Preface

Luís Torgo LTORGO@DCC.FC.UP.PT
LIAAD-INESC TEC DCC-FCUP, University of Porto
Porto, Portugal

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

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

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

This volume contains the Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications - LIDTA2017. This Workshop is co-organised by the Laboratory of Artificial Intelligence and Decision Support - INESC TEC and Department of Computer Science, Faculty of Sciences, University of Porto, Portugal and the Department of Computer Science, Virginia Commonwealth University, Richmond VA, USA. The Workshop is co-located with the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML/PKDD) 2017 and is held on the 22nd of September 2017 in the Hotel Aleksandar Palace, in Skopje, Macedonia.

Many real-world data-mining applications involve obtaining and evaluating predictive models using data sets with strongly imbalanced distributions of the target variable. Frequently, the least-common values of this target variable are associated with events that are highly relevant for end users (e.g. unusual returns on stock markets, diagnosis of rare diseases, intrusion detection, anticipation of catastrophes, crime prevention, popularity prediction in social media, etc.). This problem has been thoroughly studied in the last decade with a specific focus on classification tasks. However, the research community has started to address this problem within other contexts such as regression (Torgo et al., 2013), ordinal classification (Pérez-Ortiz et al., 2014), multi-label classification (Charte et al., 2015), association rules mining (Luna et al., 2015), multi-instance learning (Wang et al., 2013), data streams (Krawczyk et al., 2017) and time series forecasting (Moniz et al., 2017a). It is now recognized that imbalanced domains are a broader and important problem posing relevant challenges for both supervised and unsupervised learning tasks, in an increasing number of real world applications. Evidence of the new trends and challenges regarding this problem have been put forward in recent works (Branco et al., 2016; Krawczyk, 2016).

Tackling the issues raised by imbalanced domains is crucial to both academia and industry. To researchers, it is an opportunity to develop more adaptable and robust sys-

tems/approaches for very complex tasks. These tasks are in fact those that, in many cases, industry is already facing today. These are very diverse and include the ability to prevent fraud, to anticipate catastrophes, and in general to enable a more preemptive action in an increasingly fast-paced world.

This workshop received a diversity of high quality inter-disciplinary contributions discussing various aspects of learning from imbalanced domains. Overall, there were 17 paper submissions for LIDTA2017, out of which 13 papers were accepted for inclusion in workshop proceedings: 8 full papers and 5 posters. The full papers cover different aspects of learning from imbalanced domains. Namely, these contributions address pre-processing in classification and regression tasks, multi-label classification, one-class learners, unsupervised domain adaptation of Part-of-Speech taggers, and real-world applications. Let us briefly describe the accepted full papers. In Skryjomski and Krawczyk (2017) a study regarding the impact of the different types of minority class instances on SMOTE pre-processing algorithm is presented. A new method for stratification of multi-label data is proposed by Szymański and Kajdanowicz (2017). In Branco et al. (2017) a pre-processing strategy for regression tasks that combines two approaches based on data properties is introduced. A new stacking method for improving MLkNN method for multi-label classification is proposed by Pakrashi and Namee (2017). Bellinger et al. (2017) provides practical guidelines regarding when to use binary classification with some corrective measures and when to use one-class classifiers. Another work Zhu et al. (2017) explores the impact of class ratio used in the training sets on the performance of resampling-based ensembles in the context of churn prediction data. The combination of a convolutional neural network with additional static features is explored by Günnemann and Pfeffer (2017) for predicting engines damage using noise signals related with the engine internal excitation. The work of Cui et al. (2017) explores the effect of imbalanced domains in the problem of unsupervised domain adaptation of part-of-speech taggers.

Concerning the accepted posters, these contributions approach different topics on imbalanced learning, spanning from post-processing methods, new ensembles, loss functions and real-world applications. In Krasanakis et al. (2017) a framework for tuning plug-in rules while reducing the posterior certainty loss in imbalanced classification problems is being described. An evaluation of the performance of ensemble learners in imbalanced regression tasks is presented by Moniz et al. (2017b), studying the impact of characteristics such as data set size and imbalance ratio, in the performance of such tasks. A bilinear and log-bilinear loss functions which are used for controlling the error location in deep learning models are introduced in Resheff et al. (2017). The work of Fayet et al. (2017) compares three different unsupervised anomaly detection methods using different feature sets in the context of speakers profiles. Finally, Ksieniewicz and Woźniak (2017) present the notion of exposer, a tool for visualizing the data distribution. This was used as a cornerstone for a new ensemble method which is composed by a set of exposers generated on different feature subsets.

The workshop included a talk by Professor Nitesh Chawla, from the University of Notre Dame, entitled "Marking the 15-year anniversary of SMOTE: Origin, Progress and Opportunities" and a discussion table where the future challenges of imbalanced domain learning were discussed.

All of 8 accepted full papers were assigned a presentation slot, together with time for questions and answers. For authors of papers accepted as posters, we offered a lighting presentation to attract the interest of participants, and then a possibility for a more in-depth discussion during coffee breaks and lunch time.

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 2017 Workshop and Tutorial Chairs for their support in the logistics of this workshop.

Organizing Committee

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

Program Committee

- Ronaldo Prati (Universidade Federal do ABC - UFABC)
- Wojtek Kowalczyk (Leiden University)
- Colin Bellinger (University of Alberta)
- Vítor Cerqueira (LIAAD - INESC TEC)
- Isaac Velázquez (University Of Nottingham)
- Roberto Alejo (Tecnologico de Estudios Superiores de Jocotitlan)
- Inês Dutra (CRACS - INESC TEC and Faculdade de Ciências, Universidade do Porto)
- Mikel Galar (Universidad Pública de Navarra)
- Seppe Brouk (KU Leuven)
- Thomas Bäck (Leiden University)
- Alberto Cano (Virginia Commonwealth University)
- Rita P. Ribeiro (FCUP / LIAAD INESC TEC, University of Porto)
- Michal Wozniak (Wroclaw University of Science and Technology)
- Marina Sokolova (Faculty of Medicine, University of Ottawa and Institute for Big Data Analytics)

References

Colin Bellinger, Shiven Sharma, Osmar R. Zaïane, and Nathalie Japkowicz. Sampling a longer life: *Binary versus One-class classification Revisited* . In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 64–78, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

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

Paula Branco, Luís Torgo, and Rita P. Ribeiro. SMOGN: a pre-processing approach for imbalanced regression. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 36–50, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397, 2015.

Xia Cui, Frans Coenen, and Danushka Bollegala. Effect of data imbalance on unsupervised domain adaptation of part-of-speech tagging and pivot selection strategies. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 103–115, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

Cedric Fayet, Arnaud Delhay, Damien Lolive, and Pierre-François Marteau. Unsupervised classification of speaker profiles as a point anomaly detection task. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 152–163, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

Nikou Günnemann and Jürgen Pfeffer. Predicting defective engines using convolutional neural networks on temporal vibration signals. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 92–102, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Tunable plug-in rules with reduced posterior certainty loss in imbalanced datasets. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 116–128, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

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. URL http://www.sciencedirect.com/science/article/pii/S1566253516302329.

Paweł Ksieniewicz and Michał Woźniak. Dealing with the task of imbalanced, multidimensional data classification using ensembles of *exposers*. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 164–175, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

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*, pages 1–21, 2017a.

Nuno Moniz, Paula Branco, and Luís Torgo. Evaluation of ensemble methods in imbalanced regression tasks. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 129–140, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017b. PMLR.

Arjun Pakrashi and Brian Mac Namee. Stacked-mlknn: A stacking based improvement to multi-label k-nearest neighbours. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 51–63, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

María Pérez-Ortiz, Pedro Antonio Gutiérrez, and César Hervás-Martínez. Projection-based ensemble learning for ordinal regression. *Cybernetics, IEEE Transactions on*, 44(5):681–694, 2014.

Yehezkel S. Resheff, Amit Mandelbom, and Daphna Weinshall. Controlling imbalanced error in deep learning with the log bilinear loss. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 141–151, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

Przemysław Skryjomski and Bartosz Krawczyk. Influence of minority class instance types on smote imbalanced data oversampling. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 7–21, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

Piotr Szymański and Tomasz Kajdanowicz. A network perspective on stratification of multi-label data. In *Proceedings of the First International Workshop on Learning with*

Imbalanced Domains: Theory and Applications (LIDTA2017), volume 74 of *Proceedings of Machine Learning Research*, pages 22–35, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.

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

Xiaoguang Wang, Stan Matwin, Nathalie Japkowicz, and Xuan Liu. Cost-sensitive boosting algorithms for imbalanced multi-instance datasets. In *Advances in Artificial Intelligence*, pages 174–186. Springer, 2013.

Bing Zhu, Seppe vanden Broucke, Bart Baesens, and Sebastián Maldonado. Improving resampling-based ensemble in churn prediction. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA2017)*, volume 74 of *Proceedings of Machine Learning Research*, pages 79–91, ECML-PKDD, Skopje, Macedonia, 18–22 Sept 2017. PMLR.