Stream Data Processing

Jakub ŠEVCECH*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
jakub.sevcech@stuba.sk

The stream data introduces several limitations methods processing the data have to cope with. The most notable ones are the requirement for incremental processing of the data and limited memory available to store the data or derived model of the data. The stream data processing comprises multiple tasks of data analysis and data mining (similarly to the processing of static collections of data) such as classification, clustering, anomaly detection, etc. In our work we focus on various tasks of data analysis of time series data produced from potentially infinite streams of data.

The main group of data processing methods we are focusing on are time series representations and data transformations as one of the steps preceding and supporting the actual data processing. These representations have to satisfy several requirements such as highlighting important aspects of the processed data while reducing its dimensionality, handling the noise present in the data and at the same time being easy to compute and provide low reconstruction error [2]. With various specialized applications, other requirements for the time series representations emerge. Over the last years many time series representations were proposed each emphasizing different attributes of the data and fitting the needs of different applications. One of the simplest ones yet most commonly used one is the Piecewise Aggregate Approximation (PAA) [4] which transforms the time series into a sequence of aggregated values computed from running window. Another interesting representation is SAX [6] transforming the PAA coefficients into symbols. This allows the application of methods from text processing domain in time series analysis. Another symbolic representation is proposed in by Das [1] as a support method for rule discovery in time series. Unfortunately, none of these symbolic representations is able to transform the data incrementally. In our work we propose a symbolic representation based on the one proposed by Das, but we are able to transform the data on the fly as the data is incoming.

The main idea of the proposed representation is transformation of repeating sequences into symbols. The symbols are formed by clustering subsequences of the processed time series by their similarity. The representation is composed by replacing time series sequences by cluster identifiers and by remembering the dictionary of

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2015 PeWe Workshop, April 11, 2015, pp. 25–26.

clusters. The key step of the transformation is cluster formation where we use an incremental clustering algorithm instead of originally used k-means algorithm.

As previous works on parameter free data mining stated [5], big number of attributes of various methods for data processing limit their applicability. To get around this limitation, we work on methods for automatic estimation of parameters of proposed transformation by training them on the processed data. The transformation requires three parameters to be set which depends on the data and on the intended application: running window size, lag between two windows and threshold difference of sequences in the cluster. To train the optimal window size we search for length of repeating (approximately) sequences in the processed time series. To search for these lengths, we compare two methods: one based on autocorrelation and one based on repetitivity [3]. As our experiment showed on synthetic data, the autocorrelation method outperformed the second method in the result quality and the running time. However when processing real-world datasets, the repetitivity based method showed interesting properties and it was able to identify to find not only the dominant pattern, but also other minor patterns even though it is more susceptible to noise and production of false results

Extended version was published in Proc. of the 11th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2015), STU Bratislava, 123-130.

Acknowledgement. This contribution was created with the support of the Research and Development Operational Programme for the project International centre of excellence for research of intelligent and secure information-communication technologies and systems, ITMS 26240120039, co-funded by the ERDF.

References

[1] Das, G., Lin, K., Mannila, H., Renganathan, G., Smyth, P.: Rule Discovery from Time Series. In: Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), (1998), pp. 16-22.

[2] Esling, P., Agon, C.: Time-series data mining. ACM Computing Surveys, (2012), vol. 45, no. 1, pp. 1–34.

[3] Grabocka, J., Wistuba, M., Schmidt-Thieme, L.: Scalable Classification of Repetitive Time Series Through Frequencies of Local Polynomials. IEEE Trans. on Knowledge and Data Engineering, IEEE, (2014), vol. PP, no. 99, pp. 1–13.

[4] Keogh, E., Chakrabarti, K.: Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems, (2001), vol. 3, no. 3, pp. 263–286.

[5] Keogh, E., Lonardi, S., Ratanamahatana, C. A.: Towards parameter-free data mining. In: Proc. of the 2004 ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining - KDD ’04, ACM Press, (2004), pp. 206–215.

[6] Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery, (2007), vol. 15, no. 2, pp. 107–144.