Nr 51

# PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

# Projektowanie, ocena i wykorzystanie danych rynkowych

Redaktor naukowy
Józef Dziechciarz

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2009

## Spis treści

Wstęp ..... 7
**Sylwester Białowąs,** Kolejność pytań w kwestionariuszu wywiadu osobistego a zniekształcenia pomiaru wywołane heurystyką zakotwiczenia ..... 9
**Marta Dziechciarz,** Podejścia do oceny atrakcyjności segmentów rynku jako etapu kończącego proces segmentacji rynku ..... 14
**Bartłomiej Jefmański,** Rozmyta metoda *k*-średnich w identyfikacji przynależności obiektów do segmentów rynkowych – na przykładzie rynku samochodowego ..... 28
**Iwona Kasprzyk,** Wykorzystanie konfiguracyjnej analizy częstości w analizie klas ukrytych ..... 37
**Jolanta Kowal,** Wybrane teoretyczne i praktyczne aspekty metodologii badań jakościowych ..... 46
**Magdalena Kowalska-Musiał,** Relacje partnerskie w układach diadycznych – ocena i analiza danych ..... 76
**Mariusz Łapczyński,** Modele hybrydowe CART-LOGIT w analizie danych rynkowych ..... 85
**Roman Pawlukowicz,** Średnia arytmetyczna cen transakcyjnych nieruchomości a wartość rynkowa nieruchomości ..... 96
**Marcin Pełka,** Porównanie strategii klasyfikacji danych symbolicznych .... 106
**Adam Sagan,** Metaanaliza danych w marketingu zorientowanym na dowody – orientacja kliniczna w badaniach rynkowych i marketingowych ..... 114
**Piotr Tarka,** Zastosowanie analizy regresji i sztucznych sieci neuronowych w badaniach satysfakcji klientów ..... 125
**Barbara Worek,** Rzetelność i trafność w badaniach jakościowych: ocena jakości danych ..... 136

## Summaries

**Sylwester Białowąs,** The anchoring heuristic and the bias of the measurement in marketing research ..... 13
**Marta Dziechciarz,** Determining the attractiveness of market segments as the ending step of segmentation process ..... 27
**Bartłomiej Jefmański,** Fuzzy c-means in market segments membership identification – a car market example ..... 36
**Iwona Kasprzyk,** Application of configural frequency analysis in latent class analysis ..... 45

**Jolanta Kowal,** Some chosen theoretical and practical aspects of qualitative research . . . . . 75
**Magdalena Kowalska-Musiał,** Dyadic relationship – data evaluation and analysis . . . . . 84
**Mariusz Łapczyński,** The hybrid CART-LOGIT models in analysing market data . . . . . 95
**Roman Pawlukowicz,** Arithmetic mean of transactional prices of properties and property's market value . . . . . 105
**Marcin Pełka,** Comparison of symbolic data clustering strategies . . . . . 113
**Adam Sagan,** Meta-analysis in evidence-based marketing: clinical orientation in marketing research . . . . . 124
**Piotr Tarka,** Artificial neural networks and regression comparison analysis within customer satisfaction data . . . . . 135
**Barbara Worek,** Reliability and validity in qualitative research: data quality evaluation . . . . . 147

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 51 — 2009
Projektowanie, ocena i wykorzystanie danych rynkowych

**Marcin Pełka**
Uniwersytet Ekonomiczny we Wrocławiu

# PORÓWNANIE STRATEGII KLASYFIKACJI DANYCH SYMBOLICZNYCH

## 1. Wstęp

Metody klasyfikacji, należące do grupy metod statystycznej analizy wielowymiarowej, mają szerokie zastosowanie w badaniach marketingowych, a szczególnie w odniesieniu do odnośnie segmentacji rynku, pozycjonowania produktu (przedsiębiorstwa) na rynku, identyfikacji rynków czy określania struktury rynku [Walesiak 1996, s. 117, 120, 151].

W odróżnieniu od danych klasycznych w rozumieniu statystycznej analizy wielowymiarowej (SAW), gdzie obiekty opisywane są dla każdej zmiennej przez jedną liczbę rzeczywistą lub jedną kategorię, w analizie danych symbolicznych obiekty dla każdej zmiennej mogą być opisywane nie tylko przez pojedyncze liczby czy kategorie, ale również przez listy kategorii, listy kategorii z wagami, przedziały liczbowe. Dodatkowo możliwe jest zdefiniowanie zależności między typami zmiennych symbolicznych. Zależności te mogą mieć postać taksonomiczną, hierarchiczną lub logiczną.

Analiza danych symbolicznych pozwala z jednej strony na pełniejszy opis analizowanych obiektów, z drugiej zaś powoduje pewne utrudnienia w analizie zjawisk. W analizie danych symbolicznych można wykorzystać dwie strategie – strategię klasyczną, w której dokonuje się transformacji zmiennych symbolicznych w zmienne tradycyjne, a następnie do otrzymanego w ten sposób zbioru obiektów stosuje się znane z SAW metody analizy danych, oraz strategię analizy danych symbolicznych, gdzie korzysta się z pełnego zbioru informacji o obiektach.

## 2. Typy zmiennych w analizie danych symbolicznych

W przypadku obiektów symbolicznych możemy mieć do czynienia z następującymi rodzajami zmiennych [*Analysis of symbolic*… 2000, s. 2-3]:

1) ilorazowe, przedziałowe, porządkowe, nominalne,
2) kategorie, czyli dane tekstowe, np. biały, zielony,

3) przedziały liczbowe, np. kwota, którą respondent byłby skłonny wydać na dodatkowe wyposażenie samochodu (1000 zł, …, 2500 zł), co oznacza, że może ona wynieść od 1000 zł do 2500 zł,

4) lista kategorii, jak np. standard wyposażenia pewnego samochodu osobowego (obiektu) określony jako: wysoki, luksusowy, co oznacza, że oferowany jest w dwóch rodzajach standardów wyposażenia,

5) lista kategorii z wagami (prawdopodobieństwami), gdzie oprócz listy kategorii występują wagi, z jakimi obiekt posiada wybraną kategorię, np. w wyniku klasyfikacji okazało się, że segment pierwszy to konsumenci preferujący marki: Toyota (0,45); Audi (0,30); Opel (0,25) oznacza to, że w pierwszym rzędzie wybierają Toyotę, następnie samochody marki Audi, a na końcu Ople,

6) zmienne strukturalne [*Analysis of symbolic*… 2000), s. 33-37], w literaturze wyróżnia się oprócz wyżej wymienionych typów zmiennych także zmienne strukturalne:

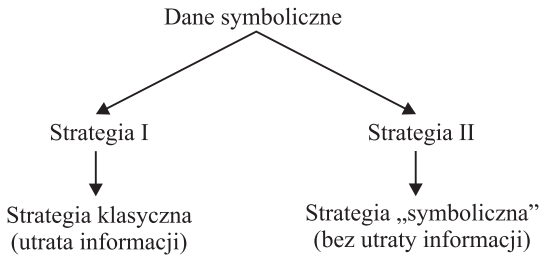
a) zmienne o zależności funkcyjnej lub logicznej pomiędzy zmiennymi, gdzie *a priori* ustalono reguły funkcyjne lub logiczne decydujące o tym, jaką wartość przyjmie dana zmienna,

b) zmienne hierarchiczne, w których *a priori* ustalono warunki, od których zależy, czy zmienna dotyczy danego obiektu, czy też nie,

c) zmienne taksonomiczne, w których *a priori* ustalono systematykę, według której przyjmuje ona swoje realizacje.

## 3. Strategie klasyfikacji danych symbolicznych

W przypadku analizy danych symbolicznych można wyróżnić dwie główne strategie analizowania tego typu danych (zob. rys. 1).




Rys. 1. Strategie analizy danych symbolicznych

Źródło: opracowanie własne.

W przypadku strategii I, która jest nazywana również strategią klasyczną, analizowanie danych symbolicznych polega na transformowaniu zmiennych symbolicznych w klasyczne, a następnie wykorzystaniu znanych z SAW metod analizy. Podej-

ście polegające na transformacji zmiennych symbolicznych w klasyczne, następnie na wykorzystaniu znanych metod analizy, a na końcu wykorzystujące sposób kodowania zmiennych do interpretacji wyników z zastosowaniem zmiennych symbolicznych zostało zaproponowane przez Didaya w 1987 r. (por. [Diday 1987]) i jest nazywane podejściem „*symboliczne – numeryczne – symboliczne*".

Strategia ta pozwala co prawda analizować dane symboliczne z wykorzystaniem dobrze znanych z SAW metod analizy, jednakże niesie za sobą ryzyko utraty części informacji o obiektach, co w konsekwencji może prowadzić do zaburzenia wyników.

Strategia II, nazywana strategią analizy danych symbolicznych, polega albo na wykorzystaniu metod opracowanych dla zmiennych i obiektów symbolicznych bazujących na tablicy danych symbolicznych, albo obliczeniu odległości miarami adekwatnymi dla danych symbolicznych (zob. [Malerba i in. 2001; Malerba, Esposito, Monopoli 2002]), a następnie wykorzystaniu metod bazujących na macierzy odległości. W tym przypadku nie ma znaczenia, czy korzystamy ze znanych z SAW metod analizy, czy ich adaptacji opracowanych dla danych symbolicznych.


Rys. 2. Strategie klasyfikacji danych symbolicznych

Źródło: opracowanie własne.

Ze strategii analizowania zbioru danych symbolicznych wynikają bezpośrednio strategie klasyfikacji danych symbolicznych (por. rys. 2).

## 4. Transformacja zmiennych symbolicznych

Istotnym problemem w pierwszej strategii klasyfikacji danych symbolicznych, która jest odbiciem strategii klasycznej analizy danych symbolicznych, jest zagadnienie transformacji zmiennych symbolicznych w klasyczne.

W przypadku zmiennych strukturalnych nie ma możliwości pozostawienia informacji o zależnościach, jakie istnieją między zmiennymi. Zmienne te należy traktować jak zwykłe zmienne symboliczne bez struktury. Takie podejście do tych zmiennych oznacza znaczny ubytek informacji o obiektach.

Dla pozostałych zmiennych sposoby transformacji zestawiono w tab. 1.

Tabela 1. Operacje transformacji dla zmiennych symbolicznych

| Zmienne symboliczne / klasyczne | Realizacje zmiennej symbolicznej / klasycznej | Typ zmiennej symbolicznej / klasycznej | Transformacja |
|---|---|---|---|
| Preferowana cena | <25000; 36000>, <28000; 37000>, <30000; 50000>, <33000; 58000>, <65000; 80000>, <66000; 90000> | przedziały liczbowe nierozłączne | usunięcie zmiennej z analizy lub rozmyte kodowanie przedziałów |
| Wybrane pojemności silnika | <1000; 1200>, (1200; 1400>, (1400; 1600>, (1600; 1800>, (1800; 2000>, (2000; 2200> | przedziały liczbowe rozłączne | kodowanie przedziałów, np.: (1000; 1200> = 1; (1200; 1400> = 2 (1400; 1600> = 3; (1600; 1800> = 4 (1800; 2000> = 5; (2000; 2200> = 6 |
| Preferowany kolor | {zielony, niebieski, biały, żółty, czarny, czerwony} | listy kategorii | wprowadzenie zmiennych binarnych w liczbie równej liczbie kategorii |
| Preferowana marka | {60% Honda, 40% Toyota} {40% Honda, 20% Škoda, 20% Toyota, 20% Audi} {80% Audi, 15% Opel, 5% Toyota} | listy kategorii z wagami | wprowadzenie zmiennych ilorazowych w liczbie równej liczbie kategorii, ich realizacjami będą prawdopodobieństwa (wagi) |
| Liczba drzwi | 2, 3, 4, 5 | zmienna klasyczna (ilorazowa) | bez zmian |

Źródło: opracowanie własne.

W przypadku przedziałów liczbowych nierozłącznych można albo całkowicie usunąć taką zmienną z analizy – co wiąże się z całkowitą utratą informacji, jaką niesie ze sobą taka zmienna, albo dokonać rozmytego kodowania przedziałów. Literatura z zakresu analizy danych symbolicznych nie podaje jednakże przykładów takich rozwiązań.

Problem rozmytego kodowania przedziałów liczbowych nierozłącznych z pewnością wymaga dalszej pogłębionej analizy. W artykule ograniczono się do usunięcia tego typu zmiennych z analizy.

## 5. Przykład empiryczny

W Banku Gospodarki Żywnościowej SA Oddział w Kłodzku zebrano informacje na temat decyzji kredytowych tego banku z 2004 r. Do badania wybrano w sposób przypadkowy 74 decyzje kredytowe, które są podzielone na dwa zbiory (klasy): k**lasa 1** – pozytywne decyzje kredytowe (wnioski zaakceptowane) – 60 obiektów, k**lasa 2** – negatywne decyzje kredytowe (odrzucone wnioski kredytowe) – 14 obiektów.

Obie grupy wniosków kredytowych opisywane są przez następujące zmienne:

1. Średnie wpływy na rachunek bieżący, liczone jako średnia przynajmniej sześciu najniższych wpływów i sześciu najwyższych wpływów – przedział liczbowy nierozłączny.

2. Staż pracy kredytobiorcy. Wskazany przez kredytobiorcę staż pracy jest w ramach scoringu kredytowego dopasowywany do *a priori* przyjętych przedziałów oceny – przedział liczbowy nierozłączny.

3. Czas trwania kredytu w miesiącach. Deklarowany we wniosku czas trwania kredytu jest oceniany w ramach przyjętych przez biuro scoringowe przedziałów – przedział liczbowy rozłączny.

4. Dochody kredytobiorcy. Podobnie jak średnie wpływy na rachunek bieżący, dochody są oceniane w ramach *a priori* przyjętych przedziałów dochodów – przedział liczbowy nierozłączny.

5. Otrzymana kwota kredytu. W przypadku osób, które kredytu nie otrzymały, jest to wnioskowana kwota kredytu oceniana w ramach przyjętych przedziałów – przedział liczbowy rozłączny.

6. Historia kredytowa, otrzymywana na podstawie danych Biura Informacji Kredytowej (BIK), Bankowego Rejestru Klientów Nierzetelnych (MIG BR) i wewnętrznych informacji prowadzonych przez BGŻ SA – lista kategorii.

7. Staż klienta w banku BGŻ SA, oceniany w przyjętych przedziałach – przedział liczbowy nierozłączny.

8. Wskazanie poręczyciela (poręczycieli) – lista kategorii.

9. Ocena poręczyciela, dokonywana na podstawie informacji Biura Informacji Kredytowej (BIK), Bankowego Rejestru Klientów Nierzetelnych (MIG BR) i wewnętrznych informacji prowadzonych przez BGŻ SA – lista kategorii.

10. Inne wskazane (proponowane) zabezpieczenia – lista kategorii.

11. Ocena klienta w BGŻ SA oznacza, czy klienta można uznać za stałego, tj. czy posiada rachunek bieżący ponad rok i czy na rachunek dokonywane są stałe wpływy – lista kategorii.

12. Lojalność klienta wobec BGŻ SA oznacza, czy klient korzysta lub korzystał z kilku produktów BGŻ SA – lista kategorii.

13. Udzielona informacja o sytuacji kredytowej, informacja udzielona przez wnioskodawcę we wniosku kredytowym – lista kategorii.

14. Przynależność obiektu do klasy 1 lub klasy 2, lub zbioru testowego – zmienna klasyczna, dychotomiczna.

W pierwszym podejściu dokonano transformacji zbioru zmiennych symbolicznych w klasyczne. Następnie dla takiego zbioru zmiennych obliczono kwadrat odległości euklidesowej i dokonano klasyfikacji wybranymi metodami bazującymi na macierzy odległości.

W drugim podejściu dokonano obliczenia macierzy odległości ważoną miarą Ichino-Yaguchiego (o jednakowych wartościach wag), która jest nazywana metryką euklidesową dla danych symbolicznych. Wynika to z tego, że jeżeli obiekty symboliczne opisywane będą tylko przez dane klasyczne, to miara Ichino-Yaguchiego przyjmuje takie samie wartości jak metryka euklidesowa (por. [*Analysis of symbolic*… 2000, s. 173]). W kolejnym kroku otrzymane wartości podniesiono do kwadratu i dokonano klasyfikacji tymi samymi metodami co w podejściu pierwszym.

Następnie dokonano porównania wyników otrzymanych w obu podejściach pod względem wskaźnika Rousseeuwa dla dwóch klas, liczby prawidłowo zaklasyfikowanych obiektów i odsetka prawidłowo zaklasyfikowanych obiektów. Dane te zestawiono w tab. 2.

Tabela 2. Wyniki symulacji

| Metoda klasyfikacji | Strategia I (z transformacją zmiennych) | | Strategia II (bez transformacji zmiennych) | |
|---|---|---|---|---|
| | wskaźnik Rousseeuwa (dla 2 klas) | liczba prawidłowo zaklasyfikowanych obiektów w klasach | wskaźnik Rousseeuwa (dla 2 klas) | liczba prawidłowo zaklasyfikowanych obiektów w klasach |
| Warda | 0,244 | Klasa 1 – 41<br>Klasa 2 – 13 | 0,761 | Klasa 1 – 59<br>Klasa 2 – 14 |
| Kompletnego połączenia | 0,343 | Klasa 1 – 3<br>Klasa 2 – 14 | 0,761 | Klasa 1 – 59<br>Klasa 2 – 14 |
| Średniej klasowej | 0,423 | Klasa 1 – 59<br>Klasa 2 – 0 | 0,722 | Klasa 1 – 58<br>Klasa 2 – 14 |
| Medianowa | 0,174 | Klasa 1 – 59<br>Klasa 2 – 0 | 0,761 | Klasa 1 – 59<br>Klasa 2 – 14 |
| Środka ciężkości | 0,423 | Klasa 1 – 59<br>Klasa 2 – 0 | 0,761 | Klasa 1 – 59<br>Klasa 2 – 14 |
| Ważonej średniej klasowej | 0,423 | Klasa 1 – 59<br>Klasa 2 – 0 | 0,722 | Klasa 1 – 58<br>Klasa 2 – 14 |
| *k*-medoidów | 0,308 | Klasa 1 – 35<br>Klasa 2 – 3 | 0,799 | Klasa 1 – 60<br>Klasa 2 – 14 |

Źródło: opracowanie własne z wykorzystaniem programu R.

W przypadku strategii I (z transformacją zmiennych) wielkość wskaźnika Rousseeuwa sugeruje, że nie odkryto struktury klas lub jest to słaba struktura klas. Należy również dodać, że metody klasyfikacji poprawnie identyfikowały zazwyczaj tylko przynależność obiektów z klasy 1 (pozytywnie rozpatrzone wnioski kredytowe).

W przypadku strategii II (bez transformacji zmiennych) wskaźnik Rousseeuwa za każdym razem wskazuje na silną strukturę klas, a większość obiektów jest poprawnie klasyfikowana. Najlepsze rezultaty w tym względzie osiąga metoda *k*-medoidów.

## 6. Wnioski

Dla zbioru danych symbolicznych można wykorzystać dwie strategie analizowania – strategię „klasyczną" i strategię „symboliczną". Strategia „klasyczna" wymaga przekształcenia zbioru zmiennych symbolicznych w klasyczne, co niesie za sobą utratę części informacji o obiektach, natomiast w strategii „symbolicznej" korzysta się z pełnego zbioru informacji o obiektach.

W klasyfikacji danych symbolicznych można zastosować dwa główne podejścia: klasyfikację bezpośrednio na podstawie tablicy danych symbolicznych lub klasyfikację na podstawie macierzy odległości.

Jeżeli podział obiektów dokonywany jest na podstawie macierzy odległości, to podstawą do jej obliczenia może być albo oryginalna tablica danych symbolicznych, albo macierz zmiennych powstałych z transformowania danych symbolicznych. Gdy w badaniu wykorzystano macierz odległości obliczoną na podstawie przetransformowanego zbioru zmiennych symbolicznych, otrzymano o wiele gorsze wyniki niż w przypadku korzystania z pełnego zbioru informacji o obiektach.

Otrzymane klasyfikacje ze zbioru transformowanego należy oceniać jako słabe z dwóch powodów. Po pierwsze, nie udało się w tym przypadku odkryć struktury klas (największa wartość wskaźnika Rousseeuwa to jedynie 0,423), a po drugie, jedynie obiekty z klasy 1 są klasyfikowane poprawnie. Najlepsze wyniki otrzymano, klasyfikując obiekty na podstawie macierzy odległości obliczanej z tablicy danych symbolicznych. Za najlepszą z przeanalizowanych metod klasyfikacji dla oryginalnego zbioru zmiennych należy uznać metodę *k*-medoidów (wskaźnik Rousseeuwa 0,799, a wszystkie obiekty zaklasyfikowano poprawnie).

Przedmiotem dalszych badań powinien się stać problem rozmytego kodowania przedziałów liczbowych nierozłącznych oraz symulacyjne porównania wyników klasyfikacji obiektów na podstawie tablicy danych symbolicznych z wynikami klasyfikacji na podstawie macierzy odległości.

## Literatura

*Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*, red. H.-H. Bock, E. Diday, Springer Verlag, Berlin-Heidelberg 2000.

Diday E., *Introduction á l'approche symbolique en analyse des données*, [w:] *Actes des journées symboliques-numériques pour l'apprentissage de connaissances á partir des données*, red. E. Diday, Y. Kodratoff, Ceremade, Université Paris IX Dauphine, Paris 1987, s. 21-56.

Malerba D., Esposito F., Giovalle V., Tamma V., *Comparing Dissimilarity Measures for Symbolic Data Analysis*, red. P. Nanopoulos, ‘‘New Techniques and Technologies for Statistcs and Exchange of Technology and Know-how” (ETK-NTTS’01) materiały konferencyjne, 2001, s. 473-481.
Malerba D., Esposito F., Monopoli M., *Comparing dissimilarity measures for probabilistic symbolic* objects, [w:] *Data Mining III*, red. A. Zanasi, C.A. Brebbia, N.F.F. Ebecken, P. Melli, “Series Management Information Systems”, vol. 6, WIT Press, Southampton 2002, s. 31-40.
Walesiak M., *Metody analizy danych marketingowych*, PWN, Warszawa 1996.

## COMPARISON OF SYMBOLIC DATA CLUSTERING STRATEGIES

### Summary

In this paper two clustering strategies for symbolic data are presented: the first strategy which requires a transformation of symbolic variables to classical ones what causes looses of some information about objects and the classification is based on distance matrix; the second strategy, where the whole information from the symbolic data table is applied and the classification can be done either from a distance matrix or directly from a symbolic data table.

In the empirical part the efficiency of those two strategies is compared to real data (with known cluster structure) regarding consumer credits of BGŻ bank clients in Kłodzko in 2004. The measurement of efficiency was done with Rousseeuw silhouette internal cluster quality index and also the number of correctly classified objects for each class.