PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 508 • 2018

Taksonomia 31
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Jerzy Korzeniewski
Uniwersytet Łódzki
e-mail: jurkor@wp.pl

SELEKCJA ZMIENNYCH W ANALIZIE SKUPIEŃ MARKETINGOWYCH ZBIORÓW DANYCH BINARNYCH

SELECTION OF VARIABLES IN MARKETING BINARY DATA CLUSTER ANALYSIS

DOI: 10.15611/pn.2018.508.09

Streszczenie W roku 2001 Desai zaproponował ciekawą miarę podobieństwa dwóch różnych wartości/wariantów tej samej cechy. Miarę tę można w dość prosty sposób wykorzystać do wyznaczenia siły dyskryminacyjnej cechy binarnej lub nominalnej wielostanowej w problemie analizy skupień. Idea oparta jest na tym, że im mniejsze podobieństwo, na przykład 1 do 0 (jako wartości zmiennej binarnej), tym większa zdolność dyskryminacyjna cechy. Ten pomysł zastosowano do skonstruowania nowej metody selekcji zmiennych binarnych w zagadnieniu analizy skupień i w zastosowaniu do dość obszernej klasy zbiorów danych binarnych, jaką są dane marketingowe. Podstawową zaletą nowej metody jest jej niezależność od konieczności grupowania danych, co wiąże się zawsze z przyjęciem jakiejś konkretnej metody grupowania oraz konkretnej wartości liczby skupień. Eksperyment przeprowadzony na 162 zbiorach danych pokazuje wysoką efektywność metody.

Słowa kluczowe: analiza skupień, dane binarne, selekcja zmiennych, dane marketingowe.

Summary: In 2011 Desai proposed an interesting measure of similarity of two different values/variants of the same variable. This measure can be easily used to assess the discrimination power of binary or multi-level nominal variable in cluster analysis. The idea is based on the fact that the smaller the similarity between e.g. 1 and 0 (treated as the binary variable values) the bigger the discrimination power of the variable. This idea was used to construct a new variable selection method for binary variables in the context of cluster analysis and for quite a broad range of binary data sets such as marketing data sets. The main advantage of the new proposal is its independence of the necessity of data grouping which is always connected with applying some grouping method and, in turn, some established number of clusters. The experiment carried out on 162 data sets shows high efficiency of the new proposal.

Keywords: cluster analysis, binary data, variable selection, marketing data.

1. Wstęp

Celem niniejszego opracowania jest zaproponowanie nowej metody selekcji zmiennych w kontekście analizy skupień dla marketingowych zbiorów danych binarnych. Ten problem jest stosunkowo słabo opracowany w literaturze przedmiotu w porównaniu z mnogością metod dla tego samego zagadnienia dla zmiennych mierzonych na skalach silniejszych. Przyczyną jest zapewne to, że typowe miary odległości dla skali binarnej nie mają takiej zdolności różnicującej obiekty jak miary odległości na skalach silniejszych. Brusco [2004] rekomenduje najpierw użycie jakiejś dobrej metody szacowania liczby skupień w zbiorze danych binarnych (np. indeksu Ratkowskiego-Lance), a następnie, gdy liczba skupień jest znana, optymalizację doboru zmiennych przy wykorzystaniu metody grupowania *k*-średnich. Funkcją dyskryminacyjną decydującą o wyniku optymalizacji jest funkcja

$$Z = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{(i<j)\in C_k} d_{ij}, \tag{1}$$

którą minimalizujemy po wszystkich podziałach $\pi = \{C_1, C_2, \ldots, C_k\}$ zbioru wszystkich rozważanych obiektów (na ogół próby z całego zbioru danych). Miarą odległości (wyrażenia d_{ij} we wzorze (1)) jest metryka Sokala-Michenera. Korzeniewski [2012] proponuje podejście filtrujące zmienne bez odwoływania się do grupowania danych, oparte na badaniu skorelowania odległościowego zmiennych, które powinno być tym silniejsze, im większy wkład zmiennych, których skorelowanie oceniamy, do tworzenia struktury skupień. Do omawianego zagadnienia można, oczywiście, stosować wiele innych metod selekcji zmiennych opracowanych pod kątem analizy skupień, jednak inne metody, przewidziane dla cech ciągłych, spisują się słabo dla cech binarnych bądź też w ogóle nie da się ich zastosować.

2. Zbiory danych

Specyfika binarnych danych marketingowych polega na tym, że istnieją grupy zmiennych wyznaczone przez specyfikę badania, które mogą być skorelowane. Cały zbiór danych składa się z kilku grup takich zmiennych. W tym artykule zostaną zbadane zbiory wygenerowane według sugestii Dimitriadou [Dimitriadou i in. 2002], zgodnie z którą każdy zbiór opisany jest przez 12 zmiennych binarnych podzielonych na 4 grupy o być może różnych liczebnościach. Przykładowy schemat takich zbiorów jest przedstawiony w tab. 1. Ogólną zasadą jest pokazanie przykładowych zależności pomiędzy grupami respondentów oraz grupami pytań w kwestionariuszu.

Symbol H oznacza wysokie prawdopodobieństwo jedynki na danej zmiennej, symbol L zaś oznacza niskie prawdopodobieństwo jedynki. Oczywiście, liczba zmiennych w poszczególnych grupach zmiennych, skorelowanie zmiennych wewnątrz grup oraz konkretna wartość H oraz L będą przyjmowały różne wartości (patrz

Tabela 1. Przykładowy schemat zbiorów danych binarnych, 12 zmiennych podzielonych na cztery grupy

	Grupa 1			Grupa 2			Grupa 3			Grupa 4		
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
Skupienie1	H	H	H	H	H	H	L	L	L	L	L	L
Skupienie2	L	L	L	L	L	L	H	H	H	H	H	H
Skupienie3	L	L	L	H	H	H	H	H	H	L	L	L
Skupienie4	H	H	H	L	L	L	L	L	L	H	H	H
Skupienie5	L	L	L	H	H	H	L	L	L	H	H	H
Skupienie6	H	H	H	L	L	L	H	H	H	L	L	L

Źródło: [Dimitriadou i in. 2002].

opis eksperymentu). Należy pamiętać o tym, że schemat przedstawiony w tab. 1 jest przykładowy, tzn. liczba zmiennych w poszczególnych grupach zmiennych nie musi być jednakowa oraz liczba skupień nie musi być równa 6 (patrz opis eksperymentu). Komentarza wymaga również możliwość uwzględnienia skorelowania pomiędzy zmiennymi wewnątrz grup. Wobec takiej konieczności do generowania zbiorów zastosowano metodę opracowaną przez Leischa i współautorów (por. [Leisch i in. 1998]), która wymaga zadania macierzy korelacyjnej oraz rozkładów brzegowych. Metoda ta została oprogramowana przez tych samych autorów w pakiecie *bindata* (por. [Leisch i in. 2015]) dostępnym w programie R. Ten pakiet zastosowano.

3. Nowa metoda selekcji zmiennych

Nowa propozycja metody selekcji zmiennych jest oparta na podobieństwie $sim(v_{ij}, v_{ik})$ dwóch różnych wartości v_{ij}, v_{ik} tej samej zmiennej o numerze i (j, k – numery dwóch różnych wartości zmiennej i). To podobieństwo definiujemy wzorem (por. [Desai i in. 2011]):

$$sim(v_{ij}, v_{ik}) = \frac{1}{d-1}\sum_{m=1,m\neq i}^{d} similarity_m, \tag{2}$$

w którym po prawej stronie występuje średnia podobieństw względem ustalonej innej zmiennej (d- liczba wszystkich zmiennych). Podobieństwa tych samych wartości przy ustalonej innej zmiennej o numerze m określa wzór;

$$similarity_m = \begin{cases} 1 - \dfrac{\left|CI[A_i: v_{ij}][A_m] - CI[A_i: v_{ik}][A_m]\right|}{Max[A_m] - Min[A_m]} & \text{gdy } A_m \text{ jest liczbowa} \\ CosProd\left(CI[A_i: v_{ij}][A_m], CI[A_i: v_{ik}][A_m]\right) & \text{gdy } A_m \text{ jest nieliczbowa} \end{cases}. \tag{3}$$

W przypadku nas interesującym, tj. wariancie drugim dla zmiennych nominalnych, przypominającym w swej postaci iloczyn skalarny dwóch wektorów, prawa strona $CosProd\big(CI[A_i: v_{ij}][A_m], CI[A_i: v_{ik}][A_m]\big)$ wzoru (3) przyjmuje postać

$$\frac{\sum_{v_{ml}, v_{mt} \in A_m} CI[A_i: v_{ij}][A_m: v_{ml}] * CI[A_i: v_{ik}][A_m: v_{mt}] * sim(v_{ml}, v_{mt})}{NormalVector1 * NormalVector2}. \quad (4)$$

We wzorze (4) wartość wyrażenia $CI[A_i: v_{ij}][A_m: v_{ml}]$ jest po prostu zliczeniem obiektów charakteryzujących się wariantem v_{ij} na zmiennej A_i oraz wariantem v_{ml} na zmiennej A_m. Wartość tego zliczenia oznaczamy symbolem rozpoczynającym się od *CI* (*categorical information*).

Czynniki normalizujące z mianownika dane są wzorami

$$NormalVector1 = \\ = \left(\sum_{v_{ml}, v_{mt} \in A_m} CI[A_i: v_{ij}][A_m: v_{ml}] * CI[A_i: v_{ij}][A_m: v_{mt}] * sim(v_{ml}, v_{mt}) \right)^{1/2}. \quad (5)$$

$$NormalVector2 = \\ = \left(\sum_{v_{ml}, v_{mt} \in A_m} CI[A_i: v_{ik}][A_m: v_{ml}] * CI[A_i: v_{ik}][A_m: v_{mt}] * sim(v_{ml}, v_{mt}) \right)^{1/2}. \quad (6)$$

Powyższe wyrażenia dla zmiennych binarnych (czyli nominalnych dwustanowych) znajduje się bardzo łatwo, gdyż podstawą jest zapamiętanie wartości wyrażeń $CI[A_i: v_{ij}][A_m: v_{ml}]$

Po znalezieniu wartości podobieństw (2) dla każdej zmiennej oddzielnie wystarczy uporządkować je i podzielić na dwie grupy, po czym jedną z grup uznać za zmienne tworzące strukturę skupień, drugą zaś za zmienne zakłócające, które należy odrzucić. Należy zatem założyć, że w zbiorze istnieją jakieś zmienne zakłócające, które powinno się odrzucić. Takie założenie nie ogranicza jednak ogólności metody, bo przed jej zastosowaniem można dołączyć sztucznie jakieś zmienne, zupełnie nie związane z tymi już istniejącymi. Dokładniej propozycję nowej metody zapiszmy w postaci następującego algorytmu.

Krok 1. Dla każdej zmiennej *i* znajdujemy wartość daną wzorem (2) dla wartości v_{ij}, v_{ik} równych 1 i 0.

Krok 2. Porządkujemy zmienne rosnąco względem wskaźnika (2).

Krok 3. Dzielimy ciąg uporządkowanych zmiennych na dwie grupy metodą *k*-średnich (dla $k = 2$). Zmienną dyskryminującą jest wskaźnik (2), punktami startowymi zaś dwie wartości tego wskaźnika dla zmiennych skrajnych z uporządkowanego ciągu zmiennych. Zmienne z pierwszej grupy (z niższymi wartościami wskaźnika) uznajemy za zmienne tworzące strukturę skupień, zmienne zaś z drugiej grupy uznajemy za zmienne zakłócające.

4. Eksperyment badawczy

W celu ocenienia nowej propozycji przeprowadzono następujący eksperyment symulacyjny. Wygenerowano 162 zbiory danych binarnych zgodnie ze schematem przedstawionym w punkcie 2. W tym celu zastosowano pakiet *bindata* dostępny w programie R. W przedstawionym schemacie zmieniały się następujące parametry:

Prawdopodobieństwa: dla H są 3 warianty 0,9, 0,8, 0,7 i dla każdego wariantu prawdopodobieństwo L ma wartość, odpowiednio, 0,1, 0,2, 0,3.

Skorelowanie wewnątrz grup: zmienne nieskorelowane, zmienne średnio silnie skorelowane (0,4), zmienne silnie skorelowane (0,8).

Liczba skupień: 4, 5, 6.

Liczebność skupień: trzy warianty, (1000, 1000, 1000, 1000, 1000, 1000), (2000, 500, 1000, 700, 700, 1100), (3000, 300, 1000, 500, 700, 500)

Liczba zmiennych w grupach: dwa warianty, (3, 3, 3, 3), (5, 4, 2, 1)

Po wymnożeniu wszystkich wariantów otrzymujemy łącznie $3 \cdot 3 \cdot 3 \cdot 3 \cdot 2 = 162$ zbiory.

W celu zbadania efektywności metody selekcji zmiennych z określonego zbioru zastosowano metodę dołączania zmiennych zakłócających do zmiennych istniejących. Liczba dołączanych zmiennych zmieniała się w zakresie od 2 do 20. Zmienne zakłócające były zmiennymi binarnymi nieskorelowanymi między sobą ani ze zmiennymi istniejącymi w zbiorze, z prawdopodobieństwem przyjęcia wartości równej 1 ustalanym dla każdej zmiennej oddzielnie przez losowanie tego prawdopodobieństwa z odcinka (0,1; 0,9).

Po uwzględnieniu zmiennych zakłócających otrzymano zatem 19 · 162 = 3078 różnych zbiorów danych.

Do każdego zbioru zastosowano metodę opisaną w punkcie 3. Jakość selekcji zmiennych była oceniana za pomocą następujących wskaźników:

- *pamięć selekcji*, tj. stosunek liczby wybranych zmiennych oryginalnych (tworzących strukturę skupień) do liczby wszystkich zmiennych oryginalnych.
- *precyzja selekcji*, tj. stosunek liczby wybranych zmiennych oryginalnych do liczby wszystkich wybranych zmiennych.

Ponadto zbadano poprawność porządkowania zmiennych.

5. Wyniki i wnioski

Na wstępie zauważmy, że nowa propozycja nie wymaga grupowania danych i ta cecha jest bardzo istotną zaletą. Olbrzymia większość metod selekcji zmiennych w analizie skupień to metody uzależnione od konkretnej metody grupowania i, co za tym idzie, na ogół również od koniecznej do określenia liczby skupień, na które trzeba pogrupować obiekty.

We wszystkich (!) 3078 przypadkach zbiorów porządkowanie zmiennych było poprawne, tzn. pierwsze 12 zmiennych z najwyższymi wskaźnikami zdolności dyskryminacyjnej były zmiennymi tworzącymi strukturę skupień.

Precyzja selekcji we wszystkich (!) 3078 przypadkach była równa 1, tzn. nigdy nie wybrano zmiennej zakłócającej jako zmiennej tworzącej strukturę skupień.

Średnia pamięć nowej metody to 90,2%, a dokładniejsze badanie tego wyniku w zależności od innych parametrów zbiorów danych przedstawione jest w tab. 2-6.

Tabela 2. Pamięć nowej metody w zależności od liczby zmiennych zakłócających

Liczba zmiennych zakłócających	2	3	4	5	6	7	8	9	10	
Średnia pamięć	,892	,891	,892	,894	,896	,899	,901	,903	,903	
Liczba zmiennych zakłócających	11	12	13	14	15	16	17	18	19	20
Średnia pamięć	,905	,905	,905	,906	,906	,907	,907	,908	,908	,909

Źródło: obliczenia własne.

Pamięć nowej metody w zależności od liczby zmiennych zakłócających jest przedstawiona w tab. 2. Zastanawiające jest to, że pamięć poprawia się minimalnie wraz ze wzrostem liczby zmiennych, czyli w trudniejszych warunkach metoda działa lepiej. Jedynym wyjaśnieniem tego paradoksu może być to, że przyjęta metoda dzielenia zbioru wszystkich zmiennych na dwie części może (w przypadkach dużego zróżnicowania wskaźnika zdolności dyskryminacyjnej) „fałszywie" przyłączać niektóre zmienne oryginalne do grupy zmiennych zakłócających z racji swojej specyfiki grupowania metodą *k*-średnich.

Tabela 3. Pamięć nowej metody w zależności od liczby skupień w zbiorze danych

Liczba skupień	4	5	6
Średnia pamięć	,919	,896	,891

Źródło: obliczenia własne.

Pamięć nowej metody w zależności od liczby skupień w zbiorze danych jest przedstawiona w tab. 3. Wyniki są zgodne z intuicją – jakość selekcji pogarsza się wraz ze wzrostem liczby skupień.

Pamięć nowej metody w zależności od siły skorelowania zmiennych jest przedstawiona w tab. 4. Trudno wyjaśnić, dlaczego najlepsze wyniki metoda uzyskała dla średniej siły korelacji, ale zgodne z intuicją jest to, że zbiory ze zmiennymi binarnymi nieskorelowanymi wypadły słabiej.

Tabela 4. Pamięć nowej metody w zależności od skorelowania zmiennych w zbiorze danych

Skorelowanie zmiennych	Brak korelacji	Średnia korelacja	Silna korelacja
Średnia pamięć	,882	,917	,906

Źródło: obliczenia własne.

Pamięć nowej metody w zależności od tego, z jakim prawdopodobieństwem występuje w zmiennych binarnych 1, a z jakim zero, jest przedstawiona w tab. 5. Wyniki są dość silnie uzależnione od rozkładu i zgodne z intuicją,

tzn. większa rozbieżność prawdopodobieństw występowania obu wariantów sprzyja lepszej selekcji zmiennych.

Tabela 5. Pamięć nowej metody w zależności od rozkładu prawdopodobieństwa zmiennych

Rozkład prawdopodobieństwa „1" oraz „0"	(0,9; 0,1)	(0,8; 0,2)	(0,7; 0,3)
Średnia pamięć	,943	,901	,861

Źródło: obliczenia własne.

Pamięć nowej metody w zależności od liczebności grup zmiennych jest przedstawiona w tab. 6. Wynik jest zgodny z intuicją, tzn. równomierne liczebności grup zdecydowanie sprzyjają dobrej selekcji (innymi słowy, obecność jednej grupy jednoelementowej zdecydowanie obniża jakość).

Tabela 6. Pamięć nowej metody w zależności od liczebności grup zmiennych

Liczby zmiennych w grupach zmiennych	(3, 3, 3, 3)	(5, 4, 2, 1)
Średnia pamięć	,989	,814

Źródło: obliczenia własne.

Podsumowując, można stwierdzić, że propozycja nowej metody okazała się bardzo udana na zbadanej klasie zbiorów. Przyczyną jest zapewne to, że miara Desai uwzględnia w swej konstrukcji cały zbiór danych – jest to miarą typu *data driven* albo *data intensive*.

Literatura

Brusco M., 2004, *A Variable-Selection Heuristic for K-Means Clustering*, Psychological Methods, vol. 9, s. 510-523.

Desai A., Singh H., Pudi V., 2011, *DISC: Data-Intensive Similarity Measure for Categorical Data*, [w:] Huang J.Z., Cao L., Srivastava J. (red.), *Advances in Knowledge Discovery and Data Mining*, PAKDD 2011. Lecture Notes in Computer Science, vol 6635. Springer, Berlin-Heidelberg.

Dimitriadou E., Dolničar S., Weingessel A., 2002, *An examination of indexes for determining the number of clusters in binary data sets*, Psychometrika, vol. 67, issue 1, s. 137-159.

Korzeniewski J., 2012, *Metody selekcji zmiennych w analizie skupień. Nowe procedury*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Leisch F., Weingessel A., Hornik K., 1998, *On the generation of correlated artificial binary data,* Working Paper Series, SFB "Adaptive Information Systems and Modelling in Economics and Management Science", Vienna University of Economics, http://www.wu-wien.ac.at/am.

Leisch F., Weingessel A., Hornik K., 2015, *Bindata package manual*, https://cran.r-project.org/web/packages/bindata/.