Engineering Applications of Artificial Intelligence 71 (2018) 113–124

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

N-dimensional extension of unfold-PCA for granular systems monitoring

Llorenç Burgas *, Joaquim Melendez, Joan Colomer, Joaquim Massana, Carles Pous

University of Girona, Campus Montilivi, P4 Building, Girona, E17071, Catalonia, Spain

ARTICLE INFO

Keywords:
Principal component analysis
Unfold-PCA
MPCA
Building energy monitoring
Data mining
Statistical process monitoring

ABSTRACT

This work is focused on the data based modelling and monitoring of a family of modular systems that have multiple replicated structures with the same nominal variables and show temporal behaviour with certain periodicity. These characteristics are present in many systems in numerous fields such as the construction or energy sector or in industry. The challenge for these systems is to be able to exploit the redundancy in both time and the physical structure.

In this paper the authors present a method for representing such granular systems using N-dimensional data arrays which are then transformed into the suitable 2-dimensional matrices required to perform statistical processing. Here, the focus is on pre-processing data using a non-unique folding–unfolding algorithm in a way that allows for different statistical models to be built in accordance with the monitoring requirements selected. Principal Component Analysis (PCA) is assumed as the underlying principle to carry out the monitoring. Thus, the method extends the Unfold Principal Component Analysis (Unfold-PCA or Multiway PCA), applied to 3D arrays, to deal with N-dimensional matrices. However, this method is general enough to be applied in other multivariate monitoring strategies.

Two of examples in the area of energy efficiency illustrate the application of the method for modelling. Both examples illustrate how when a unique data-set folded and unfolded in different ways, it offers different modelling capabilities. Moreover, one of the examples is extended to exploit real data. In this case, real data collected over a two-year period from a multi-housing social-building located in down town Barcelona (Catalonia) has been used.

1. Introduction

One of main challenges in industry's current transformation to the Industry 4.0 paradigm is to integrate, manage, process and exploit process data to benefit business. While the internet of Things (IoT) paradigm provides the infrastructure required for integration and management, data mining provides the background for processing according to the required exploitation goals. This paper focuses on the goal of such monitoring and assumes that a multivariate data mining technique is used for that purpose. In fact, the paper assumes that Principal Component Analysis (PCA) is the underlying principle to perform the monitoring and it focuses on the problem of organizing data to apply PCA. This method is also general enough to be applied to other multivariate monitoring strategies.

PCA is a well-known multivariate statistical technique which is not only widely used for dimensional reduction, but also for modelling and monitoring continuous processes based on observations provided by sensors (Russell et al., 2000; Edward Jackson and Mudholkar, 1979). PCA helps to control the processes by using the Hotelling's T^2 and SPE indices to provide charts to detect and analyse faults. The isolation of those faults is made with the contribution analysis (Kourti, 2005). However, as many other statistical methodologies, PCA requires a 2D matrix organization of data where columns represent variables and rows observations. Thus, models obtained with this technique gather correlations between the variables according to the observations (conveniently organized into rows) and assume independence between them. In monitoring applications, these observations usually refer to a single time instant (continuous processes). However, variations of PCA for monitoring include extensions for batch process monitoring based on Multiway PCA (MPCA, Nomikos and MacGregor, 1994) and other variants to address real-time (R-PCA, Yu et al., 2017), and outlier detection in an IoT context (Peter He et al., 2017).

The Multiway approach extends the concept of single instant observations to observations that have a temporal extension (typically the duration of the execution of the batch process) and consequently,

* Corresponding author.
E-mail addresses: llorenc.burgas@udg.edu (L. Burgas), joaquim.melendez@udg.edu (J. Melendez), joan.colomer@udg.edu (J. Colomer), joaquim.massana@udg.edu (J. Massana), carles.pous@udg.edu (C. Pous).

https://doi.org/10.1016/j.engappai.2018.02.013
Received 1 February 2017; Received in revised form 2 February 2018; Accepted 19 February 2018

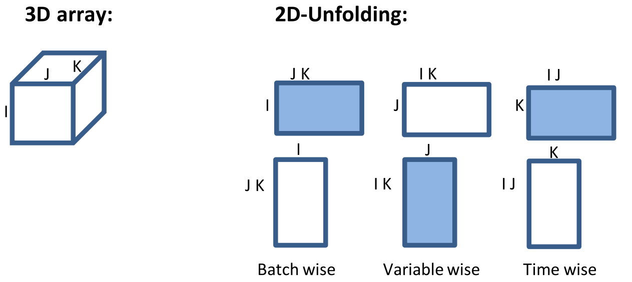

Fig. 1. Graphic representation of all the unfolding possibilities of a 3D matrix.

observations, instead of simple rows, are represented by 2D arrays (variables × samples acquired during the batch execution) and by adding one new dimension to the historic data structure, it now becomes a 3D matrix. Thus, the dimensions of this 3D matrix, containing the historic data of a batch process, are defined by the number of variables being monitored in the process, J, the number of samples acquired at each execution of the batch process, K, and the number of executions included as historic data, I. Again, the I observations represented by these 2D arrays ($I \times K$) containing the data for the monitored variables of a complete execution of a batch process, are assumed to be independent.

Independent of how complex the observations are, the fundamental principles of PCA do not change, but reorganizing (unfolding) the data under study (i.e. to be modelled) into a 2D matrix is required. This implies that, in the case of batch processes, an unfolding preprocessing of the data is required to convert the 3D matrix into a 2D array before applying PCA. This unfolding process is not unique and, depending on how it is done, the interpretations of the results after applying PCA can differ substantially. Thus, there are six known possible combinations to unfold a 3D matrix into a 2D matrix, (see Fig. 1) and not all of them provide interpretable results. (NB: in fact, for the PCA purposes, there are only really three combinations, because half of them are simply the other half transposed.) In batch process monitoring (Nomikos and MacGregor, 1994) variable-wise unfolding ($I \times J \times K \rightarrow IK \times J$, observations in rows are all the samples acquired during the execution of batches) and the batch-wise unfolding ($I \times J \times K \rightarrow K \times IJ$, where observations in rows represent completed batches and number of columns extends to the variables at every time instant, IJ, during the execution of a batch) are commonly used. In other domains, such as monitoring energy in housing buildings for example, time-wise unfolding can also be meaningful (see, for instance, Burgas et al., 2015) to identify singularities in the power consumption of dwellings.

However, there are situations where 3D arrays are not suitable for organizing historic observations and higher dimensional data arrays, or hypercubes, need to be used instead. The need to analyse and model this complex data as a whole, requires developing of a clear methodology to manage the folding/unfolding procedures (as well as other preprocessing measures) for N-dimensional arrays to make them suitable for building interpretable and exploitable PCA models. This occurs, for example, when observations contain not only information from continuous sensors, but also images or spectroscopic information evolving through time where tensor-based dimension reduction techniques are used (Lu et al., 2008; Chen and Shapiro, 2009). A similar situation transpires when considering processes, or systems in a general way, with multiple replicated structures being monitored with the same set of nominal variables (e.g. solar fields and wind farms, injection and assembly lines, cavities in a mould, inkwells in offset industrial printers, power consumers in a grid, or monitoring stores in a mall or rooms in a hotel, etc.). A new challenge appears, one that consists of monitoring not only every subsystem, but also the interactions between them and through time.

Consequently, this requires monitoring tools to be developed that are not only capable of automatically detecting the significantly differently operating elements in any subsystem (e.g. sensor faults, faulty components, performance reduction, misbehaviour detection, etc.) but that also monitor the interactions between these elements and detect any emergent behaviours. By considering modular replication as a new dimension in the data structure this analysis can be carried out, but first requires the adequate pre-treatment and management of the data. Similarly, when an operating continuous system presents a repetitive or periodic behaviour through time, this introduces a degree of redundancy that can be exploited when monitoring. This happens, for instance, in many systems that operate 24/7, but accommodate this operation accordingly due to, for example, shifts, power prices, seasons, solar illumination, etc. Examples of systems with this kind of pseudo-periodic temporal pattern (daily, weekly, seasonally, etc.) are, again, solar fields and wind farms, process industries, or hotels and tertiary buildings affected by daily variations. Such repetitive operations allow models to be built that can then be used as references for monitoring on different time scales or granularity (hourly, daily, weekly, etc.). An example of a multivariate analysis considering this temporal pattern in academic buildings is presented by the authors in Burgas et al. (2014).

Thus, organizing data into multi-dimensional arrays (usually dimension higher than four) is required for data from large systems built on the principle of repetitive modularity and periodic behaviour. This paper aims to provide a method for constructing multivariate models that will monitor such systems as a whole and allow MPCA methodology to deal with N-dimensional arrays. Because the methodology proposed is focused on a previous stage of the PCA modelling itself, then it can be useful not only for PCA modelling and monitoring, but also for other Data Mining tools, such as PLS (Partial least squares). Therefore, this work focuses on the pre-processing stage and, in particular, analysing the significance of the models obtained once specific unfolding strategies have been applied.

This introduction is followed by a background section that includes related work. Following on form that, the methodology to deal with N-dimensional arrays is introduced and the procedure to follow before applying PCA is explained step-by-step. The paper then describes an example of the application and a complete, real exploitation use case is depicted to illustrate the different models that can be obtained from an initial data set and their interpretation and use for monitoring purposes. The paper ends with a section devoted to conclusions and future work.

2. Background and related work

PCA is a method that allows linear dependencies between the variables of a system to be modelled (Russell et al., 2000; Edward Jackson and Mudholkar, 1979). Data gathered during normal operating conditions (NOC) is usually used to obtain a reference model in a new space of lower dimensionality (for instance, waste-water treatment plants as in Aguado and Rosen, 2008). Once the system has been modelled, the new observations projected onto the model's subspace can be used to verify its consistency. Usually two statistics, Hotelling's T^2 and SPE (Square Prediction Error), both defined in the model subspace, are used as the bounds of the model to check if any new observations fall inside or outside the model's thresholds. Hotelling's T^2 indicates how far an observation is from the centre of the model and SPE specifies to what extent the correlations mismatch the ones modelled. Those falling outside the model are considered faulty. Optionally, by using a contribution analysis it is possible to isolate the variables responsible for the deviation outside the statistical thresholds (Kourti, 2005). Currently, there are variations of PCA such as R-PCA (Recursive principal component analysis) in Yu et al. (2017) for sensor outlier detection or monitoring (Peter He et al., 2017) in an IoT scope, that meet the challenges that real-time presents. A complete comparison and study of PCA and its variations can be found in Camacho et al. (2008a, 2008b) and González-Martínez et al. (2014).

However, PCA itself, as with many other data modelling and mining techniques, operates over two-dimensional data matrices organized as $observations(rows) \times variables(columns)$. Some extensions of PCA (for