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.

© 2017 Elsevier Inc. All rights reserved.

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/© 2017 Elsevier Inc. All rights reserved.

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

Download English Version:

https://daneshyari.com/en/article/4944330

Download Persian Version:

https://daneshyari.com/article/4944330

Daneshyari.com