Nr rejestracyjny: 2016/23/B/ST1/00454; Kierownik projektu: dr hab. Małgorzata Bogdan

POPULARNO-NAUKOWE STRESZCZENIE PROJEKTU

Głównym celem projektu jest stworzenie nowych metod do analizy dużych zbiorów danych i określenie ich statystycznych własności. Dane o takim charakterze występują teraz w wielu dziedzinach nauki i przemysłu, a dobór odpowiedniej metody statystycznej do ich analizy jest niezwykle istotny z punktu widzenia efektywnego pozyskiwania informacji. Podstawowym problemem w analizie takich zbiorów jest ustalenie granicy oddzielającej efekt, który uznajemy za istotny, od losowego szumu. Jeżeli granica ta zostanie błędnie ustalona, wówczas badacze mogą dokonać wielu fałszywych odkryć, lub przeciwnie, nie zauważyć istotnych sygnałów. Błędy obu rodzajów redukują możliwości predykcyjne modeli stworzonych w oparciu o takie analizy i obniżają możliwe zyski. Na przykład w kontekście lokalizacji genów odpowiedzialnych za pewne cechy, fałszywe odkrycie prowadzi do zbędnych i często bardzo kosztownych biologicznych eksperymentów, zaś pominięcie istotnego genu redukuje szanse ustalenia właściwej terapii lub postawienia właściwej diagnozy.

W szczególności w naszych badaniach planujemy teoretycznie przebadać nową metodę optymalizacji wypukłej do analizy dużych zbiorw danych, SLOPE (Sorted L-One Penalized Estimation), ostatnio zaproponowaną przez autorkę projektu i współpracowników z Uniwersytetu Stanforda, oraz rozszerzyć zakres stosowalności tej metody. Metoda jest szybka obliczeniowo i spełnia precyzyjne statystyczne wymagania dotyczące kontroli frakcji fałszywych odkryć, równocześnie gwaratując relatywnie dużą moc identyfikacji rzeczywistych predykatorów. Teoretyczna analiza własności SLOPE umożliwi głębsze zrozumienie tej metody, utworzenie jej optymalnych wersji i ustalenie praktycznych granic jej stosowalności. Wśród wielu pytań na które szukamy odpowiedzi jest pytanie w jakich warunkach SLOPE umożliwia optymalną estymację parametrów w modelu i optymalną predykcję interesujących nas cech. Planujemy także opracowanie nowych Bayesowskich wersji SLOPE, które umożliwią wykorzystanie wstępnej wiedzy na temat badanego zjawiska a także ocenę stopnia niepewności uzyskanych estymatorów. Oprócz propozycji nowych wariantów SLOPE, dostosowanych do specyficznych zagadnień statystycznych, planujemy także opracowanie nowych kryteriów do porównania metod statystycznych w sytuacji gdy pojęcie prawdziwego odkrycia nie jest precyzyjnie określone. Takie kryteria pozwolą na ujednolicone porównania różnych metod i zapobiegną manipulacjom wynikami eksperymentów statystycznych.

Finalnym produktem naszych badań będzie zestaw ogólnodostępnych programów do analizy dużych zbiorów danych z implementacjami statystycznie uzasadnionych metod, które powstaną w trakcie realizacji tego projektu.