Outlier Analysis

Cam Tu Nguyen

阮锦绣

Software Institute, Nanjing University
nguyenct@lamda.nju.edu.cn
ncamtu@gmail.com

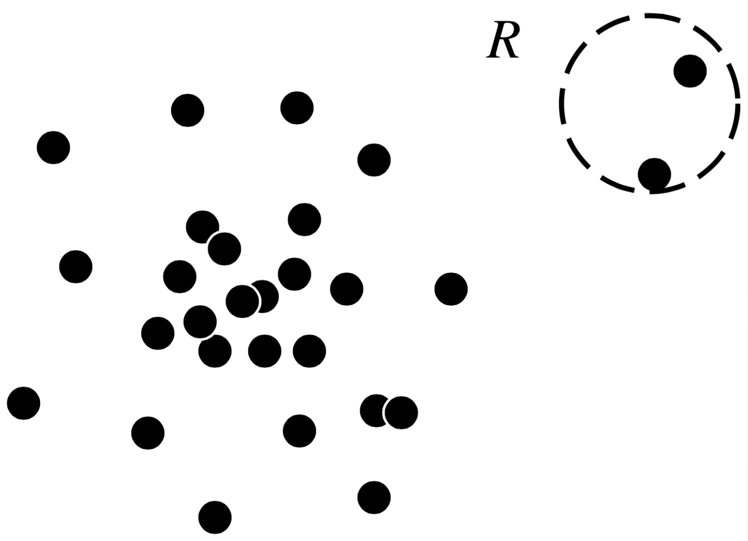
Outline

- Outliers and Outlier Analysis
- Statistical Approaches
  - Parametric methods
  - Non-parametric methods
- Proximity-Based Approaches
  - Distance-based Methods
  - Grid-based Methods
  - Density-based Methods
- Clustering and Classification-based Methods
- Additional Topics
  - Mining Contextual
  - Mining Collective Outliers

What are outliers?

- Outliers are data objects that **deviates significantly** from the rest of the objects.
- Objects that are not outliers are called **normal** or **expected data.**

- **Outliers** are different from **noise**
  - Noise is not interesting.
  - Noise is more like random error, or a variance.



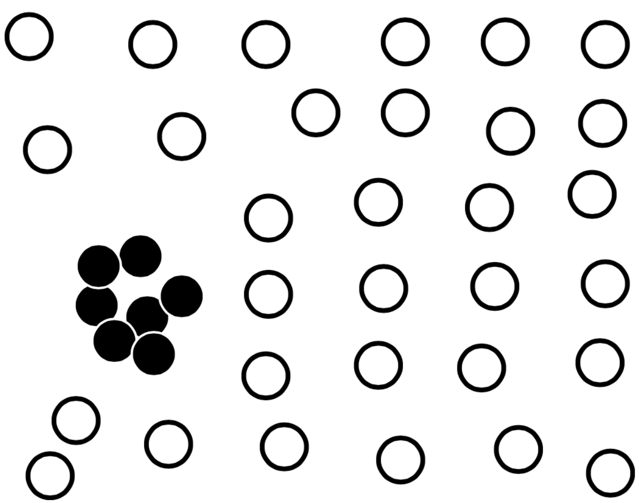
Types of outliers

- Global outliers: A data object is a **global outlier** (also called *point anomaly)* if it deviates significantly from the rest of the data.
  - Most outlier detection methods are aimed at finding global outliers.

- Contextual outliers: a data object is a **contextual outlier** (or *conditional outlier)* if it deviates significantly with respect to a specific context of the object.
  - Example: 28^{o} C is an outlier for a Toronto winter, but not an outlier in another context.

Types of outliers

- Collective outliers: a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire dataset.
  - Note that the individual data objects may not be outliers.
  - Example: DoS attack (denial-of-service attack) contains a group of DoS packages sending back and forward within several computers.



Challenges of Outlier Detection

- Modeling normal objects and outliers effectively
  - It is hard to enumerate all the possible normal behaviors in an application.
  - The border between data normality and abnormality is often not clear cut.
- Application-specific outliers
  - Choosing the similarity/distance measure and the relationship model to describe data objects is application-dependent. (There is no universally applicable method).
- Handling noise in outlier detection
  - Outlier can appear as a "disguised" noise point.
- Understandability.
  - Justify why some points are outliers

Outline

- Outliers and Outlier Analysis
- Statistical Approaches
  - Parametric methods
  - Non-parametric methods
- Proximity-Based Approaches
  - Distance-based Methods
  - Grid-based Methods
  - Density-based Methods
- Clustering-based Methods
- Classification-based Methods
- Additional Topics
  - Mining Contextual and Collective Outliers
  - Outlier Detection in High-dimensional

Statistical Approaches

- Parametric methods
  - Univariate Outlier Detection based on Normal Distribution
  - Multivariate Outlier Detection
  - Using a Mixture of Parametric Distributions

- Non-parametric methods
  - Using histogram
  - Using Kernel density estimation

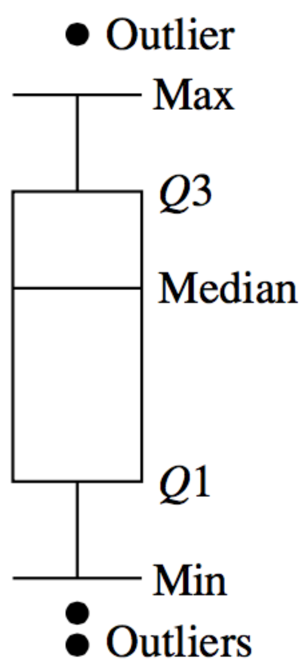
Univariate Outlier Detection based on Normal Distribution

- **Univariate data:** data involving only one attribute or variable
- Outlier Detection based on Normal Distribution
  - Assumption: data is generated from a normal distribution
  - Identify the points with low probability as outliers
  - Simple rule: points that are more than 3 standard deviations away from the mean are outliers.

- **Example:** A city's average temperature values in July in the last 10 years are 24.0; 28.9; 28.9; 29.0; 29.1; 29.1; 29.2; 29.2; 29.3; and 29.4
  - Is 24.0^{o}C is an outlier?
  - Is 25.0^{o}C is an outlier?
  - Is 33.0^{o}C is an outlier?

Univariate Outlier Detection based on Normal Distribution

- Visualization using boxplot
  - The lower quartile (Q1)
  - The upper quartile (Q3)
  - The IQR (inter-quartile-range): Q3-Q1
  - Points that are more than 1.5*IQR smaller than Q1 or 1.5* IQR larger than Q3 are considered outliers.



Univariate Outlier Detection based on Normal Distribution

- Grubb's test (*maximum normed residual test)*
  - Assume data comes from normal distribution
  - Detect one outlier at a time, remove the outlier, and repeat
    - H_0: there is no outlier in data
    - H_A: there is at least one outlier
  - For each data object x in a data set, we define a z-score $z = \frac{|x - \bar{x}|}{s}$,
    - Where $\bar{x}$, s are empirical mean, and standard deviation
  - Given a significance level alpha, an object x is an outlier if

$$z \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}},$$

$t_{\alpha/(2N),N-2}$ The upper critical value of the t-distribution with N-2 degree of freedom, with significance level alpha/(2N)

- Problem: apply Grubb's test on the previous example on the average temperature in July.

Multivariate Outlier Detection

- Multivariate data: data involving two or more attributes or variables

- Mahalanobis distance-based method:
  - Calculate the mean vector $\bar{o}$, and covariance matrix S from the multivariate data set
  - For each object **o**, calculate $MDist(o, \bar{o})$

$$MDist(\boldsymbol{o}, \bar{\boldsymbol{o}}) = (\boldsymbol{o} - \bar{\boldsymbol{o}})^T S^{-1} (\boldsymbol{o} - \bar{\boldsymbol{o}}),$$

  - Detect outliers in the transformed univariate data set $\{MDist(o, \bar{o})$ | o in *D*}

Multivariate Outlier Detection

- Multivariate data: data involving two or more attributes or variables

- Chi-square statistic-based method
  - For an object $O = \{o_1, o_2, \ldots, o_p\} \in R^p$, chi-square statistic is calculated:

$$\chi^2 = \sum_{i=1}^{p} \frac{(o_i - E_i)^2}{E_i}$$

E_i is the expected value of the ith attribute in the data set.

  - Chi-square statistic is large, the point is an outlier.

Outlier Detection using Mixture of Parametric Distribution

- Assume the data set D contains samples from a mixture of two probability distributions
  - M (majority distribution)
  - A (anomalous distribution)
- General approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log-likelihood of D at time t.
  - For each point x_t that belongs to M, move it to A
    - Let $L_{t+1}(D)$ be the new log-likelihood
    - Compute the difference, **delta=$L_t(D) - L_{t+1}(D)$**
    - If **delta > c** (some threshold), then x_t is declared as an anomaly and moved permanently from M to A.

Outlier Detection using Mixture of Parametric Distribution

- Data distribution, D = (1-lambda)M + lambda*A
- M is a probability distribution estimated from data
  - Can be based on any model (Gaussian, Mixture of Gaussian, etc.)
- A is initially assumed to be uniform distribution
- Likelihood at time t.

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

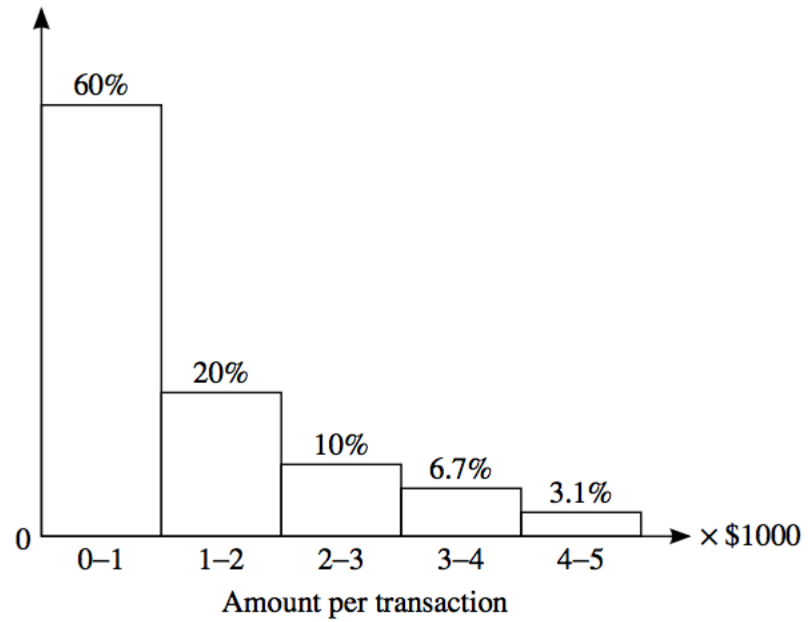
Statistical Approaches

- Parametric methods
  - Univariate Outlier Detection based on Normal Distribution
  - Multivariate Outlier Detection
  - Using a Mixture of Parametric Distributions

- Non-parametric methods
  - Using histogram
  - Using Kernel density estimation

Outlier Detection Using Histogram

- **Two-step procedure**
  - *Histogram construction*
    - Choose the type of histogram (equal width, or equal depth)
    - The number of bins and the size of each bin.
  - *Outlier detection*
    - Use histogram to assign **outlier score** (e.g. use the inverse of the volume of the bin in which the object falls)
- **Example:**
  - $7500 can be regarded as outlier because only 1-60%-20%-10%-6.7%-3.1%=0.2% of transactions have an amount higher than $5000
  - Outlier score of $7500: 1/0.2%=500



Outlier Detection using Kernel Density Estimation

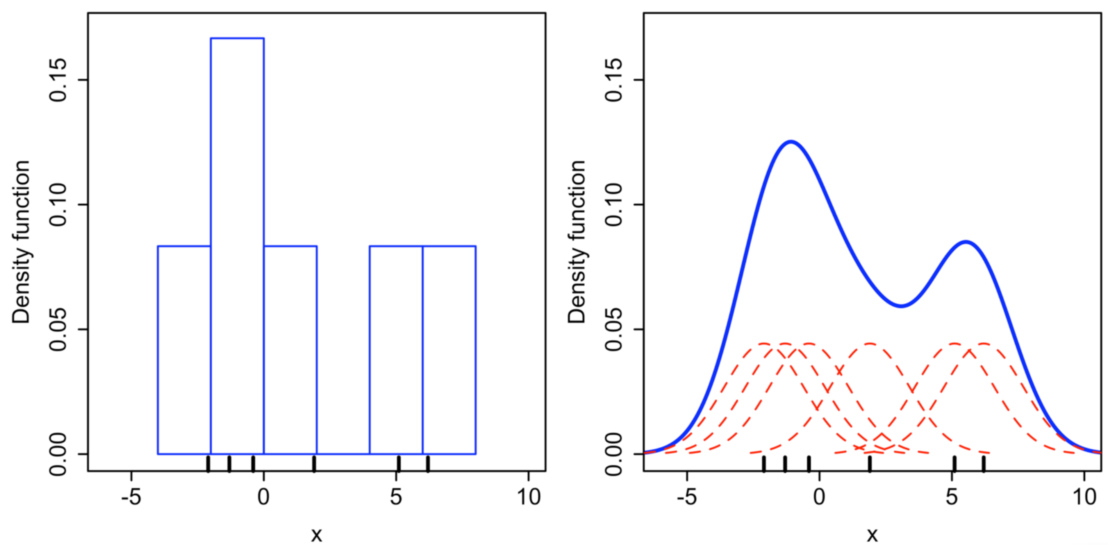
- Kernel Density Estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable.
- Let $(x_1, x_2, \ldots, x_n)$ be i.i.d sample drawn from some distribution with an unknown density f, its kernel density estimator is:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big)$$

  - K(.) is the kernel function (a non-negative function that integrates to 1 and has mean zero), h is a smoothing parameter

Outlier Detection using Kernel Density Estimation

- KDE vs Histogram: KDE result is more smooth
  - Given 6 data points: $x_1 = -2.1$, $x_2 = -1.3$, $x_3 = -0.4$, $x_4 = 1.9$, $x_5 = 5.1$, $x_6 = 6.2$.
    - Histogram: a box of height 1/12 is placed if one data point falls in a bin, if more than one data points fall into one bin, we stack 2 boxes.
    - KDE: place a normal kernel with variance 2.25 at each point, the kernels are summed to make KDE (solid blue line)



Outlier Detection using Kernel Density Estimation

- Outlier detection
  - For an object **o,** $\hat{f}(o)$ gives the estimated probability that the object is generated by the stochastic process.
    - If $\hat{f}(o)$ is high, object **o** is likely normal.
    - Otherwise, **o** is highly to be an outlier.

- A frequently used kernel is a standard Gaussian function with mean 0 and variance 1.

$$K\left(\frac{x-x_i}{h}\right)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-x_i)^2}{2h^2}}.$$

Outline

- Outliers and Outlier Analysis
- Statistical Approaches
  - Parametric methods
  - Non-parametric methods
- Proximity-Based Approaches
  - Distance-based Methods
  - Grid-based Methods
  - Density-based Methods
- Clustering and Classification-based Methods
- Additional Topics
  - Mining Contextual and Collective Outliers
  - Outlier Detection in High-dimensional

Distance-based Outlier Detection

- Let D be the set of data objects, r (r>=0) be a distance threshold and $\pi (0 < \pi \leq 1)$ be a fraction threshold
  - An object **o,** is a $DB(r, \pi)$-outlier if

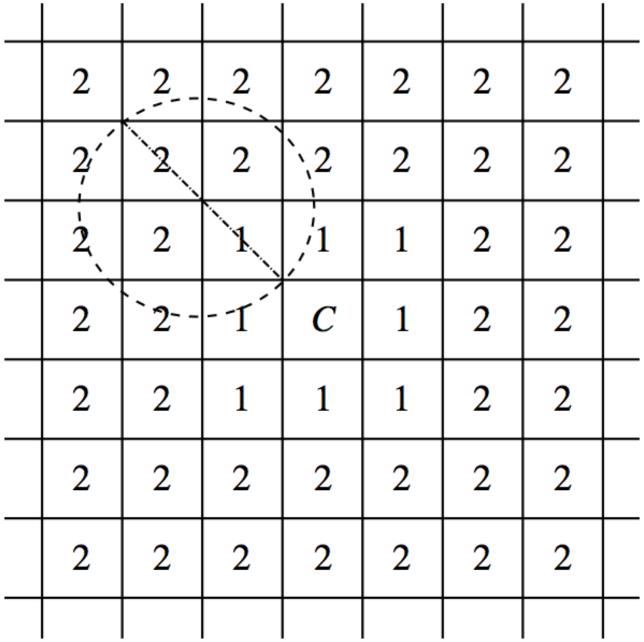
$$\frac{\|\{\mathbf{o}' \mid dist(\mathbf{o}, \mathbf{o}') \leq r\}\|}{\|D\|} \leq \pi,$$

  - Where dist(o,o') is a distance measure

- **Problem**: write an algorithm that used the above measurement to detect all $DB(r, \pi)$-outliers from D, what is the computation complexity of the algorithm?

A Grid-based method

- CELL is **a grid-based** method for distance-based outlier detection
  - Data space is partitioned into multidimensional grid
  - Each cell is a hypercube with the length of each edge is $r/2\sqrt{l}$ where l is the number of dimensions.



A Grid in 2-D dimension

C is the cell of interest.

A Grid-based method

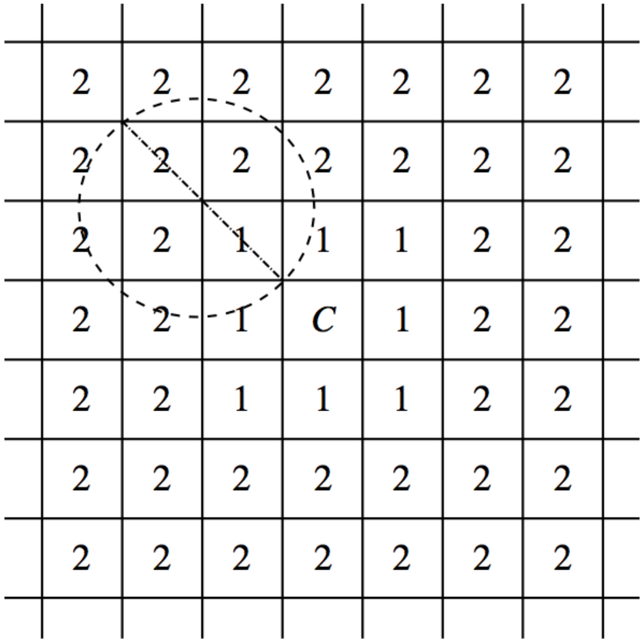
- CELL is **a grid-based** method for distance-based outlier detection
  - **Level-1 cell property**

    $\forall x \in C, y \in \text{a level 1 cell}$, then

    $$dist(x, y) \leq r$$

  - **Level-2 cell property**

    $x \in C, y$ such that $dist(x, y) \geq r$, then y is in a level-2 cell

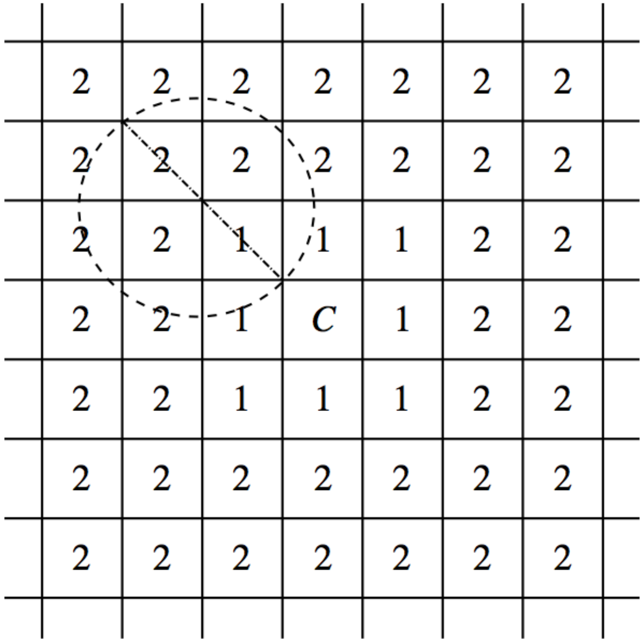


A Grid in 2-D dimension

C is the cell of interest.

A Grid-based method

- CELL is **a grid-based** method for distance-based outlier detection
  - Let a, b1, b2 be the number of points in C, level 1 cells, and level-2 cells
  - **Level-1 cell pruning rule:**
    - if $a + b_1 > \lceil \pi n \rceil$ then every object in C is not a $DB(r, \pi)$-outlier
  - **Level-2 cell pruning rule:**
    - if $a + b_1 + b_2 < \lceil \pi n \rceil + 1$, all objects in C are $DB(r, \pi)$-outliers

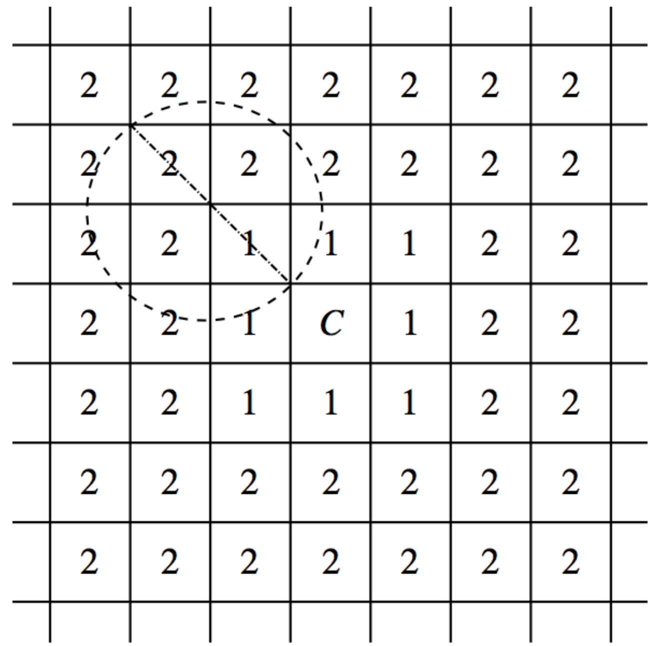


A Grid in 2-D dimension

C is the cell of interest.

A Grid-based method

- CELL is **a grid-based** method for distance-based outlier detection
- Using CELL, we only need to check for objects that can't be pruned using 2 rules
- For large data set, CELL is costly due to the need of swapping pages from disk to memory.

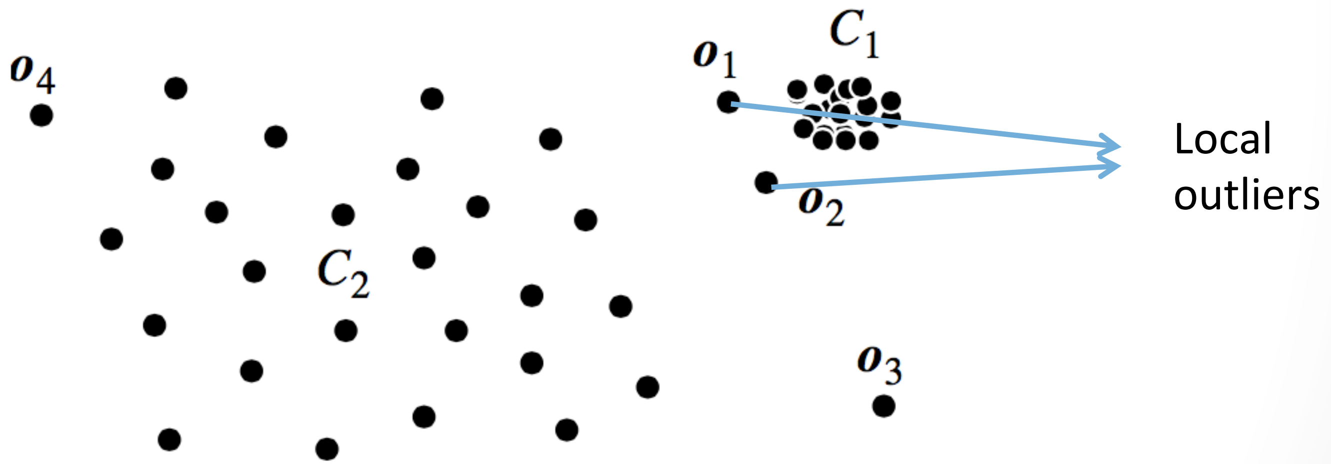


A Grid in 2-D dimension

C is the cell of interest.

Density-based Outlier Detection

- Previous distance-based outlier detection can only detect global outliers.
- We are interested in detecting outliers w.r.t their local neighborhood.



Density-based Outlier Detection

- We are interested in detecting outliers w.r.t their local neighborhood.
- We need to **compare the density around an object** to the **density around the local objects.**
- ***K-distance***
  - *$dist_k(o)$* is the distance between o and the k-nearest neighbor.
- ***K-distance neiahborhood***

$$N_k(\mathbf{o}) = \{\mathbf{o}' | \mathbf{o}' \in D, dist(\mathbf{o}, \mathbf{o}') \leq dist_k(\mathbf{o})\}.$$

$||(N_k(o)|| >= k$

- ***Reachability distance from o' to o***

$$reachdist_k(\mathbf{o} \leftarrow \mathbf{o}') = \max\{dist_k(\mathbf{o}), dist(\mathbf{o}, \mathbf{o}')\}$$

Density-based Outlier Detection

- **Local reachability density** of an object **o**

$$lrd_k(\boldsymbol{o}) = \frac{\|N_k(\boldsymbol{o})\|}{\sum_{\boldsymbol{o}' \in N_k(\boldsymbol{o})} reachdist_k(\boldsymbol{o}' \leftarrow \boldsymbol{o})}$$

Problem: Assume that our data points are on 1-D space. if we set k=4, which value ($lrd_k(orange)$ or $lrd_k(green)$) is larger? Why?

Density-based Outlier Detection

- **Local reachability density** of an object **o**

$$lrd_k(\mathbf{o}) = \frac{\|N_k(\mathbf{o})\|}{\sum_{\mathbf{o}' \in N_k(\mathbf{o})} reachdist_k(\mathbf{o}' \leftarrow \mathbf{o})}$$

- **Local outlier factor of an object o**

$$LOF_k(\mathbf{o}) = \frac{\sum_{\mathbf{o}' \in N_k(\mathbf{o})} \frac{lrd_k(\mathbf{o}')}{lrd_k(\mathbf{o})}}{\|N_k(\mathbf{o})\|} = \sum_{\mathbf{o}' \in N_k(\mathbf{o})} lrd_k(\mathbf{o}') \cdot \sum_{\mathbf{o}' \in N_k(\mathbf{o})} reachdist_k(\mathbf{o}' \leftarrow \mathbf{o}).$$

A high LOF captures a local outlier of which the local density is relatively low compared to the local densities of its k-nearest neighbors.

Outline

- Outliers and Outlier Analysis
- Statistical Approaches
  - Parametric methods
  - Non-parametric methods
- Proximity-Based Approaches
  - Distance-based Methods
  - Grid-based Methods
  - Density-based Methods
- Clustering and Classification-based Methods
- Additional Topics
  - Mining Contextual
  - Mining Collective Outliers

Clustering-based methods

- General approaches
  - Does the object belong to any cluster? If not, then it is identified as an outlier
  - Is there a large distance between the object and the cluster to which it is closest? If yes, it is an outlier
  - If the object part of a small or spares cluster? If yes, then all the objects in that clusters are outliers.
- Disadvantage:
  - Clustering may be costly

Classification-based methods

- A training set contains samples labeled as normal and others labeled as outliers
- Imbalance classification problem
  - Approaches:
    - Sampling
    - One-class classification, e.g. one-class SVM

Outline

- Outliers and Outlier Analysis
- Statistical Approaches
  - Parametric methods
  - Non-parametric methods
- Proximity-Based Approaches
  - Distance-based Methods
  - Grid-based Methods
  - Density-based Methods
- Clustering and Classification-based Methods
- Additional Topics
  - Mining Contextual
  - Mining Collective Outliers

Mining Contextual Outliers

- Two-types of attributes
  - Contextual attributes define the context
    - E.g. spatial attributes, time, network locations, etc.
  - Behavioral attributes define characteristics of an object

- Transforming contextual outlier detection to conventional outlier detection
  - Identify the context, then perform outlier detection in each context
  - Map from the model of contextual attributes to a model of behavioral attributes using statistical approaches.

Mining Collective Outliers

- Define **structured units**
  - Subsequence, a time-series segment, a local area or a subgraph

- Mining outliers in the set of structured units
  - Extract features from structured units.
  - A structure unit, which represents a group of objects in the original data set, is a collective outlier if the structure unit deviates greatly from the expected trend.

Summary

- Outliers and Outlier Analysis
- Statistical Approaches
  - Parametric methods
  - Non-parametric methods
- Proximity-Based Approaches
  - Distance-based Methods
  - Grid-based Methods
  - Density-based Methods
- Clustering and Classification-based Methods
- Additional Topics
  - Mining Contextual
  - Mining Collective Outliers