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

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.

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

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!

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.)

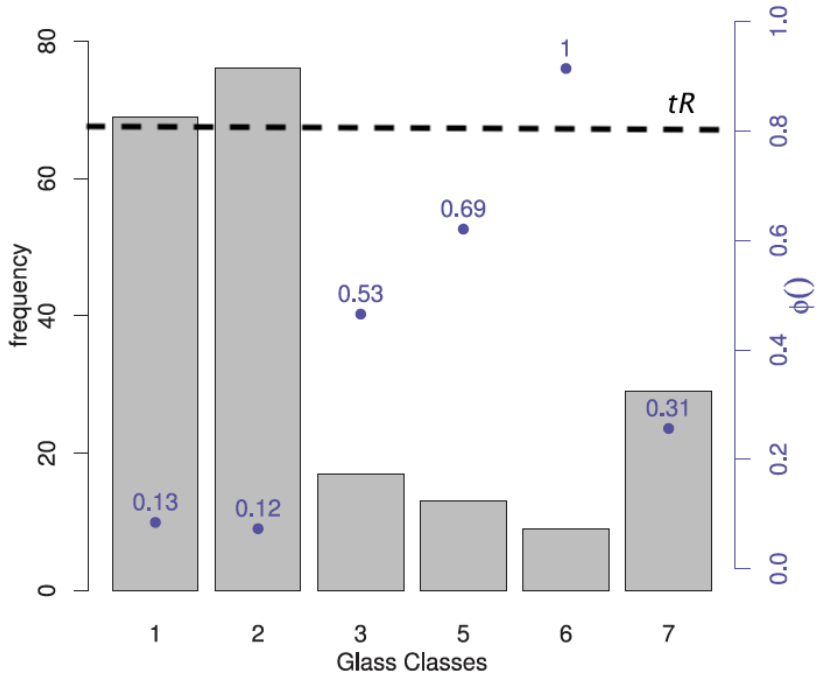
Supervised Imbalanced Domain Learning

- 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 $\phi(Y)$:

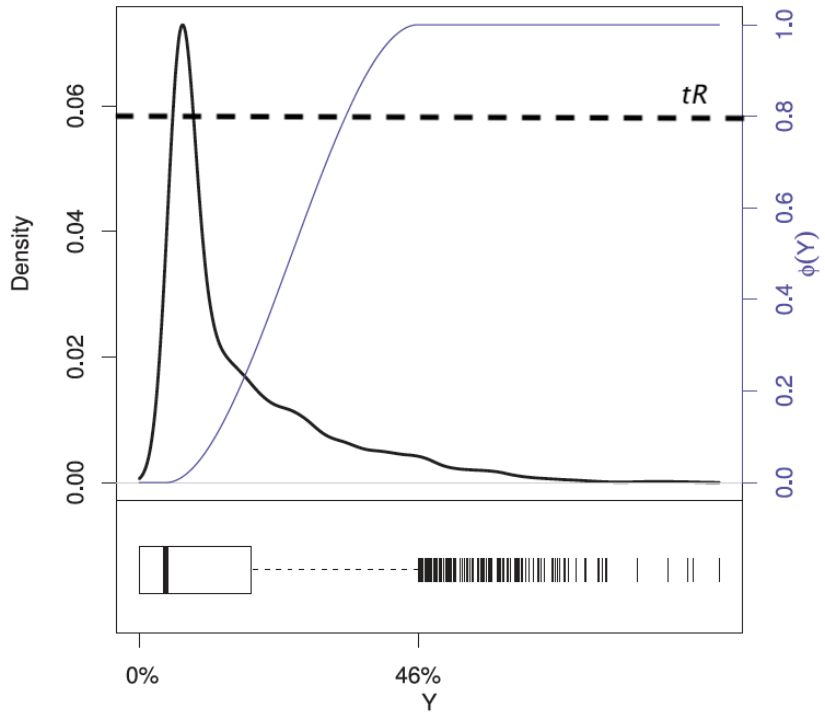
- interpolation method
  - user provides a set of interpolating points
- automatic method
  - no input required from the user;
  - it uses the target variable distribution;
  - it assumes that the most relevant cases are located at the extremes of the target variable distribution.

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.