Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Masterarbeit

Correlation Monitoring in High-Dimensional Data Streams using Bandit Theory

In predictive maintenance, one is interested in the identification of failures, sensor malfunctions or intrusions. To do so, it is often useful to extract real-time information from the data, such as its underlying correlation structure. Such information can be used as a "summary" of the stream's state and as a basis to improve the performance of data mining tasks such as clustering or outlier detection. However, the stream often is high-dimensional. By nature, data streams are infinite and evolving over time, so that dependencies that hold at some point in time might not hold in the future. This phenomenon is typically named *concept* drift. Because of that high-dimensionality, it is computationally impractical to re-estimate the dependencies between each variable pair for every time step.
Thus, one would need a **Correlation Monitoring** system to estimate the correlation structure of the stream at any time. In this context, the elements of **Bandit theory** seem to be very promising: Monitoring the correlation of n pairs can be seen as a *multi-armed bandit* problem, where the challenge consists of constructing a strategy for *pulling* each of the n *levers* (pulling lever n°1 ≡ computing the correlation of pair n°1) when one has no prior knowledge of the *payout rate* (i.e. correlation change) for any of the levers. Typically, one can only afford to *pull* a limited number of *levers* per time unit. So the problem becomes a tradeoff between *exploration* and *exploitation* of the current knowledge.

The focus of this thesis is the deployment of Bandit theory to Correlation Monitoring tasks. In particular, the following aspects are of interest:

- With limited resources, how can one maintain an estimation of the pair-wise correlation structure of a data stream? How "good" is this estimation, i.e., how close is it from an optimal computation?
- Correlation can be estimated using standard measures such as Mutual Information. What is the impact of different estimators on the execution time and quality of monitoring?
- Depending on the speed of concept drift, how can one determine a good tradeoff between exploration and exploitation?
- If we consider multivariate dependencies, the number of *levers* increases exponentially. How can we adapt the multi-armed bandits model to this setting?

This results in the following tasks:

- Exploratory analysis of Bandit theory: review of different bandit models and applications.
- Development of a framework to monitor the correlation structure of high-dimensional data streams with a focus on the pair-wise case first, and then on the multivariate case.
- Theoretical and experimental demonstration of the benefits of such framework, using synthetic and real world data.

Throughout this work, the student will acquire a deep knowledge of correlation analysis and probability theory. He/she will sharpen his/her Data Science skill. The student will learn how to conduct controlled experiments to compare results of his/her work to the state of art.

Ansprechpartner

Edouard Fouché, M. Sc. edouard.fouche@kit.edu +49 721 608-47337 Raum: 342

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34

Masterarbeit