Karlsruher Institut für Technologie

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

Master Thesis

Exploiting Neighborhood Relationships to Detect Suspicious Objects

Master Thesis

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 fraudulent 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 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. Subspace search methods rely on statistical measures to identify the relevant projections.

Improving the interpretability of outlier detection is focus of current research. A subspace provides the user context in which the object of interest appears as an outlier. Hence, a desirable result provides the user with multiple subspaces that give an explanation of an outlier. The explanation becomes more comprehensive the more diverse the subspaces are.

The focus of this thesis is to exploit relationships in local neighborhoods of data objects to increase the interpretability of outlier detection in subspaces.

In particular, the following relationships are of interest:

- The k-nearest neighbors are objects which are most similar to an outlier. Looking at additional subspaces in which the outlier has the same set of neighbors might add very little information to the user. Instead, a subspace where the outlier has different neighbors might enhance the interpretability.
- Recent literature focuses on objects that occur frequently in the k-nearest neighborhood of other objects, so-called hubs, that can be used to indicate high dimensional clusters. This property could also be exploited to traverse the search space for outlier detection. It is yet unknown, how hubs are related among different subspace. Also, the effect of correlation among the subspace dimensions on the hubness property is not studied.
- An effect that has been observed is that anti-hubs, i.e., objects that are very infrequent neighbors, seem to be correlated with the outlier score of an object. This characteristic could complement the description of outliers. On the other side, hubs can be considered as outliers themselves as they are also rare. This is different from classical outlier definitions. Subspace search approaches can be adapted to also search for (anti-)hubs.

This results in the following tasks:

- Exploratory analysis of the hubness property in subspaces and the development of similarity measures for subspaces based on nearest neighbor considerations.
- Adaption and extension of existing subspace search algorithms to exploit neighborhood relationships and to use the hubness outlier definition.
- Evaluation of algorithms and measures through experiments.

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