Imbalanced Domain Learning:

Concepts, Challenges and Strategies

Rita P. Ribeiro
(rpribeiro@fc.up.pt)

Outline

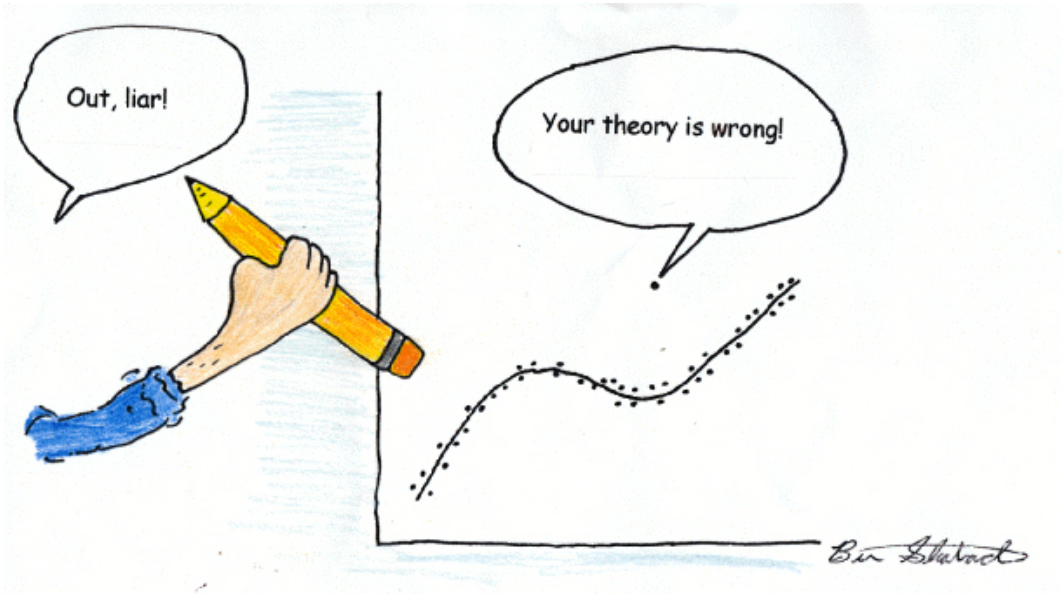
Next Subsection

Motivation

- Most data mining tasks focus on creating a model of the "normal" patterns in the data, extracting knowledge from what is common (e.g. frequent patterns).
- Still, rare patterns can also give us some important insights about data.
- Depending on the goal, those insights can be even more interesting/critical than the "normal" patterns.

What is an Outlier?

- *"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)



What is an Outlier? (cont.)

- Outliers represent patterns in data that do not conform to a well defined notion of normal.

- Initially, outliers were considered errors and their identification had data cleaning purposes.

- However, they can represent truthful deviation of data.
- For some applications, they represent critical information, which can trigger preventive or corrective actions.

Imbalanced Domain Learning

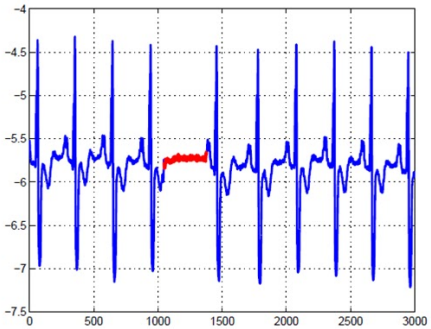
It is based on the following assumptions:

- the representativeness of the cases on the training data is not uniform;
- the underrepresented cases are the most relevant ones for the domain.

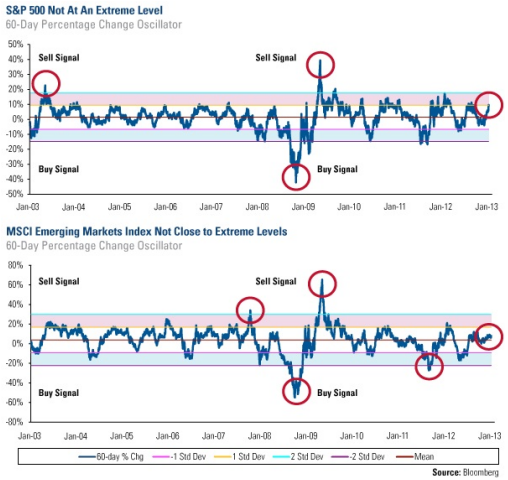
- The focus is on the identification of these scarce/outlier cases.
- But, the definition of these cases is dependent on the application domain knowledge.

Where can Outliers occur?

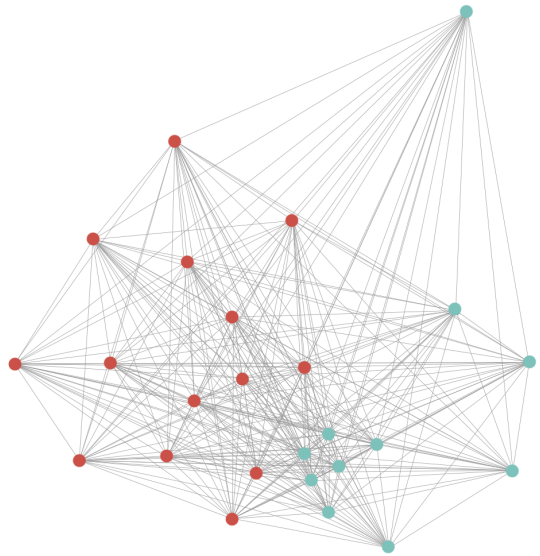
Medical Analysis



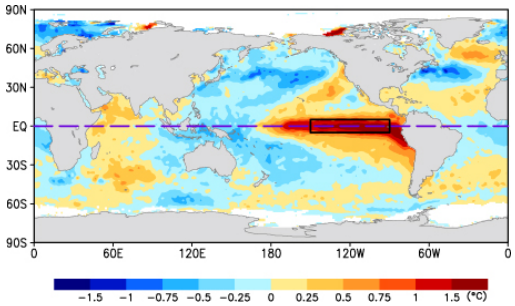
Financial Markets



Social Network Analysis

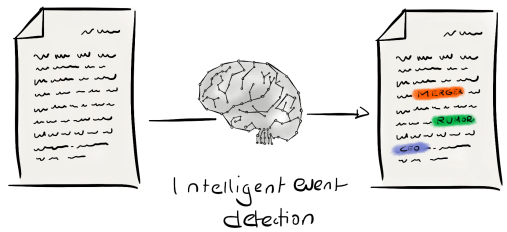


Anomalous Weather Patterns



Fraud Detection

Event Detection in Text/Social Media



Some Applications with Imbalanced Domains

- Financial Applications
  - Credit Card Fraud, Insurance Claim Fraud, Stock Market Anomalies
- (Cyber) Security Applications
  - Host-based, Network Intrusion Detection
- Medical Applications
  - Medical Sensor or Imaging for Rare Disease Diagnostics
- Text and Social Media Applications
  - Anomalous Activity in Social Networks, Fake News Detection
- Earth Science Applications
  - Sea Surface Temperature Anomalies, Environmental Disasters
- Fault Detection Applications
  - Quality Control, Systems Diagnosis, Structure Defect Detection

Challenges of Imbalanced Domain Learning

- Defining every possible "normal" behaviour is hard.
- The boundary between normal and outlying behaviour is often not precise.
- There is no general outlier definition; it depends on the application domain.
- It is difficult to distinguish real, meaningful outliers from simple random noise in data.
- The outlier behaviour may evolve with time.
- Malicious actions adapt themselves to appear as normal.
- Inherent lacks known labelled outliers for training/validation of models.

Key Aspects of Imbalanced Domain Learning

- Nature of Input Data
- Type of Outliers
- Intended Output
- Learning Task
- Performance Metrics

Nature of Input Data

- Each data instance has:
  - One attribute (univariate)
  - Multiple attributes (multivariate)
- Relationship among data instances:
  - None
  - Sequential / Temporal
  - Spatial
  - Spatio-temporal
  - Graph
- Dimensionality of data

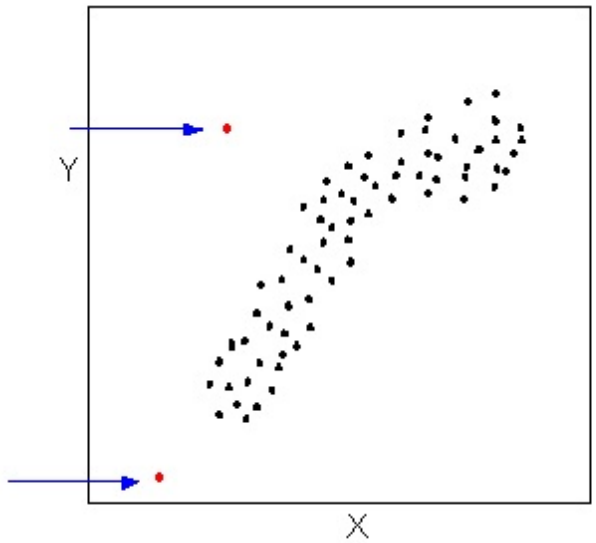
Types of Outliers

- Point (or Global) Outlier
- Contextual Outlier
- Collective Outlier

Types of Outliers (cont.)

Point Outlier

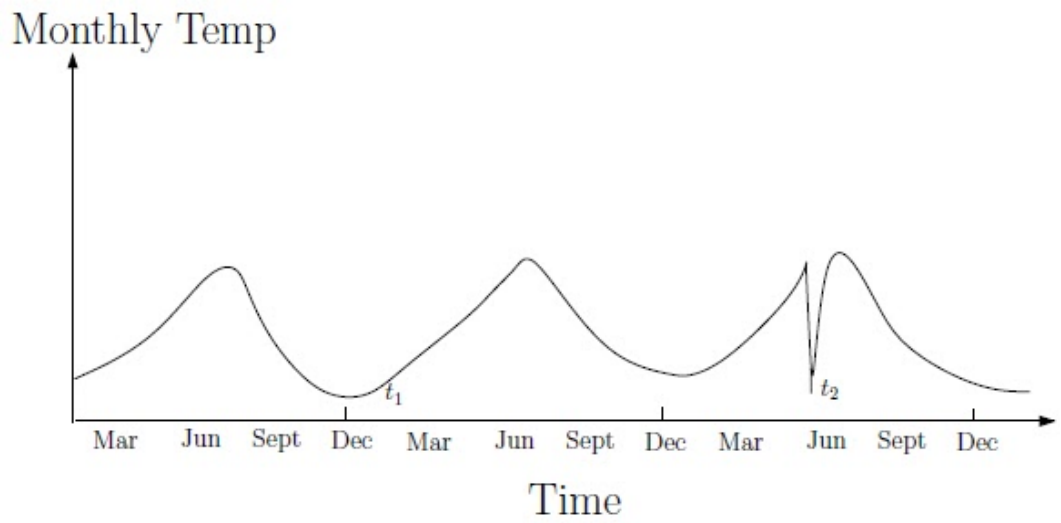
An instance that, individually or in small groups, is very different from the rest of the instances.



Types of Outliers (cont.)

Contextual Outlier

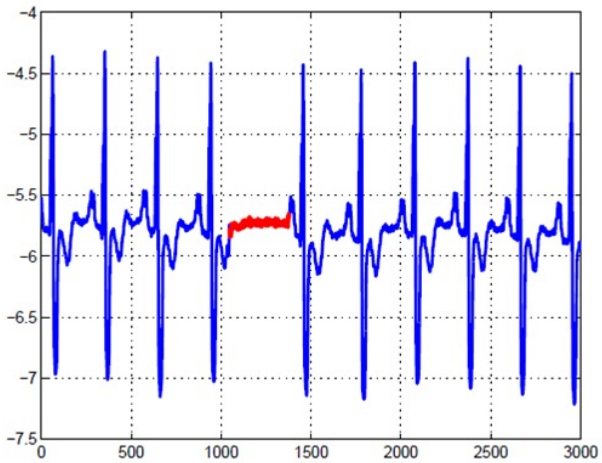
An instance that, when considered within a context, is very different from the rest of the instances.



Types of Outliers (cont.)

Collective Outlier

An instance that, even though individually may not be an outlier, inspected in conjunction with related instances and with respect to the entire data set is an outlier.



Intended Output

- Assign a label/value: identification normal or outlier instance.
- Assign a score: the probability of being an outlier.
  - allows the output to be ranked
  - requires the specification of a threshold

Learning Task

Supervised Outlier Detection

- data set has instances of both normal and outlier behaviour;
- hard to obtain such data in real-life applications.

Semi-supervised Outlier Detection

- data set has a few instances of normal or outlier behaviour;
- some real-life applications, such as fault detection, provide such data.

Unsupervised Outlier Detection

- data set has no information on the behaviour of each instance;
- it assumes that instances with normal behaviour are far more frequent;
- most common case in real-life applications.

Performance Metrics

Inadequacy of Standard Performance Metrics

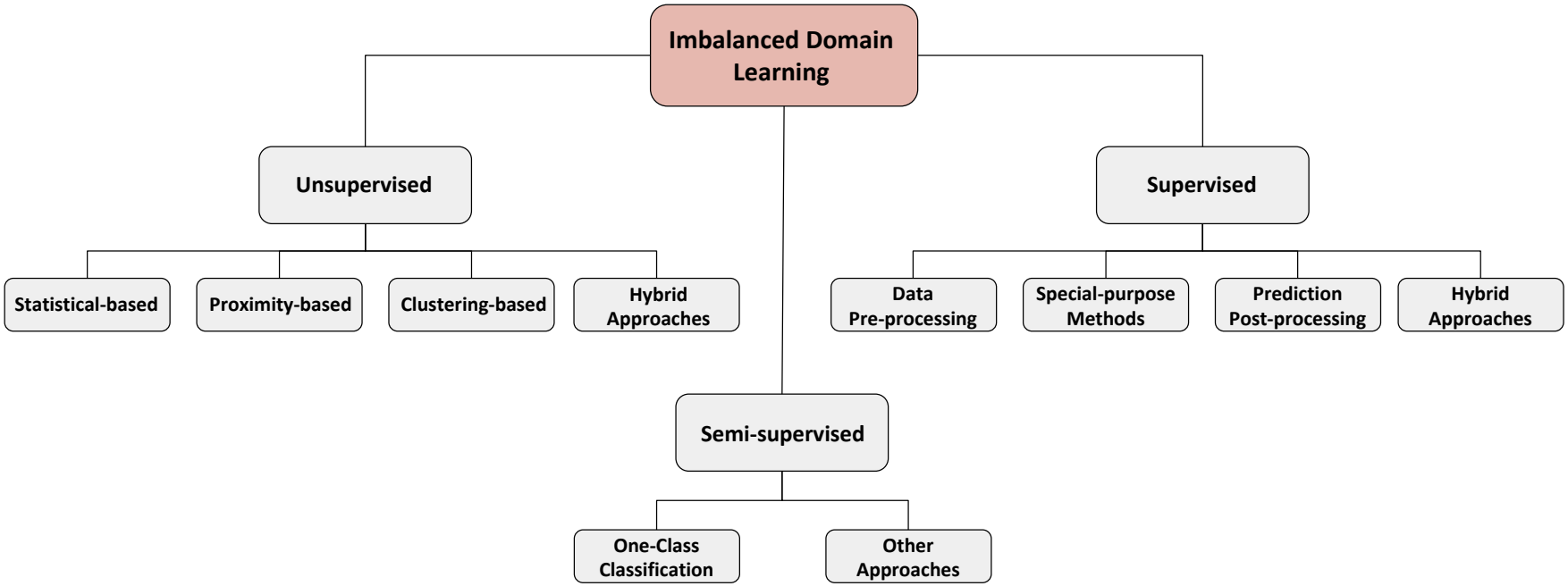
- Standard performance metrics (e.g. *accuracy*, *error rate*) assume that all instances are equally relevant for the model performance.
- These metrics would give a good performance estimation to a model that performs well in normal (frequent) cases and bad on outlier (rare) cases.

  Credit Card Fraud Detection:

  - data set D with only 1% of fraudulent transactions;
  - model M predicts all transactions as non-fraudulent;
  - M has an estimated accuracy of 99%;
  - yet, all the fraudulent transactions were missed!

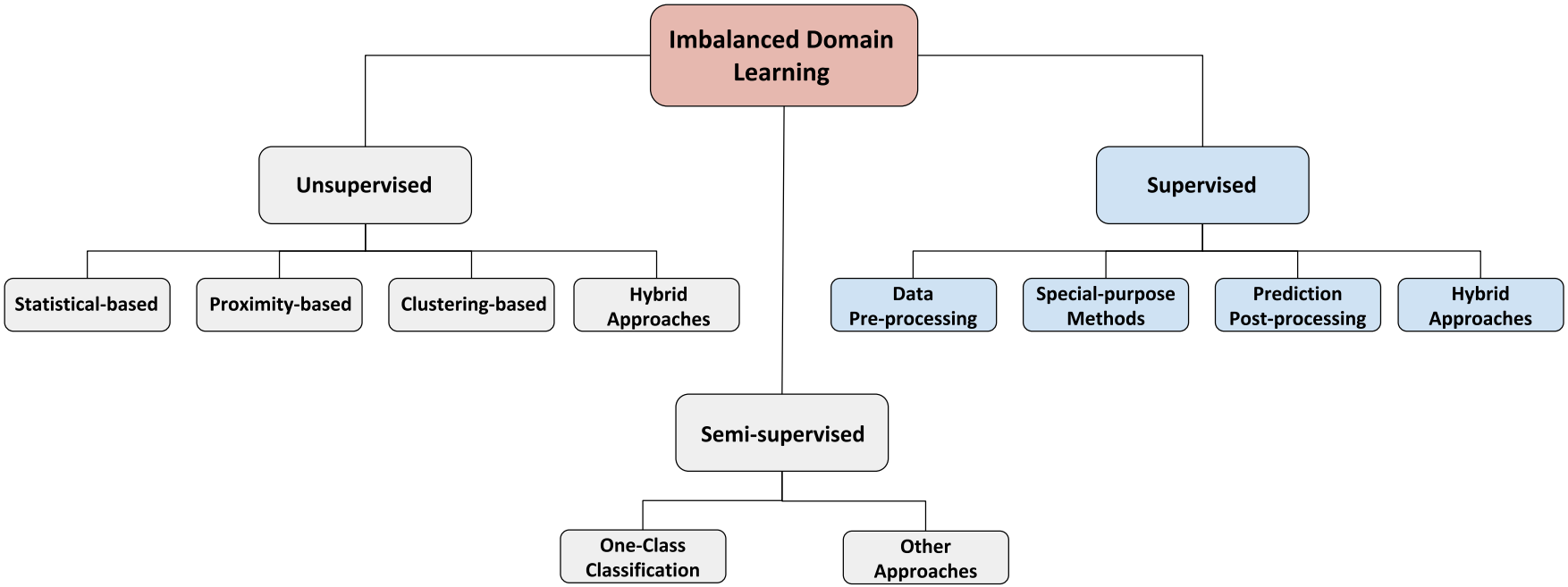
Next Subsection

Taxonomy of Imbalanced Domain Learning Approaches



Next Subsection

Taxonomy of Imbalanced Domain Learning Approaches



Imbalanced Predictive Learning

- In a supervised learning task the goal is to
  - build a model of an unknown function $Y = f(X_1, X_2, \cdots, X_p)$, based on a training sample $\{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$ with examples of this function.

- Depending on the type of target variable Y, we have:
  - classification task, if Y is nominal
  - regression task, if Y is numeric

Imbalanced Predictive Learning (cont.)

- In standard predictive learning tasks:
  - The preference is constant over all target variable values
  - The learning algorithm focuses on the most frequent cases to achieve an overall good performance
  - The cost of all similar errors and the benefit of all similar accurate predictions is the same.

Imbalanced Predictive Modelling

- **Non-uniform importance** of values **across the domain of the target variable** Y
- The **cases that are more relevant are poorly represented** in the training set.

Imbalanced Predictive Learning (cont.)

- Classification Example:
  - when analyzing credit card transactions, the fraudulent ones are of key importance
- Regression Example
  - when analyzing the air quality, the extreme values of concentration of a pollutant are of key importance

Imbalanced Predictive Learning (cont.)

Challenges in Imbalanced Predictive Learning:

1. How to specify the most important subset(s) of values of the target variable?
2. How to properly evaluate the performance of models regarding these cases?
3. How to bias the learning algorithms to these rare cases?

1-How to specify the most important cases?

Relevance function $\phi(Y)$ [Torgo and Ribeiro, 2007, Ribeiro, 2011]

A relevance function $\phi(Y) : \mathcal{Y} \rightarrow [0, 1]$ is a function that expresses the application-specific bias concerning the target variable domain $\mathcal{Y}$ by mapping it into a $[0, 1]$ scale of relevance, where 0 and 1 represent the minimum and maximum relevance, respectively.

- The notion of relevance is applicable to both classification and regression problems.
- It can be used to build sets of rare and normal cases.

1-How to specify the most important cases? (cont.)

How to define the relevance function $\phi(Y)$?

- It can be provided by domain knowledge.
- Estimated from the target variable data distribution so that rare target classes/values are assigned more importance.
- How to define relevance function in imbalanced domains for:
  - 2-class problems
  - multi-class problems
  - regression problems

1-How to specify the most important cases? (cont.)

Imbalanced 2-class Problems

- The most simple case
- Example: Credit Card Fraud Detection
  "I'm interested in accurate predictions of fraudulent credit card transactions"

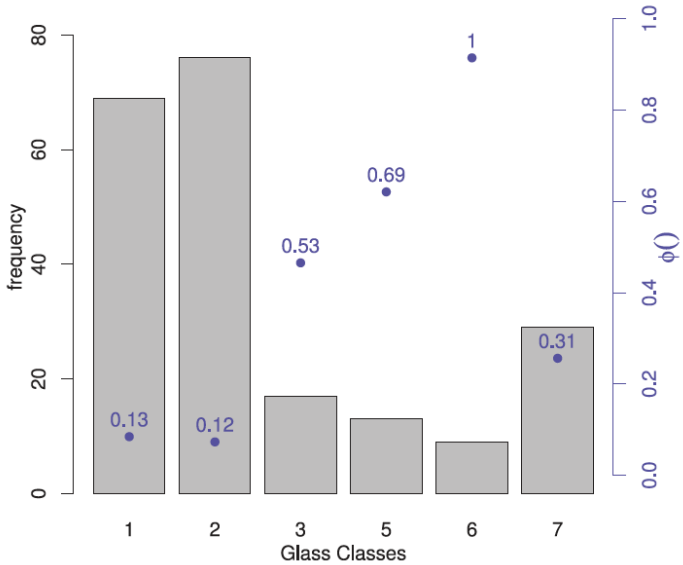
$$\phi(Y) = \begin{cases} 1 & \text{if } Y = \text{"fraud"} \\ 0 & \text{if } Y = \text{"normal"} \end{cases}$$

1-How to specify the most important cases? (cont.)

Imbalanced Multi-Class Problems

- We can ask the user to assign a relevance score to each class
- ... or use the sampling distribution of the target variable

- Example: Classification of types of glass found at crime scenes
- Cases labelled with infrequent glass type are the most important ones, as they can only be used as evidence if they are correctly identified



1-How to specify the most important cases? (cont.)

Imbalanced Regression Problems

- Given the potentially infinite nature of the target variable domain, specifying the relevance of all values is virtually impossible;
- an approximation is required.

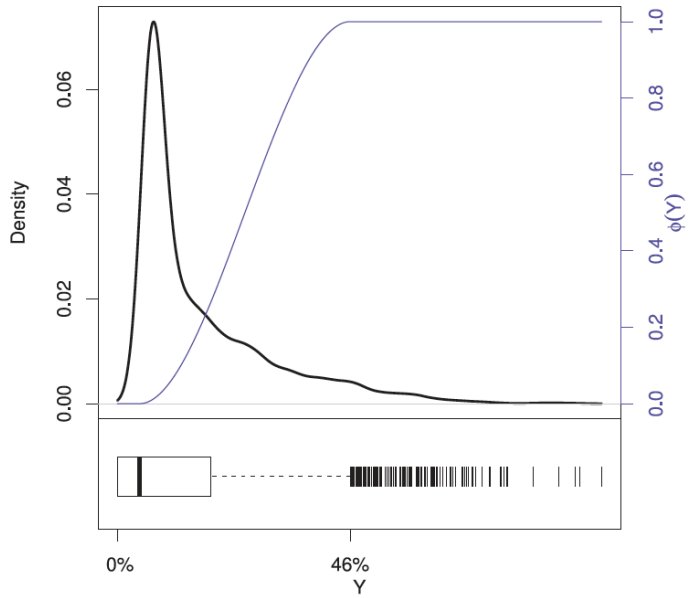
Ribeiro (2011) proposed two methods for estimating $\phi(Y)$:

- interpolation method
  - user provides a set of interpolating points
- automatic method
  - no input required from the user;
  - it uses the target variable distribution;
  - it assumes that the most relevant cases are located at the extremes of the target variable distribution.

1-How to specify the most important cases? (cont.)

- The target variable has an infinite domain, asking for relevance scores is unfeasible
- We can use the sampling distribution of the target variable

- Example: Prediction of NO_2 emissions for air quality assessment
- Cases which take rare and extreme values of NO_2 emissions are the most important ones, as it has an impact on human health.



For more details, check IRon R package available on https://github.com/nunompmoniz/IRon

2-Suitable Performance Metrics

- Why do we need new performance metrics?
  - Standard metrics (e.g. error rate or mean squared error) describe the average predictive performance of the models
  - When the user is focused on a small subset of rare values, the average is not a good idea
  - These metrics will be mostly influenced by the performance of the models on cases that are irrelevant to the user

2-Suitable Performance Metrics: Classification

- In 2-class imbalance problems, the positive class is the minority and most relevant class.

2-class Confusion Matrix				
		True		
		Negative	Positive	Total
Predicted	Negative	TN	FP	PNEG
	Positive	FN	TP	PPOS
	Total	NEG	POS	

- Standard performance metrics (e.g. *accuracy*) are not suitable for this type of problem.

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

2-Suitable Performance Metrics: Classification (cont.)

- Example: Diagnose of a rare disease

Model B Confusion Matrix			
		Disease	
		absent	present
Diagnose	negative	TN = 63	FN = 2
	positive	FP = 27	TP = 8

Model C Confusion Matrix			
		Disease	
		absent	present
Diagnose	negative	TN = 68	FN = 7
	positive	FP = 22	TP = 3

- The accuracy for both models is 71%.
- Model B correctly diagnosed 80% of the sick individuals
- Model C diagnosed only 30%
- The goal is to achieve a good performance on the rare but most important cases.

2-Suitable Performance Metrics: Classification (cont.)

- **precision**: proportion of (positive) events predicted by the model that are correct

$$precision = \frac{TP}{TP + FP}$$

- **recall**: proportion of the real (positive) events that are captured by the model

$$recall = \frac{TP}{TP + FN}$$

- But, maximizing one of them comes at the cost of the other
- It is easy to achieve 100% recall: always predict positive events
- What is difficult is to achieve high values for both precision and recall.

2-Suitable Performance Metrics: Classification (cont.)

- **F-measure**: trade-off measure between precision and recall.

$$F_\beta = \frac{(\beta^2 + 1) \cdot \textit{precision} \cdot \textit{recall}}{\beta^2 \cdot \textit{precision} + \textit{recall}}$$

where β controls the relative importance of *precision* and *recall*

  - when $\beta = 1$, F-measure is the harmonic mean between *precision* and *recall*
  - when $\beta \to 0$, the weight of *recall* decreases
  - when $\beta \to \infty$, the weight of *precision* decreases

2-Suitable Performance Metrics: Classification (cont.)

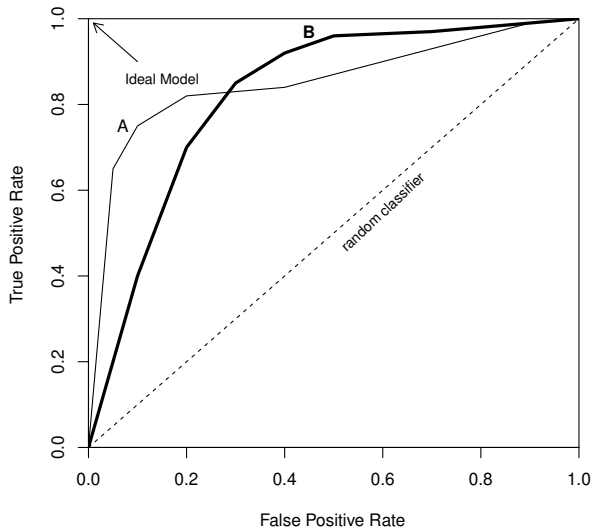
- False Positive Rate (**FPR**): proportion of normal cases wrongly predicted as (positive) events.

$$FPR = \frac{FP}{TN + FP}$$

- Receiver Operating Characteristic (**ROC**) **Curve**: trade-off between *TPR* (*recall*) and *FPR* as the discrimination threshold for the two classes varies.

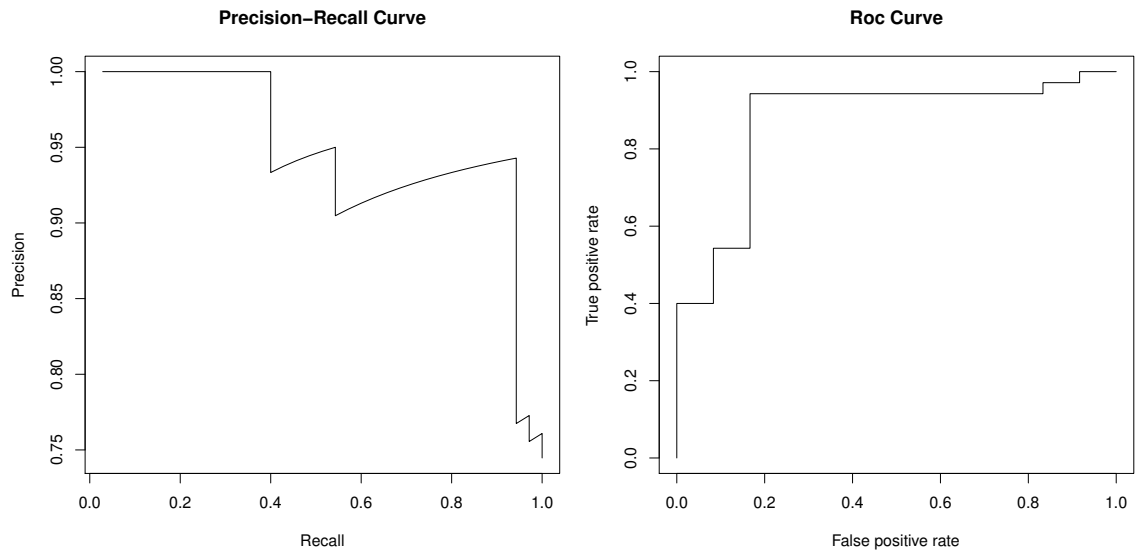
- Area Under Curve (**AUC**): performance measure that tells how good the model is in distinguishing the two classes.

  The higher the AUC, the better.



2-Suitable Performance Metrics: Classification (cont.)

- **PR Curve**: trade-off between *recall* and *precision* as the discrimination threshold for the two classes varies.
- As it does not account for TN, it is more suited for problems with class imbalance.



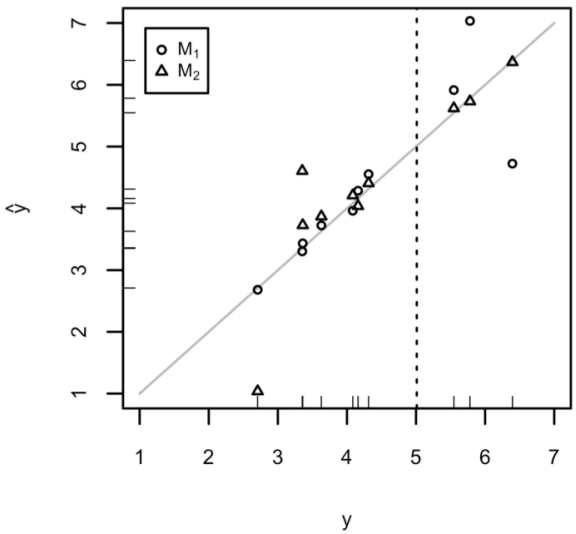
- **AUC-PR**: performance measure that tells how good the model is in distinguishing the target (positive) class.

2-Suitable Performance Metrics: Regression

- One of the most commonly used performance metrics in regression is

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Example: Prediction of NO_2 emissions



- Both M_1 and M_2 models achieve an MSE of 0.460
- Still, M_2 is more accurate at higher NO_2 concentration values, the most important to predict accurately.

- As in classification, standard performance metrics fail the goal
- The goal is to achieve a good performance on the outlier cases.

2-Suitable Performance Metrics: Regression (cont.)

- Some asymmetric loss functions have been proposed
- These proposals assume that the importance of the errors is not uniform
- Limitations
  - focused on distinguishing over- and under-predictions
  - threshold dependent (e.g. try to adapt classification metrics)
  - neglect performance on the overall target domain
  - few have been used in model optimization

2-Suitable Performance Metrics: Regression (cont.)

- Squared Error Relevance (SER_t) [Ribeiro and Moniz, 2020]

$$SER_t = \sum_{y_i \in \mathcal{D}^t} (\hat{y}_i - y_i)^2$$

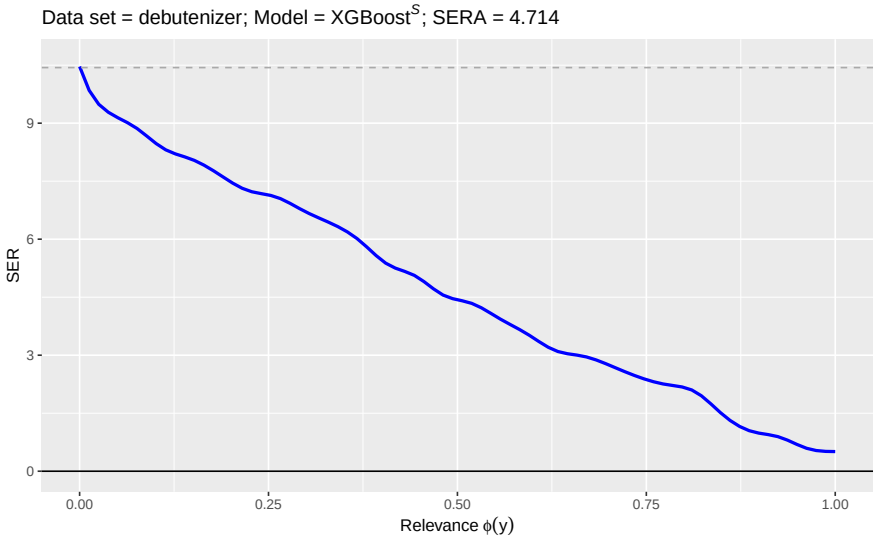
where $\mathcal{D}^t \subseteq \mathcal{D}$ defined as $\mathcal{D}^t = \{\langle \boldsymbol{x}_i, y_i \rangle \in \mathcal{D} \,|\, \phi(y_i) \geq t\}$

- Properties:
  - for any given $\delta \in \mathbb{R}^+$, s.t. $t + \delta \leq 1$: $SER_{t+\delta} \leq SER_t$
  - convex and differentiable
  - maximum: $SER_{t=0}$, minimum: $SER_{t=1}$

2-Suitable Performance Metrics: Regression (cont.)

- Squared Error Relevance Error Area (*SERA*)

$$SERA = \int_0^1 SER_t \, dt = \int_0^1 \sum_{i \in \mathcal{D}^t} (\hat{y}_i - y_i)^2 \, dt$$



Properties:

- Errors committed at relevant values $\phi(y) \approx 1$ are accounted more times than at non-relevant ones $\phi(y) \approx 0$
- Preserves (from SER_t)
  - convexity
  - differentiability

2-Suitable Performance Metrics: Regression (cont.)

The asymmetric loss function *SERA*

- reflects the notion of loss asymmetry in different ranges of the target variable
- focuses on the prediction of extreme values
- it is not threshold dependent
- penalizes severe model bias and low generalization capability

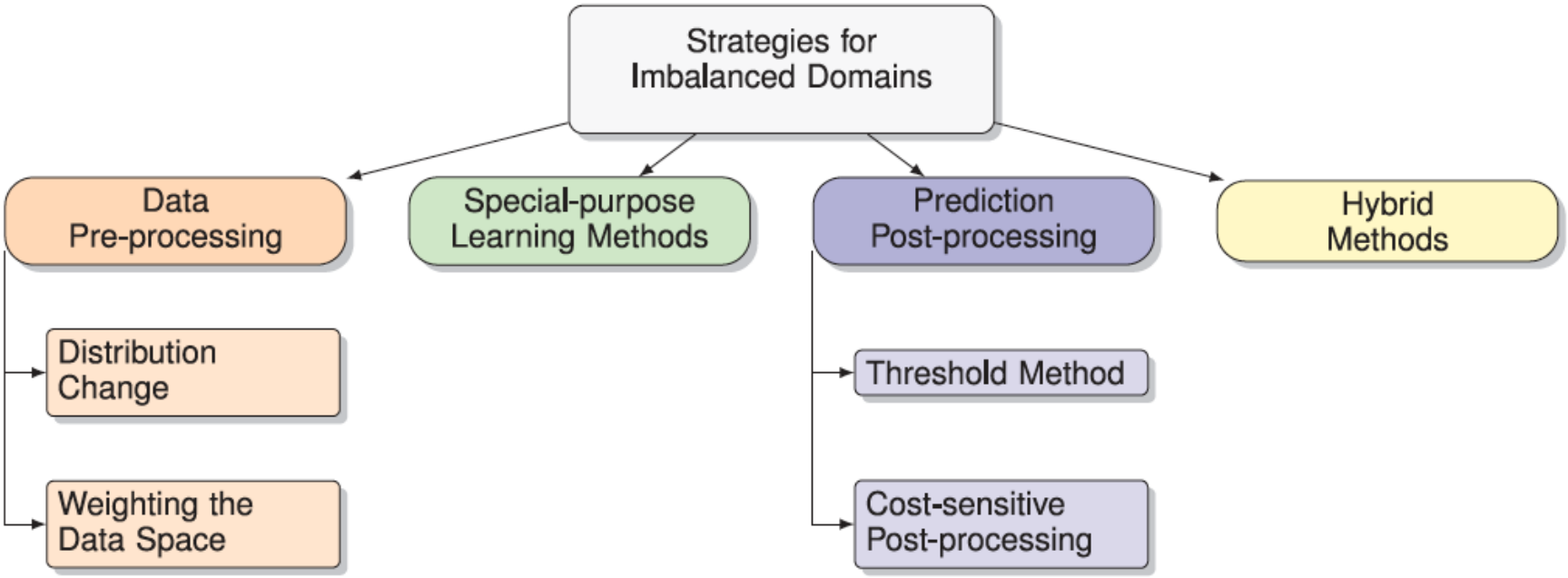
See [Ribeiro and Moniz, 2020, Silva et al., 2022] for more details

3-Learning Strategies for Imbalanced Domains

To prevent the models from being biased to the most frequent cases, it is necessary to use:

- performance metrics biased towards the performance of rare cases;
- learning strategies that focus on these rare cases.
  - Data pre-processing
  - Special-purpose Learning
  - Predictions post-processing

Learning Strategies for Imbalanced Domains



Data Pre-Processing Strategies

Proposal

Change the data distribution to make the standard algorithm focus on rare and relevant cases.

Advantages

- They allow the application of any learning algorithm
- The obtained model will be biased toward the goals of the user
- Models will be interpretable

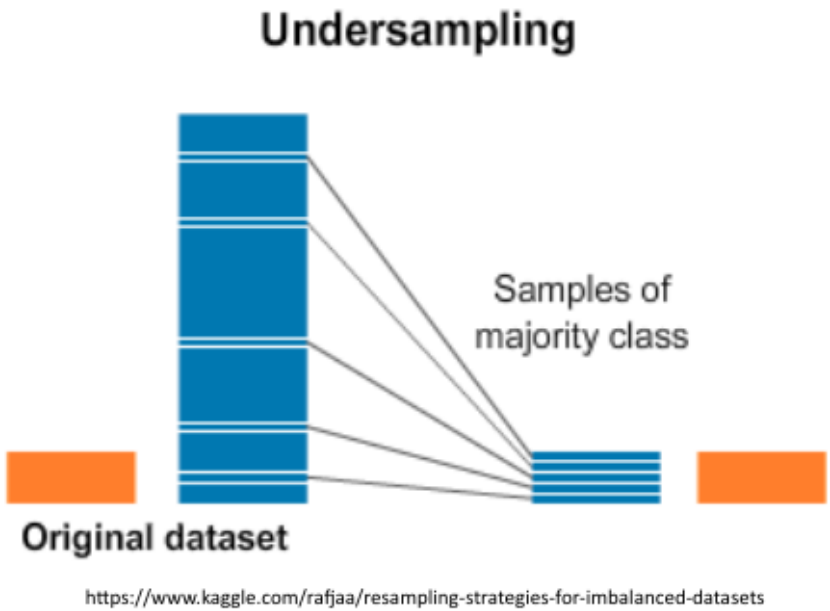
Data Pre-Processing Strategies (cont.)

Techniques

- Distribution Change
  - change the data distribution to address the issue of poor representativeness of the more relevant cases
- Weighting the data space
  - some algorithms allow different weights to be assigned to different data instances.

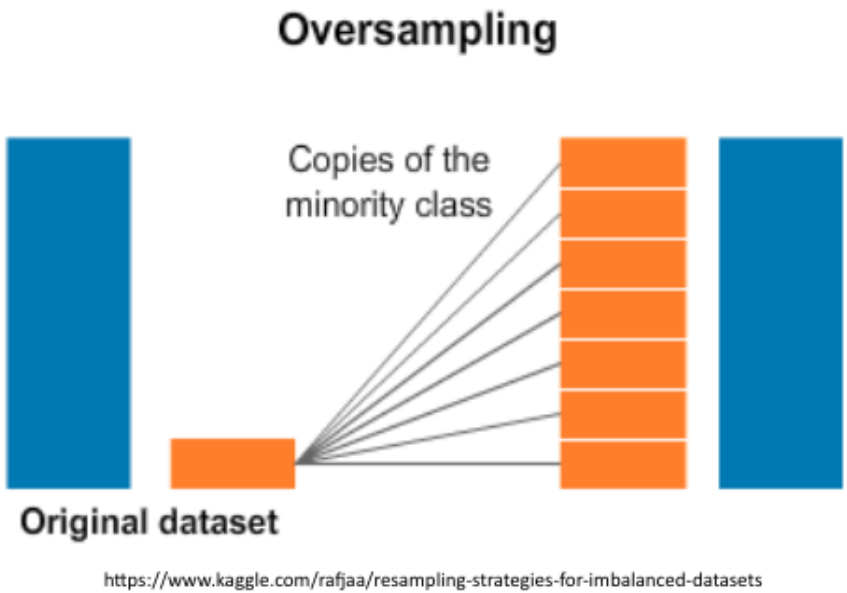
Data Pre-Processing Strategies (cont.)

- Random under-sampling
  - removes examples from the majority class or with common values from the original dataset, reducing the size of the dataset.
  - Problem: useful examples for the learning task may be discarded



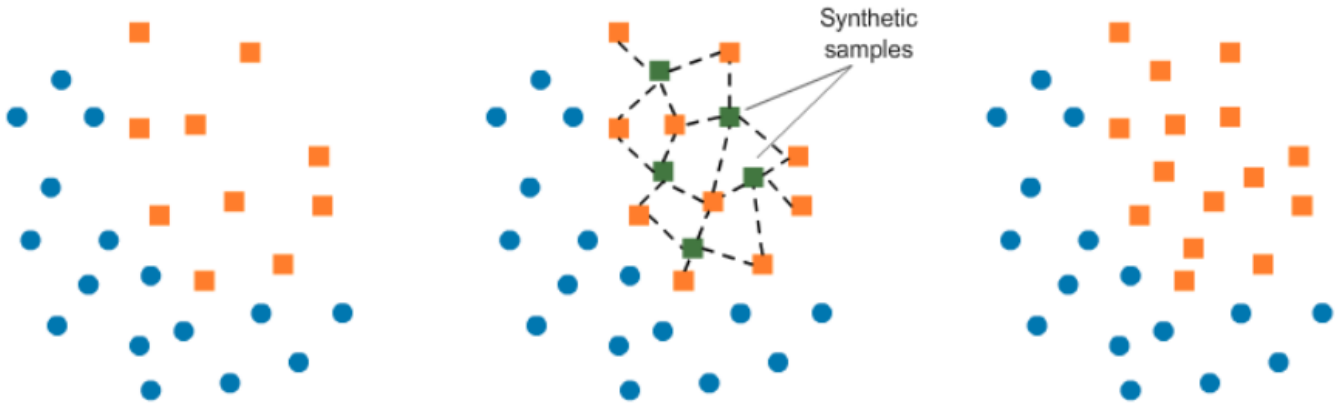
Data Pre-Processing Strategies (cont.)

- Random over-sampling
  - a random set of copies of minority class or rare values examples are added to the dataset.
  - Problem: possible over-fitting, i.e. poor generalization ability of the model



Data Pre-Processing Strategies (cont.)

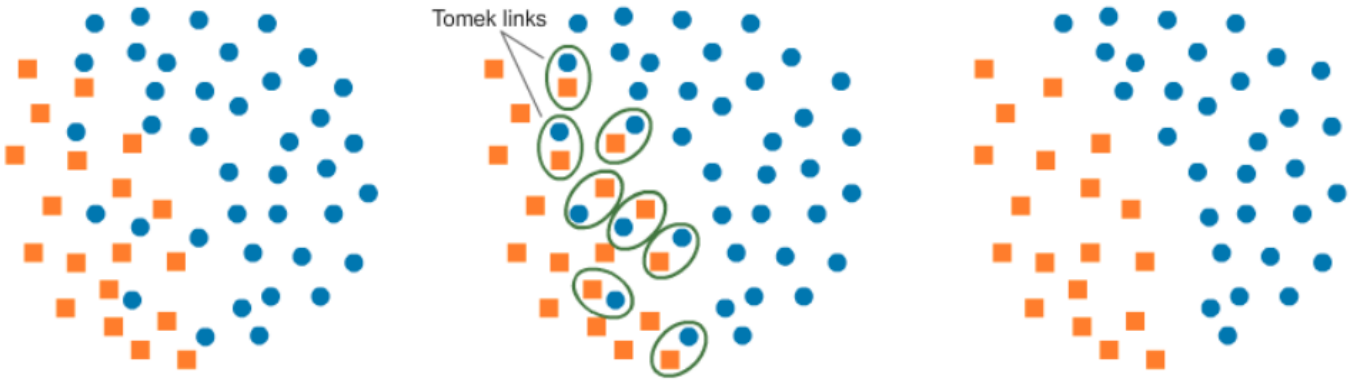
- SMOTE (Synthetic minority over-sampling technique) [Chawla et al., 2002],
  - over-samples the minority class examples by generating new synthetic data;
  - reduces the risks of under-sampling and over-sampling;
  - creates new examples by interpolating a seed minority example and one of its k minority class nearest neighbours



https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Data Pre-Processing Strategies (cont.)

- SMOTE can be combined with under-sampling of the majority class
  - random under-sampling
  - informed under-sampling (e.g. by identifying Tomek links)



https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Data Pre-Processing Strategies (cont.)

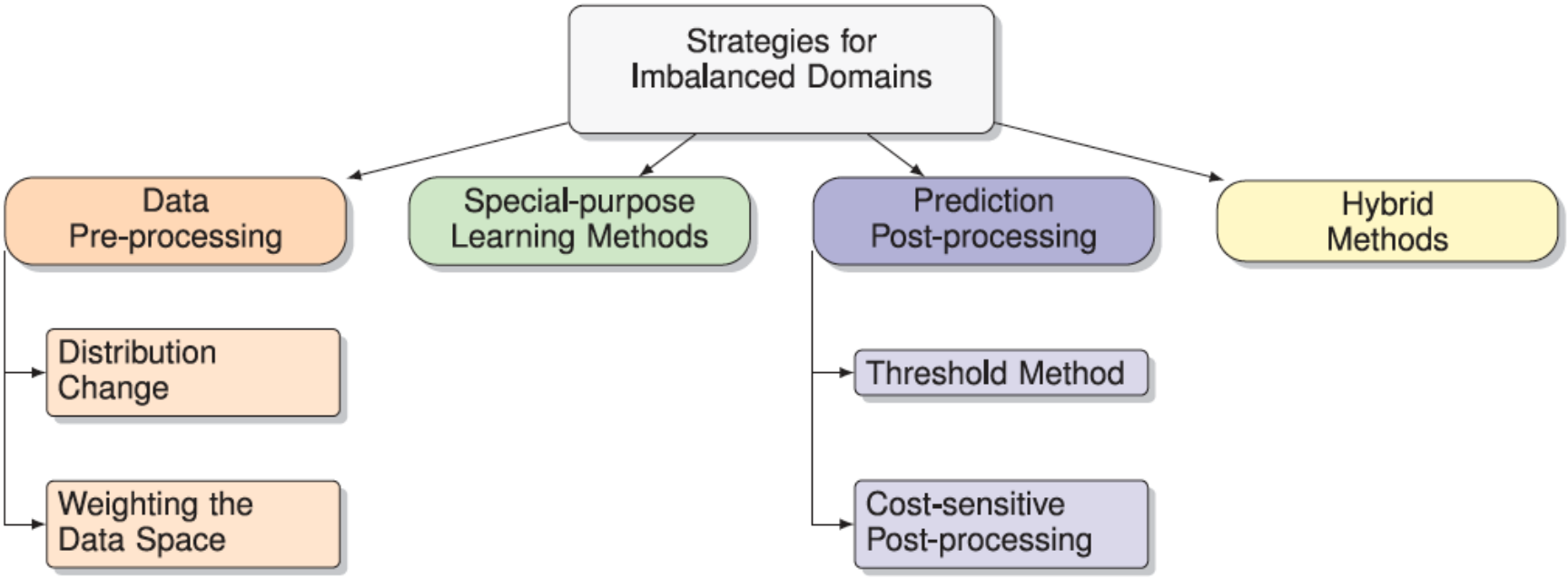
- SMOTE can be problematic depending on the distribution of minority examples (e.g. too far apart)
- Several SMOTE variants have been proposed that generate synthetic in harder-to-learn regions of the input space
  - put effort into the borders between classes (e.g. Borderline-SMOTE)
  - put effort into minority examples found in spaces dominated by the majority class (e.g. Adaysn)
  - there are many more variants
- SmoteR [Torgo et al., 2013] is a proposal for imbalanced regression scenarios

Data Pre-Processing Strategies (cont.)

Disadvantages

- difficulty of relating the modifications in the data distribution and the user preferences
- mapping the given data distribution into an optimal new distribution according to the user goals is not easy

Learning Strategies for Imbalanced Domains



Special-purpose Learning Strategies

Proposal

Change the learning algorithms so they can learn from imbalanced data.

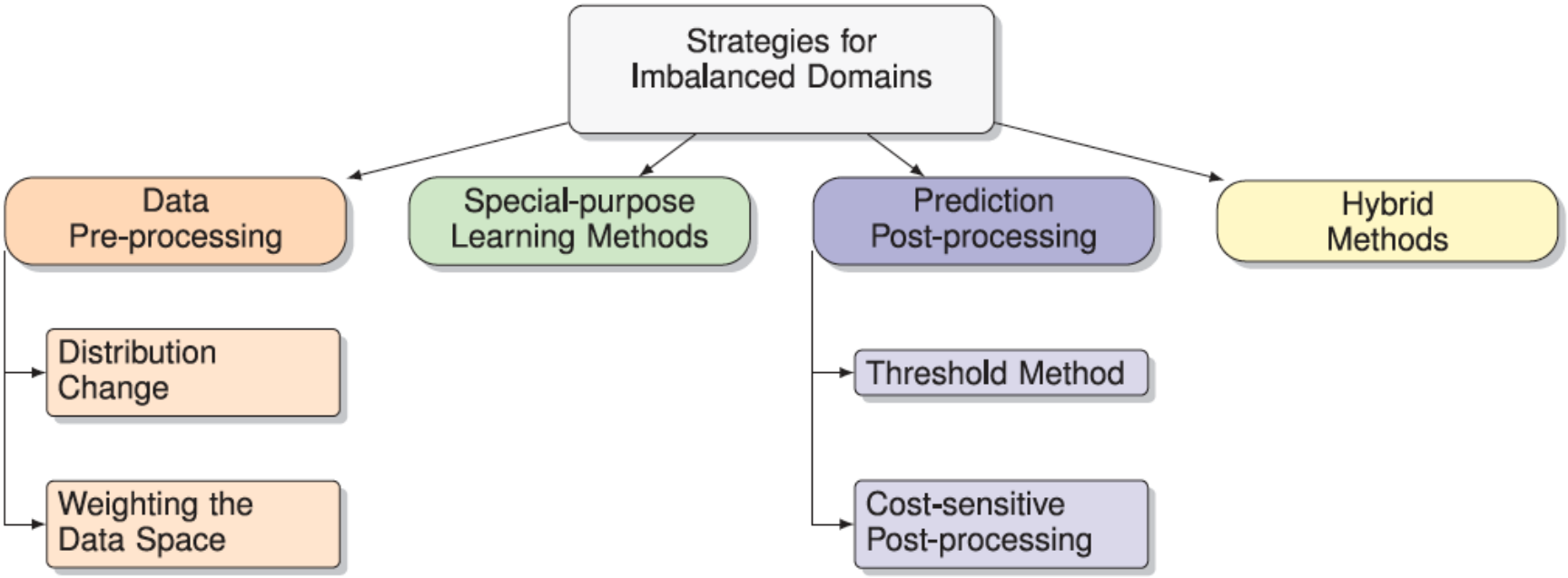
Advantages

- domain preferences incorporated as a preference criterion.
- models will be interpretable

Disadvantages

- restricted set of modified learning algorithms
- if the preference criterion changes, models have to be relearned and, possibly the algorithm has to be re-adapted
- mapping domain preferences with a suitable preference criterion is not straightforward

Learning Strategies for Imbalanced Domains



Prediction Post-processing Strategies

Proposal

Manipulate the predictions of the models according to the domain preferences (e.g. thresholding, cost-sensitive)

Advantages

- original dataset and a standard algorithm
- same model can be applied to different deployment scenarios without having to be relearned

Disadvantages

- the models do not reflect domain preferences
- models interpretability is jeopardized; they are obtained by optimizing a function that is not following the domain preference

Prediction Post-processing Strategies (cont.)

Techniques

- Threshold Method
  - obtain several models by varying the threshold on the score that expresses the degree to which an example is a member of a class (e.g. [Weiss, 2004])

- Cost-Sensitive Post-Processing
  - change the model predictions to make it cost-sensitive or to adapt it to a different operating context (e.g. [Hernández-Orallo, 2014])

Imbalanced Supervised Learning: Wrap-up

Proposal

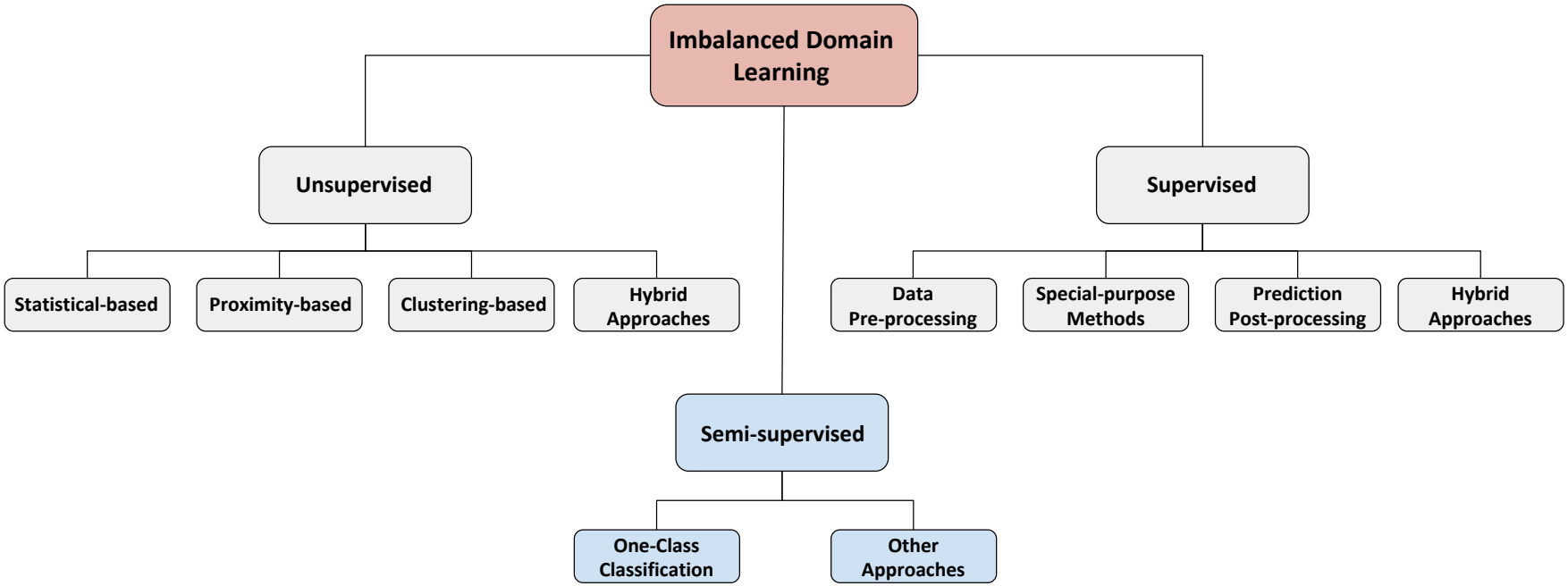
Build a prediction model for normal and rare classes (values) of the target variable.

Challenges

- Has to handle a training set with an imbalanced distribution.
- In classification relies on the availability of accurate labels for the training instances.
- In regression, it assumes that the distribution given in the training data is representative and, thus, is not expected to change in the test data.
- Cannot detect unknown or emerging outliers.

Next Subsection

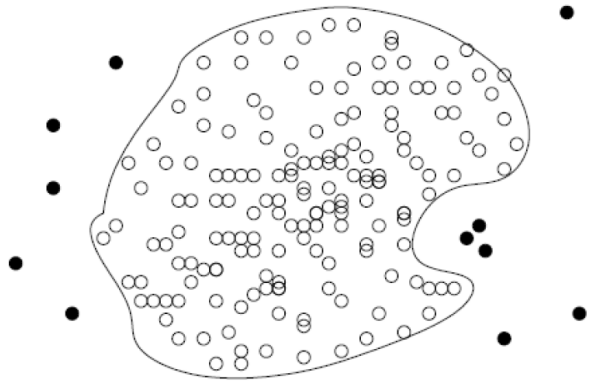
Taxonomy of Imbalanced Domain Learning Approaches



One Class Classification

Proposal

- Build a prediction model to the normal behaviour and classify any deviations from this behaviour as outliers.



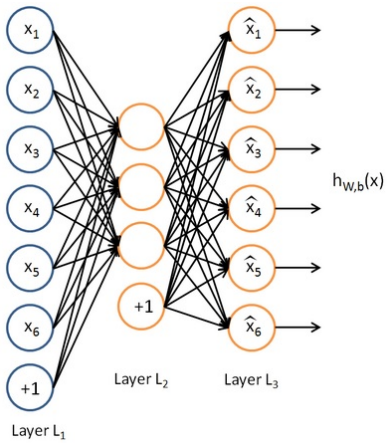
Advantages

- Models are interpretable.
- Normal behaviour can be accurately learned.
- Can detect new outliers that may not appear close to any outlier points in the training set.

One Class Classification: Techniques

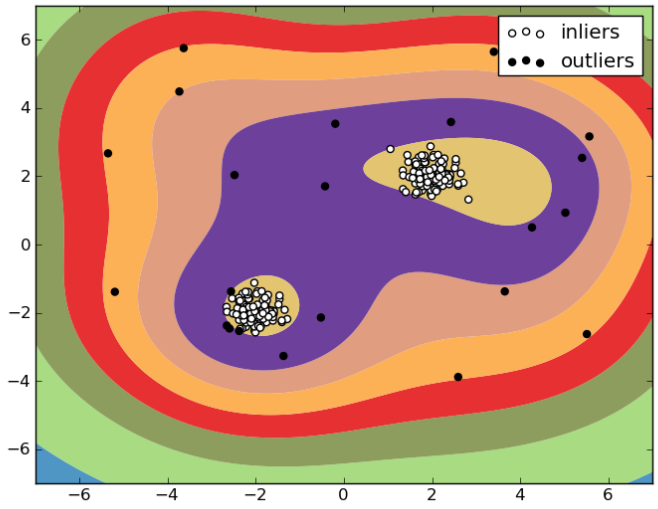
- Auto-associative neural networks / Autoencoders [Japkowicz et al., 1995]

- A feed-forward perceptron-based network is trained with normal data only.
- The network has the same number of input and output nodes and a decreased number of hidden nodes to induce a bottleneck.
- This bottleneck reduces the redundancies and focus on the key attributes of data.
- After training, the output nodes recreate the example given as input nodes.
- The network will successfully recreate normal data but will generate a high-recreation error for outlier data.



One Class Classification: Techniques (cont.)

- One-class SVM [Tax and Duin, 2004]



  - Obtain a spherical boundary in feature space around normal data.
  - The optimization problem consists of minimizing the volume of the hypersphere, so that includes all the training points.
  - Every point lying outside this hypersphere is an outlier.

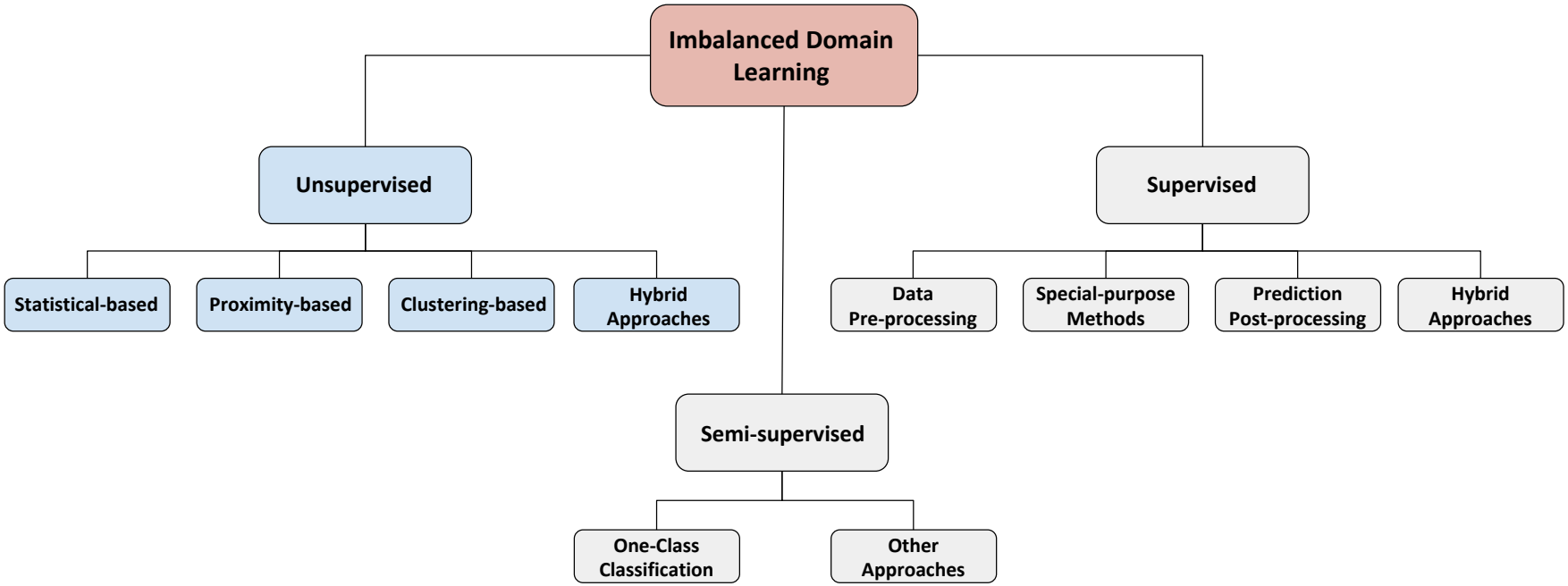
One Class Classification

Disadvantages

- Requires previous labeled instances for normal behaviour.
- Possible high false alarm rate - previously unseen normal data may be identified as an outlier.

Next Subsection

Taxonomy of Imbalanced Domain Learning Approaches



Statistical-based Approaches

Proposal

- All the points that satisfy a statistical discordance test for some statistical model are declared as outliers.

Advantages

- If the assumptions of the statistical model hold true, these techniques provide a justifiable solution for outlier detection.
- The outlier score is associated with a confidence interval.

Techniques

- Parametric
- Non-parametric

Statistical-based Approaches: Parametric Techniques

Assume one of the known probability distribution functions.

- *Grubbs' Test* (Grubbs, 1950)

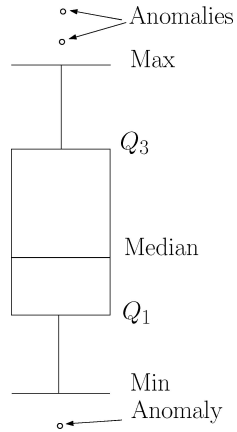
  A statistical test used to detect outliers in a **univariate** data set assumed to come from a normally distributed population.

- Boxplot (Tukey, 1977)

  It assumes a near-normal distribution of the values in a **univariate** data set, and identifies as outlier any value outside the interval

$$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$

where Q_1 (Q_3) is the 1st (3rd) quartile and IQR is the interquartile range.



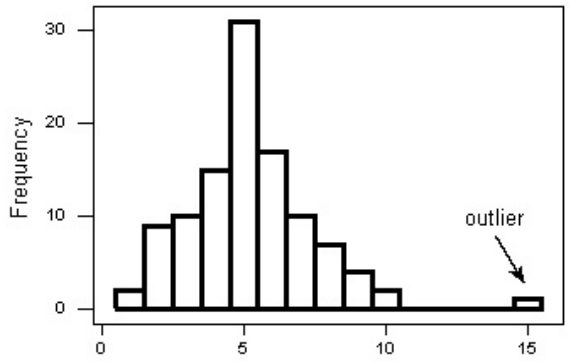
Statistical-based Approaches: Parametric Techniques (cont.)

- *Mahalanobis* distance (Mahalanobis, 1936)
  - It assumes a multivariate normal distribution of data.
  - Incorporates dependencies between attributes by the covariance matrix.
  - Transforms a **multivariate** outlier detection task into a univariate outlier detection problem.
  - All the points with a large *Mahalanobis* distance are indicated as outliers.
- Mixture of parametric distributions
- etc.

Statistical-based Approaches: Non-parametric Techniques

The probability distribution function is not assumed, but estimated from data.

- Histograms
  - Used for both univariate and multivariate data. For the later, the attribute-wise histograms are constructed and an aggregated score is obtained.
  - Hard to choose the appropriate bin size.



- Kernel functions
  - Adopt a kernel density estimation to estimate the probability density distribution of the data.
  - Ouliers are in regions of low density.

Statistical-based Approaches

Disadvantages

- The data does not always follows a statistical model.
- Choosing the best hypothesis test statistics is not straightforward.
- Capture interactions between attributes is not always possible.
- Estimating the parameters for some statistical models is hard.

Proximity-based Approaches

Proposal

- Normal instances occur in dense neighbourhoods, while outliers occur far from their closest neighbours.

Advantages

- Purely data driven technique
- Does not make any assumptions regarding the underlying distribution of data.

Some Techniques

- Distance-based
- Density-based

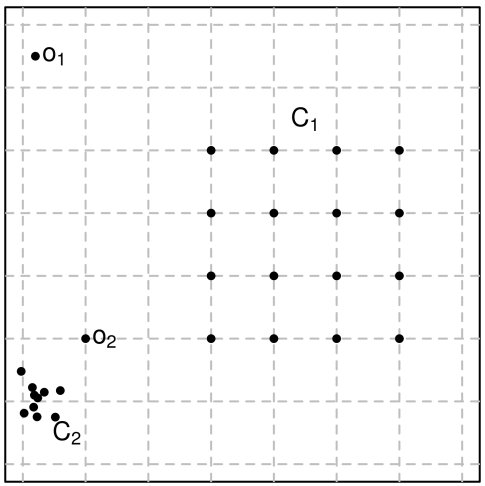
Proximity-based Approaches: Distance Techniques

A case c is an outlier if less than k cases are within a distance λ of c [Knorr and Ng, 1998]

- Outliers are points far away from other points, thus given a distance metric there should not be a lot of other points in their neighborhood.
- Define proper distance metric (e.g euclidean distance)
  - The notion of distance between cases with many variables may be distorted by different scales, different importance, different types (numerical, nominal)
- Define a "reasonable" neighborhood (λ)
- Define what is "a lot of other points" (k)

Proximity-based Approaches: Distance Techniques (cont.)

- Major cost: for each point is calculated its distance to all the other points.
- Optimization algorithms include index-based, cell-based approaches.
- The use of **global distance** measures poses difficulties in detecting outliers in data sets with different density regions.
- Example:



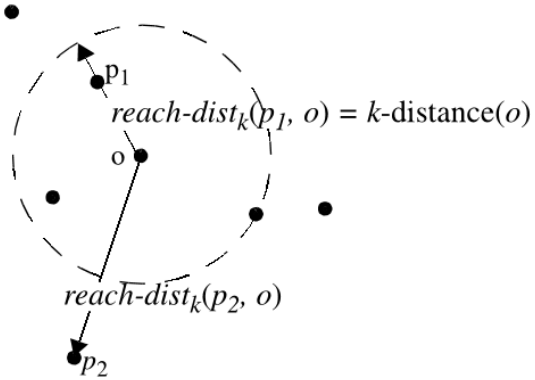
- o_1 and o_2 are outliers
- but, for the point o_2 to be identified as an outlier, all the points in C_1 would have to be identified as outliers too.

Proximity-based Approaches: Density Techniques

- Concept of outliers should be **locally** inspected.
- Compare points to their local neighborhood, instead of the global data distribution
- The density around an outlier is significantly different from the density around its neighbours.
- Use the relative density of a point against its neighbours as the indicator of the degree of the point being an outlier.
- Outliers are points in lower local density areas with respect to the density of its local neighbourhood.

Proximity-based Approaches: Density Techniques (cont.)

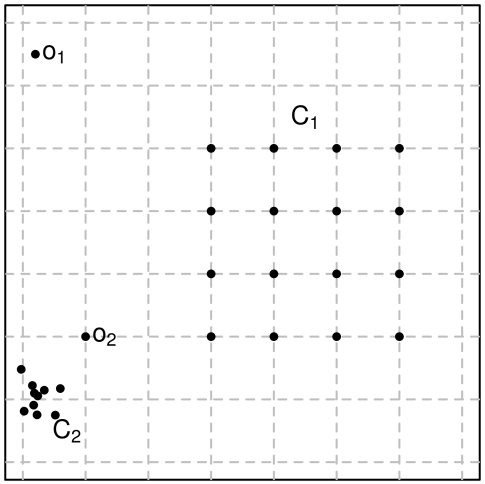
- LOF: Local Outlier Factor [Breunig et al., 2000]
  - *k-distance*: distance between p and its k-th nearest neighbour
  - *k-distance neighborhood*: all the points whose distance from p is not greater than the k-distance.
  - *reachability-distance* of p with respect to o: the maximum between their k-distance and their actual distance.



  - intuition: high values of reachability-distance between two given points indicates that they may not be in the same cluster

Proximity-based Approaches: Density Techniques (cont.)

- LOF: Local Outlier Factor [Breunig et al., 2000] (cont.)
  - *local reachability-density* of a point is inversely proportional to the average reachability-distance of its k neighbourhood.
  - *LOF* assigns high values to the points that have a much lower *local reachability-density* in comparison to its *k*-neighbourhood.
  - Example:



  - o_2 is assigned an higher LOF compared to the LOF values assigned to the points of C_1 and C_2
  - This captures a local outlier whose local density is relatively low comparing to the local densities of its *k*-neighbourhood..

Proximity-based Approaches: Density Techniques (cont.)

- Multi-granularity Deviation Factor [Papadimitriou et al., 2003]
  - finds not only outlier instances, but groups of outliers, i.e. micro-clusters
- RDF: Relative-Density Factor [Wang et al., 2004]
  - uses a vertical data structure (P-trees) to efficiently index data and prune the points which are deep in clusters, and then detects outliers only within the remaining small subset of the data

Proximity-based Approaches

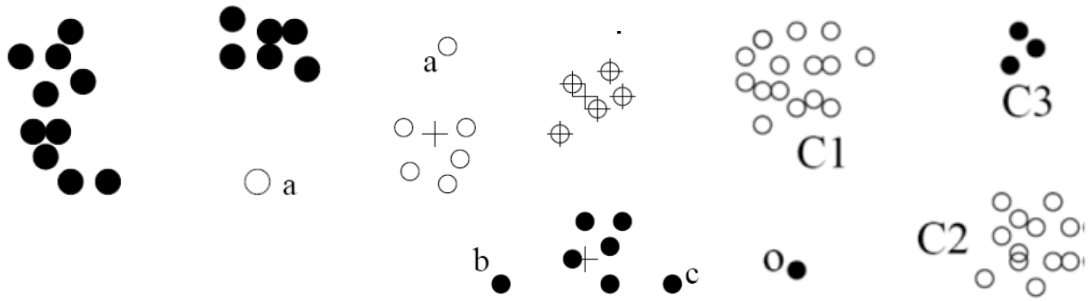
Disadvantages

- True outliers and noisy regions of low density may be hard to distinguish.
- These methods need to combine global and local analysis.
- In high dimensional data, the contrast in the distances is lost.
- Computational complexity of the test phase.

Clustering-based Approaches

Proposal

- Normal instances belong to large and dense clusters, while outlier instances are instances that:
  - do not belong to any of the clusters;
  - are far from its closest cluster;
  - form very small or low density clusters.



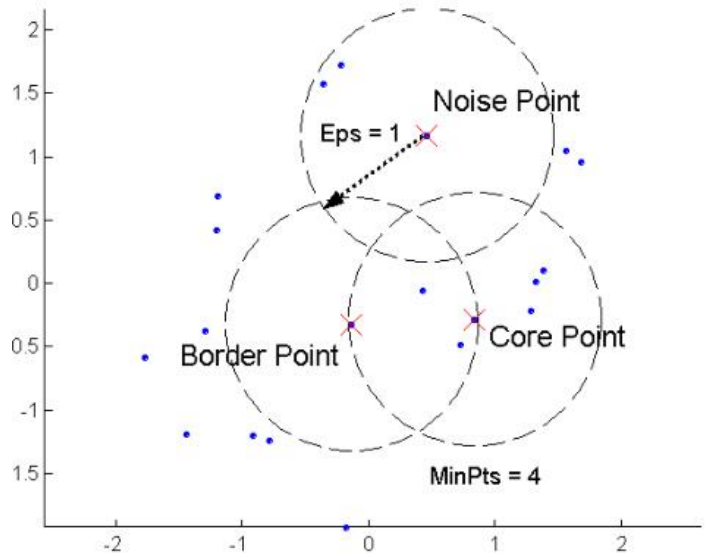
Advantages

- Easily adaptable to on-line/incremental mode.
- Test phase is fast.

Clustering-based Approaches: Techniques

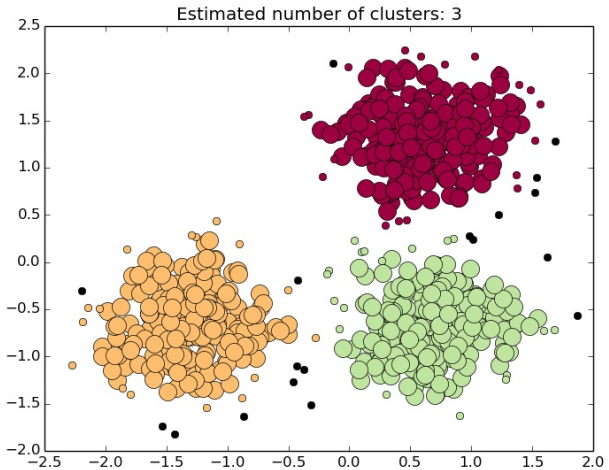
- DBSCAN [Ester et al., 1996]
  - Clustering method based on the notion of "density" of the points
  - The density of a point is estimated by the number of points that are within a certain radius.
  - Based on this idea, points can be classified as:

- *core points*: if the number of points within its radius are above a threshold
- *border points*: if the number of points within its radius are not above a threshold, but they are within a radius of a *core point*
- *noise points*: if do not have enough points within their radius, nor are sufficiently close to any *core point*.



Clustering-based Approaches: Techniques (cont.)

- DBSCAN [Ester et al., 1996] (cont.)
  - *noise points* are removed for the formation of clusters
  - all *core points* that are within a certain distance of each other are allocated to the same cluster
  - each *border point* is allocated to the cluster of the nearest *core points*
  - *noise points* are identified as outliers.

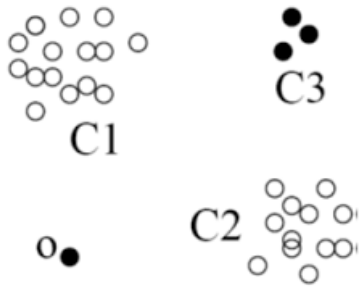


Clustering-based Approaches: Techniques (cont.)

- FindCBLOF [He et al., 2003]
  - Find clusters, and sort them in decreasing order
  - To each point, assign a *cluster-based local outlier factor* (CBLOF)
  - The CBLOF score of a point p is determined by the size of the cluster to which p belongs, and the distance between p and
    - its cluster centroid, if p belongs to a large cluster
    - its closest large cluster centroid, if p belongs to a small cluster.
  - the distance between the point and the cluster, can be the similarity measure used in the clustering algorithm.

Clustering-based Approaches: Techniques (cont.)

- FindCBLOF (cont.) [He et al., 2003]
  - Example:



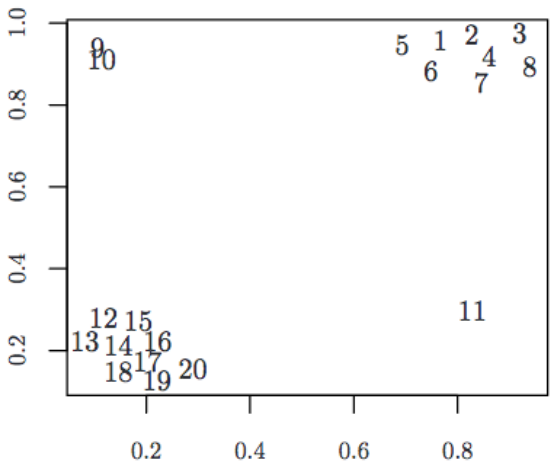
- o is outlier since its closest large cluster is C_1, and the similarity between o and C_1 is small
- for any point in C_3, its closest large cluster is C_2, but its similarity from C_2 is low, plus the size of C_3 is small

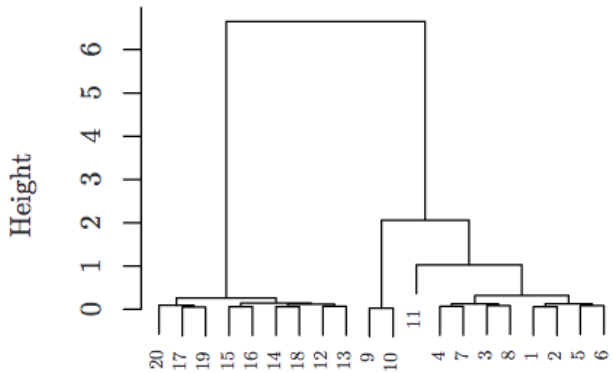
Clustering-based Approaches: Techniques (cont.)

- OR_H [Torgo, 2007]
  - Obtain an agglomerative hierarchical clustering of the data set
  - Use the information on the "path" of each point through the dendogram as a form to determine its degree of outlyingness
  - Cases that are only merged at later stages are surely very different from others

Clustering-based Approaches: Techniques (cont.)

- OR_H (cont.) [Torgo, 2007]
  - The outlier score of a point is given by the later stage of its agglomerative process and can be estimated by the size difference between the clusters being merged at that stage.
  - The higher the clusters size difference, the higher the outlier score.





Clustering-based Approaches

Disadvantages

- Computationally expensive in the training phase.
- If normal points do not create any clusters, this technique may fail.
- In high dimensional spaces, clustering algorithms may not give any meaningful clusters.
- Some techniques detect outliers as a byproduct, i.e. they are not optimized to find outliers, their main aim is to find clusters.

Next Subsection

Summary

- Outliers are not necessarily random noise.
- They can represent critical information that can trigger preventive or corrective actions.
- The interpretability of an outlier detection method is extremely important.
- The nature of the outlier detection problem is dependent on the application domain.
- Different approaches to this problem are necessary.
- Contextual and collective outliers are having increasing applicability in several real-world domains.
- Online Outlier Detection and Distributed Outlier Detection are emerging topics.
- There is much space for the development of new techniques in this area.

Open Challenges

Other Problem Settings

- Multiclass imbalance
- Regression
- Time series and data streams
- Spatiotemporal data streams
- Ordinal classification
- Multi-label classification
- Association rules mining
- Multi-instance learning
- Explainability
- etc.

Open Challenges (cont.)

Explainability with Imbalanced Domains

- Explainable AI and ML is a hot topic
- Many critical decisions are being taken based on the outcome of ML models
- This is even more critical with imbalanced domains
- Imbalanced domains have to do with rarity and high importance
- Frequently associated with rare and costly events
- Frequently used as early detection of these costly events
- Driving important (and frequently costly) decisions
- **Explaining WHY becomes even more important**

References

Aggarwal, C. (2013).

Outlier Analysis.

Springer New York.

Aggarwal, C. C. (2015).

Data Mining, The Texbook.

Springer.

Branco, P. (2014).

Re-sampling approaches for regression tasks under imbalanced domains.

Master's thesis, Dep. Computer Science, Faculty of Sciences - University of Porto.

Branco, P., Ribeiro, R. P., and Torgo, L. (2016a).

UBL: an r package for utility-based learning.

CoRR, abs/1604.08079.

Branco, P., Torgo, L., and Ribeiro, R. P. (2016b).

A survey of predictive modeling on imbalanced domains.

ACM Comput. Surv., 49(2):31:1–31:50.

References (cont.)

Breunig, M. M., Kriegel, H. P., Ng, R., and Sander, J. (2000).

Lof: Identifying density-based local outliers.

In *ACM SIGMOD 2000*. ACM Press.

Chandola, V., Banerjee, A., and Kumar, V. (2009).

Anomaly detection: A survey.

ACM Computing Surveys (CSUR), 41(3):15.

Chawla, N. V., Bowyer, K. W., Hall, O. L., , and Kegelmeyer, W. P. (2002).

Smote: Synthetic minority over-sampling technique.

Journal of Artificial Intelligence Research, 16:321–357.

AAAI Press.

Ester, M., peter Kriegel, H., S, J., and Xu, X. (1996).

A density-based algorithm for discovering clusters in large spatial databases with noise.

pages 226–231. AAAI Press.

Fan, W., Stolfo, S., Zhang, J., and Chan, P. K. (1999).

Adacost: Misclassification cost-sensitive boosting.

In *ICML'99*, pages 97–105. Morgan Kaufmann Publishers Inc.

References (cont.)

Han, J., Kamber, M., and Pei, J. (2011).

Data Mining: Concepts and Techniques.

Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Hawkins, D. M. (1980).

Identification of Outliers.

Chapman and Hall.

He, Z., Xu, X., and Deng, S. (2003).

Discovering cluster based local outliers.

Pattern Recognition Letters, 2003:9–10.

Hempstalk, K., Frank, E., and Witten, I. H. (2008).

One-class classification by combining density and class probability estimation.

In *ECML PKDD 2008 (1)*, pages 505–519.

Hernández-Orallo, J. (2014).

Probabilistic reframing for cost-sensitive regression.

ACM Trans. Knowl. Discov. Data, 8(4):17:1–17:55.

References (cont.)

Hodge, V. J. and Austin, J. (2004).

A survey of outlier detection methodologies.

Artificial Intelligence Review, 22:2004.

Japkowicz, N., Myers, C., and Gluck, M. A. (1995).

A novelty detection approach to classification.

In *IJCAI*, pages 518–523. Morgan Kaufmann.

Joshi, M. V., Agarwal, R. C., and Kumar, V. (2002).

Predicting rare classes: Comparing two-phase rule induction to cost-sensitive boosting.

In *PKDD 02*, volume 2431 of *LNCS*, pages 237–249. Springer.

Joshi, M. V., Kumar, V., and Agarwal, R. C. (2001).

Evaluating boosting algorithms to classify rare classes: Comparison and improvements.

In *ICDM 2001*, pages 257–264.

Knorr, E. M. and Ng, R. T. (1998).

Algorithms for mining distance-based outliers in large datasets.

In *VLDB'98*, pages 392–403. Morgan Kaufmann, San Francisco, CA.

References (cont.)

Kubat, M. and Matwin, S. (1997).

Addressing the curse of imbalanced training sets: one-sided selection.

In *ICML 97*, pages 179–186. Morgan Kaufmann.

Lazarevic, A. (2008).

Anomaly detection: A tutorial.

Tutorial Session on 2008 Siam Conference on Data Mining (SDM08).

Maloof, M. A. (2003).

Learning when data sets are imbalanced and when costs are unequal and unknown.

In *ICML-2003 workshop on learning from imbalanced data sets II*, volume 2, pages 2–1.

Papadimitriou, S., Kitagawa, H., Faloutsos, C., and Gibbons, P. B. (2003).

Loci: Fast outlier detection using the local correlation integral.

In *ICDE'03*, pages 315–326. IEEE Computer Society.

Ribeiro, R. P. (2011).

Utility-based Regression.

PhD thesis, Dep. Computer Science, Faculty of Sciences - University of Porto.

References (cont.)

Ribeiro, R. P. and Moniz, N. (2020).

Imbalanced regression and extreme value prediction.

Machine Learning, 109(9):1803–1835.

Silva, A., Ribeiro, R. P., and Moniz, N. (2022).

Model optimization in imbalanced regression.

In *DS'22: 25th Int. Conf. on Discovery Science*, pages to–appear. Springer.

Tax, D. (2001).

One-class classification: Concept learning in the absence of counter-examples.

PhD thesis, Technische Universiteit Delft.

Tax, D. M. J. and Duin, R. P. W. (2004).

Support vector data description.

Machine Learning, 54(1):45–66.

Torgo, L. (2007).

Resource-bounded fraud detection.

In *EPIA 2007, Workshops*, pages 449–460.

References (cont.)

Torgo, L. (2017).

Data Mining with R: Learning with Case Studies.

Chapman and Hall/CRC, 2nd edition.

Torgo, L., Branco, P., Moniz, N., Popelínský, L., Ribeiro, R. P., Japkowicz, N., and Matwin, S. (2020).

Learning with imbalanced domains and rare event detection.

Tutorial Session on ECML PKDD 2020.

Torgo, L. and Ribeiro, R. P. (2007).

Utility-based regression.

In *PKDD 2007*, pages 597–604. Springer Berlin Heidelberg.

Torgo, L. and Ribeiro, R. P. (2009).

Precision and recall in regression.

In *DS'09: 12th Int. Conf. on Discovery Science*, pages 332–346. Springer.

Torgo, L., Ribeiro, R. P., Pfahringer, B., and Branco, P. (2013).

Smote for regression.

In *Progress in Artificial Intelligence*, pages 378–389. Springer.

References (cont.)

Wang, B., Ren, D., and Perrizo, W. (2004).

Rdf: A density-based outlier detection method using vertical data representation.

In *ICDM 2004*, pages 503–506. IEEE.

Weiss, G. M. (2004).

Mining with rarity: a unifying framework.

SIGKDD Explorations Newsletter, 6(1):7–19.

Yu, D., Sheikholeslami, G., and Zhang, A. (2002).

Findout: Finding outliers in very large datasets.

Knowledge and Information Systems, 4(4):387–412.

Zhang, Y., Meratnia, N., and Havinga, P. (2007).

A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets.