Modelowanie danych Positive Unlabelled

Jan Mielniczuk

Instytut Podstaw informatyki, Polska Akademia Nauk

miel@ipipan.waw.pl, https://home.ipipan.waw.pl/j.mielniczuk/

Dane częściowo obserwowalne, w szczególności dane Positive Unlabelled (PU) są czesto spotykane i modelowane w analizie zbiorów tekstowych, medycynie i ekologii (por. np. Bekker i Davis (2020)). Jednakże konsekwencje i własności różnych sposobów modelowania często nie są szczególowo badane, co prowadzi do braku możliwości oceny jakości rozwiązań (por. Łazęcka et al (2021)). Celem projektu jest próba odpowiedzi na część istotnych pytań otwartych. Po pierwsze, własności oszacowań największej wiarogodności dla parametrycznych modeli PU przy założeniu warunku SCAR (Selected Completely at Random) zostaną przeanalizowane w sytuacji, gdy optymalizacja dokonywana jest przy użyciu metody Maximisation-Minimisation (MM) lub Convex-Concave (CC). Ponadto, konsekwencje złej specyfikacji modelu zostaną zbadane, gdy prawdziwy model dla prawdopodobieństwa aposteriori jest modelem pojedyńczego indeksu z nieznaną funkcją odpowiedzi. Po trzecie, projekt przewiduje analizę ogólnych danych PU, dla których nie zachodzi założenie SCAR i ich analizę przy wykorzystaniu modeli parametrycznych dla propensity score.
Kandydat powinien byc absolwentem studiów informatycznych lub matematycznych, być biegły w programowaniu, posiadać doświadczenie w stosowaniu metod uczenia maszynowego lub metod statystycznych i wykazywać silną motywację do rozwiązywania problemów obliczeniowych i analitycznych.

Refrencje

1. J. Bekker, J. Davis (2020) Learning from positive and unlabeled data: a survey. *Machine Learning*

2. P. Teisseyre, J. Mielniczuk, M. Łazęcka (2020) Different strategies of fitting logistic regression for positive and unlabelled data, *Proceedings of the International Conference on Computational Science ICCS'20*

3. M. Łazęcka, J. Mielniczuk, P. Teisseyre (2021) Estimating the class prior for positive and unlabelled data via logistic regression, *Advances in Data Analysis and Classification, w druku*