O

Outlier Detection

Arthur Zimek[1] and Erich Schubert[2]
[1]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
[2]Heidelberg University, Heidelberg, Germany

Synonyms

Anomaly detection; Fraud detection; Identification of outliers; Rejection of outliers

Definition

Outlier detection aims at identifying those objects in a database that are unusual, i.e., different than the majority of the data and therefore suspicious resulting from a contamination, error, or fraud. In a statistical modeling, the assessment of "being unusual" is typically based on a parametric model of the data, identifying those objects that do not fit well to the modeled distribution as outliers. In the database context, the statistical intuition of "being unusual" is typically modeled in an approximate but more efficient, nonparametric way by (local) density estimates and comparison to some reference set.

Historical Background

Filtering out those observations that look suspiciously different than the majority of observations is a procedure probably tacitly practiced since people studied data collections and tried to make sense out of observations. In the eighteenth century, Bernoulli criticizes this practice among astronomers as he does not see sufficient reason to separate those out that do not fit to the model, pointing out that those rejected observations could have possibly served best to supply corrections to the model.

Statistical reasoning typically tackles the problem with parametric approaches, that is, one assumes the data follows some specific distribution and fits a model of such distribution to the data at hand. Outliers would then be those data objects that do not fit well to the fitted distribution model. Accordingly, there is an abundance of formulations of statistical tests, for different assumptions of distributions by type of parameters, number of distributions, number of variables (dimensions), etc. [1–3].

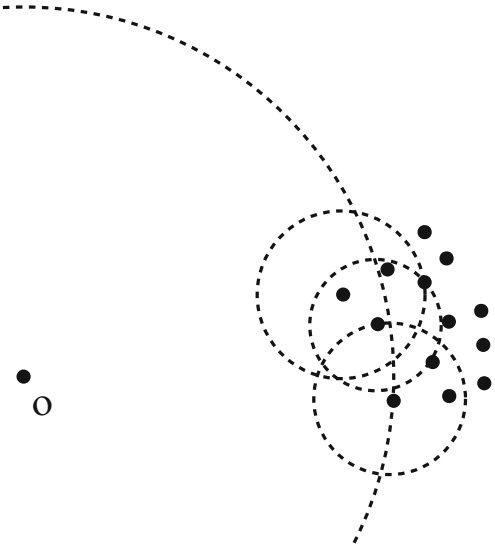
Scientific Fundamentals

Database-Oriented Efficient Outlier Models

Distance-Based Outlier Detection
The work of Knorr and Ng on the "distance-based" notion of outliers (DB-outlier) simplified

L. Liu, M.T. Özsu (eds.), *Encyclopedia of Database Systems*,
DOI 10.1007/978-1-4899-7993-3_80719-1

Outlier Detection, Fig. 1 The DB-outlier model: for a fixed neighborhood radius (ε-range query), neighbors are counted. While the amount of neighbors does not exceed $\pi = 10\%$ for point o, for point p it does. As a consequence, for the given ε and π, o will be labeled outlier; p will be labeled inlier



Outlier Detection, Fig. 2 The kNN outlier model: in areas of low density, larger radii are required to capture k neighbors than in areas of higher density (here: $k = 3$). The required radius (i.e., the k-distance) is used as outlier score. This allows to rank the data points according to their outlier characteristic. Clearly, o is more prominently an outlier than the other data points

statistical distribution-based approaches in order to enable efficient computation for large data sets [4]. Their method requires the specification of a distance threshold (ε) and a percentage (π). Those database objects that have less neighbors within the ε-range than specified by the percentage threshold π are considered outliers (cf. Fig. 1).

A method in the same spirit uses the distances to the k-nearest neighbors (kNN) of each object [5]. As a variant, the sum of distances to all points within the set of k-nearest neighbors (called the "weight") has been used as an outlier degree [6]. As a result, these methods do not deliver a binary decision (outlier vs. inlier) but a ranking of the objects w.r.t. the outlier degree or outlier score of the objects (as estimated by the outlier detection model), where outliers should be top ranked (cf. Fig. 2).

The kNN-based method could be considered Curried (or Schönfinkeled) form of the DB-outlier model as its result could be seen as a function that takes an ε parameter to map to a DB-outlier solution: the decision inherent to the DB-outlier model could be derived from the kNN ranking by specifying some ε threshold as cutoff value.

From a statistical point of view, both models, DB-outlier and kNN, relate to density estimation using a uniform/kNN kernel.

Local Outlier Factor

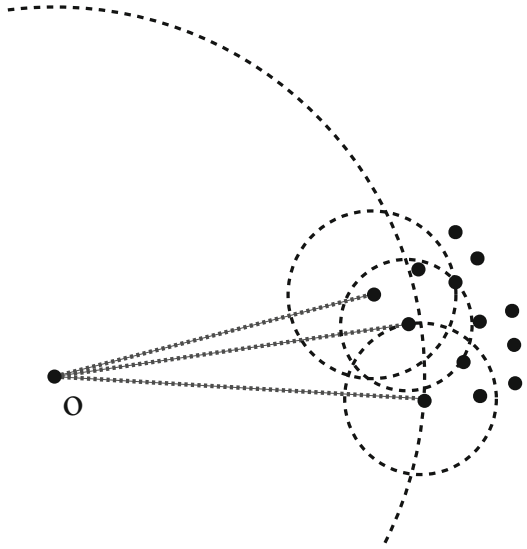
To derive rankings of outlierness instead of binary decisions remained the method of choice for the design of subsequent methods. The often so-called density-based but more accurately called "local" approaches consider ratios between the local density around an object and the local density around its neighboring objects, starting with the seminal LOF ("local outlier factor") [7] algorithm (cf. Fig. 3).

Nevertheless, the method for density estimation is also a bit more refined in LOF compared to the previous models that were using the kNN distance directly. Instead, LOF uses an asymmetric *reachability distance*:

$$\text{reach-dist}_k(o \leftarrow p) = \max\{k\text{-dist}(p), d(o,p)\}$$

where k-dist(p) is the distance from p to its kth-nearest neighbor. This distance definition has been used in hierarchical density-based clustering. Based on the reachability distances, the *local reachability density* (lrd) is estimated by

$$\text{lrd}(o) := 1 \Big/ \frac{\sum_{p \in k\text{NN}(o)} \text{reach-dist}_k(o \leftarrow p)}{|k\text{NN}(o)|}$$



Outlier Detection, Fig. 3 The local outlier model: the density estimate for some point o is compared not to the density estimates of all the other points in the database but to the density estimates of o's local neighborhood only. This allows a flexible adaptation to local variations of typical density levels

where $k\text{NN}(o)$ are the k-nearest neighbors of o.

The LOF score is a comparison of the local density estimate (lrd) for a given point o to the local density estimates (lrds) of o's neighborhood:

$$\text{LOF}(o) := \text{avg}_{n \in k\text{NN}(o)} \frac{\text{lrd}(n)}{\text{lrd}(o)}.$$

Many variants have adapted the original LOF idea in different aspects [8]. Major differences among the variants are in the method of density estimation and in the definition of "neighboring objects." In many variants of LOF, a simplified version ("simplified LOF" [8]) is used, substituting (presumably unintentionally) the regular distance $d(o, p)$ for the reachability distance.

Categories and Variants

Global and Local Outlier Detection

An important differentiation is whether an outlier score is "global" or "local." However, this is not a clear cut of methods but rather a characteristic of a "degree of locality." The differentiation of methods as "global" or "local" reflects the scale of computation and comparison of outlier score estimates for individual objects: are individual scores computed based on a local neighborhood (such as the typical so-called distance-based and the so-called "density-based" approaches) or based on the global data set (as it is typically the case in parametric statistical modeling)? Are the individual scores compared to the global database (as it is the case in the "distance-based" models DB-outlier and kNN scores) or to a local reference set only (which was the genuine novelty of the "local" outlier factor (LOF))? Different variants of outlier detection in the literature are "local" in different meaning and to different degrees [8].

Distance-Based and Density-Based

The differentiation of methods into the categories "distance-based" vs. "density-based" is traditional yet meaningless. Approaches that are termed "distance-based" or "density-based" both consider more or less simplistic density estimates as base for the computation of outlier scores.

"Distance-based" methods could be seen, however, as a subset of "density-based" methods. The virtue of methods in this subset would be the simplicity of the density estimate that allows the use of indexing-related techniques for acceleration of the computation of the outlier score.

Variants for Improved Efficiency

When computing a ranking of outlier scores, one is typically only interested in the top-ranked objects (i.e., in the most likely outliers). Many efficient variants of basic models have been proposed, typically as filter-refinement processing, refining only the outlier scores of those objects that still have a chance to be among the top-n outliers. Filters are typically based on approximate neighborhood computation, pruning of partitions (during neighborhood search or during outlier search) and ranking of (neighborhood or outlier) candidates. Orair et al. [9] analyze and categorize such variants and the techniques used.

Application-Specific Methods and Generalized Models

The concept of "local density" can be abstracted to handle more complex data types: while density – or some other measure – is computed usually on numerical attributes, the neighborhood can be based on a different data source or a non-Euclidean semantic of closeness, and standard methods can thus be applied to complex data. Examples of such adaptations include the application to geo-spatial/temporal data, graph data, video streams, or text analysis [8].

Recent Developments

The challenge of high-dimensional data has triggered development of specialized solutions for outlier detection. Many database methods for outlier detection rely on nearest neighbor retrieval. Since nearest neighbor search is negatively impacted by high dimensionality, there are two main categories of approaches to outlier detection in high-dimensional data: (i) methods to improve efficiency or effectiveness of outlier detection in a high-dimensional data space or (ii) methods identifying outliers in subspaces of a high-dimensional data space. For both directions, there are some first approaches, but there remain also many open challenges [10].

Another recent direction is to apply ensemble learning techniques, well known and studied in the area of classification or clustering, to the outlier detection problem. This transfer is nontrivial. For building effective ensembles, minimally accurate yet diverse base methods are required and their predictions need to be combined. All three requirements, accuracy, diversity, and combination procedures, have found only heuristic solutions in outlier detection so far [11].

Also methods for nonnumerical data types, such as symbolic sequences or graph data, find increasing attention in the research community [12, 13].

Key Applications

Traditionally, outlier detection was mostly used for data cleaning, as it can improve the results of other analysis methods that are sensitive to extreme deviations. With manually collected data, contamination would often stem from unintentional errors. However, not every unusual observation is an error – and non-erroneous, unusual data can often provide new insights. Furthermore, in electronic communication and automated data, an increasing amount of manipulation can be observed, such as insurance fraud. An outlier in network activity can arise from an attack. An unusual trading activity may be caused by manipulation or a software error. Outlier detection can be used when the exact nature of the effect is unknown beforehand: while an engineer may expect a device to fail eventually, it is hard to predict the exact way it will fail.

With the increasing volume of data, outliers become more and more common, and thus many applications have to deal with thousands of outliers. For example, a LiDAR sensor will yield erroneous height measurements due to reflections and refraction that need to be quickly identified and filtered. At 150,000 laser pulses a second, a 1 in a million error will occur several times per minute.

Application scenarios thus include detecting human error, or fraud, data contamination, electronic attacks, machine failures, and interviewer fabrication, finding contradicting examples for a theory, detecting rare medical conditions, and cleaning data for further analysis with other methods.

Future Directions

Since ranking methods typically abstain from deciding on a good cutting point in the ranking (i.e., they do not by default deliver a binary decision based on the ranking), an open problem is to define a good value of n for the top-n outliers or to define a meaningful threshold for outlier scores. Also, improving the interpretability of outlier scores and the explanation of outlierness is a topic that just starts to gain attention in the research community [14]. However, these decisions will always remain highly application dependent.

After all, the decision of some method about some database object being an outlier would only mean that this object is suspicious, i.e., a domain expert needs to examine the case.

Applying ensemble techniques to outlier detection is a novel direction of great potential where many open challenges remain [11].

Tackling high-dimensional data remains challenging as well. Existing methods are pointing out the challenges rather than truly tackling them [10].

URL to Code

Many standard methods for outlier detection are available in the open-source data mining framework ELKI [15] at http://elki.dbs.ifi.lmu.de/.

Cross-References

- Anomaly Detection on Streams
- Data Mining
- Density-Based Clustering
- Ensemble
- Filter/Refinement Query Processing
- Indexing and Similarity Search
- kNN Query
- Range Query

Recommended Reading

1. Hawkins D. Identification of outliers. London: Chapman and Hall; 1980.
2. Barnett V, Lewis T. Outliers in statistical data. 3rd ed. Chichester: Wiley; 1994.
3. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. Wiley Interdiscip Rev Data Min Knowl Discov. 2011;1(1):73–9.
4. Knorr EM, Ng RT, Tucanov V. Distance-based outliers: algorithms and applications. VLDB J. 2000;8(3–4):237–53.
5. Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas; 2000. p. 427–38.
6. Angiulli F, Pizzuti C. Outlier mining in large high-dimensional data sets. IEEE Trans Knowl Data Eng. 2005;17(2):203–15.
7. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. In: Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas; 2000. p. 93–104.
8. Schubert E, Zimek A, Kriegel HP. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. Data Min Knowl Disc. 2014;28(1):190–237.
9. Orair GH, Teixeira C, Wang Y, Meira Jr W, Parthasarathy S. Distance-based outlier detection: consolidation and renewed bearing. Proc VLDB Endowment. 2010;3(2):1469–80.
10. Zimek A, Schubert E, Kriegel HP. A survey on unsupervised outlier detection in high-dimensional numerical data. Stat Anal Data Min. 2012;5(5):363–87.
11. Zimek A, Campello RJGB, Sander J. Ensembles for unsupervised outlier detection: challenges and research questions. ACM SIGKDD Explor. 2013;15(1):11–22.
12. Chandola V, Banerjee A, Kumar V. Anomaly detection for discrete sequences: a survey. IEEE Trans Knowl Data Eng. 2012;24(5):823–39.
13. Akoglu L, Tong H, Koutra D. Graph-based anomaly detection and description: a survey. Data Min Knowl Disc. 2014; doi:10.1007/s10618-014-0365-y.
14. Kriegel HP, Kröger P, Schubert E, Zimek A. Interpreting and unifying outlier scores. In: Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa; 2011. p. 13–24.
15. Achtert E, Kriegel HP, Schubert E, Zimek A. Interactive data mining with 3D-parallel-coordinate-trees. In: Proceedings of the ACM International Conference on Management of Data (SIGMOD), New York; 2013. p. 1009–12.