ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 450 PRACE KATEDRY EKONOMETRII I STATYSTYKI NR 17 2006

MAREK WALESIAK
ANDRZEJ DUDEK

Akademia Ekonomiczna
Wrocław

SYMULACYJNA OPTYMALIZACJA WYBORU PROCEDURY KLASYFIKACYJNEJ DLA DANEGO TYPU DANYCH – CHARAKTERYSTYKA PROBLEMU

1. Wprowadzenie

W literaturze przedmiotu w typowej procedurze klasyfikacyjnej wyodrębnia się osiem etapów[1]: 1) wybór obiektów do klasyfikacji; 2) wybór zmiennych charakteryzujących obiekty; 3) wybór formuły normalizacji wartości zmiennych; 4) wybór miary odległości[2]; 5) wybór metody klasyfikacji; 6) ustalenie liczby klas; 7) walidacja wyników klasyfikacji; 8) opis (interpretacja) i profilowanie klas. Do newralgicznych zalicza się etapy dotyczące wyboru formuły normalizacji wartości zmiennych, miary odległości, metody klasyfikacji i ustalenia liczby klas, które mają w znacznej mierze arbitralny charakter.

W artykule zaprezentowano szczegółową charakterystykę dziewięciu ścieżek w symulacyjnej optymalizacji wyboru procedury klasyfikacyjnej dla danego typu danych, a wyodrębnionych w zależności od typu skali pomiaru zmiennych. Liczba rozpatrywanych wariantów procedury klasyfikacyjnej zależy od liczby formuł normalizacyjnych, liczby typów miar odległości i liczby

1 Por. [13], s. 342–343; [21].

2 Zob. [23].

metod klasyfikacji. Na podstawie zaproponowanego podejścia w artykule M. Wale-siaka i A. Dudka[3] scharakteryzowano program komputerowy clusterSim (opracowany w języku R oraz pomocniczo w języku C++) służący do realizacji wyodrębnionych ścieżek oraz wybrane wyniki obliczeń symulacyjnych.

1. Charakterystyka ścieżek w symulacyjnej optymalizacji wyboru procedury klasyfikacyjnej dla danego typu danych

Punktem wyjścia analizy symulacyjnej jest macierz danych. W zależności od typu skali pomiaru zmiennych wyróżniono dziewięć ścieżek w procedurze symulacyjnej. Przy opracowywaniu poszczególnych ścieżek uwzględniono następujące elementy:

- typ skali pomiaru zmiennych w macierzy danych,
- typ formuły normalizacyjnej dla zmiennych mierzonych na skali przedziałowej i (lub) ilorazowej,
- miary odległości właściwe dla poszczególnych typów skal pomiaru zmiennych,
- typ metody klasyfikacji,
- miernik oceny jakości klasyfikacji.

Liczba rozpatrywanych wariantów procedury klasyfikacyjnej dla zmiennych mierzonych na skali:

1. Ilorazowej równa się 368 (11 formuł normalizacyjnych, 7 typów miar odległości, 8 metod klasyfikacji)[4].

2. Przedziałowej (lub ilorazowej i przedziałowej) równa się 140 (5 formuł normalizacyjnych, 5 typów miar odległości, 8 metod klasyfikacji).

3. Porządkowej równa się 5 (miara odległości GDM2, 5 metod klasyfikacji).

[3] Zob. [23].

[4] Liczba wariantów w punktach 1, 2, 6, 7 nie wynika z prostego przemnożenia liczby formuł normalizacyjnych, typów miar odległości i metod klasyfikacji z uwagi na ograniczenia w ich stosowaniu.

4. Nominalnej wielostanowej równa się 5 (odległość Sokala i Michenera, 5 metod klasyfikacji).

5. Nominalnej binarnej równa się 50 (10 typów miar odległości, 5 metod klasyfikacji).

6. Ilorazowej bez normalizacji równa się 38 (7 typów miar odległości, 8 metod klasyfikacji).

7. Przedziałowej (lub ilorazowej i przedziałowej) bez normalizacji równa się 28 (5 typów miar odległości, 8 metod klasyfikacji).

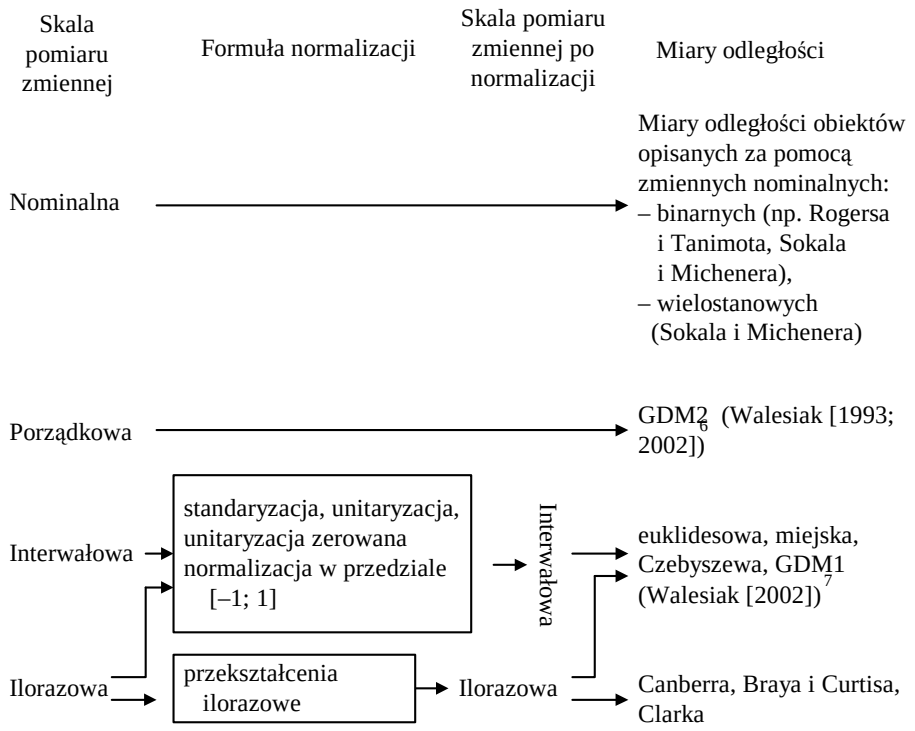
8. Ilorazowej z zastosowaniem metody *k*-średnich równa się 11 (11 formuł normalizacyjnych, 1 metoda klasyfikacji).

9. Przedziałowej (lub ilorazowej i przedziałowej) z zastosowaniem metody *k*-średnich równa się 5 (5 formuł normalizacyjnych, 1 metoda klasyfikacji).

Ścieżki 6 i 7 (dla danych metrycznych bez normalizacji) występują często w badaniach marketingowych opartych na danych otrzymanych na przykład ze skali Likerta lub semantycznej. Wprawdzie są to przykłady skal porządkowych, jednak z uwagi na to, że odstępy między kategoriami odpowiadają w przybliżeniu jednakowym interwałom, traktuje się je w badaniach jako skale metryczne[5].

Przy wyborze miar odległości obiektów opisanych zmiennymi mierzonymi na skali przedziałowej i (lub) ilorazowej należy wziąć pod uwagę zastosowaną formułę normalizacji wartości zmiennych. Klasyfikację formuł normalizacyjnych i miar podobieństwa obiektów z punktu widzenia skal pomiaru zmiennych przedstawiono na rysunku 1.

[5] Zob. [7], s. 75.



Rys. 1. Klasyfikacja formuł normalizacyjnych i miar odległości obiektów z punktu widzenia skal pomiaru zmiennych

Źródło: [22].

Rozważania w artykule ograniczono do najczęściej wykorzystywanych metod klasyfikacji, czyli metody *k-medoids* i siedmiu hierarchicznych metod aglomeracyjnych opartych na macierzy odległości oraz metody *k*-średnich opartych na macierzy danych.

Indeksy oceny jakości klasyfikacji pozwalające na wyznaczenie optymalnej liczby klas dzieli się na globalne i lokalne[8]. Indeksy globalne $G(u)$ są oparte

[6] Zob. [1]; [20].

[7] *Ibidem.*

[8] Zob. [6], s. 61.

na kompletnym zbiorze danych klasyfikacyjnych (zazwyczaj (u = 2, …, n – 1), gdzie u oznacza liczbę klas, a n – liczbę obiektów) i z ich wykorzystaniem poszukuje się optymalnego podziału badanego zbioru obiektów na klasy.

Indeksy lokalne $L(u)$ są oparte tylko na podzbiorze zbioru danych klasyfikacyjnych (zazwyczaj rozpatruje się dwa sąsiadujące podziały zbioru obiektów, czyli podziały na u oraz $u+1$ klas, lub odwrotnie) i pozwalają na ocenę, czy dana klasa powinna być podzielona na dwie klasy (lub para klas powinna być połączona w jedną). Proces podziału (łączenia) jest kontynuowany do momentu osiągnięcia określonego progu lub odrzucenia określonej hipotezy zerowej.

Ze względu na to, że w zaproponowanej optymalizacji wyboru procedury klasyfikacyjnej dla danego typu danych rozpatrywane są wszystkie warianty procedury klasyfikacyjnej danego zbioru obiektów (uzależnione od typu formuły normalizacyjnej, typu miary odległości i metody klasyfikacji), nie jest możliwe wykorzystanie kryteriów lokalnych do wyznaczenia optymalnej liczby klas,.

G.W. Milligan i M.C. Cooper[9] przetestowali na podstawie zbiorów danych o znanej strukturze klas 30 indeksów pozwalających wyznaczyć optymalną liczbę klas. Do oceny wyników symulacji w zaproponowanej optymalizacji wyboru procedury klasyfikacyjnej dla danego typu danych wprowadzono trzy najlepsze mierniki globalne z eksperymentu G.W. Milligana i M.C. Cooper: R.B. Calińskiego i J. Harabasza, F.B. Bakera i L.J Huberta, L.J. Huberta i J.R. Levine, oraz dwa indeksy często wykorzystywane w literaturze w testach porównawczych[10]: Silhouette, W.J. Krzanowskiego i Y.T. Lai:

1. Indeks Calińskiego i Harabasza[11]:

$$G1(u) = \frac{tr(\mathbf{B})/(u-1)}{tr(\mathbf{W})/(n-u)}, \quad G1(u) \in R_{+}, \tag{1}$$

gdzie:

[9] Zob. [14].

[10] Zob. np. [3]; [15]; [17]; [18].

[11] Zob. [2].

$\boldsymbol{B}$ – macierz kowariancji międzyklasowej,
$\mathbf{W}$ – macierz kowariancji wewnątrzklasowej,
tr – ślad macierzy,
u – liczba klas ($u = 2, \ldots, n - 1$),
n – liczba obiektów.

Indeks Calińskiego i Harabasza jest nazywany pseudostatystyką F[12].

2. Indeks Gamma Bakera i Huberta[13]:

$$G2(u) = \frac{s(+) - s(-)}{s(+) + s(-)},\ G2(u) \in [-1,\ 1], \tag{2}$$

gdzie:
s (+) – liczba par odległości zgodnych,
s (–) – liczba par odległości niezgodnych,
u – liczba klas ($u = 2, \ldots, n - 1$).

Przy obliczaniu indeksu Gamma[14] porównuje się wszystkie odległości wewnątrzklasowe z wszystkimi odległościami międzyklasowymi. Liczba tych porównań wynosi więc $r \cdot c$, gdzie r (c) – liczba odległości wewnątrzklasowych (międzyklasowych). Jeśli odległość wewnątrzklasowa jest mniejsza (większa) niż odległość międzyklasowa, to parę taką uznajemy za zgodną (niezgodną). Odległości wewnątrzklasowe równe międzyklasowym nie są uwzględniane.

3. Indeks Huberta i Levine[15]:

$$G3(u) = \frac{D(u) - r \cdot D_{\min}}{r \cdot D_{\max} - r \cdot D_{\min}},\ D_{\min} \neq D_{\max},\ G3(u) \in (0,\ 1), \tag{3}$$

gdzie:

[12] Zob. [12], s. 291.

[13] Zob. [1]; [8].

[14] Por. [6], s. 62.

[15] Zob. [9].

$D(u)$ – suma wszystkich odległości wewnątrzklasowych,
r – liczba odległości wewnątrzklasowych,
D_{min} – najmniejsza odległość wewnątrzklasowa,
D_{max} – największa odległość wewnątrzklasowa,
u – liczba klas ($u = 2, \ldots, n-2$).

4. Indeks Silhouette[16]:

$$S(u) = \sum_{i=1}^{n} S(i)/n\,,\ S(u) \in [-1,\ 1]\,, \tag{4}$$

gdzie:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}}\,,$$

$i, k = 1, \ldots, n$ – numer obiektu,

$a(i) = \sum_{k \in \{P_r \setminus i\}} d_{ik} \Big/ (n_r - 1)$ – średnia odległość obiektu i od pozostałych obiektów należących do klasy P_r,

$b(i) = \min_{s \neq r}\{d_{iP_s}\}$,

$d_{iP_s} = \sum_{k \in P_s} d_{ik} \Big/ n_s$ – średnia odległość obiektu i od obiektów należących do klasy P_s,

$r, s = 1, \ldots, u$ – numer klasy,

$u = 2, \ldots, n-1$ – liczba klas.

5. Indeks Krzanowskiego i Lai[17]:

$$KL(u) = \left| \frac{DIFF_u}{DIFF_{u+1}} \right|,\ KL(u) \in R_+\,, \tag{5}$$

$$DIFF_u = (u-1)^{2/m} tr\mathbf{W}_{u-1} - u^{2/m} tr\mathbf{W}_u\,,$$

[16] Zob. [10]; [16].

[17] Zob. [11].

gdzie:

W – macierz kowariancji wewnątrzklasowej,
u – liczba klas ($u = 2, \ldots, n-3$),
n – liczba obiektów, m – liczba zmiennych.

Indeksy $G1(u)$ i $KL(u)$ są oparte na macierzy danych, natomiast indeksy $G2(u)$, $G3(u)$ i $S(u)$ na macierzy odległości. Maksymalna wartość $G1(u)$, $G3(u)$, $S(u)$ i $KL(u)$ oraz minimalna $G2(u)$ wskazuje najlepszy podział zbioru obiektów, a zarazem wyznacza liczbę klas.

Szczegółowe charakterystyki ścieżek zaprezentowano w tabeli 1.

3. Podsumowanie

W typowym studium klasyfikacyjnym etapy wyboru formuły normalizacji wartości zmiennych, miary odległości, metody klasyfikacji oraz ustalenia liczby klas mają zwykle arbitralny charakter. Zaletą tego podejścia jest obiektywizacja problemu ich wyboru. Uzyskuje się to w wyniku przeprowadzenia symulacyjnej optymalizacji wyboru procedury klasyfikacyjnej dla danego typu danych z wykorzystaniem programu clusterSim[18]. Miernikami oceny wszystkich procedur klasyfikacyjnych badanego zbioru obiektów są globalne indeksy oceny jakości klasyfikacji pozwalające na wyznaczenie optymalnej liczby klas.

Zaprezentowane podejście ma pewne ograniczenia:

a) w literaturze jest ponad 40 mierników oceny jakości klasyfikacji; w zaprezentowanym podejściu możliwe było uwzględnienie tylko indeksów globalnych;
b) spośród indeksów globalnych uwzględniono pięć najważniejszych, jednak ostateczny wybór jednego z nich nadal jest arbitralny.

[18] Zob. [23].

Ścieżki w symulacyjnej optymalizacji wyboru procedury klasyfikacyjnej dla danego typu danych

Nr	Etapy typowej procedury klasyfikacyjnej	Numer ścieżki w procedurze symulacyjnej								
		1	2	3	4	5	6	7	8	9
I	Wybór obiektów i zmiennych	macierz danych $[x_{ij}]$								
II	Skala pomiaru zmiennych	ilorazowa	ilorazowa	przedziałowa[1]	porządkowa	nominalna wielostanowa	binarna	ilorazowa	przedziałowa[1]	ilorazowa	przedziałowa[1]
	Wybór formuły normalizacji[2]	n6–n11	n1–n5	n1–n5	NA	NA		bez normalizacji		n6–n11/ n1–n5	n1–n5
	Skala pomiaru zmiennych po normalizacji	ilorazowa	przedziałowa	przedziałowa	porządkowa	nominalna wielostanowa	binarna	ilorazowa	przedziałowa[1]	ilorazowa/ przedziałowa	przedziałowa
III	Wybór miary odległości[3]	d1–d7	d1–d5	d1–d5	d8	d9	b1–b10	d1–d7	d1–d5	NA	
IV	Wybór metody klasyfikacji	1. Pojedynczego połączenia 3. Średniej klasowej 5. *k-medoids* (pam) 7. Centroidalna[4] 2. Kompletnego połączenia 4. Ważonej średniej klasowej 6. Warda[4] 8. Medianowa[4]								*k*-średnich	
V	Liczba możliwości	[(6 x 7 x 5)+ (6 x 1 x 3)] + [(5 x 5 x 5) +(5 x 1 x 3)] = 368	(5 x 5 x 5) + (5 x 1 x 3) = 140	1 x 5 = 5	1 x 5 = 5	10 x 5 = 50	(7 x 5) + (1 x 3) = 38	(5 x 5) + (1 x 3) = 28	11	5
	Miernik jakości klasyfikacji	1. Caliński & Harabasz (G1) 2. Baker & Hubert (G2) 3. Hubert & Levine (G3) 4. Silhouette (S) 5. Krzanowski & Lai (KL)		1. NA 2. G2 3. G3 4. S 5. NA			1. G1 2. G2 3. G3 4. S 5. KL		1. G1 2. NA 3. NA 4. NA 5. KL	

[1] Lub ilorazowa i przedziałowa.

[2] n1 (n2) – standaryzacja klasyczna (Webera), n3 – unitaryzacja, n4 – unitaryzacja zerowana, n5 – normalizacja w przedziale [–1; 1], n6–n11 – przekształcenia ilorazowe.

[3] d1 – miejska, d2 – euklidesowa, d3 – Czebyszewa, d4 – kwadrat euklidesowej, d5 – GDM1, d6 – Canberra, d7 – Braya-Curtisa, d8 – GDM2, d9 – Sokala i Michenera dla zmiennych nominalnych; odległości dla zmiennych binarnych (dostępne w procedurze dist.binary): b1 = Jaccard; b2 = Sokal & Michener; b3 = Sokal & Sneath (1); b4 = Rogers & Tanimoto; b5 = Czekanowski; b6 = Gower & Legendre (1); b7 = Ochiai; b8 = Sokal & Sneath (2); b9 = Phi of Pearson; b10 = Gower & Legendre (2).

[4] Metody klasyfikacji przyjmujące założenie, że odległości między obiektami zostały wyznaczone za pomocą kwadratu odległości euklidesowej, tylko bowiem w tym przypadku metody te mają interpretację geometryczną, zgodną z ich nazwami.

NA – nie stosuje się.

Źródło: opracowanie własne (opisy metod znajdują się m.in. w następujących pracach: [4]; [5]; [6]).

Literatura

1. Baker F.B., Hubert L.J.: *Measuring the power of hierarchical cluster analysis*. „Journal of the American Statistical Association" 1975, vol. 70, No 349.
2. Caliński R.B., Harabasz J.: *A dendrite method for cluster analysis*. „Communications in Statistics" 1974, vol. 3.
3. Dudoit S., Fridlyand J.: *A prediction-based resampling method for estimating the number of clusters in a dataset*. „Genome Biology" 2002, vol. 3, No 7.
4. Everitt B.S., Landau S., Leese M.: *Cluster analysis*. Edward Arnold, London 2001.
5. Gatnar E., Walesiak M.: *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wyd. Naukowe Akademii Ekonomicznej we Wrocławiu, Wrocław 2004.
6. Gordon A.D.: *Classification*. Chapman and Hall/CRC, London 1999.
7. Górniak J.: *My i nasze pieniądze*. Wyd. Aureus, Kraków 2000.
8. Hubert L.J. *Approximate evaluation technique for the single-link and complete--link hierarchical clustering procedures*. „Journal of the American Statistical Association" 1974, vol. 69, No 347.
9. Hubert L.J., Levine J.R.: *Evaluating object set partitions: free sort analysis and some generalizations*, „Journal of Verbal Learning and Verbal Behaviour" 1976, vol. 15.
10. Kaufman L., Rousseeuw P.J.: *Finding groups in data: an introduction to cluster analysis*. Wiley, New York 1990.
11. Krzanowski W.J., Lai Y.T.: *A criterion for determining the number of groups in a data set using sum of squares clustering*. „Biometrics" 1985, No 44.
12. Lattin J.M., Carroll J.D., Green P.E.: *Analyzing multivariate data*. Brooks/Cole, Pacific Grove 2003.
13. Milligan G.W.: *Clustering validation: results and implications for applied analyses*. W: *Clustering and classification*. Red. P. Arabie, L.J. Hubert, G. de Soete. World Scientific, Singapore 1996.

14. Milligan G.W., Cooper M.C.: *An examination of procedures for determining the number of clusters in a data set*. „Psychometrika" 1985, No 2.
15. Mufti G.B., Bertrand P., El Moubarki L.: *Determining the number of groups from measures of cluster stability*. W: *Applied Stochastic Models and Data Analysis*. Red. J. Janssen, P. Lenca. ENST Bretagne, Brest 2005.
16. Rousseeuw P.J.: *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. „Journal of Computational and Applied Mathematics" 1987, No 20.
17. Sugar C.A., James G.H.: *Finding the number of clusters in a dataset: an information-theoretic approach*. „Journal of the American Statistical Association" 2003, vol. 98, No 463.
18. Tibshirani R., Walther G., Hastie T.: *Estimating the number of clusters in a data set via the gap statistic*. „Journal of the Royal Statistical Society" 2001, ser. B, vol. 63, part 2.
19. Walesiak M.: *Statystyczna analiza wielowymiarowa w badaniach marketingowych*. Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 654, seria Monografie i Opracowania nr 101. Wrocław 1993.
20. Walesiak M.: *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*. Wyd. Akademii Ekonomicznej, Wrocław 2002.
21. Walesiak M.: *Rekomendacje w zakresie strategii postępowania w procesie klasyfikacji zbioru obiektów*. W: *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*. Red. A. Zeliaś. Wyd. Akademii Ekonomicznej w Krakowie, Kraków 2005.
22. Walesiak M.: *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*. Wydanie II rozszerzone. Wyd. Akademii Ekonomicznej we Wrocławiu, Wrocław 2006.
23. Walesiak M., Dudek A.: *Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – oprogramowanie komputerowe i wyniki badań*. W: *Klasyfikacja i analiza danych – teoria i zastosowania*. Red. K. Jajuga, M. Walesiak. Prace Naukowe Akademii Ekonomicznej we Wrocławiu (w redakcji).

DETERMINATION OF OPTIMAL CLUSTERING PROCEDURE FOR A DATA SET – THE CHARACTERISATION OF THE PROBLEM

Summary

In typical cluster analysis study eight major steps are distinguished (see Milligan [1996], 342–343; Walesiak [2005]). Four of them represent the critical steps: decisions concerning variable normalisation formula, selection of a distance measure, selection of clustering method, determining the number of clusters.

The article presents determination of optimal clustering procedure for a data set by varying all combinations of normalization formulas, distance measures, and clustering methods. Nine paths of simulation was separated depends on variable scale of measurement in a data set. Based on this approach in article of Walesiak and Dudek [2005] the clusterSim computer program written in R and C++ languages was proposed.

Translated by Marek Walesiak, Andrzej Dudek