Methods and tools for mining multivariate time series

De Gouveia da Costa Cachucho, R.E.

English Summary

Time is a core dimension to understand the evolution of many phenomena. The economic situation of a given country, the current behavior of a customer, the diagnosis of a disease, are just a few examples that can be better understood by looking into time-ordered data.

With the consolidation of the digital development era, the collection and storage of time series data became ubiquitous. Making sense of such time series data can provide insights that help us understand better the present phenomena, extract trends and, in stationary cases, predict future events. Now, in a world increasingly connected by sensors and tractable behaviors, extracting meaningful explanations from time series can be a difficult task. This is why machine learning methods (or more generically, artificial intelligence) became so central nowadays.

Mining time series is a machine learning subfield that focuses on a particular data structure, where variables are measured over (short or long) periods of time. In this thesis we focus on multivariate time series, with multiple variables measured over the same period of time. In most cases, such variables are collected at different sampling rates. As a practical example, consider studying the health of an individual. Variables can be activity behaviors (number of steps, number of floors climbed), biometrics (weight, heart rate), medical records (historical of diseases, metabolites) and economical situation (house owner, contractual situation, income level). When combined, these variables can be explored with machine learning methods for multiple purposes.

In this Ph.D. thesis, we analyze and propose both supervised and unsupervised machine learning methods, which deal with multivariate time series. Firstly, we consider the possibility of unsupervised learning. In this case, we propose a pattern recognition method that discovers subsets of variables that show consistent behavior in a number of shared time segments. Furthermore, when in a supervised setting, given a dependent variable (target),

we propose a method that aggregates independent variables into meaningful features. This method wraps both preprocessing and model learning tasks and generalized to work both for regression and classification problems. Experimental results show the potential of both supervised and unsupervised approaches mentioned.

Additionally to the methods above, we provide two tools in the form of Software as a Service, where users without programming background can intuitively follow the learning and testing methodologies for both methods. The intuition behind these tools is to explore the present role of a data scientist. The machine learning methods we implement are just one of the components of these tools, while the focus is on the user of the tool. The users of such tools are guided trough the whole process of importing data, preprocessing, learning and evaluating results.

Finally, we present an applied study of machine learning to improve speed skating athletes performance. More concretely, we report on a cooperation with the LottoNL-Jumbo team, who have an extensive asset of detailed training data. Here, we make a deep analysis of historical data, in order to help optimize performance results. We combine training and competition results data and make a production ready implementation of a machine learning system to optimize training schedules. In other perspective, we help avoiding under- and overtraining before a given competition.