PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Marcin Pełka, Justyna Wilk
Uniwersytet Ekonomiczny we Wrocławiu

METODY SELEKCJI ZMIENNYCH SYMBOLICZNYCH W ZAGADNIENIACH KLASYFIKACJI

Streszczenie: Celem artykułu jest przedstawienie i porównanie dwóch metod selekcji zmiennych w analizie danych symbolicznych – tj. metody grafowej Ichino oraz modyfikacji metody *HINoV*. W artykule przedstawiono podstawowe pojęcia z zakresu analizy danych symbolicznych wraz z możliwymi metodami selekcji zmiennych symbolicznych.

W części empirycznej porównano wyniki badań symulacyjnych na przykładzie danych wygenerowanych za pomocą procedury cluster.Gen z pakietu cluster.Sim dla programu R.

1. Wstęp

Głównym celem klasyfikacji jest podział zbioru obiektów na klasy zawierające obiekty podobne ze względu na obserwacje na zmiennych (zob. np. [Gatnar, Walesiak 2004, s. 317]). Jednym z podstawowych etapów procesu klasyfikacji jest wybór zmiennych. Od jakości zestawu zmiennych zależy w znacznej mierze wiarygodność wyników klasyfikacji. W klasyfikacji należy uwzględnić zmienne wykazujące zdolność dyskryminacji zbioru obiektów.

Złożoność danych empirycznych (np. marketingowych) sprawia, że dane w ujęciu klasycznym (ilorazowe, przedziałowe, porządkowe, nominalne) stają się często niewystarczające do opisu badanych obiektów. Możliwość pełniejszego opisu daje wykorzystanie danych symbolicznych [Bock, Diday 2000]. Wiele aspektów dotyczących pojedynczego obiektu cechuje bowiem wielowariantowość lub przedział wartości, forma struktury udziałowej czy powiązania zależnościami, relacjami (np. logicznymi).

W literaturze prezentowane są różne opracowania dotyczące doboru zmiennych dla obiektów w ujęciu klasycznym (zob. np. [Milligan 1994; Fowlkes, Gnanadesikan, Kettenring 1988]), niestety brakuje takich opracowań dla danych w ujęciu symbolicznym.

Celem artykułu jest prezentacja metod selekcji, jakie można zastosować, gdy zbiór obiektów jest opisany zmiennymi symbolicznymi. Przedstawione są i porównane metody: grafowa Ichino (zob. [Ichino 1994]), a także metoda $HINoV^S$ (zob. [Walesiak, Dudek 2008]); do porównania wykorzystano sztucznie wygenerowane

zbiory obiektów opisywane przez zmienne symboliczne w postaci przedziałów liczbowych (z wykorzystaniem procedury `cluster.Gen` z pakietu `clusterSim` dla programu R), w tym także zbiory ze zmiennymi zakłócającymi i obserwacjami odstającymi (*outliers*).

2. Zmienne symboliczne

Obiekty symboliczne mogą być opisywane przez następujące rodzaje zmiennych symbolicznych [Bock, Diday 2000, s. 2-3]:

1) zmienne w ujęciu klasycznym, tj. ilorazowe, przedziałowe, porządkowe czy nominalne;

2) zmienne symboliczne, tj. zmienne:

- interwałowe, czyli przedziały liczbowe, rozłączne lub nierozłączne, np. kwota, którą respondent jest skłonny przeznaczyć na zakup nowego telewizora [1000; 1500];
- wielowariantowe, gdzie realizacją zmiennej jest więcej niż jeden wariant (liczba czy kategoria);
- wielowariantowe z wagami (prawdopodobieństwami), gdzie oprócz listy kategorii występują wagi (prawdopodobieństwa), z jakimi obiekt przyjmuje wybraną kategorię.

Niezależnie od typu zmiennej w analizie danych symbolicznych możemy mieć do czynienia ze zmiennymi strukturalnymi [Bock, Diday 2000, s. 2-3; 33-37]:

- zmienne o zależności funkcyjnej lub logicznej pomiędzy poszczególnymi zmiennymi, gdzie *a priori* zostały określone reguły funkcyjne czy logiczne, które decydują o tym, jakie wartości przyjmuje dana zmienna;
- zmienne hierarchiczne, w których *a priori* określono warunki, które decydują o tym, czy zmienna dotyczy danego obiektu czy też nie;
- zmienne taksonomiczne, w których *a priori* ustalono systematykę realizacji wartości takiej zmiennej.

W analizie danych symbolicznych wyróżnia się dwa rodzaje obiektów symbolicznych:

– obiekty symboliczne I rzędu – obiekt rozumiany w sensie „klasycznym" (obiekt elementarny), np. konsument, produkt, przedsiębiorstwo, pacjent, gospodarstwo domowe, gmina,
– obiekty symboliczne II rzędu – obiekty utworzone w wyniku agregacji zbioru obiektów symbolicznych I rzędu, np. grupa konsumentów preferująca określoną markę produktu, region geograficzny (jako wynik agregacji podregionów).

3. Metody selekcji zmiennych symbolicznych

W literaturze przedmiotu problematyka wyboru zmiennych podejmowana jest w pracach autorów, takich jak: Milligan [1994], Gnanadesikan, Kettenring, Tsao [1995], Makarenkov, Legendre [2001], Guyon i Elisseeff [2003] i Walesiak [2005].

W pracy Walesiaka [2005] wyróżniono trzy najważniejsze podejścia w selekcji zmiennych (cyt. za [Walesiak 2005, s. 116]):

- Fowlkes, Gnanadesikan i Kettenring [1987; 1988] zaproponowali procedurę doboru zmiennych, zwaną w analizie regresji procedurą selekcji „w przód", powiązaną z hierarchicznymi metodami klasyfikacji,
- Sokołowski [1982, s. 12-13, 50-51] zaproponował miarę zdolności grupowania dla indywidualnych zmiennych i zestawu zmiennych,
- Carmone, Kara i Maxwell [1999] zaproponowali heurystyczną procedurę doboru zmiennych (*HINoV*) powiązaną z metodami klasyfikacji i skorygowanym indeksem Randa.

Metody te mogą znaleźć zastosowanie jedynie w sytuacji, gdy w zbiorze zmiennych mamy do czynienia ze zmiennymi klasycznymi mierzonymi na mocnych skalach pomiaru (zmienne ilorazowe i przedziałowe). Dodatkowo metody te wymagają, aby wartości zmiennych poddane zostały normalizacji. Praca Walesiaka [2005] zawiera propozycję modyfikacji metody *HINoV*, która umożliwia analizowanie zmiennych niemetrycznych.

Na potrzeby selekcji zmiennych w analizie danych symbolicznych proponuje się trzy główne podejścia:

1. Przekształcenie zmiennych symbolicznych w klasyczne (o mocnej lub słabej skali pomiaru), które mają pożądane właściwości.
2. Modyfikację metod selekcji w taki sposób, aby było możliwe analizowanie zmiennych symbolicznych.
3. Zastosowanie metod opracowanych w ramach analizy danych symbolicznych.

Pierwszy sposób, choć atrakcyjny z punktu widzenia otrzymywania w efekcie zmiennych klasycznych o pożądanych właściwościach, wiąże się z utratą części lub całości informacji (zob. np. [Pełka 2009]). Dodatkowo w niektórych przypadkach transformacja zmiennych symbolicznych w klasyczne może być nieuzasadniona.

W pracy Walesiaka i Dudka [2008] zaproponowano rozszerzenie metody *HINoV* na zmienne symboliczne, których realizacją jest przedział liczbowy – $HINoV^S$. W procedurze należy zastosować metody klasyfikacji obiektów symbolicznych bazujące na:

a) tablicy danych symbolicznych, tj. metody taksonomii symbolicznej, np. SCLUST, *k*-średnich, Verde (zob. [Bock, Diday 2000]), lub

b) macierzy odległości, w tym:

- metody taksonomii numerycznej, np. metody hierarchiczne (np. metoda Warda), metody optymalizacyjne (np. metoda *k*-medoidów) bądź
- metody taksonomii symbolicznej, np. metody hierarchiczne (np. klasyfikacja Brito, klasyfikacja podziałowa oparta na kryteriach Chavent), metody optymalizacyjne (np. DCLUST).

Stosując metody klasyfikacji bazujące na macierzy odległości, do pomiaru podobieństwa obiektów należy zastosować miary odległości obiektów symbolicznych, np. miary Ichino-Yaguchiego, de Carvalha (zob. [Bock, Diday 2000]).

Wśród metod selekcji opracowanych na gruncie analizy danych symbolicznych są m.in.: metoda grafowa Ichino, metoda selekcji zaproponowana w pracy Talavery [2000]. Metoda, którą zaproponował Talavera (zob. [2000, s. 23]), pozwala analizować wyłącznie zmienne symboliczne wielowariantowe. W swojej konstrukcji wykorzystuje ona prawdopodobieństwo warunkowe (zob. [Talavera 2000, s. 23--25]).

Metoda grafowa Ichino oprócz zmiennych symbolicznych w postaci przedziałów liczbowych pozwala analizować również zmienne w postaci listy wariantów oraz listy wariantów z wagami. W dalszej części przedstawione i porównane zostaną metoda grafowa Ichino oraz metoda *HINoV*^{S}.

4. Charakterystyka metody *HINoV*^{S} oraz metody grafowej Ichino

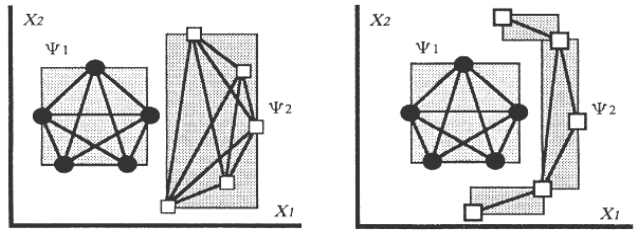
Algorytm metody *HINoV*^{S} dla danych symbolicznych można przedstawić za pomocą następujących kroków (zob. [Walesiak, Dudek 2008]):

1. Wyznacz tablicę danych symbolicznych zawierającą obserwację m symbolicznych zmiennych interwałowych w zbiorze n obiektów.
2. Dla każdej zmiennej jedną z metod bazujących na macierzach odległości (np. klasyfikacją hierarchiczną), wykorzystując miary odległości odpowiednie dla danych symbolicznych (np. miarę Hausdorffa), przeprowadź klasyfikację obiektów na ustaloną *a priori* liczbę klas.
3. Oblicz skorygowany indeks Randa dla wszystkich kombinacji par podziałów.
4. Zestaw policzone wartości skorygowanej miary Randa w macierz danych.
5. Wyznacz sumę wartości dla każdego wiersza (lub kolumny) – *topri*.
6. Uporządkuj malejąco uzyskane wartości i skonstruuj wykres osypiska.
7. Wyeliminuj zmienne zakłócające strukturę klas.
8. Przeprowadź klasyfikację zbioru obiektów symbolicznych.

Metoda grafowa Ichino bazuje na pojęciu sumy i iloczynu kartezjańskiego oraz grafie wzajemnego sąsiedztwa (zaproponowanym w artykule Ichino i Sklanskiego [1985]). Ideę iloczynu i sumy kartezjańskiej dla danych symbolicznych w polskiej literaturze przedstawiono w pracy Dudka [2004]. Dwa obiekty symboliczne A_1 i A_2 określa się jako wzajemnie sąsiadujące (*mutual neighbours*) względem zbioru obiektów symbolicznych $\mathrm{B} = \{B_1, B_2, \ldots, B_m\}$, jeśli zachodzi warunek:

$$\underset{B_i \in \mathrm{B}}{\forall} B_i \otimes \left(A_1 \oplus A_2 \right) = \varnothing . \tag{1}$$

Zbiór obiektów symbolicznych $\mathrm{A} = \{A_1, A_2, \ldots, A_l\}$ nazywamy grafem wzajemnego sąsiedztwa (*mutual neighbourhood graph*) względem zbioru obiektów symbolicznych $\mathrm{B} = \{B_1, B_2, \ldots, B_m\}$, jeśli każda para obiektów ze zbioru A jest wzajemnie sąsiadująca względem zbioru B. Przykładowe grafy wzajemnego sąsiedztwa przedstawiono na rys. 1.



a) klasy dobrze separowalne b) klasy słabo separowalne

Rys. 1. Grafy wzajemnego sąsiedztwa

Źródło: opracowanie własne.

Procedura metody grafowej Ichino składa się z następujących kroków:

1. Dla każdej kombinacji zmiennych-kandydatek znajdź wszystkie grafy wzajemnego sąsiedztwa.

2. Wyznacz liczbę wszystkich par obiektów wewnątrz grafów. Jeśli liczba obiektów w grafie wynosi n, to liczba wszystkich możliwych par jest dana wzorem:

$$\binom{n}{2} = \frac{n \cdot (n-1)}{2} . \tag{2}$$

3. Dla każdej liczby zmiennych-kandydatek znajdź taką kombinację, dla której suma obliczona w punkcie 2 jest największa.

4. Wybierz tę kombinację zmiennych, dla której przyrost wartości obliczonej w punkcie 3 w stosunku do $(k-1)$ kandydatek jest największy.

5. Przeprowadź klasyfikację zbioru obiektów, uwzględniając kombinację zmiennych wybraną w punkcie 4.

Porównanie metody grafowej Ichino z modyfikacją metody *HINoV* zawarto w tab. 1.

Tabela 1. Porównanie metody grafowej Ichino i $HINoV^S$

Wyszczególnienie	Metoda $HINoV^S$	Metoda grafowa Ichino
Rodzaj zmiennych symbolicznych	interwałowe, wielowariantowe, wielowariantowe z wagami, strukturalne	interwałowe, wielowariantowe, wielowariantowe z wagami
Metoda klasyfikacji	bazująca na tablicy danych symbolicznych bądź na macierzy odległości	brak
Liczba zmiennych	dowolna	mało liczna ze względu na złożoność obliczeniową
Kryterium	suma wartości skorygowanego indeksu Randa dla par zmiennych	liczba obiektów w grafach wzajemnego sąsiedztwa
Kryterium wyboru zmiennych	największy przyrost wartości skorygowanego indeksu Randa (*stopri)*	największy przyrost liczby par obiektów wewnątrz grafów wzajemnego sąsiedztwa

Źródło: opracowanie własne.

5. Wyniki symulacji

Na potrzeby porównania metody grafowej Ichino z metodą *HINoV*^{S} przygotowano cztery modele o znanej strukturze klas (do generowania modeli wykorzystano polecenie `cluster.Gen` z pakietu `clusterSim` dla R):

Model I – 100 obiektów tworzących dwie klasy, o wydłużonym kształcie, opisywane dwiema zmiennymi interwałowymi. Zmienne te są losowane niezależnie z dwuwymiarowego rozkładu normalnego o średnich (0, 0), (1, 5) i macierzy kowariancji $\sum\left(\sigma_{jj}=1,\sigma_{jl}=-0{,}9\right)$. Model ten nie zawiera zmiennych zakłócających czy obserwacji odstających.

Model II – 100 obiektów tworzących trzy klasy, o wydłużonym kształcie, opisywanych trzema zmiennymi interwałowymi. Zmienne te są losowane niezależnie z wielowymiarowego rozkładu normalnego o średnich (–1,5, 6, 3), (3, 12, –6), (4,5, 18, –9) oraz macierzy kowariancji, gdzie $\sigma_{jj}=1$ dla $\left(1\le j\le 3\right)$, $\sigma_{12}=\sigma_{13}=-0{,}9$ oraz $\sigma_{23}=0{,}9$. Do zbioru zmiennych dodano dwie zmienne zakłócające, model nie zawiera natomiast obserwacji odstających.

Model III – 100 obiektów tworzących pięć niezbyt dobrze separowanych klas opisywanych przez dwie zmienne interwałowe. Zmienne te są losowane niezależnie z dwuwymiarowego rozkładu normalnego o średnich (5, 5), (–3, 3), (3, –3), (0, 0), (–5, –5) oraz macierzy kowariancji $\sum\left(\sigma_{jj}=1,\sigma_{jl}=0{,}9\right)$. Do zbioru zmiennych dodano jedną zmienną zakłócającą, model dodatkowo zawiera 10 obserwacji odstających.

Model IV – 100 obiektów tworzących pięć niezbyt dobrze wyodrębnionych klas opisywanych trzema zmiennymi interwałowymi. Zmienne te są losowane niezależnie z wielowymiarowego rozkładu normalnego o średnich (5, 5, 5), (–3, 3, –3), (3, –3, 3), (0, 0, 0), (–5, –5, –5) oraz macierzy kowariancji, gdzie $\sigma_{jj}=1$ dla $\left(1\le j\le 3\right)$ i $\sigma_{jl}=0{,}9$ dla $\left(1\le j\ne l\le 3\right)$. Do zbioru zmiennych dodano dwie zmienne zakłócające, model nie zawiera natomiast obiektów odstających.

Tabela 2. Wyniki symulacji

Numer modelu	Metoda grafowa Ichino		Metoda *HINoV*^{S}	
	zmienne	wartość indeksu.S	zmienne	wartość indeksu.S
I	{1, 2}	0,8319	{1, 2}	0,8319
II	{1, 2, 3}	0,6350	{1, 2, 3}	0,6350
III	{1, 2, 3}	0,4809	{1, 2}	0,4944
IV	{1, 2, 3}	0,5543	{1, 2, 3}	0,5543

Źródło: obliczenia własne w programie R.

W procedurze klasyfikacji wykorzystano odległość Hausdorffa dla zmiennych interwałowych oraz hierarchiczne metody klasyfikacji i metodę pam. W tabeli 2 zestawiono najlepsze wyniki otrzymane w przypadku danego modelu i przekroju wszystkich zastosowanych metod.

W tabeli 2 zestawiono wyniki porównania obu metod z wykorzystaniem indeksu sylwetkowego (*silhouette index*) dla ostatecznej klasyfikacji z wykorzystaniem zmiennych wskazywanych przez daną metodę.

6. Wnioski

Przeprowadzone analizy dla zbiorów danych opisywanych wyłącznie zmiennymi interwałowymi wskazują, że metoda grafowa Ichino w większości przypadków poprawnie identyfikuje te same zmienne, które są wybierane z wykorzystaniem metody $HINoV^S$.

Obie metody wskazały podobne kombinacje zmiennych w modelach I, II i IV. W przypadku modelu II i IV obydwie metody wskazały jednakowe zmienne zakłócające. W modelach I, II, IV przy zastosowaniu kombinacji zmiennych wybranych przez obie metody, indeks sylwetkowy wskazuje na fakt wykrycia poważnej struktury klas. Natomiast w przypadku modelu III jedynie metoda $HINoV^S$ wykryła zmienną zakłócającą. Jednakże w przypadku tego modelu wykryta struktura klas jest słaba (w sensie wartości indeksu sylwetkowego).

Istotnym ograniczeniem metody grafowej Ichino jest fakt, że zbiór zmiennych powinien być mało liczny ze względu na jej złożoność obliczeniową. Z kolei w metodzie $HINoV^S$ obliczenia należy powtórzyć dla różnej liczby klas, co zwiększa pracochłonność badana. Jeśli zatem badacz nie dysponuje wiedzą na temat liczby klas, korzystniej jest zastosować metodę Ichino. Z kolei gdy zbiór obiektów jest opisany dużą liczbą zmiennych symbolicznych, należy wybrać metodę $HINoV^S$.

Literatura

Bock H.H., Diday E. (red.), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin-Heidelberg 2000.

Carmone F.J, Kara A., Maxwell S., *HINoV: a new method to improve market segment definition by indentifying noisy variables*, „Journal of Marketing Research" 1999, listopad, vol. 36, s. 501-509.

Dudek A., *Miary podobieństwa obiektów symbolicznych. Odległość Ichino-Yaguchiego*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1021, AE, Wrocław 2004.

Fowleks E.B., Gnanadesikan R., Kettenring J.R., *Variable Selection in Clustering and other Contexts*, [w:] *Design, Data, and Analysis*, red. C.L. Mallows, Wiley, New York, Toronto 1987.

Fowleks E.B., Gnanadesikan R., Kettenring J.R., *Variable selecting in clustering,* „Journal of Classification" 1988 vol. 5, s. 205-228.

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

Gnanadesikan R., Kettenring J.R., Tsao S.L., *Weighting and selection of variables for cluster analysis*, „Journal of Classification" 1995 vol. 12.

Guyon I., Elisseeff A., *An introduction to variable and feature selection*, „Journal of Machine Learning Research" 2003 no 3.
Ichino M., Sklansky J., *The relative neighborhood graph for mixed feature variables*, „Pattern Recognition" 1985 vol. 18, no 2.
Ichino M., *Feature Selection for Symbolic Data Classification*, [w:] *New approaches in Classification and Data Analysis*, E. Diday (red.), Springer-Verlag, Berlin-Heidelberg 1994.
Makarenkov V., Legendre P., *Optimal variable weighting for ultrametric and additive trees and k-means partitioning methods and software*, „Journal of Classification" 2001 vol. 18.
Milligan G.W., *Issues in applied classification: selection of variables to cluster*, „Classification Society of North America Newsletter", listopad 1994, Issue 37.
Pełka M., *Porównanie strategii klasyfikacji danych symbolicznych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu (w druku), 2009.
Sokołowski A., *Empiryczne testy istotności w taksonomii*, Zeszyty Naukowe AE w Krakowie, seria specjalna: Monografie nr 108, AE, Kraków 1982.
Talavera L., *Dependency-based feature selection for clustering symbolic data*, „Intelligent Data Analysis" 2000 vol. 4, Issue 1.
Walesiak M., *Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, AE, Wrocław 2005.
Walesiak M., Dudek A., *Identification of Noisy Variables for Nonmetric and Symbolic Data in Cluster Analysis*, [w:] C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker, *Data Analysis, Machine Learning and Applications*, *Studies in Classification, Data Analysis and Knowledge Organization*, Springer-Verlag, Berlin-Heidelberg 2008.

METHODS OF SELECTING SYMBOLIC VARIABLES FOR CLUSTERING PROBLEMS

Summary: The aim of this paper is to present and compare two symbolic variable selecting procedures – Ichino's graph feature selection and modification of *HINoV*. The paper presents basic terms of symbolic data analysis and possible selecting methods for this kind of data.

In the empirical part, simulation experiment results are compared based on artificial data generated by `cluster.Gen` procedure from `clusterSim` package for R software.