4th Workshop on Learning with Imbalanced Domains – Theory and Applications: Preface

Nuno Moniz NUNOMONIZ@ND.EDU
Lucy Family Institute for Data & Society, University of Notre Dame
Indiana, USA

Paula Branco PAULA.BRANCO@DCC.FC.UP.PT
Faculty of Engineering, University of Ottawa
Ottawa, Canada

Luís Torgo LTORGO@DAL.CA
Faculty of Computer Science, Dalhousie University
Halifax, Canada

Nathalie Japkowicz JAPKOWIC@AMERICAN.EDU
Department of Computer Science, American University
Washington DC, USA

Michal Wozniak MICHAL.WOZNIAK@PWR.EDU.PL
Wroclaw University of Science and Technology Wroclaw
Poland

Shuo Wang S.WANG.2@BHAM.AC.UK
University of Birmingham
Birmingham, UK

This volume contains the Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications - LIDTA'2022. This Workshop was co-organised by the Lucy Family Institute for Data & Society (Indiana, USA), the School of Electrical Engineering and Computer Science, Faculty of Engineering, University of Ottawa (Ottawa, Canada), the Faculty of Computer Science of the Dalhousie University (Halifax, Canada), the Department of Computer Science of the American University (Washington DC, USA), the Wroclaw University of Science and Technology (Wroclaw, Poland) and the University of Birmingham (Birmingham, UK). The Workshop was co-located with the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML/PKDD) 2022 and was held on the 23rd of September 2022 in the World Trade Center, Grenoble, France.

The LIDTA 2022 Workshop focused on theoretical and practical aspects of the problem of learning from imbalanced domains. For a diverse and vast set of real-world applications, the end-user is interested in obtaining predictive models that reflect their non-uniform preferences over the target variable domain. In imbalanced domains the target variable values that have more value to the end-user are scarcely represented in the available training data. Moreover, these least-common values are often associated with events that are highly relevant and with potentially high costs and/or benefits. Examples of real-world applications where this problem occurs include different domains such as financial (e.g. unusual returns on stock markets), medical (e.g. diagnosis of rare diseases), meteorological (e.g. anticipation of catastrophes) or social media (e.g. popularity prediction).

Over the last decade, the problem of learning from imbalanced domains has been extensively studied with a particular focus on binary classification tasks. More recently, it became clear that this problem also occurs within other predictive contexts such as regression (Torgo et al., 2013), ordinal classification (Kim et al., 2016), multi-label classification (Zhang et al., 2020), association rules mining (Luna et al., 2015), multi-instance learning (Vluymans et al., 2016), data streams (Krawczyk et al., 2017) and time series and spatio-temporal forecasting (Moniz et al., 2017; Oliveira et al., 2019). Nowadays, it is recognized that the problem of learning from imbalanced domains is a broad issue with several important challenges and widespread through a diversity of tasks including supervised and unsupervised problems.

It is crucial to both academia and industry to address the issues raised by imbalanced domains. Regarding the industry, many real-world applications are already facing this problem and seek solutions that are prompt, innovative, suitable and effective. These applications include a diversity of real problems such as fraud prevention, the anticipation of catastrophes, detection of faults in industrial systems or early diagnosis of rare diseases. Regarding the research community, this problem represents an opportunity to innovate by developing new systems and approaches that are able to deal with challenging and complex tasks. Nowadays, it becomes urgent to develop such solutions for this problem while considering the full spectrum of predictive tasks that suffer from the imbalance problem.

The 2022 edition of LIDTA was organized as a half-day workshop. The main goal of LIDTA 2022 was to provide a significant contribution to the problem of learning with imbalanced domains and to increase the interest and the contributions for solving some of its challenges. The event received a very high attendance, clearly reflecting the interest in the topic.

LIDTA 2022 was held in the morning. It included a keynote talk and the presentation and discussion of the accepted papers, both short and long. The invited talk was entitled “Fairness under class-imbalance”, by Professor Eirini Ntoutsi, from the Bundeswehr University Munich. Concerning the paper contributions, the workshop has received many high-quality inter-disciplinary contributions discussing various aspects of learning from imbalanced domains. Overall, there were 19 paper submissions, out of which 8 were accepted as long presentation and 6 as short presentation, both with inclusion in these workshop proceedings. These papers cover different aspects of imbalanced learning. Let us now briefly describe the accepted papers.

Antoniadis et al. (2022) discussed a systematic evaluation of combined algorithm selection and hyperparameter optimization search strategies for unsupervised anomaly detection. The authors analyzed how the structure of the validation set impacts the performance of different algorithm and hyperparameter optimization search strategies in anomaly detection. Block and Bekker (2022) proposed a robust bagging-based method for learning from biased positive and unlabellled (PU) data that tackles the problem frequently observed in PU data related to a labeling bias. Vasquez et al. (2022) introduced a novel technique named PU-CSBoost that integrates PU learning and the instance-dependent cost-sensitive framework. This technique minimizes the financial loss associated to fraud cases through an instance-dependent cost measure that incorporates the misclassification cost due to hidden fraudsters. Song et al. (2022) studied the adaptation to a distributed setting of SMOGN algorithm, a resampling strategy for dealing with imbalanced regression problems. The impact in the performance and time is evaluated in a variety of datasets. Shi et al. (2022) introduced a new intermediate training stage that leverages the idea of data augmentation to learn an initial representation that better fits the imbalanced distribution of the task at hands during the pre-finetuning stage. The impact of applying different cost-sensitive algorithms and resampling techniques when learning with different learners was assessed by da Silva Freitas Junior and Pisani (2022). The relative performance over different imbalance ratios as well as the influence of these techniques on the machine learning models complexities was analyzed. Wojciechowski and Lango (2022) addressed the multi-class imbalance problem by proposing a new oversampling algorithm for neural networks that changes oversampled instances during training to further expand the decision region towards the minority classes. Bonnier and Bosch (2022) experimentally assessed the robustness of various ordinal classifiers, with a focus on risk rating tasks. The authors carry out this evaluation under two scenarios: lack of training data and data distribution shift.

Rushe and Namee (2022) tackle the problem of novelty detection with contextual information by predicting this contextual information using an auxiliary prediction strategy which takes advantage of the rarity of novel examples. Bauvin et al. (2022) presented the use cases of SuMMIT, a software for running, tuning, and analyzing experiments of supervised classification tasks specifically designed for multi-view data set. In particular, the authors demonstrated the usefulness of such a platform for dealing with the complexity of multi-view benchmarking on an imbalanced setting. Agostinho and Mendes-Moreira (2022) proposed a probabilistic metric based on non-parametric approaches that, given the results of a classifier in a multi-class domain, verifies the relation between these results and the imbalance problem. Bouayed et al. (2022) studied the usefulness of the algebraically independent rotation invariant features extracted from 4th degree spherical harmonics that model the dMRI signal per voxel in the context of Alzheimer Disease identification. Komorniczak et al. (2022) showed that data oversampling techniques may lead to dataset simplification according to selected data difficulty metrics and that such simplification positively affects the quality of selected classifier learning methods. Lipska and Stefanowski (2022) studied the impact of local data characteristics and drifts on the difficulties of learning online classifiers from multi-class imbalanced data streams. The authors found that overlapping between many minority classes has a greater role than for streams with one minority class. Moreover, the presence of rare examples in the stream was the most difficult single factor.

We would like to thank all of the authors and the Program Committee members that enabled a successful workshop, for their hard work and commitment. We also want to deeply thank the ECML/PKDD 2022 Workshop Chairs for their support in the logistics of this workshop.

Organizing Committee

- **Nuno Moniz** - Lucy Family Institute for Data & Society, University of Notre Dame, Indiana, USA
- **Paula Branco** - University of Ottawa, School of Electrical Engineering and Computer Science, Canada

- **Luís Torgo** - Dalhousie University, Faculty of Computer Science, Canada
- **Nathalie Japkowicz** - American University, Department of Computer Science, USA
- **Michal Wozniak** - Wroclaw University of Science and Technology, Poland
- **Shuo Wang** - University of Birmingham, School of Computer Science, UK

Program Committee

- Colin Bellinger, University of Alberta
- Chris Drummond, NRC Institute for Information Technology
- Inês Dutra, DCC - Faculty of Sciences, University of Porto
- Alberto Fernández, Granada University
- Mikel Galar, Universidad Pública de Navarra
- Salvador Garcia, University of Granada
- Raji Ghawi, Technical University of Munich
- Bartosz Krawczyk, Virginia Commonwealth University
- Leandro Minku, University of Birmingham
- Rita Ribeiro, DCC - Faculty of Sciences, University of Porto
- Marina Sokolova, University of Ottawa
- Jerzy Stefanowski, Poznan University of Technology
- Gary Weiss, Fordham University

References

Solander Patricio Lopes Agostinho and João Mendes-Moreira. Probabilistic metric to measure the imbalance in multi-class problems. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 151–162, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/agostinho22a.html.

Ioannis Antoniadis, Vincent Vercruyssen, and Jesse Davis. Systematic evaluation of cash search strategies for unsupervised anomaly detection. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 8–22, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/antoniadis22a.html.

Baptiste Bauvin, Jacques Corbeil, Dominique Benielli, Sokol Koço, and Cecile Capponi. Integrating and reporting full multi-view supervised learning experiments using summit. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 139–150, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/bauvin22a.html.

Sander De Block and Jessa Bekker. Bagging propensity weighting: A robust method for biased pu learning. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 23–37, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/block22a.html.

Thomas Bonnier and Benjamin Bosch. Assessing the robustness of ordinal classifiers against imbalanced and shifting distributions. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 112–126, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/bonnier22a.html.

Aymene Mohammed Bouayed, Samuel Deslauriers-Gauthier, Mauro Zucchelli, and Rachid Deriche. Cnn and diffusion mri's 4th degree rotational invariants for alzheimer's disease identification. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 163–174, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/bouayed22a.html.

Jairo da Silva Freitas Junior and Paulo Henrique Pisani. Performance and model complexity on imbalanced datasets using resampling and cost-sensitive algorithms. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning*

Research, pages 83–97, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/junior22a.html.

Sungil Kim, Heeyoung Kim, and Younghwan Namkoong. Ordinal classification of imbalanced data with application in emergency and disaster information services. *IEEE Intelligent Systems*, 31(5):50–56, 2016.

Joanna Komorniczak, Paweł Ksieniewicz, and Michał Woźniak. Data complexity and classification accuracy correlation in oversampling algorithms. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 175–186, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/komorniczak22a.html.

Bartosz Krawczyk, Leandro L. Minku, João Gama, Jerzy Stefanowski, and Michał Woźniak. 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.

Agnieszka Lipska and Jerzy Stefanowski. The influence of multiple classes on learning from imbalanced data streams. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 187–198, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/lipska22a.html.

José María Luna, Cristóbal Romero, José Raúl Romero, and Sebastián Ventura. An evolutionary algorithm for the discovery of rare class association rules in learning management systems. *Applied Intelligence*, 42(3):501–513, 2015.

Nuno Moniz, Paula Branco, and Luís Torgo. Resampling strategies for imbalanced time series forecasting. *International Journal of Data Science and Analytics*, 3(3):161–181, 2017.

Mariana Oliveira, Nuno Moniz, Luís Torgo, and Vítor Santos Costa. Biased resampling strategies for imbalanced spatio-temporal forecasting. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 100–109, 2019. doi: 10.1109/DSAA.2019.00024.

Ellen Rushe and Brian Mac Namee. Deep contextual novelty detection with context prediction. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 127–138, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/rushe22a.html.

Yiwen Shi, Taha ValizadehAslani, Jing Wang, Ping Ren, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. Improving imbalanced learning by pre-finetuning with data augmentation. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 68–82, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/shi22a.html.

Xin Yue Song, Nam Dao, and Paula Branco. Distsmogn: Distributed smogn for imbalanced regression problems. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 38–52, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/song22a.html.

Luís Torgo, Rita P Ribeiro, Bernhard Pfahringer, and Paula Branco. Smote for regression. In *Progress in Artificial Intelligence*, pages 378–389. Springer, 2013.

Carlos Ortega Vasquez, Jochen De Weerdt, and Seppe vanden Broucke. The hidden cost of fraud: An instance-dependent cost-sensitive approach for positive and unlabeled learning. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 53–67, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/vazquez22a.html.

Sarah Vluymans, Dánel Sánchez Tarragó, Yvan Saeys, Chris Cornelis, and Francisco Herrera. Fuzzy rough classifiers for class imbalanced multi-instance data. *Pattern Recognition*, 53:36–45, 2016.

Adam Wojciechowski and Mateusz Lango. Adversarial oversampling for multi-class imbalanced data classification with convolutional neural networks. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak, and Shuo Wang, editors, *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2022)*, volume 183 of *Proceedings of Machine Learning Research*, pages 98–111, ECML-PKDD, Grenoble, France, 19–23 Sept 2022. PMLR. URL http://proceedings.mlr.press/v183/wojciechowski22a.html.

Min-Ling Zhang, Yu-Kun Li, Hao Yang, and Xu-Ying Liu. Towards class-imbalance aware multi-label learning. *IEEE Transactions on Cybernetics*, 2020.