Hyperellipsoidal SVM-based Outlier Detection Technique for GeoSensor Networks

Yang Zhang, Nirvana Meratnia, and Paul Havinga

Pervasive Systems Group, University of Twente,
Drienerlolaan 5, 7522NB Enschede, The Netherlands
{zhangy,meratnia,havinga}@cs.utwente.nl

Abstract. Recently, wireless sensor networks providing fine-grained spatio-temporal observations have become one of the major monitoring platforms for geo-applications. Along side data acquisition, outlier detection is essential in geosensor networks to ensure data quality, secure monitoring and reliable detection of interesting and critical events. A key challenge for outlier detection in these geosensor networks is accurate identification of outliers in a distributed and online manner while maintaining resource consumption low. In this paper, we propose an online outlier detection technique based on one-class hyperellipsoidal SVM and take advantage of spatial and temporal correlations that exist between sensor data to cooperatively identify outliers. Experiments with both synthetic and real data show that our online outlier detection technique achieves better detection accuracy compared to the existing SVM-based outlier detection techniques designed for sensor networks. We also show that understanding data distribution and correlations among sensor data is essential to select the most suitable outlier detection technique.

Key words: Geosensor networks, outlier detection, data mining, one-class support vector machine, spatio-temporal correlation

1 Introduction

Advances in sensor technology and wireless communication have enabled deployment of low-cost and low-power sensor nodes that are integrated with sensing, processing, and wireless communication capabilities. A geosensor network consists of a large number of these sensor nodes distributed in a large area to collaboratively monitor phenomena of interest. The monitored geographic space may vary in size and can range from small-scale room-sized spaces to highly complex dynamics of ecosystem regions [1]. Emerging applications of large-scale geosensor networks include environmental monitoring, precision agriculture, disaster management, early warning systems and wildlife tracking [1]. In a typical application, a geosensor network collects and analyzes continuous streams of fine-grained geosensor data, detects events, makes decisions, and takes actions.

Wireless geosensor networks have strong resource constraints in terms of energy, memory, computational capacity, and communication bandwidth. Moreover, the autonomous and self-organizing vision of these networks make them a good candidate for operating in harsh and unattended environments. Resource constraints and environmental effects cause wireless geosensor networks to be more vulnerable to noise,

faults, and malicious activities (e.g., denial of service attacks or black hole attacks), and more often generate unreliable and inaccurate sensor readings. Thus, to ensure a reasonable data quality, secure monitoring and reliable detection of interesting and critical events, identifying anomalous measurements locally at the point of action (at the sensor node itself) is a must. These anomalous measurements, usually known as outliers or anomalies, are defined as measurements that do not conform with the normal behavioral pattern of the sensed data [2].

Unlike traditional outlier detection techniques performed off-line in a centralized manner, limited resources available in sensor networks and specific nature of geosensor data necessitate outlier detection to be performed in a distributed and online manner to reduce communication overhead and enable fast respond. This implies that outliers in distributed streaming data should accurately be detected in real-time while maintaining resource consumption low. In this paper, we propose an online outlier detection technique based on one-class hyperellipsoidal Support Vector Machine (SVM) and take advantage of spatial and temporal correlation that exist between sensor data to cooperatively identify outliers. Experiments with both synthetic and real data obtained from the EPFL SensorScope System [3] show that our online outlier detection technique achieves better detection accuracy and lower false alarm compared to the existing SVM-based outlier detection techniques [4], [5] designed for sensor networks.

The remainder of this paper is organized as follows. Related work on one-class SVM-based outlier detection techniques is presented in Section 2. Fundamentals of the one-class hyperellipsoidal SVM are described in Section 3. Our proposed distributed and online outlier detection technique is explained in Section 4. Experimental results and performance evaluation are reported in Section 5. The paper is concluded in Section 6 with plans for future research.

2 Related Work

Generally speaking, outlier detection techniques can be categorized into statistical-based, nearest neighbor-based, clustering-based, classification-based, and spectral decomposition-based approaches [2], [6]. SVM-based techniques are one of the popular classification-based approaches due to the fact that they (i) do not require an explicit statistical model, (ii) provide an optimum solution for classification by maximizing the margin of the decision boundary, and (iii) avoid the curse of dimensionality problem.

One of the challenges faced by SVM-based outlier detection techniques for sensor networks is obtaining error-free or labelled data for training. One-class (unsupervised) SVM-based techniques can address this challenge by modelling the normal behavior of the unlabelled data while automatically ignoring the anomalies existed in the training set. The main idea of one-class SVM-based outlier detection techniques is to use a non-linear function to map the data vectors (measurements) collected from the original space (input space) to a higher dimensional space (feature space). Then a decision boundary of normal data will be found that encompasses the majority of the data vectors in the feature space. Those new unseen data vectors falling outside the boundary are classified as outliers. Scholkopf et al. [7] have proposed a hyperplane-based one-class SVM, which identifies outliers by fitting a hyperplane from the origin.

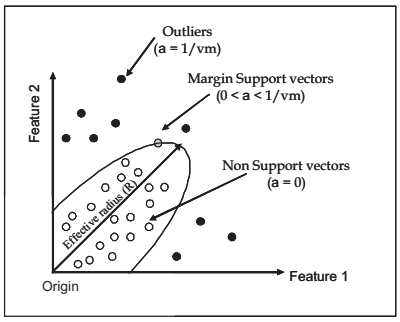

Fig. 1. Geometry of the hyperellipsoidal formulation of one-class SVM [4].

Tax et al. [8] have proposed a hypersphere-based one-class SVM, which identifies outliers by fitting a hypersphere with a minimal radius. Wang et al. [9] have proposed a hyperellipsoid-based one-class SVM, which identifies outliers by fitting multiple hyperellipsoids with minimum effective radii.

In addition to obtaining the labelled data, another challenge faced by SVM-based outlier detection techniques is their quadratic optimization during the learning process for the normal boundary. This process is extremely costly and not suitable for limited resources available in sensor networks. Laskov et al. [10] have extended work in [8] by proposing a one-class quarter-sphere SVM, which is formulated as a linear optimization problem by fitting a hypersphere centered at the origin and thus reducing the effort and computational complexity. Rajasegarar et al. [11] and Zhang et al. [5] have further exploited potential of the one-class quarter-sphere SVM of [10] for distributed outlier detection in sensor networks. The main difference of these two techniques is that unlike a batch technique of [11], the work of [5] aims at identifying every new measurement collected at a node as normal or anomalous in an online manner.

Rajasegarar et al. [4] have also extended work in [9] [10] by proposing a one-class centered hyperellipsodal SVM with linear optimization. However, this technique is neither distributed nor online. In this paper, we extend work in [4] and propose a distributed and online outlier detection technique suitable for geosensor networks, with low computational complexity and memory usage.

3 Fundamentals of the One-Class Hyperellipsoidal SVM

In our proposed technique, we exploit the one-class hyperellipsoidal SVM [9], [4] to learn the normal behavioral pattern of sensor measurements. The quadric optimization problem of the one-class hyperellipsoidal SVM has been converted to a linear optimization problem in [4] by fixing the center of the hyperellipsoidal at the origin. A hyperellipsoidal boundary is used to enclose the majority of the data vectors in the feature space. The geometries of the one-class centered hyperellipsoidal SVM-based approach is shown in Figure 1.

The constrain for optimization problem of the one-class centered hyperspherical SVM is formalized as follows:

$$\min_{R\epsilon\Re,\xi\epsilon\Re^m} R^2 + \tfrac{1}{\upsilon m}\sum_{i=1}^{m}\xi_i \tag{1}$$

$$subject\ to: \phi(x_i)\Sigma^{-1}\phi(x_i)^T \leq R^2 + \xi_i, \xi_i \geq 0, i = 1, 2, \ldots m$$

where m denotes number of data vectors in the training set. The parameter $\upsilon\ \epsilon\ (0, 1)$ controls the fraction of data vectors that can be outliers. Σ^{-1} is the inverse of the covariance matrix $\Sigma = \frac{1}{m}\sum_{i=1}^{m}(\phi(x_i) - \mu)(\phi(x_i) - \mu)^T$, $\mu = \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)$. Using Mercer Kernels [12], the dot product computations of data vectors in the feature space can be computed in the input data space. The centered kernel matrix K_c can be obtained in terms of the kernel matrix K using $K_c = K - 1_m K - K1_m + 1_m K 1_m$, where 1_m is the $m \times m$ matrix with all values equal to $\frac{1}{m}$. Finally, the dual formulation of (1) will become a linear optimization problem formulated as follows:

$$\min_{\alpha\epsilon\Re^m} -\sum_{i=1}^{m}\alpha_i \|\sqrt{m}\Lambda^{-1}P^T K_c^i\|^2 \tag{2}$$

$$subject\ to: \sum_{i=1}^{m}\alpha_i = 1, 0 \leq \alpha_i \leq \frac{1}{\upsilon m}, i = 1, 2, \ldots m$$

where Λ is a diagonal matrix with positive eigenvalues, P is the eigenvector matrix corresponding to the positive eigenvalues [13], and K_c^i is the i^{th} column of the kernel matrix K_c. From equation (2), the $\{\alpha_i\}$ value can be easily obtained using some effective linear optimization techniques [14]. The data vectors in the training set can be classified depending on the results of $\{\alpha_i\}$, as shown in Figure 1. The training data vectors with $0 \leq \alpha \leq \frac{1}{\upsilon m}$, which fall on the hyperellipsoid, are called margin support vectors. The effective radius of the hyperellipsoid $R = \|\sqrt{m}\Lambda^{-1}P^T K_c^i\|$ can be computed using any margin support vector.

4 A Distributed and Online Outlier Detection Technique for GeoSensor Networks

In this section, we will describe our distributed and online outlier detection technique. This proposed technique aims at identifying every new measurement collected at each node as normal or anomalous in real-time. Moreover, using high degree of spatio-temporal correlations that exist among the sensor readings, each node exchanges the learned normal boundary with its spatially neighboring nodes and combines their learned normal boundaries to cooperatively identify outliers. Before describing this technique in details, we present our assumptions and explain why we exploit the hyperellipsoidal SVM instead of hyperspherical SVM to learn the normal behavioral pattern of sensor measurements.

4.1 Assumptions

We assume that wireless sensor nodes are time synchronized and densely deployed in a homogeneous geosensor network, where sensor data tends to be correlated in both time and space. A sensor sub-network consists of n sensor nodes $S_1, S_2, \ldots S_n$, which are within radio transmission range of each other. This means that each node has *n-1* spatially neighboring nodes in the sub-network. At each time interval Δ_i, each sensor node in the sub-network measures a data vector. Let $x_1^i, x_2^i, \ldots, x_n^i$ denote the data vector measured at $S_1, S_2, \ldots S_n$, respectively. Each data vector is composed of multiple attributes x_j^{il}, where $x_j^i = \{x_j^{il} : j = 1 \ldots n, l = 1 \ldots d\}$ and $x_j^i \, \epsilon \, \Re^d$. Our aim is online identification of every new measurement collected at each node as normal or anomalous by means of local processing at the node itself. In addition to near real-time identification of outliers, increasing data quality, and reducing communication overhead, this local processing also has the advantage of coping with (possibly) large scale of the geosensor network.

4.2 Hyperellipsoidal SVM VS Hyperspherical SVM

In this paper, we exploit the hyperellipsoidal SVM instead of hyperspherical SVM to learn the normal behavioral pattern of sensor measurements. The reason for doing so is the fact that hyperspherical SVM assumes that the target sample points are distributed around the center of mass in an ideal spherical manner. However, if the data distribution is non-spherical, using a spherical boundary to fit the data will increase the false alarm rate and reduces the detection rate. This is because many superfluous outlier are mistakenly considered in the boundary and consequently outliers are classified as normal.

On the contrary, the hyperellipsoidal SVM is able to best capture multivariate data structures by considering not only the distance from the center of mass but also the data distribution trend, where the latter is learned by building the covariance matrix of the sample points. This feature can be used well for geosensor data, where multivariate attributes may induce certain correlation, e.g., the readings of humidity sensors are negatively correlated to the readings of temperature sensors. Unlike using the Euclidean distance in the hyperspherical SVM, the distance metric adopted in the hyperellipsoidal SVM is the Mahalanobis distance. The Mahalanobis distance takes the shape of the multivariate data distribution into account and identifies the correlations of data attributes. Thus using an ellipsoidal boundary to enclose geosensor data aims to increase outlier detection accuracy and reduce the false alarm rate. However, as a tradeoff, the hyperellipsoidal SVM has more computational and memory usage cost than the hyperspherical SVM. To correctly select the most appropriate outlier detection technique, we believe that having some understanding about data distribution and correlation among sensor data is crucial.

4.3 Hyperellipsoidal SVM-based Outlier Detection Techniques

The main idea behind our proposed Hyperellipsoidal SVM-based online outlier detection technique (OOD_E) is that each node builds a normal boundary representing normal behavior of the sensed data and then exchanges the learned normal boundary with its spatially neighboring nodes. A sensor measurement collected at a node

is identified as an outlier if it does not fit inside the boundary defined at the node and also does not fit inside the combined boundaries of the spatially neighboring nodes. We first explain the OOD_E technique in the input and feature spaces and then present the corresponding pseudocode in Table 1.

OOD_E in the Input Space Initially, each node learns the local effective radius of the hyperellipsoid using its m sequential data measurements, which may include some anomalous data. In the input space, equation (1) can be formalized as equation (3). The one-class hyperellipsoidal SVM can efficiently find a minimum effective radius R to enclose the majority of these sensor measurements in the input space. Each node then locally broadcasts the learned radius information to its spatially neighboring nodes. When receiving the radii from all of its neighbors, each node computes a median radius R_m of its neighboring nodes. We use median because in estimating the "center" of a sample set, the median is more accurate than the mean.

$$\min_{R\epsilon\Re,\xi\epsilon\Re^m} R^2 + \tfrac{1}{\upsilon m}\sum_{i=1}^{m}\xi_i \tag{3}$$

$$subject\ to: (x_i-\mu)\Sigma^{-1}(x_i-\mu)^T \le R^2+\xi_i, \xi_i \ge 0, i=1,2,\dots m$$

Sensor data collected in a densely deployed geosensor network tends to be spatially and temporally correlated [1]. When a new sensor measurement x is collected at node S_i, node S_i first compares the Mahalanobis distance of x with its local effective radius R_i. In the input space, the mean can be expressed as $\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$, and thus the Mahalanobis distance of x is formulated as follows:

$$Md(x) = \sqrt{(x-\mu)\Sigma^{-1}(x-\mu)^T} = \|\Sigma^{-\frac{1}{2}}(x-\frac{1}{m}\sum_{i=1}^{m} x_i)\| \tag{4}$$

The data x will be classified as normal if $Md(x) <= R_i$. This means that x falls on or inside the hyperellipsoid defined at S_i. If $Md(x) > R_i$, S_i further compares $Md(x)$ with the median radius R_{im} of its spatially neighboring nodes. Then if $Md(x) > R_{im}$, x will finally be classified as an outlier. The decision function to declare a measurement as normal or outlier can be formulated as equation (5), where a reading with a negative value is classified as an outlier.

$$f(x) = sgn(max(R-Md(x), R_m - Md(x))) \tag{5}$$

The computational complexity of OOD_E in the input space is low as it only depends on solving a linear optimization problem presented in equation (3) and simple computations expressed by equations (4) and (5). Once the optimization is solved, each node only keeps the effective radius value, the mean, and the covariance matrix obtained from the training data in memory. Using the radius information from adjacent nodes is to reduce high false alarm caused by unsupervised learning techniques.

OOD$_E$ in the Feature Space Each node learns the local effective radius of the hyperellipsoid using equation (1) and (2), and then exchanges the learned radius information with its spatially neighboring nodes. In the input space, the mean can be expressed as $\mu = \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)$, and thus the Mahalanobis distance of each new measurement x in the feature space can be formalized as follows:

$$Md(x) = \sqrt{(\phi(x)-\mu)\Sigma^{-1}(\phi(x)-\mu)^T} = \|\sqrt{m}\Lambda^{-1}P^T K_c^x\| \quad (6)$$

Then the data x will be classified as normal or anomalous using the same decision function as the equation (5). Due to high computational cost and memory usage required for classification of each new sensor measurement as normal or anomalous, we modify the OOD$_E$ in such a way that it does not run for every new sensor reading and it waits until a few measurements $X = \{x^n : n = 1 \ldots t\}$ are collected. The centered kernel matrix K_c^X can be obtained by using $K_c^X = K^X - 1_{tm}K - K^X 1_m + 1_{tm}K1_m$, where 1_{tm} is the $t \times m$ matrix with all values equal to $\frac{1}{m}$. This modification reduces the computational complexity and also facilitates linking outliers to actual events in later stages.

```
1 procedure LearningSVM()
2   each node collects m sensor measurements for learning its own effective radius R
    and locally broadcasts the radius to its spatially neighboring nodes;
3   each node then computes R_m;
4   initiate OutlierDetectionProcess(R, R_m);
5   return;
6 procedure OutlierDetectionProcess(R, R_m)
7   when a new measurement x arrives
8       compute Md(x);
9       if (Md(x) > R AND Md(x) > R_m)
10          x indicates an outlier;
11      else
12          x indicates a normal measurement;
13      endif;
14  return;
```

Table 1. Pseudocode of the OOD$_E$

5 Experimental Results and Evaluation

The goals of expreiments are two folds. First we evaluate performance of our distributed and online technique compared to the batch hyperellipsoidal SVM-based outlier detection technique (BOD$_E$) presented in [4] and the online quarter-sphere SVM-based outlier detection technique (OOD$_Q$) presented in [5]. Secondly, we investigated impact of data distribution and spatial/spatio-temporal correlations in performance of our outlier detection technique. In experiments, we use synthetic data as well as real data gathered from a geosensor network deployment by the EPFL [3].

5.1 Datasets

For the simulation, we use Matlab and consider a sensor sub-network consisting of seven sensor nodes. Sensor nodes are within the one-hop range of each other. Two 2-D synthetic data distributions with 10% (of the normal data) anomalous data are shown in Figure 2(a) and 4(a). It can clearly be seen that Figure 4(a) has a concentrated distribution around the origin while the data distribution shown in Figure 2(a) is not spherical but has a certain trend. The data values are normalized to fit in the [0, 1]. The BOD_E performs outlier detection when all measurements are collected at each node, while the OOD_Q operates in a distributed and online manner.

The real data is collected from a closed neighborhood by a geosensor network deployed in Grand-St-Bernard. Figure 3(a) shows the deployment area. The closed neighborhood contains the node 31 and its 4 spatially neighboring nodes, namely nodes 25, 28, 29, 32. The network records ambient temperature, relative humidity, soil moisture, solar radiation and watermark measurements at 2 minutes intervals. In our experiments, we use ambient temperature and relative humidity collected during the period of 9am-5pm on the 5th October 2007. The labels of measurements are obtained based on degree of dissimilarity between data measurements.

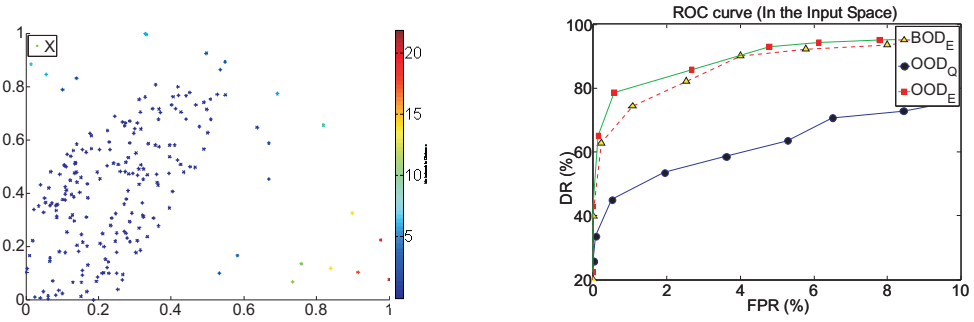

Fig. 2. (a) Plot for synthetic data; (b) ROC curves in the input space.

5.2 Experimental Results and Evaluation

We have evaluated two important performance metrics, the detection rate (DR), which represents the percentage of anomalous data that are correctly classified as outliers, and the false alarm rate, also known as false positive rate (FPR), which represents the percentage of normal data that are incorrectly considered as outliers. A receiver operating characteristics (ROC) curve is used to represent the trade-off between the detection rate and the false alarm rate. The larger the area under the ROC curve, the better the performance of the technique.

We have examined the effect of the regularization parameter υ for OOD_E, BOD_E, and OOD_Q. υ represents the fraction of data vectors that can be outliers. For synthetic dataset, we varied υ from 0.02 to 0.18 in intervals of 0.02 and evaluated the detection accuracy of the three techniques in the input space. For real dataset, we

varied υ from 0.01 to 0.10 in intervals of 0.01 and used Polynomial kernel function to evaluate the accuracy performance of three techniques in the feature space. The Polynomial kernel function is formulated as: $k_{POLY} = (x_1.x_2 + 1)^r$, where r is the degree of the polynomial.

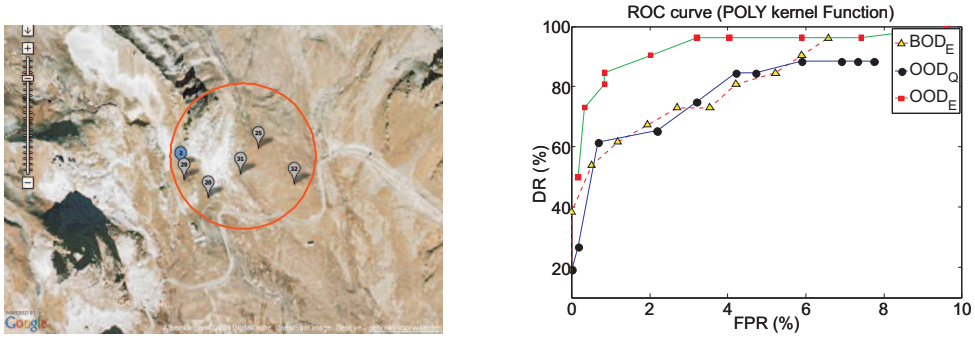

Fig. 3. (a) Grand-St-Bernard deployment in [3]; (b) ROC curves in the feature space.

Figure 2(b) and 3(b) show the ROC curves obtained for the three techniques in the input space for synthetic data and using Polynomial kernel function for real data. Simulation results show that our OOD_E always outperforms BOD_E and OOD_Q. Moreover, quarter-sphere SVM-based OOD_Q is obviously worse than the hyperellipsoidal SVM-based OOD_E and BOD_E in the input space for synthetic data. For real data in the feature space, the performance of OOD_Q and BOD_E is not very obvious to distinguish.

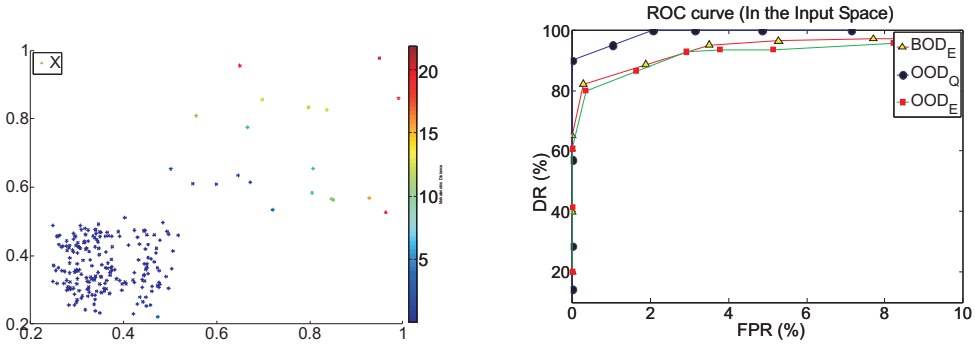

Fig. 4. (a) Plot for synthetic data; (b) ROC curves in the input space.

Although experiments show that hyperellipsoidal SVM-based techniques outperform quarter-sphere SVM-based technique, our experiments show that this greatly depends on data distribution and correlations that exist between sensor data. It can be clearly seen from Figure 4(b), the quarter-sphere SVM-based OOD_Q has a better

performance than two hyperellipsoidal SVM-based outlier detection techniques in the input space for synthetic data. The obtained results conform with our idea about the need of having some understanding about data distribution and correlation among sensor data to be able to select the most suitable outlier detection technique.

6 Conclusions

In this paper, we have proposed a distributed and online outlier detection technique based on one-class hyperellipsoidal SVM for geosensor networks. We compare the performance of our techniques with the existing SVM-based techniques using both synthetic and real data sets. Experimental results show that our technique achieves better detection accuracy and lower false alarm. Our future research includes sequentially updating the normal boundary of the sensor data, online computation of spatio-temporal correlations and online distinction outliers between events and errors.

Acknowledgments. This work is supported by the EU's Seventh Framework Programme and the SENSEI project.

References

1. Nittel, S., Labrinidis, A., Stefanidis, A.: GeoSensor Networks. Springer (2006)
2. Chandola, V., Banerjee, A., Kumar, V.: Anomaly Detection: A Survey. Technical report, University of Minnesota (2007)
3. SensorScope, `http://sensorscope.epfl.ch/index.php/Main_Page`
4. Rajasegarar, S., Leckie, C., Palaniswami, M.: CESVM: Centered Hyperellipsoidal Support Vector Machine Based Anomaly Detection. In: IEEE International Conference on Communications, pp. 1610–1614. IEEE Press, Beijing (2008)
5. Zhang, Y., Meratnia, N., Havinga, P.J.M.: An Online Outlier Detection Technique for Wireless Sensor Networks using Unsupervised Quarter-Sphere Support Vector Machine. In: 4th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp. 151–156. IEEE Press, Sydney (2008)
6. Zhang, Y., Meratnia, N. and Havinga, P.J.M.: Outlier Detection Techniques for Wireless Sensor Network: A Survey. Technical report, University of Twente (2008)
7. Scholkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the Support of a High-Dimensional Distribution. Journal of Neural Computation, 13(7), 1443–1471 (2001)
8. Tax, D.M.J., Duin, R.P.W.: Support Vector Data Description. Journal of Machine Learning, 54(1), 45–56 (2004)
9. Wang, D., Yeung, D.S., Tsang, E.C.C.: Structured One-Class Classification. IEEE Transactions on System, Man and Cybernetics, 36(6), 1283–1295 (2006)
10. Laskov, P., Schafer, C., Kotenko, I.: Intrusion Detection in Unlabeled Data with Quarter Sphere Support Vector Machines. In: Detection of Intrusions and Malware & Vulnerability Assessment, pp. 71–82. Dortmund (2004)
11. Rajasegarar, S., Leckie, C., Palaniswami, M., Bezdek, J.C.: Quarter Sphere based Distributed Anomaly Detection in Wireless Sensor Networks. In: IEEE International Conference on Communications, pp. 3864–3869. IEEE Press, Glasgow (2007)
12. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons (1998)
13. Golub, G.H., Loan, C.F.V.: Matrix Computations. John Hopkins (1996)
14. Nash, S.G., Sofer, A.: Linear and Nonlinear Programming. McGrawHill (1996)