Cost-sensitive learning for rare subtype classfication of lung cancer

*Nehir Kızılilsoley, Ezgi Tanıl, Emrah Nikerel**

*Yeditepe University, Department of Genetics and Bioengineering, Istanbul, Turkey
(Tel: +90 216 578 06 18; e-mail: emrah.nikerel@yeditepe.edu.tr)

Machine learning (ML) algorithms assume or promote that the training set is balanced among classes. For imbalanced datasets, even though the overall accuracy is high, the classical machine learning algorithms bias toward the majority class, causing the model fit poorly to the minority class [1,2] which hinders the use of these algorithms for classification of rare events. Strategies to overcome this problem including altering the training data directly to reduce the difference between classes or changing the learning procedure so that the algorithm takes also the minority class into account are proposed [2]. Usually, imbalance problem is handled with oversampling the minority or undersampling the majority class and/or generating synthetic samples from the original training data.

Gene expression data is highly valuable and popular data for cancer classification by ML. However, it is high-dimensional and severely imbalanced, making gene expression classification a cost-sensitive problem [1].

Cost-sensitive learning (CSL), uses imbalanced costs for classes while making predictions and is required when prediction of minority class is more "interesting" than the other class(es). Instead of maximizing the overall accuracy on all classes while assuming equal costs, the goal is to minimize cost (penalty of a misclassification) as classes are associated with different penalties for misclassification.

In this work, subtypes of lung cancer (AD, SC, LaC and SCLC) are classified using different CSL models that are either classical (e.g., support vector machines, naïve bayes, random forest) or ensemble learners, using imbalanced RNA-seq data from TCGA and microarray data from NCBI-GEO. Best performing model is evaluated by appropriate performance metrics (G-mean, accuracy, F-score etc.) and most important feature(s) will be extracted from this model using variable importance values.

References

[1] Lu, et al., *BMC Bioinformatics*, vol. 20, no. Suppl 25, pp. 1–10, 2019.
[2] Brownlee, "Imbalanced Classification with Python," *Mach. Learn. Mastery*, 2020.