class classification problem is to make a description of a set of training objects and to detect which objects resemble this training set (Tax and Duin, 2004). Approach to this problem taken by the standard one-class SVMs (support vector machines) is extract the regions where a certain fraction of training objects may locate, and classify an object according to whether the object resides inside the region or not. There are two standard algorithms in the literature (Müller et al., 2001; Schölkopf et al., 2001; Tax and Duin, 1999, 2004), which are equivalent for a certain type of kernel functions such as Gaussian kernel function (Müller et al., 2001). One of the standard one-class SVMs is estimate the sphere of minimum volume which encloses a given fraction of training objects (Tax and Duin, 1999, 2004). The other is extract a hyperplane in a kernel feature space such that a given fraction of training objects may reside beyond the hyperplane, while at the same time the hyperplane has maximal distance to the origin (Schölkopf et al., 2001). These standard one-class SVMs have been successfully applied for novelty detection (Deng and Xu, 2007; Ma and Perkins, 2003; Tax and Duin, 2004).

In this paper, we reformulate the standard one-class SVM in (Schölkopf et al., 2001) and derive a least squares version of the method, which is called the LS (least squares) one-class SVM. The LS one-class SVM uses a quadratic loss function and equality constraints, and extracts a hyperplane with respect to which the distances from training objects are minimized in a regularized least squares sense. This reformulation is very similar to the derivation of the LS SVM from the standard the SVM classifier (Suykens and Vandewalle, 1999; Suykens et al., 2002), in that both LS approaches use the quadratic loss functions. Hence, the proposed LS one-class SVM also loses the sparseness property of the standard one-class SVMs. One may overcome the loss of the sparseness by pruning training samples (Kruif and Vries, 2003; Kuh and De Wilde, 2007).

The hyperplane obtained from the LS one-class SVM is not the boundary of regions as in the standard one-class SVMs. Instead, it represents a hyperplane which most of training objects may lie close to. One can use the distance to the hyperplane as a proximity measure to determine which objects resemble training objects better than others. In this paper, we apply the LS one-class SVM for relevance ranking with positive examples. In the ranking problem, one should rank all documents according to the proximity to the set of training documents. This is important in modern information retrieval problems (Chakrabarti et al., 1999; Chen et al., 2001; Manevitz and Yousef, 2001; Setia et al., 2005).

There have been several attempts to use the distance from the center of sphere obtained from the standard one-class SVM (Tax and Duin, 2004) as a proximity measure to the training set (Chen et al., 2001; Manevitz and Yousef, 2001). Despite of the usefulness of these approaches, the distance to the center of sphere does not necessarily reflect the proximity to the training set. For instance, an object closer to the center might be farther from training objects. This is because in the standard one-class SVM, the training objects inside the regions may not contribute to the construction of the regions. On the other hand, the LS one-class SVM seeks to minimize the sum of distances from all training objects to the hyperplane in a regularized least squares fashion and thus most of training objects may lie close to the hyperplane. Therefore, the proximity to such hyperplane can better reflect the proximity to the training set.

* Tel.: +82 2 300 0189; fax: +82 2 3158 1419.
E-mail address: choimail@kau.ac.kr

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.05.007

There are several research works (Suykens et al., 2003; Roth, 2004) related to the proposed LS one-class SVM. Suykens et al. (2003) showed that the kernel PCA (principal component analysis) can be interpreted as a one-class modeling problem with zero target value. Roth has presented a one-class kernel Fisher discriminant which is a kernel ridge regression (Saunders et al., 1998) with a single target value with assumption of Gaussian distribution of data samples. The LS one-class SVM can be also regarded as a kind of kernel ridge regression with a single target value. Unlike Roth's approach, the proposed LS one-class SVM does not make any assumption on the distribution of data samples. Thus the LS one-class SVM can provide more flexibility on the handling of one-class problems.

In Section 2, we briefly introduce the standard one-class SVMs, and present the proposed LS one-class SVM. In Section 3, we discuss the differences between the LS and the standard one-class SVMs. Section 4 presents experimental results with several collections of Web pages, comparing the standard one-class SVMs. We make conclusions in Section 5.

2. Least squares one-class support vector machines

In this section, we briefly introduce the standard one-class SVMs and then present the LS one-class SVM. First, we define a mapping function to be used in the following description. Suppose that we are given an input data set S containing n points $\{\boldsymbol{x}_j: j = 1,\ldots,n\}$, where $S \subseteq \boldsymbol{X}$ and $X \subseteq \mathrm{R}^d$. Then, we define a map ϕ: $X \to \boldsymbol{F}$ to be the mapping of $\boldsymbol{X}$ into feature space $\boldsymbol{F}$ such that a dot product in feature space can be computed by a kernel function, i.e. $\phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

2.1. Standard one-class support vector machines

The standard one-class SVM (Schölkopf et al., 2001) can be stated as the following objective function to be minimized:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 - \rho + C\sum_j \xi_j, \tag{1}$$

subject to $\boldsymbol{w} \cdot \phi(\boldsymbol{x}_j) \geqslant \rho - \xi_j$ and $\xi_j \geqslant 0$. Here, $\boldsymbol{w} \cdot \phi(\boldsymbol{x}) = \rho$ represents a hyperplane in feature space, $\|\cdot\|$ denotes Euclidean norm, and ξ_j slack variables. The parameter C is predefined and controls the fraction of outliers (Müller et al., 2001; Schölkopf et al., 2001).

Eq. (1) seeks to extract a hyperplane which has the maximal distance $\rho/\|\boldsymbol{w}\|^2$ from the origin and beyond which most of training examples may reside. The hyperplane can be obtained by solving the following dual objective function to be maximized:

$$-\sum_{i,j} \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{2}$$

subject to $0 \leqslant \alpha_j \leqslant C$ and $\sum_j \alpha_j = 1$, where α_j denotes Lagrangian multiplier. The obtained hyperplane $f(\boldsymbol{x})$ can be written as

$$f(\boldsymbol{x}) = \sum_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) - \rho. \tag{3}$$

One can determine the values of α_j using the traditional quadratic programming with a linear constraint. The bias term ρ can be also obtained from $f(\boldsymbol{x}_s) = 0$, where $\boldsymbol{x}_s$ denotes one of the support vectors obtained. The decision function $g(\boldsymbol{x})$ for one-class classification is simply to take the sign of $f(\boldsymbol{x})$ as follows.

$$g(\boldsymbol{x}) = \mathrm{sgn}(f(\boldsymbol{x})) = \mathrm{sgn}\left(\sum_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) - \rho\right). \tag{4}$$

Another standard one-class SVM (Vapnik, 1998; Tax and Duin, 1999, 2004), which is also called support vector data description, can be formulated as the following objective function to be minimized.

$$R^2 + C\sum_j \xi_j, \tag{5}$$

subject to $\|\phi(\boldsymbol{x}_j) - \boldsymbol{a}\|^2 \leqslant R^2 + \xi_j$ and $\xi_j \geqslant 0$ for all $\boldsymbol{x}_j$, where vector $\boldsymbol{a}$ denotes the center of the sphere.

Eq. (5) seeks to extract the sphere of the minimum radius R enclosing the fraction of training objects. One can obtain the sphere by solving the following dual objective function to be maximized.

$$\sum_j \alpha_j K(\boldsymbol{x}_j, \boldsymbol{x}_j) - \sum_{i,j} \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{6}$$

with $0 \leqslant \alpha_j \leqslant C$ and $\sum_j \alpha_j = 1$, where α_j denotes Lagrangian multiplier. Note that (6) is equivalent to (2) for the kernel functions satisfying with $K(\boldsymbol{x}_j, \boldsymbol{x}_j) = 1$. The obtained center of the sphere can be written as follows:

$$\boldsymbol{a} = \sum_j \alpha_j \phi(\boldsymbol{x}_j). \tag{7}$$

The values of α_j can be determined using the quadratic programming, and the value of R^2 can be computed from $\|\phi(\boldsymbol{x}_s) - \boldsymbol{a}\|^2 = R^2$, where $\boldsymbol{x}_s$ denotes one of the support vectors. The decision function for one-class classification simply becomes

$$g(\boldsymbol{x}) = \mathrm{sgn}(R^2 - \|\phi(\boldsymbol{x}) - \boldsymbol{a}\|^2). \tag{8}$$

2.2. Least squares one-class support vector machine

To derive a LS (least squares) version of the standard one-class SVM, we reformulate the one-class SVM described in (1) by using a quadratic error function and the equality conditions. The corresponding LS one-class SVM can be written as the following objective function to be minimized:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 - \rho + \frac{1}{2}C\sum_j \xi_j^2, \tag{9}$$

subject to $\boldsymbol{w} \cdot \phi(\boldsymbol{x}_j) = \rho - \xi_j$. Now, the conditions for the slack variables, $\xi_j \geqslant 0$ in (1) no longer hold. Instead, the variable ξ_j represents an error caused by a training object $\boldsymbol{x}_j$ with respect to the hyperplane, i.e. $\xi_j = \rho - \boldsymbol{w} \cdot \phi(\boldsymbol{x}_j)$.

The LS one-class SVM described in (9) seeks to extract a hyperplane which has the maximal distance $\rho/\|\boldsymbol{w}\|^2$ from the origin, and with respect to which the sum of the squares of errors, ξ_j^2 are minimized. One can solve the problem in (9) as follows. By introducing Lagrangian multipliers α_j, the corresponding objective function can be written as the following.

$$L = \frac{\|\boldsymbol{w}\|^2}{2} - \rho + \frac{C}{2}\sum_j \xi_j^2 - \sum_j \alpha_j(\phi(\boldsymbol{x}) \cdot \boldsymbol{w} + \xi_j - \rho). \tag{10}$$

Setting to zero the first derivatives of (10) with respect to $\boldsymbol{w}$, ξ_j, ρ, and α_j leads to the following relations:

$$\begin{aligned}
&\frac{\partial L}{\partial \boldsymbol{w}} = 0 \to \boldsymbol{w} = \sum_j \alpha_j \phi(\boldsymbol{x}_j),\\
&\frac{\partial L}{\partial \xi_j} = 0 \to \xi_j = \alpha_j / C,\\
&\frac{\partial L}{\partial \rho} = 0 \to \sum_{j=1} \alpha_j = 1,\\
&\frac{\partial L}{\partial \alpha_j} = 0 \to \phi(\boldsymbol{x}_j) \cdot \boldsymbol{w} + \xi_j - \rho = 0.
\end{aligned} \tag{11}$$

Eliminating $\boldsymbol{w}$ and ξ_j through substitution in (11) yields