Chapter 10: Outlier Detection

—

Department of Computer Science and Engineering
Kathmandu University

What are outliers?

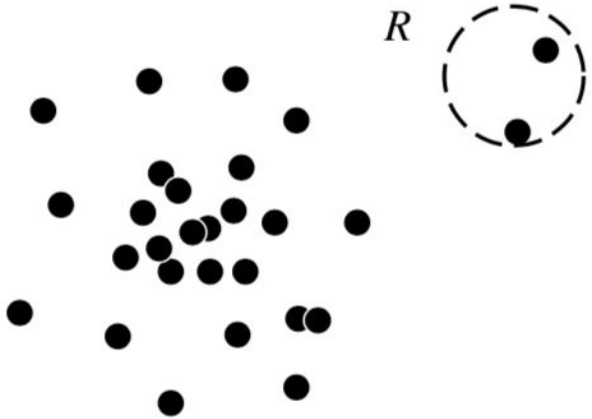
- Objects with behaviors that are very different from expectation are called outliers or anomalies.
- Applications: Fraud detection, intrusion detection, public safety and security, industry damage detection, image processing, sensor/video network surveillance etc.
- Related tasks:
  - Clustering analysis
    - Clustering finds the majority patterns in a data set and organizes the data accordingly, whereas outlier detection tries to capture those exceptional cases that deviate substantially from the majority patterns.
  - Novelty detection in evolving data sets
    - Novel items may initially appear as outliers. The follow-up instances are not treated as outliers anymore.
    - Outlier detection and novelty detection share some similarity in modeling and detection methods.

Types of outliers

1. Global outliers
2. Contextual (or conditional) outliers
3. Collective outliers

Types of outliers

1. **Global outliers**
   In a given data set, a data object is a global outlier if it deviates significantly from the rest of the data set.



Types of outliers

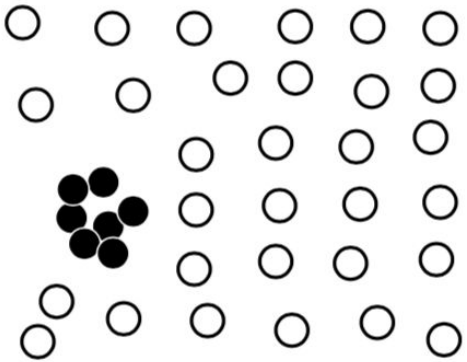
2. **Contextual (or conditional) outliers**
   In a given data set, a data object is a contextual outlier if it deviates significantly with respect to a specific context of the object.

   Example: “The temperature today is 28°C. Is it exceptional (i.e., an outlier)?”

   It depends, for example, on the time and location! If it is in winter in Toronto, yes, it is an outlier. If it is a summer day in Toronto, then it is normal.

Types of outliers

3. **Collective outliers**
   Given a data set, a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set. (The individual data objects may not be outliers.)
   Example: a large set of transactions of the same stock among a small party in a short period



Challenges

- Modeling normal objects and outliers effectively.
  - Hard to enumerate all possible normal behaviors in an application.
  - The border between data normality and abnormality (outliers) is often not clear cut.
- Application-specific outlier detection
  - Choosing the similarity/distance measure and the relationship model to describe data objects are often application specific.
    - For example, in clinic data analysis, a small deviation may be important enough to justify an outlier whereas in marketing analysis, objects are often subject to larger fluctuations.
- Handling noise in outlier detection
  - Noise can distort the data, blurring the distinction between normal objects and outliers
  - Noise and missing data may “hide” outliers and reduce the effectiveness of outlier detection
- Understandability
  - Explaining why the detected objects are outliers.

Outlier detection method

- Supervised, Semi-Supervised, and Unsupervised Methods
- Statistical Methods, Proximity-Based Methods
- Clustering-Based Methods
- Classification Based Methods

Outlier detection method: Supervised methods

Approaches

- Domain experts examine and label a sample of the underlying data. Outlier detection can then be modeled as a classification problem to recognize outliers.
- Experts may label just the normal objects, and any other objects not matching the model of normal objects are reported as outliers.
- Model the outliers and treat objects not matching the model of outliers as normal.

Outlier detection method: Supervised methods

Challenges

- Class imbalance: the population of outliers is typically much smaller than that of normal objects.
- Importance of sensitivity or recall of outlier detection: In many outlier detection applications, catching as many outliers as possible is far more important than not mislabeling normal objects as outliers.

Outlier detection method: Unsupervised Methods

- Unsupervised outlier detection methods make an implicit assumption: The normal objects are somewhat “clustered.”
- Idea: Find clusters first, and then the data objects not belonging to any cluster are detected as outliers.
  - Issues:
    - Is it an outlier or a noise?
    - It is often costly to find clusters first and then find outliers.

Outlier detection method: Semi-supervised Methods

- Semi-supervised learning methods can be applied to detect outliers.
- When some labeled normal objects are available, we can use them, together with unlabeled objects that are close by, to train a model for normal objects. Those objects not fitting the model of normal objects are classified as outliers.

Outlier detection method: Statistical methods

- Statistical methods (aka model-based methods) make assumptions of data normality.
  - They assume that normal data objects are generated by a statistical (stochastic) model, and that data not following the model are outliers.
- General idea: learn a generative model fitting the given data set, and then identify those objects in low-probability regions of the model as outliers.
- Types:
  - Parametric: assumes that the normal data objects are generated by a parametric distribution with parameter.
  - Non-parametric: tries to determine the model from the input data.

Outlier detection method: Statistical methods

Detection of Univariate Outliers Based on Normal Distribution

Data: a city's average temperature values in July in the last 10 years:

24.0°C, 28.9°C, 28.9°C, 29.0°C, 29.1°C, 29.1°C, 29.2°C, 29.2°C, 29.3°C, and 29.4°C

Assumption: the data follows a normal distribution, which is determined by two parameters: the mean, μ, and the standard deviation, σ

Using maximum likelihood method, we can estimate the parameters μ and σ.

As $\mu \pm 3\sigma$ region contains 99.7% data under the assumption of normal distribution, any object that is more than 3σ away from the mean of the estimated distribution can be considered outliers.

Outlier detection method: Statistical methods

Detection of Univariate Outliers Based on Normal Distribution (Contd.)

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

$$\hat{\mu} = \frac{24.0 + 28.9 + 28.9 + 29.0 + 29.1 + 29.1 + 29.2 + 29.2 + 29.3 + 29.4}{10} = 28.61$$

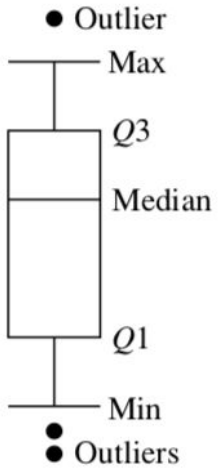
$$\begin{aligned}\hat{\sigma}^2 = ((&24.1 - 28.61)^2 + (28.9 - 28.61)^2 + (28.9 - 28.61)^2 + (29.0 - 28.61)^2 \\ &+ (29.1 - 28.61)^2 + (29.1 - 28.61)^2 + (29.2 - 28.61)^2 + (29.2 - 28.61)^2 \\ &+ (29.3 - 28.61)^2 + (29.4 - 28.61)^2)/10 \backsimeq 2.29.\end{aligned}$$

Outlier detection method: Statistical methods

Detection of Univariate Outliers Based on Normal Distribution (Contd.)

In the example, the most deviating value 24C is 4.61C (more than 3σ) away from the estimated mean. Thus, it can be considered an outlier.

This can be also visualized using a box plot.



Outlier detection method: Statistical methods

Detection of Univariate Outliers Based on Normal Distribution (Contd.)

The Grubb's test (aka the maximum normed residual test) can also be used for detecting univariate outliers detection using normal distribution.

For each object x in a data set, we define a z-score as $z = \frac{|x - \bar{x}|}{s}$,

where $\bar{x}$ is the mean, and s is the standard deviation of the input data. An object x is an outlier if

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

where $t^2_{\alpha/(2N),N-2}$ is the value taken by a t-distribution at a significance level of $\alpha/(2N)$, and N is the number of objects in the data set.

Outlier detection method: Statistical methods

Detection of Multivariate Outliers

Idea: transform the multivariate outlier detection task into a univariate outlier detection problem

Multivariate outlier detection using the Mahalanobis distance.

For an object, o, in the data set, the Mahalanobis distance from o to the mean vector $\bar{o}$ is

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

where S is the covariance matrix.

MDist(o,$\bar{o}$) is a univariate variable, and thus Grubb's test can be applied to this measure.

Outlier detection method: Statistical methods

Multivariate outlier detection using the Mahalanobis distance (Contd.)

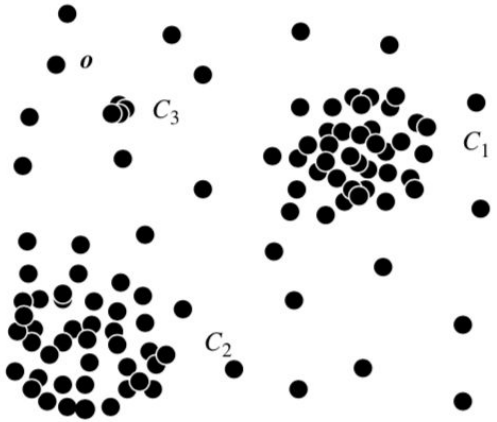
Steps:

1. Calculate the mean vector from the multivariate data set.
2. For each object o, calculate MDist(o,$\bar{o}$), the Mahalanobis distance from o to $\bar{o}$.
3. Detect outliers in the transformed univariate data set, {MDist(o,$\bar{o}$)|o $\in$ D}.
4. If MDist(o,$\bar{o}$) is determined to be an outlier, then o is regarded as an outlier as well.

Outlier detection method: Statistical methods

Using a Mixture of Parametric Distributions

When the actual data distribution is complex, we assume that the data were generated by a mixture of parametric distributions.



Outlier detection method: Statistical methods

Outlier detection using a histogram

Steps:

1. Histogram construction:
   Construct a histogram using the input data.
2. Outlier detection:
   Some approaches to determine whether an object,o, is an outlier:
   a. If the object falls in one of the histogram's bins, the object is regarded as normal. Otherwise, it is considered an outlier, or
   b. Assign an outlier score to the object. For example, score = inverse of the volume of the bin in which the object falls. The larger this score, the more likely o is an outlier.

Outlier detection method: Proximity-based approaches

Assumption:

The proximity of an outlier object to its nearest neighbors significantly deviates from the proximity of the object to most of the other objects in the data set.

Types:

- Distance-based
- Density-based

Distance-based outlier detection

- A distance-based outlier detection method consults the neighborhood of an object, which is defined by a given radius.
- An object is then considered an outlier if its neighborhood does not have enough other points.
- Let r ($r \geq 0$) be a *distance threshold* and π ($0 < \pi \leq 1$) be a *fraction threshold*. An object, o, is a DB(r,π)-outlier if

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

Distance-based outlier detection

Algorithm: Distance-based outlier detection.

Input:

- a set of objects $D = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_n\}$, threshold r ($r > 0$) and π ($0 < \pi \leq 1$);

Output: $DB(r,\pi)$ outliers in D.

Method:

for $i = 1$ to n **do**
  $count \leftarrow 0$
  for $j = 1$ to n **do**
    if $i \neq j$ **and** $dist(\boldsymbol{o}_i, \boldsymbol{o}_j) \leq r$ **then**
      $count \leftarrow count + 1$
      if $count \geq \pi \cdot n$ **then**
        exit {$\boldsymbol{o}_i$ cannot be a $DB(r,\pi)$ outlier}
      endif
    endif
  endfor
  print $\boldsymbol{o}_i$ {$\boldsymbol{o}_i$ is a $DB(r,\pi)$ outlier according to (Eq. 12.10)}
endfor;

Density-based outlier detection

- A density-based outlier detection method investigates the density of an object and that of its neighbors.
- An object is identified as an outlier if its density is relatively much lower than that of its neighbors.

Density-based outlier detection: Terminologies

- **k-distance of an object o, $\text{dist}_k(o)$:**
  Distance between o and its k-nearest neighbour
- **k-distance neighborhood of o, $N_k(o)$:**
  All objects of which the distance to o is not greater than $\text{dist}_k(o)$, i.e.,
  $N_k(o) = \{ o' \mid o' \text{ in } D, \text{dist}(o, o') \leq \text{dist}_k(o) \}$
  $N_k(o)$ may contain more than k objects because multiple objects may each be the same distance away from o.
- **Reachability distance from o' to o:**
  $\text{reachdist}_k(o \leftarrow o') = \max\{\text{dist}_k(o), \text{dist}(o,o')\}$.

Density-based outlier detection

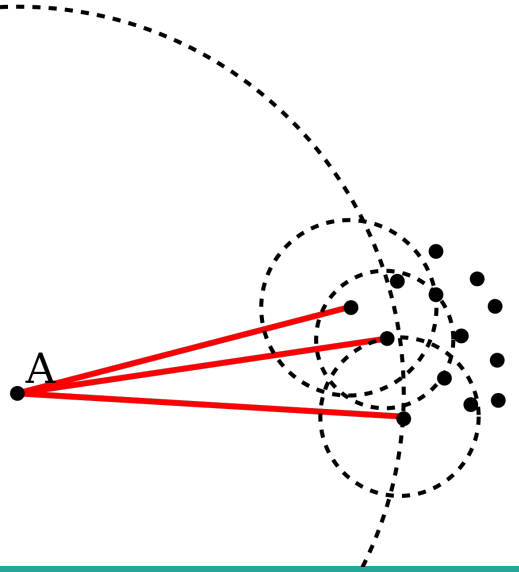
Local outlier factor of an object o:
The average of the ratio of the local reachability density of o and those of o's k-nearest neighbors.

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

$LOF_k(o) \sim 1$ means Similar density as neighbors,

$LOF_k(o) < 1$ means Higher density than neighbors (Inlier),

$LOF_k(o) > 1$ means Lower density than neighbors (Outlier)



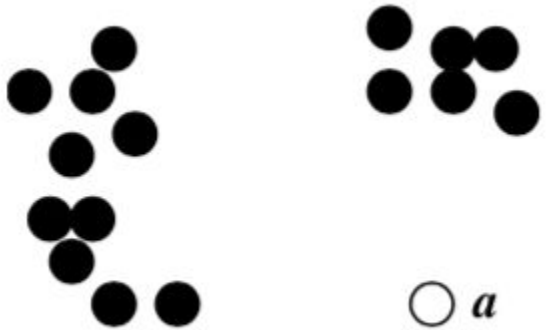
Outlier detection method: Clustering-based approaches

- Clustering-based outlier detection methods examine the relationship between objects and clusters.
- An object is an outlier if
  - It does not belong to any cluster
  - There is a large distance between the object and the cluster to which it is closest
  - It is a part of a small or sparse cluster

Clustering-based outlier detection

Detecting outliers as objects that do not belong to any cluster

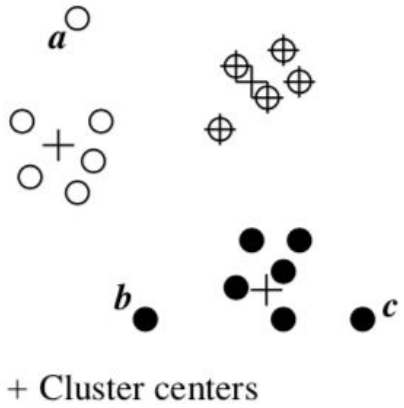
- Use a density-based clustering method, such as DBSCAN.



Clustering-based outlier detection

Clustering-based outlier detection using distance to the closest cluster

1. Partition the data using k-means
2. For each object, o, assign an outlier score to the object according to the distance between the object and the center, c_o, that is closest to the object.



$dist(o, c_o) / avg\_dist(c_o)$

Where $dist(o, c_o)$ is the distance between o and c_o and $avg\_dist(c_o)$ is the average distance between c_o and the objects assigned to o.
The larger the ratio, the more likely o is an outlier.

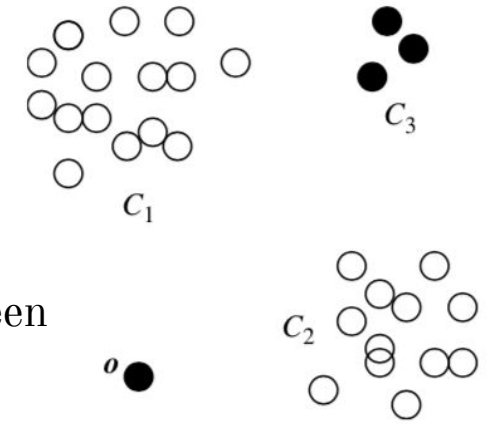
Clustering-based outlier detection

Detecting outliers in small clusters

1. Find clusters in a data set, and sort them according to decreasing size.
2. To each data point, assign a *cluster-based local outlier factor* (CBLOF).

   For a point belonging to a large cluster,
   CBLOF = The cluster's size × the similarity between the point and the cluster

   For a point belonging to a small cluster,
   CBLOF = The size of the small cluster × the similarity between the point and the closest large cluster



Outlier detection method: Classification-based approaches

Idea: Train a classification model that can distinguish normal data from outliers.

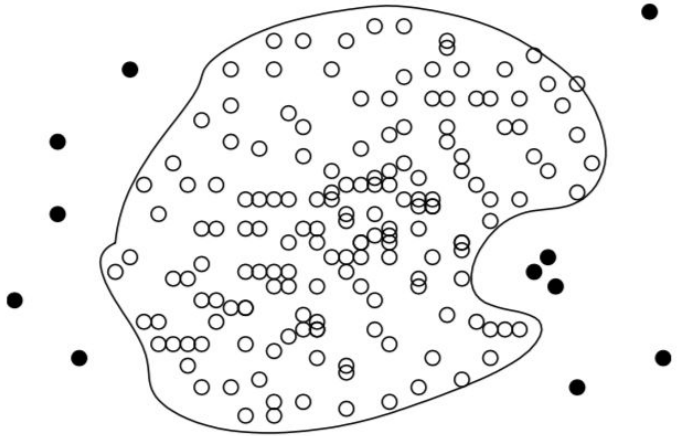
Methods:

1. One-class method
2. Semi-supervised learning

Outlier detection using a one-class model

To detect outliers using a one-class model, a classifier is built to describe only the normal class. Any samples that do not belong to the normal class are regarded as outliers.

We can learn the decision boundary of the normal class using classification methods such as SVM.



Outlier detection by semi-supervised learning

Objects are labeled as either “normal” or “outlier”, or have no label at all.

1. Using a clustering-based approach, find a large cluster, C, and a small cluster, C1.
2. Some objects in C carry the label “normal,” so we can treat all objects in this cluster (including those without labels) as normal objects.
3. We use the one-class model of this cluster to identify normal objects in outlier detection.
4. Some objects in cluster C1 carry the label “outlier,” we declare all objects in C1 as outliers.
5. Any object that does not fall into the model for C (e.g., a) is considered an outlier as well.