Joint Event Detection & Identification: A Clustering based approach for Wireless Sensor Networks

N.Shahid, S.B.Ali, K.Ali, M.A.Lodhi, O.B.Usman, I.H.Naqvi
Department of Electrical Engineering, LUMS Syed Babar Ali School of Science and Engineering, Pakistan
Email: {numans, 13100028, 13100174, 13100175, 13100026, ijaznaqvi}@lums.edu.pk

***Abstract*—Distributed clustering based techniques have been increasingly employed for *outlier detection* in Wireless Sensor Networks (WSNs). But despite its numerous advantages such as online and efficient computations and incorporation of spatio-temporal & attribute correlations, clustering has not been studied for *event detection & identification*, which is essential for smooth and reliable operations of large scale WSNs. This paper introduces the significance of clustering based event detection & identification to the research community. Further, it presents an online technique for joint event detection and identification that achieves a very high performance for synthetic and real data sets with a significant reduction in computational complexity as compared to the state-of-the-art techniques. A remarkable advantage of the proposed technique is that it can identify the key attributes in the order of their contribution towards an event without incurring any additional complexity.**

***Keywords*: Outlier detection, event detection, clustering, wireless sensor networks.**

I. Introduction

Outlier detection in the context of WSNs has attracted significant attention of research community in the recent years [1], [2], [3], [4], [5], [6]. More specifically, the classification based algorithms, such as Support Vector Machines (SVMs), which ensure a high performance, are being used extensively for this purpose [7] due to their ability to incorporate spatio-temporal and attribute correlations of data. Although SVM based algorithms have been made computationally efficient by the introduction of Quarter-Sphere SVMs (QS-SVM) [2], [3], [4], [6], which reduce the quadratic optimization problem to a linear problem, but the computational cost associated with SVMs is still high. This can be attributed to the online processing of data samples, which requires the solution of a linear optimization problem and mapping of data to a high dimensional feature space with the arrival of every new data sample, thus limiting the applicability of SVMs for environments which have strict latency requirements.

Recently, the clustering based algorithms have been found to be of significant importance in outlier detection as they are computationally inexpensive, adhere to strict latency requirements during online processing, achieve a high detection and a very low false positive rate [8], [9], [10], [11]. Moreover, these algorithms can incorporate a wide variety of multi-variate data distributions, attribute-temporal correlations and spatial data non-stationarity associated with dynamically changing environments.

A primary constraint of clustering algorithms is that they do not focus on **event detection**. For example, the algorithm presented in [12] provides a method for the estimation of boundary for sensor data. The data samples that fall outside the estimated boundary are declared as outliers. The authors of [10] present a method for incrementally updating the sensor data boundary estimated in [12] by using an exponential forgetting factor to incorporate the effect of most recently gathered data samples at a sensor node in the network. The method of [9] facilitates in global outlier detection (data sample(s) that is anomalous with respect to the entire network) by using various cluster merging strategies such as compound similarity, transformation energy similarity and focal similarity. However, none of these algorithms present a mechanism for event detection. As event detection in WSNs finds significant applications in monitoring the environments for critical situations [13], for instance industrial environments, so it should be studied in the context of energy efficient clustering algorithms.

In addition to event detection, another important property that the clustering based algorithms should exhibit is **event identification**. State-of-the-art clustering algorithms (like [10]) recommend a manual analysis of the data samples, followed by outlier or event detection, to determine the attributes involved or the 'type' of event. This is not a useful practice due to a number of reasons: a) Often the event detection algorithms do not process raw data. Data may be processed by applying some transformations, e.g normalization, principal component analysis, frequency transforms etc. Thus, the manual examination of attributes may require an inverse transformation which will pose additional computation overheads. b) Events are usually caused by a temporal or spatial change in the key attributes of the system which can be sudden or gradual. In order to perform a manual examination, continuous monitoring of the key attributes of a system is required. c) A visual analysis of data at a central location requires entire data to be transmitted to the central node. This incurs a huge communication cost. Thus, clustering based event detection and identification, which has never been studied before, solves these problems in a simple and elegant manner.

Following are the significant contributions of this paper:

- Introduction to the problem of 'clustering based event detection and identification'.
- Proposition of an online clustering based algorithm for joint event detection and identification in WSNs that achieves a very high performance and a significant reduction in computational complexity as compared to QS-SVM based algorithms.
- Event Identification strategy, i.e, identification of the

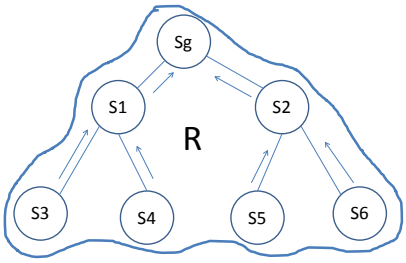


Fig. 1. A hierarchical WSN deployed in a region R; consisting of 6 sensor nodes $S1, \cdots, S6$ and one gateway node Sg

attributes in the order of their increasing contribution towards an event; a problem that has been dealt for the very first time and provides an open challenge to the research community to build and improve on the proposed algorithm.

II. Overview of Clustering based Outlier Detection

Consider a hierarchical WSN deployed in a region R which consists of N localized nodes. The network shown in fig. 1 is a hierarchical WSN in which there are six sensor nodes $S1, \cdots, S6$ and one gateway node Sg. In a hierarchical WSN the gateway node Sg is responsible for collecting the information of entire network from the parent nodes $S1, S2$. Let $X_k = \{x_1, x_2, \cdots x_k\}$ be the k samples of data collected from the q^{th} WSN node where each sample x_i is a column vector in $\Re^d$. Let m_k and S_k be the sample mean and covariance of all k samples in X_k, then upto 98% of data in X_k can be enclosed in a hyperellipsoid of effective radius t by using $t^2 = (\chi_d{}^2)^{-1}_{0.98}$ [10], if X_k is multi-variate normal [12].

$$e_k(m_k, S_k^{-1}, t) = \{x_i \epsilon \Re^d | \underbrace{\sqrt{(x_i - m_k)^T {S_k}^{-1} (x_i - m_k)}}_{D_i = mahalanobis\ \ distance\ \ of\ \ x_i} \leq t\}$$
$$\forall i = 1, 2, \cdots, k \quad (1)$$

The hyperellipsoid e_k is defined as the set of all k data samples $x_i, \forall i = 1, 2, \cdots, k$ whose mahalanobis distance $\leq t$. Whereas, those samples whose mahalanobis distance is greater than t do not lie inside the hyperellipsoid and are considered outliers.

This hyperellipsoidal boundary e_k can be updated with the arrival of every new sample x_{k+1} by using an update formula for m_k and S_k in such a way that the updated boundary can incorporate data variation in the monitored environment. The use of an exponential forgetting factor $\lambda (0 < \lambda < 1)$ in the update of m_k and S_k for every new measurement incorporates the effect of most recent data measurements in e_k [10].

$$X_{k+1} = X_k \bigcup x_{k+1}$$
$$m_{k+1,\lambda} = \lambda m_{k\lambda} + (1 - \lambda) x_{k+1}$$
$$S_{k+1,\lambda} = \frac{\lambda(k-1)}{k} S_{k\lambda} + \frac{1}{k}(x_{k+1} - m_{k+1,\lambda})(x_{k+1} - m_{k+1,\lambda})^T$$

where X_{k+1}, $m_{k+1,\lambda}$ and $S_{k+1,\lambda}$ are the updated sample set, mean and covariance of the ellipse e_k after the incorporation of new arrived sample x_{k+1}. For the ease of notation, we drop the subscripts $k+1$ and λ and use X_k for X_{k+1}, m_k for $m_{k+1,\lambda}$ and S_k for $S_{k+1,\lambda}$.

III. Proposed Projection based Event Detection & Identification

Our proposed algorithm starts with the clustering based outlier detection algorithm of [10], updates the hyperellipsoid at each time instant as discussed in the previous section and declares a newly arrived sample as an outlier or normal. **The algorithm proceeds to event detection and identification phase (discussed below) if an outlier is detected**.

A. Computations at individual nodes

We propose a projection based clustering algorithm for joint event detection and identification. The d dimensional hyperellipsoid e_k defined in eq (1) contains the vector set X_k. Let the X be the d dimensional space of X_k and let X^p be the subspace of X that contains only the p^{th} attribute present in X_k. The idea is to project the hyperellipsoid e_k along each of the subspace X^p and check the deviation of individual projections. The event detection algorithm initiates this processing at all nodes of the region R, even if an outlier is detected at one node only.

STEP 1: Defining the Projection: Let V_p be a $d \times 1$ column vector $V_p = \{v_l | v_l \epsilon \{0, 1\}, l = 1, 2, \cdots, d\}$, where $v_l = 1$, if $l = p$ and $v_l = 0 \forall l \neq p$ and $I_{d \times d}$ be a $d \times d$ identity matrix. We define the projection of e_k (containing X_k) along each of the d subspaces $X^p \forall p = 1, 2, \cdots d$ (at the q^{th} node) as

$$\begin{aligned} Proj(X_k)_q &= X_k^T I_{d \times d} \\ &= X_k^T \left[V_1 | V_2 | \cdots | V_d\right] \\ &= \left[X_k^T V_1 | X_k^T V_2 | \cdots | X_k^T V_d\right] \\ &= \left[Proj(X_k)_1 | Proj(X_k)_2 | \cdots | Proj(X_k)_d\right] \end{aligned}$$

From the above equation $Proj(X_k)_p = {X_k}^T V_p$ is the $k \times 1$ column vector and can be defined as the projection of e_k along the p^{th} subspace X^p. Thus, $Proj(X_k)_p$ contains the projection of each of the $d \times 1$ sample $x_i \epsilon X_k$ along the p^{th} subspace X^p and $Proj(X_k)_q$ contains the projection of this sample along all the d subspaces. Since there are d subspaces in total, so $Proj(X_k)_q$ is a $k \times d$ matrix.

STEP 2: Clustering on Projected Samples: Further we apply a hyperellipsoidal clustering along each of the projected sets $Proj(X_k)_p, \forall p = 1, 2, \cdots, d$ and define the resulting ellipsoid along the p^{th} subspace as

$$e_k^p(m_k^p, S_k^{p-1}, t_p) = ell(Proj(X_k)_p) = \quad (2)$$
$$\{x_i^p \epsilon \Re | \underbrace{\sqrt{(x_i^p - m_k^p)^T S_k^{p-1} (x_i^p - m_k^p)}}_{D_i^p = mahalanobis\ \ distance\ \ of\ \ x_i^p} \leq t_p, \forall i = 1, 2, \cdots, k\}$$

where S_k^{p-1} is the inverse of p^{th} diagonal entry of S_k, $x_i^p = x_i^T V_p$ and $m_k^p = m_k^T V_p$ are the projections of sample x_i and mean m_k along the p^{th} subspace. Since each of the d projected

sets $Proj(X_k)_p$ are enclosed in d ellipsoids, the set of all projected ellipsoids at the q^{th} node can be given as:

$$E_q = \{e_k^1(m_k^1, S_k^{1^{-1}}, t_1), \cdots, e_k^d(m_k^d, S_k^{d^{-1}}, t_d)\}$$

where each of these ellipsoids has an effective radius t_p given as:

$$t_p = \frac{(\frac{u_p^T V_p}{\sqrt{\lambda_p}} - m_k^p)}{\sqrt{S_k^{p^2}}} \quad (3)$$

and the matrix containing the radii of all projected ellipsoids can be defined as:

$$T_q = \begin{bmatrix} t_1 & t_2 & \cdots & t_d \end{bmatrix}^T$$

Eq (3) gives the individual effective radii of each of the projected ellipsoids e_k^p such that an enclosure of atleast 98% of entire data X_k is guaranteed in e_k. In the above equation λ_p and u_p are the p^{th} eigen value-eigen vector pair associated with S_k^{-1}. The inverse of the square root of the p^{th} eigen value $1/\sqrt{\lambda_p}$ is the p^{th} axial length of hyperellipsoid [9] and $u_p^T V_p$ is the projection of p^{th} orthonormal eigen vector u_p along the p^{th} subspace. Therefore, $(u_p^T V_p/\sqrt{\lambda_p} - m_k^p)$ is the axial length of e_k^p or the magnitude of the projection of e_k along the p^{th} subspace and $\sqrt{S_k^p}$ is the standard deviation of $Proj(X_k)_p$. The angle θ indicates the rotation of hyperellipsoid from any subspace. This is explained in fig.1 for the case of a 2D rotated and non-rotated hyperellipsoid. For both cases, the eigen-vectors u_p play a role to determine the projections along p^{th} subspace.

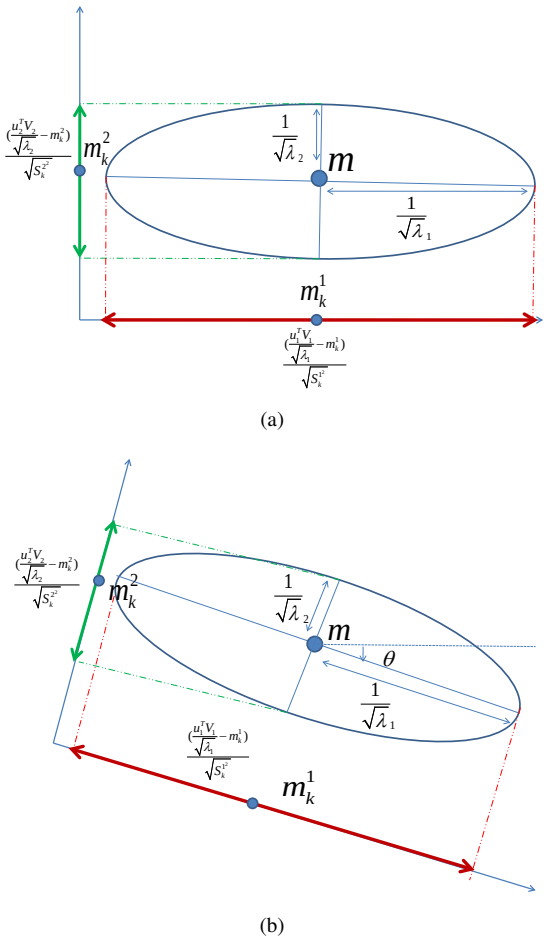
STEP 3: Decision formulation: Let D_q be the $k \times d$ matrix containing the mahalanobis distances D_i^p (of each x_i^p from e_k^p), determined from eq (2) while applying the hyperellipsoidal clustering algorithm of step 2. Let $D_q(k, 1 \cdots d)$ be the k^{th} row of D_q containing all the d distances corresponding to the most recent sample x_k. Then, the decision about x_k being an outlier or normal can be formulated as

$$Outlier_{q,k} = u([u(T_q - D_q^T(k, 1 \cdots d))]^T 1_{d \times 1} - d) \quad (4)$$

where $[u(T_q - D_q^T(k, 1 \cdots d))]$ is a $d \times 1$ matrix, each of whose entry is equal to

$$u(\underbrace{t_p - \sqrt{(x_k^p - m_k^p)^T S_k^{p^{-1}} (x_k^p - m_k^p)}}_{x}), \forall p = 1, 2, \cdots, d$$

The term 'x' gives the mahalanobis deviation of the k^{th} sample projected along the p^{th} subspace, i.e, the deviation of $x_k^p \epsilon Proj(X_k)_p$ from the effective radius t_p of its ellipsoid e_k^p and $u(x)$ gives the decision about x_k^p being a normal or outlying attribute. $u(x) = 0$ if x_k^p is an outlying attribute and 1 if it's a normal attribute. Thus, the $d \times 1$ matrix $[u(T_q - D_q^T(k, 1 \cdots d))]$ contains the decisions about all d attributes of a sample x_k. The final decision about the sample x_k being outlier is obtained by taking the dot product of this matrix with $1_{d \times 1}$, subtracting d from the result and applying unit step function on this value. A value of 0 indicates that x_k is an outlier, whereas a value 1 indicates that x_k is a normal sample.



(a)

(b)

Fig. 2. 2(a) represents the case of a 2D non-rotated hyperellipsoid. m is the mean, m_k^1 and m_k^2 are the projections of m on subspaces 1 and 2. (For 2D hyperellipsoid, there will be 2 subspaces of 1D each) The projection of hyperellipsoid along subspace 1 is shown with a red line and that on subspace 2 with a green line. 2(a) represents the case of a 2D hyperellipsoid rotated by an angle θ from the principal axis. The projection of hyperellipsoid along subspace 1 is again shown with a red line and that on subspace 2 with a green line.

As the steps 1,2 and 3 of this section are performed iff the sample has been declared an outlier (using the procedure of section II), therefore, the final decision of eq. (4) will be 0 and it does not play a role in 'outlier detection'. However, it has been computed because it will play an important role in event detection and identification (explained in section III-B).

To summarize, following important information can be derived from the decision function of eq (4):

- The matrix $[T_q - D_q^T(k, 1 \cdots d)]$ gives information about the degree of deviation of all d individual attributes of sample x_k.
- The matrix $[u(T_q - D_q^T(k, 1 \cdots d))]$ gives information about all d individual attributes of sample x_k being outlier or normal.
- The product $[u(T_q - D_q^T(k, 1 \cdots d))]^T 1_{d \times 1}$ gives the number of normal attributes in sample x_k.
- The negative of term $[u(T_q - D_q^T(k, 1 \cdots d))]^T 1_{d \times 1} - d$ gives the number of outlying attributes in sample x_k.
- The overall function $u([u(T_q - D_q^T(k, 1 \cdots d))]^T 1_{d \times 1} - d)$ gives information about x_k being an outlier or normal. x_k

will be declared as an outlier, if any one of its projections x_k^p is an outlier.

B. Computations at the gateway node of a Region

The outlier detection algorithm described in the previous section will operate on each of the nodes of hierarchical WSN. Thus, each q^{th} node $Sq, \forall q = 1, 2, \cdots, N$ will compute D_q and $Outlier_{q,k}$. In order to determine the presence of an event in the region R, the gateway node Sg tends to collect some information $I_q, \forall q = 1, 2, \cdots, N$ from each node of the network. The information I_q broadcasted by each node consists of the following statistics:

$$I_q = \{D_q(k, 1 \cdots d), Outlier_{q,k}, T_k\}$$

where $D_q(k, 1 \cdots d)$ is the k^{th} row of D_q which contains information about all the d projections of x_k being outlier or normal, $Outlier_{q,k}$ is the information about x_k being normal or outlier and $T_{q,k}$ is the time stamp at which the k^{th} sample was measured at node Sq. The leaf nodes $S3, \cdots, S6$ will broadcast their information to their parent nodes $S1, S2$. Each parents node combines its own information with its children's node. Assuming a to be the number of children nodes the combined information at a parent node in the network can be written as:

$$I_q^{'} = \underbrace{\{[D_{parent}^T(k, 1 \cdots d) | D_1^T(k, 1 \cdots d) | \cdots | D_a^T(k, 1 \cdots d)]\}}_{D_q^{'}}$$
$$\bigcup \{\underbrace{\begin{bmatrix} Outlier_{parent,k} & Outlier_{1,k} & \cdots & Outlier_{a,k} \end{bmatrix}^T}_{Outlier_q^{'}}\}$$
$$\bigcup \{\begin{bmatrix} T_{parent,k} & T_{1,k} & \cdots & T_{a,k} \end{bmatrix}^T\}$$

where $D_q^{'}$ is a matrix that contains $D_q(k, 1 \cdots d)$ of various nodes and increases in size at all the parent nodes of the network. For example, at the parent node $S1$ in fig. 1 $D_q^{'}$ will be a $3 \times d$ matrix that will contain the $D_q(k, 1 \cdots d)$ of itself and its children nodes $S3$ and $S4$. Similarly the matrix $Outlier_q^{'}$ will also be updated at all the parent nodes of the network.

This broadcast and update of information $I_q^{'}$ will continue up the hierarchy to the gateway node Sg. The gateway node of a particular region R will be able to make various decisions based on the collected information, the most important of which is to determine the presence or absence of an event in the network. Following is a description of decisions that the gateway node can make.

1) Event Detection: To detect the presence of an event in the region R, we use the definition used in [2]: "An event is said to be present in the network if more than half of the nodes in the network show outliers". The gateway node performs the following operations to determine the presence or absence of an event after an outlier is detected at any node in the region.

- The time stamp information $T_{q,k}$ of all nodes in the region R are compared with the current time instant and the information of nodes with old time stamps is discarded.
- The matrix $Outlier_q^{'}$ is extracted from the updated information after above step and the following operation is performed to determine the presence or absence of event

$$Event_R = u(\lceil \frac{N}{2} \rceil - [Outlier_q^{'T}] 1_{N \times 1})$$

where $1_{N \times 1}$ is a column vector of all 1's and $[Outlier_q^{'T}] 1_{N \times 1}$ determines the number of nodes out of N in the region R that do not indicate the presence of outliers. If this number is greater than $\lceil \frac{N}{2} \rceil$ then it implies that more than half of the nodes do not indicate an outlier and $Event_R = 0$. i.e, the region does not indicate the presence of an event. However, $[Outlier_q^{'T}] 1_{N \times 1} < \lceil \frac{N}{2} \rceil$ implies that more than half of the nodes indicate the presence of an outlier, therefore, $Event_R = 1$ and the region has been affected by an event.

2) Event Identification: Once the gateway node has declared the presence of event in the region, it can then use the information $I_q^{'}$ of the whole network to determine the type of event that has occurred in the region. This would also lead to the identification of attributes involved in the event.

- The gateway node extracts the matrix $D_q^{'}$ from the information $I_q^{'}$. $D_q^{'}$ is matrix, each of whose columns contain the information about the projections of most recent data sample, at a particular node, being outlier or normal.
- The attributes involved in the event are then identified by performing the following operation.

$$Attributes_R = Indices\{sort\{D_q^{'T} 1_{N \times 1}\}\}$$

The matrix $Attributes_R$ will return the attributes in the order of their increasing contribution towards the event.

The overall algorithm starting from outlier detection to event detection and identification is summarized in the flow chart of fig. 3.

IV. Complexity Comparison with QS-SVM

We compare the performance and complexity of our proposed clustering based event detection algorithm with five QS-SVM based algorithms presented in [2], [3] and [6] because they consider spatio-temporal and attribute correlations of data for event detection, like the proposed algorithm. Whereas, no other clustering based algorithm provides event detection strategy.

The algorithm presented in [6] (ST-QS-SVM) tends to separate the normal data from outliers using a quarter-sphere shaped boundary around the chunk of normal data in such a way that outliers remain outside the quarter-sphere. The quarter-sphere radius is determined from spatio-temporal correlations of nodes' data. The lower limit to the number of false detections in any QS-SVM algorithm is set by regularization parameter (say $u \epsilon (0, 1)$). Events are detected by using a consensus of all the nodes in the network. An event is said to occur in the network if more than half of nodes indicate the presence of an outlier. The algorithm presented in [2] (STA-QS-SVM) tends to improve the outlier and event detection performance of [6] by incorporating attribute correlations in addition to

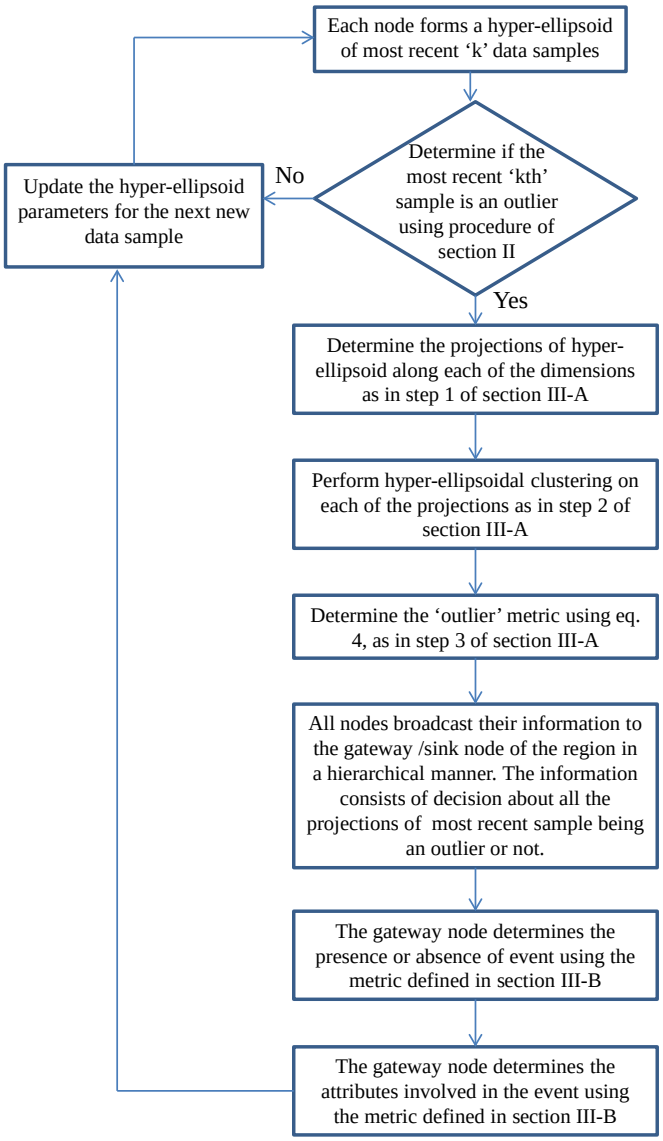


Fig. 3. Complete algorithm for outlier and event detection and event identification in a region R.

spatio-temporal correlations of nodes data for determining the quarter-sphere radius. This results in an increase in outlier and event detection performance with an added disadvantage of increase in computational complexity. Various algorithms presented in [3] (STA-TSV, STA-TASV, STA-CA) have the same underlying mechanism as that of STA-QS-SVM ([2]), therefore, their performance is comparable to STA-QS-SVM. However, these algorithms tend to reduce the communication and computational complexity of STA-QS-SVM as discussed in tab. I.

A detailed analysis of computational and communication overheads of the proposed technique along with various QS-SVM based algorithms proposed in [2], [3], [6] is given in tab. I. The analysis clearly shows that the proposed clustering based algorithm reduces the $O(n^2 + nd^2)$ complexity of various QS-SVM algorithms to $O(nd^2)$. Thus, the proposed algorithm achieves event detection as well as identification with a significant reduction in computation complexity as compared to QS-SVM algorithms which perform event detection only. The communication complexity of all the techniques are approximately same however.

TABLE I
ANALYSIS OF COMPUTATION & COMMUNICATION COMPLEXITY FOR THE PROPOSED TECHNIQUE AND QS-SVM BASED ALGORITHMS [2], [3], [6]. n IS THE TOTAL NUMBER OF MEASUREMENTS, d IS THE NUMBER OF ATTRIBUTES, $v << 1$ IS THE FRACTION OF EVENTS IN THE DATA SET AND $u\epsilon(0,1)$ IS THE REGULARIZATION PARAMETER FOR SVM. NOTE THAT $nv << n$ AND $(nv)^3 << n^3$.

Proposed Technique	**Computational Complexity**
Outlier Detection [10]	$O(nd^2)$
Projection along d dimensions	$O(nv)$
Eigenvalue decomposition	$O((nv)^3)$
Clustering along d dimensions	$O(nv)$
Decision for outlier and event	$O(1)$
Total computational complexity of proposed technique	$O(nd^2 + 2vn + (nv)^3)$ $\approx O(nd^2)$
QS-SVM Techniques	**Computational Complexity**
ST-QS-SVM [6]	$O(n^2)$
STA-QS-SVM [2]	$O(n^2 + nd^2)$
STA-TASV,STA-TSV [3]	$O(n^2 + nd^2)$
STA-CA [3]	$O(n^2 + nud^2)$
Proposed Technique	**Communication Complexity**
Maximum communication overhead for parent node close to gateway node	$O(an)$ $\approx O(n)$
QS-SVM Techniques	**Communication Complexity**
ST-QS-SVM [6]	$O(n)$
STA-QS-SVM [2]	$O(n)$
STA-TASV, STA-TSV,STA-CA [3]	$O(nu)$

V. SIMULATIONS & RESULTS

TABLE II
A COMPARISON OF EVENT DETECTION & IDENTIFICATION RATE AND FALSE POSITIVE RATE OF THE PROPOSED TECHNIQUE WITH ST-QS-SVM [6], STA-QS-SVM [2], STA-TASV, STA-TSV, & STA-CA [3]

Technique	**Detection Rate (Synthetic Data)**	**False Positive Rate (Synthetic Data)**
Proposed (Detection + identification)	98.52%	0.34%
ST-QS-SVM [6]	46.67%	0.3%
STA-QS-SVM [2]	99.2%	0.26%
STA-TASV [3]	99%	0.26%
STA-TSV [3]	98.8%	0.3%
STA-CA [3]	99%	0.3%
Technique	**Detection Rate (Real Data)**	**False Positive Rate (Real Data)**
Proposed (Detection + identification)	90%	0.1%
ST-QS-SVM [6]	16.67%	10.85%
STA-QS-SVM [2]	91.67%	0.5%
STA-TASV [3]	92.45%	0.48%
STA-TSV [3]	99%	0.6%
STA-CA [3]	91.67%	0.1%

Experimental evaluation was performed on matlab with two types of data sets, 1) synthetic and 2) real.

The synthetic data set consisted of 2500 samples belonging to a 5-dimensional Gaussian distribution with attribute

means selected from (10-60). The simulation environment for synthetic data set assumed a gateway node with 6 spatially correlated child nodes ($N = 7$) in the hierarchy. Both the synthetic and real data sets were normalized between [0,1].

Performance evaluation was done with 5% of events introduced in the tails of data set distributions along each of the attributes. For the proposed technique, the effective radius of hyperellipsoid t was set to $t^2 = ({\chi_d}^2)^{-1}_{0.98}$. For comparison with QS-SVM, the training window n for all the QS-SVM based techniques (ST-QS-SVM, STA-QS-SVM, STA-TASV, STA-TSV & STA-CA) was kept equal to 100 & RBF Kernel function was used with window width 0.2. As our proposed technique performs joint event detection & identification, so event detection & identification rates and false positive rates have been presented for the proposed technique. Whereas, for the QS-SVM based techniques, only event detection and false positive rates have been presented. Tab. II shows that our proposed technique can achieve approximately the same event detection & false positive rates as those of STA-QS-SVM, STA-TASV, STA-TSV & STA-CA with an additional advantage of event identification and significant reduction in computational complexity (from $O(n^2)$ to $O(nd^2)$ tab. I). Further, the proposed technique suggests a 111% improvement in the detection rate as compared to ST-QS-SVM.

The real data set was taken from a multi-hop ($N = 4$) WSN deployment using TelosB motes. The data consists of humidity and temperature measurements collected during 6 hour period at intervals of 5 seconds [14]. This data set has approximately 1% of events (recorded as high temperature). Tab. II shows that our proposed technique can achieve approximately the same event detection & false positive rates as those of STA-QS-SVM, STA-TASV & STA-CA with the advantage of event identification and significant reduction in computational complexity. The proposed technique also achieves a 439% improvement in the detection rate as compared to ST-QS-SVM. The detection rates for STA-TSV, however, are greater than the proposed technique.

The results for synthetic and real data sets clearly state that the performance of the proposed technique is equivalent to QS-SVM based techniques, i.e, STA-QS-SVM [2], STA-TASV, STA-TSV and STA-CA [3]. This comparable performance can be explained in terms of temporal-attribute correlations which are taken into account by STA-QS-SVM, STA-TASV, STA-TSV, STA-CA as well as the proposed technique. *Temporal correlations are taken into account by using the forgetting factor technique for outlier detection and updating the cluster parameters using the most recent data samples (section II). Attribute correlations are taken into account by the consideration of inverse covariance matrix in the calculation of mahalanobis distance (eq.* (1)*).* ST-QS-SVM has a much lower detection rate as compared to other techniques, as it does not incorporate attribute correlations of data.

VI. Conclusion

This paper introduces the problem of **clustering based event detection and identification** in WSNs that has been studied for the first time. Further, it presents a novel, online and computationally efficient clustering based strategy for joint event detection & identification that incorporates temporal-attribute correlations, specific to optimal event detection techniques. The proposed technique achieves a high detection rate (close to 99%), a very low false positive rate (close to 0.3%) and a significant reduction in computational complexity (from $O(n^2)$ to $O(nd^2)$) as compared to the state-of-the-art QS-SVM based techniques for synthetic and real data sets. A remarkable achievement of the proposed technique is its **event identification**, i.e, it identifies the key attributes involved in an event in the order of their increasing contribution towards the event; A problem that has been dealt for the very first time and provides an open challenge to the research community to build and improve on the proposed algorithm.

References

[1] Y. Zhang, N. Hamm, N. Meratnia, A. Stein, M. van de Voort, and P. Havinga, "Statistics-based outlier detection for wireless sensor networks," 2012.

[2] N. Shahid, I. H. Naqvi, and S. B. Qaisar, "Quarter-Sphere SVM: attribute and Spatio-Temporal correlations based outlier & event detection in wireless sensor networks," in *2012 IEEE Wireless Communications and Networking Conference: Mobile and Wireless Networks (IEEE WCNC 2012 Track 3 Mobile & Wireless)*, (Paris, France), Apr. 2012.

[3] N. Shahid, I. H. Naqvi, and S. B. Qaisar, "Real time energy efficient approach to outlier & event detection in wireless sensor networks," in *13th IEEE International Conference on Communication Systems 2012 (IEEE ICCS'12)*, (Singapore, Singapore), Nov. 2012.

[4] N. Shahid and I. H. Naqvi, "Energy efficient outlier detection in wsns based on temporal and attribute correlations," in *International Conference On Emerging Technologies*, 2011.

[5] E. Dereszynski and T. Dietterich, "Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns," *ACM Transactions on Sensor Networks (TOSN)*, vol. 8, no. 1, p. 3, 2011.

[6] S. Rajasegarar, C. Leckie, J. Bezdek, and M. Palaniswami, "Centered hyperspherical and hyperellipsoidal one-class support vector machines for anomaly detection in sensor networks," *Information Forensics and Security, IEEE Transactions on*, vol. 5, no. 3, pp. 518–533, 2010.

[7] V. Gomez-Verdejo, J. Arenas-Garcia, M. Lazaro-Gredilla, and A. Navia-Vazquez, "Adaptive one-class support vector machine," *Signal Processing, IEEE Transactions on*, vol. 59, pp. 2975 –2981, june 2011.

[8] S. Rajasegarar, J. Bezdek, M. Moshtaghi, C. Leckie, T. Havens, and M. Palaniswami, "Measures for clustering and anomaly detection in sets of higher dimensional ellipsoids," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pp. 1–8, IEEE, 2012.

[9] M. Moshtaghi, T. Havens, J. Bezdek, L. Park, C. Leckie, S. Rajasegarar, J. Keller, and M. Palaniswami, "Clustering ellipses for anomaly detection," *Pattern Recognition*, vol. 44, no. 1, pp. 55–69, 2011.

[10] M. Moshtaghi, C. Leckie, S. Karunasekera, J. Bezdek, S. Rajasegarar, and M. Palaniswami, "Incremental elliptical boundary estimation for anomaly detection in wireless sensor networks," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pp. 467–476, IEEE, 2011.

[11] M. Moshtaghi, S. Rajasegarar, C. Leckie, and S. Karunasekera, "An efficient hyperellipsoidal clustering algorithm for resource-constrained environments," *Pattern Recognition*, 2011.

[12] S. Suthaharan, C. Leckie, M. Moshtaghi, S. Karunasekera, and S. Rajasegarar, "Sensor data boundary estimation for anomaly detection in wireless sensor networks," in *Mobile Adhoc and Sensor Systems (MASS), 2010 IEEE 7th International Conference on*, pp. 546–551, IEEE, 2010.

[13] P. Misra, S. Kanhere, D. Ostry, and S. Jha, "Safety assurance and rescue communication systems in high-stress environments: A mining case study," *Communications Magazine, IEEE*, vol. 48, pp. 66 –73, april 2010.

[14] S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami, "Labelled data collection for anomaly detection in wireless sensor networks," in *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2010 Sixth International Conference on*, pp. 269 –274, dec. 2010.