Predictive Modeling from Data Streams

Olivier Parisot, Benoît Otjacques

Luxembourg Institute of Science and Technology, Belvaux, Luxembourg
olivier.parisot@list.lu

In order to support knowledge extraction from data streams, we propose a visual platform for quickly identifying main features and for computing predictive models in real time. To this end, we have adapted state-of-the-art algorithms in streams learning and visualization.

Nowadays, high frequency data streams are frequent in various domains and they can be used to extract meaningful insights. As a first example, textual data from social media like Twitter and Facebook can be studied to extract the hot topics and to anticipate trends. As a second example, numerical data coming from a network of environmental sensors can be inspected in order to capture events that could precede potential disaster like floods, storms or pollution peaks.

Therefore, numerous data mining techniques have been recently proposed in order to extract predictive models from data streams [1]. On the one hand, classical analytics techniques can be applied on streams by using a certain pool of observations (by using a sliding window, for example). On the other hand, specific online/incremental methods can be applied to dynamically refresh results. A clever data obsolescence strategy is necessary to consider both significant and up-to-date data windows and allow efficient methods (without accessing too much historical data).

In order to improve stream analytics, we have developed a JAVA platform to inspect data streams on-the-fly and to apply the leading predictive models. Various specific third-parties components can be integrated into the software such as WEKA for static data mining or MOA for specific stream processing.

The platform was designed to support two kinds of data sources:

- Remote streams (i.e. available through web APIs): processed on-the-fly.
- Local streams (i.e. obtained from potentially huge files): iteratively processed in a single-pass, without accessing the previous values.

The user interface was designed to be reactive (by plotting on-the-fly the continuously arriving values) and interactive (by providing a real control to the end-user like play/pause/stop the data stream or select the processing speed).

Additionally, various analytics modules were developed in order to continuously inspect the considered data streams.

Firstly, we have implemented a ‘Features similarity’ module to extract the meaningful characteristics from data. More precisely, we have designed an innovative real time Multidimensional scaling 2D projection dedicated to time series, in order to show the correlations (respectively inverse correlations) for the recent history. As an example, this module could help to determine if the IBM and ORACLE stocks quotes are following the same pattern.

Secondly, we have developed a ‘Predictive modelling’ component, to create and refresh models that continuously takes into account recent history. A multitude of techniques exists for predictive analytics, and a critical issue for the data scientist is to select the appropriate technique according to the data characteristics (completion, linear/non-linear relationships, noise, etc.) and the tasks to be carried out. Our aim is to ease the understanding of the predictive models by the user. Therefore, decision trees were selected because they allow building a model both efficient and easy to interpret.

On the one hand, we have applied VFDT, the reference method for classification tree induction. On the other hand, we have used model trees (i.e. decision trees combined to linear regressions) with the recent FIMT-DD algorithm [2] to predict numerical values.

The platform was applied on various real-world data streams (Figure 1). Initially, we have tested our approach on the live stocks quotes from the Yahoo Finances website (CAC40 index – one record per second): it helped us to check the 'Features Similarity' module with real life settings.
Then, we have processed the data from the French electricity transmission system operator (RTE) in order to analyze the energy consumption in France (oil, coal, gas, nuclear, wind, solar, bioenergy, hydraulic and pumping -- 15-min time series). In this case, we have processed heterogeneous values with different scales (for instance: how to use both gas and oil consumptions in order to produce meaningful predictions?)
Finally, we have inspected the hydrological data obtained from the hydrometric stations in Luxembourg: the considered sensors network is composed of 24 stations and produces continuously 15-min time-series [3]. Due to the poor quality of the data, we had to apply techniques that are robust to noise and missing values in sensors data: as an example, the platform was successfully used to fill data gaps in hydrological time series [3].

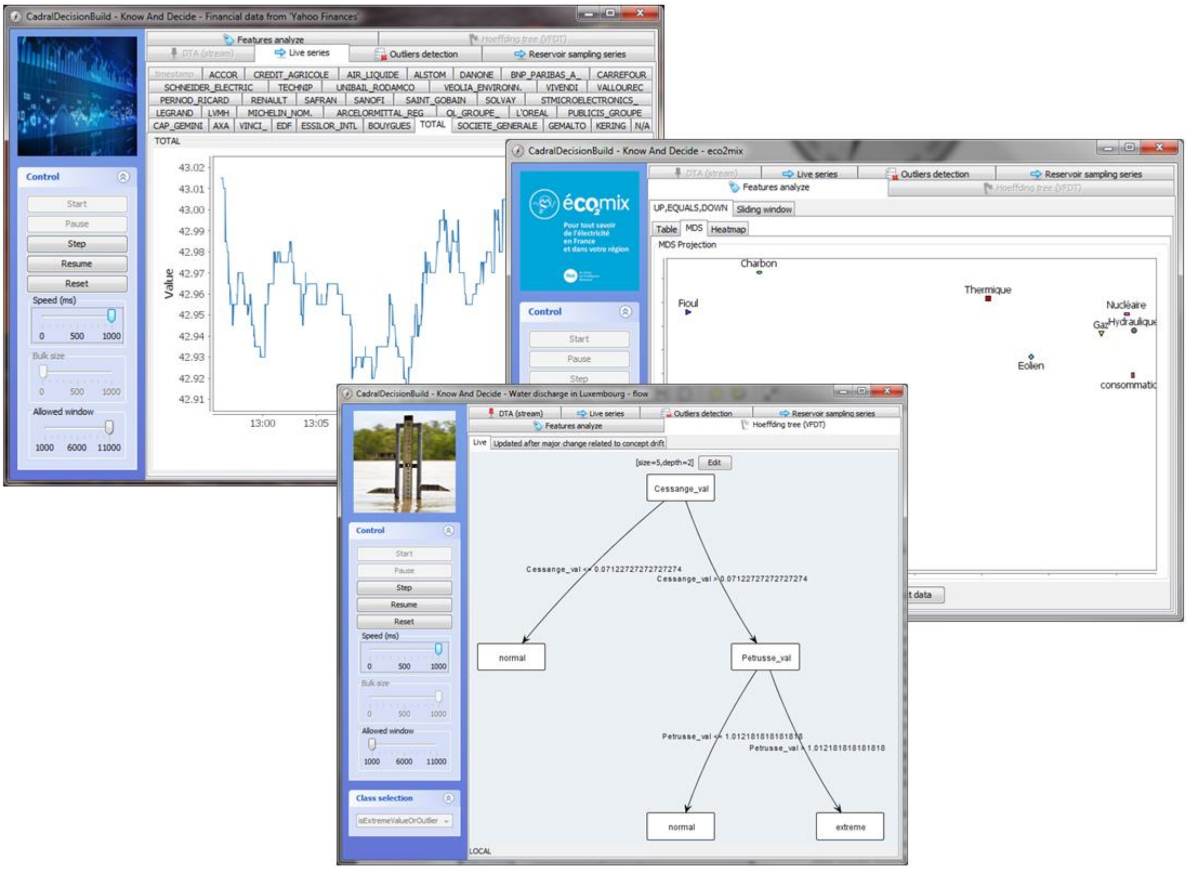

Fig. 1: live visualization of quotes (Yahoo finance API), features similarity analysis on the French energy consumptions data (RTE – eco2mix) and extreme flooding prediction using hydrological time series from Luxembourg [3].

In future works, we will extend the software in order to help the data scientists to quickly identify and eliminate bad data that pollute predictive models. To this end, we are implementing real-time modules for extreme values detection, missing data imputation and live clustering.

Links

http://www.list.lu/en/erin/

http://www.list.lu/en/erin/news/le-list-effectue-des-recherches-sur-les-precipitations-et-les-crues-extremes/

References

[1] H. Nguyen et al.: “A survey on data stream clustering and classification“, Knowledge and Information Systems, Springer, 12/2015

[2] E. Ikonomovska, J. Gama: “Learning model trees from data streams”, Discovery Science, 10/2008

[3] L. Giustarini et al.: “A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records”, Environmental Modelling and Software, 8/2016

Contacts

Olivier Parisot
Luxembourg Institute of Science and Technology, Belvaux, Luxembourg
olivier.parisot@list.lu

Benoît Otjacques
Luxembourg Institute of Science and Technology, Belvaux, Luxembourg
benoit.otjacques@list.lu