

Dr hab. Krzysztof Diks, prof. UW
Instytut Informatyki
Uniwersytet Warszawski
ul. Banacha 2
02-097 Warszawa
e-mail: diks@mimuw.edu.pl

Warszawa, 30.03.2006

Recenzja rozprawy doktorskiej
pani mgr inż. Doroty Cendrowskiej
p.t. *Konstrukcja klasyfikatora obiektów
z wykorzystaniem algorytmu badania
rozdzielności liniowej dwóch zbiorów*

Problem klasyfikowania obiektów jest jednym z podstawowych problemów pojawiających się w analizie dużej ilości danych (obiektów). Zazwyczaj celem takiej analizy jest podział zbioru danych na podzbiory o zbliżonych własnościach oraz znalezienie takiej "algorytmicznej" charakterystyki tych podzbiorów, która pozwoli na automatyczne przypisywanie (klasyfikowanie) nowych obiektów do właściwych podzbiorów. Jedno z podejść do rozwiązania tego zagadnienia polega na potraktowaniu badanych obiektów jako punktów (wektorów) w wielowymiarowej przestrzeni euklidesowej, a następnie podzielenie tej przestrzeni na rozłączne obszary zawierające tylko obiekty o podobnych charakterystykach. Najprostszy klasyfikator jest liniowym klasyfikatorem dwu-decyzyjnym, co w podejściu opisanym powyżej odpowiada wyznaczeniu hiperpłaszczyzny dzielącej dwukolorowy zbiór punktów na dwa zbiory punktów o takich samych kolorach. Niestety nie zawsze taka hiperpłaszczyzna istnieje. Wówczas poszukuje się hierarchii hiperpłaszczyzn, które dzielą przestrzeń na obszary zawierające tylko punkty o tych samych kolorach. Wyznaczona hierarchia hiperpłaszczyzn powinna także umożliwiać łatwe i efektywne klasyfikowanie nowych obiektów do odpowiednich obszarów, a tym samym zaliczanie ich do zbioru punktów tego samego koloru, charakteryzujących się zbliżonymi własnościami. Klasyfikatory dwu-decyzyjne można następnie wykorzystywać do konstrukcji klasyfikatorów wielo-decyzyjnych. Omówione powyżej zagadnienia są właśnie przedmiotem rozprawy doktorskiej pan mgr inż. Doroty Cendrowskiej.

Rozprawa składa się z wprowadzenia, pięciu rozdziałów, podsumowania oraz dwóch dodatków.

We wprowadzeniu autorka krótko omawia, co jest przedmiotem rozprawy i opisuje zawartość pozostałej części pracy.

Rozdział pierwszy rozpoczyna się od omówienia istoty samej klasyfikacji oraz ogólnych zasad budowy klasyfikatorów i oceny ich jakości. Autorka także uzasadnia przydatność algorytmu liniowego rozdzielania dwóch zbiorów w konstrukcji hierarchicznego klasyfikatora odcinkowo-liniowego. Podrozdział 1.3 zawiera kluczowe dla pracy definicje liniowej rozdzielności dwóch zbiorów oraz krótki przegląd algorytmów testowania rozdzielności liniowej, jak i znajdowania "świadectwa" tej rozdzielności, czyli hiperpłaszczyzny (lub hiperpłaszczyzn) rozdzielającej (rozdzielających). W sformułowaniu definicji 2 pojawia się pewna niezręczność. Z jednej strony żąda się, żeby $g^*(x) > 0$, dla x ze zbioru $X1$, oraz $g^*(x) < 0$, dla x ze zbioru $X2$, a z drugiej strony dopuszcza się możliwość $g^*(x) = 0$. W definicjach 4 i 5 pojawia się tajemnicze a_{n+2} . Autorka twierdzi, że najbardziej interesujące jest wyznaczenie

K. D.

świadczenia rozdzielnosci z najwiekszym przeswittem. Wymaga wyjasnienia dlaczego to jest istotne dla procesu klasyfikacji. Bardzo dobry uzupehleniem definicji sa rysunki. Jednak na rysunku 1.2b nalezy zamienic proste h_1 i h_2 .

W dalszej czesci podrozdzialu 1.3 autorka opisuje zgrubnie istniejace algorytmy badania rozdzielnosci liniowej podajac jednoczesnie ich wady i zalety. Niestety w szybkim zrozumieniu tej czesci pracy przeszkadza brak omowienia przyjetaj notacji. Na przyklad, co oznacza sumowanie punktow we wzorze na koncu strony 13. Autorka ma takze klopoty z interpunkcja, czasami przecinkow jest za duzo, a czasami nie ma ich wcale, jak na przyklad w pierwszym zdaniu na stronie 14.

Po przeprowadzeniu analizy roznych algorytmow autorka zdecydowala sie wybrac algorytm Jozwika jako podstawe konstrukcji hierarchicznego klasyfikatora binarnego. Jedna z waznych cech tego algorytmu jest to, ze dla poprawnosci jego dzialania nie jest wazne, zeby dane zbiory byly rozdzielne liniowo. Algorytm Jozwika potrafi wykrywac takie przypadki.

W rozdziale drugim autorka dokladnie przedstawia algorytm Jozwika i opisuje jego wlasnosci. Niestety i w tym rozdziale autorka nie ustrzeza sie uchybień notacyjnych utrudniajacych latwe rozumienie zaprezentowanych rozważań. Na przyklad w trzecim akapicie jest mowa o podzbiorach zbioru T , a stosuje sie notacje dla nadzbiorow. Pojawiaja sie tez potworki jezykowe w postaci "dzialanie implementacji algorytmu".

W podrozdziale 2.1 autorka opisuje dokladnie algorytm Jozwika i ilustruje go dobrze dobranym przykladem. Nadal razi nieprecyzyjne poslugiwanie sie notacja. Jesli dobrze rozumiem, w nierownosci 2.2 zbior T sklada sie ze zbioru punktow zdefiniowanych przez $f(x)$, a nie ze zbioru punktow x .

Wada algorytmu Jozwika jest to, ze nie wyznacza on zawsze hiperplaszczyny scisle rozdzielajacej, nawet w owczas, gdy taka hiperplaszczyna istnieje. Rozdzial 3 zawiera oryginalny pomysl autorki usuniecia podanej wyzej wady algorytmu Jozwika. Autorka proponuje algorytm, ktory znajduje pare hiperplaszczyn rozdzielajacych, jesli tylko istnieja, definiujacych maksymalny przeswit pomiedzy rozdzielanymi zbiorami punktow. Podstawa poprawnosci zaproponowanego algorytmu jest twierdzenie 1, a scisly dowod poprawnosci algorytmu znajduje sie w dodatku A.

W rozdziale 4 zbadano implementacyjne aspekty algorytmu zaproponowanego w pracy. Najkosztowniejszym obliczeniowo jest krok p_4 , w ktorym oblicza sie wektor prostopadly do kazdego wektora z pewnej listy, ktorego ostatnia skladowa (epsilon) ma byc maksymalna. Obliczenia w tym kroku sa przeprowadzane za pomoca odpowiednio dostosowanej eliminacji Gaussa. Autorka zaproponowala prostą heurystykę usprawniającą obliczenia w tym kroku.

Rozdzial 5 to propozycja konstrukcji klasyfikatora dwu-decyzyjnego. Jesli treningowy zbior obiektow jest rozdzielny liniowo, to w owczas wyznacza sie swiadka tej rozdzielnosci – swiadek sluzzy do klasyfikowania nowych obiektow. W przeciwnym razie dany zbior obiektow dzieli sie na trzy zbiory za pomoca dwuch rownoleglych hiperplaszczyn. Punkty z dwuch tych zbiorow sa juz dobrze rozdzielone, natomiast punkty w pasie "pomiedzy" hiperplaszczynami ulegaja dalszemu podzialowi. "Srodkowy" pas jest wyznaczany za pomoca algorytmu rozdzielania z maksymalnym przeswittem. Problemem jest wybor optymalnego pasa, lub jak to jest nazwane w pracy – optymalnej nakladki. Autorka proponuje stosowanie algorytmu obliczania maksymalnego przeswitu, az do momentu w ktorym jest to juz niemozliwe. W owczas stosuje sie ponownie ten algorytm do podzbioru, ktory nie zostal jeszcze poprawnie rozdzielony. W ten sposob cala przestrzen rozwiazań dzieli sie na czesci zawierajace tylko punkty o podobnych wlasnosciach. W dodatku 2 znajduje sie propozycja

20

