On the Evaluation of Unsupervised Outlier Detection

Arthur Zimek

Ludwig-Maximilians-Universität München

Talk at TU Vienna
18.05.2016

Campos, Zimek, Sander, Campello, Micenková, Schubert, Assent, and Houle: *On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. Data Mining and Knowledge Discovery, 2016.*

http://doi.org/10.1007/s10618-015-0444-8

Online repository with complete material (methods, datasets, results, analysis):

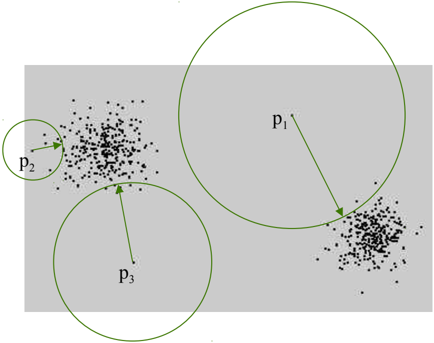
http://www.dbs.ifi.lmu.de/research/outlier-evaluation/

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".

[Hawkins, 1980]

Simple model: take the kNN distance of a point as its outlier score [Ramaswamy et al., 2000]



Motivation

- many new outlier detection methods developed every year
- some studies about efficiency [Orair et al., 2010, Kriegel et al., 2016]
- specializations for different areas [Chandola et al., 2009, Zimek et al., 2012, Schubert et al., 2014a, Akoglu et al., 2015]
- 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

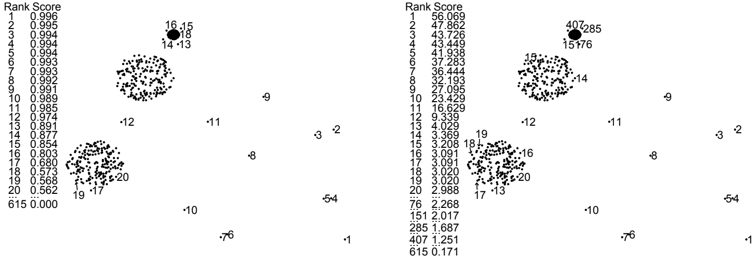
- kNN [Ramaswamy et al., 2000], kNN-weight [Angiulli and Pizzuti, 2005]
- ODIN [Hautamäki et al., 2004] (related to low hubness outlierness [Radovanović et al., 2014])
- LOF [Breunig et al., 2000]
- SimplifiedLOF [Schubert et al., 2014a], COF [Tang et al., 2002], INFLO [Jin et al., 2006], LoOP [Kriegel et al., 2009]
- LDOF [Zhang et al., 2009], LDF [Latecki et al., 2007], KDEOS [Schubert et al., 2014b]
- FastABOD [Kriegel et al., 2008]

Discussion

- all these methods have a common parameter, the neighborhood size k
- this family of kNN-based methods is popular and contains both classic and recent methods
- nevertheless, the parameter k has different interpretations and impact among the selected methods
- the selected methods comprise both 'global' and 'local' methods [Schubert et al., 2014a]
- included variants of LOF vary different components of the typical local outlier model: notion of neighborhood, distance, density estimates, model comparison
- first study to compare all these methods
- all methods are implemented in a common framework ELKI [Schubert et al., 2015]

Outline

- methods return a full ranking of database objects
- user interested in the top-ranked objects



examples taken from Campello et al. [2015]

Precision at n

- If the number of outlier candidates n is specified, the simplest measure of performance is the precision at n ($P@n$), i.e., the proportion of correct results in the top n ranks [Craswell, 2009a].

Precision at n ($P@n$)

Given a database $\mathcal{D}$ of size N, outliers $O \subset \mathcal{D}$ and inliers $I \subseteq \mathcal{D}$ ($\mathcal{D} = O \cup I$), we have

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

Precision at n — Properties

- how to fairly choose the parameter n of $P@n$?
- $n = |O|$ yields the popular R-Precision measure [Craswell, 2009b].
- $n = |O| \ll N$
  $\Rightarrow$ typically obtained values of $P@n$ can be deceptively low, and not very informative as such
- $n = |O|$ relatively large (of the same order as N)
  $\Rightarrow$ deceptively high values of $P@n$ can be obtained simply due to the relatively small number of inliers available

Precision at n — Adjustment for Chance

- general procedure due to Hubert and Arabie [1985]:

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

- maximum $P@n$: $|O|/n$ if $n > |O|$, and 1 otherwise
- expected value (random outlier ranking): $|O|/N$

Precision at n ($P@n$), adjusted for chance

If $n \leq |O|$:

$$\text{Adjusted } P@n = \frac{P@n - |O|/N}{1 - |O|/N}.$$

For larger n, the maximum $|O|/n$ must be used instead of 1.

Imbalance

- problem: imbalance between the numbers of inliers and outliers: $|I| \gg |O|$, $|I| \approx N$
- $P@n$ and Adjusted $P@n$ measures are easily interpreted
- but they are sensitive to the choice of n, particularly when n is small
- example:
  - dataset with 10 outliers and 1 million inliers
  - result with (quite high) ranks 11–20 for the true outliers
  - $P@10$ of 0
  - $P@20$ of 0.5
- Adjusted $P@n$ measure can be seen to suffer from a similar sensitivity with respect to the choice of n.

Average Precision

- even in good results, outliers typically do not form a larger fraction among the top $|O|$ ranks
- use measures that aggregate performance over a wide range of possible choices of n
- for example average precision [Zhang and Zhang, 2009] (popular in information retrieval)

Average Precision (AP)

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

Average Precision — Adjustment for Chance

- perfect ranking yields a maximum value of 1
- expected value of a random ranking is $|O|/N$

Average Precision (AP), adjusted for chance

$$\text{Adjusted }\ \mathrm{AP} = \frac{\mathrm{AP} - |O|/N}{1 - |O|/N}.$$

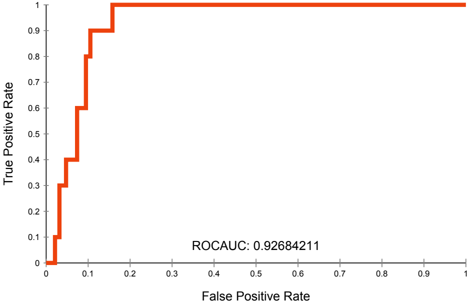
When Should we Adjust for Chance?

for $P@n$ and AP, adjustment for chance is:

- not necessary when the performance of two methods on the same dataset (that is, with the same proportion of outliers) is compared in relative terms
- helpful, if the measure is to be interpreted in absolute terms
- strictly necessary, if the performance is to be compared over different datasets with different proportions of outliers

Receiver Operating Characteristic (ROC)

- plots for all possible choices of n the true positive rate (the proportion of outliers correctly ranked among the top n) versus the false positive rate (the proportion of inliers ranked among the top n)



- based on rates, ROC inherently adjusts for the imbalance of class sizes typical of outlier detection tasks

Area under the ROC curve (ROC AUC)

- A ROC curve can be summarized by a single value known as ROC AUC, defined as the area under the ROC curve (AUC)
- $0 \leq$ ROC AUC ≤ 1
- interpretation: average of the recall at n, with n taken over the ranks of all inlier objects
- probabilistic interpretation [Hanley and McNeil, 1982]: probability of a pair (o, i), where o is some true outlier, and i is some inlier, being ordered correctly in the evaluated ranking:

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

Discussion

- all these evaluation measures are external: require annotated ground truth (outlier/inlier)
- useful for competitive evaluation of algorithms where the outliers in a dataset are actually known
- for evaluation in practical applications of methods we would need internal evaluation measures
- so far, only one paper in the literature describes an internal evaluation procedure, the work of Marques et al. [2015]

Outline

Outlier Detection Methods

Evaluation Measures

Datasets

Experiments

Conclusions

Ground Truth for Outlier Detection?

- no commonly agreed upon and frequently used benchmark data available
- UCI datasets etc.: ground truth by class labels — not readily usable for outlier evaluation
- papers on outlier detection prepare some datasets ad hoc or reuse some datasets that have been prepared ad hoc by others
- preparation involves decisions that are often not sufficiently documented
- we follow the common practice of downsampling some class in a classification dataset to produce an outlier class

Datasets Used in the Literature

Dataset	Preprocessing	N	$\lvert O \rvert$	Attributes num	Attributes cat	Version used by
ALOI	50000 images, 27 attr.	50000	1508	27		Kriegel et al. [2011], Schubert et al. [2012]
	24000 images, 27648 attr.					de Vries et al. [2012]
Glass	Class 6 (*out.*) vs. others (*in.*)	214	9	7		Keller et al. [2012]
Iono-sphere	Class 'b' (*out.*) vs. class 'g' (*in.*)	351	126	32		Keller et al. [2012]
KDDCup99	U2R (*out.*) vs. Normal (*in.*)	60632	246	38	3	Nguyen and Gopalkrishnan [2010], Nguyen et al. [2010], Kriegel et al. [2011], Schubert et al. [2012]
Lympho-graphy	Classes 1 and 4 (*out.*) vs. others (*in.*)	148	6	3	16	Lazarevic and Kumar [2005], Nguyen et al. [2010], Zimek et al. [2013]
Pen-Digits	Downsampling class '4' to 20 objects (*out.*)	9868	20	16		Kriegel et al. [2011] Schubert et al. [2012]
	Downsampling class '0' to 10% (*out.*)					Keller et al. [2012]
Shuttle	Classes 2, 3, 5, 6, 7 (*out.*) vs. class 1 (*in.*)					Lazarevic and Kumar [2005], Abe et al. [2006], Nguyen et al. [2010]
	Class 2 (*out.*) vs. downs. others to 1000 obj. (*in.*)	1013	13	9		Zhang et al. [2009]
	Downs. classes 2, 3, 5, 6, 7 (*out.*) vs. others (*in.*)					Gao and Tan [2006]
Wave-form	Downsampling class '0' to 100 objects (*out.*)	3443	100	21		Zimek et al. [2013]
WBC	'*malignant*' (*out.*) vs. '*benign*' (*in.*)					Gao and Tan [2006]
	Downs. class '*malignant*' to 10 objects (*out.*)	454	10	9		Kriegel et al. [2011], Schubert et al. [2012], Zimek et al. [2013]
WDBC	Downs. class '*malignant*' to 10 objects (*out.*)	367	10	30		Zhang et al. [2009]
	'*malignant*' (*out.*) vs. '*benign*' (*in.*)					Keller et al. [2012]
WPBC	Class 'R' (*out.*) vs. class 'N' (*in.*)	198	47	33		Keller et al. [2012]

Semantically Meaningful Outlier Datasets

Dataset	Semantics	N	$\lvert O \rvert$	Attributes	
				num.	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	

Dataset Preparation I

downsampling randomly downsample one class (as outlier class) while retaining all instances from other classes (as inlier class)

- great variation in the nature of outliers produced
- we repeat the downsampling ten times, resulting in ten different datasets
- we adopt different downsample rates where applicable (resulting in datasets with 20%, 10%, 5%, or 2% outliers)

duplicates duplicates can be problematic (e.g., density estimates) — for datasets containing duplicates, we generate two variants, one with the original duplicates, and one without duplicates

Dataset Preparation II

categorical attributes three variants of handling categorical attributes:

- categorical attributes are removed
- 1-of-n encoding: a categorical attribute with n possible values is mapped into n binary attributes (presence or absence of the corresponding categorical value)
- IDF: a categorical attribute is encoded as the *inverse document frequency*

$$IDF(t) = \ln(N/f_t)$$

(N: total number of instances, f_t: frequency of the attribute value t)

Dataset Preparation III

normalization Normalization of datasets is expected to have considerable impact on the results, but is rarely discussed in the literature. Two variants for each dataset that does not already have normalized attributes:

- unnormalized
- attribute-wise linear normalization to the range $[0, 1]$

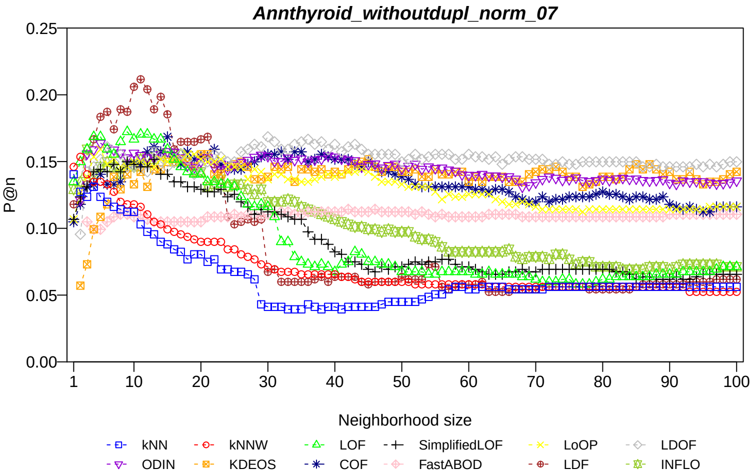
missing values If an attribute has fewer than 10% of instances with missing values, those instances are removed. Otherwise, the attribute itself is removed.

These procedures applied to most of the 23 datasets (some are taken from the literature unchanged, where they are available) results in about 1000 dataset variants.

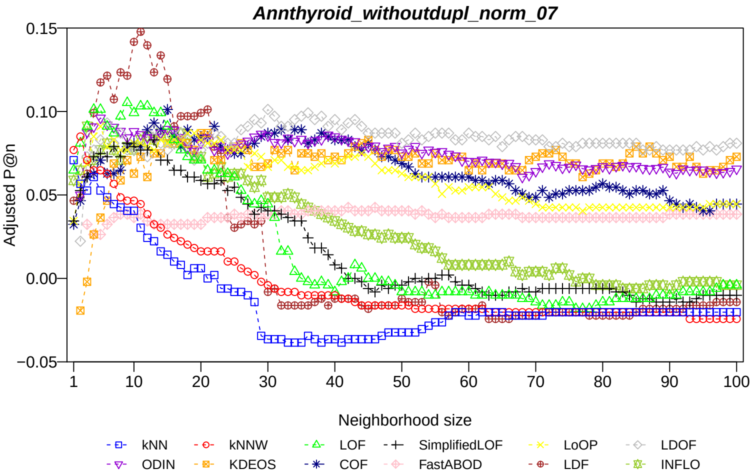
Outline

Outline

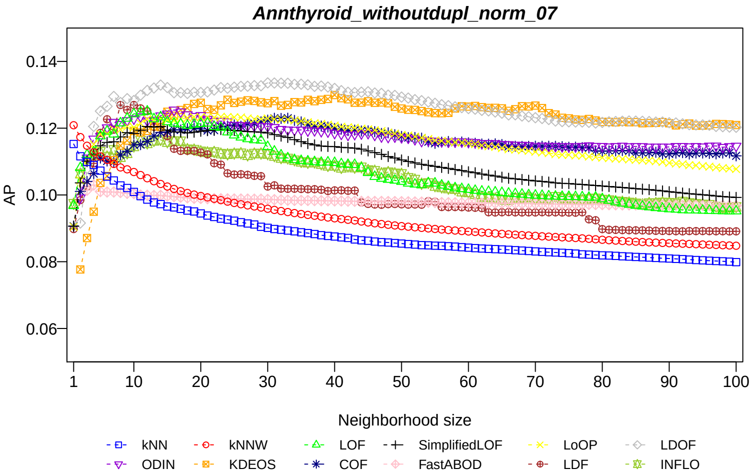
Example: Annthyroid



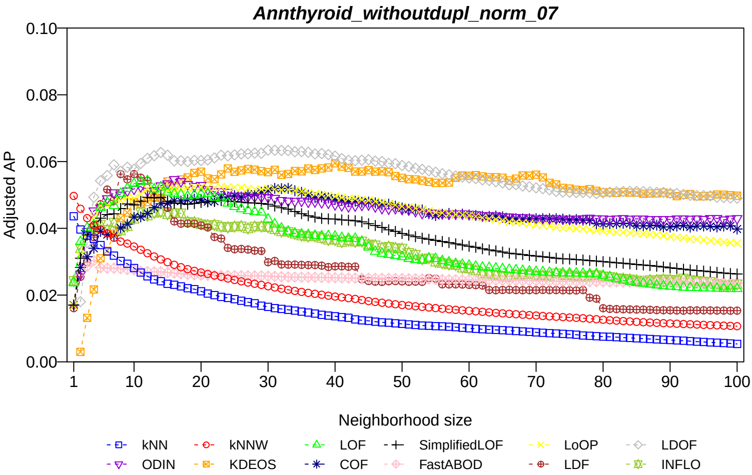
Example: Annthyroid



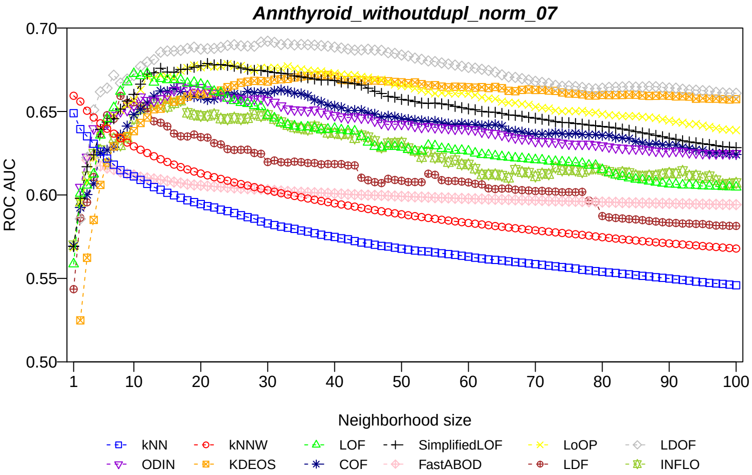
Example: Annthyroid



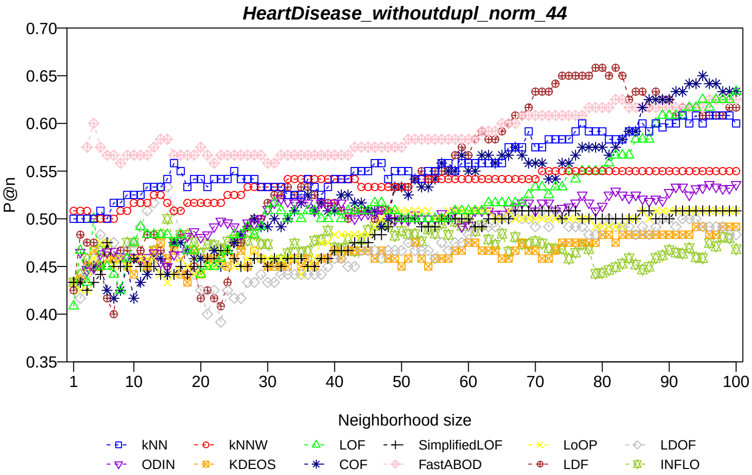
Example: Annthyroid



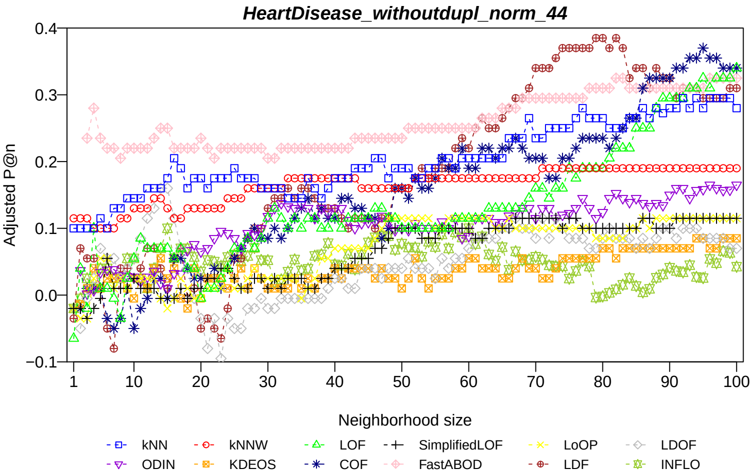
Example: Annthyroid



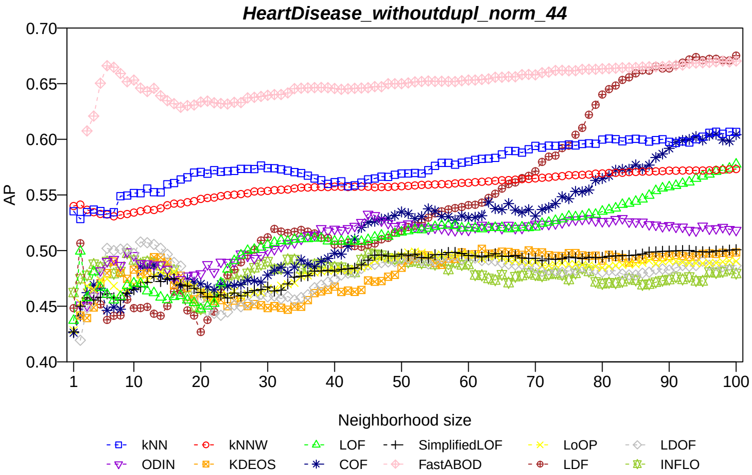
Example: HeartDisease



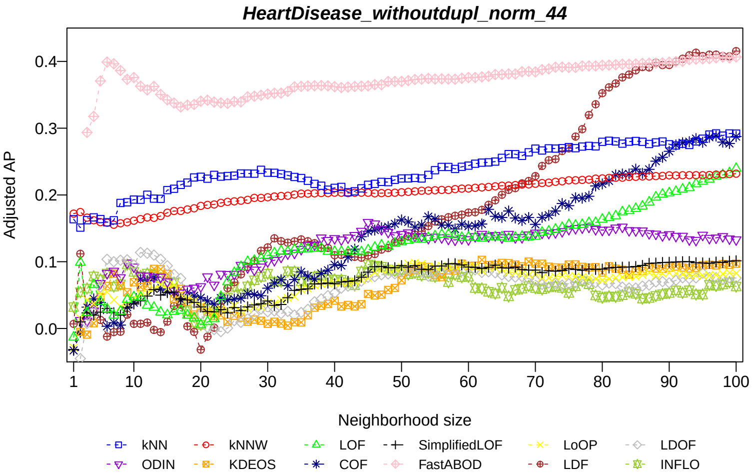
Example: HeartDisease



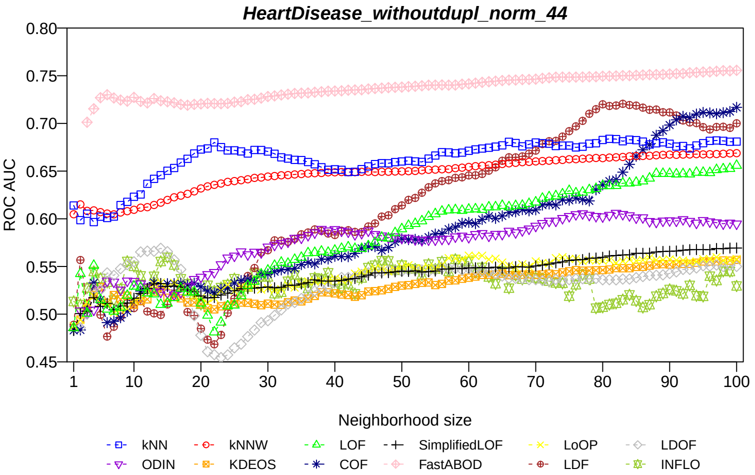
Example: HeartDisease



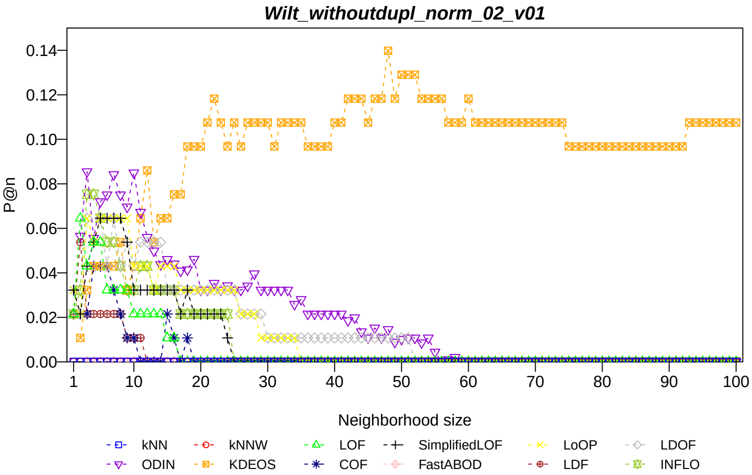
Example: HeartDisease



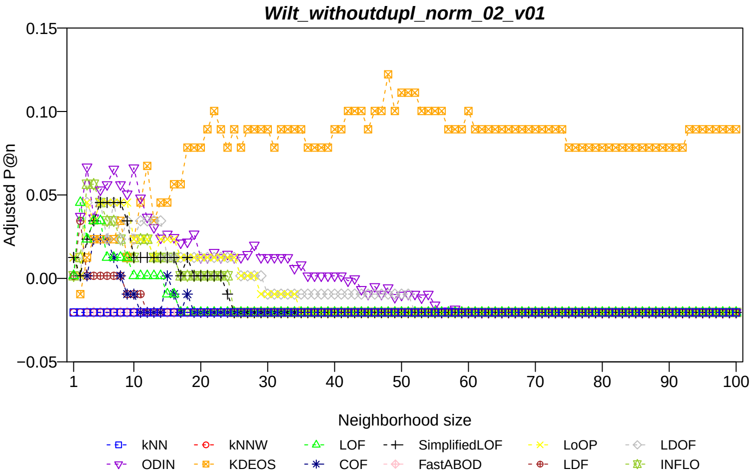
Example: HeartDisease



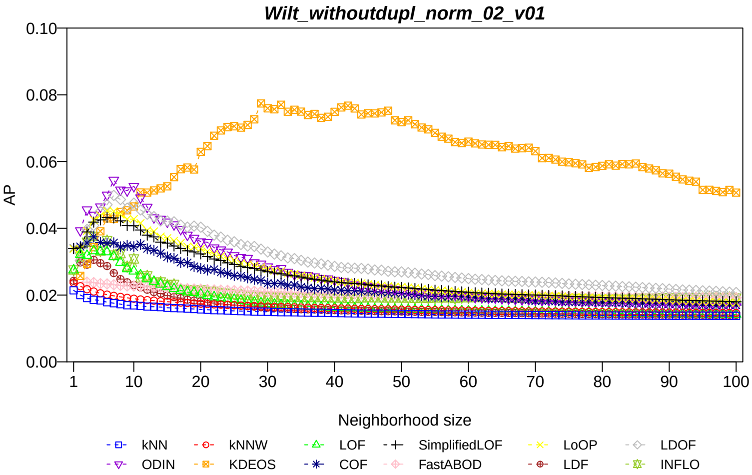
Example: Wilt



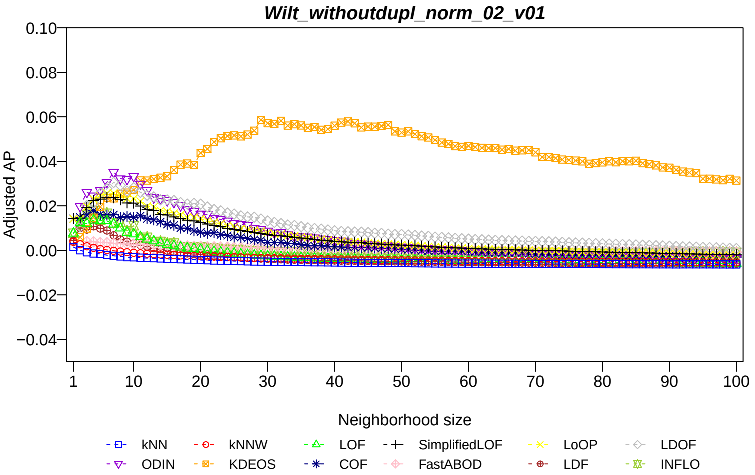
Example: Wilt



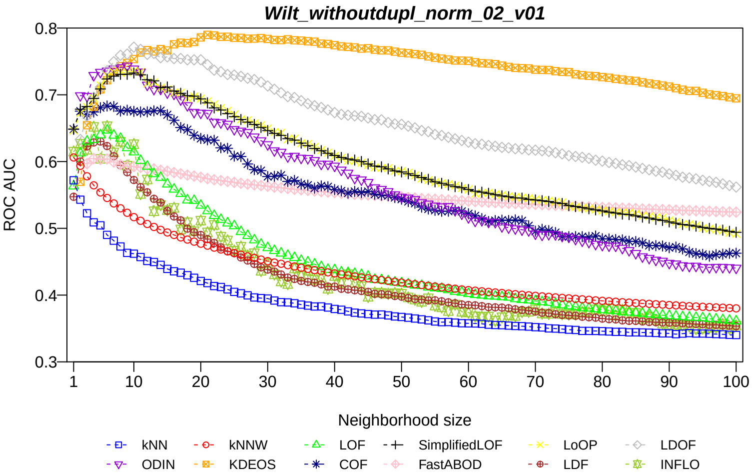
Example: Wilt



Example: Wilt



Example: Wilt



Observations

all results available in the web repository:

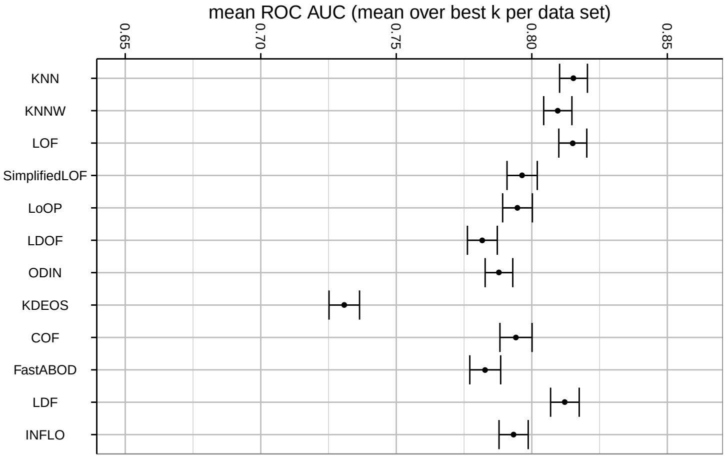
http://www.dbs.ifi.lmu.de/research/outlier-evaluation/

- 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 [Davis and Goadrich, 2006]

Outline

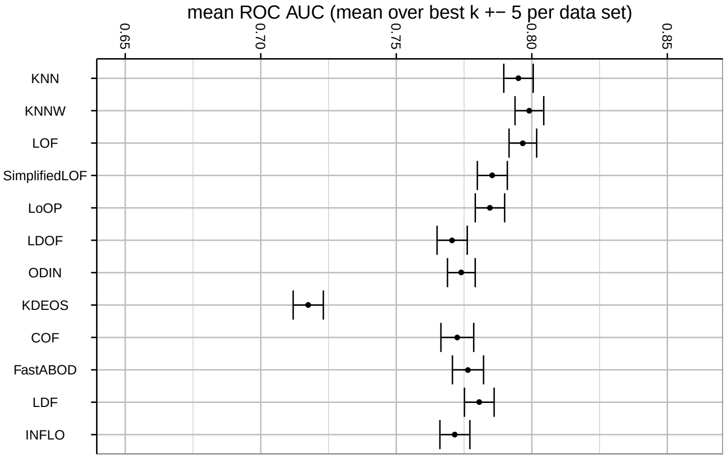
Average ROC AUC per Method

aggregated over all datasets (without duplicates, normalized)



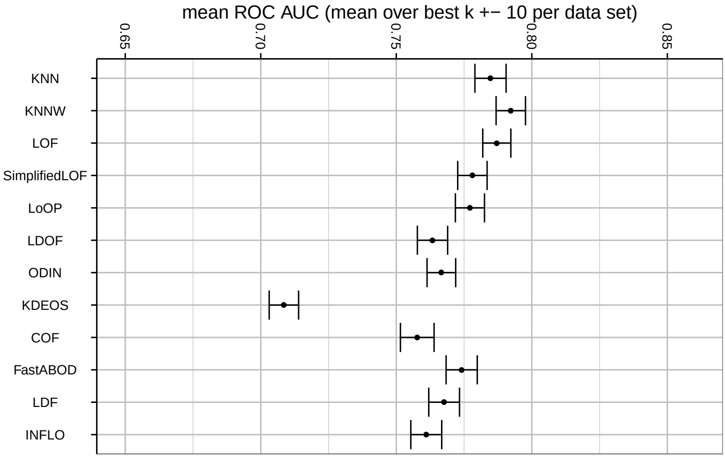
Average ROC AUC per Method

aggregated over all datasets (without duplicates, normalized)



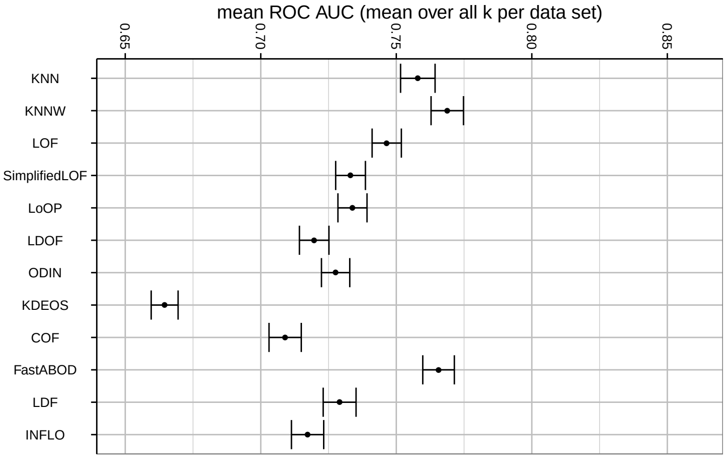
Average ROC AUC per Method

aggregated over all datasets (without duplicates, normalized)



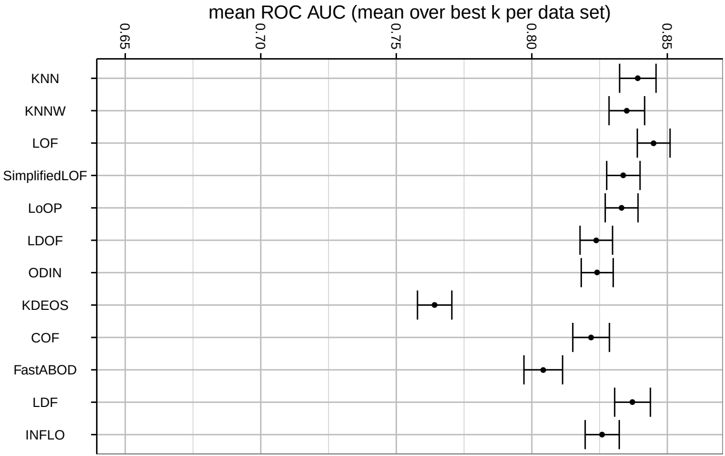
Average ROC AUC per Method

aggregated over all datasets (without duplicates, normalized)



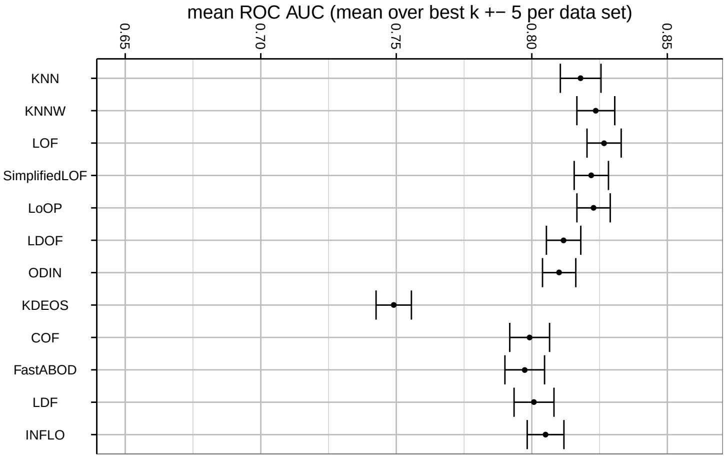
Average ROC AUC per Method

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



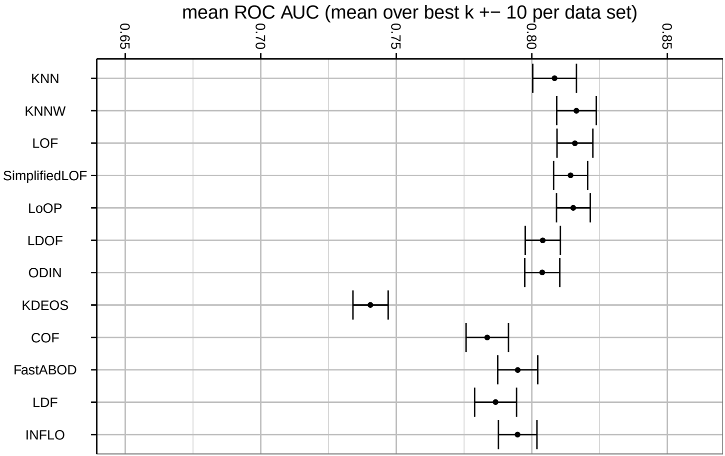
Average ROC AUC per Method

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



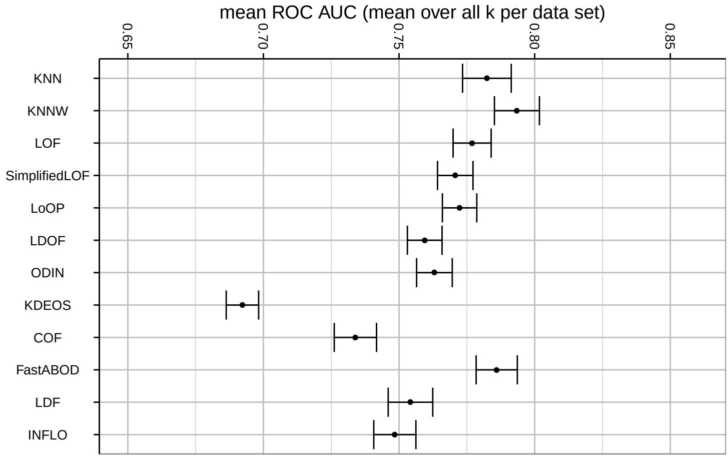
Average ROC AUC per Method

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):

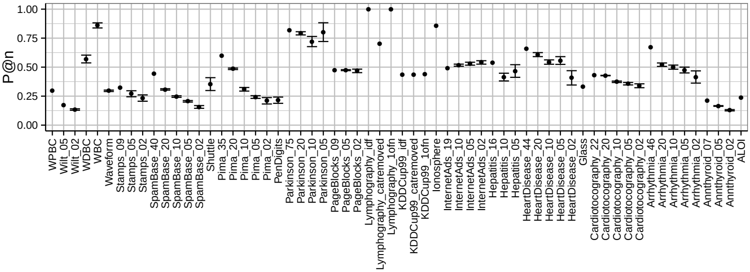
column method is better/worse than row method at 90% ('+'/'–') and 95% ('++'/'––') confidence levels.

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

Outline

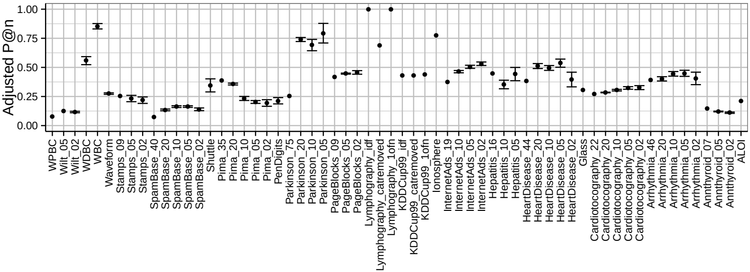
Best Results per Dataset

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



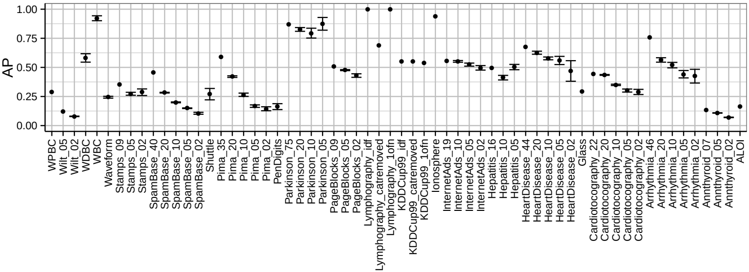
Best Results per Dataset

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



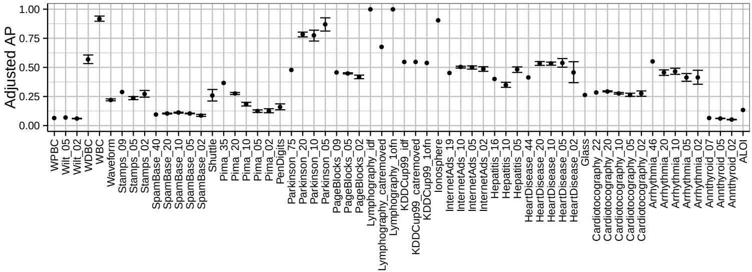
Best Results per Dataset

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



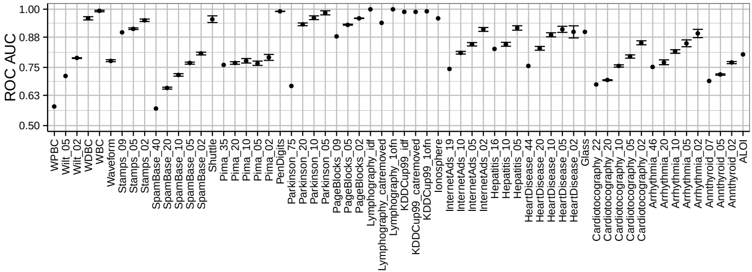
Best Results per Dataset

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

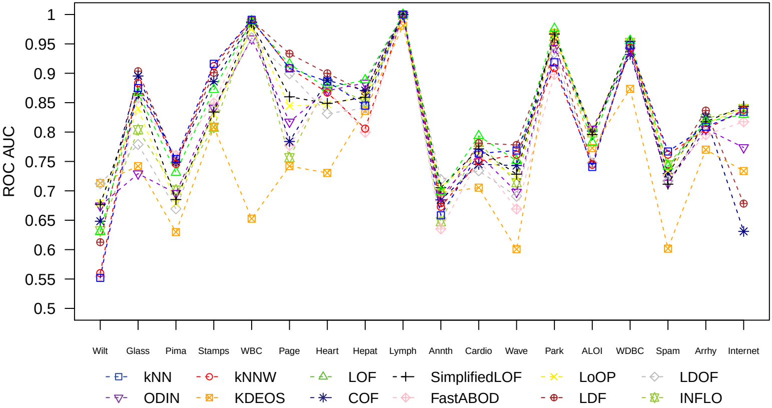


Best Results per Dataset

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



Difficulty and Dimensionality

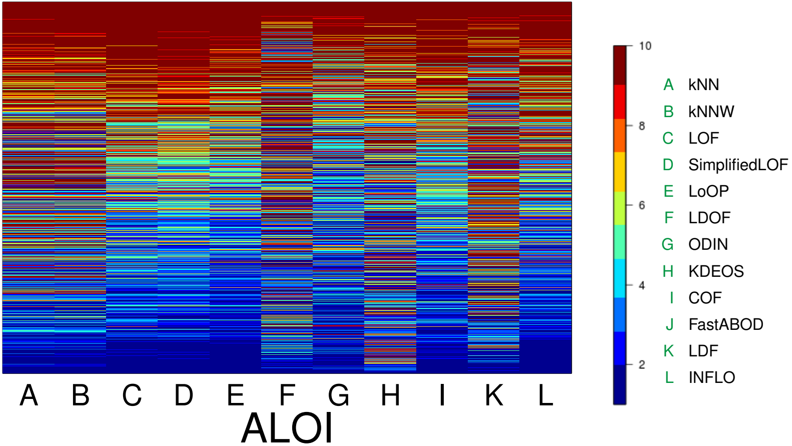


ROC AUC scores, for each method using the best k, on the datasets with 3 to 5% of outliers, averaged over the different dataset variants where available.

The datasets are arranged on the x-axis of the plot from left to right in order of increasing dimensionality.

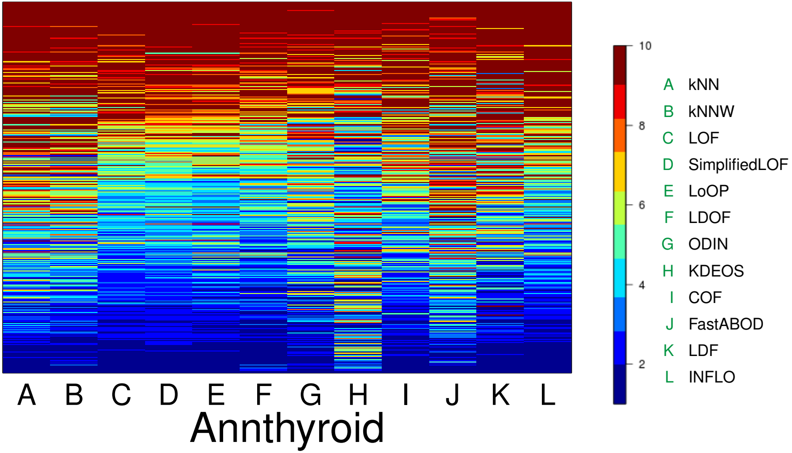
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



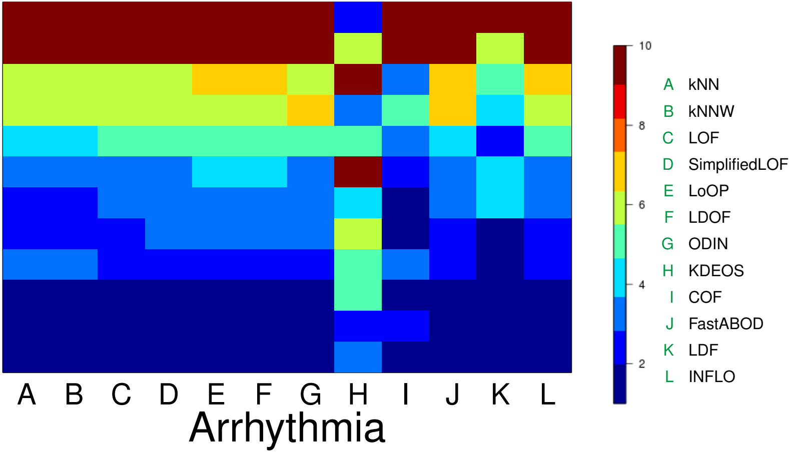
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



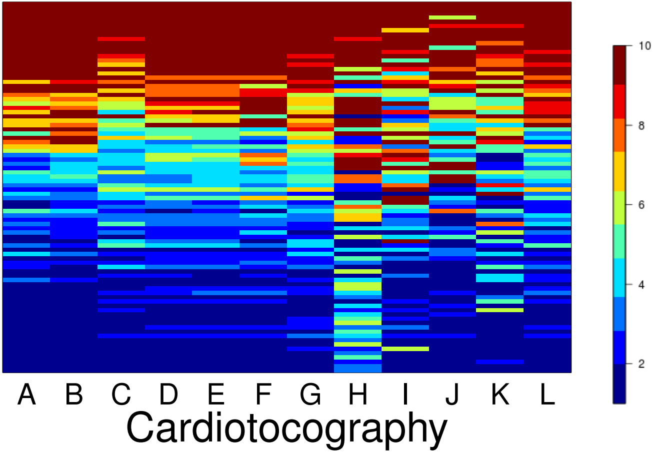
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



Diversity of Outlier Ranks

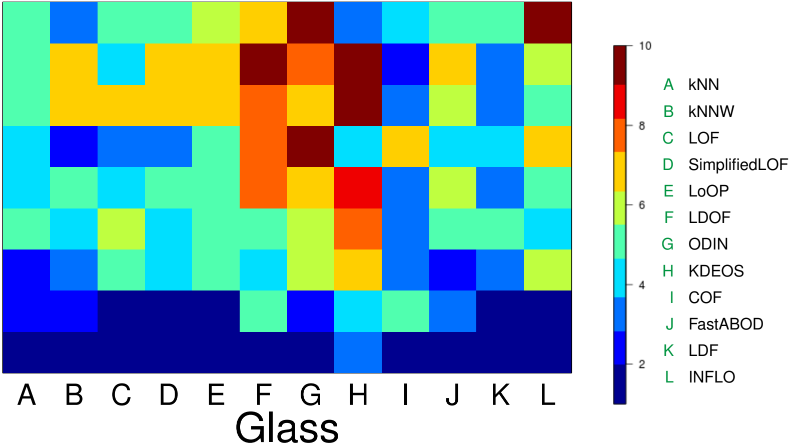
datasets without duplicates, normalized, with 3-5% outliers



A kNN
B kNNW
C LOF
D SimplifiedLOF
E LoOP
F LDOF
G ODIN
H KDEOS
I COF
J FastABOD
K LDF
L INFLO

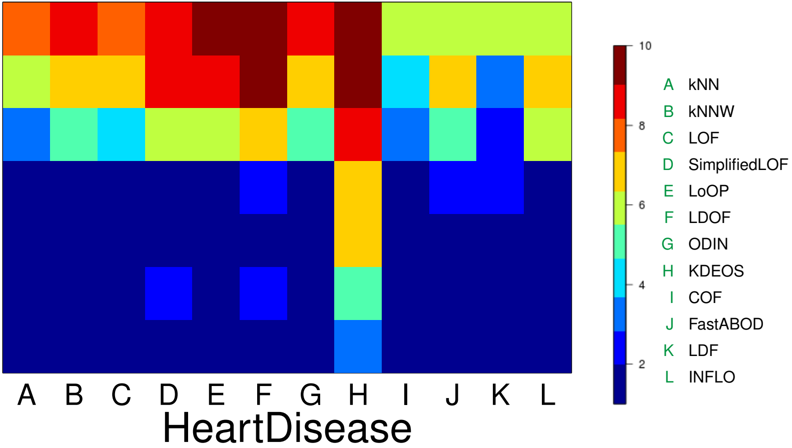
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



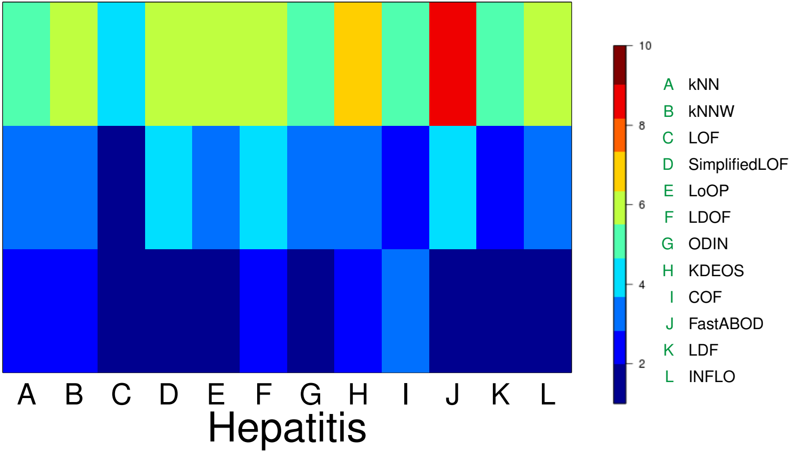
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



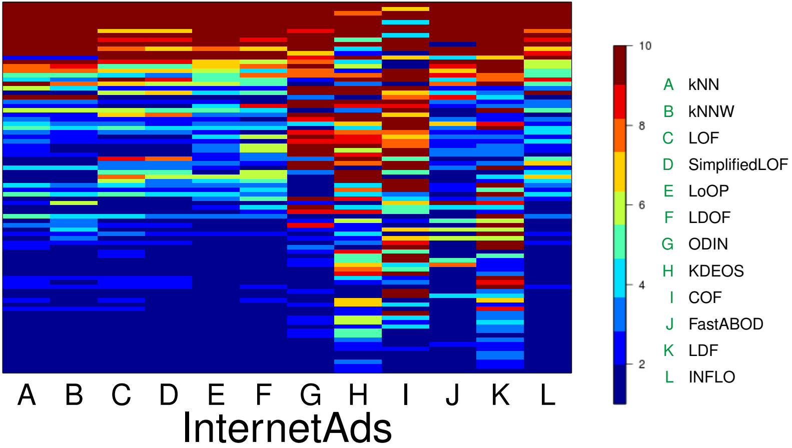
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



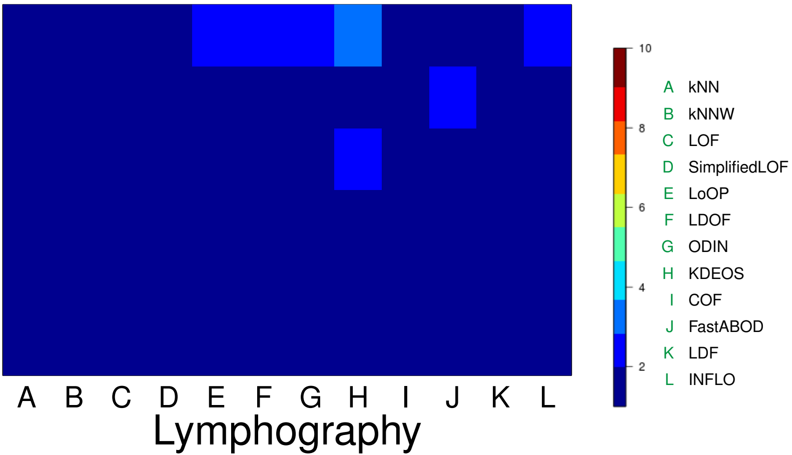
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



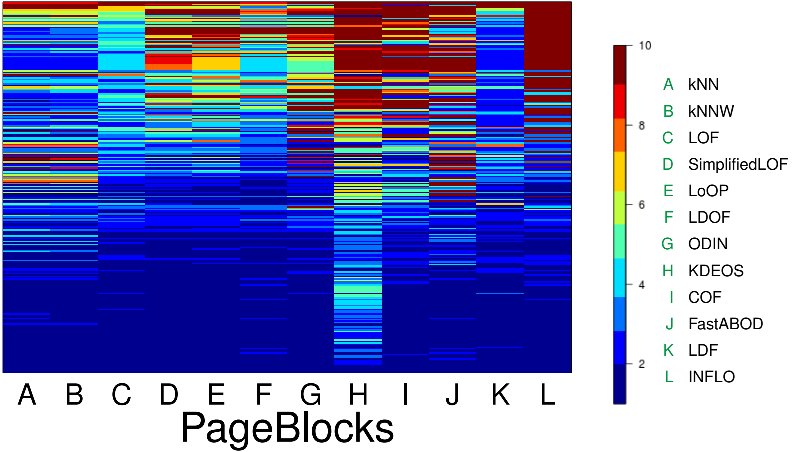
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



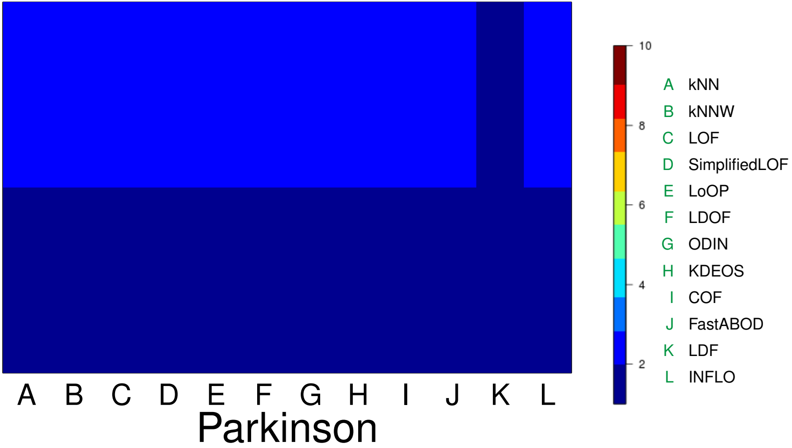
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



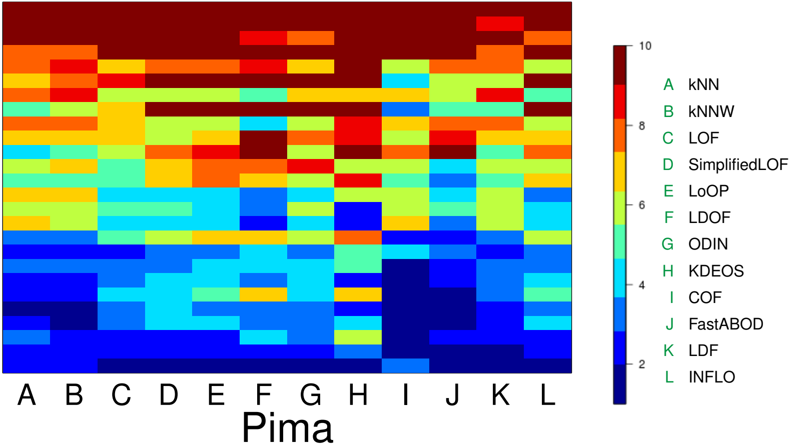
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



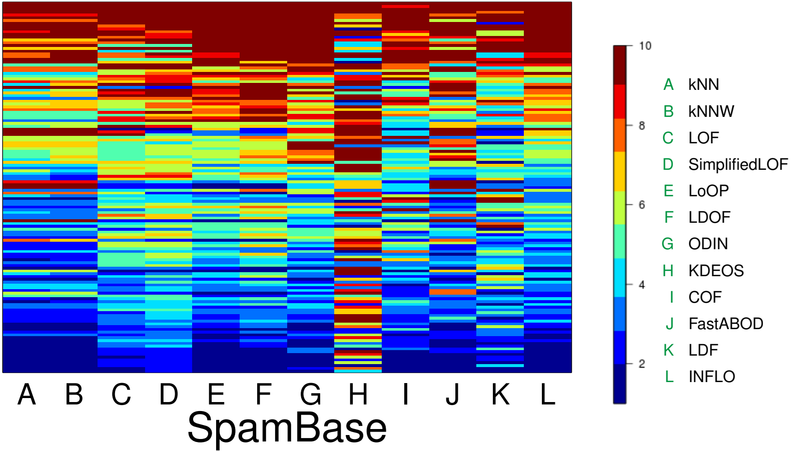
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



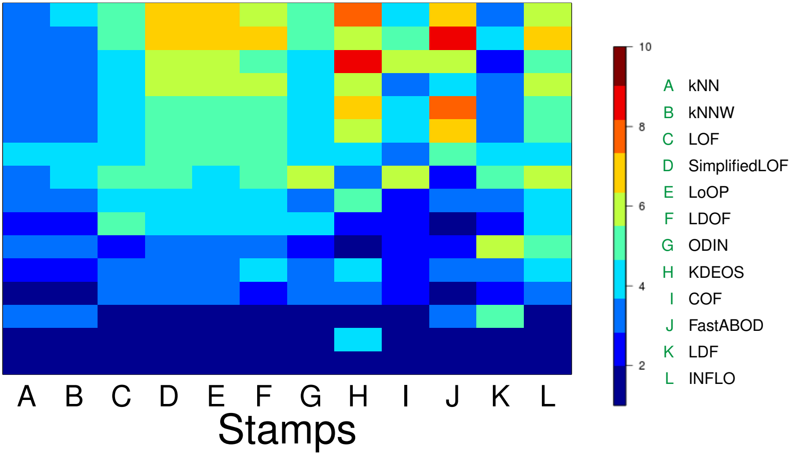
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



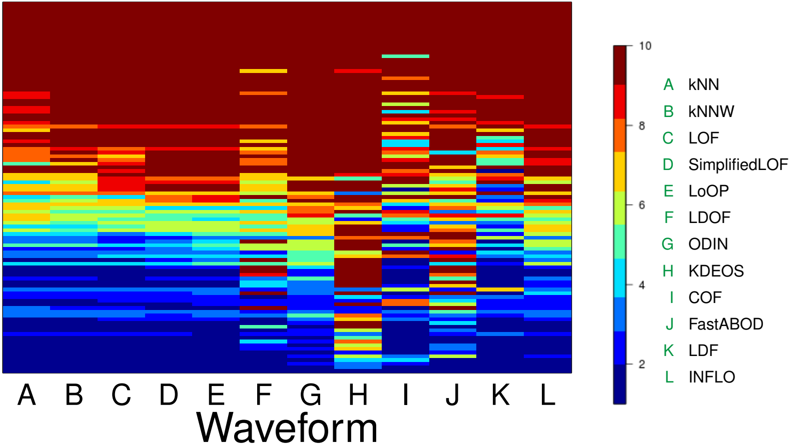
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



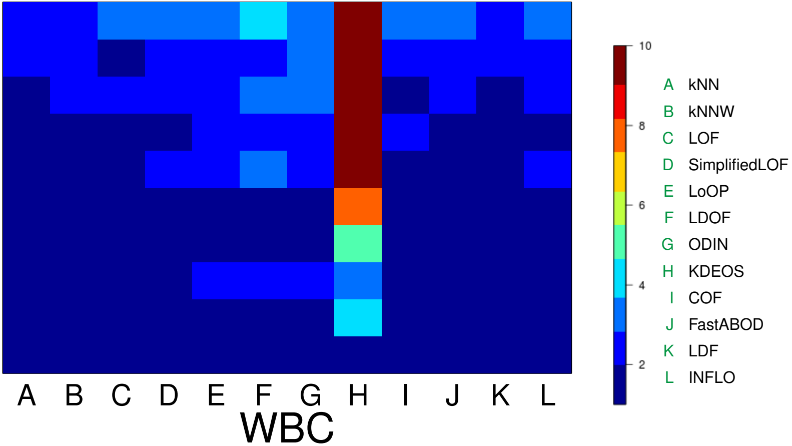
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



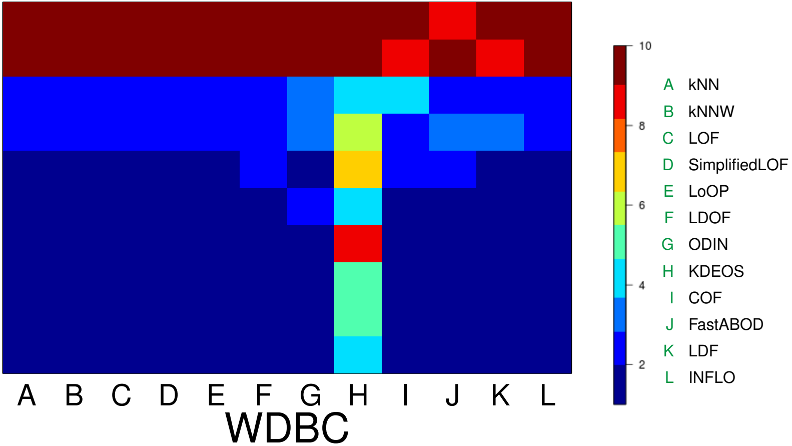
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



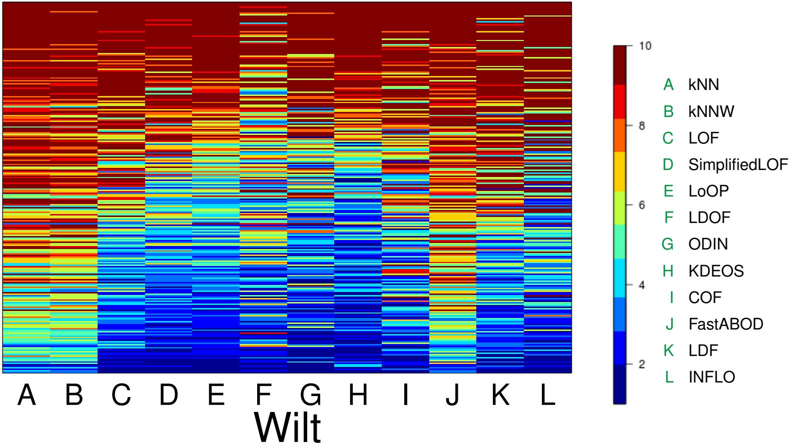
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



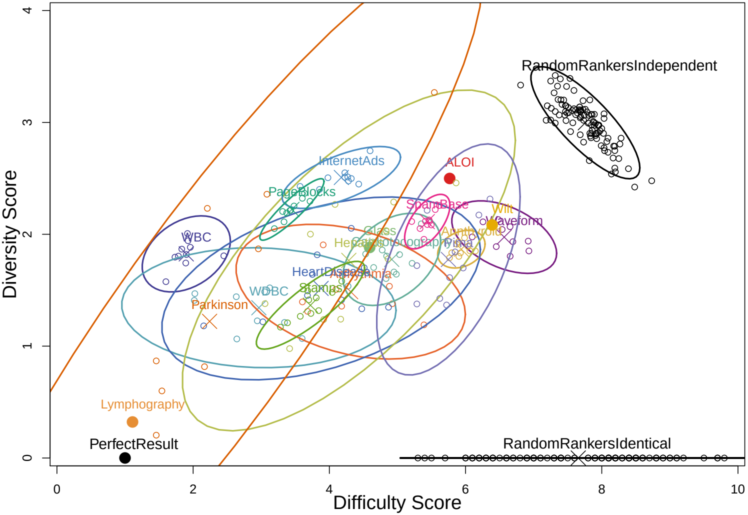
Diversity of Outlier Ranks

datasets without duplicates, normalized, with 3-5% outliers



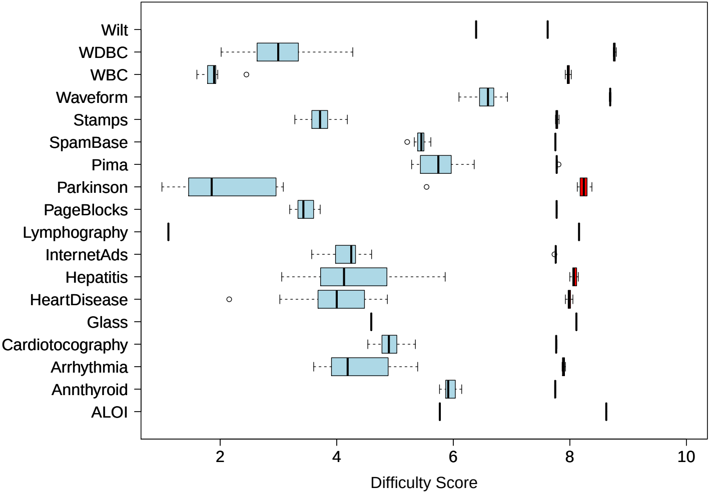
Suitability of Ground Truth Outlier Labels

Difficulty for given labels vs. random labels



Suitability of Ground Truth Outlier Labels

Difficulty for given labels vs. random labels



Outline

Conclusions

- 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
- we tested 12 outlier detection methods on these datasets with a range of choices for the neighborhood parameter $k \in [1, \ldots, 100]$

Conclusions

- 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
- Jörg Sander
- Ricardo J. G. B. Campello
- Barbora Micenková
- Erich Schubert
- Ira Assent
- Mike E. Houle

References I

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*, pages 504–509, 2006. doi: 10.1145/1150402.1150459.

L. Akoglu, H. Tong, and D. Koutra. Graph-based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015. doi: 10.1007/s10618-014-0365-y.

F. Angiulli and C. Pizzuti. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):203–215, 2005. doi: 10.1109/TKDE.2005.31.

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.

R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):5:1–51, 2015. doi: 10.1145/2733381.

References II

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. *Data Mining and Knowledge Discovery*, 2016. doi: 10.1007/s10618-015-0444-8.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):Article 15, 1–58, 2009. doi: 10.1145/1541880.1541882.

N. Craswell. Precision at n. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 2127–2128. Springer, 2009a. doi: 10.1007/978-0-387-39940-9_484.

N. Craswell. R-precision. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, page 2453. Springer, 2009b. doi: 10.1007/978-0-387-39940-9_486.

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*, pages 233–240, 2006.

T. de Vries, S. Chawla, and M. E. Houle. Density-preserving projections for large-scale local anomaly detection. *Knowledge and Information Systems (KAIS)*, 32(1):25–52, 2012. doi: 10.1007/s10115-011-0430-4.

References III

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.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

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*, pages 430–433, 2004. doi: 10.1109/ICPR.2004.1334558.

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

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1): 193–218, 1985.

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*, pages 577–593, 2006. doi: 10.1007/11731139_68.

References IV

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*, pages 1037–1048, 2012. doi: 10.1109/ICDE.2012.88.

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*, pages 444–452, 2008. doi: 10.1145/1401890.1401946.

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. doi: 10.1137/1.9781611972818.2.

H.-P. Kriegel, E. Schubert, and A. Zimek. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? submitted, 2016.

References V

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*, pages 61–75, 2007. doi: 10.1007/978-3-540-73499-4_6.

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. O. Marques, R. J. G. B. Campello, A. Zimek, and J. Sander. On the internal evaluation of unsupervised outlier detection. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management (SSDBM), San Diego, CA*, pages 7:1–12, 2015. doi: 10.1145/2791347.2791352.

H. V. Nguyen and V. Gopalkrishnan. Feature extraction for outlier detection in high-dimensional spaces. *Journal of Machine Learning Research*, Proceedings Track 10:66–75, 2010.

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 VI

G. H. Orair, C. Teixeira, Y. Wang, W. Meira Jr., and S. Parthasarathy. Distance-based outlier detection: Consolidation and renewed bearing. *Proceedings of the VLDB Endowment*, 3(2):1469–1480, 2010.

M. Radovanović, A. Nanopoulos, and M. Ivanović. Reverse nearest neighbors in unsupervised distance-based outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 2014. doi: 10.1109/TKDE.2014.2365790.

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*, pages 427–438, 2000. doi: 10.1145/342009.335437.

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. doi: 10.1137/1.9781611972825.90.

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. *Data Mining and Knowledge Discovery*, 28(1):190–237, 2014a. doi: 10.1007/s10618-012-0300-z.

References VII

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*, pages 542–550, 2014b. doi: 10.1137/1.9781611973440.63.

E. Schubert, A. Koos, T. Emrich, A. Züfle, K. A. Schmid, and A. Zimek. A framework for clustering uncertain data. *Proceedings of the VLDB Endowment*, 8(12):1976–1979, 2015. doi: 10.14778/2824032.2824115.

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*, pages 535–548, 2002. doi: 10.1007/3-540-47887-6_53.

E. Zhang and Y. Zhang. Average precision. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 192–193. Springer, 2009. doi: 10.1007/978-0-387-39940-9_482.

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.

References VIII

A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, 2012. doi: 10.1002/sam.11161.

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*, pages 428–436, 2013. doi: 10.1145/2487575.2487676.