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

Adam Kurzydłowski
Akademia Ekonomiczna we Wrocławiu

PROBLEMATYKA WIELOWYMIAROWYCH DRZEW KLASYFIKACYJNYCH

1. Wstęp

Złożoność otoczenia, w którym przychodzi nam funkcjonować, wymusza poszukiwanie i stosowanie metod badawczych uwzględniających ten charakter rzeczywistości oraz dających możliwość uzyskania miarodajnych wyników. W literaturze przedmiotu pod pojęciem wielowymiarowych metod statystycznych kryją się metody, w których do analizy zebranego materiału statystycznego wykorzystuje się przynajmniej dwie zmienne. Oczywiście w aspekcie badania zależności wymóg ten dotyczy liczby zmiennych objaśniających. Wówczas oprócz posiadania zmiennej objaśnianej konieczne jest wyodrębnienie przynajmniej dwóch zmiennych, których wpływ na badane zjawisko chcemy ustalić.

Wiele z metod oraz technik wykorzystywanych w obszarze badań wielowymiarowych zostało rozwiniętych ostatnio dzięki zwiększeniu zdolności obliczeniowych komputerów. Obecny poziom rozwoju techniki obliczeniowej pozwala na coraz śmielsze wykorzystywanie rozbudowanych baz danych oraz skraca czas potrzebny na odnalezienie rozwiązania nawet w przypadku bardzo złożonego modelu.

Drzewa klasyfikacyjne stanowią nieparametryczną metodę badania zależności, tj. metodę, w której nie wymaga się znajomości rozkładów zmiennych ani postaci analitycznej związku między nimi. Dobór zmiennych następuje na podstawie przyjętego kryterium, a uzyskany model jest relatywnie łatwy w interpretacji oraz wykazuje odporność na wartości nietypowe. Dodatkowo drzewa klasyfikacyjne należą do grupy metod eksploracyjnych, które wykorzystuje się w poszukiwaniu i wykrywaniu relacji i zależności występujących w określonym zbiorze danych. Analiza eksploracyjna, dokonywana na podstawie ustalonej procedury, ma za zadanie identyfikację istniejących powiązań oraz określenie stopnia dopasowania

modelu (1) do posiadanych danych (por. [Gatnar 2001; Hastie, Tibshirani, Friedman 2001, s. 269]):

$$Y = \sum_{k=1}^{K} a_k I(\mathbf{x} \in S_k), \tag{1}$$

gdzie: Y – zmienna objaśniana,

$\mathbf{x} = [X_1, ..., X_z]$ – wektor zmiennych objaśniających,

a_k – parametry modelu,

I – funkcja wskaźnikowa,

$k = 1, \ldots K$ – numer wyodrębnianej klasy,

S_k – rozłączne obszary (klasy) wielowymiarowej przestrzeni zmiennych objaśniających.

Niniejszy artykuł prezentuje zostanie problematykę wielowymiarowej odmiany drzew klasyfikacyjnych, w których podział zbioru obiektów na względnie jednorodne podzbiory jest przeprowadzany na podstawie kombinacji liniowej zmiennych objaśniających. Poruszone zostaną zagadnienia związane z ich praktycznym wykorzystaniem oraz wskazane zostanie pewne pośrednie, dające dobre rezultaty rozwiązanie.

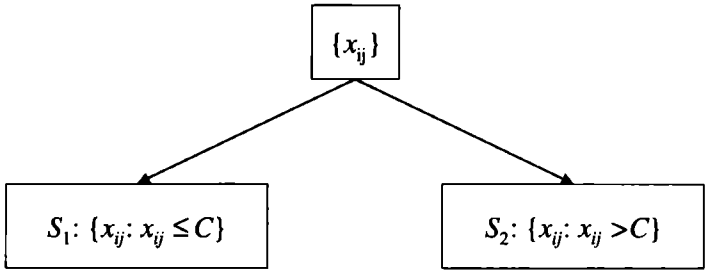
2. Drzewa klasyfikacyjne – jednowymiarowe i wielowymiarowe

Ogólnie drzewa klasyfikacyjne można podzielić na drzewa jednowymiarowe oraz wielowymiarowe. W przypadku drzew jednowymiarowych podział przestrzeni zmiennych objaśniających dokonywany jest hiperpłaszczyznami równoległymi do osi współrzędnych, a w przypadku wielowymiarowej odmiany drzew podział ten odbywa się na podstawie hiperpłaszczyzn ukośnych względem osi.

Drzewa jednowymiarowe zostały bardzo dokładnie opisane w literaturze obcojęzycznej [Breiman in. 1984; Clark, Pregibon 1992; Duda, Hart, Stork 2001; Hastie, Tibshirani, Friedman 2001; Kass 1980; Loh, Shih 1997; Quinlan 1990] oraz w polskich pracach [Gatnar 2001; Gatnar, Walesiak 2004]. Sam autor wielokrotnie przedstawiał problematykę oraz zastosowanie jednowymiarowych drzew klasyfikacyjnych tych o charakterze zarówno dyskryminacyjnym [Kurzydłowski 2002, s. 258-271], jak i regresyjnym [Kurzydłowski 2003, s. 105-113].

Ogólnie istota drzew jednowymiarowych polega na rekurencyjnym podziale zbiorowości na względnie jednorodne podzbiory, w których podział opiera się na jednej zmiennej objaśniającej. Do wyznaczenia podzbiorów wykorzystywana jest określona miara (w zależności od algorytmu) oceniająca poziom heterogeniczności wyodrębnianych podzbiorów. Niezależnie jednak od wykorzystywanej miary podstawowym celem konstrukcji drzewa klasyfikacyjnego jest poszukiwanie takiego podziału zbioru obiektów, aby każdy wydzielany podzbiór posiadał coraz to mniej-

szą heterogeniczność. Przynależność obiektów do wyodrębnionych podzbiorów na każdym poziomie drzewa jest ustalana na podstawie formuły $x_{ij} \leq C$ (rys. 1).



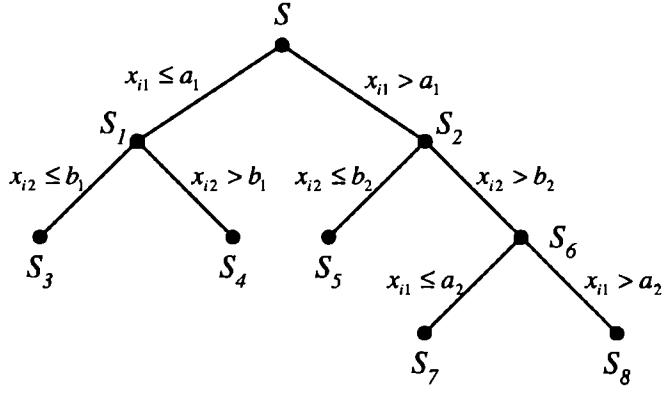
gdzie: x_{ij} – wartość j-tej zmiennej objaśniającej w i-tym obiekcie,

S_1, S_2 – klasy (podzbiory) obiektów,

C – punkt podziału.

Rys.1. Przyporządkowanie zbioru wartości j-tej zmiennej objaśniającej do podzbiorów S_1 i S_2
Źródło: opracowanie własne.

Graficzny przykład konstrukcji drzewa jednowymiarowego przedstawia rys. 2. Zbiór obiektów opisywany jest przez dwie metryczne zmienne objaśniające X_1 i X_2, stanowiące potencjalne kryteria klasyfikacji.



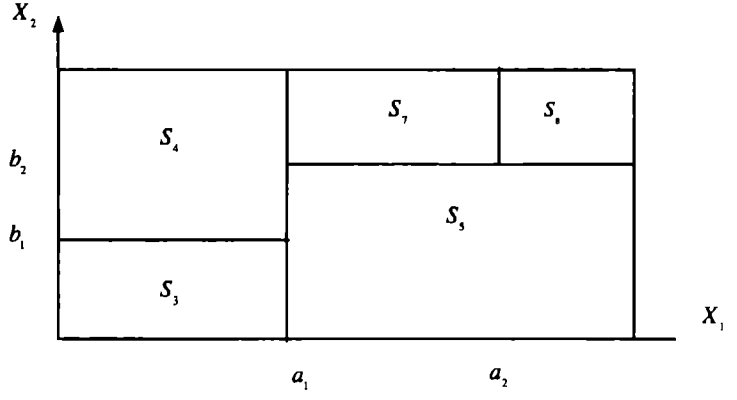
gdzie: $x_{i1}(x_{i2})$ – i-ta obserwacja na zmiennej objaśniającej $X_1(X_2)$,

a_1, a_2 – punkty podziału zbioru wartości zmiennej objaśniającej X_1,

b_1, b_2 – punkty podziału zbioru wartości zmiennej objaśniającej X_2.

Rys. 2. Realizacja podziałów w binarnym drzewie klasyfikacyjnym
Źródło: opracowanie własne.

Przedstawiony proces podziału zbioru obiektów można również zaprezentować na płaszczyźnie (rys. 3).



Rys. 3. Podział zbioru obiektów w dwuwymiarowej przestrzeni zmiennych X_1 i X_2
Źródło: opracowanie własne.

W przypadku drzew wielowymiarowych proces podziału zbioru obiektów wygląda nieco inaczej. W celu wyodrębnienia dwóch podzbiorów homogenicznych ze względu na zmienną objaśnianą stosuje się hiperpłaszczyzny ukośne względem osi współrzędnych. Jest to wynikiem przyjęcia za podstawę podziału kombinacji liniowej zmiennych objaśniających. Dzięki temu drzewa wielowymiarowe zawsze mają mniejsze rozmiary niż drzewa jednowymiarowe. Do ich wad należy zaliczyć jednak to, że są trudne do interpretacji, i to że mogą być stosowane jedynie do danych metrycznych. Dodatkowo otrzymane drzewo nie musi być optymalne, co może być spowodowane liczbą możliwych podziałów badanego zbioru obiektów.

Otóż, o ile w przypadku drzew jednowymiarowych liczba możliwych podziałów z-wymiarowej przestrzeni zmiennych, w której znajduje się n obiektów, wynosi $z \cdot n$, o tyle w przypadku drzew wielowymiarowych liczba możliwych hiperpłaszczyzn wynosi $2^z \binom{n}{z}$. W takim przypadku dosyć dużym problemem jest znalezienie wektora parametrów $\mathbf{a}$ hiperpłaszczyzny P:

$$P(x_1, \ldots, x_z) = \mathbf{a}^{\mathrm{T}} \mathbf{x}, \tag{2}$$

gdzie $\mathbf{a} = [a_0, a_1, \ldots a_z]$ – wektor parametrów.

W literaturze przedmiotu można znaleźć wiele technik deterministycznego oraz losowego wyznaczania położenia hiperpłaszczyzn. Niektóre z nich zostały przedstawione w pracy Gatnara [2001]. W niniejszym artykule zostanie zaprezentowane rozwiązanie stosowane w metodzie CART (*classification and regression trees*) Breimana i in. [1984].

W pierwszym kroku procedury dokonywana jest normalizacja wartości zmiennych poprzez odjęcie mediany i podzielenie wartości przez dwa odchylenia ćwiartkowe. W ten sposób są ustalane początkowe wartości parametrów a_z, spełniające warunek $\sum_{z=1}^{Z} a_z^2 = 1$. Następnie poszukiwany jest najlepszy podział według formuły:

$$\sum_{z=1}^{Z} a_z x_z - \delta(x_1 + \gamma) \leq C, \tag{3}$$

gdzie: $\delta = \dfrac{\sum_{z=1}^{Z} a_z x_z - C}{x_1 + \gamma}$, a γ to parametr przyjmujący jedną z trzech wartości $\{-0{,}25;0;0{,}25\}$.

W następnym kroku porównuje się trzy podziały uzyskane dla każdej wartości γ i wybiera ten, który wyodrębnił podzbiory o najmniejszej heterogeniczności. Cała procedura poszukiwania najlepszego podziału przeprowadzana jest dla wszystkich zmiennych objaśniających. Wartości parametrów δ i γ, dające najlepszy podział, wykorzystuje się do zmodyfikowania położenia hiperpłaszczyzny. W przypadku gdy podział hiperpłaszczyzną ukośną daje o wiele mniejsze zredukowanie heterogeniczności niż podział jednowymiarowy (rys. 1), wówczas wybiera się wariant drugi.

3. Dwuwymiarowe drzewa klasyfikacyjne

Przedstawione rozwiązania, stosowane w drzewach jednowymiarowych oraz wielowymiarowych, należy uzupełnić pewnym pośrednim podejściem. Jest ono wynikiem poszukiwania rozwiązania łączącego w sobie zaletę modelu jednowymiarowego (relatywnie łatwą interpretację uzyskanych wyników) oraz wielowymiarowego (niewielkie rozmiary drzewa).

Idea drzewa dwuwymiarowego polega na tym, że podział zbioru obiektów może być przeprowadzany wyłącznie na podstawie kombinacji dwóch zmiennych objaśniających. W przypadku Z-wymiarowej przestrzeni zmiennych sprawdzanych jest $z(z-1)$ sposobów podziału. W praktyce wyróżnia się dwa modele drzewa dwuwymiarowego.

W pierwszym modelu procedura rozpoczyna się od znalezienia parametrów a_j i a_l zmiennej objaśniającej x_j oraz x_l, zgodnie z formułą:

$$a_j X_j + a_l X_l \leq C. \tag{4}$$

Proces ten obejmuje wszystkie zmienne uwzględnione w badaniu oraz wszystkie możliwe wartości parametrów a_j i a_l. Do ich ustalenia wykorzystuje się najczęściej metodę CART. Podobnie jak w przypadku drzew wielowymiarowych, wartości parametrów dające najlepszy podział wykorzystuje się do zmodyfikowania położenia hiperpłaszczyzny ukośnej. Oczywiście jeżeli poszukiwanie nie przyniosłoby oczekiwanych rezultatów (zredukowanie heterogeniczności podzbiorów), wówczas stosuje się rozwiązanie z podziałem jednowymiarowym. Spotyka się również propozycję wyboru podziału dwuwymiarowego, jeżeli poprawa homogeniczności oscyluje wokół 10% [Bioch, Meer, Potharst 1996, s. 7].

W drugim modelu przyjmuje się z góry ustalone wartości parametrów a_j i a_l. Sprawia to, że proces podziału przebiega znacznie szybciej. Procedura modelu opiera się na formule $x_j + x_l \leq C$ albo $x_j - x_l \leq C$.

Niech ilustracją możliwości drzew dwuwymiarowych będzie następujący przykład [Bioch, Meer, Potharst 1996, s. 7]. W badaniu dla każdego z pięciu zbiorów danych (tab. 1) zbudowano drzewo dwuwymiarowe, wielowymiarowe oraz tradycyjne jednowymiarowe.

Tabela 1. Zbiory danych poddanych analizie

Zbiór danych	Liczebność	Liczba zmiennych	Liczba kategorii zmiennej objaśnianej
Szkło	214	9	6
Diabetycy	789	8	2
Nowotwór piersi	699	9	2
Choroba serca	270	13	2
Fale morskie	300	21	3

Źródło: [Bioch, Meer, Potharst 1996, s. 11].

Zbiór szkło (214 obserwacji) zawiera informacje o 9 cechach fizycznych oraz chemicznych szkła, które określają przynależność kawałka szkła do jednego z sześciu typów. Zbiór diabetyków to zbiór zaobserwowanych objawów cukrzycy u 768 Indianek w wieku powyżej 21 lat. Trzecim zbiorem jest zbiór 699 pacjentek cierpiących na nowotwór piersi. Zbiór czwarty to 270 pacjentów skarżących się na ból serca oraz tych, którzy takie problemy mieli. Piątym jest zbiór 300 obserwacji fal morskich scharakteryzowanych za pomocą 21 zmiennych.

Po zastosowaniu trzech typów drzew klasyfikacyjnych, tj. drzewa dwuwymiarowego z uwzględnieniem dwóch modeli oraz drzewa jednowymiarowego i wielowymiarowego, oba wygenerowane algorytmem OC1 (Oblique Classifier 1 autorstwa Murthy i in. [1993]), dla poszczególnych zbiorów danych uzyskano następujące wyniki klasyfikacji (tab. 2).

Tabela 2. Wyniki przeprowadzonych badań (dokładność klasyfikacji/rozmiar drzewa)

Metoda	Szkło	Diabetycy	Nowotwór piersi	Choroba serca	Fala morska
Drzewo dwuwymiarowe (I model)	65,3	74,3	95,4	78,5	76,1
	6,2	5,2	2,8	4,1	5,0
Drzewo dwuwymiarowe (II model)	64,8	74,7	95,4	78,5	76,2
	6,8	5,8	2,6	4,1	5,2
OC1 – drzewo jednowymiarowe	63,8	73,8	94,5	75,9	69,1
	8,1	8,4	6,4	5,2	4,8
OC1 – drzewo wielowymiarowe	64,0	74,2	95	77,4	78,2
	7,9	4,3	2,8	3,9	3,8

Źródło: [Bioch, Meer, Potharst 1996, s. 11].

W przypadku pierwszych czterech zbiorów po zastosowaniu drzewa dwuwymiarowego dokładność klasyfikacji okazała się najlepsza (najwyższa wartość). Wyższa nawet niż po zastosowaniu drzewa wielowymiarowego, uważanego za rozwiązanie prowadzące do uzyskania najlepszych rezultatów w maksymalizacji homogeniczności wyodrębnianych podzbiorów. Tylko w przypadku zbioru fal morskich dokładność klasyfikacji uzyskana na podstawie drzewa wielowymiarowego okazała się najwyższa.

Jeżeli chodzi o liczbę wyodrębnianych podzbiorów, ponownie drzewa dwuwymiarowe okazały się wyjątkowe przydatne. W pierwszym przypadku (zbiór szkła) i trzecim (nowotwór piersi) drzewa te okazały się ponownie efektywniejsze niż drzewa wielowymiarowe, doprowadzając do wygenerowania drzewa o najmniejszych rozmiarach. W zbiorze pacjentów z chorobą serca rezultat w postaci wielkości wygenerowanego drzewa w obu przypadkach był bardzo zbliżony. W zbiorze drugim i piątym liczba poziomów drzewa dwuwymiarowego była już nieco większa niż po zastosowaniu drzewa wielowymiarowego.

We wszystkich przypadkach okazało się, że drzewa wielowymiarowe i dwuwymiarowe są efektywniejsze od drzew jednowymiarowych – większa dokładność klasyfikacji i mniejsze rozmiary.

4. Zakończenie

Drzewa klasyfikacyjne należące do grupy metod eksploracyjnych dają szerokie możliwości analizy. W praktyce badacz może dokonać wyboru typu konstruowanego drzewa (ze względu na liczbę zmiennych stanowiących o podziale), algorytmu czy też miary oceny jakości podziału. To wszystko sprawia, że grupa tych metod jest coraz częściej wykorzystywana przez analityków dysponujących dużymi zbiorami danych. Oczywiście każde z wybranych rozwiązań ma swoje wady i zalety. Jednowymiarowe drzewa klasyfikacyjne charakteryzują się znacznym rozmiarem. W przypadku drzew wielowymiarowych można mówić o trudnościach w znalezieniu położenia hiperpłaszczyzn oraz o skomplikowanej interpretacji podziałów.

W takich przypadkach przydatnym rozwiązaniem staje się zastosowanie drzew dwuwymiarowych. Jak okazało się w zaprezentowanym badaniu, zastosowanie drzew dwuwymiarowych doprowadziło do redukcji zanieczyszczenia, czyli do otrzymania dokładniejszej klasyfikacji. Dodatkowym osiągnięciem po zastosowaniu modelu dwuwymiarowego było zmniejszenie rozmiarów uzyskiwanego drzewa klasyfikacyjnego. Drzewa dwuwymiarowe mogą być również wykorzystane w celu uwzględniania interakcji między zmiennymi objaśniającymi.

Wszystko to sprawia, że badacz sięgający po drzewa klasyfikacyjne otrzymuje niezwykle elastyczne narzędzie: od drzew jednowymiarowych poprzez dwuwymiarowe do wielowymiarowych. Ostateczny wybór narzędzia jak zawsze należy do badacza, ale o wiele łatwiej decyzję taką podejmuje się w warunkach pełnej wiedzy.

Literatura

Bioch J., Meer O. van der, Potharst R. (1996), *Bivariate Decision Trees*, http://www.few.eur.nl/few/research/pubs/cs/1996/eur-few-cs-96-02.pdf.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984), *Classification and Regression Trees*, Belmont, Wadsworth.

Clark L.A., Pregibon D. (1992), *Tree-Based Models,* [w:] J.M. Chambers, T.J. Hastie (red.), *Statistical Models in S,* Chapman & Hall, New York.

Duda R., Hart P., Stork D. (2001), *Pattern Classification*, Jonh Wiley & Sons Inc, New York, Chichester, Weinheim, Brisbane, Toronto.

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa.

Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław.

Hastie T., Tibshirani R., Friedman J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, New York, Berlin, Heidelberg.

Kass G.V. (1980), *An Exploratory Technique for Investigating Large Quantities of Categorical Data*, „Applied Statistics" nr 29, s.119-127.

Kurzydłowski A. (2002), *Klasyfikacja nabywców czekolady z wykorzystaniem algorytmów CHAID i C&RT,* [w:] K. Jajuga, M. Walesiak, *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 9, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 942, s. 258-271.

Kurzydłowski A. (2003), *Ocena atrakcyjności segmentów rynku z wykorzystaniem algorytmu CHAID*, [w:] E. Gatnar (red.), *Analiza i prognozowanie zjawisk rynkowych o charakterze niemetrycznym*, Materiały z VI Warsztatów Metodologicznych w Katowicach, AE, Katowice, s. 105-113.

Loh W.Y, Shih Y.S. (1997), *Split Selection Methods for Classification Trees*, „Statistica Sinica" nr 7, s. 815-840.

Murthy S., Kasif S., Salzberg S., Beigel R. (1993), *OC1: Randomized Induction of Oblique Decision Trees*, http://citeseer.ist.psu.edu/cache/papers/cs/421/http:zSzzSzwww.cs. jhu.eduzSz~murthyzSzaaai93.pdf/murthy93oc.pdf.
Quinlan J.R. (1990), *Decision Trees and Decisionmaking*, „IEEE Transactions on Systems, Man and Cybernetics" vol. 20, nr 2, s. 339-346.

MULTIVARIATE CLASSIFICATION TREES PROBLEMS

Summary

Classification trees belong to the group of the exploration methods, exploited in searching and detecting relations occurring in the data. In the article a classification multivariate trees problems is presented. Multivariate trees classify examples by testing linear combinations of the variables at each non-leaf node of the tree.