On the Evaluation of Outlier Detection: Measures, Datasets, and an Empirical Study Continued

Guilherme O. Campos[1] Arthur Zimek[2] Jörg Sander[3]
Ricardo J. G. B. Campello[1] Barbora Micenková[4] *Erich Schubert*[5,7]
Ira Assent[4] Michael E. Houle[6]

[1]University of São Paulo
[2]University of Southern Denmark
[3]University of Alberta
[4]Aarhus University
[5]Ludwig-Maximilians-Universität München
[6]National Institute of Informatics
[7]Ruprecht-Karls-Universität Heidelberg

Lernen. Wissen. Daten. Analysen.
September 12–14, 2016, Potsdam, Deutschland

On the Evaluation of Unsupervised Outlier Detection

University of São Paulo
Brazil

SDU
SYDDANSK UNIVERSITET

AARHUS UNIVERSITY

UNIVERSITY OF ALBERTA
FACULTY OF SCIENCE

NII
大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

LMU
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. "On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study". In: *Data Mining and Knowledge Discovery* 30 (4 2016), pp. 891–927. DOI: `10.1007/s10618-015-0444-8`

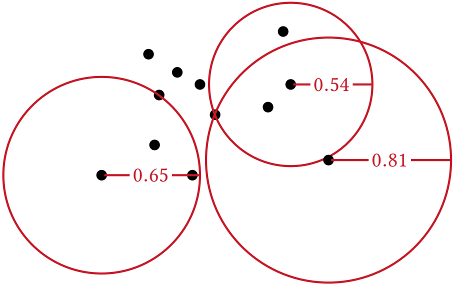
Online repository with complete material (methods, datasets, results, analysis):
`http://www.dbs.ifi.lmu.de/research/outlier-evaluation/`

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".*
>
> *[Haw80]*

Simple model example:
take the kNN distance of a point as its outlier score [RRS00]

Advanced model example:
compare the densities of neighbors (e.g. LOF [Bre+00])



Motivation

- many new outlier detection methods developed every year
  - many methods are very similar
- some studies about efficiency [Ora+10; KSZ16]
- specializations for different areas [CBK09; ZSK12; SZK14b; ATK15; SWZ15]

- evaluation of effectiveness remains notoriously challenging
  - characterisation of outlierness differs from method to method
  - lack of commonly agreed upon benchmark data
  - measure of success? (most commonly: ROC)

Outline

Selected Methods

We focus on methods based on the k nearest neighbors (same parameter k):

- kNN [RRS00], kNN-weight [AP05]
- LOF [Bre+00], SimplifiedLOF [SZK14b], COF [Tan+02], INFLO [Jin+06], LoOP [Kri+09]
- LDOF [ZHJ09], LDF [LLP07], KDEOS [SZK14a]
- ODIN [HKF04] (related to low hubness outlierness [RNI14])
- FastABOD [KSZ08] (ABOD variant using the kNN only)

The most popular classic, but also many recent methods.
Global and local methods (as defined in [SZK14b]).
All methods are implemented in the ELKI framework [Sch+15].

Additionally included in next release:

- LIC [YSW09], VoV [HS03], DWOF [MMG13], IDOS [vHZ15]

Evaluation Measures for Ranking Methods

- Precision@n (with $n = |O|$):

$$P@n = \frac{|\{o \in O \mid \mathsf{rank}(o) \leq n\}|}{n}$$

- Average Precision:

$$\mathsf{AP} = \frac{1}{|O|} \sum_{o \in O} P@\,\mathsf{rank}(o)$$

- Area under the ROC curve (ROC AUC or AUROC):

$$\text{ROC AUC} := \underset{o \in O, i \in I}{mean} \begin{cases} 1 & \text{if } \mathsf{score}(o) > \mathsf{score}(i) \\ \frac{1}{2} & \text{if } \mathsf{score}(o) = \mathsf{score}(i) \\ 0 & \text{if } \mathsf{score}(o) < \mathsf{score}(i) \end{cases}$$

- Maximum F1-Measure (newly added):

$$\text{Maximum-F1} := \underset{\mathsf{score}}{max}\ \mathsf{F1}(\text{Precision}(\mathsf{score}), \text{Recall}(\mathsf{score}))$$

- $+$ adjusted for chance versions of each.

$$\text{Adjusted Index} = \frac{\text{Index} - \text{Expected Index}}{\text{Maximum Index} - \text{Expected Index}}$$

Ground Truth for Outlier Detection?

- every author uses other data sets – no common benchmark data
- classification data (e.g. UCI) usually *not* usable:
  classes are too frequent, and expected to be similar (i.e. no outlier class)
- papers on outlier detection prepare some datasets ad hoc
- preparation involves decisions that are often not sufficiently documented (e.g. normalization, transformation)
- common problematic assumption: downsampling a class yields outliers

We produce data sets similar to existing papers,
but document preprocessing and make the resulting data sets available.

We are also interested in the question:
are these data sets suitable for outlier detection?

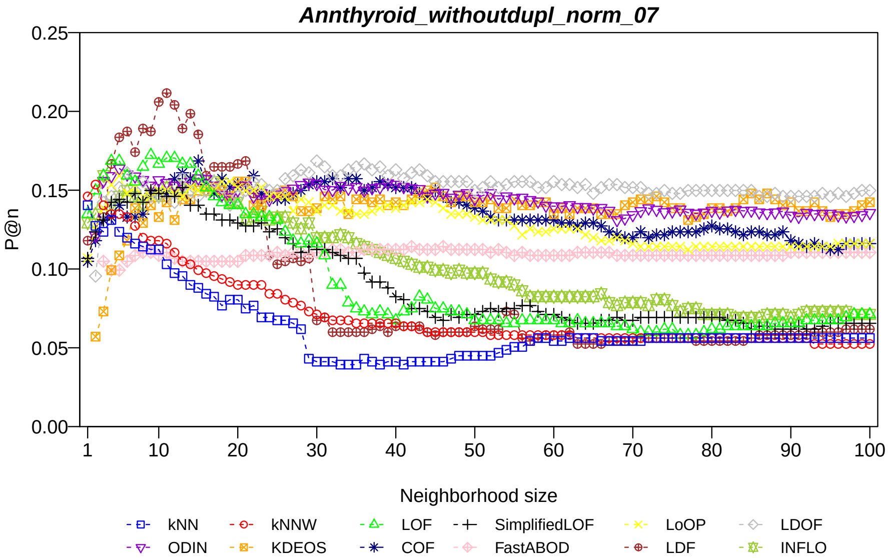
Datasets Used in the Literature

Dataset	Preprocessing	N	$\|O\|$	Attrib. num	Attrib. cat	Version used by
ALOI	50000 images, 27 attr.	50000	1508	27		[Kri+11], [Sch+12]
	24000 images, 27648 attr.					[dCH12]
Glass	Class 6 (*out.*) vs. others (*in.*)	214	9	7		[KMB12]
Ionosphere	Class 'b' (*out.*) vs. class 'g' (*in.*)	351	126	32		[KMB12]
KDDCup99	U2R (*out.*) vs. Normal (*in.*)	60632	246	38	3	[NG10], [NAG10], [Kri+11], [Sch+12]
Lympho-graphy	Classes 1 and 4 (*out.*) vs. others (*in.*)	148	6	3	16	[LK05], [NAG10], [Zim+13]
Pen-Digits	Downs. class '4' to 20 objects (*out.*)	9868	20	16		[Kri+11] [Sch+12]
	Downs. class '0' to 10% (*out.*)					[KMB12]
Shuttle	Classes 2, 3, 5, 6, 7 (*out.*) vs. class 1 (*in.*)					[LK05], [AZL06], [NAG10]
	Downs. 2, 3, 5, 6, 7 (*out.*) vs. others (*in.*)					[GT06]
	Class 2 (*out.*) vs. downs. others to 1000 (*in.*)	1013	13	9		[ZHJ09]
Waveform	Downs. class '0' to 100 objects (*out.*)	3443	100	21		[Zim+13]
WBC	'*malignant*' (*out.*) vs. '*benign*' (*in.*)					[GT06]
	Downs. class '*malignant*' to 10 obj. (*out.*)	454	10	9		[Kri+11], [Sch+12], [Zim+13]
WDBC	Downs. class '*malignant*' to 10 obj. (*out.*)	367	10	30		[ZHJ09]
	'*malignant*' (*out.*) vs. '*benign*' (*in.*)					[KMB12]
WPBC	Class 'R' (*out.*) vs. class 'N' (*in.*)	198	47	33		[KMB12]

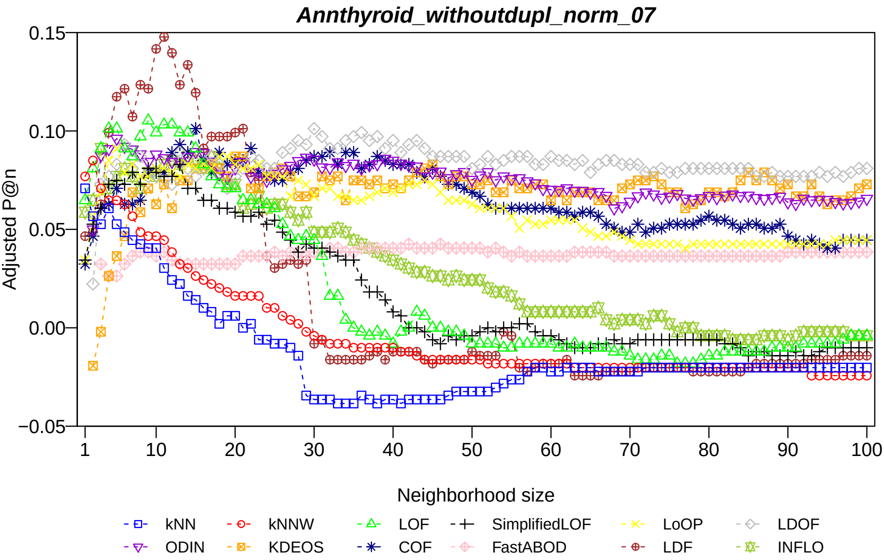
Semantically Meaningful Outlier Datasets

Dataset	Semantics	N	$\|O\|$	Attributes num.	Attributes binary
Annthyroid	2 types of hypothyroidism vs. healthy	7200	534	21	
Arrhythmia	12 types of cardiac arrhythmia vs. healthy	450	206	259	
Cardiotocography	pathologic, suspect vs. healthy	2126	471	21	
HeartDisease	heart problems vs. healthy	270	120	13	
Hepatitis	survival vs. fatal	80	13	19	
InternetAds	ads vs. other images	3264	454		1555
PageBlocks	non-text vs. text	5473	560	10	
Parkinson	healthy vs. Parkinson	195	147	22	
Pima	diabetes vs. healthy	768	268	8	
SpamBase	non-spam vs. spam	4601	1813	57	
Stamps	genuine vs. forged	340	31	9	
Wilt	diseased trees vs. other	4839	261	5	

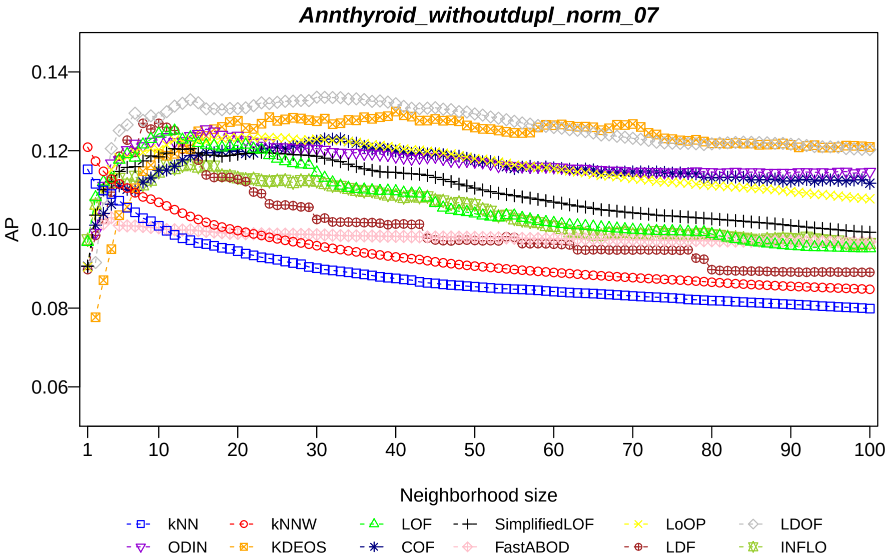
Example: Annthyroid



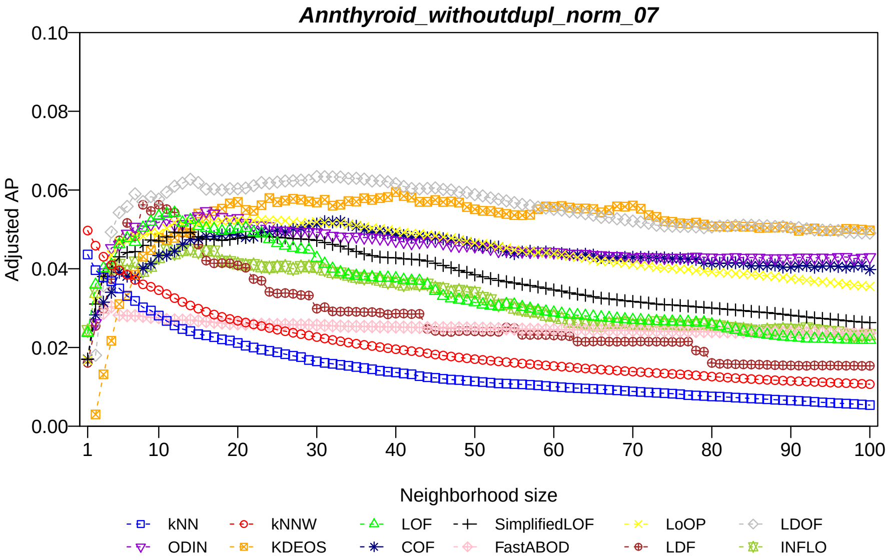
Example: Annthyroid



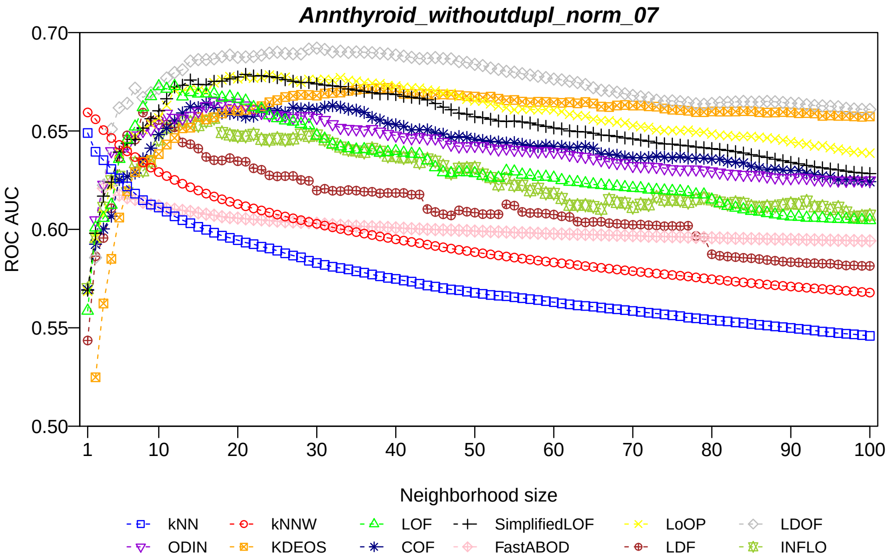
Example: Annthyroid



Example: Annthyroid



Example: Annthyroid



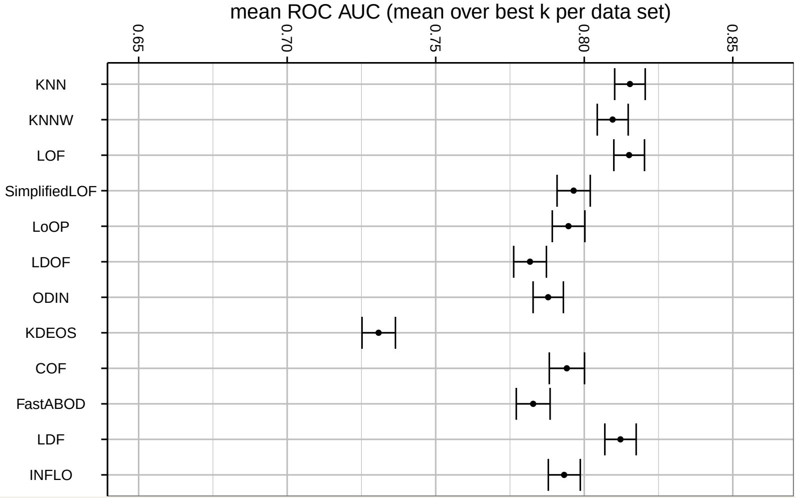
Observations

All results are available in the web repository:
`http://www.dbs.ifi.lmu.de/research/outlier-evaluation/`

- performance trends differ across algorithms, datasets, parameters, and evaluation methods
- ROC AUC less sensitive to number of true outliers
- ROC AUC scores across the datasets typically reasonably high
- $P@n$ scores considerably lower for datasets with smaller proportions of outliers
- AP resembles ROC AUC, assessing the ranks of all outliers, but tends to be lower with stronger imbalance
- $P@n$ can discriminate between methods that perform more or less equally well in terms of ROC AUC [DG06]

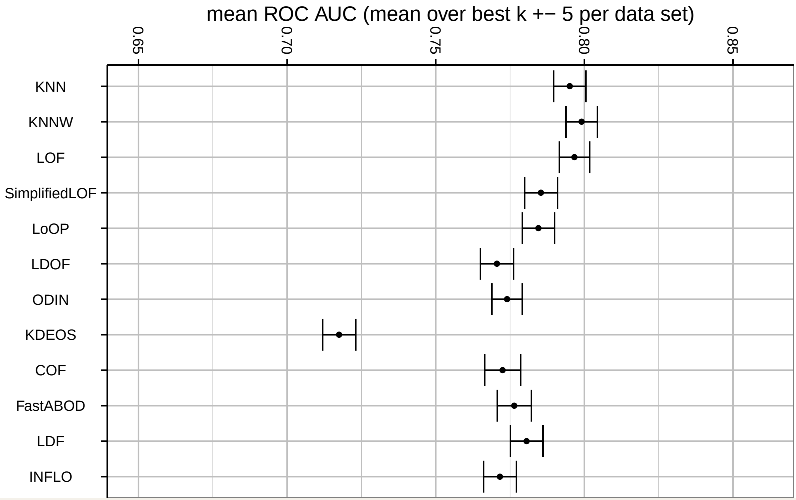
Average ROC AUC per Method

aggregated over all datasets
(without duplicates, normalized)



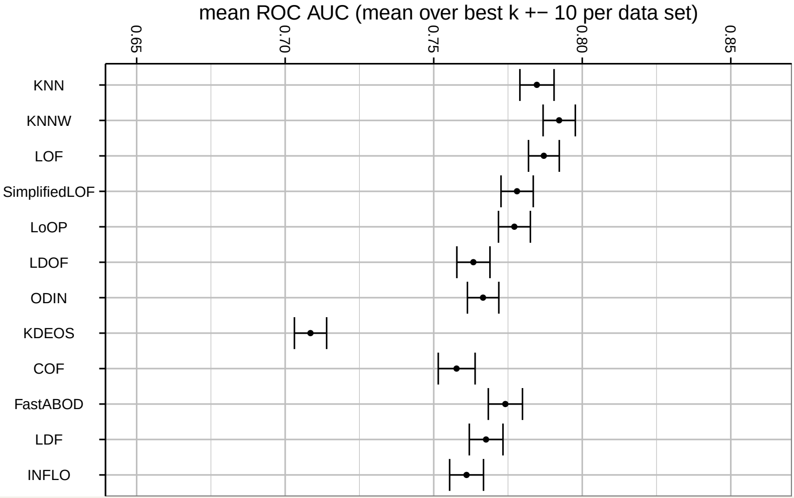
Average ROC AUC per Method

aggregated over all datasets
(without duplicates, normalized)



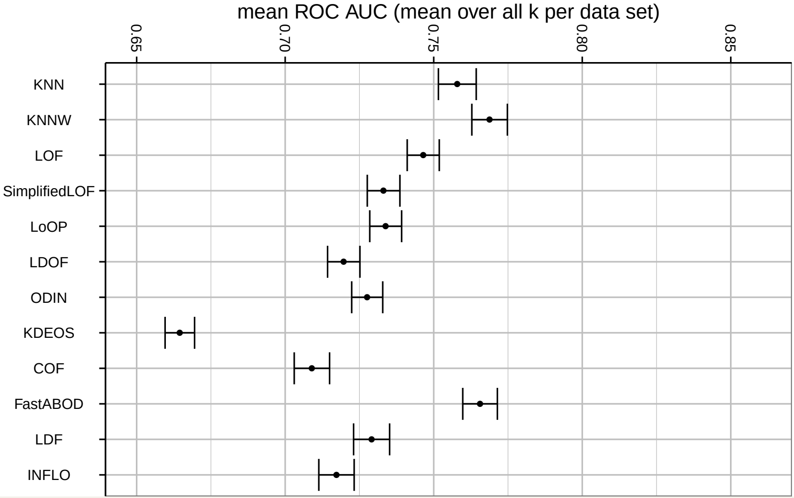
Average ROC AUC per Method

aggregated over all datasets
(without duplicates, normalized)



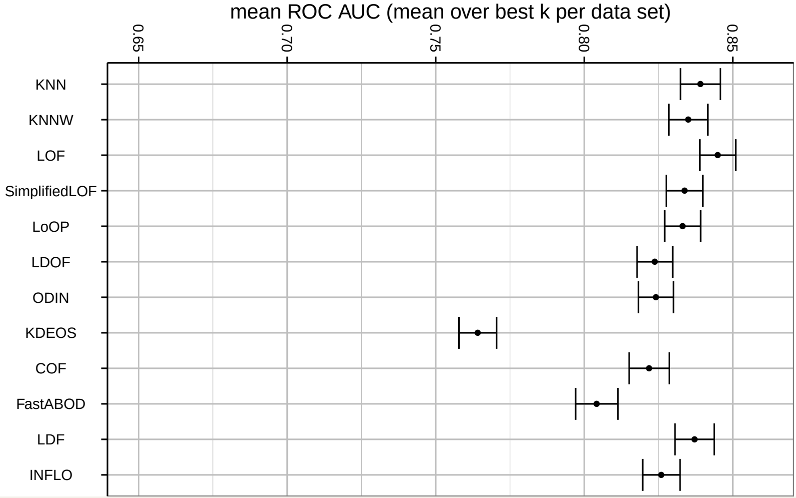
Average ROC AUC per Method

aggregated over all datasets
(without duplicates, normalized)



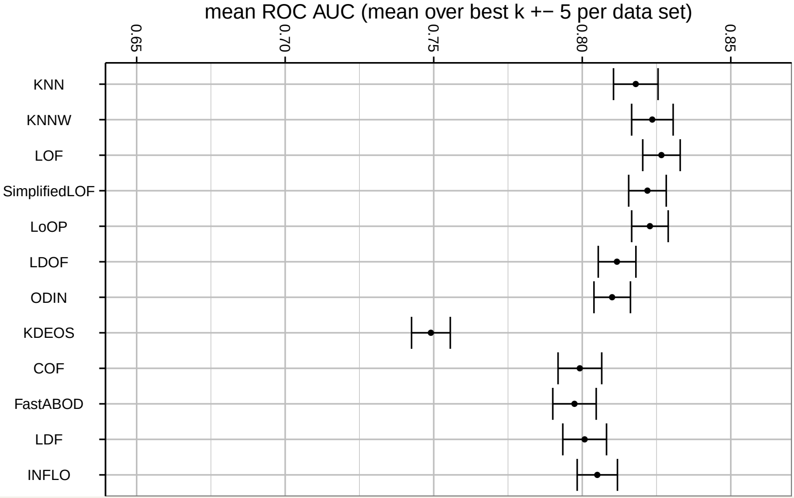
Average ROC AUC per Method

aggregated over all datasets
(without duplicates, normalized, at most 5% outliers)



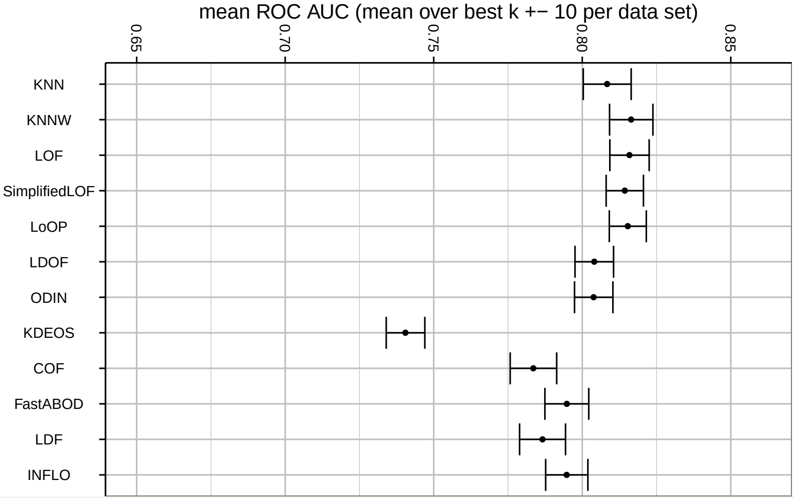
Average ROC AUC per Method

aggregated over all datasets
(without duplicates, normalized, at most 5% outliers)



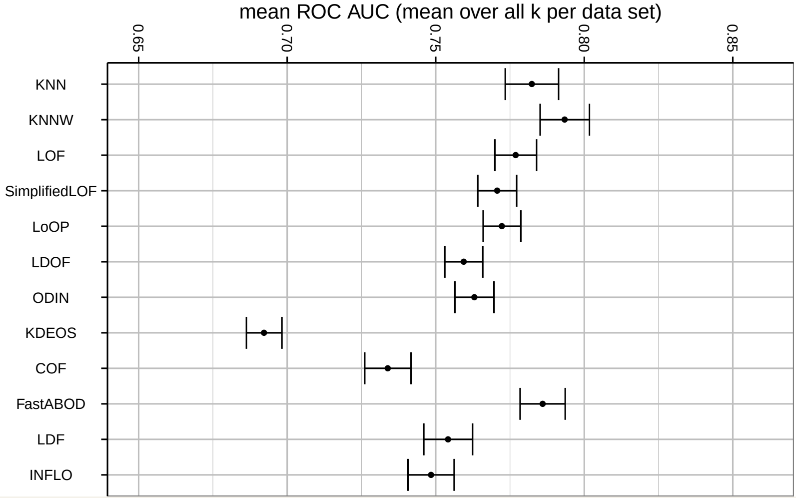
Average ROC AUC per Method

aggregated over all datasets
(without duplicates, normalized, at most 5% outliers)



Average ROC AUC per Method

aggregated over all datasets
(without duplicates, normalized, at most 5% outliers)



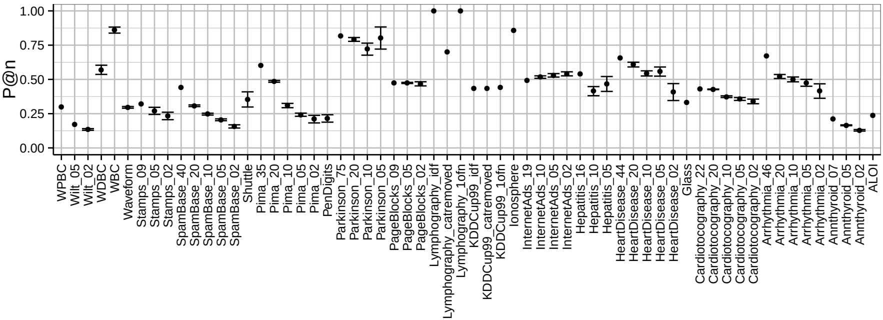
Statistical Test

Nemenyi post-hoc test (normalized datasets without duplicates, ALOI and KDDCup99 removed, best achieved quality in terms of ROC AUC chosen for each dataset independently; results for those datasets with multiple subsampled variants were grouped by averaging the best results over all variants for each method):
column method is better/worse than row method at 90% ('+'/'−') and 95% ('++'/'−−') confidence levels.

	kNN	kNNW	LOF	SimplifiedLOF	LoOP	LDOF	ODIN	KDEOS	COF	FastABOD	LDF	INFLO
kNN	=							−−				
kNNW		=						−−				
LOF			=			−	−−	−−		−−		
SimplifiedLOF				=				−−				
LoOP					=			−−				
LDOF			+			=						
ODIN			++				=					
KDEOS	++	++	++	++	++			=	++		++	++
COF								−−	=			
FastABOD			++							=	+	
LDF								−−		−	=	
INFLO								−−				=

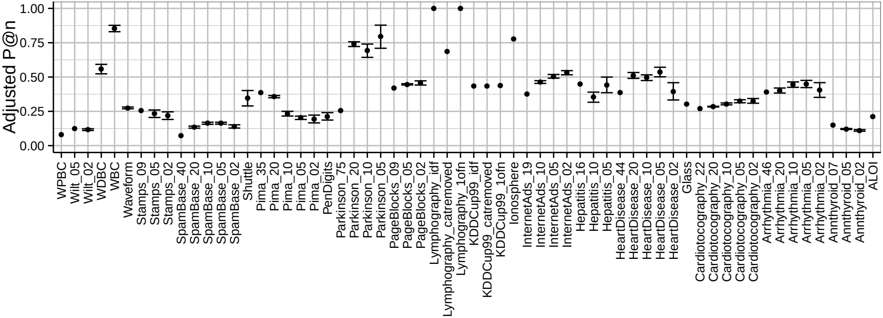
Best Results per Dataset

Average best performance of all methods, per dataset (without duplicates, normalized). Best results chosen by ROC AUC performance.



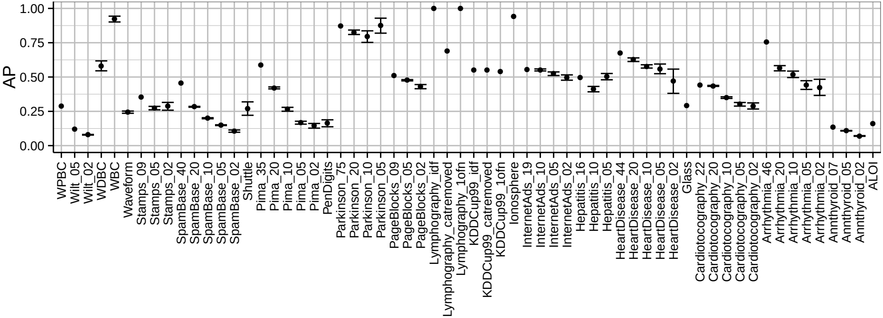
Best Results per Dataset

Average best performance of all methods, per dataset (without duplicates, normalized). Best results chosen by ROC AUC performance.



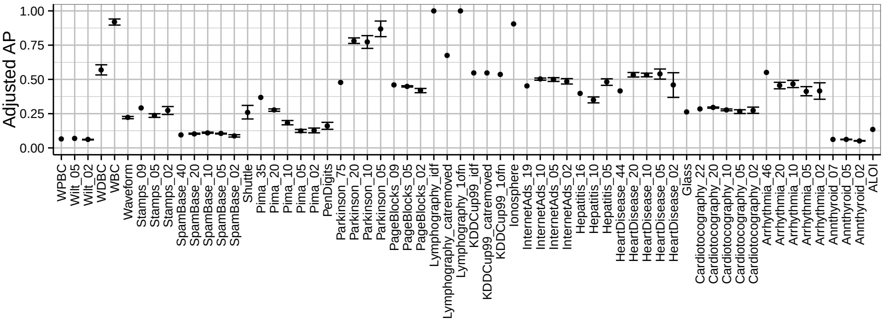
Best Results per Dataset

Average best performance of all methods, per dataset (without duplicates, normalized). Best results chosen by ROC AUC performance.



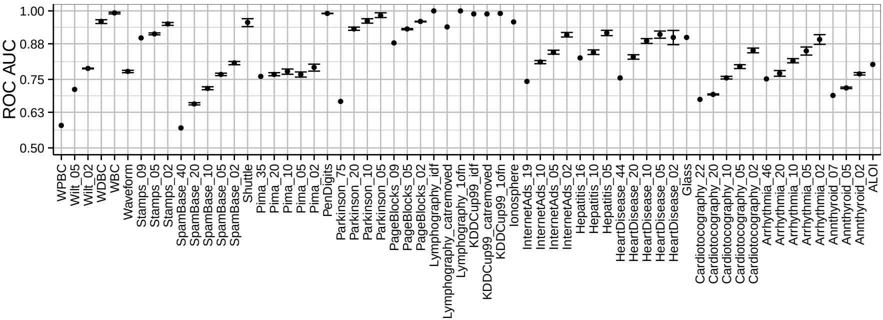
Best Results per Dataset

Average best performance of all methods, per dataset (without duplicates, normalized). Best results chosen by ROC AUC performance.

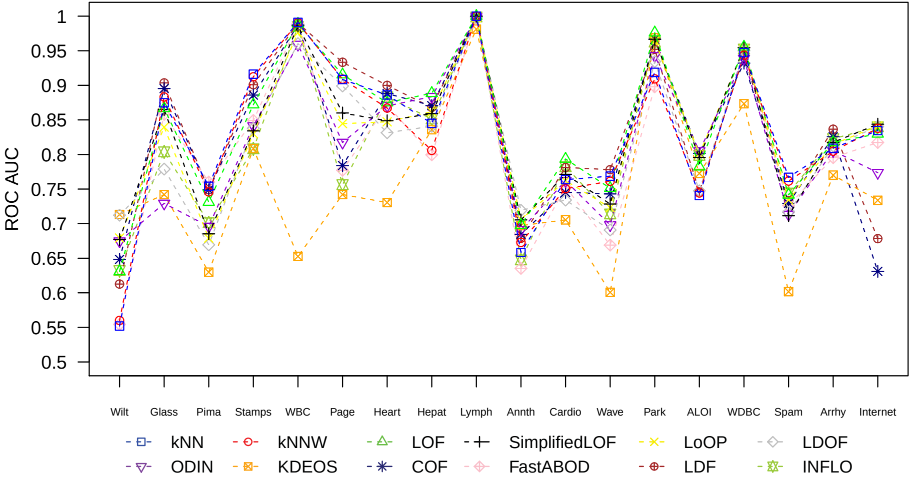


Best Results per Dataset

Average best performance of all methods, per dataset (without duplicates, normalized). Best results chosen by ROC AUC performance.



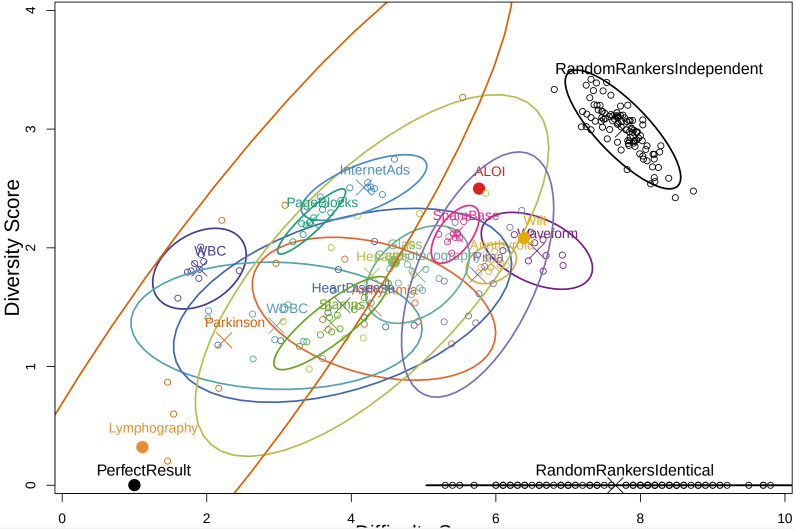
Difficulty and Dimensionality



ROC AUC scores, for each method using the best k, on the datasets with 3 to 5% of outliers, averaged over the different dataset variants where available.
The datasets are arranged on the x-axis of the plot from left to right in order of increasing dimensionality.

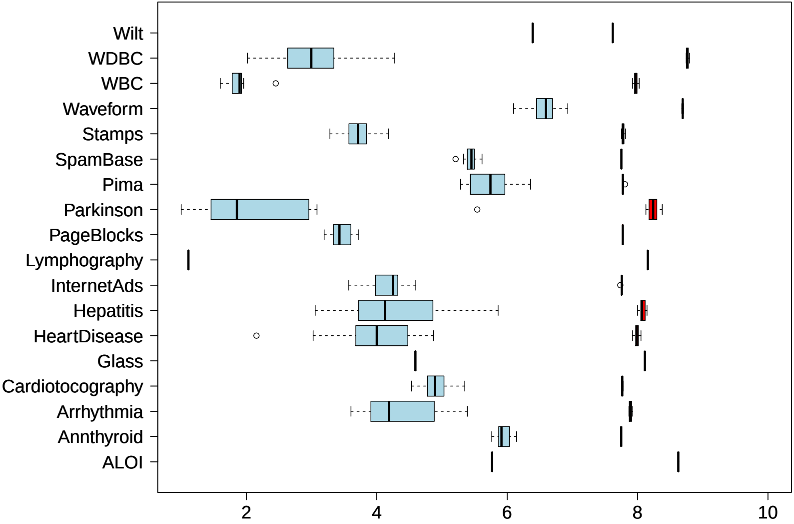
Suitability of Ground Truth Outlier Labels

Difficulty for given labels vs. random labels



Suitability of Ground Truth Outlier Labels

Difficulty for given labels vs. random labels



Conclusions

In the publication

G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. "On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study". In: *Data Mining and Knowledge Discovery* 30 (4 2016), pp. 891–927. DOI: 10.1007/s10618-015-0444-8

- we discussed evaluation measures for outlier rankings: $P@n$, AP, and ROC (AUC)
- we proposed adjustment for chance for $P@n$ and for AP
- we discussed preprocessing issues for the preparation of outlier datasets with annotatded ground truth and provide 23 datasets in about 1000 variants

Conclusions

- we tested 12 outlier detection methods on these datasets with a range of choices for the neighborhood parameter $k \in [1, \ldots, 100]$
- we aggregate and analyse the resulting $> 1,3$ million experiments and
  - summarize the effectiveness of the 12 methods
  - study the suitability of the datasets for evaluation of outlier detection
- we offer all results and analyses together with source code online:
  `http://www.dbs.ifi.lmu.de/research/outlier-evaluation/`
- experiments can be easily repeated and extended for other methods and other datasets

Thank you for your attention!

And many thanks to my collaborators:

- Guilherme O. Campos
- Arthur Zimek
- Jörg Sander
- Ricardo J. G. B. Campello
- Barbora Micenková
- Ira Assent
- Mike E. Houle

References I

[AP05] F. Angiulli and C. Pizzuti. "Outlier mining in large high-dimensional data sets". In: *IEEE Transactions on Knowledge and Data Engineering* 17.2 (2005), pp. 203–215. DOI: 10.1109/TKDE.2005.31.

[ATK15] L. Akoglu, H. Tong, and D. Koutra. "Graph-based Anomaly Detection and Description: A Survey". In: *Data Mining and Knowledge Discovery* 29.3 (2015), pp. 626–688. DOI: 10.1007/s10618-014-0365-y.

[AZL06] N. Abe, B. Zadrozny, and J. Langford. "Outlier Detection by Active Learning". In: *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA.* 2006, pp. 504–509. DOI: 10.1145/1150402.1150459.

[Bre+00] 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.* 2000, pp. 93–104. DOI: 10.1145/342009.335388.

[Cam+16] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. "On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study". In: *Data Mining and Knowledge Discovery* 30 (4 2016), pp. 891–927. DOI: 10.1007/s10618-015-0444-8.

[CBK09] V. Chandola, A. Banerjee, and V. Kumar. "Anomaly Detection: A Survey". In: *ACM Computing Surveys* 41.3 (2009), Article 15, 1–58. DOI: 10.1145/1541880.1541882.

References II

[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. DOI: 10.1007/s10115-011-0430-4.

[DG06] J. Davis and M. Goadrich. "The Relationship Between Precision-Recall and ROC Curves". In: *Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA.* 2006, pp. 233–240. DOI: 10.1145/1143844.1143874.

[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. DOI: 10.1109/ICDM.2006.43.

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

[HKF04] V. Hautamäki, I. Kärkkäinen, and P. Fränti. "Outlier Detection Using k-Nearest Neighbor Graph". In: *Proceedings of the 17th International Conference on Pattern Recognition (ICPR), Cambridge, England, UK.* 2004, pp. 430–433. DOI: 10.1109/ICPR.2004.1334558.

[HS03] T. Hu and S. Y. Sung. "Detecting pattern-based outliers". In: *Pattern Recognition Letters* 24.16 (2003), pp. 3059–3068. DOI: 10.1016/S0167-8655(03)00165-X.

References III

[Jin+06] W. Jin, A. K. H. Tung, J. Han, and W. Wang. "Ranking Outliers Using Symmetric Neighborhood Relationship". In: *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore.* 2006, pp. 577–593. DOI: 10.1007/11731139_68.

[KMB12] F. Keller, E. Müller, and K. Böhm. "HiCS: High Contrast subspaces for Density-Based Outlier Ranking". In: *Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC.* 2012, pp. 1037–1048. DOI: 10.1109/ICDE.2012.88.

[Kri+09] 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.* 2009, pp. 1649–1652. DOI: 10.1145/1645953.1646195.

[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. DOI: 10.1137/1.9781611972818.2.

[KSZ08] H.-P. Kriegel, M. Schubert, and A. Zimek. "Angle-Based Outlier Detection in High-dimensional Data". In: *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV.* 2008, pp. 444–452. DOI: 10.1145/1401890.1401946.

[KSZ16] H.-P. Kriegel, E. Schubert, and A. Zimek. "The (Black) Art of Runtime Evaluation: Are We Comparing Algorithms or Implementations?". submitted. 2016.

References IV

[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. DOI: 10.1145/1081870.1081891.

[LLP07] L. J. Latecki, A. Lazarevic, and D. Pokrajac. "Outlier Detection with Kernel Density Functions". In: *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM), Leipzig, Germany*. 2007, pp. 61–75. DOI: 10.1007/978-3-540-73499-4_6.

[MMG13] R. Momtaz, N. Mohssen, and M. A. Gowayyed. "DWOF: A Robust Density-Based Outlier Detection Approach". In: *Pattern Recognition and Image Analysis. Proceedings of the 6th Iberian Conference, IbPRIA 2013, Funchal, Madeira, Portugal*. 2013, pp. 517–525. DOI: 10.1007/978-3-642-38628-2_61.

[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. DOI: 10.1007/978-3-642-12026-8_29.

[NG10] H. V. Nguyen and V. Gopalkrishnan. "Feature Extraction for Outlier Detection in High-Dimensional Spaces". In: *Journal of Machine Learning Research* Proceedings Track 10 (2010), pp. 66–75.

References V

[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.

[RNI14] M. Radovanović, A. Nanopoulos, and M. Ivanović. "Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection". In: *IEEE Transactions on Knowledge and Data Engineering* (2014). DOI: 10.1109/TKDE.2014.2365790.

[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. DOI: 10.1145/342009.335437.

[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. DOI: 10.1137/1.9781611972825.90.

[Sch+15] E. Schubert, A. Koos, T. Emrich, A. Züfle, K. A. Schmid, and A. Zimek. "A Framework for Clustering Uncertain Data". In: *Proceedings of the VLDB Endowment* 8.12 (2015), pp. 1976–1979. DOI: 10.14778/2824032.2824115.

References VI

[SWZ15] E. Schubert, M. Weiler, and A. Zimek. "Outlier Detection and Trend Detection: Two Sides of the Same Coin". In: *1st International Workshop on Event Analytics using Social Media Data at the 15th IEEE International Conference on Data Mining (ICDM), Atlantic City, NJ*. 2015. DOI: 10.1109/ICDMW.2015.79.

[SZK14a] 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. DOI: 10.1137/1.9781611973440.63.

[SZK14b] E. Schubert, A. Zimek, and H.-P. Kriegel. "Local Outlier Detection Reconsidered: a Generalized View on Locality with Applications to Spatial, Video, and Network Outlier Detection". In: *Data Mining and Knowledge Discovery* 28.1 (2014), pp. 190–237. DOI: 10.1007/s10618-012-0300-z.

[Tan+02] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung. "Enhancing Effectiveness of Outlier Detections for Low Density Patterns". In: *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan*. 2002, pp. 535–548. DOI: 10.1007/3-540-47887-6_53.

[vHZ15] J. von Brünken, M. E. Houle, and A. Zimek. *Intrinsic Dimensional Outlier Detection in High-Dimensional Data*. Tech. rep. NII-2015-003E. National Institute of Informatics, 2015.

References VII

[YSW09] B. Yu, M. Song, and L. Wang. "Local Isolation Coefficient-Based Outlier Mining Algorithm". In: *Information Technology and Computer Science*. Vol. 2. 2009, pp. 448–451. DOI: 10.1109/ITCS.2009.230.

[ZHJ09] 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*. 2009, pp. 813–822. DOI: 10.1007/978-3-642-01307-2_84.

[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. DOI: 10.1145/2487575.2487676.

[ZSK12] A. Zimek, E. Schubert, and H.-P. Kriegel. "A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data". In: *Statistical Analysis and Data Mining* 5.5 (2012), pp. 363–387. DOI: 10.1002/sam.11161.