Anomaly Detection in Wireless Sensor Networks via Support Vector Data Description with Mahalanobis Kernels and Discriminative Adjustment

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

Anh Tuan Mai
International Training Institute for Materials Science (ITIMS),
Hanoi University of Science and Technology, Hanoi, Vietnam
Email: mtuan@itims.edu.vn

***Abstract*—In the past few years, wireless sensor networks are increasingly gaining impact in the real world with with various applications such as health-care, condition monitoring, control networks, and many other fields. Anomaly detection in a WSNs is an important aspect of data analysis in order to identify data items which do not conform to an expected pattern or other items in a dataset. This paper describes a novel anomaly detection method using support vector data description kernelized by Mahalanobis distance with adjusted discriminant threshold. The efficiency of this method to detect anomalies in wireless sensor networks is studied over a real data set. Experiment result demonstrates that the proposed approach achieved a high-level of detection accuracy and a low percentage of false alarm rate.**

***Index Terms*—Wireless sensor networks, Anomaly detection, Support vector data description, Mahalanobis kernels, Unsupervised learning.**

I. Introduction

Wireless sensor networks (WSNs) are spatially distributed autonomous sensors to monitor physical or environmental conditions, such as temperature, vibration, pressure, motion, etc. and to cooperatively pass their data through the network to a main location. WSNs are increasingly employed in a range of applications including industrial process, monitoring and control, machine health monitoring, healthcare applications and traffic control, see [1] for more details. Due to the deployment of a large number of sensor nodes in an uncontrolled or hostile environments, data measured and collected by WSNs is often unreliable. This will affect the modeling and scientific reasonable inference. Thus, it is very important that the anomaly of sensor node is detected in order to obtain accurate information and make effective decisions by information gatherers, see [2]

Anomaly detection is a method to identify when a metric is behaving differently than it has in the past, taking into account trends. Anomaly detection is implemented as one-class classification, because only one class is represented in the training data. In literature, a variety of anomaly detection techniques have been developed for certain application domains such as security systems, fraud detection and statistical process monitoring, for example, see [3], [4], [5], [6], [7], [8] and [9]. Recently, there have been growing interests in applying machine learning approaches for anomaly detection in WSNs. For further details see, for instance, [2], [1], [10] and [7]. Very recently [11] proposed a new SVDD method base Radial basis function (RBF) kernels to reduce computational complexity in the training phase and the testing phase for anomaly detection of node data in WSNs. It is important to note that the advantages of support vector machines is that we can improve generalization ability by proper selection of kernels. Mahalanobis kernels exploit the data distribution information more than RBF kernels do, see [12]. In addition, [13] shown that we can improve the performance of SVDD by using the Mahalanobis kernels.

In this paper, we propose a novel anomaly detection method using support vector data description kernelized by Mahalanobis distance with adjusted discriminant threshold. We then optimize the SVDD parameter and the kernel parameter. The remainder of the paper is organized as follows: in Section II, some necessary background definitions are introduced; in Section III, the anomaly detection approach in Wireless Sensor Networks is defined; Section IV presents an illustrative example, and finally, some concluding remarks and recommendations are made in Section V.

II. Support Vector Data Description and Preliminaries

In this section, we recall the so-called support vector data description (SVDD) originally derived in [14], which is an alternative of one-class support vector machines (OCSVM) [15] and is analogous to the known support vector machines (SVM) [16]. The uses of Malahanobis kernel and adjusted discrimination are also discussed. Notes that $\mathbf{x} \cdot \mathbf{y}$ stands for inner product of two vectors $\mathbf{x}, \mathbf{y}$ in Eucledian space.

A. Theory of SVDD

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:

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

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

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

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

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

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

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

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

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

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*:

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

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

This standard quadratic program (QP) and can be solved more efficiently than the primal problem. Evidently, (7) is feasible iff $C \geq \frac{1}{N}$ [17]. Assume that such a feasibility condition holds, the following interpretation is trivial once (7) is solved.

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

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

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

Only data samples $\mathbf{x}_k$ with $0 < \alpha_k < C$ are required in the distribution's description and these will therefore be refrered to as *support vectors*. This will be discussed later after we generalize the current linear SVDD into nonlinear version using kernel functions in the next sub-section.

B. Kernelization with Mahalanobis distance

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

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

This is the known *kernel trick* [18], aims to avoid the need of explicitly declaring a feature mapping $\phi(\cdot)$.

In this paper, we investigate the performance of SVDD using the *Mahalanobis kernel*:

$$\kappa(\mathbf{x}_k, \mathbf{x}_l) = \exp\left(-\frac{(\mathbf{x}_k - \mathbf{x}_l)'\,S^{-1}(\mathbf{x}_k - \mathbf{x}_l)}{2\sigma}\right) \tag{10}$$

where S is the estimated covariance matrix computed with the available training data $\mathbf{x}_k$, $k = 1, \ldots, N$ while parameter σ is the kernel width.

Thus, the optimisation (7) changes into following *kernelized dual optimisation*:

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

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

By solving this problem, one obtains, for example, m support vectors. Let $i, j = 1, \ldots, m$ being reserved for indices of these support vector. Then, sphere's radius is indeed the distance from any support vector, for instance $\mathbf{x}_i$, to the sphere's center:

$$R = \sqrt{\kappa(\mathbf{x}_i, \mathbf{x}_i) - 2\sum_{i=1}^{m}\alpha_i\kappa(\mathbf{x}_i, \mathbf{x}_i) + \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\kappa(\mathbf{x}_i, \mathbf{x}_j)} \tag{12}$$

C. Discrimination with threshold adjustment

This sub-section aims to establish the discriminative function for detecting anomalies. Roughly speaking, the kernel distance, namely $d(z, a)$, between a test sample z and the center $\mathbf{a}$ of sphere will be computed:

$$d(z, a) = \sqrt{\kappa(z, z) - 2\sum_{i=1}^{m}\alpha_i\kappa(z, \mathbf{x}_i) + \sum_{k=i}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\kappa(\mathbf{x}_i, \mathbf{x}_j)} \tag{13}$$

Then, this distance is compared with an appropriate threshold for discrimination. A trivial choice of such a threshold is sphere's radius R. This means that the data lies within the sphere will be categorized as anomalous. However, as suggested in [19, Section 7.1], it might be better to introduce a slack variable δ chosen *a priori* into the discriminative function. This is due to the fact that there is no guarantee that the data point outside the sphere is anomalous.

Thus, we hereafter use the following discriminative function for inclusion, where $\mathcal{H}$ stands for the Heaviside function $\mathcal{H}(x) = 1$, if $x \geq 0$ and 0 otherwise.

$$\begin{aligned} f(\mathbf{z}) &= \mathcal{H}\left(d - R^2 - \delta\right) \\ &= \mathcal{H}\left(\kappa(\mathbf{z},\mathbf{z}) - 2\sum_{i=1}^{m}\alpha_i\kappa(z,\mathbf{x}_i) - D\right) \end{aligned} \tag{14}$$

whereas the constant D can be pre-computed offline:

$$D = \kappa(\mathbf{x}_i,\mathbf{x}_i) - 2\sum_{i=1}^{m}\alpha_i\kappa(\mathbf{x}_i,\mathbf{x}_i) + \delta \tag{15}$$

A probabilistic analysis on the affect of δ upon SVDD's performance is available [19, Theorem 7.9]. Interested readers are referred to such a study and references therein for more details.

III. Description of Anomaly Detection Procedure for WSNs

This section describes the offline and online computations in our anomaly detection method. A training data set $\mathbf{x}_k^{train}$, $k = 1,\ldots,N_{train}$ is mandatory, whose most samples are assumed to be belong to an unknown distribution. While this assumptions is quite realistic, the method is obviously restricted only to a cluster of sensor nodes in WSNs (or small WSNs). In addition, we also assume that all data's feastures have been normalized in the unit range while a fixed threshold δ is given *a priori*. We hereafter describe how to select the hyperparameters (C,σ).

For performance evaluation, the false positive and false negative rates are used in this paper. A false positive occurs when a normal sample is identified as anomalous by the detector, and a false negative occurs when an anomalous measurement is identified as normal. The *false positive rate* (FPR, or false alarm) is the ratio between false positives and the actual normal measurements, and the *false negative rate* (FNR) is the ratio between false negatives and the actual anomalous measurements. The *detection rate* (DR) is given by $DR = 100 - FNR$. Also, the geometric mean accuracy (g-mean) is deployed:

$$g = \sqrt{(100 - FPR) \times DR} \tag{16}$$

Nevertheless, evaluating aforementioned performance measures requires the training samples to be labelled as either normal or anomalous. In this paper, such a labeling task will be done by means of *local outlier factor* [20], a nonparametric method to find anomalous data by measuring the local distance to its k neighbours. Then, a percentage of data, namely b, typically less than 5%, is classified as anomalous based on such local distances. Thus, the training set is splitted into the normal samples $\{\mathbf{x}_k^{normal}\}_k$ and anomalous samples $\{\mathbf{x}_k^{anomalous}\}_k$.

The major drawback of using local outlier factor for labeling is that it may leads to a typical high value of FPR which is generally prohibited. However, we will show later that by choosing a sufficiently large threshold δ, such an issue can be alleviated.

Once the data is roughly classified, the hyperparameters (C,σ) can be selected to optimize the g-mean criteria.

The whole procedure is summarized as below.

Algorithm 1 (Anomaly detection procedure) Assume a training set $\{\mathbf{x}_k^{train}\}_k$, a threshold δ and kernel $\kappa(\cdot,\cdot)$ are given. This algorithm produces the decision function $f(\cdot)$ defined by the constant D, the support vectors $\mathbf{x}_i$ and corresponding Lagrange mutipliers α_i. It also requires the local outlier factor's parameters k and b to be set *a priori*.

▷ Preprocessing phase:

1 Split the training set $\{\mathbf{x}_k^{train}\}_k$ using local outlier factor into $\{\mathbf{x}_k^{normal}\}_k$ and $\{\mathbf{x}_k^{anomalous}\}_k$.

▷ Training phase:

2 For each candidate pair (C_s,σ_s)

3 Solve (7) using only normal samples $\{\mathbf{x}_k^{normal}\}_k$.

4 Obtain decision function (14).

5 Evaluate $g(C_s,\sigma_s)$ using both $\{\mathbf{x}_k^{normal}\}_k$ and $\{\mathbf{x}_k^{anomalous}\}_k$.

6 Set hyperparameters as:

$$(C^\star,\sigma^\star) = \text{argmax } g(C,\sigma) \tag{17}$$

▷ Decision phase:

7 For a new sample $\mathbf{z}$, classify it according to (14), then raise an alarm if $f(\mathbf{z}) = 1$.

Generally, the candidate pair (C_s,σ_s) at the step 2 is belongto a finite set of candidates. Such a candicates are produced either by heuristic grid search [21] or Bayesian optimization [22].

IV. Illustrative Example

This section investigates the efficiency of anomaly detection algorithm using Mahalanobis kernel and control limit adjustment over a real data set. All computation was performed on a platform with 2.6 GHz Intel(R) Core(TM) i7 and 16GB of RAM.

A. Data description

We consider a data set gathered from a wireless sensor network deployment at the Intel Berkeley Research Laboratory (IBRL) [23]. A wireless sensor network consisting of 54 *Mica2Dot* sensor nodes was deployed in the IBRL for a 30 day (720 hour) period between 28th Feb 2004 and 5th April 2004 [10]. Fig. 1 shows the sensor deployment in the laboratory. The sensors collect five measurements: light in Lux, temperature

in degrees celsius, humidity (temperature corrected relative humidity) ranging from 0% to 100%, voltage in volts and network topology information in each 30 second interval. Node 0 is the gateway node. Other nodes transmit their data in multiple hops to the gateway node. The furthest node in the network is about 10 hops away from the gateway node. During the 30 day period, the 54 nodes collected about 2.3 million readings.

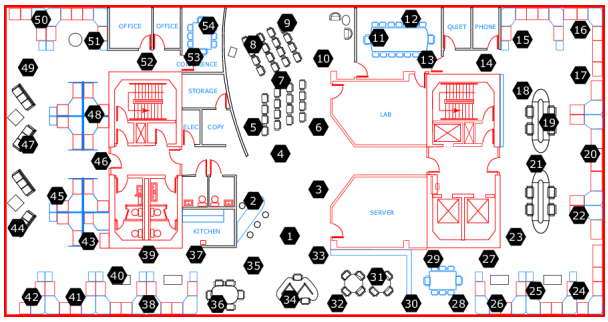

Fig. 1: A map of sensors' location. (Source: [23])

In this paper we consider the IBRL data set obtained from 5 close nodes, $1, 2, 33, 35, 37$. Also, only two features, namely temperature and humidity, are taken into account.

The data during the first 10 days period on March 2004 will be used as the training set. This training set contains more than 82000 samples.

In order to evaluate performance of the proposed method, we also use a testing set in some concrete time intervals. Since the original data did not contain any labels as to which data is normal and anomalous, we visually identify and label them as normal and anomalous. This data set contains about 10000 normal and 4000 anomalous samples.

B. Results

First, we use the Matlab's routine *consolidator(·)*[1] to reduce the cardinality of training set into only 55421 samples. Then, by using local outlier factor method with $k = 50$, $b = 1\%$, 54866 normal and 555 anomalous samples are obtained. The inverse of covariance matrix corresponding to normal samples is:

$$S^{-1} = \begin{bmatrix} 27,468 & 14,46 \\ 14,46 & 25,27 \end{bmatrix} \tag{18}$$

With different values as

$$\delta \in \{0.001, 0.005, 0.01, 0.02, 0.03, 0.05, 0.1\} \tag{19}$$

the Alg. 1 gives the hyperparameters (C, σ) with corresponding decision function. Fig. 2 illustrates obtained decision boundaries. It is evident that with $0.01 \leq \delta \leq 0.05$, the distribution of considered data can be well estimated.

[1]https://mathworks.com/matlabcentral/fileexchange/8354-consolidator

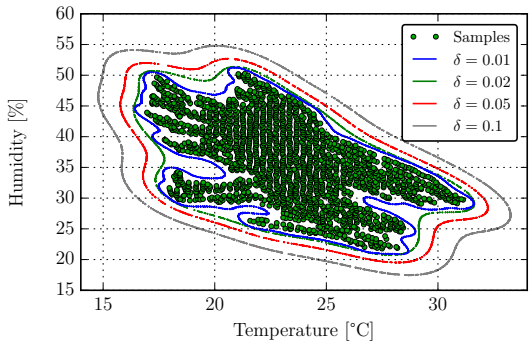

Fig. 2: Discriminative boundaries with some δ's values.

δ	0.001	0.005	0.01	0.02	0.03	0.05	0.1
DR [%]	100	100	100	100	100	100	100
FPR [%]	29.61	5.66	3.45	0	0	0	0

TABLE I: DR and FPR versus δ.

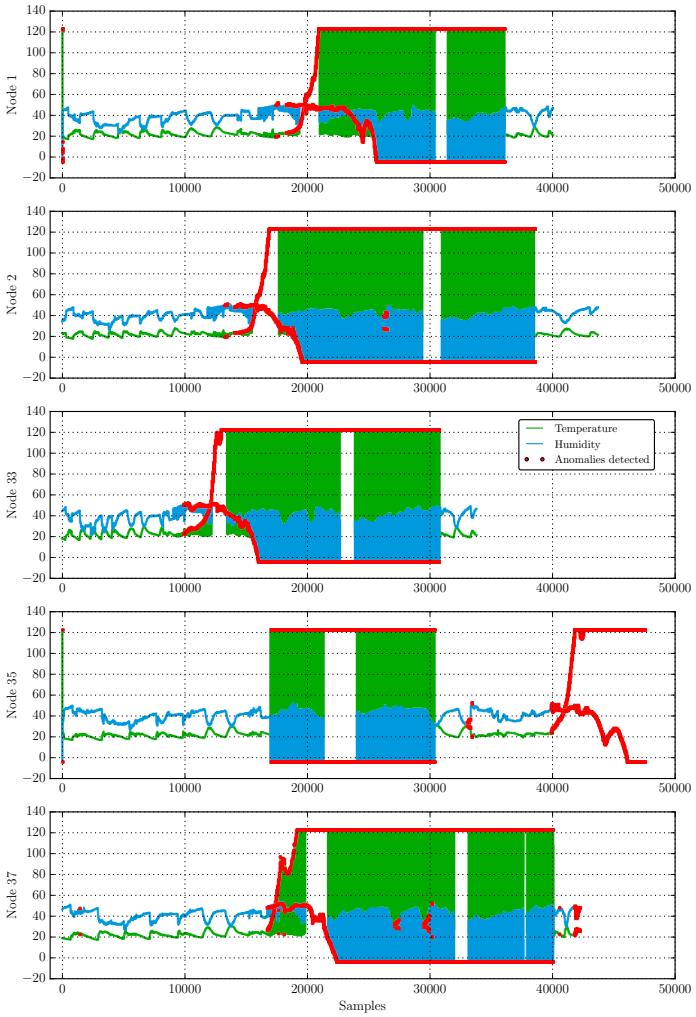

Fig. 3: Time-domain validation.

Table 1 presents the DR and FPR. Although with $\delta \geq 0.02$, these performance indices are surpringly good. However, this results might be unreliable and conservative. Since, the FPR is generally assumed to be about 5%, the choices of $\delta = 0.01$ or 0.02 may be appropriate.

Fig. 3 depicts detection result on considered nodes over time with $\delta = 0.02$, $C = 0.5025$ and $\sigma = 0.5039$. While the perfect DR is admitted as in the Table 1, one can observe some false alarm which is acceptable.

V. Concluding remarks

We presented a novel anomaly detection approach using SVDD in WSNs. Unlike the existing techniques which are all based on the RBF kernels, we developed an even more robust approach by incorporating the Mahalanobis distance-based kernel to SVDD and discriminative adjustment. Experiment result shown that the proposed approach achieved a high-level of detection accuracy and a low percentage of false alarm rate.

In the future, we would like to address the problem of anomaly detection using autoencoder and control charts and for uncertain data. We also focus on the detection ability of our proposed approach for large stream data.

References

[1] M. Xie, S. Han, B. Tian, and S. Parvin, "Anomaly detection in wireless sensor networks: A survey," *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1302–1325, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.jnca.2011.03.004

[2] A. Sharma, L. Golubchik, and R. Govindan, "Sensor faults: Detection methods and prevalence in real-world datasets," *ACM Transactions on Sensor Networks (TOSN)*, vol. 6, no. 3, p. 23, 2010.

[3] J. Ilonen, P. Paalanen, J. Kamarainen, and H. Kalviainen, "Gaussian mixture pdf in one-class classification: computing and utilizing confidence values," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 2. IEEE, 2006, pp. 577–580.

[4] D. A. Clifton, S. Hugueny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," *Journal of signal processing systems*, vol. 65, no. 3, pp. 371–389, 2011.

[5] K. Tran, P. Castagliola, and G. Celano, "Monitoring the Ratio of Two Normal Variables Using Run Rules Type Control Charts," *International Journal of Production Research*, vol. 54, no. 6, pp. 1670–1688, 2016.

[6] K. Tran, P. Castagliola, and G. Celano, "Monitoring the Ratio of Two Normal Variables Using EWMA Type Control Charts," *Quality and Reliability Engineering International*, 2015, in press, DOI: 10.1002/qre.1918.

[7] V. Chandola, A. Banerjee, and V. Kumar, *Anomaly Detection*. Boston, MA: Springer US, 2016, pp. 1–15.

[8] K. Tran, P. Castagliola, and G. Celano, "Monitoring the Ratio of Population Means of a Bivariate Normal distribution using CUSUM Type Control Charts," *Statistical Papers*, 2016, in press, DOI: 10.1007/s00362-016-0769-4.

[9] K. Tran, "The efficiency of the 4-out-of-5 Runs Rules scheme for monitoring the Ratio of Population Means of a Bivariate Normal distribution," *International Journal of Reliability, Quality and Safety Engineering*, 2016, in press, DOI: 10.1142/S0218539316500200.

[10] S. Rajasegarar, C. Leckie, and M. Palaniswami, "Hyperspherical cluster based distributed anomaly detection in wireless sensor networks," *Journal of Parallel and Distributed Computing*, vol. 74, no. 1, pp. 1833–1847, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jpdc.2013.09.005

[11] Z. Feng, J. Fu, D. Du, F. Li, and S. Sun, "A new approach of anomaly detection in wireless sensor networks using support vector data description," *International Journal of Distributed Sensor Networks*, vol. 13, no. 1, p. 1550147716686161, 2017.

[12] S. Abe, "Training of support vector machines with mahalanobis kernels," *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005*, pp. 750–750, 2005.

[13] E. Maboudou-Tchao, I. Silva, and N. Diawara, "Monitoring the mean vector with mahalanobis kernels," *Quality Technology & Quantitative Management*, pp. 1–16, 2016.

[14] D. M. J. Tax and R. P. W. Duin, "Support Vector Data Description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[15] 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.

[16] V. N. Vapnik, *Statistical Learning Theory*, 1998, vol. pp.

[17] W.-C. Chang, C.-P. Lee, and C.-J. Lin, "A revisit to support vector data description," *Dept. Comput. Sci., Nat. Taiwan Univ., Taipei, Taiwan, Tech. Rep*, 2013.

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

[19] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[20] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[21] A. Theissler and I. Dear, "Autonomously determining the parameters for svdd with rbf kernel from a one-class training set."

[22] J. Mockus, *Bayesian approach to global optimization: theory and applications*. Springer Science & Business Media, 2012, vol. 37.

[23] 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.