How Useful are Hand-crafted Data? Making Cases for Anomaly Detection Methods

Len Du
Australian National University
len.du.public@gmail.com

Marcus Hutter
Australian National University
www.hutter1.net

Motivation

- Small data is good
- Anomaly detection is good for small data.

Small data is good

- The importance of small data has been acknowledged in principle.
- Hidden hypothesis in real papers: bigger data are always better.
- Sustainability of the field.

Anomaly detection as an example.

- high-dimensional large datasets from real world are hard to comprehend
- If we have always been thinking “motivating examples” are good practice…… Why not systematically try multiple algorithms on the same small examples?

Outlier detection methods

- Basic kNN methods
  - kth nearest neighbour
  - kNNw
  - LIC
  - Outlier Detection using Indegree Number
- Methods based on LOF
  - Local outlier factor
  - LDOF
  - sLOF

Basic kNN methods

- distance to its kth nearest neighbor
- sum of distances to its k nearest neighbors
- LIC:

$$\mathrm{OL}(p) = \mathrm{LIC}(p) = \text{k-th-dist}(p) + \frac{\text{k-dist}(p)}{k}$$

- Outlier Detection using Indegree Number:

$$\mathrm{Ind}(p) = |\{q \in \{1..P\} \mid p \in \mathrm{kNN}(q)\}|$$

$$\mathrm{OL}(p) = \mathrm{ODIN}(p) = \frac{-\mathrm{Ind}(p)}{k}$$

Methods based on LOF

- Local outlier factor (LOF)

$$\mathrm{OL}(p) = \mathrm{LOF}(p) = \frac{1}{|\mathrm{kNN}(p)|} \sum_{q \in \mathrm{kNN}(p)} \frac{\mathrm{lrd}(q)}{\mathrm{lrd}(p)}$$

$$\mathrm{lrd}(p) = \frac{|\mathrm{kNN}(p)|}{\sum_{q \in \mathrm{kNN}(p)} \text{reach-dist}(p, q)},$$

$$\text{reach-dist}(p, q) = \max\{\text{k-th-dist}(q), D_{p,q}\} .$$

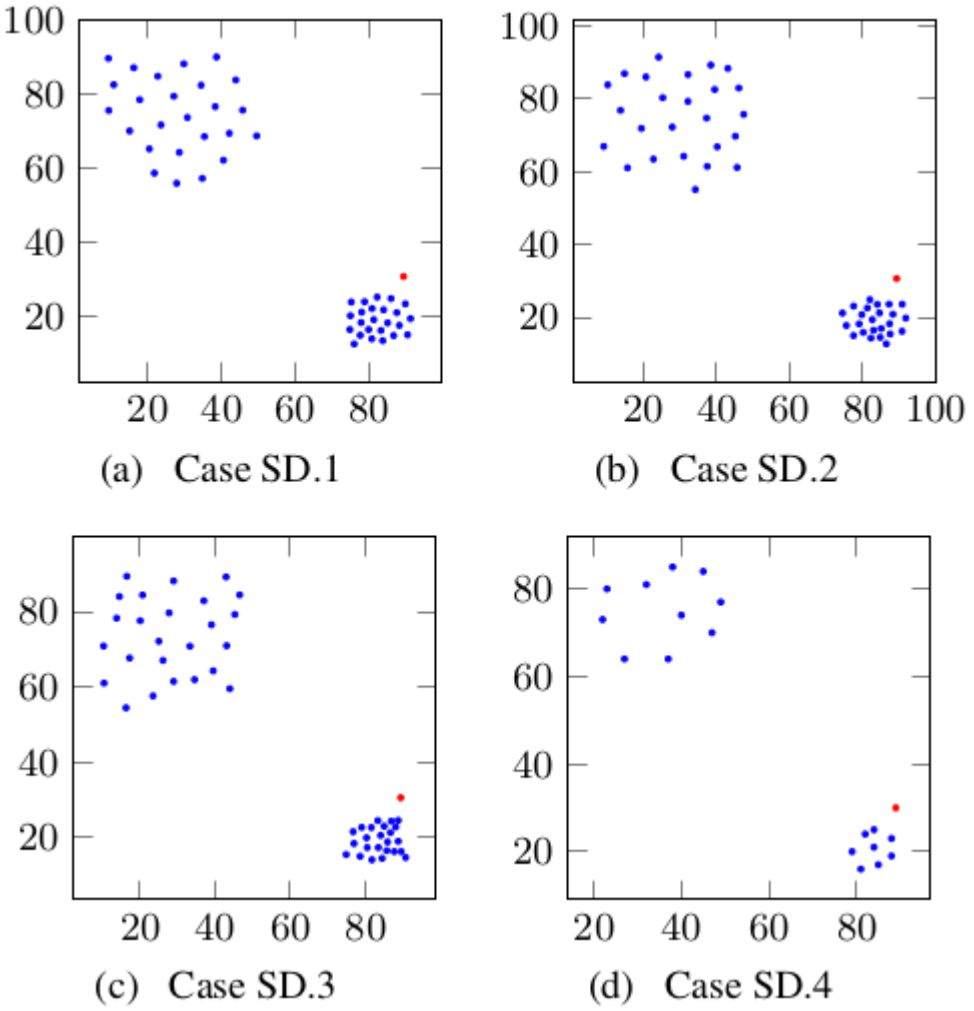
Methods based on LOF

- LDOF
  - Improved LOF
- sLOF
  - mistaking the reachability “distance” with the actual distance.

Challenging the detection algorithms

- No other hyperparameters than k ; letting each algorithm do its best regardless of k.
- Sparse and dense clusters
- Straight line and outlier
- Case where LDOF was claimed to have advantages over LOF
- Grid patterns

Sparse and dense clusters

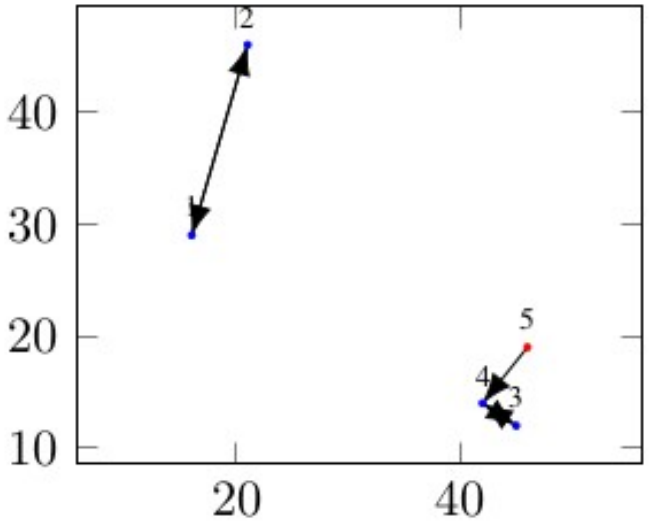


Sparse and dense clusters

	k^{th}NN	kNNw	LIC	ODIN	LOF	sLOF	LDOF
Case SD.1	✗	✗	✗	✓	✓	✓	✓
Case SD.2	✗	✗	✗	✓	✓	✓	✓
Case SD.3	✗	✗	✗	✓	✓	✓	✓
Case SD.4	✗	✗	✗	✓	✓	✓	✓

ODIN also succeeded in disambiguating these outliers, despite being a “global” method along with k th NN and kNNw.

Sparse and dense clusters

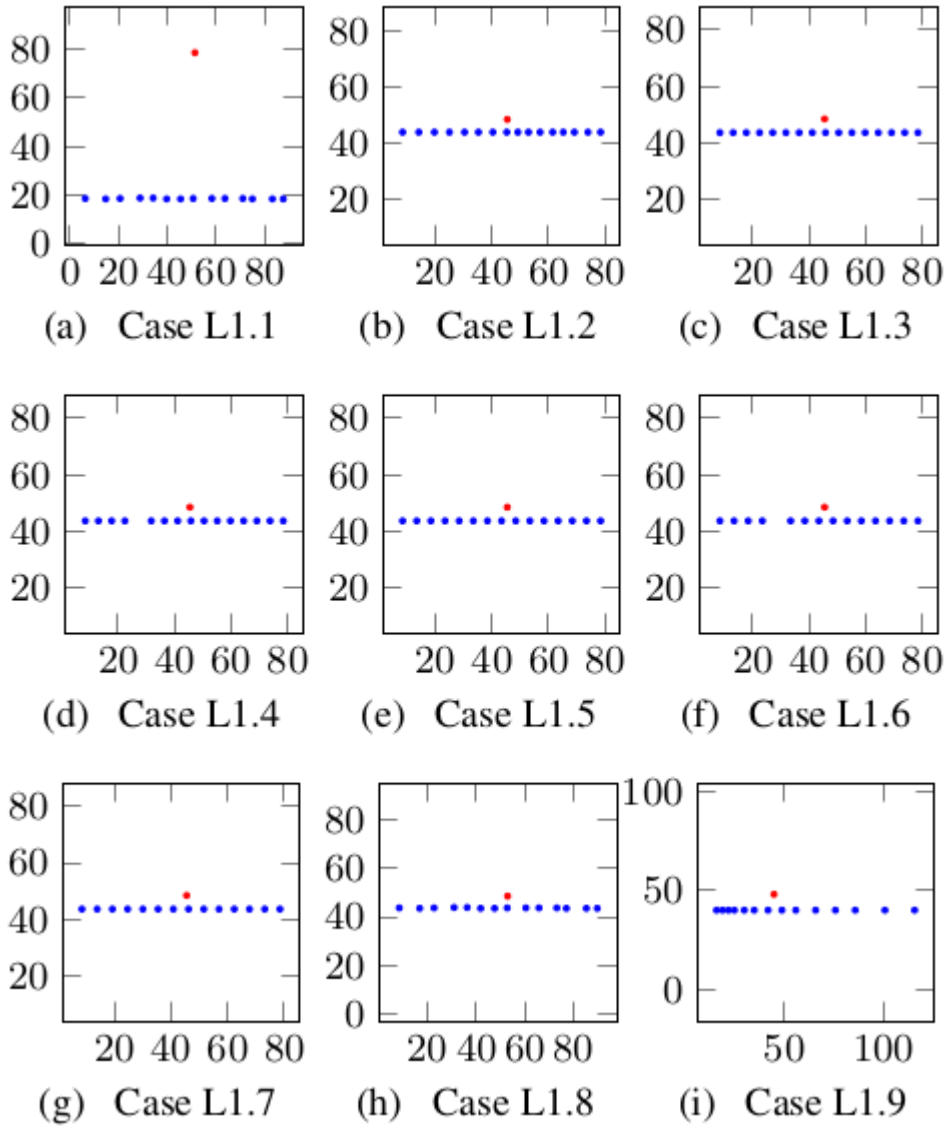


- the kNN set of both inliers and outliers are composed of inliers, so inliers are likely to be in at least one point’s kNN set, yet the outlier can be in no point’s kNN set.

Straight line and outlier

- Inliers form (approximately) a line (segment) or more generally, lie in a subspace of lower dimensions.

Straight line and outlier

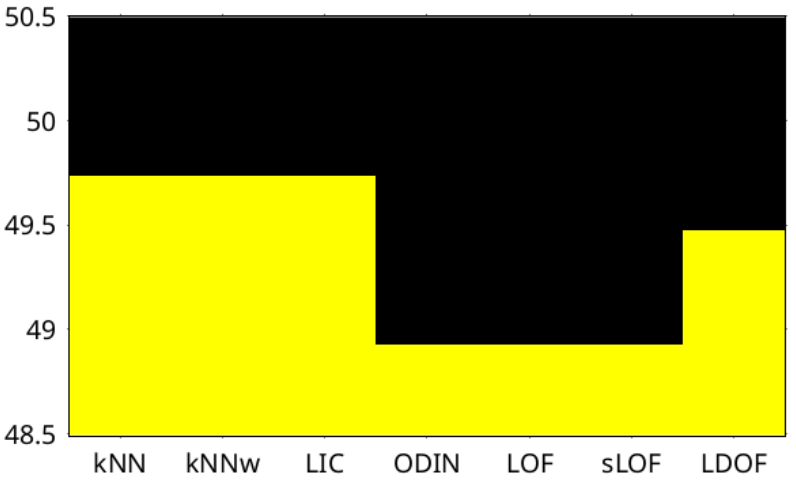


Straight line and outlier

	k^{th} NN	kNNw	LIC	ODIN	LOF	sLOF	LDOF
Case L1.1	✓	✓	✓	✓	✓	✓	✓
Case L1.2	✗	✗	✗	✗	✗	✗	✗
Case L1.3	✓	✓	✓	✓	✓	✓	✗
Case L1.4	✓	✓	✓	✓	✓	✓	✗
Case L1.5	✓	✓	✓	✓	✓	✓	✗
Case L1.6	✓	✓	✓	✓	✓	✓	✗
Case L1.7	✗	✗	✗	✗	✓	✗	✗
Case L1.8	✗	✗	✗	✗	✗	✗	✗
Case L1.9	✗	✗	✗	✓	✗	✗	✗

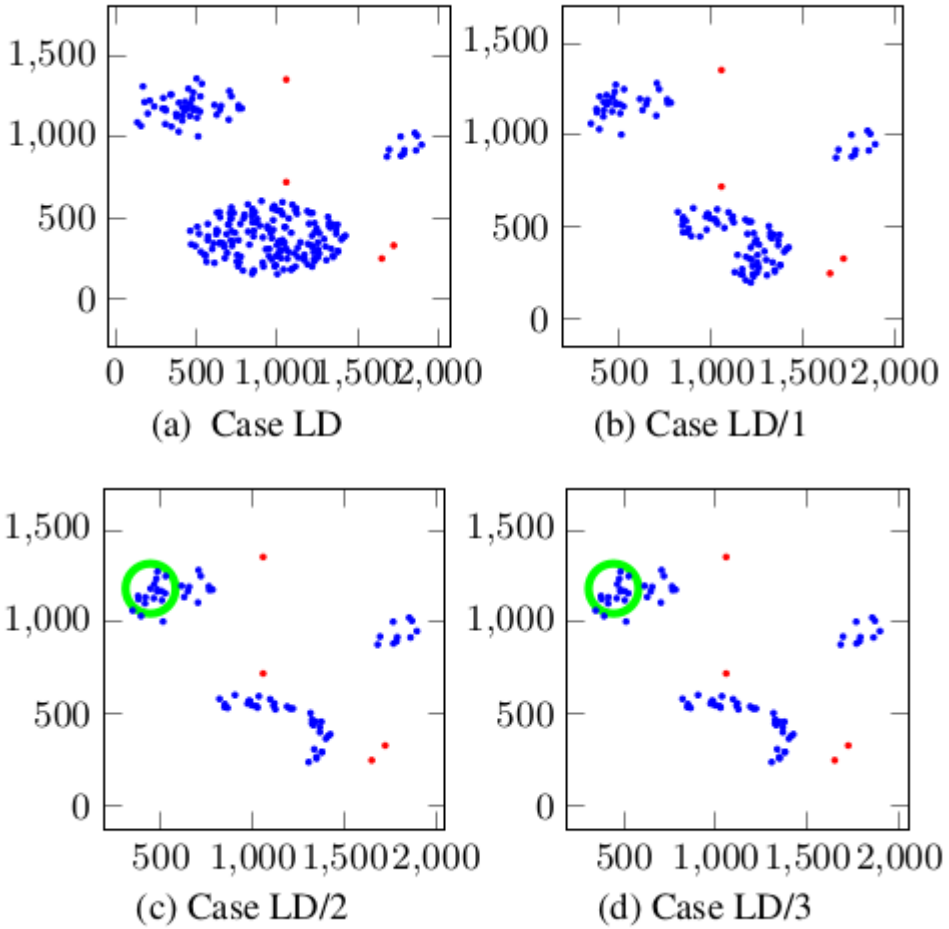
Straight line and outlier

- Only cases where outlier-to-line distance and gap sizes are comparable, are interesting. Otherwise, any algorithm works reasonably well.
- Move the outlier point against the line.
  - Monotonicity!
  - yellow (failure) black (success)



Case where LDOF was claimed to have advantages over LOF

- Recovered by parsing PDF, and with variants.



Case where LDOF was claimed to have advantages over LOF

	k^{th}NN	kNNw	LIC	ODIN	LOF	sLOF	LDOF
Case LD	✓	✓	✓	✗	✓	✓	✓
Case LD/1	✓	✓	✓	✗	✓	✓	✓
Case LD/2	✓	✓	✓	✗	✓	✓	✓
Case LD/3	✓	✓	✓	✓	✓	✓	✓

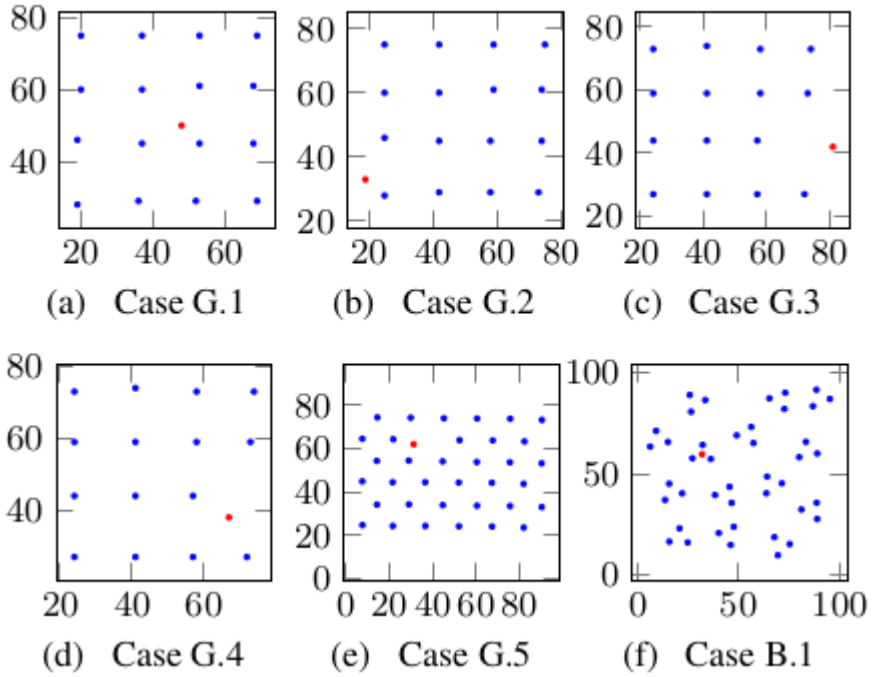
- It was only claimed that the case defeated kNN , kNNw and LOF when k > 10. In our tests all the algorithms did pass with k < 10. No conflict technically……
- Why ODIN alone failed ?

Case where LDOF was claimed to have advantages over LOF

- (Jumping to conclusion) The ODIN method is scale-invariant within one configuration, like fractals.
- But too scale-invariant. we cannot separate small-scale features from large-scale ones using the indegree numbers under a single k.
- Possible new algorithms: combine Ind(p) with other methods to filter out small-scale “outliers” ?

Grid patterns

- Algorithm abuse!
- Probably not intended for most unsupervised anomaly detection algorithms.



Grid patterns

- But it works (somewhat).

	k^{th} NN	kNNw	LIC	ODIN	LOF	sLOF	LDOF
Case G.1	✗	✗	✗	✗	✓	✗	✗
Case G.2	✓	✗	✓	✓	✓	✗	✗
Case G.3	✓	✓	✓	✓	✓	✓	✗
Case G.4	✗	✗	✗	✗	✓	✗	✗
Case G.5	✗	✗	✗	✗	✓	✗	✗

- LOF certainly has great potential to be adapted for grid-like patterns.

Discussion

- Both the size of each case and the total number of cases are much smaller than a typical review.
- Yet we have found an unexpected use case for outlier detection methods.
- Small, well designed examples provoke insights.