Nr rejestracyjny: 2016/21/D/ST6/00980; Kierownik projektu: dr Marek Śmieja

Dodatkowa informacja w grupowaniu danych i zagadnieniach pokrewnych

Grupowanie jest jedną z kluczowych technik nauczania maszynowego i analizy danych. Mimo, iż grupowanie może być stosowane bez żadnego nadzoru, to znaczy nie są wymagane przykłady uczące oznaczone przez eksperta, to jednak użycie dodatkowej wiedzy może wpłynąć znacząco na jego końcowy efekt. Przypomnijmy, że jeden z paradygmatów nauczania maszynowego mówi, że w procesie uczenia powinno uwzględnić się każdą dostępną informację o danych. Niniejszy projekt ma na celu rozwinięcie i zastosowanie tego paradygmatu w teorii grupowania danych, a także powiązanych zagadnieniach klasyfikacji, czy teorii reprezentacji.

W przeciągu ostatnich lat nastąpił gwałtowny rozwój technik częściowo nadzorowanych (korzystających z wiedzy dodatkowej) w grupowaniu danych oraz klasyfikacji. Niestety istnieje niewiele metod grupowania, które potrafią wykorzystać dodatkową informację do wyboru odpowiedniego skomplikowania modelu. Innymi słowy, opracowane dotychczas metody bazują na bezpośredniej adaptacji istniejących algorytmów grupowania do sytuacji częściowo nadzorowanej. W konsekwencji wynikowa grupa jest opisywana prostym modelem, nawet jeśli jest to sprzeczne z dodatkową informacją. Wyróżniającym elementem projektu, będzie wykorzystanie wiedzy dodatkowej do automatycznego wyboru odpowiedniej klasy modelu grupowania, co jest podejściem nowatorskim w tej dziedzinie.

Klasyczny sposób specyfikacji informacji dodatkowej polega na zdefiniowaniu etykiet wybranych elementów (częściowe etykietowanie) bądź więzów równoważności, mówiących czy zadana para elementów powinna zostać zgrupowana razem czy oddzielnie. Niestety w rzeczywistych problemach niejednokrotnie dodatkowa informacja jest zadana w sposób odmienny od określenia przynależności wybranych elementów, co uniemożliwia bezpośrednie zastosowanie klasycznych metod częściowo nadzorowanych. Dla przykładu, w cheminformatyce aktywność związku, którą chcemy przewidzieć, jest zadana za pomocą niedokładnego pomiaru, co nie przekłada się na przynależność związku do klasy. Ponadto, w badaniach medycznych oprócz danych o pacjentach zdrowych i chorych mamy również informacje o tych z pośrednią diagnozą. Zgodnie z paradygmatem przytoczonym we wstępie, użycie wszelkiej informacji o danych może wpłynąć na ich lepszą analizę. Te oraz inne przykłady występujące w rzeczywistych sytuacjach pokazują że istnieje niewystarczająca ilość opracowanych dotychczas narzędzi pracy z danymi opatrzonymi informacją dodatkową. Bazując na istniejących problemach, zdefiniujemy nowe modele grupowania oraz klasyfikacji, które będą mogły wypełnić lukę w dziedzinie uczenia częściowo nadzorowanego.

Pozyskiwanie etykiet danych (bądź definiowanie więzów równoważności) wiąże się w praktyce ze wzrostem kosztów przeprowadzanej analizy. Dlatego w wielu przypadkach dysponujemy jedynie niewielką ilością przykładów uczących. Istniejące metody grupowania często nie potrafią wykorzystać tak niewielkiej ilości wiedzy dodatkowej w odpowiedni sposób i tworzą podział podobny jak w przypadku nienadzorowanym. W ramach tego projektu podejmiemy się opracowania modeli grupowania, które będą mogły być kierowane niewielką ilością poetykietowanych elementów, bądź więzów równoważności.

O ile motywacja naszych badań wynika w wielu przypadkach z typowych problemów praktycznych, to projekt nie ogranicza się w żaden sposób do jednej tylko dziedziny. Będziemy dążyć do stworzenia teoretycznych podstaw tworzonych narzędzi oraz przetestowania ich w różnych praktycznych problemach. Mamy nadzieją, że przełoży się to na uzyskanie interesujących wyników w informatyce oraz rozwiązanie rzeczywistych problemów występujących w innych dziedzinach nauki.