LOF(local outlier factor): Identifying Density-Based Local Outliers

JAIMIN PRASHANTKUMAR OZA*,

TU-Dortmund,

Germany

Outlier Detection can be more useful in finding anomalies in a given dataset rather than some clustering or pattern finding algorithms. Here we would dive deep into the LOF (Local Outlier Factor) algorithm. Local Outlier Factor helps us understand how outlying the object is regarding its neighboring data points. We would discuss the pros and cons of the LOF algorithm and compare it with other outlier detection methods like isolation forest etc.

Additional Key Words and Phrases: Data Mining, Outlier Detection, Anomaly Detection, text tagging

ACM Reference Format:

Jaimin PrashantKumar Oza. 2022. LOF(local outlier factor): Identifying Density-Based Local Outliers. 1, 1 (February 2022), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Outliers are the points that are far-off from other data points in the dataset and display different behavior from the rest of the data. Most methods consider outliers to be noise elements. We should focus on occasional or unfamiliar events rather than the frequent ones for outlier detection. Outlier detection approaches help us investigate various fraud detection scenarios in Banks, Internet, etc. In contrast, clustering techniques look for the common attributes in input dataset. The outlier which are not driven in terms of global distribution in the dataset are also known as local outliers. Local outlier articulations are a variation of global outlier articulations, in which global outliers are generally also local outliers, but not the other way around.The computational cost of local outlier strategies is larger than that of global techniques[14]. Being an outlier is considered as binary property by various algorithms. Here we discuss the Local outlier factor method that helps identify local outliers by displaying how distant the data point is from other surrounding entities. It assigns a LOF score to each and every object in dataset.This LOF score represents degree of outlier-ness of the objects. This approach identifies local outliers as it considers only restricted neighbors, which will be explained in detail in the next section. I will explain LOF in detail along with its pros and cons in the next section. I will provide an example and also compare LOF with another outlier detection method to show the use of the LOF method.

[*] Matriculation Number:229984

Author's address: Jaimin PrashantKumar Oza, jaimin.oza@tu-dortmund.de,
TU-Dortmund,
Dortmund, North Rhine-Westphalia,
Germany, 44227.

2 LOF DEFINITION AND PROPERTIES

The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection technique.It evaluates a data point's local density deviation from its neighbors to find local outliers from the given dataset. The LOF method is implemented using K-nearest neighbour algorithm for identifying local density and such density is determined by the distance between data points. It helps to find areas with similar densities and places with significantly lower densities by comparing an object's local density to the local densities of its neighbors. Outliers are those that deviate from the norm.Tracing Local outliers can be difficult as density might vary inside a dataset. Because some clusters are dense and others are sparse, determining whether a point is a local outlier or not becomes challenging. The LOF approach overcomes these issues and accurately detects local outliers within a dataset.The notions of "core distance" and "reachability distance," which are employed for local density estimation in LOF, are similar to those used in DBSCAN -[5],CLIQUE -[6] and OPTICS -[7].

2.1 Working of LOF

As already seen above the k-nearest neighbor method is used to determine local density between data points that are neighbors. The local density for each data point is calculated separately. We can see which data points have similar densities and which have densities that are lower than their neighbors by comparing their local densities. Outliers are those with low density. To begin, k-distances are measured for each data point to discover their k-nearest neighbors. Nk(A) denotes a set of points that lie within or on the circle of radius K-distance. K-neighbors can be greater than or equal to K. The reachability distance is further computed using this k-distance. It's the maximal of the distance between two points and that point's k-distance.

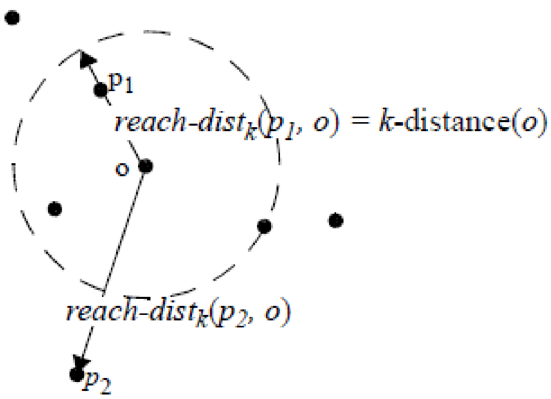
p_1
$reach\text{-}dist_k(p_1, o) = k\text{-distance}(o)$
o
$reach\text{-}dist_k(p_2, o)$
p_2

Fig. 1. Reachability distance

As shown in[Figure 1][1] The reachability distance is K-distance if a point is within the K-neighborhood; otherwise, it is the distance between the two points. Now, in order to estimate the local reachability density(LRD), reachability distances needs to be determined for all the nearest neighbors. The LRD is the inverse of the sum of all the reachability distances of all the k-nearest surrounding points. It is a measure of the density of k-nearest points around a point[16].

The equation uses inverse because the closer the points are, the smaller the distance and the greater the density. Steps to compute the Local Reachability Density of a data point:

- Determine the reachability distance from all of its k nearest neighbors.
- Average that number.
- Local Reachability Density is inverse of that average calculated in previous step.

The local reachability density indicates how far we must travel from our current location to the next point (or cluster of points). We'll have to travel longer if Local Reachability Density is lower since it's less dense. To compute LOF[1], the Local Reachability Density of each point is compared to the average Local Reachability Density of its K neighbors. The Local Reachability Density of a point is divided by the average Local Reachability Density of the point's K neighbors to get the LOF.

$$LOF_{Minpts^{(p)}} = \frac{\sum_{\mathbf{o} \in N_{Minpts^{(p)}}} \frac{lrd_{Minpts^{(p)}}}{lrd_{Minpts^{(o)}}}}{|N_{Minpts^{(p)}}|} \tag{1}$$

If the point isn't an outlier (inlier), the ratio of average Local Reachability Density of neighbors approximately near to the Local Reachability Density of the point. (Here, in this case, LOF is 1). Outlier can be spotted, if the average of LRD of the neighbors is greater than LRD of a point. The LOF value will thereafter be very high. In general, if the LOF is more than one, the point is deemed an outlier, however, this is not always the case.

2.2 LOF Properties

In this section below, we'll go over some key LOF characteristics that enable us accurately locate local outliers[1].

- Objects deep within the cluster have a LOF score of around 1 and cannot be called outliers.This property shows that for objects inside cluster with low or high density the LOF score will be close to 1,thus they cannot be considered as outliers.
- A point's LOF is just a function of the distances between its direct and indirect neighbors.The direct/indirect ratio has a bearing on the LOF spread.The LOF spread is affected by the direct/indirect ratio. Reachability distances are used to calculate the reach-dist-min and reach-dist-max values. This minimum and maximum are solely determined by the objects in the MinPts-nearest neighborhoods. As a result, the LOF has tighter borders.
- The relative deviation of the LOF is determined solely by the ratios of the underlying reachability distances in direct and indirect neighborhoods, not by their absolute values.If reachability distances in direct and indirect neighborhoods change by the same amount, the relative fluctuation of the LOF remains constant.
- The LOF-min and LOF-max values may differ considerably if an object's MinPts-nearest neighbors are from different clusters with different densities. This property provides better boundaries on the LOF when an object's MinPts-nearest neighborhood coincides with multiple clusters. Each of the groups contributes significantly to LOF point while MinPts- nearest neighbors are aggregated into several groups.
- MinPtsLB can be thought of as the least number of objects that a "cluster" must have for other objects to be local outliers relative to this cluster. The maximum number of "nearby" objects that could be local outliers is MinPtsUB.These are the key guidelines for the MinPts selection.

When the data points in a cluster are homogeneously distributed, the LOF values for the objects in the cluster are frequently near to 1. The LOF values of data points in clusters is close to 1 even if they consist of different densities. Two outliers, named ’A’ and ’B,’ are seen in the [Figure 2][15]. Furthermore, there are two clusters in the image, one of which is significantly sparser than the other. It is very difficult for distance based approaches to identify the outlier ’A’ unless it considers the smaller distance threshold value. If a smaller distance criterion is employed, numerous data points in the sparser cluster may be mistakenly classified as outliers.LOF algorithm works perfectly here in identifying Both outliers ’A’ and ’B’ as LOF scores calculated for both points would be greater than 1.Here the Local reachability density for points in both clusters would be different and hence LOF scores for points inside the clusters would be around 1.

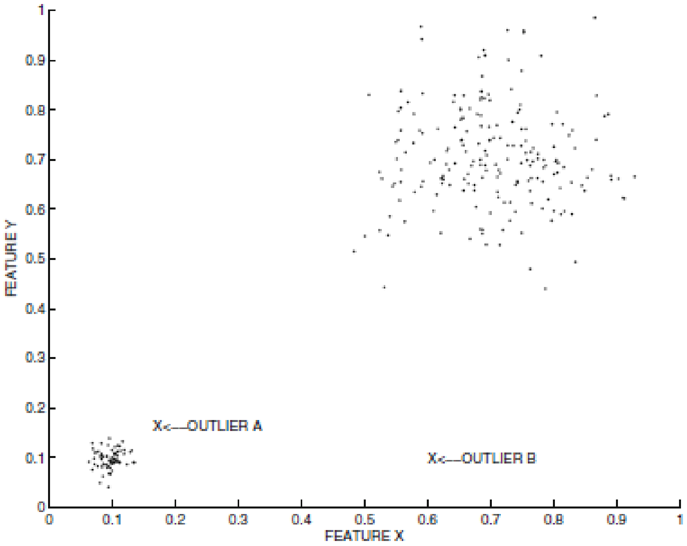

Fig. 2. LOF Example

In contrast, LOF values are much larger since LOF values of both outlying points are calculated as ratios to average neighbor reachability distances. The normalization factor is mean of the locality’s reachability distances, therefore the LOF scores can simply be considered as a point’s normalized reachability distance[15]. In the literature, the LOF method is known as a density-based approach, although it can also be thought of as a relative distance-based approach with smoothing[1]. The correlated distances is being calculated based on the local distribution of reachability distance. The LOF method was first offered as a density-based strategy because of its ability to respond to regions of varying density. The density is calculated as the inverse of a point’s smoothed reachability distances in this methodology.

3 LOF ADVANTAGES AND DISADVANTAGES

Benifits of Local Outlier factor algorithm:

- Identifying outliers might be difficult at times. If a point is close enough to the highly dense cluster, it will be classified as an outlier. That point might not be seen as an outlier by the global approach. The LOF, on the other hand, is capable of detecting outliers in the immediate vicinity[16].
- The LOF approach can be used to tackle problems of finding outliers in a variety of disciplines, including Fraud detection, Crime investigation,sports data,geographic data and so on[1].
- It also outperforms a variety of other anomaly detection techniques[16].

- While the LOF algorithm's geometric insight is restricted to low-dimensional vector spaces, it can be used in any scenario where a dissimilarity function can be specified. It has been found to work very well in a variety of settings, often beating competitors, such as network intrusion detection[19].

Drawbacks of Local Outlier factor algorithm:

- The LOF score used to evaluate whether or not a point is an outlier varies. It could be different for different data sets.
- The LOF algorithm's detection accuracy suffers as the dimensions increase.Because the computations are memory-intensive, scalability is a major concern, and larger datasets demand a powerful computing infrastructure[18].Apart from this, at Higher dimensions data points become equidistant from each other.So, finding correct local ouliers becomes difficult as the Local Reachability Density is almost equal and LOF score become near to 1.
- Since the LOF score might be any number that the ratio generates, it can be difficult to discern inliers and outliers based on it as it depends on the density of neighborhood.
- It's difficult to interpret LOF because it's a ratio. An outlier is defined as a point that exceeds a certain threshold value. The problem and the user both determine how an the point is interpreted as outlier[17].
- Duplicate points might cause problems in calculating LOF scores since all points in close proximity of duplicate points risk having their scores set to infinity[15].

4 LOF EXTENTIONS AND ALTERNATIVES

There are many extentions applied on LOF to make it more useable for large dataset and streaming dataset[19].

- Feature bagging for outlier detection- It basically applies bagging on the dataset to randomly select and reduce the dimensionality of the high dimension dataset. It applies the LOF algorithm on each random subset of the dataset thus obtaining multiple outliers and assigning each of them with a score and combining the result. Thus, for increased detection capabilities in high dimensions, run LOF on numerous projections and integrate the results[2].
- LoOP: Local Outlier Probabilities- This method directly provides the probability score between [0:1] to a point for being an outlier. It also introduces the probabilistic distances rather than K-distances. The probabilistic distance's reciprocal can be thought of as density estimation. It computes the Probabilistic Local Outlier Factor with a value between [0:1][3].
- Interpreting and unifying outlier scores- It is an approach to unifying outlier models through regularization and normalization, with goals of increasing the contrast between outlier and inlier scores and generating a rough probability value for being an outlier or not being an outlier[10].
- Outlier Detection for High Dimensional Data- This method works by detecting lower-dimensional projections that are locally scattered and are tough to find using brute force techniques due to the large number of available options. Simple distance-based outliers are unable to overcome the dimensionality curse's consequences[11]. Hence the earlier discussed technique is more efficient than simple distance based outliers.
- Isolation Forest- It provides an intuition of anomalies as data points that are "rare and different". In this method, on the basis of randomly picked features random sub-sampled data is processed in a tree structure. The samples that reach deeper into the tree require more cuts to isolate them, thus they are less likely to be abnormalities. Similarly, samples that end up on shorter branches represent anomalies since the tree found it easier to distinguish

them from other data.Isolation forest approach makes use of feature bagging for outlier detection and hence has capabilities of detecting outliers at higher dimension[9].

5 SUMMARY

We have already discussed the benefits of outlier detection in various fields and the role of LOF in finding local outliers.The local outlier factor approach aids in the detection of local outliers by indicating each point's degree of isolation from its neighbors. The LOF method has been proven to be more accurate in detecting anomalous points at lower dimensioins[13].LOF is an unsupervised outlier detection method that helps in differentiating between inliers and local outliers by calculating LOF scores for each point. We discussed the in-depth working of the LOF method from calculation of K-distance to reachability distance to density calculation and finally the LOF score. We have seen how the density of points is calculated related to the density of nearby data points which helps in calculating the Local Outlier Factor (degree of being outlier) of that point. We have seen the important properties of the Local outlier factor and how the LOF score varies in different scenarios for points deep inside a cluster and local outliers. The properties of LOF helped us understand how the LOF score is dependent on the ratio of direct and indirect reachability distances rather than actual values. It helped us in understanding how LOF-min and LOF-max values are dependent on the point's MinPts-nearest neighbors. We have seen how the MinPtsLB and MinPtsUB can be selected for the computation of LOF scores. As previously mentioned, because it accurately detects local outliers, LOF may be applied in a variety of disciplines such as fraud detection,criminal investigation etc. Even if there are multiple local clusters with varied densities, the LOF perfectly recognizes local outliers.. We know now that LoF struggles at Higher dimensions due to computational complexity and generalization problem at a higher level. We have also looked at many extensions of LOF like Feature bagging for outlier detection etc which help in outlier detection at a higher dimension. We can also use Local Outlier Probabilities which converts LOF scores into probabilities which could help normal users understand the point is an outlier or not. To summarise, we can say that LOF is accurate in computing the degree of outlying of data points at lower dimensions, and extensions like Feature bagging, Isolation Forest, etc. for outlier detection can be used to detect outliers at higher dimensions.

REFERENCES

[1] LOF: Identifying Density-Based Local Outliers ,Markus M. Breunig and Hans-Peter Kriegel and Raymond T. Ng and Jörg Sander (2000).

[2] Feature Bagging for Outlier Detection,Aleksandar Lazarevic and Vipin Kumar(2005).

[3] LoOP: Local Outlier Probabilities,Hans-Peter Kriegel and Peer Kröger and Erich Schubert and Arthur Zimek(2009).

[4] A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams,Omar Alghushair and Raed Alsini and Terence Soule and Xiaogang Ma(2020).

[5] A density-based algorithm for discovering clusters in large spatial databases with noise ,Martin Ester and Hans-Peter Kriegel and Jörg Sander and Xiaowei Xu(1996).

[6] Automatic subspace clustering of high dimensional data for data mining applications ,Rakesh Agrawal and Johannes Gehrke and Dimitrios Gunopulos and Prabhakar Raghavan(1998).

[7] OPTICS: Ordering Points To Identify the Clustering Structure ,Ankerst, Mihael and Breunig, Markus M. and Kriegel, Hans-Peter and Sander(1999).

[8] Identification of Outliers,Hawkins, D Chapman and Hall, London (1980).

[9] Isolation Forest ,Fei Tony Liu and Kai Ming Ting and Zhi-Hua Zhou(2009).

[10] Interpreting and Unifying Outlier Scores ,Hans-Peter Kriegel and Peer Kr¨oger and Erich Schubert and Arthur Zimek(2011),https://doi.org/10.1137/1.9781611972818.2.

[11] Outlier Detection for High Dimensional Data ,Charu C. Aggarwal and Philip S. Yu(2002),10.1145/376284.375668.

[12] Advancements of Outlier Detection: A Survey,Ji Zhang(2013),10.4108/trans.sis.2013.01-03.

[13] A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection,Aleksandar Lazarevic, Levent Ertöz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava(2003).

[14] Finding Local Anomalies in Very High Dimensional Space,Timothy de Vries, Sanjay Chawla and Michael E. Houle(2010).

[15] Outlier Analysis,Charu C. Aggarwal,978-3-319-47577-6.

[16] Local outlier factor. GeeksforGeeks. (2020, September 5). Retrieved February 6, 2022, from https://www.geeksforgeeks.org/local-outlier-factor/.

[17] Local outlier factor(LOF)-algorithm for Outlier identification. Retrieved February 6, 2022, from https://towardsdatascience.com/local-outlier-factor-lof-algorithm-for-outlier-identification-8efb887d9843.

[18] Anomaly/outlier detection using local outlier factors. Data Science Central. Retrieved February 6, 2022, from https://www.datasciencecentral.com/anamoly-outlier-detection-using-local-outlier-factors/.

[19] Wikimedia Foundation,(2021, July 12), Local outlier factor,Wikipedia, from https://en.wikipedia.org/wiki/Local_outlier_factor.