The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/2398-7294.htm

An anomaly detection method to improve the intelligent level of smart articles based on multiple group correlation probability models

Received 17 September 2019
Revised 13 October 2019
Accepted 16 October 2019

Xudong Lu
School of Software, Shandong University, Jinan, China

Shipeng Wang
School of Software, Research Center of Software and Data Engineering, Shandong University, Jinan, China

Fengjian Kang
School of Software, Shandong University, Jinan, China

Shijun Liu
School of Computer Science and Technology, Shandong University, Jinan, China, and

Hui Li, Xiangzhen Xu and Lizhen Cui
Shandong University, Jinan, China

Abstract

Purpose – The purpose of this paper is to detect abnormal data of complex and sophisticated industrial equipment with sensors quickly and accurately. Due to the rapid development of the Internet of Things, more and more equipment is equipped with sensors, especially more complex and sophisticated industrial equipment is installed with a large number of sensors. A large amount of monitoring data is quickly collected to monitor the operation of the equipment. How to detect abnormal data quickly and accurately has become a challenge.

Design/methodology/approach – In this paper, the authors propose an approach called Multiple Group Correlation-based Anomaly Detection (MGCAD), which can detect equipment anomaly quickly and accurately. The single-point anomaly degree of equipment and the correlation of each kind of data sequence are modeled by using multi-group correlation probability model (a probability distribution model which is helpful to the anomaly detection of equipment), and the anomaly detection of equipment is realized.

Findings – The simulation data set experiments based on real data show that MGCAD has better performance than existing methods in processing multiple monitoring data sequences.

© Xudong Lu, Shipeng Wang, Fengjian Kang, Shijun Liu, Hui Li, Xiangzhen Xu and Lizhen Cui. Published in *International Journal of Crowd Science*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

This work is partially supported by National Key R&D Program No.2017YFB1400100.

International Journal of Crowd Science
Vol. 3 No. 3, 2019
pp. 333-347
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-09-2019-0024

Originality/value – The MGCAD method can detect abnormal data quickly and accurately, promote the intelligent level of smart articles and ultimately help to project the real world into cyber space in CrowdIntell Network.

Keywords Internet of things, Clustering, Anomaly detection, Intelligent level

Paper type Research paper

1. Introduction

As human beings enter the network era, the phenomenon of intelligence is becoming more and more extensive and complex. The individuals, enterprises, governmental agencies, smart equipment and articles in the physical space are becoming more and more intelligent in future Web-based industrial operation systems and social operation management patterns (Chai *et al.*, 2017). The smart article is a kind of intelligent subject in CrowdIntell Network (Wang *et al.*, 2019). And the smart articles have one or more functions and abilities to solve problems and complete tasks, which can assist other intelligent subjects (such as people) to complete certain transactions. In CrowdIntell Network, the smart articles mainly include intelligent monitoring equipment, intelligent transportation equipment, intelligent communication equipment and other intelligent auxiliary equipment.

Unlike the other three Digital-selfs in CrowdIntell Network, the Digital-self of smart article has no mental model (Wang *et al.*, 2019) and its intelligent level is reflected in the level of the ability to complete the task. For example, in CrowdIntell Network, sensors can collect data and assist to project the real world into cyber space. The ability of sensors to collect data and project raw data into cyber space reflects their level of intelligence. For example, sensors that can collect heterogeneous data may have higher level of intelligence than sensors that can only collect homogeneous data. In the process of collecting real world data and projecting it into cyber space, whether the sensors can actively detect abnormal data also reflect the intelligent level of the smart articles. The more intelligent the sensor is, the more abnormal data can be detected; then the abnormal results will be fed back to the individual. In the meantime, CrowdIntell Network needs comprehensive, real, correct and synchronous projection, so the sensor should be able to achieve rapid detection of anomalies in various monitoring data.

The Internet of Things technology (Atzori *et al.*, 2010) is an important application of smart articles in CrowdIntell Network. It can be observed that the Internet of Things technology has developed rapidly in recent years. The Internet of Things collects information through various information sensors and monitors objects in real time. The working condition of equipment is generally monitored by the condition monitoring system. The condition monitoring system generates multiple monitoring data while working. If the device runs abnormally, it can typically affect the collected monitoring data during the fault period. The identification of abnormal monitoring data can help us identify abnormal equipment and avoid incorrect data projection into CrowdIntell Network as far as possible.

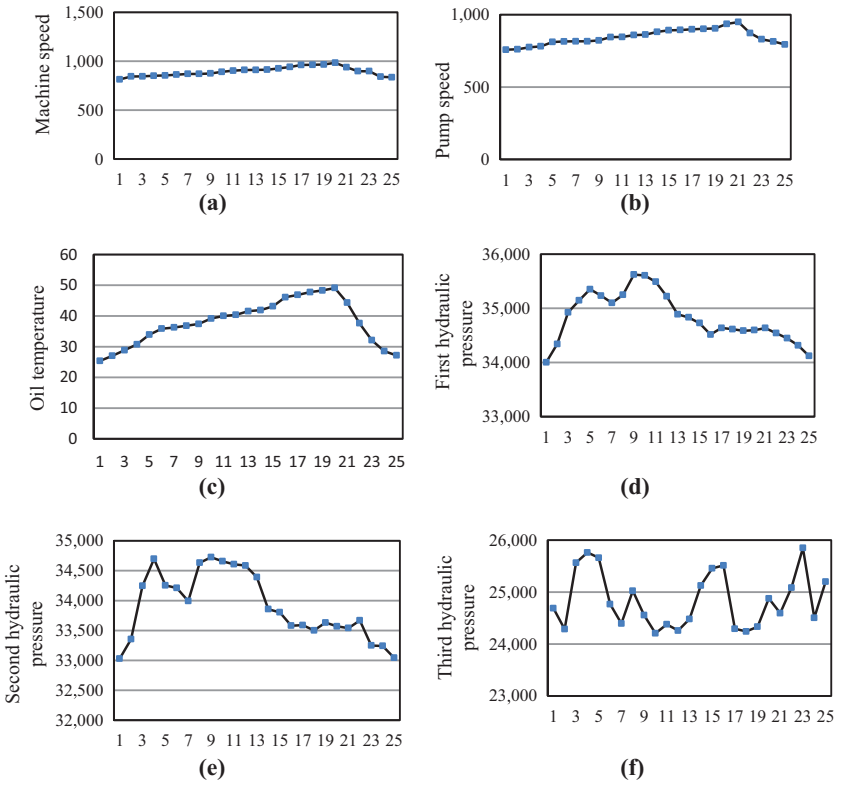
However, due to the rapid increase in the amount of monitoring data and monitoring data types continue to increase, the rapid detection of anomaly in various monitoring data has the following several challenges:

- If a condition monitoring system is to monitor multiple equipment, each equipment is fitted with multiple sensors. With the time increasing, the steady stream of data is collected by the sensor. It is very difficult to detect abnormal data effectively and accurately in a large amount of data.

- Because equipment is fitted with multiple sensors, a large number of isomerous data are generated. In the multiple monitoring data sequences, some have similar trend. The original method is no longer adapted to so many of the monitoring data sequence types.

Different sensors produce a large number of isomerous monitoring data while the same sensors produce a large number of isomorphic data. First, observe a group of monitoring data sequences from equipment, which will be used as a point of discussion. The group of monitoring data sequence from the equipment is displayed Figure 1. It can be observed that the curves of Figure 1(a), Figure 1(b) and Figure 1(c) are similar; the curves of Figure 1(d) and Figure 1(e) are similar; the curve of Figure 1(f) is not similar to the other curves. Thus, it can be seen that monitoring data sequence can be clustered by using correlation.

The group of monitoring data from the machine speed, pump speed of equipment, is displayed in Figure 4. Observation of the machine speed and pump speed in the normal operation of the equipment, the pump speed is increased with the increase of the machine speed, and reduced with the reduction of the machine speed. In the Figure 4, seeing a part of the green rectangle, the pump speed and the speed of the machine are in the normal range,



Notes: (a) Machine speed; (b) pump speed; (c) second hydraulic pressure; (d) third hydraulic pressure; (e) oil temperature; (f) first hydraulic pressure

Figure 1.
The group of monitoring data sequence from the equipment

but with the significant increase in the speed of the machine, the pump speed is not obvious. The correlation of the pump speed and the machine speed is abnormal, which can indicate the section of green rectangle is abnormal.

Isomerous data correlation is normal or a certain extent normal, if there is data in isomorphic data beyond its normal range, it may also appear abnormal. The correlations of data sequences compress several data sequences into a few numbers. The anomaly of data transience is possibly weakened. A set of monitoring data is derived from equipment's machine speed, pump speed, and driving device for lubricating oil temperature, which is shown in Figure 5. Observing the part of the green rectangle enclosed in the Figure 5, the isomerous data correlation is normal or a certain extent normal, but the temperature is beyond the normal range (60°C), which also can be identified as an anomaly about the part of the green rectangle.

In Figure 2, there are tendency charts from two data sequences of equipment. The middle part of the line is the normal scope of their respective. It can be observed that the first hydraulic pressure and second hydraulic pressure have a similar trend. But the two curves of the rectangular part are abnormal, and they are beyond the normal range. In Figure 2(a), the part of the green rectangle is enlarged in Figure 3. It can be observed that the first hydraulic pressure is beyond the normal range (34000-36000 KPa). We quantify the anomaly degrees of the single point according to the magnitude of the excess.

In this paper, we propose Multiple Group Correlation-based Anomaly Detection (MGCAD) which can quickly and accurately detect anomalies of equipment. In MGCAD, we first use correlation coefficient to cluster monitor data sequence. If there are more than one kind of monitoring data sequence in the class, we use latent correlation vector to quantify the latent correlation of multiple monitoring data sequences in the class, and get the normal range of each monitoring data sequence; if there is only one monitoring data sequence in the class, we can only obtain the normal range of the monitoring data sequence. Finally, we use Multiple Group Correlation Probability Models (MGCPM) to model latent correlation vectors (lcv) and abnormal degree of single points. Using MGCPM we can quickly and accurately detect the anomaly of the equipment.

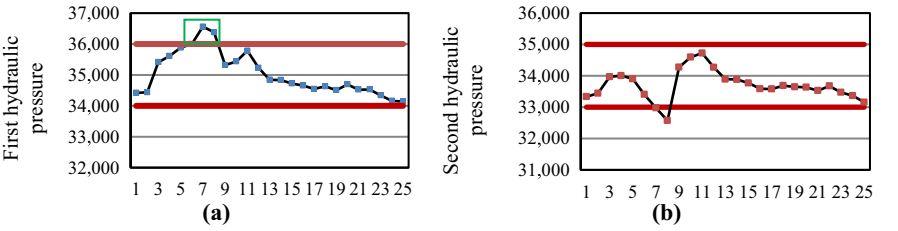


Figure 2.
There are tendency chart from two data sequences of equipment

Notes: (a) First hydraulic pressure; (b) second hydraulic pressure

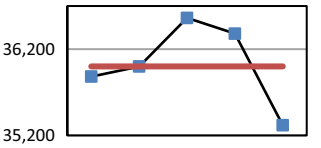


Figure 3.
Local chart of Figure 2(a)

We propose Multiple Group Correlation-based Anomaly Detection, which can quickly and accurately detect anomalies of equipment. Our main contributions are as follows:

- We use correlation to cluster all of the monitoring data sequences, and the sequence of monitoring data in each class is related to each other. We find out the normal range of each monitoring data sequence, according to the extent of the data sequence over or below the normal range, to quantify the anomaly degree of the single point.
- The correlation coefficient between the monitoring data sequence is used to show the correlation between the monitoring data, finally, the correlation of each monitoring data sequence with other monitoring data sequence is expressed as the square sum of each column vector element.
- We propose Multiple Group Correlation-based Anomaly Detection method about a large number of monitoring data sequences. We model the anomaly degree of single point and correlation about each class of data sequence using Multiple Group Correlation Probability Models, which makes the anomaly detection work very well. We reduce the dimension of the monitoring data sequence by clustering, and then we use the method of the respective detection.

The rest of the paper is shown below. Section 2 is related to work. In Section 3, we set the problem, show the outline of our method. Section 4 is a detailed introduction of our approach. Section 5 shows the test evaluation. Section 6 gives a summary of the full text.

2. Related work

To the best of our knowledge, there are varied abnormal detection methods, mainly as classification based, nearest neighbor based, clustering based, statistical and others. They are displayed in Figure 6:

- Classification-based. Classification based is an anomaly detection method, using classification rules to classify data into normal data and abnormal data, to carry out anomaly detection. Classification-based anomaly detection approaches consist of support vector machine-based approaches (Ma and Perkins, 2003) and neural network-based approaches (Schlechtingen and Santos, 2011).
- Nearest neighbor-based. Nearest neighbor-based anomaly detection method, using the distance between the data by which the normal data can be distributed in the dense region and the abnormal data away from their Nearest Neighbor. Nearest neighbor-based approaches consist of density-based (Pokrajac *et al.*, 2007) and distance-based.
- Cluster-based. Cluster-based anomaly detection method (Izakian and Pedrycz, 2013), which divides the data into multiple clusters, among them, normal data in a cluster and abnormal data do not belong to any cluster.
- Statistical. Statistical approaches consider normal data as the high probability region, and the abnormal data are in the low probability region. These approaches mainly consist of parametric (Aggarwal and Yu, 2008) and nonparametric approaches.

There are some existing methods for anomaly detection of multiple time series. Ding *et al.* (2014, 2016) proposed LCAD that only considers the correlation of the data, and no correlation is used for clustering. Zhang *et al.* (2009) proposed abnormal trends detection method for multiple data streams, but their method is very time-consuming.

All of the aforementioned methods focus on the isomorphic data, without considering the correlation between the monitoring data sequence. For a large number of monitoring data, it is not easy to detect abnormal. If only consider the correlation, there may be uncorrelated section in a large number of monitoring data sequence as shown in Figure 1. If only consider the correlation of the isomerous data, while ignoring the isomorphic data, the anomaly may not be detected as shown in Figure 4. In short, the existing methods cannot well solve the anomaly detection based on a large number of relevant time series.

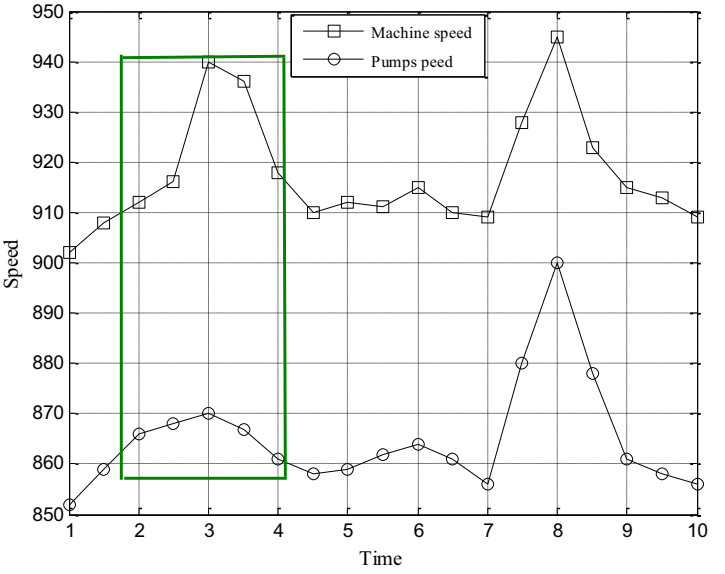


Figure 4.
The above line represents the collection of machine speed; the following line represents the Collection of pump speed

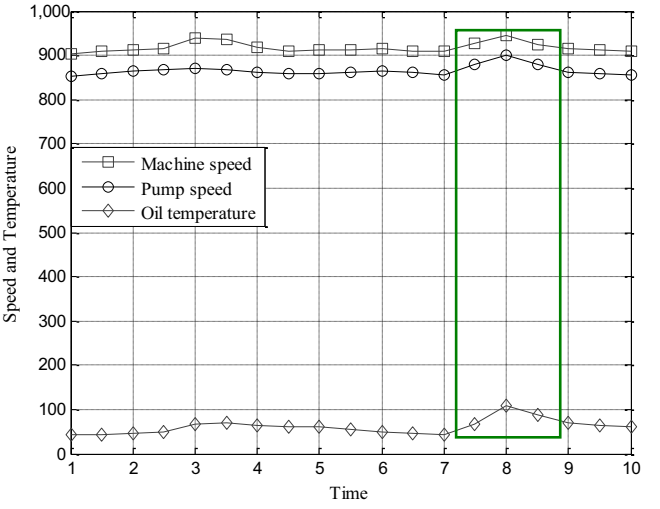


Figure 5.
The above line represents the collection of machine speed, the middle line represents the collection of pump speed, and the following line represents the collection of oil temperature

3. Problem settings and outline of proposed approach

3.1 Collection plan of monitoring data

A large number of equipment for the same type are given, defined as O = $\{E_1, E_2, E_3, \cdots, E_N\}$, where E_N representsthe N-th equipment. Each equipment is fitted with K sensors, which is expressed as S = $\{S_1, S_2, S_3, \cdots S_K\}$, where S_K represents the K-th sensor. Each equipment generates K types of monitoring data sequence, and the sequence of monitoring data for the N-th equipment is represented as $W^N = \left\{ W_1^N, W_2^N, W_3^N, \cdots, W_K^N, \right\}$, where $W_K^N = \left\{ V_K^N(1), V_K^N(2), V_K^N(3), \cdots, V_K^N(T) \right\}$ represents the K-th sensor monitoring data for the N-th equipment, and $V_K^N(T)$ represents the collected data at the T time.

3.2 Outline of proposed approach

We have N same types of equipment and each equipment is equipped with K sensors. Let this equipment to work with the sensor. Equipment E_i produced a total of L_i working cycle sequence groups; make $L = L_1 + L_2 + L_i + \cdots L_N$ this N equipment have produced a total of L working cycle sequence groups.

The MGCAD is mainly composed of six parts: Data pre-processing, latent correlation extraction, monitoring data sequence clustering, set the normal range of each data sequence, Multiple Group Correlation Probability Models training, and abnormal detection.

In the data pre-processing phase, our main job is to carry out data collation and cleaning, to maintain the same dimensions of the isomerous data in a work cycle, and to prepare for the second part of the job.

In the latent correlation extraction phase, extracting correlation from the data generated from the first part using correlation coefficients and sum of squares.

In the monitoring data sequence clustering phase, we use correlation coefficient to cluster all the monitoring data sequence. Each class of monitoring data sequence is related to each other, and the correlation of the monitoring data sequence between the classes is very small.

In the setting the normal range of each data sequence phase, we set the normal range of each data sequence according to observation data, access to information, consulting engineers.

In the multiple group correlation probability models training phase, we model the anomaly degree of single point and correlation about each class of data sequences using multiple group correlation probability models.

In the abnormal detection phase, we use the model of the fifth parts to carry out anomaly detection.

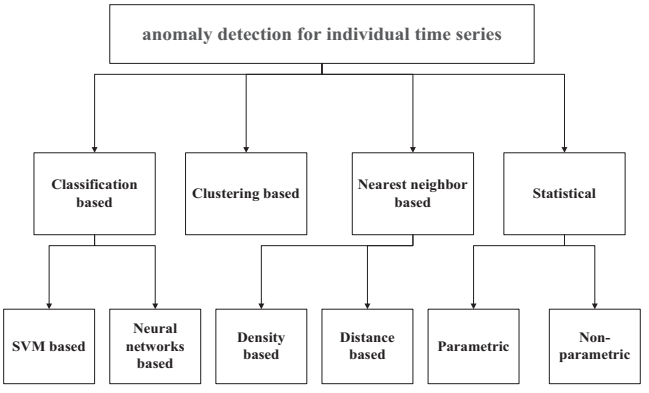


Figure 6.
Classification of the anomaly detection method for a single time series

4. MGCAD: our anomaly detection approach

4.1 Data pre-processing

Due to the frequency of different sensors to collect data is different and monitoring data may become dirty, to facilitate the extraction of latent correlations between isomerous monitoring data sequence, we need to reconstruct the data, so that monitoring data sequence to maintain the same dimension and to maximize the meaning of the original data in a working cycle.

Piecewise aggregate approximation (PAA) is a famous dimension reconstruction technology, which is widely used in data processing. As shown in Figure 7, after the processing of the PAA technology (Chakrabarti *et al.*, 2002; Faloutsos, 2000), the original data is eventually represented by 10 data.

In the l work cycle, the k monitoring data sequence $V_k^1 = v(1)_k^1, v(2)_k^1, \cdots, v(n)_k^1$ Which can be expressed as: $W_k^1 = \left\{w(1)_k^1, w(2)_k^1, \cdots, w(m)_k^1\right\}$, use formula:

$$W(i)_k^1 = \frac{m}{n} \sum_{j=\frac{n}{m}(i-1)+1}^{\frac{n}{m}i} v(j)_k^1 \tag{1}$$

4.2 Latent correlation extraction

First, we consider a work cycle, in a work cycle sequence group, there are K work cycle sequences, in this part, and we define the latent correlation between the work cycle sequences.

Let (X, Y) be a two dimensional random variable, and $\mathrm{Var(X)} = \sigma_X^2 > 0$, $\mathrm{Var(Y)} = \sigma_Y^2 > 0$, then said:

$$Corr(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{2}$$

Which is the correlation coefficient between X and Y.

If $0 < |Corr(X, Y)| < 1$, X and Y have a certain degree of linear relationship. $|Corr(X, Y)|$ is close to 1, and the linear degree is higher; $|Corr(X, Y)|$ is close to 0, then the linear degree is lower. However, the covariance is not that, if the covariance is small and the two standard deviations are small, the ratio is not necessarily very small (Fisz and Bartoszyński, 2018).

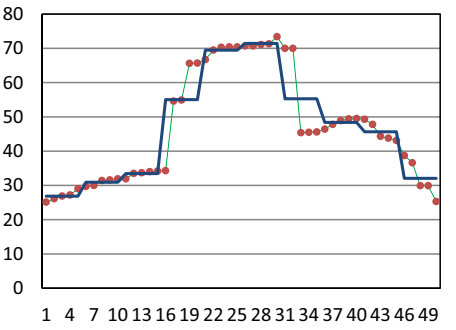


Figure 7.
The raw data is represented by the dotted line, the data through the PAA technology to reconstruct is indicated by real lines

In this paper, we use correlation coefficients to represent latent correlations between isomerous data sequences. In the l-th work cycle, We define a latent correlation coefficient matrix to measure the latent correlation of the l-th work cycle sequences, which is expressed as $LCCM^1$, Calculation formula:

$$\mathrm{LCCM}^l = \begin{pmatrix} C_{11}^l & \cdots & C_{1k}^l \\ \vdots & \ddots & \vdots \\ C_{k1}^l & \cdots & C_{kk}^l \end{pmatrix} \tag{3}$$

In $LCCM^1$ elements C_{ij}^1 is latent correlation parameter in between the i-th cycle sequence and the j-th work cycle sequence that is computed as:

$$C_{ij}^1 = \mathrm{Corr}^1(i,j) = \frac{\mathrm{Cov}^1(i,j)}{\sqrt{\mathrm{Var}(i)}\sqrt{\mathrm{Var}(j)}} = \frac{\mathrm{Cov}^1(i,j)}{\sigma_i \sigma_j} \tag{4}$$

We use correlation coefficient to extract the correlation between the monitoring data sequence of each work cycle, which is expressed as: $LCCM^{1,}\ LCCM^2 \cdots LCCM^L$

4.3 Monitoring data sequence clustering

K sensors are installed on one equipment, and each equipment can collect K types of monitoring data sequence. We use the correlation between monitoring data sequence to cluster the monitoring data sequence (Guha *et al.*, 1998; Kanungo *et al.*, 2002; Cacciari *et al.*, 2008). Monitoring data sequences of each class is related to each other, and the correlation of the monitoring data sequence between the classes is very small. We find a correlation coefficient matrix about monitoring data sequence of each working cycle, using the formula (4). Therefore, we obtain L correlation coefficient matrix: $LCCM^{1,}\ LCCM^2 \cdots LCCM^L$. Then, these L correlation coefficient matrix is used to calculate the average correlation coefficient matrix: $\overline{LCCM}$, Calculated as:

$$\overline{LCCM} = \frac{\left(LCCM^1 + LCCM^2 +, \cdots, LCCM^L\right)}{L} \tag{5}$$

We cluster monitoring data sequence according to equation (5) Clustering results are shown in Table I.

4.4 Set the normal range of each data sequence

We set the normal range of each data sequence according to observation data, access to information, and consulting engineers. In the modeling, the single point data can be more than or less than the normal range, which can help us quantify the anomaly degree of the

Table I. Clustering results of monitoring data sequences

First class	W1, W2, ...
Second class	W3,W5 ,W6
⋮	⋮
Last class	W4

single point. For example, through investigating data we know: when concrete pump truck in uninterrupted running time, oil temperature does not exceed 70 degrees Celsius, otherwise it should stop to test. Therefore, we design the oil temperature $< = 60$.

4.5 Multiple group correlation probability models

In Section 4.3, we have clustered the monitoring data sequences, and each class of the monitoring data sequence is related to each other, and the correlation of the monitoring data sequence between classes is very small. We use the method of the respective detection to model each class of data sequence. There are two cases after clustering. The number of species of monitoring data sequence in the first case is more than one, and the other is equal to one.

First, we consider the first case. As in Table I, second class has three kinds of monitoring data sequence: W_3, W_5, W_6. We extracted the three monitoring data sequence of L cycle: $W_3^1, W_3^2, \cdots, W_3^L$; $W_5^1, W_5^2, \cdots, W_5^L$; $W_6^1, W_6^2, \cdots, W_6^L$. Therefore, we can get L correlation coefficient matrix: $LCCM_2^1, LCCM_2^2, LCCM_2^3, \cdots, LCCM_2^L$. Where $LCCM_2^1$ is expressed as:

$$LCCM_2^J = \begin{pmatrix} C_{33}^1 & C_{35}^1 & C_{36}^1 \\ C_{53}^J & C_{55}^1 & C_{56}^1 \\ C_{63}^1 & C_{65}^1 & C_{66}^1 \end{pmatrix} \tag{6}$$

We find the sum of square of each column vector element of the l-th latent correlation coefficient matrix, that is:

$$\lambda_3^{\,1} = C_{33}^{1\ 2} + C_{53}^{1\ 2} + C_{63}^{1\ 2} \tag{7}$$

Expressed as $1cv_2^1 = \{\ \lambda_3^{\,1}, \lambda_5^{\,1}, \lambda_6^{\,1}\}$, in which element $\lambda_i^{\,1}$ is defined as latent correlation factors, it represents the correlation between the i-th work cycle sequence and the sequence of every other work cycle. $LCCM_2^1, LCCM_2^2, LCCM_2^3, \cdots, LCCM_2^L$ is expressed as: $1cv_2^1, 1cv_2^2, 1cv_2^3, \cdots, 1cv_2^L$.

$$\bigcup_{l=1}^{L} 1cv^1 = \{1cv^1, \cdots, 1cv^L\} = \begin{Bmatrix} \lambda_3^{\,1} & \cdots & \lambda_3^{\,l} & \cdots & \lambda_3^{\,L} \\ \lambda_5^{\,1} & \cdots & \lambda_5^{\,l} & \cdots & \lambda_5^{\,L} \\ \lambda_6^{\,1} & \cdots & \lambda_6^{\,l} & \cdots & \lambda_6^{\,L} \end{Bmatrix}_{3\times L} = \{\forall_3.\forall_5.\forall_6\}^T \tag{8}$$

Where $\forall_k = \left(\lambda_k^{\,1}, \cdots, \lambda_k^{\,l}, \cdots, \lambda_k^{\,L}\right)$.

Vector $\forall_k$ represents the k-th latent correlation factor vector, which reflects the latent correlation between the k-th monitoring data sequence and each sequence of other monitoring data sequence in each work cycle.

We assume that all the latent correlation factor λ of the vector $\forall_k$ is satisfied with a Gauss distribution. To verify our idea, we use IBM SPSS Statistics tool tests the results of our experimental data distribution.

We use the characteristics of the distribution of Gauss to carry out anomaly detection. We can get three Gaussian distributions, each Gaussian distribution corresponding to a group (μ, σ). Because our method uses: $\lambda_3^{\,1} = C_{33}^{1\ 2} + C_{53}^{1\ 2} + C_{63}^{1\ 2} \leq 3$. Then all anomalies

are on the left, and the closer to the left, the more abnormal. If the λ_k^1 is calculated and satisfying $\lambda_k^1 = \mu_k - 3\sigma_k$ then we think that the equipment is an anomaly in the l-th cycle.

It is possible that the correlation between isomerous data is normal or a certain extent normal, but only considering the correlation between the isomerous data, the abnormal monitoring results may be missing. The correlations of data sequences compress several data sequences into a few numbers. The anomaly of data transience is possibly weakened. The correlation of a few data sequences to a certain extent is normal. We must also consider abnormal degree of a single point. The method of amplifying the abnormal probability of λ_i by adding the anomaly factor that is to change λ_i. Expressed as:

$$W_k = \{V_k(1), V_k(2), V_k(3), \ldots, V_k(T)\} \tag{9}$$

$$r_k = 0 \ for \ t = 1 : T \begin{cases} \text{if} \quad V_k(t) > \beta_k^1 \quad \text{then} \quad r_k = r_k + (V_k(t) - \beta_k^1) * p\sigma_k \\ \text{if} \quad V_k(t) < \beta_k^2 \quad \text{then} \quad r_k = r_k + (\beta_k^2 - V_k(t)) * p\sigma_k \end{cases} \tag{10}$$

$$\begin{cases} \text{If} \quad \mu_k > \lambda_k \quad \text{then} \quad \lambda_k = \lambda_k - r_k \\ \text{if} \quad \mu_k < \lambda_k \quad \text{then} \quad \lambda_k = \lambda_k + r_k \end{cases} \tag{11}$$

Among them, λ_i is the k-th latent correlation factor of the test data, W_k indicates that the collected monitoring data sequence of the k-th sensor. β_k^1 as higher limiting value. β_k^2 as lower limiting value.

At this time, we can set up the model of second class of monitoring data sequence:

$$\begin{cases} \begin{cases} N_3(\forall_3; \mu_3; \sigma_3) \\ N_3(\forall_5; \mu_5; \sigma_5) \\ N_3(\forall_6; \mu_6; \sigma_6) \end{cases} \\ \begin{cases} if \ \ \mu_k > \lambda_k \ then \ \ \lambda_k = \lambda_k - r_k (k = 3, 5, 6) \\ if \ \ \mu_k < \lambda_k \ then \ \ \lambda_k = \lambda_k + r_k (k = 3, 5, 6) \end{cases} \end{cases} \tag{12}$$

Then we consider the second cases, as shown in Table I, last class has one monitoring data sequence: W4. At this point, we cannot use the correlation anomaly to detect the abnormal data, we can only consider the anomaly degree of the single point.

$$W_k = \{V_k(1), V_k(2), V_k(3), \ldots, V_k(T)\} \tag{13}$$

$$r_k = 0 \ for \ t = 1 : T \begin{cases} \text{if} \quad V_k(t) > \beta_k^1 \quad \text{then} \quad r_k = r_k + (V_k(t) - \beta_k^1) * p\sigma_k \\ \text{if} \quad V_k(t) < \beta_k^2 \quad \text{then} \quad r_k = r_k + (\beta_k^2 - V_k(t)) * p\sigma_k \end{cases} \tag{14}$$

$$\begin{cases} \text{If} \quad \mu_k > \mu \quad \text{then} \quad \mu = \mu - r_k \\ \text{if} \quad \mu_k < \mu \quad \text{then} \quad \mu = \mu + r_k \end{cases} \tag{15}$$

Among them, μ_k is the k-th mean value of the monitoring data, σ_k is the k-th standard deviation of the monitoring data, W_k indicates that the collected monitoring data sequence of the k-th sensor. β_k^1 as higher limiting value. β_k^2as lower limiting value.

Above two aspects, we build a model that is multiple group correlation probability models (MGCPM). MGCPM is expressed as:

$$\begin{cases} \textit{The first class} \begin{cases} \begin{cases} N_g(\forall_g; \mu_g, \sigma_g) \\ \vdots \\ N_h(\forall_h; \mu_h, \sigma_h) \end{cases} \\ \begin{cases} \textit{if } \mu_\kappa > \lambda_k \textit{ then } \lambda_k = \lambda_k - r_k \\ \textit{if } \mu_\kappa < \lambda_k \textit{ then } \lambda_k = \lambda_k + r_k \end{cases} \end{cases} \\ \vdots \\ \textit{The s class} \begin{cases} \begin{cases} N_x(\forall_x; \mu_x, \sigma_x) \\ \vdots \\ N_y(\forall_y; \mu_y, \sigma_y) \end{cases} \\ \begin{cases} \textit{if } \mu_\kappa > \lambda_k \textit{ then } \lambda_k = \lambda_k - r_k \\ \textit{if } \mu_\kappa < \lambda_k \textit{ then } \lambda_k = \lambda_k + r_k \end{cases} \end{cases} \\ \vdots \\ \textit{The last class} \begin{cases} \textit{if } \mu_k > \mu \textit{ then } \mu = \mu - r_k \\ \textit{if } \mu_k < \mu \textit{ then } \mu = \mu + r_k \end{cases} \end{cases} \tag{16}$$

4.6 Abnormal detection

The last of MGCAD is detecting anomaly Based on the MGCPM. First, the monitoring data sequence tested would we be segmented into a plurality of work cycles, and then we can get dimension reconstruction after PAA. By the model MGCPM, the final LCV $=\{\lambda_1, \lambda_1, \cdots, \lambda_K)$ of each work cycle. We define the anomaly detection equation for LCV:

$$F(LCV; \mu_k, \sigma_k, \alpha) = \begin{cases} 0 & \text{if } \lambda_k > \mu_k - \alpha\sigma_k \\ 1 & \text{if } \lambda_k \leq \mu_k - \alpha\sigma_k \end{cases} \tag{17}$$

In the anomaly detection equation, the parameters μ_k are the mean value of $\forall_k$ and σ_k are the standard deviation of $\forall_k$. In the experiment, we set the parameters α equal to 3. In the equation, the input is α, and the output is kBoolean types (where 1 represents the equipment in this cycle is abnormal, and the 0 represents the equipment in this cycle is normal).If there is a 1 in the output of k Boolean values, the equipment is an anomaly in this cycle.

5. Experimental evaluations

5.1 Data preparation

To display the anomaly detection method effectively, we carried out scientific and reasonable experiment. Our experimental data is to use a truthful data set as a seed and use the MATLAB tool to simulate the synthesis. To a large extent, it represents the real data.

Our data set consists of six parts: machine speed, pump speed the temperature of the transmission equipment lubricating oil between the machine and the pump, first hydraulic pressure, second hydraulic pressure, third hydraulic pressure, these six types of data were collected by six kinds of sensors. Our data set contains 1,000 work cycles, and each cycle contains six types of monitoring data sequence, as shown in Table II.

5.2 Experiment

In this section, we will demonstrate the performance of our method on the above data set. Using the above data set we got Multiple Group Correlation Probability Models. Six responding monitoring data sequences are collected. In addition, each monitoring data sequence is divided into ten work cycles. We use our model and anomaly detection equation, respectively, to detect these ten work cycles. In Table III, it showed that six kinds of monitoring data were clustered into three classes.

5.3 Experimental results and analysis

We use our anomaly detection method to detect the results of the above ten work cycles in turn, as shown in Table IV. Using the correlation to detect the abnormal data without taking the correlation to cluster in to account may lead to unexpected results. If you only consider the correlation, and ignore the abnormal degree of single point, the detection results will be missing. The LCAD method only considers the correlation of the data, and no correlation is

Table II. Description of experimental data

Data type	Monitoring data sequence
Machine speed (r/min)	1000
Pump speed(r/min)	1000
Oil temperature (° C)	1000
First hydraulic pressure (Kpa)	1000
Second hydraulic pressure (Kpa)	1000
Third hydraulic pressure (Kpa)	1000

Table III. Description of clustering results

The 1st class	Third hydraulic pressure
The 2nd class	Machine speed, Pump speed, Oil temperature
The 3rd class	First hydraulic pressure, Second hydraulic pressure

Table IV. Using our method detects the results of the above ten cycle data

Cycle	First class	Second class	Third class	MGCAD	LCAD	EUC-KNN
The 1-th cycle	normal	abnormal	abnormal	abnormal	abnormal	abnormal
The 2-th cycle	normal	normal	normal	normal	normal	normal
The 3-th cycle	normal	normal	normal	normal	normal	normal
The 4-th cycle	normal	abnormal	normal	abnormal	abnormal	normal
The 5-th cycle	abnormal	normal	normal	abnormal	normal	abnormal
The 6-th cycle	normal	abnormal	normal	abnormal	normal	normal
The 7-th cycle	normal	normal	normal	normal	normal	normal
The 8-th cycle	normal	normal	normal	normal	normal	normal
The 9-th cycle	normal	normal	normal	normal	normal	normal
The 10-th cycle	normal	normal	normal	normal	normal	normal

used for clustering. We use the LCAD method to model and detect our data sets. Experimental results are shown in Table IV. After we calculate the *lcv*of each cycle, we use the Euc-KNN method (Peterson, 2009) to detect the anomaly, and the experimental results are in Table IV. Euc-KNN method is a classifier based on k-nearest neighbor and Euclidean distance.

In the sixth work cycle, the oil temperature is higher than 60 degrees Celsius, which appears two times. And machine speed increased significantly; pump speed was not obvious. In the fifth work cycle, first hydraulic pressure and second hydraulic pressure have multiple single point anomalies. Through the experimental results, we can see that our method is time saving and accurate.

6. Summary

In CrowdIntell Network, the smart articles such as sensors can assist to realize a comprehensive, real, correct and synchronous projection from the real world to the cyber space, and ultimately realize the transaction between Digital-selfs. To better understand the operation status of equipment, it is very important for equipment managers to detect abnormal equipment quickly and accurately. At the same time, the faster and more accurately the sensors can detect the abnormal data, the higher its intelligent level is. In this paper, we have employed MGCPM to model the latent correlation between the monitoring data sequences and the anomaly degree of the single point. Extensive experimental results show that our method has better performance than previous methods.

References

Aggarwal, C.C. and Yu, P.S. (2008), "Outlier detection with uncertain data", *Proceedings of the 2008 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, pp. 483-493.

Atzori, L., Iera, A. and Morabito, G. (2010), "The internet of things: a survey", *Computer Networks*, Vol. 54 No. 15, pp. 2787-2805.

Cacciari, M., Salam, G.P. and Soyez, G. (2008), "The anti-k_t jet clustering algorithm", *Journal of High Energy Physics*, Vol. 4 No. 4, pp. 403-410.

Chai, Y., Miao, C., Sun, B., Zheng, Y. and Li, Q. (2017), "Crowd science and engineering: concept and research framework", *International Journal of Crowd Science*, Vol. 1 No. 1, pp. 2-8.

Chakrabarti, K., Keogh, E., Mehrotra, S. and Pazzani, M. (2002), "Locally adaptive dimensionality reduction for indexing large time series databases", *ACM Transactions on Database Systems (TODS)*, Vol. 27 No. 2, pp. 188-228.

Ding, J., Liu, Y., Zhang, L. and Wang, J. (2014), "LCAD: a correlation based abnormal pattern detection approach for large amount of monitor data" in *Asia-Pacific Web Conference*, Springer, pp. 550-558.

Ding, J., Liu, Y., Zhang, L., Wang, J. and Liu, Y. (2016), "An anomaly detection approach for multiple monitoring data series based on latent correlation probabilistic model", *Applied Intelligence*, Vol. 44 No. 2, pp. 340-361.

Faloutsos, C. (2000), "Fast time sequence indexing for arbitrary Lp norms", *Proc International Conference on Vldb*.

Fisz, M. and Bartoszyński, R. (2018), *Probability Theory and Mathematical Statistics*, J. wiley.

Guha, S., Rastogi, R. and Shim, K. (1998), "CURE:an efficient clustering algorithm for large databases", *Information Systems*, Vol. 26 No. 1, pp. 35-58.

Izakian, H. and Pedrycz, W. (2013), "Anomaly detection in time series data using a fuzzy c-means clustering", *Ifsa World Congress and Nafips Meeting, IEEE*.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y. (2002), "An efficient k-means clustering algorithm: analysis and implementation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24 No. 7, pp. 881-892.

Ma, J. and Perkins, S. (2003), "Online novelty detection on temporal sequences", *Acm sigkdd International Conference on Knowledge Discovery and Data Mining*.

Peterson, L. (2009), "K-nearest neighbor", *Scholarpedia*, Vol. 4 No. 2, p. 1883.

Pokrajac, D., Lazarevic, A. and Latecki, L.J. (2007), "Incremental local outlier detection for data streams", *IEEE Symposium on Computational Intelligence and Data Mining*.

Schlechtingen, M. and Santos, I.F. (2011), "Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection", *Mechanical Systems and Signal Processing*, Vol. 25 No. 5, pp. 1849-1875.

Wang, S., Cui, L., Liu, L., Lu, X. and Li, Q. (2019), "Projecting real world into CrowdIntell network: a methodology", *International Journal of Crowd Science*.

Zhang, C., Weng, N., Chang, J. and Zhou, A. (2009), "Detecting abnormal trend evolution over multiple data streams", in *Advances in Data and Web Management*, Springer, Berlin, pp. 285-296.

Corresponding author
Lizhen Cui can be contacted at: clz@sdu.edu.cn

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**