Karlsruher Institut für Technologie

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

Masterarbeit

Masterarbeit

Relevance-Driven Feature Engineering

Feature Engineering is an essential task of any data analysis pipeline. It means enriching the current feature set by discovering useful hidden variables. This process can be hypothesis-driven, i.e., data analysts derive new features based on the existing ones or by taking domain knowledge into account. It can also be automated, e.g., by the systematic computation of aggregates and statistical indicators or by using feature transformation methods such as Principal Component Analysis. This is an expensive process, because the relevance of constructed features is not known in advance. In general, the relative performance improvement of underlying learning algorithms is used as proxy to estimate the relevance of newly created features. This is biased and time-consuming. Next, the space of possibly relevant features is virtually unbounded.

Orthogonally, Feature Selection is the relevance estimation process of features and of feature subsets. This also is a difficult problem, because the number of feature subsets to be considered grows exponentially with the number of dimensions. However, a wide range of heuristics exists to at least mitigate this problem.

In a Data Science pipeline, one would be interested in determining the relevance of features during the construction process, as early as possible. This would shorten the process and offer guidance towards building a minimal relevant feature set. **The focus of this thesis is the development of algorithms to improve the Feature Engineering process in this spirit**. In particular, the following aspects are of interest:

- One is interested in building a compact set of relevant non-redundant features in a reasonable time. Can one theoretically define *relevance* and *minimality* of a feature set, to advise the Feature Engineering process? Can we speed up Feature Engineering by ignoring parts of the search space without a significant loss of quality?
- A major risk in the Feature Engineering process is the construction of too many features, leading to the *curse of dimensionality* and *overfitting*. How can we couple Feature Selection and Feature Engineering processes in order to mitigate these effects?
- In predictive maintenance scenario, the data is often available as a stream. By nature, data streams are infinite and evolving over time, such that the relevance of feature might change over time. How can we generalize the methods developed so far accordingly?

This results in the following tasks:

- Exploratory analysis of Feature Engineering/Selection methods and development of algorithms to integrate some kind of relevance evaluation in the construction process.
- Theoretical study and definition of relevance/minimality criteria in Feature Engineering and extrapolation to the data stream setting.
- Evaluation of algorithms and measures through experiments, including the comparison with the current state of the art.

Throughout this work, the student will get a deep knowledge of the Feature Engineering/Selection processes. He/she will learn how to articulate theoretical concepts and approach unsolved problems of the Data Science community.

Ansprechpartner

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

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34