Expert Systems With Applications 57 (2016) 324–336

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Self-adaptive statistical process control for anomaly detection in time series

Dequan Zheng[a], Fenghuan Li[b,a,*], Tiejun Zhao[a]

[a] MOE-MS Key Laboratory of Natural Language Processing and Speech, Harbin Institute of Technology, Harbin 150001, PR China
[b] School of Software Engineering, South China University of Technology, Guangzhou 510006, PR China

ARTICLE INFO

Article history:
Received 5 January 2015
Revised 15 March 2016
Accepted 16 March 2016
Available online 24 March 2016

Keywords:
Anomaly detection
Self-adaptive
Statistical process control
Fuzzy set theory
Statistical testing

ABSTRACT

Anomaly detection in time series has become a widespread problem in the areas such as intrusion detection and industrial process monitoring. Major challenges in anomaly detection systems include unknown data distribution, control limit determination, multiple parameters, training data and fuzziness of 'anomaly'. Motivated by these considerations, a novel model is developed, whose salient feature is a synergistic combination of statistical and fuzzy set-based techniques. We view anomaly detection problem as a certain statistical hypothesis testing. Meanwhile, 'anomaly' itself includes fuzziness, therefore, can be described with fuzzy sets, which bring a facet of robustness to the overall scheme. Intensive fuzzification is engaged and plays an important role in the successive step of hypothesis testing. Because of intensive fuzzification, the proposed algorithm is distribution-free and self-adaptive, which solves the limitation of control limit and multiple parameters. The framework is realized in an unsupervised mode, leading to great portability and scalability. The performance is assessed in terms of ROC curve on university of California Riverside repository. A series of experiments show that the proposed approach can significantly increase the AUC, while the false alarm rate is improved. In particular, it is capable of detecting anomalies at the earliest possible time.

1. Introduction

Anomaly detection in time series provides significant information for numerous applications. For example, it can be used to detect intrusions in network data (Abadeh, Mohamadi, & Habibi, 2011), fraud detection (Ahmed, Mahmood, & Islam, 2016), incident faults in industrial process (Brighenti & Sanz-Bobi, 2011). Anomalies in time series can manifest in terms of the changes in the amplitude of data, or can be associated with the changes in the shape of temporal waveforms. In light of this, we categorize anomalies into two types: anomalies in amplitude and anomalies in shape. For example, it is an anomaly in amplitude that is a premature ventricular contraction in electrocardiogram (ECG) signals in Fig. 1 and it is an anomaly in shape that is a premature Poppet withdrawal in a Space Shuttle Marotta Valve time series shown in Fig. 2. These anomalous parts are highlighted in red in both figures.

Anomalies are time series that are the least similar to all other time series and depart from the bounds of the state of statistical control which exists when certain critical process variables remain close to their target values and do not change perceptibly. Time series that stay in a state of statistical control are called in-control data (normal data), otherwise, are called out-of-control data (anomaly). In statistical process control, control charts are used to determine if a process is in a state of statistical control. As shown in Fig. 3, a control chart consists of:

(1) Points representing a statistic of measurements of a quality characteristic in samples taken from the process at different times or different data.
(2) The mean of this statistic using all the samples at which a center line is drawn.
(3) Upper control limits (UCL) and lower control limits (LCL) that indicate the threshold at which the process output is considered statistically 'unlikely'.

Anomaly detection in time series is more challenging due to several reasons. First, it makes control limits very important decision aids. Control limits provide information about the process behavior and have no intrinsic relationship to any specification targets. In practice, the process mean (the center line) may not coincide with the specified value of the quality characteristic, because the process design simply cannot deliver the process characteristic

[*] Corresponding author. Tel.: +8645186412449606; fax: +8645186412449608.
E-mail addresses: dqzheng2007@gmail.com (D. Zheng), finelee2012@gmail.com (F. Li), tjzhao@hit.edu.cn (T. Zhao).

http://dx.doi.org/10.1016/j.eswa.2016.03.029
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

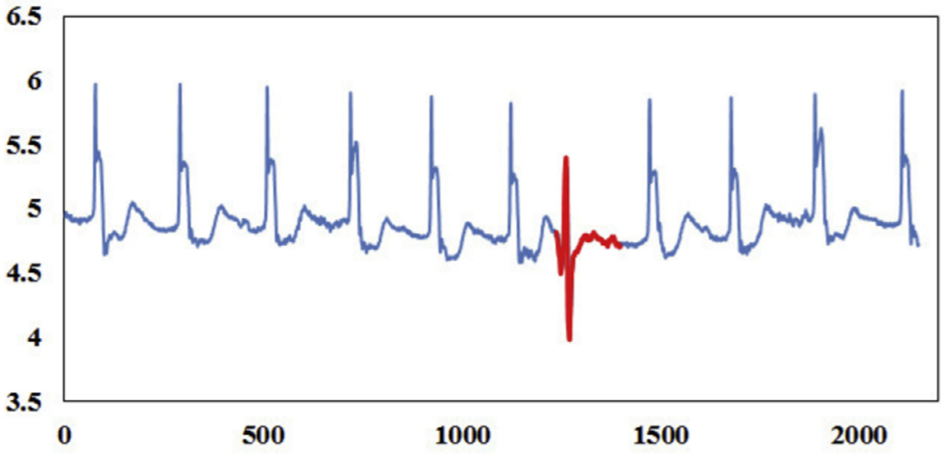

Fig. 1. The time series anomaly found in an excerpt of electrocardiogram.

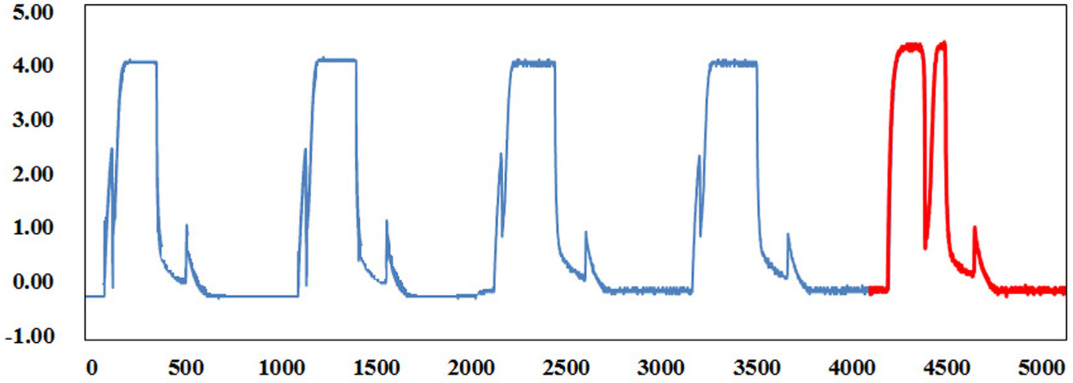

Fig. 2. An example of an annotated Marotta Valve time series.

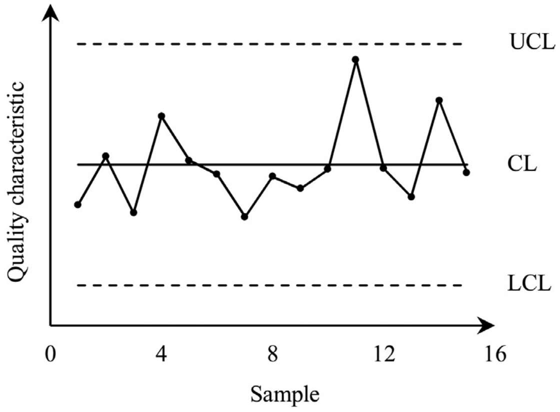

Fig. 3. An example of control chart.

at the desired level. It is also a key challenge to select a threshold instead of process mean. Second, anomaly is a more complex concept. For example, if one sample's characteristic is equal to UCL - ε (ε is an infinitesimal positive number), it is normal. But if one sample's characteristic is equal to UCL + ε, it becomes difficult to determine whether it is normal or abnormal. Third, many other algorithms require several parameters whose values are to be determined. This requires to acquire large amounts of training data, therefore, most of algorithms are realized in the supervised mode.

Due to these major challenges including unknown data distribution, control limit determination, multiple parameters, training data and fuzziness of 'anomaly' in anomaly detection systems, a synergistic combination of statistical and fuzzy set-based technique is proposed in this paper. We view anomaly detection as a statistical hypothesis testing and introduce a definition based on control chart in statistical process control. Because the process mean may not coincide with the specified value of the quality, we do not adopt the mean of samples' characteristic, but a threshold. Anomaly could be a more complex concept, so the threshold should be fuzzy. Fuzzy set theory is taken into account to provide a better characterization of the boundary between normal and abnormal. What's more, the inequality ($>$, $\leq$) in statistical hypothesis test is treated as a fuzzy predicate (the degree of inclusion). Intensive fuzzification process is adopted to realize related parameters determination which is self-adaptive. Therefore, the values of parameters are not required to be specified by the user. Due to the use of fuzzy set theory, statistical hypothesis testing in this paper is a distribution-free and totally unsupervised model. What's more, the overall scheme is self-adaptive. The utility is demonstrated using synthetic and real data sets. We have conducted a number of studies that show the effectiveness of our algorithm to detect anomalies in time series data.

The paper is structured as follows. Section 2 reviews some previous works on anomaly detection. Section 3 illustrates how anomaly detection can be viewed as a statistical hypothesis testing. A fuzzy-statistical algorithm for detecting anomalies is described in Section 4. We present some applications and perform extensive