Temporal Data Mining.

Vera Shalaeva

Université Grenoble Alpes/Laboratoire d'Informatique de Grenoble (AMA group), UMR5217, Bâtiment IMAG, 700 avenue Centrale, CS 40700 - 38058 GRENOBLE CEDEX

E-mail: vera.shalaeva@imag.fr

Abstract.

Nowadays almost each object in the world has sensors and able to emit enormous amount of temporal data. With growing quantity of data, the needs to efficiently process and analyze time series also increased. There are a lot of applications where it's necessary to mine temporal data such that genomic analysis, information retrieval, finance, energy data analytics, airplane tracking and so forth. There are machine learning algorithms that were modified to deal with temporal data. However, there is few general purpose tools to deal with Big temporal data both the machine learning experts and non-experts can use.

Keywords:

temporal big data, time series, classification

Building a complete software to deal with time series is the relevant industrial and scientific goal of IKATS (Innovative ToolKit for Analysing Time Series) project.[1] In the scope of this project the aim is to set tools for preprocessing, classification and clustering of temporal data. To complete the development of an end-user oriented software, one requires interactive tools for visualization of data, results and workflows.

To achieve all these goals, the architecture of the project was defined as three primary blocks: pre-processing, machine learning, visualization.

Pre-processing step includes the extraction the data from database with data cleaning, dimension reduction and so forth.

Machine learning block of the project has to be able to accomplish the

[1] IKATS (Innovative ToolKit for Analysing Time Series) is research and development project. The consortium is composed of LIG (Laboratoire d'Informatique de Grenoble), CS (Concepteur, intégrateur & opérateur de systèmes critiques), Airbus (Leading aircraft manufacturer) and EDF (electricity generator).

clustering, classification, pattern search tasks on time series.

And the last block is responsible for visualization. All results and models generated by the machine learning tools have to be graphically represented. The visual representation must be interpretable and reveal the link between dataset and the models for the targeted users that are either data scientists, engineers or domain experts.

One important task in temporal data mining and for IKATS project is time series classification. All methods in temporal classification can be divided in the following groups:

- feature based classification, where first some features are extracted from time series on which conventional classification methods can be applied next;
- distance based classification. These methods measure and use pairwise similarity distances between all input time series. The most of the popular method in this group is 1-NN method with DTW distance;
- model based classification, where assumed the classes of time series are generated under some model. During the training, the parameters of this model are learned. On the classification step, model assigns a probability to each class and the time series is associated with the class having the highest likelihood. Neural network algorithm are related to this category.

In the context of temporal classification task, we are working with the algorithm Classification Trees for Time Series [A. Douzal-Chouakria, C. Amblard 2012]. This method modifies conventional decision tree algorithm which split the dataset at each node by using features of data. Instead of feature extraction from temporal dataset, we use distances between time series. At each node of a tree the algorithm searches for the best split pair of series according to an evaluation criterion such as Gini impurity index. Each sub-node of tree node comprises the time series that more similar with one time-series from the split pair than with another. The algorithm allows also to explore different distance

functions at each node and to find the most split significant time interval by dichotomy search. The split process continues while sub-node is not represented by one pure class called a leave. In the classification step, time series traverses the tree and class of achieved leave is assigned.

Besides the accuracy of classification, the advantage and interest of this method is its high level of interpretability. During the visualization step it's possible to have clear representation of learned model. However, to be able to include this method in the project tool we face the challenge of scalability. The current version of the algorithm has the high complexity $O\left(\log_{\frac{1}{\alpha}}(T)\, KN^3\right)$, where N is number of time series, K - number of explored distances, T - the time series length and α is the cover rate for dichotomous search of the most significant time interval. Therefore, in our PhD work, we focus on different approaches to decrease the algorithm complexity without losing in classification accuracy.

References:

1. A. Camerra,T. Palpanas, J. Shieh and E. Keogh. *iSAX 2.0: Indexing and Mining One Billion Time Series* (2010). The 10th {IEEE} International Conference on Data Mining.
2. P. Chaudhari and D. P. Rana and R. G. M. N. J. Mistry and M. M. Raghuwanshi. *Discretization of Temporal Data: A Survey* (2014). International Journal of Computer Science and Information Security (IJCSIS).
3. A. Douzal-Chouakria, C. Amblard. *Classification Trees for Time Series* (2012). Pattern Recognition Journal.
4. P. Esling, C. Agon. *Time-series Data Mining* (2012). ACM Computing Surveys (CSUR).
5. Z. Xing, J. Pei and E. Keog. *A Brief Survey on Sequence Classification* (2010). ACM SIGKDD Explorations Newsletter.

Université Grenoble Alpes is a public university in Grenoble, France. http://www.univ-grenoble-alpes.fr/

Laboratoire d'Informatique de Grenoble is research laboratory of informatics in Grenoble, France. https://www.liglab.fr/

The **AMA** team (dAta analysis, Modeling and mAchine learning) was created in LIG in January 2011 to work on machine learning and information modeling for complex data. http://ama.liglab.fr/