HOCHSCHULE HANNOVER
UNIVERSITY OF APPLIED SCIENCES AND ARTS
–
Fakultät IV Wirtschaft und Informatik

Finding the Difference: Anomaly Detection in Computer Science Teaching and Network Security Research

Volker Ahlers

University of Applied Sciences and Arts Hannover, Germany

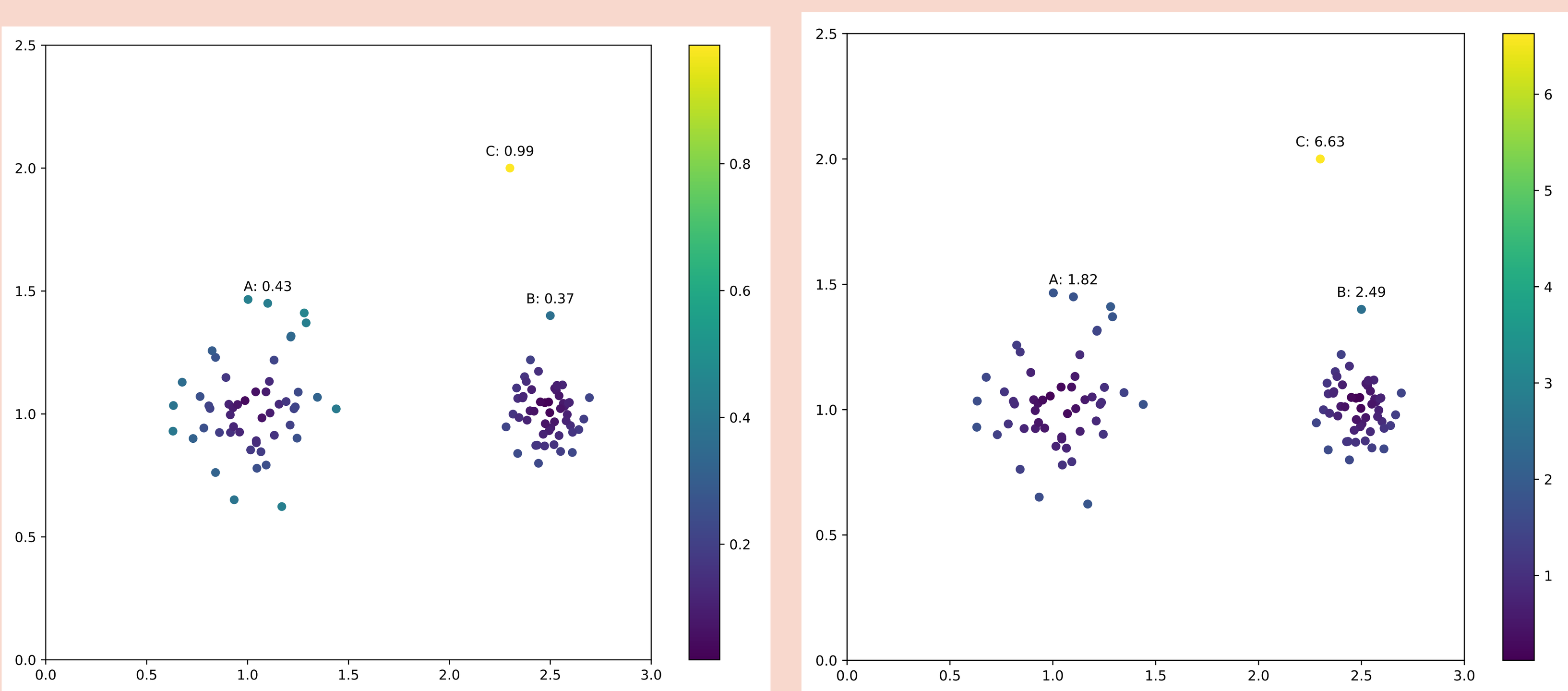
Outliers vs. Anomalies

- **Outliers**:
  - data points deviating from "normal" data
  - rare events, measurement errors, noise, . . .
- **Anomalies**:
  - deviating data points that are really interesting
  - "real interesting" depends on the application
  - point anomalies, contextual anomalies, collective anomalies
- Possible challenges in anomaly detection:
  - How to model normal behavior?
  - no sharp boundary between normal behavior and anomalies
  - **imbalanced data**: typical datasets contain only few anomalies

Teaching Anomaly Detection

- Part of master module *Machine Learning*
- Study different **anomaly detection methods**:
  - supervised (for labeled data) vs. unsupervised
  - model-based vs. model-free
  - **evaluation** of results
- Programming tutorials using Python and Scikit-Learn

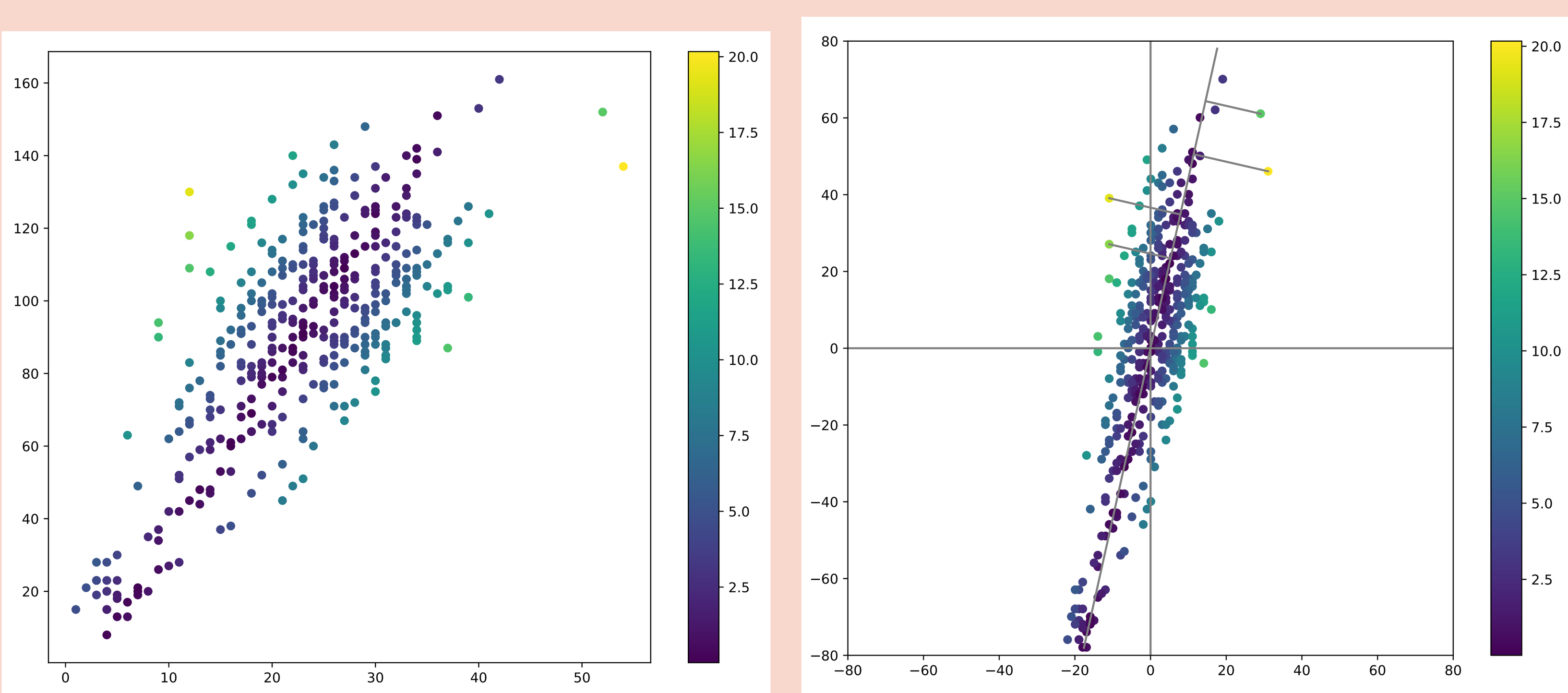
Clustering Approach (e. g., k-Means)



Figures: Felix Heine

- unsupervised, model-based method (model: clusters of normal data)
- left: **anomaly score** based on distance to closest cluster center
- right: **anomaly score** based on distance relative to average distance of points within same cluster

Reconstruction-based Approach



Figures: Felix Heine

- unsupervised, model-based method (statistical model)
- Use principal component analysis (PCA) or auto-encoders to compress the data by eliminating non-important axes.
- **anomaly score** based on reconstruction error of original data from compressed data

- Further methods: local outlier factor, one-class SVM, isolation forests

Anomaly Detection for Network Security

- Anomaly-based **intrusion detection system** (IDS)
- **Collective anomalies** in network traffic data streams:
  - High number of connections from server X and high amount of traffic over port P are unsuspicious separately.
  - Combination (X, P) within short time might indicate an attack.

OLAP Cube Approach (Supervised, Model-based)

- dimensional attributes: protocol, host/IP address, port
- metric attribute: packet count
- **cell** $c = (a_1, \ldots, a_n)$
  - apex cell: $(*, \ldots, *)$ ($*$ indicates "aggregated")
  - base cell: ('tcp', 'Google', 53)
- **cuboid**: set of cells with common pattern, e. g., $(A, *, C)$
- **normality model** for each cell: normal distribution $N(\mu, \sigma^2)$ of packet counts over time slices
- **anomaly score**: deviation of actual packet count c from mean in terms of standard deviations, $|c - \mu|/\sigma$

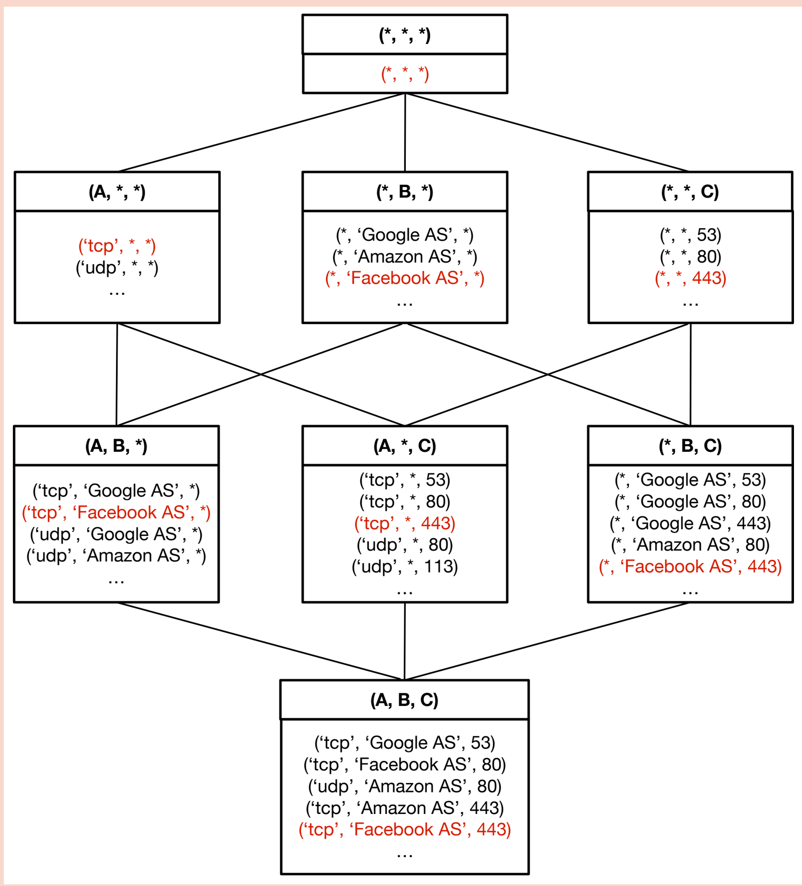
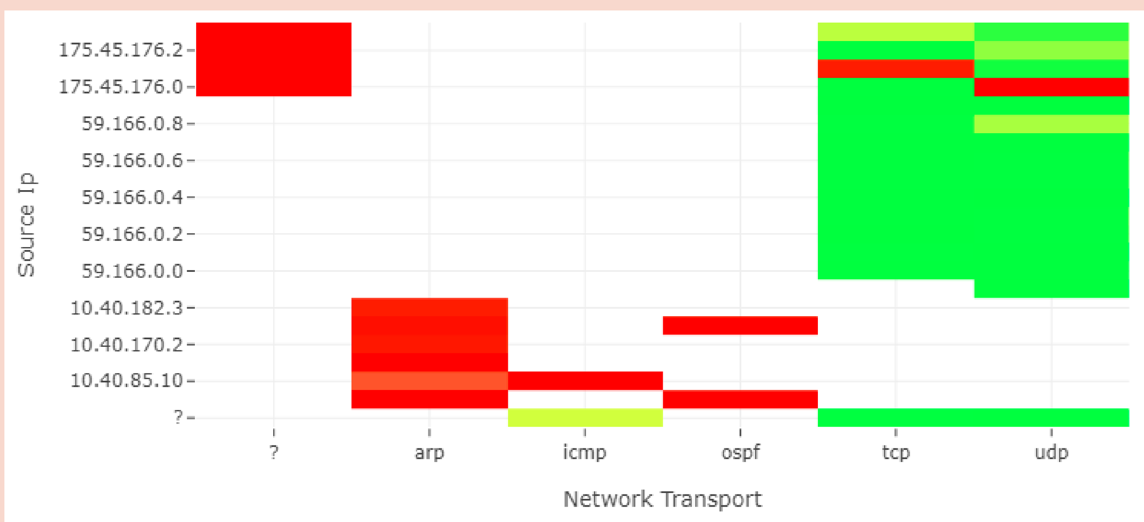


Figure: Felix Heine

Visualization of Anomaly Scores (Student Project, TypeScript/Plotly.js)

- GUI with attribute filters and color-coded visualization of anomaly scores
- Heat map example: abnormal cells in IP range 175.45.176.* in combination with TCP/UDP (test data: UNSW-NB15)



References

V. Ahlers, T. Laue, N. Wellermann, F. Heine: Visualization of data cubes for anomaly detection in network traffic data streams. In *Proceedings of IDAACS 2021*. 272–277. IEEE, 2021. doi:10.1109/IDAACS53288.2021.9660978

F. Heine, C. Kleiner, P. Klostermeyer, V. Ahlers, T. Laue, N. Wellermann: Detecting attacks in network traffic using normality models: the cellwise estimator. In *Foundations and Practice of Security (Proceedings of FPS 2021)*. 265–282. Springer, 2022. doi:10.1007/978-3-031-08147-7_18

N. Moustafa: The UNSW-NB15 dataset. 2015. https://researchdata.edu.au/unsw-nb15-dataset/1425943

Funded by

Federal Ministry of Education and Research

Grant no. 16KIS0950

GLACIER project
www.glacier-project.de

Contact: volker.ahlers@hs-hannover.de