Karlsruher Institut für Technologie

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

Bachelor Thesis

Bachelor Thesis

Improving Outlier Scores to Identify Suspicious Objects in Subspaces

Outlier detection has the goal to reveal unusual patterns in data. Typical scenarios that fall under the area of outlier detection are the identification of unusual energy consumption, of suspicious financial transactions or of peaks and odd sequences in high dimensional data. Such data sets require to rethink well established outlier detection concepts, as some of them lose their effectiveness in a high dimensional data space. Recent approaches search for low-dimensional projections of the data space, where outliers become visible. The number of possible low-dimensional projections increases exponentially with the number of dimensions. Therefore, subspace search methods rely on statistical measures to identify the relevant projections.

Outlier detection models are often applied to each of the identified subspaces. However, the user is interested in an overall ranking of outliers and not the individual subspace rankings. Hence, different subspaces rankings are aggregated.

The focus of this thesis is to study the effects of different scaling and combination functions on subspace outlier rankings. In particular, the following research questions are of interest:

- Different scaling and combination functions have advantages and disadvantages and can highlight different outliers. The question is, which functions are best suited for specific data sets that show certain characteristics, such as correlation.
- Some outliers scores might be robust against transformation while others might only become visible after a certain score transformation. It is an open question, how this information can support the user in distinguishing between obvious and less obvious outliers.
- Preliminary experiments indicate that the overlap of subspaces can negatively impact the outlier detection quality. Removing redundant subspaces from an outlier ranking allows some outliers to become more prominent. Quantification of redundancy is not properly defined and the impact of redundancy in subspace search results not studied.
- Currently, the results are presented to the user as a single outlierness score. With multiple aggregation possibilities, the presentation of outliers needs to be accommodated.

This results in the following tasks:

- Exploration of effects of different outlier detection models, scaling functions, combination functions on outlier ranking. This includes to also think about additional functions beyond what has already been presented in literature.
- Recommendations of score transformations based on data set characteristics including exploration and description of (dis-)advantages.
- Development of evaluation metrics for score robustness, subspace redundancy and the difference between final scorings.

In this thesis you gain deep insight and knowledge on large scale data analytics. You train highly demanded skills in development and evaluation of data mining algorithms. Knowledge from the lecture “Big Data Analytics” is no prerequisite. However, statistical knowledge and the ability to accomplish conceptual work is desired.

Contact

Holger Trittenbach	holger.trittenbach@kit.edu	+49 721 608-44066	Room: 338
Am Fasanengarten 5	76131 Karlsruhe	Building: 50.34	