Knowledge-Based Systems 90 (2015) 129–137

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Robust support vector data description for outlier detection with noise or uncertain data

Guijun Chen [a], Xueying Zhang [a,*], Zizhong John Wang [a,b], Fenglian Li [a]

[a] College of Information Engineering, Taiyuan University of Technology, Taiyuan, Shanxi, China
[b] Department of Mathematics and Computer Science, Virginia Wesleyan College, Norfolk, VA, USA

ARTICLE INFO

Article history:
Received 11 July 2015
Revised 23 September 2015
Accepted 25 September 2015
Available online 9 October 2015

Keywords:
Outlier detection
Support vector data description
Local density
ε-insensitive loss

ABSTRACT

As an example of one-class classification methods, support vector data description (SVDD) offers an opportunity to improve the performance of outlier detection and reduce the loss caused by outlier occurrence in many real-world applications. However, due to limited outliers, the SVDD model is built only by using the normal data. In this situation, SVDD may easily lead to over fitting when the normal data contain noise or uncertainty. This paper presents two types of new SVDD methods, named R-SVDD and εNR-SVDD, which are constructed by introducing cutoff distance-based local density of each data sample and the ε-insensitive loss function with negative samples. We have demonstrated that the proposed methods can improve the robustness of SVDD for data with noise or uncertainty by extensive experiments on ten UCI datasets. The experimental results have shown that the proposed εNR-SVDD is superior to other existing outlier detection methods in terms of the detection rate and the false alarm rate. Meanwhile, the proposed R-SVDD can also achieve a better outlier detection performance with only normal data. Finally, the proposed methods are successfully used to detect the image-based conveyor belt fault.

1. Introduction

Detecting outliers from the available data has been an important task in many diverse applications, such as fault detection, reliability analysis, disease diagnosis, hazard prediction, etc. [1,2]. The goal of outlier detection is to find the abnormal data with inconsistent characteristics that are generated by a different mechanism. In practice, the abnormal data are expensive to obtain or even not available at all, for instance, possible defect features in the fault detection or non-healthy data in the medical diagnosis; however, we can usually acquire a large number of normal ones. Consequently, one-class classification (OCC) has attracted much attention in such situations [3–5], which allows for describing the situation of positive (normal) data and identifies the negative data as outliers.

Support vector data description (SVDD) is one of widely used OCC methods [6,7]. It is capable to build a flexible description boundary in the high-dimensional feature space by kernel trick [8,9]. The constructed boundary tends to enclose most of normal data in the hyper-sphere and simultaneously minimize the chance of accepting outliers. The outliers can be distinguished from normal data in the following way: the data within the hyper-sphere are considered as normal, while the data outside the hyper-sphere are outliers. Another advantage to use SVDD is the detection strategy without need of any prior knowledge about the detected object [10,11].

Depending on the kernel-based distance between the hyper-sphere and the training data, SVDD may easily lead to over fitting when the training data contain noise or uncertainty [12]. The noise data may behave like normal, and be enclosed inside the hyper-sphere in the training processes [13]. In this case, the spherical boundary may not be optimal and the detection performance will become deteriorated, especially for some applied sensor data with sampling errors and transmission noise. Thus, it is necessary to develop a robust SVDD method to deal with noise or uncertain data [14].

Earlier studies concentrated on adopting some distribution characteristics of the target data in the training phase of SVDD [15]. Lee et al. [16] proposed the density-induced SVDD by introducing new distance measurements based on the nearest neighborhood and Parzen-window approaches, which reflected the relative density degree for each data point. Furthermore, the kernel-based class center method was used to generate the confidence level [17] and the position-based weighting [18] respectively. Both parameters indicated the likelihood of input data belonging to the normal. Besides, there were some other methods to calculate the likelihood value, i.e. the *k*-nearest neighbor (k-NN) method [19], the kernel *k*-means clustering and kernel LOF-based method [20]. Though so much progress has been made

* Corresponding author. Tel.: +863516014864.
E-mail address: xyztut@163.com (X. Zhang).

http://dx.doi.org/10.1016/j.knosys.2015.09.025
0950-7051/

to improve the detection performance of SVDD, most of the studies equally reflect the distribution characteristics of all data. From our observation on the training process of SVDD, taking the two dimensional space for example, the data points located on the boundary of the real space would usually be the support vectors (SVs) (i.e. boundary points of the feature space), and they have a great impact on the performance of data description. In addition, most of the above mentioned distribution characteristics were calculated by using kernel-based distance, which would be directly affected by selected kernel parameters [21].

On the other hand, most of contemporary SVDD algorithms are used only with normal training data, which is similar to the unsupervised learning. But the abnormal data do exist even though the number is small. The abnormal data could refine the description boundary of SVDD if they are used in the training processes, such as SVDD with negative examples (N-SVDD) [7]. However, the margin between the normal data and the abnormal data is zero in N-SVDD, which would result in poor generalization ability. To overcome this drawback, the margin between the hyper-sphere and the abnormal (negative) data was maximized to restructure the optimization problems in [23,24]. In [12], the rough SVDD including a lower hyper-sphere and an upper hyper-sphere was constructed by using the rough set principle. Due to the class imbalance problem between two-class data, these improved methods are subject to the hyper-sphere shift and the classification deviation.

In this paper, a robust SVDD is proposed with the introduction of a cutoff distance-based local density for each data point [22], which is used as the penalty weight of input data towards the noise data. Moreover, the cutoff distance-based density can effectively indicate the characteristics of those boundary data with noise. Furthermore, we investigate the margin between normal data and abnormal data, and find out that a ring-shaped band containing the inseparable data would be formed around the margin. Inspired by the ε-SVR [25], we construct another robust SVDD model with abnormal data by adding two ε bands (i.e. ε-insensitive loss) on both sides of the description boundary. The two developed models are named R-SVDD and εNR-SVDD respectively. In order to assess their detection performance, ten benchmark datasets from UCI are used for experiments. The proposed methods are also applied to detect image-based conveyor belt fault.

Compared with the previous work on the robustness improvement of SVDD, the main contribution of our work can be indicated as follows. First, the cutoff distance-based local density is introduced that can mitigate the effect of noise towards SVDD, especially that of the boundary noise. Second, the ε-insensitive loss is used to refine the description boundary combing with the limited abnormal data, which can improve generalization performance and avoid the hyper-sphere shift. Finally, incorporating above two strategies to the SVDD optimization framework, two robust SVDD models are built to detect outliers.

This paper is organized as follows. In Section 2, we briefly introduce the original SVDD and the cutoff distance-based local density. Section 3 presents our proposed methods and the theoretical analyses related to the methods. Section 4 demonstrates the empirical study about the robustness to noise, including UCI datasets and image-based conveyor belt fault dataset. Finally, the conclusion and further study are drawn in Section 5.

2. Fundamentals

2.1. Support vector data description

As a one-class classification method, the goal of support vector data description (SVDD) is to find the minimum hyper-sphere that can enclose most of normal (target) data in the feature space. Given the target dataset $X = \{x_1, x_2, \ldots, x_l\}$, where $x_i \in R^n$, the optimization problem is constructed as Eq. (1).

$$\begin{aligned} \min_{R,a,\xi} \quad & R^2 + C\sum_{i=1}^{l}\xi_i \\ s.t. \quad & \|\phi(x_i) - a\|^2 \le R^2 + \xi_i, \\ & \xi_i \ge 0,\ i = 1, 2 \ldots, l \end{aligned} \tag{1}$$

where R and a are the radius and center of the hyper-sphere respectively in the feature space, ξ_i is the error term to allow the data point x_i to locate outside the hyper-sphere, $C>0$ is the penalty parameter of ξ_i, and $\phi(.)$ is the mapping function that makes point x_i mapped onto a high-dimensional feature space. We can obtain Eq. (2) by solving the Lagrange dual problem. The resolving process of the dual problem can be derived from [7] in details.

$$\begin{aligned} \min_{\alpha} \quad & \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\left(\phi(x_i)\cdot\phi\left(x_j\right)\right) - \sum_{i=1}^{l}\alpha_i(\phi(x_i)\cdot\phi(x_i)) \\ s.t. \quad & \sum_{i=1}^{l}\alpha_i = 1, \\ & 0 \le \alpha_i \le C,\ i = 1, 2 \ldots, l \end{aligned} \tag{2}$$

where $\alpha_i > 0$ is the Lagrange multiplier. Generally, $K(x_i, x_j) = (\phi(x_i)\cdot\phi(x_j))$ is defined as the kernel function. Because the Gaussian radial basis function (RBF) can approximate most kernel functions if the kernel parameter is chosen appropriately [26], the Gaussian RBF kernel: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\sigma^2)$ is adopted in this paper. The data with $\alpha_i > 0$ constitute the support vectors (SVs). And then we can obtain the following:

$$a = \sum_{i=1}^{l}\alpha_i\phi(x_i) \tag{3}$$

$$R^2 = \frac{1}{|SVs|}\sum_{x_i\in SVs}\|\phi(x_i) - a\|^2 \tag{4}$$

To test an object x, the decision function $f(x)$ is defined as Eq. (5).

$$f(x) = \|\phi(x) - a\|^2 - R^2 \tag{5}$$

When $f(x)\le 0$, x is classified as normal; otherwise, it is classified as an outlier.

As shown in Fig. 1, 50 blue star-shaped points are generated randomly with a banana shape in the two-dimensional space. The black dot 51 represents the outlier. The green dashed line is the data description boundary of SVDD. Under normal condition, the outlier can be detected by the description boundary as shown in Fig. 1(a). However, when the normal data are corrupted by noise, such as the triangular point 52 and 53, the outlier is misclassified as normal data shown in Fig. 1(b). Although decreasing the C value can reduce the interference of boundary noise, the likelihood of each data point to be an outlier is taken the same by using the same parameter C for all points. It makes most normal boundary points excluded from the data description region. Therefore, it is important to make full use of the distribution characteristic of each point, especially the boundary points. Herein, a robust modified strategy combining with the cutoff distance-based local density has been proposed to mitigate the effect of individual noise point towards SVDD.

2.2. The cutoff distance-based local density

The cutoff distance-based local density was proposed to make cluster analysis in [22]. This quantity depends only on the distances d_{ij} between data i and data j, which indicates the number of points within the range of a cutoff distance. The local density ρ_i of data point i is defined as Eq. (6).

$$\rho_i = \sum_j \chi\left(d_{ij} - d_c\right) \tag{6}$$