2014 28th International Conference on Advanced Information Networking and Applications Workshops

Kernel Principal Subspace based Outlier Detection Method in Wireless Sensor Networks

Oussama Ghorbel, Mohamed Abid
CES Research Unit
National Engineers School of Sfax
Sfax University, Tunisia
oussama.ghorbel@ceslab.org; mohamed.abid@enis.rnu.tn

Hichem Snoussi
LM2S Research Unit
University of Technology of Troyes
Troyes, France
hichem.snoussi@utt.fr

Abstract— **An emerging class of Wireless Sensor Networks (WSNs) applications involves the acquisition of large amounts of sensory data from battery-powered, low computation and low memory wireless sensor nodes. The accuracy of sensor readings is without a doubt one of the most important measures to evaluate the quality of a sensor and its network. For this case, the task amounts to creating a useful model based on KPCA to recognize data as normal or outliers. Over the last few years, Kernel Principal Component Analysis (KPCA) has found several applications in outlier detection. Within this setting, we propose a new outlier detection method based on Kernel Principal Component Analysis (KPCA) 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 wireless sensor networks.**

Keywords— ***Wireless Sensor Networks, Outlier Detection, Kernel methods, Mahalanobis kernel, Kernel Principal Component Analysis (KPCA), Mahalanobis Distance (MD), Reconstruction Error (RE).***

I. INTRODUCTION

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 [1]. 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 [2]. KPCA has found a new field which is integrated in application of novelty 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 [3, 4]. KPCA has been used in several applications, such as voice recognition, image segmentation, face detection, feature extraction [14], data denoisin and etc. 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 [5]. 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 [6].

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 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.

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,

978-1-4799-2652-7/14 $31.00 © 2014 IEEE
DOI 10.1109/WAINA.2014.120

showcases the obtained experimental results, and section 6 concludes and summarizes the main outcomes of the paper.

II. Related Work

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 [7].

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 eigen-values 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 [8], [9]. The detail introduction of the basic KPCA can be viewed in [8] and [10]. 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). The feature space is constructed such that a nonlinear operation can be applied in the input space by applying a linear operation in the feature space. Consequently, standard PCA (a linear operation) can be applied in feature space to perform nonlinear PCA in the input space.

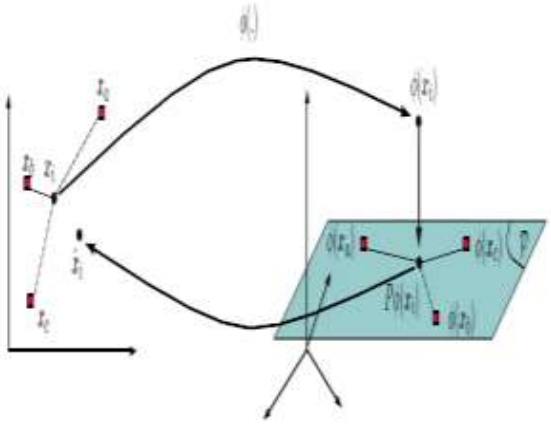

Fig. 1. Kernel-PCA representation techniques in wireless sensor networks.

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 as shown in figure (Fig 2). 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 [11].

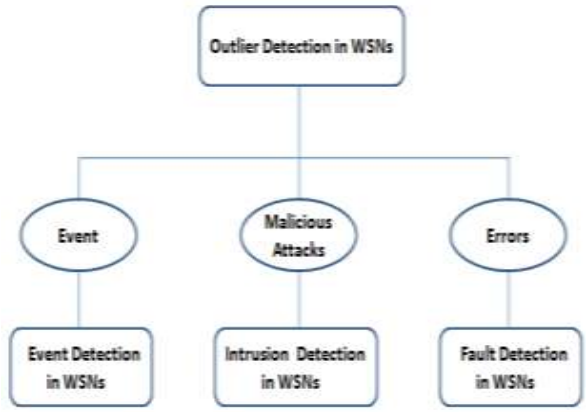

Fig. 2. Different types of outlier source in Wireless Sensors Networks

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. 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 [12]. 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:

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

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. For

more detail on clustering and partitioning methods in WSN see [10].

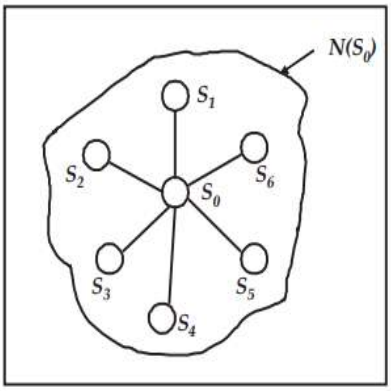

Fig. 3. Cluster example: S_0 is the cluster head.

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 where $X = (X_1,...,X_m)$ 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: $x_k^i = (x_{k1}^i, x_{k2,}^i,..., x_{kj}^i)$. During each time window t, s_i captures a set of data measurements $X_k^i = \{x_k^i(t), k = 1...n_i\}$.

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. In our centralized model, the learning phase is the first step of every learning task which is performed in the cluster head's 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. Then the data matrix is normalized and global mean and global covariance matrix are calculated [13]. To establish the detection model, the cluster head executes Kernel PCA on this data matrix: first by calculating global principal components. 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 this three parameters: the global mean, the global principal components and the maximal distance $d_{\max}$. 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. Finally, for the detection phase, 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 $d_p \le d_{\max}$, the observation are considered as normal, otherwise outlier.

B. 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. 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 $\{x_1,...,x_N\}$ be a dataset composed of N data points of dimension m, we define the data center c and the covariance matrix Q:

$$c = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1)$$

$$Q = \frac{1}{N}\sum_{i=1}^{N}(x_i - c)(x_i - c)^T \qquad (2)$$

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

$$d(x) = \sqrt{(x-c)^T Q^{-1}(x-c)} \qquad (3)$$

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

$$A(x, x') = \exp(-(x-x')^T H(x-x')) \qquad (4)$$

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:

$$A(x,x') = \exp(-\frac{\delta}{m}(x-x')^T Q^{-1}(x-x')) \quad (5)$$

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 (5) is chosen to resolve the nonlinearity of data distribution. This type of kernel depends on kernel width σ and number of principal components q [15].

V. Experimental Results

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 [16]. 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 $d_{\max}$. Our method has been tested on real data (See Table1) and on synthetic distribution. The Intel Berkeley data is used as real sensor data to validate the proposed method. The real data are collected from a closed neigh-borhood from a WSN deployed in Intel Berkeley. The Kernel PCA was tested on the Berkley's sensor dataset. We choose to work on the cluster composed of six nodes including Mote 45 as a cluster head and motes 43,44,46,47,48 as members as shown in Figure 4.

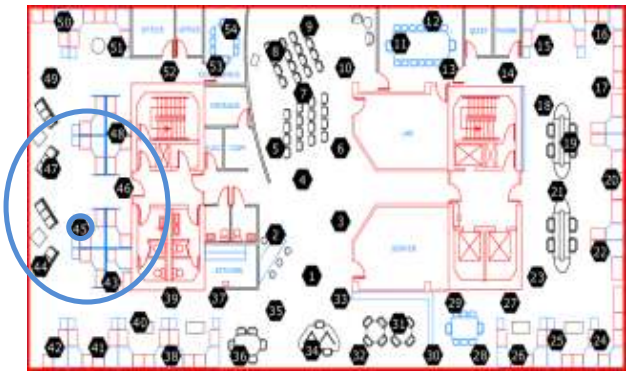

Fig. 4. WSN of Berkley Lab

The network recorded ambient temperature, relative humidity, soil moisture, solar radiation and watermark measurements. In our experiments, we use a period of data recorded on 15 days from 4 to 18 September 2007 with two attributes: ambient temperature and relative humidity for each sensor measurement. 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. The tables below present the difference between Mahalanobis Distance based KPCA and Reconstruction Error based KPCA.

TABLE I. Experimental Results for MD-KPCA, RE-KPCA on the Real World Datasets.

	MD	RE
Intel Berkeley	96.53%	92.41%
Grand-St-Bernard	95.12%	93.76%

TABLE II. Experimental Results for MD-KPCA, RE-KPCA on Synthetic Datasets.

	MD	RE
Sine	90.26%	86.77%
Square	91.37%	89.52%
Ring-Line Square	92.51%	88.29%

When comparing the results given on our experimentation by KPCA-MD and KPCA-RE, we see that using mahalanobis distance is more beneficial to detect outliers. Then, this comparison reveals that the RE 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 that detects perfectly the outliers as observed in our experiments and as mentioned by the two tables. However, the MD induced boundary seems to capture much better the overall structure of the normal data.

The ROC curve (Fig 5) shows that the Detection Rate (DR) of KPCA-MK using Mahalanobis Distance is much better than that of KPCA-MK using Reconstruction Error. 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.

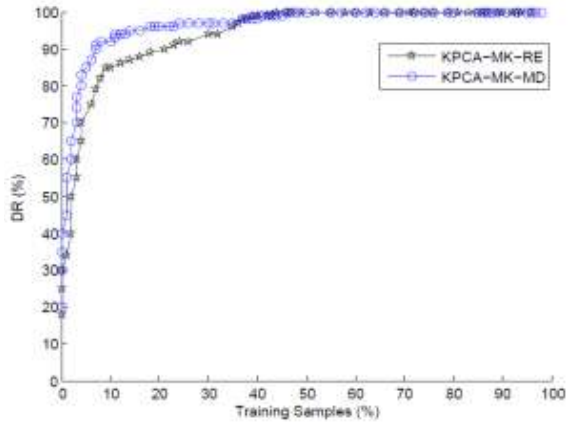

Fig. 5. Comparative ROC curves based KPCA using Mahalanobis Distance and Reconstruction Error.

The following figure presents a comparison between the two kernels. It is clear that KPCA using the Mahalanobis distance is more sensitive to the detection of FPR (Fig 6) and DR (Fig 7) than KPCA using reconstruction error varied by sigma in our experiments. To show the robustness of our work using Mahalanobis kernel, we are referred to the work of Heiko Hoffmann [17] entitled "Kernel PCA for Novelty Detection". After the following figures we see that Mahalanobis kernel is more efficient either by simulation on MATHLAB or in Wireless Sensors Networks using Mahalanobis distance.

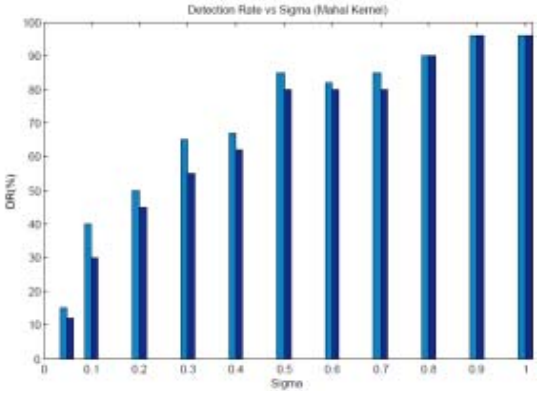

Fig. 6. Detection Rate based KPCA using Mahalanobis Distance and Reconstruction Error for real data.

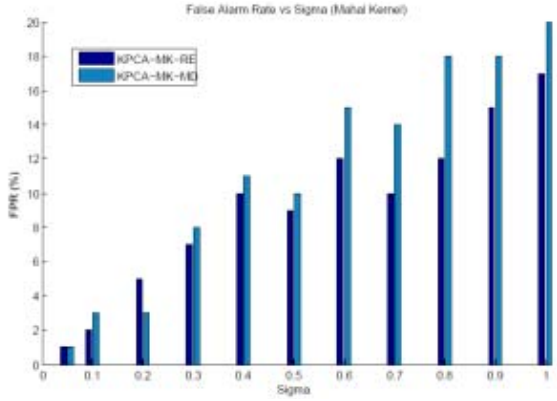

Fig. 7. False Positive Rate based KPCA using Mahalanobis Distance and Reconstruction Error for real data.

VI. CONCLUSION

This article is a comparative study based KPCA between using Mahalanobis Distance (MD) and Reconstruction Error (RE) 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 synthetic and real database used compared with KPCA using Reconstruction Error. So, compared to KPCA-RE, our method demonstrated to be more robust against outlier detection within the training set. 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.

REFERENCES

[1] M.L. Braun, J M. Buhmann, and K.R Muller. On relevant dimensions in kernel feature spaces. Journal of Machine Learning Research, pp 1875–1908, 2008.

[2] T.Naumowicz, R.Freeman, A. Heil, M. Calsyn, E. Hellmich, A. Brandle, T. Guilford et J. Schiller. "Autonomous monitoring of vulnerable habitats using a wireless sensor network". In : Proceedings of the Workshop on Real-World Wireless Sensor Networks, REALWSN'08. Glasgow, Scotland, 2008.

[3] Ian F. Akyildiz, T. Melodia, Kaushik R. Chowdhury. "A survey on wireless multimedia sensor networks". Journal Computer Networks: The International Journal of Computer and Telecommunications Networking, Volume 51, Issue 4, Inc . New York, NY, USA, United State, 2007.

[4] Y. Zhang, N. Meratnia, P. Havinga, "Outlier detection Techniques for wireless sensor networks: A survey", pp .11-20, 2008.

[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, January 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, January 2005.

[7] 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, July 1998.

[8] K. Kapitanova, S.H. Son, and K. D. Kang. Event Detection in Wireless Sensor Networks. Second International Conference, ADHOCNETS 2010, Victoria, BC, Canada, August 2010.

[9] Y. Zhang, N. Meratnia, P. J.M. Havinga. Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine. Ad Hoc Networks , December 2012.

[10] Y. Zhang, N.A.S. Hammb, N. Meratniaa, A. Steinb, M. van de Voorta & P.J.M. Havingaa. Statistics-based outlier detection for wireless sensor networks, Volume 26, Issue 8, 2012.

[11] Bernhard Schölkopf , Alexander Smola , Klaus-Robert Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation, v.10 n.5, pp:1299-1319, 1998.

[12] Chakour et al. 2012. Adaptive kernel principal component analysis for nonlinear time-varying processes monitoring ICEECA2012.

[13] WenmingZheng, CairongZou, and Li Zhao. An Improved Algorithm for Kernel Principal Component Analysis. Neural Process, pp: 49-56, 2005.

[14] Mingtao Ding, ZhengTian, and HaixiaXu. Adaptive kernel principal component analysis. Signal Process, pp 1542-1553, 2010.

[15] Minh Hoai Nguyen, Fernando de la Torre. Robust Kernel Principal Component Analysis, 2008.

[16] http://sensorscope.epfl.ch/index.php/MainP age.

[17] Heiko Hoffmann. Kernel PCA for novelty detection. Pattern Recognition, pp 863–874, 2007.