Good and Bad Neighborhood Approximations for Outlier Detection Ensembles

Evelyn Kirner, Erich Schubert, Arthur Zimek

October 4, 2017, Munich, Germany

LMU Munich; Heidelberg University; University of Southern Denmark

Outlier Detection

The intuitive definition of an outlier would be "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism".

Hawkins [Haw80]

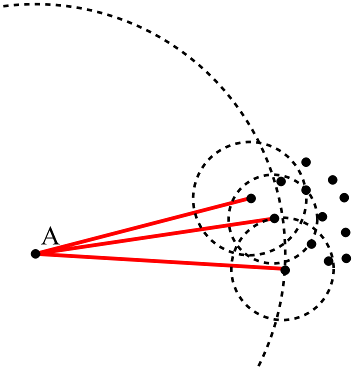
An outlying observation, or "outlier," is one that appears to deviate markedly from other members of the sample in which it occurs.

Grubbs [Gru69]

An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data

Barnett and Lewis [BL94]

Outlier Detection



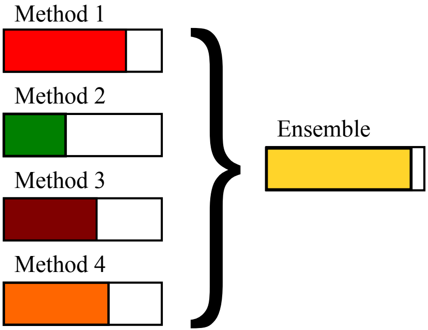
- Estimate density $= \frac{\text{Number of neighbors}}{\text{Distance}}$ (or e.g. KDEOS [SZK14])
- Least dense points are outliers (e.g. kNN outlier [RRS00])
- Points with relatively low density are outliers (e.g. LOF [Bre+00])

Ensembles

Assume a binary classification problem
(e.g., “does some item belong to class ‘A’ or to class ‘B’?”)

- in a “supervised learning” scenario, we can learn a model (i.e., train a classifier on training samples for ‘A’ and ‘B’)
- some classifier (model) decides with a certain accuracy
- error rate of the classifier: how often is the decision wrong?
- “ensemble”: ask several classifiers, combine their decisions (e.g., majority vote)

Ensembles

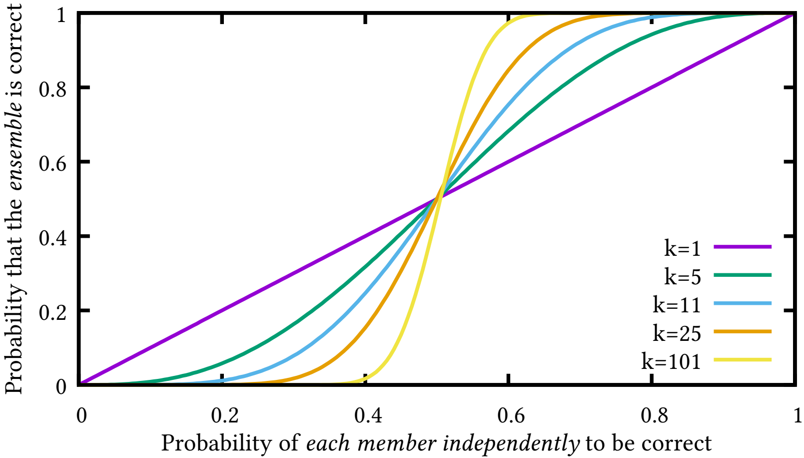


The ensemble will be much more accurate than its components, *if*

- the components decide independently,
- and each component decides more accurate than a coin.

In supervised learning, a well developed theory for ensembles exists in literature.

Error-Rate of Ensembles



$$P(k,p) = \sum_{i=\lceil k/2 \rceil}^{k} \binom{k}{i} p^i (1-p)^{k-i}$$

Diversity for Outlier Detection Ensembles

Different ways to get diversity:

- feature bagging: combine outlier scores learned *on different subsets of attributes* [LK05]
- use the same base method with *different parameter choices* [GT06]
- combine *different base methods* [NAG10; Kri+11; Sch+12]
- use *randomized base methods* [LTZ12]
- use *different subsamples* of the data objects [Zim+13]
- learn on data *with additive random noise* components (“perturbation”) [ZCS14]
- use approximate neighborhoods (this paper)

Approximate Methods for Outlier Detection

Approximate nearest neighbor search has often been used for *accelerating* outlier detection, but in a fundamentally different way:

- Find *candidates* using approximation, then refine the top candidates with exact computations [Ora+10; dCH12]
- Ensemble of approximate nearest neighbor methods, then detect outliers using the ensemble neighbors [SZK15]

- In this paper, we study building the ensemble *later*:
  1. Find approximate nearest neighbors
  2. Compute outlier scores for each set of approximate neighbors
  3. Combine resulting scores in an ensemble

Embrace the Uncertainty of Approximate Neighborhoods

Ensembles need to have diverse members to work.

Other ensemble methods try to (occasionally quite artificially)
induce diversity in the outlier score estimates,
often by changing the neighborhoods.

We take advantage of the "natural" variance in neighborhood estimations
delivered by approximate nearest neighbor search.

Different approximate nearest neighbor methods have different bias,
which can be beneficial or not for outlier detection.

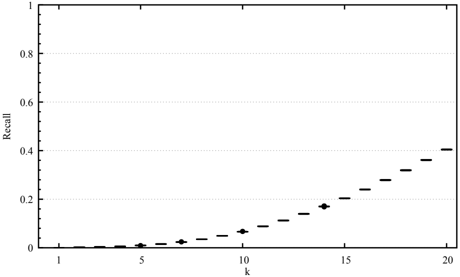
Approximate Nearest-Neighbors

We experimented with the following ANN algorithms:

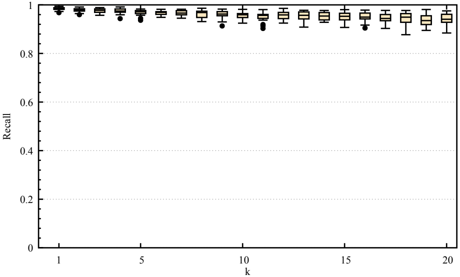
- NN-Descent [DCL11]
  Begin with random nearest neighbors, refine via closure.
  (We use only 2 iterations, to get enough diversity.)
- Locality Sensitive Hashing (LSH) [IM98; GIM99; Dat+04]
  Discretize into buckets using random projections
- Space filling curves (Z-order [Mor66])
  With random projections; project onto a one-dimensional order
  (similar to [SZK15], but with Z-order only)

Experiments: Recall of ANN

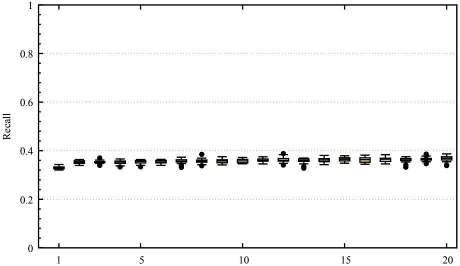
NN-Descent



LSH



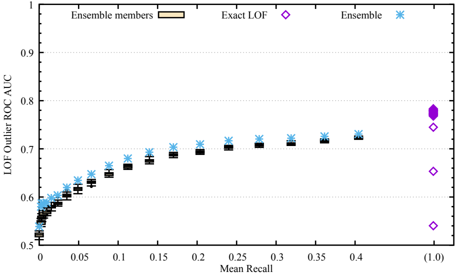
SFC



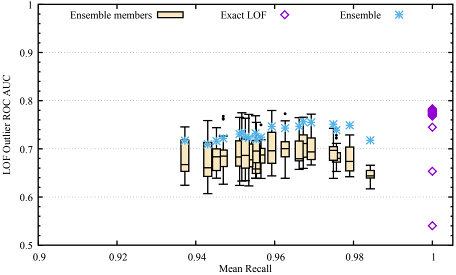
But is *nearest neighbor recall* what we need?

Experiments: Outlier ROC AUC

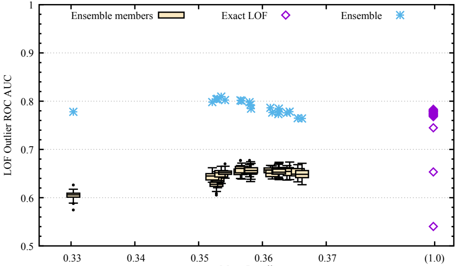
NN-Descent



LSH



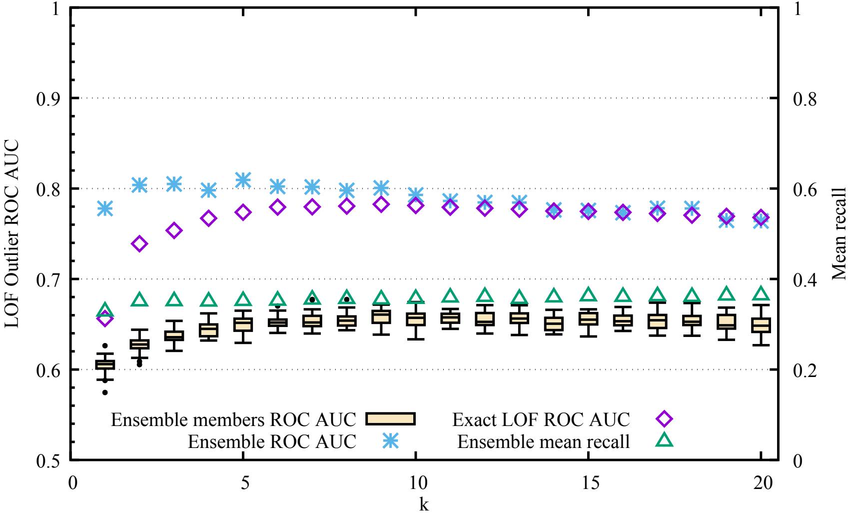
SFC



There is *no strong correlation* between neighbor recall and outlier ROC AUC.

Experiments: Space-Filling-Curves

Space-Filling-Curves worked surprisingly well (also in [SZK15]):



Observations

NN-descent: recall improves a lot with k (larger search space).
But we observed very little variance (diversity),
and thus only marginal improvement.

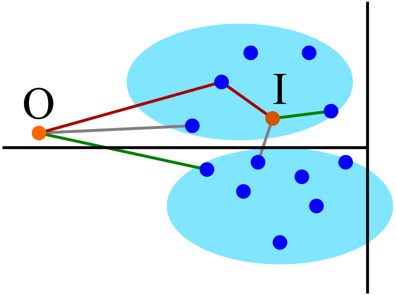
LSH: very good recall, in particular for small k.
Ensemble better than most members, but not as good as exact.

SFC: Intermediate recall – but very good ensemble performance.

- If we have too high recall, we lose diversity.
- If we have too low recall, the outliers are not good enough.
- A working ensemble needs to balance these two.

Beneficial Bias of Space-Filling Curves

Why approximation is good enough (or even better):



Approximation error caused by a space filling curve:

Black lines: neighborhoods not preserved

Grey lines: real nearest neighbor

Green lines: real 2NN distances

Red lines: approximate 2NN distances

The effect on cluster analysis is substantial, while for outlier detection it is minimal but rather beneficial.

- Since outlier scores are based on density *estimates* anyway – why would we need *exact* scores (that are still just some approximation of an inexact property)?
- Essentially the same motivation as for ensembles based on perturbations of neighborhoods (e.g., by noise, subsamples, or feature subsets) would also motivate to base an outlier ensemble on approximate nearest neighbor search.

Conclusions

When is the bias of the neighborhood approximation beneficial?

Presumably when the approximation error leads to a stronger underestimation of the local density for outliers than for inliers.

➡ We should study the bias of NN approximation methods.

Thank You!

Questions?

References i

[BL94] V. Barnett and T. Lewis. *Outliers in Statistical Data.* 3rd. John Wiley&Sons, 1994.

[Bre+00] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. "LOF: Identifying Density-based Local Outliers". In: *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX.* 2000, pp. 93–104.

[Dat+04] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. "Locality-sensitive hashing scheme based on p-stable distributions". In: *Proceedings of the 20th ACM Symposium on Computational Geometry (ACM SoCG), Brooklyn, NY.* 2004, pp. 253–262.

[dCH12] T. de Vries, S. Chawla, and M. E. Houle. "Density-preserving projections for large-scale local anomaly detection". In: *Knowledge and Information Systems (KAIS)* 32.1 (2012), pp. 25–52.

[DCL11] W. Dong, M. Charikar, and K. Li. "Efficient K-Nearest Neighbor Graph Construction for Generic Similarity Measures". In: *Proceedings of the 20th International Conference on World Wide Web (WWW), Hyderabad, India.* 2011, pp. 577–586.

[GIM99] A. Gionis, P. Indyk, and R. Motwani. "Similarity Search in High Dimensions via Hashing". In: *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB), Edinburgh, Scotland.* 1999, pp. 518–529.

[Gru69] F. E. Grubbs. “Procedures for Detecting Outlying Observations in Samples”. In: *Technometrics* 11.1 (1969), pp. 1–21.

[GT06] J. Gao and P.-N. Tan. “Converting Output Scores from Outlier Detection Algorithms into Probability Estimates”. In: *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China.* 2006, pp. 212–221.

[Haw80] D. Hawkins. *Identification of Outliers.* Chapman and Hall, 1980.

[IM98] P. Indyk and R. Motwani. “Approximate nearest neighbors: towards removing the curse of dimensionality”. In: *Proceedings of the 30th annual ACM symposium on Theory of computing (STOC), Dallas, TX.* 1998, pp. 604–613.

[Kri+11] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. “Interpreting and Unifying Outlier Scores”. In: *Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ.* 2011, pp. 13–24.

[LK05] A. Lazarevic and V. Kumar. “Feature Bagging for Outlier Detection”. In: *Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL.* 2005, pp. 157–166.

[LTZ12] F. T. Liu, K. M. Ting, and Z.-H. Zhou. "Isolation-Based Anomaly Detection". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.1 (2012), 3:1–39.

[Mor66] G. M. Morton. *A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing.* Tech. rep. International Business Machines Co., 1966.

[NAG10] H. V. Nguyen, H. H. Ang, and V. Gopalkrishnan. "Mining Outliers with Ensemble of Heterogeneous Detectors on Random Subspaces". In: *Proceedings of the 15th International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan.* 2010, pp. 368–383.

[Ora+10] G. H. Orair, C. Teixeira, Y. Wang, W. Meira Jr., and S. Parthasarathy. "Distance-Based Outlier Detection: Consolidation and Renewed Bearing". In: *Proceedings of the VLDB Endowment* 3.2 (2010), pp. 1469–1480.

[RRS00] S. Ramaswamy, R. Rastogi, and K. Shim. "Efficient algorithms for mining outliers from large data sets". In: *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX.* 2000, pp. 427–438.

[Sch+12] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. "On Evaluation of Outlier Rankings and Outlier Scores". In: *Proceedings of the 12th SIAM International Conference on Data Mining (SDM), Anaheim, CA.* 2012, pp. 1047–1058.

References iv

[SZK14] E. Schubert, A. Zimek, and H.-P. Kriegel. “Generalized Outlier Detection with Flexible Kernel Density Estimates”. In: *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA.* 2014, pp. 542–550.

[SZK15] E. Schubert, A. Zimek, and H.-P. Kriegel. “Fast and Scalable Outlier Detection with Approximate Nearest Neighbor Ensembles”. In: *Proceedings of the 20th International Conference on Database Systems for Advanced Applications (DASFAA), Hanoi, Vietnam.* 2015, pp. 19–36.

[ZCS14] A. Zimek, R. J. G. B. Campello, and J. Sander. “Data Perturbation for Outlier Detection Ensembles”. In: *Proceedings of the 26th International Conference on Scientific and Statistical Database Management (SSDBM), Aalborg, Denmark.* 2014, 13:1–12.

[Zim+13] A. Zimek, M. Gaudet, R. J. G. B. Campello, and J. Sander. “Subsampling for Efficient and Effective Unsupervised Outlier Detection Ensembles”. In: *Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL.* 2013, pp. 428–436.