Anomaly Detection in Stream Data

Jakub ŠEVCECH*

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

During the last few years we can hear all over us a buzzword *Big Data*. The definition of this term is rather fuzzy, but one of the most frequent one is that it is a common name for techniques for processing data which are characteristic by its big amount, big velocity and/or big variability. The most common technique for dealing with such data is batch processing. However in many applications this type of processing is not viable mainly due to delays caused by the batch job processing time. When we require real time processing we have consider the data as stream and to reach for various stream processing methods.

The domain of stream data processing caught increased attention in recent years, but the majority of works published so far focused on search and analysis of stream data collected in time series databases, hence static collections of stream samples. There were achieved substantial successes in search and pattern comparison, association mining, time series prediction and so forth [5]. The majority of these works focused on processing of static collections of data and they did not focused on problems associated with processing of stream data in real or near-real time. Until recently minor attention was dedicated to stream processing but with the rising interest in Big Data, this domain is becoming the topic of interest for many researchers and practitioners.

Real time processing of stream data introduces new restrictions and problems, methods for processing of static collection of time series samples are not designed for [3]. The main restriction is fast, potentially unbound stream of data, which is often accompanied with great variability of the data and the restriction of a single pass through the data during the processing. The main directions of current stream data processing research are change detection, clustering [1], classification [4], pattern detection and association rules mining [2] and time series analysis.

In our work we focus on processing stream of data where we are working on methods for anomaly detection. The main challenge is to be able to process big number of various metrics running on the streamed data and to be able to do so in a single pass through the data. Our aim is to create and evaluate (precision and performance) a

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 39–40.

method for anomaly detection in stream data based on pattern detection on individual metrics running on the data and on classification of stream state using detected patterns and supervised learning. The first restriction we have to face is the problem of frequent patterns counting. The patterns comparison with current stream state is rather expensive operation. When we want to detect patterns, we have to select a set of patterns that have the biggest probability to be frequent and to limit the set of patterns we will look for in the data. This is thorough problem as we can only use single pass through the data and it is hard to predict future pattern frequency in time of their construction. A work described in [1] is elaborating this problem in further detail and it propose some viable methods for dealing with the problem. In our work, we build on these methods and we use them in anomaly detection in evolving data stream. We propose a method for anomaly detection composed of two phases:

- **Patterns detection** in metrics running on the data stream. The result of this phase is a set of patterns matched on the stream in every time. The stream of source data is reduced into the stream of matched patterns.
- **Anomaly detection** in the stream of patterns phase uses common machine learning algorithms for anomaly detection and categorization of stream state in specified time window.

We will evaluate the proposed method on various sources of stream data such as Twitter or Bit.li and on various applications such as application log analysis or standard datasets used in machine learning applications.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Aggarwal, C. C., Han, J., Wang, J., & Yu, P. S.: A framework for clustering evolving data streams. In: Proc. of the 29th international conference on Very large data bases, VLDB Endowment, (2003), vol. 29, pp. 81-92.

[2] Cheng, J., Ke, Y., Ng, W.: Maintaining frequent itemsets over high-speed data streams. In: Advances in Knowledge Discovery and Data Mining, Springer, (2006), pp. 462-467.

[3] Ling, C., Ling-Jun, Z., Li, T.: Stream Data Classification Using Improved Fisher Discriminate Analysis. In: Journal of Computers, (2009), vol. 4 no. 3.

[4] Wang, H., Fan, W., Yu, P. S., Han, J.: Mining concept-drifting data streams using ensemble classifiers. In Proc. of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, (2003), pp. 226-235.

[5] Zhao, Q., Bhowmick, S. S.: Sequential pattern mining: A survey. In: ITechnical Report CAIS Nayang Technological University Singapore, (2003), pp. 1-26.