IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON URBAN COMPUTING AND INTELLIGENCE

Received March 13, 2019, accepted April 2, 2019, date of publication April 11, 2019, date of current version April 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2909967

Anomaly Detection Approach for Urban Sensing Based on Credibility and Time-Series Analysis Optimization Model

HONG ZHANG[1,2] AND ZHANMING LI[1]

[1]College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China
[2]Networks and Information Administration Center, Gansu University of Chinese Medicine, Lanzhou 730000, China

Corresponding author: Hong Zhang (532337136@qq.com)

This work was supported in part by the Project of Gansu Province for Industrial and Information Development under Grant 23051358, and in part by the Project of Gansu Province for Guiding Scientific and Technological Innovation and Development under Grant 2018ZX-05.

ABSTRACT Urban sensor networks often consist of a large number of low-cost sensor nodes. Due to the constrained resource devices and hazardous deployment, urban sensing is vulnerable to interference and destruction of external factors or the impact of external environmental emergencies. Abnormal data, outliers, or anomalies have affected the utility in various domains seriously. Timely and accurate detection of unexpected events, monitoring of network performance, and anomaly detection of data flow are of great significance to improve the decision-making ability of the system. In this paper, we propose an anomaly detection method for urban sensing based on sequential data and credibility. First, based on Bayesian methods, a reputation model is established for the selection of credible sample points. Second, aiming at the problem that the threshold range is difficult to determine in the traditional method, the pivot quantity is defined by using the median of the credible sample, and the confidence interval can be estimated to quantify the deviation degree of the sensor data. Finally, an anomaly data identification and source verification approach is proposed to distinguish errors and events accurately. The evaluation results on both the detection rate and the false positive rate demonstrate a better performance of our approach than the other existing methods.

INDEX TERMS Anomaly detection, urban sensing, credibility, spatio-temporal correlation, smart city.

I. INTRODUCTION

By utilizing information about city-scale processes extracted from heterogeneous data sources, the goal of smart cities is to improve the quality of life of citizens collected from city-wide deployment [1]. As one of the important supporting technologies of smart city, the Internet of Things (IoT) have been rapidly developed and widely applied in the fields of military, health care, environmental monitoring and manufacturing. Quantity of data captured by urban sensing are considered to contain highly useful and valuable information. However, for a variety of reasons, received sensor data often appear abnormal. Therefore, effective anomaly detection methods are required to guarantee the quality of data collected by those sensor nodes [2].

Accurate analysis and decision-making rely on the quality of sensor data as well as on additional information and context [3]. With the increase of network scale, anomaly detection in sensor networks becomes more and more critical, especially for some emergencies, such as chemical leaks and fires, timely warning and response mechanisms are often needed imminently [4]. It puts forward high requirements for reliability and accuracy of data collected by sensor nodes. However, due to the uncertainty of the environmental monitoring and the limited resources of sensor nodes, they are vulnerable to interference and destruction of external factors or the impact of external noise and obvious error. The quality of data cannot be effectively guaranteed, and some samples may be likely to deviate significantly from the actual features. This is one of the main barriers to urban sensing, because it is difficult to distinguish the anomalies caused by structural damage from those related to incorrect data. Those unreliable and inaccurate data can easily devour the benefits of the intelligent systems with high precision control demands in terms of usage and performances. The detection of incorrect data requires expertise and is very time-consuming.

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

The main reasons for generating abnormal data include: specific events occur in the wanted region, device fault or energy exhaustion results in node's failure, and the influence of external factors [5], [6]. For example, when forest fires occur, the temperature readings of the sensors will increase significantly. The illumination intensity of sensor nodes in shaded areas is significantly lower than that of nodes directly exposed to sunlight. Comparatively, outliers caused by events usually tend to demonstrate an extremely smaller probability of occurrence rather than the erroneous data. The sensing data from densely deployed nodes usually exhibit spatial-temporal correlated, which can be exploited by outliers' identification techniques. Existing spatial outlier detection mechanisms can be divided into the following steps [7], [8]: Firstly, the minimum neighborhood number is used to determine the neighborhood relationship based on spatial attributes. Secondly, spatial attributes will be applied to estimate the degree of outliers and to capture the differences between objects and their spatial neighborhoods. Finally, the objects with maximum spatial metrics are output as outliers. It is necessary to exchange information with neighbor nodes, which results in additional communication and accelerate the energy consumption of sensor nodes.

The rest of this paper is organized as follows. Section 2 discusses related work in the field of data outlier detection for urban sensing. In Section 3, we present the formal statement of the problem we address. Section 4 is dedicated to our approach for outlier detection. We then illustrate Simulation results in Section 5. A summary and conclusions are presented in Section 6.

II. RELATED WORKS

In order to improve the accuracy of data in urban sensing, anomaly detection is particularly important. Traditional event detection methods are mainly used to distinguish whether events or errors occur among the sensor nodes. Outlier detection is becoming critically important. In the literature, most of the outlier detection techniques focus on differentiating between outliers or events. Those outlier detection solutions for urban sensing can be categorized into statistical-based, nearest neighbor-based and clustering-based approaches.

Palpanas *et al*. [9] proposed a method for on-line detection of anomalous data by using nuclear density estimation. Based on the analysis of the difference between the current data in the sliding window and the prior model, it can identify anomalous data and does not require prior knowledge of data distribution. Wei and Li [10] proposed an outlier detection method based on reduction strategy and support vector data description. The decision-making model is updated adaptively based on data distribution density criterion and time correlation of data stream. Chen *et al*. [11] proposed a new definition of the traditional distance metrics by considering neighborhood information.

According to the characteristics of collected data, Fei and Li [12] presented an improved k-means algorithm to employ Euclidean distance as an indicator to estimate anomalous sensor data. Bi and Li [13] proposed an energy-efficient Top-k query algorithm in wireless sensor networks. The algorithm filters and ranks data sets according to the anomalous degree of observations, and selects k values of maximum or minimum data to users as well as their location. Samparthi and Verma [14] designed a renewal strategy with nuclear density estimation to renovate the data distribution model. Subramaniam *et al*. [15] proposed an online outlier detection in sensor data using non-parametric models, which can be applied for real-time identification for events of interest. Yi *et al*. [16] compare three boundary detection methods to retain critical samples for two-class supervised outlier detection, including nearest neighbors' distribution, relative density degree, and local geometrical information. In three boundary detection methods, the nearest neighbors' distribution is more suitable than others. Pan *et al*. [17] proposed an anomaly data detection method based on data set density. According to the historical samples and predicted values, the anomaly data can be analyzed ahead of time. This method can achieve real-time detection effectively, but the time complexity of the algorithm is obviously too high especially for dealing with high-dimensional data. Lee and Choi [18] proposed a hypothesis-based mathematical statistical method to distinguish the observations with positive or negative gain effects on the integrity and consistency of data sets as normal or abnormal data, respectively. However, this method has some shortcomings in real-time due to its centralized processing. Beggel *et al*. [19] proposed a novel method for detecting anomalous time series in the scenario where the training set contains no labels and is contaminated with an unknown amount of anomalies. However, requires recurring short temporal patterns to indicate normal behavior or class membership, and a smooth behavior of the time series. Wang *et al*. [20] presented MCRT, a multichannel real-time communication protocol that utilizes both multiple channels and transmission power adaptation to achieve real-time communications.

Sensor data can be viewed as a large volume of real-valued observations, and the characteristics depend on the attributes of their spatial and temporal correlations. Moreover, spatial and temporal correlation with the collected data also can play an important role in dealing with anomaly detection of sensor data. Considering that special events often occurs in spatial and temporal correlation and sensor faults are relatively independent, Krishnamachari and Iyengar [21] proposed a Bayesian fault identification algorithm, which employs message exchange among sensor nodes to calculate the probability of events. This method is clustering-based and it is difficult to identify and distinguish the types and causes of errors effectively. By using statistical comparisons to analyze the observations at event boundary, Li *et al*. [22] proposed a distributed fault-tolerant event boundary detection algorithm. The method has low complexity and high fault-tolerance, but it depends on the sufficient density of sensor networks. Lee and Kim [23] proposed a method to calculate the probability of estimated errors in real time based on the

difference between the measured values and their mathematical expectations, and to distinguish the types of anomalies according to the spatial-temporal correlation. However, this approach is not suitable for high-dimensional data sets. Xie and Chen [2] proposed an algorithm for sensor data anomaly detection based on principal statistical analysis and Bayesian networks. From the result of comparing with other traditional anomaly detection algorithms it can be seen that our algorithm can improve the precision of anomaly detection while ensuring the result of recall. Ren *et al*. [24] introduced a fault-tolerant event monitoring mechanism, which can effectively estimate the occurrence range and detection boundary of abnormal events and has a certain fault-tolerant ability. Zhang *et al*. [25] presents a monitoring framework to extract the common features and construct local monitoring statistics.

According to the temporal-spatial correlation of events, Zhang *et al*. [26] presented a fault-tolerant detection algorithm. By constructing a fusion tree distributed, each node sends the collected data to the corresponding nearest root node, thus achieving robust fault-tolerant detection of single/multiple events. Cao *et al*. [27] proposed a distributed fault-tolerant algorithm for event detection based on the temporal and spatial correlation of events. The algorithm uses the statistical characteristics of the event random process and analysis the coincidence degree of the data with time series. However, with the increase of error nodes, the event detection rate also decreases.

In this paper, an anomaly detection method for urban sensing based on credibility and time-series analysis is proposed. By exploiting the observations' distribution, an improved confidence interval is designed for outlier detection. The concept of cumulative degree of discrepancy is proposed to diagnosis the source of anomalies based on statistical methodology.

III. PROPOSED METHODOLOGY

A. OVERVIEW

Generally, when the data perceived by the sensor node deviates from the actual features several times in a row, it can be considered that anomalous data arise in certain equipment units. However, there are many reasons for the anomalous data generated in urban sensing [28]. The event nodes caused by special events in the region are called event nodes. Moreover, the nodes that fail to collect normal data due to their own faults or external attacks are identified as fault nodes [29]. The sensors are deployed in a homogeneous environment, in which the measurements taken have the same unknown distribution. All the sensor nodes are time synchronized. Note that, there are several time synchronization algorithms available for sensor networks that can be utilized for this purpose [30]. In addition, it is possible to make errors between the data and the actual values due to the influence of external factors in the sampling process. Therefore, the analysis about the dynamic streaming nature of sensing data is significant to build the normal reference model identify new data point as normal or outliers. However, we are also interested in cases where the majority of measurements at a sensor node are anomalous in comparison to other nodes in the network. In practice, when including a time dimension, the concept of spatial process can be extended to the spatio-temporal case. It will be overcome the computational complexity of non-separable models, some simplifications are introduced. For example, under the separability hypothesis the space-time covariance function is decomposed into the sum of a purely spatial and a purely temporal term. In this case, the temporal evolution could be introduced assuming that the spatial process evolves in time following an autoregressive dynamics [31].

As a statistical based approach, the spatio-temporal correlation happens when the nature of the collected data has both spatial and temporal correlations, nodes close geographically have the same reading that is similar to the previous one. In this case, solutions that use both correlations can take advantage of the nature of the detected event to decrease the number of reported data [32]. Of course, the spatio-temporal correlations also can determine the presence of an outlier by a sudden change in the data distribution. Outlier data often demonstrates historical correlation, and the length of sampling period has a greater impact on this correlation. In addition, under the condition of dense deployment, the readings collected by any sensor node and its neighbors also can show a certain degree of approximation. Therefore, the spatio-temporal correlation of sensor data flow provides a theoretical basis for anomaly detection.

In traditional threshold methods, measurement above the critical threshold will be considered an outlier. However, the threshold range is difficult to determine, and inappropriate threshold will cause false alarm or missed alarm. Data being collected often have different characteristics in different deployment environments. For example, the system operates over a broad temperature range, even large variation in data stream should be considered normal. Nevertheless, in some stable environments, small fluctuations may be identified as abnormal events. Accordingly, the actual measurements from each node that lie in or outside the confidence interval can be applied to deal with the problem of outlier detection.

B. NODES'S CREDIBILITY

Due to environmental changes or error comings from hardware with unreliability, faulty Sensor readings will be produced easily thus affecting anomaly detection. Therefore, it is necessary to quantify the reliability of sensor nodes [33]. By choosing the samples of the nodes with high confidence, the subsequent confidence intervals can be calculated. The data measurements above the above expected interval will be considered as outliers, and the fluctuation interval of total samples even in a small range can be estimated.

In this paper, the trend of current monitoring readings relative to historical data is analyzed to determine whether the current monitoring data is credible or not. Generally, when the variance of data is small, the data should be stable.

Suppose that sensor nodes have limited storage capacity and can store up to k data. Let $x^{(1)}, x^{(2)}, \cdots, x^{(k)}$ denote a sensor observation at time $t-k, t-k+1, \cdots, t-1$, respectively. Hence, the average value of the k samples is $\bar{x}_{t-1} = \sum_{i=1}^{k} \frac{x^{(i)}}{k}$, and the variance can be given as

$$\sigma_{t-1}^2 = \frac{\sum_{i=1}^{k} (x^{(i)} - \bar{x}^{(t-1)})}{k} \tag{1}$$

Let $x^{(t)}$ denote the observation at time t, the mean value of all samples can be represented as $\bar{x}_{t-1} = (x_2 + x_3 + \cdots + x_k + x_t)/k$, and the corresponding variance is

$$\sigma_t^2 = \frac{\sum_{i=2}^{k} (x^{(i)} - \bar{x}^{(t)})^2 + (x^{(i)} - \bar{x}^{(t-1)})^2}{k} \tag{2}$$

If $\left|\sigma_t^2 - \sigma_{t-1}^2\right| < \sigma$ for given parameter $\sigma > 0$, the variance varies slightly and the current observation may be similar to historical samples. Otherwise, it indicates that the current data has a sudden change. However, it is easy to ignore sensor malfunction by variance analysis to identify outliers. For example, a sensor node fails at a time and its monitoring data increases abnormally, and then the subsequent data will show certain deviation from normal level. However, $\left|\sigma_t^2 - \sigma_{t-1}^2\right|$ may gradually decrease or even be less than σ over time, and it results in an increase in term of credibility. Therefore, the variance and mean should be considered as both factors to measure the node's credibility.

For densely deployed sensor networks, there is a certain similarity between the data collected by the geographically adjacent nodes [34]. Therefore, the data correlation of the neighbor nodes can be implemented to measure the sensor's credibility. In order to save energy, each node only interacts directly with its neighbors within the communication radius. In this paper, Beta probability density function is used to evaluate the reputation of nodes dynamically. As a conjugate distribution of binomial distribution, Beta distribution is appropriate to calculate the posterior probability of binary events [35].

The reliability parameters α and β are defined for each sensor node, and the acceptable and unacceptable times of data deviation collected by each sensor node will be counted separately. If $\left|\sigma_t^2(i) - \sigma_{t-1}^2(i)\right| < \sigma$ and $x_i^{(t)} - \bar{x}_i^{(t-1)} < d$, α plus 1. otherwise, β will plus 1.

The Beta distribution has two parameters of (α, β), and gamma function Γ can be used to represent the Beta distribution $f(\zeta|\alpha, \beta)$ as follows: $\zeta \in [0, 1], \alpha > 0, \beta > 0$. Then, the probability expectation of Beta distribution can be given as

$$E(\zeta) = \frac{\alpha}{\alpha + \beta} \tag{3}$$

with the probability density function:

$$\begin{aligned} f(\zeta|\alpha, \beta) &= \frac{\chi^{\alpha-1}(1-\zeta)^{\beta-1}}{\int_0^1 \zeta^{\alpha-1}(1-\zeta)^{\beta-1} dp} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \zeta^{\alpha-1}(1-\zeta)^{\beta-1}. \end{aligned}$$

Assuming that in a sampling period p denotes the number of times that data deviations can be accepted and q denotes the number of times that data deviations exceed the deviation. Then, the posterior distribution still obeys the Beta distribution after $p+q$ events, and the parameter (α, β) will satisfy with $\alpha = p+1, \beta = q+1$.

The recommended credibility T_{ij} will be calculated as:

$$T_{ij} = E\left(Beta(p+1, q+1)\right) = \frac{p+1}{p+q+2} \tag{4}$$

Let $X \sim beta(\alpha_{ij}+1, \beta_{ij}+1)$ represent that the prior probability X of the credibility distribution of node i obeys the Beta distribution with respect to node j. In the next sampling period, the distribution of node's credibility will still obey Beta distribution after performing $p+q$ events, and the new Beta distributed random variable $X' \sim beta(\alpha_{ij}+p+1, \beta_{ij}+q+1)$.

During the course of urban sensing, the collected data by sensor nodes will record the variation of the surrounding environment with time and demonstrate the temporal correlation between previous observations and the current data [36]. That is to say, the event will last for a period of time after it happens, and the characteristics of the event show some superficial characteristics over time. Therefore, time forgetting factor θ is introduced to adjust the impact of historical reputation on recent credibility, and then the expression of direct credibility updating can be written as

$$X' \sim beta(\alpha_{ij}\theta + s + 1, \beta_{ij}\theta + f + 1) \tag{5}$$

The parameters of Beta distribution can be calculated recursively, and after occurrence of $p+q$ events, the parameters will be calculated as

$$\begin{cases} \alpha'_{ij} = \alpha_{ij}\theta + p \\ \beta'_{ij} = \beta_{ij}\theta + q \end{cases} \tag{6}$$

Therefore, the recommended credibility from node j will be modified to

$$T_{ij} = E\left(Beta(\alpha_{ij}+1, \beta_{ij}+1)\right) = \frac{\alpha'_{ij}+1}{\alpha'_{ij}+\beta'_{ij}+2} \tag{7}$$

Therefore, the overall credibility of the observations of node i can be estimated as:

$$DT_j = \frac{1}{|Neighbor(j)|} \sum_{i \in Neighbor(j)}^{T_{ij}} \tag{8}$$

where $Neighbor(j)$ denotes the neighbor nodes set of node j, and $|Neighbor(j)|$ is the number of member nodes.

Therefore, after a certain accumulation of sampling, the credibility of each node can be obtained. The threshold of credibility may be set according to the specific requirements of outlier accuracy.

C. CONFIDENCE INTERVAL

Using trusted samples, the mean square deviation of samples is calculated. Then, according to the sample mean, sample mean square deviation and given confidence level, the confidence interval can be obtained. Consequently, the observations fall outside the confidence interval will be labeled as suspicious outliers for further process.

Firstly, the samples of n credible nodes at a given time point constitutes as a set $Y = \{y_1, y_2, \cdots, y_n\}$, and $\tilde{y}$ denotes the median of samples. The median instead of the mean is employed to estimate the confidence interval, which can minimize the influence of the outliers. When the variance is unknown, the confidence interval of the median can be approximated by mean square deviation [37].

Suppose the density function of population Y be $f(y)$, and y_p^* is the p-th quantile. $f(y)$ can keep continuity at y_p^* and $f(y_p^*) > 0$. Hence, when the size of samples is large enough, the asymptotic distribution of the sample p-th quantile can be given as

$$m_p \xrightarrow{L} \mathcal{N}\left(y_p^*, \frac{p(1-p)}{f^2(y_p^*)}\right) \tag{9}$$

Since $y_1, y_2, \cdots, y_n$ comes from a uniformly distributed population $Y \sim U(0, \varphi)$ and φ denotes the general parameter, we can deduce that

$$2\sqrt{n}\left(\frac{\tilde{y}}{\varphi} - \frac{1}{2}\right) \xrightarrow{L} \mathcal{N}(0, 1) \tag{10}$$

It can be observed that in large sample situation, $2\sqrt{n}\left(\frac{\tilde{y}}{\varphi} - \frac{1}{2}\right)$ can be used as pivot. For given λ $(0 < \lambda < 1)$, by using the quantile $u_{1-\frac{\lambda}{2}}$ with lower level $1 - \frac{\lambda}{2}$ of the standard normal distribution, it can be formally formulated as

$$P\left\{2\sqrt{n}\left|\frac{\hat{y}}{\varphi} - \frac{1}{2}\right| \le u_{1-\frac{\lambda}{2}}\right\} \approx 1 - \lambda \tag{11}$$

Therefore, the approximate confidence interval of confidence degree $1 - \lambda$ with parameter φ can be obtained as:

$$\left[\frac{2\tilde{y}}{1 + u_{1-\frac{\lambda}{2}}/\sqrt{n}}, \frac{2\tilde{y}}{1 - u_{1-\frac{\lambda}{2}}/\sqrt{n}}\right] \tag{12}$$

Obviously, the accuracy of interval estimation is greatly affected by the number of samples and confidence level. The quantile can resist the interference of outlier data. Especially, the closer the quantile is to the median of the samples, the effect is more obvious.

Next, by using time-series analysis, outlier identification and anomaly source verification will be conducted. After detecting the anomalous data points, it is necessary to distinguish errors and events accurately.

IV. OUTLIER DETECTION

A. CUMULATIVE DEGREE OF DISCREPANCY

As discussed above, the observations differ significantly from the measured value and deviate from the upper or lower bounds, which should be identified as an outlier. On the other hand, as the sensor itself fails, e.g., energy exhaustion or damage cannot work properly, the readings may be generated continuously as same as the previous in different sampling times, i. e., $x_i^{(t)} = x_i^{(t-1)}$. In this paper, the two cases mentioned above should be regarded as the criteria for identifying outlying sensors.

Let $P_i(t)$ denote the probability of the observations from node i being detected as outliers at time period t. $P_i(t)$ should be a cumulative value and reflect the occurrence of anomalous data in a consecutive time sequence. We use statistic w to record the number of abnormalities that may occur in data stream. If the reading satisfies the judgment condition at several sampling times continuously, w will increase and appear exponential relationship to $P_i(t)$. Otherwise, if $x_i^{(t)}$ does not meet all the condition, w,$P_i(t-1)$ and $P_i(t)$ should return to 0. Therefore, the expression of $P_i(t)$ can be written as:

$$P_i(t) = \frac{1 - P_i(t-1)e^{-w}}{1 + P_i(t-1)e^{-w}} \tag{13}$$

By taking into account of the difference of data sets and the robustness of the method, it is not appropriate to detect abnormal data only by confidence intervals. Spatial correlation should be exploited to verify the origin of anomalies based on the statistical characteristics of data sets. The consistency degree $R_i(t)$ is defined to measure the spatial similarity between the observation node i and those of all adjacent trusted samples at time t as

$$R_i(t) = \sum_{j=1}^{n} r_{i,j}(t)/n, \tag{14}$$

and the nearness degree between the sensor i and j can be calculated as

$$r_{i,j} = \frac{\min_{1 \le j \le n}\{x_i^{(t)}, x_j^{(t)}\}}{\max_{1 \le j \le n}\{x_i^{(t)}, x_j^{(t)}\}} \tag{15}$$

Consequently, for sample point $x_i^{(t)}$, the degree of discrepancy with confidence interval can be obtained by following function:

$$g(x_i, t) = \begin{cases} R_i(t)\dfrac{c_{\min} - x_i^{(t)}}{c_{\max} - c_{\min}}, & \text{if } x_i^{(t)} < c_{\min} \\ R_i(t)\dfrac{x_i^{(t)} - c_{\max}}{c_{\max} - c_{\min}}, & \text{if } x_i^{(t)} > c_{\max} \end{cases} \tag{16}$$

where $c_{\max}$ and $c_{\min}$ represent the upper confidence limit and lower confidence limit in formula (12), respectively.

Next, the cumulative degree of discrepancy can be expressed as:

$$G(x_i, t) = \sum_{m=1}^{k} g(x_i, t - m) \tag{17}$$

The larger the value of the cumulative degree is, it indicates that the readings of sensor deviate a lot from the representative data, and the data will be considered outliers with high probability. Therefore, the probability of the sensor node being considered as outlier can be modified to:

$$P_i(t) = \frac{1 - P_i(t-1)e^{-w}G(x_i, t-1)}{1 + P_i(t-1)e^{-w}G(x_i, t-1)} \tag{18}$$

B. VERFICATION OF OUTLIER SOURCE

In order to identify the source of anomalies, spatial correlation can be exploited to verify the anomalies [38]. The spatial and temporal model represents the transition dynamics of the process over time, as well the correlative structure between the different sensor stations within a deployment. We manually set the parameters of the sensor state variables and observation variables [39]. When a node detects a suspected anomaly, it sends a request message to its neighbor through the wireless channel, and receives the anomaly probability from its neighbors. Suppose that μ_0 and σ_0 are the mean and standard deviation of the sensed data as outlier among neighbor nodes, respectively. According to the Grubbs criterion [40], [41], $K_G(\delta)$ represents the discriminant value of the criterion and δ is the probability of exceeding tolerance, if $P_i(t)$ will be satisfied with:

$$|P_i(t) - \mu_0| < K_G(\delta)\sigma_0 \tag{19}$$

It means that the error originates from the random error in the event process, and the state of the node is consistent with that most of its neighbors. Otherwise, it is considered that the state of the node is inconsistent with that of the adjacent representative nodes, and may result in faults or measurement errors. The parameter δ should be selected according to the specific situation. Since the event process can be regarded as a Bernoulli process of random variable with normal distribution, it can be simplified to a random variable in the standard normal distribution.

Finally, it can be summarized as follows: (i) if $x_i^{(t)} \neq x_i^{(t-1)}$ and $K_G(\delta)\sigma_0$, the sequence can be determined as an event and indicates a change in the normal behavior of the sensor nodes. Otherwise, $x_i^{(t)} = x_i^{(t-1)}$ and $|P_i(t) - \mu_0| \geq \delta\sigma_0$, the measurements in the sequence will be detected as outliers and with errors being labeled.

V. RESULTS AND DISCUSSION

In this paper, a mathematical model is established by using MATLAB, and the proposed anomaly detection algorithm for wireless sensor networks is simulated and analyzed. Assuming that each node is equipped with temperature or humidity sensors, a set of data can be collected at the same time in each sampling period. In the simulation experiment, the size of the region is 500×500 m, and the sensor node's communication radius is limited within 25 m. The sampling frequency of the sensor is 10Hz and the duration of each round is 100 s. The faulty nodes, event nodes, normal nodes and error nodes are selected and randomly deployed in the region, and the simulated data sets are injected into all nodes. In addition, the value of θ is 0.5, and we take $\mu = 100$, $\sigma = 10$.

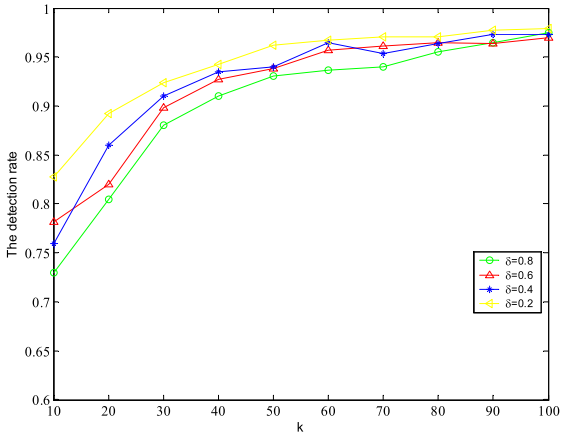

FIGURE 1. The detection rate with different λ.

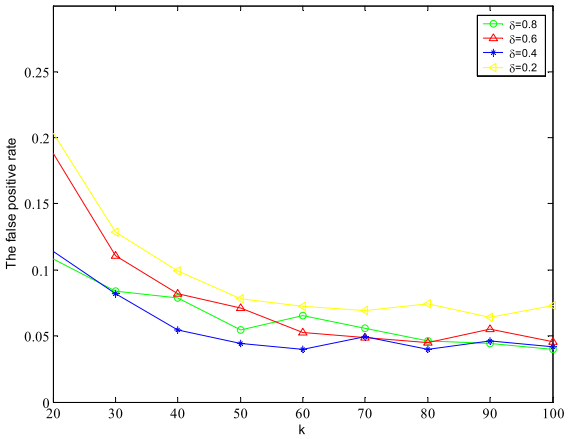

FIGURE 2. The false positive rate with different λ.

Firstly, the influence of parameter setting on experimental results is analyzed. To evaluate the detection accuracy, we employ the metrics to capture the performance, which includes the detection rate and false positive rate [42], [43]. The detection rate is the ratio of outliers being detected to all outliers, and indicates the rate of the detection accuracy. The false positive rate is the ratio of the number of non-outlier data diagnosed as outlier to the total number of non-outlier data. To obtain at a reliable level, small false positive rate should be ensured. According to the experimental results of Fig. 1 and 2, it can be seen that the detection rate will obtain optimum value with $\delta = 0.2$, but the corresponding false alarm rate demonstrates relatively high. Our method makes use of spatial correlation to verify the source of anomalies, and the conditions tend to be stringent with neighbors' observations by decreasing the value of δ. At this time, the abnormal data in the region can be quickly identified by the algorithm, but at the same time, it is easy to misjudge some normal data in the region as abnormal values. At this time, the abnormal data

in the region can be quickly identified. However, it is easy to mistaken some normal observations in the region as outlier data simultaneously. Comparatively, when $\delta = 0.4$, higher detection rate and lower false alarm rate can be obtained.

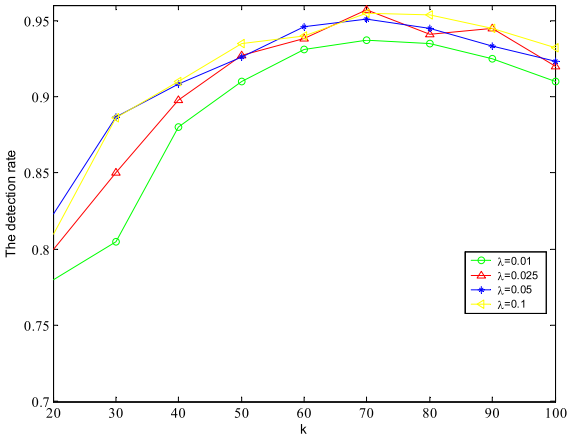

FIGURE 3. The detection rate with different significance level.

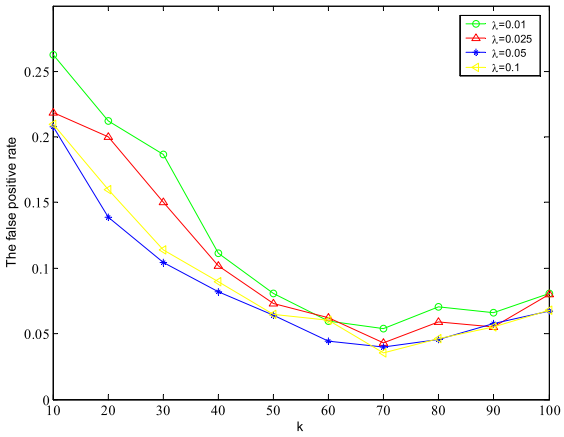

FIGURE 4. The false positive rate with different significance level.

Further, the effect of confidence level on the detection rate and false positive rate is investigated [44], [45]. According to the definition of confidence interval, the accuracy of data deviation will be greatly influenced by the number of samples and confidence level λ. Fig. 3 and 4 show the results of detection rate and false positive rate with different λ, respectively. The threshold range in traditional threshold methods is difficult to determine, which leads to high false alarm about the occurrence of errors or events. From the experimental results, it can be seen that when the number of samples is large enough, the interval difference measure can solve the above problems effectively. However, under the condition of different confidence levels, the detection rate increases first and then decreases with the number of samples. That is due to the fact of the cumulative difference will increase with the number of credible observations, and the excessive sample size is easy to increase the probability of anomaly occurs and result in high false alarm rate. It is worth noting that a relatively low confidence level will result in a high the detection rate and high FPR, and vice versa. To reduce the error rate caused by confidence interval estimation, $k = 70$ and 0.05 significance level were selected as the parameters of subsequent experiments.

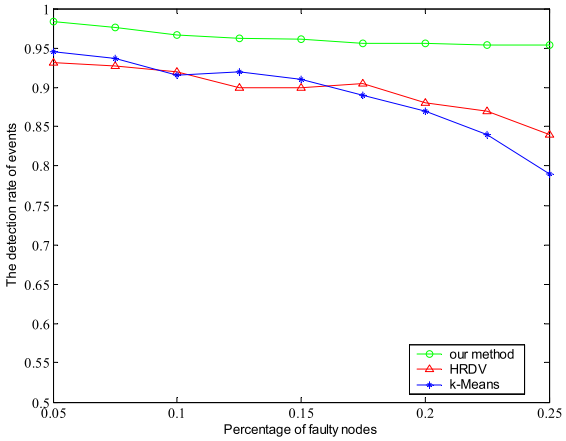

FIGURE 5. Comparison of the detection rate of events.

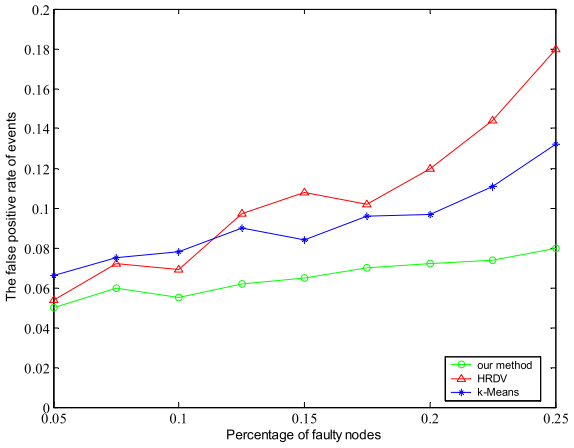

FIGURE 6. Comparison of the false positive rate of events.

Subsequently, we make experimental comparisons with the traditional methods HRDV [23], and k-Means [43] in aspect of the detection rate and false positive rate of events and errors. In the scenarios, the proportion of fault nodes in the network ranges from 0.05 to 0.25. Fig. 5 and 6 illustrate the detection rate and false positive rate of events and errors with different percentage of faulty sensor nodes. HRDV does not fully consider the state of neighbor nodes, and the reply of suspected neighbor nodes should not be included in the reference sample set. In k-Means, according to the diagnosis and response of the neighbor nodes, the classification of the situation is not accurate, which result in unreliable outlier detection in all cases. As can be seen from Fig. 6, the false positive rate of events in k-Means increases sharply as well as the percentage of faulty nodes. During the process of anomaly source verification in our method, the nodes with

measurement errors will be checked as errors or events of outliers primarily. Then, the spatial correlation is used to verify the anomaly, and confidence intervals and deviation degree are utilized to identify the source of anomalies. As can be seen from the experimental results, even with high percentage of faulty nodes, our method can obtain better performance in aspects of the event's identification.

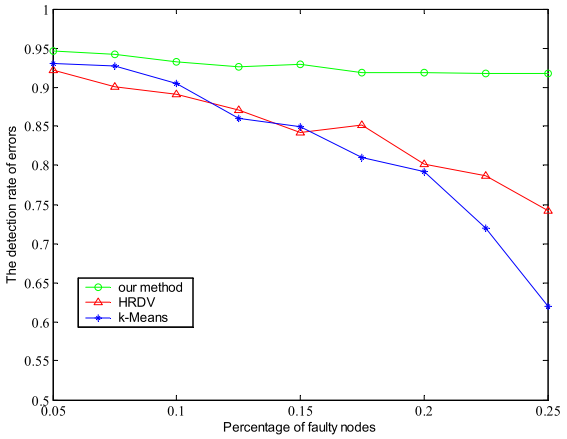

FIGURE 7. Comparison of the detection rate of errors.

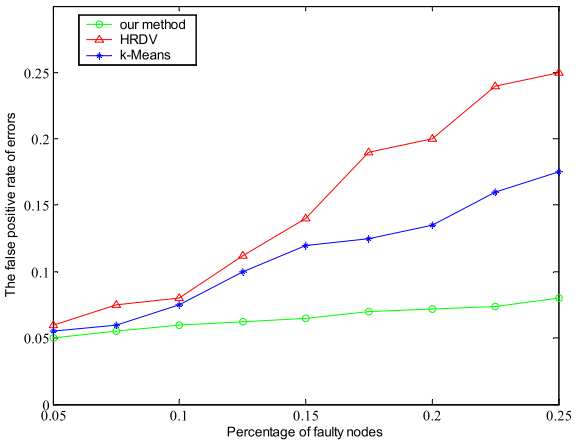

FIGURE 8. Comparison of the false positive rate of errors.

Fig. 7 and 8 show the comparison of detection rate and false positive rate of errors, respectively. In HRDV, by using of the variance of historical data to estimate the reliability level of nodes, the value of variance may decrease gradually once the observations maintains a stationary value within an erroneous range even when anomaly events occur. In this case, the anomaly may be ignored. k-Means conducts the outlier detection base on the distance to cluster center. However, when the deviated readings lower or higher than the threshold being generated at the same time, the average value of outlier data will counteract the influence on confidence interval of prediction and decrease the accuracy of detection. Moreover, event or fault diagnosis only by the mean will produce too many suspicious nodes. With the increase of percentage of faulty nodes, it cannot well resist the interference of outliers and demonstrate sharp decline in error detection rate. In contrast, by employing the median of the credible sample and estimating the confidence interval to quantify the deviation degree of the sensor data, our method distinguishes errors and events more accurately. Experiments show that even with the increase of error nodes, the proposed algorithm can maintain a high detection rate, and the false alarm rate will not increase with the increase of error nodes.

VI. CONCLUSIONS

Recent technological advances have enabled urban sensor networks to be widely used in civil areas both technically and economically, and sensors embedded in mobile devices or electrical devices can be applied to collect and report sensing data. However, abnormal data, outliers or anomalies have affected the utility in various domains seriously. Timely and accurate detection of unexpected events, monitoring of network performance and anomaly detection of data flow are of great significance to improve the decision-making ability of the system. In this paper, we proposed an anomaly detection method for urban sensing based on sequential data and credibility. The evaluation results on both the detection rate and the false positive rate demonstrate better performance of our approach than existing methods. Future work will focus on excellent design of anomaly detection method in aspect of tradeoff among security and delivery ratio of heterogeneous data sources collected on city-wide deployments.

ACKNOWLEDGMENT

The authors would like to thank Miss Jun Ma for useful discussions and consultations.

REFERENCES

[1] S. Bae and H. Kim, "Unlimited cooperative sensing with energy detection for cognitive radio," *J. Commun. Netw.*, vol. 16, no. 2, pp. 172–182, Apr. 2014.

[2] S. Xie and Z. Chen. (Mar. 2017). "Anomaly detection and redundancy elimination of big sensor data in internet of things." [Online]. Available: https://arxiv.org/abs/1703.03225

[3] U. Lee, E. Magistretti, M. Gerla, P. Bellavista, and A. Corradi, "Dissemination and harvesting of urban data using vehicular sensing platforms," *IEEE Trans. Veh. Technol.*, vol. 58, no. 2, pp. 882–901, Feb. 2009.

[4] L. Martí, N. Sanchez-Pi, J. M. Molina, and A. C. B. Garcia, "Anomaly detection based on sensor data in petroleum industry applications," *Sensors*, vol. 15, no. 2, pp. 2774–2797, Jan. 2015.

[5] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[6] A. E. Ibor, F. A. Oladeji, and O. B. Okunoye, "A survey of cyber security approaches for attack detection, prediction, and prevention," *Int. J. Secur. Appl.*, vol. 12, no. 4, pp. 15–28, Jul. 2018.

[7] A. Fawzy, H. M. O. Mokhtar, and O. Hegazy, "Outliers detection and classification in wireless sensor networks," *Egyptian Inform. J.*, vol. 14, no. 2, pp. 157–164, Jul. 2013.

[8] S. Rajasegarar, C. Leckie, and M. Palaniswami, "Anomaly detection in wireless sensor networks," *IEEE Wireless Commun.*, vol. 15, no. 4, pp. 34–40, Aug. 2008.

[9] T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Distributed deviation detection in sensor networks," *ACM SIGMOD Rec.*, vol. 32, no. 4, pp. 77–82, Dec. 2003.

[10] C. Wei and G. Li, "Outlier detection in wireless sensor networks based on reduction strategy and adaptive SVDD," *Chin. J. Sens. Actuators*, vol. 30, no. 9, pp. 1388–1395, Sep. 2017.
[11] Y. Chen, D. Miao, and H. Zhang, "Neighborhood outlier detection," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8745–8749, Dec. 2010.
[12] H. Fei and G. Li, "Abnormal data detection algorithm for WSN based on k-means clustering," *Comput. Eng.*, vol. 41, no. 7, pp. 124–128, Jul. 2015.
[13] R. Bi and J. Li, "Energy efficient Top-k monitoring algorithm in wireless sensor networks," *J. Comput. Res. Dev.*, vol. 51, no. 11, pp. 2361–2373, Nov. 2014.
[14] V. S. K. Samparthi and H. K. Verma, "Outlier detection of data in wireless sensor networks using kernel density estimation," *Int. J. Comput. Appl.*, vol. 5, no. 7, pp. 28–32, Aug. 2010.
[15] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *Proc. 32nd Int. Conf. Very large Data Bases*, 2006, pp. 187–198.
[16] Y. Yi, W. Zhou, Y. Shi, and J. Dai, "Speedup two-class supervised outlier detection," *IEEE Access*, vol. 6, pp. 63923–63933, Oct. 2018.
[17] Y. Pan, G. Li, and Y. Xu, "Abnormal data detection method for environment wireless sensor networks based on DBSCAN," *Comput. Appl. Softw.*, vol. 29, no. 11, pp. 69–72, Nov. 2012.
[18] M.-H. Lee and Y.-H. Choi, "Fault detection of wireless sensor networks," *Comput. Commun.*, vol. 31, no. 14, pp. 3469–3475, Sep. 2008.
[19] L. Beggel, B. X. Kausler, M. Schiegg, M. Pfeiffer, and B. Bischl, "Time series anomaly detection based on shapelet learning," in *Computational Statistics*. Berlin, Germany: Springer, 2018. doi: 10.1007/s00180-018-0824-9.
[20] X. Wang, X. Wang, X. Fu, and G. Xing, "MCRT: Multichannel real-time communications in wireless sensor networks," *ACM Trans. Sensor Netw.*, vol. 8, no. 1, p. 2, Aug. 2011.
[21] B. Krishnamachari and S. Iyengar, "Distributed Bayesian algorithms for fault-tolerant event region detection in wireless sensor networks," *IEEE Trans. Comput.*, vol. 53, no. 3, pp. 241–250, Mar. 2004.
[22] G. Li, Y. Sun, L. Liu, Q. Zhang, and T. Yang, "Distributed fault-tolerant event boundary detection in wireless sensor networks," *Comput. Eng.Appl.*, vol. 45, no. 17, pp. 28–32, Jun. 2009.
[23] D.-W. Lee and J.-H. Kim, "High reliable in-network data verification in wireless sensor networks," *Wireless Pers. Commun.*, vol. 54, no. 3, pp. 501–519, Aug. 2010.
[24] Q.-Q. Ren, J.-Z. Li, and S.-Y. Cheng, "Fault-tolerant event monitoring in wireless sensor networks," *Chin. J. Comput.*, vol. 35, no. 3, pp. 581–590, Mar. 2012.
[25] C. Zhang, H. Yan, S. Lee, and J. Shi, "Multiple profiles sensor-based monitoring and anomaly detection," *J. Qual. Technol.*, vol. 50, no. 4, pp. 344–362, Oct. 2018.
[26] S.-K. Zhang, Y.-H. Wang, Z.-M. Cui, and J.-X. Fan, "Event region fault-tolerant detection algorithm based on aggregation tree," *J. Commun.*, vol. 31, no. 9, pp. 74–87, Sep. 2010.
[27] D.-L. Cao, J.-N. Cao, and B.-H. Jin, "A fault-tolerant algorithm for event region detection in wireless sensor networks," *Chin. J. Comput.*, vol. 30, no. 10, pp. 1770–1776, Oct. 2007.
[28] U. Fiore, F. Palmieri, A. Castiglione, and A. De Santis, "Network anomaly detection with the restricted Boltzmann machine," *Neurocomputing*, vol. 122, pp. 13–23, Dec. 2013.
[29] C. Alippi, S. Ntalampiras, and M. Roveri, "A cognitive fault diagnosis system for distributed sensor networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 8, pp. 1213–1226, Aug. 2013.
[30] S. Rajasegarar, C. Leckie, and M. Palaniswami, "Hyperspherical cluster based distributed anomaly detection in wireless sensor networks," *J. Parallel Distrib. Comput.*, vol. 74, no. 1, pp. 1833–1847, Jan. 2014.
[31] L. A. Villas, A. Boukerche, D. L. Guidoni, H. A. B. F. de Oliveira, R. B. de Araujo, and A. A. F. Loureiro, "An energy-aware spatio-temporal correlation mechanism to perform efficient data collection in wireless sensor networks," *Comput. Commun.*, vol. 36, no. 9, pp. 1054–1066, May 2013.
[32] M. Blangiardo, M. Cameletti, G. Baio, and H. Rue, "Spatial and spatio-temporal models with R-INLA," *Spatial Spatio-Temporal EpideMiol.*, vol. 4, pp. 33–49, Jan. 2013.
[33] A. H. Mohajerzadeh, M. H. Yaghmaee, and A. Zahmatkesh, "Efficient data collecting and target parameter estimation in wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 57, pp. 142–155, Nov. 2015.
[34] S. B. Chandanapalli, E. S. Reddy, and D. R. L. Davuluri, "Efficient design and deployment of aqua monitoring systems using WSNs and correlation analysis," *Int. J. Comput. Commun. Control*, vol. 10, no. 4, pp. 471–479, Aug. 2015.
[35] M. C. Jones, "Kumaraswamy's distribution: A beta-type distribution with some tractability advantages," *Stat. Methodol.*, vol. 6, no. 1, pp. 70–81, Jan. 2009.
[36] A. Ghaddar, T. Razafindralambo, I. Simplot-Ryl, D. Simplot-Ryl, S. Tawbi, and A. Hijazi, "Investigating data similarity and estimation through spatio-temporal correlation to enhance energy efficiency in WSNs," *Ad Hoc Sensor Wireless Netw.*, vol. 16, no. 4, pp. 273–295, 2012.
[37] J. J. Jeong, S. H. Kim, G. Koo, and S. W. Kim, "Mean-square deviation analysis of multiband-structured subband adaptive filter algorithm," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 985–994, Feb. 2016.
[38] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 2, pp. 159–170, Second Quarter 2010.
[39] E. W. Dereszynski and T. G. Dietterich, "Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns," *ACM Trans. Sensor Netw.*, vol. 8, no. 1, p. 3, Aug. 2011.
[40] B. Wu, X. Yan, Y. Wang, and C. Guedes Soares, "An evidential reasoning-based CREAM to human reliability analysis in maritime accident process," *Risk Anal.*, vol. 37, no. 10, pp. 1936–1957, Oct. 2017.
[41] K. K. L. B. Adikaram, M. A. Hussein, M. Effenberge, and T. Becker, "Data transformation technique to improve the outlier detection power of grubbs' test for data expected to follow linear relation," *J. Appl. Math.*, vol. 2015, pp. 1–9, Jan. 2015.
[42] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.
[43] M. Wazid and A. K. Das, "An efficient hybrid anomaly detection scheme using k-means clustering for wireless sensor networks," *Wireless Pers. Commun.*, vol. 90, no. 4, pp. 1971–2000, Oct. 2016.
[44] Z. Huang, G. Shan, J. Cheng, and J. Sun, "TRec: An efficient recommendation system for hunting passengers with deep neural networks," *Neural Comput. Appl.*, vol. 31, pp. 209–222, Jan. 2019. doi: 10.1007/s00521-018-3728-2.
[45] B. Wu, L. Zong, X. Yan, and C. G. Soares, "Incorporating evidential reasoning and TOPSIS into group decision-making under uncertainty for handling ship without command," *Ocean Eng.*, vol. 164, pp. 590–603, Sep. 2018.

HONG ZHANG was born in Tianshui, Gansu, China, in 1977. She received the B.Sc. degree from the Lanzhou University of Technology, in 1999, and the M.D. degree from Lanzhou University, in 2011. She became an Associate Professor of computer science and technology with the Gansu University of Chinese Medicine, in 2010. Her current research interests include intelligent information systems and computer networks.

ZHANMING LI was born in Xi'an, Shanxi, China, in 1962. He received the B.E. degree from Xi'an Jiaotong University, in 1984, and the M.D. degree from the University of Strathclyde, in 2011. He is currently a Professor with the Lanzhou University of Technology. His current research interests include the modeling and control of complex systems, neuro-fuzzy systems and soft computing, theory and engineering of computer control systems, and the research and development of embedded network single-chip microcomputer systems.