Expert Systems With Applications 85 (2017) 25–32

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Lightly trained support vector data description for novelty detection

Rekha A.G.[a,*], Mohammed Shahid Abdulla[b], Asharaf S.[b]

[a] *IT & Systems Area, Indian Institute of Management Kozhikode, 673570, Calicut, Kerala, India*
[b] *Indian Institute of Information Technology and Management-Kerala, 695581, Trivandrum, India*

ARTICLE INFO

Article history:
Received 3 December 2015
Revised 10 April 2017
Accepted 5 May 2017
Available online 5 May 2017

Keywords:
SVDD
Outlier detection
One-class classification
Scaling

ABSTRACT

Anomaly (or outlier) detection is well researched objective in data mining due to its importance and inherent challenges. An outlier could be the key discovery to be made from large datasets and the insights gathered from them could be of significance in a wide variety of domains like information security, business intelligence, clinical decision support, financial monitoring etc. Recently, Support Vector Data Description (SVDD) driven approaches are shown as having good predictive accuracy. This paper proposes a novel low-complexity anomaly detection algorithm based on Support Vector Data Description (SVDD). The proposed algorithm reduces the complexity by avoiding the calculation of Lagrange multipliers of an objective function, instead locates an approximate pre-image of the SVDD sphere's center, within the input space itself. The crux of the training algorithm is a gradient descent of the primal objective function using Simultaneous Perturbation Stochastic Approximation (SPSA). Experiments using datasets obtained from UCI machine learning repository have demonstrated that the accuracies of the proposed approach are comparable while the training time is much lesser than Classical SVDD.

1. Introduction

Outlier detection, the problem of finding patterns in data that do not conform to expected behavior has attracted lot of attention due to its applicability in a wide variety of domains. One-class classification problem is one of the classical problems in data analysis and has got its original application in outlier detection (Bishop, 1994; Ritter & Gallegos, 1997). The difference of one-class classification from conventional two-class or multi-class classification is that the information on only one of the classes (called the target class) will be available for training. One-class classification has been applied in various scenarios like text classification (Liu, Lee, Yu, & Li, 2002), medical analysis (Gardner, Krieger, Vachtsevanos, & Litt, 2006), machine fault detection (Shin, Eom, & Kim, 2005) etc. Moreover, it has also been applied to various business domains like financial credit scoring (Wang, Wang, & Lai, 2005) and supplier selection (Guo, Yuan, & Tian, 2009). One classical approach for One-Class Classification is Support Vector Data Description (SVDD). SVDD algorithm has been used in scenarios where single class information is available in high quality and resolution, and a few outliers exist. SVDD has also been applied to cases where the problem has to scale to a multi-class environment with information of other classes only gradually becoming available, e.g. in Munoz-Mari, Bruzzone, and Camps-Valls (2007).

SVDD was proposed by Tax and Duin (2004) to solve the original one-class classification problem. The basic idea is to construct a spherically shaped decision boundary that envelops most of the data of interest, with a smaller set of support vectors describing the boundary. This technique is first motivated without using the concept of support vectors.

Given a set of data points, $\mathrm{x_i}$: i = 1:N in the d-dimensional real (or input) space R^d, the objective is to minimize an objective function that depends on the radius R of a sphere and its center a.

$$\begin{aligned} &\mathrm{O(R, a, \xi) = R^2 + C\sum_i \xi_i} \\ &\text{s.t. } \|x_i - a\|^2 \le R^2 + \xi_\mathrm{i},\ \xi_\mathrm{i} \ge 0\ \forall \mathrm{I} \end{aligned} \tag{1}$$

Here the parameter C controls the trade-off between the volume and the errors while ξ_i are slack variables which make the classifier 'soft-margin', i.e. allow some possibility of outliers in the training set. Object z is accepted by the description (i.e. z is within the (a, R) sphere) when the Euclidean distance is s.t. : $\|\mathrm{(z-a)^2}\| \le \mathrm{R^2}$

1.1. *Dual & primal SVDD*

Normally, for computational convenience and adaptation to the 'Kernel Trick', (1) is solved in the dual space by introducing its Lagrangian function. A description is given in Tax and Duin (2004) as also (4) below. For now, we assume we have the Lagrangian mul-

[*] Corresponding author.
E-mail addresses: agrekha64@gmail.com (R. A.G.), shahid@iimk.ac.in (M.S. Abdulla), asharaf.s@iiitmk.ac.in (A. S.).

http://dx.doi.org/10.1016/j.eswa.2017.05.007
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

tipliers α_i's corresponding to each pattern x_i. The x_i which have an associated $\alpha_i > 0$ are called support vectors (*SVs*). In particular, the *SVs* with $0 < \alpha_i < C$ are called unbounded *SVs* and the *SVs* with $\alpha_i = C$ as the bounded *SVs*. In all calculations within the dual formulation, patterns x_i appear only in the form of inner products with other patterns (x_i. x_j). These inner products can be replaced by a kernel function *K* to obtain more flexible methods. This kernel function *K*is analogous to inner product in a possibly infinite dimensional hyper-space, and represents the 'kernel trick' of Classical SVDD (C-SVDD). The centre of the minimum enclosing ball a_F and the radius R are represented as

$$a_F = \sum_{i=1}^{N_s} \alpha_i \phi(x_i)$$
$$R^2 = 1 - 2\sum_{x_i \in SVs} \alpha_i K(x_i, x_k) + \sum_{x_i \in SVs} \sum_{x_j \in SVs} \alpha_i \alpha_j K\left(x_{i,} x_j\right)$$

of these the latter quantity is calculable due to K being known. The former quantity is not needed explicitly, the decision function for checking a pattern *z* now becomes:

$$1 - 2\sum_{i} \alpha_i K(z, x_i) + \sum_{i,j} \alpha_j \alpha_j K\left(x_i, x_j\right) \le R^2$$

Thus the testing time complexity for C-SVDD is linear in the number of support vectors. However, solving the dual optimization problem (that yields the lagrange multipliers) is also of high-complexity, typically $O(N^3)$. Studies suggest that primal optimization will be superior for large scale optimization (Chapelle, 2007), due to the observation that when the number of training points N is large, the number of support vectors will also likely be large, and this results in updates of nearly N lagrange multiplier parameters during optimization and a complicated decision function during the testing of the algorithm. Hence it is advisable to directly minimize the primal objective function. We give an reference for this below:

While solving the SVDD problem, (Pauwels & Ambekar, 2011) proposes solving an unconstrained optimization problem in the primal:

$$\textit{Minimize} \qquad O'_p(a, R) = R^2 + C\sum_{i=1}^{N} \left(d_i^{\,2} - R^2\right)_+ \tag{2}$$

where $d_i = \|x_i - a\|$ and
$(.)_+$ is the ramp function, i.e. if$X \ge 0$ then $(x)_+ = x$, else $(x)_+ = 0$.

While solving the above, no transformation $\phi(.)$ is applied, and hence the generalization power of the kernel trick is not available in this arrangement.

2. Related work

The C-SVDD discussed above, as well as its variants that rely on expanding spatial resolution at the support vector locations (a method known as Conformal Kernel SVDD or CK-SVDD), as seen in Liu, Weng, Kang, Teng, and Huang (2010) have found applications like the P300 Speller Brain-Computer Interface. The current best complexity to solve the C-SVDD training problem is $O(N)$, an improvement from the original $O(N^3)$ as demonstrated in the core vector application of Chu, Tsang, and Kwok (2004). Even in this work, C-SVDD applies for small cases of the original problem and hence the LT-SVDD algorithm presented here applies there too. Also, the work in Chu et al. (2004) relies crucially for termination on a pre-identified fraction of the expected number of outliers: we do not need this in our algorithm. As is explained in Lee and Wright (2012), while the dual problem in 2-class SVMs is convex, the worst case space complexity is one dual variable per example/pattern. In order to assure that our algorithm does obtain an optimal point, the convexity of the primal problem in SVDD being obtained for a minor modification in Wang, Chung, and Shitong (2011) is our reference. The actual progress towards optimum is done using stochastic gradient methods which are considered popular only in linear SVMs (e.g.Lin (2013)). However, here we introduce an algorithm that adapts stochastic gradient to a method that uses the kernel trick.

3. Proposed work

This work proposes a novel low-complexity anomaly detection algorithm based on Support Vector Data Description (SVDD). For N patterns of dimension d, the current best complexity to solve SVDD training problem is $O(N)$ as demonstrated in Chu et al. (2004). The proposed algorithm reduces the complexity of both training and testing to $O(N + d)$ by avoiding the calculation of the Lagrange multipliers α_i, by locating an approximate pre-image of the SVDD sphere's center in the input space during the training phase itself. The proposed algorithm retains the benefit of the kernel trick: i.e. a minimum enclosing space is more descriptive of the data when calculated in a higher-dimensional feature space. The crux of the training algorithm is a gradient descent of the primal objective function using Simultaneous Perturbation Stochastic Approximation (SPSA) adapted to sub-gradients (He, Fu, & Marcus, 2003) and a recast form of the primal problem suggested in Pauwels & Ambekar (2011) that does away with slack variables.

The rest of this paper is organized as follows. Section 4 reviews the Fast-SVDD (F-SVDD) and then Section 5 describes our proposed procedure LT-SVDD. Experimental results on five UCI benchmark datasets and real-world credit datasets from the literature are presented in Section 6, while Section 7 gives concluding remarks.

4. Fast svdd (f-svdd)

The authors of Liu, Liu, and Chen (2010) propose a method called Fast SVDD (F-SVDD) to reduce the computational burden in the testing phase by replacing the kernel expansion in the decision function by a single kernel term. This work relies on calculating the pre-image $\hat{x}$ of a point termed as the 'agent of the SVDD sphere's centre a_F' and denoted by ψ_a. Note that $\hat{x}$ is in the input space whilst a_F and ψ_a are in the feature space. F-SVDD then uses a simple relationship between ψ_a and a_F, i.e. ψ_a is a scalar multiple of $a_{F,}$ to re-express the centre with a single vector. Hence the decision function of FSVDD contains only one kernel term, and thus the complexity of the FSVDD decision function during testing is a constant, no longer linear in the support vectors.

F-SVDD solves the pre-image problem to find a pattern $\hat{x} \in R^d$ such that $\psi_a = \phi(\hat{x})$ and $\psi_a = \gamma a_F$. In particular, F-SVDD solves as first step the dual of this problem:

$$\text{Minimize}\, O_p(R,\ a_F,\ \xi_i) = R^2 + C\sum_{i=1}^{N} \xi_i \tag{3}$$

$$\text{Subject to} \qquad \begin{matrix} \|\phi(x_i) - a_F\|^2 \le R^2 + \xi_i \\ \xi_i \ge 0,\ \forall i \in \{1..N\} \end{matrix},$$

where a_F is the center of the minimum enclosing ball, R is its radius and ξ_i are slack variables that allow the enclosing ball to have a soft margin. Here a_F and the kernel-trick based transformation of input pattern x_i, $\phi(x_i)$, are potentially vectors in the infinite dimensional feature space. All N vectors are assumed to belong to one, non-anomalous, class.

Since it is convenient for computational purposes, it is the dual of this problem that is solved:

$$\text{Maximize}\ O_d(\alpha) = 1 - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) \tag{4}$$