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