ID: 307

Multi-Criteria Benchmarking Methodology for Evaluation of Anomaly Detection Algorithms

DAVID RENAUDIE, MARIA A. ZULUAGA, RODRIGO ACUNA-AGOST
Research, Innovation and Ventures Division
Amadeus SAS

amadeus

Abstract

In the scope of fraud detection for IT systems such as in Travel IT, it is crucial to build an efficient methodology to compare performance of unsupervised or semi-supervised anomaly detection algorithms. In an industrial context, where very few labelled data are available and anomalies are usually of multiple types, it is needed to unambiguously rank algorithms in order to choose which one(s) to implement in a production environment - a costly task.

We propose a concrete approach to evaluate algorithms' performance, by first building multiple benchmarking datasets from limited available positives, and using a "safe random" sampling approach to get negatives from unlabeled production data. Then, we train and test several candidate algorithms in novelty detection mode, using a kx2 cross-validation method. For comparing algorithms' performance, we rely on a metric used in ranked retrieval tasks, and aggregate scores obtained on different benchmarking datasets into a single ordering of all algorithms. We present results obtained on real industrial data, showing exposure to genuine anomalies.

Objective

Industrial problem: detect new domain-specific fraud patterns in production-level data of a reservation system.
→ What are the **best unsupervised algorithms** to be developed and deployed in a **production** environment (costly) ?

Questions

1. How to build a **benchmarking** dataset with **scarce labeled data** ?
2. How to perform **training** / **testing** in an unsupervised setup?
3. How to **evaluate** performance of unsupervised algorithms ?
4. How to **combine** several evaluation metrics into a unique ranking ?

Materials & Setup

Materials: Data

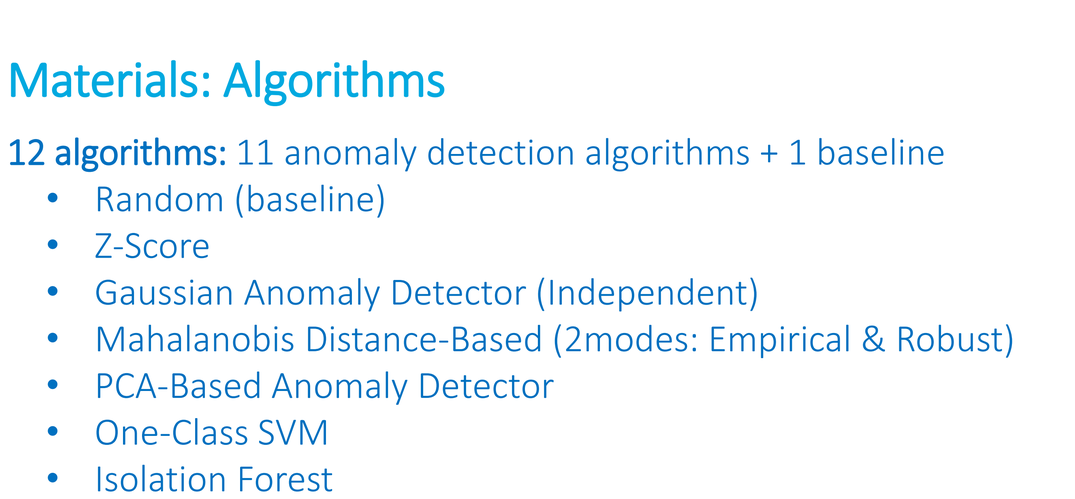
- **Before ETL:** Daily raw passenger records updates (anonymized) from a reservation system → Several Mios structured data records
- **After ETL:** two geographical markets, expert-designed features → 350k travelers records, 35-100+ features

Materials: Algorithms

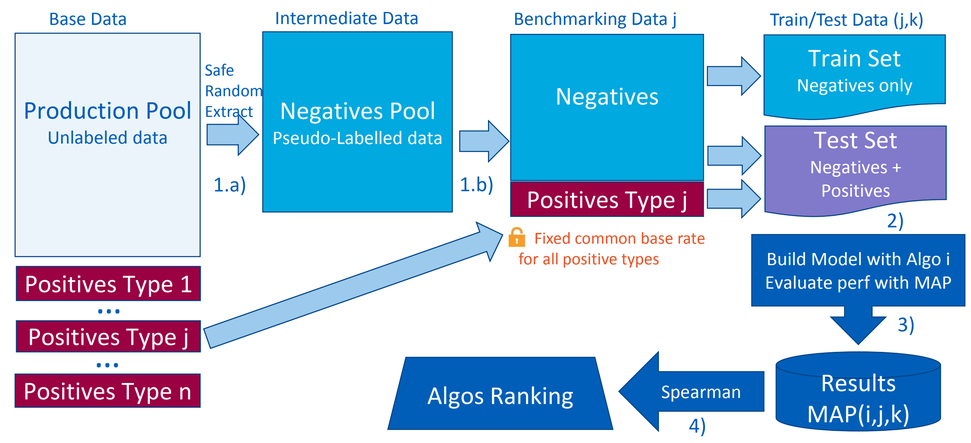
12 algorithms: 11 anomaly detection algorithms + 1 baseline

- Random (baseline)
- Z-Score
- Gaussian Anomaly Detector (Independent)
- Mahalanobis Distance-Based (2modes: Empirical & Robust)
- PCA-Based Anomaly Detector
- One-Class SVM
- Isolation Forest
- Gaussian Mixture Model
- Deep Auto-Encoder
- Local Outlier Factor
- Least Squares Anomaly Detector

Setup: Benchmarking / Production



Methodology



1) Benchmarking Dataset Construction

a) Build **Negatives Pool** dataset with a "Safe Random" approach
b) Build **1 dataset per positive type,** with fixed common base rate

2) Training/Testing: k×2 Cross-Validation

- **Novelty detection mode**: learn on uncontaminated training datasets to simulate discovering of never-seen-before occurrences
- **Increase k** to reduce evaluation metric **variability**

3) Choice of Evaluation Metric

- Precision, Recall, F1-Score are set-based, but order matters
- Need ranking evaluation → **Mean Average Precision** (MAP)

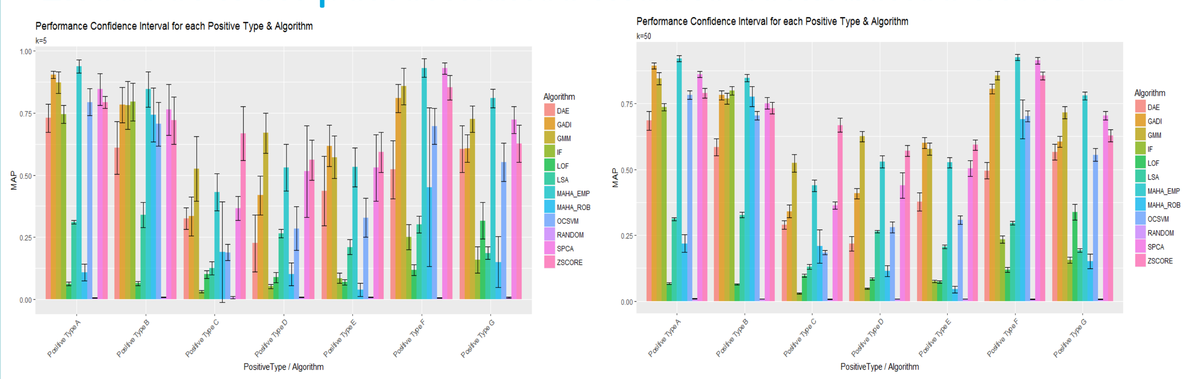
4) Multi-Criteria Aggregation

- From 1 MAP value per algorithm per positive type, to one ranking

→ **SpearMan** weighted rank aggregation with relative importance

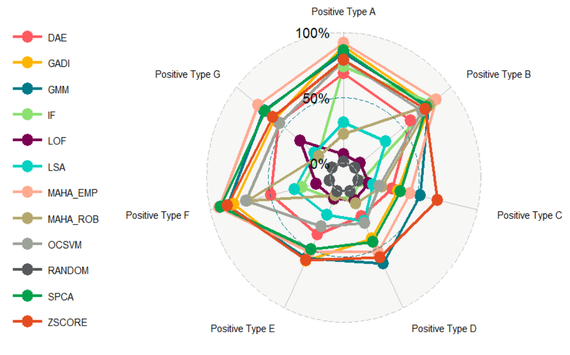
Experiments and Results

Effect of k on performance metric confidence interval



Benchmarking Results

12 algorithms / 7 positive types



Radar plot of Benchmarking Results

Production Results

Expert task force 300 top anomalies

- Found new patterns: fake names, abusive re-bookings, test records, end-of-ops, rail sync flows…
- Raised relevant functional/business
- Identified & fixed a few data issues

Implemented in Production: GMM

PNR Record Locator	Owner Office ID	Airlines	Score
WSB[redacted]	PAR[redacted]	[redacted]	1.00
WTS[redacted]	PAR[redacted]		0.94
WRS[redacted]	NCE[redacted]		0.68

Screenshot of Production-running UI