A new approach of anomaly detection in wireless sensor networks using support vector data description

Zhen Feng[1,2], Jingqi Fu[1], Dajun Du[1], Fuqiang Li[1] and Sizhou Sun[1]

Abstract

Anomaly detection is an important challenge in wireless sensor networks for some applications, which require efficient, accurate, and timely data analysis to facilitate critical decision making and situation awareness. Support vector data description is well applied to anomaly detection using a very attractive kernel method. However, it has a high computational complexity since the standard version of support vector data description needs to solve quadratic programming problem. In this article, an improved method on the basis of support vector data description is proposed, which reduces the computational complexity and is used for anomaly detection in energy-constraint wireless sensor networks. The main idea is to improve the computational complexity from the training stage and the decision-making stage. First, the strategy of training sample reduction is used to cut back the number of samples and then the sequential minimal optimization algorithm based on the second-order approximation is implemented on the sample set to achieve the goal of reducing the training time. Second, through the analysis of the decision function, the pre-image in the original space corresponding to the center of hyper-sphere in kernel feature space can be obtained. The decision complexity is reduced from $O(l)$ to $O(1)$ using the pre-image. Eventually, the experimental results on several benchmark datasets and real wireless sensor networks datasets demonstrate that the proposed method can not only guarantee detection accuracy but also reduce time complexity.

Keywords

Wireless sensor networks, support vector data description, anomaly detection, sequential minimal optimization, pre-image

Introduction

Wireless sensor networks (WSNs) are composed of a large number of distributed autonomous sensors, which monitor the environmental conditions, such as temperature, humidity, sound, vibration, pressure, motion, and pollutants.[1] WSNs have been extensively applied to many different fields, such as smart city,[2] smart grid, battlefield reconnaissance, environmental monitoring,[3,4] medical sensing,[5] traffic control, and other industrial applications. Due to the characteristics of WSNs, a sensor node is vulnerable to anomaly by some resource constraints, including energy, memory, bandwidth, computing capability, and transmission channel. Anomaly may be caused by not only faulty sensor node but also security threats in the network or

[1]School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
[2]College of Mechatronics and Control Engineering, Hubei Normal University, Huangshi, China

Corresponding author:
Jingqi Fu, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200072, China.
Email: jqfu@staff.shu.edu.cn

unusual phenomena in the monitoring scope. Therefore, 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.

Anomaly detection techniques from the aspect of data analysis could be categorized as[1] rule-based methods, statistical techniques, machine learning, and data mining approaches.[6–8] Among them, classification method is an important and systematic approach in the data mining and machine learning domains. It needs to acquire a classification model using a kind of samples and classify a new incoming sample into one of the class. Abnormal data, as a general rule, are difficult to obtain compared with the normal data. Thus, anomaly detection belongs to one-class classification problems. This method obtains a model by learning the normal samples and then uses the model to detect any abnormal sample difference from normality.

Recently, there have been growing interests in applying machine learning and data mining approaches for anomaly detection in WSN.[9–14] Anomaly detection based on data analysis in WSNs has been surveyed by O'Reilly et al.[15] An efficient algorithm is presented in Moshtaghi et al.,[11] which is a novel adaptive model for anomaly detection in a decentralized manner. This method mainly achieves the lower communication burden of WSNs and the higher detection precision. A distributed approach to outlier detection is performed in a principal component analysis (PCA)–based technique proposed by M Ahmadi Livani et al.[16] The scheme reduces communication complexity and achieves comparable accuracy in WSNs. Two outlier detection techniques based on distributed and online are presented in Zhang et al.[17] These techniques are achieved using a hyper-ellipsoidal one-class support vector machine (SVM) combined with the spatiotemporal correlation between sensor data. The objective of all above schemes is to improve detection accuracy and reduce false alarm. A robust and scalable mechanism is proposed in Kumarage et al.,[18] which can accurately and efficiently detect malicious anomalies in industrial WSNs, and achieves high detection accuracy and less communication overheads. In general, these literatures present anomaly detection methods in WSNs, which mainly consider detection accuracy and communication complexity of the algorithm. However, the computational complexity of the algorithm is less taken into account. In this article, a new method of anomaly detection is proposed in view of the computational complexity and can achieve comparable accuracy and less communication cost.

The support vector data description (SVDD)[19] is perhaps one of the most well-known one-class classification techniques for anomaly detection, and it has attracted extensive interests.[20] Given a target datasets, SVDD is to find a minimum hyper-sphere such that all or most normal data samples are enclosed into the hyper-sphere. The hyper-sphere boundary is the decision boundary, which is used to identify outliers different from the target data. By introducing kernel function, the nonlinear data in the original can be mapped into a high-dimensional feature space to achieve linear separable. SVDD can get a more flexible boundary to adapt irregularly shaped target datasets, which is able to be effectively applied to the field of anomaly detection.[21–24] However, in the training phase, SVDD is required to solve the quadratic programming problem with the strength of calculation and obtain the decision boundary of target data. If the number of training samples is M, then its computational complexity will be up to $O(M^3)$. Meanwhile, when an unknown sample needs to be evaluated in testing phase, the decision function requires all support vectors (SVs) to participate in the computation. That is, the complexity of decision making will be up to $O(|SVs|)$ for an unknown sample. Thus, the complexity of decision making for N unknown samples is $O(N|SVs|)$. If the number of N or SVs is quite large in WSNs, then this will inevitably lead to the testing phase of SVDD with large computational complexity.

Therefore, our goal is to propose a new SVDD method to reduce computational complexity in the training phase and the testing phase. Meanwhile, this method is applied to anomaly detection of node data in WSNs. First, combined with the strategy of training sample reduction, sequential minimal optimization (SMO) algorithm based on two-order approximation is used to reduce the computational complexity in the training phase. Next, the pre-image, which is corresponding to the center of hyper-sphere in kernel feature space, can be acquired in the original feature space. A fast decision-making method is presented by the pre-image in the testing phase, so that the complexity of decision making in SVDD for a single sample is reduced from $O(|SVs|)$ to $O(1)$. In the end, the proposed method is verified using University of California Irvine (UCI) datasets, the real Intel Berkeley Research Lab (IBRL) datasets, and the labeled WSNs datasets.

SVDD

The basic idea of SVDD classifier[19,25,26] is to find the minimum hyper-sphere containing all possible target data in the feature space. Given a set of training data $X = \{x_1, x_2, \ldots, x_l\}$, where $x_i \in R^d (1 \leq i \leq l)$ represents d-dimensional data and l is the size of the training data. The primal optimization problem of SVDD is then defined as

$$\text{Min } R^2 + C \sum_{i=1}^{l} \xi_i$$
$$s.t. \quad \|x_i - a\|^2 \leq R^2 + \xi_i, \quad i = 1, 2, \ldots, l$$
$$\xi_i \geq 0, \quad i = 1, 2, \ldots, l \tag{1}$$

where R and a are the radius and center of the hyper-sphere, respectively, in the feature space; ξ_i is the slack variable to allow for a few training data outside the hyper-sphere;[27,28] and the penalty parameter C controls the trade-off between the volume of the hyper-sphere and the number of target data outside the hyper-sphere.

In SVDD, the normal class is mapped from the input space into a feature space via a mapping function $\phi(\cdot)$. In this feature space F, the normal class is denoted as

$$\phi(x_1), \phi(x_2), \ldots, \phi(x_l) \tag{2}$$

where $\phi(x_i)$ is the image of sample x_i. The purpose of mapping function $\phi(\cdot)$ is to make the patterns much more compact in the feature space than in the input space, so as to enhance the performance. Furthermore, in the feature space, the inner products of two vectors can be calculated by a kernel function

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \tag{3}$$

where K satisfies the Mercer theorem and $\phi(x_i)$ and $\phi(x_j)$ represent two vectors in the feature space. In ϕ space, the primal optimization problem of SVDD is then defined as

$$\text{Min } R^2 + C \sum_{i=1}^{l} \xi_i$$
$$s.t. \quad \|\phi(x_i) - a\|^2 \leq R^2 + \xi_i, \quad i = 1, 2, \ldots, l$$
$$\xi_i \geq 0, \quad i = 1, 2, \ldots, l \tag{4}$$

In order to solve the optimization problem of equation (4) with these constraints, Lagrangian function is constructed as follows

$$L(R, a_\phi, \xi_i, \alpha_i, \gamma_i) = R^2 + C \sum_{i=1}^{l} \xi_i$$
$$- \sum_{i=1}^{l} \alpha_i \left(R^2 + \xi_i - \|\phi(x_i) - a_\phi\|^2 \right) - \sum_{i=1}^{l} \gamma_i \xi_i \tag{5}$$

where the Lagrange multipliers are $\alpha_i = (\alpha_1, \ldots, \alpha_l)^T \geq 0$ and $\gamma_i = (\gamma_1, \ldots, \gamma_l)^T \geq 0$. To find the stationary point of the Lagrange function, set partial derivatives to 0

$$\frac{\partial L}{\partial R} = 0 \rightarrow \sum_{i=1}^{l} \alpha_i = 1 \tag{6}$$

$$\frac{\partial L}{\partial a_\phi} = 0 \rightarrow a_\phi = \sum_{i=1}^{l} \alpha_i \phi(x_i) \tag{7}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \rightarrow a_i + \gamma_i = C \tag{8}$$

Equation (8) shows that $\alpha_i = C - \gamma_i$. When $0 \leq \alpha_i \leq C$, $\gamma_i \geq 0$ will be set up, so it can omit this constraint. Then, by substituting equations (6)–(8) into equation (5), dual form of the original problem (4) can be expressed as follows

$$\text{Max } \sum_{i=1}^{l} \alpha_i K(x_i, x_i) - \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j K(x_i, x_j)$$
$$s.t. \quad \sum_{i=1}^{l} \alpha_i = 1$$
$$0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, l \tag{9}$$

Here, the optimal solution is $\alpha' = (\alpha'_1, \alpha'_2, \ldots, \alpha'_l)^T$ about dual problem (9). Practically, most elements in the vector of α' are zero. The corresponding sample points are called nonsupport vectors (NSVs). The scope of hyper-sphere is determined by the sample points, which correspond to the value of the optimum solution $\alpha'_i > 0 | \quad \forall i \in \{1, 2, \ldots, l\}$. These sample points are called SVs. The sample points corresponding to the value of the optimum solution $C > \alpha'_i > 0$ are called margin support vectors (MSVs). The value of the optimum solution $\alpha'_i = C$ corresponding sample points are called nonmargin support vectors (NMSVs). The radius R of the hyper-sphere can be obtained by calculating the distance from the center of the hyper-sphere to one of the MSV. The sketch map of SVDD is shown in Figure 1.

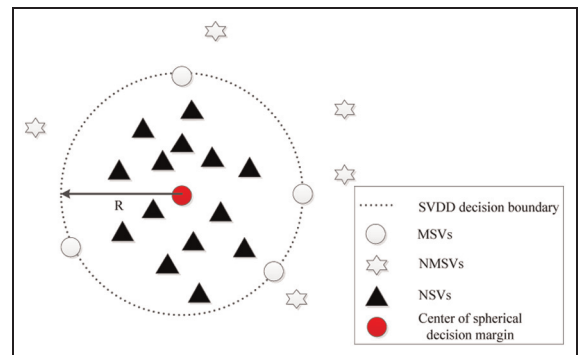Assume x_k is one of the MSVs, and $0 < \alpha'_k < C$ holds true, R can be calculated as follows

Figure 1. The sketch map of SVDD.

$$R^2 = \left\| x_k - a_\phi \right\|^2 = K(x_k, x_k) - 2 \sum_{i=1}^{l} \alpha_i K(x_k, x_i)$$
$$- \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j K(x_i, x_j) \tag{10}$$

To judge whether a test sample x_z is in the target class or not, it is assigned into the normal class if the distance between it and the sphere center is smaller than or equal to the radius R; on the contrary, sample x_z is then classified as outliers.

Proposed method of anomaly detection in WSNs

In WSNs, raw sensor observations often have low accuracy due to the limited energy and harsh deployment environments. This often results in outlying observations and affects the utility of WSNs for reliable decision making and situation awareness. In order to effectively utilize the data of WSNs, it is necessary to have anomaly detection for the sensor observations. An efficient algorithm for outlier detection based on SVDD was proposed in section "SVDD." However, SVDD has high computational complexity in the training and testing phases. According to this problem, this section proposes a method to reduce computational complexity of SVDD in the training and testing phases. The combination of the strategy of training set reduction and SMO algorithm based on second-order approximation is used to improve training speed. Meanwhile, a fast decision approach for an unseen sample is proposed in the testing phase, so as to accelerate testing speed.

Training set reduction strategy

For the dual problem (9), its solution has characteristic of sparse. That is, the decision boundary is obtained from the minimum hyper-sphere, which is composed of a fraction of SVs. A large number of sample points close to the center of the hyper-sphere do not contribute to the determination of the sphere, but conventional SVDD learning is performed over the entire training sample. Thus, this process will consume a significant amount of time and memory space. Here, considering the principle of SVDD and the relevant documents,[29,30] a kind of evaluation standard for sample based on Euclidean distance is presented in this article. Using the standard to evaluate all samples, the reduced set of training sample is obtained by removing a certain percentage of the sample near the center of all samples. Eventually, hyper-sphere boundary is obtained using the algorithm of the SVDD in the reduced set of training sample.

Given the training sample set $X = \{x_1, x_2, \ldots, x_l\}$, where $x_i \in R^d (1 \le i \le l)$, X is mapped into a feature space F by nonlinear mapping function $\phi(\cdot)$. The center of the training sample set X on space F is shown in formula (11)

$$\mu_F = \frac{1}{l} \sum_{i=1}^{l} \phi(x_i) \tag{11}$$

The Euclidean distance in the feature space F is shown in formula (12) between sample points x_i and x_j

$$d_F = \left\| \phi(x_i) - \phi(x_j) \right\|$$
$$= \sqrt{K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j)} \tag{12}$$

The distance is shown in equation (13) between the sample $\phi(x_i)$ and the center of μ_F in the feature space F

$$\left\| \phi(x_i) - \mu_F \right\|^2 = K(x_i, x_i) - \frac{2}{l} \sum_{j=1}^{l} K(x_i, x_j)]$$
$$+ \frac{1}{l^2} \sum_{j=1}^{l} \sum_{k=1}^{l} K(x_j, x_k) \tag{13}$$

Because the last item is a constant in equation (13), $\theta_F(i)$ is defined as evaluation criteria of sample and formula (14) is used to express

$$\theta_F(i) = K(x_i, x_i) - \frac{2}{l} \sum_{j=1}^{l} K(x_i, x_j) \tag{14}$$

The greater $\theta_F(i)$ indicates that the distance is farther from the center of μ_F. These values are arranged in a descending order. The reducing datasets of the sample are made up of the preceding τl values corresponding to the samples, where τ is the reduction factor based on the experimental results.

Remark 1. In the training phase, the conventional SVDD method needs to spend a large amount of time training NSVs of the training sample set. Thus, it is necessary to cut down the number of NSVs. Due to NSVs generally located near the center of sample set, a reduction strategy based on Euclidean distance is proposed. The strategy is implemented by removing a certain proportion of samples near the center of the training sample set.

SMO algorithm based on the second-order approximation

Similar to SVM, in SVDD, the key problem of training SVMs is how to solve quadratic programming (QP) optimization problem. Due to its immense size, the QP problem (9) that arises from SVs cannot be easily solved by standard QP techniques. SMO was presented

by Platt,[31] which is an extreme case of the decomposition algorithm where the size of working set is restricted to two elements. In each iteration, it does not require any optimization software in order to solve a simple two-variable problem. The SMO algorithm is mainly to solve two problems. One is to optimize the Lagrange multiplier of violating Karush–Kuhn–Tucker (KKT) conditions and meet the KKT conditions. The other is the problem of working set selection, which is the decision of the first to optimize the Lagrange multiplier. Certainly, working set selection is a key step in the convergence rate of SMO algorithm. There have been many literature studies on this work. Existing methods mainly rely on the violation of the optimality condition, which also corresponds to first-order information of the objective function. Fan et al.[32] proposed a simple working set selection using second-order approximation, which further improves the convergence rate of SMO algorithm. According to this idea, the SMO algorithm of the SVDD is derived using second-order approximation.

Remark 2. SVDD algorithm needs to solve QP optimization problem, which has high computational complexity. Thus, the SMO algorithm based on second-order approximation is proposed to improve the computational complexity of the QP problem.

Stop criterion. According to the optimization principle, when α_i satisfies the KKT condition of the objective function, it is a solution to the optimization problem. Therefore, a criterion is given to judge whether α_i is a violation of the KKT condition and is a stopping criterion. Rewriting the dual problem (9) into a matrix form

$$\begin{aligned} &\operatorname{Min} f(\alpha) = \alpha^T Q \alpha - P^T \alpha \\ &s.t. \quad e^T \alpha = 1, \quad e^T \alpha \geq 0, \, Ce - \alpha \geq 0 \end{aligned} \quad (15)$$

where Q is $l \times l$ matrix, $Q_{ij} = k(x_i, x_j)$; α, P, and e are column vectors of l dimension; $P_i = k(x_i, x_i)$; and $e_i = 1$. Equation (15) of the Lagrange function is

$$\overline{L}(\alpha, \lambda, \mu) = f(\alpha) - \lambda^T \alpha - \mu^T (Ce - \alpha) + b(e^T \alpha - 1) \quad (16)$$

where $\lambda_i \geq 0$, $\mu_i \geq 0$, and $b \geq 0$, and they are Lagrange multipliers.

For any α_i, if the problem (16) of the KKT condition is satisfied, then it is equivalent to meet the following conditions

1. $-f(\alpha)_i \leq b$, if $\alpha_i < C$
2. $-f(\alpha)_i \geq b$, if $\alpha_i > 0$
3. $e^T \alpha = 1$

It is defined that the index set are $I_{up}(\alpha) = \{t | \alpha_t < C\}$ and $I_{low}(\alpha) = \{t | \alpha_t > 0\}$. Take $i \in I_{up}(\alpha), j \in I_{low}(\alpha)$, if $-\nabla f(\alpha)_i \leq -\nabla f(\alpha)_j$ was established, then it is indicated that α_i and α_j are satisfied with the KKT condition of the problem (16). Otherwise, α_i and α_j are called a violation pair of the KKT condition.

Remark 3. In consequence, the iteration termination conditions are $m(\alpha) \equiv \max_{i \in I_{up}(\alpha)} -\nabla f(\alpha)_t$ and $M(\alpha) \equiv \min_{i \in I_{low}(\alpha)} -\nabla f(\alpha)_t$. If $m(\alpha) \leq M(\alpha) + e$, then α satisfies the KKT condition, where $e \geq 0$ is considered as a very small training accuracy in practical application.

Work set selection strategy based on the second-order approximation. The feasible direction is defined as $d^T \equiv [d_B^T, 0_N^T]$. In order to make the algorithm achieve faster convergence, at each iteration, the objective function $f(\alpha^k)$ needs to have the maximum reduction along the feasible direction. After $k + 1$ iterations, it will use $\alpha^k + d$ instead of α^k and carry out the second-order Taylor expansion of $f(\alpha^k + d)$ in α^k. Thus, the result can be expressed as follows

$$\begin{aligned} f(\alpha^k + d) - f(\alpha^k) &= \nabla f(\alpha^k)d + \frac{1}{2}d^T \nabla^2 f(\alpha^k)d \\ &= \nabla f(\alpha^k)_B d_B + \frac{1}{2}d_B^T \nabla^2 f(\alpha^k)_{BB} d_B \end{aligned} \quad (17)$$

Since $B(i, j)$ is a working set and $N = (1, 2, \ldots, l)/B$ is a nonworking set. In order to make the objective function $f(\alpha^k)$ obtain the largest descent in d, it is equivalent to solve the following optimization problem

$$\begin{aligned} &\operatorname{Min} Sub(B) = \nabla f(\alpha^k)_B d_B + \frac{1}{2}d_B^T \nabla^2 f(\alpha^k)_{BB} d_B \\ &s.t. \qquad e^T d_B = 0 \\ &\qquad d_t \geq 0, \quad if \, \alpha_t^k = 0, t \in B \\ &\qquad d_t \leq 0, \quad if \, \alpha_t^k = C, t \in B \end{aligned} \quad (18)$$

$d_i = -d_j$ can be obtained by $e^T d_B = 0$ and then it is substituted into the objective function of $Sub(B)$ to obtain the following formula

$$Sub(B) = p_{ij}d_j + \frac{1}{2}\eta_{ij}d_j^2 \quad (19)$$

where $p_{ij} = -\nabla f(\alpha^k)_i + \nabla f(\alpha^k)_j$ and $\eta_{ij} = k_{ii} + k_{jj} - 2k_{ij}$, k_{ij} indicates kernel function of $k(x_i, x_j)$. If $i \neq j$ and (i, j) is a violation pair, then $\eta_{ij} > 0$ and $p_{ij} > 0$. Now, $Sub(B)$ has the minimum value of $-p_{ij}^2/2\eta_{ij}$ when $\widehat{d}_i = -\widehat{d}_j = p_{ij}/\eta_{ij}$.

Remark 4. Hence, based on the second-order approximation, the working set selection strategy is as follows:

1. Get $i \in \arg\max_t \{-\nabla f(\alpha^k)_t | t \in I_{up}(\alpha^k)\}$;

2. Get $j \in \arg\min_t \{-\frac{p_{it}^2}{2\eta_{it}} | t \in I_{low}(\alpha^k), -\nabla f(\alpha^k)_t$
$< -\nabla f(\alpha^k)_i\}$;

3. Return $B = (i, j)$.

where k indicates the number of iterations, $p_{it} = -\nabla f(\alpha^k)_i + \nabla f(\alpha^k)_t$, and $\eta_{it} = k_{ii} + k_{tt} - 2k_{it}$. The k_{it} indicates kernel function of $k(x_i, x_t)$.

Optimization of two Lagrange multipliers. Let α_i^k and α_j^k be two multipliers in violation of the KKT condition. For their optimization, the rest of the multipliers are considered as constant. The value of α_i^{k+1} and α_j^{k+1} are separately the optimal values of α_i^k and α_j^k. It is obtained that $\alpha_i^{k+1} + \alpha_j^{k+1} = \alpha_i^k + \alpha_j^k = \varepsilon$ according to linear constraint condition of $e^T \alpha = 1$, where ε is a constant and $\varepsilon = 1 - \sum_{t=1, t \neq i, j}^{l} \alpha_t^k$. Without loss of generality, α_j^{k+1} was first calculated and then it is used to draw α_i^{k+1}.

Remark 5. The feasible region of α_j^{k+1} is $L \leq \alpha_j^{k+1} \leq C$, where $L = \max(0, \alpha_i^k + \alpha_j^k - C)$ and $H = \min(C, \alpha_i^k + \alpha_j^k)$.

Remark 6. The values of α_i^k and α_j^k are optimized, respectively

$$\alpha_i^{k+1} = \alpha_i^k - \left(\alpha_j^{k+1} - \alpha_j^k\right)$$

$$\alpha_j^{k+1} = \begin{cases} H, & \alpha_j^k - \frac{p_{ij}}{2\eta_{ij}} > H \\ \alpha_j^k - \frac{p_{ij}}{2\eta_{ij}}, & L \leq \alpha_j^k - \frac{p_{ij}}{2\eta_{ij}} \leq H \\ L, & \alpha_j^k - \frac{p_{ij}}{2\eta_{ij}} < L \end{cases}$$

where $p_{ij} = -\nabla f(\alpha^k)_i + \nabla f(\alpha^k)_j$ and $\eta_{ij} = k_{ii} + k_{jj} - 2k_{ij}$, k_{ij} indicates kernel function of $k(x_i, x_j)$.

Therefore, the optimal solution $\alpha^* = (\alpha_1^*, \ldots, \alpha_l^*)^T$ and SV set are obtained by the above training method. In order to reduce the errors in the operation, R is calculated by adopting the mean value of the MSVs, as follows

$$R^2 = \frac{1}{N} \sum_{x_k \in MSVs} \|x_k - a_\phi\|^2$$

$$= K(x_k, x_k) - \frac{2}{N} \sum_{x_k \in MSVs} \sum_{i=1}^{l} \alpha_i K(x_k, x_i) \qquad (20)$$

$$+ \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j K(x_i, x_j)$$

Given the unknown sample of $x \in R^d$, it is normal or not according to the following function

$$f(x) = R^2 - \|\phi(x) - a_\phi\|^2 \qquad (21)$$

When given a kernel function, such as the Gauss function $K(x_i, x_j) = \exp(-x_i - x_j^2 / 2h^2)$ (h is bandwidth parameters for the Gauss kernel), equation (21) can be rewritten as

$$f(x) = 2 \sum_{i=1}^{l} \alpha_i K(x_i, x) - v \qquad (22)$$

where

$$v = \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j K(x_i, x_j) + 1 - R^2 \qquad (23)$$

Obviously, v is a computable constant. If $f(x) \geq 0$, the sample x is the target sample, otherwise it is an abnormal sample. According to formula (22), the computational complexity of decision making for an unknown sample is O(l). However, there is a part of $\alpha_i = 0$ in the Lagrange multiplier which does not participate in the calculation of equation (22), and only the SVs corresponding to $\alpha_i > 0$ involve in the calculation. Hence, the computational complexity of decision making for an unknown sample is O($|SVs|$). In general, the number of SVs is not too small. Otherwise, the target sample is more likely to be error partition. Thus, the decision computation is very large when the amount of SVs and N is very large. To this end, on the basis of improving the training speed of SVDD, this article further proposed a new method to improve the decision complexity of SVDD, so as to improve the abnormal detection performance of SVDD in WSN.

SVDD decision approach

By observing the decision function of formula (21), if the pre-image of a_ϕ is a, then $a_\phi = \phi(a)$. Decision functions can be expressed as follows

$$f(x) = R^2 - \|\phi(x) - \phi(a)\|^2 = 2K(x, a) - v' \qquad (24)$$

where $v' = K(x, x) + K(x, a) - r^2$ is a constant.

From equation (24), it can be seen that the computational complexity of $K(x, a)$ is O(1). That is, the computational complexity of decision making for an unknown sample is O(1). However, the computational complexity with formula (22) is O($|SVs|$). If the pre-image of a_ϕ can be found in the original space, then this will significantly reduce the decision complexity.

The following sections describe how to obtain the pre-image. It is well known that a point in space can be represented approximately as the linear combination of

its neighbors, for example, locally linear embedding.[33,34] Hence, there is $a \approx \sum_i \beta_i \delta_i$ in certain δ neighborhood region of the sample a, where $\delta_i \in \delta$ and $\beta = (\beta_1, \ldots, \beta_{|\delta|})$. β is the weight vector, where $\beta_i > 0$ and $\sum_i \beta_i = 1$. Since a is located in all the sample points, the corresponding δ neighborhood can be composed of MSVs, namely $\delta_i \in MSVs$. It has also been reasonably assumed that the pre-image a can be estimated by

$$\hat{a} = \sum_{x_i \in MSVs} \beta_i x_i \qquad (25)$$

How to select the weight vector $\beta = (\beta_1, \ldots, \beta_{|\delta|})$ to minimize the value of the loss function $\hat{a} - a$. According to the mean value theorem,[35] the following formula is obtained

$$\begin{aligned}
\phi(\hat{a}) &\approx \phi(a) + \phi'(\varsigma)(\hat{a} - a) \\
&\Leftrightarrow \phi(\hat{a}) - \phi(a) \approx \phi'(\varsigma)(\hat{a} - a) \\
&\Rightarrow \|\phi(\hat{a}) - \phi(a)\| \geq \|\hat{a} - a\| \min(\phi'(\varsigma))
\end{aligned} \qquad (26)$$

From the formula, we can easily show that the lower bound for the smallest value of $\hat{a} - a$ can approximately be obtained by solving the lower bound for smallest value of $\phi(\hat{a}) - \phi(a)$. Therefore, β can be obtained by constructing the integrated squared error (ISE) and made as small as possible. Namely

$$\hat{\beta} = \min_\beta ISE(\beta) = \min\left(\sum_{x_i \in MSVs} \sum_{x_j \in MSVs} \beta_i \beta_j K(x_i, x_j) - 2 \sum_{x_i \in MSVs} \beta_i \sum_{x_j \in SVs} \alpha_j K(x_i, x_j) + \sum_{x_i \in SVs} \sum_{x_j \in SVs} \alpha_i \alpha_j K(x_i, x_j) \right) \qquad (27)$$

Since the last item in formula (27) is independent of β, the optimized model can be expressed by the following formula

$$\begin{aligned}
\hat{\beta} = \max_\beta &\sum_{x_i \in MSVs} \beta_i \sum_{x_j \in SVs} \alpha_j 2K(x_i, x_j) \\
&- \sum_{x_i \in MSVs} \sum_{x_i \in MSVs} \beta_i \beta_j K(x_i, x_j) \\
s.t. \; \beta^T 1 = 1, &\quad \beta_i \geq 0, 1 \leq i \leq |MSVs|
\end{aligned} \qquad (28)$$

It is obvious that formula (28) is a QP problem. The direct method is used to solve the QP problem. That is, the partial derivative is obtained with respect to β_k, and the result is equal to zero with the following expression

$$\frac{\partial ISE(\beta)}{\partial \beta_k} = 2 \sum_{x_j \in SVs} \alpha_j K(x_i, x_j) - 2 \sum_{x_j \in MSVs} \beta_k K(x_i, x_j) = 0$$

$$\Rightarrow \beta_k = \frac{\sum_{x_j \in SVs} \alpha_j K(x_j, x_k)}{\sum_{x_j \in MSVs} K(x_j, x_k)} \qquad (29)$$

Obviously, the weight vector is calculated by equation (29), which can effectively obtain the pre-image a. Thus, the pre-image a can be replaced by equation (24),

so that the computational complexity of the decision for the unknown sample is reduced to $O(1)$.

Remark 7. By analyzing the decision function of formula (24), and using formula (29) to obtain pre-image of the center of hyper-sphere, the aim is to reduce the computational complexity from $O(|SVs|)$ to $O(1)$ in the decision-making process of the SVDD algorithm.

New SVDD implementation process:

Training phase:
Step 1: initialize the kernel width parameter and the error penalty parameter C of SVDD;
Step 2: solve the QP problem using the SMO algorithm based on the second-order approximation in this article;
Step 3: compute the radius R by equation (20);
Step 4: compute the weight vector β by equation (29);
Step 5: estimate the pre-image $\hat{a}$ of the hyper-sphere center $\phi(a)$ by equation (25) which can be realized by the pre-image finding method.
Decision phase:
Step 6: for an unknown sample, classify it according to equation (24).

Experimental results

This article analyzes and compares the experimental results which are mainly reflected in the following three algorithms: the proposed new SVDD, SMO2-SVDD, and the traditional SVDD. SMO2-SVDD is an algorithm to improve the training speed by adding the strategy described above on the basis of SVDD. First, experimental parameter setting and performance evaluation metrics are presented in the following sections. Experimental results on the UCI datasets[36] are given in view of the three algorithms. Meanwhile, three kinds of algorithms are applied to anomaly detection of the node data in WSNs and are compared and analyzed. All algorithms are implemented in MATLAB 2013a on Windows 7 running on a PC.

On UCI datasets

The results about these algorithms running on publicly available UCI datasets are analyzed, which are widely used in the test of machine learning algorithms and recorded in Table 1. The first column and the second column in this table, respectively, represent name and

Table 1. Datasets used in the experiments.

Datasets	Dimension	Class	Samples
Balance scale (BS)	4	1	288
		2	49
		3	288
Breast cancer (BC)	30	1	357
		2	212
Wine (W)	13	1	71
		2	59
		3	48
Iris (I)	4	1	50
		2	50
		3	50
Liver (L)	6	1	200
		2	145
Connectionist bench (CB)	60	1	111
		2	97
Spambase (SB)	57	1	2788
		2	1813
Waveform (WF)	21	1	1696
		2	1647
		3	1657
Landsat satellite (LS)	36	1	1533
		2	1508
Blood transfusion service center (BSC)	4	1	570
		2	178

dimension of datasets. The corresponding target classes are shown in the third column, and the number of samples in every target class is indicated in the fourth column. It can be seen from Table 1 that spambase (SB), waveform (WF), and Landsat satellite (LS) are relatively larger than others. Here, this is intended to select the purpose of these datasets which conform to the feature of large data in WSNs.

For all experiments, the Gaussian kernel is applied in the process of simulation, and the cross-validation strategy is employed. The Gaussian kernel parameter h and the error penalty parameter C are, respectively, selected from the grid of $\{s^2/128, s^2/64, s^2/32, s^2/16, s^2/8, s^2/4, s^2/2, s^2, 2s^2, 4s^2, 8s^2\}$ and $\{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$, where s is an average of the 2 norm of all training samples. For a given method, these two parameters (h and C) are selected on the basis of the best classification performance, and the selected parameters are utilized in all the runs. In each run, 80% of the samples randomly acquired from the target class 1 are used for training. The remaining 20% of the samples together with all samples of other target classes are used for testing. For example, when class 1 is elected as target in balance scale datasets, 80% of the samples in class 1 are used for training, and the remaining 20% of the samples together with all samples of other classes are used for testing. That is to say, all other classes are regarded as outliers.

In order to make a comparative analysis reasonably, UCI datasets are used, and the research of algorithm mainly relates to three aspects of performance indicators, which are average accuracy, training time, and testing time. New SVDD, SMO2-SVDD, and SVDD are, respectively, run 10 times with the same training sample, testing sample, and parameters. Eventually, the mean and the standard deviation of running 10 times are the result of evaluation. In view of the imbalanced training datasets,[37,38] the geometric mean (g-mean) metric is employed in evaluating the accuracy performance of our algorithms

$$g = \sqrt{Acc^+ \cdot Acc^-} \qquad (30)$$

where Acc^+ and Acc^- are the classification accuracy on positive and negative classes, respectively

$$Acc^+ = \frac{Number\ of\ target\ samples\ correctly\ classified}{Number\ of\ total\ target\ samples\ classified} \times 100\%$$

$$Acc^- = \frac{Number\ of\ nontarget\ samples\ correctly\ classified}{Number\ of\ total\ nontarget\ samples\ classified} \times 100\%$$

In this section, by choosing appropriate parameters from the given grids, the accuracy of the three algorithms is described and compared. Table 2 lists the average g-means and the standard deviation with the 10-fold cross-validation method on these datasets. As shown in Table 2, the average g-means of the proposed new SVDD is slightly lower than the normal SVDD because our proposed method improves the training time. However, the gap between them is very small. The proposed new SVDD may be compared with the other algorithms on a majority of datasets.

The results shown in Table 3 indicate that the training and testing central processing unit (CPU) time of new SVDD compares favorably to the other algorithms. It usually achieves the best performance among all the methods. The results show that the training time of SVDD is slightly longer than that of new SVDD and SMO2-SVDD on these datasets. This is because the latter two methods use second-order approximation on working set selection to train target samples. Meanwhile, the results indicate that the proposed method of new SVDD obtains an extremely fast testing speed than SMO2-SVDD and SVDD. Obviously, this typical result originates from the fact that our methods here can cut down the decision complexity of SVDD from O(|SVs|) to O(1).

On IBRL datasets

In this experiment, the proposed new SVDD is evaluated with real IBRL datasets in WSNs. The IBRL datasets contain information collected from 54 sensors deployed in the IBRL, between 28 February and 5 April 2004. Mica2Dot sensors with weatherboards collected time-stamped topology information, along with

Table 2. Average g-means (%) and the standard deviation (%).

Datasets	New SVDD		SMO2-SVDD		SVDD	
	g-means	Standard deviation	g-means	Standard deviation	g-means	Standard deviation
Balance scale (BS)	78.31	0.65	79.67	0.51	80.52	0.43
Breast cancer (BC)	78.87	3.83	80.82	4.12	81.38	3.62
Wine (W)	87.64	4.96	88.53	5.18	90.31	4.54
Iris (I)	91.78	3.26	92.75	3.45	95.18	3.78
Liver (L)	58.13	4.68	57.13	4.24	58.84	4.87
Connectionist bench (CB)	50.02	3.45	50.54	3.56	51.77	3.24
Spambase (SB)	73.12	1.19	73.35	1.02	75.65	0.85
Waveform (WF)	88.24	0.46	89.87	0.52	91.24	0.38
Landsat satellite (LS)	90.18	0.62	90.58	0.67	92.31	0.32
Blood transfusion service center (BSC)	77.25	2.67	77.84	2.53	78.54	2.86

SVDD: support vector data description.
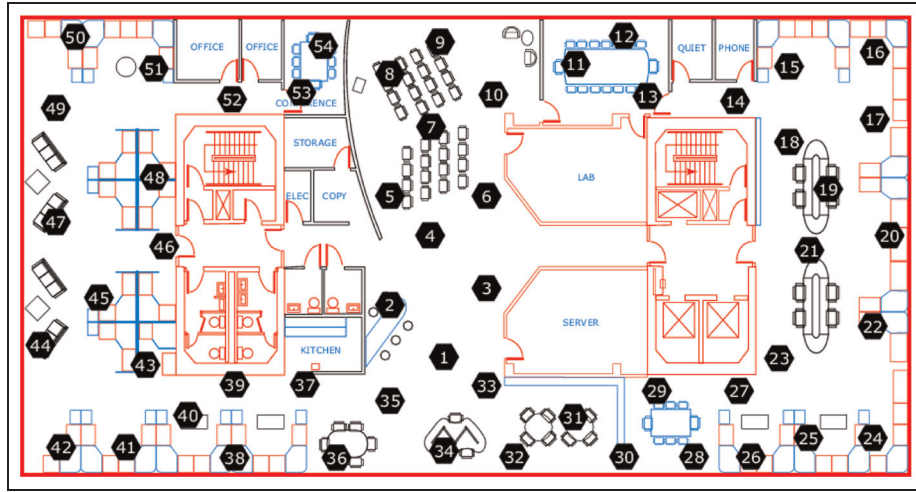
Figure 2. Sensor node location in the IBRL deployment.

humidity, temperature, light, and voltage values once every 31 s. The data were collected using the TinyDB in-network query processing system, built on the TinyOS platform.[39] The sensors were arranged in the lab, according to the diagram shown in Figure 2. Let us first consider a small sensor sub-network, which can be easily extended to a cluster-based or a hierarchal network topology. This sub-network consists of densely deployed n nodes $\{s_1, \ldots, s_n\}$. In Figure 2, the nodes 1, 2, 33, 35, and 37 are closed to each other, which consists of a sub-network $\{s_1, s_2, s_{33}, s_{35}, s_{37}\}$. According to the requirements of applications, local outlier detection is the most important task of anomaly detection in WSNs. Local outliers represent those outliers that are detected at individual sensor node only using its local data. Considering the computational complexity of anomaly detection, a method of new SVDD is proposed to identify local outliers at individual sensor node. This article utilizes the IBRL dataset

to be collected by 5 sensor nodes, which are 6, 12, 18, 24, and 30 hours of partial data, respectively, recorded on March 6, 2004. These data are used in our evaluation. In total, three attributes of each data vector are used, including temperature, humidity, and light measurements. Since time is different in terms of the obtained data, the distribution of the data is not the same. It is conducive to verify the effectiveness of the algorithm.

Robustness analysis. In this section, the robustness of the three algorithms to Gaussian kernel parameter h and the error penalty parameter C is researched in WSNs. In this experiment, the datasets of node 33 with 18 h were adopted, in which 80% of the data obtained randomly from the datasets are selected as target class, and the remaining 20% of the data together with the generated of artificial data accounted for 30% of the datasets are used for testing. Here, the artificial data randomly

Table 3. Average training time and testing time on the datasets (s).

Datasets	New SVDD				SMO2-SVDD				SVDD			
	Training time		Testing time		Training time		Testing time		Training time		Testing time	
	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
Balance scale (BS)	0.7682	0.2857	0.0043	0.0001	0.7682	0.2857	1.5482	0.0041	1.6527	0.3541	1.5482	0.0041
Breast cancer (BC)	4.6894	0.3128	0.0079	0.0001	4.6894	0.3128	2.4700	0.0086	12.8554	0.2533	2.4700	0.0086
Wine (W)	1.0875	0.3451	0.0012	0.0000	1.0875	0.3451	0.1800	0.0015	2.5423	0.3768	0.1800	0.0015
Iris (I)	0.4251	0.1845	0.0010	0.0000	0.4251	0.1845	0.1500	0.0013	1.0767	0.1287	0.1500	0.0013
Liver (L)	1.9542	0.5436	0.0034	0.0001	1.9542	0.5436	1.3200	0.0062	4.2959	0.6183	1.3200	0.0062
Connectionist bench (CB)	5.0547	0.4537	0.0025	0.0000	5.0547	0.4537	0.8125	0.0024	13.2151	0.5142	0.8125	0.0024
Spambase (SB)	402.6985	48.6858	0.0364	0.0002	402.6985	48.6858	63.5684	0.1026	987.5320	75.8364	63.5684	0.1026
Waveform (WF)	286.8542	32.5784	0.0265	0.0002	286.8542	32.5784	45.6528	0.0624	659.1806	52.9859	45.6528	0.0624
Landsat satellite (LS)	187.6428	28.5279	0.0182	0.0001	187.6428	28.5279	25.5268	0.1639	440.5286	46.6439	25.5268	0.1639
Blood transfusion service center (BSC)	2.9248	0.5143	0.0084	0.0001	2.9248	0.5143	3.2856	0.0123	7.2718	0.7293	3.2856	0.0123

SVDD: support vector data description.
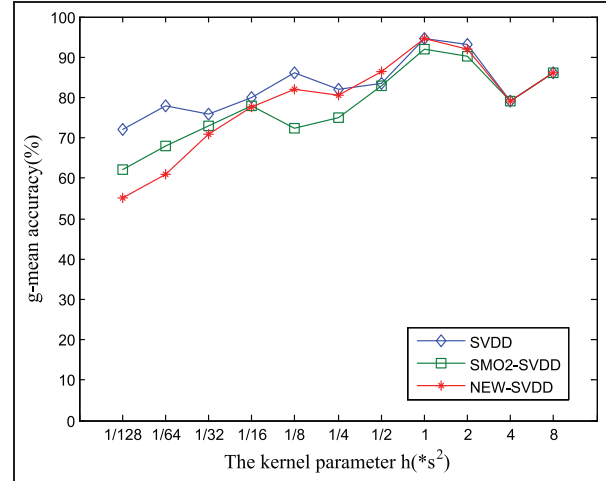
Figure 3. Robustness performance to parameter h.
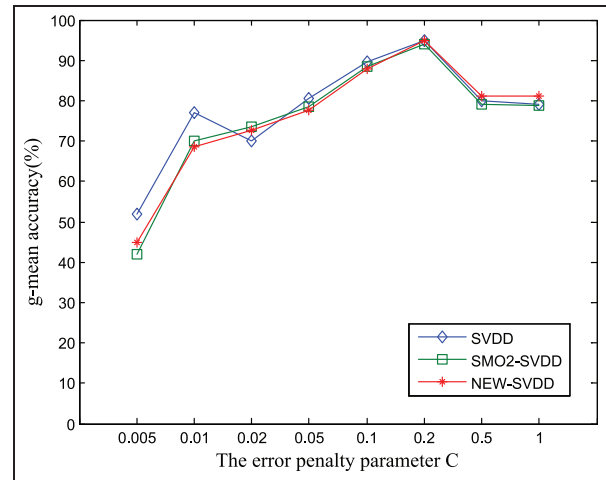
Figure 4. Robustness performance to parameter C.

generated the abnormal values different from the node data. The value of the error penalty parameter C is set to be 0.1 when the influence of the kernel parameter h on the g-mean accuracy is studied. The experimental results are shown in Figure 3. Similarly, the value of the kernel parameter h is set to be s^2 when the influence of the error penalty parameter C on the g-mean accuracy is studied. The experimental results are shown in Figure 4. As can be seen from Figures 3 and 4, the new SVDD is closest to the SVDD, and the SMO2-SVDD is slightly worse. Therefore, the SMO2-SVDD and new SVDD can obtain good performance when h is, respectively, equal to s^2 and $2s^2$, as well as when C is, respectively, equal to 0.1 and 0.2.

Accuracy analysis. In this section, the performance of WSNs is demonstrated on the g-mean accuracy of the
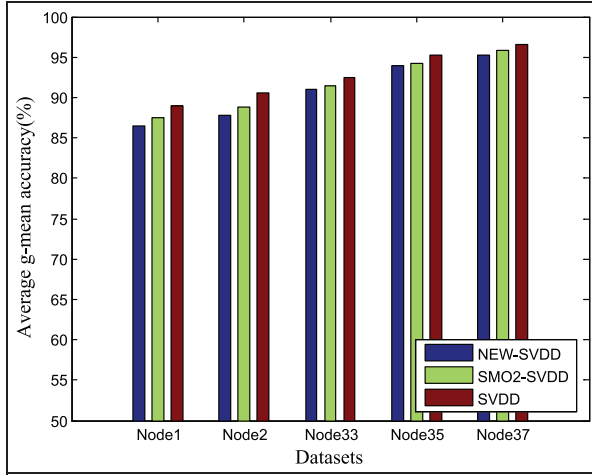
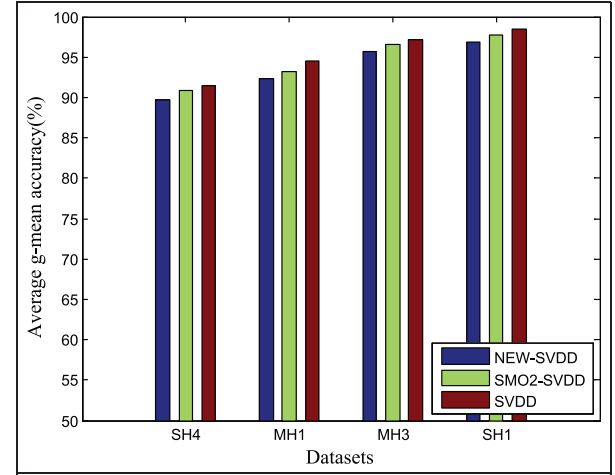**Figure 5.** Average *g*-means (%) on the node.

Figure 6. Average *g*-means (%) on the node.

three methods where appropriate kernel parameter h and the error penalty parameter C have been searched from the given grids. In each run of experiment, 80% of the data obtained randomly from each node are used for training, and the remaining 20% of the data together with the generated artificial outliers are used for testing. The generated artificial outliers are 30% of the node data. The experiment was repeated 10 times, and the average of *g*-mean accuracy is listed in Figure 5. It can be seen from Figure 5 that the proposed new SVDD may be compared with SVDD on the datasets of node. Moreover, with the increasing amount of sample in the datasets, new SVDD improves the *g*-mean accuracy of anomaly detection. This result demonstrates that new SVDD can give quite good detection performance for large amount of data in WSNs.

On the labeled WSN datasets

In this experiment, the proposed new SVDD is evaluated on the labeled WSN datasets.[40] The datasets consist of humidity and temperature measurements collected during 6-h period at intervals of 5 s. Single-hop data are collected on 9 May 2010, and the multi-hop data are collected on 10 July 2010. Label "0" denotes normal data and label "1" denotes an introduced event or outlier. Similarly, this article utilized a portion of the data from the labeled WSN datasets in our evaluation, namely, single-hop data of nodes 1 and 4, as well as multi-hop data of nodes 1 and 3. In the evaluation, the used single-hop data of node 1 (SH1) consist of 115 abnormal data with label "1" and 3200 normal data with label "0." The used single-hop data of node 4 (SH4) consist of 30 abnormal data with label "1" and 1200 normal data with label "0." The used multi-hop data of node 1 (MH1) consist of 50 abnormal

data with label "1" and 1800 normal data with label "0." The used multi-hop data of node 3 (MH3) consist of 100 abnormal data with label "1" and 3000 normal data with label "0."

Robustness analysis. In this section, the robustness of the three algorithms to Gaussian kernel parameter h and the error penalty parameter C is researched in the labeled WSNs. In this experiment, the datasets of SH1 were adopted, in which 80% of the normal data obtained randomly from the datasets are selected as target class, and the remaining 20% of the data together with the abnormal data are used for testing. The values of Gaussian kernel parameter h and the error penalty parameter C are the same as those in the IBRL datasets when the SMO2-SVDD and new SVDD can obtain good performance.

Accuracy analysis. In this section, the performance of WSNs is demonstrated on the *g*-mean accuracy of the three methods where appropriate kernel parameter h and the error penalty parameter C have been searched from the given grids. In each run of experiment, 80% of the data obtained randomly from each node are used for training, and the remaining 20% of the data together with the abnormal data are used for testing. The experiment was repeated 10 times, and the average of *g*-mean accuracy is listed in Figure 6. It can be seen from Figure 6 that the proposed new SVDD may be compared with SVDD on the datasets of node. Moreover, with the increasing amount of sample in the datasets, new SVDD improves the *g*-mean accuracy of anomaly detection. This result demonstrates that new SVDD can give quite good detection performance for the labeled datasets in WSNs.

Complexity analysis

In this section, the complexity of the proposed new SVDD is analyzed aiming at the problem of anomaly detection in WSNs and compared with SMO2-SVDD and SVDD. Since SMO2-SVDD improves the training speed on the basis of SVDD, the training time of the target sample is reduced. New SVDD applies the method for obtaining pre-image to improve the testing speed on the basis of SMO2-SVDD, so the testing time of an individual sample is reduced from $O(|SVs|)$ to $O(1)$. Tables 4 and 5 show the experimental results about the training time and testing time in WSNs. The results are that the average value and the standard deviation are calculated after 10 times of running. As can be seen from Tables 4 and 5, the training time of SVDD is the longest. New SVDD and SMO2-SVDD approximately obtain the same training time because they take the same training algorithm. Meanwhile, since new SVDD adopts the method for obtaining pre-image, new SVDD obtains the shortest time in the decision phase which does not change with the increase in sample.

Conclusion

Anomaly detection on data is challenging and demanding issue in WSNs, due to its increase diverse applications such as fault detection, incident or intrusion detection. This article has presented a new method of SVDD for anomaly detection of large data in WSNs. This method mainly solves two aspects of the traditional SVDD method. The first aspect is QP problem in the training phase which involves highly complicated calculations. In order to reduce the training complexity, the SMO algorithm based on the second-order approximation is adopted. The second aspect is testing complexity for anomaly detection. A pre-image finding approach based on ISE criterion is proposed to reduce the complexity of decision making. Finally, using UCI datasets and IBRL datasets of WSNs, the three algorithms, namely, new SVDD, SMO2-SVDD, and SVDD, are compared in terms of performance. Experimental results show that the proposed new SVDD method can reduce the computational complexity compared with SMO2-SVDD and SVDD method and maintain similar accuracy performance of detection.

In this article, the Gaussian kernel is only addressed, and other kernels such as the dot product kernels will be developed in the corresponding fast SVDD algorithms in the future. Meanwhile, the proposed method needs to be further improved in terms of detection accuracy.

Table 4. Average training time and testing time on the IBRL datasets (in second).

Datasets	New SVDD				SMO2-SVDD				SVDD			
	Training time		Testing time		Training time		Testing time		Training time		Testing time	
	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
Node 1	0.8154	0.2763	0.0016	0.0001	0.8154	0.2763	0.8526	0.0026	1.7425	0.2651	0.8526	0.0026
Node 2	1.8253	0.3251	0.0029	0.0001	1.8253	0.3251	1.5261	0.0035	3.2861	0.3128	1.5261	0.0035
Node 33	2.6538	0.4015	0.0042	0.0001	2.6538	0.4015	2.1962	0.0038	4.8432	0.3825	2.1962	0.0038
Node 35	3.4862	0.3528	0.0063	0.0001	3.4862	0.3528	3.0258	0.0029	6.5243	0.2684	3.0258	0.0029
Node 37	4.3258	0.5634	0.0086	0.0001	4.3258	0.5634	3.8247	0.0031	8.1685	0.4265	3.8247	0.0031

SVDD: support vector data description.

Table 5. Average training time and testing time on the labeled WSN datasets (in second).

Datasets	New SVDD						SMO2-SVDD						SVDD					
	Training time		Testing time				Training time		Testing time				Training time		Testing time			
	Average	Standard deviation	Average	Standard deviation			Average	Standard deviation	Average	Standard deviation			Average	Standard deviation	Average	Standard deviation		
SH4	1.5648	0.1686	0.0012	0.0001			1.5648	0.1686	0.6384	0.0019			2.8658	0.1862	0.6384	0.0019		
MH1	2.0672	0.2547	0.0023	0.0001			2.0672	0.2547	0.8758	0.0028			3.7856	0.2072	0.8758	0.0028		
MH3	3.6374	0.3292	0.0068	0.0001			3.6374	0.3292	1.4583	0.0032			6.8475	0.3185	1.4583	0.0032		
SH1	4.0165	0.3863	0.0075	0.0001			4.0165	0.3863	1.6026	0.0037			7.9649	0.3831	1.6026	0.0037		

SVDD: support vector data description; SH4: single-hop data of node 4; MH1: multi-hop data of node 1; MH3: multi-hop data of node 3; SH1: single-hop data of node 1.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

References

1. Xie M, Han S, Tian B, et al. Anomaly detection in wireless sensor networks: a survey. *J Netw Comput Appl* 2011; 34(4): 1302–1325.
2. Leccese F, Cagnetti M and Trinca D. A smart city application: a fully controlled street lighting isle based on Raspberry-Pi card, a Zigbee sensor network and WiMAX. *Sensors* 2014; 14(12): 24408–24424.
3. Dajun D, Bo Q, Minrui F, et al. Multiple event-triggered H2/H ∞ filtering for hybrid networked systems with random network-induced delays. *Inform Science* 2015; 325: 393–408.
4. Dajun D, Bo Q, Minrui F, et al. Quantized control of distributed event-triggered networked control systems with hybrid wired-wireless networks communication constraints. *Inform Sciences* 2017; 380: 74–91.
5. Yang Y, Liu Q, Gao Z, et al. Data fault detection in medical sensor networks. *Sensors* 2015; 15(3): 6066–6090.
6. Zamani M. Machine learning techniques for intrusion detection (arXiv preprint arXiv:1312.2177v2), 2015, pp.1–11, https://arxiv.org/pdf/1312.2177.pdf
7. Dua S and Du X. *Data mining and machine learning in cybersecurity*. Boca Raton, FL: CRC Press, 2014.
8. Butun I, Morgera SD and Sankar R. A survey of intrusion detection systems in wireless sensor networks. *IEEE Commun Surv Tutor* 2014; 16(1): 266–282.
9. Siripanadorn S, Hattagam W and Teaumroong N. Anomaly detection in wireless sensor networks using self-organizing map and wavelets. *Int J Commun* 2010; 4(3): 74–83.
10. Branch JW, Giannella C, Szymanski B, et al. In-network outlier detection in wireless sensor networks. *Knowl Inf Syst* 2013; 34(1): 23–54.
11. Moshtaghi M, Leckie C, Karunasekera S, et al. An adaptive elliptical anomaly detection model for wireless sensor networks. *Comput Netw* 2014; 64: 195–207.
12. Salem O, Guerassimov A, Mehaoua A, et al. Anomaly detection scheme for medical wireless sensor networks. In: Furht B and Agarwal A (eds) *Handbook of medical and healthcare technologies*. New York: Springer, 2013, pp.207–222.
13. Rajasegarar S, Leckie C and Palaniswami M. Hyperspherical cluster based distributed anomaly detection in wireless sensor networks. *J Parallel Distr Com* 2014; 74(1): 1833–1847.

14. Salmon HM, De Farias CM, Loureiro P, et al. Intrusion detection system for wireless sensor networks using danger theory immune-inspired techniques. *Int J Wireless Inform Network* 2013; 20(1): 39–66.

15. O'Reilly C, Gluhak A, Ali Imran M, et al. Anomaly detection in wireless sensor networks in a non-stationary environment. *IEEE Commun Surv Tutor* 2014; 16(3): 1413–1432.

16. Ahmadi Livani M, Abadi M, Alikhany M, et al. Outlier detection in wireless sensor networks using distributed principal component analysis. *J AI Data Min* 2013; 1(1): 1–11.

17. Zhang Y, Meratnia N and Havinga PJM. Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine. *Ad Hoc Netw* 2013; 11(3): 1062–1074.

18. Kumarage H, Khalil I, Tari Z, et al. Distributed anomaly detection for industrial wireless sensor networks based on fuzzy data modelling. *J Parallel Distr Com* 2013; 73(6): 790–806.

19. Tax DMJ and Duin RPW. Support vector data description. *Mach Learn* 2004; 54(1): 45–66.

20. Wu M and Ye J. A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE T Pattern Anal* 2009; 31(11): 2088–2092.

21. Liu Y-H, Lin S-H, Hsueh Y-L, et al. Automatic target defect identification for TFT-LCD array process inspection using kernel FCM-based fuzzy SVDD ensemble. *Expert Syst Appl* 2009; 36(2): 1978–1998.

22. Park J, Kang D, Kim J, et al. SVDD-based pattern denoising. *Neural Comput* 2007; 19(7): 1919–1938.

23. Nanni L. Machine learning algorithms for T-cell epitopes prediction. *Neurocomputing* 2006; 69(7–9): 866–868.

24. Banerjee A, Burlina P and Diehl C. A support vector method for anomaly detection in hyperspectral imagery. *IEEE T Geosci Remote* 2006; 44(8): 2282–2291.

25. Tax DMJ and Duin RPW. Support vector domain description. *Pattern Recogn Lett* 1999; 20(11–13): 1191–1199.

26. Tax DMJ. *One-class classification: concept-learning in the absence of counter-examples*. Delft: Delft University of Technology, 2001.

27. Kang W-S and Choi JY. Domain density description for multiclass pattern classification with reduced computational load. *Pattern Recogn* 2008; 41(6): 1997–2009.

28. Lee S-W, Park J and Lee S-W. Low resolution face recognition based on support vector data description. *Pattern Recogn* 2006; 39(9): 1809–1812.

29. Rico-Juan JR and Inesta JM. Adaptive training set reduction for nearest neighbor classification. *Neurocomputing* 2014; 138: 316–324.

30. Zhu F and Wei JF. A new SVM reduction strategy of large-scale training sample sets. In: *Proceedings of the 4th international conference on manufacturing science and technology (ICMST 2013)*, Dubai, UAE, 3–4 August 2013, pp.816–817, pp.512–515. Zurich: Trans Tech Publications Ltd.

31. Platt JC. Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges C and Smola A (eds) *Advances in kernel methods—support vector learning*. Cambridge, MA: MIT Press, 1999, pp.185–208.

32. Fan R-E, Chen P-H and Lin C-J. Working set selection using second order information for training support vector machines. *J Mach Learn Res* 2005; 6: 1889–1918.

33. Roweis ST and Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000; 290(5500): 2323–2326.

34. Tenenbaum JB, Silva V and Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000; 290(5500): 2319–2323.

35. Jeffreys H and Jeffreys BS. Mean-value theorems. In: Jeffreys H and Jeffreys BS (eds) *Methods of mathematical physics*. 3rd ed. Cambridge: Cambridge University Press, 1988, pp.49–50.

36. Frank A and Asuncion A. UCI machine learning repository, 2010, http://mlearn.ics.uci.edu/MLRepository.html

37. Kubat M and Matwin S. Addressing the curse of imbalanced training sets: one-sided selection. In: *Proceedings of the 14th international conference on machine learning*, Nashville, TN, 8–12 July 1997. Burlington, MA: Morgan Kaufmann.

38. Wu G and Chang EY. Aligning boundary in kernel space for learning imbalanced dataset. In: *Proceedings of the 4th IEEE international conference on data mining (ICDM 2004)*, Brighton, 1–4 November 2004, pp.265–272. New York: IEEE Computer Society.

39. IBRL dataset, 2012, http://db.lcs.mit.edu/labdata/labdata.html

40. Suthaharan S, Alzahrani M, Rajasegarar S, et al. Labelled data collection for anomaly detection in wireless sensor networks. In: *Proceedings of the 2010 6th international conference on intelligent sensors, sensor networks and information processing (ISSNIP)*, Brisbane, QLD, Australia, 7–10 December 2010. New York: IEEE.