Online Distributed Distance-based Outlier Clearance Approaches for Wireless Sensor Networks

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):

Published in:
Pervasive and Mobile Computing

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester’s Takedown Procedures [http://man.ac.uk/04Y6B6] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Online Distributed Distance-based Outlier Clearance Approaches for Wireless Sensor Networks

Tianwei Dai\textsuperscript{a}, Zhengtao Ding\textsuperscript{a,*}

\textsuperscript{a}Control Systems, Department of Electrical and Electronic Engineering, The University of Manchester, Manchester, M13 9PL, U.K.

Abstract

One key challenge for sensor networks is to provide real-time high reliable sensor data with the minimum resource consumption. Outlier clearance in sensor networks can ensure the quality of sensor data and dependable monitoring. In this paper, we propose two online distributed outlier clearance approaches with low computational complexity and memory usage that can identify and remove the spurious sensor data. The proposed approaches are operated locally and thus save communication overhead as well as possess good scalability. The evaluation performance of proposed approaches and existing widely used methods on synthetic and real-life dataset illustrates that our Adaptive Top-n WAD approach achieves remarkable outlier clearance performance as compared to these methods.

Keywords: Outlier clearance, Wireless sensor networks, Nearest neighbors

1. Introduction

At present, the rapid progress of MEMS technology, sensor technology, embedded computing technology, distributed computation technology, wire-
less communication technology, etc., has promoted the researches and applications of wireless sensor networks (WSNs) forward to a greater and broader scope. Therefore, we are able to witness the successful application and enormous potential of WSNs in multi-disciplinary fields such as target tracking, environmental monitoring, industrial manufacture, smart home, intelligent medical system, traffic control and localization systems [1–3].

A vast quantity of smart sensor nodes comprise WSNs, which are low-power terminal devices with the ability of remote communication and control. In actuality, data acquisition of states of interest and event detection are the main tasks of WSNs. Besides, in most applications of WSNs, it is required to monitor targets and make event classification based on multi-sensor fusion results in real-time. However, owing to the imperfect nature of WSNs, the quality of raw sensor measurements collected by sensor nodes is usually inaccurate and unreliable in practical applications [4]. For one thing, out of cost consideration, a large number of cheap sensor nodes would be employed in networks, the capability and resources of which are stringently limited [5], such as energy supply, internal memory, processing capacity and communication bandwidth, that may lead to collecting questionable and imprecise data. Especially, when the energy supply is out of order or exhausted, the possibility of producing inaccurate data will rise sharply. For another thing, in real applications, numbers of sensor nodes in networks are often randomly or methodically deployed in extremely harsh environments with the enormous complex terrain, which may make some sensor nodes suddenly fail resulting in noise data, spurious data, conflict data, out of sequence data or redundant data [6]. Additionally, sensor nodes may face malicious attacks from the enemy, such as sensor nodes capture, black hole attacks, communication denial of service, sybil and clone ID attacks [7].
Therefore, those issues as mentioned above cause a sharp decline in the quality of raw sensor measurements, which further adversely affect multisensor fusion results and decision-making process. We note that detecting and clearing those spurious and imprecise raw sensor measurements in WSNs become two of the first and foremost prerequisites for successfully ensuring the high-accuracy and high-reliability in monitoring target system and being able to make effective and sound decisions with multisensor fusion data.

In WSNs, the outlier, which is the sensor node that appears to deviate considerably from other members of the sample or is inconsistent with the remainder of the set of data [8], is one kind of erroneous data that would tremendously adversely affect the quality of multisensor fusion results. From these well-known definitions, it is relatively straightforward to detect and clear outliers by defining a normal behavior model of sensor measurement to identify outliers, which do not conform to this model. However, in reality, WSNs is often employed to process real-time streaming sensor measurements, which is unbounded and incapable of being modeled precisely, the previous approach would be too laborious to be implemented. Accordingly, a series of critical challenges faced by WSNs is to design a model-free and online outlier clearance approach with low communication overhead, low memory usage, and low computation cost.

In this paper, we propose two online distributed outlier clearance approaches with low resource consumption. In these approaches, we define a weighted average distance-based outlier factor criterion, inspired by the nearest neighbor rule, exploiting sensor measurements from a particular node and its spatially nearest neighbors in the same sample time, to identify outliers in real-time. And also, the false alarm probability of the proposed criterion is derived, which suggests parameter settings in practical applica-
tions. The evaluation performance with the widely used metric of Receiver Operator Characteristic (ROC) curve [9] on real-life dataset for our proposed approaches and the other typical methods shows that the proposed Top-n WAD and Adaptive Top-n WAD approaches can handle the outliers efficiently and reliably in an online fashion, even if multiple sensor nodes fail at the same given time.

The remainder of the paper is organized as follows. Related works on outlier detection and clearance method in WSNs are presented in Section 2. In Section 3, the problem formulation and principles of the nearest neighbor rule are described. The formal definitions and related properties of our outlier factor, as well as the details of proposed two online distributed outlier clearance approaches, are explained in Section 4 and Section 5, respectively. In Section 6, the processed results on both synthetic and real-life dataset, using our proposed approaches and the other widely used methods, provide the detailed performance evaluation of all methods. Last, we make the concluding remarks in Section 7.

2. Related work

Outlier detection and clearance approaches in WSNs aim to clear and ameliorate the raw sensor measurements in order to deliver the optimal information to the decision maker. According to [4], these approaches can be sorted as centralized approaches, decentralized approaches and hybrid approaches based on the architectural structure. Moreover, on account of the disciplines, we can classify them as nearest neighbor-based approaches, statistical-based approaches, classification-based approaches and artificial intelligence based approaches.
The central type method is one of the most commonly used methodologies in outlier detection and clearance within the small-scale WSNs. It identifies the outliers through the entire sensor measurements stored at a base station. Following this architecture, Curiac et al. build an ensemble system with multiple classifiers, which are selected to implement a complex decisional system in the base station, to clear the sensing anomaly in WSNs [10]. In addition, a central type method, presented in [11], utilizes complex transforms of multivariate time-series to produce accurate models and further enhance its performance in fault detection and clearance.

However, traditional central type methods, such as [10] and [11], cannot deal with the scalability issues, high communication cost and huge memory overhead in the exceptionally large-scale WSNs. Because of the increasing sensor nodes deployed in WSNs, it requires a large amount of storage space and costly processors with great computational capacity as the fusion center. Nevertheless, distributed approaches generally demand less computational power, lower communication cost, and smaller memory footprint [12]. In [13], Branch et al. propose a global and semi-global nearest neighbor-based outlier detection and clearance approaches for the sake of identifying Top-n outliers. Despite requiring the prior knowledge about outliers to define Top-n value, it demands less energy supply and bandwidth compared with the well-known centralized outlier detection algorithm, AODV [14]. Concerning the local outlier in WSNs, some nearest neighbor-based approaches and classification-based techniques can be employed to identify them. A prime example of nearest neighbor-based approaches is that of the Local Outlier Factor (LOF) [15]. LOF-based approaches with the hierarchical network structure, proposed in [16], are employed to identify anomalous sensor measurements. Also, a simplified variation of LOF applied in WSNs is described
in [17]. Abid et al. create one distinctive nearest neighbor-based outlier detector, DNOD [18]. This unsupervised detector can identify outliers without defining the number of nearest neighbors and the distance threshold of accepted neighbors. Nonetheless, its high accuracy of detection is guaranteed by the large size of learning window, which significantly increases the storage and computing. Zhang et al. propose different distributed outlier detection approaches in [19] and [20] using the support vector machine (SVM). They can make each sensor node sequentially update the normal behavior model of itself using meaningful information derived from the spatial and temporal correlations between itself and each of its spatially nearest neighbors in order to identify local outlier. Furthermore, an innovative semi-supervised SVM classifier described in [21] can identify the local outlier of solar power in the large photovoltaic power stations. Although these SVM-based approaches prove a high detection accuracy and low false alarm rate, it has the high computational complexity and memory costs, which restricts the application of real-time scene.

3. Preliminaries

3.1. Problem statement

Let us consider that a homogeneous multisensor networks monitor a target signal $r(t)$, in which there are $H$ identical time synchronized sensor nodes that are stochastically distorted by noise $W$ and outlier $O$. The noise and the outlier are the zero-mean white Gaussian noise and stochastic erroneous data, respectively. Let a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with its adjacency matrix $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{H \times H}$, describe the topology of this homogeneous multisensor networks. $\mathcal{V} = \{1, \ldots, H\}$ denotes the sensor nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents
the set of edges, in which the edge \((v_j, v_i)\) describes the information flow from node \(j\) to node \(i\). \(A\) is defined such that \(a_{ii} = 0\), \(a_{ij}\) is positive if \((v_j, v_i) \in \mathcal{E}\) and \(a_{ij} = 0\) otherwise. Note that \(a_{ij}\) indicates the weight for the information flow, the edge \((v_j, v_i) \in \mathcal{E}\). Then, set \(a_{ij} = 1\) when the weights are not relevant, if \((v_j, v_i) \in \mathcal{E}\).

Figure 1 displays a small multisensor sub-network centered at node \(i\), in which they can directly communicate with each other. Let \(N_i = \{j \in \mathcal{V} : (v_j, v_i) \in \mathcal{E}\}\) denote the set of spatially nearest neighbors of sensor node \(i\), in-degree of which is also the cardinality of \(N_i\), and \(J_i = N_i \cup \{i\}\) denote the set of this multisensor sub-network. In addition, set the degree matrix, \(\mathcal{D} = \text{diag}(d_1, \ldots, d_H)\), for this homogeneous multisensor networks with \(d_i\) as the in-degree number of node \(i\). Our objective is to online distributed identify outliers from every new sensor measurement collected by each sensor node in networks and substitute proper estimated data for them.

![Figure 1: Sensor node i with its five nearest neighbors.](image)

3.2. Nearest neighbor rule

The most popular approaches in the distance-based domain of outlier clearance for WSNs are the nearest neighbor-based methods. The nearest
neighbor rule is a non-parametric method used for classification, which is generally regarded as one of the most mature and straightforward classifiers [22]. Given a set of training data and cases, we can assign a class label to a specific case, if a point is the closest to the case based on the rule.

More specifically, in \( k \) nearest neighbors (KNN) classification approach, a majority vote of its \( k \) spatially nearest neighbors, which are labeled by using the Euclidean distance measure between test data and training data, can optimally estimate the label of unknown case. There are a set of training data \( s_t = (s_{t1}, s_{t2}, ..., s_{tn}) \), each of which has \( b \) features \((s_{t11}, s_{t12}, ..., s_{t1b})\), and a set of input sample data \( s_i = (s_{i1}, s_{i2}, ..., s_{im}) \) with the same features. The Euclidean distance between \( s_{i1} \) and \( s_{t1} \) is defined as:

\[
Edist(s_{i1}, s_{t1}) = \sqrt{\sum_{z=1}^{b} ||s_{i1z} - s_{t1z}||^2}
\]  

(1)

4. Formal definitions and properties of weighted average distance-based outlier factor

In this section, firstly, we introduce a series of formal definitions of the weighted average distance-based outlier factor criterion, which is the fundamental of our proposed approaches. We then analyze several properties of the proposed outlier factor criterion, which can provide the guidance for threshold selection in practical applications.

4.1. Formal definitions of WADOF

Definition 1. Given, \( \mathcal{D} = \text{diag}(d_1, ..., d_H) \), the degree matrix of multisensor networks, the nearest neighbors matrix of sensor nodes is:

\[
\mathcal{K} := \mathcal{D},
\]  

(2)
\( \mathcal{K} = \text{diag}(k_1, \ldots, k_i, \ldots, k_H) \) with \( k_i \) as the number of nearest neighbors of sensor node \( i \).

**Definition 2.** Given the adjacency matrix of multisensor networks, \( A = [a_{ij}] \in \mathbb{R}^{N \times N} \), the set of nearest neighbors of sensor node \( i \), i.e., \( N_i \), and the measurement of sensor node, \( u \), the weighted average distance of sensor measurement of node \( i \) is defined as:

\[
\overline{\text{wad}}_i := \frac{1}{k_i} \sum_{j \in N_i} \text{dist}(u_i, a_{ij} u_j).
\] (3)

**Definition 3.** In one sub-network, i.e., \( J_i \) (\( J_i = N_i \cup \{i\} \)), set \( A^2_{k_i} \) as the 2-permutation of \( k_i \). The internal weighted average distance of sensor measurements in this sub-network, is defined as:

\[
\overline{\text{WAD}}_i := \frac{1}{A^2_{k_i}} \sum_{j,l \in N_i, j \neq l} \text{dist}(a_{ij} u_j, a_{il} u_l).
\] (4)

**Definition 4.** The weighted average distance-based outlier factor of sensor measurement from node \( i \) is:

\[
\overline{\text{WADO}}_i := \frac{\overline{\text{wad}}_i}{\overline{\text{WAD}}_i}.
\] (5)

**Remark 1.** In reality, within the sub-network, \( J_i \), \( \overline{\text{WADO}}_i \) describes the degree that one certain sensor measurement of node \( i \) diverges from those of its \( k_i \) spatially nearest neighbors in a specific sampling time. If \( \overline{\text{WADO}}_i \gg 1 \), it indicates that sensor measurement from node \( i \) lies outside its sub-network system and it is an obvious outlier. Furthermore, the higher \( \overline{\text{WADO}}_i \) is, the farther sensor measurement of node \( i \) is away from those of its \( k_i \) spatially nearest neighbors. Oppositely, the measurement of sensor node \( i \) is encompassed by those of its neighbors when \( \overline{\text{WADO}}_i \lesssim 1 \).
4.2. Properties of WADOF

Undoubtedly, a universal and particular threshold of the proposed outlier factor criterion can offer great convenience to users when distinguishing spurious sensor measurements adopting our criterion in reality. However, the acquisition of the ideal threshold is quite a laborious and time-consuming process due to the sophisticated and diversified application environments and the instability in networks. We can develop an approaching lower bound of outlier in sensor measurements, LWADOF, with few assumptions, though.

**Theorem 1.** In one sub-network, let \( N_i \) be the set of nearest neighbors of sensor node \( i \), which has \( k_i \) nearest neighbors. For \( k_i \to \infty \), we can obtain the lower bound of outlier:

\[
LWADOF = \frac{1}{2}. \quad (6)
\]

**Proof.** Let \( u_i \in U = \mathbb{R} \) and \( \pi_i := \frac{1}{k_i} \sum_{m \in N_i} a_{im} u_m \). Using the squared Euclidean distance \( \| \cdot \|^2 \), the definitions above can be rewritten as:

\[
\overline{wad}_i = \frac{1}{k_i} \sum_{m \in N_i} \| u_i - a_{im} u_m \|^2 \\
= \frac{1}{k_i} \left[ \sum_{m \in N_i} (a_{im} u_m)^2 + k_i u_i^2 - 2u_i \sum_{m \in N_i} a_{im} u_m \right] \\
= \frac{1}{k_i} \left[ \sum_{m \in N_i} (a_{im} u_m)^2 + k_i u_i^2 - 2k_i u_i \pi_i \right] \\
= \frac{1}{k_i} \left[ \sum_{m \in N_i} (a_{im} u_m)^2 + k_i u_i^2 - 2k_i u_i \pi_i + k_i \pi_i^2 - k_i \overline{u}_i^2 \right] \\
= \| u_i - \overline{u}_i \|^2 + \frac{1}{k_i} \sum_{m \in N_i} [(a_{im} u_m)^2 - \overline{u}_i^2] \quad (7)
\]
\[ WAD_i = \frac{1}{A_{ki}^{2}} \sum_{m,n \in N_i; m \neq n} \|a_{im}u_m - a_{in}u_n\|^2 \]

\[ = \frac{1}{k_i(k_i-1)} \left[ 2(k_i - 1) \sum_{m \in N_i} (a_{im}u_m)^2 - 4(a_{i1}u_1a_{i2}u_2 + \cdots + a_{i(k_i-1)}u_{k_i} - a_{i1}u_1a_{i2}u_2) \right] \]

\[ = \frac{2}{k_i(k_i-1)} \left[ k_i \sum_{m \in N_i} (a_{im}u_m)^2 - 2(a_{i1}u_1a_{i2}u_2 + \cdots + a_{i(k_i-1)}u_{k_i} - a_{i1}u_1a_{i2}u_2) \right] \]

\[ = \frac{2}{k_i(k_i-1)} \left[ k_i \sum_{m \in N_i} (a_{im}u_m)^2 - \left( \sum_{m \in N_i} a_{im}u_m \right)^2 \right] \]

\[ = \frac{2}{k_i(k_i-1)} \left[ k_i \sum_{m \in N_i} (a_{im}u_m)^2 - (k_i \overline{u_i})^2 \right] \]

\[ = \frac{2}{k_i - 1} \sum_{m \in N_i} [(a_{im}u_m)^2 - \overline{u_i}^2] \tag{8} \]

\[ WADOF_i = \frac{1}{A_{ki}^{2}} \sum_{m,n \in N_i; m \neq n} \|a_{im}u_m - a_{in}u_n\|^2 \]

\[ = \frac{2}{k_i - 1} \sum_{m \in N_i} \|u_i - \overline{u_i}\|^2 - \frac{1}{2k_i} + \frac{1}{2} \tag{9} \]

Meanwhile, \( \|u_i - \overline{u_i}\|^2 \to 0 \), if one is located in the centroid area of its neighbors, which is definitely not an outlier. Taking the ratio, we obtain the lower bound of outlier, \( LWADOF \).

\[ \square \]

**Remark 2.** Theorem 1 shows that the measurement \( u_i \) is not the outlier, once the \( WADOF_i \approx 0.5 \). The approaching lower bound, \( LWADOF \), provides a potential trimming rule, which can considerably reduce the computational complexity in outlier detection proceeding. In practical applications, since \( WADOF \) is always positive, one certain sensor measurement can be
directly ignored if its WADOF is smaller than LWADOF. Only those sensor measurements whose WADOFs are larger than LWADOF are possible to be outliers.

In real-world applications, it is laborious and time-consuming to achieve proper parameters through the process of trial-and-error. Instead, given the nearest neighbors matrix and the absolute error of sensor nodes, which is the accuracy of this type of sensor node, we can theoretically ascertain the false detection probability for specific outlier threshold with some assumptions.

Theorem 2. Assume that, in the homogeneous sensor networks, all sensor observations are independent and identically distributed (i.i.d.) and variance of measurement for each sensor node is $\sigma^2$. Given the nearest neighborhood size of $N_i$, $k_i$, and the absolute error, $E$, of specific type sensor node in-service, for the outlier threshold $\epsilon > 1$, the probability of false detecting sensor measurement of node $i$ as an outlier is:

$$P[WADOF_i \geq \epsilon] \leq \exp\left(-\frac{k_i \sigma^4 (\epsilon - 1)}{2E^4 (1 + 2\epsilon)^2}\right).$$  \hspace{1cm} (10)

Proof. For simplicity of notation, we assume that networks are not described by a weighted graph, then weights are not relevant and set $a_{ij} = 1$. By means of the computational formula for the variance, we can obtain the expectation of both $\overline{\text{wad}}_i$ and $\overline{\text{WAD}}_i$:

$$\mathbb{E}(\overline{\text{wad}}_i) = \frac{1}{k_i} \mathbb{E}\left(\sum_{m \in N_i} \|u_i - u_m\|^2\right) = 2\sigma^2$$  \hspace{1cm} (11)

$$\mathbb{E}(\overline{\text{WAD}}_i) = \frac{1}{A_{k_i}^2} \mathbb{E}\left(\sum_{m,n \in N_i; m \neq n} \|u_m - u_n\|^2\right) = 2\sigma^2$$  \hspace{1cm} (12)

We define that $f_T := f_d - \epsilon f_D$, $f_d := \overline{\text{wad}}_i$ and $f_D := \overline{\text{WAD}}_i$. 12
As we know that McDiarmid’s Inequality presents the values of a function of i.i.d. variables concentrate around its mean. More exactly, let \( f : \mathcal{X}^k \to \mathbb{R} \) satisfies that for all \( t = 1, \ldots, k \) there is a \( c_t > 0 \) such that for all \( x_1, \ldots, x_k, x'_t \in \mathcal{X} \),

\[
|f(x_1, \ldots, x_t, \ldots, x_k) - f(x_1, \ldots, x'_t, \ldots, x_k)| \leq c_t \tag{13}
\]

then for all \( \zeta > 0 \),

\[
P[f - \mathbb{E}(f) \geq \zeta] \leq \exp\left(- \frac{2\zeta^2}{\sum_{t \in N_i} c_t^2}\right). \tag{14}
\]

For \( f_d \), we have

\[
|f_d(u_1, \ldots, u_d, \ldots, u_{k_i}) - f_d(u_1, \ldots, u'_d, \ldots, u_{k_i})| \leq \frac{1}{k_i} \|u_i - u_d\|^2 - \|u_i - u'_d\|^2 \leq \frac{4E^2}{k_i} = c_d. \tag{15}
\]

Set \( C_{k_i}^2 \) as the 2-combination of \( k_i \) and for \( f_D \), we have

\[
|f_D(u_1, \ldots, u_D, \ldots, u_{k_i}) - f_D(u_1, \ldots, u'_D, \ldots, u_{k_i})| \leq \frac{1}{C_{k_i}^2} \sum_{m \in N_i, m \neq D} \|u_m - u_D\|^2 - \sum_{n \in N_i, n \neq D} \|u_n - u'_D\|^2 \leq \frac{8E^2}{k_i} = c_D. \tag{16}
\]

For the function \( f_T = f_d - \epsilon f_D \), we have

\[
|f_T(u_1, \ldots, u_T, \ldots, u_{k_i}) - f_T(u_1, \ldots, u'_T, \ldots, u_{k_i})| \leq c_d + \epsilon c_D = \frac{4E^2(1 + 2\epsilon)}{k_i} = c_T. \tag{17}
\]

then, the probability of false alarm is

\[
P[WADOF_i \geq \epsilon] = P[\overline{wad}_i - \epsilon \overline{WAD}_i \geq 0] = P[f_T - \mathbb{E}(f_T) \geq \zeta] \tag{18}
\]
where \( \zeta = \mathbb{E}(f_T) = 2\sigma^2(\epsilon - 1) \). Using the McDiarmid Inequality, we have

\[
P[W_{ADOF_i} \geq \epsilon] \leq \exp\left(-\frac{2\zeta^2}{\sum_{T \in N_i} c_T^2}\right) = \exp\left(-\frac{2\zeta^2}{k_i \epsilon T^2}\right) = \exp\left(-\frac{k_i \sigma^4 (\epsilon - 1)}{2E^4(1 + 2\epsilon)^2}\right).
\]

(19)

**Remark 3.** Theorem 2 provides the enlightenment for users to select the outlier threshold in practical applications. In actual fact, most sensor measurements with WADOFs less than 1 and greater than 0.5 can be considered as normal data with noise. In most real applications, users just set certain outlier threshold to clear such obviously unreliable and inaccurate sensor measurements, the WADOFs of which are much greater than 1.

5. **Online distributed outlier clearance approaches**

The proposed approaches empower every sensor node in networks to exploit sensor data information of itself concerning those of its spatially nearest neighbors, rather than its latest historical data information alone, to identify the new arriving measurements in a real-time fashion.

5.1. **Top-n WAD**

It is extremely onerous to determine a general and specific outlier threshold to identify spurious data in the arbitrary dataset, even though we make the full theoretical analysis in the previous section. Consequently, we adopt the well-accepted Top-n style for each sensor node to discern outliers in its measurement and received measurements from its neighbors by sorting
them with larger WADOFs than LWADOFS in the descending order. Then, in each sensor node, the Top-n measurements with the largest WADOFs are regarded as outliers replaced with estimated data, as shown in Algorithm 1, where we label a closed neighborhood of sensor node $i$ from 1 to $k_i + 1$ for ease of presentation. The sensor nodes are not aware of such a local labeling, but each sensor node can identify its nearest neighbors.

### Algorithm 1 Top-n WAD runs in sensor node $i$

1: **Input**: $u_1, ..., u_t, ..., u_{k_i+1}$ (measurements from node $i$ and its neighbors)
2: Initialize memory variables $N$, $M$, $W$ with zeros
3: $N = [u_1, ..., u_{k_i+1}]$
4: for each measurement $u_t \in J_i$ do
5: Calculate $WADOF_t$
6: if $WADOF_t > LWADOFS$ then
7: $M = N_t$
8: else
9: $W = N_t$
10: end if
11: end for
12: Sort the first $n$ measurements with the highest WADOF in $M$ and replace by the mean of the remainder measurements
13: **Output**: $u_1P, ..., u_{(k_i+1)P}$ (processed measurements)

5.2. Adaptive Top-n WAD

Nevertheless, the main drawback of traditional Top-n style outlier detection is that it could not entirely deal with such cases, in which there are more outliers than pre-defined Top-n value. In real applications, we cannot easily obtain the precise number of outliers in each step. For the sake of adaptiveness to such uncertainty, we introduce the second approach. In this approach, we employ the sensor measurement with the lowest WADOF, which is also smaller than LWADOFS, as a gauge of the measurement quality. Then, in one step, any measurement with WADOF larger than LWADOFS would be regarded as an outlier when its distance from the measurement with the lowest WADOF is greater than twice the absolute error, $E$. However, in
extreme cases, new arriving sensor measurements may involve large numbers of outliers, or measurements may be excessively scattered. Thus, the sensor measurement with the lowest $WADOF$ can be greater than $LWADOF$. If such rare cases are encountered, we still incorporate traditional Top-n style into the second approach, which is outlined in Algorithm 2.

**Algorithm 2** Adaptive Top-n WAD runs in sensor node $i$

1: **Input**: $u_1, ..., u_t, ..., u_{k_{i+1}}$ (measurements from node $i$ and its neighbors)
2: **Initialize** memory variables $N, M, W$ with zeros
3: $N = [u_1, ..., u_{k_{i+1}}]$
4: **for** each measurement $u_t \in J_i$ **do**
5: Calculate $WADOF_t$
6: **if** $WADOF_t > LWADOF$ **then**
7: $M = N_t$
8: **else**
9: $W = N_t$
10: **end if**
11: **end for**
12: Find the measurement with the lowest $WADOF$, $W_{\text{lowest}}$ in $W$.
13: **if** $WADOF_{W_{\text{lowest}}} > 0$ **then**
14: **for** each measurement $u_y \in M$ **do**
15: **if** $\|u_y - W_{\text{lowest}}\| > 2E$ **then**
16: $u_y$ is replaced by the mean of normal measurements in $W$.
17: **end if**
18: **end for**
19: **else**
20: Sort the first $n$ measurements with the highest $WADOF$ in $M$ and replace by the mean of the remainder measurements
21: **end if**
22: **Output**: $u_{1P}, ..., u_{(k+1)_P}$ (processed measurements)

Since the fact that outliers account for the minority of the entire dataset, the proposed lower bound saves the computation for the identification of abundant normal sensor measurements in the outlier detection process. Thus, it can greatly save energy consumption to extend the life of networks as far as possible. Besides, the demand for memory usage of each sensor node is relatively compact. Because we only require the measurements from its $k$ neighbors in one sampling time without storing historical measurements. By comparison, SVM-based methods, [19] and [20], require a large amount of historical data to model their own hyper-ellipsoid SVMs. Moreover, su-
supervised and semi-supervised methods, such as [23] and [21], require labeled training sets to create system models. Our approaches, in contrast, can deal with the outlier clearance problem without training sets and long time learning, which are ideal for real-time applications without the prior knowledge about sensor measurements. We make the Table 1 using the Big-O notation [24] to display the comparison of different local outliers detection methods in computational complexity, memory usage and adaptiveness. Calling $U$ the number of nearest neighbors of one sensor node and itself, $b$ the number of features, $sw$ the size of sliding window used by SVM-based approaches and $lw$ the size of learning window used by the DNOD approach, we have the following approximations as shown in Table 1. One should notice that for the sake of high detection accuracy, SVM-based methods and the DNOD approach need to keep the size of their sliding window and learning window relatively large. In other words, $sw$ and $lw$ are much larger than $U$. Moreover, in the interest of the adaptiveness, SVM-based methods need to update their model of targets regularly, which significantly increases the computation complexity. The computational complexity of our approaches mainly depends on two parts. The first is the calculation of $WADOF$, which is straightforward and this step is linear in $U$. The second is that we sort measurements based on their $WADOF$ values, which can be finished in $O(U \log U)$. Since measurements with $WADOF < LWADOF$ are flushed (i.e., they are undoubtedly not outliers), the number of measurements needed to sort in this step is smaller than $U$. In conclusion, the maximum computation complexity of our approaches is $O(U \log U)$.
Table 1: Different local outliers detection methods comparison

<table>
<thead>
<tr>
<th>Methods</th>
<th>References</th>
<th>Model</th>
<th>Identify</th>
<th>Memory</th>
<th>Adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM-based</td>
<td>[19][20]</td>
<td>(O(sw^2b + sw^3))</td>
<td>(O(b))</td>
<td>(sw)</td>
<td>Yes</td>
</tr>
<tr>
<td>LOF-based</td>
<td>[15]</td>
<td>-</td>
<td>(O(bU^2))</td>
<td>(U)</td>
<td>No</td>
</tr>
<tr>
<td>DNOD</td>
<td>[18]</td>
<td>-</td>
<td>(O(2bUlw))</td>
<td>(lw)</td>
<td>Yes</td>
</tr>
<tr>
<td>T-WAD</td>
<td>-</td>
<td>-</td>
<td>(O(bU\log U))</td>
<td>(U)</td>
<td>No</td>
</tr>
<tr>
<td>AT-WAD</td>
<td>-</td>
<td>-</td>
<td>(O(bU\log U))</td>
<td>(U)</td>
<td>Yes</td>
</tr>
</tbody>
</table>

6. Synthetic and experimental results analysis

We investigate the performance of proposed approaches in online clearing outliers through tests, in which we adopt other five widely used nearest neighbor-based approaches with Top-n style: LOF [15], KNN [9], INFLO [25], ODIN [26] and RDOS [27] as well as two typical methods: SVM-based EAOD [20] and DNOD [18]. In these comparative tests, we process synthetic sensor measurements and real experimental measurements collected by the SensorScope networks on the EPFL campus [28].

The Receiver Operator Characteristic (ROC) curve is the most popular evaluation metric for outlier detection approaches [9]. A ROC curve is established through plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TPR is the probability of real outlier detection. And the FPR is the probability of false alarm. The larger the area under the ROC curve, the better the accuracy of the method. Recent years, the metric of ROC curve has been demonstrated to be the most effective measure compared with many others, such as precision and recall, and widely used for the performance evaluation of outlier detection approaches [9].
6.1. Synthetic sensor measurements

Firstly, in the synthetic data tests, we consider that a closed neighborhood of the undirected homogeneous multisensor networks, which is centered at a sensor node $i$ with its 9 spatially nearest neighbors, measure the reference signal with noise and outliers. We use the following test signals:

$$r(t) = \sin(t + 9) + \sin(2t + 5) + \sin(4t + 2).$$  \hspace{1cm} (20)
For all sensor nodes, we set different measurement noise with the same variance, 0.01 and random outliers that there are no more than 4 outliers in the same sampling step. Further, the sampling frequency is 100 Hz and the running time is 4 seconds. According to the given noise, we can roughly treat absolute sensor measurement error, $E$, as 0.1.

Figure 2 shows the synthetic raw sensor measurements containing the noise only as well as both noise and outliers. Figure 3 shows the processed results with the Top-2 WAD approach and Adaptive Top-2 WAD approach. Obviously, the Top-2 WAD cannot cope with such cases that there are more than 2 outliers in the same sampling time, while the Adaptive Top-2 WAD can handle those cases, even, in each of which there would be 40 percentage outliers in one step. Moreover, the TPR and FPR values for the Top-2 WAD are 54.05% and 13.73%, while the TPR and FPR values for the Adaptive Top-2 WAD are 100% and 4.19%.

### 6.2. Real experiments sensor measurements

The real experimental sensor measurements are gathered from 97 networked sensing stations deployed on the EPFL campus from July 2006 to May 2007. This measurement campaign aims at better understanding meteorology and atmospheric transport in the urban environment. Thus, its high temporal and spatial density measurements, covering the macro-homogeneous areas, are extremely suitable for evaluating the performance of outlier clearance approaches. In evaluation works, the real-life dataset are collected via a randomly selected cluster of neighboring nodes centered at node 105 with its 9 spatially nearest neighbors: nodes 12, 13, 14, 15, 18, 24, 39, 44, 45 as illustrated in Figure 4. Due to the gigantic amount and multi-attributes of dataset, for the sake of simplification, we randomly
choose parts of their measurements, 5380 samples in total, with two attributes: ambient temperature and relative humidity, which are collected by Sensirion SHT75 from 1am to 4am on 16th April 2007, avoiding the bias effects on measurements of these two attributes from the shadow and sunlight. Figure 5 displays the Sensirion SHT75 and Table 2 describes the detailed information of Sensirion SHT75.

### Table 2: Sensirion SHT75

<table>
<thead>
<tr>
<th>Measurement Type</th>
<th>Unit</th>
<th>Range</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature Digital (14 bits)</td>
<td>Temperature: °C</td>
<td>Temperature: -20 - 60</td>
<td>Temperature: ±0.3°C</td>
</tr>
<tr>
<td>Humidity Digital (12 bits)</td>
<td>Humidity: % (relative humidity)</td>
<td>Humidity: 0 - 100</td>
<td>Humidity: ±2%</td>
</tr>
</tbody>
</table>

![Figure 4](image1.png)

**Figure 4:** The deployment of closed neighborhood on the EPFL campus.

![Figure 5](image2.png)

**Figure 5:** Sensirion SHT75.

![Figure 6](image3.png)

(a) Original temperature dataset  
(b) Full evaluation temperature dataset

**Figure 6:** Evaluation dataset in temperature
The original ambient temperature and relative humidity dataset, collected by ten sensor nodes, are shown in Figure 6(a) and Figure 7(a), respectively. More specifically, there are 10 and 20 outliers in total for each attribute, which are thoroughly distinct with others or impossible in reality, and the obvious outliers are labeled. On account of the excessive naivety of these outlier cases, they can not be adequately used to evaluate the outlier clearance approaches. For this reason, we randomly inject more outliers into these original dataset, which are similar to the outliers existing in the original dataset. Detailed information of the original and injected outlier cases for both dataset are described in Table 3 and Table 4. Then, Figure 6(b) and Figure 7(b) display the full evaluation dataset, in which there would be more than five outliers in the same step, with the number of outliers, 49 and 46, respectively.

Figures 8, 9 and 10 show the TPR, FPR and ROC curves on these evaluation dataset obtained via proposed approaches Top-n WAD and Adaptive Top-n WAD as well as other widely used methods, in which outliers are identified through their outlier detection criteria and then replaced with
Table 3: Outliers cases for temperature dataset

<table>
<thead>
<tr>
<th>Case</th>
<th>Outliers/Sample</th>
<th>Original Occurrences</th>
<th>After Injection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4: Outliers cases for relative humidity dataset

<table>
<thead>
<tr>
<th>Case</th>
<th>Outliers/Sample</th>
<th>Original Occurrences</th>
<th>After Injection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

the mean of remainders. Moreover, the TPR and FPR are produced with different Top-n values ranging from 1 to 5. Perspicuously, the Adaptive Top-n WAD approach exhibits superior outlier clearance performance with the metric of ROC in both dataset. And the Top-n WAD approach shows competitive performance which is slightly better than the other five highly applied methods; meanwhile, the TPR and FPR of these six approaches increase gradually as the Top-n value rising. In addition, even though the Top-n value is 1, the Adaptive Top-n WAD approach can still achieve more than 90 percent TPR and low FPR in diverse dataset containing many outlier cases that two or more outliers exist at the same sample time, which thoroughly verifies the high adaptability of the Adaptive Top-n WAD. However, under the identical circumstance, the other six approaches cannot offer superb outlier clearance performance as the Adaptive Top-n WAD. Particularly, the Adaptive Top-n WAD approach would achieve exceedingly high
TPR as almost 100 percent on these dataset and very low FPR at the same time, if we adopt 2 as its Top-n value. As a matter of fact, in real applications, the high Top-n value may obtain the acceptable TPR while it would also lead to the high FPR and information lost owing to the complex unpredictability of the number of outliers in each step. Nevertheless, our Adaptive Top-n WAD approach with the smaller Top-n value can not only offer high TPR performance but also keep very low FPR and information integrity without a prior knowledge of sensor measurements.

Besides, the outlier clearance performances of different typical methods are described by Figure 11. In these tests, the Top-n value for proposed approaches and T-LOF is 3. With consideration of the evaluation dataset, we follow the settings in [20] and [18] for EAOD and DNOD such that the regularization parameter of EAOD is 0.07 and the size of learning window for DNOD is 100. Obviously, the Adaptive Top-n WAD and SVM-based EAOD achieve the remarkable performance and the Top-n WAD approach demonstrates its competitive capabilities in outlier clearance. However, SVM-based EAOD method requires much higher memory and computing

24
resources, which may cause the outlier detection delay.

7. Conclusion

In this paper, we propose two online distributed outlier clearance approaches with low computational complexity and memory usage. The evaluation works demonstrate that our approaches achieve better outlier clearance results, even if multiple sensor nodes fail at the same given time, compared with other widely used methods. Particularly, without the prior knowledge
of sensor measurements, the Adaptive Top-n WAD approach secures high TPR and keeps low FPR at the same time bringing great convenience to practical applications. Future works include the appliance of outlier clearance approaches in multivariate dataset and event classification.

References


[26] V. Hautamaki, I. Karkkainen, P. Franti, Outlier detection using \( k \)-nearest neighbour graph, in: Proceedings of the 17th International
