Expert Systems with Applications 41 (2014) 3343–3350

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Density weighted support vector data description

Myungraee Cha, Jun Seok Kim, Jun-Geol Baek *

School of Industrial Management Engineering, Korea University, Anam-dong, Seongbuk-gu, 136-701 Seoul, Republic of Korea

ARTICLE INFO

Keywords:
One-class classification (OCC)
Support vector data description (SVDD)
Density weighted SVDD (DW-SVDD)
k-Nearest neighbor approach

ABSTRACT

One-class classification (OCC) has received a lot of attention because of its usefulness in the absence of statistically-representative non-target data. In this situation, the objective of OCC is to find the optimal description of the target data in order to better identify outlier or non-target data. An example of OCC, support vector data description (SVDD) is widely used for its flexible description boundaries without the need to make assumptions regarding data distribution. By mapping the target dataset into high-dimensional space, SVDD finds the spherical description boundary for the target data. In this process, SVDD considers only the kernel-based distance between each data point and the spherical description, not the density distribution of the data. Therefore, it may happen that data points in high-density regions are not included in the description, decreasing classification performance. To solve this problem, we propose a new SVDD introducing the notion of density weight, which is the relative density of each data point based on the density distribution of the target data using the *k*-nearest neighbor (*k*-NN) approach. Incorporating the new weight into the search for an optimal description using SVDD, this new method prioritizes data points in high-density regions, and eventually the optimal description shifts to these regions. We demonstrate the improved performance of the new SVDD by using various datasets from the UCI repository.

1. Introduction

One-class classification (OCC) is a response to the data classification problem in which there is an absence of suitable negative cases that can be used for training. The subject of a great deal of past research, OCC thus aims to find the best description of a data set using only objects from one class, known as the target data. If the target data is described accurately, it can be used to classify other classes when there are insufficient non-target data (Tax & Duin, 2002); thus, OCC has attracted attention for use in exceptional situations where it is difficult to gather datasets for other classes or where no other classes exist (Khan & Madden, 2010; Mazhelis, 2006).

Support vector data description (SVDD) is a widely used example of OCC. The objective of SVDD is to find a set of support vectors (SVs) describing the spherical boundary of the target data by mapping it into high-dimensional feature space. Since the process occurs in feature space, SVDD has a flexible description boundary. SVDD has been developed from support vector machines as a way to compensate for weaknesses in previous OCC research. Before the use of support vectors, many classification methods were based on the estimation of the probability distribution of the target data set, and this produced severe limitations for data sets that did not follow a specific distribution (Tax & Duin, 2002). In contrast, SVDD is easily applicable to data generated in the real world with no assumptions regarding the data distribution (Grinblat, Uzal, & Granitto, 2013; Sjöstrand, Hansen, Larsson, & Larsen, 2007).

In addition to not requiring assumptions of data distribution, SVDD is also used in various fields for its flexible description boundaries. In terms of feature extraction methodology, it can be used to produce a representative set of target data for image retrieval (Lai, Tax, Duin, Pękalska, & Paclík, 2004), facial images (Lee, Park, & Lee, 2006), and pattern recognition (Dong, Zhaohui, & Wanfeng, 2001; Zhao, Wang, & Xiao, 2013). SVDD has also been used in outlier detection for image sensory devices (Bovolo, Camps-Valls, & Bruzzone, 2010; Guo, Chen, & Tsai, 2009), intrusion detection (Kang, Jeong, & Kong, 2012), and mura inspection of thin-film transistor liquid–crystal displays (TFT-LCDs) (Liu, Lin, Hsueh, & Lee, 2009; Liu, Liu, & Chen, 2011). With outliers recognized as fault in the process, it is possible to identify faults in the dataset using SVDD (Liu, Liu, & Chen, 2010; Luo, Cui, & Wang, 2011; Zhang, Liu, Xie, & Li, 2009).

However, even though conventional SVDD has advantages in data domain description, a major limitation exists. To decide the optimal description of target data, SVDD takes into account only the kernel-based distance between the spherical boundary and the data points, not the distribution of the data. When SVDD sets the description boundary without considering the density distribution of the data, it is possible that the boundary will pass through

* Corresponding author. Tel.: +82 2 3290 3396; fax: +82 2 929 5888.
E-mail address: jungeol@korea.ac.kr (J.-G. Baek).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.11.025

the highest density area. Hence, the algorithm could misjudge outliers and this weakness could decrease classification performance (Lee, Kim, Lee, & Lee, 2005).

A great deal of research has been conducted seeking to overcome this weakness by applying additional characteristics of the target dataset. Lee, Kim, Lee, and Lee (2007) offered density-induced SVDD by introducing relative density based on the nearest neighborhood and Parzon-window approaches to reflect the density distribution of a dataset (Lee et al., 2007). Based on the degree of density for each data point, they also proposed a new geometric distance strategy, the density-induced distance measure, for positive and negative data in the search for an optimal SVDD.

Liu, Xiao, Cao, Hao, and Deng (2013) also proposed a new SVDD by introducing a confidence score for target data, a score that indicates the likelihood of an example belonging to the normal class by using kernel-based distance. The data is mapped into high-dimensional feature space and the score decided based on the distance from the centroid of the dataset to each data point in feature space. By introducing this score when searching for the optimal spherical description, the description first includes data points with high confidence scores, which are those near the centroid of the dataset in feature space. The effect of confidence score is therefore to apply the characteristics of data distribution in feature space to find the optimal SVDD. However, when mapping datasets into high-dimensional space, the SVDD algorithm uses the kernel function so that the centroid of the dataset in feature space may not correspond with the centroid in real space. Therefore, a characterized description of the dataset in high-dimensional feature space may not produce an optimal SVDD.

To solve this problem, we introduce a density weight into the search for an optimal SVDD. Density weight reflects the density distribution of a dataset in real space using the k-nearest neighbor (k-NN) approach and each data point is assigned this weight according to the data density distribution. By applying the density weight of each data point into the search process, the description prioritizes data points in high-density regions. Eventually the optimal description shifts toward these dense regions. As a result, the introduction of density weight compensates for the weakness in the SVDD by reducing the risk of not including data points in high-density regions. We therefore propose a new SVDD named density weighted SVDD (DW-SVDD) and we examine the performance of this algorithm using UCI repository datasets.

The structure of the paper is organized as follows. In Section 2, we give an overview of conventional SVDD. Section 3 introduces the density weight method, and the steps for applying this to SVDD will be explained. In Section 4, experimental results for DW-SVDD will be presented, including a performance comparison with other algorithms using UCI repository datasets. Finally, Section 5 contains our concluding remarks and suggestions about future research opportunities.

2. Support vector data description

The objective of SVDD is to find the best data description of target data in OCC. Assume a data set $\{\mathbf{x}_i,\ i = 1, ..., l\}$ where l is the number of target data. The objective function of SVDD is as follows:

$$\begin{aligned} &\min R^2 + C\sum_{i=1}^{l}\xi_i \\ &\text{s.t. } \|\mathbf{x}_i - \mathbf{a}\|^2 \leqslant R^2 + \xi_i, \xi_i \geqslant 0 \quad \forall_i \end{aligned} \tag{1}$$

The basic idea of SVDD is to determine the data description as the smallest sphere containing all possible target data in feature space (Tax & Duin, 2004). Usually, some data points are allowed outside the sphere and are treated as outliers. As shown in Fig. 1, the variable ξ_i is used to incorporate the effect of data not included in the spherical description. R is the radius of the sphere used in SVDD.

The variable C represents the trade-off between sphere volume and the number of target data outside the sphere, allowing the relative importance of each term to be adjusted. Adding the slack variable ξ_i into the constraints allows for soft boundaries (Kang & Choi, 2008; Lee et al., 2006).

To solve the optimization problem with these constraints, we construct a Lagrangian function as follows:

$$\begin{aligned} L(R, \mathbf{a}, \alpha_i, \gamma_i, \xi_i) = R^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i\left\{R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\right\} \\ - \sum_{i=1}^{l}\gamma_i\xi_i \end{aligned} \tag{2}$$

where the Lagrange multipliers are $\alpha_i \geqslant 0$ and $\gamma_i \geqslant 0$. To find the stationary point of the Lagrange function, set partial derivatives to 0.

$$\frac{\partial L}{\partial R} = 0 : 2R - 2R\sum_{i=1}^{l}\alpha_i = 0 \qquad \therefore \sum_{i=1}^{l}\alpha_i = 1 \tag{3}$$

$$\frac{\partial L}{\partial \mathbf{a}} = 0 : 2\mathbf{a} - 2\sum_{i=1}^{l}\alpha_i\mathbf{x}_i = 0 \qquad \therefore \mathbf{a} = \sum_{i=1}^{l}\alpha_i\mathbf{x}_i \tag{4}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \qquad \therefore C - \alpha_i - \gamma_i = 0 \quad \forall_i \tag{5}$$

Adjusting to the new constraints (3)–(5), the rearranged function is as follows:

$$\begin{aligned} &\max \sum_{i=1}^{l}\alpha_i\langle x_i, x_i\rangle - \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\langle x_i, x_j\rangle \\ &\text{s.t. } 0 \leqslant \alpha_i \leqslant C \quad i = 1, 2, \ldots, l \\ &\sum_{i=1}^{l}\alpha_i = 1 \end{aligned} \tag{6}$$

As a result, we can calculate each value of α_i and the centroid of the optimal sphere is determined by the linear combination of any non-zero α_i known as support vectors (SVs) and according to these SVs, the best description is determined.

Because the problem is related to the inner products between vectors, it could be mitigated by replacing the inner products with a kernel function $K(x_i, x_i)$ that satisfies Mercer's theorem (Tax & Duin, 2004). Replacing inner products with the kernel function, the search for the optimal data description is equivalent to:

$$\begin{aligned} &\max \sum_{i=1}^{l}\alpha_i K(x_i, x_i) - \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i, x_j) \\ &\text{s.t. } 0 \leqslant \alpha_i \leqslant C \quad i = 1, 2, \ldots, l \\ &\sum_{i=1}^{l}\alpha_i = 1 \end{aligned} \tag{7}$$

Despite SVDD providing a flexible description boundary suitable for the dataset, there are some drawbacks inherent to the search for the description boundaries in feature space. In the process of extracting the best description of the dataset, SVDD decides whether each data point is located inside or outside the sphere in feature space. To make this decision, SVDD primarily considers the kernel-based distance, which represents how far the data point is from the sphere in high-dimensional space, expressed as the slack variable ξ_i. However, only using kernel-based distance is insufficient in describing every relevant characteristic of the dataset.