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

Truong Thu Huong
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) [**?**]. OCSVM is used to estimate the support of a distribution. Notationally, let us consider a data set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\}$, with each $\mathbf{x}_i \in \mathcal{R}^D$ belonging to a given class of interest (named target class). The basic idea behind the OCSVM is to separate data from the origin by finding a hyperplane with maximum margin separation from the origin. In order to deal with nonlinearly problems, the hyperplane is defined in a high-dimensional Hilbert feature space $\mathcal{F}$ where the samples are mapped through a nonlinear transformation $\Phi(.)$. We will work only a kernel function $k(\mathbf{x}, \mathbf{y})$ instead of the scalar product $(\mathbf{\Phi}(\mathbf{x}).\mathbf{\Phi}(\mathbf{y}))$. To separate the data set from the origin, [1] solved the following quadratic program:

$$\underset{\mathbf{w} \in \mathcal{F}, \mathbf{a}, \boldsymbol{\xi} \in \mathcal{R}^N, \rho \in \mathcal{R}}{\text{Minimize}} \quad \frac{1}{2}\left|\left|\mathbf{w}\right|\right|^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho \qquad \text{(1a)}$$

$$\text{Subject to } (\mathbf{w}.\Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0 \quad \forall i = 1, \ldots, N \qquad \text{(1b)}$$

Here, $\mathbf{w}$ is a vector perpendicular to the hyperplane in $\mathcal{F}$, and ρ is the distance to the origin. Since the training data distribution may contain outliers, a set of slack variables $\xi_i \geq 0$ is introduced to deal with them. The parameter $\nu \in (0, 1]$ controls the tradeoff between the number of examples of the training set mapped as positive by the decision function

$$f(\mathbf{z}) = \text{sgn}((\mathbf{w}.\Phi(\mathbf{x})) - \rho) \qquad (2)$$

and having a small value of $||\mathbf{w}||$ to control model complexity.

Using multipliers $\alpha_i, \beta_i \geq 0$, [1] introduced a Lagrangian

$$L(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \rho) = \frac{1}{2}\left|\left|\mathbf{w}\right|\right|^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho - \sum_{i=1}^{N}\alpha_i((\mathbf{w}.\Phi(\mathbf{x}_i)) - \rho + \xi_i) - \sum \qquad (3)$$

and set the derivatives with respect to the primal variables $\mathbf{w}$, $\boldsymbol{\xi}$, ρ equal to zero, i.e.

$$\frac{\partial L(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \rho)}{\partial \mathbf{w}} = 0: \quad \mathbf{w} = \sum_{i=1}^{N}\alpha_i\Phi(\mathbf{x}_i), \qquad (4)$$

$$\frac{\partial L(R, \mathbf{a}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \rho)}{\partial \xi_i} = 0: \quad \alpha_i = \frac{1}{\nu N} - \beta_i \leq \frac{1}{\nu N}, i = 1, \ldots, N \qquad (5)$$

$$\frac{\partial L(R, \mathbf{a}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \rho)}{\partial \rho} = 0: \quad \sum_{i=1}^{N}\alpha_i = 1 \qquad (6)$$

$$(7)$$

In (4), all patterns $\{\mathbf{x}_i : i \in [1, \ldots, N], \alpha_i > 0\}$ are called Support Vectors. From (4), using the kernel function, the decision function (2) is transformed into a kernel expansion

$$f(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{N}\alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho\right) \qquad (8)$$

Substituting (4), (5) and (6) into (3), and using the kernel function, we have

$$L(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j k(\mathbf{x}_i, \mathbf{x}_j). \qquad (9)$$

Then, we obtain the dual problem

$$\underset{\alpha}{\text{Minimize}} \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{10a}$$

$$\text{Subject to } 0 \le \alpha_i \le \frac{1}{\nu N}, \quad \forall i = 1 \dots N, \sum_{i=1}^{N} \alpha_i = 1 \tag{10b}$$

[1] shown that at the optimum, the two inequality constraints (1b) became equalities if α_i and β_i are nonzero, which implies $0 < \alpha_i < \frac{1}{\nu N}$. Thus, the value of ρ can be recovered by exploiting that for any such α_i, the corresponding pattern $\mathbf{x}_i$ satisfies

$$\rho = \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle = \sum_{j=1}^{N} \alpha_j k(\mathbf{x}_j, \mathbf{x}_i) \tag{11}$$

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.

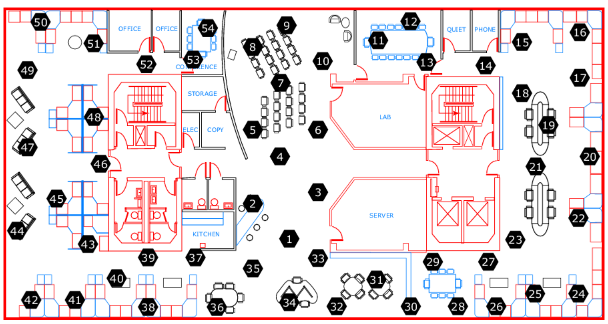

Fig. 1: A map of sensors' location. (Source: [2])

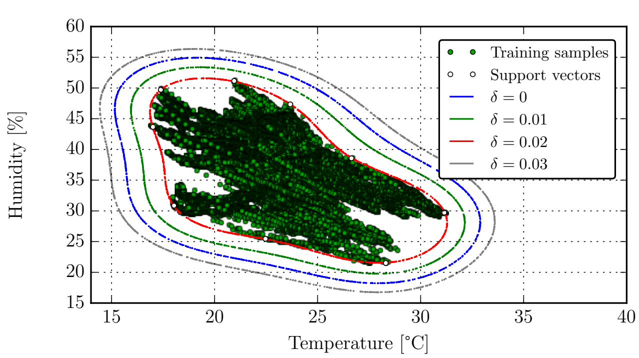

Fig. 2: Discrimination boundaries with some δ's values.

δ	0.001	0.005	0.01	0.02	0.03	0.05	0.1
DR [%]	100	100	100	100	100	100	100
FPR [%]	0	0	0	0	0	0	0

TABLE I: DR and FPR versus δ.

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] P. Buonadonna, D. Gay, J. M. Hellerstein, W. Hong, and S. Madden, "TASK: Sensor network in a box," *Proceedings of the Second European Workshop on Wireless Sensor Networks, EWSN 2005*, vol. 2005, pp. 133–144, 2005.

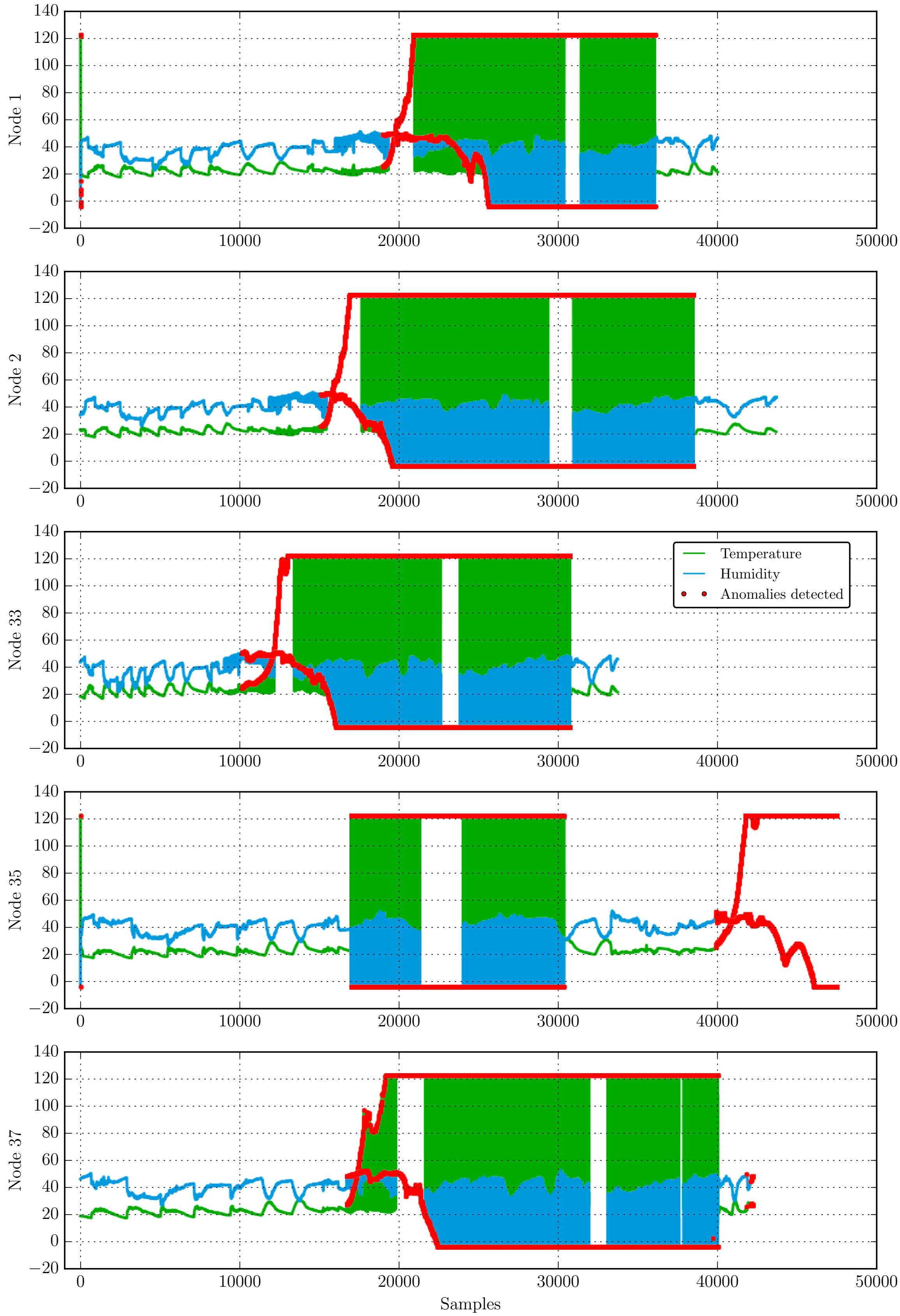

Fig. 3: Anomaly detection validation upon whole IBRL data set on 5 nodes. Almost apparent anomalies, i.e. temperature and humidity measurements that are too high or too low, are detected.