Outlier Detection

Zuzana Pekarčíková

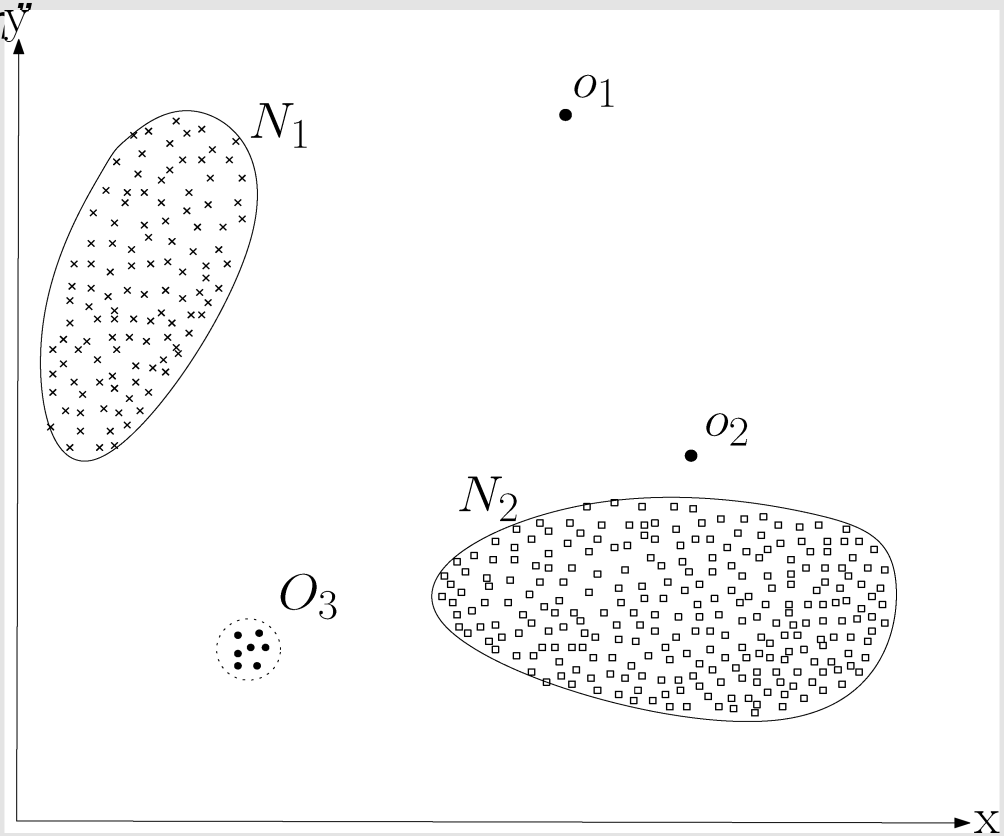
Outline

- What is an Outlier ?
- Applications of Outlier Detection
- Types of Outliers
- Outlier Detection Methods Types
- Basic Outlier Detection Methods
- High-dimensional Outlier Detection Methods
- Class Outlier Detection – Random Forests
- Context-based Approach

What is an Outlier ?

- Definition of Hawkins [Hawkins 1980]:

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”



Applications of Outlier Detection

- Fraud detection

Purchasing behavior of a credit card owner usually changes when the card is stolen

- Medicine

Unusual symptoms or test results may indicate potential health problems of a patient

Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, …)

- Detecting measurement errors

Data derived from sensors may contain measurement errors

Removing such errors can be important in other data mining and data analysis tasks

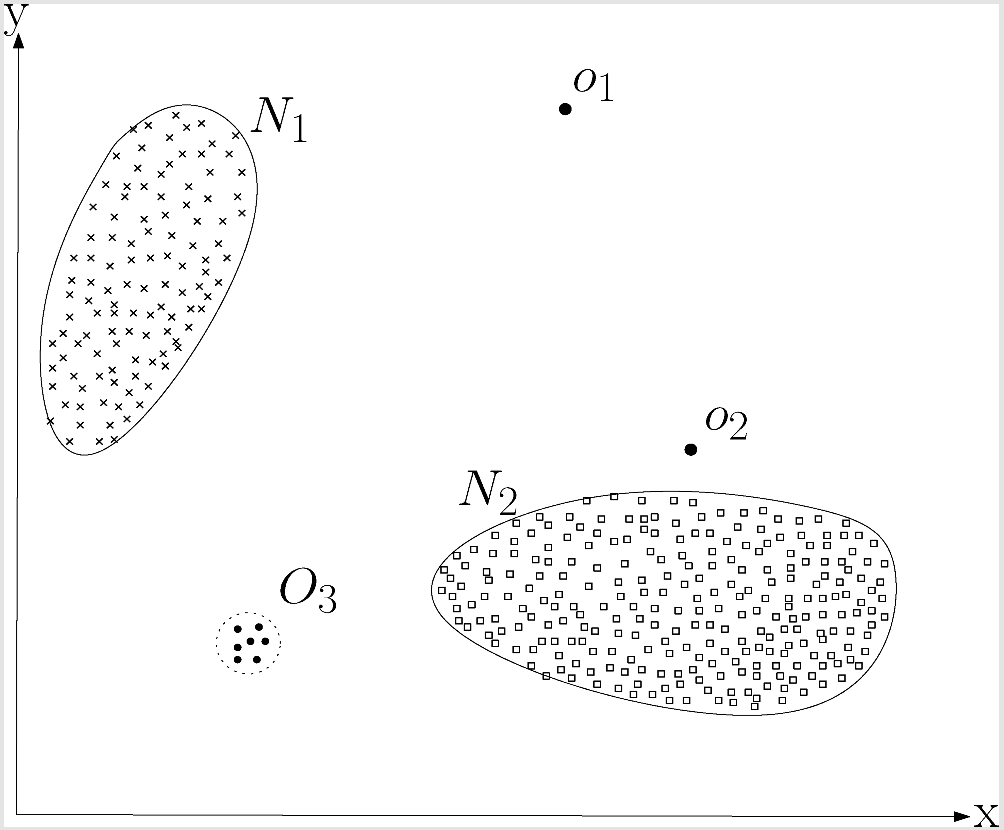
Types of Outliers

- Point Anomalies

An individual data instance can be considered as anomalous with respect to the rest of data.

The simplest type of Outliers.

Example: credit card fraud detection



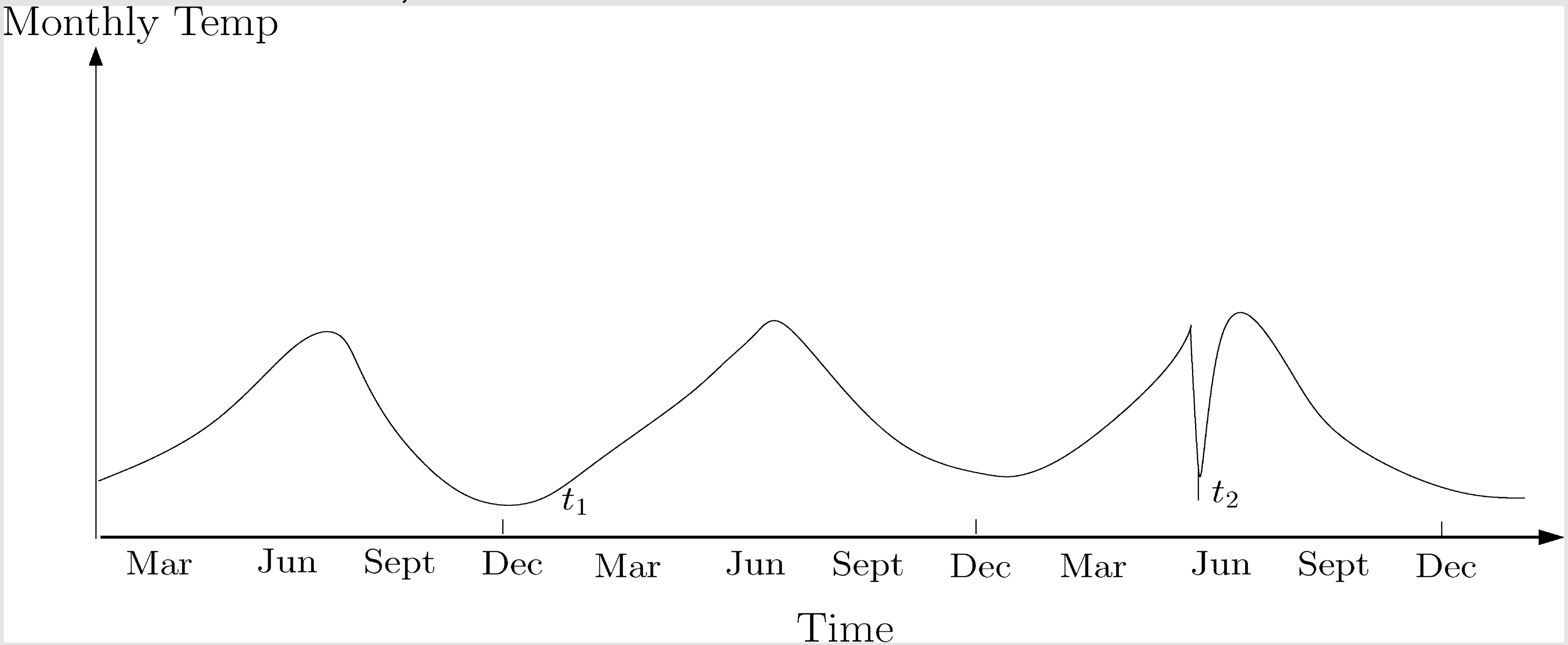
Types of Outliers

- Contextual Anomalies

If a data instance is anomalous in a specific context, but not otherwise.

Example: temperature time-series

- t1 = t2, however t1 is in winter whereas t2 in summer



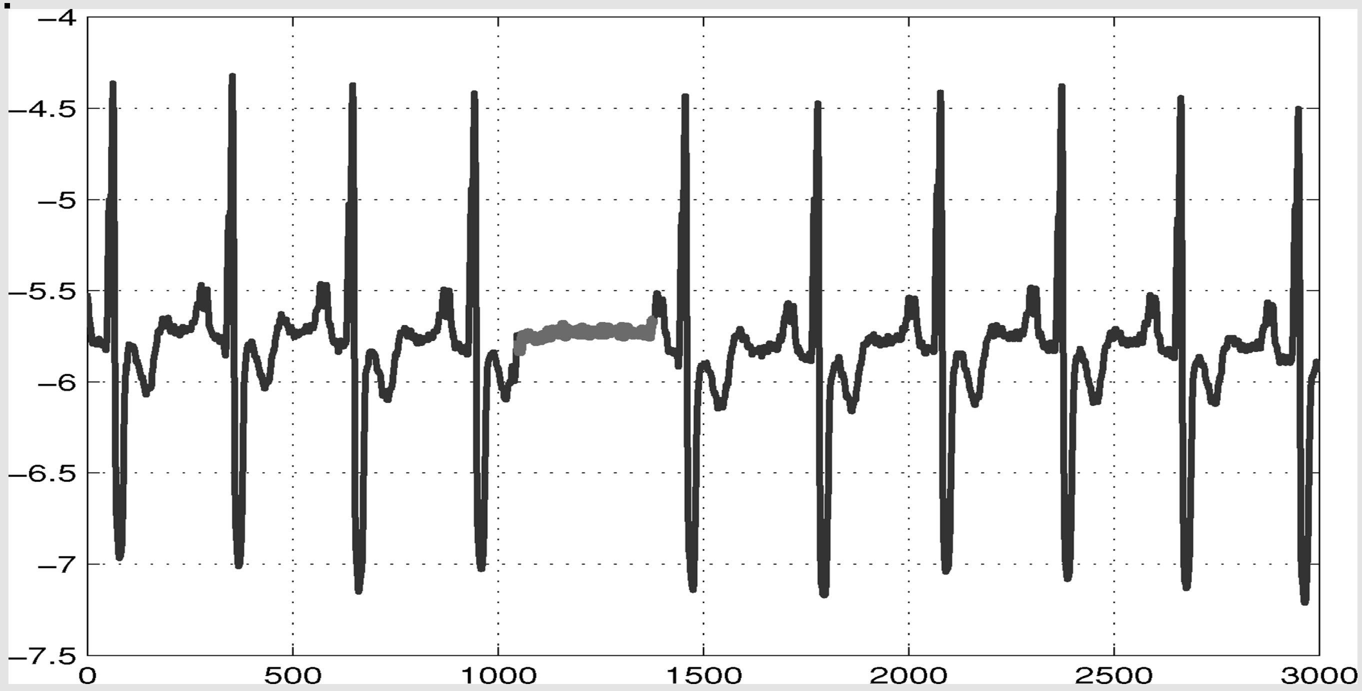
Types of Outliers

- Collective Anomalies

A collection of related data instances is anomalous with respect to the entire data set.

The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous.

Example: human cardiogram



Outlier Detection Methods Types

- The *labels* associated with a data instance denote whether that instance is normal or anomalous.
- Supervised Methods
  - availability of a training data set that has labeled instances for normal

as well as anomaly classes

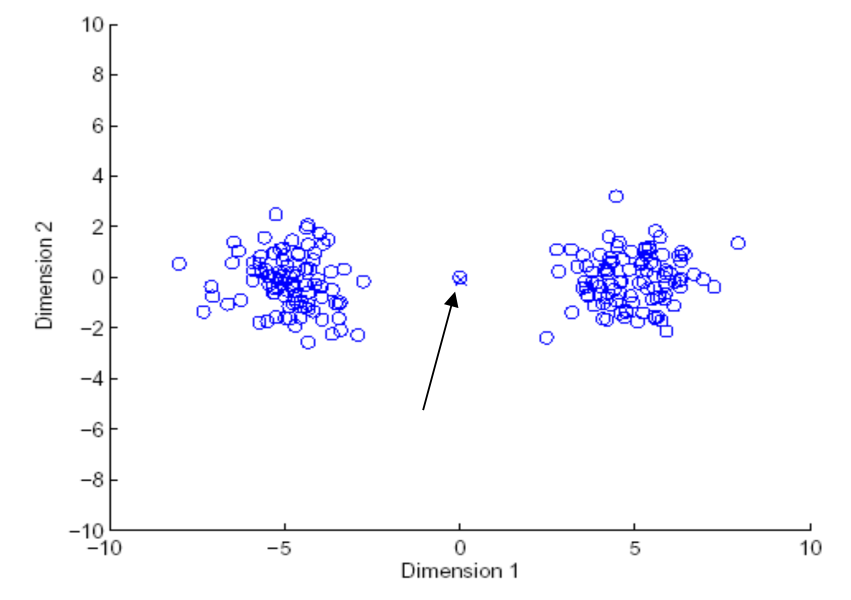
  - building a predictive model for normal vs. anomaly classes – problem is transformated to classification problem
  - Problems:
    - anomalous instances are far fewer than normal instances
    - obtaining acurate labels for the anomaly class is challenging
- Semi-supervised Methods
  - training data has labeled instances only for the normal class
- Unsupervised Methods
  - no labels, most widely used
  - assumption: normal instances are far more frequent than anomalies in the test data and they make clusters

Outlier Detection Methods

- Statistical Methods
  - normal data objects are generated by a statistical (stochastic) model, and data not following the model are outliers
  - Example: statistical distribution: Gaussina

    Outliers are points that have a low probability to be generated by Gaussian distribution
  - Problems: Mean and standard deviation are very sensitive to outliers

    These values are computed for the complete data set (including potential outliers)
  - Advantage: existence of statistical proof why the object is an outlier

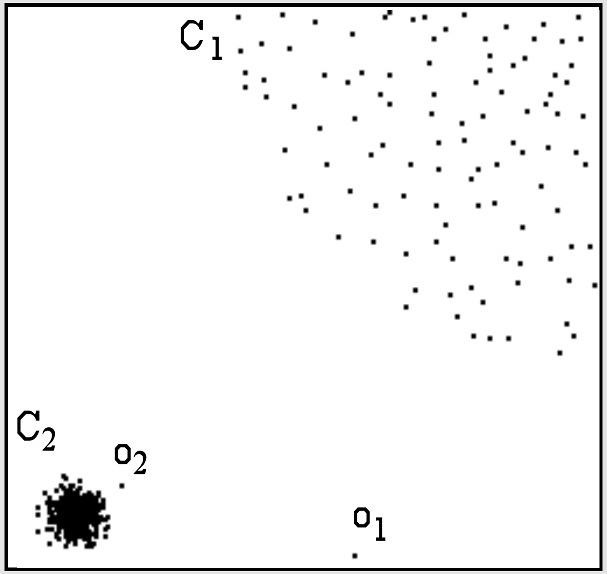


Outlier Detection Methods

- Proximity-Based Methods
- An object is an outlier if the proximity of the object to its neighbors significantly deviates from the proximity of most of the other objects to their neighbors in the same data set.
  - Distance-based Detection
    - Radius *r*
    - *k* nearest neighbors
  - Density-based Detection
    Relative density of object counted from density of its neighbors

- Clustering-Based Methods

Normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters.



Outlier Detection Methods

- Classification-Based Methods
  - Main idea: training a classification model that can distinguish normal data from outliers
  - Problem: imbalanced classes
  - Solution: using one-class model – classifier describe only the normal class and samples that do not belong to the normal class are regarded as outliers

Hight-dimensional Outlier Detection Methods

Problems in high-dimensional:

- Relative contrast between distances decreases with increasing dimensionality
- Data are very sparse, almost all points are outliers
- Concept of neighborhood becomes meaningless

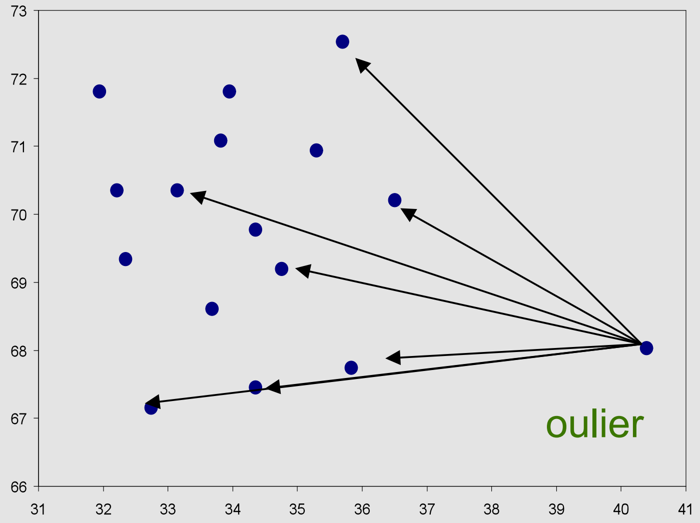
Solutions:

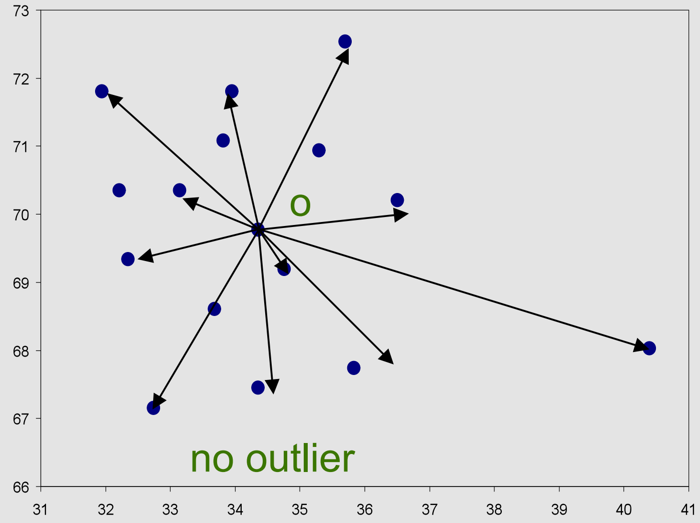
- Use more robust distance functions and find full-dimensional outliers
- Find outliers in projections (subspaces) of the original feature space

High-dimensional Outlier Detection Methods

ABOD – angle-based outlier degree

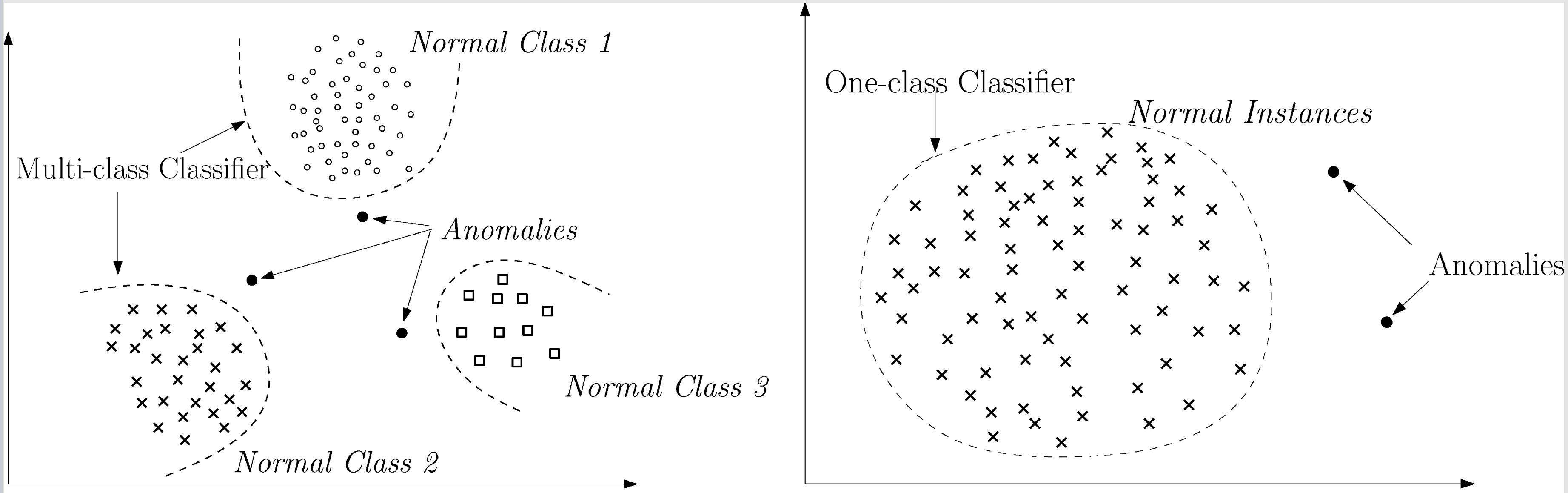
- Object o is an outlier if most other objects are located in similar directions
- Object o is no outlier if many other objects are located in varying directions





Class Outlier Detection

- ‘semantic outlier’
- A semantic outlier is a data point, which behaves differently with other data points in the same class, while looks normal with respect to data points in another class.



(a) Multi-class Anomaly Detection

(b) One-class Anomaly Detection

Class Outlier Detection

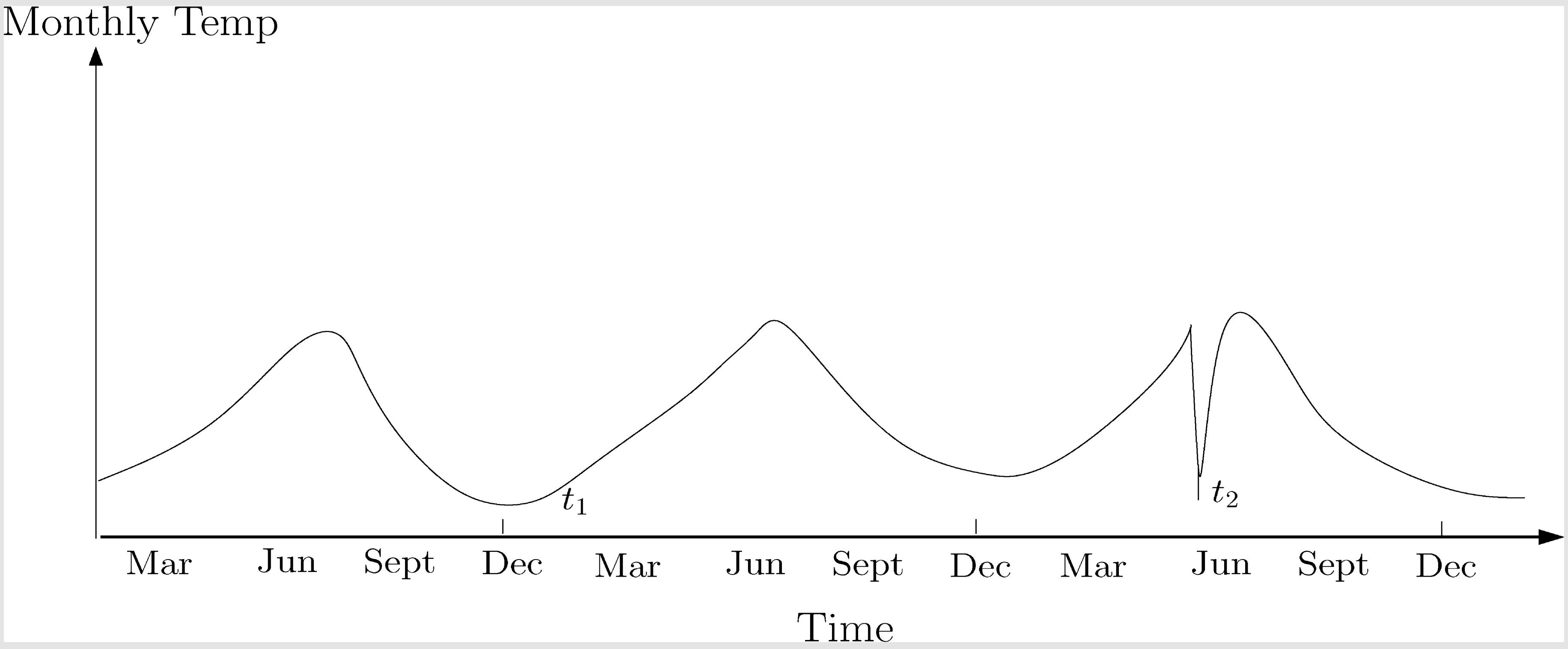
- Multi-class classification based anomaly detection techniques assume that the training data contains labeled instances belonging to multiple normal classes
- Anomaly detection techniques teach a classifier to distinguish between each normal class and the rest of the classes.

Class Outlier Detection – Random Forests

- ***Random Forests*** is an ***enensemble*** classification and regression approach.
- *Ensemble methods* use multiple models to obtain better predictive performance than could be obtained from any of the constituent models.
- Random Forests:
  - consists of many classification trees
  - 1/3 of all samples are left out – ***OOB (out of bag) data*** – for classification error
  - each tree is constructed by a different bootstrap sample from the original data
  - all data are run down the tree and proximities are computed for each pair of cases – These proximities are used for outlier detection.
  - Outliers are cases whose proximities to all other cases in the data are generally small.
  - Used in outliers relative to their class – an outlier in class j is a case whose prosimities to all other class j cases are small.

Context-based Approach

Is the temperature 28°C outlier?
If we are in Brno in summer NO
If we are in Brno in winter YES
→ it dependes on the location and time – CONTEXT



Context-based Approach

Contextual outlier significantly deviates from model with respect to a specific context of the object.

Generally the attributes of the data objects are divided into two groups:

- **Contextual attributes**: Define the object’s context. In the example, the contextual attributes may be date and location.
- **Behavioral attributes**: Define the object’s characteristics, and are used to evaluate whether the object is an outlier in the context to which it belongs. In the example, the behavioral attributes may be the temperature, humidity.

Contextual outlier detection methods can be devided into two categories according to whether the contexts can be clearly identified:

- Transforming Contextual Outlier Detection to Conventional Outlier Detection

  The context can be easily identified.
- Modeling Normal Behavior with Respect to Contexts

  The context identification is more difficult

Context-based Approach

Transforming Contextual Outlier Detection to Conventional Outlier Detection

General Idea:

Evaluation wheater the object is an outlier is done in two steps:

- identifycation the context of the object using the contextual attributes
- calculation the outlier score for the object in the context using a conventional outlier detection method

Example:

In customerrelationship management, we can detect outlier customers in the context of customer groups.

3 attributes:

- contextual: *age group (25, 25-45, 45-65, and over 65), post code*
- behavioral: *number of transactions per yer*

Is customer c outlier?

- locate the context of *c* using the attributes *age group* and *post code*
- compare *c* with the other customers in the same group, and use a conventional outlier detection method

Context-based Approach

Modeling Normal Behavior with Respect to Contexts

Context is not easy to identify

Example:

An online store records the sequence of products seached for by each customer. Outlier behavior is when customer suddenly purchased a product that is unrelated to those he/she recently browsed.

→ contexts cannot be easily specified because it is unclear how many products browsed earlier should be considered as the context, and this number will likely differ for each product

General idea:

modelation of normal behaviour with respect to contexts

With using a training data set a method trains a model that predicts the expected behavior attribute values with respect to the contextual attribute values.

Is an object outlier?

We apply the model to the contextual attributes of the object. If the behavior attribute values deviate from the values predicted by the model, then the object is a contextual outlier