Karlsruher Institut für Technologie

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

Masterarbeit

Designing Games on Data-Driven Fraud Detection and Sabotage

Masterarbeit

In general, outliers are data objects that deviate from other (usual) data objects. This deviation typically only occurs in some attribute subspaces. Consider for example data on the temperature of an engine and the speed the engine is running at. These two measures should clearly be highly correlated. I. e., a combination of a high temperature and a low speed is an outlier. However, when we examine both one-dimensional subspaces on their own, the values might be in usual ranges of temperature and speed. Depending on the subspaces an algorithm searches for outliers, this is what we refer to as hidden outlier: Its outlierness in the two-dimensional space is hidden by the one-dimensional subspaces. If outliers are hidden, an outlier-detection algorithm may miss them, which might have drastic consequences. For example, the situation above might lead to an engine failure that could have been avoided.

A purely analytical analysis of data characteristics and subspace selections resulting in many hidden outliers is impractical, leastwise due to the variety of outlier-detection algorithms. Hence, we recently developed an approach to place hidden outliers in datasets. These placed objects can be used to estimate effects of different data and subspace. However, being able to place hidden objects raises the question of how to obtain subspace selections robust to such. This gives way to the design of a competitive two-player game. Think of two researchers in the field of subspace search

1. The honest one: Develops new heuristics to select meaningful subspaces to detect outliers.
2. The cheater: Creates data on which the heuristics of the honest researcher do not work well.

Here the cheater could make use of hidden outliers, i.e., outliers that are hard to detect if the honest researcher is ‘on a budget’. In other words, his resources for outlier detection are bounded. In this assignment we are interested in the consequences of such games. In particular, can the honest researcher win, and under which circumstances exactly is this the case? The work can be divided into smaller task as follows

- Assess and if necessary improve algorithms that place synthetic hidden outliers in data
- Analyze the vulnerability of different subspace search methods in regards of hidden outliers
- Define meaningful strategies for
    - The honest one: How does he adapt his heuristics to outliers which have been missed?
    - The cheater: How does he adapt the placement of hidden outliers to new outlier detection algorithms?
- Simulate the game using various benchmark datasets and outlier detection mythologies
- Based on the results of these simulations try to find criteria for a subspace search heuristic that makes it robust to hidden outliers

With this work you will develop your skills in Big Data Analytics as well as regarding the design and evaluation of algorithms.

Ansprechpartner

Georg Steinbuß, M. Sc. georg.steinbuss@kit.edu +49 721 608-43911 Raum: 363

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34