Automatic evaluation of labels in trajectories with anomaly detection

Christof Leutenegger
Master Thesis, FS 2020
Supervisors: Henry Martin, Jannik Hamper, Prof. Dr. Martin Raubal

IKG
Institut für Kartografie
und Geoinformation

geoinformation
engineering
ikg eth zurich

1 Introduction

Nowadays, the measurement of human movement is done with the help of trajectories based on GPS. The trajectories generated in this way contain a lot information on various problems, such as to make transportation more sustainable and using transport infrastructure efficiently. In order to derive knowledge from trajectories, it is important that the data they hold is reliable. To measure whether they are reliable we need to find anomalies in trajectories and communicate their quality accordingly.

2 Methodolgy

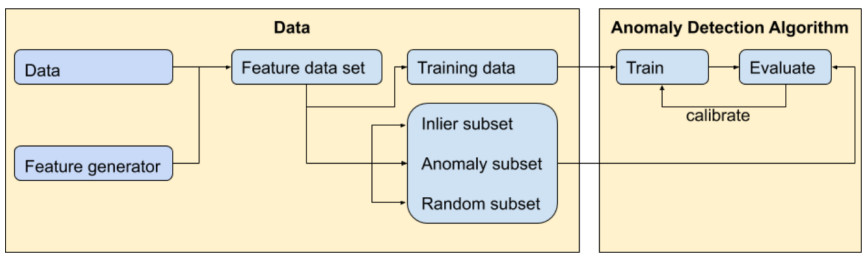

Fig. 1. Workflow for testing and comparing anomaly detection algorithms.

In order to measure and detect anomalies, features must be generated to compare anomalies with normal instances. We selected features to represent the properties of the trajectories in term of velocity, distance and geometric shape. We then transformed the feautures into log-space to obtain symmetric features distributions.

Tab. 1. Features generated for anomaly detection.

Feature	Description
Length	Sum of distance between all successive points.
Average Speed	Length divided by the duration.
Baseline/Length Ratio	Distance between start and end point divided by the length.
Corner Count	Sum of all turning points above a threshold.

For this thesis we used the car trajectories of the SBB Green Class study[1]. To evaluate the selected algorithms we generated three subsets: a subset containing only normal trajectories, a subset containing only anomalies, and â subset of mixed, randomly drawn trajectories.

Three anomaly detection algorithm were used to generate an anomaly score: Local Outlier Factor[2], Isolation Forest[3], and Minimum Covariance Determinant[4].

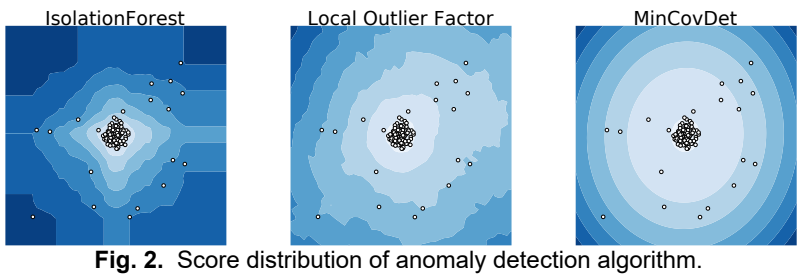

Fig. 2. Score distribution of anomaly detection algorithm.

As testing we used binary classification metrics to measure the precision at which the algorithms can differentiate between normal instances and anomalies. True positive rate (TPR) for the normal instances, true negative rate (TNR) for the anomaly subset, and negative predictive value (NPV) and AUC for the random subset.

3 Results

In our findings, we can see that the three algorithms perform mostly similar across all three test sets. Only the Isolation Forest has problems detecting anomalies in the anomaly subset (TNR) but is then again the algorithm with the best performance in the random subset (NPV and AUC).

Tab. 2. Testing results of the different algorithms.

Algorithm	TPR	TNR	NPV	AUC
LOF	0.83	0.74	0.66	0.83
MCD	0.84	0.70	0.62	0.87
IForest	0.8	0.57	0.73	0.90

In Fig. 3 we see the ROC to the corresponding AUC. We can see that the ability to distinguish between normal and abnormal instances is good up to the threshold we set. From then on, the algorithm is no longer able to distinguish, at least with the given features, whether the next included point is an anomaly or not.

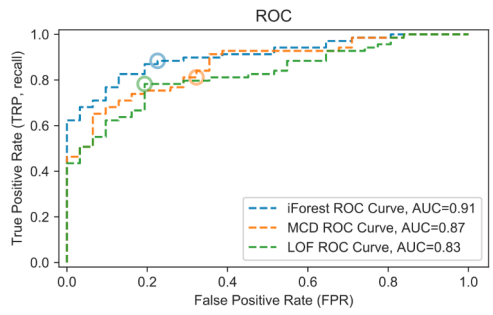

Fig. 3. ROC with thresholds encircled.

4 Discussion

The detection of anomalies with this techniques is satisfactory for such a simple approach. We see potential for improvement in the following points:

- Adding additional features to better distinguish between anomalies and normal instances.
- Extend IForest to handle better clustered anomalies, which are likely responsible for the low detection rate in the anomaly test set.
- Increase the number of instances in the test sets to increase the reliability of the analysis

5 References

[1] P. Rousseeuw und K. Driessen, „A Fast Algorithm for the Minimum Covariance Determinant Estimator“, Technometrics, Bd. 41, S. 212–223, Aug. 1999, doi: 10.1080/00401706.1999.10485670.

[2] H. Martin, H. Becker, D. Bucher, D. Jonietz, M. Raubal, und K. W. Axhausen, „Begleitstudie SBB Green Class - Abschlussbericht“, IVT, ETH Zürich, Working Paper, 2019. Available under: https://www.research-collection.ethz.ch/handle/20.500.11850/353337.

[3] F. T. Liu, K. M. Ting, und Z.-H. Zhou, „Isolation Forest“, in Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, USA, Dez. 2008, S. 413–422, doi: 10.1109/ICDM.2008.17.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, und J. Sander, „LOF: identifying density-based local outliers“, in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dallas, Texas, USA, Mai 2000, S. 93–104, doi: 10.1145/342009.335388.