Karlsruher Institut für Technologie

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

Masterarbeit

Masterarbeit

High-Dimensional Neural-Based Outlier Detection

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, especially when the data is high-dimensional, because outliers become “hidden” and are visible only in particular subspaces.

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 in low-dimensional data. Also, neural networks were often trained on labelled 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 – 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 significantly sped up.

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

- The capacity of neural networks (i.e., number of parameters) should be reduced or controlled in order to prevent overfitting. To this end, many regularization methods have been developed. However, overfitting stays a major concern. Can we find data-driven methods to infer an appropriate number of parameters in a neural network?
- Neural networks generally are considered black-box systems, while one is interested to know in which subspaces the outlierness of points is visible. How can we derive such information from a neural network?
- Networks are generally trained on "normal" data only. This is problematic for two reasons: First, a ground truth, i.e. whether a point is an anomaly or not, may be non-existent or imperfect. Second, it restricts the learning scope only to anomalies that are known. Since anomalies shall be assumed to be unknown beforehand, the question that arises is how anomaly contamination does affect the learning process?

This results in the following tasks:

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

Throughout this work, the student twill 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 high-dimensional data. 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