Outlier detection

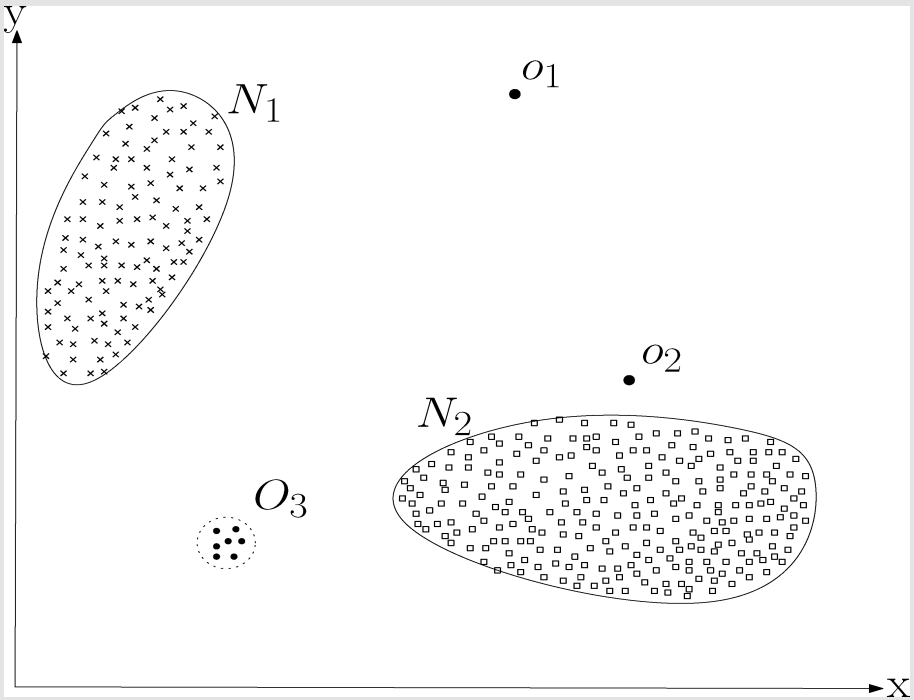
"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism" [Hawkins 1980]

Outlier factor
= dissimilarity with other instances

Two needs for OD

1. Detect THAN
   Remove & Run again

2. Detect THAN
   Analyze



Applications of Outlier 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.

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.

Example:

online-shop, planning marketing campaigne

Example:

online-shop, planning marketing campaigne

To which clients you should sent a new offer?

Example:

online-shop, planning marketing campaigne

To which clients you should sent a new offer?

monitoring two groups of clients

Group **PLUS** : buying products more or less often

Example:

online-shop, planning marketing campaigne

To which clients you should sent a new offer?

monitoring two groups of clients

Group **PLUS** : buying products more or less often

Group **MINUS** : just browsing list of offers/products more or less often
but (almost) have not bought anything so far

Example:

online-shop, planning marketing campaigne

To which clients you should sent a new offer?

monitoring two groups of clients

Group **PLUS** : buying products more or less often

Group **MINUS** : just browsing list of offers/products more or less often
but (almost) have not bought anything so far

To which clients you should sent a new offer?

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

Example:

online-shop, planning marketing campaigne

To which clients you should sent a new offer?

monitoring two groups of clients

Group **PLUS** : buying products more or less often

Group **MINUS** : just browsing list of offers/products more or less often
but (almost) have not bought anything so far

To which clients you should sent a new offer?

On class-based outlier detection

Luboš Popelínský, popel@fi.muni.cz
DML & KDLab
Faculty of Informatics, FI MU Brno

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

dcc f.sciencias up 20 nov 2019

Class-based Outliers

each example belongs to a class

Class-based outliers are those cases that

look anomalous when the class labels are taken into account,

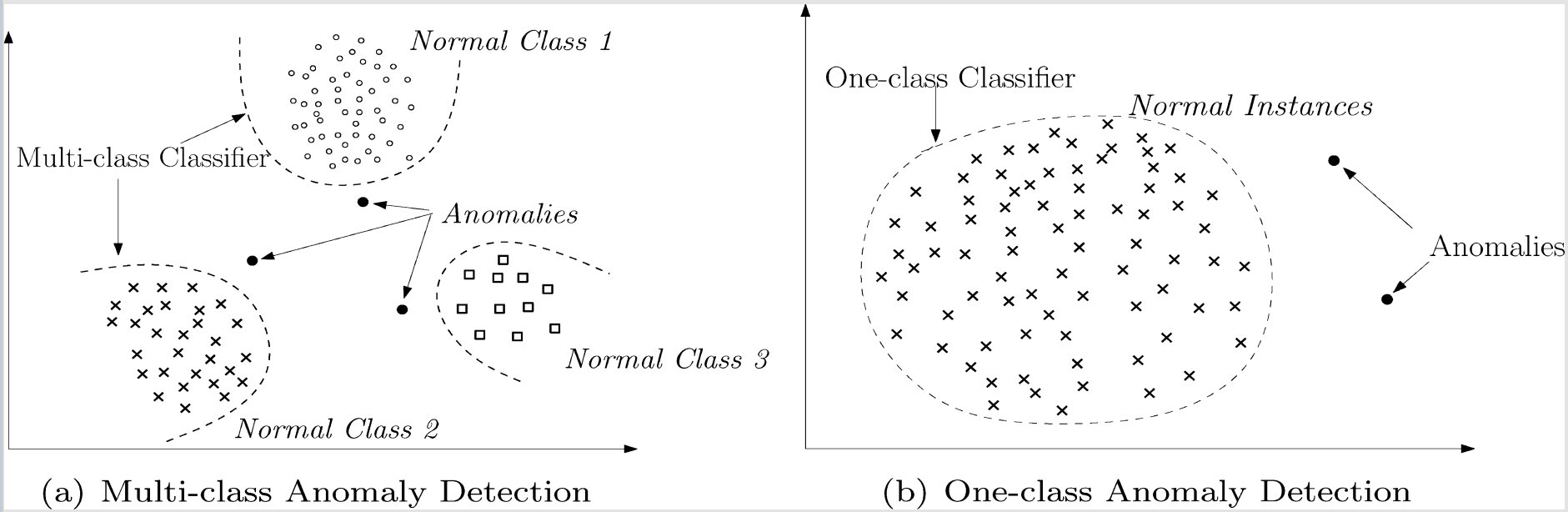
but they

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-based Outlier Detection

- sometimes called ‘semantic outlier’



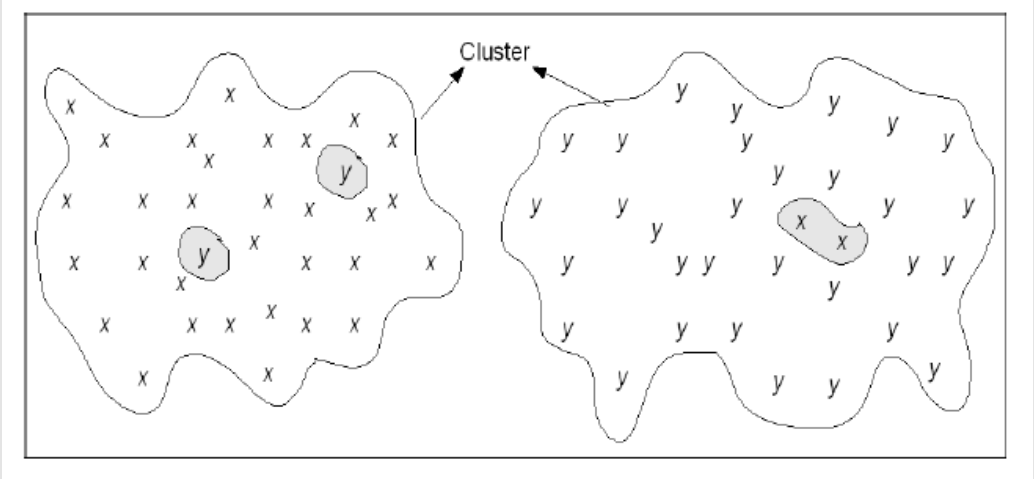
Multi-class outlier detection

[Han et al. 2012] *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

+ easy to use – some outliers cannot be detected



Class-based outlier factor. How to compute

[He et al. 2004]
Mining Class Outliers: Concepts, Algorithms and Applications in CRM.

Class outlier factor (COF)

Semantic outliers; clustering-based

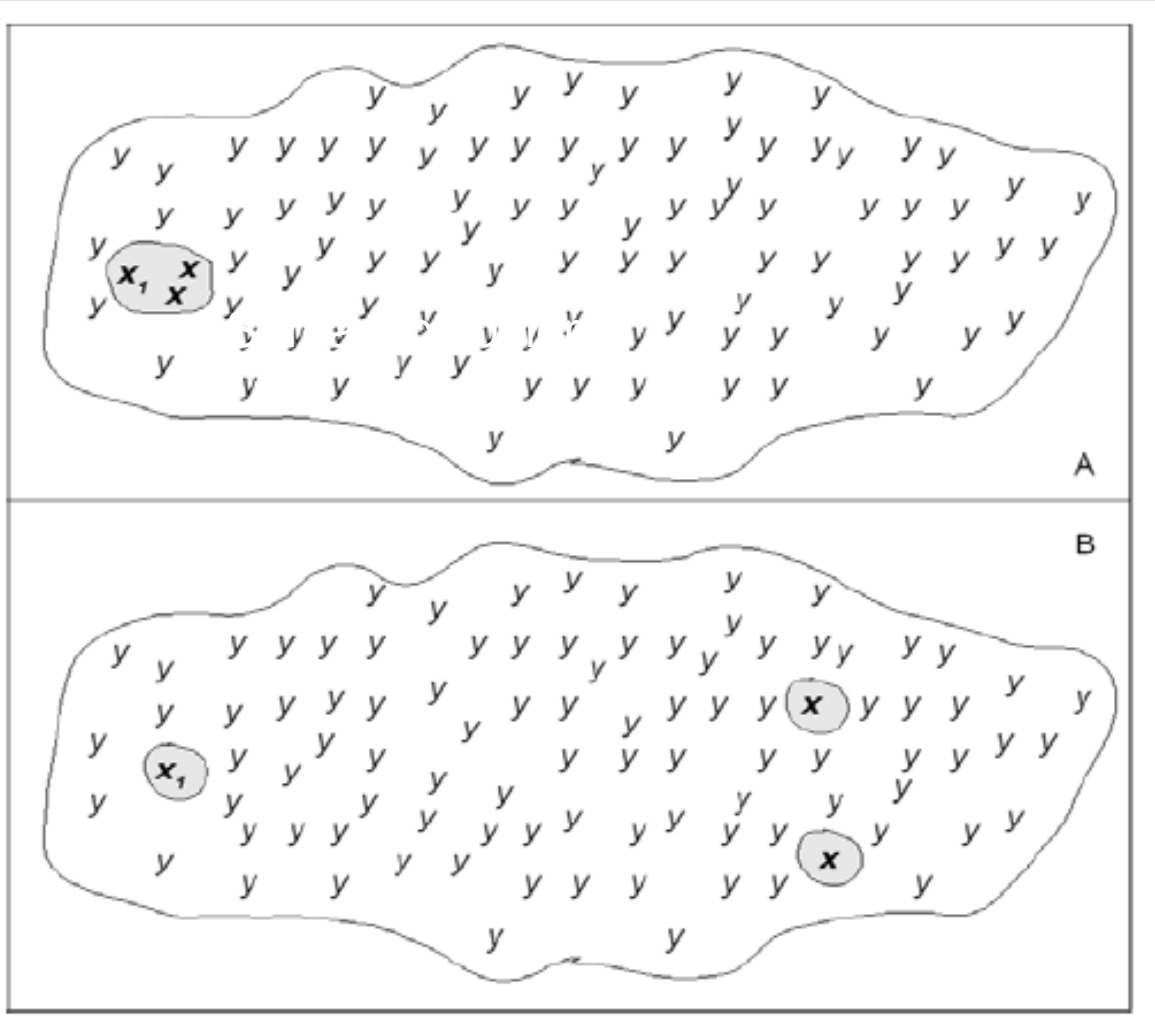
COF = OF w.r.t. own class (+) OF w.r.t. the other class/es

Pros & Cons

Semantic outliers (cont.)

x_1 has the same rank

To fix it . . .



[Hewahi and Saad 2007]

use a supervised machine learning algorithm

ROBUST-C4.5 [John 1995]

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

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).

CODB [Hewahi and Saad 2007]

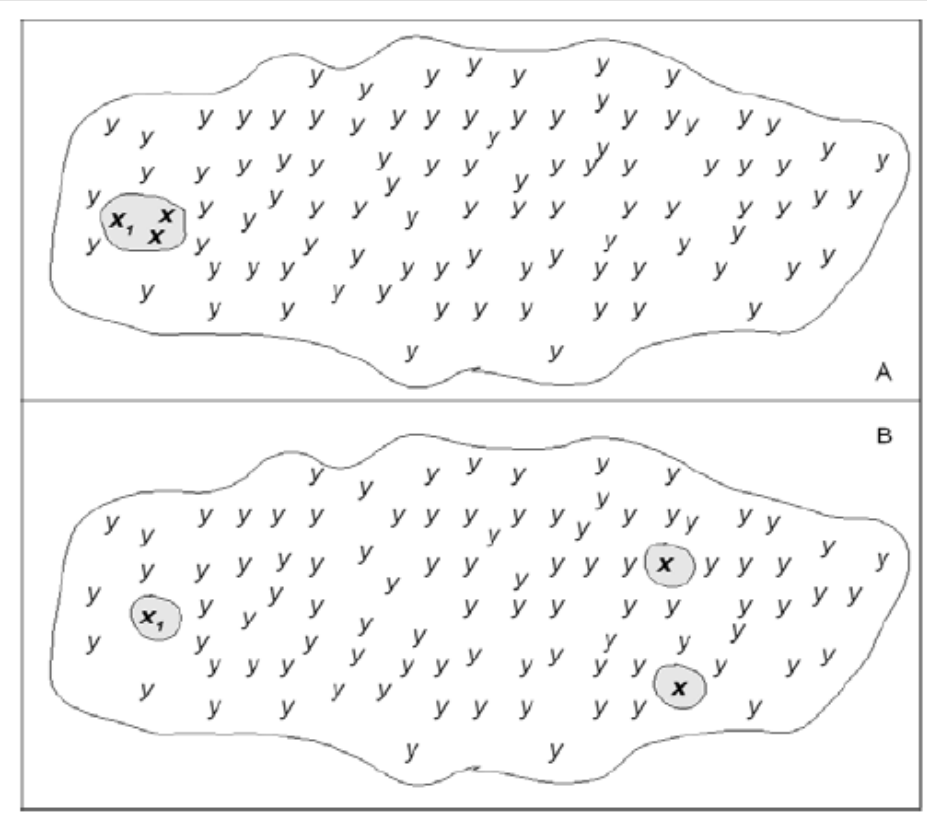
- combination of

  distance-based

  and density-based approach

  w.r.t class attribute

no need for clustering



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

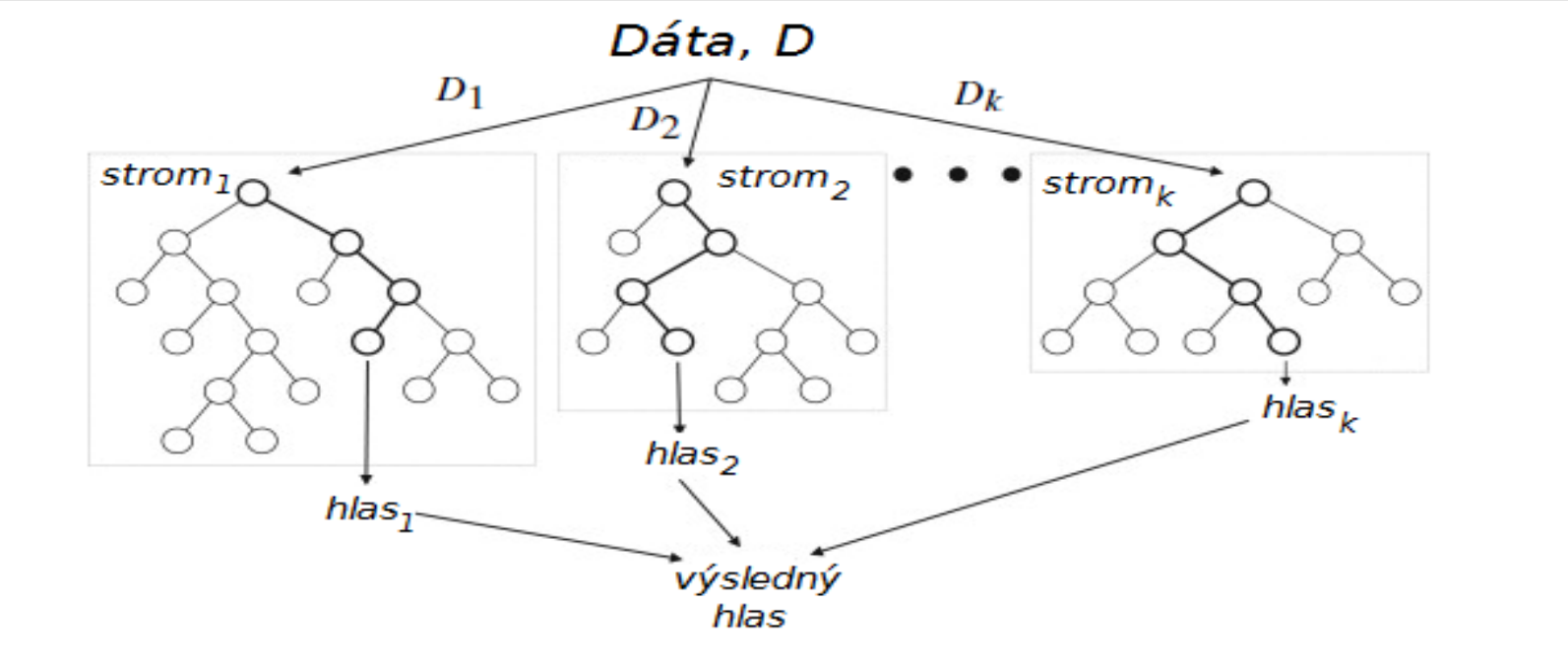
Dev(T) ... ... the sum of distance from all other elements from the same class

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 [Breiman 2000] is an ***ensemble*** classification and regression approach

employs 1. bootstraping and 2. random tree learning



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.

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

+

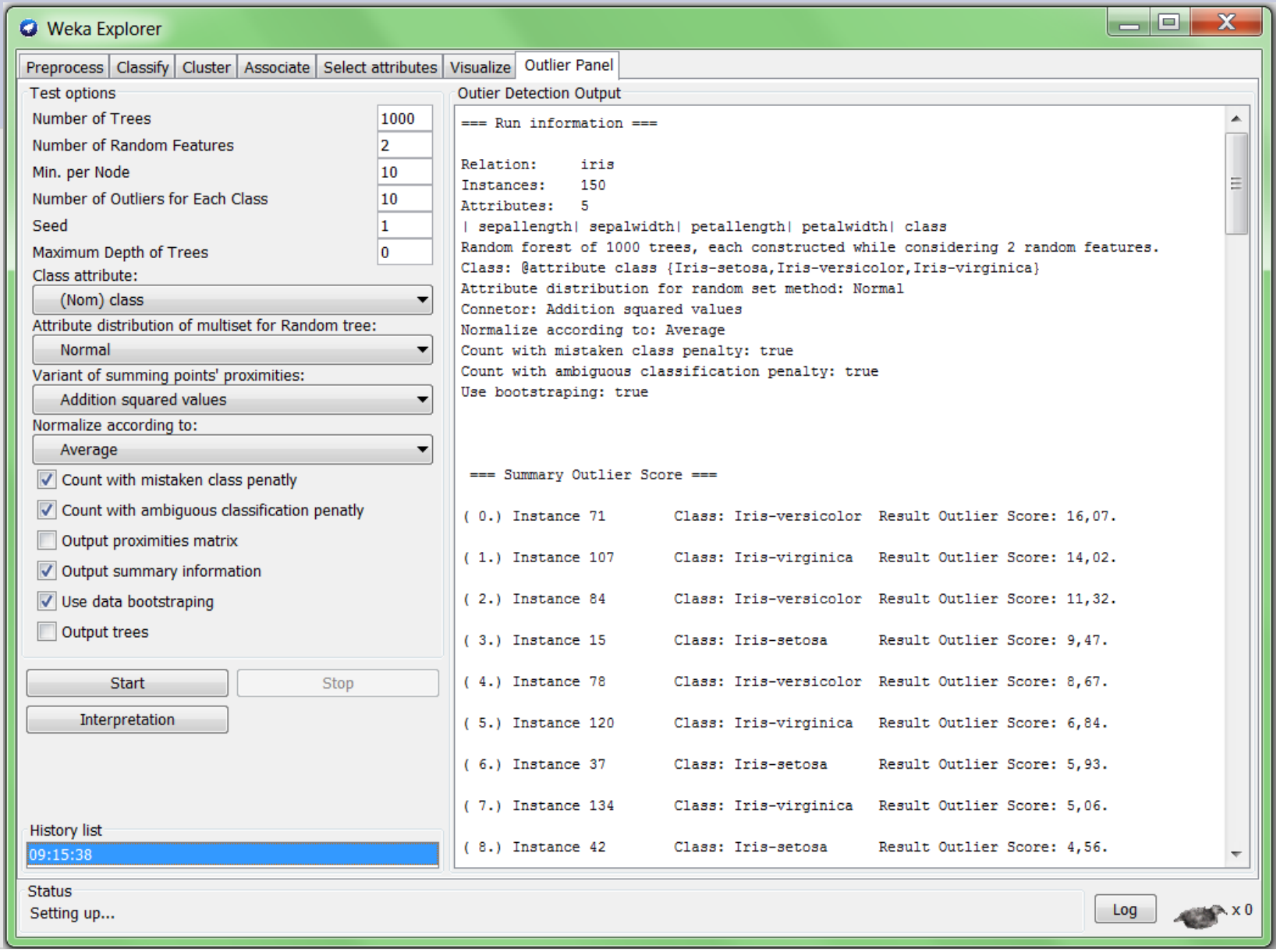
Ambiguity measure – a percentage of ambiguous classification

RF-OEX

Detection

+

explanation



Applications

E-shop: Clients vs. potential clients

ZOO

Educational data mining:

Intro to logic: Finding anomalous solutions

Students with standard/non-standard study interval

IMDb

Czech Parliament

Data pre-processing

..


JUERGEN FREUND

or·ni·tor·rin·co

(*ornito*-+ grego *rhúgkos, -eos*, focinho)

Género de monotremos de corpo alongado e cujo focinho se assemelha a um bico de pato.

ptakopysk

vtákopysk

dziobak

ornithorynque

schnabeltier

Applications

E-shop: Clients vs. potential clients

ZOO

Educational data mining:

Intro to logic: Finding anomalous solutions

Students with standard/non-standard study interval

IMDb

Czech Parliament

Data pre-processing

..

Teaching Logic: Finding student anomalous solutions

Task: Build a resolution proof, 400 students, at least 3 tasks to solve

Automated evaluation: error detection

Two classes CORRECT, INCORRECT

If a serious error appeared, the solution is classified as incorrect (ignoring typos)

Teaching logic: Finding anomalous solutions (cont.)

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

frequent pattern mining, frequent subgraphs

One attribute for each higher-level generalized pattern
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

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

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

IMDb Movie database: Funny reviews

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 *… ****
positive *******…***

IMDb: Ambiguous or funny reviews

A positive review with very poor actor ratings

> Tsui Hark's visual artistry is at its peek in this movie. **Unfortunately the terrible acting by Ekin Cheng and especially Cecilia Cheung (I felt the urge to strangle her while watching this, it's that bad :) made it difficult to watch at times.**
>
> This movie is a real breakthrough in the visual department. When I first saw this, my jaw dropped repeatedly and I thought to myself that I've never seen

IMDb: Ambiguous or funny reviews

A positive review of a film about extreme human poverty

> Kurosawa is a proved humanitarian. This movie is totally about people living in poverty. You will see nothing but angry in this movie. It makes you feel bad but still worth. All those who's too comfortable with materialization should spend 2.5 hours with this movie.

Applications

E-shop: Clients vs. potential clients

ZOO

Educational data mining:

Intro to logic: Finding anomalous solutions

Students with standard/non-standard study interval

IMDb

Czech Parliament

Data pre-processing

..

Applications

Current: Coming back to cleaning data

Outlier filtering followed by supervised learning algorithm

Can we improve performance by outlier filtering?

Thanks for your attention

popel@fi.muni.cz
www.fi.muni.cz/~popel

Literature

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

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

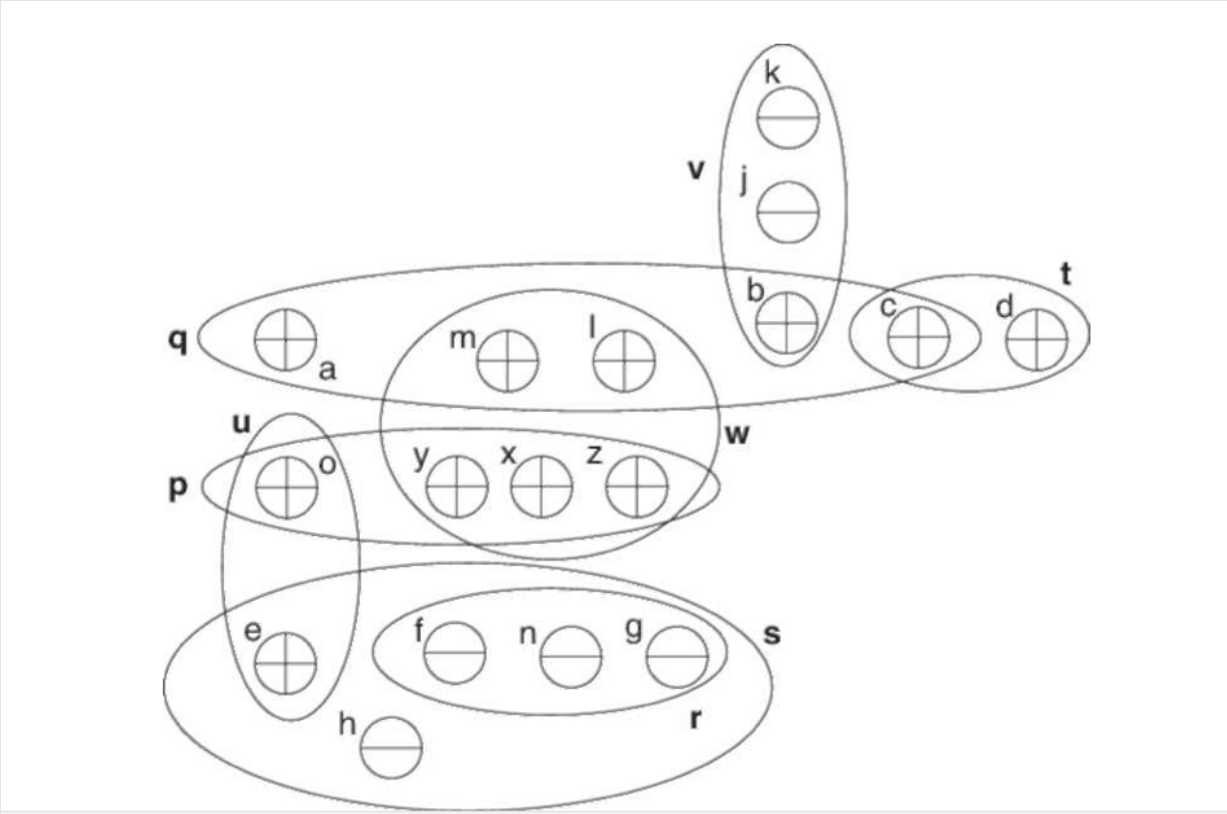
He Z. et al. *Mining Class Outliers: Concepts, Algorithms and Applications in CRM.* Expert Systems and Applications, ESWA 2004, 27(4), pp. 681-697, 2004.

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, 2012

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

Future/Ideas

(Naive) bayes classifier, 2 classes,
Sureness around 50% => outlier

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)
```

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:

E-shop: Clients vs. potential clients

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

- The raw outlier measure for case n is defined as

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

Own the rights?

Buy it at Amazon

More at IMDb Pro

Discuss in Boards

Add to Watchlist

Update Data

Quicklinks

reviews

Reviews & Ratings for

The Lion King II: Simba's Pride (V)

More at IMDbPro »

Write review

Filter: Best Hide Spoilers:

Page 1 of 16: [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11]

Index 160 reviews in total

41 out of 56 people found the following review useful:

Why is this Movie Given So Much Crap?

★★★★★★★★★★

Author: apeclaw2011 from United States

7 October 2005

I don't understand why this movie is regarded to as trash. Of course it is not as good as the first movie but it comes pretty stinkin close! The animation is actually equal too the quality of the original movie. I think that it is the most perfect Disney sequel ever! It is a very interesting story that shows Simba as a father. It is cool because you get to see Simba has now become basically, like his father. Every time I see this movie, I can feel that Simba has the same sense of power that Mufasa had. It has a fun and sweet story line and a great ending. When this movie was being made, the goal was to create a sequel to a movie that everyone loves so that they could spend more time with the characters. I think (despite what everyone say's) they created an awesome, spectacular Disney film!