Metody selekcji cech w klasyfikacji wieloetykietowej

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

1. Opis projektu

Klasyfikacja wieloetykietowa (KW) jest bardzo dynamicznie rozwijającą się częścią uczenia maszynowego. W klasyfikacji wieloetykietowej (w odróżnieniu od tradycyjnej klasyfikacji) rozważa się wiele binarnych zmiennych decyzyjnych (etykiet) jednocześnie. Celem jest zbudowanie modelu, który umożliwia przewidywanie etykiet na podstawie cech obiektów. W ostatnich latach, problem KW wzbudził bardzo duże zainteresowanie w wielu dziedzinach, takich jak automatyczna anotacja obrazów, kategoryzacja tekstów, marketing, genomika, medycyna, projektowanie leków (Gibaja, E. and Ventura, S. (2015) „A tutorial on multilabel learning"; Zhang, M. and Zhou, Z. (2013) „A review on multi-label learning algorithms"). Istotnym zadaniem w KW jest selekcja cech, tzn. odkrycie które cechy mają wpływ na wartości etykiet. Celem projektu jest opracowanie i implementacja metod wyboru cech w KW, w sytuacji dużego wymiaru przestrzeni cech.

W ostatnich latach opracowano szereg metod umożliwiających predykcję dla wielu etykiet jednocześnie. Większość metod bazuje na wykorzystaniu zależności między etykietami. Brakuje jednak wyników (zarówno teoretycznych jak i empirycznych), które pokazują jaki jest wpływ wyboru cech na działanie klasyfikatorów. Celem projektu jest zbadanie tego obszaru tematycznego. Opracowane podejścia mogą przyczynić się do poprawienia istniejących metod w kilku aspektach. Po pierwsze pozwolą zwiększyć moc predykcyjną stosowanych metod. Po drugie, metody selekcji umożliwiają odkrycie skomplikowanych struktur zależności w danych, a co za tym idzie, lepsze zrozumienie tego, które zmienne wpływają na etykiety oraz jakie są zależności między etykietami. Wreszcie, redukcja wymiaru przestrzeni zmiennych i przestrzeni etykiet jest istotna ze względu na zmniejszenie czasu obliczeń. Ma to szczególne znaczenie w przypadku danych o ogromnym wymiarze (takich jak dane tekstowe czy genomiczne). Dodatkowo, podczas projektu zostanie stworzona otwarta biblioteka, zawierająca implementację powyższych procedur. Przeprowadzimy również eksperymenty, w których zaproponowane metody zostaną porównane z istniejącymi algorytmami.

2. Wymagania (oczekiwania)

a. Ukończone studia drugiego stopnia z informatyki, matematyki lub fizyki
b. Znajomość podstaw uczenia maszynowego i statystyki
c. Programowanie: znajomość R, dodatkowym atutem będzie znajomość języka Java
d. Entuzjazm w rozwiązywaniu problemów matematycznych i analitycznych
e. Dobra znajomość języka angielskiego