LOF: Identifying Density-Based Local Outliers

M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, ACM SIGMOD Int. Conf. on Management of Data, 2000.

Outline

- Background - Outlier Definitions
- Develop LOF
  - k-nearest neighborhood
  - Reachability distance
  - …
  - LOF
- Properties of LOF
  - LOF Estimation
  - Bounds
- Experimental Results
  - Soccer data
  - Hockey Data
  - Larger Datasets
- Advantage and Disadvantages
- Conclusion

Background

- Outliers in Clustering
  - Often regarded as a confounding factor
  - Generally discarded as noise
- Outliers may be interesting!
  - Fraud detection
  - Intrusion detection

Background

- Pre-existing outlier definitions
  - Largely based on statistical models
  - Distribution-based [1]
  - Depth-based [2]
  - Distance-based [3]
- Some Examples:
  - Hawkins-Outlier [4]
  - DB(pct, dmin)-Outlier
- Problems
  - Outlier as binary property
  - Dependent on fitting data to distribution
  - Global outliers

[1] Barnett, V., Lewis, T. *Outliers in statistical data*, John Wiley, 1994.
[4] Hawkins, D. *Identification of Outliers*, Chapman and Hall, London, 1980.
[3] Knorr, E.M., Ng R.T., *Algorithms for Mining Distance-Based Outlier in Large Datasets*. Proc. 24th Int. Conf. on VLDB, NY, NY, 1998.
[2] Tukey, J.W., *Exploratory Data Analysis*, Addison-Wesley, 1977.

Develop LOF

- To develop an LOF, we will have to define a "local outlier"
- Will need a number of definitions:
  - K-distance of an object p
  - K-distance neighborhood of p
  - Reachability distance
  - Local reachability density
  - Finally . . . LOF!

Develop LOF

- To develop an LOF, we will have to define a "local outlier"
- Will need a number of definitions:
  - K-distance of an object p
  - K-distance neighborhood of p
  - Reachability distance
  - Local reachability density
  - Finally . . . LOF!

For any integer k >0, k-distance(p) is the distance d(p,o) between p and o (an object in D) such that:

There are at least k other objects (o') in D | d(p, o') <= d(p, o)
and
There are at most k -1 objects (o') in D | d(p, o') < d(p, o)

So: The distance to the farthest of the k objects nearest to p.
k-distance neighborhood is composed of points within k-distance

Develop LOF

- To develop an LOF, we will have to define a "local outlier"
- Will need a number of definitions:
  - K-distance of an object p
  - K-distance neighborhood of p
  - Reachability distance
  - Local reachability density
  - Finally . . . LOF!

Reachability distance helps to smooth statistical fluctuations

reach-distk(p, o) = max{k-distance(o), d(p, o)}

lrd of an object p is: inverse of average reachability distance on MinPts nearest neighbors of p

Develop LOF

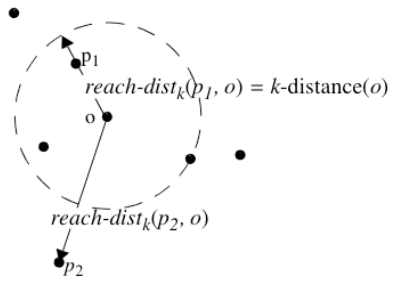


Figure 2: *reach-dist*(p_1,o) and *reach-dist*(p_2,o), for k=4

reach-distk(p, o) = max{k-distance(o), d(p, o)}

Develop LOF

- To develop an LOF, we will have to define a "local outlier"
- Will need a number of definitions:
  - K-distance of an object p
  - K-distance neighborhood of p
  - Reachability distance
  - Local reachability density
  - Finally . . . LOF!

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

LOF is: average of ratio of (lrd of p) : (lrd of p's MinPts-nearest neighbors)
Higher LOF -> more of an outlier

Intuitively: if density around p is much lower then around p's neighbors, p must be an outlier!

Properties of LOF

- LOF approximation for items "deep" in clusters
- LOF bounds for all items
  - Bounds for items with all neighbors in one cluster
  - Bounds for other items
  - Tighness analyses
- Impact of MinPts

Approximation of LOF

LOF for items in a cluster is approx. 1

Lemma 1: C is a collection of objects, reach-dist-min is minimum reachability distance of objects in C, and reach-dist-max is max reachability distance of objects in C.

Then: define ε = (reach-distmax/reach-dist-min) -1

So: for all objects p in C such that all neighbors and 2nd-degree neighbors are also in C

It is true that: $1/(1+\varepsilon) <= LOF(p) <= (1+\varepsilon)$

Approximation of LOF

Intuitive interpretation of previous slide:

if:
C is a cluster
p are objects deep in a cluster

If C is a "tight" cluster, then ε will be very small, and from:

$1/(1+\varepsilon) <= LOF(p) <= (1+\varepsilon)$

We can see that LOF(p) will be about 1

An example:

Approximation of LOF

LOF of objects deep in a cluster are approx 1

Remember: ε = (reach-dist-max/reach-dist-min) -1

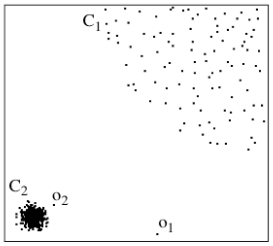


Figure 1: 2-*d* dataset DS1

Approximation of LOF

This approximation is useful for objects deep in a cluster - but what about other objects?

-> We need bounds on LOF!

More notation required:

direct_min(p) = min {reach-dist(p,q) | q is in p's MinPts neighborhood}
direct_max(p) = max {reach-dist(p,q) | q is in p's MinPts neighborhood}

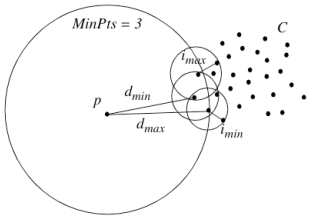
indirect_min(p) = min {reach-dist(p,o) | q is in p's MinPts neighborhood, o is in q's MinPts neighborhood}
indirect_max(p) = max {reach-dist(p,o) | q is in p's MinPts neighborhood, o is in q's MinPts neighborhood}

An Example

Bounds on LOF

direct_min(p) = min {reach-dist(p,q) | q is in p's MinPts neighborhood}
direct_max(p) = max {reach-dist(p,q) | q is in p's MinPts neighborhood}

indirect_min(p) = min {reach-dist(p,o) | q is in p's MinPts neighborhood, o is in q's MinPts neighborhood}
indirect_max(p) = max {reach-dist(p,o) | q is in p's MinPts neighborhood, o is in q's MinPts neighborhood}



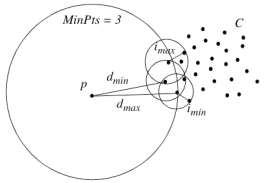
$d_{min} = 4 * i_{max}$ $\Rightarrow LOF_{MinPts}(p) \geq 4$

$d_{max} = 6 * i_{min}$ $\Rightarrow LOF_{MinPts}(p) \leq 6$

Bounds on LOF

Bounds on LOF $\frac{direct_{min}(p)}{indirect_{max}(p)} \leq LOF(p) \leq \frac{direct_{max}(p)}{indirect_{min}(p)}$

An Example:



$d_{min} = 4 * i_{max}$ $\Rightarrow LOF_{MinPts}(p) \geq 4$

$d_{max} = 6 * i_{min}$ $\Rightarrow LOF_{MinPts}(p) \leq 6$

How tight are the bounds?

Bounds on LOF

How tight are the bounds?
It depends on the nature of the point under consideration.

If fluctuation of average reachability distance in the direct and indirect neighborhoods is small, then bounds are tight
This occurs when all MinPts nearest neighbors are in one cluster.

For object with neighbors in multiple clusters, other bounds must be developped - based on “fractional impact” of each cluster on the LOF of a point

Impact of MinPts

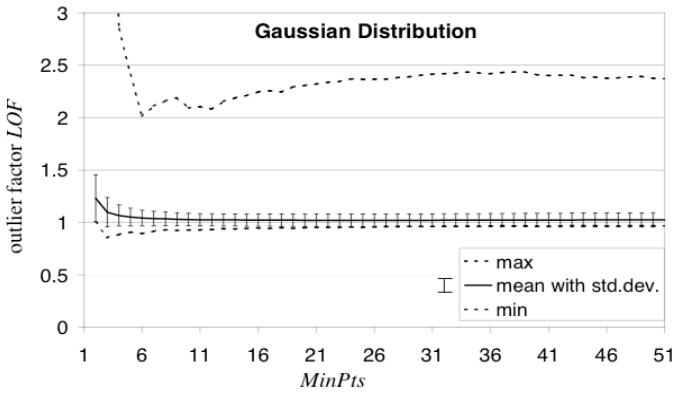
Impact of MinPts

Reminder:
MinPts is the minimum number of points in a neighborhood
MinPts is the only parameter for LOF

It is therefore important to understand how changing MinPts impacts LOF

Consider the result of changing MinPts over a Gaussian cluster dataset.

Impact of MinPts



Impact of MinPts

- Unpredictable impact of MinPts is a potential problem for LOF
- The authors suggest a heuristic:
  - Determine reasonable bounds for MinPts, test all MinPts in those bounds

- MinPtsLB
  - Min number of objects in cluster
- MaxPtsUB
  - Max number of objects in an object’s neighborhood such that that item might still be an outlier

Experimental Results

- Two types of experiments done:
  - Correctness
  - Efficiency
- Correctness
  - Hockey and Soccer data
  - Identified meaningful outlier
- Efficiency
  - Large synthetic data sets
  - Two phases of computation
  - Second phase independent of dimensionality of original data set

Experimental Results

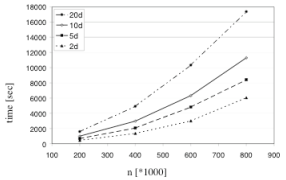


Figure 10: Runtime of the materialization of the 50-nn queries for different dataset sizes and different dimensions using an index

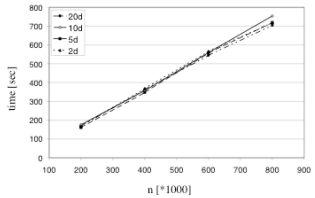


Figure 11: Runtime for the computation of the LOFs for different dataset sizes

Pros and Cons of LOF

- Advantages
  - Finds meaningful local outliers
  - Only one parameter
  - Second computation step independent of dimensions of initial data set
  - May “handshake” with clustering algorithms
  - Good bounds
  - Fairly intuitive interpretation
- Disadvantages
  - Unpredicatable impact of MinPts
  - Must find all MinPts-neighborhoods - take care to choose right approach!
  - Doesn’t indicate why an outlier might be interesting

Conclusion

LOF finds local outliers using density, and is viable for large data sets.

Questions?