Karlsruher Institut für Technologie

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

Bachelorarbeit

Statistical Generation of High-Dimensional Data Streams with Complex Dependencies

Bachelorarbeit

Extracting knowledge from data streams is challenging: in contrast to static data, they are infinite and always evolving, so concepts learned at a certain time cannot be expected to hold in the future. For example, a particular value may be considered an anomaly at time t_1 but a normal value at time t_2. If such an anomaly goes undetected, it can have disastrous consequences in the context of say, fraud detection or a production plant monitoring system. The problem becomes even more difficult with increasing dimensionality. Because of the so-called "*curse of dimensionality*", outliers become "*hidden*" and are visible only in particular subspaces. Obviously, stream mining is not restricted to outlier detection, but other tasks such as *clustering* or *classification* in data streams belong to current research.
A major obstacle is the unavailability of benchmark data. To assess the quality of a new algorithm, researchers require a ground truth, which may be expensive or even impossible to obtain. A solution is the generation of synthetic benchmark data where the characteristics of the data and the labels are known. The generation of such data should be based on statistical properties that are controllable, such as the dependencies between subspaces and the presence or absence of noise and outliers. **The topic of this bachelor thesis is the design of statistical methods to generate high-dimensional data streams with complex dependencies**. In particular, the following aspects are of interest:

- Which different kinds of dependencies can we model in multiple dimensions? How can we parameterize them to let them vary over time?
- Certain kinds of dependencies are not easy to detect via traditional correlation measures such as Pearson Correlation or Spearman Rho coefficients. How do we generate data featuring these dependencies?
- How can we model different dependencies on distinct but overlapping subspaces?
- Can we simulate concept drifts both in the supervised and the unsupervised setting?

This results in the following tasks:

- Literature review of statistical dependencies and concept drift in data streams.
- Development of efficient algorithms to generate controlled dependencies in high-dimensional data streams.
- Implementation of the algorithms in a unified framework in R and evaluation of the produced results with state-of-the-art algorithms.

Throughout this work, the student will get a deeper understanding of statistical dependencies and become familiar with high-dimensional data. The student will train highly valuable Data Science skills and get the opportunity to build software that will be used extensively in research. A good knowledge of statistics, R (or another comparable programming language) and software development is required.

Ansprechpartner

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

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34