Queen Mary
University of London

ECS766P Data Mining

Week 9: Outlier Detection

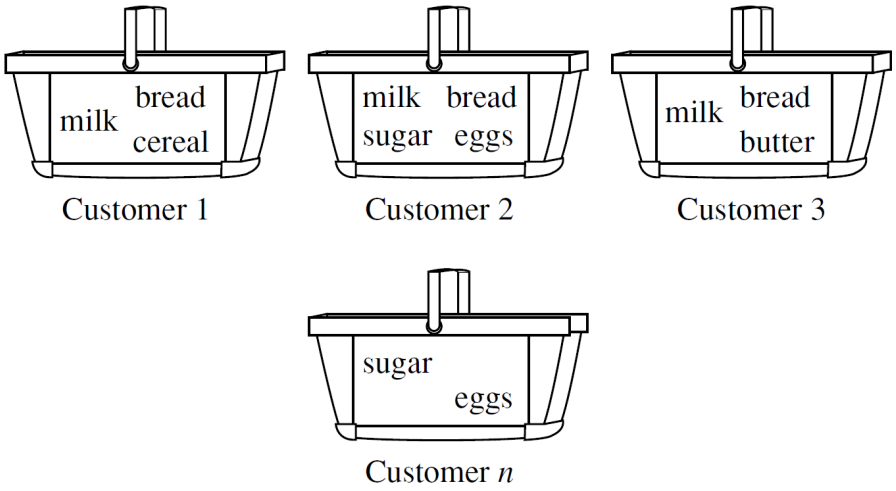
Emmanouil Benetos
emmanouil.benetos@qmul.ac.uk

November 2021

School of EECS, Queen Mary University of London

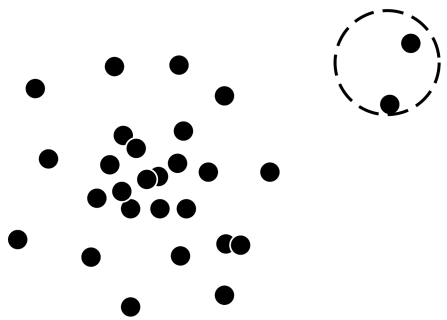
Last week: Association Analysis

- Frequent itemsets - basic concepts
- Frequent itemset mining methods
- Association rule mining
- Pattern evaluation methods



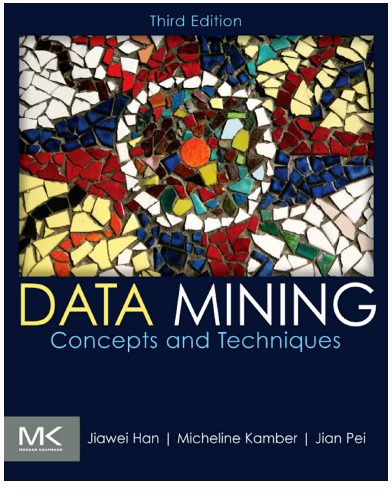
This week’s contents

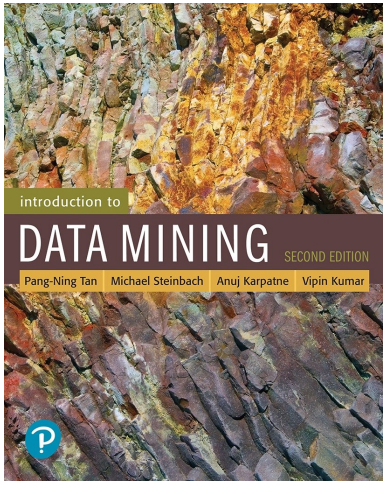
- Outliers and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Based Approaches
- Clustering-Based Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers

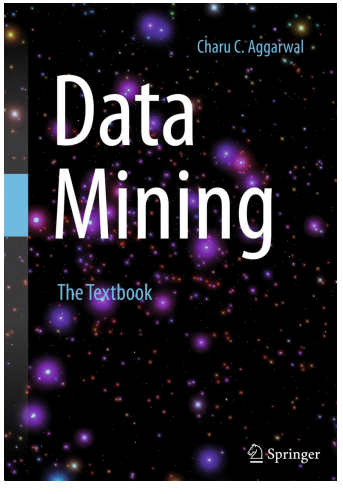


Reading

- **Chapter 12 of J. Han, M. Kamber, J. Pei, "Data Mining: Concepts and Techniques", 3rd edition, Elsevier/Morgan Kaufmann, 2012**
- Chapter 9 of P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, "Introduction to Data Mining", 2nd edition, Pearson, 2019
- Chapter 8 of C. C. Aggarwal, "Data Mining: The Textbook", Springer, 2015







Outliers and Outlier Analysis

What Are Outliers?

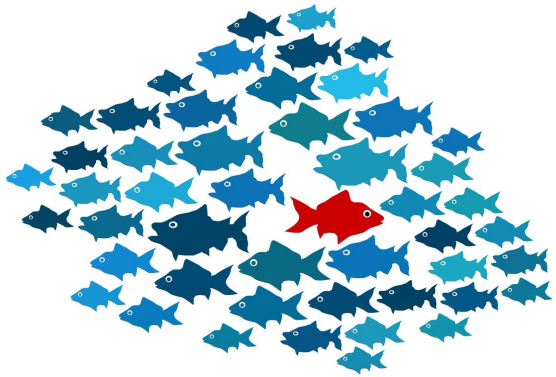
Outlier

A data object that deviates **significantly** from the normal objects as if it were **generated by a different mechanism**.

- Outliers are different from noise
- Outliers are interesting: they violate the mechanism that generates normal data
- Outlier detection vs. novelty detection

Applications of Outlier Detection

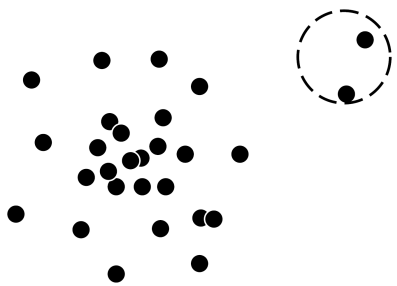
- Fraud detection
- Intrusion detection
- Ecosystem disturbances
- Medicine and public health
- Aviation safety



Types of Outliers

Global Outliers

- Also called point anomaly
- Object is a global outlier if it significantly deviates from the rest of the data set
- **Issue**: finding an appropriate measurement of deviation wrt the application in question.
- Simplest type of outlier; most outlier detection methods are aimed at finding global outliers.



Types of Outliers

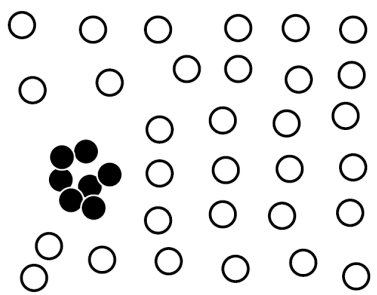
Contextual Outliers

- Also called conditional outliers
- Object is a conditional outlier if it deviates significantly based on a selected context.
- Attributes of data objects are divided into two groups:
  - Contextual attributes: defines the object's context
  - Behavioural attributes: define the object's characteristics, and are used in outlier evaluation
- **Issue**: How to define or formulate meaningful context?
- Global outlier detection can be viewed as a special case of contextual outlier detection.

Types of 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
- **Example**: intrusion detection – when a number of computers keep sending denial-of-service packages to each other
- Detection of collective outliers:
  - Consider not only behaviour of individual objects, but also that of groups of objects
  - Need to have background knowledge on the relationship among data objects (e.g. distance, similarity)



Challenges of Outlier Detection

Outlier detection is useful in many applications yet faces many challenges such as the following:

- Modeling normal objects and outliers properly
- Application-specific outlier detection
- Handling noise in outlier detection
- Interpretability

Outlier Detection Methods

Outlier Detection Methods

Two ways to categorize outlier detection methods:

- Based on whether user-labeled examples of outliers can be obtained:
  Supervised, semi-supervised vs. unsupervised methods
- Based on assumptions about normal data and outliers:
  Statistical, proximity-based, and clustering-based methods

Supervised Methods

- **Modeling outlier detection as a classification problem**
  - Samples examined by domain experts used for training & testing
- **Methods for learning a classifier for outlier detection effectively**:
  - Model normal objects & report those not matching the model as outliers, or
  - Model outliers and treat those not matching the model as normal
- **Challenges**
  - Imbalanced classes, i.e. outliers are rare: Boost the outlier class and make up some artificial outliers
  - Catch as many outliers as possible, i.e. recall is more important than accuracy (not mislabeling normal objects as outliers)

Unsupervised Methods

- Assume the normal objects are somewhat **clustered** into multiple groups, each having some distinct features
- An outlier is expected to be far away from any groups of normal objects
- **Weakness**: cannot detect collective outliers effectively
- Many clustering methods can be adapted for unsupervised methods
  - Find clusters, then outliers: not belonging to any cluster
  - **Problem 1**: hard to distinguish noise from outliers
  - **Problem 2**: it is often costly to find clusters first and then find outliers

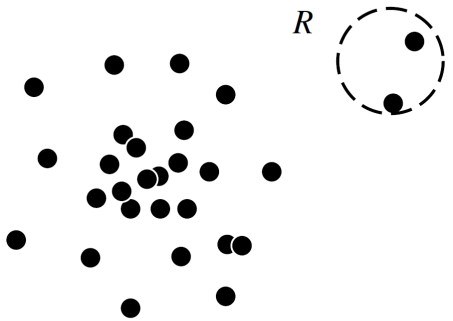
Semi-Supervised Methods

- **Situation**: In many applications, the number of labeled data is often small: labels could be on outliers only, normal objects only, or both.
- Semi-supervised outlier detection: Regarded as applications of semi-supervised learning
- If some labeled normal objects are available:
  - Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
  - Those not fitting the model of normal objects are detected as outliers
- If only some labeled outliers are available, a small number of labeled outliers many not cover the possible outliers well

Statistical Methods

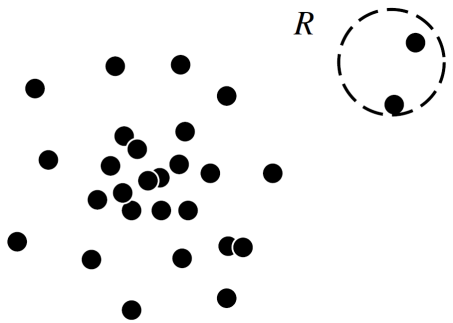
Statistical methods assume that the normal data follow some statistical model. The data not following the model are outliers.

- **Example**: first use a Gaussian distribution to model the normal data
  - For each object y in region R, estimate $g_D(y)$, the probability that y fits the Gaussian distribution.
  - If $g_D(y)$ is very low, y is unlikely to have been generated by the Gaussian model, thus is an outlier.
- **Effectiveness of statistical methods**: highly depends on whether the assumption of statistical model holds in the real data



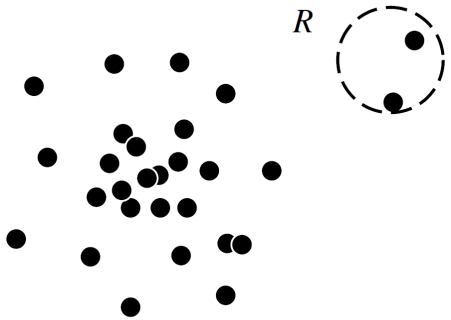
Proximity-Based Methods

- An object is an outlier if the **nearest neighbours** of the object are far away, i.e. 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 neighbours
  - Objects in region R are substantially different from other objects in the data set.
  - Thus the objects in R are outliers.
- The effectiveness of proximity-based methods highly relies on the proximity measure used.



Clustering-Based Methods

- **Assumption**: Normal data belong to large and dense clusters, whereas outliers belong to **small or sparse clusters**, or **do not belong to any clusters**.
- **Example**: two clusters
  - All points not in R form a large cluster
  - The two points in R form a tiny cluster, thus are outliers
- Clustering is expensive: straightforward adaption of a clustering method for outlier detection can be costly and does not scale up well for large data sets.



Statistical Approaches

Statistical Approaches

- 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** vs. **non-parametric**

Statistical Approaches

Parametric methods

- Assume that the normal data is generated by a distribution with parameters θ.
- The probability density function of the parametric distribution $f(x, \theta)$ gives the probability that object x is generated by the distribution.
- The smaller this value, the more likely x is an outlier.

Non-parametric methods:

- Not assume an a-priori statistical model and determine the model from the input data
- Not completely parameter free but consider the number and nature of the parameters are flexible and not fixed in advance
- **Examples**: histogram and kernel density estimation

Parametric Methods: Detection of Univariate Outliers

- Univariate data: a data set involving only one attribute or variable.
- Often assume that data is generated from a normal distribution.
- **Process**: learn the parameters from the input data, and identify the points with low probability as outliers.

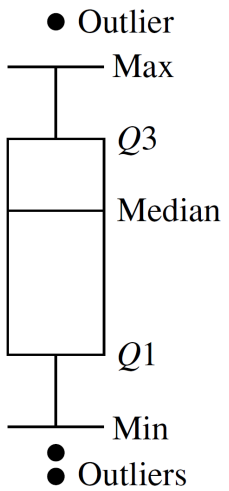
Parametric Methods: Detection of Univariate Outliers

Example: Suppose a city's average temperature values in July in the last 10 years are 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. Let's assume that the average temperature follows a normal distribution, which is determined by two parameters: the mean μ and the standard deviation σ.

- Using the maximum likelihood method, we estimate $\hat{\mu} = 28.61$ and $\hat{\sigma} = 1.51$ from the data.
- The most deviating value, 24.0°C, is 4.61°C away from the estimated mean.
- We know that the $\mu \pm 3\sigma$ region contains 99.7% of the data under the assumption of a normal distribution.
- Because $\frac{4.61}{1.51} = 3.04 > 3$, the probability that the value 24.0°C is generated by the normal distribution is less than 0.15%, and thus can be identified as an outlier.

Parametric Methods: Detection of Univariate Outliers

- The previous example elaborates a simple yet practical outlier detection method. It simply labels any object as an outlier if it is more than 3σ away from the mean of the estimated distribution.
- Such straightforward methods for statistical outlier detection can also be used in visualization, for example using **boxplots**.



Parametric Methods: Detection of Univariate Outliers

Grubb's Test

- Another statistical method for univariate outlier detection using normal distributions.
- For each object x in a data set, compute its z-score z. 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 **Student's t-distribution** at a significance level of $\alpha/(2N)$, and N is the number of objects in the data set.

Parametric Methods: Detection of Multivariate Outliers

- Multivariate data: A data set involving two or more attributes or variables
- **Goal**: Transform the multivariate outlier detection task into a univariate outlier detection problem
- **Method 1: Compute Mahalanobis distance**:
  - Let $\bar{\mathbf{x}}$ be the mean vector for a multivariate data set.
  - For object $\mathbf{x}$, the Mahalanobis distance from $\mathbf{x}$ to $\bar{\mathbf{x}}$ is:

$$MDist(\mathbf{x}, \bar{\mathbf{x}}) = (\mathbf{x} - \bar{\mathbf{x}})^T S^{-1} (\mathbf{x} - \bar{\mathbf{x}})$$

    where S is the covariance matrix.
  - $MDist(\mathbf{x}, \bar{\mathbf{x}})$ is a univariate variable, and thus Grubb's test can be applied to this measure.

Parametric Methods: Detection of Multivariate Outliers

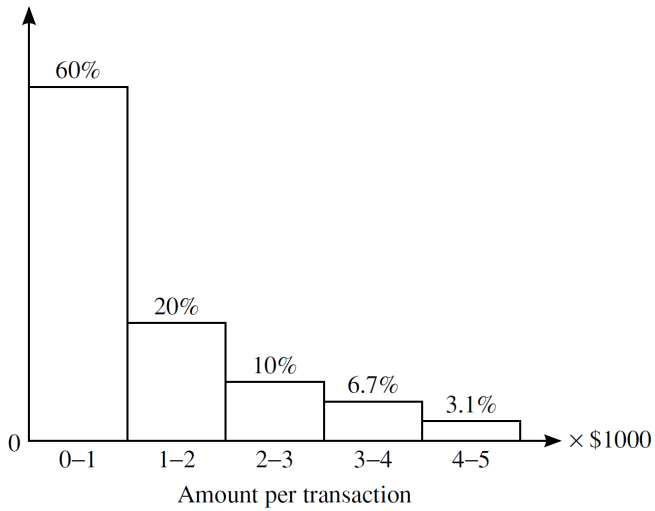
- **Method 2: Use χ^2 statistic**:
  - The χ^2-statistic can also be used to capture multivariate outliers under the assumption of a normal distribution.
  - For an object **o**, the χ^2 statistic is:

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

    where where o_i is the value of **o** on the i-th dimension, e_i is the mean of the i-th dimension among all objects, and n is the dimensionality.
  - If the χ^2-statistic is large, the object is an outlier.

Nonparametric Methods: outlier detection using histograms

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



Nonparametric Methods: outlier detection using histograms

- Problem: Hard to choose an appropriate bin size for the histogram.
- Too small bin size $\rightarrow$ normal objects fall in empty/rare bins, false positives
- Too large bin size $\rightarrow$ outliers in some frequent bins, false negatives
- **Statistical methods summary**:
  - Statistical methods for outlier detection learn models from data to distinguish normal data objects from outliers.
  - Advantage: the outlier detection may be statistically justifiable.
  - Disadvantage: The data distribution of high-dimensional data is often complicated and hard to fully understand.

Proximity-Based Approaches

Proximity-Based Approaches

- **Intuition**: Objects that are far away from other objects are outliers
- **Assumption**: The proximity of an outlier deviates significantly from that of most of the others in the data set.
- Two types of proximity-based outlier detection methods:
  - Distance-based outlier detection: An object **o** is an outlier if its **neighbourhood** 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 neighbours.

Distance-Based Outlier Detection

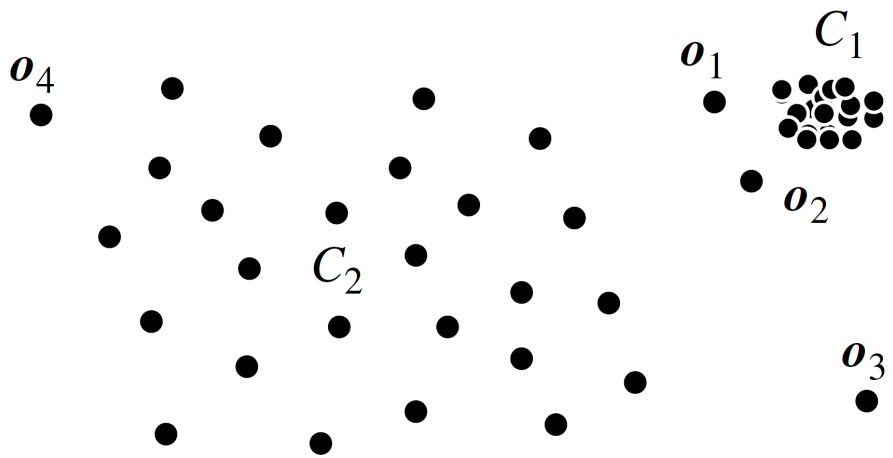
- For each object **o**, examine the number of other objects in the r-neighbourhood of **o**, where r is a user-specified **distance threshold**.
- An object **o** is an outlier if most (taking π as a **fraction threshold**) of the objects in dataset D are far away from **o**, i.e. not in the r-neighbourhood of **o**.
- Formally, an object **o** is a $DB(r, \pi)$ outlier if:

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

  where $dist(\cdot)$ is a distance measure and $o' \in D$.
- Efficient computation: Nested loop algorithm

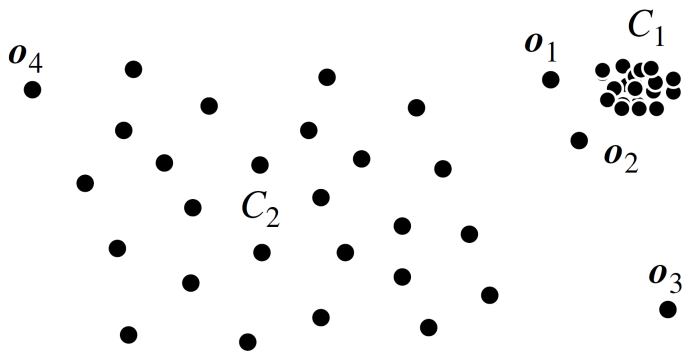
Density-Based Outlier Detection

- Local outliers: Outliers compared to their local neighbourhoods, instead of the global data distribution.
- In the below figure, o_1 and o_2 are local outliers to cluster C_1. However, distance-based methods cannot capture local outliers such as o_1 and o_2.



Density-Based Outlier Detection

- **Intuition**: the density around an outlier object is significantly different from the density around its neighbours.
- **Method**: use the relative density of an object against its neighbours as the indicator of the degree of the object being outliers.
- Relative density is measured based on the concept of the k-distance of **o**, defined as the distance between **o** and its k-th nearest neighbour.



Clustering-Based Approaches

Clustering-Based Approaches

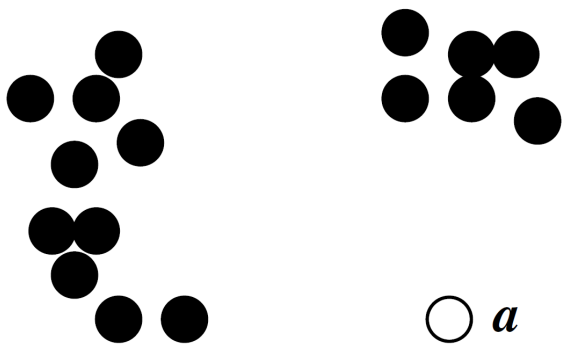
Three general approaches to clustering-based outlier detection:

1. Object does not belong to any cluster.
2. There is a large distance between the object and its closest cluster.
3. Object belongs to a small or sparse cluster.

Clustering-Based Approaches

Case 1: Object not belonging to any cluster

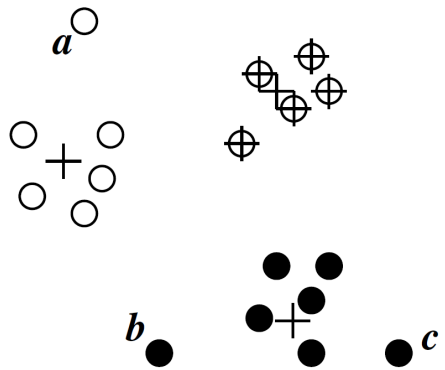
- **Approach**: Using a density-based clustering method (e.g. DBSCAN algorithm), we identify objects that belong to clusters.



Clustering-Based Approaches

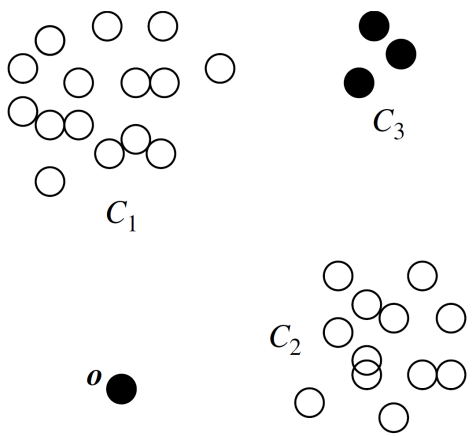
Case 2: There is a large distance between the object and its closest cluster.

- Using k-means, partition data points of into clusters.
- For each object $\mathbf{o}$, assign an outlier score based on its distance from its closest center $\mathbf{c}_o$, called $dist(\mathbf{o}, \mathbf{c}_o)$.
- Compute $l(\mathbf{c}_o)$ as the average distance between $\mathbf{c}_o$ and the objects assigned to $\mathbf{o}$.
- If $\frac{dist(\mathbf{o},\mathbf{c}_o)}{l(\mathbf{c}_o)}$ is large, $\mathbf{o}$ is likely an outlier.



Case 3: Detecting Outliers in Small Clusters

- FindCBLOF algorithm
- Find clusters, and sort clusters in decreasing size.
- To each data point, assign a cluster-based local outlier factor (CBLOF).
- If object **o** belongs to a large cluster, CBLOF = cluster_size $\times$ similarity between **o** and cluster
- If **o** belongs to a small cluster, CBLOF = cluster_size $\times$ similarity between **o** and the closest large cluster.



Clustering-Based Approaches: Pros and Cons

Advantages:

- Detect outliers without requiring any labeled data
- Work for many types of data
- Once the clusters are obtained, we only compare a new object against the clusters

Disadvantages:

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

A method to reduce the cost: Fixed-width clustering

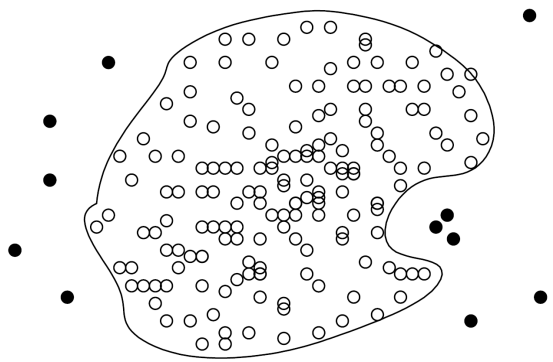
Classification Approaches

Classification-Based Methods

- **Idea**: Train a classification model that can distinguish "normal" data from outliers.
- A brute-force approach: Consider a training set that contains samples labeled as "normal" and others labeled as "outlier"
- However, the training set is typically heavily biased: the number of "normal" samples likely far exceeds the number of outlier samples.
- Cannot detect unseen outliers.

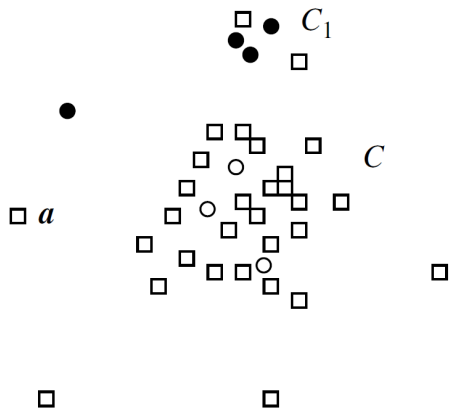
Classification-Based Methods: One-Class Model

- One-class model: A classifier is built to describe only the normal class.
- Learns the decision boundary of the normal class using classification methods.
- Any samples that do not belong to the normal class (not within the decision boundary) are declared as outliers.
- **Advantage**: can detect new outliers that may not appear close to any outlier objects in the training set.



Classification-Based Methods: Semi-Supervised Learning

- Combining **classification-based** and **clustering-based** methods
- Using a clustering-based approach, find a large cluster C and a small cluster C_1.
- We can treat all objects in C as “normal ”.
- Use the one-class model of this cluster to identify normal objects in outlier detection.
- Since some objects in cluster C_1 carry the label “outlier”, we declare all objects in C_1 as outliers.
- Any object that does not fall into the model for C (such as object **a**) is considered an outlier as well.



○ Objects with label “normal” ● Objects with label “outlier” □ Objects without label

Classification-Based Methods: Pros and Cons

Advantages:

- Methods incorporate domain knowledge into the detection process by learning from labeled samples.
- Once system is trained, outlier detection is fast.

Disadvantages:

- Quality heavily depends on the availability and quality of the training set.
- It is often difficult to obtain representative and high-quality training data.

Mining Contextual and Collective Outliers

Transforming into Conventional Outlier Detection

If the context can be clearly identified, transform problem to conventional outlier detection:

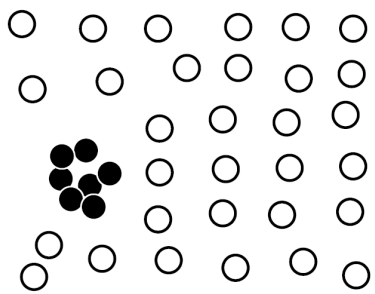
- Identify the context of the object using the contextual attributes.
- Calculate the outlier score for the object in the context using a conventional outlier detection method.
- **Example**: Detecting outlier customers in the context of customer groups.
  Contextual attributes: age group, postal code
  Behavioural attributes: # of transactions/year, annual total transaction amount.

Modeling Normal Behaviour with Respect to Contexts

- In some applications, one cannot clearly partition the data into contexts
- Model the “normal” behaviour with respect to contexts.
- Using a training data set, train a model that predicts the expected behaviour attribute values with respect to the contextual attribute values.
- An object is a contextual outlier if its behaviour attribute values significantly deviate from the values predicted by the model.
- Using a prediction model that links the contexts and behaviour, these methods avoid the explicit identification of specific contexts.

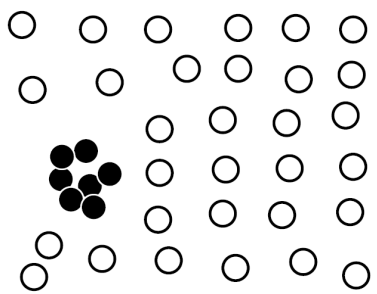
Mining Collective Outliers

- Collective outliers: if the objects as a whole deviate significantly from the entire data set, even though each individual object in the group may not be an outlier.
- Need to examine the **structure of the data set**, i.e. the relationships between multiple data objects.
- Each of these structures is inherent to its respective type of data
  - For **temporal data**, we explore structures which occur in segments of the time series or subsequences.
  - For **spatial data**, explore local areas.
  - For **graph and network data**, we explore subgraphs.



Mining Collective Outliers

- **Difference from contextual outlier detection**: the structures are often not explicitly defined, and have to be discovered as part of the outlier detection process.
- **Methods**:
  1. Reduce the problem to conventional outlier detection
  2. Models the expected behaviour of structure units directly
- Collective outlier detection is subtle due to the challenge of exploring the structures in data.



Summary

- Types of outliers: global, contextual & collective outliers
- Outlier detection: supervised, semi-supervised, or unsupervised
- Statistical (or model-based) approaches
- Proximity-based approaches
- Clustering-based approaches
- Classification approaches
- Mining contextual and collective outliers

Questions?
also please use the forum on QM+