Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches

Density-based approaches 1/2

- Outliers are objects in regions of low density
- General idea:
  - Compare the density around a point with the density around its local neighbors
  - The relative density of a point compared to its neighbors' density is computed as an outlier score
  - Approaches essentially differ on how they estimate density
- Basic assumption
  - The density around a normal data object is similar to the density around its neighbors
  - The density around an outlier is considerably different from the density around its neighbors
- Closely related to distance-based methods, since density is usually defined in terms of proximity.

Density-based approaches 2/2

The outlier score of an object is the inverse of the density around this object

- Different definitions of density:
  - e.g., # points within a specified distance *d* from the given object
  - The choice of *d* is critical
    - Too small *d* → many normal points will be considered outliers
    - Too larger *d* → many outlier points will be considered normal
- A global definition of density is problematic (recall our discussion on the clustering lectures)
  - Fail when data contains regions of different densities
  - Solution: use a notion of density that is relative to the neighborhood of the object

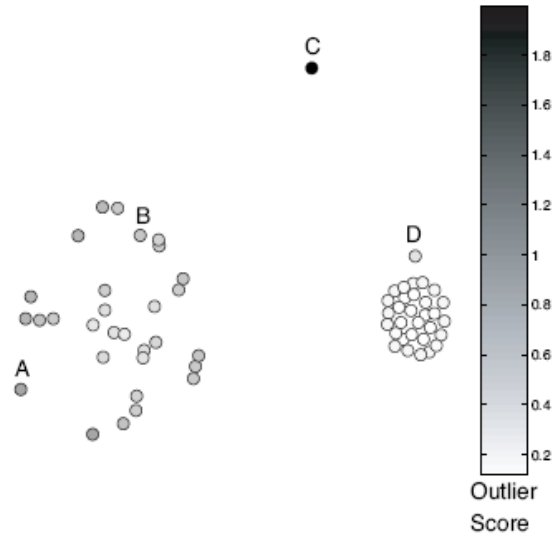
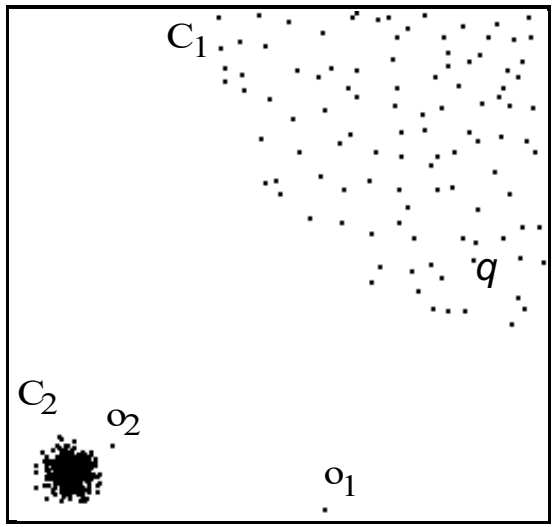


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

D has a higher absolute density than *A*, but comparing to its neighborhood its density is lower.

LOF(Local Outlier Factor) 1/4

- Local Outlier Factor (LOF) [Breunig et al. 1999], [Breunig et al. 2000]
- Motivation:
  - Distance-based outlier detection models have problems with different densities
  - How to compare the neighborhood of points from areas of different densities?
  - Example
    - *DB*(ε,π)-outlier model
      - Parameters ε and π cannot be chosen so that o_2 is an outlier but none of the points in cluster C_1 (e.g. q) is an outlier
    - Outliers based on kNN-distance
      - kNN-distances of objects in C_1 (e.g. q) are larger than the kNN-distance of o_2
  - Solution: consider relative density

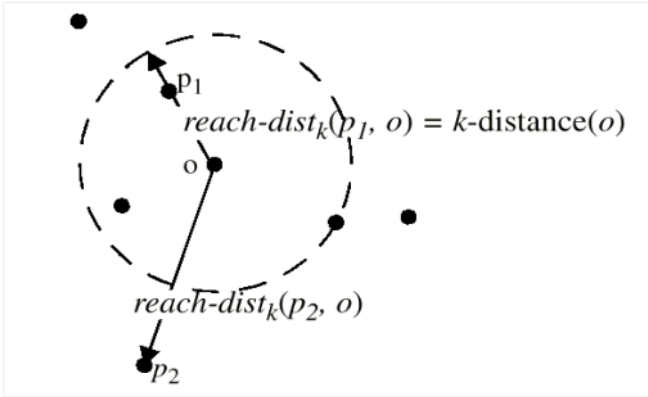


LOF model 2/4

- Reachability distance of an object *p* w.r.t. an object *o*

$$reach\text{-}dist_k(p,o) = \max\{k\text{-distance}(o), dist(p,o)\}$$

  - This is not symmetric!



- Local reachability density (*lrd*) of point *p*
  - Inverse of the average reach-dists of the kNNs of *p*

$$lrd_k(p) = 1/\left(\frac{\sum_{o\in kNN(p)} reach\text{-}dist_k(p,o)}{|kNN(p)|}\right) \quad \Rightarrow \frac{1}{lrd_k(p)} = \frac{\sum_{o\in kNN(p)} reach\text{-}dist_k(p,o)}{|kNN(p)|}$$

- Local outlier factor (LOF) of point *p*
  - Average ratio of *lrd*s of neighbors of *p* and *lrd* of *p*

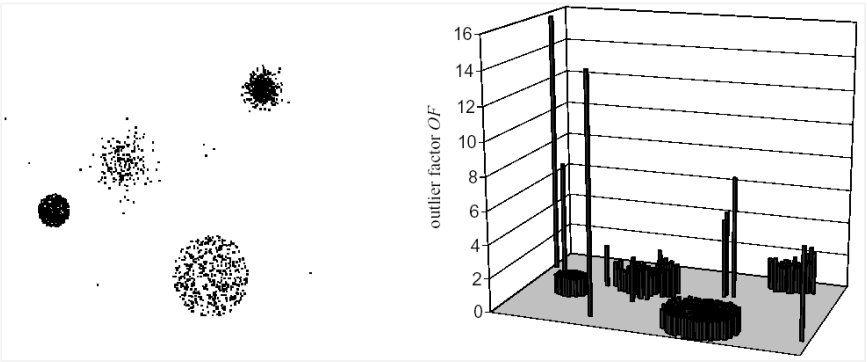
$$LOF_k(p) = \frac{1}{|kNN(p)|} * \sum_{o\in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}$$

relative density

average

LOF 3/4

- Properties
  - LOF ≈ 1: point is in a cluster (region with homogeneous density around the point and its neighbors)
  - LOF >> 1: point is an outlier
  - So, outliers are points with the largest LOF values



LOFs (MinPts = 40)

- Discussion
  - Choice of *k* (*MinPts* in the original paper) specifies the reference set
  - Implements a local approach (resolution depends on the user's choice for *k*)
  - Outputs a scoring (assigns a LOF value to each point)

LOF example 4/4

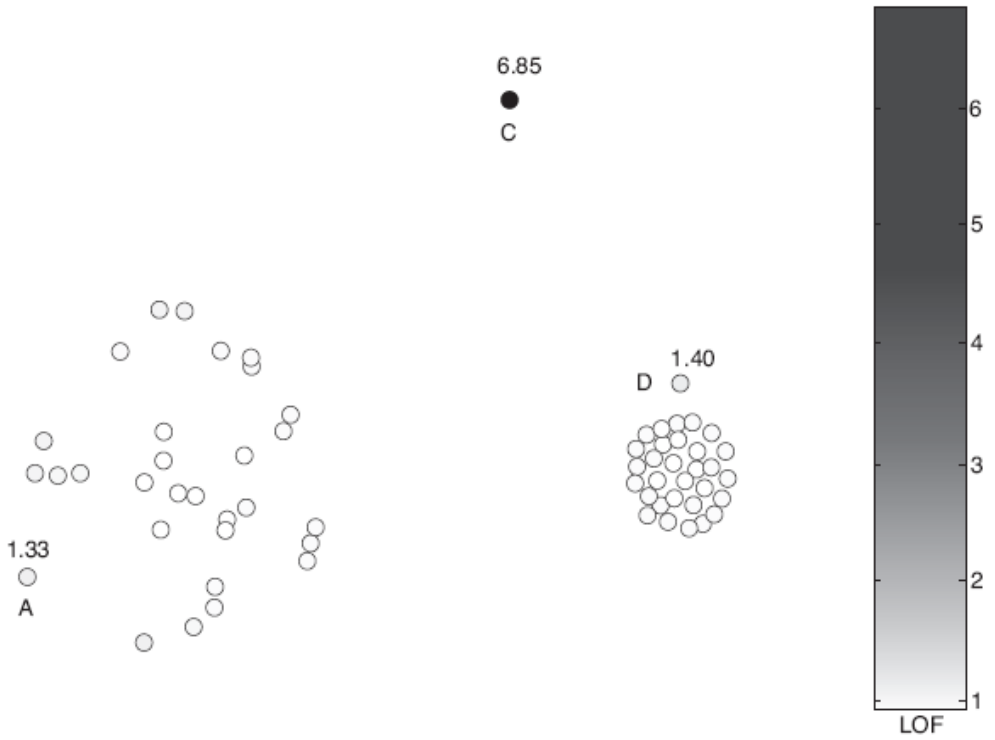


Figure 10.8. Relative density (LOF) outlier scores for two-dimensional points of Figure 10.7.

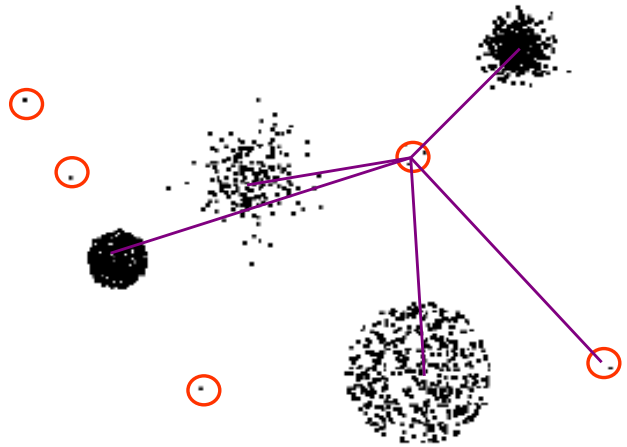
Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches

Clustering-based approaches

An object is a cluster-based outlier if it does not strongly belong to any cluster.

- Basic idea:
  - Cluster the data into groups
  - Choose points in small clusters as candidate outliers. Compute the distance between candidate points and non-candidate clusters.
    - If candidate points are far from all other non-candidate points, they are outliers
- A more systematic approach
  - Find clusters and then assess the degree to which a point belongs to any cluster
    - e.g. for *k*-Means distance to the centroid
  - In case of *k*-Means (or in general, clustering algorithms with some objective function), if the elimination of a point results in substantial improvement of the objective function, we could classify it as an outlier
    - i.e., clustering creates a model of the data and the outliers distort that model.



Prototype-based clusters

- Methods like *k*-Means, *k*-Medoids
- Several ways to assess the extent to which a point belongs to a cluster
  - Measure the distance of the object to the cluster prototype and take this as the outlier score
  - Or, if the clusters are of different densities, the outlier score could be the relative distance of an object from the cluster prototype w.r.t. the distances of the other objects in the cluster.

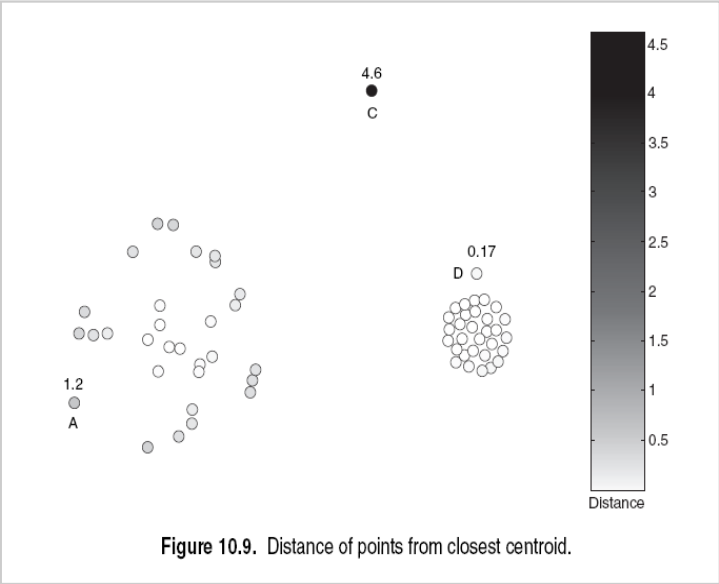


Figure 10.9. Distance of points from closest centroid.

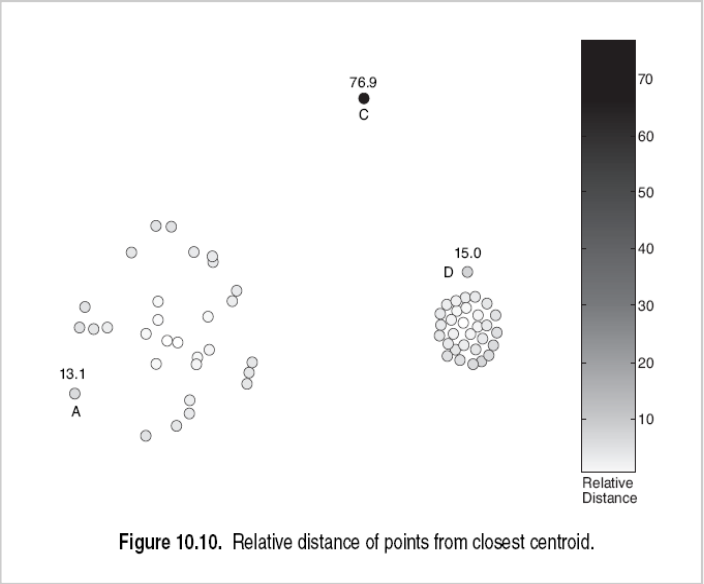


Figure 10.10. Relative distance of points from closest centroid.

Outlier evaluation

- If there are class labels
  - Similar to classifier evaluation, but outlier class is typically smaller than the normal class
  - Measures such as precision and recall are more appropriate than e.g., accuracy or error rate
- In the absence of class labels
  - More difficult
  - For model-based approaches, one could check model improvement after outlier removal