Propozycja tematu pracy dla doktoranta:
Metody uczenia maszynowego uwzględniające informacje o kosztach zmiennych

prof. dr Jan Mielniczuk
Instytut Podstaw Informatyki PAN
miel@ipipan.waw.pl

dr Paweł Teisseyre
Instytut Podstaw Informatyki PAN
teisseyrep@ipipan.waw.pl

Opis projektu:
Celem projektu jest opracowanie modeli uczenia maszynowego, które uwzględniają informacje o kosztach zmiennych. Współcześnie, w wielu zastosowaniach dysponujemy ogromną liczbą zmiennych i dlatego zwykle konieczne jest przeprowadzenie ich selekcji, która może być elementem budowy modelu lub być wykonana przed dopasowaniem właściwego modelu, jako wstępny krok. Istotną wadą istniejących rozwiązań jest założenie o tym że wszystkie zmienne mają ten sam koszt. Założenie to może być błędne ponieważ w pewnych sytuacjach pozyskanie wartości zmiennych jest związane z pewnym kosztem. W przypadku diagnostyki medycznej uzyskanie pewnych informacji jest łatwe (np. płeć pacjenta), natomiast wyniki testów diagnostycznych wiążą się z wykonaniem nieraz bardzo drogich badań. Pominięcie kosztów przy dopasowaniu modelu może oznaczać że nasz model, choć dokładny, stanie się bezużyteczny w praktyce, ponieważ wykonanie prognozy dla pacjenta będzie związane ze zbyt dużymi kosztami. W takich przypadkach lepiej znaleźć zbiór zmiennych umożliwiających akceptowalną dokładność modelu, przy znacznie mniejszych kosztach. Uwzględnienie kosztów w procesie uczenia to trudne zadanie, przede wszystkim dlatego że wymaga znalezienia kompromisu między użytecznością danej zmiennej a jej kosztem. Podczas projektu zostanie stworzona otwarta biblioteka, zawierająca implementację powyższych procedur. Przeprowadzimy również eksperymenty, w których zaproponowane metody zostaną porównane z istniejącymi algorytmami.

Oczekiwania:

1. Wykształcenie wyższe: informatyka lub(i) matematyka
2. Dobra znajomość uczenia maszynowego i statystyki
3. Programowanie: znajomość R, dodatkowym atutem będzie znajomość języków Python i Java
4. Dobra znajomość języka angielskiego
5. Entuzjazm w rozwiązywaniu problemów matematycznych i analitycznych

Literatura:

1. V. Bolon-Canedo et al., A framework for cost-based feature selection, Pattern Recognition, 2014.
2. Q. Zhou et al., Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features, Knowledge-based systems, 2016
3. P. Teisseyre et al, Cost-sensitive classifier chains: selecting low-cost features in multi-label classification, Pattern Recognition, 2019.