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

Guilherme O. Campos1, Arthur Zimek2, Jörg Sander3, Ricardo J. G. B. Campello1
Barbora Michenkova1, Erich Schubert5,7, Ira Assent1, Michael E. Houle6
1University of São Paulo 2University of Southern Denmark 3University of Alberta 4Aarhus University
5Ludwig-Maximilians-Universität München 6National Institute of Informatics 7Ruprecht-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
  - characterization 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
- LDIF, LDF, KDEOS

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 \( n = [10] \):
  \[
  \text{Precision} = \frac{|\{i \in \text{rank}(o) \leq n \}|}{n}
  \]
- Average Precision:
  \[
  \text{AP} = \frac{1}{|O|} \sum_{o \in O} \text{rank}(o)
  \]
- Area under the ROC curve (ROCAUC or AUROC):
  \[
  \text{ROCAUC} = \frac{1}{|O|} \sum_{o \in O} \text{score}(o)
  \]
- Maximum F1-Measure
  \[
  \text{Max F1} = \frac{\text{F1}(\text{score}, \text{ROC}(\text{score}))}{\text{score}}
  \]
- And adjusted for chance versions of each
  Adjusted Index = \[
  \frac{\text{Maximum Index} - \text{Expected Index}}{\text{Expected Index}}
  \]

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

Dataset Used in the Literature

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

Datasets Studied

Conclusions

- we discussed evaluation measures for outlier rankings: Precision@n, AP, and ROC (AUC)
- we proposed adjustment for chance for Precision@n and for AP
- we discussed preprocessing issues for the preparation of outlier datasets with annotated 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 \( k \in [1, \ldots, 10] \)
- we aggregate and analyse the resulting \( > 1,3 \text{ml} \) 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/