Fast and Efficient Outlier Detection Method in Wireless Sensor Networks

Oussama Ghorbel, Walid Ayedi, Hichem Snoussi, and Mohamed Abid

***Abstract*— Outlier detection is the task of classifying test data that differ in some respect from the data that are available during training. This may be seen as one-class classification, in which a model is constructed to describe normal training data. In wireless sensor networks (WSNs), the outlier detection process is a necessary step in building sensor network systems to assure data quality for perfect decision making. For this case, the task amounts to create a useful model based on kernel principal component analysis (KPCA) to recognize data as normal or outliers. Recently, KPCA has used for nonlinear case which can extract higher order statistics. KPCA mapping the data onto another feature space and using nonlinear function. On account of the attractive capability, KPCA-based methods have been extensively investigated, and it have showed excellent performance. Within this setting, we propose KPCA-based Mahalanobis kernel as a new outlier detection method using Mahalanobis distance to implicitly calculate the mapping of the data points in the feature space so that we can separate outlier points from normal pattern of data distribution. The use of KPCA-based Mahalanobis kernel on real-word data obtained from Intel Berkeley are reported showing that the proposed method performs better in finding outliers in WSNs when compared with the original reconstruction error-based variant and the one-class support vector machine detection approach. All computation are done in the original space, thus saving computing time using Mahalanobis kernel.**

***Index Terms*— Wireless sensor networks, outlier detection, kernel methods, mahalanobis kernel, kernel principal component analysis (KPCA), feature extraction, reconstruction error (RE), mahalanobis distance (MD), one-class support vector machine (OCSVM).**

I. Introduction

WIRELESS sensor networks have become an important source of data for such applications. These collaborated and interconnected sensors in wireless sensor networks produce data continuously which is uncertain and unreliable. Hence an effective processing and analysis of data streams becomes our utmost importance for various applications like outlier detection [1]. Wireless sensor networks are widely used and have gained attention in various fields including traffic control, health care, precision agriculture, etc [2]. Most WSN's applications require precise and accurate data to provide reliable information to the end user. Although the importance of information quality provided from WSNs, collected sensor data may be of low quality and reliability due to the low cost nature and harsh deployments of WSNs [3]. To ensure the quality of sensor measurements, outlier detection methods allow cleaning and refinement of collected data and let providing the most useful information to end users, while maintaining low energy consumption and preserve high computational efforts due to the limited energy resources of sensor nodes. To detect outliers, a detection model is built upon historical data structure of WSN. This model should be able to detect outliers among new observations with good precision [4]. The advantage of using WSNs is that they are cheaper and more practical then wired networks. However WSNs are vulnerable to intruders and faults. In general, WSNs need to be data mined to find anomalies as efficient as possible and then send these data for the base station or the central location for further processing. By means of an alternative way of computing the principal axes through the use of inner product evaluations, Principal Component Analysis has been extended to a kernel-based PCA. Dimensionality reduction by principal component analysis (PCA) is a trusted machine learning workhorse, kernel based methods for non-linear dimensionality reduction are only starting to find application. The use of non-linear dimensionality reduction to expand in many applications as recent research has shown that kernel principal component analysis (KPCA) can be expected to work well as a pre-processing device for pattern recognition [5]. The use of KPCA is a new field on wireless sensor networks (WSN) which are composed of interconnected micro-sensors that are able to collect, store, process and transmit data over the wireless channel. KPCA has found a new field which is integrated in application of outlier detection. Compared with the conventional data collection techniques, wireless sensor networks can provide continuous measurements of physical phenomena by means of dense deployments of sensor nodes. Wireless sensor networks are widely used and have gained attention in various fields including traffic control, health care, precision agriculture, etc [6]. KPCA has been used in several applications, such as voice recognition, image segmentation, face detection, feature extraction, data denoising and etc [7]. The main contribution of our work is the uses of Mahalanobis kernel based KPCA for outlier detection method in wireless sensor networks. To identify outliers, we use Mahalanobis distance induced feature subspace spanned by principal components as obtained by Kernel PCA. If the distance of a new data point is above a prefixed

Manuscript received December 1, 2014; accepted December 27, 2014. Date of publication January 7, 2015; date of current version April 22, 2015. The associate editor coordinating the review of this paper and approving it for publication was Dr. Amitava Chatterjee.

O. Ghorbel, W. Ayedi, and M. Abid are with the Ecole Nationale d'Ingénieurs de Sfax, Sfax 3038, Tunisia (e-mail: oussama.ghorbel@ceslab.org; ayediwalid@yahoo.fr; mohamed.abid@enis.rnu.tn).

H. Snoussi is with the University of Technology of Troyes, Troyes 10010, France (e-mail: hichem.snoussi@utt.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSEN.2015.2388498

threshold, the observation is considered as an outlier. We define now reconstruction error which is a measure of deviation from the principal subspace. It assumes that the principal subspace represents the normal data. When reconstruction error exceeds a certain threshold, test data are identified as outlier, which is also established experimentally. Therefore, we advocate the use of Mahalanobis distance within the principal subspace as an alternative to the reconstruction error. The model is tested on real data from Intel Berkeley. The obtained results are competitive and the proposed method can achieve high detection rate with the lowest false alarm rate. However, our proposed approach using Mahalanobis distance can deal with the false alarm problem efficiently than using reconstruction error and one class SVM. For the latter ones, we draw comparisons in the form of tables and figures based on the Area Under the Receiver Operating Characteristic (ROC) Curve.

The remainder of this article is organized as follows. Section 2, present the related work for KPCA. Section 3, describes outliers detection and its different category in wireless sensor networks. Section 4, describes adopted method. Section 5, showcases the obtained experimental results, and section 6 concludes and summarizes the main outcomes of the paper.

II. Related Work

Principal Component Analysis (PCA) is a data analysis technique used for reducing the dimensionality of correlated data. It transforming them into a set of uncorrelated variables called Principal Components (PCs) [8], [9]. The work done by [9] introduce how detecting outliers and identifying faulty nodes using PCA. This model showed two types of analysis; offline analysis and real time analysis. Many other works using PCA based outlier detection scheme is introduced by [10] in which PCA was used with a fixed width clustering algorithm to establish the global normal profile. More recently, an outlier aware data aggregation technique using distributed PCA has been proposed by [11] in the field of Wireless Sensor Networks. The work done by the authors [12] describes a data fault detection method for WSNs using multi-scale PCA. They integrate wavelet analysis with PCA which is used to capture the time frequency information while PCA was used to detect the faulty data.

Kernel based principle components analysis is a non-linear PCA created using the kernel trick. KPCA maps the original inputs into a high dimensional feature space using a kernel method [13]. Mathematically, we transform the current features into a high-dimensional space and the calculate eigenvectors in this space. We ignore the vectors with really low eigenvalues and then do learning in this transformed space. KPCA is computationally intensive and takes a lot more time compared to PCA. The reason being that the number of training data points in KPCA is much higher than PCA. So number of principle components that need to be estimated is also much larger. The KPCA method has exhibited superior performance compared to linear PC analysis method in processing nonlinear systems [14], [15]. The detail introduction of the basic KPCA can be viewed in [16]. Kernel PCA (KPCA), as presented by Scholkopf et al., is a technique for nonlinear dimension reduction of data with an underlying nonlinear spatial structure. A key insight behind KPCA is to transform the input data into a higher-dimensional feature space (Fig 1).

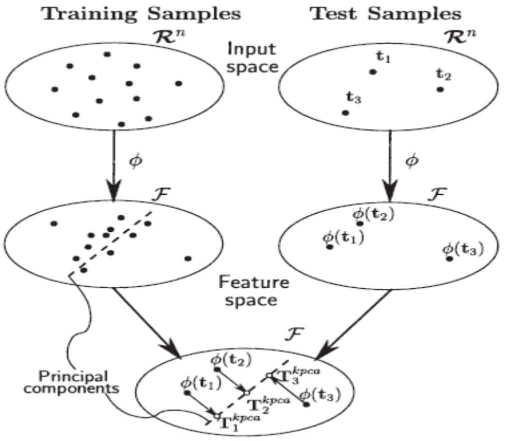

Fig. 1. One Kernel-PCA representation technique. The left-hand-side frame corresponds to the Training Samples (input space) and the right-hand-side to the Test Samples (feature space).

In [6], the authors used kernel PCA with Gaussian kernel for fault detection and identification of process monitoring in the field of chemical engineering. In [18] and [19], the authors used the greedy KPCA which essentially works by *filtering* or *sampling* the original training set for a lesser but representative subset of vectors which span approximately the same subspace as the subspace in the kernel induced feature space spanned by the training set. The training set is then projected onto the span of the lesser subset, where PCA is carried out. Other sampling-based methods exist [20]. Current KPCA reconstruction methods equally weigh all the features; it is impossible to weigh the importance of some features over the others.

Other existing methods also have limitations. Some works only considers robustness of the principal subspace; they do not address robust fitting. Lu et al present an iterative approach to handle outliers in training data. At each iteration, the KPCA model is built, and the data points that have the highest reconstruction errors are regarded as outliers and discarded from the training set. However, this approach does not handle intra-sample outliers. Several other approaches also considering Berar et al propose to use KPCA with polynomial kernels to handle missing data. However, it is not clear how to extend this approach to other kernels. Furthermore, with polynomial kernels of high degree, the objective function is hard to optimize. Sanguinetti & Lawrence propose an elegant framework to handle missing data. The framework is based on the probabilistic interpretation inherited from Probabilistic PCA. However, Sanguinetti & Lawrence do not address the problem of outliers.

Rassam et al [29] propose an OCPCC model that is able to track the dynamic normal changes of data streams in the monitored environment. The efficiency and effectiveness of the proposed models are demonstrated using real life datasets collected by real sensor network projects. Experimental results show that the proposed models have advantages over existing

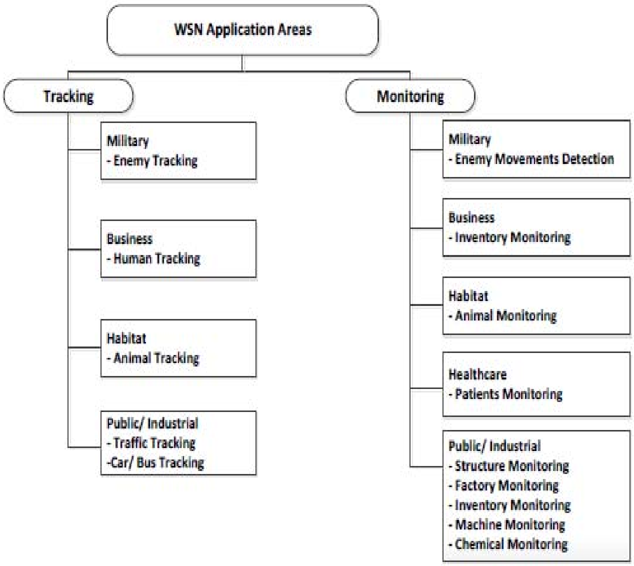

Fig. 2. Different application areas in WSNs.

models in terms of efficient utilization of sensor limited resources. The results further reveal that the proposed models achieve better detection effectiveness in terms of high detection accuracy with low false alarms.

This paper presents a novel cost function based on mahalanobis kernel using mahalanobis distance that unifies the treatment of outliers by KPCA in WSNs domain. Some previous work use KPCA with his standard form wish is not applicable in WSNs domain. So, our approach based on KPCA using mahalanobis kernel with mahalanobis distance can be used to detect outlier in WSNs whish, as we know, not used yet. Experiments show that our algorithm outperforms existing approaches. So, its performance is in terms of accuracy, detection rate and false alarm rate.

III. Outlier Detection in Wireless Sensor Networks

Outlier is used for finding errors, noise, missing values, inconsistent data, or duplicate data. This abnormal value may affect the quality of data and reduces the system performance. There are three sources of outliers occurred in WSNs: errors, events, and malicious attacks. The use of Outlier detection technique is very important in several real life applications, such as, environmental monitoring, health and medical monitoring, industrial monitoring, surveillance monitors and target tracking [13] as shown in figure 2.

In wireless sensor networks, the sensors have low cost and low energy, so to improve the quality and performance, the better solution is to use outlier detection technique [29]. Evaluation of an outlier detection technique for WSNs depends on whether it can satisfy the mining accuracy requirements while maintaining the resource consumptions of WSNs to a minimum. Outlier detection techniques are required to maintain a high detection rate while keeping the false alarm rate (number of normal data that are incorrectly considered as outliers) low [22]. A receiver operating characteristic (ROC) curves usually is used to represent the trade-off between the detection rate and false alarm rate. For the problem, we can summarize many problems in detection of outliers in WSNs as follows:

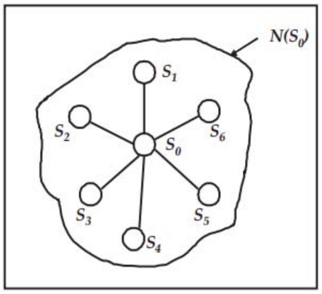

Fig. 3. Example of a closed neighborhood $N(S_i)$ of the sensor node S_i.

- High communication cost
- Modeling normal objects and outliers effectively
- Application specific outlier detection
- Identifying outlier source
- Distributed data
- Communication failures frequently
- Dynamic network topology.

A. Outlier Detection Method

There are many outlier detection method used in Wireless sensor networks mentioned as follow:

- Statistics-based outlier detection techniques
- Distance-based outlier detection techniques
- Clustering-based outlier detection techniques
- Classification-based outlier detection techniques.

B. Categories of Outlier Detection Technique

We classify also outlier detection techniques according to the following five general categories:

- Probabilistic,
- Distance-based,
- Reconstruction-based,
- Domain-based,
- Information-theoretic techniques.

IV. Adopted Method

A sensor network consist a collection of sensor that can measure characteristics of their local environment from real world physical phenomenon. It performs certain computation, and transmits the collected data samples to base station. Then it is partitioned onto groups or clusters. Each group consists of a cluster head and a number of members. Nodes which belong to the same cluster are geographically close and monitoring generally similar phenomenon (Fig 3). In this work, we will not take into consideration clustering details. We assume that the network is pre-partitioned and the clusters are predefined: every cluster is defined by his cluster head and members.

A. Problem Formulation

Let's consider a set of n sensor nodes measuring each one a single real valued attribute X_i at each time instant

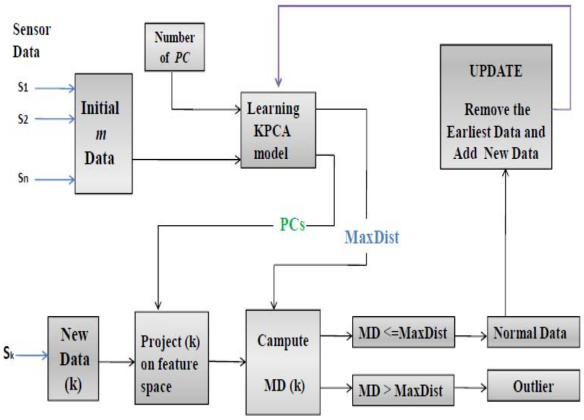

Fig. 4. Architecture of our model of outlier detection in WSN.

where $X = (X_1, \ldots, X_m)$ is an m-dimensional random variable. In every time interval Δt_k every node $s_i \in n$ captures a data vector $x_k^i \in \mathbb{R}^d$ composed of j dimensions such that: $x_k^i = (x_{k1}^i, x_{k2,}^i, \ldots, x_{kj}^i)$. During each time window t, s_i captures a set of data measurements $X_k^i = \{x_k^i(t),\ k = 1 \ldots n_i\}$.

Our goal is detecting outlier observations among data vectors collected by sensor nodes. Every one sends his local measurements to the cluster head which collects all data vectors and combines it with his data vector. First, outlier detection algorithm is performed in the cluster head's node using Kernel PCA with Mahalanobis kernel. The detection model is built based on first stream of received data vectors on activation of detection feature. Below, the architecture of our model to detect outlier in wireless sensor networks.

Based on figure 4, our outlier detection model is described as follow: first, every sensor $(S1, \ldots, S6)$ send his data measurement (presented as initial(m) data) to the cluster head. This latter add his collected data to the previous collected sensor data. Combined with the number of principal component PC, the KPCA model is built to create the initial model which define the maximal distance (MaxDist). Then, when a new data arrived, it will be projected on the feature space and his mahalanobis distance is calculated. After that, this distance is compared to the threshold (MaxDist) to determine if this new data are classified as normal or outlier data. However, our model is updated, so the earliest data are removed and replaced by the new data. Finally, the learned KPCA mode is also updated.

B. Training Phase

In our centralized model, the training phase is the first step of every training task which is performed in the cluster head's (CH) node. This step is stored on the cluster head, the next data streams received are subject to outlier detection using the initial model. When the cluster head receiving data from sensors, it combines his normalized data vector with all received data vectors in a global data matrix:

$$X(0) = \begin{bmatrix} X_1(0) \\ X_2(0) \\ \\ X_n(0) \end{bmatrix}$$

where $X_i(0)$ is a $n \times d$ matrix of data vectors collected by members $s_i, i = 1 \ldots s$. Then the data matrix is normalized and global mean (GM) and global covariance matrix (GCM) are calculated [23].

To establish the detection model, the cluster head executes Kernel PCA on this data matrix: first by calculating global principal components (GPC). Then the projection distance of every observation of the global data vector on the subspace spanned by the maintained principal components is calculated to get the maximal distance.

The global model is defined by these three parameters: the global mean, the global principal components and the maximal distance $MaxDist$. For the Learning Procedure, we describe the process: First, we start with Collect data vector and normalize it. Second, send the data vector to the cluster head cluster head. Third, execute Kernel PCA on normalized data and finally, establish global model.

C. Detection Phase

In the detection step, the cluster head receives periodically data vectors from sensors. It calculate the projection distance of every data vector on the subspace defined by global principal component. Based on d_p, the cluster head could decide if an observation is outlier or normal using the maximal threshold of mahalanobis distance. So, if $d_p \leq MaxDist$, the observation are considered as normal, otherwise outlier.

$$\begin{cases} (d_p \leq MaxDist) : Class = Normal \\ (d_p > MaxDist) : Class = Outlier. \end{cases}$$

D. Mahalanobis Kernel

In the literature, many types of kernels were employed in the nonlinear transformation of data points (polynomial kernel, sigmoid kernel, RBF, etc...) but as we know, mahalanobis kernel was not used yet in the field of wireless sensor networks. The Mahalanobis kernel (MK) is defined as:

$$K(x_i, x_j) = \exp\left(\frac{-1}{2\sigma^2}(x_i - x_j)^T Q^{-1}(x_i - x_j)\right) \tag{1}$$

Transformation results of such a kernel are similar to those of a density estimator as it gives a weighted value w_i for every sample x_i of input space. This weighting is not defined for each variable separately although some variables may be more relevant than others in the practice [14]. Let $\{x_1, \ldots, x_N\}$ be a dataset composed of N data points of dimension m, we define the data center c and the covariance matrix Q:

$$c = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

$$Q = \frac{1}{N}\sum_{i=1}^{N} (x_i - c)(x_i - c)^T \tag{3}$$

The Mahalanobis distance between a point and the center is defined as:

$$d(x) = \sqrt{(x - c)^T Q^{-1} (x - c)} \tag{4}$$

We define the Mahalanobis kernel function as follow, where H is a positive semi definite matrix:

$$A(x, x') = \exp(-(x - x')^T H(x - x')) \tag{5}$$

In this case, the Mahalanobis distance is calculated between every data point pair x and x'. The Mahalanobis kernel is an extension of the RBF kernel when $H = \gamma I$ with $\gamma > 0$ is a parameter that controls the depth of the kernel and I is the identity matrix. In practice, the Mahalanobis kernel (MK) is calculated only for one class:

$$A(x, x') = \exp(-\frac{\delta}{m}(x - x')^T Q^{-1}(x - x')) \tag{6}$$

where $\delta > 0$ is a scale factor to control the Mahalanobis distance.

The MK kernel differs from the Gaussian kernel in the fact that for every dimension of the input space data it defines a specific depth value or weight. This makes the calculated decision boundary has a non-spherical shape relative to the center of data points. Using kernel PCA in a learning task has to be well carried out. Choosing the better parameters is important in order to establish the best model with higher accuracy and lower false alarm rate. The outlier detection method of kernel PCA depends generally on kernel type and kernel parameters. In this work, Mahalanobis kernel given by (6) is chosen to resolve the nonlinearity of data distribution. This type of kernel depends on kernel width and number of principal components [7], [28], [30]. We present below the pseudo algorithm of the training phase and the pseudo algorithm of the detection phase.

The pseudo algorithm of the training phase:

```
Input: Cluster $G = \{S_i; i = 1...n\}$
Output: Global model (GM, GPC, DMAX)
    For all members
    Collect data vector and normalize it
    Send the data vector to the cluster head (CH)
    End
    For every (CH)
    Execute Kernel PCA on normalized data and establish
    global model(GM)
    End
End.
```

The pseudo algorithm of the detection phase:

```
Input: Cluster $G = \{S_i; i = 1...n\}$
Output: Class index of the real time measurement
    For every time interval (t)
      For all members
      Collect data vector and normalize it (using global
      minimum and global maximum).
      Send the data vector to the cluster head (CH)
      End
      For every (CH)
      Detect outliers using detection model
      End
    End
  End
```

V. Experimental Results

A. Datasets

To validate the proposed models, some data samples were extracted from three WSN deployments which represent static and dynamic environments. The next subsections introduce the datasets and explain the data labeling procedure. The datasets that are used in this paper are extracted from the following WSN deployments:

- *Intel Berkeley Research Lab (IBRL):* IBRL dataset [24] was collected from the WSN deployed at Intel Berkeley Research Laboratory, University of Berkeley. The network consists of 54 Mica2Dot sensor nodes and was deployed in the period of 30 days from 15/04/2004 until 14/05/2004. Two types of measurements were collected which are: temperature and humidity. The measurements were collected in 31 s intervals. In this research, subsets of this dataset were chosen for evaluating the proposed outliers detection model. The observations of the small cluster which have 5 sensors namely, Node-25, Node-28, Node-29, Node-31 and Node-32 were used to evaluate the proposed model.
- *Grand St. Bernard (GStB):* GStB dataset [25] is one of sensorscope project deployment dataset was gathered using WSN deployment at the Grand St. Bernard pass that is located between Italy and Switzerland. The network is formed of 23 sensors that record metrological environmental data that include temperature and humidity. In our experiments, the ambient temperature observations in the period 6am-14pm of data recorded on 5th March 2007 were used.
- *Sensorscope Lausanne Urban Canopy Experiment (LUCE):* Sensorscope-LUCE dataset [26] was collected by a sensorscope project in the École Polytechnique Fédérale de Lausanne (EFPL) campus between July 2006 and May 2007. The experiments aimed at better understanding of micrometeorology and atmospheric transport at the urban environments. The measurement system was based on a WSN of 110 sensor nodes deployed on the EPFL campus to measure key environment quantities which include; ambient temperature, surface temperature, and relative humidity.

B. Performance Analysis Based KPCA Using Reconstruction Error and Mahalanobis Distance

Mahalanobis kernel is used recently in the field of WSN, specially based outlier detection, was introduced in several works. Kernel PCA performance was showcased in comparison to other established kernel-based methods [27]. To compute the Kernel PCA transform of a set of test patterns, this approach chooses a training set and a suitable projection dimensionality p, and, finally, computes the reconstruction error (RE) for each of these test patterns. Given the projection dimensionality p, outliers are identified as data points, whose RE exceeds an appropriately established threshold value R_{th}. To measure the precision of our method, we calculate the detection rate and the false positive rate for all data points of the test database. In figure 5, we compare the use as

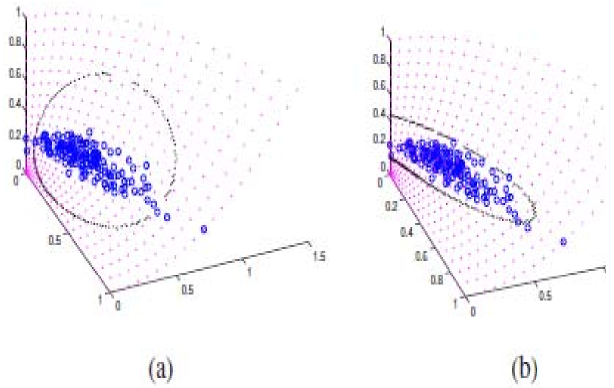

Fig. 5. Constant RE (left figure) and MD (right figure) contours based on the first two principle component. Figure (a) present KPCA using RE which generate a spherical contour and Figure (b) present KPCA using MD which generate an ellipsoidal contour.

outlier discriminant of the Mahalanobis Distance (MD) in the principal subspace and the Reconstruction Error (RE) in the principal subspace's orthogonal complement. We experimentally examine the outlier detection ability of our proposed MD-based method by comparing it with the RE-based method. When comparing the results given on our experimentation by KPCA-MD and KPCA-RE, we see that using mahalanobis distance is more benefit to detect outliers. From Figure 5a, the contour with RE value is considered as the largest RE of the training points. Then, from Figure 5b, the contour with MD value is indicated and it is considered as the largest Mahalanobis distance in the principal subspace of the training points. As we see from the figure, RE produces a decision boundary that is overly broad which many potential outliers would not be detected. However, it does not satisfactorily fit the normal (training) data. The MD induced boundary seems to capture much better the overall structure of the normal data. However, this comparison demonstrates that the RE may not be an effective measure of deviation from normalcy, when compared to using the MD.

It is therefore noted that according to the results of the experiment, the Mahalanobis Distance based KPCA is more beneficial than Reconstruction Error in terms of outliers detection.

C. Comparative Study Between KPCA Using Reconstruction Error, Mahalanobis Distance and OCSVM

This section specifies the performance evaluation of our technique based KPCA using Mahalanobis distance compared to KPCA using reconstruction error and one class SVM. In our experiments, we have used a real data gathered from a deployment of WSN in the Intel Berkeley Research Laboratory, Grand St. Bernard and Sensorscope Lausanne Urban Canopy Experiment. We simulate our protocol in Matlab and consider a closed neighborhood as shown in Figure 2, which is centered at a node with its 6 spatially neighboring nodes. Mahalanobis kernel is used recently in the field of WSN, specially based outlier detection, was introduced in several works. Kernel PCA performance was showcased in comparison to other established kernel-based methods [23]. To compute the Kernel PCA transform of a set of test patterns, this approach chooses a training set and a suitable projection dimensionality p, and, finally, computes the Mahalanobis distance (MD) for each of these test patterns. Given the projection dimensionality p, outliers are identified as data points, whose MD exceeds an appropriately established threshold value $MaxDist$. Our method has been tested on real data as seen in Table I.

TABLE I
KPCA-MD, KPCA-RE AND OCSVM ON THE REAL WORLD DATASETS

	MD	RE	OCSVM
Intel Berkeley (IBRL)	**0.9764** 0.9635 **0.9727** **0.9651** **0.9760**	0.9170 **0.9264** 0.8533 0.8997 0.9685	0.8937 0.9043 0.6152 0.6242 0.6396
Grand-St-Bernard (GStB)	**0.9282** **0.8522** 0.9075 **0.8686** 0.8959	0.3077 0.7167 0.2055 0.2224 0.2658	0.7528 0.2334 0.4432 0.8593 **0.9409**
Sensor-scope (LUCE)	**0.8437** 0.8206 0.8155 **0.8336** 0.8027	0.7466 **0.8541** **0.8652** 0.7242 0.7896	0.7641 0.8460 0.7533 0.7997 **0.8473**

TABLE II
DETECTION RATE AND FALSE ALARM-BASED KPCA USING MAHALANOBIS DISTANCE ON IBRL REAL DATASET

Nodes						
	N25	N28	N29	N31	N32	Average
DR (%)	100	94	97	89	100	97
FPR (%)	14	0	8	1	0	6.3

When comparing the results given on our experimentation by KPCA-MD, KPCA-RE and one class SVM, we see that using mahalanobis distance is more beneficial to detect outliers. Then, this comparison reveals that twice the RE and OCSVM may not be an effective measure of deviation from normalcy, when compared to using the MD. From previous experiments, we see that RE produces a decision boundary that is overly broad. Thus, it does not satisfactorily fit the normal data because many potential outliers would not be detected. So, our proposed MD based method has an important advantage compared to the RE-based method and OCSVM that detects perfectly the outliers as observed in our experiments and as mentioned by the table. Then, it is clear that KPCA using the Mahalanobis distance (MD) is more sensitive to the detection of FPR and DR (as shown in table 2) than KPCA using reconstruction error (RE) and OCSVM. However, the MD induced boundary seems to capture much better the overall structure of the normal data.

Based on the following figures (Fig 6, Fig 7 and Fig 8), we presents a comparison between MD, RE and OCSVM.

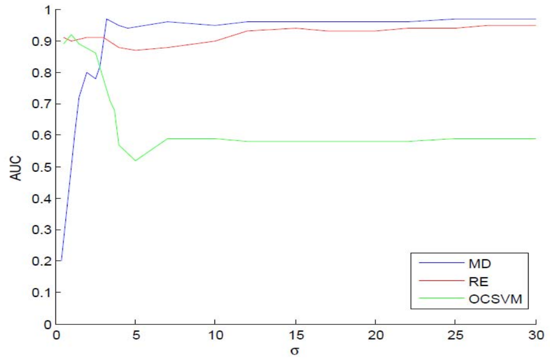

Fig. 6. IBRL dataset: Maximum AUC value versus kernel parameter value.

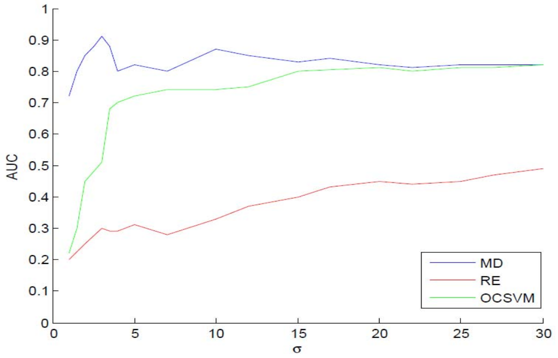

Fig. 7. GStB dataset: Maximum AUC value versus kernel parameter value.

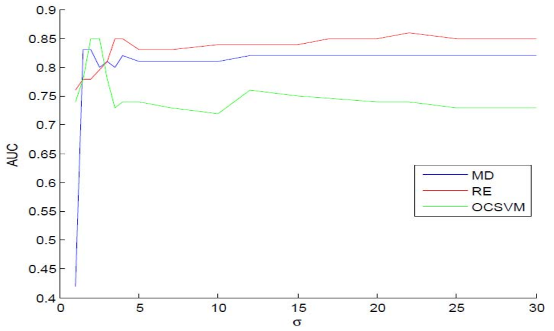

Fig. 8. LUCE dataset: Maximum AUC value versus kernel parameter value.

TABLE III

AVERAGE EFFECTIVENESS OF KPCA-MK MODEL FOR SIMULATED-BASED DATASETS

	E1	E2	E3
Accuracy (%)	98.7	96.5	95.4
DR (%)	98.2	97.1	96.2
FNR (%)	1.8	2.9	3.8

TABLE IV

EXPERIMENTAL RESULTS OF KPCA-MD ON E1 DATASET

S	Training set size	Testing set size	DR (%)	Accuracy (%)	FPR (%)	FNR (%)
1	1000	2000	100	99	1.94	0
2	1000	2000	100	98.2	1.8	0
3	1000	2000	100	89.7	10.3	0

TABLE V

EXPERIMENTAL RESULTS OF KPCA-MD ON E2 DATASET

S	Training set size	Testing set size	DR (%)	Accuracy (%)	FPR (%)	FNR (%)
1	1300	19000	100	98.1	1.8	0
2	3000	22700	100	97.4	3.6	0

The ROC curve shows that KPCA based Mahalanobis Kernel using Mahalanobis Distance are much better than that of KPCA-MK using Reconstruction Error and OCSVM. It is therefore noted that according to the results of the experiment, the Mahalanobis Distance based KPCA is more beneficial than Reconstruction Error and OCSVM in terms of outliers detection varied by sigma in our experiments on Matlab or on real dataset in Wireless Sensors Networks.

For simulated-based datasets, the KPCA-MK model is tested and compared with some existing outlier detection models. The results are analyzed in terms of detection effectiveness which is represented by detection rate (DR), detection accuracy, and false negative rate (FNR) as described in Table III. To present the performance of our model to detect outlier, three dataset samples E1, E2, and E3 are used were extracted from nodes N25, N29, and N31 in the IBRL, LUCE and GStB deployment. E1 dataset contains temperature readings extracted from IBRL dataset. Three scenarios were investigated for this dataset based on the size of training set size used to obtain the normal reference model. The results on this dataset using the three scenarios are provided in Table IV. In each scenario, the size of training and testing sets were varied. E2 dataset is ambient temperature readings extracted from the LUCE deployment. Two scenarios were examined for this dataset which varied in training and testing set sizes as shown in Table V. In E3 dataset which exhibit constant or long period anomalies, extracted from GStB deployment to form E3 dataset. Two scenarios were considered and shown in Table VI. Two variables that report temperature and humidity were chosen from each node. 120 artificial anomalies were randomly generated using normal distribution and injected in each dataset sequentially. Since anomalies were generated using normal distribution random function, the experiments were repeated 10 times with different testing set and the same training set to show the stability of the performance. The overall performance in terms of effectiveness measures is the average over the 10 runs. In the experiments, the size of training set was 210 instances while the size of each testing set was 110 instances that contain 40 normal instance and 70 artificially generated anomalous instances. The average of effectiveness measures over 10 runs for all datasets is shown in Table III.

TABLE VI
EXPERIMENTAL RESULTS OF KPCA-MD ON E3 DATASET

S	Training set size	Testing set size	DR (%)	Accuracy (%)	FPR (%)	FNR (%)
1	2000	3500	100	96.7	3.3	0
2	2000	3500	100	97.8	2.2	0

TABLE VII
COMPARISON RESULTS BETWEEN KPCA-BASED MAHALANOBIS DISTANCE, KPCA-RE AND OCSVM

Data-set	Adopted method	DR (%)	Accuracy (%)	FPR (%)	FNR (%)
E1	KPCA-MD	100	99	1.3	0
	KPCA-RE	99	96.5	3.5	0
	OCSVM	97	95.4	4.6	0
E2	KPCA-MD	100	97.4	1.60	0
	KPCA-RE	97	95.1	3.9	0
	OCSVM	99	96	4	0
E3	KPCA-MD	99	95.8	4.2	0
	KPCA-RE	96	93.7	7.25	0
	OCSVM	95	90.4	10.6	0

As can be seen in Table III, an average of 98.7%, 97.6%, and 95.4% detection accuracy was achieved for dataset E1, E2 and E3, respectively. However, the proposed model achieves higher detection rate for all datasets. The highest detection rate was reported for dataset E1 while the lowest was reported for E3. As a result, the FNR was affected by the detection rate such that it decreases with the increase of the detection rate and vice versa.

The results in Table VII indicate that false positives are of main concern for these kinds of datasets. The amount of false positives increases with the increase of data dynamicity. The FPR increases when the training set is not a representative of the data behaviour.

Finally, from the different table and different figure of real dataset, we can see that KPCA-based outlier detection using Mahalanobis Distances outperforms the other two methods KPCA-RE and OCSVM. Our methodology can be applied on both small and large datasets. Our Approach is scalable and very efficient in the WSN application because when the dataset used are large, this give a better accuracy, increase the percentage of detection rate and decrease the fals alarm rate. Our method is tested on control of fire in a wheat field, so, it gives us a good detection rate with a minimum false alarm rate compared to Great-Duck-Island: [31] or Volcano Monitoring: [32] and Sensorscope [33].

D. Complexity Analysis

To investigate the computational complexity of our proposed model using KPCA based mahalanobis kernel, only the testing phase which is conducted online should be considered. In this phase, the calculation of the reconstruction error measure based on mahalanobis distance for each new observation is needed by dividing its projection on the PC space by the corresponding eigenvectors which are already calculated in the training phase. The upper bound computational complexity involved in this process is O (M), where M is the number of observed variables. The training phase which involves the calculation of the PCs has a time complexity of O (NM^2) where N and M are the size and the dimension of the training set respectively. For the training phase of the OCSVM, the relative complexity is O (N^3) as it needs to solve an optimization problem to compute the hyper plan that separates normal and outlier data. The complexity of online testing phase in the theme of our model structure is O (M) where M is the number of observed variables which is equivalent to the complexity incurred by our proposed model. The retraining of the OCSVM will cause a high power consumption which makes it unsuitable for anomaly detection in these types of environments compared to our model using KPCA based mahalanobis kernel.

VI. CONCLUSION

In our work, we presents a comparative study based KPCA between using Mahalanobis Distance (MD), Reconstruction Error (RE) and One Class Support Vector Machine (OCSVM) for outlier detection. A principal subspace in an infinite-dimensional feature space described the distribution of training data. The Mahalanobis distance of a new data point with respect to this subspace was used as a measure to decide if this new point is considered as a normal point or outliers. The use of the KPCA using Mahalanobis Distance demonstrated a higher classification performance on a real database used compared with KPCA using Reconstruction Error (RE) and OCSVM. So, our method demonstrated to be more robust against outlier detection within the training set. In order to showcase the merits of our proposed approach, we performed a number of experiments that compared the capability of detecting outliers in data of the One-Class SVM, the RE-based, and the MD-based KPCA detection methods. Nevertheless, the outcomes indicate that the MD-based KPCA method is competitive in detecting true outliers, when compared to the other two approaches. As a future work, we focus on improving the performances of the proposed model and extending it to be able to detect events that may occur instead of only considering outliers in an adaptive method.

REFERENCES

[1] M. L. Braun, J. M. Buhmann, and K.-R. Muller, "On relevant dimensions in kernel feature spaces," *J. Mach. Learn. Res.*, vol. 9, pp. 1875–1908, Jun. 2008.

[2] T. Naumowicz *et al.*, "Autonomous monitoring of vulnerable habitats using a wireless sensor network," in *Proc. Workshop Real-World Wireless Sensor Netw. (REALWSN)*, Glasgow, Scotland, 2008, pp. 51–55.

[3] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," *Comput. Netw.*, vol. 51, no. 4, pp. 921–960, 2007.

[4] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 12, pp. 159–170, 2010.

[5] J.-M. Lee, C. Yoo, S. W. Choi, P. A. Vanrolleghem, and I.-B. Lee, "Nonlinear process monitoring using kernel principal component analysis," *Chem. Eng. Sci.*, vol. 59, no. 1, pp. 223–234, Jan. 2004.

[6] S. W. Choi, C. Lee, J.-M. Lee, J. H. Park, and I.-B. Lee, "Fault detection and identification of nonlinear processes based on kernel PCA," *Chemometrics Intell. Lab. Syst.*, vol. 75, no. 1, pp. 55–67, Jan. 2005.

[7] H. N. Minh and T. Fernando, "Robust kernel principal component analysis," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., 2008.

[8] I. K. Fodor, "A survey of dimension reduction techniques," Ph.D. dissertation, U.S. Dept. Energy, Lawrence Livermore Nat. Lab., Livermore, CA, USA, 2002.

[9] M. A. Rassam, A. Zainal, and M. A. Maarof, "An adaptive and efficient dimension reduction model for multivariate wireless sensor networks applications," *Appl. Soft Comput.*, vol. 13, no. 4, pp. 1978–1996, 2013.

[10] M. A. Livani and M. Abadi, "Distributed PCA-based anomaly detection in wireless sensor networks," in *Proc. Int. Conf. Internet Technol. Secured Trans. (ICITST)*, Nov. 2010, pp. 1–8.

[11] N. Chitradevi, V. Palanisamy, K. Baskaran, and U. B. Nisha, "Outlier aware data aggregation in distributed wireless sensor network using robust principal component analysis," in *Proc. Int. Conf. Comput. Commun. Netw. Technol.*, Jul. 2010, pp. 1–9.

[12] Y.-X. Xie, X.-G. Chen, and J. Zhao, "Data fault detection for wireless sensor networks using multi-scale PCA method," in *Proc. 2nd Int. Conf. Artif. Intell., Manage. Sci. Electron. Commerce (AIMSEC)*, Aug. 2011, pp. 7035–7038.

[13] B. Scholkopf, A. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[14] K. Kapitanova, S. H. Son, and K.-D. Kang, "Event detection in wireless sensor networks—Can fuzzy values be accurate?" in *Proc. 2nd Int. Conf. ADHOCNETS*, Victoria, BC, Canada, Aug. 2010, pp. 1–16.

[15] Y. Zhang, N. Meratnia, and P. J. M. Havinga, "Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine," *Ad Hoc Netw.*, vol. 11, no. 3, pp. 1062–1074, Dec. 2013.

[16] Y. Zhang, N. A. S. Hamm, N. Meratnia, A. Steinb, M. van de Voort, and P. J. M. Havinga, "Statistics-based outlier detection for wireless sensor networks," vol. 26, no. 8, pp. 1373–1392, Apr. 2012.

[17] X. Liu, U. Kruger, T. Littler, L. Xie, and S. Wang, "Moving window kernel PCA for adaptive monitoring of nonlinear processes," *Chemometrics Intell. Lab. Syst.*, vol. 96, no. 2, pp. 132–143, 2009.

[18] V. Franc and V. Hlavac, "Greedy algorithm for a training set reduction in the kernel methods," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2003, pp. 426–433.

[19] J.-H. Cho, J.-M. Lee, S. W. Choi, D. Lee, and I.-B. Lee, "Fault identification for process monitoring using kernel principal component analysis," *Chem. Eng. Sci.*, vol. 60, no. 1, pp. 279–288, 2005.

[20] Q. W. Kilian, S. Fei, and K. S. Lawrence, "Learning a kernel matrix for nonlinear dimensionality reduction," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, pp. 839–846.

[21] S. Bernhard, S. Alexander, and M. Klaus-Robert, "Kernel principal component analysis," in *Advances in Kernel Methods—Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999, pp. 327–352.

[22] C. Chakour, M. F. Harkat, and M. Djeghaba, "Adaptive kernel principal component analysis for nonlinear dynamic process monitoring," in *Proc. 9th Asian Control Conf. (ASCC)*, Istanbul, Turkey, 2013.

[23] D. Mingtao, T. Zheng, and X. Haixia, "Adaptive kernel principal component analysis," *Signal Process.*, vol. 90, no. 5, pp. 1542–1553, 2010.

[24] Intel Berkeley Research Lab. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html, accessed Apr. 5, 2004.

[25] (2007). *GStB, Grand-St-Bernard Dataset*. [Online]. Available: http://lcav.epfl.ch/cms/lang/en/pid/86035

[26] SensorScope. [Online]. Available: http://sensorscope.epfl.ch/index.php/MainPage

[27] H. Hoffmann, "Kernel PCA for novelty detection," *Pattern Recognit.*, vol. 40, no. 3, pp. 863–874, 2007.

[28] M. H. Nguyen and F. Torre, "Robust kernel principal component analysis," Ph.D. dissertation, School Comput. Sci., Robot. Inst., Pittsburgh, PA, USA, 2008.

[29] M. A. Rassam, A. Zainal, and M. A. Maarof, "One-class principal component classifier for anomaly detection in wireless sensor network," in *Proc. 4th Int. Conf. Comput. Aspects Soc. Netw. (CASoN)*, Nov. 2012, pp. 271–276.

[30] W. Zheng, C. Zou, and L. Zhao, "An improved algorithm for kernel principal component analysis," *Neural Process. Lett.*, vol. 22, no. 1, pp. 49–56, 2005.

[31] R. Szewezyk, A. Mainwaring, J. Polastre, and D. Culler, "An analysis of a large scale habitat monitoring application," in *Proc. 2nd ACM Conf. Embedded Netw. Sensors Syst. (SenSys)*, Baltimore, MD, USA, Nov. 2004, pp. 214–226.

[32] G. Werner-Allen *et al.*, "Deploying a wireless sensor network on an active volcano," *IEEE Internet Comput.*, vol. 10, no. 2, pp. 18–25, Mar./Apr. 2006.

[33] G. Barrenetxea, F. Ingelrest, G. Schaefer, M. Vetterli, O. Couach, and M. Parlange, "SensorScope: Out-of-the-box environmental monitoring," in *Proc. 7th Int. Conf. Inf. Process. Sensor Netw.*, Apr. 2008, pp. 332–343.

Oussama Ghorbel is currently pursuing the Ph.D. at the Ecole Nationale d'Ingénieurs de Sfax (ENIS), Sfax, Tunisia. His research activity is conducted within CES research unit. He received the M.S. degree from ENIS, in 2010. He is also an invited Ph.D. Student with the Laboratory of Systems Modeling and Dependability, University of Technology of Troyes, Troyes, France. His current research interests are in the field of wireless sensor networks and image compression. He served on national and international conference organization: ICM, TWESD, and SensorNets-09.

Walid Ayedi received the M.S. degree from the Ecole Nationale d'Ingénieurs de Sfax, Sfax, Tunisia, in 2008, where he is currently pursuing the Ph.D. degree at the Computer and Embedded Systems Laboratory. He is also an invited Ph.D. Student with the Laboratory of Systems Modeling and Dependability, University of Technology of Troyes, Troyes, France. His research interests include image analysis, computer vision, and machine learning. He has served on national and international conference organizations, such as IDT, ICM, TWESD, and Sensor Nets.

Hichem Snoussi was born in Bizerte, Tunisia, in 1976. He received the Diploma degree in electrical engineering from Supélec, Gif-sur-Yvette, France, in 2000, and the D.E.A. and Ph.D. degrees in signal processing from the University of Paris-Sud, Orsay, France, in 2000 and 2003, respectively. Since 2010, he has been a Full Professor with the University of Technology of Troyes, Troyes, France.

Mohamed Abid was the Head of the Computer Embedded System Laboratory with the Ecole Nationale d'Ingénieurs de Sfax, University of Sfax, Sfax, Tunisia, where he is currently a Professor. He received the Ph.D. degree from the National Institute of Applied Sciences, Toulouse, France, in 1989, and the Doctorate degree in computer engineering and microelectronics from the National School of Engineering of Tunis, Tunis, Tunisia, in 2000. His current research interests include hardware–software co-design, system-on-a-chip, reconfigurable system, and embedded system.