Outlier Identification in Multivariate Time Series

Joana Patrícia Bordonhos Ribeiro

September, 2017

With the technological development, there is an increasing availability of data. Usually representative of day-to-day actions, the existence of large amounts of information has its own interest when it allows to extract value to the market. In addition, it is important to analyze not only the available values but also their association with time.

The existence of abnormal values is inevitable. Usually denoted as *outliers*, the search for these values is commonly made in order to exclude them from the study. However, outliers often represent a goal of study. For example, in the case of bank fraud detection or disease diagnosis, the central objective is to identify the abnormal situations.

Throughout this dissertation we present a methodology that allows the detection of outliers in multivariate time series, after application of classification methods. The chosen approach is then applied to a real data set, representative of boiler operation. The main goal is to identify faults. It is intended to improve boiler components and, hence, reduce the faults.

The implemented algorithms allow to identify not only the boiler faults but also their normal operation cycles. We aim that the chosen methodologies will also be applied in future devices, allowing to improve real-time fault identification.

The main difficulty of the proposed problem is that the data is from multivariate time series with several specificities that common machine learning algorithms have difficulty to deal with. Some variables are represented in the format of strings, like the boiler name, and others as numerical, such as temperatures. Also, there is a big importance in capturing the evolution of each variable, in order to understand the behaviors, and relate all the variables. Several algorithms have been proposed to deal with time series data but we could not found any able to solve our specific problem. The algorithm studied in this section is responsible for performing a representation of the information present in a time series in a simpler form, enabling an easier implementation of common machine learning algorithms.

Briefly, the chosen algorithm has the purpose of compactly represent the variables evolution of each time series, also performing a reduction of dimensionality without a high loss of information. This representation is made through the construction of vectors representing the region of values assumed by the variable and its tendency in each time segment of the time series. After the construction of the new representation, usual techniques of machine learning that are able to solve the problem of supervised classification under study were applied.

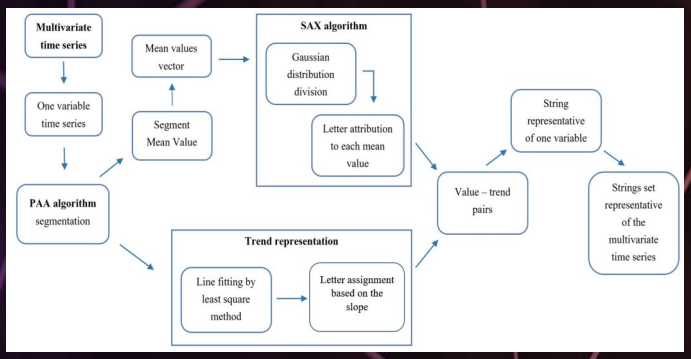

Figure 1: General vision of the value-trend approach.

The studied technique consists of two main steps: the transformation of the time series into sets of vectors through the piecewise aggregate approximation algorithm (PAA) and the conversion of those vectors into a set of letters by the symbolic aggregate approximation algorithm (SAX) [1]. An additional phase where a trend analysis is performed as an improvement of the SAX algorithm is also discussed[2].

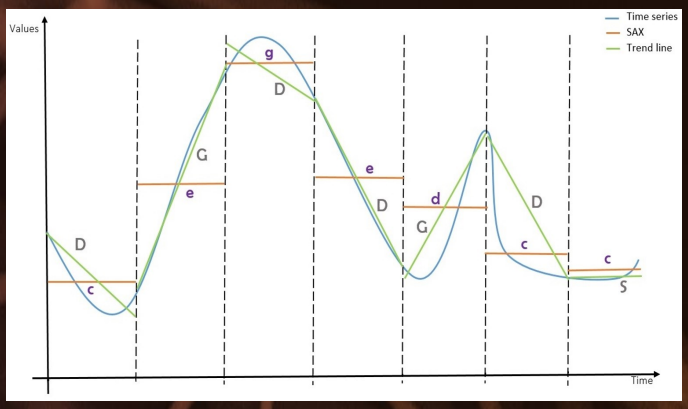

Figure 2: Value-trend approach into a univariate time series.

Using the value-trend algorithm, each variable of the time series was coded as one string. In this way, it was possible to express all the different types of predictive variables into a single one (strings). Also, the behaviors previously represented in a set of matrix lines were transformed into a single string (one line matrix). Therefore, each time series became a single matrix of strings, and so, a multivariate classification problem of string type variables. After that, decision tree and *k*-nearest neighbors models were constructed. Some variations were considered, like the splitting rules and the measures of similarity, in order to test different model parameters and obtain the best results for the performance measures. The decision tree model was the one able to solve the problem.

However, since the results were not satisfying, a final variation was made. Instead of the original 42 labels, the problem was transformed into a binary classification. Therefore, the last algorithm is able to identify the faults, however not being capable of distinguish them into each fault code. This variation presented much better results since the studied data set suffer from number of observations per class.

Processing the data in order to better express the real behaviors of the boilers had became an unexpected difficulty of this work (for example, the representation of the continuity of values in the data matrices). Future work includes the prediction of faults in a space time of one week before. A new metric should have been chosen for the *k*-nearest neighbors model and some techniques to overcome the unbalanced data set can be considered to improve the results of the multi-class problem.

1. J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: A novel symbolic representation of time series", Data Mining and Knowledge Discovery, vol. 15, pp. 107—114, 2 Oct. 2007.

2. B. Esmael, A. Arnaout, R. K. Fruhwirth, and G. Thonhauser, "Multivariate time series classification by combining trend-based approximations", in *Proceedings of the 12th International Conference on Computacional Science and Its Applications—Volumen Part* IV, Springer-Verlage, 2012, pp. 392-403.