Advanced anomaly analysis

Luboš Popelínský Masaryk University

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

Data science and anomaly detection

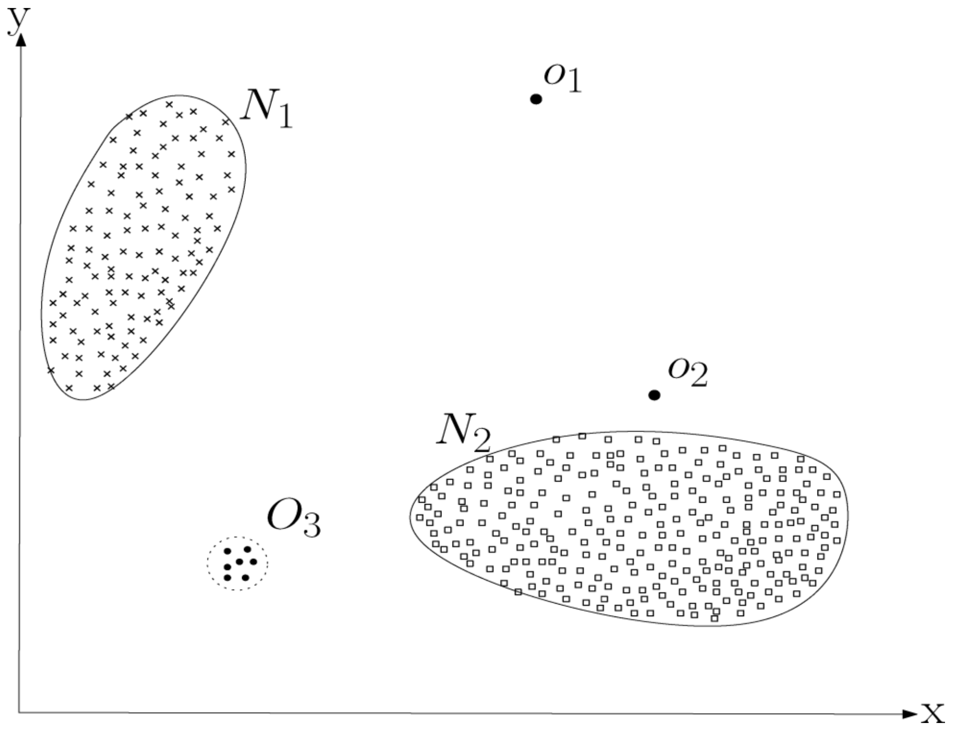
- Machine learning techniques – four main categories:
  - Clustering
  - Classification
  - Frequent pattern mining and

  - **Anomaly detection**

    “Unlike the first three main tasks, which aim to find patterns that characterize the majority of the data, the fourth task focuses on identifying patterns that represent only the minority data."

Anomaly detection

- "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" [Hawkins 1980]



- Outlier factor
- = dissimilarity with other instances
- Two needs for outlier detection (OD):
  1) Detect, Remove & Run again
  2) Detect, Analyze

Applications of anomaly detection

- 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
- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
- Education: detection of unexpected solutions
  - E.g., constructive tasks in logic
- Intrusion detection
  - Attacks to a network, or to a blog
- Plagiarism detection
  - A part of text has been written by somebody else

Applications of anomaly detection

- Language "irregularities" PT: *Ser casado, estar morte*

  We'll begin with a box, and the plural is boxes; but the plural of ox became oxen not oxes.

  One fowl is a goose, but two are called geese, yet the plural of moose should never be meese.

  You may find a lone mouse or a nest full of mice; yet the plural of house is houses, not hice.

  If the plural of man is always called men, why shouldn't the plural of pan be called pen?

  If I spoke of my foot and show you my feet, and I give you a boot, would a pair be called beet?

  If one is a tooth and a whole set are teeth, why shouldn't the plural of booth be called beeth?

  We speak of a brother and also of brethren, but though we say mother, we never say methren.

  Then the masculine pronouns are he, his and him, but imagine the feminine, she, shis and shim.

- Medicine
  - Unusual symptoms/test results may indicate potential health problems
  - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g., gender, age, …)
- …

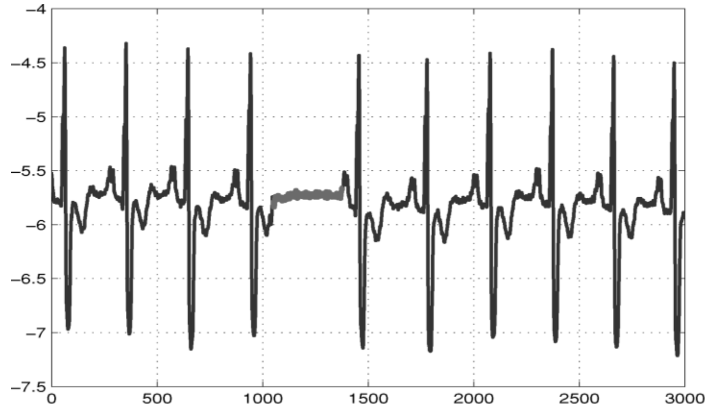
Anomaly detection and Novelty detection

- What is the difference between novelty detection and anomaly detection?

- Anomaly detection encompasses two broad practices: outlier detection and novelty detection
- Outliers are abnormal or extreme data points that exist only in training data
- In contrast, novelties are new or previously unseen instances compared to the original (training) data

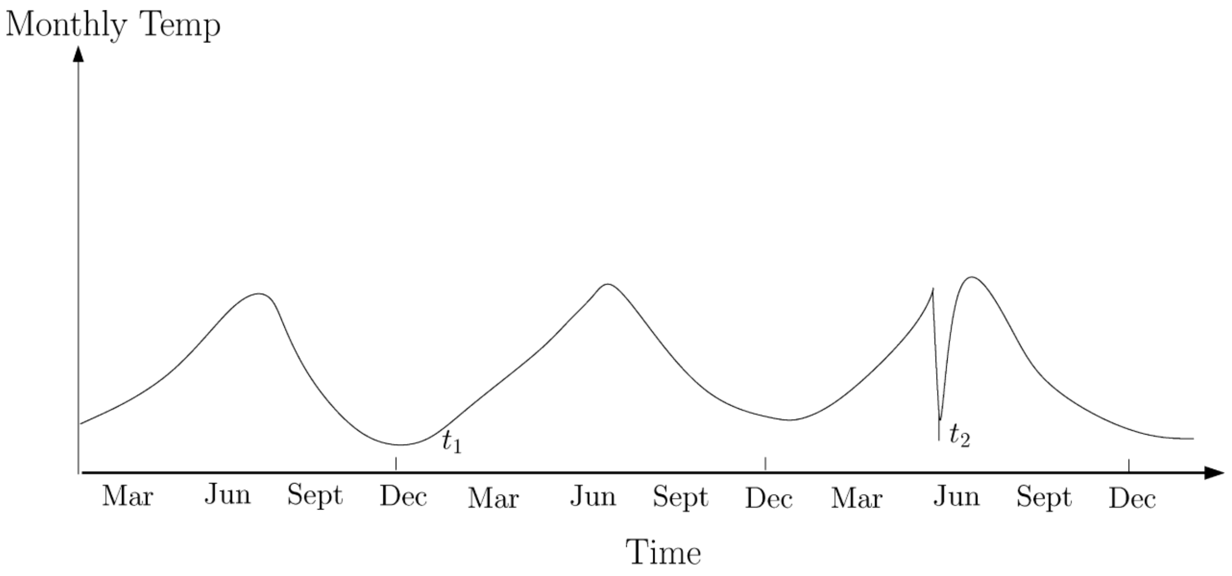
Types of outliers

- 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



Contextual outliers

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



- Solution: find contextual features
- Example: temperature time-series
- Is the temperature 28°C outlier?
  - If we are in Brno in summer NO
  - If we are in Brno in winter YES

  → it depends on the location and time – CONTEXT
- Any other solution?

Types of anomaly detection methods

- Supervised methods
  - Building a predictive model for normal vs. anomaly classes

- Semi-supervised methods
  - Training data has labeled instances only for the normal class
  - Example: accidents in nuclear power stations

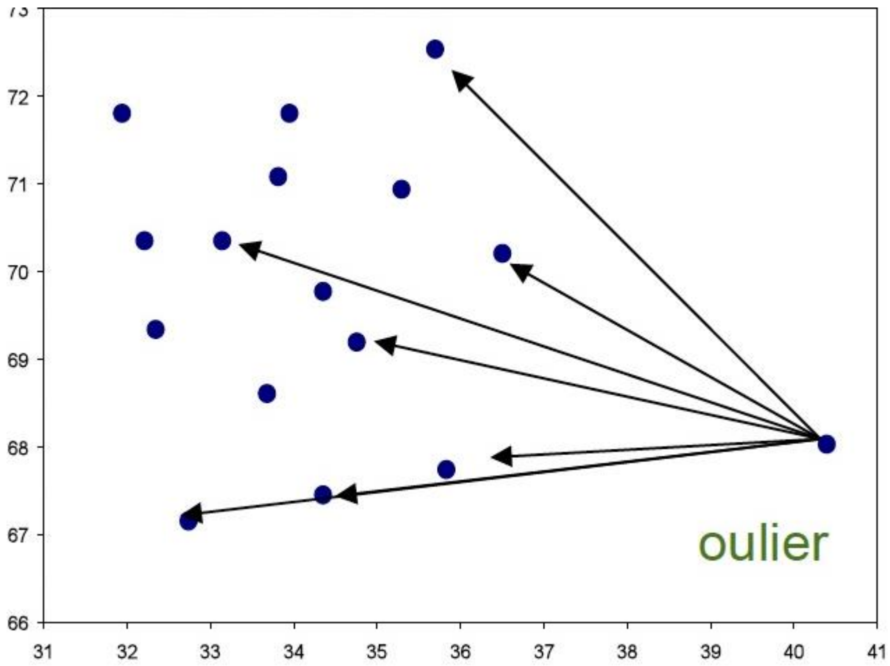
- Unsupervised methods
  - No labels, most widely used

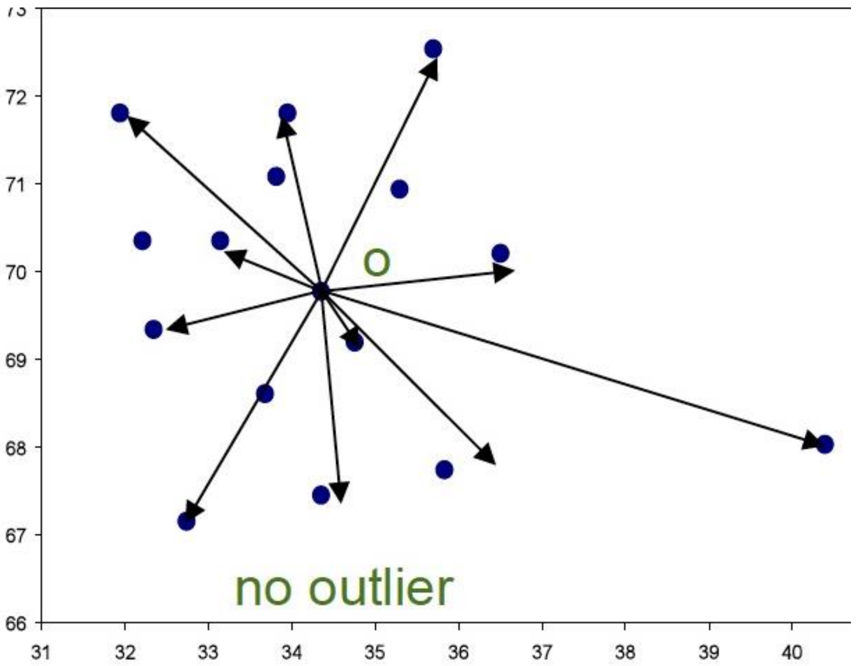
Anomaly detection methods

- Statistical 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 detection
  - 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

- 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





Local and global anomalies/methods

- A global anomaly
  - Is an object which has a significantly large distance to its *k*-th nearest neighbor (usually greater than a global threshold) whereas
  - = can be used for sorting anomalies w.r.t. the outlier factor

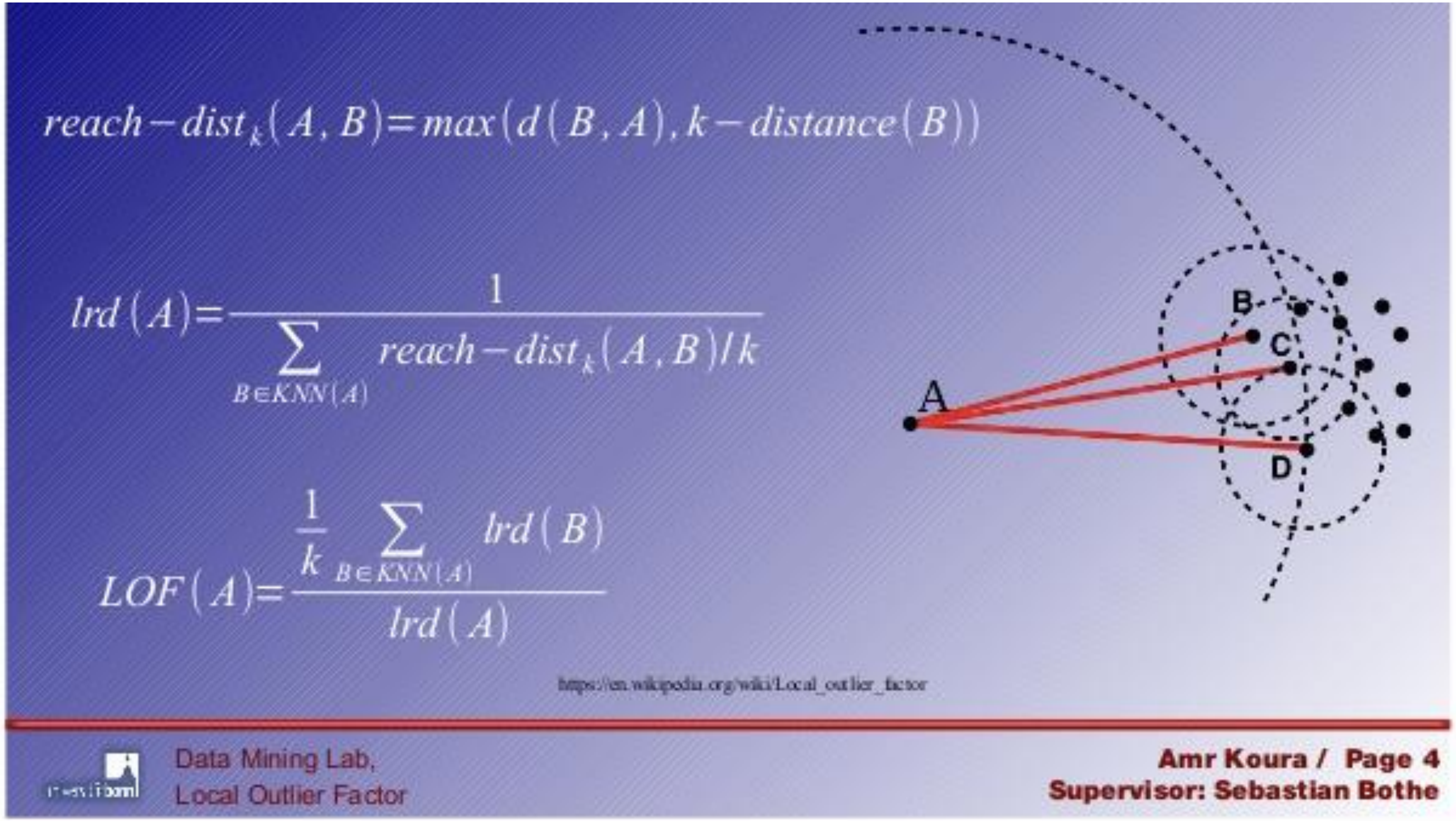
  - Example: *k*-NN

- A local anomaly
  - Has a distance to its *k*-th neighbor that is large relatively to the average distance of its neighbors to their own *k*-th nearest neighbors

  - Example: LOF

Local Outlier Factor (LOF)

- Only one parameter, *k*, a number of neighbors



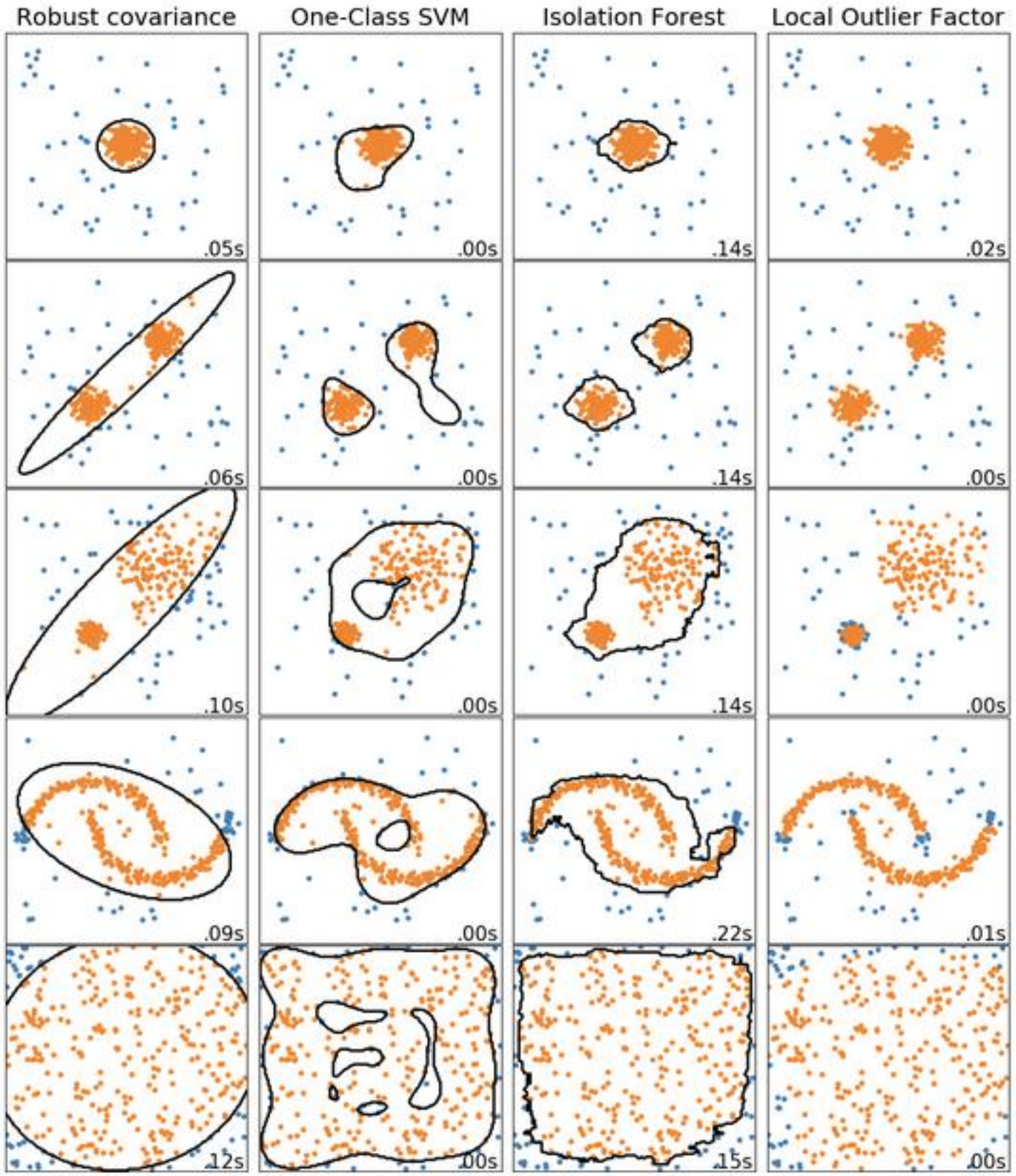
Local Outlier Factor (LOF)

- $dist_k(o)$ – *k*-distance of an object o – distance from o to its *k*-th nearest neighbor
- $N_k(o)$ – *k*-distance neighborhood of o – set of *k* nearest neighbors of o
- $reach\text{–}dist_k(o,p) = \max\{dist_k(p), dist(o,p)\}$ – 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

Example of Scikit-learn

- Black – border between inliers and outliers
- 15% samples generated as random uniform noise
- 15% is also a parameter of one class-SVM and the contamination for other algorithms



Deep learning and anomaly detection

- Autoencoders
  - Two multilayered perceptrons (MLP) – encoder $X \rightarrow Z$ + decoder $Z \rightarrow X$
  - Reconstruction loss = outlier factor

- Variational autoencoders
  - Model conditional probabilities $Z|X$ and $X|Z$, assuming Gaussian distribution

- Generative adversarial networks
  - Two adversaries (MLP) – generator + discriminator
  - Generator creates samples that resemble the real data, while the discriminator is trying to recognize the fake samples from the real ones

Which OD algorithm is better?

- Hyperparameter settings can be a problem
- [Škvára et al. Are generative deep models for novelty detection truly better? 2018]

	kNN	IForest	AE	VAE	GAN	fmGAN
test auc	3.94	5.63	3.47	2.07	3.90	1.99
train auc	3.13	4.61	3.63	2.84	4.46	2.33
top 5%	2.57	4.07	3.24	2.73	4.90	3.49
top 1%	2.14	3.53	3.13	2.93	4.97	4.30

Class-based outliers

Class-based outliers. Why do we need a new concept?

Class-based outliers

- Example: e-shop planning marketing campaign to increase income
- Which clients to be sent with a new offer?
  - Monitoring two groups of clients:
  - Group PLUS: buying products more or less often
  - Group MINUS: browsing list of offers/products more or less often but (almost) have not bought anything so far
- Which clients to be sent with a new offer?
- Other examples?

ROBUST-C4.5

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

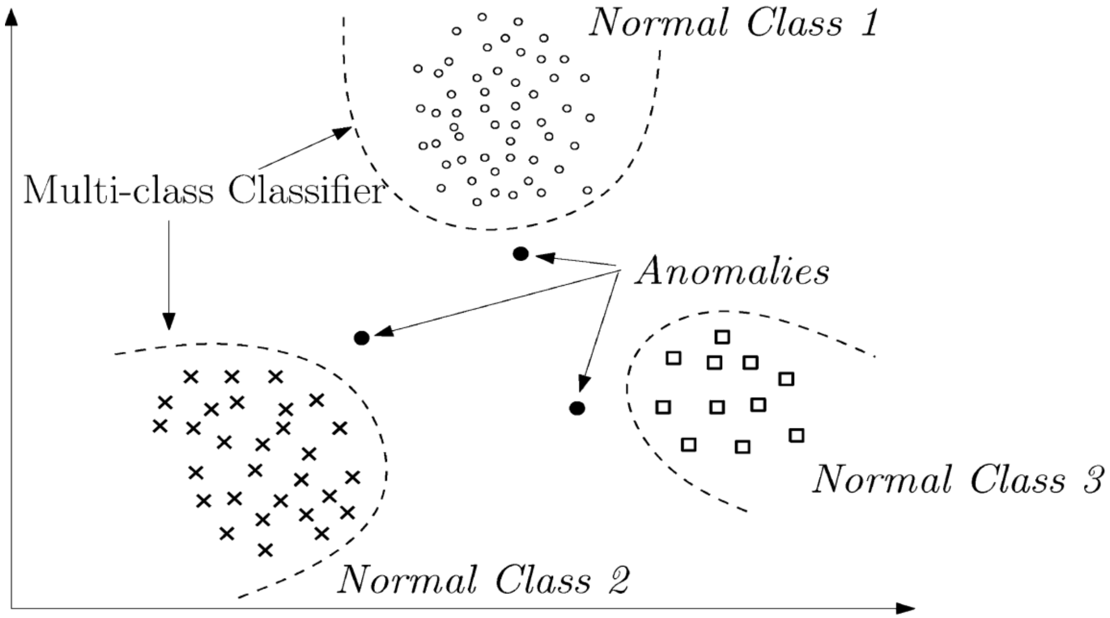
- ROBUST-C4.5 (John 1995)
- Extending the pruning method to fully remove the effect of outliers

```
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
```

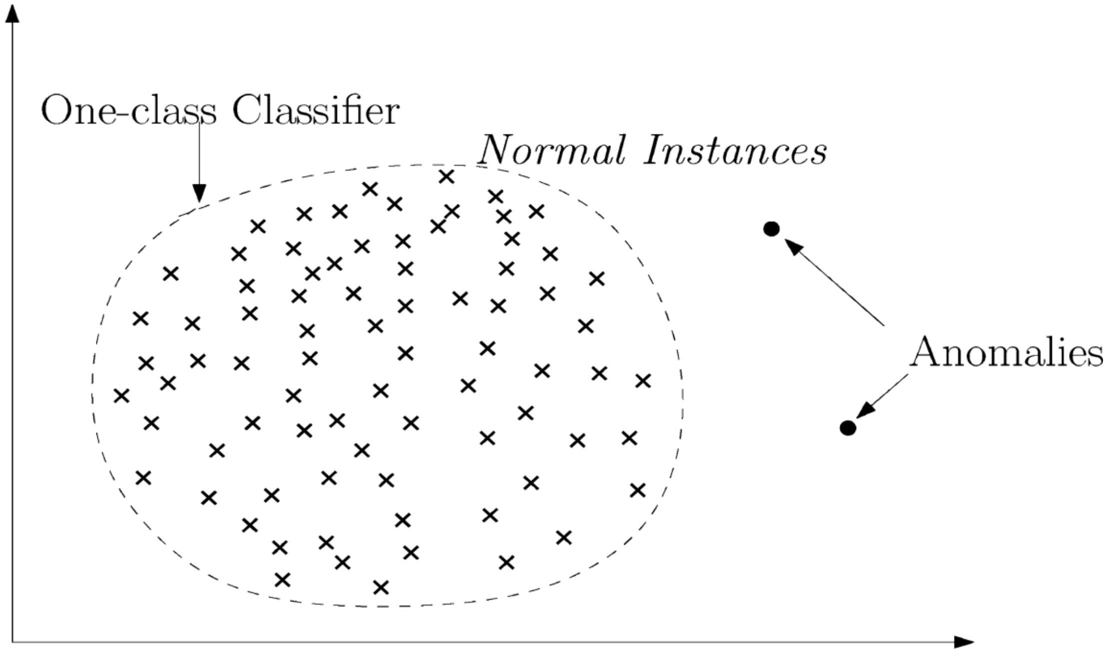
- This results in a smaller tree without decrease of accuracy (average and st. dev. on 21 datasets)

Class-based outlier detection

- Sometimes called “semantic outlier”



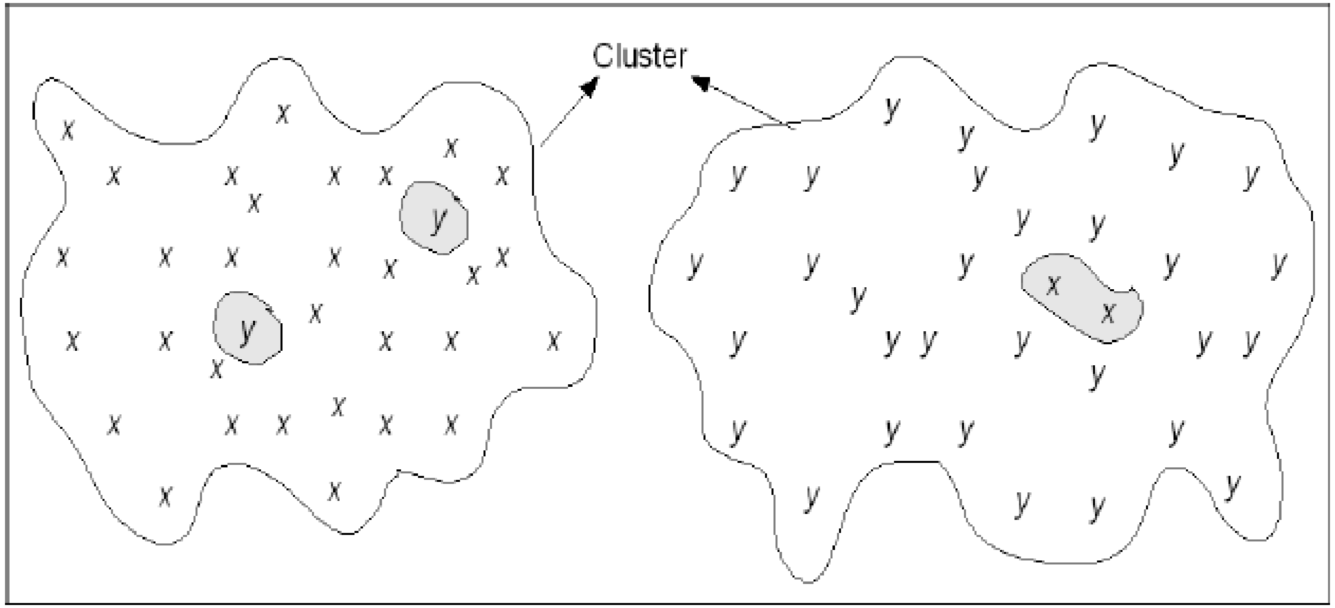
(a) Multi-class Anomaly Detection



(b) One-class Anomaly Detection

Multi-class outlier detection

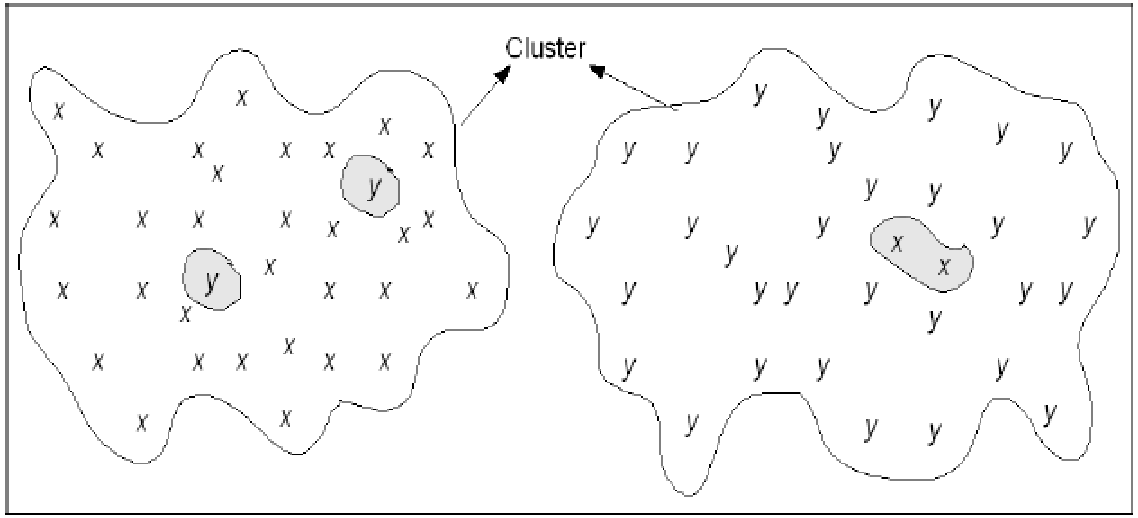
- [Han, Data Mining. Principle and Techniques, 3rd edition]
- Learn a model for each normal class
  - If the data point does not fit any of the model, then it is declared an outlier
- Advantage – easy to use
- Disadvantage – some outliers cannot be detected



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



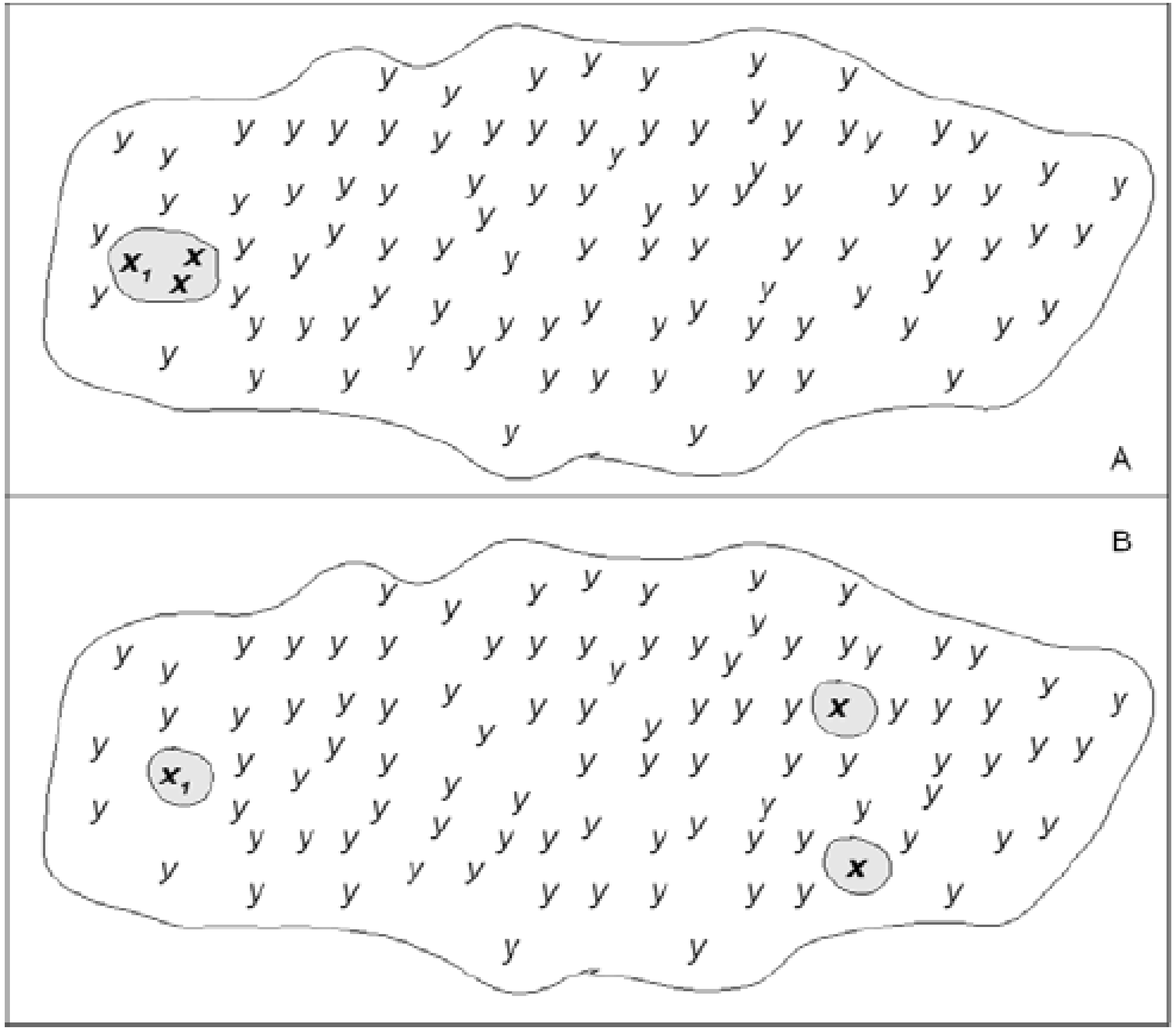
How to compute class-based outlier factor

- [He et al., 2004]
- COF = OF w.r.t. own class (+) OF w.r.t. the other classes
- Pros & Cons

Semantic outliers (cont.)

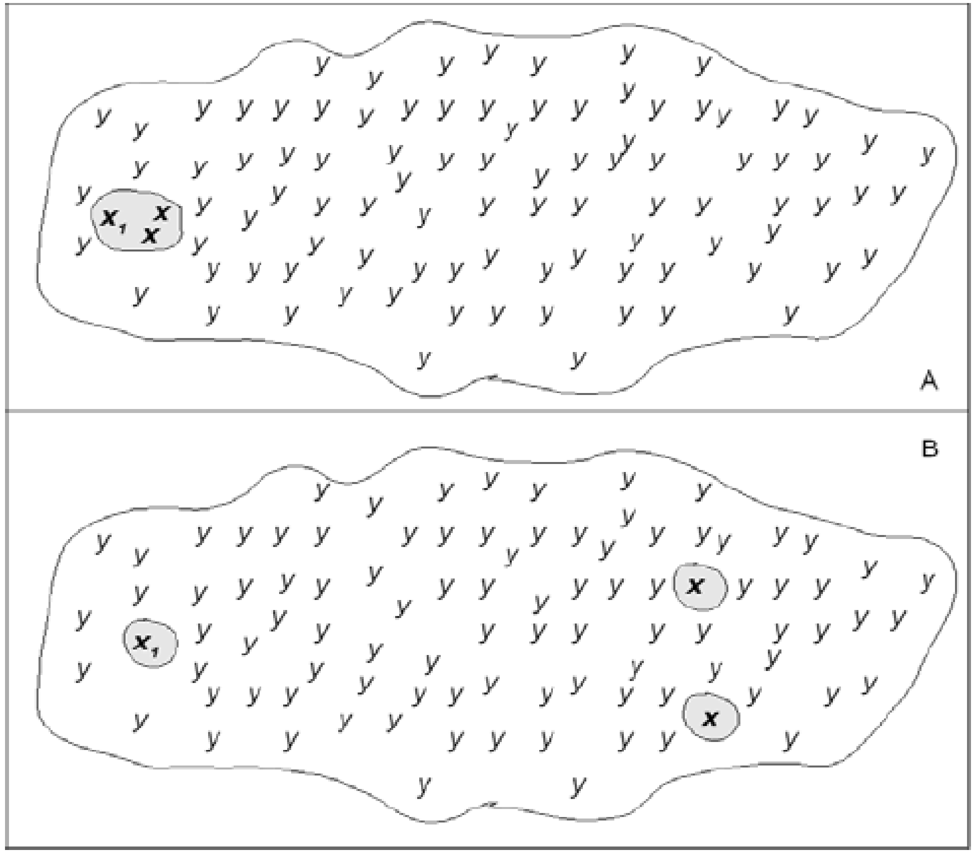
- x_1 has the same rank

- To fix it:



CODB (Class Outlier Distance-Based)

- [Hewahi and Saad, 2007]
- Combination of distance-based and density-based approach w.r.t. class attribute
- No need for clustering



CODB

- $COF(T) = similarityToKNearestNeighbors + \alpha \cdot 1/distanceFromOtherElementsOfTheClass + \beta \cdot distanceFromTheNearestNeighbors$

- $COF(T) = k \cdot PCL(T,k) + \alpha \cdot 1/Dev(T) + \beta \cdot dist_k(T)$
  - $PCL(T,K)$ – the probability of the class label of T w.r.t. the k nearest neighbors
  - $Dev(T)$ – the sum of distances from all other elements from the same class
  - $dist_k(T)$ – the distance between T and its k nearest neighbor

RF-OEX: COD with Random Forests

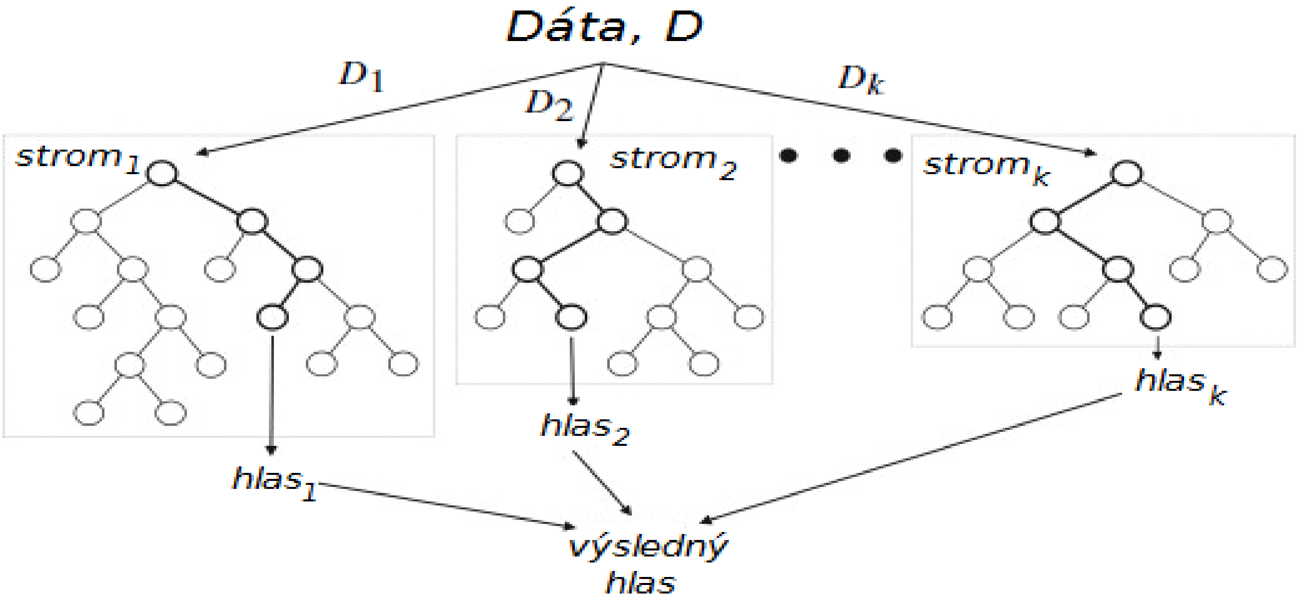
- Random Forests is an ensemble classification and regression approach

- 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)

- Bootstrapping

- Random tree



Class Outlier Detection – Random Forests

- After 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:

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

- The raw outlier measure for case n is defined as: $nsample/\bar{P}(n)$

Proximity matrix

	Example 1	*Example 2*	*Example 3*	*Example 4*	*Example 5*
Example 1		0	1	1	2
Example 2	0		0	1	1
Example 3	1	0		4	3
Example 4	1	1	4		3
Example 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

  \+

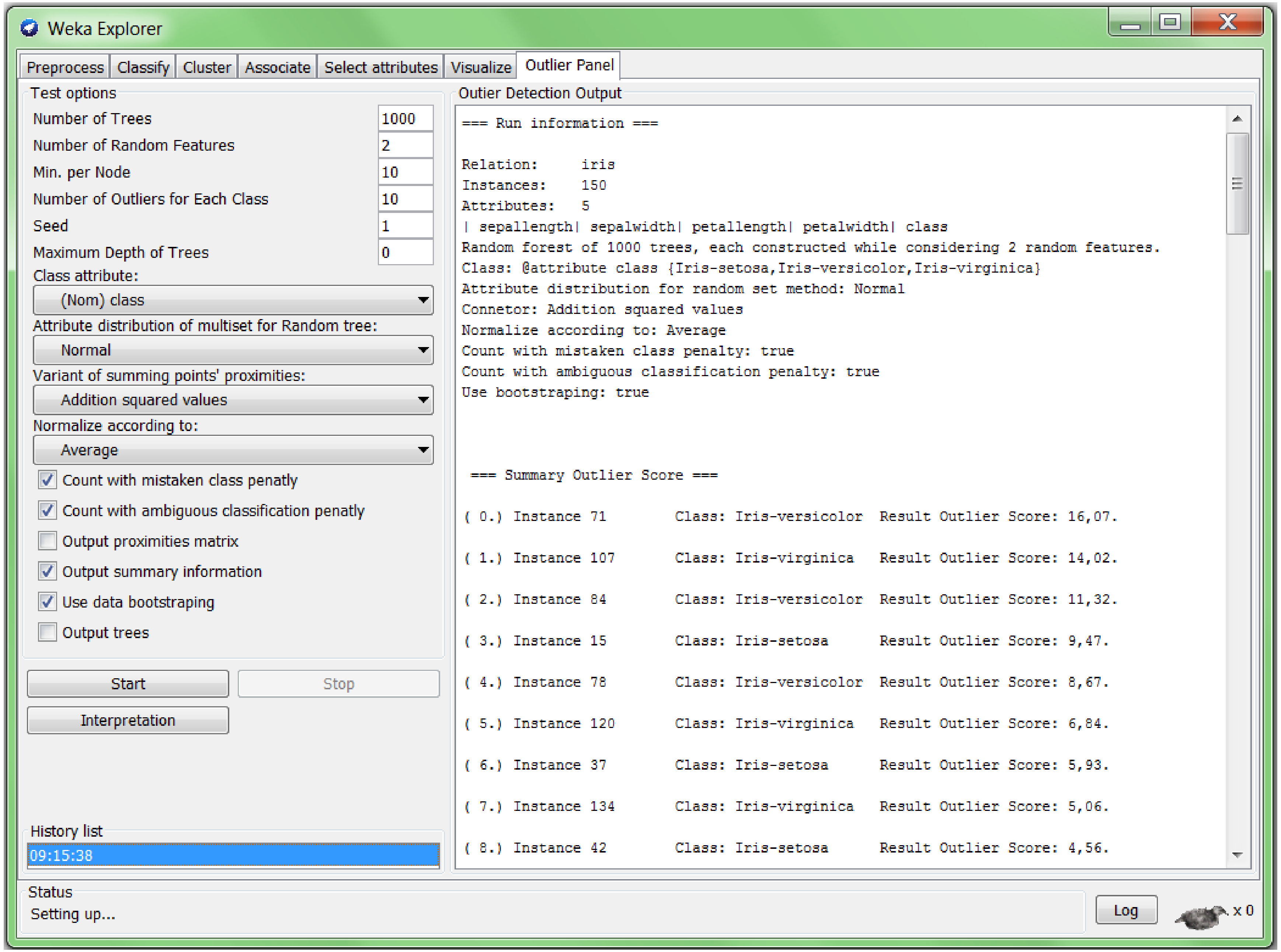
  Misclassification – proximity to the members of other classes and

  \+

  Ambiguity measure – a percentage of ambiguous classification

RF-OEX

- Detection
- +
- Explanation



Applications

- ZOO
- E-shop: clients vs. potential clients
- Educational data mining:
  - Students with standard/non-standard study interval
  - Intro to logic: finding anomalous solutions
- IMDb
- Czech Parliament
- Data pre-processing
- ... and more?

Teaching logic: finding anom. solutions

- Task: Build a resolution proof, 400 students, at least 3 task to solve
- Automated evaluation: error detection
- Two classes: CORRECT, INCORRECT
- If an error appeared, the solution is classified as incorrect
- Find solutions that was classified as CORRECT and not, and opposite

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

- Class outlier detection can help

Finding anom. solutions

- 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 recognized with the tool used

IMDb: Funny/unusual reviews

Finding anom. solutions

- Search/discover reviews that do not correspond to positive or negative star evaluation

- Large Movie Review Dataset

- Each review represented as a list of word appearance
- Only 68 most frequent words in the dataset used

- Class negative *… ****
- Class positive *******…***

Finding anom. solutions

Branca de Neve (2000)

User Reviews

Review this title

9 Reviews

Hide Spoilers   Filter by Rating: Show All   Sort by: Helpfulness

★ 6/10

one of the most interesting movies of the past couple of years, but perhaps for all the wrong reasons.

Z_cm   1 October 2004

João César Monteiro was known for his excruciatingly lengthy movies and awkward humour, but nothing could prepare both the audiences and the critics for his outrageous 'Branca de Neve'! A huge debate followed its debut, it has been labeled everything, from a masterpiece to a fraud and four years later it still angers and baffles a great deal of people. The first shocker is the movie itself. All of us have heard of and may recall with fondness the silent movie era, but 'Branca de Neve' introduces us to the 'radiophonic movie' concept, that is, a movie that has no image at all! Most of the movie leaves the viewer staring at a monotonous black canvas, interrupted only by a few occasional and might I add, very brief still shots. The story itself is an adaptation of Robert Walser's 'Schneewittchen' and the dialog between the characters happens in complete darkness, like a radio play. But a very strangely acted one, like some weird cross between the

References

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

Appendix: ILP

- 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 than a threshold
  - = normal examples