

Outlier Detection of Generalized Deduplication Compressed IoT Data

Something

Morten Lyng Rosenquist
Faculty of Technical Sciences
Aarhus University
Aarhus, Denmark
201706031

May 17, 2022

Abstract— Index Terms—

I. INTRODUCTION

II. BACKGROUND

A. Generalized Deduplication

Deduplication is a technique to perform compression in storage systems. The technique works by utilizing the similarity of file chunks. Each unique file chunk is stored once. Subsequent copies of the chunks are then replaced with a reference to the stored chunk. The method is established and shown to have good compression gain on various practical scenarios [2]. However, if there are minor discrepancies in the file chunks, the technique will not leverage any of the similarities. Resulting in the near-identical chunks being stored in full. Sensor data from IoT devices is one example of the data potentially being near-identical.

To utilize the similarities in the almost identical data, a generalization of deduplication has been studied. This method consider the chunks at the bit level and splits them into two parts, the *base* and *deviation*. The *base* is the identical part that is to be stored once and hereafter referenced with pointers. The *deviation* is the disparity between the chunks. Looking at a simple example with four 6-bit numbers, 100000, 100001, 100010 and 100011. It can be identified that the four most significant bits of the numbers are identical. Hence, leading to all having a shared *base* of 1000. The two least significant bits are then the *deviation*[4].

B. Isolation Forest

Anomaly detection is a combination of outlier- and novelty detection. Including both identifying outliers in the training data and determining if unseen observations are outliers. Isolation Forest (iForest) is an anomaly detection method. It differs from other popular techniques in the way that it identifies anomalies explicitly instead of profiling ordinary

data points[1]. IForest utilizes decision trees similar to other tree ensemble methods. The main principle is to recursively split each data point, and then evaluate the amount of splits necessary to split each data point. The logic is that anomalies will requires less splits to be isolated than an ordinary point. Trees are built by selecting a random feature and then selecting a random value between the minimum and maximum value of that feature. The process is then repeated untill all data points are isolated or a maximum height of the tree is reached. An illustration can be seen on figure....

III. RELATED WORK

Performing analytics on compressed data is not an untouched subject.

A collection of models and algorithms are developed to perform classification and anomaly detection within network communication on compressed data[3].

Another paper looks into anomaly detection based on compressed data. They do it on the edge of the cloud on compressed data. Lots of formulas regarding rate vs. distortion. Dont know compression or anomaly detection method[7].

Direct analytics on data compressed using generalized deduplication has been carried out. It was studied how clustering(K-Means, IMM, DTC) could be performed on synthetic, synthetic with noise and a power consumption data set [6].

Since isolation forest only performs horizontal and vertical splits certain anomalies will not be detected. Imagining a two dimensional data set. Then the ones having the same x and y values might not be isolated correctly. This is extended by allowing diagonal splits in the extended isolation forest[5].

IV. METHODS

Write
ab-
stract

Write
key-
words

write
intro-
duction

Write
back-
ground

Tilføj
figur
med
plot
af
data
points
og
tilhørende
træ.

Write
re-
lated
work

write
method

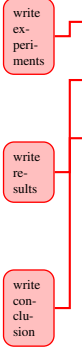
A. Count Isolation Forest

V. EXPERIMENTS

VI. RESULTS

VII. CONCLUSION

REFERENCES

- 
- [1] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation Forest”. In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 413–422. DOI: [10.1109/ICDM.2008.17](https://doi.org/10.1109/ICDM.2008.17).
 - [2] Wen Xia et al. “A Comprehensive Study of the Past, Present, and Future of Data Deduplication”. In: *Proceedings of the IEEE* 104.9 (2016), pp. 1681–1710. DOI: [10.1109/JPROC.2016.2571298](https://doi.org/10.1109/JPROC.2016.2571298).
 - [3] Christina Ting et al. “Compression Analytics for Classification and Anomaly Detection Within Network Communication”. In: *IEEE Transactions on Information Forensics and Security* 14.5 (2019), pp. 1366–1376. DOI: [10.1109/TIFS.2018.2878172](https://doi.org/10.1109/TIFS.2018.2878172).
 - [4] Rasmus Vestergaard, Daniel Enrique Lucani Rötter, and Qi Zhang. “Generalized Deduplication: Lossless Compression for Large Amounts of Small IoT Data”. English. In: *European Wireless Conference*. VDE Verlag GmbH, 2019, pp. 67–71. ISBN: 978-3-8007-4948-5. URL: <https://www.vde-verlag.de/proceedings-en/564948016.html>.
 - [5] Sahand Hariri, Matias Carrasco Kind, and Robert J. Brunner. “Extended Isolation Forest”. In: *IEEE Transactions on Knowledge and Data Engineering* 33.4 (Apr. 2021), pp. 1479–1489. DOI: [10.1109/tkde.2019.2947676](https://doi.org/10.1109/tkde.2019.2947676). URL: <https://doi.org/10.1109/tkde.2019.2947676>.
 - [6] Aaron Hurst et al. “Direct Analytics of Generalized Deduplication Compressed IoT Data”. English. In: *IEEE Global Communications Conference (GLOBECOM)*. IEEE Global Communications Conference (GLOBECOM). IEEE Conference and Exhibition on Global Telecommunications, GLOBECOM ; Conference date: 07-12-2021 Through 11-12-2021. IEEE, 2021. DOI: [10.1109/GLOBECOM46510.2021.9685589](https://doi.org/10.1109/GLOBECOM46510.2021.9685589). URL: <https://globecom2021.ieee-globecom.org/>.
 - [7] Alex Marchioni et al. “Anomaly Detection based on Compressed Data: an Information Theoretic Characterization”. In: (Oct. 2021). DOI: [10.36227/techrxiv.16738171.v1](https://doi.org/10.36227/techrxiv.16738171.v1).