Measurement 95 (2017) 280–292

Contents lists available at ScienceDirect

Measurement

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

Anomaly detection based on uncertainty fusion for univariate monitoring series

Jingyue Pang, Datong Liu *, Yu Peng, Xiyuan Peng

Department of Automatic Test and Control, Harbin Institute of Technology, Harbin 150080, China

ARTICLE INFO

Article history:
Received 6 January 2016
Received in revised form 22 September 2016
Accepted 12 October 2016
Available online 13 October 2016

Keywords:
Anomaly detection
Monitoring data series
Gaussian Process Regression
Model uncertainty
k-Fold cross validation

ABSTRACT

Detecting the anomalies timely in the condition monitoring data, which are highly relevant to the potential system faults, has become a research focus in many domains. Among the various detection methods available, the prediction-based algorithms are popular without using prior knowledge and expert labels. Additionally, these methods can take the time-ordered specialty into account which is highly significant for time-series-based anomaly detection. However, the detected feedback is binary, especially, due to the influence of inaccurate confidence interval (CI), the false alarm phenomena occur frequently based on predicted models. Thus this paper proposes an Uncertainty Fusion method to realize anomaly detection. Firstly, in order to estimate the data uncertainty, the Gaussian Process Regression (GPR) is applied to perform the prediction with uncertainty presentation. Then, based on the GPR model, the improved k-fold cross-validation is combined to represent the model uncertainty. Moreover, the quantitative error index is designed to provide more detecting information for decision-making. Eventually, the effectiveness of the proposed method are verified by different simulated and open-source data sets, as well as the real application in mobile traffic data detecting. The quantitative results on simulated data experiments show the proposed method can largely eliminate the false alarms without sacrificing much detection rate compared with the basic GPR model. Especially, the experiments on periodic data sets with high Signal Noise Ratio have the better performance. And the mobile traffic data detecting proves the Uncertainty Fusion method can expand the basic GPR model to meet the real industrial requirements.

1. Introduction

With the promotion and recent advances in Internet of Things, Cyber-Physical System and Industrial 4.0, condition monitoring has been widely emphasized to monitor the working performance of the objects and complex systems, such as the devices, components and humans in many domains [1,2]. In this case, analyzing the collected data from the condition monitoring system is an essential part to improve the sensory capability of the objective system [3,4]. Moreover, compared with monitoring the normal condition, detecting abnormal events is more significant which can reflect variations in specified technical performance of the related systems [5,6]. Especially, with more and more practical applications such as ECG Anomaly Detection [7], electronic components detection [8], battery capacity anomaly detection [9], the anomaly detection has become the focus of considerable research.

Furthermore, compared with the static data, the monitoring data is generated in the form of successive measurements in a time-ordered fashion, giving rise to time series data. In addition, some monitoring series are acquired by some sensors to track the performance of the related devices. Particularly, there are some key sensors which represent the important health information of the object, such as the temperature in the industrial system, the current and voltage in electrical devices, the vibration signal in mechanical system, and so on. Undoubtedly, the data from these sensors should be monitored independently and timely. So the focus of this paper is on performing abnormal detection for univariate monitoring series.

As to the wide attention for anomaly detection, there are many algorithms available for identifying the deviating points and modes in monitoring series. Some previous work had reviewed the methods [10–12], and which can be roughly divided into six categories referring to simple threshold, statistical analysis, Nearest Neighbor-based methods, cluster-based analysis, classification-based methods, and prediction-based algorithms. They are developed to meet different application requirements.

* Corresponding author.
E-mail addresses: jypang@hit.edu.cn (J. Pang), liudatong@hit.edu.cn (D. Liu), pengyu@hit.edu.cn (Y. Peng), pxy@hit.edu.cn (X. Peng).

http://dx.doi.org/10.1016/j.measurement.2016.10.031
0263-2241/© 2016 Elsevier Ltd. All rights reserved.

Simple threshold approach can quickly detect anomalies requiring virtually no extra CPU resource which can meet the simple application needs. But the thresholds are always set by the users, rather than learned. As a result, some abnormal evolvements within the limits are difficult to be discovered [13]. Certainly, statistical method has better learning ability than simple threshold based on computing the statistical features of the training data. Currently, most statistical methods have been applied to solve a wide variety of issues including intrusion detection [14], fault detection and diagnosis [15], environmental anomaly detection [16], and so on. However, in the real-world application, it is very difficult to model any data sets with certain distribution. Regarding the methods based on Nearest Neighbor, including the Density-based algorithms and Distance-based method, assume that the normal data occur in dense neighborhoods, while the abnormal data are far from the normal data [1]. Cluster-based analysis realizes anomaly detection by determining on the data whether belongs to the normal clusters or not [17]. For Nearest Neighbor-based and cluster-based method, it is very sensitive to the distance calculation function and the noise. Moreover, it is not suitable for detecting the case without enough normal data samples. Classification-based algorithms mainly refer to rule-based [18], neural networks based [19] and support vector machine (SVM) method [20], etc. As supervised methods, they need the corresponding normal and abnormal labels. Nevertheless, in most applications, the labeling process is often costly which needs to perform manually. At the same time, the labels may be still incomplete. As requiring no prior knowledge and consumed labels, as well as detecting timely, prediction-based methods have been widely used for anomaly detection by comparing the real data and the model predicted output. For example, Naïve predictor (Naïve), nearest cluster (NC), Single-layer linear network (LN) and Multilayer perception (MLP) were utilized to realize anomaly detection for streaming environmental data [21]. Certainly, the effective anomaly detection based on predicted models largely depends on the prediction accuracy. Given that there is often not enough prior knowledge and it is costly to label the monitoring data manually. In addition, condition monitoring series are always acquired and stored in time and arrive steamily, the traditional statistic-based, classification-based and Nearest Neighbor methods are not suitable with ignoring the temporal feature. Therefore, this paper makes focus on the prediction-based methods.

However, these monitoring data are often acquired remotely from targets and may suffer from severe noise contamination [22]. And the predicted results will inevitably contain a variety of errors, such as system noise, the error of model and sensor data errors. Moreover, any kind of prognostic methods, whether it is model-driven or data-driven, cannot eliminate all the factors, thus the predicted results must be with uncertainty [23]. There are two ways to realize the prediction with uncertainty output, one is to give the confidence interval of the prediction by combing with the other means, as k-fold cross validation shown in [21]. In another way, several methods support uncertainty presentation as Particle Filter (PF) [24], Relevance Vector Machine (RVM) [25] and Gaussian Process Regression (GPR). Given the nonparametric model and a small number of setting parameters, GPR model, which supports uncertainty presentation with mean and variance output, is adopted to perform anomaly detection to take the data uncertainty into consideration [26].

Even with the GPR model, the detection still faces the challenge of high false alarms due to the influence of model error [27]. So how to quantify the model error is the key to realize anomaly detection. In [28], the authors reviewed state-of-the-art measures of surrogate model error, and developed a good error quantification method for surrogate models named Predictive Estimation of Model Fidelity (PEMF). But the PEMF is designed for quantifying the surrogate model error that can not meet the application of abnormal detection.

In order to further reduce the false detection rate with the inaccurate model accuracy, this paper analyzes the corresponding reason and proposes an improved detection framework to realize the uncertainty fusion of multiple sources. Firstly, based on prediction uncertainty, the model uncertainty is combined into the basic GPR detection framework with the modified 10-fold cross-validation. Furthermore, considering the quantitative feedback can provide more decision-information, the improved anomaly detection strategy is developed by the error index quantification on the basis of the combination of the data and model uncertainty. Finally, some simulated and public data sets are utilized for verifying the effectiveness of the proposed method. In addition, the experiments on real application in mobile traffic data set can demonstrate the detection capability of our proposed method.

2. Anomaly detection based on GPR model

2.1. GPR model

A Gaussian process defines a distribution over functions. For instance, given input data set $D = \{\boldsymbol{x}_i\}_{i=1}^{N}, \boldsymbol{x} \in R^d$, and its corresponding function $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)$ constitute a collection of limited random variables which obey to joint Gaussian distribution, and then $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)$ form a GP described as Eq. (1):

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}_i, \boldsymbol{x}_j)) \tag{1}$$

As to single variable which obeys a Gaussian distribution, whose features depend on the mean and variance, the characteristics of GP are determined by its mean function and covariance function described as follows:

$$m(\boldsymbol{x}) = E[f(\boldsymbol{x})] \tag{2}$$

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = E[(f(\boldsymbol{x}_i) - m(\boldsymbol{x}_j))(f(\boldsymbol{x}_i) - m(\boldsymbol{x}_j))] \tag{3}$$

where $m(\boldsymbol{x})$ and $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ are the mean function and covariance function, respectively. And $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the d-dimension inputs. Among some kinds of covariance function, the most commonly used covariance function is the square exponential function [22]:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \upsilon_0 \exp\left\{-\frac{1}{2}\sum_{l=1}^{d}\omega_l(\boldsymbol{x}_i - \boldsymbol{x}_j)^2\right\} \tag{4}$$

where υ_0 is the setting variance of the model, d is the dimension of input data, and ω_l is the distance size. Moreover, we can define the covariance function which should comply with the nonnegative conditions. The parameters in the mean and covariance function, named hyper-parameters, need to be identified. Generally, it is assumed that the mean of GP is zero everywhere. In this case, the relationship between one observation and another just depends on the covariance function. Therefore, the prior distribution of GP is determined with the setting of initial hyper-parameters and the form of covariance function.

In the view of regression problem, the functional relationship between the d dimensional input variables $\boldsymbol{x}$ and the target variable y should be modeled. Compared with some parametric models which restrict the explicit form of $f(\boldsymbol{x})$ with some unknown parameters, the GPR model just assumes that $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)$ constitute a collection of limited random variables which obey to joint Gaussian distribution. With this assumption, $f(\boldsymbol{x})$ forms a GP as described in Eq. (1). Especially in real-word applications, the target y is always with noise, so the regression problem is showed in the following equation:

$$y = f(\boldsymbol{x}) + \varepsilon \tag{5}$$