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

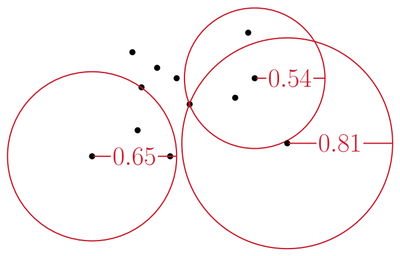
Outlier Detection

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 example: take the kNN distance of a point as its outlier score

Advanced model example: compare the densities of neighbors (e.g. LOF)



Motivation

- new outlier detection methods developed every year
- some studies about efficiency
- specializations for different areas
- evaluation of effectiveness remains challenging
  - characterisation of outlierness differs
  - lack of common benchmark data
  - measure of success? (most commonly: ROC)

Benchmark Parameters

Methods in Benchmark

- kNN, kNN-weight
- LOF
- SimplifiedLOF, COF, INFLO, LoOP
- LDOF, LDF, KDEOS
- ODIN (related to low hubness outlierness)
- FastABOD
- In next iteration: LIC, VoV, DWOF, IDOS

Selection Criteria for Methods

- 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
- both 'global' and 'local' methods
- included variants of LOF vary different components of the typical local outlier model: notion of neighborhood, distance, density estimates, model comparison
- all methods are available in ELKI

Evaluation Measures Studied

Ranking evaluation measures used:

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

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

- Average Precision:

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

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

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

- Maximum F1-Measure

$$\text{Max-F1} := \max_{\text{score}} \text{F1}(\text{Prec.}(\text{score}), \text{Rec.}(\text{score}))$$

- And *adjusted for chance* versions of each.

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

Datasets Studied

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:

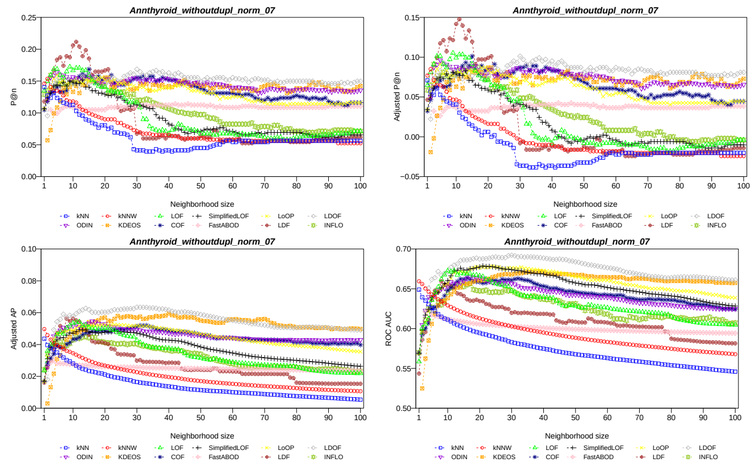
Dataset	Preprocessing	N	$\lvert O\rvert$	Attrib. num	Attrib. cat	Version used by
ALOI	50000 images, 27 attr.	50000	1508	27		[Kri+11], [Sch+12]
	24000 images, 27648 attr.					[dCH12]
Glass	Class 6 (*out.*) vs. others (*in.*)	214	9	7		[KMB12]
Ionosphere	Class 'b' (*out.*) vs. class 'g' (*in.*)	351	126	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	$\lvert O\rvert$	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	

Observations

Example: Correlation of Measures on Annthyroid

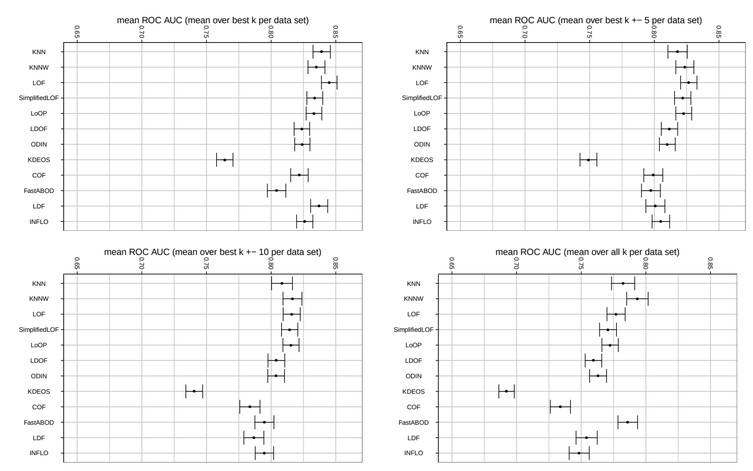


Observations on Measures

- performance trends differ across algorithms, datasets, parameters, and evaluation methods
- ROC AUC is 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, and have low numerical precision
- 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

Characterization of the Methods

Average ROC AUC per Method aggregated over all datasets
(without duplicates, normalized, at most 5% outliers)



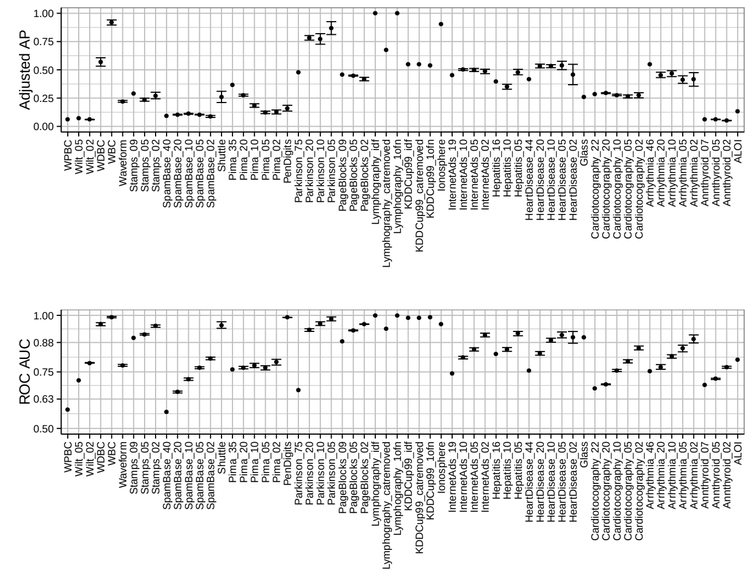
Statistical Test

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

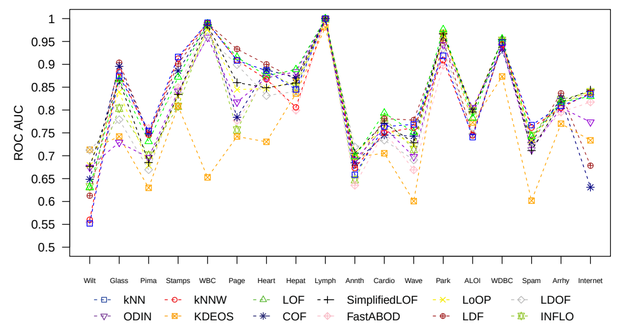
Nemenyi post-hoc test using optimal parameters each
90% ('+'/'−') and 95% ('++'/'−−') confidence levels.

Characterization of the Datasets

Average best performance of all methods, per dataset (without dupl. norm.)



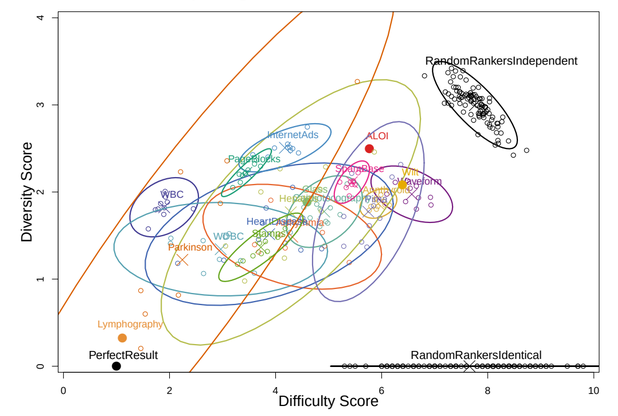
Difficulty and Dimensionality

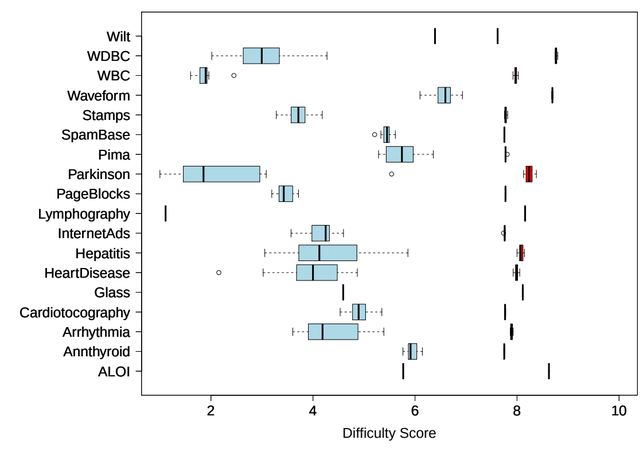


ROC AUC scores, for each method using the best k, on the datasets with 3 to 5% of outliers, averaged over the different dataset variants where available.
The datasets are arranged on the x-axis of the plot from left to right in order of increasing dimensionality.

Suitability of Ground Truth Labels

Difficulty for given labels vs. random labels





Conclusions

- we disussed evaluation measures for outlier rankings: $P@n$, AP, and ROC (AUC)
- we proposed adjustment for chance for $P@n$ and for AP
- we disussed preprocessing issues for the preparation of outlier datasets with annotatded ground truth and we 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 $k \in [1, \ldots, 100]$
- we aggregate and analyse the resulting $> 1,3$ mil. experiments and
  - summarize the effectiveness of the 12 methods
  - study the suitability of the datasets for evaluation
- 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

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

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 Min. Knowl. Disc.* 30 (4 2016), pp. 891–927