Pattern Recognition Letters 85 (2017) 15–20

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Ramp Loss based robust one-class SVM

Yingchao Xiao[a,b], Huangang Wang[a,c,*], Wenli Xu[a]

[a] Department of Automation, Tsinghua University, Beijing, 100084, China
[b] State Key Laboratory of Air Traffic Management System and Technology, Nanjing, 210014, China
[c] National Engineering Laboratory for E-Commerce Technologies, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Article history:
Received 1 June 2016
Available online 27 November 2016

Keywords:
Ramp Loss function
Outliers
One-class SVM
One-class classification

ABSTRACT

One-class SVM (OCSVM) is widely adopted in one-class classification (OCC) fields. However, outliers in the training set negatively influence the classification surface of OCSVM, degrading its performance. To solve this problem, a novel method is proposed in this paper. This proposed method introduces Ramp Loss function into OCSVM optimization, so as to reduce outliers' influence. Then the outliers are identified and removed from the training set. The final classification surface is obtained on the remaining training samples. Various experiments verify the effectiveness of this proposed method.

1. Introduction

In many practical manufacturing processes, it is convenient to obtain plenty of target data in the normal operating condition, whereas the non-target data in the abnormal operating condition is too difficult to characterize or too expensive to obtain. In this case, general classification methods are no longer suitable. To achieve better classification performance, one-class classification (OCC) methods can be adopted to describe the target data, which is usually the case in anomaly detection [6,18,24] and novelty detection [5,7].

Among the OCC methods, one-class SVM (OCSVM) avoids the estimation of the distribution density of the target class, is able to deal with nonlinear data, and inherits the sparseness from SVM. Therefore it is widely used in OCC application fields [12,14,19,23]. However, there are some drawbacks in OCSVM. Because some training samples are allowed to be located outside the classification surface, outliers are apt to become support vectors (SVs), negatively influencing the surface.

In order to reduce the outliers' influence on OCC method, researchers have tried to remove the outliers by data preprocessing methods. Tax and Duin [25] propose to use the distances of a sample to its k nearest neighbors to detect and remove outliers. Breunig et al. [4] try to assign to each sample in the data set a degree of being an outlier by estimating its local density, and the degree is called LOF. Zheng et al. [30] use the LOF to filter the raw samples to remove outliers. Khan et al. [16] and Andreou and Karathanassi [2] calculate the inter quartile range (IQR) of the training samples, and it provides a means to indicate the boundary beyond which the samples will be labeled as outliers and be removed. However, due to the diversity of sample distribution, it is not easy to remove outliers through the data preprocessing.

In OCSVM, researchers have tried to modify the OCSVM method to improve its robustness to outliers. Yin et al. [28] propose to weigh training samples according to their distances to the sample center in the feature space, hoping to alleviate the penalty on outliers, so as to make them more likely to be located outside the classification surface. This method is hereafter referred to as weight OCSVM. However, due to the absence of prior knowledge of outlier distribution, it is difficult to weigh samples properly, and thus the results of this method are not satisfactory. Different from the above method, which contains two steps: weighing samples and training OCSVM, Amer et al. [1] propose to identify outliers while optimizing the OCSVM hyper-plane. This method is hereafter referred to as eta OCSVM. They modified the OCSVM optimization object by introducing 0-1 variables η_i, which indicate whether $\mathbf{x}_i$ is an outlier. The introduction of these discrete variables makes it more difficult to solve the optimization problem. They relax this prime optimization problem and thus put forward an iterative algorithm. In each iteration, the samples with $\eta_i = 1$ are constrained to be located above the hyper-plane, and those with $\eta_i = 0$ are excluded from this training stage. If the η_i does not indicate outliers correctly in some iteration, then the result of this iteration will be negatively influenced by outliers, and the subsequent results will also be influenced. Therefore, this method does not perform as expected.

In this paper, we propose a novel method to reduce the influence of outliers. First, a surface enclosing the cluster core [3] of the target sample distribution is learned from the outlier

[*] Corresponding author. Fax: +86 10 62786911.
E-mail addresses: xiaoyc1016@163.com (Y. Xiao), hgwang@tsinghua.edu.cn (H. Wang), xuwl@tsinghua.edu.cn (W. Xu).

http://dx.doi.org/10.1016/j.patrec.2016.11.016

contaminated training sample set. The cluster core refers to the samples located at the core of the target sample cluster, i.e. the more representative target samples. This part is done by the proposed method named “Ramp Loss OCSVM”, where the Ramp Loss function is used to replace the Hinge Loss function to avoid outliers from becoming support vectors, such that the surface enclosing cluster core can be obtained. Next, this surface is used to identify outliers and remove them from the training set, so as to train the final OCSVM. Combining the above two parts, the whole algorithm is formed, ROCSVM.

The remainder of this paper is arranged as follows: the second section states the basic idea of the proposed method; the third section proposes the Ramp Loss OCSVM problem; the fourth section gives the algorithm of the proposed method; the fifth section compares the proposed method with other relevant methods on outlier contaminated data sets; the sixth section concludes this paper.

2. The basic idea of the proposed method

We first review conventional OCSVM briefly. The basic idea of OCSVM is to find a hyper-plane $\langle \mathbf{w}, \varphi(\mathbf{x}) \rangle - \rho = 0$ in the feature space that separates sample images from the origin with maximum margin. The primal optimization problem is as follows [22],

$$\begin{array}{ll} \min\limits_{\mathbf{w},\xi,\rho} & \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\upsilon n}\sum\limits_{i=1}^{n}\xi_i \\ \text{s.t.} & \langle \mathbf{w}, \varphi(\mathbf{x}_i)\rangle \geq \rho - \xi_i, \ \ \xi_i \geq 0. \end{array} \tag{1}$$

where $\mathbf{x}_i$ denote training samples, n is the total number of training samples, υ is a trade-off parameter and ξ_i are slack variables. This problem is a convex optimization and can be solved by its dual problem, shown in Eq. (2).

$$\begin{array}{ll} \max\limits_{\boldsymbol{\alpha}} & -\frac{1}{2}\sum\limits_{i,j=1}^{n}\alpha_i\alpha_j k\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ \text{s.t.} & 0 \leq \alpha_i \leq \frac{1}{\upsilon n}, \\ & \sum_i \alpha_i = 1 \end{array} \tag{2}$$

where $k(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function. Gaussian kernel is adopted in this paper, $k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2/s\right)$, where s is the width parameter.

If the image of a sample $\mathbf{x}_B$ is below the hyper-plane, i.e. the sample is outside the classification surface, then its $\xi_B > 0$. According to the KKT condition, $\beta_B = 0$ and thus $\alpha_B = 1/n\upsilon$. This means, samples outside the surface become SVs in conventional OCSVM. It is known that the normal vector $\mathbf{w}$ can be expressed in terms of the mappings of SVs, $\mathbf{w} = \sum_i \alpha_i \varphi(\mathbf{x}_i)$. If υ is set large to make outliers located outside the surface, the outliers become SVs and their α_i's are $1/n\upsilon$, so they can influence the normal vector and the surface with more magnitude; if υ is set small, outliers are enclosed by the surface, so the surface is influenced by outliers. Therefore, although υ is used to reduce outliers' influence in conventional OCSVM, its ability is limited.

The method proposed in this paper contains two parts. The first part is to obtain the surface enclosing the target cluster core; the second one is to identify outliers by this surface and to exclude them from the training set, before training the final OCSVM model. To better enclose the cluster core of the target class, it is expected that the surface should locate the outliers outside, and more importantly, should not be influenced by these outside-located outliers. The first expectation can be realized by tuning υ. The fraction of outliers in the OCC training set is usually small, and it is reasonable to give it a small prior estimate $R_{outlier}$ (e.g. $R_{outlier} = 2\%, 5\%$). Because υ is the upper bound of training samples located outside the classification surface [22], more outliers are likely to be located outside under a large υ (e.g. $\upsilon = 3R_{outlier}$). To achieve the second expectation, we propose the Ramp Loss OCSVM, which will be discussed in the next section.

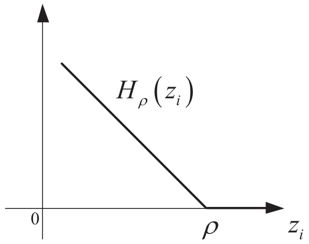

Fig. 1. Hinge Loss function. The x-axis stands for z_i and the y-axis stands for the value of $H_\rho(z_i)$.

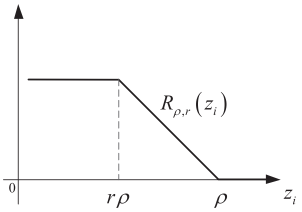

Fig. 2. Ramp Loss function. The x-axis stands for z_i and the y-axis stands for the value of $R_{\rho,\,r}(z_i)$.

3. Ramp Loss OCSVM

Similar to OCSVM, the binary-class and multi-class SVM are also influenced by the outliers in training sets [21,26]. To solve this problem for SVM, researchers propose to use Ramp Loss function [8,10,15,27]. Inspired by this, we propose to introduce Ramp Loss function into OCSVM, attempting to reduce outliers' negative influence.

To facilitate the subsequent discussion, the conventional OCSVM is rewritten as follows.

$$\min_{\mathbf{w},\rho} J(\mathbf{w}, \rho) = \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{n\upsilon}\sum_{i=1}^{n} H_\rho(z_i) \tag{3}$$

where $z_i = \langle \mathbf{w}, \varphi(\mathbf{x}_i)\rangle$. The slack variable ξ_i and its corresponding constraints $\langle \mathbf{w}, \varphi(\mathbf{x}_i)\rangle \geq \rho - \xi_i$, $\xi_i \geq 0$ in Eq. (1) are integrated into one function $H_\rho(z_i)$ in Eq. (3). This function is the Hinge Loss function, and as shown in Fig. 1, its formula is $H_\rho(z_i) = \max(0, \rho - z_i)$. Specifically, for samples satisfying $z_i = \langle \mathbf{w}, \varphi(\mathbf{x}_i)\rangle \geq \rho$, they are located above the OCSVM hyper-plane and no penalty is inflicted, so $H_\rho(z_i) = 0$; for samples satisfying $z_i = \langle \mathbf{w}, \varphi(\mathbf{x}_i)\rangle < \rho$, they are located below the OCSVM hyper-plane and some penalty should be inflicted, so $H_\rho(z_i) = \rho - z_i > 0$, which means the penalty increases as the samples move far away from the hyper-plane.

Based on the above analysis, the drawback of OCSVM comes from the Hinge Loss, because it does not bound the influence of outliers effectively [10]. To overcome this drawback, the influence of outliers should be bounded. Therefore, the Ramp Loss is adopted in this paper and its formula is as follows.

$$R_{\rho,r}(z_i) = \begin{cases} 0, & z_i \geq \rho \\ \rho - z_i, & r\rho < z_i < \rho \\ \rho - r\rho, & z_i \leq r\rho \end{cases} \tag{4}$$

where $0 < r < 1$. As shown in Fig. 2, the influence of outliers is bounded by the Ramp Loss. When $z_i > r\rho$, the Ramp Loss is identical to the Hinge Loss; when $z_i \leq r\rho$, the value of Ramp Loss is a constant, different from the Hinge Loss whose value increases as z_i decreases. Therefore, it is reasonable to use Ramp Loss to overcome the drawback of conventional OCSVM caused by Hinge Loss.