IoT 네트워크에서 다중 스케일 PCA를 사용한 트렌드 적응형 이상 탐지

Thien-Binh Dang, Manh-Hung Tran, Duc-Tai Le, Hyunseung Choo
e-mail: {dtbinh, hungtm, ldtai, choo}@skku.edu

Trend-adaptive Anomaly Detection with Multi-Scale PCA in IoT Networks

Thien-Binh Dang, Manh-Hung Tran, Duc-Tai Le, Hyunseung Choo
The College of Software, Sungkyunkwan University

Abstract

A wide range of IoT applications use information collected from networks of sensors for monitoring and controlling purposes. However, the frequent appearance of fault data makes it difficult to extract correct information, thereby sending incorrect commands to actuators that can threaten human privacy and safety. For this reason, it is necessary to have a mechanism to detect fault data collected from sensors. In this paper, we present a trend-adaptive multi-scale principal component analysis (Trend-adaptive MS-PCA) model for data fault detection. The proposed model inherits advantages of Discrete Wavelet Transform (DWT) in capturing time-frequency information and advantages of PCA in extracting correlation among sensors' data. Experimental results on a real dataset show the high effectiveness of the proposed model in data fault detection.

1. Introduction

The advances in mobile computing, sensor industry and the rapid growth in the number of Internet applications and services are paving the way to next generation of the Internet of Things (IoT) [1]. There everything in the physical world is integrated into virtual world that enables the Internet to touch all aspects of human life. An IoT network typically consists of a significant number of IoT end nodes including sensors for monitoring and actuators for controlling purposes. Sensors and actuators has been widely employed for various IoT applications such as smart house, smart building, smart city, smart agriculture, smart grid and so on. Many real deployment experiences of IoT applications show that data collected from sensors is prone to being faulty due to various reasons including internal factors or external factors such as environment interference (Elnahrawy, E., et al., 2003) [2], sensor ageing, battery drain (Toll, G., et al., 2005) [3], hardware malfunctions (Ramanathan, N., et al., 2012) [4], malicious attacks (Roy, S., et al., 2012) [5] and so on. The faulty data collected from sensors not only provides wrong information about phenomenons but also leads to improper decisions.

In order to keep sensory data accurate and reliable, it is necessary to develop efficient data fault detection algorithms. In reality, measurements from sensors can have various types of failures such as outliers, spikes, excessive noise, offset fault, stuck-at fault, degradation fault, drift fault, ect [4]. These faults appear in various forms which are different in lengths, magnitudes or patterns. For example, some faults show abrupt changes in sensor readings in short duration and quite easy to identify, whereas others show slowly and relatively gradual changes in long duration and hard to detect. Because of the variety of fault types, there is no single method to perfectly detect all type of faults. For this reason, a hybrid approach combining advantages of different techniques is a sensible direction to take into account.

PCA is a powerful tool to analyze multivariate data collected from IoT networks. There are many works using PCA for data fault detection [6]-[8]. However, conventional PCA cannot capture time-frequency information of time-series data which is generated by almost practical IoT applications. In addition, PCA is a single-scale analyzing model, while data generated by sensors is usually dynamic and in multi-scales. Similarly, Oussama Ghorbel et al propose kernel-PCA using Mahalanobis distance to detect outliers in sensory data [9]. Even though the simulation results show that proposed approach can effectively detect outliers with high precision rate, it still cannot overcome the limitations of PCA technique that is unable to capture time-frequency information and multi-scales property of time-series data. To overcome the limitations of PCA, Xie Ying-

xin and Z. Wang use MS-PCA to detect fault data collected by wireless sensor networks (WSN). Experimental results show that MS-PCA can capture timefrequency information and be able to detect fault data at multi-resolutions. Nevertheless, conventional MS-PCA is not sensitive with fault data having small changes or low noise intensity.

The rest of the paper is organized as follows. Section 2 provides preliminaries of DWT and MS-PCA. Proposed model for data fault detection is introduced in section 3. Section 4 describes evaluation methods and experiment results on a real dataset. Eventually, some conclusions and future works are drawn in section 5.

2. Preliminaries

A. Discrete Wavelet Transform

Discrete Wavelet Transform(DWT) is one of the most powerful technique and widely used in analyzing time-series data. In practical situations, we encounter anomalous timeseries data which contains noises and abrupt changes because of faults on sensors. The multi-resolution property and time frequency localization of DWT can sensitively reveal anomalous sensory data. In addition, the time complexity of DWT takes only O(N) which is very important in dealing with huge data generated from the IoT networks. Haar filter is one of the most popular tool in Discrete Wavelet Transform (DWT) family used to analyze the time series data. In this paper, we propose the idea of using Haar filter to detect anomalies in sensor data. In more detail, the sensor data is represented by a vector $x = [x_1\ x_2\ \cdots\ x_N]$. This vector will be decomposed into approximation coefficients and detail coefficients by formulars (1) (2) as below:

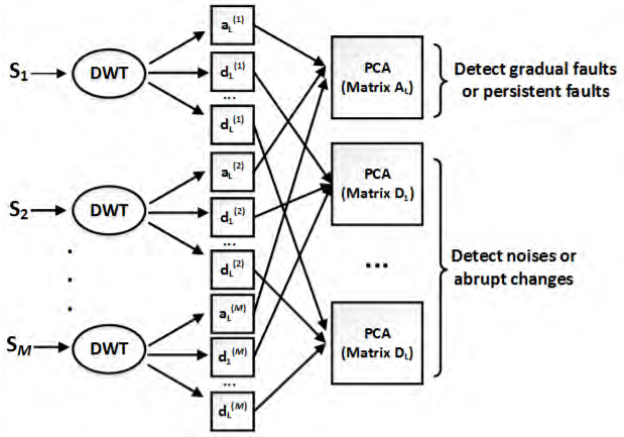
$$y_{high}[k] = \sum_n x[n]\, g[2k-n] \qquad (1)$$

$$y_{low}[k] = \sum_n x[n]\, h[2k-n] \qquad (2)$$

, where g[n] and h[n] are high pass filter and low pass filter respectively. While approximation coefficients represent the trend, the detail coefficients describe the transient of the signal. For this reason, while the noises or abrupt anomalies are shown clearly in detail coefficients, gradual and persistent anomalies can be detected effectively by approximation coefficients. In order to detect anomalies, we calculate the Euclidean Distance (ED) between training data vector and testing data vector and then compare with control limits which are obtained in training phase.

B. Multi-Scale Principal Component Analysis

In IoT networks, while the major advantage of DWT is the ability of capturing time-frequency localization of individual sensor's data at different scales, conventional Principal Component Analysis (PCA) can efficiently extract correlation across data generated by a network of sensors. Therefore, Multi-Scale Principal Component Analysis (MS-PCA) model is the idea of combining DWT and PCA to extract maximum information from data. The work-flow of MS-PCA model is shown in the Fig. 1. Data of M sensors S_1, S_2, ..., S_M is collected



(Figure 1) MS-PCA model for fault detection

into M columns of a matrix S. DWT technique is then used to decompose each column i (i = 1...M) of matrix S into approximation coefficients $a^{(i)}_k$ and detail coefficients $d^{(i)}_k$, where L is the decomposition level and k = 1...L. Approximation coefficients $a^{(i)}_L$ are collected into columns of the matrix A_L which represents the trends within all sensors in the network. Matrix D_k which captures the deviations of sensors is constructed by combining detail coefficients $d^{(i)}_k$ at level k of all sensors. Therefore, a total of L + 1 matrices (A_L and $D_1 \ldots D_L$) are formed representing multi-time scales of sensors' data. Eventually, correlations across the sensors at each scale is extracted by applying PCA to each of these L + 1 matrices.

3. Proposed Model For Data Fault Detection

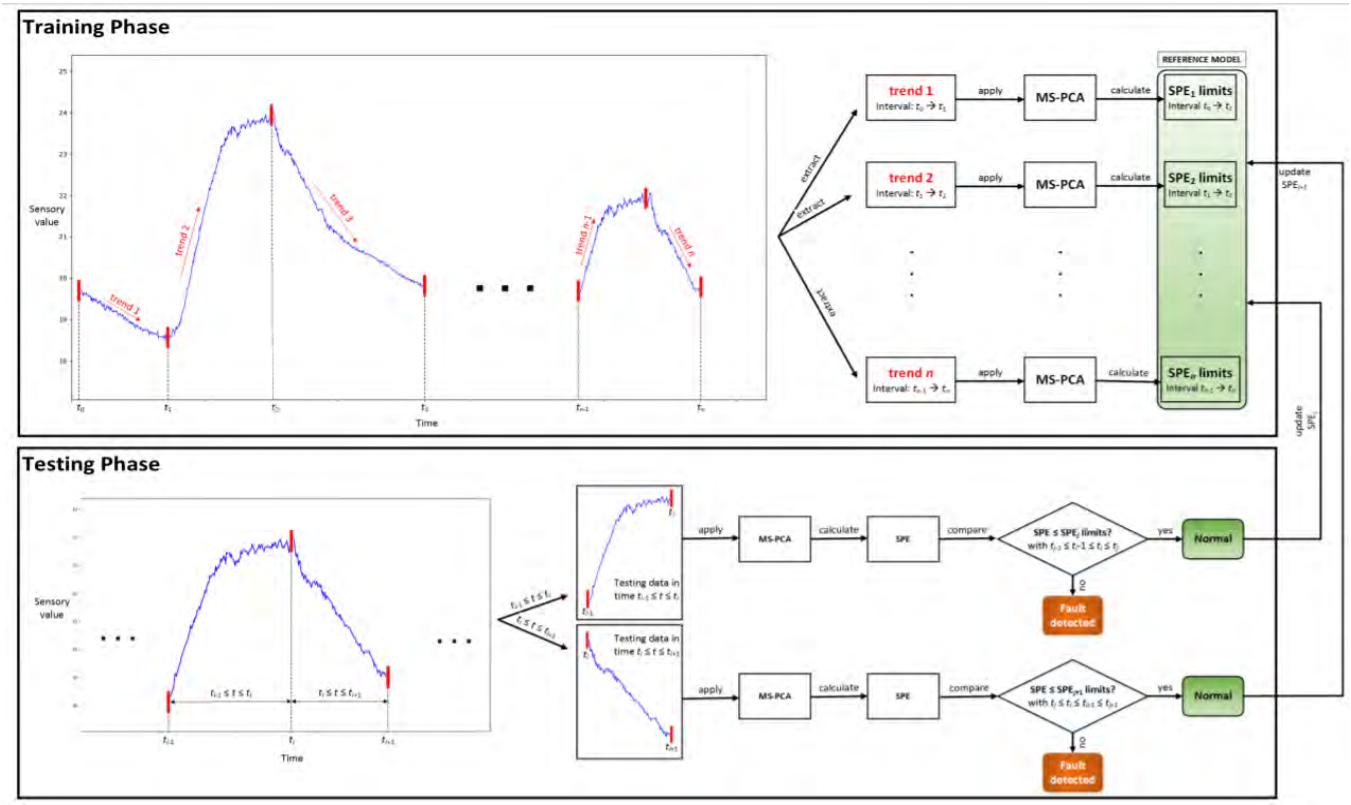
As discussed in the previous section, MS-PCA model shown that it is an effective tool for data fault detection, especially on multivariate data generated by a network of sensors. However, its sensitiveness of detection can be improved by applying suitable training strategy based on characteristics of data. In this section, we propose the Trend-adaptive MS-PCA model for data fault detection based on the MS-PCA model.

The proposed scheme has two phases: training phase and testing phase. In training phase, the normal time-series data is split into trends 1...n corresponding to time intervals ($t_0 \rightarrow t_1$), ($t_1 \rightarrow t_2$), etc. The minimal number of samples of each trend is predefined which ensures enough input data for training. In case the trend is too short, it will be merged with the next trend or the previous trend (if the considering trend is the last trend). MS-PCA is then applied for each trend to calculate SPE limits for each interval including the SPE limits of different resolutions of the last approximation part A_L and the detail parts D_1, D_2, ..., D_L. In testing phase, first the data is split into different sets corresponding to time intervals which are established in training phase. Then, MS-PCA is applied on these sets of data to calculate SPE values for each sample (one sample has $L + 1$ SPE values corresponding to A_L and D_1, D_2, ..., D_L). The model checks whether a sample s is fault or normal by comparing its SPE values to corresponding SPE limits whose time interval contains sampling time of s. If the SPE values of s is less or equal to the SPE limits, the sample s is normal and its SPE values will be used to update the reference model in order to ensure the self-adaptability of the detection system; otherwise s is detected as fault data.

4. Performance Evaluation

A. Datasets acquisition

In this research, a real dataset taken from Intel Berkeley Research lab (IBRL) is carried out in order to evaluate the efficiency of the proposed model. IBRL dataset was collected from a network of 54 Mica2Dot sensors which are deployed in the Intel Berkeley Research lab between February 28th and April 25th, 2004. Four sensory measurements were collected in 31s intervals include humidity, temperature, light and voltage values. With no loss of generality, we choose temperature measurements from seven sensors whose IDs are 1, 2, 33, 34, 35, 36 and 37 for experiment (as shown inside the red circle in Fig. 4). In our experiment, temperature readings from these seven sensors are re-sampled at 108s interval so a total of 800 samples are taken in one day. We use 800 samples of February 28th for training and 200 samples in the time period from 0 AM to 6:40 AM of February 29th (these samples are in the first trend's interval) for testing. Also, we manually remove error readings to ensure the accuracy for training data.

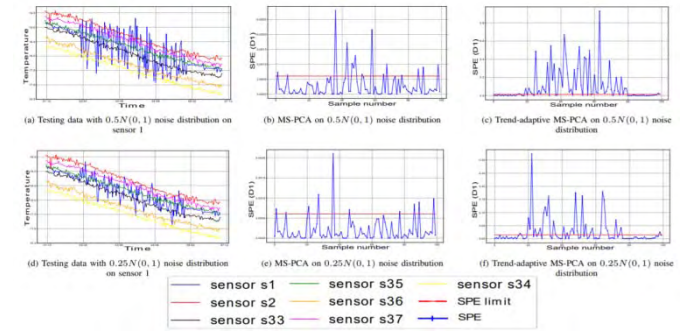


(Figure 2) Trend-adaptive MS-PCA model for fault detection

B. Evaluation Method and Experimental Results

Precision degradation fault

Precision degradation fault is a fault which contains high amount of unexpected noises in sensor data. These noises may be caused by low battery or hardware failure on sensors. In this experiment, we add random noises following the normal distributions $0.5\mathcal{N}(0, 1)$ (Fig. 3a) and $0.25\mathcal{N}(0, 1)$ (Fig. 3d) respectively into the range of samples from 40 to 160 of testing data of sensor 1. As can be seen from the Fig. 3b and Fig. 3c, both MS-PCA and Trend-adaptive MS-PCA can detect noises but Trend-adaptive MS-PCA shown an enhanced performance over MS-PCA. In the next experiment, we reduce the noise intensity to 0.25. The obtained results shown that it does not affect Trend-adaptive MS-PCA's performance. Nevertheless, a significant number of false negatives appear in MS-PCA. It can be concluded that Trend-adaptive MS-PCA is more sensitive with data fault than MS-PCA.



(Figure 3) Experimental results on 0.5N(0, 1) noise distribution and 0.25N(0, 1) noise distribution

5. Conclusion

In this paper, we proposed a data fault detection model based on Trend-adaptive MS-PCA for sensory data collected from IoT networks. Since Trend-adaptive MS-PCA adopts full characteristics of DWT and PCA techniques, it is capable for detecting fault data in multi-resolutions and capturing the correlation among sensors' data. Moreover, splitting data into trends makes the proposed model more sensitive with fault data. In order to evaluate the proposed model, we use a real dataset collected from a real development of sensors and different types of faults with different magnitudes are artificially injected into testing data. The obtained results show that Trend-adaptive MS-PCA not only performs well in detecting abrupt faults and high noise faults, but also it is still sensitive with small changes and low noise intensity in data. Moreover, the simulation results prove that Trend-adaptive MS-PCA outperforms MS-PCA in detection of precision degradation fault. In our future work, we continue to test Trend-adaptive MS-PCA with other types of faults such as offset fault, drift fault or stuck-at fault. Last but not least, a study on an algorithm for automatically splitting data into trends is necessary to automate the process steps of Trend-adaptive MS-PCA.

ACKNOWLEDGEMENT

본 논문은 기초연구사업 (NRF-2010-0020210)과 과학기술정보통신부 및 정보통신기술진흥센터의 Grand ICT 연구센터지원사업 (IITP-2018-2015-0-00742)의 연구결과로 수행되었음

References

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," Comput. Netw., vol. 54, no. 15, pp. 2787–2805, Oct. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2010.05.010

[2] E. Elnahrawy and B. Nath, "Cleaning and querying noisy sensors," in Proceedings of the 2Nd ACM International Conference on Wireless Sensor Networks and Applications, ser. WSNA '03. New York, NY, USA: ACM, 2003, pp. 78–87. [Online]. Available: http://doi.acm.org/10.1145/941350.941362

[3] G. Tolle, J. Polastre, R. Szewczyk, D. Culler, N. Turner, K. Tu, S. Burgess, T. Dawson, P. Buonadonna, D. Gay, and W. Hong, "A macroscope in the redwoods," in Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems, ser. SenSys '05. New York, NY, USA: ACM, 2005, pp. 51–63. [Online]. Available: http://doi.acm.org/10.1145/1098918.1098925

[4] K. Ni, N. Ramanathan, M. N. H. Chehade, L. Balzano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen, and M. Srivastava, "Sensor network data fault types," ACM Trans. Sen. Netw., vol. 5, no. 3, pp. 25:1–25:29, Jun. 2009. [Online]. Available: http://doi.acm.org/10.1145/1525856.1525863

[5] S. Roy, M. Conti, S. Setia, and S. Jajodia, "Secure data aggregation in wireless sensor networks," IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 1040–1052, June 2012.

[6] X. L. Zhang, F. Zhang, J. Yuan, J. lan Weng, and W. h. Zhang, "Sensor fault diagnosis and location for small and medium-scale wireless sensor networks," in 2010 Sixth International Conference on Natural Computation, vol. 7, Aug 2010, pp. 3628–3632.

[7] M. Livani and M. Abadi, "Distributed pca-based anomaly detection in wireless sensor networks," in Internet Technology and Secured Transactions (ICITST), 2010 International Conference for, 2010, pp. 1–8.

[8] 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 2010 Second International conference on Computing, Communication and Networking Technologies, July 2010, pp. 1–9.

[9] O. Ghorbel, W. Ayedi, H. Snoussi, and M. Abid, "Fast and efficient outlier detection method in wireless sensor networks," IEEE Sensors Journal, vol. 15, no. 6, pp. 3403–3411, June 2015.

[10] Y. x. Xie, X. g. Chen, and J. Zhao, "Data fault detection for wireless sensor networks using multi-scale pca method," in 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Aug 2011, pp. 7035–7038.