Data Driven Hyperparameters Optimization of One-Class Support Vector Machines for Anomaly Detection in Wireless Sensor Networks

Van Vuong Trinh and Kim Phuc Tran
Division of Artificial Intelligence,
Faculty of Information Technology,
Dong A University, Danang, Vietnam
Email: vanvuong.trinh@gmail.com, phuctk@donga.edu.vn

Thu Huong Truong
Department of Telecommunication Systems,
School of Electronics and Telecommunications,
Hanoi University of Science and Technology, Hanoi, Vietnam
Email: huong.truongthu@hust.edu.vn

***Abstract*—One-class support vector machines (OCSVM) have been recently applied to detect anomalies in wireless sensor networks (WSNs). Typically, OCSVM is kernelized by radial bais functions (RBF, or Gausian kernel) whereas selecting hyperparameters is based upon availability of labelled anomalous, which is rarely applicable in practice. This article investigates the application of OCSVM with data-driven hyperparameters optimization. Specifically, a kernel distance based optimization criteria is used instead of labelled data based metrics such as geometric mean accuracy (g-mean) or area under the receiver operating characteristic (AUROC). The efficiency of this method is illustrated over a real data set.**

***Index Terms*—one-class support vector machines, anomaly detection, wireless sensor networks, Gaussian kernel, parameters selection.**

I. Introduction

II. One-Class Support Vector Machines and Preliminaries

In this section, we briefly recall one-class support vector machines (OCSVM) [1].

A. Theory

Given N samples $\mathbf{x}_k$, $k = 1, \ldots, N$, SVDD method aims to estimate a sphere with minimum volume that contains all (or most of) these data. It is also assummed that almost these training samples belong to an unknown distribution. Let $\mathbf{a}$ and R being reserved for the center and the radius of the sphere, we define the objective function to minimize the volume of the sphere and the number of outliers as:

$$R^2 + C\sum_{k=1}^{N}\xi_k \tag{1}$$

where $C > 0$ is a regularization parameter with constraints that almost data points are within the sphere:

$$||\mathbf{x}_k - \mathbf{a}||^2 \leq R^2 + \xi_k,\ \xi_k \geq 0 \quad \forall k \tag{2}$$

To adapt with nonspherical distribution, a conventional approach is to map given data into a higher dimensional feature space, then learning a sphere in such a new space. This results into the so-called *primal optimisation* as follows:

$$\underset{R,\mathbf{a},\xi}{\text{Minimize}}\ R^2 + C\sum_{k=1}^{N}\xi_k \tag{3a}$$

$$\text{Subject to}\ \ ||\phi(\mathbf{x}_k) - \mathbf{a}||^2 \leq R^2 + \xi_k,\ \xi_k \geq 0 \quad \forall k \tag{3b}$$

where $\phi(\cdot)$ is the aforementioned feature mapping. The Lagrangian is hereafter written as:

$$\mathcal{L} = R^2 + C\sum_{k=1}^{N}\xi_k - \sum_{k=1}^{N}\alpha_k\left[R^2 + \xi_k - ||\phi(\mathbf{x}_k) - \mathbf{a}||^2\right] - \sum_{k=1}^{N}\gamma_k\xi_k \tag{4}$$

with the Lagrange multipliers $\alpha_k, \gamma_k \geq 0$. $\mathcal{L}$ should be minimized w.r.t. $R, \mathbf{a}, \xi_k$ and maximized w.r.t. α_k, γ_k.

Setting partial derivatives w.r.t. R, a, ξ gives:

$$\frac{\partial\mathcal{L}}{\partial R} = 0: \quad \sum_{k=1}^{N}\alpha_k = 1 \tag{5a}$$

$$\frac{\partial\mathcal{L}}{\partial \mathbf{a}} = 0: \quad \mathbf{a} = \sum_{k=1}^{N}\alpha_k\phi(\mathbf{x}_k) \tag{5b}$$

$$\frac{\partial\mathcal{L}}{\partial \xi_k} = 0: \quad \alpha_k + \gamma_k = C \quad \forall k \tag{5c}$$

Obviously, Lagrange multipliers γ_k can be eliminated by imposing bound constraints on α_k as:

$$0 \leq \alpha_k \leq C \quad \forall k \tag{6}$$

Substituting (5a)-(5c) into (4) leads to the following *dual*

optimisation:

$$\underset{\alpha}{\text{Maximize}} \sum_{k=1}^{N} \alpha_k \left(\phi\left(\mathbf{x}_k\right) \cdot \phi\left(\mathbf{x}_k\right)\right) - \sum_{k=1}^{N}\sum_{l=1}^{N} \alpha_k \alpha_l \left(\phi\left(\mathbf{x}_k\right) \cdot \phi\left(\mathbf{x}_l\right)\right) \tag{7a}$$

$$\text{Subject to } \sum_{k=1}^{N} \alpha_k = 1,\ 0 \leq \alpha_k \leq C \quad \forall k \tag{7b}$$

$$\left|\left|\phi\left(\mathbf{x}_k\right) - \mathbf{a}\right|\right|^2 < R^2 \rightarrow \alpha_k = 0 \tag{8a}$$

$$\left|\left|\phi\left(\mathbf{x}_k\right) - \mathbf{a}\right|\right|^2 = R^2 \rightarrow 0 < \alpha_k < C \tag{8b}$$

$$\left|\left|\phi\left(\mathbf{x}_k\right) - \mathbf{a}\right|\right|^2 > R^2 \rightarrow \alpha_k = C \tag{8c}$$

B. Kernelization

Instead of using inner product, an alternative kernel product can also be adopted:

$$\kappa\left(\mathbf{x}_k, \mathbf{x}_l\right) = \phi\left(\mathbf{x}_k\right) \cdot \phi\left(\mathbf{x}_l\right) \tag{9}$$

This is the known *kernel trick* [2], aims to avoid the need of explicitly declaring a feature mapping $\phi\left(\cdot\right)$.

$$\kappa\left(\mathbf{x}_k, \mathbf{x}_l\right) = \exp\left(-\frac{\left(\mathbf{x}_k - \mathbf{x}_l\right)'\left(\mathbf{x}_k - \mathbf{x}_l\right)}{2\sigma}\right) \tag{10}$$

where parameter σ is the kernel width.

$$\underset{\alpha}{\text{Maximize}} \sum_{k=1}^{N} \alpha_k \kappa\left(\mathbf{x}_k, \mathbf{x}_k\right) - \sum_{k=1}^{N}\sum_{l=1}^{N} \alpha_k \alpha_l \kappa\left(\mathbf{x}_k, \mathbf{x}_l\right) \tag{11a}$$

$$\text{Subject to } \sum_{k=1}^{N} \alpha_k = 1,\ 0 \leq \alpha_k \leq C \quad \forall k \tag{11b}$$

III. Description of Anomaly Detection Procedure for WSNs

IV. Illustrative Example

This section investigates the efficiency of anomaly detection algorithm over a real data set. The source code is freely available at https://github.com/trinhvv/wsn-ocsvm-dfn. All computation was performed on a platform with 2.6 GHz Intel(R) Core(TM) i7 and 16GB of RAM.

References

[1] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[2] B. Scholkopf, "The kernel trick for distances," *Advances in Neural Information Processing Systems 13*, vol. 13, pp. 301–307, 2001.