Rare Event Detection

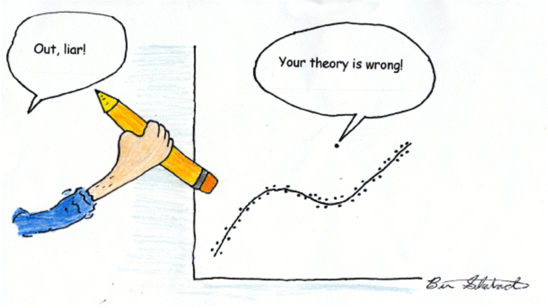
Principles

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

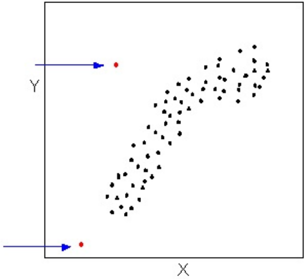
- Each data instance has:
  - One attribute (univariate)
  - Multiple attributes (multivariate)
- Relationship among data instances:
  - None
  - Sequential/Temporal
  - Spatial
  - Spatio-temporal
  - Graph
- Dimensionality of data

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.

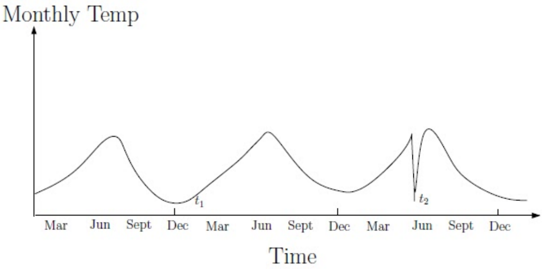


Types of Outliers (cont.)

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.

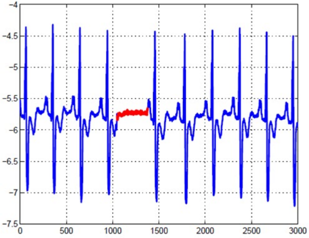


Types of Outliers (cont.)

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.

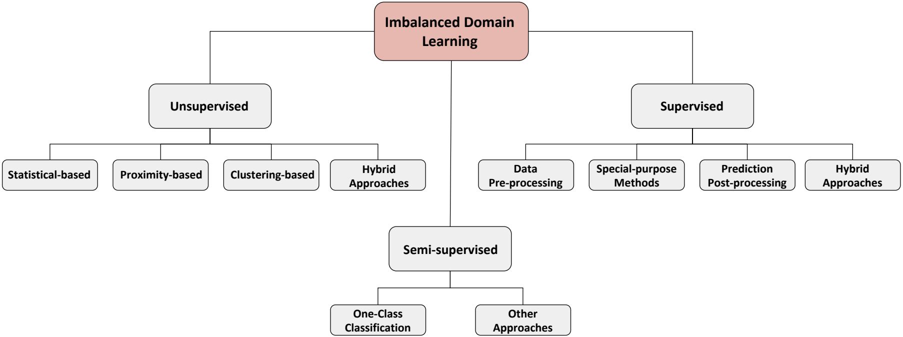
Credit Card Fraud Detection:

- 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!

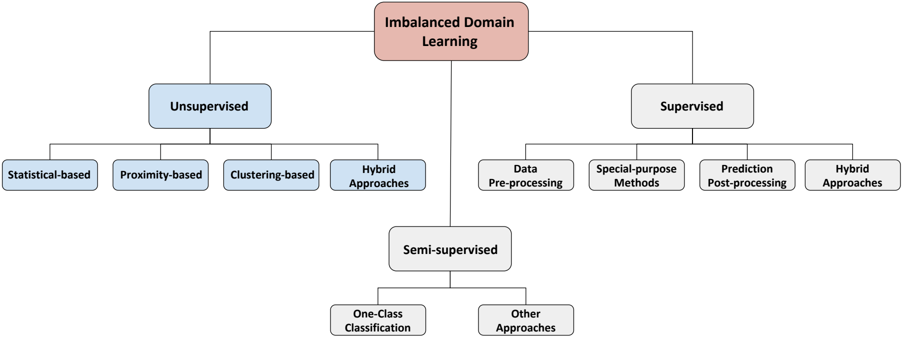
Taxonomy of Approaches

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

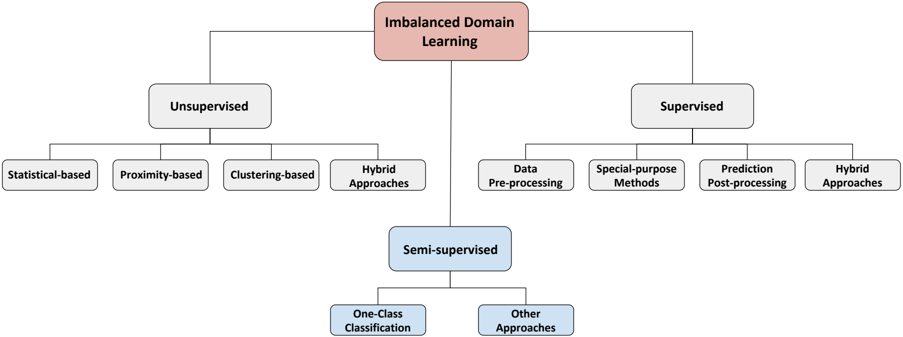
- Easily adaptable to on-line/incremental mode.
- Test phase is fast.

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

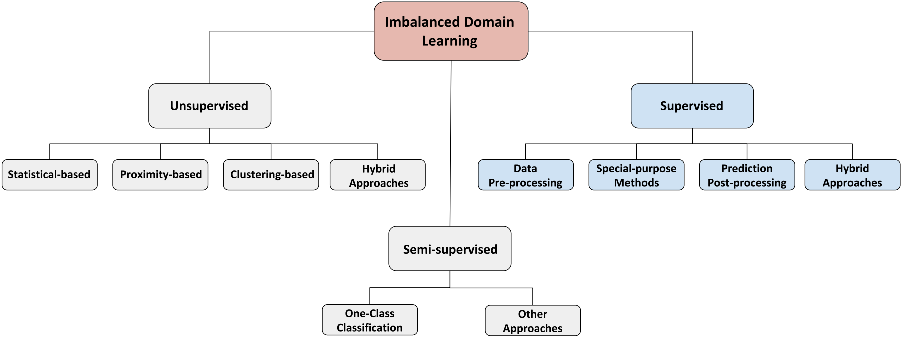
- Models are interpretable.
- Normal behaviour can be accurately learned.
- Can detect new outliers that may not appear close to an outlier point in the training set.

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



Imbalanced Domain Learning

Unsupervised

Statistical-based

Proximity-based

Clustering-based

Hybrid Approaches

Supervised

Data Pre-processing

Special-purpose Methods

Prediction Post-processing

Hybrid Approaches

Semi-supervised

One-Class Classification

Other Approaches

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 $\phi(Y) : \mathcal{Y} \rightarrow [0,1]$ 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 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.)

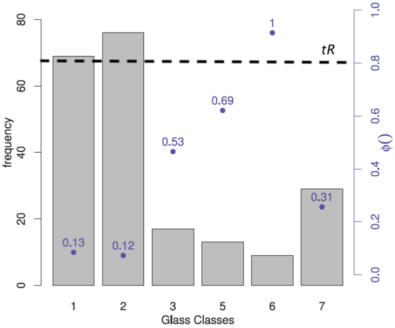
Supervised Imbalanced Domain Learning

- 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 $\phi(Y)$:

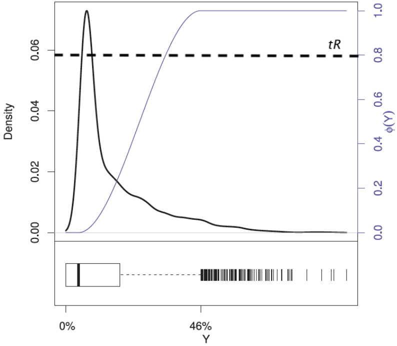
- interpolation method
  - user provides a set of interpolating points
- automatic method
  - no input required from the user;
  - it uses the target variable distribution;
  - it assumes that the most relevant cases are located at the extremes of the target variable distribution.

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