PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Mariusz Kubus
Politechnika Opolska

O PEWNEJ METODZIE DYSKRYMINACJI OBIEKTÓW JAKOŚCIOWYCH

1. Wstęp

Nieklasyczne i nieparametryczne metody analizy dyskryminacyjnej mają wiele zalet. Można je stosować, w przeciwieństwie do metod klasycznych (liniowe funkcje dyskryminacyjne, Bayesowska reguła klasyfikacji, metoda współrzędnych dyskryminacyjnych), do cech zarówno ilościowych, jak i jakościowych, które bardzo często są pozyskiwane w badaniach społeczno-ekonomicznych. Metody te nie wymagają znajomości rozkładów zmiennych ani postaci związku między nimi. Mają charakter adaptacyjny, tzn. dobór zmiennych odbywa się jednocześnie z budowaniem modelu, tak aby poprawić jego jakość. Wykazują także odporność na braki danych oraz na obserwacje nietypowe. Generowane wyniki – charakterystyki klas – są łatwe w interpretacji, gdyż wykorzystują język zbliżony do naturalnego. Reguły klasyfikacji są postaci „jeśli *opis klasy*, to przypisz *klasę*", gdzie *opis klasy* jest koniunkcją wartości cech. Ponieważ w praktyce wiele informacji ankietowych jest nieostrych czy niepełnych, można zauważyć wzrost zainteresowania tymi metodami.

Niniejszy artykuł jest propozycją zastosowania drzew logicznych do konstrukcji reguł klasyfikacyjnych. Uzyskany na minimalnych drzewach logicznych porządek zmiennych jakościowych wytycza kierunek przeszukiwania przestrzeni charakterystyk klas. Wyniki dokładności klasyfikacji będą porównane z algorytmami CN2 Clarka, Nibletta oraz C4.5 Quinlana.

2. Nieparametryczne metody analizy dyskryminacyjnej

Metoda drzew klasyfikacyjnych jest metodą rekurencyjnego podziału (*recursive partitioning*) zbioru uczącego na coraz bardziej jednorodne względem zmiennej objaśnianej podzbiory. Podziału dokonuje się względem wybranej zmiennej, a podzbiory są wyzna-

czone przez wartości zmiennej. Zmienne ilościowe poddaje się dyskretyzacji. Zmienne oceniane są funkcją kryterium, którym jest różnica heterogeniczności podzbiorów przed podziałem i po nim. Szeroki przegląd stosowanych miar jakości podziału można znaleźć w pracy Gatnara [2001]. W algorytmie C4.5 Quinlana stosuje się w tym celu entropię:

$$J(S,x) = E(\mathbf{p}) - \frac{1}{n}\sum_{j=1}^{m} n_j E(\mathbf{p}_j), \tag{1}$$

gdzie: $J(S,x)$ – jakość podziału zbioru S, jakiego dokonuje zmienna x,
E – funkcja entropii (2),
$\mathbf{p}$ – wektor prawdopodobieństwa przynależności do klas w zbiorze S,
$\mathbf{p}_j$ – wektory prawdopodobieństw przynależności do klas w podzbiorach powstałych w wyniku podziału,
n_j – liczebności tych podzbiorów,
m – liczba wartości zmiennej x.

Miara $J(S, x)$ podlega maksymalizacji.

Oprócz drzew klasyfikacyjnych wyróżnić można grupę metod opartych na analizie obiektów. Najbardziej znane algorytmy tego nurtu to AQ [Michalski, Larson 1978] oraz CN2 [Clark, Niblett 1989]. W metodach tych proces uczenia polega na przeszukiwaniu zbioru obserwacji i łączeniu obiektów w podzbiory o dostatecznym stopniu homogeniczności. O ile AQ wymaga całkowitej homogeniczności podzbiorów, o tyle CN2 dopuszcza do pewnego stopnia heterogeniczność, przypominając swą ideą porządkowanie drzew klasyfikacyjnych (*tree pruning*). Ponieważ metody te opracowane są dla cech jakościowych, więc przeszukiwanie obiektów polega na analizie ich cech wspólnych. Charakterystyki klas wyznaczane są na podstawie wartości cech wspólnych w otrzymanych podzbiorach. Do oceny jakości podzbiorów obiektów CN2 wykorzystuje dwie funkcje. Entropia (2) ocenia stopień homogeniczności podzbiorów, a statystyka I (3) mierzy statystyczną istotność podzbioru, tzn. sprawdza, czy rozkład liczebności klas w ocenianym podzbiorze nie mógłby się pojawić losowo:

$$E = -\sum_{i=1}^{k} p_i \log_2 p_i, \tag{2}$$

$$I = 2N\sum_{i=1}^{k} p_i \log_2 \frac{p_i}{r_i}, \tag{3}$$

gdzie: p_i – prawdopodobieństwo występowania *i-tej* klasy w ocenianym podzbiorze,
k – liczba klas,
r_i – prawdopodobieństwo występowania *i-tej* klasy w zbiorze uczącym,
N – liczebność ocenianego podzbioru.

Późniejsza modyfikacja CN2 [Clark, Boswell 1991] zastąpiła dwie przedstawione wyżej miary jedną – funkcją Laplace'a (4):

$$L = \frac{n_i + 1}{N + k}, \tag{4}$$

gdzie: n_i – liczebność klasy opisywanej w ocenianym podzbiorze.

Autorzy otrzymali większą dokładność klasyfikacji.

3. Reguły klasyfikacji

Ponieważ metody przedstawione w artykule dotyczą obiektów symbolicznych, których cechy reprezentowane są przez zmienne mierzone na słabych skalach pomiaru (nominalnej i porządkowej), naturalnym sposobem przedstawiania charakterystyk klas jest koniunkcja wartości cech. Koniunkcją posługuje się język naturalny, gdy mówimy: „najchętniej kupowanymi samochodami są srebrne kombi". Oznacza to regułę klasyfikacji: jeśli obiekt samochód ma cechy: kolor srebrny i nadwozie typu kombi, to przypisuje się go do klasy NAJCHĘTNIEJ KUPOWANE, co można zapisać formalnie:

[kolor = srebrny] $\wedge$ [nadwozie = kombi] ::> NAJCHĘTNIEJ KUPOWANE.

Do opisu obiektów symbolicznych oraz charakterystyk klas wykorzystuje się logikę wielowartościową Michalskiego VL_1 (*variable-valued logic* [Michalski 1977]). Wyrażenia koniunkcyjne nazywane są w tej logice kompleksami. Kompleks może opisywać pojedynczy obiekt (np. C_1 opisuje obiekt w czterowymiarowej przestrzeni cech) lub zbiór obiektów (np. C_2 opisuje wszystkie białe samochody z silnikiem 1.8, niezależnie od wartości pozostałych cech).

C_1: [silnik = 1.8] $\wedge$ [kolor = biały] $\wedge$ [nadwozie = kombi] $\wedge$ [ABS = tak].

C_2: [silnik = 1.8] $\wedge$ [kolor = biały].

Mówi się, że kompleks C_2 jest ogólniejszy od kompleksu C_1, ponieważ zbiór obiektów opisanych przez C_2 zawiera w sobie zbiór obiektów opisanych przez C_1. Przykład kompleksu C_3 pokazuje, że w składowych koniunkcji (tzw. selektorach) mogą występować różne symbole relacji oraz wewnętrzne alternatywy wartości cech.

C_3: [silnik > 1.7] $\wedge$ [kolor $\neq$ czarny] $\wedge$ [nadwozie = kombi $\vee$ hatchback] $\wedge$ [ABS = tak].

Charakterystyka klasy w postaci alternatywy kompleksów nazywana jest dysjunkcyjnym opisem klasy:

$C_1 \vee C_2 \vee ... \vee C_n$ ::> K.

Każdy otrzymany w algorytmach AQ czy CN2 podzbiór o akceptowanym stopniu homogeniczności opisany jest kompleksem, co ostatecznie daje dysjunkcję.

4. Metoda drzew logicznych

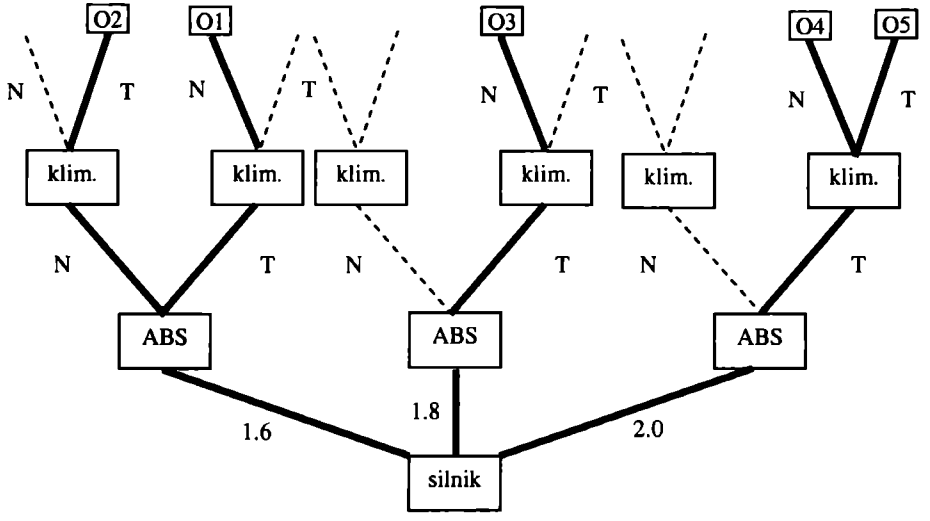
Drzewo logiczne jest graficzną reprezentacją skończonej przestrzeni cech jakościowych (rys.1). Podobnie jak na drzewach klasyfikacyjnych, w węzłach znajdują się testy wartości zmiennych, a krawędzie kodowane są wartościami zmiennych. Różnica polega na tym, że wszystkie węzły jednego poziomu testują tę samą zmienną. Ponadto liczba liści na drzewie logicznym jest równa liczbie obiektów, ponieważ ścieżki od korzenia do liści reprezentują współrzędne obiektów. Zmienne można porządkować dowolnie na po-

ziomach drzewa (oznacza to jedynie zamianę kolejności kolumn w macierzy danych), czego efektem jest różna liczba krawędzi (węzłów).

Tabela 1. Przykład macierzy danych jakościowych

	X_1	X_2	X_3
	silnik {1.6 ; 1.8 ; 2.0 }	ABS {NIE ; TAK }	klimatyzacja {NIE ; TAK }
O_1	1.6	TAK	NIE
O_2	1.6	NIE	TAK
O_3	1.8	TAK	NIE
O_4	2.0	TAK	NIE
O_5	2.0	TAK	TAK

Źródło: opracowanie własne.



Rys. 1. Drzewo logiczne ilustrujące zbiór obiektów z tab.1
Źródło: opracowanie własne.

Przyjmuje się, że minimalne drzewo logiczne ma najmniejszą liczbę krawędzi (węzłów). Minimalizacja polega zatem na znalezieniu odpowiedniego uporządkowania zmiennych na poziomach drzewa. W tym celu wykorzystuje się założenia heurystyczne [Kubus 2005].

Minimalne drzewo logiczne dla zbioru obiektów wybranej klasy wytycza kierunek przeszukiwania przestrzeni opisu tej klasy. Proponuje się następujący algorytm.

1. Znajdź minimalne drzewo logiczne dla zbioru obiektów klasy opisywanej.
2. Przedstaw na drzewie obiekty pozostałych klas.
3. Odczytaj maksymalnie ogólny dysjunkcyjny opis klasy z dopuszczalnym progiem heterogeniczności.
4. Otrzymany opis uprość, stosując równoważne przekształcenia logiczne.
5. Procedurę powtarzaj dla każdej klasy.

Krok trzeci oznacza przeszukiwanie węzłów drzewa (reprezentujących podzbiory), począwszy od korzenia (a więc od opisu ogólnego do szczegółowego), i ocenianie podzbiorów ze względu na homogeniczność. Do oceny wykorzystać można np. funkcję entropii. Przyjmując pewną wartość progową wybranej miary, tworzy się liście, a ścieżki łączące je z korzeniem są kompleksami opisującymi klasę.

Opisane w ten sposób podzbiory nie zawsze są rozłączne. Oznacza to niejednoznaczność na etapie klasyfikacji. Rozwiązaniem, jakie można zaproponować, jest uporządkowanie reguł według klas. Oznacza to, że jeśli klasyfikowany obiekt spełnia charakterystykę dwóch klas, to zostaje przypisany do wcześniejszej z nich na liście reguł. Kolejny problem powstaje w sytuacji, gdy obiekt nie spełnia żadnej reguły klasyfikacyjnej. Najprostsze rozwiązanie polega na dołączeniu reguły domyślnej, polegającej na klasyfikowaniu takiego obiektu do klasy najliczniej, reprezentowanej w zbiorze uczącym (takie rozwiązanie przyjęto np. w CN2).

5. Wyniki

Dokładność klasyfikacji porównana będzie na powszechnie uznawanych zbiorach danych, wykorzystywanych do testowania metod klasyfikacji wzorcowej. Charakterystykę zbadanych zbiorów przedstawia tab. 2.

Tabela 2. Zbiory danych

Zbiory	Liczba obiektów	Liczba zmiennych		Liczba klas	Brak danych
		jakościowych	ilościowych		
Echocardio	131	-	7	2	39
Glass	194	-	9	7	-
Heart-disease C	303	7	6	2	-
Heart-disease H	294	7	6	2	746
Hepatitis	157	13	6	2	167
Thyroid	1960	22	7	3	3092
Voting-records	435	16	-	2	392

Źródło: [Clark, Boswell 1991].

Tabela 3. Porównanie dokładności klasyfikacji (w %)

Zbiory	Algorytm				
	CN2			C4.5	Drzewa logiczne
	entropia	Laplace	Laplace*		
Echocardio	63,9	62,3	66,6	63,6	71,1
Glass	45,2	58,5	65,5	64,2	64,7
Heart-disease C	66,3	75,4	76,7	76,4	73,9
Heart-disease H	73	75	78,8	78	81,6
Hepatitis	71,3	77,6	80,1	79,3	82,7
Thyroid	95,6	96,3	96,6	96,4	97,4
Voting-records	93,6	94,8	94,8	95,6	95

* Algorytm generujący nieuporządkowaną listę reguł klasyfikacyjnych.

Źródło: obliczenia własne oraz [Clark, Boswell 1991].

W zbiorze VOTING-RECORDS obiektami są przedstawiciele dwóch ugrupowań politycznych: republikanów i demokratów. Scharakteryzowani są 16 zmiennymi binarnymi określającymi ich poglądy polityczne. ECHOCARDIO, HEPATITIS, HEART-DISEASE i THYROID zawierają dane medyczne. Obiektami są pacjenci, a zmienne to na ogół wiek pacjenta, płeć oraz różne wyniki badań (np. ciśnienie krwi, odczyty kardiogramu, występowanie anoreksji ...). Klasy określają kondycję pacjenta. W zbiorze GLASS identyfikuje się gatunek szkła (7 rodzajów), a zmienne opisują zawartość różnych pierwiastków chemicznych.

Zmienne ilościowe zostały poddane dychotomizacji miarą jakości podziału (1). Brakujące dane potraktowano jako dodatkowe kategorie zmiennych jakościowych (takie rozwiązanie przyjął np. Kass w algorytmie CHAID).

Dokładność klasyfikacji rozumiana jako procent obiektów poprawnie sklasyfikowanych na zbiorze testowym została porównana w trzech algorytmach: CN2 (w trzech odmianach), C4.5 oraz w zaproponowanej metodzie drzew logicznych. Algorytm CN2 zastosowano z funkcją entropii, z funkcją Laplace'a oraz modyfikację polegającą na przekształceniu reguł klasyfikacyjnych do postaci nieuporządkowanej. W metodzie drzew logicznych oceny podzbiorów dokonano funkcją entropii. Dokładność klasyfikacji oszacowano metodą sprawdzania krzyżowego, dzieląc zbiór uczący na 10 części. Wyniki przedstawiono w tab. 3.

6. Podsumowanie

Metoda wykorzystująca drzewa logiczne daje dosyć wysoką dokładność klasyfikacji na tle prezentowanych metod. W czterech zbiorach otrzymano nawet najlepsze wyniki. Ponadto, dzięki uporządkowaniom zmiennych na minimalnych drzewach logicznych, przestrzeń przeszukiwania charakterystyk klas jest znacznie zawężona (np. CN2 przeszukuje całą przestrzeń charakterystyk klas). Charakterystyki klas w postaci uporządkowanej listy reguł mogą prowadzić w przypadku wielu klas do mniejszej czytelności wyników i trudności interpretacyjnych. Należy zauważyć jednak, że tę samą niedogodność mają wcześniejsze wersje CN2, przy jednocześnie mniejszej dokładności klasyfikacji na zbadanych zbiorach. Wyniki przedstawione w artykule pokazują, że mniej znane drzewa logiczne są zasługującym na uwagę narzędziem analizy danych jakościowych.

Literatura

Clark P., Boswell R. (1991), *Rule Induction with CN2: some Recent Improvements*, [w:] Y. Kodratoff (red.), *Machine Learning – EWSL-91, European Working Session on Learning*, Springer Verlag, Berlin.

Clark P., Niblett T. (1989), *The CN2 Induction Algorithm*, Machine Learning, t. 3.

Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.
Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa.
Kubus M. (2005), *Zastosowanie funkcji logicznych do porządkowania cech jakościowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, AE, Wrocław.
Michalski R.S. (1977), *Variable-Valued Logic and its Applications to Pattern Recognition and Machine Learning*, [w:] D.C. Rine (red.), *Computer Science and Multiplevalued Logic*, North-Holland, Amsterdam.
Michalski R.S., Larson J. (1978), *Selection of Most Representative Training Examples and Incremental Generation of VL1 Hypothesis: The Underlying Methodology and the Descriptions of Programs ESEL and AQ11*, Technical Report Report No. 867, Department of Computer Science, University of Illinois, Urbana, Illinois.
Quinlan J.R. (1993), *C4.5 Programs for Machine Learning*, Morgan Kaufmann, San Mateo.

ON A METHOD OF DISCRIMINATION OF THE NOMINAL OBJECTS

Summary

This paper contains an application of minimal logical trees in discriminant analysis. The logical tree is a graph of nominal attributes in a finite space. The minimal tree of every class – with the minimal number of nodes – is used for searching the class description space. Finally, an ordered list of classification rules is obtained. Using well known data sets, the classification accuracy is compared of the logical tree heuristic to CN2 and C4.5 algorithms.