Universität
Augsburg
University

Disputation:

From Statistics to Representation Learning: A Comparative Study on Outlier Detection Methods

Studiengang: Betriebswirtschaftslehre

Lehrstuhl für Statistik

Eingereicht bei:	Prof. Dr. Yarema Okhrin
Betreuer:	Herr Philipp Haid
Vorgelegt von:	Thore Johannsen
Adresse:	Salomon-Idler-Straße 4 Augsburg
Matrikel-Nr.:	2231165
Email:	thore.johannsen@outlook.com

Augsburg, im September 2010

Thesis Proposal

This thesis proposes a comparative analysis of FAST-MCD, KNN and a mothod based on contrastive learning proposed by (Shenkar und Wolf, 2022)
Outlier detection plays a significant role in scientific analysis and in private Enterprises. For instance, outliers influence the variance and standard deviation of variables. Hence, confidence intervals become wider and you might fail to reject the null hypothesis even if a true effect exists. In some contexts outliers are the main findings of dection methods. Banks use outlier detection to distinguish between fraudulent and normal useage of credit cards (Ramaswamy et al., 2000) Thus outliers tend to interfere with the results of data analysis. In medicine its been used to detect diseases like cervical cancer (Ijaz et al., 2020) In my thises I want to examine wich method will be the most effective one. I hypothesize, that the most effective method will be the KNN-algorithm. In addition I hypothesize, that contrastive learning algorithm will be more efficient than FAST-MCD but not as efficient than KNN.
The minimum covariance determinant (MCD) will be the first algorithm I will describe in my thesis. Therefore I will describe the Mahalanobis distance (Mahalanobis, 2018) bescause it will be used in MCD algorithm. (Rousseeuw, 1984) After that I will illustrate the expansion of MCD to FAST-MCD wich will be implemented into my thises. FAST-MCD has the advantages, that it is
- Was versprechen Sie sich von verschiedenen empirischen Methoden, welche Sie in ihrer Arbeit anwenden werden?

- Was sagt die wissenschaftliche Literatur zu den Vorzügen und Nachteilen verschiedener Methoden, welche Sie in ihrer Arbeit anwenden werden?

- Welche Daten werden Sie verwenden, um ihre Fragestellung zu untersuchen?

- Von welcher wissenschaftlichen Literatur werden Sie in Ihrer Arbeit ausgehen?

Literatur

Ijaz, Muhammad Fazal, Attique, Muhammad und Son, Youngdoo: Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. In: *Sensors*, Band 20(10):S. 2809, 2020.

Mahalanobis, Prasanta Chandra: On the generalized distance in statistics. In: *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, Band 80:S. S1–S7, 2018.

Ramaswamy, Sridhar, Rastogi, Rajeev und Shim, Kyuseok: Efficient algorithms for mining outliers from large data sets. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, S. 427–438.

Rousseeuw, Peter J: Least median of squares regression. In: *Journal of the American statistical association*, Band 79(388):S. 871–880, 1984.

Shenkar, Tom und Wolf, Lior: Anomaly detection for tabular data with internal contrastive learning. In: *International conference on learning representations*. 2022.