ARTICLE IN PRESS

Information Sciences xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry

Jun Zhao *, Kai Liu, Wei Wang, Ying Liu

School of Control Science and Engineering, Dalian University of Technology, China

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Energy system in steel industry
Anomaly detection
Fuzzy clustering
Dynamic time warping

ABSTRACT

The accuracy of the acquired data is very significant for the decision-making process for the purpose of the safety and reliability of the energy system in steel industry. However, owing to the instability and vulnerability of industrial system of supervisory control and data acquisition (SCADA), the anomaly data usually exist in practice. In this study, considering the data feature of the energy system, we classify the anomalies as the trend anomaly for the pseudo-periodic data and the deviants for the generic data. As for the trend anomaly, a dynamic time warping (DTW) based method combining with adaptive fuzzy C means (AFCM) is proposed by referencing the similar industrial processes; while, as for the deviants detection, a *k*-nearest neighbor AFCM algorithm (KNN-AFCM) is designed here for the local anomaly detection for the generic data. To verify the effectiveness of the proposed method, the real-world energy data coming from a steel plant are employed to perform the experiments, and the results indicate that the proposed method exhibits a higher precision compared to the other methods for the anomaly detection.

1. Introduction

The energy system of steel plant is very complicated, which involves a large number of energy generation units, consumption units and storage units. Due to the instability of the industrial SCADA system and the vulnerability of the data transportation equipments, the data anomaly usually exists in various categories of data, such as flow data, volume data and pressure data, which largely impact on the decision-making process of the energy operation, even might lead to serious safety accidents or economic losses. With the advancement of data-driven modeling and optimization technology, the completeness and the correctness of the acquired data play a very significant role on the industrial modeling and decision making process [8], especially the data-based approaches are becoming a research concentration nowadays [5,1,11]. Importantly, the data-driven based energy system modeling for steel industry had been proposed in [24,12,17,23], most of which aimed to optimize the energy operations under the assumption that the acquired data from SCADA system are accurate and reliable. In fact, a large number of the anomaly data are widely existent in the industrial system, which usually results in the negative impacts on the accuracy and stability of the data-driven methodologies. In current practice, the identification of the anomaly relies mainly on the personal experiences of the energy operators, which make this task rather intricate and time-consuming.

In literature, there were some mentioned anomaly detection approaches for time series data. Typically, an immune system based method inspired by a negative selection mechanism was proposed in [6]. However, the reported data coding procedure would lose the meaningful patterns of the original time series; furthermore, the negative set could be useless

* Corresponding author. Tel.: +86 411 84707582.
E-mail address: zhaoj@dlut.edu.cn (J. Zhao).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.05.018

when the diversity of the normal set largely increases. Subsequently, another kind of detection for time series data was also proposed in [10], where the sample data was converted into a symbolic string, and a suffix tree was constructed by a Markov model to obtain a quantitative evaluation criterion. Yet, the similar pattern loss was also generated in the procedure of discretization, which might lose the useful information of the data. Besides, a class of kernel learning based detection was presented in [14] by transforming the detection into a support vector machine based prediction process. In that method, since the solution depended directly on the prediction accuracy and the computation was rather time-consuming, the detection results might be unsatisfactory for the real-time application. Similarly, a wavelet-based detection was designed in [16]. Although the above studies claimed the good performance for a sort of periodic sequence data, they were inapplicable to the irregular industrial time series. Recently, a fuzzy clustering-based model was reported in [22] by considering the internal connectivity feature of the data points, and that method paid more attentions to improving the clustering outcomes and mining the outliers in the data, which exhibited a weak ability to detect the anomaly for time series. Besides, one can reference the related works in [13,18–21].

Considering the anomaly characteristics of industrial time series, this study classifies the anomalies of the energy system in steel industry into (1) the trend anomaly of the pseudo-periodic data and (2) the deviants anomaly of the general data, and the corresponding AFCM-based detection methods are then designed. As for the trend anomaly of the pseudo-periodic data, a dynamic time warping (DTW) based sequence stretching is proposed to transform the similarity of the sequences with unequal length into the Euclidean distance metric. As for the deviants anomaly of the general data, a *k*-nearest neighbor AFCM algorithm (KNN-AFCM), in which the neighborhood effect and the sample weight coefficient are introduced to the objective function, is designed to improve the detection accuracy. To verify the performance of the proposed method, the real-world energy data coming from a steel plant that includes the two kinds of the mentioned abnormal data are then employed, and the experimental results show that the proposed method exhibits the better performance for the anomaly detection of industrial data.

The rest of this paper is organized as follows. In Section 2, the data anomaly problem in the energy system is briefly depicted and analyzed. The detection methods are then proposed for the two categories of anomalies in the energy system in Section 3. The experiments and the comparative analysis are detailed in Section 4 by using a number of practical industrial data from the energy center of a steel plant. Finally, Section 5 draws the conclusions of this study.

2. Problem descriptions

With the development of industrial information technology, most of the steel enterprises have built the SCADA systems to monitor the internal energy conditions so as to optimize the energy operations for saving the production costs. However, due to the malfunctions such as sensor failure, communication interruption, storage exception and shutdown of acquisition program, a large number of data anomalies exist in the industrial system, which might misguide the energy scheduling operators in the decision-making process.

Since the industrial data collected from the energy units exhibit a variety of manufacturing characteristics, their dynamics features are rather different. For example, the generation flow of the blast furnace gas (BFG) and its consumption by the hot blast stove always show a pseudo-periodic feature, whose amplitude and frequency fluctuate in a certain range, illustrated in Fig. 1a and b; and the pressure of the coke oven gas (COG) pipeline networks usually exhibits irregular trend without violent fluctuation, shown in Fig. 1c and d. With respect to the data anomaly, a number of typical anomalies generated in the process of data acquisition can be illustrated as Fig. 2, in which the trend anomalies can be presented like Fig. 2a–c, and the anomalies coming from the irregular data with amplitude deviation from its neighborhood can be usually regarded as the contextual anomaly [4] or deviants [9], see Fig. 2d.

3. Anomaly detection based on AFCM

With a long term data analysis on-site, it is discovered that the acquired data by the SCADA system exhibit obvious classifiable features. As for the pseudo-periodic data in the energy system, its dynamics reflect the similar production operation in manufacturing process; while as for the generic energy data, the anomaly deviants can be effectively identified by its neighborhood tendency. As such, aiming at the data anomaly of the energy system, the proposed AFCM based methods transform the detection process into a clustering problem on the basis of the data characteristics of the industrial system.

3.1. Adaptive fuzzy C means

AFCM aimed originally at the data clustering with the objective function formulated as [22],

$$\begin{aligned} &J_{AFCM} = \sum_{j=1}^{n} w_j^p \sum_{i=1}^{c} u_{ij}^m d_{ij}^2 \\ &s.t. \sum_{j=1}^{n} u_{ij} > 0, \quad \sum_{i=1}^{c} u_{ij} = 1, \quad \prod_{j=1}^{n} w_j = 1 \end{aligned} \tag{1}$$

Please cite this article in press as: J. Zhao et al., Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry, Inform. Sci. (2013), http://dx.doi.org/10.1016/j.ins.2013.05.018