LoOP: Local Outlier Probabilities

Hans-Peter Kriegel, Peer Kröger, Erich Schubert, Arthur Zimek
Institut für Informatik, Ludwig-Maximilians Universität München
Oettingenstr. 67, 80538 München, Germany
http://www.dbs.ifi.lmu.de
{kriegel,kroegerp,schube,zimek}@dbs.ifi.lmu.de

ABSTRACT

Many outlier detection methods do not merely provide the decision for a single data object being or not being an outlier but give also an outlier score or "outlier factor" signaling "how much" the respective data object is an outlier. A major problem for any user not very acquainted with the outlier detection method in question is how to interpret this "factor" in order to decide for the numeric score again whether or not the data object indeed is an outlier. Here, we formulate a local density based outlier detection method providing an outlier "score" in the range of $[0, 1]$ that is directly interpretable as a probability of a data object for being an outlier.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database applications—Data mining

General Terms: Algorithms

Keywords: Outlier Detection

1. INTRODUCTION

The problem of identifying outliers has been addressed by different approaches that can be roughly classified as global versus local outlier models. This distinction refers to the scope of a database being considered when a method decides on the "outlierness" of a given object. While some methods take always the complete database into account, others consider only a local selection of database objects, e.g., the ε-neighborhood or the k nearest neighbors of a point. At a different axis, one can distinguish "labeling" versus "scoring" outlier detection methods. The former are leading to a binary decision of whether or not a given object is an outlier while the latter are rather assigning a degree of "outlierness" to each object. Such an "outlier factor" is a value characterizing each object in "how much" this object is an outlier. In the literature, in most cases global outlier detection schemas produce binary labels whereas local outlier approaches often assign scores to the database objects. However, this relation does not appear to be necessary. For a more detailed discussion of outlier detection methods and their relations to and differences between each other, see [5]. Here, we are primarily interested in the nature and meaning of outlier scores provided by "scoring" outlier detection methods because these approaches are usually more flexible since they can also produce labels. In addition, we focus on a local scope of the database to decide about the outlier score. Indeed, the scores provided by different methods differ widely in their scale, their range, and their meaning. In some cases, high values of an outlier score mean, the corresponding database object is *not at all* an outlier, in other cases, a higher value indicates more "outlierness". For many methods, the scaling of occurring values of the outlier score even differ within the same method from data set to data set, i.e., outlier score x in one data set means, we have an outlier, in another data set it is not extraordinary at all. In many cases, even within one and the same data set, the identical outlier score x for two different database objects can denote substantially different degrees of outlierness, depending on different local data distributions around the two objects.

Here, we propose a scoring that includes a normalization to become independent from the specific data distribution in a given data set as well as a statistically sound motivation for a mapping into the range of $[0, 1]$ readily interpretable as the probability of a given database object for being an outlier. In the remainder, we will introduce this new local scoring scheme LoOP in Section 2. The behavior of LoOP in comparison to existing approaches is evaluated in Section 3. Section 4 concludes the paper.

2. FORMAL DEFINITION OF LOCAL OUTLIER PROBABILITY

In this paper, we introduce a new outlier model that combines the idea of local, density-based outlier scores like LOF [4], its variants, and LOCI [8] with probabilistic concepts to model the "outlierness" of a point. A probabilistic approach means to offer a natural tolerance to noise effects in the data. Traditional approaches such as LOF and LOCI may even emphasize such effects. For example, LOF is based on comparing the k-distances of points, i.e., the distances of points to their respective kth nearest neighbor. A (locally) inappropriate choice of k can cause instable results.

In the following, we assume $\mathcal{D}$ being a set of n objects and d being a distance function used to distinguish outliers. To become more reliable, we introduce the *probabilistic distance* of $o \in \mathcal{D}$ to a context set $S \subseteq \mathcal{D}$, referred to as $\text{pdist}(o, S)$.

CIKM'09, November 2–6, 2009, Hong Kong, China.

This distance has the following property:

$$\forall s \in S: \quad \mathcal{P}[d(o,s) \leq \text{pdist}(o,S)] \geq \varphi.$$

Intuitively, a sphere around o with radius pdist covers any element in the context set S with a probability of φ. The probabilistic distance $\text{pdist}(o,S)$ of o to S can be interpreted as the *statistical extent* of the context set S. The main difference to the (normal) extent of a set of points is that the statistical extent deliberately allows for some error. The reciprocal of the probabilistic distance can be seen as an estimation for the density of S, i.e.,

$$\text{pdens}(S) = \frac{1}{\text{pdist}(o,S)},$$

and, thus, for the local context of o in terms of S. If we use $\lambda = \sqrt{2} \cdot \text{erf}^{-1}(\varphi)$ instead of φ, where erf denotes the Gaussian error function, in the estimation of the density of S, we can simulate the classical and sound statistical notion of outliers defined as objects that deviate more than a given λ times the standard deviation σ from the mean. Values of λ are those of the empirical 68-95-99.7 ("three sigma") rule, e.g. $\lambda = 1 \Leftrightarrow \varphi \approx 68\%$, $\lambda = 2 \Leftrightarrow \varphi \approx 95\%$, and $\lambda = 3 \Leftrightarrow \varphi \approx 99.7\%$.

Assuming that o is the center of S and the set of distances of $s \in S$ to o is approximately half-Gaussian, one can compute a *standard distance* of the object in S to o similar to the standard deviation:

$$\sigma(o,S) = \sqrt{\frac{\sum_{s \in S} d(o,s)^2}{|S|}}$$

Note that this is subtly different from $\text{Stddev}(d(o,S))$, since it uses a mean of $d(o,o) = 0$ instead of $E[d(o,S)]$. In particular, the difference is that we cannot assume the distance values to be normally distributed, but instead we assume S to be approximately normally distributed around o. When we determine S for a given object o this assumption should be taken into account. As a consequence, we propose to obtain the context set S via a k nearest neighbor query around o. As such, the assumption of S being centered around o is usually reasonable. Based on these considerations, we define the *probabilistic set distance* of o to S with significance λ as

$$\text{pdist}(\lambda, o, S) := \lambda \cdot \sigma(o,S).$$

Intuitively, this probabilistic set distance estimates the density around o based on S. The parameter λ gives control over the approximation of the density. It is, however, just a normalization factor solely affecting contrast in the resulting scores. The ranking of outliers will not be affected by λ.

Naturally any statistical modeling is based on certain assumptions. The local probabilistic modeling of outliers is based on the assumptions (i.) that the context set S is centered around the query object o, and (ii.) that the distances behave like the positive leg of a normal distribution. Regarding assumption (i.), we admit that not every space will offer the option to compute a centroid c_S for the set S to be used instead of o in order to compute the standard distance and the probabilistic set distance, so that this would not be a generally viable solution. However, we need to investigate what happens if o differs substantially from c_S and in which situations this can occur. By definition of the centroid, $E[d(c_S,s)] \leq E[d(o,s)]$, we will usually obtain higher distances and thus a higher value for σ compared to the case when using the centroid c_S of S. This situation occurs if the context set S is asymmetric to o, especially if o is an outlier point w.r.t. S. Thus, by using o as context center, we overestimate σ in particular for outlier points. This however is a beneficial effect, since it will only increase this points outlier score. Assumption (ii.) is not directly assuming a certain distribution of the data. Rather, we make an assumption only about the distribution of the distances from the point o. Intuitively, this means that the full-dimensional vector space is projected onto a one-dimensional subspace of distances. To this projection, the central limit theorem can be applied which states the following: Assuming that there are sufficiently many independent components involved in computing the distances, these distances behave approximately normally distributed. Note that this will not necessarily hold for dimension-selecting distances; it does however apply to L_p-norms such as Euclidean or Manhattan distances. However, based on this theorem, we can assume that the distances are normally distributed without limiting our approach to any fixed type of distribution. In other words, our method works with any kind of data distribution. Thus, it combines the advantages of two worlds, the density-based approaches that do not assume any specific data distribution and the sound mathematical concepts of statistical methods.

Based on this reasoning for estimating the density around an object w.r.t. a context set, the *Probabilistic Local Outlier Factor (PLOF)* of an object $o \in \mathcal{D}$ w.r.t. a significance λ, a context set $S(o) \subseteq \mathcal{D}$, can be defined as follows:

$$\text{PLOF}_{\lambda,S}(o) := \frac{\text{pdist}(\lambda, o, S(o))}{E_{s \in S(o)}[\text{pdist}(\lambda, s, S(s))]} - 1.$$

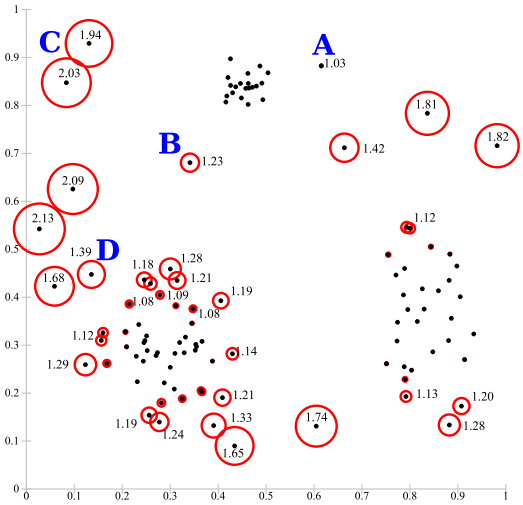
The PLOF value of an object $o \in \mathcal{D}$ calculates the ratio of the estimation for the density around o which is based on $S(o)$ and the expected value of the estimations for the densities around all objects in the context set $S(o)$. Let us note that the resulting value—which scales similar to the LOF score (minus 1)—is not yet a probability and not normalized. A value ≤ 0 is not an outlier, while higher values indicate an increasing outlierness. Similar to existing outlier models, these values still cannot be easily compared between data sets. To achieve a normalization making the scaling of PLOF independent of the particular data distribution, the aggregate value nPLOF is obtained during PLOF computation.

$$\text{nPLOF} := \lambda \cdot \sqrt{E[(\text{PLOF})^2]}$$

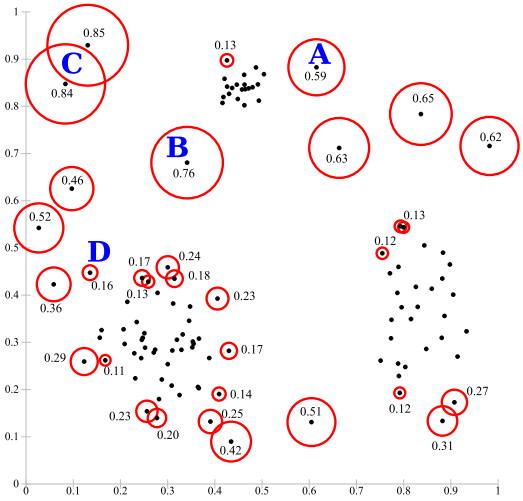
This value can be seen as a kind of standard deviation of PLOF values, i.e., $\lambda \cdot \text{Stddev}(\text{PLOF})$ assuming a mean of 0. To convert the not yet normalized PLOF value into a probability value, we assume that the values are normally distributed around 1 with a standard deviation of nPLOF. We then apply the Gaussian Error Function to obtain a probability value, the *Local Outlier Probability (LoOP)*, indicating the probability that a point $o \in \mathcal{D}$ is an outlier:

$$\text{LoOP}_S(o) := \max\left\{0, \text{erf}\left(\frac{\text{PLOF}_{\lambda,S}(o)}{\text{nPLOF} \cdot \sqrt{2}}\right)\right\}$$

The LoOP value will be close to 0 for points within dense regions and close to 1 for density based outliers. Hence, while traditional local density based outlier scores are not readily comparable with each other even within one single data set we have obtained means to directly derive the probability of a database object of any given data set for being a density



(a) LOF values.



(b) LoOP values.

Figure 1: Comparison of the interpretability of local, density-based outlier scoring (here: LOF) values and LoOP values on 2D synthetic data. Both algorithms were run with $k = 20$, for LoOP $\lambda = 3$.

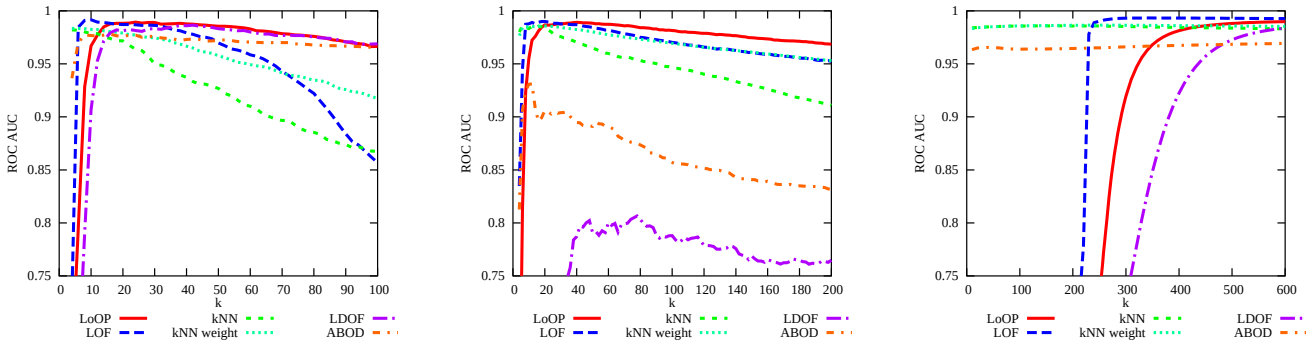
based local outlier. In other words, while an outlier score of x computed by any local approach like LOF will designate different degrees of outlierness for different regions of the same data set and cannot be compared to a score of x on a different data set, LoOP values are consistent over the complete data set and over multiple data sets.

Figure 1 illustrates some differences between LOF and LoOP on a synthetic 2D data set. In the respective figures, the interesting outlier scores are given. Additionally, LOF values are converted to a circle radius with non-trivial scaling, while LoOP values can be directly used as circle radius. For example, object A (obviously an outlier) is not recognized by LOF due to an unlucky choice of $k = 20$: the cluster left of object A together with object A makes 21 objects, and the 20-nearest-neighbor distance of any point in the cluster will be the distance to A. Thus, the difference in their k-distance is small and A gets a low LOF score. LoOP assigns an outlier probability of 59% to this object: due to the averaging effects, the pdist for cluster points just covers the cluster itself, and object A gets a significantly higher pdist value. The situation with object B is similar. Although LOF values below 3 are usually considered "insignificant", even clear outliers such as the objects at C are just at approximately 2. LoOP assigns them an outlier probability of about 85%. Object D highlights another weakness of LOF: it is designed for clusters of uniform density. The cluster around D shows a Gaussian distribution. LOF assigns object D an outlier score higher than e.g. object B while this point was in fact generated by the cluster it is adjacent to. When modeling the data set using three Gaussian distributions and uniform background noise, the probability of object D being generated by the cluster is higher than that of the noise distribution. The LoOP value of 16% is much more useful here: there is a clear chance the the point is an outlier, but it is also very likely it is just an outer point of the clusters normal distribution.

3. EXPERIMENTS

We compare the accuracy of our novel LoOP model with several competitive algorithms: We used the LOF [4] and one of its latest variants LDOF [10] as representative algorithms for local, density-based outlier models. We also used the angle-based ABOD [6] as recent non-density-based proposal. As a baseline, we used k-NN based outlier detection as defined in [9] (distance to k-th nearest neighbor, or "kNN" in short) and [2] (sum of the distances to the k nearest neighbors, or "kNN weight" in short). Since we are not interested in efficiency but in the recall and precision of the methods, we did not use the efficient since approximate solutions for "kNN" and "kNN weight" as proposed along with the model in the corresponding papers but the exact implementations of the corresponding outlier-models. All competitors have been implemented in the unified framework ELKI [1]. We chose $\lambda = 3$ for LoOP throughout all experiments that are reported here. However, results for different λ values ($\lambda = 1$ and $\lambda = 2$) are identical because they give the same ranking. This confirms our statement above, that LoOP is robust against the choice of λ.

We used three real-world data sets known from classification and prepared them for unsupervised outlier detection by sampling one of the classes to become sparse, and using this class as outliers. The first data set is the "Wisconsin Breast Cancer" diagnosis set [3], which consists of 357 "benign" and 212 "malignant" medical diagnosis records (31 dimensions). We removed the malignant records except for the first 10 records, which we consider outliers. Thus, the data set consists of 367 records. The second data set is called "Pen-Based Recognition of Handwritten Digits" training set [3], which consists of 7494 records, 719 to 780 for each of the classes (which correspond to the digits 0 to 9). The dimensionality is 16, each dimension resembling a pixel value in a 4x4 grid. We chose the digit 4 to be our outlier class, and again only kept the first 10 records of this digit, resulting in a data set size of 6724 records. A third set consists of metabolic data records [7] measuring the concentration of 43 metabolites in the blood of newborns. The largest share in this data set is a control group with 19,730 instances. We removed all atypical records except for "Phenylketonuria" entries. These 306 records were kept as outliers in the data set. The final data set contains 20,036 records with 43 measurements each.



(a) Wisconsin Breast Cancer data set. (b) Pen Digits data set. (c) Metabolic data set.

Figure 2: Accuracy of different outlier models on different data sets.

Using the outlier scores as a ranking criterion, ROC curves (receiver operating characteristic) are the means of choice to compare the performance of different methods. A ROC curve compares the true positive rate with the false positive rate. Given a ranking of the objects according to their outlier score, a perfect outlier detection method would first return all outliers followed by all remaining objects. A single ROC curve can then be summarized by computing the area under curve (AUC). An AUC value of 1.0 means a perfect separation because first the true positive rate goes to 1.0 and then the false positive rate increases. An AUC value of 0.5 corresponds to random sampling, i.e., the true positive rate and the false positive rate both increase simultaneously. An AUC value significantly below 0.5 indicates that the method prefers non-outliers. Finally, an AUC value of 0.0 is a perfect separation again, although with non-outliers coming first. Since all competitors rely on the parameter k specifying the scope of the database, we will focus on this parameter. For that purpose, we use a two-dimensional plot comparing AUC and k for a wide range of values in k.

Figure 2 shows the results for different algorithms on different data sets. On the Wisconsin Breast Cancer (cf. Figure 2(a)) all algorithms receive high scores. For a well-chosen value of k, LOF performs best. However, it is very sensitive to the choice of k. For a large range of values, LoOP performs best (and close to the top result of LOF). Both the kNN and the kNN weight outlier detection approaches drop quickly on this data set. ABOD is also quite stable, but at a lower level than LoOP or LDOF. Results on the Pen Digits data set are shown in Figure 2(b). Again LOF is performing very well. However, LoOP is slightly better and in particular much more stable in its results. Both the kNN and the kNN weight models also work very well, while ABOD and LDOF struggle for this data set. A slightly different observation can be made from the results on the Metabolic data set in Figure 2(c). Here the two kNN variants and ABOD perform consistently good. The local, density-based methods produce good or even better results but only for large values of k. Anyway, for high values of k LOF, LDOF and LoOP outperform the other approaches.

In summary, LoOP performs good on all the used real-world data sets. In most cases, it achieves the best results for a specific choice of k and shows to be more robust against this choice than the competitors.

4. CONCLUSION

The state-of-the-art approaches for unsupervised outlier detection usually rely on computing an outlier score for each database object based on a local scope of the database as a reference. However, for different approaches, the calculated scores are not standardized and often hard to interpret. Thus, scores of objects from different data sets and even scores of objects from the same data set can not be compared. In this paper, we propose the novel LoOP (Local Outlier Probability) outlier detection model that combines the idea of local, density-based outlier scoring with a probabilistic, statistically-oriented approach. The benefit of our model is that it provides for each data object an outlier probability as score that is easily interpretable and can be compared over one data set and even over different data sets. Our experimental evaluation confirms the competitive behavior of LoOP on several synthetic and real-world data sets.

5. REFERENCES

[1] E. Achtert, T. Bernecker, H.-P. Kriegel, E. Schubert, and A. Zimek. ELKI in time: ELKI 0.2 for the performance evaluation of distance measures for time series. In *Proc. SSTD*, 2009.

[2] F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In *Proc. PKDD*, 2002.

[3] A. Asuncion and D. J. Newman. UCI Machine Learning Repository. `http://www.ics.uci.edu/~mlearn/MLRepository.html`, 2007.

[4] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proc. SIGMOD*, 2000.

[5] H.-P. Kriegel, P. Kröger, and A. Zimek. Outlier detection techniques. Tutorial at PAKDD, 2009.

[6] H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proc. KDD*, 2008.

[7] B. Liebl, U. Nennstiel-Ratzel, R. von Kries, R. Fingerhut, B. Olgemöller, A. Zapf, and A. A. Roscher. Very high compliance in an expanded MS-MS-based newborn screening program despite written parental consent. *Preventive Medicine*, 34(2):127–131, 2002.

[8] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos. LOCI: Fast outlier detection using the local correlation integral. In *Proc. ICDE*, 2003.

[9] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proc. SIGMOD*, 2000.

[10] K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. In *Proc. PAKDD*, 2009.