ISA Transactions ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

ISA Transactions

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

Research article

# A hybrid clustering approach for multivariate time series – A case study applied to failure analysis in a gas turbine

Cristiano Hora Fontes [a,*], Hector Budman [b]

[a] Graduate Program in Industrial Engineering, Polytechnic School, Federal University of Bahia, Brazil
[b] Department of Chemical Engineering, University of Waterloo, Canada

ARTICLE INFO

*Article history:*
Received 6 July 2016
Received in revised form
4 July 2017
Accepted 5 September 2017

*Keywords:*
Multivariate time series
Fuzzy clustering
Fault detection
Gas turbine
PCA-based similarity
Oversampling

ABSTRACT

A clustering problem involving multivariate time series (MTS) requires the selection of similarity metrics. This paper shows the limitations of the *PCA* similarity factor (*SPCA*) as a single metric in nonlinear problems where there are differences in magnitude of the same process variables due to expected changes in operation conditions. A novel method for clustering MTS based on a combination between *SPCA* and the average-based Euclidean distance (*AED*) within a fuzzy clustering approach is proposed. Case studies involving either simulated or real industrial data collected from a large scale gas turbine are used to illustrate that the hybrid approach enhances the ability to recognize normal and fault operating patterns. This paper also proposes an oversampling procedure to create synthetic multivariate time series that can be useful in commonly occurring situations involving unbalanced data sets.



## 1. Introduction

In industrial processes, technologies for the acquisition of large sets of data are mature and encourage the use of Data Mining (DM) approaches to extract knowledge from these data [1]. Time series are widely used in the fields of process engineering, medicine, finance, among others [2,3] and they are an important class of data objects. One of the mining tasks associated to this kind of data comprises pattern discovery and clustering [4].

Pattern recognition in univariate time series has been investigated [5,6] using standard approaches [4,5,7]. This kind of clustering problem can be solved using batch clustering models [8] and the theory of similarity in univariate time series, generally based on the Euclidean or the Dynamic Time Warping (DTW) distances [9,10] is fairly well developed. Pattern recognition in multivariate time series (MTS) is a more complex problem (non-point prototyping problem) with intrinsic features such as similarity measures, cluster validity and domain knowledge [11,12]. This kind of problem cannot be solved directly using classical algorithms of batch clustering such as Fuzzy C-Means. The challenge for analyzing multivariate time series lies in its multivariate nature because the adoption of dimensionality reduction approaches may lead to a loss of information, reducing the ability to detect joint features or information hidden in the variables as a whole [9,12,13].

Some studies on MTS are based on the measurement of similarity between two multivariate time series. Considering that a multivariate time series can be stored in a matrix, some works propose new similarity metrics based on extended versions of the Frobenius norm [12] also by classifying similarity measures into internal factors and external structure [9]. Plant et al. [13] present a model-based approach and consider the similarity between mathematical models that describe the relationship among variables in each MTS. Other works highlight the use of PCA-based similarity metrics (*SPCA*) in pattern recognition involving MTS and illustrate them in nonlinear dynamic case studies. Singhal and Seborg [14] apply modified *SPCA* in a pattern matching approach for fault detection. Other strategies comprise the development of similarity metrics also based on the PCA to improve the performance of this technique or to address its shortcomings in applications involving nonlinear systems. Khediri et al. [15] and Deng et al. [16] propose alternative to use the Kernel Principal Component Analysis (KPCA). Dobos and Abonyi [17] apply dynamic PCA in problem involving time series segmentation. Harrou et al. [18] propose a metric that combines PCA and Multivariate Cumulative Sum (MCUMSUM). Singhal and Seborg [11] present a hybrid approach for clustering multivariate time series data applying a linear combination of the *SPCA* and another similarity metric based on the Mahalanobis distance leading to a modified version of the K-means algorithm. The method comprises a moving-window approach in which patterns (snapshot data) are matched in time-series databases. Therefore,

* Corresponding author.
*E-mail addresses:* cfontes@ufba.br (C.H. Fontes), hbudman@uwaterloo.ca (H. Budman).

http://dx.doi.org/10.1016/j.isatra.2017.09.004
0019-0578/

the method does not follow a clustering approach and the patterns must be arbitrarily chosen by the user or previously determined through an expert support. A recent work [3] proposes the use of a Fuzzy clustering approach (Fuzzy C-Means and Fuzzy C-Medoids) combined with a Dynamic Time Warping (DTW) technique to compare MTS [2,19]. The method of [3] is based on univariate time series but does not consider multivariate time series as considered in the current study. Surveys of times series data clustering and analysis are reported in [4,5,20] and [21]. Bankó and Abonyi [2] consider two similarity metrics, namely, DTW and SPCA, and the correlation among Multivariate Time Series is based on a segmentation approach. Since this method does not consider the AED of the multivariate series, it is limited for a certain class of problems as shown in our first case study. Two works present the application of the classical FCM algorithm to cluster Multivariate Time Series [22,23]. In these works the points of the variables at each time instant are clustered using the classical Euclidean distance. The authors propose instantaneous classification of MTS which in turn requires the use of an additional strategy/method to cluster the different instantaneous classifications and their correspondence over time.

The various methods employed in problems involving Fault Detection and Diagnosis (FDD) can be classified into three general categories, namely, quantitative model-based methods, qualitative model-based methods, and process history based methods [24]. Additionally, techniques based on fuzzy logic and Artificial Neural Networks, or a hybrid of these [25], provide useful alternatives to cope with the fault detection problem. A large body of work have been carried out in recent years into the reliability and operational availability in industrial plants such as power generation plants [26–35], but few works present the use of clustering approaches applied to time series in order to detect and/or prevent failures in a process or piece of equipment. In the specific case of failure analysis in gas turbines, empirical model-based approaches have been proposed [36,37]. However, empirical models are limited in terms of their ability to correctly capture complex interactions between variables and the system's nonlinearity. In this case, approaches based on historical process data [38] can provide a feasible and useful alternative to recognize failures and/or normal operation patterns.

When dealing with clustering of multivariate time series (MTS) with fault detection as the primary goal, the selection of a suitable measure of similarities/dissimilarities between objects (MTS) is necessary so as to enhance the resulting classification ability [3,39]. A single similarity metric is only capable of computing differences based on one of the features associated to the objects (MTS) thus limiting the clustering ability. This paper presents a novel method for clustering multivariate time series based on a combination between two similarity metrics, namely, the modified PCA similarity factor [11,14] and the average-based Euclidean distance (*AED*) which are combined within a fuzzy clustering approach. The use of this novel hybrid metric is driven by our findings that the similarity PCA or the AED do not work well for a particular class of nonlinear problems. The proposed hybrid metric compares different MTS based on the weighted sum of these two different metrics, i.e. the weighted sum of the direction of principal components and the averages. Beyond its Fault Detection (FD) capabilities as illustrated later in the case studies, the proposed algorithm could be used to generate desired reference trajectories for control to ensure good operation. It should be emphasized that the proposed algorithm has been devised, in view of the target application of a startup process, for batch processes where the entire MTS during the batch are used for assessing similarity or dissimilarity between MTSs between different batches. The procedure could be modified to address continuous processes by considering moving windows and considering together the MTSs corresponding to different time periods but this is beyond the scope of the current work.

Two case studies related to the fault detection problem are presented. The first case involves three simple numerical examples which were tailored to elucidate the relevance of the hybrid approach. These numerical examples are used to show the limitations of either the similarity PCA or the AED when used separately for particular nonlinear problems. The combination of the two metrics can therefore improve detection. The second case study is a real industrial scenario which involves the recognition of patterns that result in either successful or faulty start-ups of a gas turbine of commercial scale. This piece of equipment is the main section of a thermoelectric unit operating in the industrial park of a Brazilian Oil Company (Brazil).

This paper is structured as follows. In Section 2, basic definitions about multivariate time series, modified PCA similarity factor and Fuzzy C-Means method are reviewed. In Section 3, the hybrid metric and a formulation for the clustering problem are presented. Section 4 presents the application of the hybrid metric on three simple examples that differ in terms of the direction between principal components and the average of the corresponding time series in the data. Section 5 presents the industrial application of fault detection in a gas turbine followed by discussion and conclusions.

## 2. Preliminaries

### 2.1. Multivariate time series and PCA similarity factor

Two kinds of objects are generally considered for clustering and pattern recognition of time series, namely, the Univariate Time Series (UTS) and Multivariate Time Series (MTS) [40,41]. Considering a general series of observations over time (time series) associated with a specific process variable $\boldsymbol{z}_j(t)\,(j = 1, \ldots, p;\ t = 1, \ldots, m)$ where $p$ is the number of variables (number of sensors), $m$ is the number of observations and $t$ indexes the measurements made at each time instant, an MTS object comprises the case in which $p \geq 2$ and can be represented by the following $m \times p$ matrix:

$$\boldsymbol{Z}_i = \begin{bmatrix} z_{i1}(1) & \cdots & z_{ip}(1) \\ \vdots & \ddots & \vdots \\ z_{i1}(m) & \cdots & z_{ip}(m) \end{bmatrix} \tag{1}$$

where $\boldsymbol{Z}_i$ is the object, $z_{ij}(t)$ is the measurement of variable $j$ ($j = 1,\ldots,p$) at time instant $t$ ($t = 1,\ldots,m$) in the object $\boldsymbol{Z}_i$ ($i = 1,\ldots,n$ objects). The column $j$ contains the time series related to the variable $j$.

The PCA (Principal Component Analysis) similarity metric (*SPCA* index, [12,17]) measures the level of similarity between two different objects (MTS represented by two matrices $m \times p$, respectively) with the same number of variables ($p$) but not necessarily the same number of observations ($m$). *SPCA* is based on the principal components of each MTS. The number of principal components associated to each MTS (respectively $k_1$ and $k_2$) is chosen in order to represent at least 95% of the total variance in each object [42,43]. The variance related to each component can be computed directly by the eigenvalues of the covariance matrix associated to MTS. The PCA performed for each time series is mean centered. The original version of *SPCA* index comprises the following expression:

$$SPCA(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{k_0} \cdot \sum_{i=1}^{k_0} \sum_{j=1}^{k_0} \cos^2\theta_{ij} \tag{2}$$

Please cite this article as: Fontes CH, Budman H. A hybrid clustering approach for multivariate time series – A case study applied to failure analysis in a gas turbine. ISA Transactions (2017), http://dx.doi.org/10.1016/j.isatra.2017.09.004