PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 84
Ekonometria 27 2009

Andrzej Dudek, Justyna Wilk
Uniwersytet Ekonomiczny we Wrocławiu

METODY DOBORU ZMIENNYCH W PROCESIE KLASYFIKACJI OBIEKTÓW SYMBOLICZNYCH

Streszczenie: Jednym z kluczowych etapów procedury klasyfikacyjnej jest wybór zmiennych. W procedurze klasyfikacji należy uwzględnić tylko te zmienne, które mają zdolność do grupowania obiektów w jednorodne klasy. Zagadnienie doboru zmiennych jest szczególnie utrudnione w sytuacji, gdy zbiór obiektów jest opisany zmiennymi symbolicznymi. Złożona struktura zmiennych symbolicznych uniemożliwia aplikację klasycznych procedur. Problematyka wyboru zmiennych symbolicznych jest podejmowana w niewielu pracach. Celem artykułu jest przedstawienie podejść w procesie selekcji zmiennych symbolicznych, a także prezentacja metody selekcji zmiennych symbolicznych opracowanej przez Ichino.

Słowa kluczowe: klasyfikacja, dobór zmiennych, analiza danych symbolicznych.

1. Wstęp

W typowej procedurze klasyfikacyjnej wyodrębnia się następujące etapy (por. [Walesiak 2004]):

1. Wybór obiektów do klasyfikacji.
2. Wybór zmiennych charakteryzujących obiekty.
3. Wybór formuły normalizacji wartości zmiennych.
4. Wybór miary odległości i metody klasyfikacji.
5. Ustalenie liczby klas i ocena wyników klasyfikacji.
6. Opis (interpretacja) i profilowanie klas.

Proces klasyfikacji polega na zbadaniu podobieństwa obiektów i połączeniu w jednorodne klasy obiektów do siebie podobnych. Klasy tworzone są na podstawie zestawu zmiennych opisujących obiekty. Jednym z najważniejszych etapów procedury klasyfikacyjnej jest zatem wybór zmiennych. W procedurze klasyfikacji należy uwzględnić tylko te zmienne, które mają zdolność do grupowania obiektów w jednorodne klasy.

Zagadnienie doboru zmiennych jest szczególnie utrudnione w sytuacji, gdy zbiór obiektów jest opisany zmiennymi symbolicznymi. Złożona struktura tego rodzaju zmiennych uniemożliwia aplikację klasycznych metod doboru zmiennych (zob. np.

[Walesiak 2005; Carmone, Kara, Maxwell 1999; Gnanadesikan, Kettenring, Tsao 1995]).

Problematyka selekcji zmiennych symbolicznych podejmowana jest w niewielu pracach [Ichino 1994; Walesiak, Dudek 2008]. Celem artykułu jest przedstawienie podejść w selekcji zmiennych symbolicznych i metody grafowej Ichino.

W pierwszej części artykułu przedstawiono pojęcia obiektu symbolicznego i zmiennej symbolicznej. W kolejnej części zaproponowano podejścia w procesie doboru zmiennych symbolicznych. W części trzeciej zaprezentowano metodę selekcji zmiennych symbolicznych opracowaną przez Ichino, bazującą na grafie wzajemnego sąsiedztwa. W ostatniej części zaprezentowano przykład empiryczny. Porównano rezultaty klasyfikacji obiektów symbolicznych w przypadku przeprowadzania doboru zmiennych symbolicznych metodą grafową Ichino oraz z pominięciem etapu wyboru zmiennych.

2. Pojęcie obiektu symbolicznego i zmiennej symbolicznej

Bock, Diday i in. [2000, s. 2] wśród podstawowych typów zmiennych symbolicznych wymieniają zmienne, których realizacją jest:

- przedział liczbowy (*interval-valued variable*),
- lista wariantów (tzw. zmienna wielowariantowa, *multi-valued variable*),
- lista wariantów z wagami (*multi-valued variable with weights*).

Realizacją zmiennej symbolicznej w postaci przedziału liczbowego może być przedział rozłączny, np. miesięczny dochód netto respondenta w zł (np. (0, 3000), [3000, 6000), [6000, ∞)) lub nierozłączny, np. planowana wysokość wydatków na zakup samochodu w tys. zł (np. respondent 1: [10, 30], respondent 2: [30, 60], respondent 3: [50, 70]). Szczególnym przypadkiem tej zmiennej jest klasyczna zmienna metryczna.

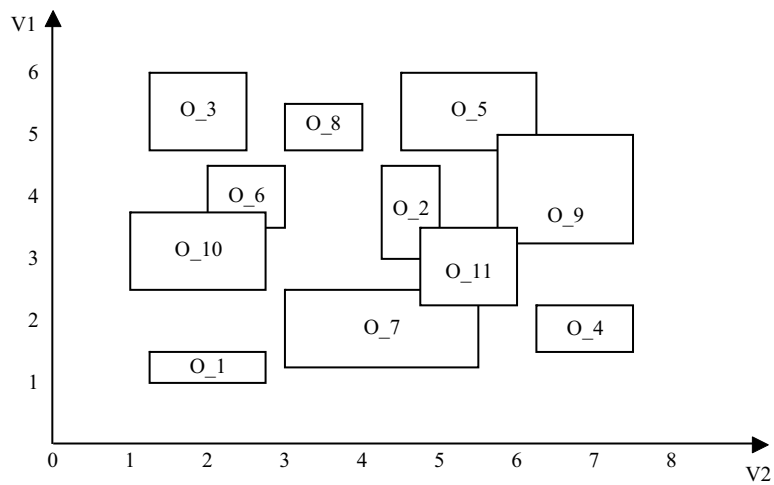
Zmienna wielowariantowa dopuszcza występowanie wielu kategorii lub wartości dla pojedynczego obiektu, np. znajomość języków obcych (np. respondent 1 włada językami: angielskim, francuskim i niemieckim, natomiast respondent 2 − językami francuskim i włoskim). Zmienne w ujęciu klasycznym, mierzone na słabych skalach pomiaru, są szczególnym przypadkiem zmiennej wielowariantowej.

Podobną zmienną jest lista wariantów z wagami, z tym że wariantom nadane są stopnie ważności, prawdopodobieństwa, częstości lub udziały. Przykładem zmiennej jest struktura wydatków konsumenta na zakupy w hipermarketach, np. respondent 1: {Auchan (50%), Real (30%), Tesco (20%)}, a respondent 2: {Tesco (80%), Carrefour (20%)}.

Analiza danych symbolicznych daje możliwość uwzględnienia powiązań między zmiennymi. Zmienne pozostające w relacji tworzą tzw. zmienną strukturalną (*dependent variable*). Można wskazać zmienne o strukturze taksonomicznej, hierarchicznej i logicznej. Różnią się one charakterem relacji, jaka zachodzi między poziomami zmiennej.

Obiekty opisane zmiennymi symbolicznymi określa się obiektami symbolicznymi. Wyróżnia się dwa rodzaje obiektów symbolicznych:

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



Rys. 1. Zbiór obiektów symbolicznych w przestrzeni dwuwymiarowej

Źródło: opracowanie własne.

Przykładowy zbiór 11 obiektów symbolicznych opisanych dwoma zmiennymi symbolicznymi, których realizacją jest przedział liczbowy (V1 i V2), prezentuje rys. 1.

3. Podejścia w selekcji zmiennych symbolicznych

Problematyka doboru zmiennych w procedurze klasyfikacji zbioru obiektów podejmowana jest w pracach: [Milligan 1994; Gnanadesikan, Kettenring, Tsao 1995; Makarenkov, Legendre 2001; Guyon, Elisseeff 2003; Walesiak 2005]. W zagadnieniu wyboru zmiennych wyróżnia się trzy podejścia; są nimi (por. [Walesiak 2005]):

1. Wprowadzenie zróżnicowanych wag zmiennym wyrażających ich relatywną ważność.

2. Selekcja zmiennych, tzn. dobór mniejszej liczby zmiennych przez eliminację tych, które nie mają zdolności dyskryminacji zbioru obiektów.

3. Zastąpienie oryginalnych zmiennych nowymi „sztucznymi" zmiennymi o pożądanych właściwościach (wykorzystuje się tutaj analizę czynnikową i analizę głównych składowych).

Walesiak [2005] wśród najważniejszych metod selekcji zmiennych wyodrębnia:
- procedurę doboru zmiennych, zaproponowaną przez Fowlkesa, Gnanadesikana i Kettenringa [1987; 1988], zwaną w analizie regresji procedurą selekcji „w przód", powiązaną z hierarchicznymi metodami klasyfikacji,
- metodę wykorzystującą miarę zdolności grupowania dla indywidualnych zmiennych i zestawu zmiennych [Sokołowski 1992, s. 12-13, 50-51],
- heurystyczną procedurę doboru zmiennych (*HINoV*) [Carmone, Kara, Maxwell 1999] powiązaną z metodą *k*-średnich i ze skorygowanym indeksem Randa.

Wymienione procedury selekcji zmiennych są adekwatne w sytuacji, gdy w zbiorze znajdują się wyłącznie zmienne klasyczne. Większość tych metod odnosi się do zmiennych mierzonych na mocnych skalach pomiaru. Ponadto metody te wymagają, aby wartości zmiennych poddane zostały uprzednio normalizacji. W pracy Walesiaka [2005] zaproponowana została modyfikacja metody *HINoV* umożliwiająca analizę zmiennych niemetrycznych.

Wspomniane ograniczenia stwarzają trudność stosowania tych metod, gdy zbiór obiektów charakteryzowany jest zmiennymi symbolicznymi. W rozwiązaniu problemu doboru zmiennych symbolicznych można zaproponować następujące podejścia:

1. Przekształcenie zmiennych symbolicznych na zmienne klasyczne (mierzone na mocnych lub słabych skalach pomiaru).
2. Modyfikacja metody selekcji zmiennych tak, aby możliwe było analizowanie zmiennych symbolicznych.
3. Zastosowanie metod opracowanych w ramach analizy zmiennych symbolicznych.

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

Wyszczególnienie	Metoda $HINoV^S$	Metoda grafowa Ichino
Rodzaj zmiennych symbolicznych	Przedziały liczbowe	Przedziały liczbowe, zmienne wielowariantowe, listy wariantów z wagami
Metoda klasyfikacji	Bazująca na macierzy odległości	Brak
Rozpatrywane wartości	Suma wartości skorygowanego indeksu Randa dla par zmiennych	Liczba obiektów w grafach wzajemnego sąsiedztwa
Kryterium wyboru zmiennych	Kryterium osypiska	Największy przyrost liczby par obiektów wewnątrz grafów wzajemnego sąsiedztwa

Źródło: opracowanie własne na podstawie prac: [Walesiak, Dudek 2008; Ichino 1994].

Sposób pierwszy łączy się z utratą części informacji zawartych w zmiennych symbolicznych. Ponadto w niektórych przypadkach transformacja zmiennych symbolicznych może się okazać niemożliwa lub nieuzasadniona.

Możliwość zastosowania podejścia drugiego jest ściśle związana z algorytmem danej metody. W pracy Walesiaka i Dudka [2008] zaproponowano rozszerzenie me-

tody *HINoV* na zmienne symboliczne, których realizacją jest przedział liczbowy – *HINoV*^{S}. Przeprowadzone badania symulacyjne potwierdzają użyteczność tej metody w analizie danych symbolicznych.

Wśród metod opracowanych na gruncie analizy danych symbolicznych jedyną propozycją procedury selekcji zmiennych jest metoda grafowa Ichino. Oprócz zmiennych symbolicznych w postaci przedziałów liczbowych pozwala ona analizować również zmienne w postaci listy wariantów oraz listy wariantów z wagami. Nie uwzględnia natomiast zmiennych strukturalnych. Porównanie metody *HINoV*^{S} i metody grafowej Ichino prezentuje tab. 1.

4. Charakterystyka metody grafowej Ichino

Metoda grafowa Ichino bazuje na pojęciu sumy i iloczynu kartezjańskiego oraz grafie wzajemnego sąsiedztwa. Suma kartezjańska (*cartesian join*) oznaczona jest symbolem $\oplus$, a sposób jej wyznaczenia zależy od rodzaju zmiennych symbolicznych. Jeśli realizacją zmiennej symbolicznej jest:

a) przedział liczbowy, to:

$$A \oplus B = (a_1, a_2) \oplus (b_1, b_2) = \left(\min(a_1, b_1),\ \max(a_2, b_2)\right), \tag{1}$$

gdzie: A, B – zbiór realizacji zmiennej symbolicznej dla i-tego, j-tego obiektu symbolicznego, $A = (a_1, a_2)$, $B = (b_1, b_2)$ – przedział liczbowy,

b) lista wariantów, to:

$$A \oplus B = A \cup B, \tag{2}$$

gdzie: $A = \{a_1, a_2, ..., a_f\}$, $B = \{b_1, b_2, ..., b_f\}$ – zbiór wariantów $(r = 1, ..., f)$,

c) lista wariantów z wagami, to:

$$A \oplus B = A \cup B, \tag{3}$$

gdzie: $A = \{a_1(p(a_1)), a_2(p(a_2)), ..., a_f(p(a_f))\}$, $B = \{b_1(p(b_1), b_2(p(b_2)), ...,$ $b_f(p(b_f))\}$ – zbiór wariantów z wagami, $p(a_f)$, $p(b_f)$ – wagi wariantów, $p(a_1) + p(a_2) + ... + p(a_f) = 1$, $p(b_1) + p(b_2) + ... + p(b_f) = 1$.

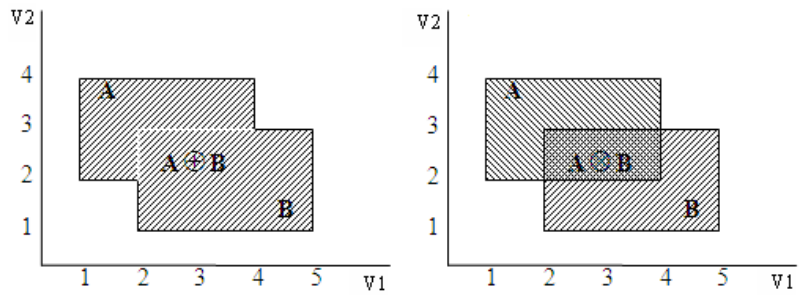
Iloczyn kartezjański (*Cartesian meet*) jest oznaczony symbolem $\otimes$. Dla każdego rodzaju zmiennych symbolicznych jest definiowany wzorem:

$$A \otimes B = A \cap B. \tag{4}$$

Ilustrację sumy i iloczynu kartezjańskiego dla dwóch obiektów symbolicznych opisanych dwiema zmiennymi symbolicznymi, których realizacją jest przedział liczbowy, prezentuje rys. 2.

Dwa obiekty symboliczne X_1 i X_2 określa się wzajemnie sąsiadującymi (*mutual neighbours*) względem zbioru obiektów symbolicznych $Y = \{Y_1, Y_2, \ldots, Y_m\}$ $(i = 1, ..., m)$, jeśli zachodzi warunek:

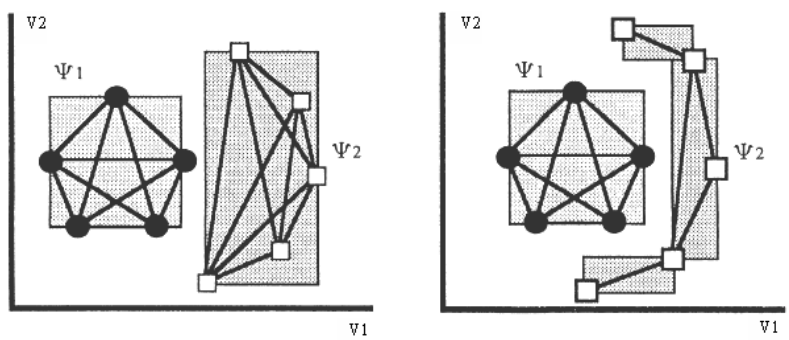
$$\underset{Y_i \in Y}{\forall} Y_i \otimes \left(X_1 \oplus X_2 \right) = \varnothing. \tag{5}$$



Rys. 2. Suma kartezjańska i iloczyn kartezjański

Źródło: opracowanie własne.

Zbiór obiektów symbolicznych $X = \{X_1, X_2, \ldots, X_l\}$ $(i = 1, ..., l)$ nazywamy grafem wzajemnego sąsiedztwa (*mutual neighbourhood graph*) względem zbioru obiektów symbolicznych $Y = \{Y_1, Y_2, \ldots, Y_m\}$, jeśli każda para obiektów ze zbioru X jest wzajemnie sąsiadująca względem zbioru Y. Ilustrację grafów wzajemnego sąsiedztwa prezentuje rys. 3.



a) klasy dobrze separowalne b) klasy słabo separowalne

Rys. 3. Grafy wzajemnego sąsiedztwa

Źródło: [Ichino 1994].

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:

$$\binom{n}{2} = \frac{n \cdot (n-1)}{2}. \tag{6}$$

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 $k-1$ zmiennych-kandydatek ($k = 1, ..., z$) jest największy.

5. Przykład empiryczny selekcji zmiennych symbolicznych

Za pomocą funkcji `cluster.Gen` pakietu `cluster.Sim` środowiska R wygenerowano zbiór 150 obiektów symbolicznych opisanych sześcioma zmiennymi symbolicznymi, których realizacją jest przedział liczbowy. W zbiorze znalazły się trzy zmienne zakłócające strukturę klas. W selekcji zmiennych symbolicznych zastosowano metodę grafową Ichino. Liczbę par obiektów wewnątrz grafów wzajemnego sąsiedztwa dla różnych kombinacji zmiennych-kandydatek prezentuje tab. 2.

Tabela 2. Wyniki uzyskane w kroku 3 algorytmu metody grafowej Ichino

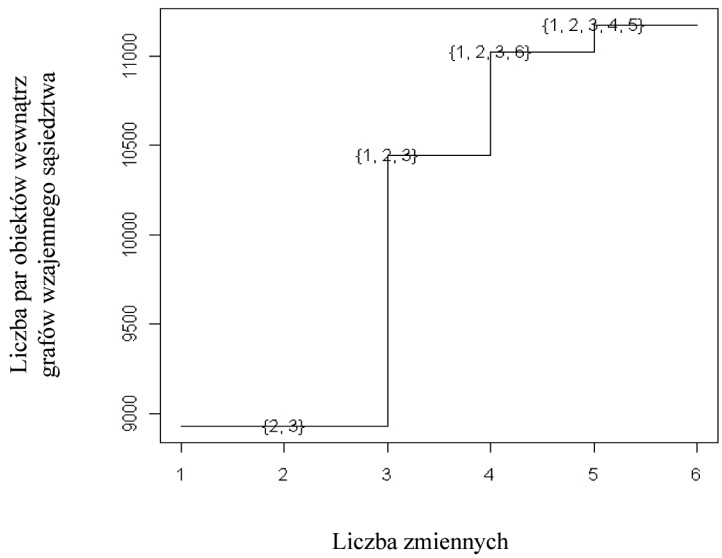
Liczba zmiennych	Kombinacja zmiennych	Liczba par obiektów wewnątrz grafów wzajemnego sąsiedztwa
2	{2,3}	8 929
3	{1,2,3}	10 443
4	{1,2,3,6}	11 026
5	{1,2,3,4,5}	11 175
6	{1,2,3,4,5,6}	11 175

Źródło: opracowanie własne.

Przyrosty liczby par obiektów wewnątrz grafów wzajemnego sąsiedztwa ilustruje rys. 5.

Największy przyrost można zaobserwować dla kombinacji trzech zmiennych. Zatem najwyższą moc dyskryminacji zbioru obiektów mają zmienne V_1, V_2 i V_3. Identyczne wyniki uzyskano stosując metodę *HINoV*^{S}.

W dalszej części badania przeprowadzono klasyfikację zbioru obiektów symbolicznych z wykorzystaniem odległości Ichino-Yaguchiego (zob. [Bock, Diday (red.) 2000]) i metody *k*-medoidów. Klasyfikację przeprowadzono w dwóch wariantach – na podstawie zestawu sześciu zmiennych oraz na podstawie trzech zmiennych wskazanych przez metodę grafową Ichino. Otrzymaną strukturę klas porównano z ich rzeczywistą strukturą za pomocą skorygowanego indeksu Randa. Wartość indeksu dla danych zawierających wszystkie sześć zmiennych wynosiła 0,52, natomiast w przypadku trzech zmiennych V_1, V_2 i V_3 wyniosła 1,00, co oznacza pełną zgodność.



Rys. 4. Przyrosty liczby par obiektów wewnątrz grafów wzajemnego sąsiedztwa

Źródło: opracowanie własne.

6. Podsumowanie

W artykule wskazano podejścia w doborze zmiennych symbolicznych w procesie klasyfikacji zbioru obiektów. Zaprezentowano metodę selekcji zmiennych symbolicznych opracowaną przez Ichino. Metoda grafowa Ichino, w przeciwieństwie do metody *HINoV*^{S}, daje możliwość analizowania nie tylko zmiennych symbolicznych, których realizacją jest przedział liczbowy, ale również zmiennych w postaci listy wariantów i listy wariantów z wagami. Należy jednak zaznaczyć, iż ze względu na sposób wyboru optymalnej kombinacji zmiennych metoda grafowa Ichino ma wykładniczą złożoność obliczeniową i jest odpowiednia w sytuacji mało licznych zbiorów danych symbolicznych.

Literatura

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

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

Fowlkes E.B., Gnanadesikan R., Kettenring J.R., *Variable selection in clustering and other contexts*, [w:] C.L. Mallows (red.), *Design, Data, and Analysis*, Wiley, New York, Toronto1987.

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

Gnanadesikan R., Kettenring J.R., Tsao S.L., *Weighting and selection of variables for cluster analysis*, „Journal of Classification" 1995, vol. 12, s. 113-136.

Guyon I., Elisseeff A., *An introduction to variable and feature selection*, „Journal of Machine Learning Research" 2003, 3, s. 1157-1182.

Ichino M., *Feature selection for symbolic data classification*, [w:] E. Diday (red.), *New Approaches in Classification and Data Analysis*, Springer-Verlag, Berlin-Heidelberg 1994, s. 387-394.

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, s. 245-271.

Milligan G.W., *Issues in applied classification: selection of variables to cluster*, „Classification Society of North America Newsletter" 1994, issue 37.

Sokołowski A., *Empiryczne testy istotności w taksonomii*, Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, seria: Monografie nr 108, AE, Kraków 1992.

Walesiak M., *Problemy decyzyjne w procesie klasyfikacji zbioru obiektów*, [w:] J. Dziechciarz, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1010, AE, Wrocław 2004, s. 52-71.

Walesiak M., *Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, AE, Wrocław 2005, s. 106-118.

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 (red.), *Data Analysis, Machine Learning and Applications*, *Studies in Classification, Data Analysis and Knowledge Organization*, Springer-Verlag, Berlin-Heidelberg 2008, s. 84-92.

FEATURE SELECTION METHODS IN SYMBOLIC OBJECTS CLASSIFICATION PROCESS

Summary: Selecting variables is one of the most important steps in cluster analysis procedure. Variables used to characterize a set of objects should be able to discriminate the data. Identifying noisy variables is particularly complicated in case of symbolic data analysis. The complexity of symbolic data makes the direct application of classical procedures for variables selection impossible.

The problem of selecting symbolic variables is not often mentioned in the literature of subject. The aim of the paper is a presentation of symbolic variables selection approaches and a description of Ichino method dedicated for symbolic data analysis.