

Classification Data using Outlier Detection Method in Wireless Sensor Networks

Oussama Ghorbel, Aya Ayadi, Kais Loukil

CES Research Unit

National Engineers School of Sfax, Sfax University, Tunisia
National Engineers School of Gabes, Gabes University, Tunisia
Email: oussama.ghorbel@ceslab.org; ing.aya.ayadi@gmail.com;
kais_loukil@yahoo.fr

Mohamed S. Bensaleh and Mohamed Abid

Digital Research Center (CRNS), Technopark Sfax, Tunisia
King Abdulaziz City for Science and Technology, KSA
Email: mbensaleh@kacst.edu.sa; mohamed.abid@enis.rnu.tn

Abstract—For classification data, we use Wireless sensor networks (WSNs) as hardware for collecting data from harsh environments and controlling important events in phenomena. To evaluate the quality of a sensor and its network, we use the accuracy of sensor readings as surely one of the most important measures. Therefore, for anomalous measurement, real time detection is required to guarantee the quality of data collected by these networks.

In this case, the task amounts to create a useful model based on KPCA to recognize data as normal or outliers. On account of the attractive capability, KPCA-based methods have been extensively investigated, and have shown excellent performance. So, to extract relevant feature for classification and to prevent from the events, we use KPCA based on Mahalanobis kernel as a preprocessing step. In the original space, the totality of computation is done thus saving computing time. Then the classification was done on real Intel Berkeley data collecting from urban area. Compared to a standard KPCA, the results show that our method are specially designed to be used in the field of wireless sensor networks (WSNs).

Keywords—component; Classification data, Wireless Sensor Networks (WSNs); Mahalanobis kernel; Kernel Principal Component Analysis (KPCA); Outlier Detection

I. INTRODUCTION

Nowadays, wireless sensor networks are considered as an important source of data for different applications. These interconnected sensors in wireless sensor networks provide data continuously which is uncertain and unreliable [1]. Hence, an effective processing task of data streams are considered as very important for several applications like outlier detection. Wireless sensor networks can provide continuous measurements of physical phenomena by means of dense deployments of sensor nodes, compared with the conventional data collection techniques. They are widely used and they have gained attention in various fields including health care, precision agriculture, traffic control, etc [2, 3]. For KPCA, it has been used in multiple applications, for example: feature extraction, face detection, voice recognition, image segmentation, data denoising and etc. To provide the user with reliable information, most WSN's applications require precise and accurate data. Although the importance of information quality is derived from WSNs. The collected sensor data may be of low quality and reliability because of the low cost nature and harsh

deployments of WSNs [4]. Outlier detection methods allow the cleaning and the refinement of collected data. These latter, providing the most useful information to end users, while maintaining low energy consumption and preserving high computational efforts due to the limited energy resources of sensor nodes. Also, outlier detection method ensure the quality of sensor measurements.

The detection model is built upon historical data structure of WSN to detect outliers. This latter, should be able to detect outliers among new observations with good precision [5]. Principal Component Analysis, considered as an alternative way of computing the principal axes through the use of inner product evaluations, has been extended to a kernel-based PCA. In fact, in many applications, the use of non-linear dimensionality reduction was expanded. For pattern recognition field, recent research has shown that kernel principal component analysis (KPCA) can be expected to work well as a pre-processing device. Also, KPCA is used in the field of wireless sensor networks (WSNs) 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 novelty detection.

One class outlier detection methods are used to resolve various problems in wireless sensor networks domains. This paper untitled classification data using outlier detection method, is a comparative study based on one class outlier detection method in wireless sensor networks. So, our work deals mainly with the uses of Mahalanobis kernel based KPCA for outlier detection method in wireless sensor networks. We use Mahalanobis distance induced feature subspace spanned by principal components as obtained by Kernel PCA to identify outliers. The observation is classified as an outlier, if the distance of a new data point is above a prefixed threshold which is also established experimentally. We assume that the principal subspace represents the normal data. The model is tested on real data from Grand Saint Bernard, Luce and Intel Berkeley. We show that the obtained results are important and the proposed method can provide high detection rate with the lowest false alarm rate.

This article is detailed as follows: First, the related work for KPCA is described in Section 2. Second, the outliers detection and its different categories in wireless sensor networks are presented in Section 3. Third, the adopted method is described in Section 4. Then, Section 5 showcases

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

II. RELATED WORK

The kernel based PCA is considered as a non linear principle components analysis (PCA) created using the kernel trick. Kernel PCA maps the original inputs into a high dimensional feature space using a kernel method [6]. Based on Mathematical theory, the current features are transformed into a high-dimensional space. Then, eigenvectors are calculated in this space. The vectors with really low Eigen-values are ignored and then the learning are doing in this transformed space. Compared to PCA, kernel principle components analysis (KPCA) is computationally intensive. Also, it need much more time of the training data points task, which refers that kernel KPCA process is much higher than PCA. However, much larger number of principle components that need to be estimate. In processing nonlinear systems, KPCA method gives a higher performance compared to linear PCA method [7], [8]. The basic detail of the KPCA are described in [7], and [9]. As presented by Scholkopf et al., Kernel PCA (KPCA), is a nonlinear dimension reduction technique of data with an underlying nonlinear spatial structure. Behind KPCA a strong point that allow it to transform the input data into a larger functionality space. The latter, is constructed so that a nonlinear operation can be applied by applying a linear operation in the feature space. Consequently, the nonlinear feature extraction method based on KPCA can perfectly extract the main components of the information. Another benefit of KPCA, that can eliminate data correlation, which is effecttively used in wireless sensor networks field, to reduce the dimensions of the sample space.

In [21], the authors used kernel PCA with Gaussian kernel for fault detection and identification of process monitoring in the field of chemical engineering. In [23] and [24], the authors present a comparative study of One Class outlier detection method in wireless sensor networks. So, they use Mahalanobis kernel based KPCA for outlier detection method in wireless sensor networks. To identify outliers, they 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. The proposed method achieve a high detection rate with the lowest false alarm rate.

In [26] and [27], 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 [20]. The training set is then projected onto the span of the lesser subset, where PCA is carried out. Other sampling-based methods exist [25]. Current KPCA reconstruction methods equally weigh all the features; it is impossible to weigh the importance of some features over the others.

III. OUTLIER DETECTION IN WIRELESS SENSOR NETWORKS

Sensor data are sensitive to multiples sources of errors such as changing environmental conditions which may generate noise coming from other sources. These noises can severely affect data transmitted to central base. These abnormal data are untitled outliers. It is used to find errors, noise, missing values, inconsistent data, or duplicate data. The quality of data may affect by this abnormal value. This latter may also reduce the system performance. We present differents sources of outliers founded in wireless sensor networks as follow:

- Errors: these inaccurate observations reduce the quality of data analysis. They may lead to erroneous results. So, they need to be identified, immediately discarded and be corrected or removed.
- Events: is a member of outliers where it is useful to identify an abnormal situation in order to perform an action/decision-making when an event is detected.
- Malicious attacks: classified into two major categories. The first category is an attack obtains data exchanged in the network without interrupting the communication and the second category implies the disruption of the normal functionality of the network.

Outlier detection techniques are very important in various real life applications such as environmental monitoring, medical monitoring, target tracking, industrial monitoring and surveillance monitors [11]. This latter are categorised into two groups cited as follow:

- Predefined patterns
- Automatic methods.

Outlier detection techniques presents the node limitations and describe detection performances. Predefined patterns are characterized by a threshold which depend on a user while automatic methods using environmental data are based on learning techniques presented in the figure 1:

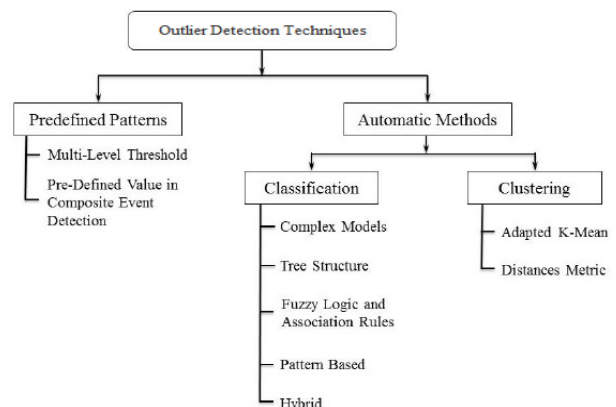


Fig. 1. Outlier Detection Techniques

Sensors is characterized by its low cost and low energy. To improve the quality and performance in wireless sensor networks, the better solution is to use outlier detection techniques. To satisfy the mining accuracy requirements while maintaining the resource consumptions to a minimum, an evaluation of an outlier detection techniques for wireless sensor networks (WSNs) are required. This outlier detection techniques are used to maintain a high detection rate while keeping the false alarm rate low (number of normal data that are incorrectly considered as outliers) [12]. to represent the trade-off between the detection rate and false alarm rate, we use a receiver operating characteristic (ROC). Many problems in detection of outliers data in wireless sensor networks can be cited as follows:

- Specific Application of outlier detection
- Communication cost are Higher
- Effectively modeling of normal objects and outliers
- Classify outlier source
- Network topology are dynamic
- Frequent Communication failures
- Network topology are dynamic
- Centralized and Distributed data

IV. ADOPTED METHOD

Outlier detection is considered as an important technique for data analyses in wireless sensor networks. For classification data, we use this technique to identify an abnormal situation where it is useful, for reason to perform a decision-making when a strange event is detected. Sensors nodes are used to collect data samples from real world physical phenomenon. As shown in figure 2, we present an example of a closed neighborhood $N(S_i)$ of the sensor node (S_i) that measure a multi-real valued attribute at each time instant. These value are an m-dimensional random variable [12].

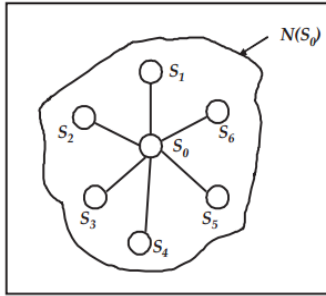


Fig. 2. Example of a closed neighborhood $N(S_i)$ of the sensor node (S_i)

A. Problem Formulation

Let $G = \{s_i = 1 \dots s\}$ be a set of sensor nodes. In every time interval Δt_k every node $s_i \in G$ captures a data vector $x_k^i \in \mathbb{R}^d$ composed of j dimensions such that:

$x_k^i = (x_{k1}^i, x_{k2}^i, \dots, x_{kj}^i)$. During each time window t , S_i captures a set of data measurements $X_k^i = \{x_k^i(t), k = 1 \dots n_i\}$.

Identifying an outlier is important as it can cause the detection to be false or unreliable from other data vectors. Outlier detection method aim at detecting abnormal observations among data vectors collected by sensor nodes [13].

Every sensors sends his local measurements to its cluster-head. The cluster-head collects data vectors from different sensors nodes and combine it with his data vector.

Using Kernel PCA with Mahalanobis kernel, the outlier detection algorithm is performed in the cluster head's node.

Based on activation of detection feature, the detection model is built based on first stream of received data vectors. however, the next data streams received are subject to outlier detection using the initial model once the model is stored on the cluster head.

The process is described as follows: First, in every problem formulation, we present a learning phase which is important step. This latter, is the first step of every learning task. Then, the procedure is performed in the cluster head's node. The cluster head receive the local measurements as a normalized data vector from each member. After that, the cluster head combines his normalized data vector with all received data vectors in a global data matrix where $X_i(0)$ is a $n \times d$

matrix of data vectors collected by members $s_i, i = 1 \dots s$.

However, the data matrix is normalized, global mean and global covariance matrix are calculated [14]. So, the cluster head executes Kernel PCA on this data matrix to establish the detection model.

Secondly, global principal components are calculating. To get the maximal error, every observation of the global data vector on the subspace spanned by the maintained principal components absed on the projection distance is calculated [15]. Three parameters of the global model are defined as follows: the global mean, the global principal components and the maximal error.

Thirdly, in the Learning Procedure, we start by collecting and normalizing data vector. So, the data vector will be sent to the cluster head. Kernel PCA will be executed on normalized data, and the global model will be established.

Then, the cluster head receives periodically data vectors from member nodes in the detection phase. After that, global model are used to detect outliers. Defined by global principal components, the Reconstruction error RE of every data vector on the subspace is calculated.

Finally, based on Reconstruction error RE , the cluster head could decide if an observation is outlier or normal.

B. Kernel Selection

Many types of kernels, in the literature, were employed in the nonlinear transformation of data points (polynomial kernel, sigmoid kernel, etc...). The common choice for most of classification tasks was a Radial Basis Function (RBF) kernel defined as follows:

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

Where $\sigma^2 \in \mathbb{R}^+$ is the depth of the kernel. For every sample of input space, transformation results of such a kernel are similar to those of a density estimator as they give a weighted value. Although some variables may be more relevant than others in practice, this weighting is not defined for each variable separately [16]. A specific depth attributed for each variable, is considered as a possible solution but this is not applicable in practice. A nonlinear principal component analysis based on Mahalanobis kernel function is integrated in detecting outliers in wireless sensor networks. As we know, they was not yet used in the field of data classification. However, the Mahalanobis kernel (MK) is defined as follows [17]:

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

Where Q is the covariance matrix of a data vector. For every dimension of the input space data, mahalanobis kernel differs from the standard RBF kernel in the fact that it defines a specific depth value or weight [18]. Relative to the center of data points, the calculated decision boundary have a non-spherical shape.

C. Parameter selection

In a learning task, using kernel PCA has to be well carried out. So, in order to establish the best model with higher accuracy and lower false alarm rate, choosing the better parameters is important. Kernel PCA based on outlier detection method depends generally on kernel type and kernel parameters. To resolve the non-linearity of data distribution, Mahalanobis kernel, represented by (2), is chosen in this work. It depends on kernel width and number of principal components q [17].

V. EXPERIMENTAL RESULTS

In the field of wireless sensor networks, mahalanobis kernel is used recently and it was introduced in several works, specially based outlier detection for classification data. In comparison to other established kernel-based methods, Kernel PCA performance was showcased [17]. To computes the reconstruction error (RE), kernel PCA transform of a set of test patterns. Then, a training set and a suitable projection dimensionality p , are choosen. Finally, based on the projection, outliers are identified as data points, whose RE exceeds an appropriately established threshold value R_{th} .

Based on real data and on synthetic distribution (See Table1), our method has been tested to show her performance in the field of WSNs.

In our work, we use three synthetic distributions named: sine, square and ring-line-square, described as follow:

Square: The Square consists of four lines, 2.2 long and 0.2 wide distributed with equal probability.

Ring-line-square: The ring-line-square distribution is composed of a ring with an inner diameter of 1.0 and an outer diameter of 2.0, a square with the size as described above, and a 1.6 long and 0.2 wide line connecting the two parts distributed with equal probability.

Sine: The sine distribution consists of a sine-wave. These points are surrounded by 200 points that were distributed randomly with equal probability.

To validate the proposed method, we use Intel Berkeley data as the first real sensor data. These latter, are collected from a closed neighbor-hood from a WSN deployed in Intel Berkeley. Based on the Berkley's sensor dataset, the Kernel PCA was tested. A cluster composed of six nodes are choosen to work on. Mote 45 is a cluster head and motes 43,44,46,47,48 are the members as shown in Figure 3. Grand-St-Bernard is used as the second real data collected from a closed neighborhood from a wireless sensor networks. The node 2 is the cluster head and the nodes 3, 4, 8, 12, 20, 14 are the 6 spatially neighboring nodes.

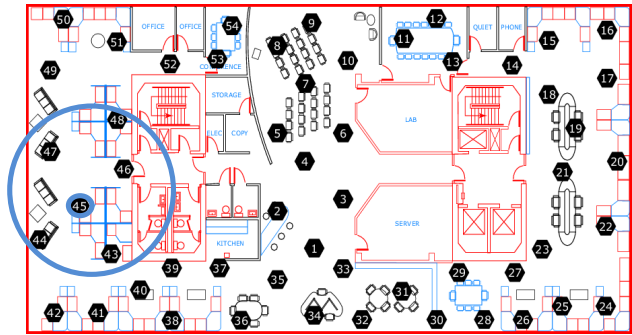


Fig. 3. WSNs of Intel-Berkley Lab

Ambient temperature, soil moisture, relative humidity, watermark and solar radiation measurements are recorded by the network. We use a period of data recorded on 15 days in our experiments, from 4 to 18 September 2007 with two attributes: ambient temperature and relative humidity for each sensor measurement. We calculate the detection rate and the false positive rate for all data points of the test database to measure the precision of our method. The difference between Mahalanobis Distance based KPCA and Reconstruction Error based KPCA are presented in the tables below.

TABLE I. EXPERIMENTAL RESULTS FOR MD-KPCA, RE-KPCA ON THE REAL WORLD DATASETS.

	MD	RE
Intel Berkeley	98.47%	92.38%
Grand-St- Bernard	97.23%	93.59%

TABLE II. EXPERIMENTAL RESULTS FOR MD-KPCA, RE-KPCA ON SYNTHETIC DATASETS.

	MD	RE
Sine	91.26%	87.64%
Square	90.49%	88.73%
Ring-Line Square	91.94%	89.51%

Using mahalanobis distance, is more beneficial to detect outliers when comparing the results given on our experimentation with KPCA-MD and KPCA-RE. Then, the RE may not be an effective measure of deviation from normalcy compared with using the mahalanobis distance. Reconstruction Error produces a decision boundary that is overly broad seen from previous experiment.

Because many potential outliers would not be detected, Reconstruction Error does not satisfactorily fit the normal data. However, compared with the RE-based method, our proposed MD method have an important advantage in terms of performance. As mentioned by the two tables, it detect perfectly the outliers as observed in our experiments. The overall structure of the normal data are much better induced boundary using mahalanobis distance.

The Detection Rate (DR) of KPCA-MK using Mahalanobis Distance is much better than that of KPCA-MK using Reconstruction Error shown in the ROC curve and demonstrated in figure 4. In the field of classification data, Mahalanobis Distance based KPCA is more beneficial than Reconstruction Error in terms of outliers detection, which is noted according to the results of our experiment.

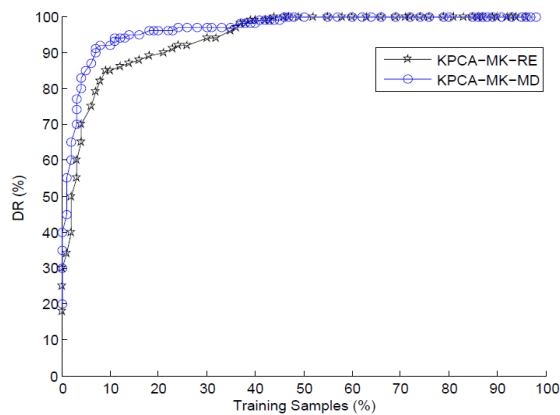


Fig. 4. Comparative ROC curves based KPCA using Mahalanobis Distance and Reconstruction Error.

A comparison between the two kernels is presented as follows. Figure 5, is a detection rate (DR) based KPCA

using MD and RE for real data. Figure 6, is a false positive rate (FPR) based KPCA using MD and RE for real data. In our experiments we acknowledge that, varied by sigma, KPCA using the Mahalanobis distance is more sensitive to the detection of FPR and DR than KPCA using reconstruction error. We are referred to work of Heiko Hoffmann [20] entitled "Kernel PCA for Novelty Detection", to show the robustness of our work using Mahalanobis kernel. After that, we can compare our results Hoffman's job results. We see, after considering the following figures, that Mahalanobis kernel using Mahalanobis distance is more efficient for classification data either by simulation on MATLAB or in Wireless Sensors Networks.

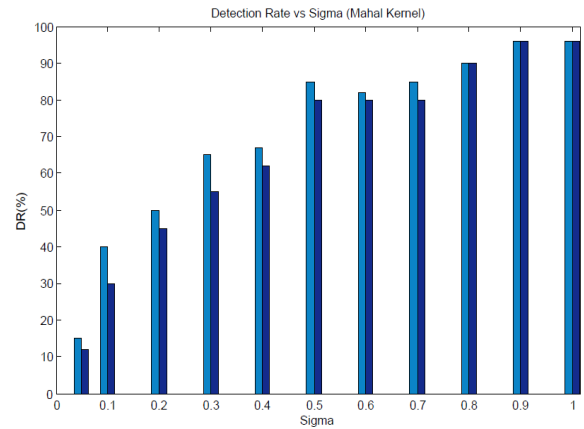


Fig. 5. Detection Rate based KPCA using MD and RE for real data.

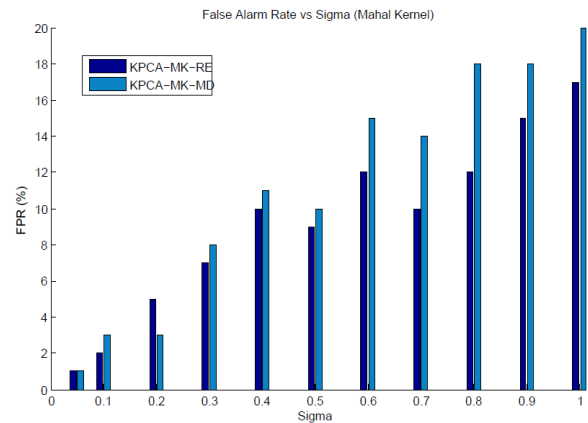


Fig. 6. False Positive Rate based KPCA using MD and RE for real data.

VI. CONCLUSION

In this article, we present our work as a comparative study based on KPCA for outlier detection. This latter, between using Mahalanobis Distance (MD) and Reconstruction Error (RE) in the field of classification data. A principal subspace in an infinite-dimensional feature space described the distribution of training data. To decide if a new

data point is considered as a normal point or outliers, the mahalanobis distance was used as a measure with respect to this subspace. A higher classification performance on a synthetic and real database used compared with KPCA using Reconstruction Error, show that our model based on KPCA using Mahalanobis Distance are more robust against outlier detection within the training set for the classification data. We will focus, as a future work, to improving the performances of the our proposed model to be able to detect events that may occur instead of only considering outliers.

ACKNOWLEDGMENT

This work was supported by King Abdulaziz City for Science and Technology (KACST) and the digital Research Center of Sfax (CRNS) under a research grant (project no. 35/1012).

REFERENCES

- [1] 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.
- [2] F. Marcelloni et M. Vecchio. "An Efficient Lossless Compression Algorithm for Tiny Nodes of Monitoring Wireless Sensor Networks". The Computer Journal 52(8), pp 969-987, 2009.
- [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] I. Enesi, E. Zanjaj, B. Kamo, V. Kolici, O. Shurd. "Image Compression for Wireless Outdoor Sensor Networks". BALWOIS 2010 - Ohrid, Republic of Macedonia, 2010.
- [5] Y. Zhang, N. Meratnia, P. Havinga, "Outlier detection Techniques for wireless sensor networks: A survey", pp .11-20, 2008.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] Y. Zhang, N. Meratnia, P. J.M. Havinga. Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machinel. Ad Hoc Networks , December 2012.
- [11] S. Rajasegarar, C. Leckie, M. Palaniswami., J. C. Bezdek. "Quarter sphere based distributed anomaly detection in wireless sensor networks", Proceedings of IEEE International Conference on Communications, pp. 3864-3869, 2007.
- [12] Y. Zhang, N.A.S. Hamm, N. Meratnia, A. Steinb, M. van de Voorta & P.J.M. Havingaa. Statistics-based outlier detection for wireless sensor networks, Volume 26, Issue 8, 2012.
- [13] 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.
- [14] Chakour et al. 2012. Adaptive kernel principal component analysis for nonlinear time-varying processes monitoring ICEECA2012.
- [15] Harsh K Verma and Kumar V S Samparathi. Article: "Outlier Detection of Data in Wireless Sensor Networks Using Kernel Density Estimation". International Journal of Computer Applications. Published By Foundation of Computer Science, pp 28-32, 2010.
- [16] WenmingZheng, CairongZou, and Li Zhao. "An Improved Algorithm for Kernel Principal Component Analysis". Neural Process, pp: 49-56, 2005.
- [17] Mingtao Ding, ZhengTian, and HaixiaXu. "Adaptive kernel principal component analysis". Signal Process, pp 1542-1553, 2010.
- [18] Minh Hoai Nguyen, Fernando de la Torre. Robust Kernel Principal Component Analysis, 2008.
- [19] <http://sensorscope.epfl.ch/index.php/MainPage>.
- [20] Heiko Hoffmann. Kernel PCA for novelty detection. Pattern Recognition, pp 863-874, 2007.
- [21] X. Liu, U. Kruger, T. Littler, L. Xie, and S. Wang, "Moving window kernel PCA for adaptive monitoring of nonlinear processes," Chemometrics and Intelligent Laboratory Systems, vol. 96, no. 2, pp. 132-143, 2009.
- [22] Nauman Shahid, Naqvi Ijaz Haider and Saad Bin Qaisar, "Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: a survey". Artificial Intelligence Review, 43(2), 193-228, 2015.
- [23] Oussama Ghorbel, Walid Ayedi, Hichem Snoussi, and Mohamed Abid, "Fast and Efficient Outlier Detection Method in Wireless Sensor Networks", IEEE journal, 2015.
- [24] Oussama Ghorbel, Abdulfattah M. Obeid, Mohamed Abid and Hichem Snoussi, "One Class Outlier Detection Method in Wireless Sensor Network:Comparative Study". SoftCOM, 2016.
- [25] Yang Zhang, Nirvana Meratnia, Paul J.M. Havinga, "Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine", Ad Hoc Networks, Volume 11, Issue 3, Pages 1062- 1074, 2013.
- [26] V. Franc and V. Hlavac. "Greedy algorithm for a training set reduction in the kernel methods," in Proc. Int. Conf. Computer Analysis of Images and Patterns, pp. 426-433, 2003.
- [27] 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, Chemical Engineering Science, vol. 60, no. 1, pp. 279-288, 2005.