2nd Workshop on Learning with Imbalanced Domains: Preface

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

Stan Matwin STAN@CS.DAL.CA
Faculty of Computer Science, Dalhousie University
Halifax, Canada

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

Bartosz Krawczyk BKRAWCZYK@VCU.EDU
Department of Computer Science, Virginia Commonwealth University
Richmond, VA 23284, USA

Nuno Moniz NMMONIZ@INESCTEC.PT
LIAAD-INESC TEC DCC-FCUP, University of Porto
Porto, Portugal

Paula Branco PAULA.BRANCO@DCC.FC.UP.PT
LIAAD-INESC TEC DCC-FCUP, University of Porto
Porto, Portugal

This volume contains the Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications - LIDTA2018. This Workshop was co-organised by the Faculty of Computer Science of the Dalhousie University, Halifax, Canada, the Department of Computer Science of the American University, Washington DC, USA, the Department of Computer Science of the Virginia Commonwealth University, Richmond VA, USA and the Laboratory of Artificial Intelligence and Decision Support - INESC TEC and Department of Computer Science, Faculty of Sciences of the University of Porto, Portugal. The Workshop was co-located with the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML/PKDD) 2018 and was held on the 10th of September 2018 in the Croke Park Conference Centre in Dublin, Ireland.

The LIDTA 2018 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 her/his 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., 2015), association rules mining (Luna et al., 2015), multi-instance learning (Vluymans et al., 2016), data streams (Krawczyk et al., 2017) and time series forecasting (Moniz et al., 2017). 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. Recent survey papers (Branco et al., 2016b; Krawczyk, 2016) have presented the new trends and open challenges of the problem of learning under imbalanced domains.

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 2018 edition of LIDTA was organized as a combined one day workshop and tutorial. The main goal of LIDTA 2018 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 tutorial component was designed to target researchers and professionals who have a recent interest on the subject, but also those who have previous knowledge and experience concerning this problem. The workshop part received inter-disciplinary contributions to tackle the problems that many real-world domains face nowadays. Both the tutorial and the workshop had a very high attendance, clearly reflecting the interest of the topic. The tutorial took place in the morning while the afternoon was reserved for the workshop.

In the tutorial, several important issues related to the problem of learning with imbalanced domains were discussed. The main goals were twofold: i) to provide a fast introduction to this problem presenting the main developments achieved by the research community, and ii) to present the most recent trends and topics that researchers wanting to address this problem may target. These two objectives allowed to reach a diverse audience that included researchers with less experience in this problem but also researchers with a higher expertise in this domain that could discuss the new trends and open problems when learning from imbalanced domains. The tutorial was organized into five slots covering the following topics: fundamentals of imbalanced domains learning; strategies for imbalanced domains learning; imbalanced regression; evaluation and pitfalls - case studies; and imbalanced time series and challenges. The tutorial was presented by Luís Torgo, Paula Branco and Nuno Moniz.

The first part of the tutorial was presented by Luís Torgo. After welcoming the audience, the presentation started with a brief introduction and motivation to the problem of learning from imbalanced domains, followed by the description of the fundamental properties of the problem. The relationship with other predictive tasks was debated and the main challenges of imbalanced domains were summarized. Then, the main types of strategies for dealing with imbalanced domains were presented. Paula Branco discussed the imbalanced regression problem showing how this problem can be formalized into a framework that is applicable to both classification and regression problems. Several solutions for the imbalanced regression problem were presented. This part of the tutorial involved practical illustrations of several solutions with the help of `UBL` R package (Branco et al., 2016a), an open source software tool freely available to the research community. The problem of performance assessment under imbalanced domains was discussed. Suitable evaluation measures were presented with the help of case studies. Finally, Nuno Moniz discussed the problem of imbalanced domains in the context of time series. Several solutions were presented that explored the use of the different dependencies of the data. These solutions were illustrated with the help of the open source R package `rewind`[1]. Overall, this was a very dynamic and interactive tutorial with a large participation of the audience.

The Workshop component of LIDTA 2018 was held in the afternoon. This part included a keynote talk followed by the presentation and discussion of the accepted papers. The invited talk was entitled "Novelty Detection: Beyond one-class Classification", by Professor João Gama, from the University of Porto. This talk was followed by several interesting questions. 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 11 paper submissions for LIDTA 2018, out of which 8 were accepted for inclusion in the workshop proceedings. All the accepted papers were assigned a presentation slot, together with time for questions and answers. These papers cover different aspects of imbalanced learning. Let us now briefly describe the accepted papers.

Bekker and Davis (2018) proposed the incorporation of the Selected At Random (SAR) assumption in the problem of learning from Positive and Unlabeled data. The learning problem is then solved through an Expectation-Maximization approach. Lango et al. (2018) presented a new method for tackling the class imbalance problem using local and neighbourhood information. The proposed algorithm, LmWeights, weights the training examples according to their local difficulty (safety) and the vicinity of larger minority clusters (gravity). A non-linear version of gradient boosting method was proposed by Frery et al. (2018) for dealing with two-class imbalanced tasks. This method improves the performance of the minority class cases by favouring the minority class instances faster than the standard gradient boosting and it also alleviates the over-fitting problem by reducing the model complexity. In Ksieniewicz (2018) a new hybrid method, Undersampled Majority Class Ensemble (UMCE) was presented. This method starts by dividing the majority class into k randomly created folds which are then paired with the minority class cases to create k balanced problems. The predictions of the ensemble are obtained using the k balanced problems and different fusing mechanisms. A new extension of bagging for imbalanced regression problems was provided by Branco et al. (2018). The proposed REBAGG algorithm is an

1. `rewind` R package is available at https://github.com/nunompmoniz/rewind.

adaptation of the bagging method that incorporates different data pre-processing methods for tackling the imbalanced regression problem. A multi-label k Nearest Neighbor that uses a self-adjusting memory for drifting data streams was proposed by Roseberry and Cano (2018). The presented algorithm exploits both short and long-term memories in order to predict the data streams' current and evolving states. Maletzke et al. (2018) proposed an approach for concept drift detection in the feature space that is insensitive to the class imbalance problem. This method also provides an estimate for the class distribution, that is used to dynamically adjust the classifiers' decision threshold. Finally, the work of Hepburn et al. (2018) addresses the problem of learning with example dependent costs. The presented solution explores the use of linear combinations of proper losses.

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 2018 Workshop and Tutorial Chairs for their support in the logistics of this workshop.

Organizing Committee

- Luís Torgo (Faculty of Computer Science, Dalhousie University)
- Stan Matwin (Faculty of Computer Science, Dalhousie University)
- Nathalie Japkowicz (Department of Computer Science, American University)
- Bartosz Krawczyk (Department of Computer Science, Virginia Commonwealth University)
- Nuno Moniz (Department of Computer Science, Faculty of Sciences, University of Porto; LIAAD - INESC TEC)
- Paula Branco (Department of Computer Science, Faculty of Sciences, University of Porto; LIAAD - INESC TEC)

Program Committee

- Roberto Alejo (Tecnologico de Estudios Superiores de Jocotitlan)
- Gustavo Batista (Universidade de São Paulo)
- Colin Bellinger (University of Alberta)
- Seppe Brouk (KU Leuven)
- Alberto Cano (Virginia Commonwealth University)
- Inês Dutra (CRACS - INESC TEC and Faculdade de Ciências, Universidade do Porto)
- Tom Fawcett (Apple)
- Mikel Galar (Universidad Pública de Navarra)

- Salvador García (Granada University)
- Francisco Herrera (Granada University)
- Jose Hernandez-Orallo (Universitat Politecnica de Valencia)
- Ronaldo Prati (Universidade Federal do ABC - UFABC)
- Rita P. Ribeiro (FCUP / LIAAD INESC TEC, University of Porto)
- José Antonio Saez (University of Salamanca)
- Shengli Victor Sheng (University of Central Arkansas)
- Marina Sokolova (Faculty of Medicine, University of Ottawa and Institute for Big Data Analytics)
- Jerzy Stefanowski (Poznan University of Technology)
- Isaac Velázquez (University Of Nottingham)
- Anibal R. Figueiras-Vidal (Universidad Carlos III de Madrid)
- Shuo Wang (University of Birmingham)
- Michal Wozniak (Wroclaw University of Science and Technology)

References

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data under the selected at random assumption. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2018)*, volume 94 of *Proceedings of Machine Learning Research*, pages 8–22, ECML-PKDD, Dublin, Ireland, 2018.

Paula Branco, Rita P. Ribeiro, and Luis Torgo. UBL: an R package for utility-based learning. *CoRR*, abs/1604.08079, 2016a.

Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):31, 2016b.

Paula Branco, Luis Torgo, and Rita P Ribeiro. REBAGG: REsampled BAGGing for imbalanced regression. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2018)*, volume 94 of *Proceedings of Machine Learning Research*, pages 67–81, ECML-PKDD, Dublin, Ireland, 2018.

Jordan Frery, Amaury Habrard, Marc Sebban, and Liyun He-Guelton. Non-linear gradient boosting for class-imbalance learning. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2018)*, volume 94 of *Proceedings of Machine Learning Research*, pages 38–51, ECML-PKDD, Dublin, Ireland, 2018.

Alexander Hepburn, Ryan McConville, Raúl Santos-Rodríguezo, Jesús Cid-Sueiro, and Dario García-García. Proper losses for learning with example-dependent costs. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2018)*, volume 94 of *Proceedings of Machine Learning Research*, pages 52–66, ECML-PKDD, Dublin, Ireland, 2018.

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.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

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.

Pawel Ksieniewicz. Undersampled majority class ensemble for highly imbalanced binary classification. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2018)*, volume 94 of *Proceedings of Machine Learning Research*, pages 82–94, ECML-PKDD, Dublin, Ireland, 2018.

Mateusz Lango, Dariusz Brzezinski, and Jerzy Stefanowski. Imweights: Classifying imbalanced data using local and neighborhood information. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2018)*, volume 94 of *Proceedings of Machine Learning Research*, pages 95–109, ECML-PKDD, Dublin, Ireland, 2018.

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.

André Maletzke, Denis dos Reis, Everton Cherman, and Gustavo Batista. On the need of class ratio insensitive drift tests for data streams. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2018)*, volume 94 of *Proceedings of Machine Learning Research*, pages 110–124, ECML-PKDD, Dublin, Ireland, 2018.

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.

Martha Roseberry and Alberto Cano. Multi-label knn classifier with self adjusting memory for drifting data streams. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2018)*, volume 94 of *Proceedings of Machine Learning Research*, pages 13–37, ECML-PKDD, Dublin, Ireland, 2018.

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

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.

Min-Ling Zhang, Yu-Kun Li, and Xu-Ying Liu. Towards class-imbalance aware multi-label learning. In *IJCAI*, pages 4041–4047, 2015.