Karlsruher Institut für Technologie

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

**Masterarbeit**

# Prediction-Oriented Correlation Monitoring in Data Streams

Masterarbeit

In predictive maintenance, one is interested in the identification of failures, sensor malfunctions, intrusions – or the prediction of a target variable. In such scenarios, data is often available as streams. By nature, data streams are infinite; they are evolving over time and can be aggregated at multiple time scales. Also, data streams can be high-dimensional. The effects of the "curse of dimensionality" lead to the degradation of the performance of conventional prediction approaches. Furthermore, in critical applications, a short response time is required, so that the efficiency of learning algorithms is crucial. As a result, data analysis for predictive maintenance is challenging.

In this setting, Correlation Monitoring – i.e., a natural extension of correlation analysis for data streams – is promising to cope with these issues. By keeping track of the most interesting features or feature subsets over time, one can improve the prediction algorithms, by reducing the effects of the curse of dimensionality and of overfitting.

**The focus of this thesis is the deployment of existing Correlation Monitoring approaches with an emphasis on efficiency and parsimonious resource consumption.** In particular, the following aspects are of interest:

- Many Correlation Monitoring approaches target at pairs of streams. How can we query in an efficient way the top-k pairs of attributes that show the highest correlation in a multivariate data stream?
- Can we infer from a set of bivariate correlation coefficients a – possibly approximate – measure for multivariate correlation and reciprocally?
- How can the information from a correlation monitoring system be used to improve the performance of underlying machine learning algorithms? For example, what could be the impact of a significant increase or decrease in correlation of a set of attributes?

This results in the following tasks:

- Exploratory analysis of bivariate and multivariate correlation estimators on data streams.
- Development of a framework to monitor the correlation relationships of multivariate data streams.
- Theoretical and experimental demonstration of the benefits of such framework, including the comparison with state of the art methods for correlation estimation in data streams.

Throughout this work, the student will acquire a deep knowledge of correlation analysis and its manifold applications in Data Science. He/she will sharpen his/her Data Science skills and learn how to articulate theoretical concepts. 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