Anomaly Detection Approaches

Huiping Cao

Additional Anomaly Detection Techniques

- Visual approaches
- Proximity-based
  - Anomalies are points far away from other points
  - Can detect this graphically in some cases
  - The proximity of outliers to their neighbors are different from the proximity of most other objects to their neighbors
  - Distance-based
  - Density-based
    - Low density points are outliers
- Clustering-based
  - Normal objects belong to large and dense clusters
  - Outliers belong to small or sparse clusters, or belong to no cluster

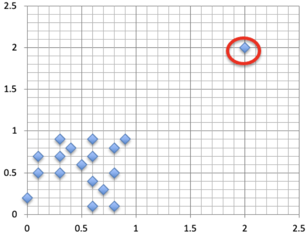
Proximity-based Approaches

- Data is represented as a vector of features
- Based on the neighborhood
- Major approaches
  - Distance based
  - Density based

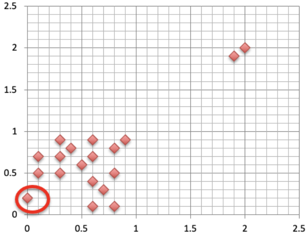
Distance-based approach

- Anomaly: if an object is distant from most points.
- Several different techniques
  - An object is an outlier if a specified fraction of the objects is more than a specified distance away (Knorr, Ng 1998)
  - Distance to k-Nearest Neighbor: the outlier score of an object is given by the distance to its k-nearest neighbor.
- Problem: hard to decide k (see next slides) or the threshold
- Improvement: average of the distances to the first k-nearest neighbors

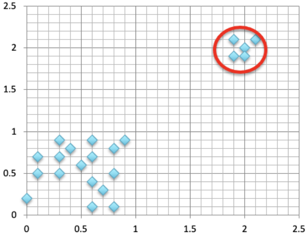
Distance-based approach



k=1, outlier is O



k=1, outlier is O



k=5, all points at the right upper corner are outliers

Distance-based outlier detection

- Given a dataset D with n data points, a distance threshold r
- r-neighborhood: about outliers vs. the rest of the data
- Object o is a DB(r, π)-outlier

$$\frac{\{o' | dist(o, o') \leq r\}}{n} \leq \pi$$

- Approach:
  - Compute the distance between every pair of data points
  - $O(n^2)$
  - Practically, $O(n)$

A grid-based method implementation

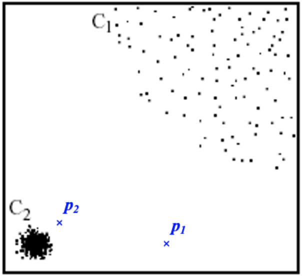
- Cell diagonal length: $\frac{r}{2}$
- Cell edge length: $\frac{r}{2\sqrt{d}}$ where d is the number of dimensions.
- Level-1 cell
  - Direct neighbor cells of a cell C
  - Any point o' in such cells has $dist(o, o') \leq r$
- Level-2 cell
  - One or two cells away from a cell C
  - Any point with $dist(o, o') > r$ must be in level-2 cell

A grid-based method implementation

- Pruning
  - n_0 total number of objects in a cell C
  - n_1 total number of objects in a cell C's level-1 cells
  - n_2 total number of objects in a cell C's level-2 cells
- Level-1 cell pruning
  - If $(n_0 + n_1) > \pi n$, o is NOT an outlier
- Level-2 cell pruning
  - If $(n_0 + n_1 + n_2) < \pi n + 1$, all the points in C are outliers

Distance-based outlier detection

Global outliers: cannot handle data sets with regions of different densities.



Strengths/Weaknesses of Distance-Based Approaches

- Simple
- Expensive $O(n^2)$
- Sensitive to parameters
- Sensitive to variations in density
- Distance becomes less meaningful in high-dimensional space

Density-Based Approaches

- **Local** proximity-based outlier
- Compare the density around one object with the density around its local neighborss
- Density-based outlier: the outlier score of an object is the inverse of the density around the object.
  - Can be defined in terms of the k nearest neighbors
  - Definition 1: Inverse of distance to kth neighbor
  - Definition 2: Inverse of the average distance to k neighbors
  - DBSCAN definition

Density-Based Approaches

- D: a set of objects
- Nearest neighbor of o: $d(o, D) = min\{d(o, o')|o' \ in \ C\}$
- Local outliers: relative to their local neighborhoods, particularly with respect to the densities of the neighborhoods.
- Density based outlier: the outlier score of an object is the inverse of the density around an object.

Concepts

- *k-distance of an object o* $d_k(o)$ (or $d(o,k)$): measure the relative density of an object o.
- Formally, $d_k(o) = d(o,k)$ s.t.
  - at least k objects o' in $D/\{o\}$, $d(o,o') \leq d(o,p)$
  - at least k-1 objects o' in $D/\{o\}$, $d(o,o') < d(o,p)$
- *k-distance neighborhood* of an object o
  - $N_k(o) = N(o,k) = \{o' | o' \ in \ D, d(o,o') \leq d_k(o)\}$
  - $N_k(o)$ may contain more than k objects
- Measure local density: average distance from o to $N_k(o)$
  - Problem: fluctuations

Concepts (cont.)

- Reachable distance
  - $reachdist(o' \to o) = max\{d_k(o), d(o, o')\}$
  - Alleviate fluctuations
  - Not symmetric, $reachdist(o' \to o) \neq reachdist(o \to o')$
- Local density of o: average reachability distance from o to $N_k(o)$

$$density_k(o) = density(o, k)$$

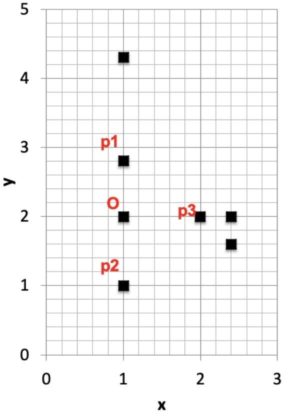
$$= \frac{|N_k(o)|}{\sum_{o' \in N_k(o)} reachdist(o \to o')}$$

$$= \frac{|N_k(o)|}{\sum_{o' \in N_k(o)} maxdist\{d_k(o'), d(o, o')\}}$$

- Different from density definition in density-based clustering
  - Global/local

Example

- k=2, use Euclidean distance
- Distance from o to o's 2NN is 1
- $d_k(o)=1$
- $N_k(o)=\{p1,p2,p3\}$
  - $d_k(p1)$ = sqrt(0.64+1.0) = 1.28, dist(o,p1)=0.8
  - $d_k(p2)$ = sqrt(2) =1.41, dist(o,p2)=1
  - $d_k(p3)$ = sqrt(0.32) = 0.57, dist(o,p3)=1
  - reachdist(o->p1) = 1.28
  - reachdist(o->p2) = 1.41
  - reachdist(o->p3) = 1
- $density_k(o)$=3/(1.28+1.41+1) = 0.813



Local outlier factor (LOF)

- Local outlier factor (LOF) (or average relative density of o)
  - Average ratio of local reachability density of o and local reachability density of the k-nearest neighbors of o

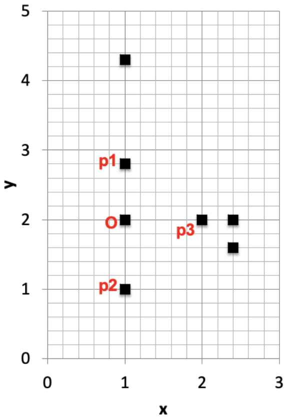
$$LOF_k(o) = relative\ density(x, k) = \frac{\sum_{o' \in N_k(o)} \frac{density_k(o')}{density_k(o)}}{|N_k(o)|}$$

  where $density_k(o) = density(o, k)$ and $density_k(o') = density(o', k)$.

- The lower $density_k(o)$ and the higher $density_k(o') \rightarrow$ higher LOF $\rightarrow$ higher probability to be outlier

Example

- k=2, use Euclidean distance
- Distance from o to o's 2NN is 1
- $d_k(o)=1$
- $N_k(o)=\{p1,p2,p3\}$
  - $d_k(p1)$ = sqrt(0.64+1.0) = 1.28, dist(o,p1)=0.8
  - $d_k(p2)$ = sqrt(2) =1.41, dist(o,p2)=1
  - $d_k(p3)$ = sqrt(0.32) = 0.57, dist(o,p3)=1
  - reachdist(o->p1) = 1.28
  - reachdist(o->p2) = 1.41
  - reachdist(o->p3) = 1
- $density_k(o)$=3/(1.28+1.41+1) = 0.813
- Then, calculate $density_k$ (p1), $density_k$ (p2), $density_k$ (p3)



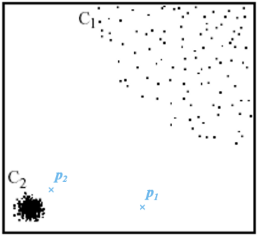
Relative Density

Algorithm 10.2 Relative density outlier score algorithm.

{k is the number of nearest neighbors}
for all objects $\mathbf{x}$ **do**
    Determine $N(\mathbf{x}, k)$, the k-nearest neighbors of $\mathbf{x}$.
    Determine $density(\mathbf{x}, k)$, the density of $\mathbf{x}$, using its nearest neighbors, i.e., the objects in $N(\mathbf{x}, k)$.
end for
for all objects $\mathbf{x}$ **do**
    Set the $outlier\ score(\mathbf{x}, k) = relative\ density(\mathbf{x}, k)$
end for

Density-based: LOF approach

- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample p as the average of the ratios of the density of sample p and the density of its nearest neighbors
- Outliers are points with largest LOF value



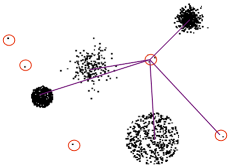
In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Strengths/Weaknesses of Density-Based Approaches

- Simple
- Expensive $O(n^2)$
- Sensitive to parameters
- Density becomes less meaningful in high-dimensional space

Clustering-Based Approaches

- Clustering-based Outlier: an object is a cluster-based outlier if it does not strongly belong to any cluster
  - For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
  - For density-based clusters, an object is an outlier if its density is too low
  - For graph-based clusters, an object is an outlier if it is not well connected
- An outlier
  - an object belonging to a small and remote cluster
  - or not belonging to any cluster



Clustering-Based

- Basic steps: Cluster the data into groups of different density
- Three general approaches
  - Approach 1: An object does not belong to any cluster $\rightarrow$ outlier object
  - Approach 2: There is a large distance between an object and the cluster to which it is closest $\rightarrow$ outlier
  - Approach 3: The object is part of a small and sparse cluster $\rightarrow$ all the objects in that cluster are outliers

Approach 2

- Approach 2: There is a large distance between an object and the cluster to which it is closest $\rightarrow$ outlier
- Calculate ratio, the larger the ratio, the farther away o is from its closest cluster C_o, whose center is c_o.

$$ratio = \frac{d(o, c_o)}{\frac{\sum_{o' \in C_o} d(o', c_o)}{|C_o|}}$$

Outliers in Lower Dimensional Projection

- In high-dimensional space, data is sparse and notion of proximity becomes meaningless
  - Every point is an almost equally good outlier from the perspective of proximity-based definitions
- Lower-dimensional projection methods
  - A point is an outlier if in some lower dimensional projection, it is present in a local region of abnormally low density.

Strengths/Weaknesses of Clustering-Based Approaches

- Simple
- Many clustering techniques can be used
- Can be difficult to decide on a clustering technique
- Can be difficult to decide on number of clusters
- Outliers can distort the clusters

References

- Chapter 9: Introduction to Data Mining (2nd Edition) by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar
- Unsupervised Outlier Detection using the Local Outlier Factor (LOF): `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html`