3rd Workshop on Learning with Imbalanced Domains: Preface

Nuno Moniz NMMONIZ@INESCTEC.PT
INESC TEC / Faculty of Sciences, University of Porto
Porto, Portugal

Paula Branco PBRANCO@UOTTAWA.CA
School of Electrical Engineering and Computer Science, University of Ottawa
Ontario, 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

Michał Woźniak 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 Third International Workshop on Learning with Imbalanced Domains: Theory and Applications - LIDTA 2021. This Workshop was co-organised by INESC TEC, the Department of Computer Science at the Faculty of Sciences of the University of Porto (Portugal), the School of Electrical Engineering and Computer Science at the University of Ottawa (Canada), the Faculty of Computer Science at Dalhousie University (Canada), the Department of Computer Science of the American University (USA), the Wroclaw University of Science and Technology (Poland) and the University of 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) 2021 and was held as an online event on the 17th of September 2021.

The LIDTA 2021 Workshop focused on both theoretical and practical aspects of the problem of learning from imbalanced domains. In multiple real-world applications, the end-user aims at obtaining predictive models that are able to reflect her/his domain preferences. When these preferences are not uniform over the target variable domain, this causes a problem. Non-uniform preferences are critical in imbalanced domains where we observe that the most relevant target variable values for the end-user are scarcely represented. This problem is evident in many real-world domains such as financial (Kamalov, 2020), medical (Cao et al., 2018), meteorological (Troncoso et al., 2018), cybersecurity (Wheelus et al., 2018) or social media (Li and Liu, 2018).

The problem of imbalanced domains has been extensively studied on the last decade for binary classification tasks. The study of this problem in other predictive contexts has been gaining more attention in the recent years. In fact, several researchers are now focused on tackling this problem in the context of multiclass problems (Koziarski et al., 2020), regression tasks (Torgo et al., 2013), multi-label classification (Charte et al., 2019), association rules mining (Luna et al., 2015), multi-instance learning (Vluymans et al., 2016), data streams (Krawczyk et al., 2017), and time series (Moniz et al., 2017), among other. This is in fact a broad issue involving multiple challenges which is common to a diversity of tasks.

The LIDTA 2021 workshop is focused on these problems. Following the trend of the previous editions of this workshop, LIDTA 2021 received a diverse set of contributions. The selected papers are high quality, inter-disciplinary articles that discuss numerous aspects of the problem of learning form imbalanced domains. Overall, there were 13 paper submissions, out of which 8 papers were accepted for inclusion in the workshop proceedings. The high workshop attendance reflected the great interest of the research community in the topic. The workshop included a morning session and an afternoon session. Each session included a keynote talk. After the welcoming, the morning session started with an invited talk entitled "Learning with Imbalanced Data Streams", by Professor Bartosz Krawczyk, from the Department of Computer Science, Virginia Commonwealth University. The afternoon session started with our second invited talk from Professor Nathalie Japkowicz from the Department of Computer Science, American University, Washington DC, USA, entitled "Class Imbalances and Deep Learning". Both talks raised several interesting questions and remarks from the audience. The success of this workshop edition which builds on the previous workshops accomplishments enabled a follow-up Special Issue on Imbalanced Learning, hosted by the Machine Learning Journal.

All the papers accepted in LIDTA 2021 workshop were assigned a 15-minute presentation slot, followed by 5 minutes for questions and answers. Four papers were presented in each one of the morning and afternoon sessions. More details about the accepted papers are described next.

Nardi et al. (2021) address the anomaly detection problem in decentralized scenarios. The authors propose an unsupervised ensemble method for this problem in which the base learners are lightweight autoencoders. Sadeghi and Viktor (2021) present Online-MC-Queue algorithm, a novel solution for online learning in the context of multi-class imbalanced problems. The authors use a queue-based resampling method that creates an instance queue for each problem class. This queue maintains the number of instances balanced. A novel algorithm named Multi-label Neighbourhood Component Analysis (ML-NCA) is proposed by Pakrashi et al. (2021). The ML-NCA is designed for addressing issues on multi-label classification problems. ML-NCA performs a supervised linear transformation of the input space to obtain a new space where KNN-based algorithms are expected to perform well. Draghi et al. (2021) explore methods to improve synthetic data generators. The authors use probabilistic methods to identify difficult to predict data samples, and then use these methods to boost these types of data when generating synthetic samples. Limnios et al. (2021) present a new method for outlier detection in the presence of vast amounts of normal data. The authors propose the learning of a data-driven scoring function that reflects the degree of abnormality of the observations. Nazari and Branco (2021) present an analysis of the impact of using CGANs as an oversampling strategy as a method to

tackle the class imbalance and other data difficult factors. Naklicka and Stefanowski (2021) provide a contribution related to the extension of the BRACID rule-based classifier to a multi-class imbalanced scenario. Two solutions are proposed: the first uses BRACID in the OVO ensemble along with modifications of the prediction aggregation strategy while the second changes the rules induction for multiple classes simultaneously. Finally, Bougaham et al. (2021) present a solution for an application domain involving the use of intermediate patches, after a WGAN training procedure. The key goal of this approach is to enable the detection of anomalies on full size Printed Circuit Board Assembly (PCBA) images.

We would like to thank all of the authors and the Program Committee members for their hard work and commitment that allowed to carry out a successful and interesting workshop. We also want to deeply thank the ECML/PKDD 2021 Workshop and Tutorial Chairs for their support.

Organizing Committee

Program Committee

- Inaki Inza, University of the Basque Country
- Michał Koziarzki, AGH University of Science and Technology
- Bartosz Krawczyk, Virginia Commonwealth University
- Leandro Minku, University of Birmingham
- Ronaldo Prati, Universidade Federal do ABC - UFABC
- Rita Ribeiro, DCC - Faculty of Sciences, University of Porto
- Marina Sokolova, University of Ottawa
- Jerzy Stefanowski, Poznan University of Technology
- Herna Viktor, University of Ottawa
- Gary Weiss, Fordham University

References

Arnaud Bougaham, Adrien Bibal, Isabelle Linden, and Benoit Frenay. Ganodip - gan anomaly detection through intermediate patches: a pcba manufacturing case. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak, and Shuo Wang, editors, *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2021)*, volume 154 of *Proceedings of Machine Learning Research*, pages 104–117, ECML-PKDD, Bilbao, Basque Country, Spain, 13–17 Sept 2021. PMLR. URL http://proceedings.mlr.press/v154/bougaham21a.html.

Peng Cao, Fulong Ren, Chao Wan, Jinzhu Yang, and Osmar Zaiane. Efficient multi-kernel multi-instance learning using weakly supervised and imbalanced data for diabetic retinopathy diagnosis. *Computerized Medical Imaging and Graphics*, 69:112–124, 2018.

Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Remedial-hwr: Tackling multilabel imbalance through label decoupling and data resampling hybridization. *Neurocomputing*, 326:110–122, 2019.

Barbara Draghi, Zhenchen Wang, Puja Myles, and Allan Tucker. Bayesboost: Identifying and handling bias using synthetic data generators. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak, and Shuo Wang, editors, *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2021)*, volume 154 of *Proceedings of Machine Learning Research*, pages 49–62, ECML-PKDD, Bilbao, Basque Country, Spain, 13–17 Sept 2021. PMLR. URL http://proceedings.mlr.press/v154/draghi21a.html.

Firuz Kamalov. Forecasting significant stock price changes using neural networks. *Neural Computing and Applications*, 32(23):17655–17667, 2020.

Michał Koziarski, Michał Woźniak, and Bartosz Krawczyk. Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise. *Knowledge-Based Systems*, 204:106223, 2020.

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.

Chaoliang Li and Shigang Liu. A comparative study of the class imbalance problem in twitter spam detection. *Concurrency and Computation: Practice and Experience*, 30(5): e4281, 2018.

Myrto Limnios, Nathan Noiry, and Stephan Clémençon. Learning to rank anomalies: Scalar performance criteria and maximization of two-sample rank statistics. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak, and Shuo Wang, editors, *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2021)*, volume 154 of *Proceedings of Machine Learning Research*, pages 63–75, ECML-PKDD, Bilbao, Basque Country, Spain, 13–17 Sept 2021. PMLR. URL http://proceedings.mlr.press/v154/limnios21a.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.

Maria Naklicka and Jerzy Stefanowski. Two ways of extending bracid rule-based classifiers for multi-class imbalanced data. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak, and Shuo Wang, editors, *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2021)*, volume 154 of *Proceedings of Machine Learning Research*, pages 90–103, ECML-PKDD, Bilbao, Basque Country, Spain, 13–17 Sept 2021. PMLR. URL http://proceedings.mlr.press/v154/naklicka21a.html.

Mirko Nardi, Lorenzo Valerio, and Andrea Passarella. Centralised vs decentralised anomaly detection: when local and imbalanced data are beneficial. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak, and Shuo Wang, editors, *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2021)*, volume 154 of *Proceedings of Machine Learning Research*, pages 7–20, ECML-PKDD, Bilbao, Basque Country, Spain, 13–17 Sept 2021. PMLR. URL http://proceedings.mlr.press/v154/nardi21a.html.

Ehsan Nazari and Paula Branco. On oversampling via generative adversarial networks under different data difficulty factors. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak, and Shuo Wang, editors, *Proceedings of the Third*

International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2021), volume 154 of *Proceedings of Machine Learning Research*, pages 76–89, ECML-PKDD, Bilbao, Basque Country, Spain, 13–17 Sept 2021. PMLR. URL http://proceedings.mlr.press/v154/nazari21a.html.

Arjun Pakrashi, Sayel Sadhukhan, and Brian Mac Namee. Ml-nca: Multi-label neighbourhood component analysis. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak, and Shuo Wang, editors, *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2021)*, volume 154 of *Proceedings of Machine Learning Research*, pages 35–48, ECML-PKDD, Bilbao, Basque Country, Spain, 13–17 Sept 2021. PMLR. URL http://proceedings.mlr.press/v154/pakrashi21a.html.

Farnaz Sadeghi and Herna L. Viktor. Online-mc-queue: Learning from imbalanced multi-class streams. In Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak, and Shuo Wang, editors, *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2021)*, volume 154 of *Proceedings of Machine Learning Research*, pages 21–34, ECML-PKDD, Bilbao, Basque Country, Spain, 13–17 Sept 2021. PMLR. URL http://proceedings.mlr.press/v154/sadeghi21a.html.

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

A Troncoso, P Ribera, Gualberto Asencio-Cortés, I Vega, and D Gallego. Imbalanced classification techniques for monsoon forecasting based on a new climatic time series. *Environmental Modelling & Software*, 106:48–56, 2018.

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.

Charles Wheelus, Elias Bou-Harb, and Xingquan Zhu. Tackling class imbalance in cyber security datasets. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 229–232. IEEE, 2018.