Information Sciences 411 (2017) 52–65

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Anomaly detection based on a dynamic Markov model

Huorong Ren[a,b], Zhixing Ye[a,b,*], Zhiwu Li[c,a]

[a] School of Electro-Mechanical Engineering, Xidian University, Xi'an 710071, China
[b] The Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xi'an 710071, China
[c] Institute of Systems Engineering, Macau University of Science and Technology, Taipa 999078, Macau, China

ARTICLE INFO

Article history:
Received 4 August 2016
Revised 12 May 2017
Accepted 14 May 2017
Available online 15 May 2017

Keywords:
Sequence data
Anomaly detection
Markov model
Higher order Markov model

ABSTRACT

Anomaly detection in sequence data is becoming more and more important in a wide variety of application domains such as credit card fraud detection, health care in medical field, and intrusion detection in cyber security. In the existing anomaly detection approaches, Markov chain techniques are widely accepted for their simple realization and few parameters. However, the short memory property of a classical Markov model ignores the interaction among data, and the long memory property of a higher order Markov model clouds the relationship between the previous data and current test data, and reduces the reliability of the model. Besides, both of these models cannot successfully describe the sequences changing with a tendency. In this paper, we propose an anomaly detection approach based on a dynamic Markov model. This approach segments sequence data by a sliding window. In the sliding window, we define the states of data according to the value of the data and establish a higher order Markov model with a proper order consequently, to balance the length of the memory property and keep up with the trend of sequences. In addition, an anomaly substitution strategy is proposed to prevent the detected anomalies from impacting the building of the models and keep anomaly detection continuously. The experimental results using simulated datasets and real-world datasets have demonstrated that the proposed approach improves the adaptability and stability of anomaly detection in sequence data.

1. Introduction

Anomaly detection, as an important problem in data mining, has been studied in a variety of research fields and applications such as intrusion detection in cyber security [2,33], fraud detection of credit cards [33] and safety systems [27], insurance [19], and health care [15]. As far as anomaly is concerned, there is still no uniformly acceptable definition. One commonly used definition in statistics is that the data, which do not obey sequence distributions and position far away from other objects, are regarded as abnormal [10,14]. Sequence data can be found in extensive application domains such as networks, information biology, weather forecast, and system management [3]. Usually, most of them exhibit two important characteristics: dynamics and trends [36], and as such are hard to detect [11]. Anomaly detection in those sequence data is a challenging task, and one has to refer to the usage of sequential properties of data in order to detect anomalies [23,40,41].

* Corresponding author.
E-mail addresses: 18702979131@163.com, 925977035@qq.c (Z. Ye).

http://dx.doi.org/10.1016/j.ins.2017.05.021
0020-0255/

Anomaly detection in sequence data is a focus of a deluge of studies. Quite commonly, most of the existing techniques are classified into the following three categories [6,8,28]: distance-based anomaly detection; clustering-based anomaly detection; and prediction-based anomaly detection.

The distance-based anomaly detection techniques focus on calculating the distance among the data points in the data space by accepting a certain distance function [13]. When a data object exhibits a large distance with other objects, it is regarded as abnormal. For example, Chandola et al. [4,20] propose a kNN-based (k-nearest neighbor) technique in which the k-nearest neighbor distances of all objects are calculated and treated as the anomaly scores of objects. Two disadvantages of distance-based techniques are found, i.e., the choice of the distance measure directly determines their performance and the time complexity is up to $O(n^2)$ when computing the distance among n points [7,25].

The clustering-based anomaly detection techniques directly or indirectly utilize a clustering approach (e.g., Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and K-means) [12,16,21] to cluster data. The data points that cannot be easily clustered will be considered as abnormal. This methodology is simple and can make use of a large number of existing research results. However, there is a big difference between cluster analysis and anomaly detection. The purpose of the former is to seek for the category of clusters; and the latter is to find the abnormal data. Anomaly detection is just an "ancillary products" of clustering [17,24]. The fact that general approaches are not particularly optimized for anomaly detection leads to low detection efficiency. Besides, in most cases, the definitions of anomaly and detection criteria are implicit and cannot be clearly reflected in the process of clustering.

In the prediction-based anomaly detection techniques, many studies use mathematical models (e.g. Bayesian networks, Markov models, neural networks, and support vector machines) [18,19,26,31] to formally decide the unknown quality of sequence data, and then build the prediction models. Finally the anomaly will be found according to the deviation between the predicted value and the actual value at each time. These methods have better performance on the sequence of lower dimensionality. However, Bayesian networks have an assumption that attributes are independent of each other, which is usually not true in practical applications [18]. Neural networks require a large number of parameters, such as network topology, weights and threshold values. Besides, the learning time is too long, and may even fail to achieve the purpose of learning [19]. Support vector machines are difficult to implement large scale training samples. It will consume a lot of memory and computing time [30].

A Markov model is a powerful finite state machine, which is widely used in sequence modeling. The main advantage of the Markov techniques is that each event can be analyzed. Therefore, the techniques are able to detect anomalies even if they are located in a long sequence [35]. In this paper, we concentrate on the anomaly detection based on Markov models. Ozkan and Kozat [22] propose an online anomaly detection under Markov statistics with controllable false alarm rate for fast streaming temporal data. This algorithm learns the nominal attributes under possibly varying Markov statistics. Then, an anomaly is declared at a time instant, if the observations are statistically sufficiently deviant. Sha et al. [29] present a multi-order Markov chain based scheme for anomaly detection in server systems. This approach takes a higher order Markov chain and multivariate sequences into account to produce several indicators of anomalies.

In Markov chain approaches, classical Markov chain techniques mostly utilize the short memory property (a single step) of classical Markov models. The short memory property essentially comes with the two basic assumptions [34]: (1) The state probability distribution of time t is only related to the state of time $t-1$. (2) The transformation from the state of time $t-1$ to the state of time t is time independent.In practical applications, however, these two basic assumptions cannot be strictly satisfied. The state probability distribution of time t is usually not only related to the state of time $t-1$, but also related to the states of a period of time before time $t-1$. Therefore, the short memory property of classical Markov models is not applicable to real-world data [1].

A higher order Markov model [5,32] is presented with its long memory property by taking the interaction among states into account, such that the model can better describe the characteristics of sequence data than classical Markov models. In theory, the memory time can be infinitely long by increasing the order of the Markov model. Besides, in Markov chain approaches, once the Markov models are established in training phase, the order of Markov models is fixed to detect anomaly in testing phase. However, the fact, that the fixed Markov models (n-order) force each state of a sequence to be conditioned on the fixed previous n states, may not be sufficient to provide a reliable estimate of the detecting state. With the decrease of the correlation between old and new data, the fixed Markov models are no longer applicable to the entire sequence. At the same time, both models mentioned above cannot completely describe the characteristics of whole sequence with a trend yet. They will be invalid, when the value of a sequence data exceeds the area covered by the training data.

In cognitive science, as is known to all that the reliability and accuracy of memory will be lower and lower over time. Thus an appropriate length of memory time is useful to cognize current events. Besides, as time goes by, the events in cognitive memory are constantly updated to keep up with the changing of the current events. Motivated by this theory, a dynamic Markov model is proposed in this paper to balance the length of the memory property of Markov models and keep the strong correlation between the memory (or the Markov model) and current test data. This dynamic model first makes use of a sliding window to segment a sequence data. Then the correlation analysis of data in the sliding window is used to find out a proper order of a Markov model. And the order of the Markov model is continuously updated with the sliding window sliding to keep the relationship between the Markov model and current test data. Besides, when the current test data exceed the scope of the previously defined states, the states of data in the sliding window will be redefined, and the model will be retrained to follow the changes of the sequence. At the same time, in order to detect anomalies continuously and prevent anomaly points detected from infection to the building of the models, an anomaly substitution strategy is

proposed. Therefore this research presents a robust anomaly detection approach based on a dynamic Markov model. In addition, this paper focuses on the sequences type on the data with dynamics and trend, such as electrocardiograms (ECGs) data in medical, seasonal data, and quarterly data.

The paper is organized as follows: Section 2 reviews classical Markov models and higher order Markov models. In Section 3, we develop an anomaly detection approach based on a dynamic Markov model. Section 4 focuses on the comparison results due to the proposed approaches. Finally, the conclusions of this paper are drawn in Section 5.

2. Classical Markov models and higher order Markov models

In order to better understand the proposed dynamic Markov model, we introduce classical Markov models and higher order Markov models in this section.

2.1. Classical Markov models

For a sequence $X(T) = \{x_1, x_2, x_3, \cdots, x_t, \cdots, x_T\}$, where x_t is the data present at time t, the complete parameter set of a classical Markov model can be represented by, as shown in [18,22,29,32]:

$$\lambda = \{\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{P}\} \tag{1}$$

where:

(1) λ represents the classical Markov model.
(2) $\boldsymbol{S}$, the state space of the sequence $X(T)$, includes all of the possible states of each data, i.e., $\boldsymbol{S}$ can be represented by $\boldsymbol{S} = \{1, 2, 3, \cdots, N\}$, where N is the number of the states present in the sequence. Note that we use s_t ($s_t \in \boldsymbol{S}$) to denote that the data x_t is in the state s_t at time t.
(3) $\boldsymbol{Q} = \{q_1, q_2, \cdots, q_i, \cdots, q_N\}$ is an initial probability distribution set of the sequence $X(T)$, where q_i is the initial probability of the state i ($i \in \boldsymbol{S}$) with $\sum_{i=1}^{N} q_i = 1$. q_i can be calculated as follows:

$$q_i = \frac{M_i}{T} \tag{2}$$

where M_i is the number of the data in the state i in the sequence $X(T)$, and T is the number of data in the sequence $X(T)$.
(4) $\boldsymbol{P}$, a state transition probability matrix, can be expressed in the form $P = |p_{s_{t-1}s_t}|_{N\times N}$, where $p_{s_{t-1}s_t}$ is the state transition probability from the state s_{t-1} to s_t after a single step. For example, if the states at time $t-1$ and t are $s_{t-1} = 1$ and $s_t = 3$, the state transition probability between these two states is p_{13}. For a state s_{t-1}, we have $\sum_{s_t=1}^{N} p_{s_{t-1}s_t} = 1$.

$$p_{s_{t-1}s_t} = \frac{M_{s_{t-1}s_t}}{\sum_{s_t=1}^{N} M_{s_{t-1}s_t}} \tag{3}$$

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix} \tag{4}$$

where $M_{s_{t-1}s_t}$ is the number of state transition from the states s_{t-1} to s_t after a single step.

Classical Markov models, mostly utilizing the short memory property, can record the states of data in the short term. However, it is well known that short memory time cannot effectively support the cognition of current events, since the real-world sequence data are often correlated with each other.

2.2. Higher order Markov models

For the sequence $X(T) = \{x_1, x_2, x_3, \cdots, x_t, \cdots, x_T\}$, a higher order Markov model (n-order) can be expressed as [1,5,34,38]:

$$\lambda(n) = \{\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}, \cdots, \boldsymbol{P}^{(n)}\} \tag{5}$$

where:

(1) $\lambda(n)$ represents a higher order Markov model.
(2) $\boldsymbol{S}$ and $\boldsymbol{Q}$ have the same meaning as in the classical Markov model λ.

(3) $P^{(n)}$, an n-order state transition probability matrix, can be expressed as $P^{(n)} = |p^{(n)}_{s_{t-n}s_t}|_{N\times N}$, where N is the number of states, and $p^{(n)}_{s_{t-n}s_t}$ is the state transition probability from states s_{t-n} to s_t after n steps. It can be calculated as follows:

$$p^{(n)}_{s_{t-n}s_t} = \frac{M^{(n)}_{s_{t-n}s_t}}{\sum_{s_t=1}^{N} M^{(n)}_{s_{t-n}s_t}} \tag{6}$$

where $M^{(n)}_{s_{t-n}s_t}$ is the number of state transitions from the states s_{t-n} to s_t after n step. In particular, when $n = 1$, the higher order Markov model is the same as the classical Markov model.

From the process of building the n-order Markov model, it can be found that a higher order Markov model can record state characteristics in arbitrarily long term by increasing the order of the Markov model in theory. This feature is described as the long memory property of higher order Markov models. However, with the increase of the order, the reliability of old data will be getting lower and lower. This paper tries to address this issue by allowing the order n to be adjusted in the dynamic Markov model.

3. Anomaly detection approach based on a dynamic Markov model

This section develops the anomaly detection approach based on a dynamic Markov model. The main phases of the proposed approach are shown as follows: (1) a sliding window $W(l)$ is used to segment the sequence data, where l is the length of the sliding window. Then the states of data in the sliding window are defined by an equal width interval segmentation method; (2) an n-order Markov model is established in the sliding window, where n is determined by the Pearson correlation analysis approach [31,37]; (3) the current test data are evaluated whether they exceed the scope of the defined states and detected by the n-order Markov model; (4) an anomaly substitution strategy is used to keep the detection continuously and prevent the detected anomalies from infecting the building of the models.

3.1. State definition

The nature of the Markov model requires that there should be a clear state definition. In this phase, for a sequence $X(T) = \{x_1, x_2, x_3, \cdots, x_t, \cdots, x_T\}$, a sliding window $W_t(l)$ is used to segment this sequence data. For example, a sliding window is $W_t(l) = \{x_{t-l}, x_{t-l+1}, x_{t-l+2}, \ldots, x_{t-1}\}$, where $t = l+1, l+2, \ldots, T$, and l is the length of sliding window $W_t(l)$. Then we apply an equal width interval segmentation method that segments the value range of the sliding window into N intervals with the identical size as the states of the data. The width of each state interval is calculated as follows:

$$\omega = \frac{max(W_t(l)) - min(W_t(l))}{N} \tag{7}$$

where N is the number of states, which remains unchanged. $\max(W_t(l))$ and $\min(W_t(l))$ are the maximum and minimum values of points in the sliding window $W_t(l)$, respectively. There are N states 1, 2, 3, $\cdots$, and N. Each data point in this sliding window has its state tag.

3.2. Establishment of an n-order Markov model

In order to balance the length of the memory property and maintain the reliability of the established model, an n-order Markov model whose order depends on the data in the sliding window $W_t(l)$ is built. Owing to the fact that the memory property of higher order Markov models utilizes the correlation among data, a stronger correlation means a longer effective memory time and a higher order of the Markov model. The order of the Markov model can be determined by the correlation among data in the sliding window to ensure an effective length of the memory property and a reliability model.

In the developed approach, Pearson correlation [31,37] is used to determine the order of higher Markov models in the sliding window. For example, for a sliding window $W_t(l) = \{x_{t-l}, x_{t-l+1}, x_{t-l+2}, \ldots, x_{t-1}\}$, the Pearson correlation coefficient $r(n)$ between the two sliding windows $W_t(l)$ and $W_{t+n}(l)$, can be calculated as follows [31,37]:

$$r(n) = \frac{\frac{1}{l}\sum_{i=1}^{l} x_{t-i}x_{t+n-i} - \left(\frac{1}{l}\right)^2 \sum_{i=1}^{l} x_{t-i} \sum_{i=1}^{l} x_{t+n-i}}{\sqrt{\frac{1}{l}\sum_{i=1}^{l} x_{t-i}^2 - \left(\frac{1}{l}\sum_{i=1}^{l} x_{t-i}\right)^2}\sqrt{\frac{1}{l}\sum_{i=1}^{l} x_{t+n-i}^2 - \left(\frac{1}{l}\sum_{i=1}^{l} x_{t+n-i}\right)^2}} \tag{8}$$

where $1 \le n \le T - t + 1$. Generally, there is a consensus that the two vectors with n dimensions have strong correlation when their Pearson correlation coefficient $r(n) \ge 0.8$ [31]. If there is a strong correlation ($r(n) \ge 0.8$) between the two sliding windows data with as large n as possible, an n-order Markov model will be built with the most effective long memory property. Then an n-step transition probability matrix $\boldsymbol{P}^{(n)}$ of the normal sequence data will be calculated by Eq. (6).

Concretely, an n-order Markov model can be established in the sliding window as follows:

(1) Segment the sequence $X(T)$ using a sliding window $W_t(l)$, and then define the states of data in the sliding window with the number of states N.
(2) Use the Pearson correlation analysis approach to determine the order of the higher order Markov model n in the sliding window.
(3) Establish an n-order Markov model $\lambda(n) = \{\boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}, \cdots, \boldsymbol{P}^{(n)}\}$ in the sliding window as done in Section 2.2.

3.3. Detection and retraining phase

Once an n-order Markov model $\lambda(n)$ has been established in the sliding window $W_t(l)$, for example, in the sliding window $W_t(l) = \{x_{t-l}, x_{t-l+1}, \ldots, x_{t-1}\}$, a support probability of current test data point x_t can be calculated by using the following formula:

$$P\big(x_t|\lambda(n)\big) = q_{s_{t-n}} \prod_{i=1}^{n} p^{(i)}_{s_{t-i}s_t} \neq 0 \tag{9}$$

where $t > 1$ and $1 \leq n \leq l$. If (9) does not hold, there is at least one impossible case in all transition from the states of data points $x_{t-n}, x_{t-n+1}, \ldots,$ and x_{t-1} to x_t. In other words, the state of current test data point x_t cannot be obtained through the transition probability matrixes $\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}, \cdots,$ and $\boldsymbol{P}^{(n)}$. Therefore, the data point x_t is regarded as an anomaly.

In addition, it is necessary to take into account the situation that the state of the current test data point x_t cannot be found in the defined states in the sliding window $W_t(l)$, owing to the value of x_t not belonging to the collection $[max(W_t(l)) - min(W_t(l))]$ (e.g., the tendency sequence data). Thus the states of data and the model in this sliding window should be redefined and retrained as done in Sections 3.1 and 3.2 to catch up with the changes of the sequence data. Besides, it can be seen from (9) that there is an assumption that the l data points which are used to train the n-order Markov model in the sliding window before current test data x_t, should be normal at least. If there are some anomalies in the n data points, it is very likely for the data x_t to be infected and detected as an anomaly.

3.4. Anomaly substitution strategy

In order to ensure that the detected anomalies cannot infect the detection results of the rest data points in the future, an anomaly substitution strategy is used to prevent the rest data from the infection of detected anomalies. For example, data point x_t is detected as an anomaly, and then one point x'_t selected from the sliding window whose n-step support probability $P\big(x'_t|\lambda(n)\big)$ is the largest will be adopted to substitute the point x_t. The largest support probability means that data point x'_t has the largest probability to appear in the location of anomaly x_t through the n-order Markov model $\lambda(n)$. The n-step support probability can be calculated as follows:

$$P\big(x'_t|\lambda(n)\big) = \mathbf{max}\Big(\sum_{j=1}^{n} p^{(j)}_{s_{t-j}s'_t}\Big), \quad (s'_t = 1, \quad 2, \quad 3, \cdots, N) \tag{10}$$

Thus, the detection can be carried out dynamically and continuously by substituting the detected anomalous points and sliding the window step by step.

4. Experiments and analysis of results

In this section, we conduct experiments on both synthetic and real-world data to exhibit the performance of the proposed approach. In addition, the true positive rate (TP) and the false alarm rate (FA) are used to evaluate the performance of the anomaly detection approaches. They can be calculated as follows:

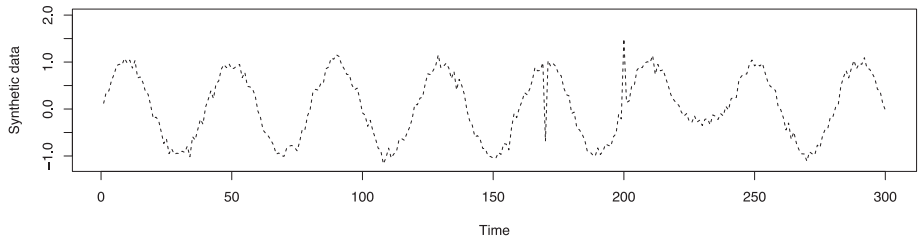
$$TP = \frac{DO}{AO} \tag{11}$$

$$FA = \frac{FDO}{ADO} \tag{12}$$

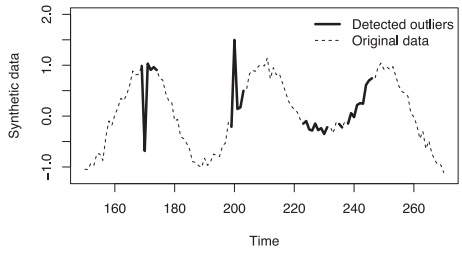
where DO, AO, FDO, and ADO denote the numbers of detected real anomalies, all real anomalies, falsely detected anomalies, and all detected anomalies, respectively.

4.1. Experiment and analysis of synthetic data

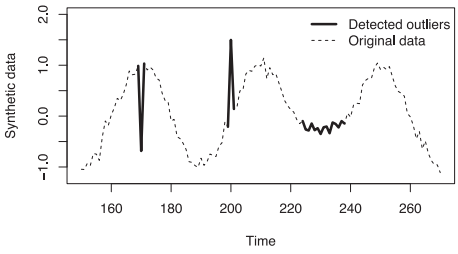
In this part, we first use synthetic sequence data to evaluate the proposed anomaly detection approach based on higher order Markov models. The data are periodic, composed of a sine signal (cycle $T = 40$) with amplitude equal to one and a zero mean normal noise with the variance of 0.1. At $t = 170$ and $t = 200$ two impulses are added with the amplitude of -1 and 1 that could be regarded as two anomalies. In addition at $t = 220$–240, the amplitude of the sine signal is 0.2, which also can be regarded as anomaly [32].



(a) Synthetic sequence data

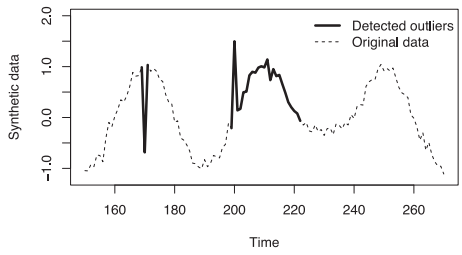


(b) No anomaly substitution strategy

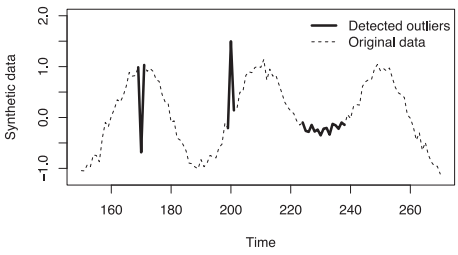


(c) Using anomaly substitution strategy

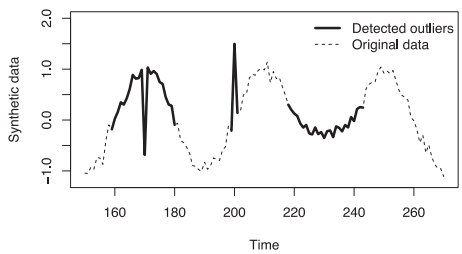
Fig. 1. Experimental results obtained for synthetic data using and not using anomaly substitution strategy.



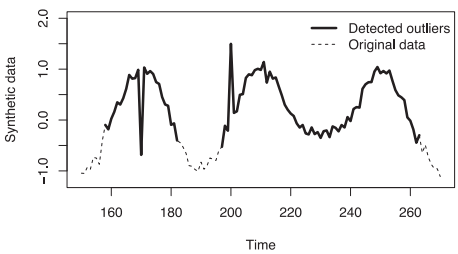
(a) $N = 4$



(b) $N = 5$



(c) $N = 6$



(d) $N = 7$

Fig. 2. Experimental results obtained for synthetic data with different states numbers.

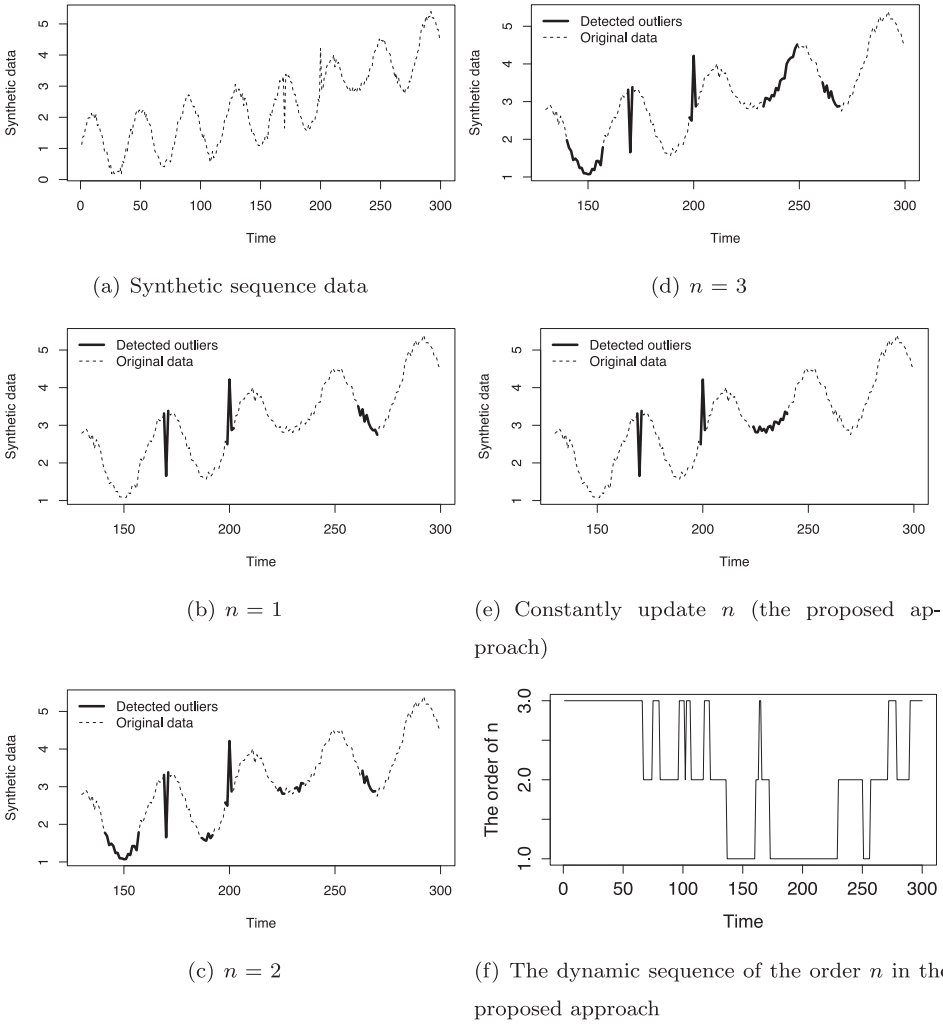
First, the effect of the anomaly substitution strategy is discussed in this experiment. The experimental results obtained on synthetic data by using and not using the anomaly substitution strategy are shown in Fig. 1.

It can be seen that the majority of anomalies at $t = 170$, 200, 220–240 can be detected in these two cases. However, in Fig. 1(b), since there is no a timely replacement of anomalies, the subsequent detection is infected (i.e. the points at $t =$ 171–174, 201–202, 241–247), and a high false alarm rate ($TP = 94.6\%$ with $FA = 53.1\%$) is obtained. In Fig. 1(c), the anomaly substitution strategy reduces the false alarm rate by 48.4% ($TP = 95.2\%$ with $FA = 4.7\%$). This experience indicates that the anomaly substitution strategy does prevent the rest data from the infection of anomaly detection and effectively restrain the false alarm rate.

In Fig. 2, the performance impact of the number of states N is discussed in this experiment. The TP and FA values of the experimental results are shown in Table 1. From Fig. 2(a) only the anomalous points at $t = 170$ and 200 can be detected.

Table 1
The *TP* and *FA* values of the experimental results for different orders *n*.

	TP(%)	*FA*(%)
$n = 1$	8.6	81.8
$n = 2$	65.2	46.1
$n = 3$	62.1	53.9
Constantly update n (the proposed approach)	95.7	8.3



(a) Synthetic sequence data

(b) $n = 1$

(c) $n = 2$

(d) $n = 3$

(e) Constantly update n (the proposed approach)

(f) The dynamic sequence of the order n in the proposed approach

Fig. 3. Experimental results for different orders *n*.

In Fig. 2(b) it can be seen that the number of states $N = 5$ leads to the best detection result ($TP = 95.2\%$ and $FA = 4.7\%$). For the number of states $N = 6$ and 7, the anomalies at $t = 170$, 200, 220–240 are detected, but the false alarm rate ($FA = 52.1\%$ and 86.7%) becomes too high. The choice of the number of states N has certain influence on the detection results. By running a large number of simulation experiments, it has been determined that the number of states N is generally in the range 5–11.

In many cases, a sequence data has a tendency. For example, the global temperature sequence data is increased in the past few decades. For this issue, an increasing trend is added to the synthetic data. The order of the higher order Markov model n in the proposed approach is also discussed by using the synthetic data. These experimental results are shown in Fig. 3 and Table 1. In Fig. 3 and Table 1, we can observe that the proposed approach (i.e., constantly update n with $N = 8$) has the best detection result with lower false alarm rate ($TP = 95.7\%$ and $FA = 8.3\%$). From Fig. 3(f), the order n changes in 1–3 with the sliding window sliding. In Fig. 3(b), the anomaly points at $t = 220$–240 cannot be detected. The reason is that when the order $n = 1$, the transfer of only two adjacent points will be considered, which ignores the relationship

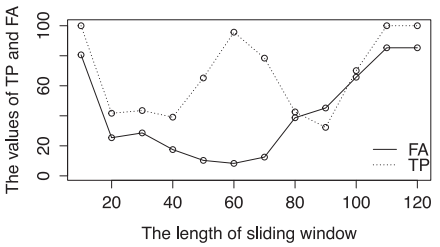


Fig. 4. Experimental results for different length (l) of sliding window.

Table 2
The *TP* and *FA* values of the experimental results on different data.

Detection methods	Passenger traffic data		Ann gun CentroidA		Chfdb chf13 45590.3	
	TP(%)	*FA*(%)	*TP*(%)	*FA*(%)	*TP*(%)	*FA*(%)
The LOF approach	66.7	66.7	62.8	16.3	58.3	10.5
The classical Markov chain techniques	100	90.5	54.3	3.7	62.5	43.3
The higher order Markov chain	50.0	94.7	89.0	32.0	64.7	0
The proposed approach	100	25.0	93.2	2.4	94.8	6.7

among data. In Figs. 3(c) and (d), a part of the anomaly points at $t = 220$–240 can be detected by increasing the order n. However, the fixed order n also causes a high false positive rate for a strict judgment standard by considering the transfer of three and four adjacent points, respectively. Thus this experience indicates that the method of constantly updating n in the proposed approach does improve the detection accuracy by keeping proper relationship with the previous data.

In the proposed approach, a sliding window $W(l)$ is used to segment a sequence data to provide adequate data for modeling. Thus in the next experiment, the effect of the length (l) of the sliding window is discussed by using the synthetic data with a trend. The experimental results are shown in Fig. 4.

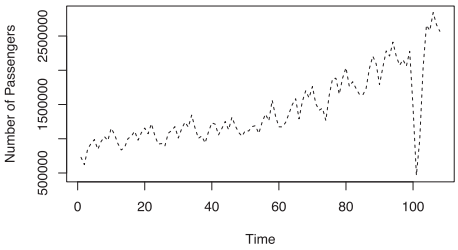
Fig. 4 shows the variation curves of *TP* and *FA* for the values of length (l). It can be seen that when $l = 60$, the experimental result is the best with $TP = 95.7\%$ and $FA = 8.3\%$. The value of *TP* reaches 100% when $l < 10$ or $l > 100$, but that *FA* is too high (greater than 80%) to be acceptable. The reason is that according to Eq. (7), in the case of a certain number of states, a too small or too large length (l) leads to the state definition too strict or too loose, respectively. Thus it is necessary to select a suitable value l. In the light of a large number of simulation experiments, it has been determined that a suitable value of l is generally between $T - 2T$ (T is the periodic or quasi periodic of sequence data).

4.2. Experiment and analysis of real-world data

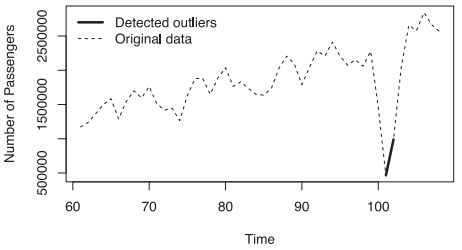
Applications of the developed adaptive anomaly detection approach based on a Markov model are tested on real-world datasets. These datasets include the passenger traffic data from an airport in Shanghai from January 1995 to January 2004 (http://robjhyndman.com/tsdldata/data/fancy.dat) representing non-stationary sequence data, the video surveillance dataset and the Electrocardiograms (ECGs) datasets from UCR Time Series Data Mining Archive (http://www.cs.ucr.edu/%7Eeamonn/discords/) representing dynamic sequence data [5,9]. In addition, In order to make the experiment more convincing, a classical anomaly detection method, the local outlier factor (LOF) [39], is used to compare with the proposed method with the classical Markov chain techniques and the higher order Markov chain.

First, the classical Markov chain techniques, the higher order Markov chain, the LOF method and the proposed approach are analyzed and compared in detail using the passenger traffic data. The anomalies in the passenger traffic data are located in January 2003 (data points 101–103 in Fig. 4), due to the outbreak of SARS (Severe Acute Respiratory Syndrome) in China and then a sharp drop in airport passenger traffic. These experimental results are shown in Fig. 5 and Table 2.

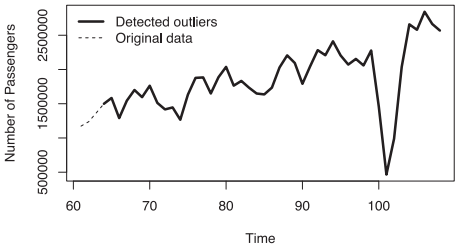
Fig. 5(a) is the passenger traffic data and the first 60 data points are used as training data. In Fig. 5(b), the classical Markov chain techniques cannot normally detect data points from 64 to 108 because of the amplitude of the data points after 63 beyond the value range of training data, where no corresponding states in the classical model are found. Therefore, all of these data points are regarded as anomalies. In Fig. 5(c) the higher order Markov model ($n = 5$) is used to replace the classical Markov model, where the order of the Markov model is $n = 5$. Similar to the result in Fig. 5(b), the higher order Markov model can neither normally detect the data points from 61 to 108. Besides, in Fig. 5(d), the LOF method only detect the data points 101 and 102 with $TP = 66.7\%$ and $FA = 66.7\%$, where $k = 6$ (k indicates the k-distance neighborhood). However, in Fig. 5(e), the result indicates that the approach proposed in this paper is well able to detect all the abnormal data points 100–103, where the length of sliding window is $l = 24$, the number of states is $N = 5$, and the dynamic order of Markov model n is shown in Fig. 5(f). According to Eq. (9), the support probability of the data points 100–103 are 0, and then these points are determined as abnormal data points. All of these experiments results show that the proposed approach has better adaptability in anomaly detection especially for tendency sequence data.



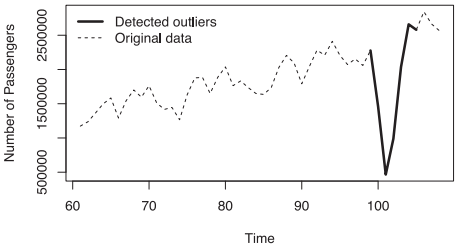
(a) The passenger traffic data



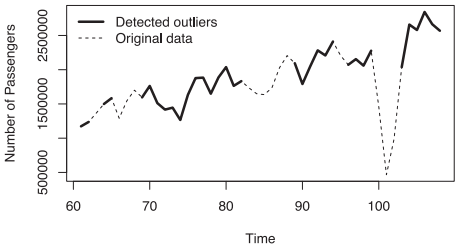
(d) The detection result using the LOF approach



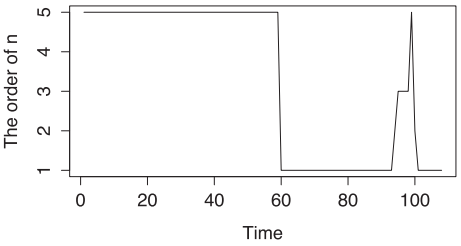
(b) The detection result using the classical Markov chain techniques



(e) The detection result using proposed approach



(c) The detection result using the higher order Markov chain ($n = 5$)



(f) The dynamic order n in the proposed approach

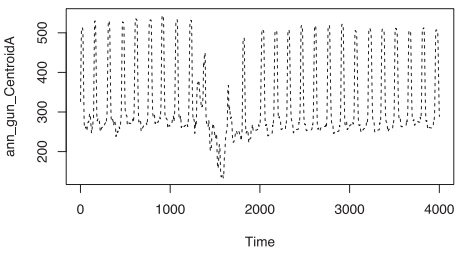
Fig. 5. Comparison of the experimental results produced by the four approaches.

In addition, the video surveillance dataset ann_gun_CentroidA extracted from a video of an actor performing various actions with and without a replica gun is detected using the four approaches mentioned above. According to the description of the data source, the anomalies in the sequence are data points 1340–1792. The results of this experiment in which the first 1000 data points are regarded as training data are depicted in Fig. 6 and Table 2.

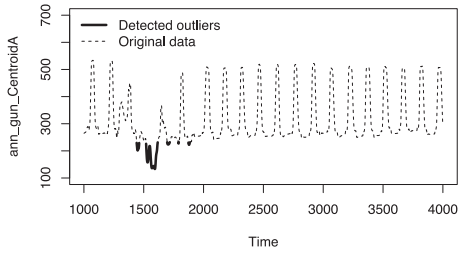
In Fig. 6(b), we can observe that the classical Markov chain techniques can only detect a small part of anomalies, which is impossible to accept in many cases. In Fig. 6(c), the higher order Markov chain techniques ($n = 5$) can successfully detect all of the anomalies. However, there are too much false alarms ($FA = 32.0\%$). In Fig. 6(d), the LOF method also only detected part of anomalies with $TP = 62.8\%$ and $FA = 16.3\%$. In comparison, the approach proposed ($N = 7$, $l = 220$) in this paper detects almost the anomalies (data points 1378–1769) with almost no false alarms ($FA = 2.4\%$). It shows that the proposed approach has better detection performance.

In the next experiments, the ECG dataset chfdb_chf13_45590.3 coming from UCR Time Series Data Mining Archive is used for anomaly detection using the four approaches mentioned above. It is known that the anomalies in the sequence data are data points 676–733 and 1119–1190. Besides, the first 600 data points are regarded as training data. The results are shown in Fig. 7 and Table 2.

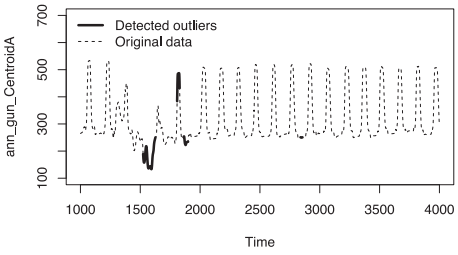
In Fig. 7, the classical Markov chain techniques and the higher order Markov chain techniques ($n = 4$) can successfully detect parts of anomalies. However, for the classical Markov chain techniques specially, there also are some false alarms ($FA = 43.3\%$). In Fig. 7(f), the LOF method detect fewer anomalies than the above two methods with $TP = 58.3\%$. In Fig. 7(e), the approach proposed in this paper ($N = 10$, $l = 200$) detects almost the anomalies (data points 676–733 and 1119–1177) with fewer false alarms (data points 843–845 and 991–993, $FA = 6.7\%$). Thus comparing with other two approaches, the



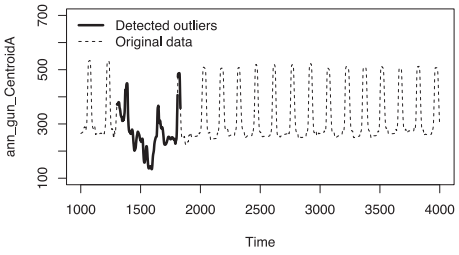
(a) The video surveillance data



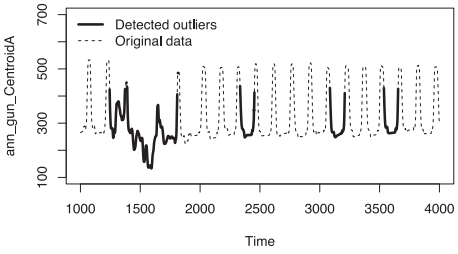
(d) The detection result using the LOF approach



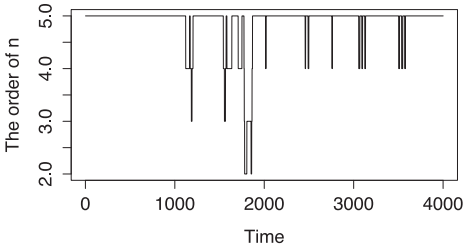
(b) The detection result using the classical Markov chain techniques



(e) The detection result using proposed approach



(c) The detection result using the higher order Markov chain ($n = 5$)



(f) The dynamic order n in the proposed approach

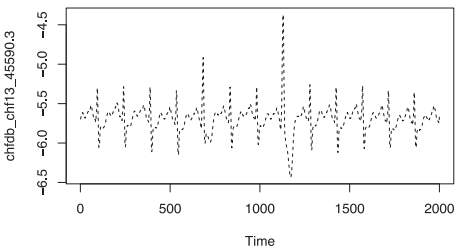
Fig. 6. Comparison of the experimental results of the four approaches.

proposed one greatly improves the detection accuracy. The results of other experiments on eight ECG datasets representing the dynamic sequence data are shown in Table 3.

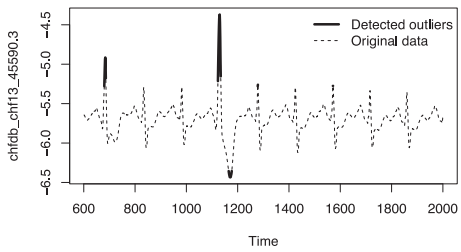
In Table 3, the proposed approach achieves highest true positive rate and lowest false alarm rate on all of the eight ECG datasets and outperforms the other three benchmark approaches. For the classical Markov chain techniques, the average true positive rate *TP* is 63.7% with the average false alarm rate *FA* of 26.6%. For the higher order Markov chain techniques ($n = 5$), the average *TP* is 75.3% and the average *FA* is 15.8%. For the LOF method, the average *TP* and *FA* are 79.3% and 13.9%. For the proposed approach, the average *TP* is 89.1% and the average *FA* is 6.5%. The best performance for the proposed approach appears on the ltstdb_20321_240_2 dataset with the true positive rate of 92.5% and the false alarm rate of 6.4%, which is better than the classical Markov chain techniques ($TP = 62.1\%$ and $FA = 23.8\%$), the higher order Markov chain techniques ($TP = 68.2\%$ and $FA = 10.9\%$), and the LOF method ($TP = 72.8\%$ and $FA = 10.9\%$). The worst performance of the proposed approach appears on the stdb_308_0_2 dataset ($TP = 85.4\%$ and $FA = 4.8\%$) and is still better than the other three approaches. The obtained results indicate that the proposed approach, constantly updating the order n of the higher order Markov model to keep the reliability of the model with the change of the sequences, has better performance than other two established approaches and the classical LOF method with higher true positive rate and lower false alarm rate. The true positive rate is increased by 13.8%, 25.4%, and 9.8% than the higher order Markov chain technique, classical Markov chain technique, and LOF method, respectively. At the same time, the false alarm rate is reduced by 9.3%, 20.1%, and 7.4% than the three techniques, respectively.

Table 3
The results of anomaly detection on eight ECG datasets using the four approaches.

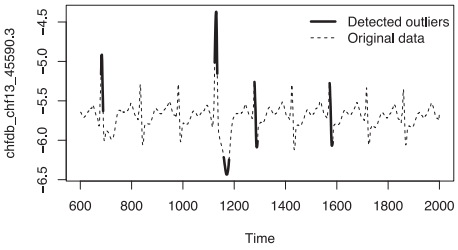
ECG datasets	datasets length	Classical Markov chain techniques			Higher order Markov chain techniques ($n = 5$)			Proposed approach			LOF method		
		N	*TPR*(%)	*FPR*(%)	*N*	*TPR*(%)	*FPR*(%)	*N*	*TPR*(%)	*FPR*(%)	*k*	*TPR*(%)	*FPR*(%)
stdb_308_0_1	2000	8	70.8	30.0	9	80.0	13.3	8	87.4	4.8	200	82.7	18.5
stdb_308_0_2	2000	7	52.6	37.4	10	84.9	16.7	7	85.6	5.8	200	78.3	13.3
ltstdb_20321_240_1	2500	10	56.9	33.7	6	67.6	21.7	10	88.5	8.3	250	80.5	14.7
ltstdb_20321_240_2	2500	10	62.1	23.8	9	68.2	10.9	10	92.5	6.4	250	72.8	10.9
chfdb_chf13_45590_1	3000	5	64.4	25.4	8	78.9	19.2	5	87.9	5.2	280	85.4	13.8
chfdb_chf13_45590_2	3000	7	60.6	23.3	9	67.1	14.8	7	89.6	8.8	280	79.3	9.4
chfdb_chf01_275_1	3500	9	69.2	18.7	10	73.5	12.6	9	91.3	5.4	300	81.5	20.8
chfdb_chf01_275_2	3500	6	72.8	20.6	8	82.5	17.3	6	90.7	7.5	300	74.1	10.3
Average			63.7	26.6		75.3	15.8		89.1	6.5		79.3	13.9



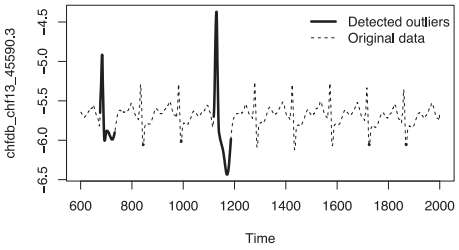
(a) The ECG data



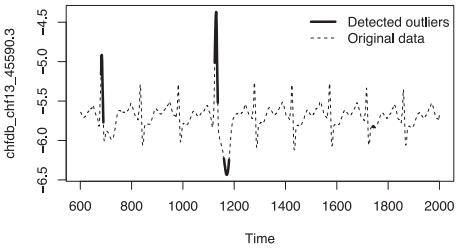
(d) The detection result using the LOF approach



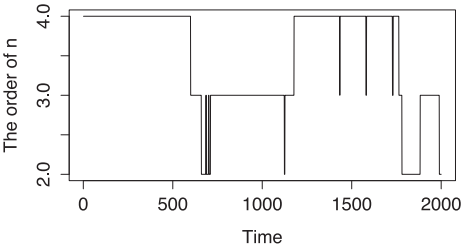
(b) The detection result using the classical Markov chain techniques



(e) The detection result using proposed approach



(c) The detection result using the higher order Markov chain ($n = 4$)



(f) The dynamic order n in the proposed approach

Fig. 7. Comparison of the experimental results of the four approaches.

5. Conclusions

This paper presents an anomaly detection approach based on a dynamic Markov model. The short memory property of a classical Markov model and the long memory property of a higher order Markov model are analyzed. The proposed approach balances the length of the memory property by repeatedly utilizing the Pearson correlation analysis approach to find a proper order of the Markov model in a sliding window. Moreover, the proposed approach maintains the established model reliable and successfully detects the tendency sequences by dynamically defining the states of data and training the models in the sliding window. For keeping detection continuously, a substitution strategy of anomalies is reported to protect the building of models from the infection of detected anomalies. The comparison of the proposed approach with other benchmark approaches shows that the proposed approach performs better in terms of the true positive rate and the false alarm rate for all of dynamic and tendency sequence data. Another interesting topic will use Petri nets [42,43] to model the anomaly detection process.

Acknowledgments

This work is partially supported by the Fundamental Research Funds for the Central Universities with Grant Nos. K50510040013 and K5051304007, the Natural Science Foundation of China under Grant No. 61374068, and the Science and Technology Development Fund, MSAR, under Grant No. 078/2015/A3.

References

[1] Y. Cao, Y. Li, S. Coleman, A. Belatreche, T. McGinnity, Adaptive hidden markov model with anomaly states for price manipulation detection, IEEE Trans. Neural Netw. Learn. Syst. 26 (2) (2015) 318–330, doi:10.1109/TNNLS.2014.2315042.
[2] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, ACM Comput. Surv. 41 (3) (2009) 1–58, doi:10.1145/1541880.1541882.
[3] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection for discrete sequences: a survey, IEEE Trans. Knowl. Data Eng. 24 (5) (2012) 823–839, doi:10.1109/TKDE.2010.235.
[4] D.Q. Chen, Anomaly detection boundary based on the moving averages of markov chain model, in: Proc. IEEE Intl Conf. Fuzzy Systems and Knowledge Discovery, 2015, pp. 1532–1536, doi:10.1109/FSKD.2015.7382172.
[5] Y.X. Chen, X. Dang, H. Peng, H. Bart, Outlier detection with the kernelized spatial depth function, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 288–305, doi:10.1109/TPAMI.2008.72.
[6] F. Esponda, S. Forrest, P. Helman, A formal framework for positive and negative detection schemes, IEEE Trans. Syst. Man Cybern.Part B (Cybernetics) 34 (1) (2004) 257–273, doi:10.1109/TSMCB.2003.817026.
[7] G. Florez, Z. Liu, S. Bridges, A. Skjellum, R. Vaughn, Lightweight monitoring of MPI programs in real time, Concurrency Comput. 17 (13) (2005) 1547–1578, doi:10.1002/cpe.889.
[8] B. Gao, H.Y. Ma, Y.H. Yang, HMMs (hidden markov models) based on anomaly intrusion detection method, in: Proc. IEEE Intl Conf. Machine Learning and Cybernetics, volume 1, 2002, pp. 381–385, doi:10.1109/ICMLC.2002.1176779.
[9] B.C. Geiger, T. Petrov, G. Kubin, H. Koeppl, Optimal kullback–leibler aggregation via information bottleneck, IEEE Trans. Autom. Control 60 (4) (2015) 1010–1022, doi:10.1109/TAC.2014.2364971.
[10] F.A. Gonzalez, D. Dasgupta, Anomaly detection using real-valued negative selection, Genet. Program. Evolvable Mach. 4 (4) (2003) 383–403, doi:10.1023/A:1026195112518.
[11] M. Gupta, J. Gao, C. Aggarwal, J. Han, Outlier Detection for Temporal Data, in: Synthesis Lectures on Data Mining and Knowledge Discovery, 5, Morgan & Claypool, 2014, pp. 1–129, doi:10.2200/S00573ED1V01Y201403DMK008.
[12] R. Gwadera, M. Atallah, W. Szpankowski, Reliable detection of episodes in event sequences, Knowl. Inf. Syst. 7 (4) (2005) 415–437, doi:10.1007/s10115-004-0174-5.
[13] V. Jumutc, J. Suykens, Multi-class supervised novelty detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (12) (2014) 2510–2523, doi:10.1109/TPAMI.2014.2327984.
[14] E. Keogh, J. Lin, A. Fu, H. Van, Finding unusual medical time-series subsequences: algorithms and applications, IEEE Trans. Inf. Technol.Biomed. 10 (3) (2006) 429–439, doi:10.1109/TITB.2005.863870.
[15] E.M. Knorr, R.T. Ng, V. Tucakov, Distance-based outliers: algorithms and application, Int. J. Very Large Data Bases 8 (3–4) (2000) 237–253, doi:10.1007/s007780050006.
[16] J. Lin, E. Keogh, W. Li, S. Lonardi, Experiencing SAX: a novel symbolic representation of time series, Data Min. Knowl. Discov. 15 (2) (2007) 107–144, doi:10.1007/s10618-007-0064-z.
[17] B. Liu, Y. Xiao, P. Yu, L. Cao, Y. Zhang, Z. Hao, Uncertain oneclass learning and concept summarization learning on uncertain data streams, IEEE Trans. Knowl. Data Eng. 26 (2) (2014) 468–486, doi:10.1109/TKDE.2012.235.
[18] Z. Liu, J. Yu, L. Chen, D. Wu, Detection of shape anomalies: a probabilistic approach using hidden markov models, in: Proc. IEEE Intl Conf. Data Engineering, 2008, pp. 1325–1327, doi:10.1109/ICDE.2008.4497544.
[19] M. Markou, S. Singh, Novelty detection: a review part 2: neural network based approaches, Signal Process. 83 (12) (2003) 2499–2521, doi:10.1016/j.sigpro.2003.07.019.
[20] B.A. Moser, T. Natschlager, On stability of distance measures for event sequences induced by level-crossing sampling, IEEE Trans. Signal Process. 62 (8) (2014) 1987–1999, doi:10.1109/TSP.2014.2305642.
[21] H. Ozkan, A. Akman, S. Kozat, A novel and robust parameter training approach for HMMs under noisy and partial access to states, Signal Process. 94 (2014) 490–497, doi:10.1016/j.sigpro.2013.07.015.
[22] H. Ozkan, F. Ozkan, S. Kozat, Online anomaly detection under markov statistics with controllable type-i error, IEEE Trans. Signal Process. 64 (6) (2016) 1435–1445, doi:10.1109/TSP.2015.2504345.
[23] H. Ozkan, O. Pelvan, S. Kozat, Data imputation through the identification of local anomalies, IEEE Trans Neural Netw Learn Syst 26 (10) (2015) 2381–2395, doi:10.1109/TNNLS.2014.2382606.
[24] P. Protopapas, J. Giammarco, L. Faccioli, M. Struble, R. Dave, C. Alcock, Finding outlier light curves in catalogues of periodic variable stars, Mon. Not. R. Astron. Soc. 369 (2) (2006) 677–696, doi:10.1111/j.1365-2966.2006.10327.x.
[25] Y. Qiao, X. Xin, Y. Bin, S. Ge, Anomaly intrusion detection method based on HMM, Electron. Lett. 38 (13) (2002) 663–664, doi:10.1049/el:20020467.
[26] U. Rebbapragada, P. Protopapas, C. Brodley, C. Alcock, Finding anomalous periodic time series, Mach. Learn. 74 (3) (2009) 281–313, doi:10.1007/s10994-008-5093-3.
[27] V. Saligrama, J. Konrad, P. Jodoin, Video anomaly identification, IEEE Signal Process. Mag. 27 (5) (2010) 18–33, doi:10.1109/MSP.2010.937393.
[28] W.Y. Sha, Y.X. Zhu, M. Chen, T. Huang, Statistical learning for anomaly detection in cloud server systems: a multi-order markov chain framework, IEEE Trans. Cloud Comput. (99) (2015) 1, doi:10.1109/TCC.2015.2415813.
[29] W.Y. Sha, Y.X. Zhu, T. Huang, M. Qiu, A multi-order markov chain based scheme for anomaly detection, in: Proc. IEEE Intl Conf. Computer Software and Applications Conference Workshops, 2013, pp. 83–88, doi:10.1109/COMPSACW.2013.12.
[30] S.F. Tian, S.M. Mu, C.H. Yin, Sequence-similarity kernels for SVMs to detect anomalies in system calls, Neurocomputing 70 (4–6) (2007) 859–866, doi:10.1016/j.neucom.2006.10.017.
[31] I.I.N. Azha, N.E.A. Khalid, A. Ismail, N. Sakamat, R.A. Latif, Pattem recognition using pearson correlation on neuron values, in: Proc. IEEE Intl Conf. Control and System Graduate Research Colloquium, 2016, pp. 24–30, doi:10.1109/ICSGRC.2016.7813299.
[32] N.D. Vanli, S.S. Kozat, A comprehensive approach to universal piecewise nonlinear regression based on trees, IEEE Trans. Signal Process. 62 (20) (2014) 5471–5486, doi:10.1109/TSP.2014.2349882.
[33] L.T. Wang, M.G. Mehrabi, E. Kannatey-Asibu, Hidden markov model based tool wear monitoring in turning, J. Manuf. Sci. Eng. 124 (3) (2002) 651–658, doi:10.1115/1.1475320.
[34] H. Wang, M. Tang, Y. Park, C. Priebe, Locality statistics for anomaly detection in time series of graphs, IEEE Trans. Signal Process. 62 (3) (2014) 703–717, doi:10.1109/TSP.2013.2294594.
[35] M. Wang, C. Zhang, J.J. Yu, Native API based windows anomaly intrusion detection method using SVM, in: Proc. IEEE Intl Conf. Sensor Networks, Ubiquitous, and Trustworthy Computing, Vol. 1, 2006, pp. 514–519, doi:10.1109/SUTC.2006.1636219.
[36] H.H. Wei, Y.F. Jia, L. Wang, Spectrum anomalies autonomous detection in cognitive radio using hidden markov models, in: Proc. IEEE Intl Conf. Advanced Information Technology, Electronic and Automation Control Conference, 2015, pp. 388–392, doi:10.1109/IAEAC.2015.7428581.
[37] N. Ye, X. Li, Q. Chen, S. Emran, M. Xu, Probabilistic techniques for intrusion detection based on computer audit data, IEEE Trans. Syst. Man Cybern. Part A Syst. Humans 31 (4) (2001) 266–274, doi:10.1109/3468.935043.
[38] J. Zhang, I.C. Paschalidis, An improved composite hypothesis test for markov models with applications in network anomaly detection, in: Proc. IEEE Intl Conf. Conference on Decision and Control, 2015, pp. 3810–3815, doi:10.1109/CDC.2015.7402811.
[39] M.M. Breunig, H. Kriegel, T.N. Raymond, J. Sander, LOF: identifying density-based local outliers, in: Proc. ACM SIGMOD Intl Conf. On Management of Data, 2000, pp. 93–104, doi:10.1145/342009.335388.
[40] M. Bhuyan, D.K. Bhattacharyya, J.K. Kalita, A multi-step outlier-based anomaly detection approach to network-wide traffic, Inf. Sci. 348 (20) (2016) 243–271, doi:10.1016/j.ins.2016.02.023.

[41] A. Forestiero, Self-organizing anomaly detection in data streams, Inf. Sci. 373 (10) (2016) 321–336, doi:10.1016/j.ins.2016.09.007.
[42] M. Uzam, Z.W. Li, G. Gelen, R.S. Zakariyya, A divide-and-conquer-method for the synthesis of liveness enforcing supervisors for flexible manufacturing systems, J. Intell. Manuf. 27 (5) (2016) 1111–1129, doi:10.1007/s10845-014-0938-z.
[43] Y.F. Chen, Z.W. Li, A. Al-Ahmari, N.Q. Wu, T. Qu, Deadlock recovery for flexible manufacturing systems modeled with petri nets, Inf. Sci. 381 (2017) 290–303, doi:10.1016/j.ins.2016.11.011.