Karlsruher Institut für Technologie

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

Masterarbeit

Masterarbeit

Neural-Based Outlier Detection in Data Streams

Outlier detection has the goal to reveal unusual patterns in data. Typical scenarios in predictive maintenance are the identification of failures, sensor malfunctions or intrusions. This is a challenging task. Predictive maintenance data is often available as a stream. By nature, data streams are infinite; they are evolving over time and can be aggregated at multiple time scales. Since most existing methods for outlier detection only apply to static data, they cannot accommodate data streams. Furthermore, in real-time applications, short response time is required, so the efficiency of algorithms is crucial.

Neural-based unsupervised methods have been developed and used to detect outliers, such as Auto-encoder (Replicator Networks) and Self-Organizing Maps. However, their performance in existing studies has only been demonstrated with static data. Also, neural networks were often trained on labeled data, which is unrealistic. Due to the lack of labels and the unknown characteristics of anomalies, outlier detection should be considered an unsupervised problem. A recent interest of the scientific community – characterized as the "neural network renaissance" – has led to the development of methods to optimize the learning quality of neural networks and has proven to be very effective. Also, thanks to the improvement of available hardware, training can be sped up significantly.

The focus of this thesis is the development of neural-based algorithms to tackle the problem of outlier detection in streams. In particular, the following aspects are of interest:

- Since data streams are evolving, a particular point may be considered an outlier at time t_1 but an inlier at time t_2. How can we deploy neural networks in such situations? Data-driven methods for continuous neural learning need to be designed to control the speed of neural networks adapting to distribution changes.
- A point may be considered an outlier w.r.t. a particular time scale, e.g., an hour, but may be seen as an inlier in other time scales, e.g., a month. How can we detect such time-dependent outliers, and how can we design metrics to compare them?
- High-dimensionality brings additional challenges, i.e., the „curse of dimensionality". How does high-dimensionality affect the learning in data streams?

This results in the following tasks:

- Exploratory analysis of neural networks for unsupervised outlier detection and development of various outlier scores in data streams.
- Development of data-driven methods for the optimization of neural-based learning and for the interpretation of their output through time.
- Evaluation of algorithms and measures through experiments, including the comparison with other state of the art approaches for outlier detection in data streams.

Throughout this work, the student will get a deep understanding of neural networks and their application in the field of outlier detection. He/she will sharpen his/her Data Science skills and become familiar with theoretical and practical aspects of handling data streams. The student will acquire highly relevant experience with neural network frameworks such as Tensorflow, Theano or Caffee.

Ansprechpartner

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

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34