Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 139 – 146

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

# The Entropy and PCA Based Anomaly Prediction in Data Streams

Daocheng Hong[a], Deshan Zhao[a]*, Yanchun Zhang[a,b]

[a]*Shanghai Key Laboratory of Intelligent Information Processing & Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China*

[b]*Centre for Applied Informatics, Victoria University, Melbourne, Australia*

**Abstract**

With the increase of data and information, anomaly management has been attracting much more attention and become an important research topic gradually. Previous literatures have advocated anomaly discovery and identification ignoring the fact that practice needs anomaly detection in advance (anomaly prediction) but anomaly detection with post-hoc analysis. Given this apparent gap, this research proposes a new approach for anomaly prediction based on PCA (principle component analysis) and information entropy theory, and support vector regression. The main idea of anomaly prediction is to train the historical data and to identify and recognize outlier data according to previous streams patterns and trends. The explorative results of $SO_2$ concentration of exhaust gas in WFGD (Wet Flue Gas Desulfurization) demonstrate a good performance (efficient and accurate) of the target data prediction approach. This robust and novel method can be used to detect and predict the anomaly in data streams, and applied to fault prediction, credit card fraud prediction, intrusion prediction in cyber-security, malignant diagnosis, etc.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of KES International

*Keywords:* data streams analysis; anomaly prediction; data management

## 1. Introduction

In this modern era of Internet + with the online and offline depth fusion, the data assimilation process has changed significantly into the form of data streams. The data stream is a continuous, unbounded sequence of data

* Corresponding author. Tel.: +86+21+65654549; fax: +86+21+65654253.
*E-mail address:* 15210240108@fudan.edu.cn, hongdc@fudan.edu.cn

1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.115

points accompanied special characteristics, such as transiency, uncertainty, dynamic data distribution, multidimensionality, and dynamic relationship. The arrival rate of data stream is usually high and the distributions of data stream often change over time. The most but not the last point, these multiple data streams are not independent. Actually, these data streams frequently demonstrate high correlations with each other in the latent layer. Furthermore, much useful information and knowledge is lost if each stream only is analyzed individually. Outlier or anomaly detection refers to automatic identification and recognization of unforeseen or abnormal phenomena embedded in a large amount of normal data [1].

Predicting anomalies or outliers can be more useful than finding common patterns in data streams, even it's really hard to predict anomalies exactly. It is desirable to find particular aspects of the current stream which are indicative of events of significance to the future event. One of most attractive application scenarios of anomaly prediction is when multiple data streams are targeted since it provides significant information for these applications. For instance, anomaly prediction in medical stream data, for example the electrocardiogram (ECG) signals, can save lives, identify outliers in disease event provides useful information about the possibility of occurring outbreaks. And in industrial processes, anomaly prediction may help to diagnose the incident faults and promote the performance for the organization.

Well known abnormalities can be modeled and detected according to the literatures, but unforeseen problems are not defined and hence are much harder to detect even in a single data stream. In this particular study, we understand that for many scenarios, it is more meaningful to predict abnormal for multiple data streams instead of finding anomaly in individual stream. More specifically, our goal is to monitor multiple data streams so that we observe the value of each stream at every time-tick and to automatically detect anomalies for targeted data stream. The main contributions to the data streams and anomaly management include:

(1) We propose a new approach for anomaly prediction based on the main idea of training the historical data and to identify and recognize outliers according to previous streams patterns and trends which provide a novel and integrated perspective both for academia and industry practice.

(2) We design the features extraction algorithm of anomaly detection based on PCA (principle component analysis) and information entropy theory which is applied to anomaly prediction with support vector regression in data streams.

(3) Within the proposed method, we conduct an experimental study (cooperative research between Fudan University and ABC Company) on the $SO_2$ concentration prediction of exhaust gas in WFGD (Wet Flue Gas Desulfurization) of power plant which shows a good performance (efficient and accurate), and illustrates that the new approach of outlier prediction can be extended to other anomaly management.

After the introduction, we will present the theoretical concepts such as definitions of the theoretical terms and relevant research. And then the novel approach of anomaly prediction is delineated with strong and rigorous logic. After that we will apply this new method to the $SO_2$ concentration prediction of exhaust gas in WFGD of power plant and verify the effectiveness. Finally, the future research directions and conclusions are drawn for both academia and industry practice.

## 2. Related Work

### *2.1. Anomaly Analysis Techniques*

Anomaly analysis aims to detect a small set of observations that deviate considerably from other observations in data [1]. Anomaly detection has been applied to various applications, such as fault detection, credit card fraud detection, intrusion detection of cyber space, and malignant diagnosis [2-5]. Practically, there is only a small amount of labeled data available for the real world applications. Therefore, many researchers of data mining and machine learning communities have been interested in and devoted to the detecting anomaly of unseen data [6-9]. Many anomaly analysis techniques have been proposed in the related works [9–17]. These studies can be divided into three groups roughly: statistical, distance and density-based techniques for anomaly detection. The statistical techniques presume that the data has some predetermined distributions, and then they use the deviation of instance from these predetermined distributions to find the anomaly. However, most distribution models are presumed

univariate, and thus they cannot be used for multidimensional data. Furthermore, it is really a challenge to find and determine the data distribution for the practical problems of real world.

The distance-based techniques of anomaly detection consider a target instance as an outlier if the distances between each target point and its neighbors are above some predetermined threshold. Although these techniques can find anomaly without prior knowledge of data distribution, they might face a problem when the data distribution is complex, for example the multi-clustered structure. In this scenario, anomaly cannot be correctly determined due to determining improper neighbors with these techniques. Due to the limitations of statistical and distance-based techniques, the density-based methods are proposed to improve the effectiveness of anomaly management. A density-based local outlier factor (LOF) is presented to assess the outlierness of each data instance. The anomaly degree for all samples is determined based on the local density of each data instance. The LOF is able to estimate local data structure by using density estimation. Therefore this technique is able to identify anomalies that are hidden under a global data structure. The limitation of this method is that it has high computational complexity since it needs to estimate the local data density for each instance.

Because of the above techniques implemented in batch mode, they cannot be easily extended to identify the anomaly in applications with streaming data or online settings. Recently, some incremental or online anomaly management techniques have been advocated. However, they might not satisfy online detection scenarios in terms of computational cost or memory requirements.

### *2.2. Principle Component Analysis*

Principal component analysis (PCA) is a traditional tool for dimension reduction that projects a multidimensional dataset onto a lower dimensional subspace. Theoretically, PCA reduces the number of variables (dimensionality of the dataset), while keeping most of the variance in original dataset. PCA has been used in variety of applications such as pattern learning of subspace representations, classification for recognition systems and anomaly detection [18-20]. The conventional PCA are carried out in batch mode which means that all data have to be available for calculating the PCA which involves computing the eigenvectors and eigenvalues of the covariance matrix of data [21]. The batch PCA is inefficient for stream data applications, since it has to be retrained whenever a new data arrives under the data streams context. To overcome this problem, many iterative algorithms, such as recursive least square (RLS), stochastic gradient ascent (SGA) and generalized Hebbian algorithm (GHA), have been used to calculate the principal components in sequentially input data [22-24]. These algorithms with lower time complexity than batch PCA converge outputs to the principal components iteratively by using multiple passes over training data. When the new data is added to previous training data and the iterations are restarted to update the principal components.

The anomaly detection method based on PCA is firstly proposed in 2003 [20]. The result of that study shows that the proposed method outperforms density-based local outliers approach in reality. Lakhina et al. [25] proposed the subspace method based on PCA to identify anomaly in network traffic data. This method separates a high dimensional space of network traffic into two subspaces respectively representative of the normal and anomalous components of the original data. Then, new traffic data sample is projected on to the normal and anomalous subspace and classified as normal or anomalous based on different thresholds. This research motivates other works in [26, 27]. Huang et al. [26] proposed network anomaly detection based on approximate PCA analysis. An online oversampling principal component analysis (osPCA) algorithm is proposed to detect outliers from a large amount of data via an online updating technique [27]. The proposed osPCA identifies the anomaly of the target instance according to the variation of the resulting dominant eigenvector by oversampling the target instance and extracting the principal direction of the data.

In summary, while all of the above anomaly detection methods are useful in specific applications, none of them can satisfy all of the specifications that requires for online anomaly management including outlier detection and prediction in WFGD processing. Many techniques are proposed to identify anomaly in univariate data, and thus they cannot be used for multidimensional data. Moreover, they have high computational complexity which might not satisfy online detection scenarios in terms of computational cost or memory requirements. The existing work focus on anomaly detection and to the best of our knowledge no published method can automatically identify, recognize and predict anomalies in Wet Flue Gas Desulfurization.

### 2.3. Information Entropy

In information theory, a mathematical measure of the degree of randomness in a set of data, is with greater randomness implying higher entropy and greater predictability implying lower entropy [28]. In a more technical sense, there are reasons to define information as the negative of the logarithm of the probability distribution. The probability distribution of the events, coupled with the information amount of every event, forms a random variable whose expected value is the average amount of information, or entropy, generated by this distribution. Obviously, it's easy to compute the entropy with the definition in the next part.

Entropy is a measure of the uncertainty of a random variable in information theory. Let $Y$ be a discrete random variable with $r$ states, $y_i$ ($i=1...n$), and probability function $p_i = P\{Y = y_i\}$, $y \in Y$, $\sum p_i = 1$, $0 \leqslant p_i \leqslant 1$. The classic Shannon entropy $H(Y)$ of the discrete random variable $Y$ is defined as follows:

$$H(\mathrm{Y}) = -\sum_{i=1}^{\mathrm{n}} p_i \log p_i \tag{1}$$

And when all data value are the same, $H(Y)=0$; when all data are totally different extremely, $H(Y)=\log_2 N$. Therefore the entropy equation can be easily adopted to analyse attribute distributions of data streams which can be considered discrete random variables.

## 3. Anomaly Prediction Algorithm

The proposed approach of anomaly detection was validated following the steps: (1) matrix generation including data streams attributions (feature) matrix and corresponding entropy matrix; (2) PCA based anomaly detection; (3) selected feature-based anomaly prediction. Step 1, in the data streams, the dimensions of attributions matrix $X$ are defined as $t \times p$ because there are $t$ rows and $p$ columns. Apparently, all feature data shall be normalized when we construct the feature matrix. $X_{ij}$ means the measured value of feature $p$ at the $t$ time. The generated matrix of data features will be inputted new observed value with the assistance of slide window as shown Fig. 1. Entropy matrix is constructed similarly.

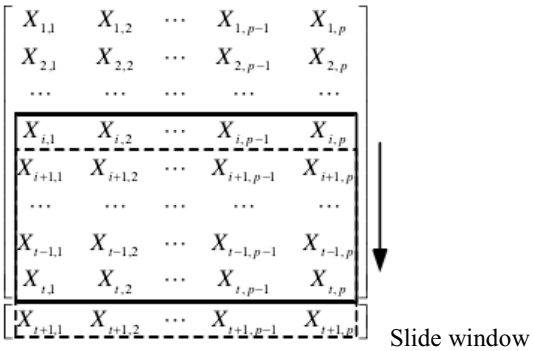


Fig. 1. Matrix Generation of Data Attributes

Step 2, attributions matrix and entropy matrix are going to be processed with PCA method. We construct normal subspace $s$ ($p \times k$) with the first $k$ eigenvectors of principle components, and the other outliers subspace $s\tilde{}$ ($p \times (p-k)$) with the left $p$-$k$ eigenvectors. $x$ is the vectors at time $t$, and $x=\hat{x}+\tilde{x}$. $\hat{x}$ is the projection of $x$ in the normal subspace, $\tilde{x}$ is the projection of $x$ in anomaly subspace and also named residuals. The matrix $P$ ($p \times k$) is composed of the $k$ eigenvectors of normal subspace $s$, and $x$ can be computed as:

$$\hat{x} = PP^T x \quad (2)$$

$$\tilde{x} = (I-PP^T)x \quad (3)$$

Then, anomaly detection is conducted by comparing the squared prediction error (*SPE*) with threshold $\delta_\alpha^2$. The $\delta_\alpha^2$ is calculated with this method [25]. In this step, anomaly data can be detected and eliminated.

$$SPE=|| \tilde{x}||^2 =|| (I-PP^T)x||^2 \quad (4)$$

Step 3, feature selection is carried out. We choose the top *r* components with the greatest amounts data variance that PCA captures as principal components. SVR (support vector regression) prediction model is trained based on *r* principal components. $f(x_i)$ are predicting outcomes, $w^T$ and $b$ are parameters based on training set. $\varphi(x_i)$ is kernel function.

$$f(x_i) = w^T \varphi(x_i) + b \quad (5)$$

We can get final prediction model by constructing Lagrange function，Duality (optimization), kernel function mapping and other steps. $\alpha$ and $\alpha_i^*$ are Lagrange multiplier， $K(x, x_i)$ is kernel function selected according to practical problems.

$$f(x) = \sum_{i=1}^{t} (\alpha - \alpha_i^*) K(x, x_i) + b \quad (6)$$

The proposed method requires an *O(t^3)* training time. When the trained model is deployed, newly collected data is mapped to the principal components for data prediction and anomaly prediction. The online prediction time complexity is *O(1)*, guaranteeing real-time feedback.

## 4. Experiment and Results

The dataset is collected by various sensors in WFGD (Wet Flue Gas Desulfurization) process of power plant A. The background of this experiment is that restrict regulations and supervision by the environment supervisory authority require power plants to predict the $SO_2$ concentration of exhaust gas, to refine the WFGD procedure avoiding heavy fines. Considering the environment protection and regulations the managers would like to control the $SO_2$ concentration of exhaust gas. They are eager to detect outlier of $SO_2$ concentration of flue gas in advance according to the procedure data of WFGD.

In this paper, a case study with the new approach of anomaly prediction as the above session was carried out. Feature or attribution selection of data is very important and shall be processed according to the practical demands. After consulting the managers and a few preliminary tests, 10 features were selected finally as the Table 1.

Table 1. Selected Data Features.

| Code Number | Name |
|---|---|
| V000 | Flow velocity of flue gas at the entrance |
| V002 | $O_2$ concentration of flue gas at the entrance |
| V003 | $SO_2$ concentration of flue gas at the entrance |
| V005 | Static pressure of flue gas at the entrance |
| V007 | Temperature of flue gas at the entrance |
| V023 | $SO_2$ concentration of exhaust gas |
| V040 | PH value of slurry |
| V072 | Current of slurry circulating pump |

| | |
|---|---|
| V109 | Air rate of WFGD units |
| V136 | Flow rate of coal feeder |

With the new method in previous session, feature matrix and entropy matrix were constructed, PCA based outlier detection was conducted on the data streams with 20000 samples. The scree plot on the figure 2 only shows the first six selected components. They explain 95% of the total variance so that it can capture most of the information of the data in a space of lower dimension. The results also show time series plots of the *SPE* on the figure 3. Note that the anomalies (marked with red cross) are more distinct after the data is mapped to the outlier subspace $s\tilde{}$. This makes the prediction easier and more effective.

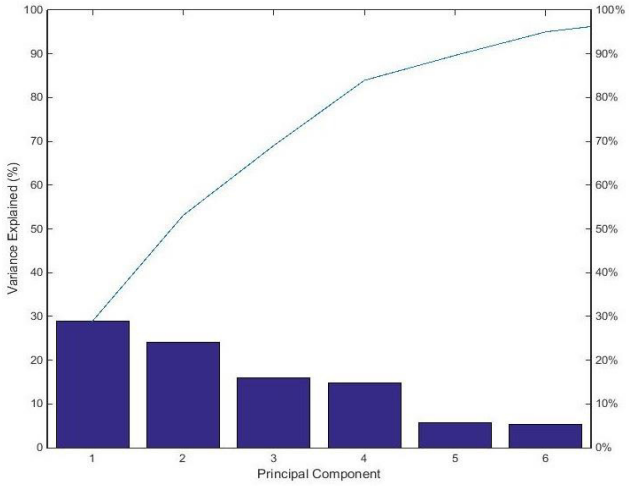


Fig. 2. Principal Component Selection

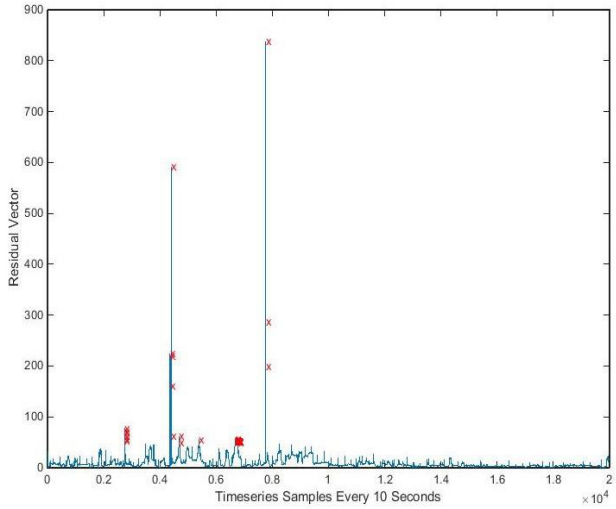


Fig. 3. Residual Vector Squared Magnitude (*SPE*)

A prediction model (SVR based on linear kernel function) is built after outlier elimination. Figure 4 shows 10000 measured values collected every ten seconds and the corresponding predicted values.

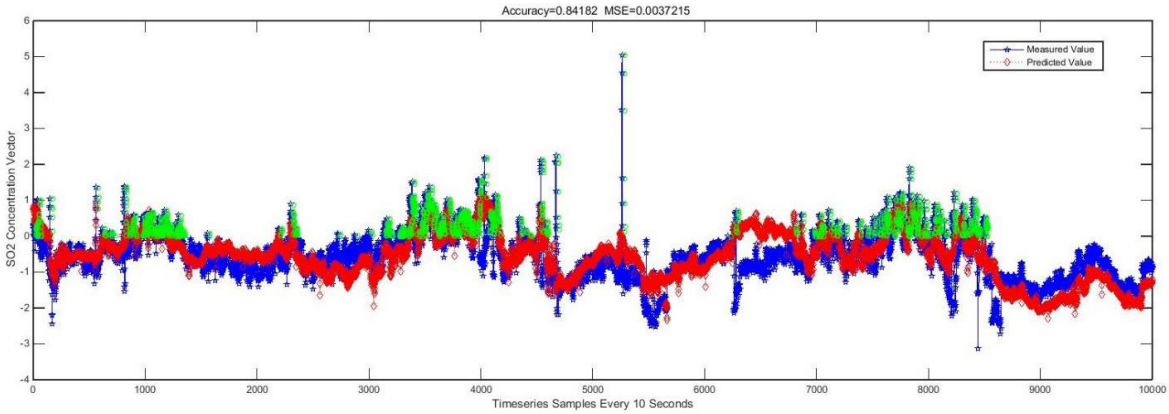


Fig. 4. Anomaly Prediction of $SO_2$ Concentration for WFGD

The anomalies (marked with green circle) can be predicted by analysing the relative errors between the measured values and the corresponding predicted values. In terms of the accuracy value displayed on the top of Fig. 4, we consider a prediction accurate if the percentage of relative error does not exceed 0.2 (0.2 is a threshold advised by domain experts). The high accuracy and low mean square error indicates that our method has a good performance for $SO_2$ concentration anomaly prediction in WFGD.

## 5. Conclusion

Previous literatures paid much more attention on anomaly discovery and identification ignoring the fact that practice needs anomaly detection in advance but anomaly detection post hoc analysis. Considering this apparent gap, this study proposes a new approach for anomaly prediction based on information entropy theory and PCA theory. The main idea of this method is to construct data feature matrix and corresponding entropy matrix, and then analyse matrix with PCA. Further, the support vector regression prediction model is trained based on previous principal components. The explorative results of $SO_2$ concentration anomaly prediction in WFGD demonstrate a good performance of the proposed approach. Therefore, this robust and novel method can be used to detect and predict the anomaly in variety of data streams scenarios.

The aims of this paper are to develop a new approach of anomaly prediction for data stream and make out the main contributions to the data management. The proposed novel model of anomaly prediction in advance by focusing on the data feature matrix and corresponding entropy matrix, PCA and support vector regression is a comprehensive perspective ignored by most researchers in data management. With the new approach, an experimental study of $SO_2$ concentration prediction of WFGD has been conducted in the mainland of China which shows the model validity and reliability, and illustrates that this new approach can be extended to other data streams prediction in practice. In the near future, we will apply this novel anomaly prediction approach to credit card fraud prediction, malignant diagnosis. We also would like to conduct comparative study on public research-quality datasets and verify the performance of our new method.

## 6. Acknowledgements

The authors thank the anonymous referees for their valuable comments and suggestions. This work was supported partly by NSFC (61332013, 61472086), Humanities and Social Sciences Foundation of MOE (15YJC630032), Senior Visiting Professor Project of Fudan University and IIPL (IIPL2014-001).

## References

1. Hawkins D. M. *Identification of outliers*. Chapman and Hall, London, 1980
2. Rawat S., A. K. Pujari, and V. P. Gulati. On the use of singular value decomposition for a fast intrusion detection system, *Electronic Notes in Theoretical Computer Science*, 2006, 142. 215–228.
3. Huang L., X. Nguyen, M. Garofalakis, M. I. Jordan, A. Joseph, and N. Taft. In-network PCA and anomaly detection, *Advances in Neural Information Processing Systems*, 2007, 19. 617.
4. Lazarevic A., L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava. A comparative study of anomaly detection schemes in network intrusion detection, *Proc. SIAM*. 2003.
5. Wang W., X. Guan, and X. Zhang. A novel intrusion detection method based on principle component analysis in computer security, in *Advances in Neural Networks*,2004, 657–662.
6. Kriegel H.-P., M. Schubert and A. Zimek. Angle-based outlier detection in high-dimensional data, in Proceedings of *ACM SIGKDD*,2008, 444–452.
7. Chandola V., A. Banerjee, and V. Kumar. Anomaly detection: A survey, *ACM Computing Surveys (CSUR)*,2009,41, 3. 15.
8. Song X., M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection, *Knowledge and Data Engineering, IEEE Transactions on*, 2007, 19(5): 631–645.
9. Breunig M. M., H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers, in *ACM Sigmod Record*, 2000,29,2, 93–104.
10. Knox E. M. and R. T. Ng. Algorithms for mining distance-based outliers in large datasets, in Proceedings of the *International Conference on Very Large Data Bases*, 1998.
11. Angiulli F., S. Basta, and C. Pizzuti. Distance-based detection and prediction of outliers, *Knowledge and Data Engineering, IEEE Transactions on*, 2006, 18, 2. 145–160.
12. Kriegel H.-P., P. Kröger, E. Schubert, and A. Zimek. Outlier detection in axis-parallel subspaces of high dimensional data, in *Advances in Knowledge Discovery and Data Mining*, Springer, 2009, 831–838.
13. Barnett V. and T. Lewis. Outliers in statistical data, vol. 3. Wiley New York, 1994.
14. Jin W., A. K. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship, in *Advances in Knowledge Discovery and Data Mining*, Springer, 2006, 577–593.
15. Khoa N. L. D. and S. Chawla. Robust outlier detection using commute time and eigenspace embedding, in *Advances in Knowledge Discovery and Data Mining*, Springer, 2010, 422–434.
16. Ma J.，L. Sun，H. Wang，Y. Zhang. Supervised Anomaly Detection in Uncertain Pseudo-Periodic Data Streams，A*CM Transactions on Internet Technology*，2016,16(1)：4。
17. Zhou X.，J. He，G. Huang，Y. Zhang. SVD-Based Incremental Approaches for Recommender Systems, *Journal of Computer and System Sciences*，2015,81(4):717-733
18. Skovcaj D. and A. Leonardis. Incremental and robust learning of subspace representations, *Image and Vision Computing*, 2008, 26, 1. 27–38,
19. Zhao H., P. C. Yuen, and J. T. Kwok. A novel incremental principal component analysis and its application for face recognition, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 2006, 36, 4. 873–886.
20. Shyu M.-L., S.-C. Chen, K. Sarinnapakorn, and L. Chang. A novel anomaly detection scheme based on principal component classifier, *IEEE Foundations and New Directions of Data Mining Workshop, in Conjunction with ICDM03*, 2003,172-179.
21. Li Y., L.-Q. Xu, J. Morphett, and R. Jacobs. An integrated algorithm of incremental and robust pca, *Proceedings of International Conference on Image Processing*, 2003. 1, I–245.
22. Bannour S. and M. R. Azimi-Sadjadi. Principal component extraction using recursive least squares learning, *Neural Networks, IEEE Transactions on*, 1995,6, 2. 457–469.
23. Oja E. Simplified neuron model as a principal component analyzer, *Journal of mathematical biology*, v1982,15, 3. 267–273.
24. Sanger T. D. Optimal unsupervised learning in a single-layer linear feed forward neural network, *Neural networks*, 1989,2, 6. 459–473.
25. Lakhina A, Crovella M, Diot C. Diagnosing Network-wide Traffic Anomalies. Proc of the ACM SIGCOMM 2004. New York: 219-230.
26. Huang L., Xuan Long Nguyen, Minos Garofalakis, Joseph M. Hellerstein, Michael I. Jordan, Anthony D. Joseph, and Nina Taft. Communication-efficient online detection of network-wide anomalies. In I*NFOCOM 2007*. 134-142..
27. Lee Y. J., Y. R. Yeh, and Y. C. F. Wang. Anomaly detection via online over-sampling principal component analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2012.
28. Gray, R. M. *Entropy and Information Theory*, Springer, NY, 2011