Subsampling for Efficient and Effective Unsupervised Outlier Detection Ensembles

Arthur Zimek[1,2], Matthew Gaudet[1],
Ricardo J. G. B. Campello[1,3], Jörg Sander[1]

[1]University of Alberta, Edmonton, AB, Canada

[2]Ludwig-Maximilians-Universität München, Munich, Germany

[3]University of São Paulo, São Carlos, Brazil

SIGKDD 2013, Chicago, IL

Outline

Outline

What is an Outlier?

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, 1980]

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

[Grubbs, 1969]

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

[Barnett and Lewis, 1994]

What is an Outlier?

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, 1980]

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

[Grubbs, 1969]

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

[Barnett and Lewis, 1994]

Ensembles for Outlier Detection?

- statistical reasoning about outliers: rich literature, results accumulated over centuries
- database/data mining research: $\approx$ 15 years, several models and variants, many variants for efficiency (top-k, filter-refinement-approaches)
- efficiency variants aim at approximating the basic models, not the statistical intuition
  They are approximating approximations!
- each model has strengths and weaknesses (bias, assumptions)
- combination of models for outlier detection is as promising as for classification or clustering but did not gain much attention so far

Outline

Existing Ensemble Methods for Outlier Detection

- feature bagging: combine outlier scores learned on different subsets of attributes [Lazarevic and Kumar, 2005]
  - problem: combination of scores when scores may scale differently (e.g., due to different dimensionality)
- normalize scores (also aiming at combination of scores from different base learners)
  - by trained Sigmoid functions or mixture modeling [Gao and Tan, 2006]
  - by normalization by standard deviation [Nguyen et al., 2010]
  - based on properties of the score distribution [Kriegel et al., 2011]
- greedy combination of *diverse* base learners [Schubert et al., 2012]

Methods for Inducing Diversity

- feature bagging [Lazarevic and Kumar, 2005]
  - requires proper normalization
- using different parameters [Gao and Tan, 2006]
  - requires proper normalization
  - results usually in rather correlated models [Schubert et al., 2012]
- using different base methods [Nguyen et al., 2010, Kriegel et al., 2011, Schubert et al., 2012]
  - requires proper normalization

Here...

...we discuss an ensemble based on learning diverse models on different *subsamples* of the data.

Theoretical Insights

- impact of diversity of models (empirical study) [Schubert et al., 2012]
- algorithmic patterns (position paper) [Aggarwal, 2012]
  - sequential vs. independent learning of models
  - data centered vs. model centered ensembles

But...

...why should, what has a clear theoretical background in supervised learning, also work in unsupervised outlier detection?

Outline

Motivation

Related Work

Outlier Detection Ensembles Based on Subsampling

Evaluation

Conclusion

Basic Considerations

- probability density function f represents the process "generating" the majority of the data (at least the inliers)
- data set X is a sample drawn from the true but unknown density distribution
- outlier methods, in order to compute outlier scores, try to estimate $f(x)$ around points x, using a more or less "rough" density estimate $\hat{f}_X(x)$
- assuming the correctness of the underlying outlier model of some method, the quality of the method's results crucially depends on the quality of the density estimate $\hat{f}_X$
- we show formally and empirically that a diverse ensemble of such outlier detectors is expected to show and does in fact show an improved performance over the individual ensemble members

Benefits of Ensembles for Outlier Detection

- given a true, smooth p.d.f. $f(x)$ and a data set X:

$$\hat{f}_X(x) = f(x) + v_X(x)$$

  where $v_X(x)$ is a random variable describing the error of the estimate due to the finite sample
- averaging multiple density estimates for each point x

$$\begin{aligned} E\{\hat{f}_X(x)\} &= E\{f(x)\} + E\{v_X(x)\} \\ &= f(x) + E\{v_X(x)\} \end{aligned}$$

- ranking of objects w.r.t. $E\{\hat{f}_X(x)\}$ is the same as the ranking w.r.t. the true density $f(x)$ (the “ideal ranking”), if just the *expectation* of the error $v_X(x)$ in the individual estimates is the same for every point x

Expected Error and Ranking Quality

- the expected error would obviously be the same for all points if the random variable that describes the error would not depend on x
- the expected error does not need to be the same in order to achieve the "ideal ranking" (i.e., the ranking due to the unknown true p.d.f. f)
  - $E\{v_X(x_1)\}$ and $E\{v_X(x_2)\}$ may differ for two points x_1 and x_2, as long as the difference does not cause an inversion between the actual ranks $E\{\hat{f}_X(x_1)\}$ and $E\{\hat{f}_X(x_2)\}$
- we do not even need to achieve the "ideal ranking" for all points as we only need to distinguish between outliers and inliers
  - rank inversions can happen among outliers and among inliers but should not swap inliers for outliers or vice versa

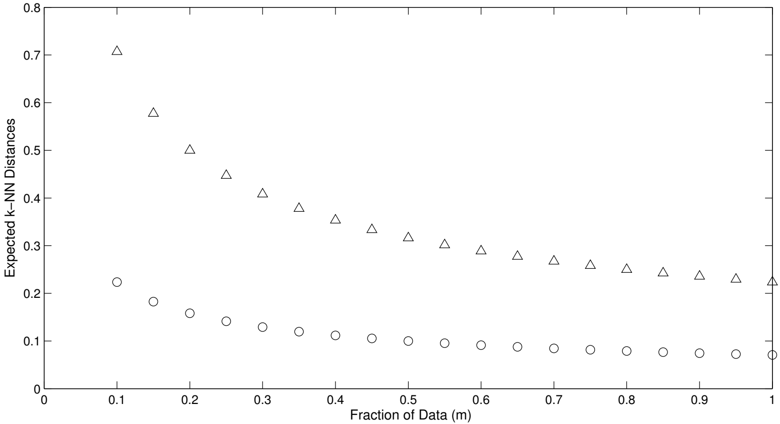
Benefits of Subsampling

- expected kNN distances decrease proportionally as a function of the subsampling rate m, $0 < m \leq 1$
- the *relative* contrast between densities remains constant between areas of different densities for subsamples of the data
  - justifies a subsampling procedure with even sampling probabilities in all areas of the data space
- ensemble setting: get different (diverse) estimates but keep the same expected density profile as the full data set
- the *absolute* contrast is increasing with smaller sample sizes

As a consequence. . .

the gap between inliers and outliers is increasing.

Expected Behaviour of k-NN Distances with Different Sample Sizes



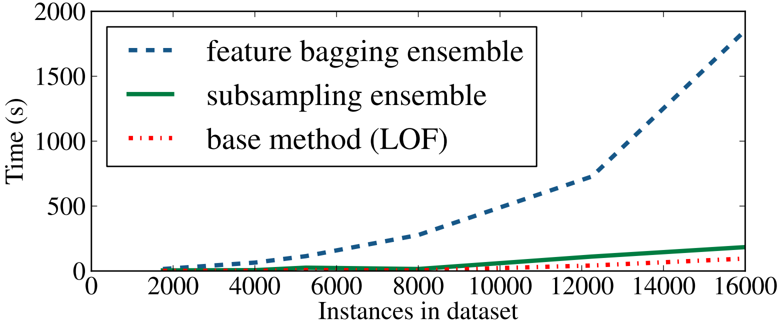
Behaviour of the expected 5-NN distances for two spheres with radius $r = 1$, in a 2D Euclidean space, containing $1000m$ (circles) and $100m$ (triangles) objects uniformly distributed (m is a fraction of the data).

Efficiency

- typical complexity of outlier detection methods: $\mathcal{O}(n^2)$ (due to kNN queries)
- common ensemble $\mathcal{O}(s \cdot n^2)$ for s ensemble members
- subsampling ensemble:
  - For each data object (n): kNN query on subsample ($m \cdot n$ for sample rate $0 < m < 1$)
  - repeat on s subsamples $\Rightarrow \mathcal{O}(n \cdot mn \cdot s)$
- example:
  - sample rate 10%
  - ensemble size 10 members
  - requires roughly the same runtime than a single base learner on the full data set
  - common ensemble: runtime 10 times the base learner's runtime

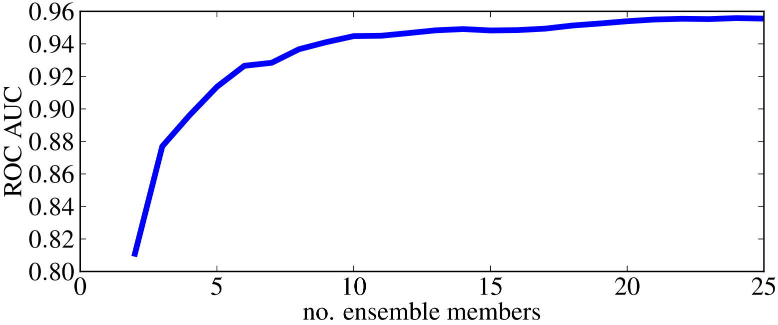
Outline

Efficiency



Runtime of LOF [Breunig et al., 2000], subsampling ensemble (10%, 25 members), and feature bagging ensemble (25 members) with increasing database size. (Implementation of ensemble, base methods, and competitor in ELKI [Achtert et al., 2013])

Ensemble Size



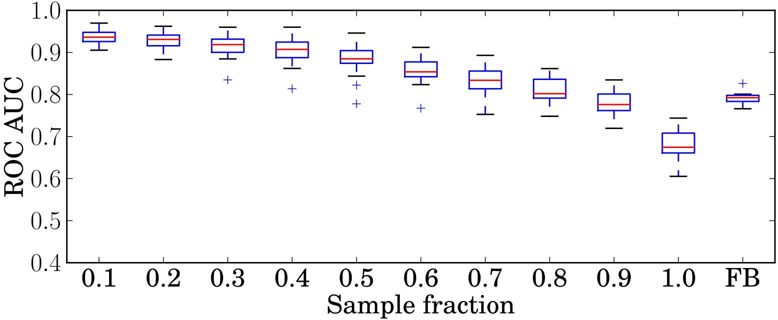
Quality with increasing ensemble size (example synthetic data set).

Synthetic Datasets

- two batches of 30 data sets
- dimensionality $d \in [20, \ldots, 40]$
- number of clusters $c \in [2, \ldots, 10]$
- for each cluster independently: number of points $n_{c_i} \in [600, \ldots, 1000]$
- points are generated from a Gaussian distribution, choosing for each cluster, for each attribute independently μ and σ
- resulting clusters are rotated, computing the corresponding covariance matrix Σ [Soler and Chin, 1985]
- annotate as outliers which Mahalanobis distance from the cluster center (using Σ) are larger than the theoretical 0.975 quantile
- expected amount of 2.5% outliers per dataset

Distribution of Results over 30 Datasets (batch1)

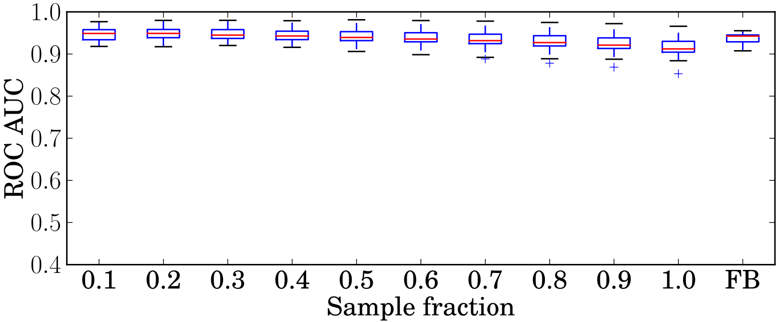
Base method: LDOF [Zhang et al., 2009]



ROC AUC for ensembles—different sample sizes as well as feature bagging (FB)—and base method (sample size=1.0), on the 30 datasets of batch1.

Distribution of Results over 30 Datasets (batch1)

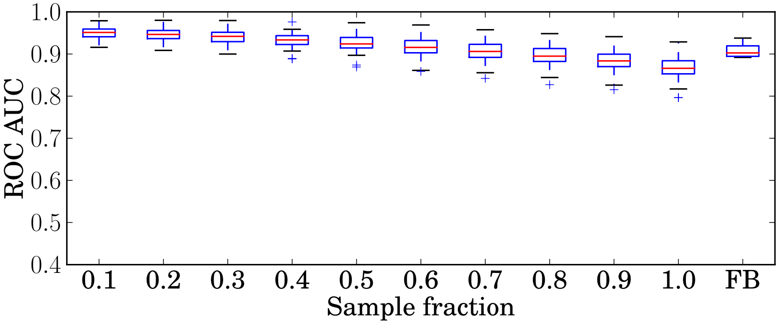
Base method: LOF [Breunig et al., 2000]



ROC AUC for ensembles—different sample sizes as well as feature bagging (FB)—and base method (sample size=1.0), on the 30 datasets of batch1.

Distribution of Results over 30 Datasets (batch1)

Base method: LoOP [Kriegel et al., 2009]



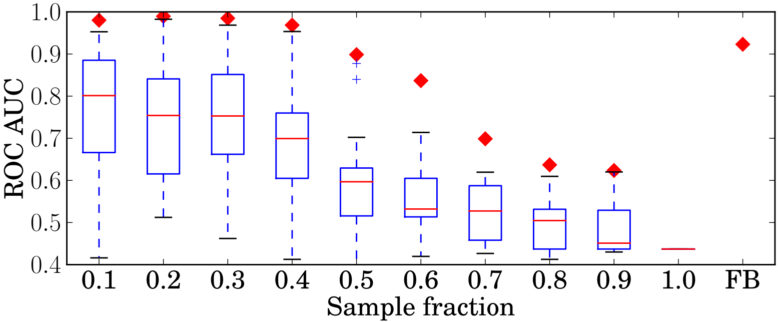
ROC AUC for ensembles—different sample sizes as well as feature bagging (FB)—and base method (sample size=1.0), on the 30 datasets of batch1.

UCI Datasets, Preprocessing for Evaluation

- Satimage, Lymphography, Segment (used also by Lazarevic and Kumar [2005])
- additional UCI data [Frank and Asuncion, 2010]: Wisconsin breast cancer (WBC) and Waveform Database Generator (waveform)
- take a sample from the most distinct class as outliers against rest (Segment: take GRASS, PATH, SKY in turn as outliers)

"Lymphography" – Different Sample Sizes

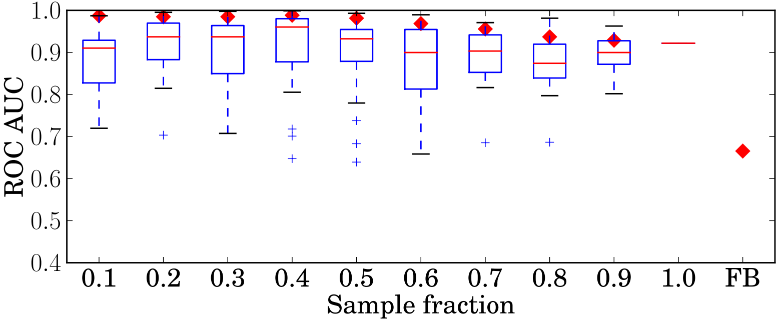
Base method: LDOF [Zhang et al., 2009]



ROC AUC for ensemble members of the subsampling ensemble for different sample sizes (boxes), the base method (sample size=1.0), and ensembles (diamonds)—on top of subsamples and feature bags (FB)—on dataset lymphography.

“Lymphography” – Different Sample Sizes

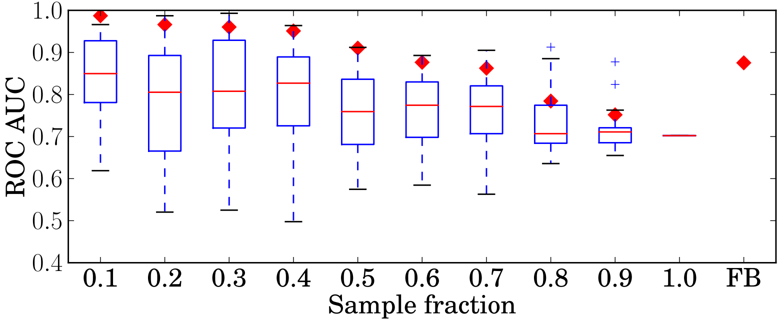
Base method: LOF [Breunig et al., 2000]



ROC AUC for ensemble members of the subsampling ensemble for different sample sizes (boxes), the base method (sample size=1.0), and ensembles (diamonds)—on top of subsamples and feature bags (FB)—on dataset lymphography.

“Lymphography” – Different Sample Sizes

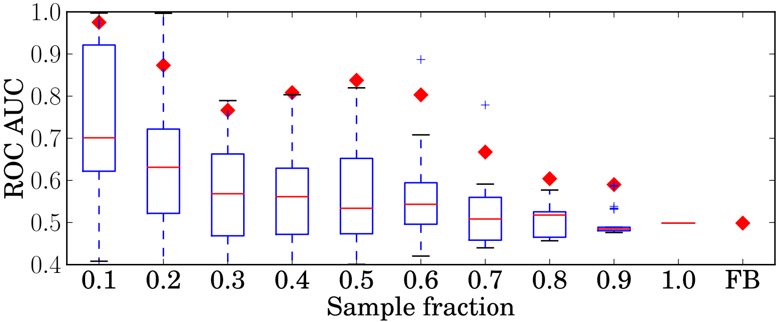
Base method: LoOP [Kriegel et al., 2009]



ROC AUC for ensemble members of the subsampling ensemble for different sample sizes (boxes), the base method (sample size=1.0), and ensembles (diamonds)—on top of subsamples and feature bags (FB)—on dataset lymphography.

“Wisconsin Breast Cancer” – Different Sample Sizes

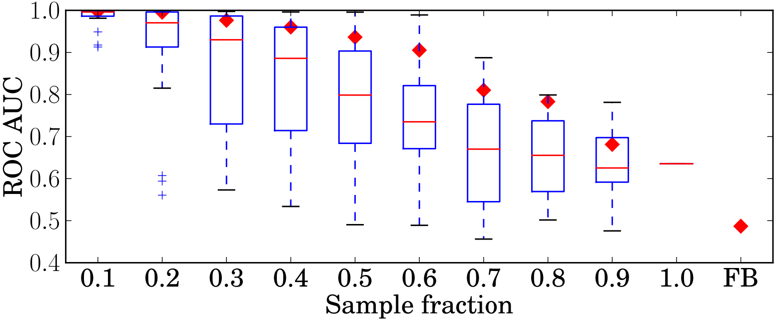
Base method: LDOF [Zhang et al., 2009]



ROC AUC for ensemble members of the subsampling ensemble for different sample sizes (boxes), the base method (sample size=1.0), and ensembles (diamonds)—on top of subsamples and feature bags (FB)—on dataset WBC.

"Wisconsin Breast Cancer" – Different Sample Sizes

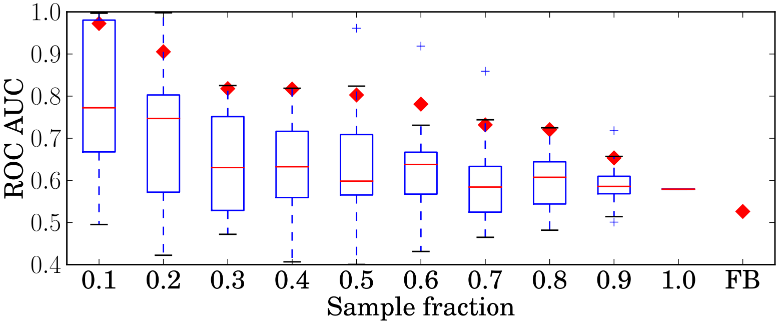
Base method: LOF [Breunig et al., 2000]



ROC AUC for ensemble members of the subsampling ensemble for different sample sizes (boxes), the base method (sample size=1.0), and ensembles (diamonds)—on top of subsamples and feature bags (FB)—on dataset WBC.

“Wisconsin Breast Cancer” – Different Sample Sizes

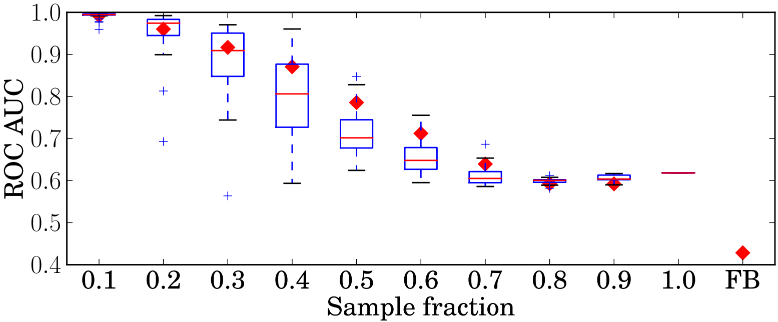
Base method: LoOP [Kriegel et al., 2009]



ROC AUC for ensemble members of the subsampling ensemble for different sample sizes (boxes), the base method (sample size=1.0), and ensembles (diamonds)—on top of subsamples and feature bags (FB)—on dataset WBC.

“sat.-image 2” – Different Sample Sizes

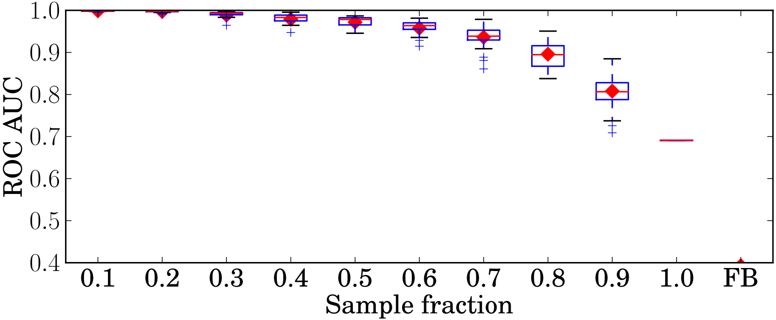
Base method: LDOF [Zhang et al., 2009]



ROC AUC for ensemble members of the subsampling ensemble for different sample sizes (boxes), the base method (sample size=1.0), and ensembles (diamonds)—on top of subsamples and feature bags (FB)—on dataset Satimage-2.

“sat.-image 2” – Different Sample Sizes

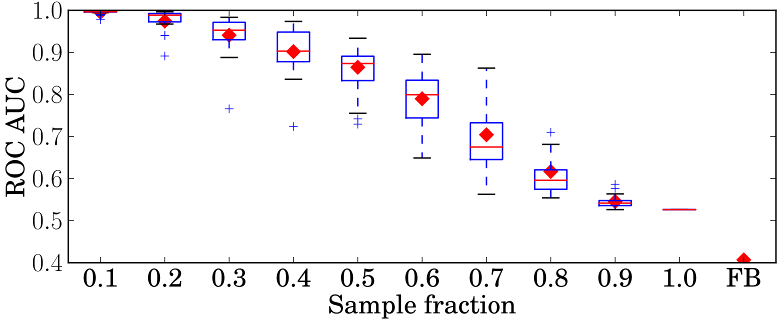
Base method: LOF [Breunig et al., 2000]



ROC AUC for ensemble members of the subsampling ensemble for different sample sizes (boxes), the base method (sample size=1.0), and ensembles (diamonds)—on top of subsamples and feature bags (FB)—on dataset Satimage-2.

“sat.-image 2” – Different Sample Sizes

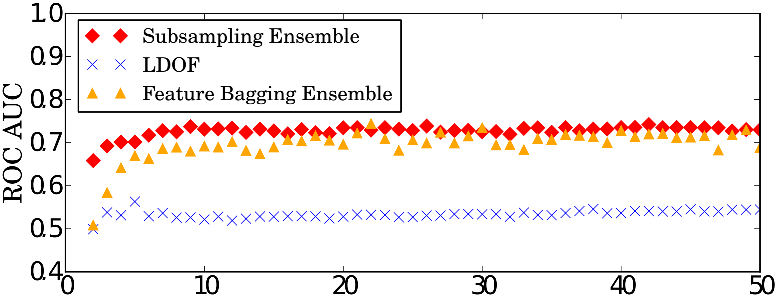
Base method: LoOP [Kriegel et al., 2009]



ROC AUC for ensemble members of the subsampling ensemble for different sample sizes (boxes), the base method (sample size=1.0), and ensembles (diamonds)—on top of subsamples and feature bags (FB)—on dataset Satimage-2.

"waveform" – Different Parameters for Base Method

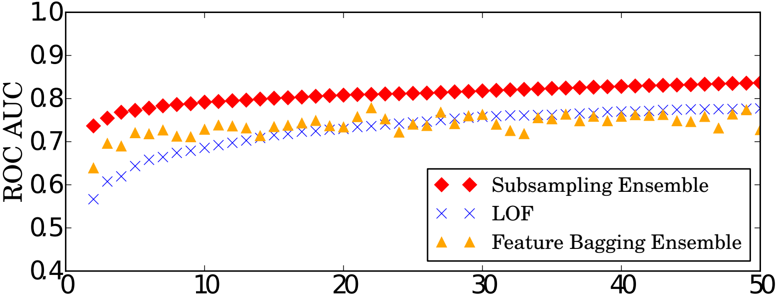
Base method: LDOF [Zhang et al., 2009]



ROC AUC for base methods and corresponding ensembles varying k on dataset waveform.

“waveform” – Different Parameters for Base Method

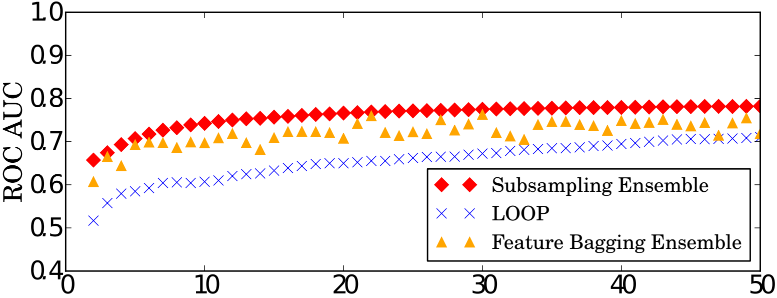
Base method: LOF [Breunig et al., 2000]



ROC AUC for base methods and corresponding ensembles varying k on dataset waveform.

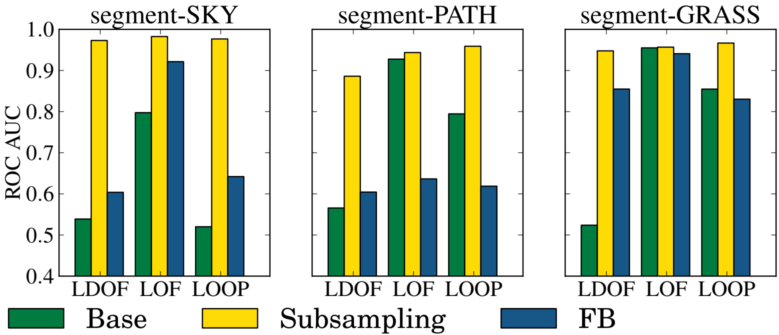
"waveform" – Different Parameters for Base Method

Base method: LoOP [Kriegel et al., 2009]



ROC AUC for base methods and corresponding ensembles varying k on dataset waveform.

All Base-Methods, Several Datasets (Variants of "segment")



ROC AUC for all methods, $k = 20$, on different datasets (variants of segment).

Outline

Summary

- theoretical reasoning and empirical demonstration that it is possible to construct ensemble members for outlier detection which perform individually already better than the base method
- combining these outlier detectors into an ensemble renders this performance gain more robust and can improve the performance even further
- using small sample sizes leads to considerable speed-up compared to a standard ensemble and, using small ensemble sizes, even compared to the base method
- the proposed principle is fundamental and flexible:
  - does not rely on specific data types
  - can be combined with various conventional outlier detection techniques

Future Work

- canonical competitor was feature bagging but actually, feature bagging and subsampling are not strictly competitors
- feature bagging: preference for high dimensional data
- subsampling: preference for large data sets
- both, feature bagging and subsampling, could be applied simultaneously (potential future work)
- both could be applied in combination with other diversity methods (different models, different parameters)
- more future work: study the diversity of ensemble members that is actually achieved by subsampling

Thank you for your attention!

References I

E. Achtert, H.-P. Kriegel, E. Schubert, and A. Zimek. Interactive data mining with 3D-Parallel-Coordinate-Trees. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), New York City, NY*, pages 1009–1012, 2013.

C. C. Aggarwal. Outlier ensembles [position paper]. *ACM SIGKDD Explorations*, 14(2):49–58, 2012.

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

M. M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, pages 93–104, 2000. doi: 10.1145/342009.335388.

A. Frank and A. Asuncion. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2010.

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*, pages 212–221, 2006. doi: 10.1109/ICDM.2006.43.

References II

F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

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

H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. LoOP: local outlier probabilities. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China*, pages 1649–1652, 2009. doi: 10.1145/1645953.1646195.

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*, pages 13–24, 2011.

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*, pages 157–166, 2005. doi: 10.1145/1081870.1081891.

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*, pages 368–383, 2010. doi: 10.1007/978-3-642-12026-8_29.

References III

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*, pages 1047–1058, 2012.

T. Soler and M. Chin. On transformation of covariance matrices between local Cartesian coordinate systems and commutative diagrams. In *ASP-ACSM Convention*, pages 393–406, 1985.

K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand*, pages 813–822, 2009. doi: 10.1007/978-3-642-01307-2_84.