Class-based Outlier Detection

Class-based outliers

Why we need a new concept?

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 – subsets of groups PLUS and MINUS – to be sent with a new offer?

Class-based outliers

Definition

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

Multi-class outliers

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

introduce COF, a class outlier factor
COF = OF w.r.t. own class $(+)$ OF w.r.t. the other classes

disadvantage: how to define $(+)$ addition

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.

CODB

combination of distance-based and density-based approach w.r.t class attribute

in RapidMiner

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

$COF(T) =$

SimilarityToTheK-NearestNeighbors

... compare a class of T to classes of the neighbors

$+ \alpha$ * 1/**DistanceFromOtherElementsOfTheClass** ... Distance

$+ \beta$ * **DistanceFromTheNearestNeighbors** ... Density

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

RF-OEX

Random Forest-based method

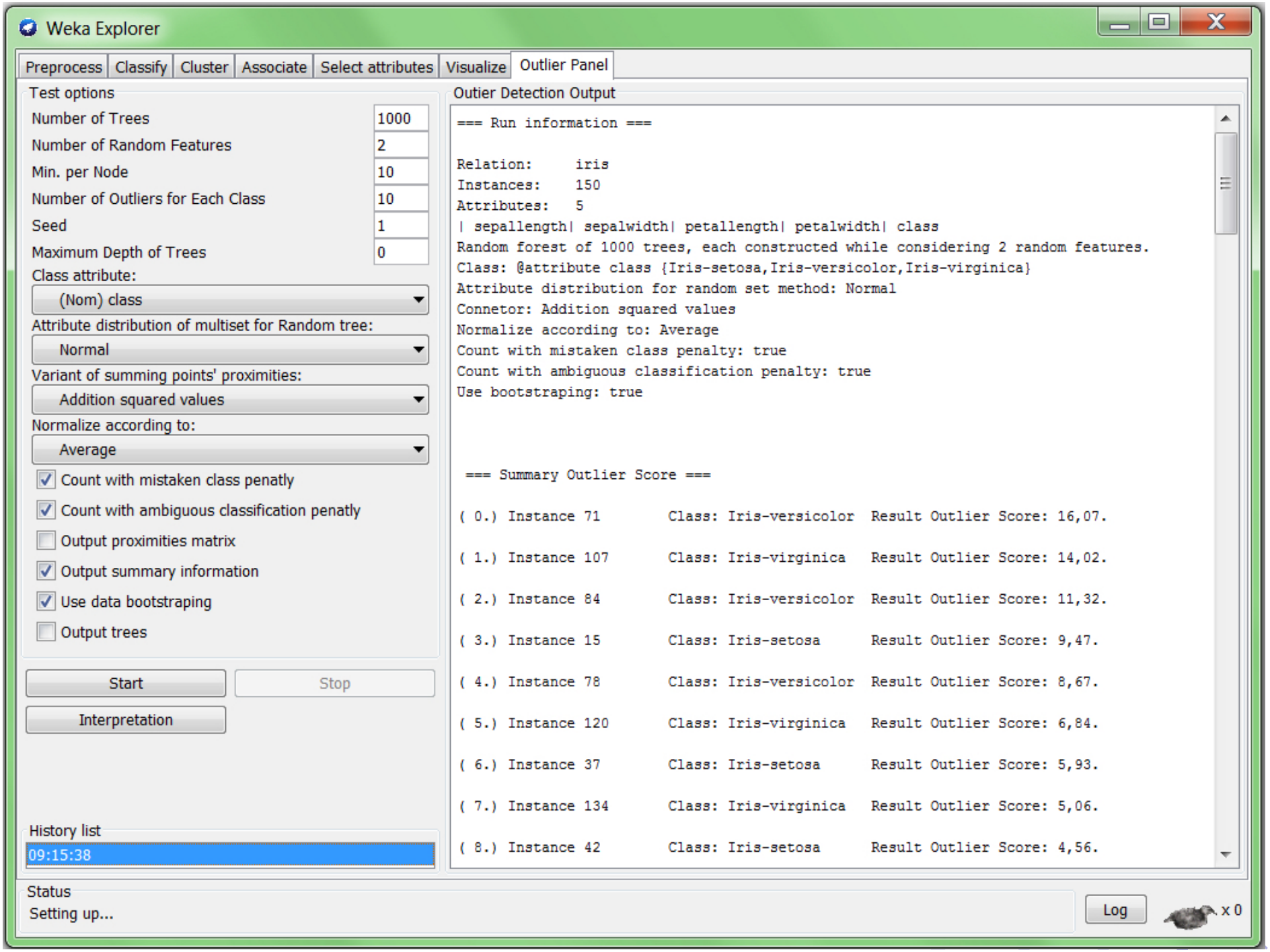
use proximity matrix for class outlier factor computation

COF = sum of three different measures of proximity or outlierness
COF =

- **Proximity** to the members of the same class
- + **Misclassication** - proximity to the members of other classes and
- + **Ambiguity** measure – a percentage of ambiguous classification

More: https://www.fi.muni.cz/~popel/685269/

NEZVALOVÁ, Leona, Lubomír POPELÍNSKÝ, Luis TORGO a Karel VACULÍK. Class-Based Outlier Detection: Staying Zombies or Awaiting for Resurrection? In Proceedings of IDA 2015.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Outlier Panel
Test options
Number of Trees
1000
Number of Random Features
2
Min. per Node
10
Number of Outliers for Each Class
10
Seed
1
Maximum Depth of Trees
0
Class attribute:
(Nom) class
Attribute distribution of multiset for Random tree:
Normal
Variant of summing points' proximities:
Addition squared values
Normalize according to:
Average
Count with mistaken class penatly
Count with ambiguous classification penatly
Output proximities matrix
Output summary information
Use data bootstraping
Output trees
Start
Stop
Interpretation
History list
09:15:38
Outier Detection Output
=== Run information ===
Relation: iris
Instances: 150
Attributes: 5
| sepallength| sepalwidth| petallength| petalwidth| class
Random forest of 1000 trees, each constructed while considering 2 random features.
Class: @attribute class {Iris-setosa,Iris-versicolor,Iris-virginica}
Attribute distribution for random set method: Normal
Connetor: Addition squared values
Normalize according to: Average
Count with mistaken class penalty: true
Count with ambiguous classification penalty: true
Use bootstraping: true
=== Summary Outlier Score ===
(0.) Instance 71 Class: Iris-versicolor Result Outlier Score: 16,07.
(1.) Instance 107 Class: Iris-virginica Result Outlier Score: 14,02.
(2.) Instance 84 Class: Iris-versicolor Result Outlier Score: 11,32.
(3.) Instance 15 Class: Iris-setosa Result Outlier Score: 9,47.
(4.) Instance 78 Class: Iris-versicolor Result Outlier Score: 8,67.
(5.) Instance 120 Class: Iris-virginica Result Outlier Score: 6,84.
(6.) Instance 37 Class: Iris-setosa Result Outlier Score: 5,93.
(7.) Instance 134 Class: Iris-virginica Result Outlier Score: 5,06.
(8.) Instance 42 Class: Iris-setosa Result Outlier Score: 4,56.
Status
Setting up...
Log
x 0

ILP. Rule-based approach.

Given E^+ positive and E^- negative examples and the background knowledge B, **learn concept** C **and dual concept** C_1 (swap positive and negative examples). C and C_1 are pure logic programs.

Look for examples that if removed from the learning set **change** the description (logic program) of C **and** C_1 **significantly**
i.e. difference of coverage is greater then a threshold.

= outliers

ANGIULLI, Fabrizio; FASSETTI, Fabio. Exploiting domain knowledge to detect outliers. Data Mining and Knowledge Discovery. 2014, vol. 28, no. 2,

Case studies

- **Educational Data mining** Correct vs incorrect student solutions in logic
- **Czech Parliament** 44 most important votings. Deputies that looks anomalous if compared with other members of the same party
- **Small and medium enterprises** (growing/non-growing)
- ...

IMDb

- **Star ratings** vs. **sentiment of a review**
- transform 0..10 stars into positive/negative rating
- perform 2-class sentiment analysis of a review
- used RF-OEX, CODB and LOF (LOF for each class separately)

IMDb. Example of results

positive review, actors horrible

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. ...

positive review to a realy bad horror that cannot be taken seriously

People are seeing it as a typical horror movie that is set out to scare us and prevent us from getting some sleep. Which if it was trying to do then it would deservedly get a 1/10.
The general view on this movie is that it has bad acting, a simple script that a 10 year old could produce and that it cant be taken seriously...

...

Open challenges

- two groups A, B, a member of A pretends to be in B
- Filtering outliers to improve (classifier) accuracy
- Anomalies in multi-modal data

Updated version of this part and the next one can be found here

https://www.fi.muni.cz/~popel/685269/

Explanation of rare events

Need for explanation of outliers

- A user need to understand why an instance is detected as an outlier
- For many applications, **explanation** (interpretation, description, outlying property detection, characterization) of outliers is as important as identification
- Outlier factor (degree) and ranking is only quantitative information
- Not only for high-dimensional data we need qualitative information

Based also on *ODD v5.0: Outlier Detection De-constructed ACM SIGKDD 2018 Workshop* keynote speeches, namely Making sense of unusual suspects - Finding and Characterizing Outliers (Ira Assent) and Outlier Description and Interpretation (Jian Pei)

How to generate explanation?

- Compare with inlying data as well as confirmed outlying data
- Find outlier explanatory component / outlying property / outlier context / outlier characteristic
- Help domain expert in verifying outliers and understanding how the outlier method works

What is meaningfull explanation

A method for finding of explanation must be

- **helpful** for a user, namely easy to understand. E.g. the smallest subset of attributes
- **efficient**, scalable

Most frequent approaches

- visual
- look for **a subset of attributes** where each outlier has its own explanatory subspace

Finding the most important attributes

For an object q, find the subspaces where q is most unusual compared to the rest of the data

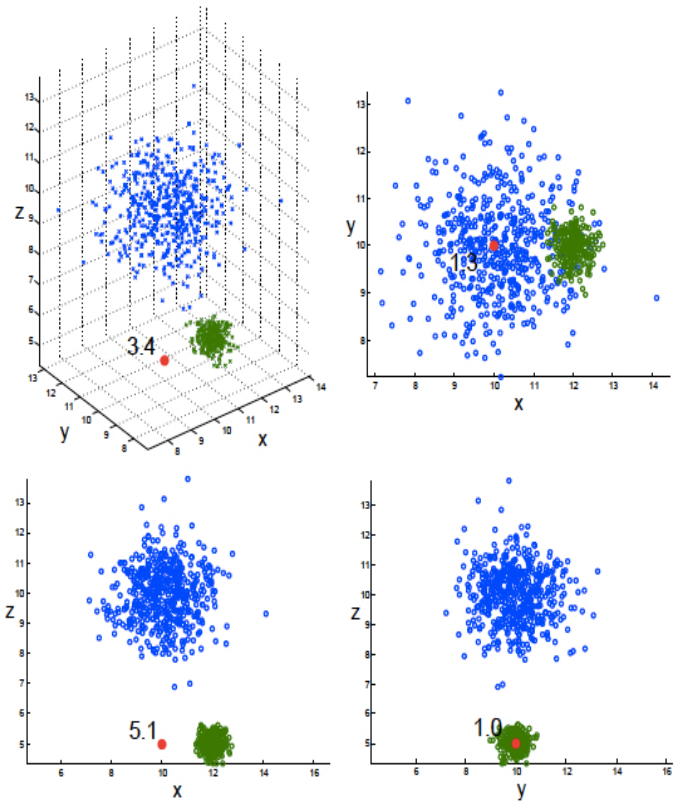


Figure 4.1: A 3D space $\{x, y, z\}$ and all its 2D projections. $\{x, z\}$ is an explanatory subspace.

A 3D space $\{x, y, z\}$ and all its 2D projections. $\{x, z\}$ is an explanatory subspace (Micenkova 2015)

Strongest, weak and trivial outliers

Knorr and Ng 1998

Non-trivial outliers

P is a *non-trivial outlier* in space A if P is not an outlier in any subspace of A.

Strongest outlier

The space A containing one or more outliers is called a *strongest outlying space* if no outlier exist in any subspace of A.

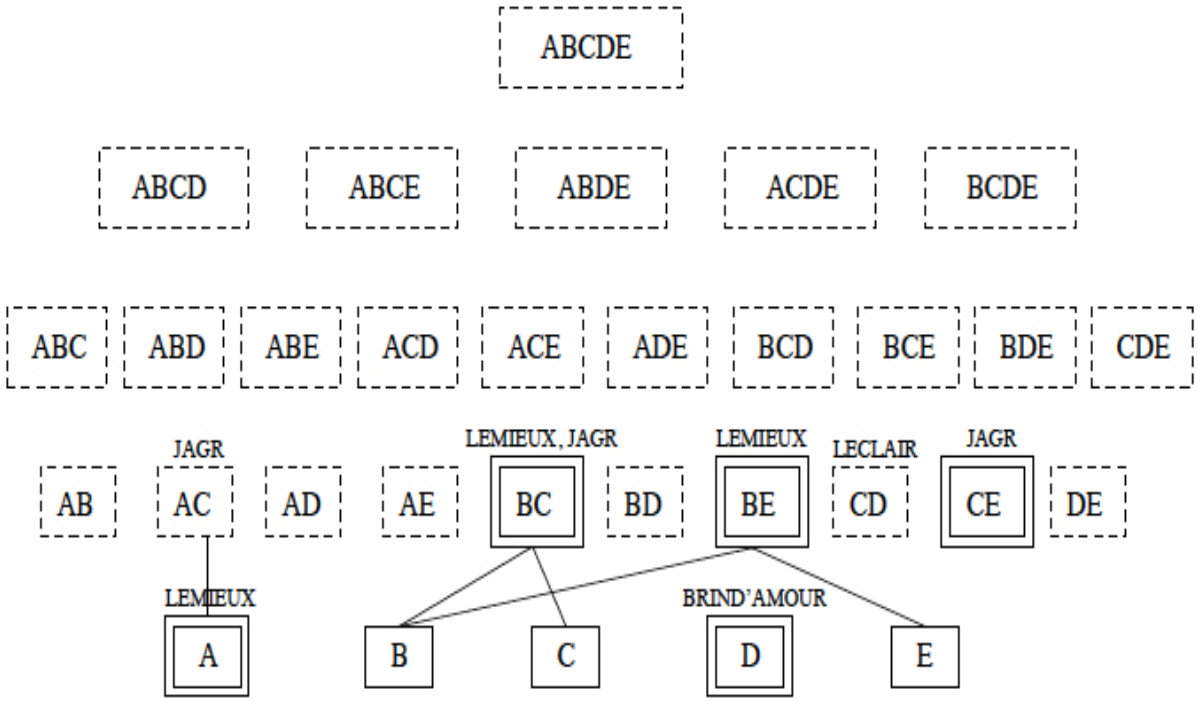
Any P that is an outlier in A is called a *strongest* outlier.

Any non-trivial outlier that is not strongest is called *weak* outlier.

Example: NHL ice hockey players

Knorr and Ng 1999

5-D space $\{A, B, C, D, E\}$ of power-play goals, short-handed goals, game-winning goals, game-tying goals, and game played



Lattice representation

Explaining outliers by subspace separability

(Micenkova and Ng 2013)

- Cannot derive explanatory subspace just by analyzing vicinity of the point in full space $\Rightarrow$ need to consider different subspace projections
- no monotonicity property for outliers wrt. subspaces
- need for heurstics because of exponential complexity,

look for a subspace A where the outlier factor is high and the dimension of A is low

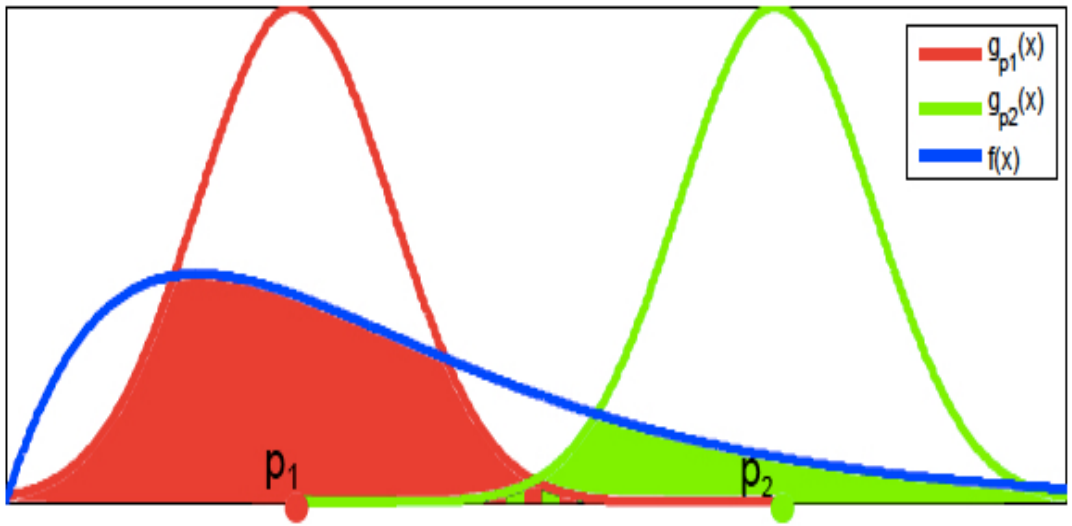
- separability - instance outlierness is related to its separability from the rest of the data

B. Micenková, R. T. Ng, X. H. Dang, and I. Assent. Explaining outliers by subspace separability. In IEEE ICDM 2013

Outlierness as accuracy of classification

(Micenkova and Ng 2013)

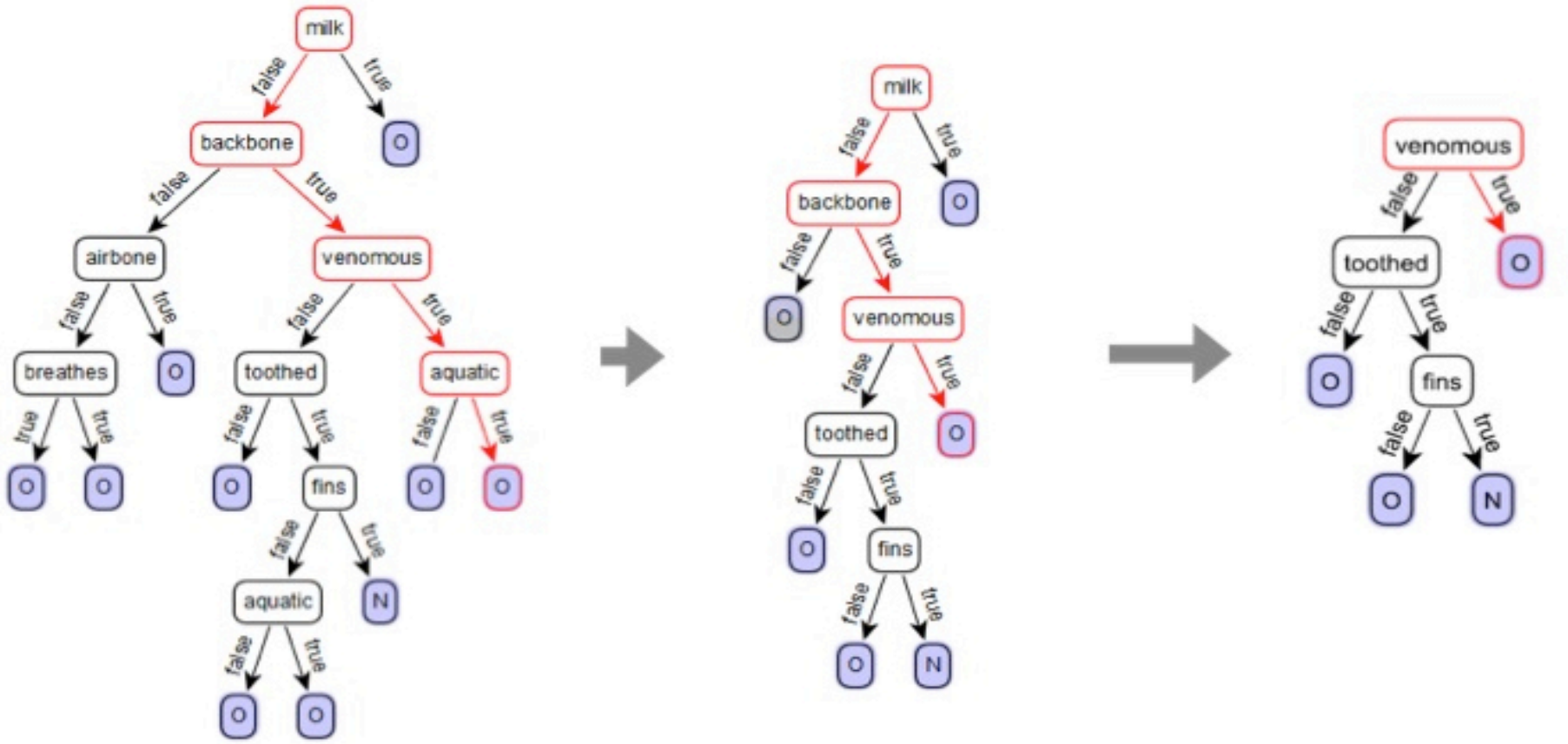
- separablity as error at classification. Assume that the data follows a distribution f
- original data = inlierclass; outlier + artificial points = outlierclass
- use standard feature selection methods to find explanatory subspaces



Measuring outlierness by separability. $p1, p2$ are points from the distribution $f(x)$ and the normal distributions $g_{p1}(x)$ and $g_{p2}(x)$ were artificially generated.

RF-OEX: Analysis of Random Forest

two methods: 1. search for frequent branches and **2. reduction of trees**



NEZVALOVÁ, Leona et al. Class-Based Outlier Detection: Staying Zombies or Awaiting for Resurrection? In Proceedings of IDA 2015.

RF-OEX

Examples of explanation

Form: (Condition, certainty factor)

Zoo dataset

Instance number: 64, Class: mammal

eggs=true, 0.51

toothed=false, 0.49

Iris dataset

Instance number: 19, Class: Iris-setosa

sepallength >= 5.5 && sepalwidth < 4, 0.53

sepallength >= 5.5, 0.47

Recent work

Beyond Outlier Detection: LookOut for Pictorial Explanation, ECML PKDD. (Gupta et al. 2018)

Explaining anomalies in groups with characterizing subspace rules.Data Mining and Knowledge Discovery (2018) 32 (Macha and Akoglu 2018)

Oui! Outlier Interpretation on Multi-dimensional Data viaVisual Analytics Eurographics Conference on Visualization (EuroVis) (Xun Zhao et al. 2019)

Sequential Feature Explanation for Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, Vol. 13, No. 1, (Siddiqui et al. 2019)

Towards explaining anomalies. A deep Taylor decomposition of one-class models. Pattern Recognition 101 (2020) 1071098 (Kauffmann et al. 2020)