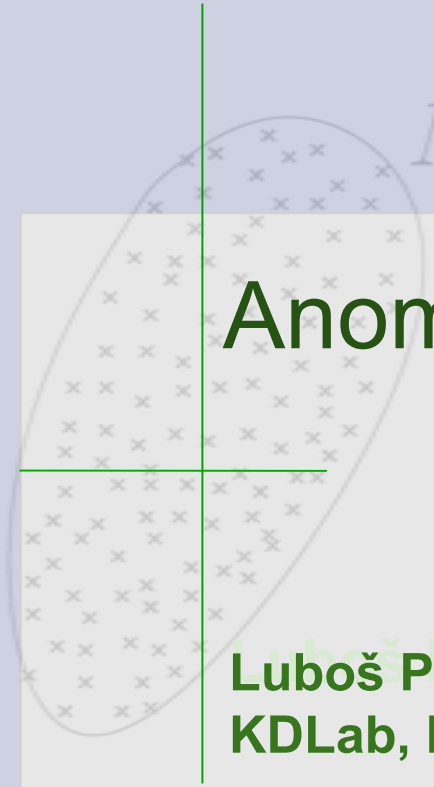


Anomaly detection

Luboš Popelínský
KDLab, Faculty of Informatics, MU

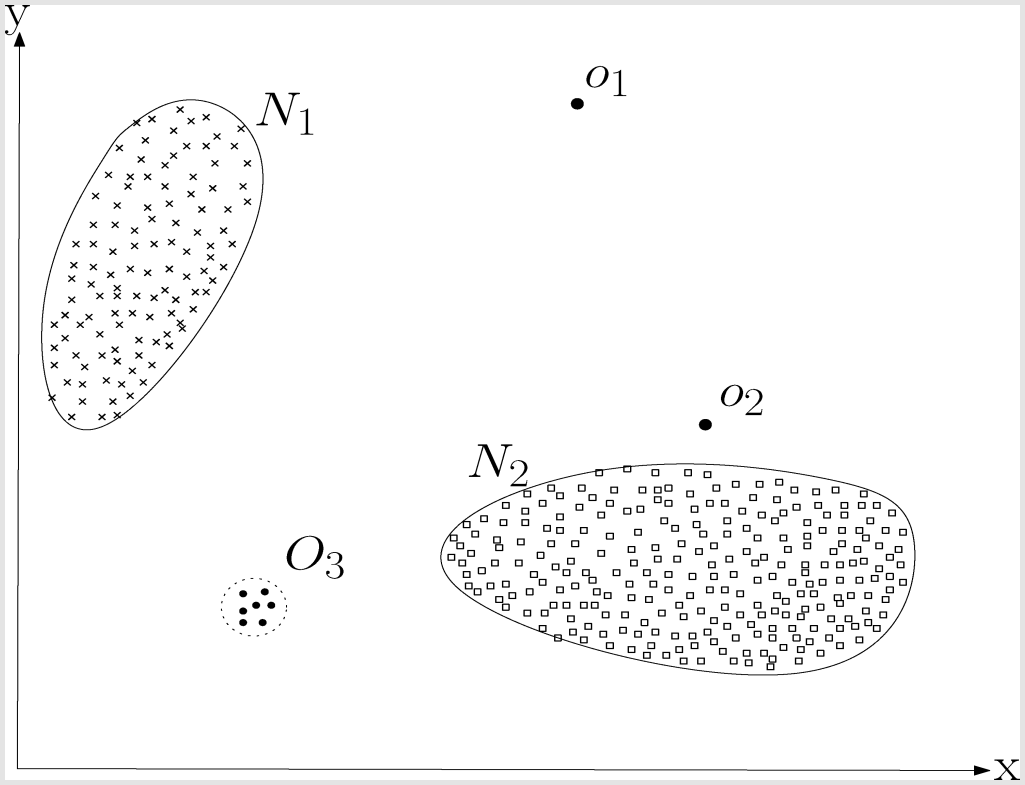
Thanks to Luis Torgo, Lea Nezvalová, Karel Vaculík and other members of the KDLab

KEG VŠE Mar 1, 2018

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

- Intrusion detection

- Language learning "irregularities"

Jedu do Porta. Jedu do hor. VS. *Jedu na hory.*

Types of Outliers

Point outliers

Cases that either individually or in small groups are very different from the others.

Contextual outliers

Cases that can only be regarded as outliers when taking the context where they occur into account.

Collective outliers

Cases that individually cannot be considered strange, but together with other associated cases are clearly outliers.

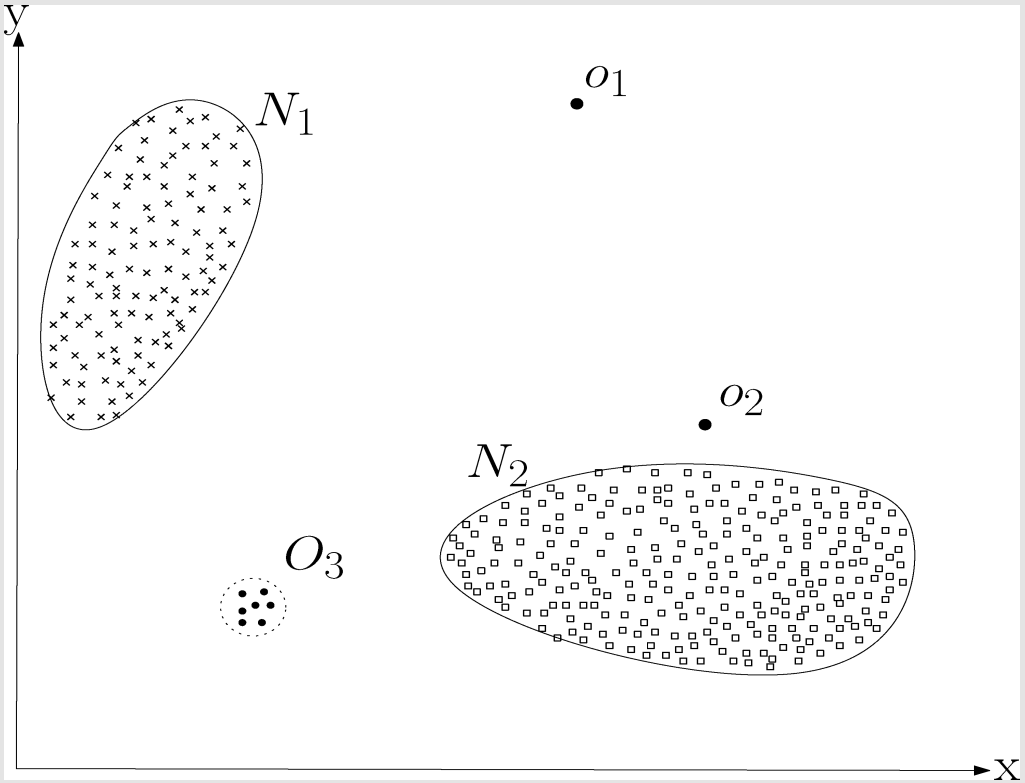
Types of Outliers

- 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



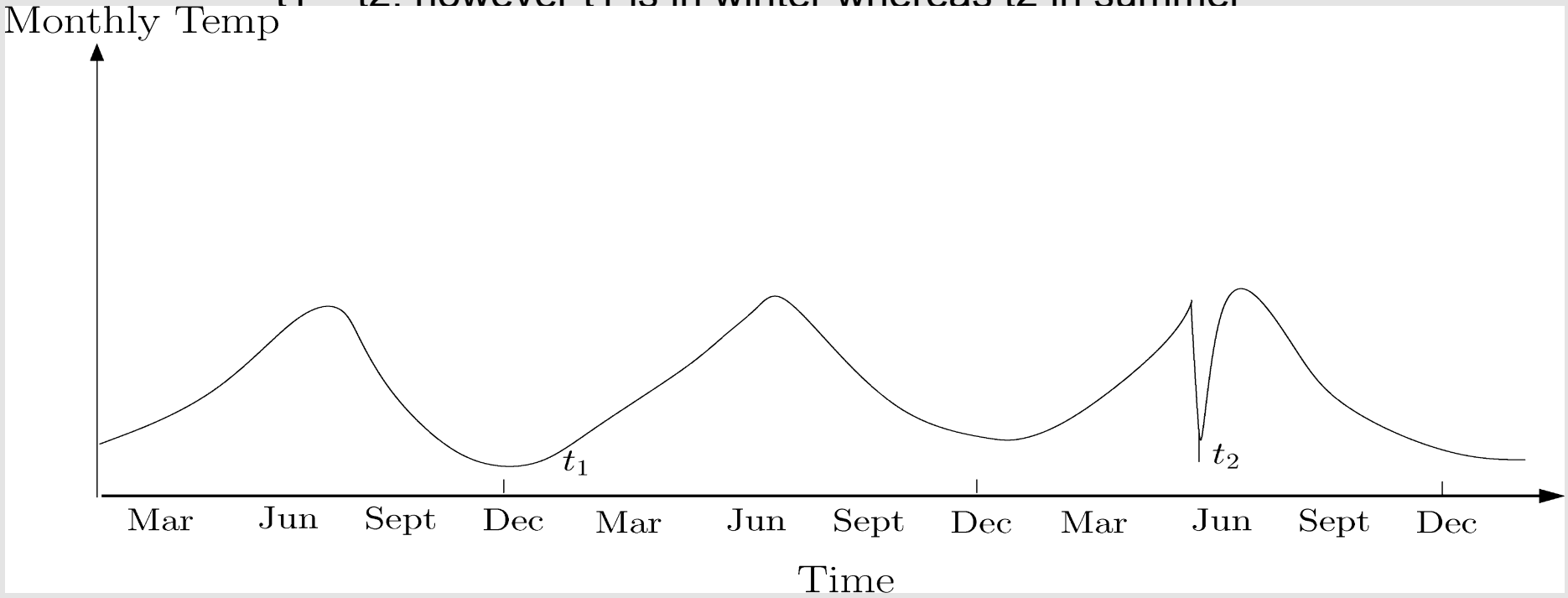
Types of Outliers

- Contextual Anomalies

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

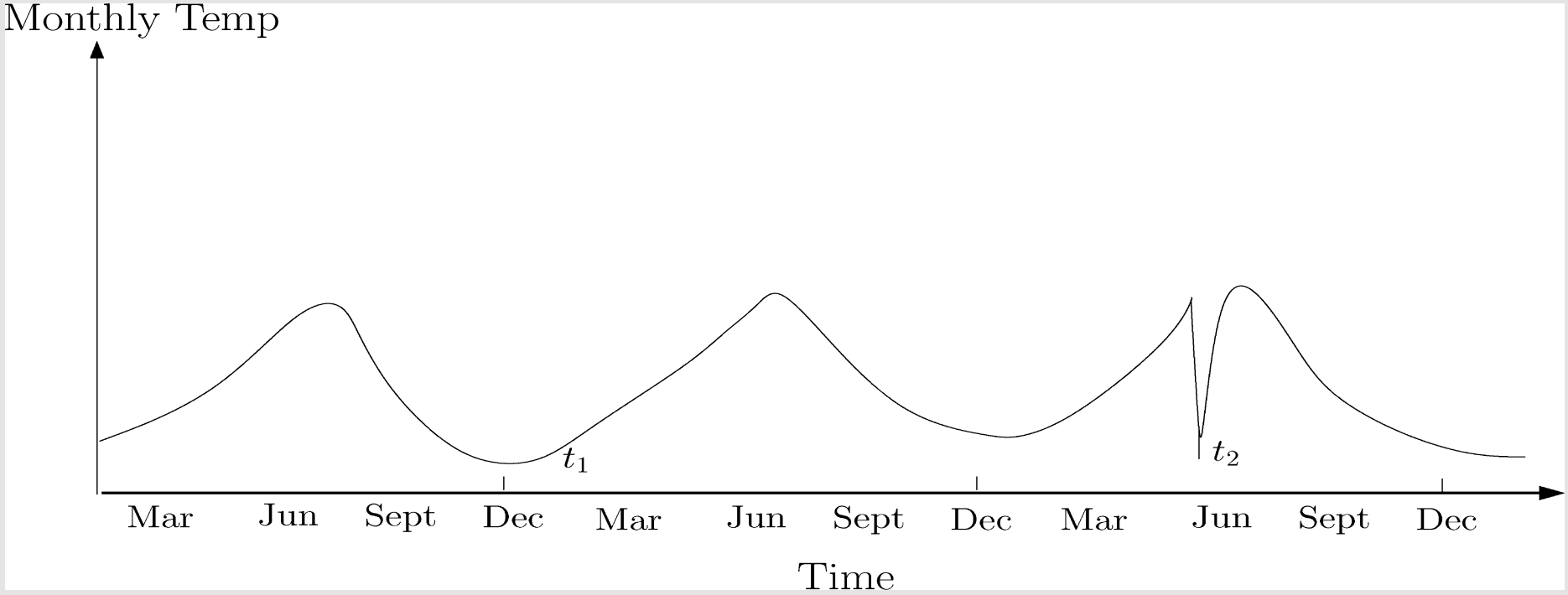
Example: temperature time-series

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



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



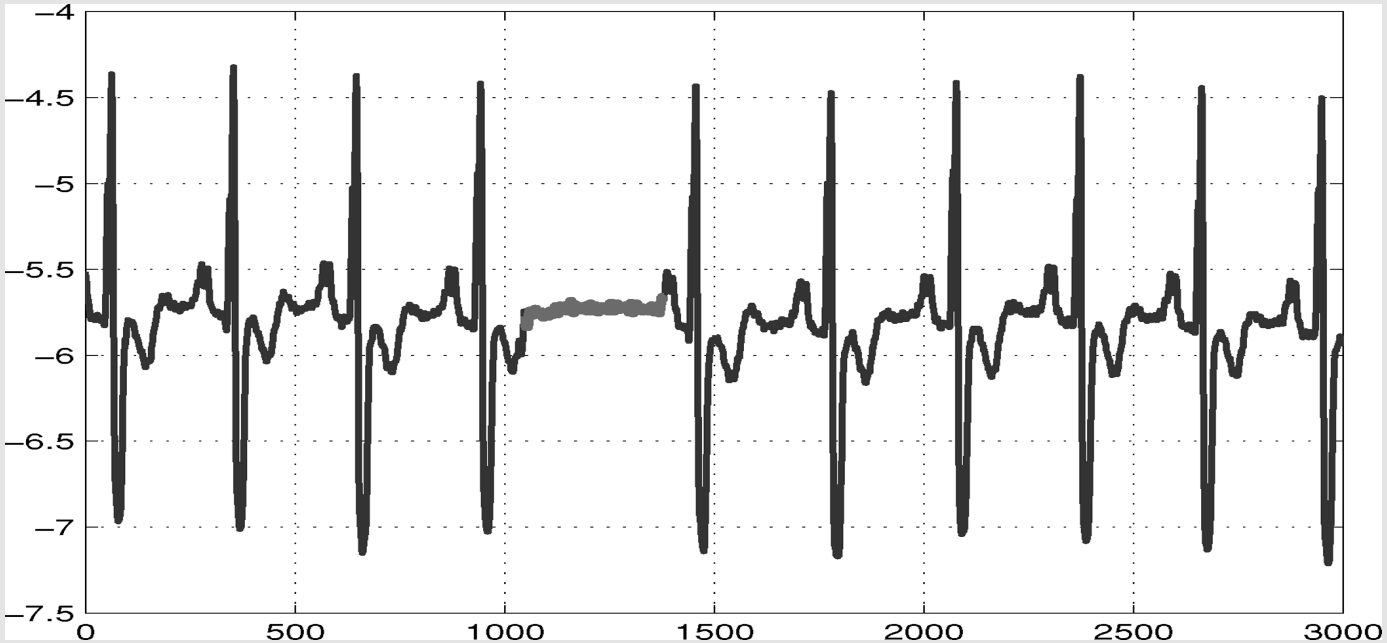
Types of Outliers

- Collective Anomalies

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

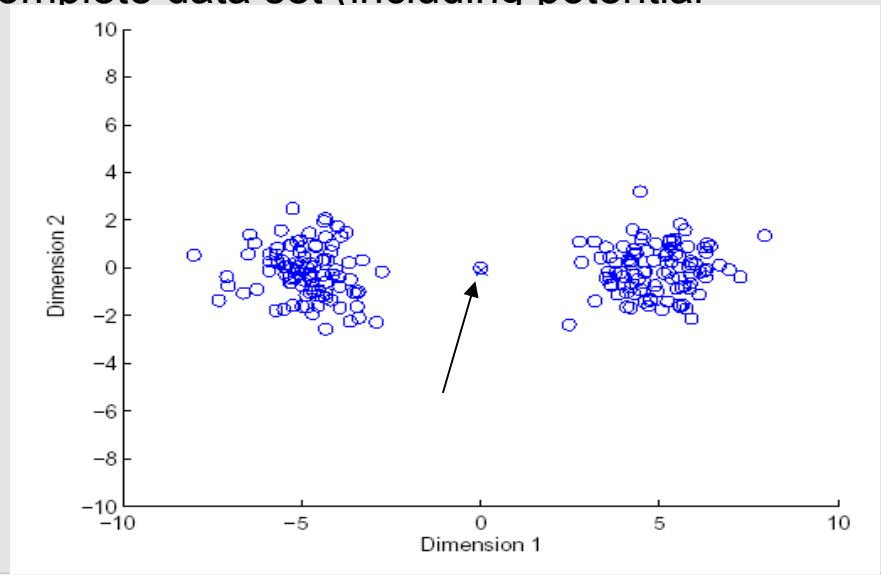
The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous.

Example: human cardiogram



Outlier Detection Methods

- 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

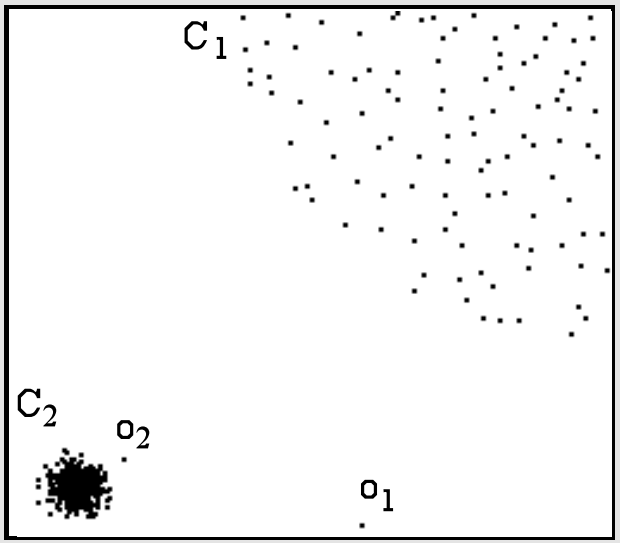
Radius *r, k* nearest neighbors

Density-based Detection

Relative density of object counted from density of its neighbors

Clustering-Based Methods

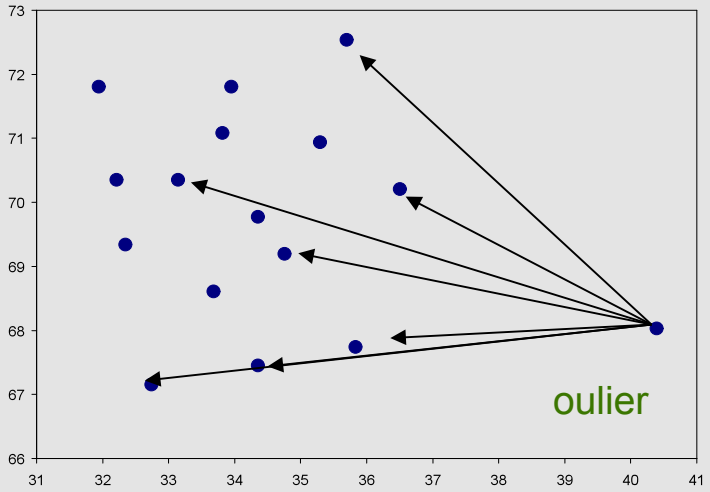
Normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters.

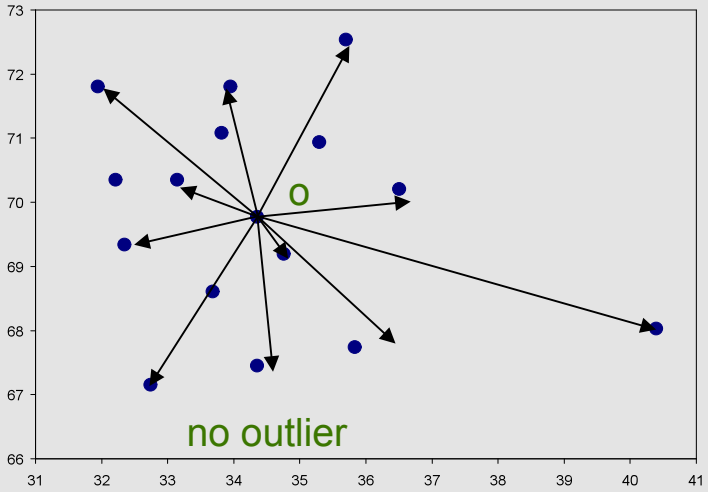


High-dimensional Outlier Detection Methods

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





Outlier Detection Methods Types

- Supervised Methods
  building a predictive model for normal vs. anomaly classes
- Semi-supervised Methods
  training data has labeled instances only for the normal class
- Unsupervised Methods
  no labels, most widely used

Supervised Methods

building a predictive model for normal vs. anomaly classes

problem is transformated to **classification problem**

Any supervised learning algorithm

E.g. a decision tree

how to detect outliers

Supervised Methods (cont.)

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

one-class learning

e.g. One-class SVM

Clustering (e.g. EM algorithm)

Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid.

Unsupervised Methods

no labels, most widely used

assumption:
normal instances are far more frequent than anomalies in the test data and they make clusters

Proximity-based methods, clustering

LOF, Local Outlier Factor

Local Outlier Factor (LOF)

$dist_k(o)$ . . . k-distance of an object o . . . **distance from o to its kth nearest neighbor**

$N_k(o)$ k-distance neighborhood of o . . . set of k nearest neighbors of o

reach.distk(o, p) = max{$dist_k(p)$, dist(o,p)} . . .
reachability-distance of an object o with respect to another object p

The local reachability-distance is the inverse of the average reachability-distance of its k-neighborhood.

LOF is the average of the ratio between the local reachability-distance of o and those of its k-nearest neighbors.

Evaluation of anomaly detection methods

Supervised settings – easy, precision/recall

Semi-supervised, unsupervised methods:

Need for classified data

1. Two class data, e.g. from UCI,
   1st class aka normal, the 2nd is a source of anomalies

2. Artificial data generator
   more flexible

Implementations

R

e.g. mvoutliers, DMwR and many others

WEKA

http://www.cs.waikato.ac.nz/ml/weka/

OutRules: A Framework for Outlier Descriptions in Multiple Context Spaces,
Univ. Saarbruecken
http://www.ipd.kit.edu/~muellere/OutRules/

And now

Similar to supervised methods

VOC Pascal data, 2048 features by Resnet

```
att504 <= 0.291603
|   att1746 <= 0.653862: aeroplane (95.0/4.0)
|   att1746 > 0.653862: person (5.0)
att504 > 0.291603
|   att456 <= 1.082573
|   |   att268 <= 1.711109
|   |   |   att1195 <= 1.121543: person (148.0/2.0)
|   |   |   att1195 > 1.121543
|   |   |   |   att1142 <= 0.340023: person (12.0/4.0)
|   |   |   |   att1142 > 0.340023: aeroplane (2.0)
|   |   att268 > 1.711109
|   |   |   att521 <= 1.855855
|   |   |   |   att365 <= 0.007182: aeroplane (5.0)
|   |   |   |   att365 > 0.007182: person (48.0/14.0)
|   |   |   att521 > 1.855855: aeroplane (13.0)
|   att456 > 1.082573
|   |   att1928 <= 0.140609: aeroplane (21.0/1.0)
|   |   att1928 > 0.140609: person (3.0/1.0)
```

Similar to supervised methods (cont.)

```
att504 <= 0.291603
|   att1746 <= 0.653862: aeroplane (95.0/4.0)
|   att1746 > 0.653862: person (5.0)
att504 > 0.291603
|   att456 <= 1.082573
|   |   att268 <= 1.711109
|   |   |   att1195 <= 1.121543: person (148.0/2.0)
|   |   |   att1195 > 1.121543
|   |   |   |   att1142 <= 0.340023: person (12.0/4.0)
|   |   |   |   att1142 > 0.340023: aeroplane (2.0)
|   |   att268 > 1.711109
|   |   |   att521 <= 1.855855
|   |   |   |   att365 <= 0.007182: aeroplane (5.0)
|   |   |   |   att365 > 0.007182: person (48.0/14.0)
|   |   |   att521 > 1.855855: aeroplane (13.0)
|   att456 > 1.082573
|   |   att1928 <= 0.140609: aeroplane (21.0/1.0)
|   |   att1928 > 0.140609: person (3.0/1.0)
```

ROBUST-C4.5

C4.5 incorporates a pruning scheme that partially addresses the outlier removal problem.

ROBUST-C4.5 (John 1995)
extending the pruning method to fully remove the effect of outliers

```
ROBUSTC45(TrainingData)
  repeat {
    T <- C45BuildTree(TrainingData)
    T <- C45PruneTree(T)
    foreach record in TrainingData
       if T misclassifies Record then
          remove Record from TrainingData
  } until T correctly classifies all
    Records in TrainingData
```

this results in a smaller tree without decrease of accuracy (average and st.dev.on 21 datasets).

Mining with rarity [Weiss 2004]

Experience from outlier detection: rare objects may be of a great interest

here we focus on labeled (classified) data

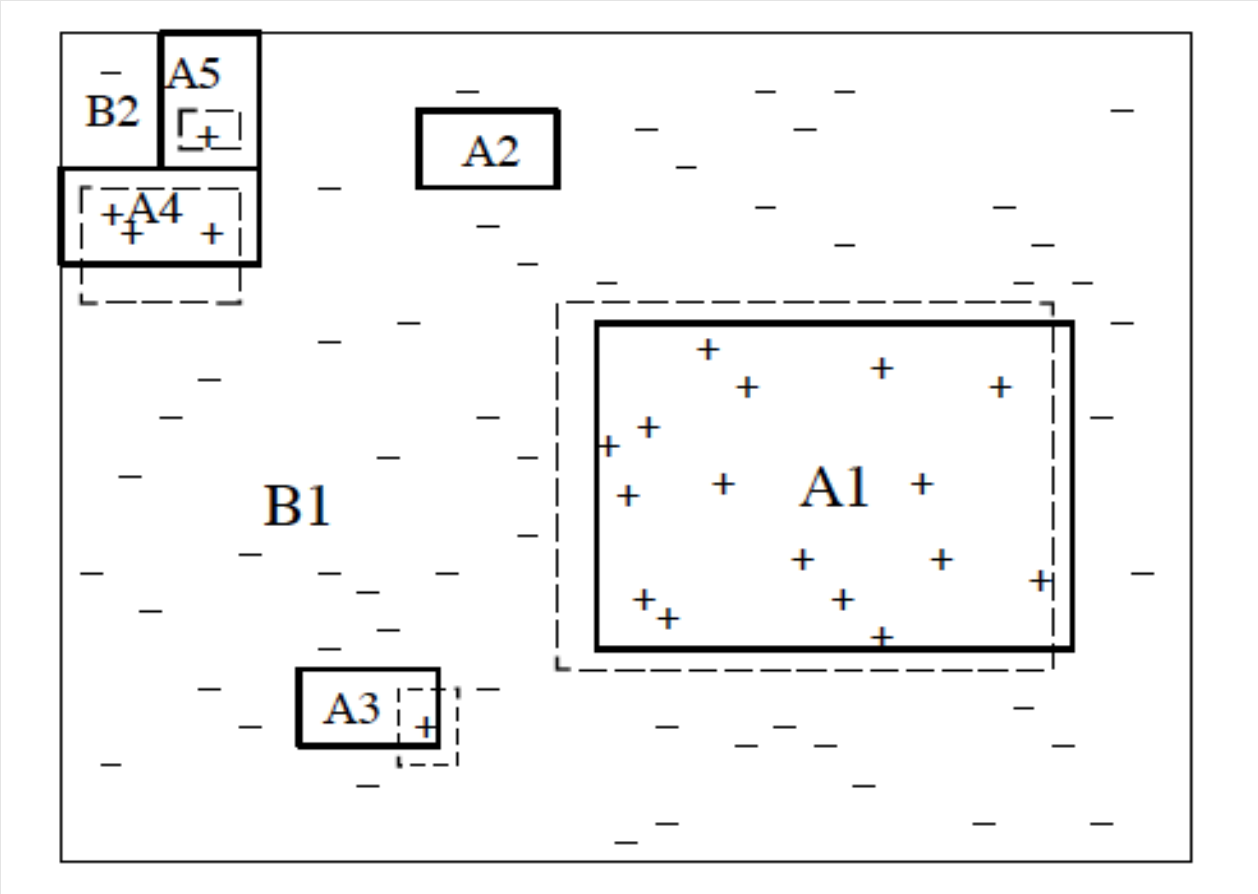
Rarity (*raridade - qualidade do que é raro; sucesso raro; objecto raro)*

of two kinds

rare classes

rare cases

Rare classes and rare cases (Weiss 2004)



Why rarity make data mining difficult

lack of the data – absolute, relative

data fragmentation – also in the case of divide-and-conquer strategy of the learner

noise

inappropriate inductive bias – maximum generality bias vs. maximum specificity bias

improper evaluation metrics – accuracy

Improper evaluation metrics

accuracy - rare classes have less impact on accuracy than common classes

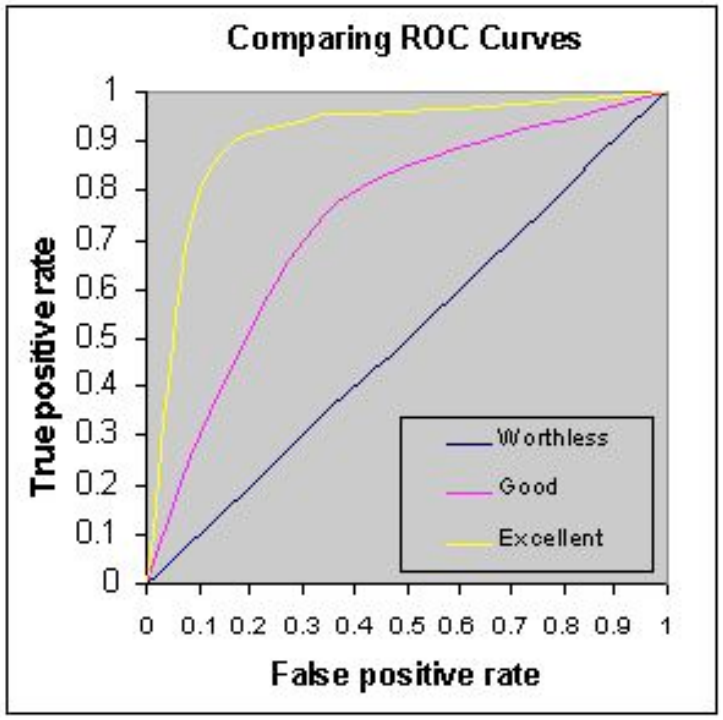
Weiss & Provost: error rate of minority class classification rules is 2-3 times that of the rules for majority-class examples

solution

AUC (Area Under ROC curve)

precision and recall and there combinations, e.g. F-measures

cost-sensitive learning



Class Outlier Detection

- each example belongs to a class

- Class-based outliers are those cases that

  look anomalous when the class labels are taken into account,

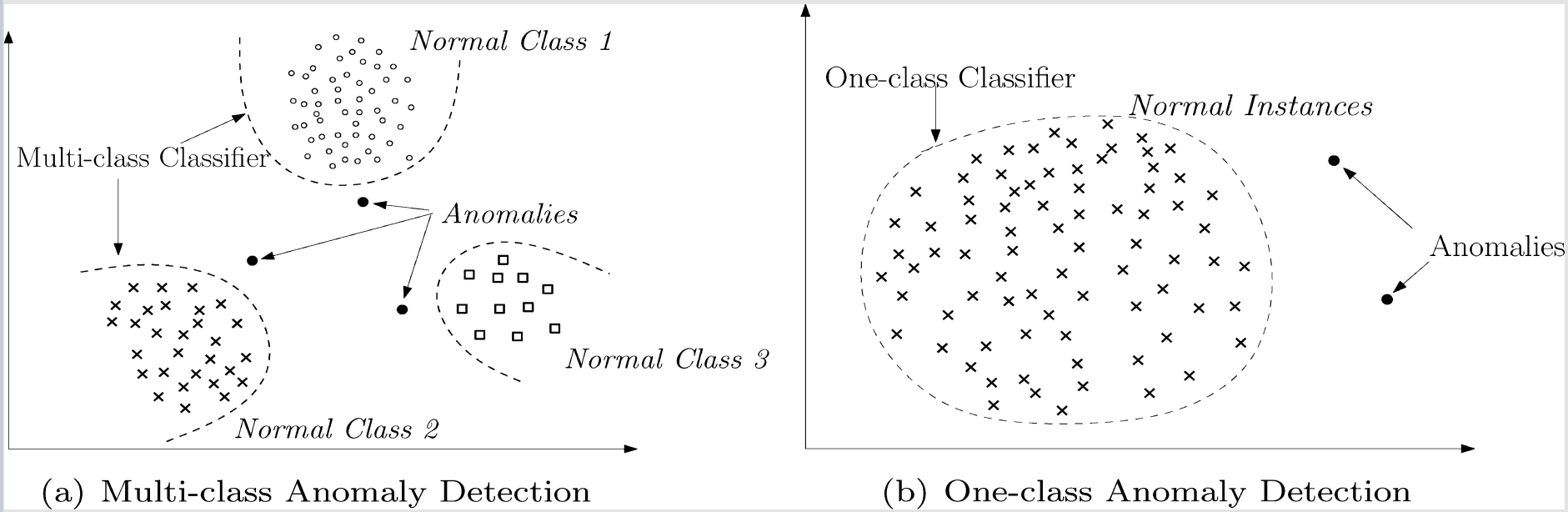
  but they

  do not have to be anomalous when the class labels are ignored.

- outliers = data point which behaves differently with other data points in the same class
- may look normal with respect to data points in another class

Class Outlier Detection

- sometimes called ‘semantic outlier’

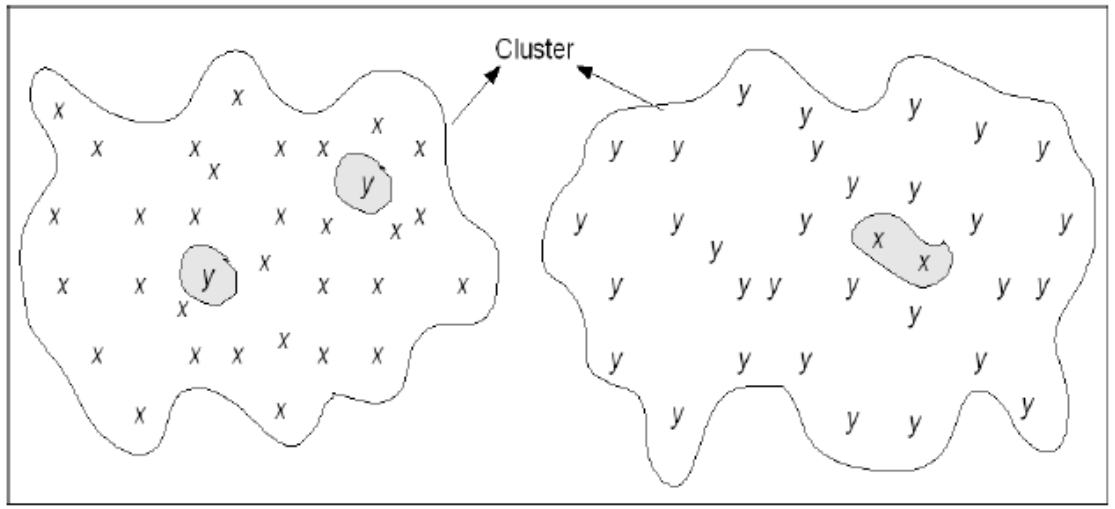


Multi-class outlier detection

- [Han 3rd edition]

- learn a model for each normal class
- if the data point does not fit any of the model, then it is declared an outlier

- advatage - easy to use
- disadvantage – some outliers cannot be detected

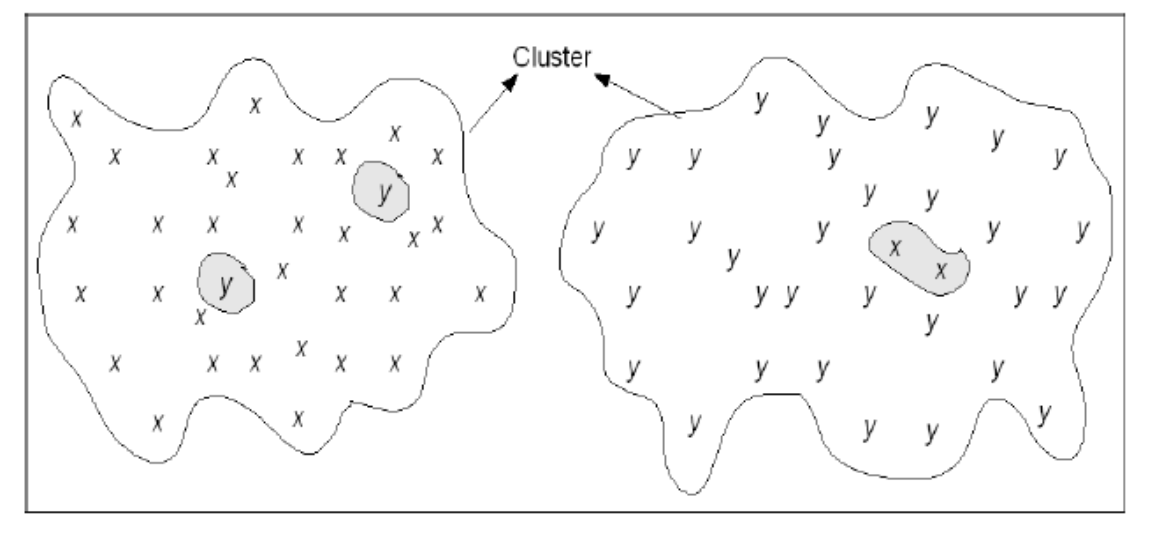


Semantic outliers (He et al. 2004)

- solve the problem
- cluster and then compute

- the probability of the class label of the example with respect to other members of the cluster

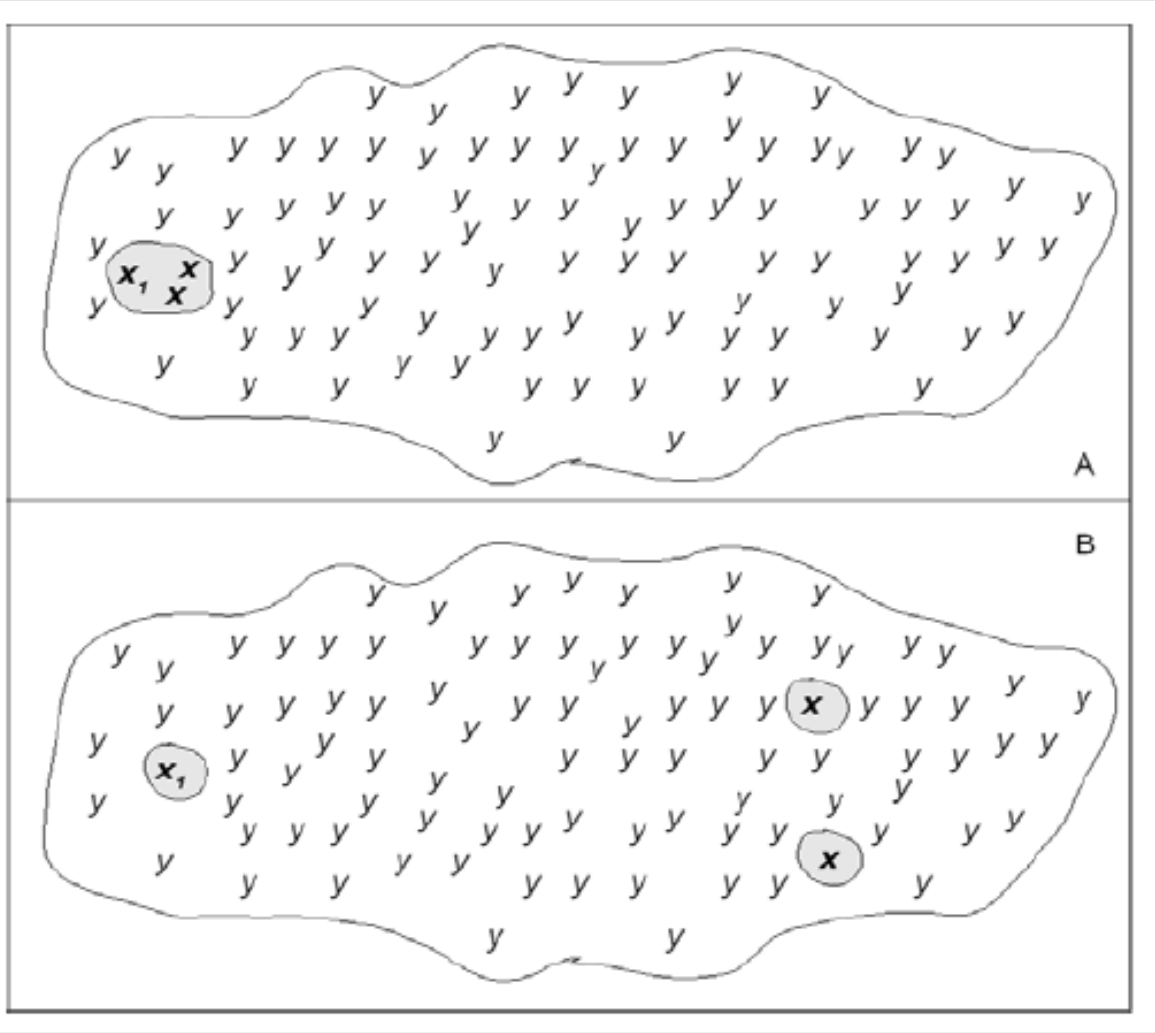
- the similarity between the example and other examples in the class



Semantic outliers (cont.)

x1 has the same rank

to fix it:



CODB (Hewahi and Saad 2007)

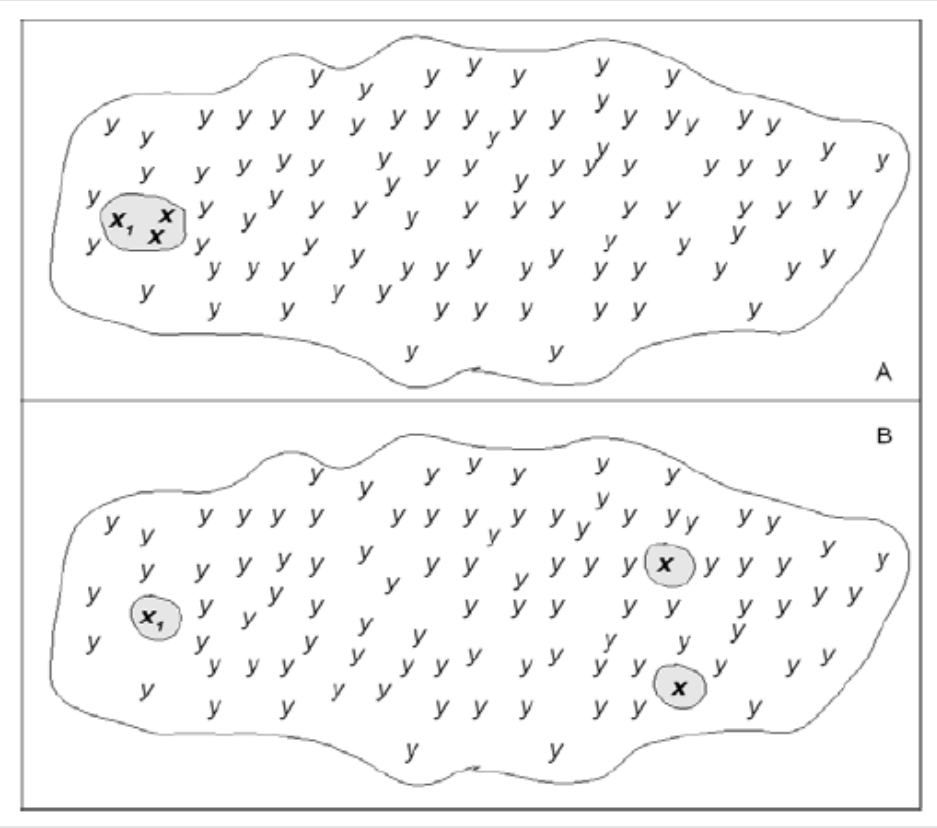
- combination of

  distance-based

  and density-based approach

  w.r.t class attribute

no need for clustering



CODB

Class Outlier Factor (COF)

T ... instance
K ... a number fo nearest neighbors
α, β ... parameters

COF(T) = SimilarityToTheK-NearestNeighbors *Similarity to the K neighbors*
+
α * 1/DistanceFromOtherElementsOfTheClass *Distance*
+
β * DistanceFromTheNearestNeighbors *Density*

CODB

COF(T) = SimilarityToTheK-NearestNeighbors +
α * 1/DistanceFromOtherElementsOfTheClass +
β * DistanceFromTheNearestNeighbors

COF(T) = K*PCL(T,K) + α * 1/Dev(T) + β *Kdist(T)

PCL(T,K) ... the probability of the class label of T w.r.t. the K nearest neighbors

Dev(T) ... ... the sum of distance from all other elements from the same class

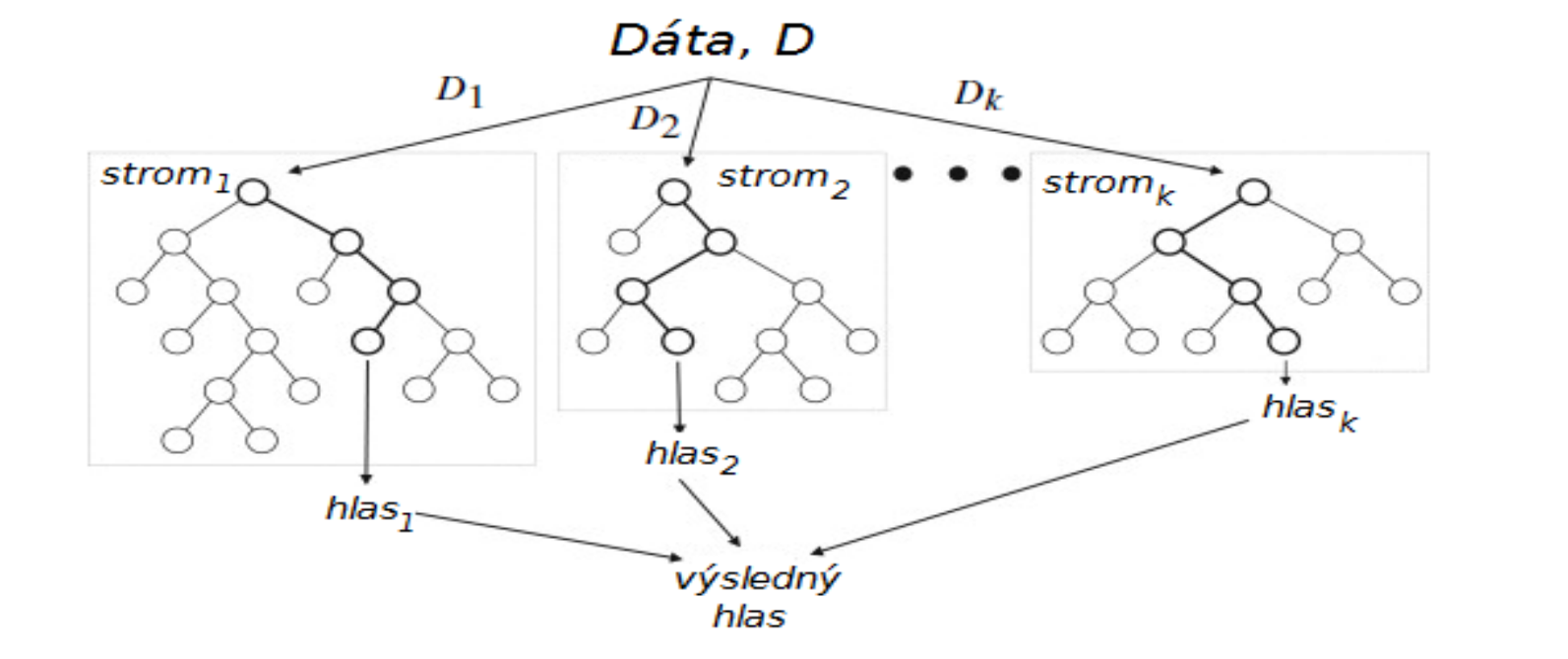
Kdist(T) ... the distance between T and its K nearest neighbors

RF-OEX: Class Outlier Detection with Random Forests (Nezvalová et al. IDA 2015)

- ***Random Forests*** is an ***ensemble*** classification and regression approach.
- 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
  - and with different subset of attributes

Random forest (Breiman 2000)

1. Bootstraping
2. Random tree



Class Outlier Detection – Random Forests

- *A*fter each tree is built, all of the data are run down the tree, and **proximities** are computed for each pair of cases:

- If two cases occupy the same terminal node, their proximity is increased by one.
- At the end of the run, the proximities are normalized by dividing by the number of trees.
- Define the average proximity from case n in class j to the rest of the training data class j as:

$$\bar{P}(n) = \sum_{cl(k)=j} \text{prox}^2(n,k)$$

- The raw outlier measure for case n is defined as

$$\text{nsample}/\bar{P}(n)$$

Proximity matrix

	Príklad 1	***Príklad 2***	***Príklad 3***	***Príklad 4***	***Príklad 5***
Príklad 1		0	1	1	2
Príklad 2	0		0	1	1
Príklad 3	1	0		4	3
Príklad 4	1	1	4		3
Príklad 5	2	1	3	3	

Class Outlier Factor

Outlier factor

=

sum of three different measures of proximity or outlierness

=

Proximity to the members of the same class

+

Misclassication - proximity to the members of other classes and

+

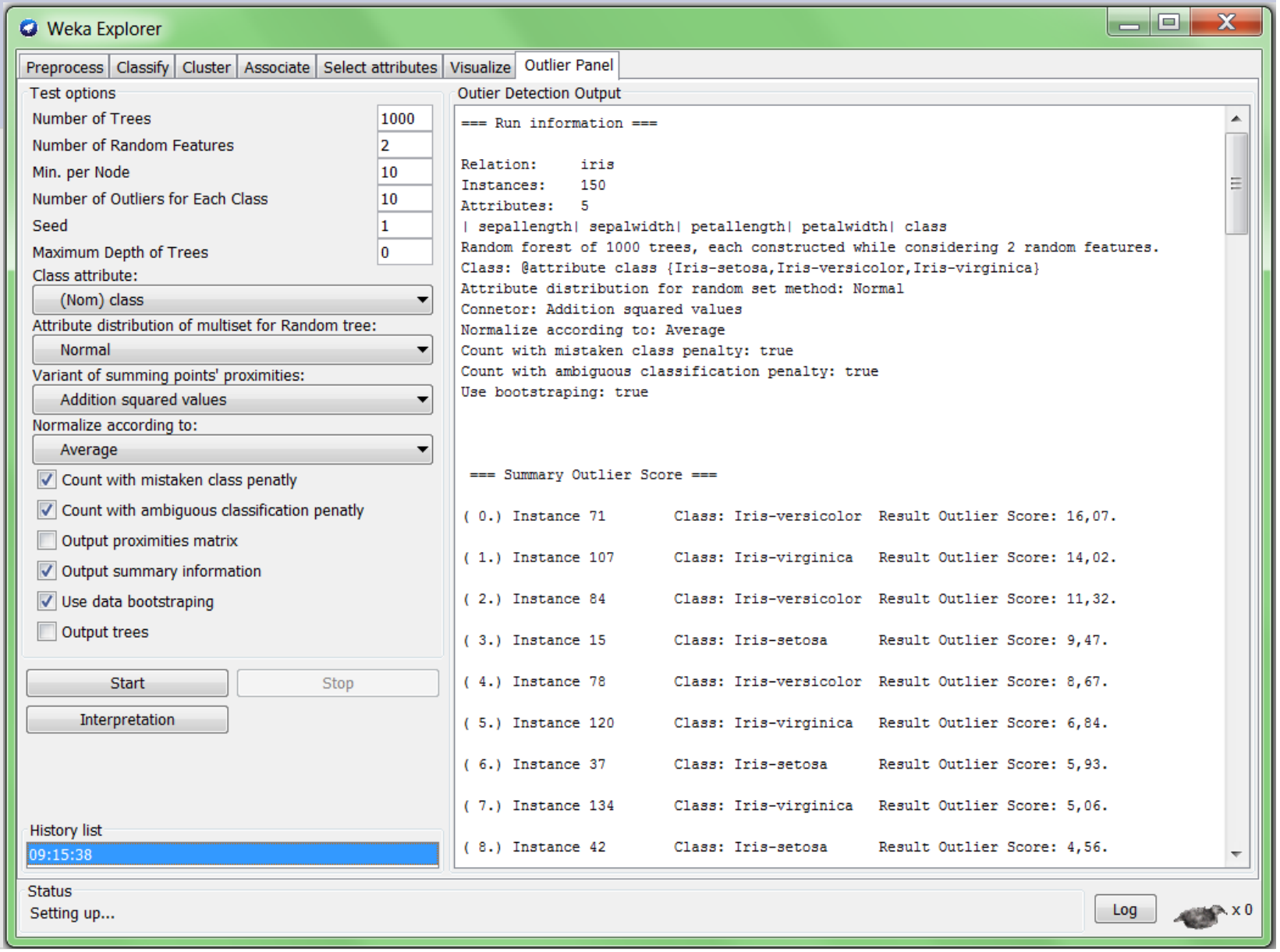
Ambiguity measure – a percentage of ambiguous classification

RF-OEX

Detection

+

explanation



Applications

Small and medium enterprises – growing/not-growing

Logic: Finding anomalous solutions– correct/incorrect
- resolution proofs
- graph mining employed

IMDb

PS ČR

ZOO, House Votes (Republicans vs Democrats), Student Loans

Students with standard/non-standard study interval

Rooms in MU campus

Logic: Finding anomalous solutions

1. to learn classifier that would discriminate between correct and incorrect solutions and
2. detect solutions that are incorrectly classified - outliers

Or directly detect outliers without learning a classifier

We cannot use a common outlier detection methods because data are labeled as correct and incorrect solutions.

Class outlier detection can help

Logic: Finding anomalous solutions (cont.)

Search/**discover students' solutions which are unusual**

We need data in attribute-value representation

frequent pattern mining, frequent subgraphs
One attribute for each higher-level generalized pattern; values are true (occurrence of the pattern) and false (non-occurrence of the pattern).

Class: occurrence or non-occurrence of the error of resolving on two literals at the same time (***we call it*** E3 error)

Novel „solutions" found, not recognised with the tool used

And more . . .

Anomaly explanation

Anomaly detection in text

Since Guthrie (Shefield) 2009…

Research directions

Evaluation of anomaly detection (AD) methods
    KDD 2013 Outlier Detection And Description Workshop

Explanation of outliers
    Micenková & Ng

Feature selection for anomaly detection

Ensembles for anomaly detection

Anomaly detection in structured data (graphs. . .)

ILP-based anomaly detection
    Angiulli F. and Fasetti F, ICDM 2009, DMKD J. 2014

Autoencoders etc.

Literature

L. Nezvalová, L. Torgo, K. Vaculík, L. Popelínský [AIMSA 2014] [IDA 2015]

Han j. et al. *Data Mining. Principles and Techniques.* 3rd edition.

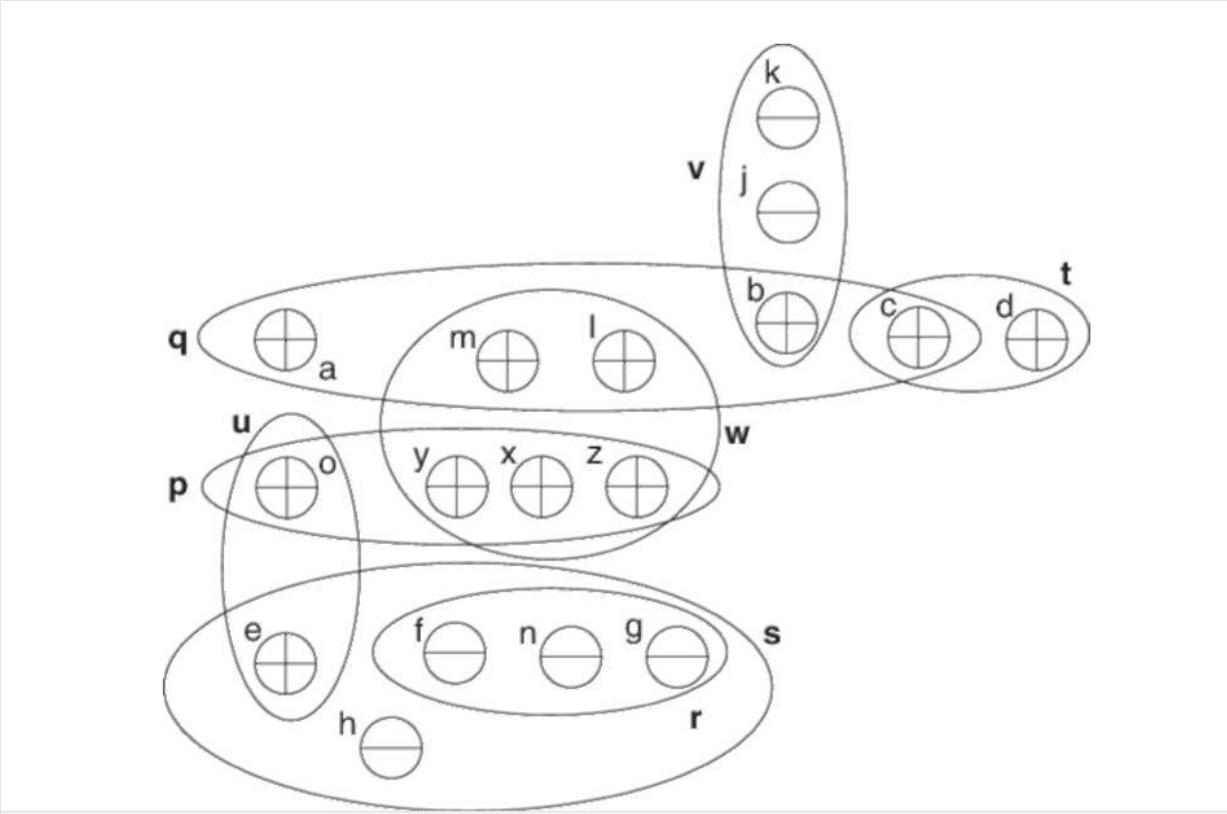
He Z. et al. *Mining Class Outliers: Concepts, Algorithms and Applications in CRM.* Expert Systems and Applications, ESWA 2004, 27(4), pp. 681-697, 2004.

Hewahi N.M. and Saad M.K. *Class Outliers Mining: Distance-Based Approach*. International Journal of Intelligent Systems and Technologies, Vol. 2, No. 1, pp 55-68, 2007.

John G.H. *Robust Decision Trees: Removing Outliers from Databases.* Knowledge Discovery and Data Mining - KDD , pp. 174-179, 1995

Weiss G.M. Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter 6 (1), 7-19

ILP



Idea

Given E+ positive and E- negative examples and the
background knowledge B,
learn concept C and dual Concept C'
(swap positive and negative examples)

Look for examples that if removed from the learning set do not change the description (logic program) of C and C' significantly

i.e. difference of coverage is smaller then a threshold

= normal examples

Idea

Suppose A, a set of normal examples, is a subset of E+
E+ \ A = A’ … abnormal examples

Given the k-max, the number of outliers,
find the abnormal subset A’ of examples not greater than k-max.

Explanation of an outlier: two theories.

1. – rules that cover some of abnormal examples
   A^ - examples outside of A’ covered only by clauses that cover an example from A’
2. - rules induced in absence of A’ and
   covers some of examples from A^

Literature

Angiulli F. and Fasetti F. *Outlier detection using Inductive Logic Programming*. Proceedings of ICDM 2009.

Han j. et al. Data Mining. Principles and Techniques. 3rd edition.

He Z. et al. *Mining Class Outliers: Concepts, Algorithms and Applications in CRM.* Expert Systems and Applications, ESWA 2004, 27(4), pp. 681-697, 2004.

Hewahi N.M. and Saad M.K. *Class Outliers Mining: Distance-Based Approach*. International Journal of Intelligent Systems and Technologies, Vol. 2, No. 1, pp 55-68, 2007.

John G.H. *Robust Decision Trees: Removing Outliers from Databases.* Knowledge Discovery and Data Mining - KDD , pp. 174-179, 1995

Weiss G.M. *Mining with rarity: a unifying framework*. ACM SIGKDD Explorations Newsletter 6 (1), 7-19