Learning with Imbalanced Domains and Rare Event Detection

Luis Torgo, Stan Matwin, Nathalie Japkowicz, **Nuno Moniz**, **Paula Branco**, **Rita P. Ribeiro** and **Lubomir Popelinsky**

Dalhousie University, Canada
INESC TEC, University of Porto, Portugal
American University, USA
University of Ottawa, Canada
Masaryk University, Czech Republic

September, 2020

Outline

1. Welcome
2. Rare Event Detection - Principles
3. Methods and Evaluation
4. Class-based Outlier Detection
5. Explanation of rare events
6. Open Challenges

Rare Event Detection

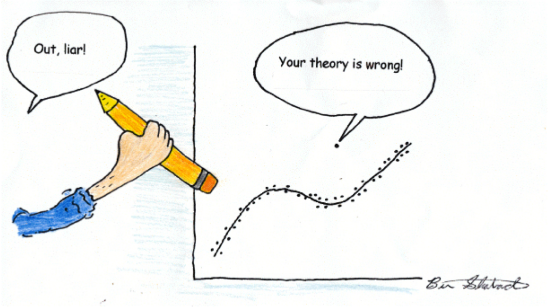
Principles

Motivation

- Most of machine learning 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 import insights about data.
- These patterns represent rare events, i.e. outliers.
- Depending on the goal, those insights can be even more interesting 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/noise and their identification had data cleaning purposes.
- However, they can represent a truthful deviation of data.
- For some applications, they represent critical information, which can trigger preventive or corrective actions.

Learning with Imbalanced Domains

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.

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

- Define 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 lack of known labelled outliers for training/validation of models.

Key Aspects of Imbalance Domain Learning

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

Nature of Input Data

Key Aspects of Imbalanced Domain Learning

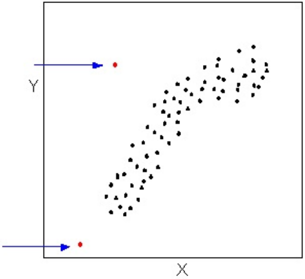
- Each data instance has:
  - One attribute (univariate)
  - Multiple attributes (multivariate)
- Relationship among data instances:
  - None
  - Sequential/Temporal
  - Spatial
  - Spatio-temporal
  - Graph
- Dimensionality of data

Types of Outliers

Key Aspects of Imbalanced Domain Learning

Point Outlier

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

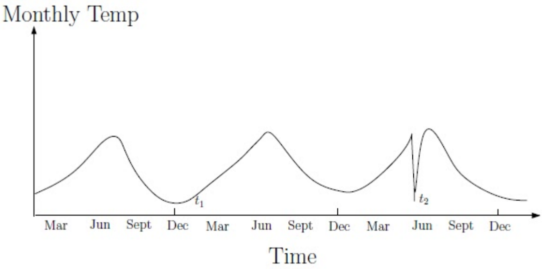


Types of Outliers (cont.)

Key Aspects of Imbalanced Domain Learning

Contextual Outlier

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

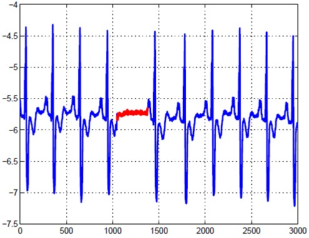


Types of Outliers (cont.)

Key Aspects of Imbalanced Domain Learning

Collective Outlier

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



Intended Output

Key Aspects of Imbalanced Domain Learning

Value

- A label / numeric value identifying normal or outlier instance.

Score

- The probability of being an outlier.
- This allows the output to be ranked.
- But, requires the specification of a threshold.

Learning Task

Key Aspects of Imbalanced Domain Learning

Unsupervised Learning

- 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.

Semi-supervised Learning

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

Supervised Learning

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

Performance Metrics

Key Aspects of Imbalanced Domain Learning

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

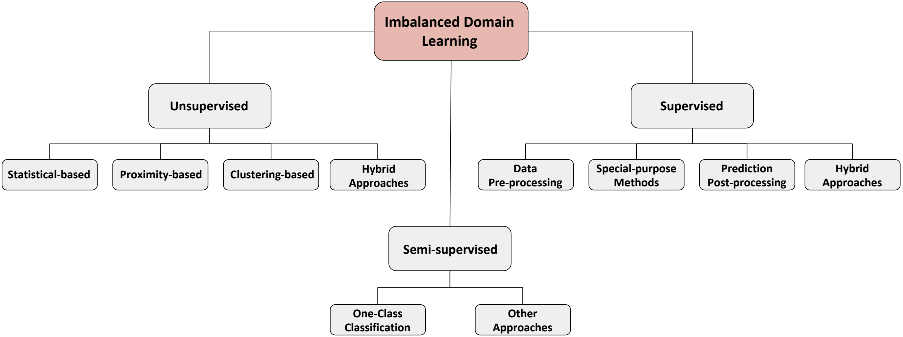
Credit Card Fraud Detection:

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

- Standard performance metrics are not suitable!

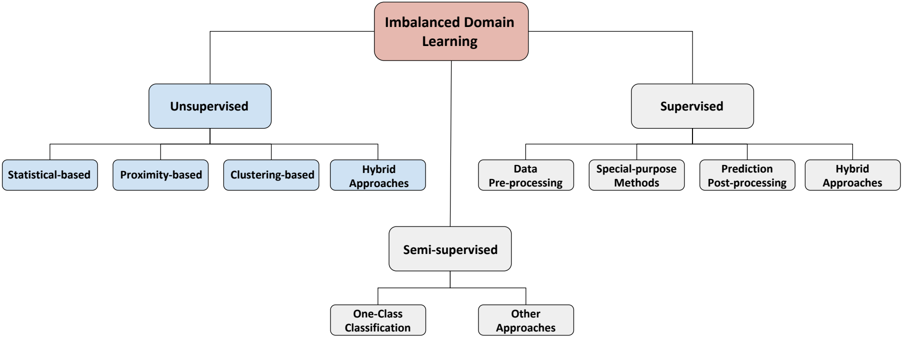
Taxonomy of Approaches

Imbalanced Domain Learning



Taxonomy of Approaches

Imbalanced Domain Learning



Statistical-based Approaches

Unsupervised Imbalanced Domain Learning

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, these techniques provide an acceptable solution for outlier detection.
- The outlier score is associated with a confidence interval.

Disadvantages

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

Proximity-based Approaches

Unsupervised Imbalanced Domain Learning

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.

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.
- Computationally expensive in the test phase.

Clustering-based Approaches

Unsupervised Imbalanced Domain Learning

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 or form very small or low-density clusters.

Advantages

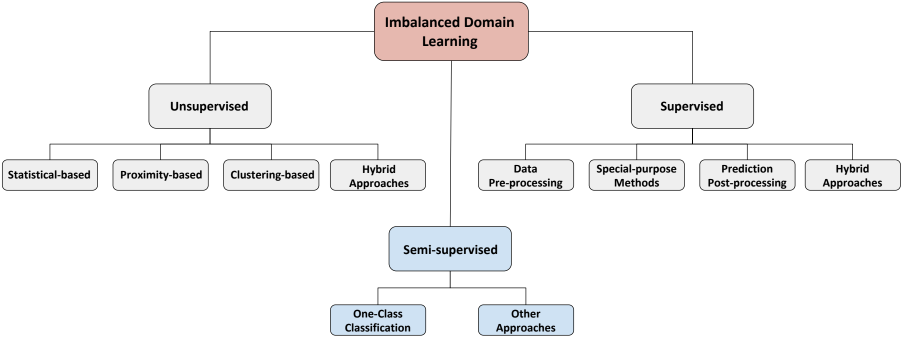
- Easily adaptable to on-line/incremental mode.
- Test phase is fast.

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.

Taxonomy of Approaches

Imbalanced Domain Learning



One Class Classification Approach

Semi-supervised Imbalanced Domain Learning

Proposal

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

Advantages

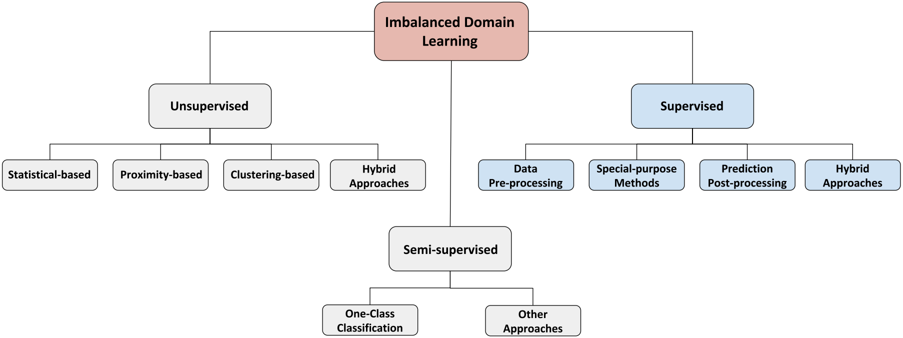
- Models are interpretable.
- Normal behaviour can be accurately learned.
- Can detect new outliers that may not appear close to an outlier point in the training set.

Disadvantages

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

Taxonomy of Approaches

Imbalanced Domain Learning



Predictive Modelling

Supervised Imbalanced Domain Learning

- In a supervised learning task the goal is:
  - given an unknown function $Y = f(X_1, X_2, \cdots, X_p)$,
  - use a training set $\mathcal{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$ with examples of this function
  - to obtain the best approximation to the function f, i.e. the model, $h(X_1, X_2, \cdots, X_p)$.

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

Imbalanced Predictive Modelling

- More importance is assigned to a subset of target variable Y domain.
- The cases that are more relevant are poorly represented in the training set.

- How to specify these non-uniform importance values?

Predictive Modelling: Notion of Relevance

Supervised Imbalanced Domain Learning

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

A relevance function $\phi(Y) : \mathcal{Y} \to [0, 1]$ 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 applicable to both classification and regression problems.
- It can be used to build the sets of rare and normal cases.

Torgo, L. and Ribeiro, R. (2007). “Utility-based Regression”. In: Proceedings of 11th ECML/PKDD 2007. Springer.

Predictive Modelling: Notion of Relevance (cont.)

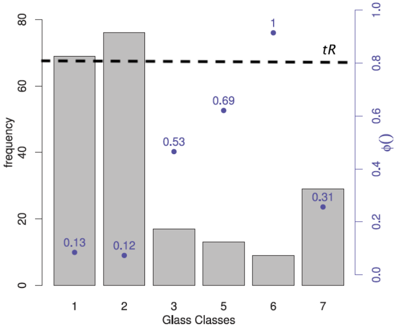
Supervised Imbalanced Domain Learning

- With an user-defined threshold on the relevance values t_R.
- Partition the training set $\mathcal{D}$ in two complementary subsets:
  - $\mathcal{D}_R = \{\langle \mathbf{x}, y \rangle \in \mathcal{D} : \phi(y) \geq t_R\}$
  - $\mathcal{D}_N = \mathcal{D} \setminus \mathcal{D}_\mathcal{R}$
- In this case, we have that $|\mathcal{D}_R| << |\mathcal{D}_N|$
- How to define the relevance function $\phi(Y)$?
  - It can be provided by the domain knowledge.
  - Estimated from the target variable data distribution, so that rare target classes/values are assigned more importance.

Predictive Modelling: Imbalanced Classification

Supervised Imbalanced Domain Learning

- In imbalanced classification specifying the relevance of a target variable for each class is feasible.
- The most important cases are the cases labelled with infrequent classes in the target variable Y, i.e. the cases for which $\phi(y) \geq t_R$.



Predictive Modelling: Imbalanced Regression

Supervised Imbalanced Domain Learning

- In imbalanced regression, given the potentially infinite nature of the target variable domain, specifying the relevance of all values is virtually impossible, requiring an approximation.

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

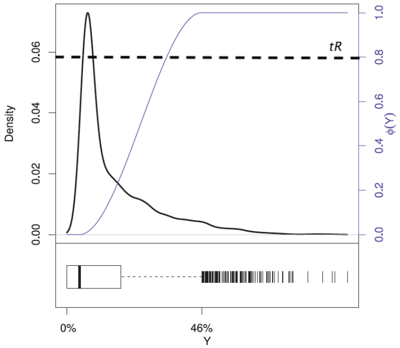
- interpolation method
  - user provides a set of interpolating points
- automatic method
  - no input required from the user;
  - it uses the target variable distribution;
  - it assumes that the most relevant cases are located at the extremes of the target variable distribution.

Ribeiro, Rita P. "Utility-based regression". PhD thesis, Dep. Computer Science, Faculty of Sciences, University of Porto, 2011.

Predictive Modelling: Imbalanced Regression (cont.)

Supervised Imbalanced Domain Learning

- The automatic method interpolates the boxplot statistics to obtain a continuous relevance function that maps the domain of the target variable Y to the relevance interval $[0, 1]$, so that the extreme values of Y are most important ones, i.e. the cases for which $\phi(y) \geq t_R$.



Predictive Modelling Challenges

Supervised Imbalanced Domain Learning

- It is of key importance that the obtained models are particularly accurate at the sub-range of the domain of the target variable for which training examples are rare.

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

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

Branco P, Torgo L, Ribeiro RP (2016). "A survey of predictive modeling on imbalanced domains". In: ACM Computing Surveys (CSUR) 49 (2), 1–35

Data Pre-Processing Strategies

Supervised Imbalanced Domain Learning

Proposal

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

Advantages

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

Disadvantages

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

Special-purpose Learning Strategies

Supervised Imbalanced Domain Learning

Proposal

- Change the learning algorithms so they can learn from imbalance data.

Advantages

- The domain goals are incorporated directly into the models by setting an appropriate preference criterion.
- Models will be interpretable.

Disadvantages

- It is restricted to that specific set of modified learning algorithms.
- It requires a deep knowledge of algorithms.
- If the preference criterion changes, models have to be relearned and, possibly the algorithm has to be re-adapted.
- It is not easy to map the domain preferences with a suitable preference criterion.

Prediction Post-processing Strategies

Proposal

- Use the original data set and a standard learning algorithm, only manipulating the predictions of the models according to the domain preferences and the imbalance of the data

Advantages

- It is not necessary to be aware of the domain preferences at learning time.
- The same model can be applied to different deployment scenarios without having to be relearned.
- Any standard learning algorithm can be used.

Disadvantages

- the models do not reflect the domain preferences.
- models interpretability is jeopardized as they were obtained by optimizing a function that does not follow the domain preference bias.

Up next ...

- Suitable Performance Metrics
- Specific Learning Methods

Methods and Evaluation

Imbalanced Domains and Rare Event Detection

Performance Evaluation

Why is performance evaluation a challenge?

- 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 for the user

An Example from Classification

Fraud Detection

- Two classes: Fraud and Normal
- Fraudulent cases are roughly 1% of the training sample
- A classifier that always predicts Normal would achieve on average 99% accuracy!
- This classifier is completely useless!
- Because frauds are very rare, failing them or correctly predicting them will have a minor impact on the accuracy (or error rate) metric.

An Example from Regression

Forecasting Stock Market Returns

- Very high or low returns (% variations of prices) are interesting
- Near-zero returns are very common but uninteresting for traders - unable to cover transaction costs
- Examples:
  - Forecasting a future return of 3% and then it happens -5% is a very bad error!
  - Forecasting a return of 3% and then it happens 11% has the same error amplitude but it is not a serious error
  - Forecasting 0.2% for a true value of 0.4% is reasonably accurate but irrelevant!
  - Forecasting -7.5% for a true value of -8% is a good an useful prediction
- Because near 0 returns are very common a model that always forecasts 0 is hard to beat in terms of Mean Squared Error. But this model is useless!

Metrics and the Available Information

- Different applications may involve different type of information on the user preferences
- This may have an impact on the metrics you can and/or should calculate
- Independently, there are two classes of metrics: scalar and graphical

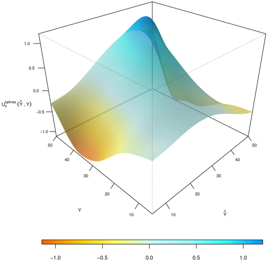
Evaluation with Full Utility Information

Utility Matrices

Table where each entry specifies the cost (negative benefit) or benefit of each type of prediction

		Pred.		
		c_1	c_2	c_3
Obs.	c_1	$B_{1,1}$	$C_{1,2}$	$C_{1,3}$
	c_2	$C_{2,1}$	$B_{2,2}$	$C_{2,3}$
	c_3	$C_{3,1}$	$C_{3,2}$	$B_{3,3}$

- Models are then evaluated by the total utility of their predictions, i.e. the sum of the benefits minus the costs.
- Similar setting for regression using Utility Surfaces (Ribeiro, 2011)



R. Ribeiro (2011). “Utility-based Regression”. PhD on Computer Science, Univ. Porto.

The Precision/Recall Framework

Classification

- Problems with two classes
- One of the classes is much less frequent and it is also the most relevant

		Preds.	
		Pos	Neg
Obs.	Pos	True Positives (TP)	False Negatives (FN))
	Neg	False Positives (FP)	True Negatives (TN)

The Precision/Recall Framework

Classification - 2

		Preds.	
		P	N
Obs.	P	TP	FN
	N	FP	TN

- *Precision* - proportion of the signals (events) of the model that are correct

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

- *Recall* - proportion of the real events that are captured by the model

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

The F-Measure

Combining Precision and Recall into a single measure

- Useful to have a single measure - e.g. optimization within a search procedure
- Maximizing one of them is easy at the cost of the other (it is easy to have 100% recall - always predict "P").
- What is difficult is to have both of them with high values

The F-Measure

Combining Precision and Recall into a single measure

- Useful to have a single measure - e.g. optimization within a search procedure
- Maximizing one of them is easy at the cost of the other (it is easy to have 100% recall - always predict "P").
- What is difficult is to have both of them with high values
- The F-measure is a statistic that is based on the values of precision and recall and allows establishing a trade-off between the two using a user-defined parameter (β),

$$F_\beta = \frac{(\beta^2 + 1) \cdot Prec \cdot Rec}{\beta^2 \cdot Prec + Rec}$$

where β controls the relative importance of *Prec* and *Rec*. If $\beta = 1$ then F is the harmonic mean between *Prec* and *Rec*; When $\beta \rightarrow 0$ the weight of *Rec* decreases. When $\beta \rightarrow \infty$ the weight of *Prec* decreases.

The G-Mean and Adjusted G-Mean

$$Gm = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} = \sqrt{sensitivity \times specificity}$$

$$AGm = \begin{cases} \frac{Gm + Specificity \times N_n}{1+N_n} & sensitivity \geq 0 \\ 0 & sensitivity = 0 \end{cases}$$

where N_n is the proportion of majority class examples in the data set.

M. Kubat and S. Matwin. "Addressing the curse of imbalanced training sets: one-sided selection." In Proc. of 14th Int. Conf. on Machine Learning, 1997, Nashville, USA, pp.179-186

R. Batuwita and V. Palade. "A new performance measure for class imbalance learning. Application to bioinformatics problems." In ICMLA'09, pp.545–550. IEEE, 2009.

Metrics for Multiclass Imbalance Problems

- $\phi(i)$ is the relevance of class i.
- Different ways to obtain $\phi()$ depending on the available domain information (Branco, 2017).

$$Rec^{\phi} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \phi(i) \cdot recall_i \qquad Prec^{\phi} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \phi(i) \cdot precision_i$$

$$F_{\beta}^{\phi} = \frac{(1+\beta^2) \cdot Prec^{\phi} \cdot Rec^{\phi}}{(\beta^2 \cdot Prec^{\phi}) + Rec^{\phi}} \qquad AvF_{\beta}^{\phi} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \frac{\phi(i) \cdot (1+\beta^2) \cdot precision_i \cdot recall_i}{(\beta^2 \cdot precision_i) + recall_i}$$

$$CBA^{\phi} = \sum_{i=1}^{C} \phi(i) \cdot \frac{mat_{i,i}}{max\left(\sum_{j=1}^{C} mat_{i,j}, \sum_{j=1}^{C} mat_{j,i}\right)}$$

P. Branco, L. Torgo, and R. Ribeiro. "Relevance-based evaluation metrics for multi-class imbalanced domains." PAKDD. Springer, Cham, pp.698-710 (2017).

The Precision/Recall Framework

Regression

For forecasting rare extreme values, the concepts of Precision and Recall were also adapted to regression (Torgo and Ribeiro, 2009; Branco, 2014),

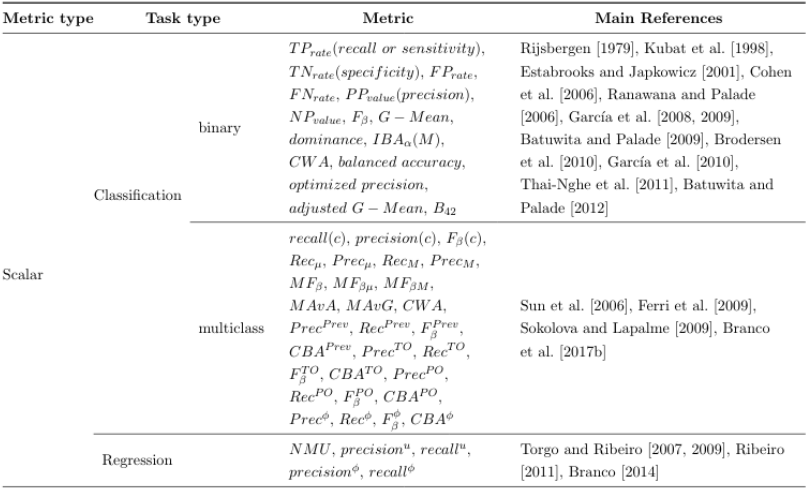
$$prec^{\phi} = \frac{\sum_{\phi(\hat{y}_i) > t_R} (1 + U(\hat{y}_i, y_i))}{\sum_{\phi(\hat{y}_i) > t_R} (1 + \phi(\hat{y}_i))}$$

$$rec^{\phi} = \frac{\sum_{\phi(y_i) > t_R} (1 + U(\hat{y}_i, y_i))}{\sum_{\phi(y_i) > t_R} (1 + \phi(y_i))}$$

L. Torgo and R. P. Ribeiro (2009). "Precision and Recall for Regression". In: Discovery Science'2009. Springer.

P. Branco (2014). "Re-sampling Approaches for Regression Tasks under Imbalanced Domains". MSc on Computer Science, Univ. Porto.

Summary of Scalar Metrics for Imbalanced Domains



Metric type	Task type		Metric	Main References
Scalar	Classification	binary	TP_{rate}(*recall or sensitivity*), TN_{rate}(*specificity*), FP_{rate}, FN_{rate}, PP_{value}(*precision*), NP_{value}, F_β, $G-Mean$, *dominance*, $IBA_\alpha(M)$, CWA, *balanced accuracy*, *optimized precision*, *adjusted* $G-Mean$, B_{42}	Rijsbergen [1979], Kubat et al. [1998], Estabrooks and Japkowicz [2001], Cohen et al. [2006], Ranawana and Palade [2006], García et al. [2008, 2009], Batuwita and Palade [2009], Brodersen et al. [2010], García et al. [2010], Thai-Nghe et al. [2011], Batuwita and Palade [2012]
		multiclass	$recall(c)$, $precision(c)$, $F_\beta(c)$, Rec_μ, $Prec_\mu$, Rec_M, $Prec_M$, MF_β, $MF_{\beta\mu}$, $MF_{\beta M}$, $MAvA$, $MAvG$, CWA, $Prec^{Prev}$, Rec^{Prev}, F_β^{Prev}, CBA^{Prev}, $Prec^{TO}$, Rec^{TO}, F_β^{TO}, CBA^{TO}, $Prec^{PO}$, Rec^{PO}, F_β^{PO}, CBA^{PO}, $Prec^\phi$, Rec^ϕ, F_β^ϕ, CBA^ϕ	Sun et al. [2006], Ferri et al. [2009], Sokolova and Lapalme [2009], Branco et al. [2017b]
	Regression		NMU, $precision^u$, $recall^u$, $precision^\phi$, $recall^\phi$	Torgo and Ribeiro [2007, 2009], Ribeiro [2011], Branco [2014]

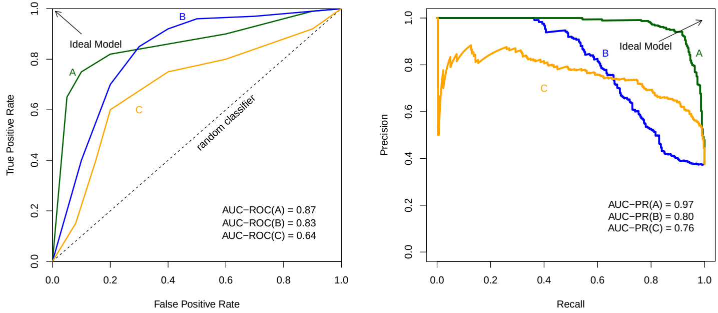
Adapted from:

P. Branco, L. Torgo and R. Ribeiro. “A Survey of Predictive Modeling on Imbalanced Domains”. In: ACM Comput. Surv. 49-2, 1–31 (2016).

P. Branco (2018). "Utility-based Predictive Analytics". PhD on Computer Science, Univ. Porto.

ROC curve and Precision-Recall Curve

Classification

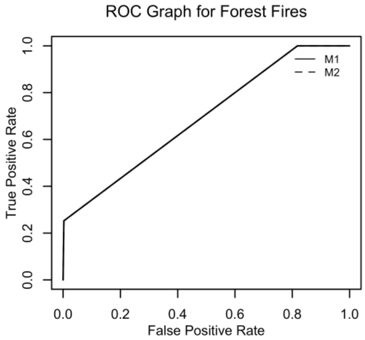


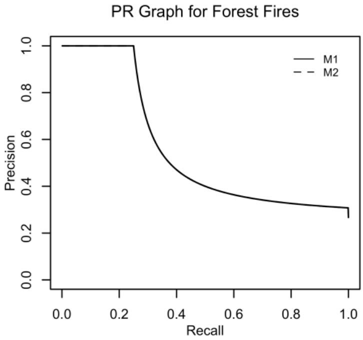
Taken from:

P. Branco (2018). "Utility-based Predictive Analytics". PhD on Computer Science, Univ. Porto.

ROC curve and Precision-Recall Curve

Regression

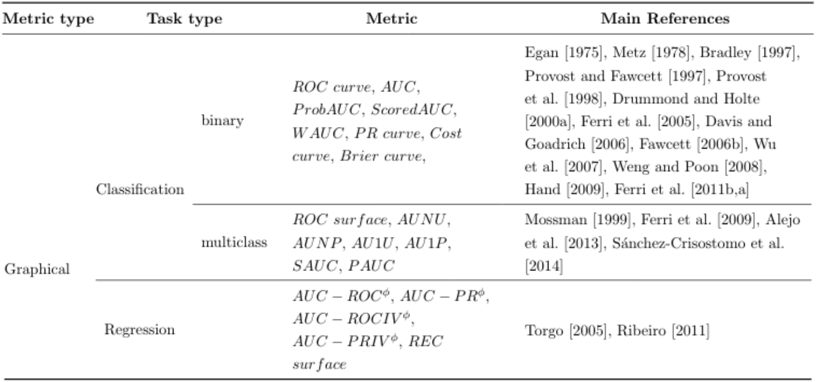




Taken from:

R. Ribeiro (2011). “Utility-based Regression”. PhD on Computer Science, Univ. Porto.

Summary of Graphical Metrics for Imbalanced Domains



Metric type	Task type		Metric	Main References
Graphical	Classification	binary	$ROC\ curve$, AUC, $ProbAUC$, $ScoredAUC$, $WAUC$, $PR\ curve$, $Cost\ curve$, $Brier\ curve$,	Egan [1975], Metz [1978], Bradley [1997], Provost and Fawcett [1997], Provost et al. [1998], Drummond and Holte [2000a], Ferri et al. [2005], Davis and Goadrich [2006], Fawcett [2006b], Wu et al. [2007], Weng and Poon [2008], Hand [2009], Ferri et al. [2011b,a]
		multiclass	$ROC\ surface$, $AUNU$, $AUNP$, $AU1U$, $AU1P$, $SAUC$, $PAUC$	Mossman [1999], Ferri et al. [2009], Alejo et al. [2013], Sánchez-Crisostomo et al. [2014]
	Regression		$AUC-ROC^{\phi}$, $AUC-PR^{\phi}$, $AUC-ROCIV^{\phi}$, $AUC-PRIV^{\phi}$, $REC\ surface$	Torgo [2005], Ribeiro [2011]

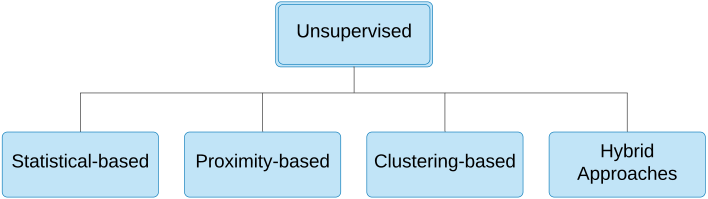
Adapted from:
P. Branco, L. Torgo and R. Ribeiro. “A Survey of Predictive Modeling on Imbalanced Domains”. In: ACM Comput. Surv. 49-2, 1–31 (2016).

P. Branco (2018). "Utility-based Predictive Analytics". PhD on Computer Science, Univ. Porto.

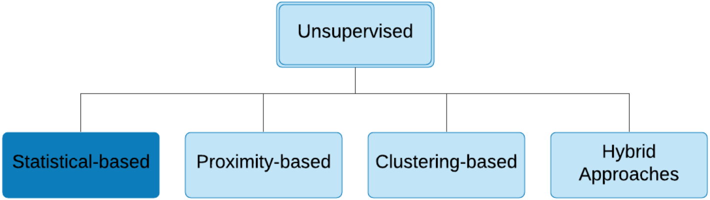
Imbalanced Domains and Rare Event Detection

Unsupervised Methods

Unsupervised Methods



Unsupervised Methods



Statistical-based Methods

Parametric

Assumption

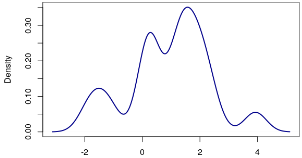
Normal instances occur in high probability regions of a stochastic model. Anomalies occur in the low probability regions of the stochastic model.

- Gaussian Model Based
- Regression Model Based
- Mixture of Parametric Distributions Based

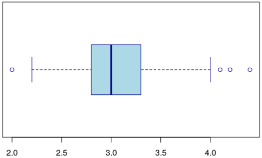
Statistical-based Methods

Parametric

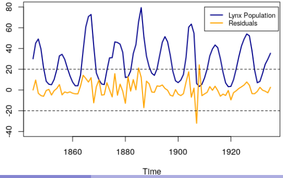
- Grubb's test
  $z = \frac{|x-\mu|}{\sigma}$



- Box Plot Rule



- Regression model based
  - fit a regression model
  - use the residuals to determine the anomaly score



Statistical-based Methods

Non-parametric

Assumption

The model structure is not determined a priori but is determined from the given data. Few assumptions regarding the data when compared to parametric techniques.

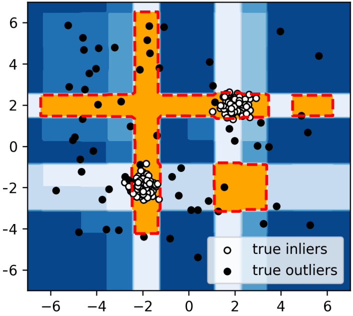
- Histogram Based
- Kernel Function Based

Statistical-based Methods

Non-parametric

Histogram Based

- build histogram
- for a new test instance, check if it falls in a bin of the histogram. If it does: normal, otherwise: anomaly.
- Variant: assign an anomaly score based on the bin frequency



Statistical-based Methods

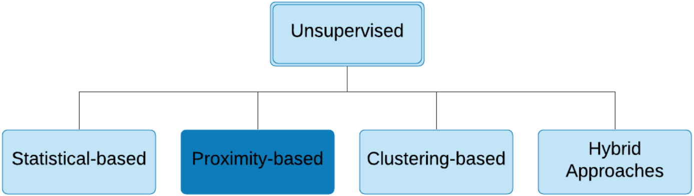
Non-parametric

Kernel Function Based

- Non-parametric techniques for probability density estimation
- Example: parzen windows estimation (Parzen, 1962)
- Use kernel functions to approximate the actual density.
- Similar to parametric methods. Difference: the density estimation technique used

Parzen, E. (1962) On the estimation of a probability density function and mode. Annals of Mathematical Statistics 33, 1065–1076.

Unsupervised Methods



Proximity-based Methods

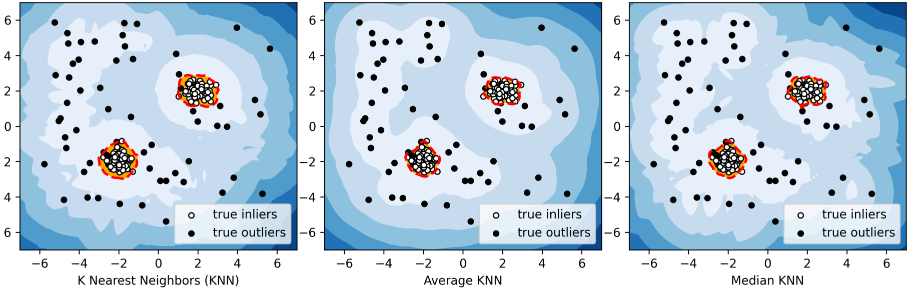
Distance-based: Nearest Neighbors (NN) Approach

- The anomaly score of a case is its distance to the k^{th} nearest neighbor
- Apply a threshold on the anomaly score to determine is a case is anomalous or not.
- Examples of applications: land mines detection from satellite ground images, detect anomalies in large synchronous turbine-generators

Ramaswamy, S.,Rastogi, R. and Shim., K. "Efficient algorithms for mining outliers from large data sets." Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000.

Proximity-based Methods

Distance-based: Nearest Neighbors (NN) Approach



Proximity-based Methods

Distance-based: Nearest Neighbors (NN) Approach

- Alternative way for computing the anomaly score: count the number of nearest neighbors that are not more than *d* distance apart from the case.
- Can be viewed as a way to obtain an estimate of the global density for each case.

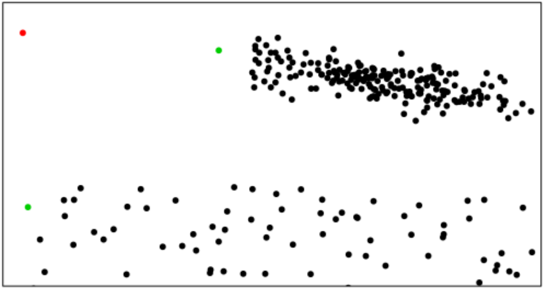
Knorr, E. M., and Ng, R. T. "Algorithms for mining distance based outliers in large datasets." Proceedings of the international conference on very large data bases. 1998.

Proximity-based Methods

LOF-based

Local Outlier Factor (LOF) (Breunig et al., 2000)

Each point has a score that captures the relative degree of isolation of the point from its surrounding neighbourhood.



Breunig, M. M., Kriegel, H. P., Ng, R., and Sander, J. (2000). "LOF: Identifying density-based local outliers." In Chen, W., Naughton, J. F., and Bernstein, P. A., editors, Proceedings of ACM SIGMOD 2000 International Conference on Management of Data. ACM Press.

Proximity-based Methods

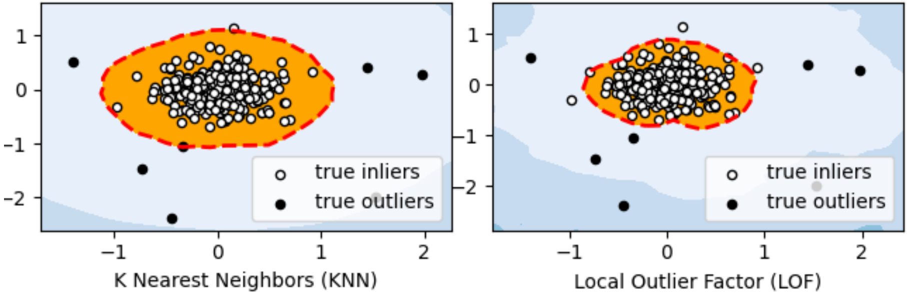
LOF-based

LOF Approach

- MinPts: number of nearest neighbors used in defining the local neighborhood
- For each point x compute distance to the k^{th} nearest neighbor $(k - dist)$
- Compute reachability distance: $reach - dist_k(x, p) = max\{k - dist(p), d(x, p)\}$
- Compute local reachability density: $lrd_{MinPts}(x) = \frac{MinPts}{\sum_p reach - dist_{MinPts}(x,p)}$
- Compute LOF score: $LOF_{MinPts}(x) = \frac{1}{MinPts} \cdot \sum_p \frac{lrd_{MinPts}(p)}{lrd_{MinPts}(x)}$

Proximity-based Methods

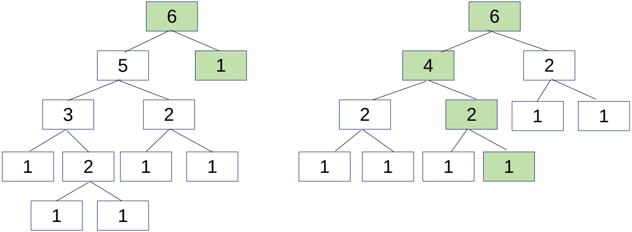
KNN-based vs LOF-based



Proximity-based Methods

Isolation Forest

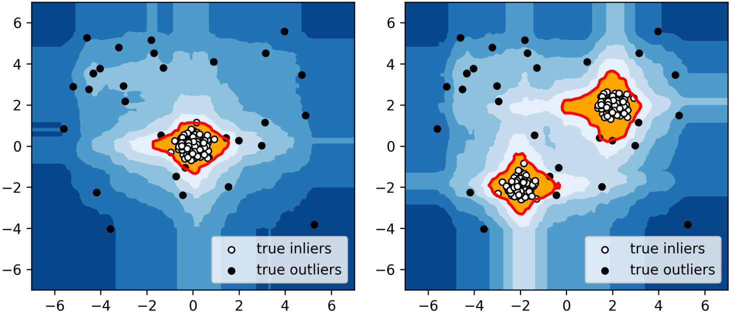
- Anomalies are **few and different**.
  - Collection of isolation trees (iTrees)
  - Each iTree isolates every case from the remaining cases for a given sample
  - Anomalies should be more susceptible to isolation, i.e., they exhibit a shorter average path
  - $Score(x) = \frac{1}{t} \sum_{i=1}^{t} l_i(x)$, where $l_i(x)$ is the path length of observation x in tree i



[1] Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua (2008). "Isolation Forest". 2008 Eighth IEEE International Conference on Data Mining: 413–422.

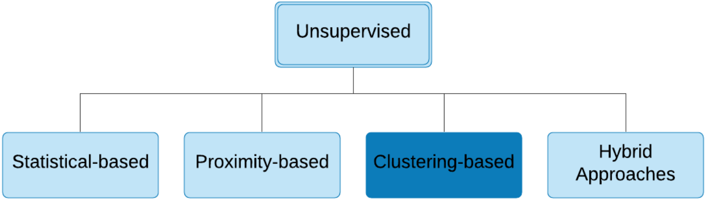
Proximity-based Methods

Isolation Forest



[1] Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua (2008). "Isolation Forest". 2008 Eighth IEEE International Conference on Data Mining: 413–422.

Unsupervised Methods



Clustering-based Methods

DBSCAN

- Idea: find the areas that satisfy a simple minimum density level, and which are separated by areas with lower density.
- Parameters: *MinPts*: threshold for the number of neighbors, ϵ: radius
- Objects with more than *MinPts* neighbors within a radius of ϵ (including the query point) are considered to be core points.

Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." KDD. Vol. 96. No. 34. 1996.

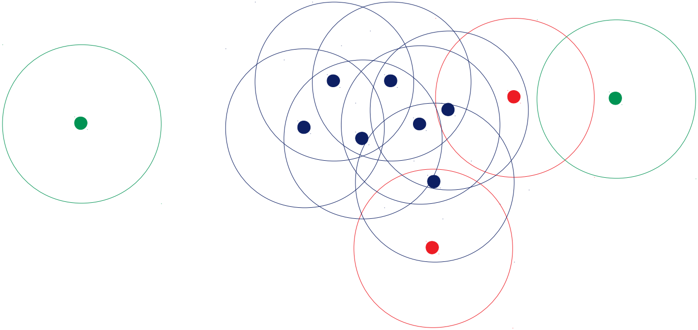
Schubert, Erich, et al. "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN." ACM TODS 42.3 (2017): 1-21.

Clustering-based Methods

DBSCAN

Steps

- Compute neighbors of each point and identify core points
- Join neighboring core points into clusters
- for each non-core point
  - Add to a neighboring core point if possible
  - Otherwise, add to noise

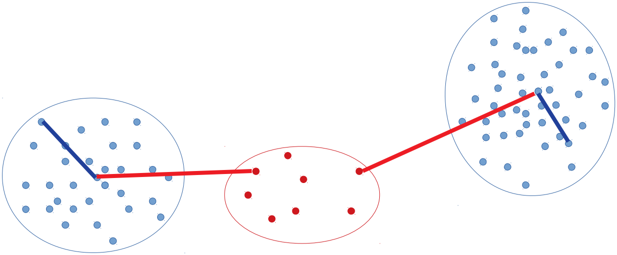


Clustering-based Methods

CBLOF

Cluster-based Local Outlier Factor (He, 2003)

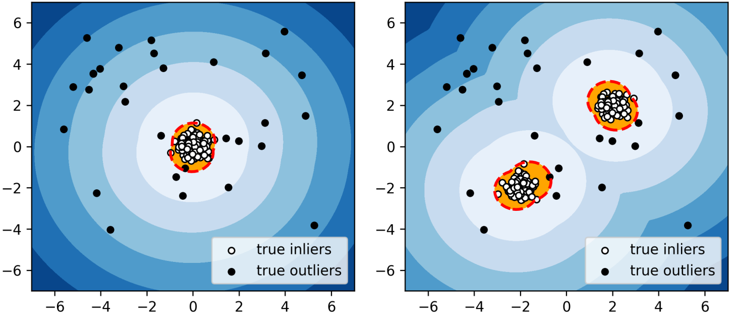
- Two parameters:
  - α: ratio of the data set that is expected to be normal
  - β: minimum ratio of the size of the large cluster to the small clusters
- Idea: Anomaly score of a case is equal to the distance to the nearest large cluster multiplied by the size of the cluster the case belong to.



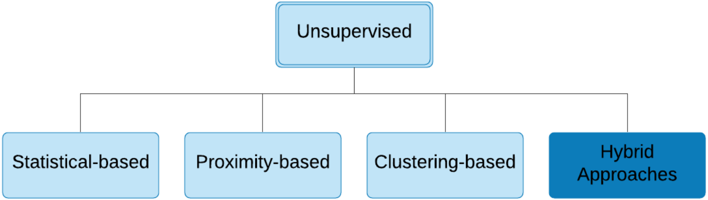
He, Zengyou, Xiaofei Xu, and Shengchun Deng. "Discovering cluster-based local outliers."

Clustering-based Methods

CBLOF



Unsupervised Methods



Unsupervised

Statistical-based

Proximity-based

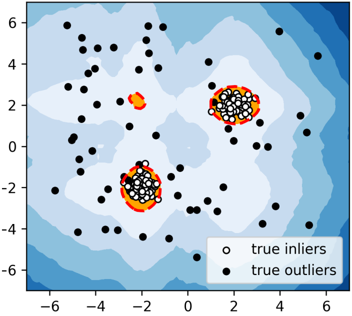
Clustering-based

Hybrid Approaches

Hybrid Approaches

Feature Bagging for Outlier Detection

- Feature Bagging for Outlier Detection runs LOF method on multiple projections of the data and combines the results for improved detection qualities in high dimensions.
- First ensemble learning approach to outlier detection.



Lazarevic, A. and Kumar, V., 2005, Feature bagging for outlier detection. In KDD '05. 2005.

Up next ...

- Semi-supervised Methods
- Class-based Anomaly Detection
- Explanation of Rare Events

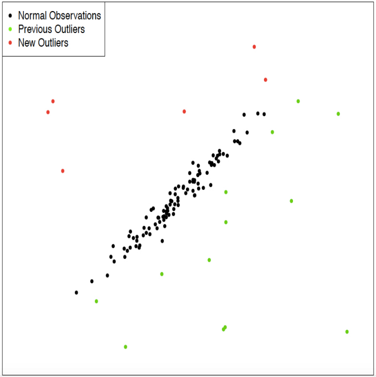
Semi-supervised outlier detection

training data has labeled instances only for one class

the most common - only normal data available, labeled outliers are missing

more robust than unsupervised methods

can outperform supervised ones if we are not sure about representativness of labeled outliers



Jason Sopheap Tun, Semi-Supervised Outlier Detection Algorithms, U. California 2018

Methods

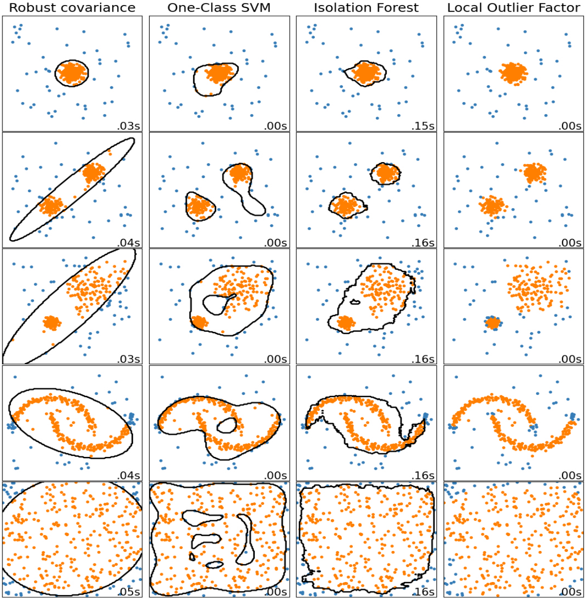
One-class learning: disadvantage: can be sensitive (as One-class SVM) to outliers and thus does not perform very well (see also Aggarwal for details)

any unsupervised anomaly detection algorithm can be used

- learning set contains only normal instances, test set both
- = (a sort of) novelty detection
- Evaluation: outliers lie outside the area

Novelty detection is more general: can result even in a dense cluster that is far from normal points.

Example: sckit-learn



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

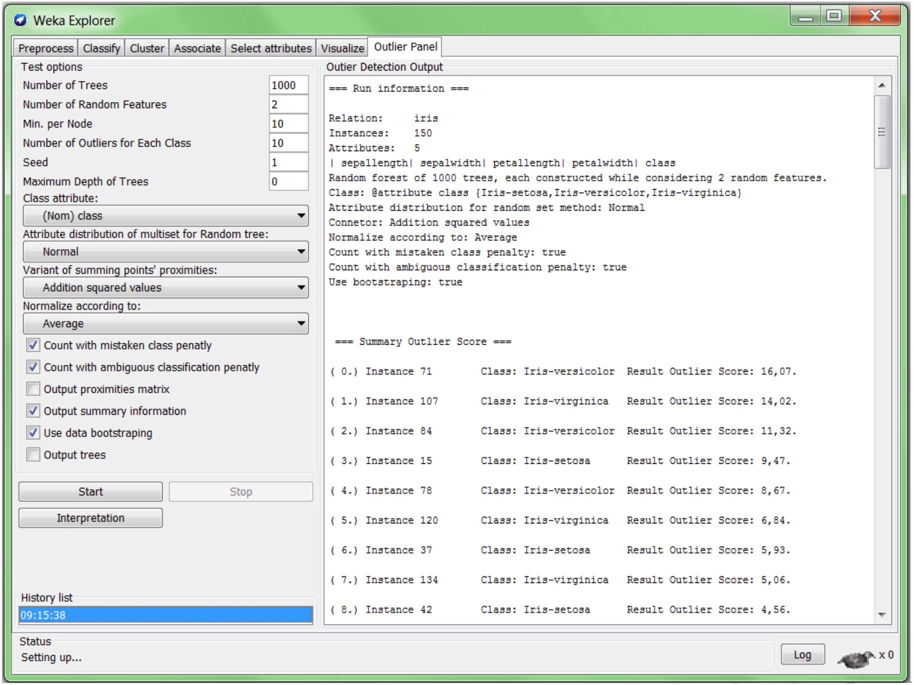
use proximity matrix for class outlier factor computation

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.



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

History list

09:15:38

Status

Setting up...

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.
```

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

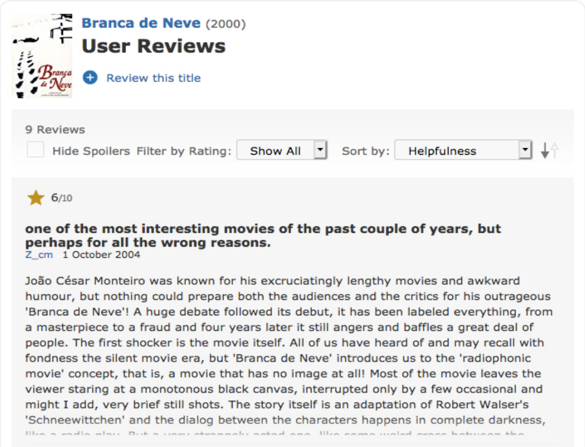
ANGIULLI, Fabrizio; FASSETTI, Fabio. Exploiting domain knowledge to detect outliers. Data Mining and Knowledge Discovery. 2014, vol. 28, no. 2,

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)



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. 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

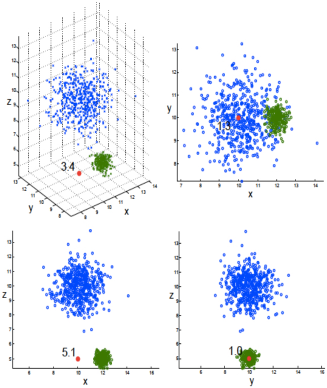


Figure 4.1: A 3D space $\{x, y, z\}$ and all its 2D projections. $\{x, z\}$ is an explanatory subspace.

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.

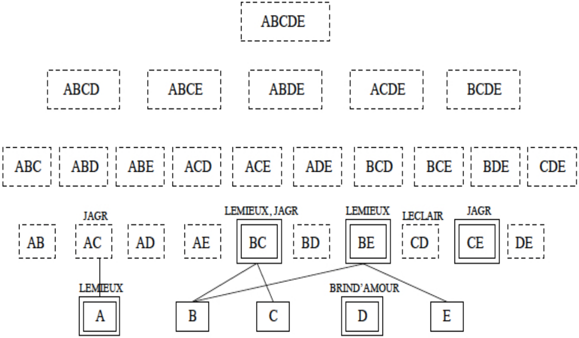
Any P that is an outlier in A is called a *strongest* outlier.

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 ⇒ 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

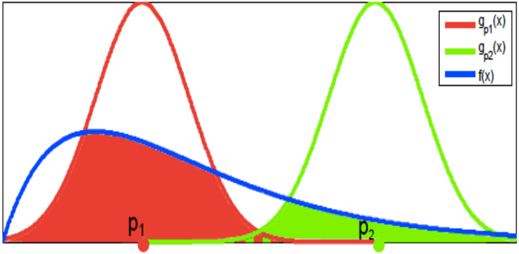
- separability - instance outlierness is related to its separability from the rest of the data

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)

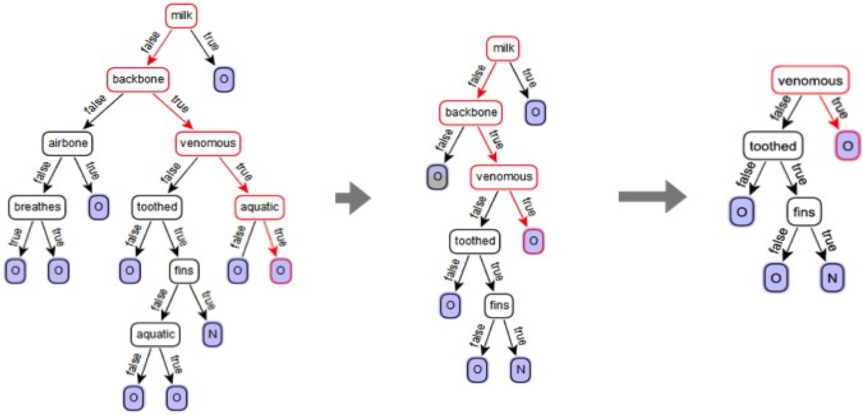
- separablity as error at classification. Assume that the data follows a distribution f
- original data = inlierclass; outlier + artificial points = outlierclass
- use standard feature selection methods to find explanatory subspaces



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)

Open Challenges

Imbalanced Domains and Rare Events

The future

- More and more human activities are being monitored through data collection
- For a large set of critical application domains, **stability** is a key factor
- Deviations from *normality* are undesirable and their timely anticipation may be highly rewarding
- **This is the central playing field of imbalanced domains and rare event detection**

Some Examples

- Monitoring critical ecosystems (e.g. Ocean, Marine Protected Areas, etc.)
- Autonomous vehicles
- Health data science (e.g. monitoring elderly people, ICU's)
- Financial domains
- etc.

Some Critical Aspects of Imbalanced Domains

- We face Imbalanced Domains when the following two conditions (**both**!) are true:
    - Not all values of the target variable are equally important
    - The more important values are scarcely represented in the training data

- How to differentiate importance?
- How to define rarity?
- How to make algorithms focus on what is important (being rare)?
- How to properly evaluate the performance of the algorithms on what it matters to the end user?

Some Critical Aspects of Imbalanced Domains (cont.)

- Established answers exist for some special cases.
  - Problem definition
    - The most established case is **Binary Classification** with one class value being rare and more important
  - Focus of algorithms
    - The most established methods revolve around resampling
  - Evaluation
    - The most common is to use the Precision+Recall/ F-measure setting

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.

Multiclass imbalance

Classification problems with more than two classes, with differentiated importance among classes, with several of the relevant class values being rare

- How to express which classes are relevant?
- How to differentiate the different classes (maybe some values are more important than others) ?
- How to bias the algorithms taking into account the importance information?
- How to evaluate the performance in these contexts?

Multiclass imbalance

How to express which classes are relevant?

- Branco et. al (2017) proposed to use the concept of relevance to specify the importance of each class value.
- How to derive this relevance information depends on the type of information we can get from the user:
  - Informal - the user just let us know that the rarer the class the more important
  - Intermediate - the user is able to establish a partial order of importance between class values
  - Full - the user is able to provide full quantification of the importance of each class value
- The authors provide means of estimating the relevance of each class for different user information settings
- Based on the estimated relevance the authors propose a series of evaluation metrics for the performance of the models

P. Branco, L. Torgo, and R. Ribeiro. "Relevance-based evaluation metrics for multi-class imbalanced domains." PAKDD. Springer, Cham, pp.698-710 (2017).

Multiclass imbalance

Open Challenges

- How to bias the models?
  - Resampling based on relevance?
  - Algorithms directly optimizing the proposed metrics?
  - Other approaches to multiclass? Multiple binary problems (e.g. Fernández et al, 2013)?

Fernández, A., López, V., Galar, M., del Jesús, M.J., Herrera, F.: Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches. Knowledge Based Systems, 42, 97–110 (2013)

Imbalance in Time Series Forecasting

In time series observations are ordered by time and the goal is to obtain models able to forecast future values of the series

- Imbalance occurs when some of the observed values are more important, but rare.
- Time series are usually numeric, so can we re-use the methods of imbalanced regression ?

Imbalance in Time Series Forecasting

Specifying differentiated importance

Moniz et al. (2017) used the concept of relevance by automatically deriving it assuming extreme values of the time series are rare and thus more important (e.g. returns of stock market)

- Standard resampling algorithms were applied to numeric time series forecasting tasks (under- and over-sampling and SMOTE)
- Several biased resampling methods were proposed
  - Time bias - preferring to keep important values that are more recent
  - Time and relevance bias - also allowing older values if they are extremely relevant

N. Moniz, P. Branco, and L. Torgo. Resampling strategies for imbalanced time series forecasting. International Journal of Data Science and Analytics, 3(3):161–181, 2017.

Imbalance in Time Series Forecasting

Open Challenges

- How to address other importance criteria (not only extreme values)?
- What about time series of symbolic values? Can we also adapt classification approaches?
- Are there other forms of biasing resampling using other properties of time series (e.g. taking into account seasonality)?

Imbalance in Data Streams

Data Streams share some of the characteristics of Time Series but are usually too big for normal batch learning and frequently have serious concept drift effects

- Imbalance ratio may change with time
- What was rare may become common and vice versa
- New types of values not seen in the past may appear in the data
- Full past data access is typically impossible

Imbalance in Data Streams (cont.)

- Several authors describe some of the efforts in this area (e.g. Krawczyk et al., 2017; or Hoens et al,, 2012)
- A frequent strategy to fight problems raised by the properties of data streams the use of ensembles

B. Krawczyk, L. Minku, J. Gama, J. Stefanowski, and M. Wozniak. Ensemble learning for data stream analysis: A survey. Information Fusion, 37:132 – 156, 2017. ISSN 1566-2535. doi: http://dx.doi.org/10.1016/j.inffus.2017.02.004.

Hoens, T.R., Polikar, R., Chawla, N.V.: Learning from streaming data with concept drift and imbalance: an overview. Progress AI 1(1), 89–101 (2012)

Imbalance in Data Streams

- With the advances of mobile computing more and more data with both spatial and temporal properties is being collected
- How to resample a dataset that is a moving target?
- How to properly evaluate the performance in these settings?

Imbalance in Spatiotemporal Datasets

In spatiotemporal forecasting you have to cope not only with temporal correlation but also spatial correlation.

- Are standard resampling strategies applicable?
- Is there any change on the way we should evaluate the models?
- Is it worth to think about special purpose learning algorithms?

Imbalance in Spatiotemporal Datasets (cont.)

Oliveira et al., 2019 are among the first to address the issue of imbalance on spatiotemporal forecasting

- The authors propose resampling approaches tuned for this type of data
- They proposed biased resampling that takes into account the temporal and spatial correlation of the data
- The bias is calculate through temporal and spatial weights
  - temporal - favour more recent observations
  - spatial 1 - favour spatially isolated rare cases
  - spatial 2 - decrease the weights on cases that are faraway from rare cases

M. Oliveira, N. Moniz, L. Torgo and V. Santos Costa, "Biased Resampling Strategies for Imbalanced Spatio-Temporal Forecasting," 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Washington, DC, USA, 2019, pp. 100-109, doi: 10.1109/DSAA.2019.00024.

Imbalance in Spatiotemporal Datasets

Open Challenges

- Cope with situations of real time data (spatiotemporal data streams)
- New forms of sampling bias
- Cope with rarity that may be location dependent

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**

Learning with Imbalanced Domains and Rare Event Detection

Luis Torgo, Stan Matwin, Nathalie Japkowicz, **Nuno Moniz**, **Paula Branco**, **Rita P. Ribeiro** and **Lubomir Popelinsky**

Dalhousie University, Canada
INESC TEC, University of Porto, Portugal
American University, USA
University of Ottawa, Canada
Masaryk University, Czech Republic

September, 2020