PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Iwona Kasprzyk
Akademia Ekonomiczna w Katowicach

ZASTOSOWANIE DRZEW KLASYFIKACYJNYCH W ANALIZIE KLAS UKRYTYCH

1. Wstęp

W badaniach ekonomicznych często spotyka się dane mierzone na słabych skalach pomiaru. Jedną z metod pozwalających na analizę tego typu danych jest analiza klas ukrytych (*latent class analysis*), która została zaproponowana przez Lazarsfelda [Lazarsfeld, Henry 1968]. Obecnie tą metodą zajmuje się Vermunt [1997].

Drzewa klasyfikacyjne oraz regresyjne są nieparametryczną metodą. Zostały one wprowadzone przez Breimana i in. [1984]. Wyraźną zaletą tego rodzaju drzew jest to, że w przeciwieństwie do analizy klas ukrytych, pozwalają analizować dane mierzone zarówno na słabych, jak i na mocnych skalach pomiaru.

Celem artykułu będzie zaprezentowanie, w jaki sposób drzewa klasyfikacyjne mogą usprawnić analizę klas ukrytych oraz jak tego typu drzewa można wykorzystywać do ograniczenia liczby zmiennych wprowadzanych do analizy klas ukrytych.

W końcowym etapie artykułu zostanie również pokazana graficzna prezentacja danych za pomocą tzw. wykresu współrzędnych barycentrycznych, który to pozwoli na przedstawienie zależności między kategoriami zmiennych zawartych w tablicy kontyngencji.

2. Etapy postępowania w proponowanej procedurze

2.1. Redukcja zbioru zmiennych przeprowadzana za pomocą drzew klasyfikacyjnych

Drzewa są nieparametryczną postacią modelu dyskryminacyjnego lub regresyjnego [Breiman i in. 1984]. Drzewa umożliwiają budowę modelu w postaci:

$$y = \sum_{k=1}^{K} a_k I(x \in R_k), \tag{1}$$

gdzie: R_k – rozłączne obszary w przestrzeni cech,
a_k – parametry modelu,
x – wektor cech.

W zależności od tego, czy y jest zmienną mierzoną na skali mocnej, czy zmienną mierzoną na skali nominalnej, mówimy w pierwszym przypadku o *drzewach regresyjnych*, natomiast w drugim – o *drzewach klasyfikacyjnych*.

Celem budowy takiego drzewa jest uzyskanie podzbiorów maksymalnie jednorodnych ze względu na wartość zmiennej zależnej.

Do podstawowych miar jakości podziału drzewa stosowanych w drzewach klasyfikacyjnych należą: wskaźnik Giniego, miara entropii, reguła podziału na dwie części (*twoing rule*). Więcej miar można znaleźć w pracy [Gatnar 2001].

W przypadku natomiast drzew regresyjnych miarą jakości podziału drzewa jest wariancja.

2.2. Przeprowadzenie analizy klas ukrytych

Załóżmy, że mamy daną tablicę kontyngencji z trzema zmiennymi obserwowalnymi: A $(i=1,2, ...,I)$, B $(j=1,2, ...,J)$ oraz C $(k=1,2, ...,K)$. Zmienna ukryta X przyjmuje wartości $w=1,2, ...,W$, gdzie W oznacza liczbę klas. Model z jedną zmienną ukrytą X można przedstawić za pomocą poniższego równania:

$$\pi_{ijk} = \sum_{w=1}^{W} \pi_{ijkw}^{ABCX}, \tag{2}$$

gdzie: π_{ijkw}^{ABCX} – prawdopodobieństwo warunkowe tego, że i-ta, j-ta, k-ta kategoria zmiennej A, B oraz C znajdzie się w opisie klasy ukrytej w.

Wykorzystanie wzoru (2) wymaga spełnienia założenia o lokalnej niezależności zmiennych:

$$\pi_{ijkw}^{ABCX} = \pi_{w}^{X} \pi_{iw}^{A\backslash X} \pi_{jw}^{B\backslash X} \pi_{kw}^{C\backslash X}, \tag{3}$$

gdzie: π_{w}^{X} – prawdopodobieństwo przynależności danych obserwacji do klasy w zmiennej ukrytej X,

$\pi_{iw}^{A\backslash X}$ – prawdopodobieństwo warunkowe tego, że i-ta kategoria zmiennej A znajdzie się w opisie klasy ukrytej w.

Prawdopodobieństwa po prawej stronie równania (3) wymagają spełnienia odpowiedniego założenia:

$$\sum_{w=1}^{W} \pi_{w}^{X} = \sum_{i=1}^{I} \pi_{iw}^{A\backslash X} = \sum_{j=1}^{J} \pi_{jw}^{B\backslash X} = \sum_{k=1}^{K} \pi_{kw}^{C\backslash X} = 1. \tag{4}$$

Wybór optymalnej liczby klas

Jednym ze sposobów wyboru modelu z odpowiednią liczbą klas są kryteria informacyjne. W celu wyboru liczby klas w analizie klas ukrytych można się posłużyć wieloma kryteriami informacyjnymi. Do najbardziej popularnych w tego typu analizie zaliczamy kryterium informacyjne Akaike (AIC) oraz bayesowskie kryterium informacyjne (BIC). Również w analizie klas ukrytych często stosuje się kry-

terium informacyjne CAIC. Szeroko stosowana w analizie klas ukrytych jest statystyka BIC, którą można zapisać za pomocą poniższej formuły:

$$BIC = -2LL + \ln(n) \cdot df, \tag{5}$$

gdzie: LL – iloraz wiarygodności,
n – liczba obserwacji,
df – liczba parametrów modelu.

Sclove [1987] skorygował próbę w kryterium informacyjnym BIC, przyjmując za liczbę obserwacji $n = (n+2)/24$. W dalszej części to kryterium będzie oznaczane jako ABIC.

Kryterium informacyjne Akaike [1974] ma następującą postać:

$$AIC = -2LL + 2 \cdot df. \tag{6}$$

Kryterium informacyjne CAIC zaproponowane przez Bozdogana [1987] można zapisać za pomocą poniższej formuły:

$$CAIC = -2LL + (\ln(n) + 1) \cdot df. \tag{7}$$

Zgodnie z przyjętym kryterium informacyjnym wybierany jest ten model, w którym wartość danego kryterium jest najmniejsza. Wówczas taki model będzie wskazywał optymalną liczbę klas, jaką należy przyjąć do dalszej analizy.

Wizualizacja danych

Van der Heijden, Gilula oraz van der Ark [1999] pokazali, że modele klas ukrytych można przedstawić za pomocą wykresu współrzędnych barycentrycznych (*barycentric coordinates*). Wykresem tego rodzaju współrzędnych jest trójkąt równoboczny, którego wierzchołki (trzy niewspółliniowe punkty P_1, P_2, P_3) tworzą bazę przestrzeni. Każdemu z tych punktów przyporządkowana jest pewna waga (w tym przypadku zakładamy, że wszystkie wagi są równe 1). Dowolny punkt w przestrzeni trójwymiarowej można wyrazić jako sumę ważoną:

$$P = xP_1 + yP_2 + zP_3.$$

Współczynniki x, y, z są nazywane współrzędnymi barycentrycznymi. Przyjmują wartości z przedziału [0,1], a ich suma jest równa 1 ($x + y + z = 1$). Kategorię danej zmiennej można potraktować jako taki punkt P o współrzędnych x, y, z.

Poprzez odpowiednie przekształcenie poszczególnych prawdopodobieństw warunkowych $\pi_{iw}^{A\backslash X}, \pi_{jw}^{B\backslash X}, \pi_{kw}^{C\backslash X}$ model klas ukrytych może być przedstawiony za pomocą poniższego równania:

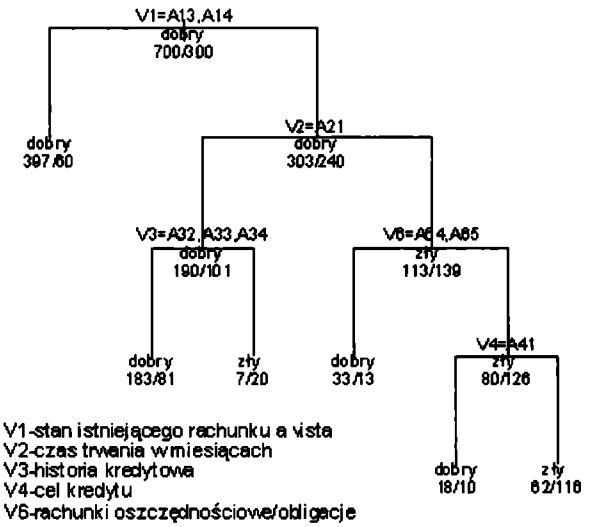
$$\frac{\pi_{ijk}}{\pi_{i...}} = \sum_{w=1}^{W} \pi_{wi}^{X/A} \pi_{jw}^{B/X} \pi_{kw}^{C/X}, \tag{8}$$

$$\text{gdzie: } \pi_{wi}^{X/A} = \frac{\pi_w^X \pi_{iw}^{A/X}}{\sum_{w=1}^{W} \pi_w^X \pi_{iw}^{A/X}} = \frac{\pi_w^X \pi_{iw}^{A/X}}{\pi_{i..}}. \tag{9}$$

W ten sposób przekształcone prawdopodobieństwa warunkowe pozwalają na naniesienie punktów na wykresie współrzędnych barycentrycznych i odczytanie zależności między poszczególnymi kategoriami zmiennych obserwowalnych.

3. Zastosowanie

Do ilustracji obydwu metod wykorzystano zbiór danych *German Credit*, który jest udostępniany przez repozytorium Uniwersytetu Kalifornijskiego [Blake i in. 1998]. W zbiorze tym jest 1000 obserwacji, które opisane są 21 zmiennymi. Zmienną y jest zmienna opisująca rodzaj klienta: 700 wniosków oceniono jako dobre, a 300 jako złe. Znajdujące się w zbiorze zmienne ciągłe poddano dyskretyzacji zgodnie z wartością średniej albo mediany.



Rys. 1. Drzewo klasyfikacyjne dla zbioru *German Credit*
Źródło: opracowanie własne.

Klasyfikacji dokonano za pomocą funkcji `rpart`, którą można znaleźć w programie R. W wyniku przeprowadzonej klasyfikacji algorytm CART wyodrębnił 5 zmiennych (zob. rys. 1), które to wykorzystano następnie do analizy klas ukrytych. Zmienną decydującą o podziale drzewa jest stan istniejącego rachunku *a vista*. Następną zmienną decydującą o podziale drzewa jest *czas trwania kredytu*. Jeżeli czas trwania kredytu był poniżej 21 miesięcy, to obserwacje kierowane są do lewej gałęzi

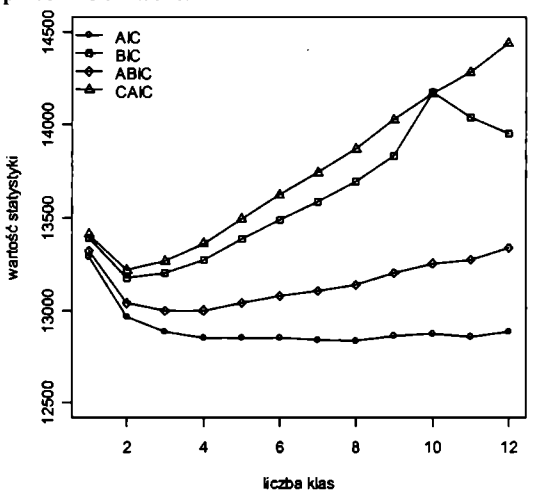
drzewa, gdzie następuje podział zgodnie ze zmienną *historia kredytowa*. Jeżeli czas trwania kredytu jest powyżej 21 miesięcy, to obserwacje kierowane są do prawej gałęzi drzewa, gdzie następuje podział ze względu na zmienną *rachunki oszczędnościowe/obligacje*, a w kolejnym kroku ze względu na zmienną *cel kredytu*.

W wyniku przeprowadzonej analizy klas ukrytych (zob. tab. 1 i rys. 2), zgodnie z kryterium informacyjnym BIC oraz CAIC należy przyjąć, że najlepiej dopasowanym modelem jest model z dwiema klasami ukrytymi, podczas gdy skorygowane kryterium BIC (ABIC) wskazuje na model z trzema klasami. Kryterium informacyjne Akaike (AIC) wybiera modele znaczniej rozbudowane. Dla tego typu danych kryterium to wskazało model z ośmioma klasami. Poniżej znajdują się wartości poszczególnych kryteriów informacyjnych wraz z ilustracją graficzną.

Tabela 1. Wartości kryteriów informacyjnych

Liczba klas	BIC	AIC	CAIC	ABIC
1	13391,7	13288,7	13412,7	13325,0
2	**13176,8**	12965,8	**13219,8**	13040,3
3	13204,8	12885,8	13269,8	**12998,4**
4	13277,1	12850,1	13364,1	13000,7
5	13387,7	12852,8	13496,7	13041,5
6	13494,5	12851,6	13625,5	13078,5
7	13591,5	12840,6	13744,5	13105,5
8	13696,6	**12837,7**	13871,6	13140,8
9	13830,9	12864	14027,9	13205,2
10	14175,7	12876,4	14170,2	13255,7
11	14041,9	12859,1	14282,9	13276,4
12	13951,2	12884,9	14438,7	13340,4

Źródło: opracowanie własne.



Rys. 2. Wartości kryteriów informacyjnych – ilustracja graficzna

Źródło: opracowanie własne.

Tabela 2. Wyniki przeprowadzonej segmentacji dla dwóch i trzech klas ukrytych

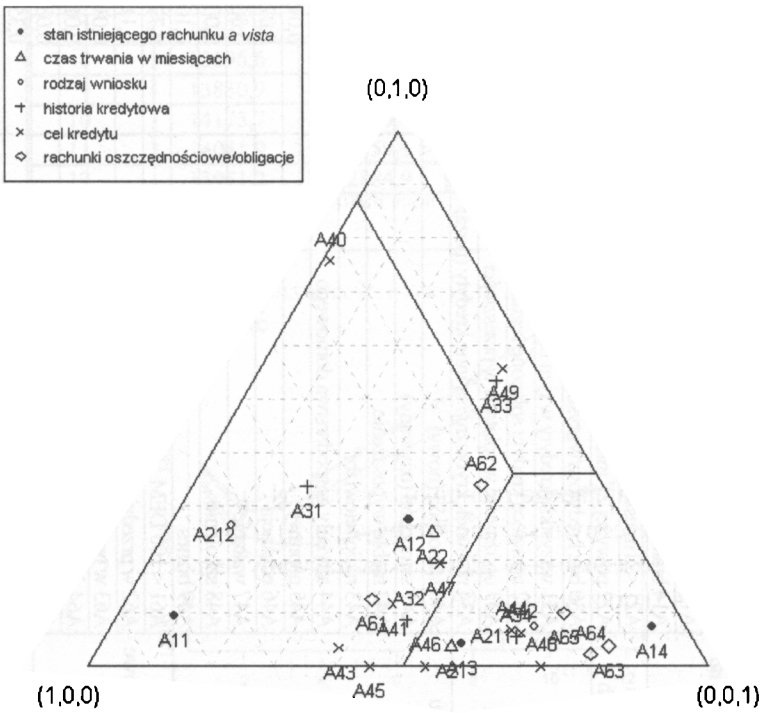
		Podział zgodnie z kryterium BIC i CAIC		Podział zgodnie z kryterium ABIC		
Nazwa zmiennej	Kategorie zmiennej	Klasa I	Klasa II	Klasa I	Klasa II	Klasa III
		0,394	0,606	0,360	0,132	0,508
Stan istniejącego rachunku *a vista*	A11 <0 DEM	**0,533**	0,106	**0,619**	0,199	0,049
	A12 w przedziale (0,200> DEM	0,358	0,211	0,257	**0,562**	0,201
	A13 ≥ 200 DEM/pensja na rachunek min. rok	0,044	0,076	0,066	0,020	0,072
	A14 brak rachunku	0,066	**0,608**	0,058	0,219	**0,677**
Czas trwania w miesiącach*	A21 <21miesięcy	0,481	**0,602**	**0,606**	0,155	**0,621**
	A22 powyżej 21 miesięcy	**0,519**	0,398	0,394	**0,845**	0,379
Historia kredytowa	A30 bez kredytów/wszystkie kredyty spłacone	0,097	0,023	0,069	0,133	0,019
	A31 wszystkie kredyty w tym banku spłacone	**0,632**	**0,502**	**0,676**	0,358	**0,512**
	A32 istniejące kredyty spłacone regularnie	0,089	0,094	0,019	**0,383**	0,073
	A33 opóźnienie w spłatach w przeszłości	0,183	0,381	0,236	0,125	0,396
	A34 rachunek krytyczny/są inne kredyty (poza)	0,097	0,023	0,069	0,133	0,019
Cel kredytu	A40 samochód (nowy)	**0,279**	0,204	**0,293**	0,207	0,199
	A41 samochód (używany)	0,064	0,128	0,077	0,049	0,136
	A42 meble/wyposażenie	0,026	0,003	0,291	0,048	0,138
	A43 radio/telewizor	0,232	0,148	0,219	0,144	**0,358**
	A44 artykuły gospodarstwa domowego	0,194	**0,336**	0,018	0,000	0,011
	A45 naprawy	0,015	0,010	0,028	0,000	0,024
	A46 edukacja	0,025	0,020	0,046	0,073	0,047
	A47 wakacje	0,060	0,044	0,007	0,000	0,013
	A48 szkolenie	0,005	0,012	0,014	**0,410**	0,075
	A49 biznes	0,100	0,095	0,008	0,069	0,000
Rachunki oszczędnościowe /obligacje	A61 < 100 DEM	**0,773**	**0,492**	**0,803**	**0,573**	**0,469**
	A62 w przedziale <100,500) DEM	0,111	0,098	0,056	0,265	0,094
	A63 w przedziale <500,1000) DEM	0,016	0,094	0,025	0,017	0,102
	A64 ≥ 1000 DEM	0,015	0,070	0,024	0,008	0,076
	A65 brak danych/brak rach. oszczędnościowego	0,085	0,247	0,093	0,137	0,259
Rodzaj klienta	A211 klient dobry	0,354	**0,925**	0,470	0,397	**0,942**
	A212 klient zły	**0,646**	0,075	**0,530**	**0,603**	0,058

* Dokonano dyskretyzacji zmiennej.

Źródło: opracowanie własne.

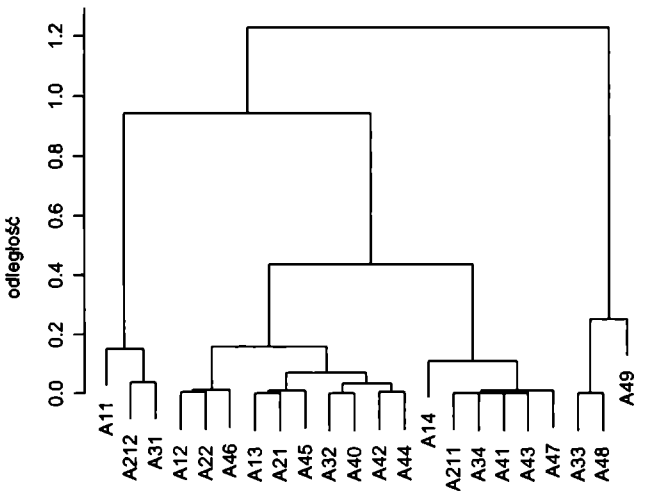
W wyniku dokonanego podziału zgodnie z kryterium informacyjnym BIC oraz CAIC można wskazać, że do klasy I należy 60,6% wszystkich wniosków, natomiast do klasy II 39,4% wniosków kredytowych. Klasę I opisują klienci, którzy zostali przez bank uznani za niewiarygodnych. Ta klasa charakteryzuje się tym, że rachunek *a vista* tych osób jest poniżej 0 DEM (53,3% wniosków). Osoby ubiegające się o kredyt chciały go wziąć na więcej niż 21 miesięcy (51,9%). Główny cel, na jaki chciały przeznaczyć kredyt, to nowy samochód (27,9%). Większość przyszłych kredytobiorców (77,3%) miało rachunek oszczędnościowy poniżej 100 DEM. Natomiast w II klasie, w której 92,5% wszystkich wniosków zostało uznanych za dobre, kredyt ma być przeznaczony na artykuły gospodarstwa domowego (33,6%). Jeżeli chodzi o stan rachunku, to osoby te w większości (60,8%) nie miały rachunku *a vista*, kredyt chciały wziąć na mniej niż 21 miesięcy. Zarówno w klasie I, jak i II u większości osób ubiegających się o kredyt wszystkie poprzednie kredyty w tym banku zostały spłacone.

W tab. 2 znajduje się rozkład prawdopodobieństw dla trzech klas, gdybyśmy kierowali się kryterium informacyjnych ABIC. Zgodnie ze wzorem (4) należy przekształcić prawdopodobieństwa warunkowe tak, aby można było nanieść poszczególne kategorie zmiennych (jako punkty) na tzw. wykresie współrzędnych barycentrycznych. Tego rodzaju wykres pozwoli w efekcie na interpretację zależności między kategoriami zmiennych (rys. 3).



Rys. 3. Wykres współrzędnych barycentrycznych dla danych *German Credit*
Źródło: opracowanie własne.

Ponieważ na wykresie znajduje się bardzo duża liczba kategorii i często nie jesteśmy w stanie odczytać bezpośrednio zależności między kategoriami zmiennych, można posłużyć się (podobnie jak w analizie korespondencji) hierarchiczną analizą skupień. W tym przypadku będzie to metoda Warda, która okazuje się niezwykle przydatna przy tego rodzaju wykresach. W programie *R* tę metodę realizuje m.in. funkcja `agnes` w pakiecie `cluster`. Analizowane kategorie należy poddać grupowaniu na podstawie przekształconych prawdopodobieństw warunkowych, które to są współrzędnymi tych kategorii na omawianym wykresie (rys. 3).



Rys. 4. Dendrogram uzyskany za pomocą metody Warda dla danych *German Credit*
Źródło: opracowanie własne.

W artykule skupiono się na interpretacji zależności między kategoriami zmiennej *rodzaj klienta* a pozostałymi kategoriami analizowanych zmiennych wziętych pod uwagę w analizie klas ukrytych. W celu łatwiejszej interpretacji rys. 3 posłużono się dendrogramem przedstawionym na rys. 4.

Z uzyskanego dendrogramu wynika, że osoba, której wniosek uznano za dobry, charakteryzowała się:

- brakiem stanu istniejącego rachunku *a vista* (A14),
- rachunkiem krytycznym/są inne kredyty (poza) (A34),
- kredyt będzie przeznaczany głównie na używany samochód (A41), radio bądź telewizor (A43) oraz wakacje (A47).

Wniosek osób, który został przez bank odrzucony, charakteryzował się tym, że:

- stan rachunku *a vista* był poniżej 0 DEM (A11),
- wszystkie kredyty w tym banku zostały spłacone (A31).

4. Podsumowanie

Drzewa klasyfikacyjne, jako metoda, mają wyraźną przewagę nad analizą klas ukrytych – wykorzystują zarówno zmienne niemetryczne, jak i metryczne, podczas gdy analiza klas ukrytych pozwala na analizę danych mierzonych na słabych skalach pomiaru.

Celem artykułu była przede wszystkim redukcja liczby zmiennych wprowadzanych do analizy klas ukrytych. Bardzo użyteczną metodą okazały się w tym wypadku drzewa klasyfikacyjne, które pozwoliły wybrać 6 zmiennych spośród 21 analizowanych w przykładzie *German Credit*.

Dodatkowo, za pomocą analizy klas ukrytych, pokazano na wykresie współrzędnych barycentrycznych kategorie zmiennych, które pomogły w scharakteryzowaniu wniosku. Został on następnie uznany za dobry lub został odrzucony przez bank. Analiza klas ukrytych może się okazać uzupełnieniem dla drzew klasyfikacyjnych, gdyż pozwala na identyfikację współwystępowań między kategoriami rozpatrywanych zmiennych, gdy posługujemy się graficzną wizualizacją tej metody.

Wyraźną zaletą przedstawionej procedury jest to, że za pomocą drzew klasyfikacyjnych zostały wybrane te zmienne, które miały istotny wpływ na zmienną zależną. W końcowym etapie został pokazany wykres współrzędnych barycentrycznych, który wskazał najbardziej istotne kategorie zmiennych mające decydujące znaczenie w przypadku przyjęcia lub odrzucenia wniosku kredytowego.

Literatura

Akaike H. (1974), *A New Look at the Statistical Model Identification*, „IEEE Transactions on Automatic Control" 19 (6), s. 716-723.

Blake C., Keogh E., Merz C.J. (1998). *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, www.ics.uci.edu /~mlearn/MLRepository.html.

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

Bozdogan H. (1987), *Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions*, „Psychometrica" 52, s. 345-370.

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

Haberman S.J. (1979), *Analysis of Qualitative Data*, vol. 2, *New Developments*, Academic Press, New York.

Heijden P.G.M. van der, Gilula Z., van der, Ark L.A. (1999), *An Extended Study into the Relationship between Correspondence Analysis and Latent Class Analysis*, „Sociological Methodology" nr 29, s. 147-186.

Lazarsfeld P.F., Henry N.W. (1968), *Latent Structure Analysis*, Houghton Mill, Boston.

Sclove L.S. (1987), *Application of Model-selection Criteria to Some Problems in Multivariate Analysis*, „Psychometria" nr 52, s. 333-343

Vermunt J.K. (1997), *Log-linear Models for Event Histories*, Sage, Thousand Oaks.

AN APPLICATION OF THE CLASSIFICATION TREE IN THE LATENT CLASS ANALYSIS

Summary

The latent class analysis is one of multivariate analysis techniques of the contingency table which is based on discrete data. This method was introduction by Lazarsfeld [1968]. The classification and regression tree is a nonparametric technique which was proposed by Breiman et al. [1984].

The main aim of this article is to show how we can use the classification tree to reduce the number of variables which can be put into the latent class analysis. Additionally the article shows the barycentric coordinates graph which permits to describe the relations among categories of variables.