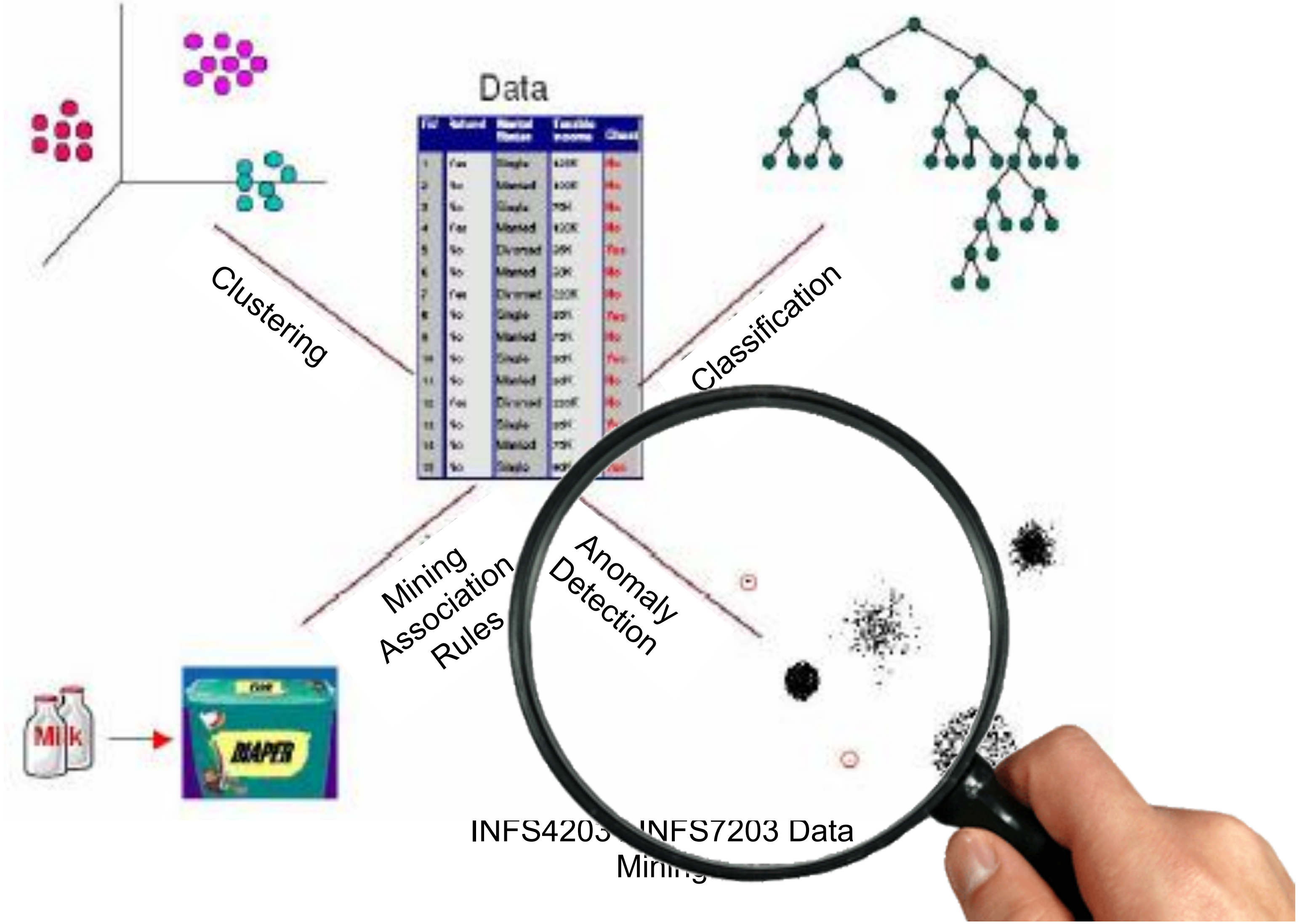
Data Mining Tasks

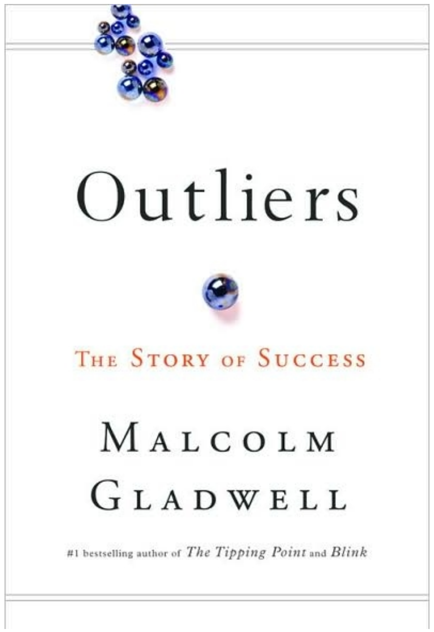


Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are **considerably different** than the remainder of the data

- Given a database D, find all the data points **x** $\in$ D with anomaly scores greater than some threshold t

- Applications:
  - Credit card fraud detection, fault detection, telecommunication fraud detection, network intrusion detection, …



Beyond Outliers

- Outliers are different from the noise data
  - Noise is random error
  - Noise should be removed before outlier detection

- Outliers are interesting: they violate the mechanism that generates the normal data

- Outlier detection vs. *novelty detection*: early stage, outlier; but later merged into the model

Anomaly Detection

- Challenges
  - Method is **unsupervised**
    - Validation can be quite challenging
      - just like for clustering
  - **How many** outliers are there in the data?
    - Finding needle in a haystack

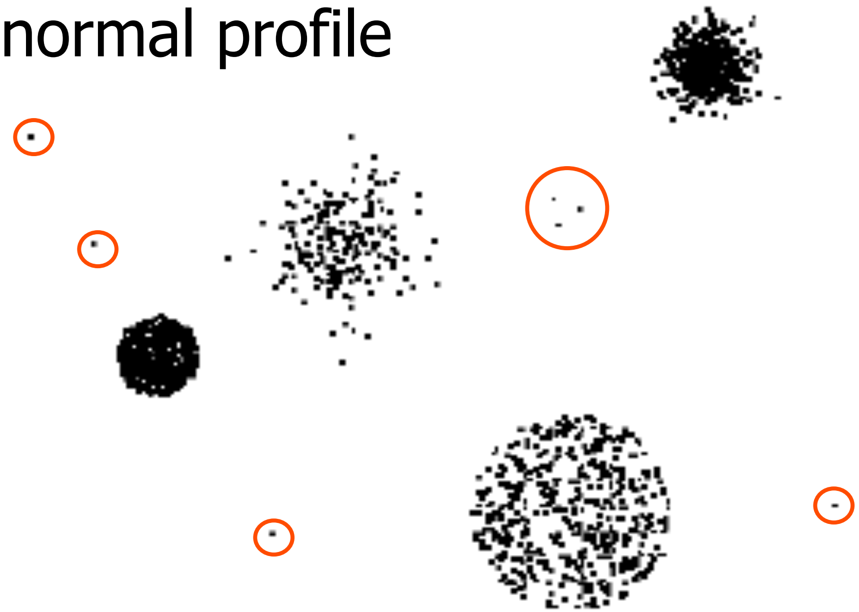
- Working assumption:
  - There are considerably **more** "normal" observations than "abnormal" observations (outliers/anomalies)

Anomaly Detection Schemes

- Build a **profile** of the “normal” behavior

- Use the “normal” profile to **detect** anomalies

  - Anomalies are observations whose characteristics differ significantly from the normal profile



Kinds of Outliers

- Global Outliers
- Contextual Outliers
- Collective Outliers

Global Outlier

- **Global outlier** (or point anomaly)
  - Object is O_g if it significantly deviates from the rest of the data set
  - E.g.: Intrusion detection in computer networks
  - Issue: Find an appropriate measurement of deviation

Contextual Outlier

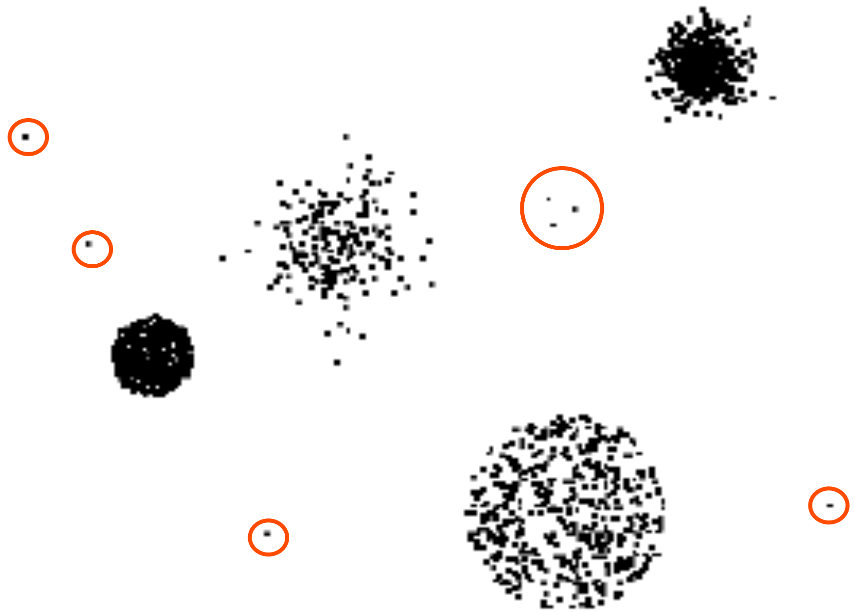
- **Contextual outlier** (or *conditional outlier*)
  - Object is O_c if it deviates significantly based on a **selected context**
  - Ex. 40^{o}C: outlier?
  - Attributes of data objects should be divided into two groups
    - Contextual attributes: defines the context, e.g., time & location
    - Behavioral attributes: characteristics of the object, used in outlier evaluation, e.g., temperature

Collective Outliers

- **Collective Outliers**
  - A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers
  - Applications: E.g., *intrusion detection*:
    - When a number of computers keep sending denial-of-service packages

Anomaly Detection Schemes

- Types of anomaly detection schemes
  1. Statistical-based
  2. Proximity-based
  3. Cluster-based

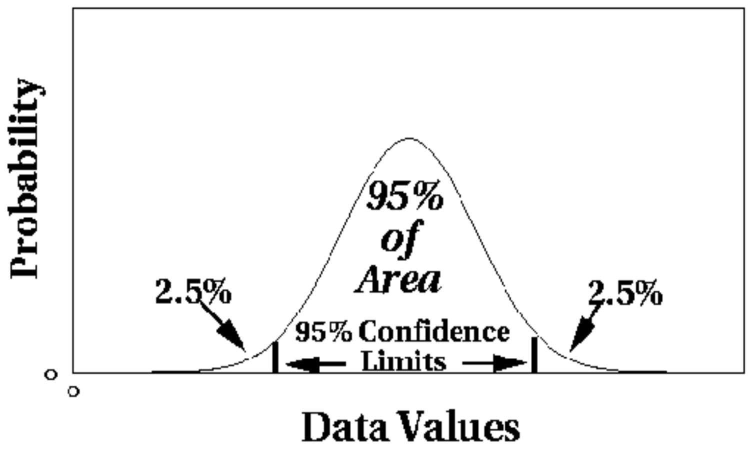


Statistical Schemes

- Statistical approaches assume that the objects in a data set are generated by a stochastic process (a generative model)
- Idea: learn a generative model fitting the given data set, and then identify the objects in low probability regions of the model as outliers
- Methods are divided into two categories:
  - Parametric
  - Non-parametric

Statistical Schemes - Parametric

- Assume a parametric model describing the distribution of the data
  - Example: normal distribution
- Apply a statistical test that depends on
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)



Example

- Ex: Avg. temp.: {24.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4}
  - Use the maximum likelihood method to estimate μ and σ
  - Taking derivatives with respect to μ and σ^2, we derive the following maximum likelihood estimates

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  - For the above data with n = 10, we have $\hat{\mu} = 28.61$ $\hat{\sigma} = \sqrt{2.29} = 1.51$
  - Then |24 – 28.61| /1.51 = 3.04 > 3, 24 is an outlier since $\mu \pm 3\sigma$ region contains 99.7% data

Limitations of Parametric Schemes

- In many cases, data distribution may not be known

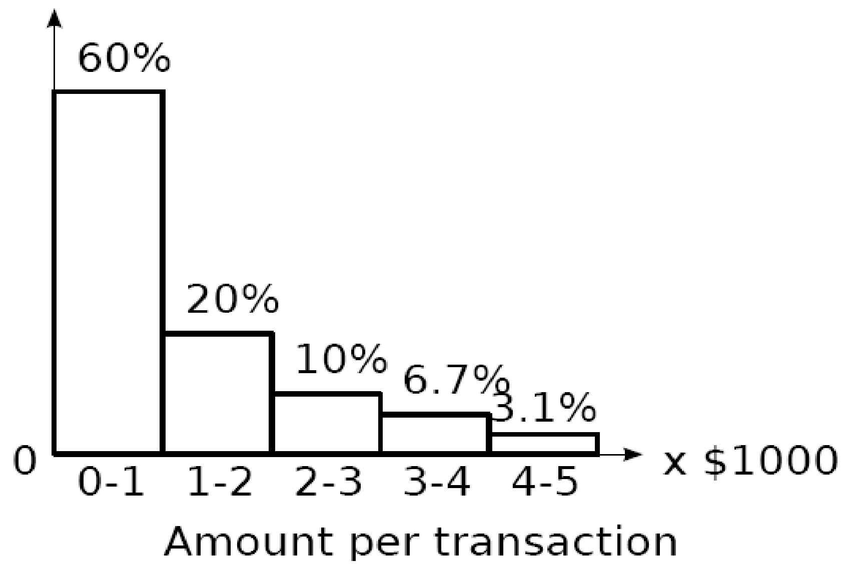
- Most of the tests are for a single attribute
  - For high dimensional data, it may be difficult to estimate the true distribution

Statistical Schemes – Non-Parametric

- The model of normal data is learned from the input data without any a priori structure.
- Often makes fewer assumptions about the data, and thus can be applicable in more scenarios
- Outlier detection using histogram

Example

- Figure shows the histogram of purchase amounts in transactions
- A transaction in the amount of $7,500 is an outlier:
  - only 0.2% transactions have an amount higher than $5,000

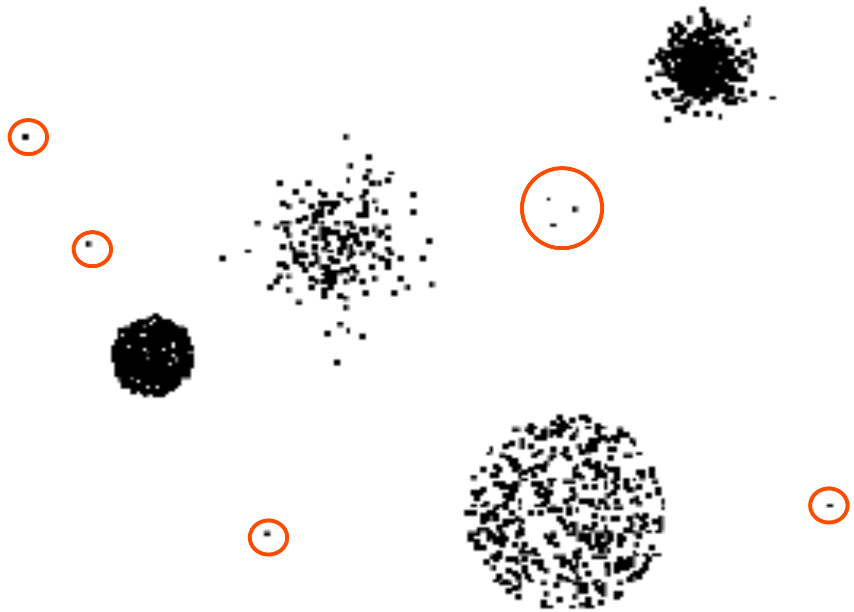


Histogram Challenge

- Problem: Hard to choose an appropriate bin size for histograms
- Too small bin size:
  - normal objects in empty/rare bins: false positive
- Too big bin size:
  - outliers in some frequent bins: false negative

Anomaly Detection Schemes

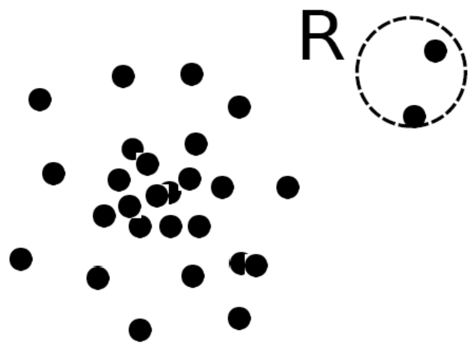
- Types of anomaly detection schemes
  1. Statistical-based
  2. Proximity-based
  3. Cluster-based



Distance-Based Schemes

- An object is an outlier if the nearest neighbors of the object are far away
  - the **proximity** of the object **significantly deviates** from the proximity of most of the other objects in the same data set

- Example: Model the proximity of an object using its **3** nearest neighbors
  - Objects in region **R** are substantially different from other objects in the data set.
  - Thus the objects in **R** are outliers



Distance-Based vs. Density-Based Outlier Detection

- Two types of proximity-based outlier detection methods
  - **Distance-based outlier detection:**
    - An object **o** is an outlier if its neighborhood does not have enough other points
  - **Density-based outlier detection:**
    - An object **o** is an outlier if its density is relatively much lower than that of its neighbors

Distance-Based Outlier Detection

- For each object o,
  - examine the number of other objects in the *r*-neighborhood of o
    - *r* is a user-specified **distance threshold**
  - an object o is an outlier if
    - **most** of the objects in D are far away from o (i.e., not in the r-neighborhood of o)

Distance-Based Outlier Detection

- An object o is a **DB(r, π)** outlier if $\frac{\|\{o' | dist(o, o') \leq r\}\|}{\|D\|} \leq \pi$

where π is a fraction threshold

- Equivalently, one can check the distance between ***o*** and its *k*-th nearest neighbor o_k

where $k = \lceil \pi \|D\| \rceil$

o is an outlier if dist(o, o_k) > r

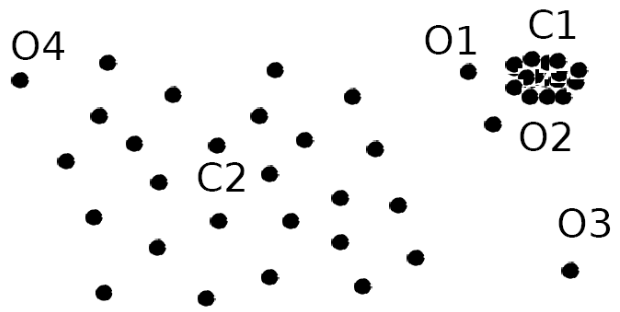
Computation

- Efficient computation: Nested loop algorithm
- For any object o_i:
  1. calculate its distance from other objects, and
  2. count the number of objects in its r-neighborhood.
- If πD other objects are within r distance, then terminate the inner loop
- Else, o_i is a DB(r, π) outlier

- Efficiency: Actually CPU time is not $O(n^2)$ but linear to the data set size since for most non-outlier objects, the inner loop terminates early

Density-Based Outlier Detection

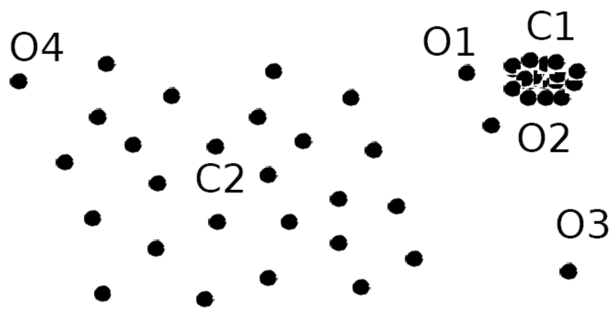
- **Local outliers:**
  - Outliers compared to their local neighborhoods, instead of the global data distribution



- Example: o_1 and o_2 are local outliers to C_1, o_3 is a global outlier, but o_4 is not an outlier.
- However, proximity-based clustering <u>cannot find</u> o_1 and o_2 are outliers

Density-Based Outlier Detection

- **Intuition:**
  - The **density** around an outlier object is significantly different from the density around its neighbors



- **Method:**
  - Use the **relative** density of an object against its neighbors as the indicator of the degree of the object being an outlier

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 ratio of the density of sample *p* and the density of its nearest neighbors

- Outliers are points with lowest LOF value

Density

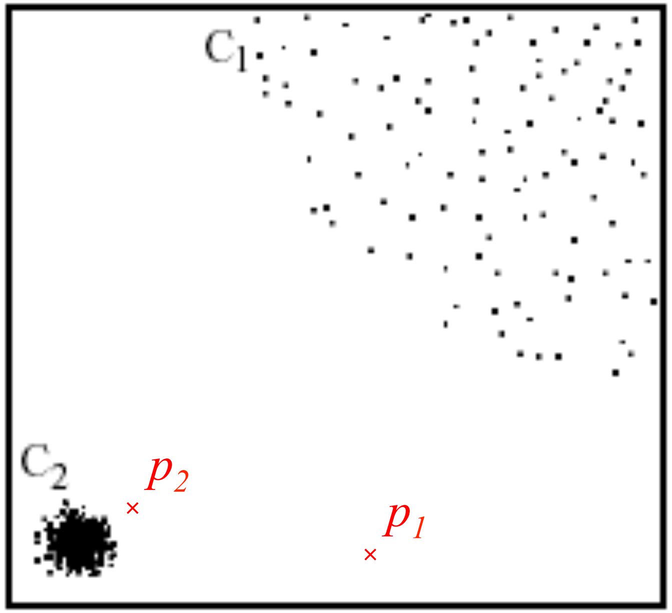
- Several methods to measure the density of a point *x*

  - Density = k / distance to the k nearest neighbors, or

$$density(x,k) = \left(\frac{\sum_{y \in N(x,k)} \text{distance}(x, y)}{|N(x,k)|} \right)^{-1}$$

Where *N(x, K)* is the set containing the k nearest neighbors of *x*

Density-based: LOF approach



In the **distance-based** approach:

•p_2 is <u>not considered</u> as outlier,

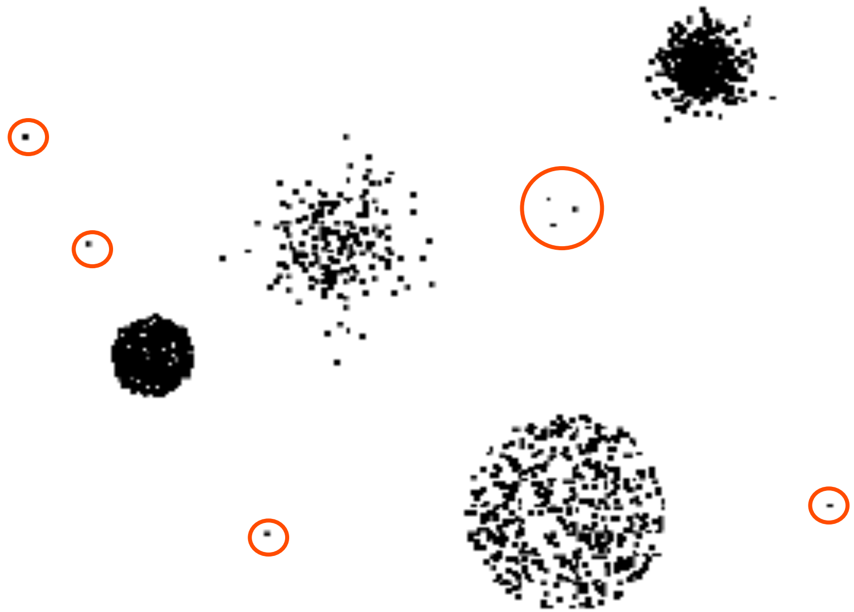
•LOF approach finds both p_1 and p_2 as outliers

Proximity-Based Methods

- The effectiveness of proximity-based methods highly <u>relies on the proximity measure</u>
- In some applications, proximity or distance measures cannot be obtained easily
- Often have a difficulty in finding a group of outliers which stay close to each other

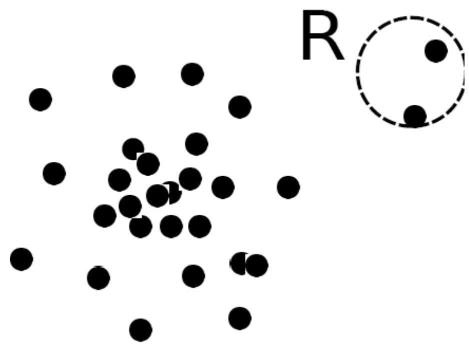
Anomaly Detection Schemes

- Types of anomaly detection schemes
  1. Statistical-based
  2. Proximity-based
  3. Cluster-based



Clustering-Based Methods

- Normal data belong to large and dense clusters
- An object is an outlier if:
  1. it does not belong to any cluster,
  2. there is a large distance between the object and its closest cluster , or
  3. it belongs to a small or sparse cluster

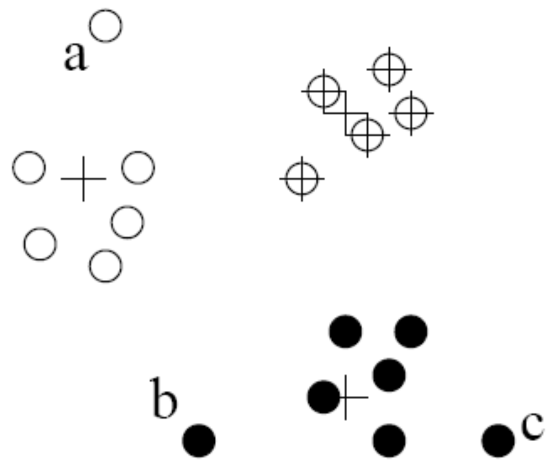


- Since there are many clustering methods, there are many clustering-based outlier detection methods as well

Clustering-Based Methods

- **Case 1:** Far from its closest cluster

- Using k-means, partition data points of into clusters
- For each object o, assign an outlier score based on its distance from its closest center
  - If dist(o, c_o)/avg_dist(c_o) is large, likely an outlier



Clustering-Based Methods

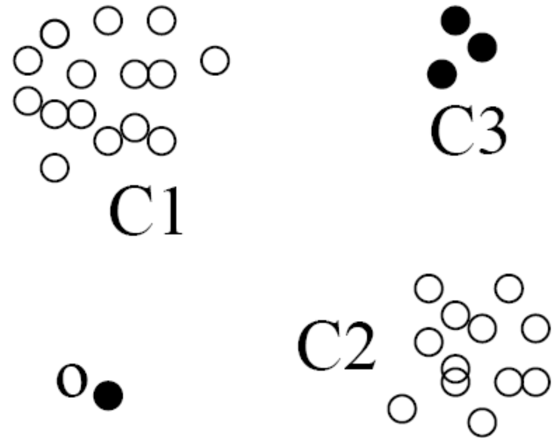
- **Case 2:** outliers in small clusters

- Find clusters, and sort them in decreasing size
- To each data point, assign a **cluster-based local outlier factor (CBLOF)**:

- If p belongs to a large cluster:
  - CBLOF = cluster size X similarity between p and cluster
- If p belongs to a small cluster:
  - CBLOF = cluster size X similarity between p and the closest large cluster

- The points with the <u>lowest</u> CBLOF scores are suspected outliers

Clustering-Based Methods



- For any point in C_3:
  - its closest large cluster is C_2, but its similarity from C_2 is low,
  - plus $|C_3| = 3$ is small

Clustering-Based Method: limitations

- Effectiveness depends highly on the clustering method used
- High computational cost: Need to first find clusters

Data Mining Tasks