Comprehensive Outlier Detection in Wireless Sensor Network with Fast Optimization Algorithm of Classification Model

Haiqing Yao, Heng Cao, and Jin Li

School of Mechanical and Power Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

Correspondence should be addressed to Heng Cao; hengcao@ecust.edu.cn

Received 24 April 2015; Revised 16 June 2015; Accepted 17 June 2015

Copyright © 2015 Haiqing Yao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Since the nonstationary distribution of the detected objects is general in the real world, the accurate and efficient outlier detection for data analysis within wireless sensor network (WSN) is a challenge. Recently, with high classification precision and affordable complexity, one-class quarter-sphere support vector machine (QSSVM) has been introduced to deal with the online and adaptive outlier detection in WSN. Regarding the one-sided consideration of optimization or iterative updating algorithm for QSSVM model within current techniques, we have proposed comprehensive outlier detection methods in WSN based on the QSSVM algorithm. To reduce the complexity of optimization algorithm for QSSVM model in existing techniques, a fast optimization algorithm based on average Euclidean distance has been developed and employed to the comprehensive outlier detection method. Evaluated by real and synthetic WSN data sets, our methods have shown an excellent outlier detection performance, and they have been proved to meet the requirements of online adaptive outlier detection in the case of nonstationary detection tasks of WSN.

1. Introduction

Mainly due to the development of modern information and electronics technology, WSN has been widely used in smart home, logistics, industrial detection, and automation fields with the advantages of low cost and miniaturization. Its major limitations lie in the finite processing and storage capacity, limited energy, and short wireless communication distance and bandwidth [1, 2]. As a result, WSN is extremely sensitive to the interference and attack in complex application environments and thus generates outlier. Outlier or anomaly (these words are often used interchangeably) refers to the observations that seem to be inconsistent with the rest of a data set [3, 4]. Outlier detection is a process of finding data patterns that deviate from expected behaviors [5]. In the real world, the statistical properties of monitoring objects drift over time and lead to unpredictable changes in the collections of WSN. To ensure high outlier detection accuracy and efficiency under these situations, online adaptive outlier detection algorithms with model update, optimization capability, and low computation complexity are required. Additionally, the collection, processing, and storage of the unlabeled data in WSN should be in real time. Therefore, the optimization and adaptive update of unsupervised outlier detection model in WSN become popular in recent researches.

As an unsupervised learning algorithm, QSSVM has excellent classification accuracy and moderate computation complexity and has been introduced into the adaptive outlier recognition in WSN. In QSSVM, parameter \( \nu \) can significantly affect the classification accuracy and further represent the outlier rate of training set (ORTS) [6]. Therefore, the optimal QSSVM model requires searching for optimal parameter \( \nu \) or ORTS. Additionally, the optimal parameter \( \nu \) can be obtained by Golden Section Search algorithm on the standard deviation (SD) of normal data in training set [7]. Considering that outlier is minority in WSN data set and has temporal continuity, accurate and efficient model updating strategies should be involved in outlier detection algorithm. So far, a strategy taking parameter of previous classification model as the updating criteria of QSSVM model has been proposed, and it has proven good performance [8]. Although the two strategies still have some limitations [7, 8], they represent the latest development of outlier detection algorithms based on QSSVM.
2. Related Works

As an outlier detection technology based on classification idea [9], support vector machine (SVM) is a supervised learning method. It classifies normal data and outlier in training set with maximum classification interval by constructing optimal hyperplane. Based on SVM, one-class SVM is an unsupervised learning algorithm. It introduces nonlinear function mapping the training set into a high dimensional feature space and then constructs hyperplane [10] or hypersphere [11] in the feature space to classify the mapped training set. One disadvantage of building one-class SVM model is that it takes up lots of computing resources and time to solve a quadratic optimization problem. Taking into account the one-sided nature of WSN data, Laskov et al. further developed one-class SVM to QSSVM. By locating the center of hypersphere at the origin of feature space, QSSVM requires solving a linear optimization problem to obtain classification model, which greatly reduces the computation complexity [6].

To further meet the challenges of outlier detection in WSN monitoring nonstationary distribution objects, recent researches have been focused on the efficient adaptive optimization algorithm and update strategy of QSSVM model. Rätsch et al. proposed that the optimal parameter $\nu$, which can be obtained when the QSSVM model classifies the normal data and outlier with maximum separation, represents the proportion of outliers in training set [12]. Rajasegarar et al. introduced QSSVM to WSN outlier detection [13] and researched the relationship between the performance of QSSVM classification model and parameter $\nu$ [14]. Inspired by these works, O’Reilly et al. proposed an optimization algorithm of parameter $\nu$ and extended it into an adaptive optimization algorithm of QSSVM model [7]. However, this method does not involve an iterative update strategy and has high computation complexity. Zhang et al. proposed an effective iterative update strategy of QSSVM model, which is an adaptive outlier detection (AOD) technique [8]. However, AOD does not involve an adaptive optimization algorithm of QSSVM model, and it has low outlier detection accuracy when the monitoring objects have nonstationary distribution.

In this paper, we focus on the outlier detection at the node level of WSN. Firstly, we have proposed comprehensive outlier detection methods with respect to previous literatures by considering both adaptive parameter optimization and iterative update strategy of QSSVM classification model. Secondly, to further improve the efficiency of searching for optimal QSSVM model, a fast optimization algorithm employing the relationship between the statistical characteristics and the ORTS of training set has been proposed. The state-of-the-art optimization algorithm of QSSVM model proposed by O’Reilly et al. repeatedly solves the linear optimization problem to get the optimal QSSVM model and requires a lot of computation time. Finally, we have verified these comprehensive outlier detection methods by real and synthetic data sets with different dimension and distribution. The results show that they have excellent outlier detection performance and satisfy the outlier detection requirements of WSN with nonstationary distribution.

3. Problem Statement and Formulations

3.1. Assumption and Problem Statements. At the node level, the nonstationary distribution of monitoring objects can affect outlier detection from two aspects [15]. On the one hand, fluctuation of normal data distribution can affect the classification boundary of QSSVM model. On the other hand, change in the percentage of normal data or outlier can affect the choice of QSSVM model parameters. In order to simulate the two nonstationary distributions of WSN data set, changes in ORTS and the mean of normal data are introduced in our estimations.

3.2. QSSVM. As the basic algorithm of our comprehensive outlier detection strategies, the theoretical derivation and calculation equations of QSSVM have been detailed in the previous studies [6–8]. These theories would be cited in this paper and not detailed again.

3.3. Adaptive Outlier Detection (AOD). AOD is an online iterative updating algorithm of QSSVM and employs the spatiotemporal correlation of WSN [8]. Here, we only introduce the strategy of AOD at the node level. As shown in Figure 1, sensor node collects $n$ samples $\{x_1, x_2, \ldots, x_n\}$ as the initial training set. With a certain parameter $\nu$, the initial classification hypersphere and radius $R$ of QSSVM model on this training set can be obtained. $x_m$ is the sample value at

![Figure 1: Sequence diagram of AOD. AOD: adaptive outlier detection.](image)
subsequent time interval \( m \) \((m > n)\). Based on the previous training set \([x_1, x_2, \ldots, x_m]\), \(d(x_m)\) can be calculated by (1). It is the distance from origin to \(x_m\) in feature space formed by previous QSSVM model:

\[
d(x_m) = \sqrt{k(x_m, x_m) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) - 2 \sum_{i=1}^{n} k(x_m, x_i)},
\]

where the linear kernel \(k(x_i, x_j)\) is \((\phi(x_i) \cdot \phi(x_j))\). Whether \(x_m\) is an outlier can be identified as follows.

1. \(x_m\) is normal data: if \(R > d(x_m)\).
2. \(x_m\) is a support vector or outlier: if \(R \leq d(x_m)\).

When \(x_m\) is normal data, sensor node will continue to determine the status of subsequent data based on the previous classification model and keep the original classification model and training set unchanged. When \(x_m\) is an outlier or support vector, new classification model based on new training set \([x_{m-n+1}, x_{m-n+2}, \ldots, x_m]\) will be constructed. The classification model will continue to be updated with continuous measurement. Compared to normal data, outlier is very rare in WSN [11]. Therefore, AOD improves the computation efficiency with the excellent updating strategy. However, employing a fixed parameter \(v\), the outlier detection accuracy of AOD will degenerate when the monitoring objects have nonstationary distribution.

3.4. Online and Adaptive QSSVM Optimization Algorithm.

To continuously keep the optimal classification accuracy, an adaptive optimization algorithm of QSSVM model should be introduced. When a training set has been collected, the SD of distance from the normal data to the origin in feature space is calculated at each construction of QSSVM model with \(0 \leq v \leq 0.5\) and \(\Delta v = 0.01\) [7, 8, 14]. O’Reilly et al. found that a knee point appears in the SD curve with the linear kernel and RBF kernel functions [7]. The corresponding \(v\) of the knee point is nearby the ORTS, and it is the optimal parameter \(v\). A unimodal relationship of converted SD-\(v\), where the converted SD reaches its maximum at the optimal parameter \(v\), can be obtained by coordinate transformation. To quickly find the optimal parameter \(v\), O’Reilly et al. introduced the Golden Section Search algorithm into the unimodal relationship of converted SD-\(v\) and constructed an online and adaptive QSSVM optimization (OAQO) algorithm. However, OAQO algorithm needs to repeatedly solve linear programming problem at each training set to search for optimal QSSVM model and has high computation complexity. Besides, with a fixed sliding window, which has been verified to be inefficient [8], to update QSSVM model, OAQO algorithm will miss part of outlier when the fixed sliding window is too wide and mistakenly judge part of normal data as outlier introducing unnecessary model update when the fixed sliding window is too narrow [15]. Therefore, OAQO algorithm lacks efficient iterative update algorithm for outlier classification model.

4. Comprehensive Outlier Detection Algorithm in WSN

First of all, we make a statement that all the programs in this paper were developed and evaluated using MATLAB; hardware of computer is Inter Core i3 CPU with 2.5 GHz clock.

4.1. Comprehensive Outlier Detection (COD) Algorithm.

Firstly, we proposed a comprehensive outlier detection (COD) algorithm based on QSSVM with optimization and iterative update capabilities uniting AOD and OAQO algorithms. Using the model update strategy of AOD, COD adopts the Golden Section Search algorithm of OAQO to search for the optimal classification model of training set at each iteration. In addition, we convert the original SD-\(v\) curve to a unimodal relationship, where the converted SD reaches its minimum at optimal parameter \(v\). This will be detailed in the next section. The pseudocode of COD algorithm is detailed in Pseudocode 1.

4.2. Fast Comprehensive Outlier Detection (FCOD) Algorithm.

Inspired by O’Reilly et al. [7], we have found the relevance between the mean of Euclidean distance (MED) of one piece of data and the rest of the data in training set and ORTS. First of all, for a training set \(X\) with a data capacity of \(n\), the MED \(_j\) of each piece of data \(x_i\) \((x_i \in X, 1 \leq i \leq n)\) is arranged in descending order. Then, for each MED \(_j\), MED\(_{\text{num}}\), is calculated as the number of MED \(_j\) \((1 \leq j \leq n, i \neq j)\) whose value is greater than or equal to MED\(_j\). The proportion MED\(_{\text{Rate}}\) of corresponding MED\(_j\) is MED\(_{\text{num}}/n\). Then a descending relationship of MED-MED\(_{\text{Rate}}\) is acquired, where a knee point appears. This phenomenon can be explained as follows: taking into account that outlier is always away from the densely distributed area of normal data, so the average distance of outlier to other data is always farther than that of normal data, which causes a knee point in the MED-MED\(_{\text{Rate}}\) curve appearing nearby the ORTS. Then, a converted relationship of MED-MED\(_{\text{Rate}}\), where the converted MED gets a minimum value nearby the ORTS, can be obtained by introducing the coordinate transformation idea of OAQO. The original MED\(_{\text{Rate}}\) corresponding to the minimum converted MED is the estimated optimal parameter \(v\) of QSSVM model. With the descending relationship of MED-MED\(_{\text{Rate}}\), two data sets \((1/n, \text{MED}\(_{\text{max}}\))\) and \((1, \text{MED}\(_{\text{min}}\))\) are constituted, and then the rotation angle \(\theta = \arctan((1-1/n)/(\text{MED}\(_{\text{max}}-\text{MED}\(_{\text{min}}\)))\) is calculated. Finally, the conversion relation is detailed in

\[
\begin{bmatrix}
\text{Converted MED} \\
\text{Converted MED}_{\text{Rate}}
\end{bmatrix} =
\begin{bmatrix}
\cos\theta & -\sin\theta \\
\sin\theta & \cos\theta
\end{bmatrix}
\begin{bmatrix}
\text{MED} \\
\text{MED}_{\text{Rate}}
\end{bmatrix}.
\]

As shown in Figure 2(a), we introduce two-dimensional synthetic data set \(s\) with a data capacity of 100. The mean of normal data and outlier is 0.4 and 0.6, respectively. Both normal data and outlier are normally distributed with
the same SD of 0.03. All synthetic data has been aligned in a random order. The ORTS of sl is 0.05. Then, a descending relationship of MED-MED Rate is obtained (see Figure 2(b)), where a knee point appears. In Figure 2(c), the converted MED reaches its minimum when its original MED is equal to 0.051. Moreover, 0.051 is in the vicinity of correct ORTS, and it can be used as the optimal parameter \( \nu \) to optimize the QSSVM classification model. By introducing this optimization algorithm of QSSVM model based on the mean of Euclidean distance (OQMED) into COD, a fast online and adaptive outlier detection (FCOD) algorithm with optimization and efficient update of classification model is obtained, whose pseudocode is detailed in Pseudocode 2. As shown in Pseudocode 1, the Golden Section Search algorithm employs many solutions of linear programming problem with a computation complexity of \( O(n^3) \), where \( n \) is the total number of data in training set [13]. However, the proposed OQMED algorithm employs some simple solutions to rearrange all MED in a descending order, convert them, and find the minimum converted MED with a computation complexity of \( O(n^2) \) (see Pseudocode 2). Therefore, the proposed OQMED algorithm has better computational efficiency than the Golden Section Search algorithm.

To estimate the accuracy of model optimization, we employed two methods based on real and synthetic WSN data sets. The first strategy, as an extension of Golden Section Search adopted by COD, sets the range of parameter \( \nu \) within (0.01, 0.51), whose iterative step is \( \Delta \nu = 0.02 \). Then, a SD-\( \nu \) curve can be obtained, where the optimal estimation of ORTS can be found near a parameter \( \nu \) corresponding to the minimum converted SD. Linear kernel function is adopted in the QSSVM model. The other method adopts the OQMED to search for the optimal estimation of ORTS.

We have used the data sets from SensorScope WSN system deployed in Grand-st-Bernard [16]. In these experiments, we intercept a set of data recorded during 6:00

---

PSEUDOCODE 1: Pseudocode of COD algorithm. COD: comprehensive outlier detection algorithm; ORTS: the outlier rate of training set; SD: the standard deviation.

(1) Sensor node collects \( n \) data measurements as training set;
(2) Golden Section Search function is called to search for optimal parameter \( \nu \);
(3) The optimal model is constructed;
(4) The radius \( R \) of classification sphere is obtained;
(5) Sensor node collects a new measurement \( x_m \) (\( m > n \));
(6) \( d(x_m) \) is calculated;
(7) if \( d(x_m) \geq R \) go to (2);
(8) \( x_m \) is an outlier or support vector;
(9) A new training set is constructed by \( x_m \) with its previous and continuous (\( n - 1 \)) measurements
\[ [x_{m-n+1}, x_{m-n+2}, \ldots , x_{m-1}] \];
(11) Returns to (2);
(12) else
(13) \( x_m \) is a normal data;
(14) Returns to (5);
Function Golden Section Search(\( X \))
(1) Set the ORTS within [0.01, 0.5] and the tolerance \( \nu_T = 0.01 \);
(2) \( \nu_{min} = 0.01, \nu_{max} = 0.5, \nu_1 = \nu_{min} + 0.382(\nu_{max} - \nu_{min}), \nu_2 = \nu_{min} + 0.618(\nu_{max} - \nu_{min}) \);
(3) The SD\_min and SD\_max of normal data at \( \nu_{min} \) and \( \nu_{max} \) are calculated by solving two linear programming problems;
(4) The transformation angle: \( \theta = \arctan((\text{SD}\_\text{max} - \text{SD}\_\text{min})/\text{(SD}_{\text{max}} - \nu_{\text{min}})) \) is calculated;
(5) The transformation relationship:
\[
\begin{align*}
\text{Converted SD} &= \left[ \begin{array}{c}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta 
\end{array} \right] \text{SD} \\
\text{Converted } \nu &= \left[ \begin{array}{c}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta 
\end{array} \right] \nu
\end{align*}
\]
(6) if \( (\nu_{max} - \nu_{min}) > \nu_T \)
(7) The SD\_1 and SD\_2 of normal data at \( \nu_1 \) and \( \nu_2 \) are calculated by solving two linear programming problems;
(10) The converted SD\_1 and SD\_2 are calculated;
(11) if converted SD\_1 > converted SD\_2
(12) \( \nu_{min} = \nu_1 \);
(13) \( \nu_1 = \nu_{min} + 0.382(\nu_{max} - \nu_{min}), \nu_2 = \nu_{min} + 0.618(\nu_{max} - \nu_{min}) \);
(14) Returns to (7);
(15) else if converted SD\_1 < converted SD\_2
(16) \( \nu_{max} = \nu_2 \);
(17) \( \nu_1 = \nu_{min} + 0.382(\nu_{max} - \nu_{min}), \nu_2 = \nu_{min} + 0.618(\nu_{max} - \nu_{min}) \);
(18) Returns to (7);
(19) else
(20) Go to (21);
(21) The optimal parameter \( \nu_{\text{optimal}} = \nu_{min} + (\nu_{\text{max}} - \nu_{\text{min}})/2 \) is calculated and returned.
a.m.–6:00 p.m. on 20 September 2007 with two attributes: ambient temperature and relative humidity of sensor nodes 2, 3, and 4. Temperature and humidity data in each node have been normalized within (0, 1) and constituted to a two-dimensional data set. As shown in Figures 3(a) to 3(c), optimized by the first strategy, the estimated ORTS of sensor nodes 2, 3, and 4 are all around 0.2. In Figures 3(d) to 3(f), the converted MED-MED_Rate curve corresponding to each sensor node is obtained by OQMED, from which the estimated ORTS of sensor nodes 2, 3, and 4 is 0.28, 0.19, and 0.25, respectively.

As shown in Figure 4, the ORTS of three two-dimensional synthetic data sets s2, s3, and s4 with the same data capacity of 300 is estimated by the two methods. The mean of normal data and outlier are normally distributed with a SD of 0.03. Each synthetic data set has been aligned in a random order. The ORTS of synthetic data sets s2, s3, and s4 is 0.05, 0.2, and 0.35, respectively. The estimated ORTS of synthetic data sets s2, s3, and s4 is 0.05, 0.2, and 0.29, respectively, after they are optimized by the first strategy (see Figures 4(a) to 4(c)). The converted MED-MED_Rate curve corresponding to each synthetic data set is obtained by OQMED, from which the estimated ORTS of synthetic data sets s2, s3, and s4 is 0.05, 0.21, and 0.39, respectively (see Figures 4(d) to 4(f)).

Obviously, for the estimation of ORTS, the first optimization algorithm is more accurate than the OQMED algorithm. However, the estimation accuracy of the first method using linear kernel function will degenerate when the ORTS is greater than or equal to 0.35 [7]. Finally, the converted
MED-MED_Rate curve calculated by the OQMED algorithm is not always a unimodal relationship, but it can always acquire a minimum converted MED nearby the ORTS. The original MED_Rate of the minimum converted MED is the approximate optimal estimation of ORTS and can be used as the optimal parameter ν of QSSVM model.

5. Experimental Results and Evaluation

In this section, we will further evaluate the performance of COD and FCOD algorithms based on QSSVM with linear kernel function from the aspects of outlier detection rate (DR), false positive rate (FPR), and computation complexity. Outlier detection rate is the rate of true positives to outlier, and false positive rate is the rate of false positives to normal data. In the real world, the nonstationary distribution of monitoring objects is continuous, and the ORTS of WSN data set is usually less than 0.15–0.35 [17]. Therefore, synthetic data sets with continuously changing mean and ORTS are introduced, and the ORTS of synthetic data sets introduced in the paper are all within [0, 0.35].

5.1. Further Validation of the Assessment Accuracy of ORTS

To further validate the estimation accuracy of ORTS with Golden Section Search and OQMED algorithms when the ORTS of WSN data set is within [0, 0.35], we have introduced synthetic data sets s5–s13 with normal distribution and synthetic data sets s14–s22 with uniform distribution. Synthetic data sets s5–s13 have a fixed SD of 0.03 with normal distribution. Each synthetic data set has the same data capacity of 550, and it is divided into 11 data segments with 50 pieces of data in each data segment. Each data segment of these synthetic data sets has been aligned in a random order. Initially, for s5–s13, the ORTS of data segment one is 0.05, and the mean of normal data and outlier of data segment one is 0.4 and 0.6, respectively. For s14–s22, the ORTS of data segment one is 0.05, the normal data of data segment one is uniformly distributed within [0.3, 0.5], and the outlier of data segment one is uniformly distributed within [0.5, 0.7]. Finally, in data segment 11 of s5–s22, the ORTS is 0.35. The constitution rules of s5–s22 are detailed in Table 1. The change rate of ORTS is +0.03 per subsequent data segment to change the value of ORTS from 0.05 to 0.35. The change rate of mean of the normal data and outlier is ± 0.02 to keep the value of the normal data and outlier within [0, 1]. To remove the effect of random factors, a total of 50 constructions and computations are repeated for each synthetic data set.

As shown in Figure 5, the simultaneous change of mean of normal data and outlier cannot affect the estimation accuracy of the two methods. For s5–s13 with normal distribution, the estimation accuracy of Golden Section Search is more accurate than that of QMED in most cases, and the estimation accuracy for the two algorithms will be more precise with higher data dimension. For Golden Section Search algorithm, its estimation accuracy degenerates when detecting low dimension data with high ORTS. However, this phenomenon has not appeared in the proposed QMED algorithm. Finally, the estimated optimal ORTS of OQMED algorithm is always larger than the correct ORTS and that of Golden Section Search algorithm.

As shown in Figure 6, similarly to normal distribution data sets s5–s13, for s14–s22, the estimation accuracy of
Figure 3: Optimal estimation of ORTS in real WSN data sets: ((a), (b), and (c)) original and converted SD-\(\nu\) curve by iterative calculation of QSSVM; ((d), (e), and (f)) original and converted MED-MED\_rate curve by OQMED. SD: standard deviation; MED: mean of Euclidean distance; MED\_rate: corresponding rate of MED; ORTS: outlier rate of training set; QSSVM: one-class quarter-sphere support vector machine; OQMED: optimization of QSSVM model based on the mean of Euclidean distance.
Figure 4: Optimal estimation of ORTS in synthetic data sets: ((a), (b), and (c)) original and converted SD-\(\nu\) curve by iterative calculation of QSSVM; ((d), (e), and (f)) original and converted MED-MED_rate curve by OQMED. SD: standard deviation; MED: mean of Euclidean distance; MED_rate: corresponding rate of MED; ORTS: outlier rate of training set; QSSVM: one-class quarter-sphere support vector machine; OQMED: optimization of QSSVM model based on the mean of Euclidean distance.
Figure 5: Continued.
Figure 5: Optimal estimation of ORTS in synthetic data sets $s_5$–$s_{13}$ by Golden Section Search and OQMED algorithms: (a)–(i) optimal estimation of ORTS of synthetic data sets $s_5$–$s_{13}$. ORTS: outlier rate of training set; QSSVM: one-class quarter-sphere support vector machine; OQMED: optimization of QSSVM model based on the mean of Euclidean distance.

Table 1: Constitution rules of synthetic data sets $s_5$–$s_{13}$ with normal distribution and synthetic data sets $s_{14}$–$s_{22}$ with uniform distribution.

| Synthetic data set | $\Delta$ mean of normal data and outlier/each subsequent data segment | $\Delta$ ORTS/each subsequent data segment | Dimension |
|--------------------|------------------------------------------------|---------------------------------|----------|
| $s_5, s_{14}$      | Unchanged                                      | +0.03                           | 2        |
| $s_6, s_{15}$      | +0.02                                          | +0.03                           | 2        |
| $s_7, s_{16}$      | −0.02                                          | +0.03                           | 2        |
| $s_8, s_{17}$      | Unchanged                                      | +0.03                           | 6        |
| $s_9, s_{18}$      | +0.02                                          | +0.03                           | 6        |
| $s_{10}, s_{19}$   | −0.02                                          | +0.03                           | 6        |
| $s_{11}, s_{20}$   | Unchanged                                      | +0.03                           | 10       |
| $s_{12}, s_{21}$   | +0.02                                          | +0.03                           | 10       |
| $s_{13}, s_{22}$   | −0.02                                          | +0.03                           | 10       |

ORTS: outlier rate of training set.
Golden Section Search is much better than that of OQMED in most cases, and the estimation accuracy for the two algorithms would be more precise with higher data dimension. However, the estimation accuracy for both Golden Section Search algorithm and OQMED algorithm degenerates when detecting low dimension data with high ORTS. For two synthetic data sets (one with normal distribution and the other with uniform distribution), the number of data within the boundary between normal data and outlier of uniformly distributed random data set is more than that of normally distributed random data set. This can lead to the knee point (minimum converted MED point in OQMED) deviate from the correct ORTS and cause performance degradation.

Therefore, the estimation performance of the two algorithms changes with different distribution data. Finally, for synthetic data sets s14~s22 with uniform distribution, the estimated optimal ORTS of OQMED algorithm is always larger than that of Golden Section Search algorithm.

As a result, the QSSVM classification model obtained by OQMED algorithm is always smaller and tighter (in the high dimensional feature space) than that obtained by Golden Section Search algorithm, and more data in the boundary between normal data and outlier is classified as outlier by OQMED algorithm. Considering that most of the false positive outlier is in the boundary between normal data and outlier, the DR performance of OQMED algorithm is...
Table 2: Constitution rules of synthetic data sets s23~s46 with normal distribution and synthetic data sets s47~s70 with uniform distribution.

(a) 

| Synthetic data set | Δmean of normal data and outlier/each subsequent data segment | ΔORTS/each subsequent data segment |
|--------------------|---------------------------------------------------------------|----------------------------------|
| s23, s31, s39, s47, s55, s63 | Unchanged | +0.03 |
| s24, s32, s40, s48, s56, s64 | Unchanged | −0.03 |
| s25, s33, s41, s49, s57, s65 | +0.05 | Unchanged |
| s26, s34, s42, s50, s58, s66 | −0.05 | Unchanged |
| s27, s35, s43, s51, s59, s67 | +0.05 | +0.03 |
| s28, s36, s44, s52, s60, s68 | +0.05 | −0.03 |
| s29, s37, s45, s53, s61, s69 | −0.05 | +0.03 |
| s30, s38, s46, s54, s62, s70 | −0.05 | −0.03 |

(b) 

| Synthetic data set | Dimension |
|--------------------|-----------|
| s23, s24, s25, s26, s27, s28, s29, s30, s47, s48, s49, s50, s51, s52, s53, s54 | 2 |
| s31, s32, s33, s34, s35, s36, s37, s38, s55, s56, s57, s58, s59, s60, s61, s62 | 6 |
| s39, s40, s41, s42, s43, s44, s45, s46, s63, s64, s65, s66, s67, s68, s69, s70 | 10 |

ORTS: outlier rate of training set.

better than that of Golden Section Search algorithm. Similarly, considering that more normal data in the boundary is misclassified as outlier by OQMED algorithm, the FPR performance of OQMED algorithm is worse than that of Golden Section Search algorithm. Therefore, the proposed OQMED algorithm has a better classification accuracy than the Golden Section Search algorithm when detecting outlier as much as possible is highly valued.

5.2. Assessment of the Comprehensive Performance of the New Algorithm

We have introduced synthetic data sets s23~s46 with normal distribution and synthetic data sets s47~s70 with uniform distribution to analyze the detection performance of COD and FCOD algorithms. Synthetic data sets s23~s46 have a fixed SD of 0.03 with normal distribution. Each synthetic data set has the same data capacity of 300, and it is divided into 6 data segments with 50 pieces of data in each data segment. Each data segment of these synthetic data sets has been aligned in a random order. Initially, for s23~s46, the ORTS of data segment one is 0.2, and the mean of normal data and outlier of data segment one is 0.4 and 0.6, respectively. For s47~s70, the ORTS of data segment one is 0.2 too, the normal data of data segment one is uniformly distributed within [0.3, 0.5], and the outlier of data segment one is uniformly distributed within [0.5, 0.7]. The constitution rules of s23~s70 are detailed in Table 2. The change rate of ORTS is ±0.03 per subsequent data segment to keep the value of ORTS within [0, 0.35]. The change rate of mean of the normal data and outlier is ±0.05 to keep the value of the normal data and outlier within [0, 1].

Firstly, we introduce the estimation of ORTS in s23~s30 (synthetic data sets s31~s70 have similar properties) to determine whether these synthetic data sets can explain the continuous change of ORTS of WSN in the real world. Then, the ORTS of synthetic data sets is evaluated by calculating the outlier rate within 50 pieces of data before each piece of data. A total of 50 constructions and computations are repeated for each synthetic data set to remove the effect of random factors.

As shown in Figure 7, the ORTS of s23~s30 change continuously and can explain a variety of distribution changes of WSN data set in the real world (the same as s31~s70). Based on the analysis, we adopt COD and FCOD algorithms to detect the outlier of s23~s70. Similarly, a total of 50 constructions and computations are repeated for each synthetic data set to remove the effect of random factors.

As shown in Figures 8 and 9, for these synthetic data sets with both normal distribution and uniform distribution, FCOD has a much better DR performance of outlier than COD, but its FPR performance is worse than that of COD. This can be explained as follows: the optimal parameter γ of QSSVM model calculated by OQMED of FCOD is always much larger than that computed by Golden Section Search of COD (see Figures 5 and 6). Therefore, the classification hypersphere of QSSVM model calculated by FCOD is always much smaller than that obtained by COD. As a result, more data in the boundary between the normal data and outlier is classified as outlier by FCOD algorithm. Considering that most of the false positive outlier is in the boundary between normal data and outlier, the DR performance of FCOD algorithm is better than that of COD algorithm. Similarly, considering that more normal data in the boundary is misclassified as outlier by FCOD algorithm, the FPR performance of FCOD algorithm is worse than that of COD algorithm. Therefore, the FCOD algorithm has a better classification accuracy than the COD algorithm when detecting outlier as much as possible is highly valued. Additionally, considering the oversampling in data collection of WSN, a small part of
Figure 7: Continued.
normal data can be considered as outlier without a serious impact on the overall detection of objects.

Simultaneously, for both COD and FCOD algorithms, the DR performance of synthetic data sets, whose mean has downward trend, is worse than that of other synthetic data sets (see the average detection rate of outlier in Figures 8 and 9). Additionally, the DR performance of these synthetic data sets becomes even worse on high dimension data. This phenomenon can be explained as follows: firstly, both the COD and FCOD algorithms have the same online iterative updating algorithm of QSSVM based on the AOD algorithm. New data is determined whether it is an outlier or support vector by previous QSSVM model. If the new data is an outlier or support vector, a new QSSVM model is constructed by the new data with its previous and continuous \((n - 1)\) data. Secondly, we set the value of outlier larger than that of normal data in all synthetic data sets. Simultaneously, for the synthetic data sets whose mean has downward trend, the value of subsequent data is smaller than the previous data. Therefore, the value of new outlier tends to be the same as the value of previous normal data, and the new outlier cannot be distinguished from the previous normal data. Additionally, the admixture of the previous normal data and the new outlier becomes even worse on high dimension data set. As a result, the DR performance of COD and FCOD algorithms on these data sets is low.

For FCOD algorithm, much bigger FPR than COD algorithm can also lead to more constructions of QSSVM model (see the average training number in Figures 8 and 9). However, without repeatedly solving the linear programming problem in model optimization process, FCOD spends less time on computing than COD on the same synthetic data sets (see the average training time in Figures 8 and 9). Therefore, compared with COD algorithm, FCOD is a more efficient and fast outlier detection algorithm for WSN.

5.3. Computation Complexity. The average training time is quite different between FCOD and COD (see the average training time in Figures 8 and 9), which indicates the complexity of FCOD algorithm is much lower than that of COD algorithm. The complexity of linear programming algorithm in QSSVM is \(O(n^3)\), where \(n\) is the total number of data in training set [13]. Therefore, both COD and FCOD based on QSSVM algorithm have a complexity of \(O(n^3)\). However, the Golden Section Search algorithm of COD needs to repeatedly solve QSSVM model with a complexity of \(O(n^3)\) in model optimization process, while the OQMED algorithm of FCOD only has a complexity of \(O(n^2)\). As a result, we can save the computational cost by using FCOD algorithm.

6. Conclusions

For the WSN system monitoring objects with nonstationary distribution, we proposed two online and adaptive outlier detection algorithms (COD and FCOD) based on QSSVM with model optimization and iterative update strategies. The characteristics of both algorithms are evaluated by real and synthetic WSN data sets. COD algorithm is an integration of existing advanced outlier detection algorithms. The COD algorithm acquires excellent outlier detection accuracy with outstanding FPR performance. As an expansion of the COD algorithm, FCOD algorithm combines the proposed fast optimization algorithm of QSSVM model. The FCOD algorithm can perform more efficient and more precise outlier detection task than the COD algorithm. However, the FCOD algorithm has degeneration in the FPR performance. As for the computational efficiency, the FCOD algorithm takes less resources and time than the COD algorithm. It can be concluded that the FCOD algorithm would be the recommended algorithm when the computational efficiency

Figure 7: Average ORTS of synthetic WSN data sets: (a)–(h) synthetic data sets s23–s30. ORTS: outlier rate of training set.
Figure 8: Continued.
Figure 8: Outlier detection performance in synthetic WSN data sets $s_{23}$–$s_{46}$ with COD and FCOD: ((a), (e), and (i)) average detection rate of outlier, ((b), (f), and (j)) average false positive rate, ((c), (g), and (k)) average training number, and ((d), (h), and (l)) average training time. DR: detection rate; FPR: false positive rate; COD: comprehensive outlier detection algorithm; FCOD: fast comprehensive outlier detection algorithm.
Figure 9: Continued.
Figure 9: Outlier detection performance in synthetic WSN data sets s47~s70 with COD and FCOD: ((a), (e), and (i)) average detection rate of outlier, ((b), (f), and (j)) average false positive rate, ((c), (g), and (k)) average training number, and ((d), (h), and (l)) average training time. DR: detection rate; FPR: false positive rate; COD: comprehensive outlier detection algorithm; FCOD: fast comprehensive outlier detection algorithm.
and DR performance of WSN outlier detection are highly valued. Finally, our future work will focus on the development of WSN outlier detection algorithms with more balanced and comprehensive performance.

**Conflict of Interests**

The authors declare that there is no conflict of interests regarding the publication of this paper.

**Acknowledgment**

This work was funded in part by the National Natural Science Foundation of China (Grant 51275170).

**References**

[1] J. Yick, B. Mukherjee, and D. Ghosal, “Wireless sensor network survey,” *Computer Networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[2] H. Kim and J. K. Min, “An energy-efficient outlier detection based on data clustering in WSNs,” *International Journal of Distributed Sensor Networks*, vol. 2014, Article ID 619313, 7 pages, 2014.

[3] M. A. Rassam, A. Zainal, and M. A. Maarof, “Advancements of data anomaly detection research in wireless sensor networks: a survey and open issues,” *Sensors*, vol. 13, no. 8, pp. 10087–10122, 2013.

[4] V. J. Hodge and J. Austin, “A survey of outlier detection methodologies,” *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[5] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: a survey,” *ACM Computing Surveys*, vol. 41, no. 3, article 15, 58 pages, 2009.

[6] P. Laskov, C. Schäfer, I. Kotenko, and K. Müller, “Intrusion detection in unlabeled data with quarter-sphere Support Vector Machines,” *Praxis der Informationsverarbeitung und Kommunikation*, vol. 27, no. 4, pp. 228–236, 2004.

[7] C. O’Reilly, A. Gluhak, M. Imran, and S. Rajasegarar, “Online anomaly rate parameter tracking for anomaly detection in wireless sensor networks,” in *Proceedings of the 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON ’12)*, pp. 191–199, June 2012.

[8] Y. Zhang, N. Meratnia, and P. J. M. Havinga, “Ensuring high sensor data quality through use of online outlier detection techniques,” *International Journal of Sensor Networks*, vol. 7, no. 3, pp. 141–151, 2010.

[9] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, “Outlier detection for temporal data: a survey,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2014.

[10] 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.

[11] D. M. J. Tax and R. P. W. Duin, “Support vector data description,” *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[12] G. Rätsch, S. Mika, B. Schölkopf, and K.-R. Müller, “Constructing boosting algorithms from SVMs: an application to one-class classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1184–1199, 2002.
Submit your manuscripts at
http://www.hindawi.com