School of Mathematics and Statistics
Faculty of Science, Technology, Engineering and Mathematics

The Open University

2024 PhD Projects

Project title	Resampling Methods for Supervised Learning from Extremely Class-imbalanced Data
Principal supervisor	Daniel Berrar
Discipline	Machine Learning
Research area/keywords	machine learning, data resampling, supervised learning, estimation
Suitable for	Full time applicants, Part time applicants

Project background and description

Data resampling plays a pivotal role in machine learning where it is used to estimate the variability of statistical estimators, to tune model hyperparameters, to prevent overfitting, and to estimate the generalization error of predictive models. There exist various resampling methods; however, when the data are extremely class-imbalanced, existing methods are not sufficient. A labeled data set is considered "extremely imbalanced" if at least one class is represented by fewer than 0.1% of all cases. Such imbalanced data sets are not at all uncommon in the real world; in fact, they are the norm in countless domains. Supervised learning from such data is very challenging: when the available data set is split into subsets for fitting the model (training sets) and subsets for evaluating the trained model (validation or test sets), the subsets contain too few cases of the minority classes for effective model training and evaluation. The goal of this project is (*i*) to investigate existing data resampling methods, such as k-fold cross-validation and bootstrapping, for the estimation of the generalization error of supervised learning algorithms when the data are extremely class-imbalanced; and (*ii*) to develop and evaluate (both analytically and empirically) new approaches to data resampling. The new resampling methods will be evaluated based on synthetically generated data and real-world data sets. This project requires excellent programming skills (in either R or Python, ideally both), a solid knowledge of statistics, and practical experience in developing machine learning models.

Background reading/references

1. D. Berrar. Cross-validation. In S. Ranganathan, K. Nakai, C. Schönbach, and M. Gribskov, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 542–545. Elsevier, 2018.
2. D. Berrar. Introduction to the non-parametric bootstrap. In S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 766–773. Elsevier, 2018.
3. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on AI (IJCAI), pages 1137–1143, 1995.
4. J. Wainer and G. Cawley. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, 182:115222, 2021.