MODELOWANIE „DŁUGICH" ZBIORÓW DANYCH Z PRZYKŁADAMI ZASTOSOWAŃ

Magdalena TOPCZEWSKA[1], Tomasz ŁUKASZUK[1], Leon BOBROWSKI[1,2]

[1]Politechnika Białostocka, Wydział Informatyki
[2]Instytut Biocybernetyki i Inżynierii Biomedycznej PAN, Warszawa

Analiza dyskryminacyjna jest jedną z metod eksploracyjnej analizy danych, służącą budowaniu obszarów decyzyjnych pozwalających przydzielać obiekty do odpowiednich klas. W przypadku, gdy nie są znane warunkowe gęstości prawdopodobieństwa w klasach możliwe jest ich szacowanie za pomocą metod nieparametrycznych bądź założenie dotyczące ogólnej klasy modelu klasyfikatora. W pracy założoną postacią funkcji dyskryminacyjnej jest hiperpłaszczyzna. Rozpatrywane są przypadki dwóch i większej ilości klas.

Dane wejściowe reprezentowane są w postaci wektorów cech $x_j[n]=[x_{j1},\ldots, x_{jn}]^T$ (j=1,…, m). Wartości przyjmowane przez cechy mogą być mieszane - część cech może być binarnych ($x_i \in \{0,1\}$), pozostałe mogą być liczbami rzeczywistymi ($x_i \in R^1$). Modelując dane wejściowe, uzyskujemy na wyjściu informację, do której klasy należy przyporządkować dany obiekt. Przyporządkowanie do poszczególnych klas uzyskujemy poprzez podział przestrzeni na obszary decyzyjne za pomocą hiperpłaszczyzn, których parametry znajdowane są poprzez minimalizację wypukłych i odcinkowo liniowych funkcji kryterialnych typu CPL [1].

Do minimalizacji rozpatrywanej funkcji kryterialnej w zadanej przestrzeni cech stosujemy algorytm wymiany rozwiązań bazowych. Rozważamy ponadto dwie strategie poszukiwań hiperpłaszczyzny rozdzielającej zbiory. Podstawowa strategia - *Sekwem* operuje na pełnej przestrzeni cech jednokrotnie stosując algorytm wymiany rozwiązań bazowych. Strategia *Genet* bazuje na stopniowym powiększaniu przestrzeni cech aż do osiągnięcia minimum funkcji kryterialnej. Wynikiem jest otrzymanie parametrów opisujących hiperpłaszczyznę w znacznie zredukowanej przestrzeni cech. Dzięki temu uzyskuje się poprawę efektywności w przypadku wysokowymiarowych zbiorów danych. Możliwe jest również wyodrębnienie cech najbardziej istotnych w procesie dyskryminacyjnym.

W pracy zostaną przedstawione wyniki eksperymentów wykonanych na syntetycznych i rzeczywistych zbiorach danych porównujące działanie klasyfikatorów opartych na minimalizacji funkcji kryterialnych oraz ich zmodyfikowanych wersjach. Porównane zostaną także strategie poszukiwań minimów funkcji kryterialnych zarówno pod względem efektywności numerycznej jak również jakości klasyfikacji.

Bibliografia

1. Bobrowski L., Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych (Data mining based on convex and piecewise linear (CPL) criterion functions) (in Polish), Białystok Technical Univeristy, 2005.
2. Bobrowski L., Łukaszuk T., *Selection of the linearly separable feature subsets*, The 7th International Conference on Artificial Intelligence and Soft Computing (ICAISC), Zakopane Poland, June 2004, Eds. L. Rutkowski, J. Siekemann, R. Tadeusiewicz, L. Zadeh, Lecture Notes in AI Vol. 3070, pp. 544-549.

Praca finansowana w ramach W/WI/1/05.