Karlsruher Institut für Technologie

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

Bachelor Thesis

Bachelor Thesis

Estimating Mutual Information on Data Streams in Finance

Capturing associations and dependencies is an everyday task of a financial quantitative analyst. It is of interest in questions like: Which stocks behave similarly, and what is the degree of that similarity? The analysts may also want to identify time periods in which associations are highest or lowest. Answering these questions helps in successful portfolio optimization.

Several dependence measures are utilized to capture associations. They include, but are not limited to: Pearson and Spearman correlation, Kendall's τ and Mutual Information (MI). It has been shown that Mutual Information is a very general measure, which does not make any assumptions on data distribution, in contrast to many other measures.

Conventional MI estimators cannot be adapted in a straightforward manner to data streams, as they must deal with:

- the dynamic nature of the stream,
- its infinite and multiscale characteristic,
- large numbers of queries.

Recently the MISE framework has been proposed to cope with these requirements. However, this approach is not always optimal in terms of memory consumption, accuracy, or processing/calculation speed. We are aware of at least two alternative approaches. They differ in the relative share of calculations performed during stream processing, as compared to those performed in order to answer the query.

The focus of this assignment is to study theoretically and experimentally the quality characteristics of the approaches. Specifically, it is unknown so far whether one of the approaches is always superior to the other ones in terms of accuracy, calculation speed and memory consumption. It is also unclear to which extent the choice of the best algorithm depends on the underlying data. If the dependency is strong or changes rapidly, which factors do affect the quality of each algorithm?

This implies the following tasks:

- Exploration of design alternatives and implementation of all three algorithms.
- Provide theoretical considerations on quality measures for the algorithms.
- Propose a methodology to compare the algorithms.
- Design synthetic data and find real world data which covers a wide range of possible situations (different levels of MI, rapid/slow changes in dependence).
- Experimentally assess the quality of the algorithms.

In this work you will learn about advantages and disadvantages of different measures for revealing associations in financial data. You will gain skills regarding the development and evaluation of streaming algorithms.
R and C++ are preferable programming languages for implementation. Some programming experience is desired.

Contact

Vadim Arzamasov vadim.arzamasov@kit.edu Room: 340

Am Fasanengarten 5 76131 Karlsruhe Building: 50.34