PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176

Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Andrzej Dudek

Uniwersytet Ekonomiczny we Wrocławiu

ANALIZA DANYCH SYMBOLICZNYCH W ŚRODOWISKU R. PODSTAWY METODOLOGICZNE I PRZYKŁADY ZASTOSOWAŃ

Streszczenie: Analiza danych symbolicznych to gałąź wielowymiarowej analizy statystycznej, zajmująca się dużymi zbiorami danych (głównie pochodzącymi z komputerowych baz danych) zagregowanych w obiekty symboliczne, mogące zawierać dane w postaci liczb, tekstu, przedziałów liczbowych, zbiorów kategorii, zbiorów kategorii z wagami. Artykuł zawiera opis autorskiego pakietu symbolicDA, służącego do analizy danych symbolicznych w popularnym środowisku **R** i składa się z dwóch części. Pierwsza jest próbą umiejscowienia podejścia symbolicznego w badaniach statystycznych ze szczególnym uwzględnieniem zastosowań ekonomicznych, druga zaś prezentuje funkcje pakietu wraz z przykładami ich zastosowań.

Słowa kluczowe: dane symboliczne, metody symboliczne, środowisko **R**.

1. Wstęp

Gwałtowny rozwój nauk informatycznych w ostatnich latach XX wieku, a przede wszystkim powstawanie coraz większych baz danych spowodowało, że badacze dysponują coraz większą ilością danych wejściowych dla procedur taksonomicznych. Jednak nie zawsze informacje zawarte w dużych bazach danych mają postać umożliwiającą zastosowanie klasycznych metod klasyfikacji i analizy danych. W szczególności rzadko zdarza się, aby dane przybierały postać tabeli liczb, częściej bowiem występują jako dane nienumeryczne: tekstowe czy w postaci listy wartości. Dziedziną wielowymiarowej analizy statystycznej, wychodzącą naprzeciw tym wyzwaniom, jest analiza danych symbolicznych. Dyscyplina ta nie tylko inaczej określa dane używane w algorytmach i metodach w jej ramach opracowanych, ale również ze względu na charakter danych oraz przyjęte założenia metodologiczne trochę inaczej definiuje sam sposób ich analizy, co powoduje, że można

wyróżnić podejście symboliczne w analizie danych jako uzupełniające w stosunku do powszechnie przyjętego podziału na podejście ilościowe i podejście jakościowe.

Artykuł jest próbą zdefiniowania miejsca podejścia symbolicznego wśród metod analizy danych oraz opisu autorskiego narzędzia wspomagającego analizę danych symbolicznych w środowisku statystycznym **R**. Składa się z czterech części. W pierwszej z nich porównano podejście symboliczne w analizie danych z tradycyjnymi podejściami: ilościowym i jakościowym oraz przedstawiono pojęcia obiektu symbolicznego i zmiennych symbolicznych, a także scharakteryzowano dane symboliczne. Część druga to opis pakietu symbolicDA służącego do analizy danych symbolicznych, zaimplementowanych algorytmów oraz funkcji je realizujących. Części trzecia i czwarta zawierają przykłady zastosowań praktycznych z wykorzystaniem pakietu symbolicDA. Całość kończy podsumowanie i wskazanie kierunków popularyzacji podejścia symbolicznego.

2. Podejście symboliczne a podejście jakościowe i ilościowe

Rozróżnienie na podejście jakościowe[1] i podejście ilościowe w badaniach jest obecne w naukach społecznych od końca XIX wieku. Początkowo można było wręcz mówić o istnieniu konfliktu między badaczami „jakościowymi" a „ilościowymi", natomiast obecnie uważa się te podejścia za komplementarne metodologie badań, a wiele metod jakościowych jest wspomaganych częścią ilościową lub semiilościową i odwrotnie, wiele metod ilościowych może być potwierdzonych lub odrzuconych poprzez badania jakościowe.

Do podstawowych cech rozróżniających te dwa podejścia w naukach społecznych[2] należą (por. np. [Silverman 2007, 2008; Nikodemska-Wołowik 1999; Neil, *Qualitative*...]):

Wykorzystywanie obliczeń w badaniach. W podejściu jakościowym nie dokonuje się obliczeń lub obliczenia są prowadzone w ostatniej części badań i mają

[1] W literaturze przedmiotu istnieją przynamniej dwa znaczenia słowa „jakościowe" (*qualitative*). Tradycyjnie w naukach społecznych oznaczają one badania, w których nie dokonuje się obliczeń lub obliczenia pełnią funkcję czysto utylitarną w końcowej fazie badań. O badaniach jakościowych, jako o badaniach bez obliczeń, piszą np. Silverman [2007, 2008], Marshall i Rossman [2009], Wolcott [2010]. Jednakże w ekonometrii i wielowymiarowej analizie statystycznej dane jakościowe, to dane niemetryczne, mierzone na słabych skalach pomiarowych. W takim znaczeniu sformułowania „jakościowe" używają np. Manski, McFadden [1984], Maddala [1986] czy Gatnar [1993]. Aby uniknąć nieporozumień interpretacyjnych, w dalszej części artykułu podejście jakościowe będzie oznaczało − zgodnie na przykład z definicją Nikodemskiej-Wołowik [1999] − podejście badawcze, w którym nie są dokonywane obliczenia, natomiast „podejście jakościowe" w rozumieniu Maddali nazywane będzie analizą danych kategorialnych.

[2] W naukach ekonomicznych najbardziej wyraźnie widać rozróżnienie na podejście jakościowe i podejście ilościowe w badaniach marketingowych (por. np. [Nikodemska-Wołowik 1999; Kaczmarczyk 2007, s. 86, 93, 121, 132]).

funkcję czysto użytkową, nacisk zaś jest położony na odpowiednie wnioskowanie dotyczące przedmiotów badań. W podejściu ilościowym, z samej natury, obliczenia odgrywają kluczową rolę w analizie.

Wielkość próby. W podejściu jakościowym wielkość próby jest zazwyczaj ograniczona. W badaniach ilościowych powinna być na tyle duża, aby umożliwiała tworzenie modeli stochastycznych.

Typ danych. W podejściu jakościowym są to dane tekstowe (lub w postaci multimediów, obrazów, dźwięków), w podejściu ilościowym danymi są liczby lub kategorie przekodowane na liczby.

Orientacja badania. Badania jakościowe są nastawiane zazwyczaj na eksplorację, odkrywanie zależności logicznych i prawidłowości o charakterze deterministycznym. Badania ilościowe służą do weryfikacji postawionych twierdzeń lub do znajdowania zależności stochastycznych.

Analiza danych symbolicznych, którą w najprostszej postaci można rozumieć jako zbiór technik i algorytmów dotyczących danych reprezentowanych w szerszej niż numeryczna postaci, jest równocześnie próbą odpowiedzi na wiele wyzwań dotyczących konstrukcji nowoczesnych metod badawczych związanych z rozwojem technik informatycznych. Dodatkowo, oprócz samych obliczeń, definiuje ona określanie zależności między zmiennymi i uzyskiwanie reguł symbolicznych, co Diday [Diday, Noirhomme-Fraiture 2008]) nazywa przejściem od akwizycji danych do akwizycji wiedzy (*from data acquisition to knowledge acquisition*). Wraz z nową postacią danych definiuje ona również inne podejście w ich analizie, łączące cechy podejścia ilościowego i jakościowego. Wykorzystuje też w swoich metodach obliczenia, jednak równie silny nacisk jak na wyniki liczbowe jest kładziony na reguły symboliczne otrzymywane w trakcie badania. Badania są przeprowadzane na dużych próbach, ale z możliwością agregacji danych pierwotnych w mniejszą liczbę obiektów symbolicznych (tzn. obiektów symbolicznych II rzędu). Dane w podejściu symbolicznym odzwierciedlają rzeczywistość dużo pełniej niż dane numeryczne, choć są na tyle uporządkowane, aby możliwa była ich analiza przez odpowiednie oprogramowanie. Wreszcie badania w tym podejściu mogą być zorientowane zarówno na weryfikację zadanych tez, jak i na znajdowanie zależności logicznych w modelowanym fragmencie rzeczywistości.

Berry i Linoff [1997] definiują *Data Mining* jako eksplorację i analizę w sposób automatyczny i semiautomatyczny dużych zbiorów danych w celu odkrycia znaczących wzorów i reguł w nich zawartych. W tym rozumieniu analizę danych symbolicznych można traktować jako część *Data Mining*. Z drugiej jednak strony *Data Mining* (por. np. [Olson, Delen 2008]) jest ściśle związane z komputerowymi bazami danych, natomiast analiza danych symbolicznych może zajmować się danymi z baz danych, zagregowanymi w obiekty symboliczne, ale może również analizować dane z innych źródeł.

Jajuga [1993] wyróżnia w metodach wielowymiarowej analizy porównawczej podejście opisowe i stochastyczne. Z tego punktu widzenia analiza danych symbolicznych jest również nową jakością, gdyż łączy w sobie cechy obu podejść (metody podejścia opisowego dla danych, które same w sobie zawierają informację o rozkładach zmiennych).

Diday [Diday, Noirhomme-Fraiture 2008], odwołując się do koncepcji idei Platona i Organona Arystotelesa (1994 r. p.n.e. – porównaj [Smith, Su 2008]), przedstawia pojęcie obiektu symbolicznego opisującego poszczególne obiekty jednostkowe. Obiekt symboliczny ma jednocześnie wnętrze w postaci zbioru reguł lub danych w złożonych strukturach, a także zewnętrze zawierające opisywane obiekty jednostkowe.

Diday [2002] wyróżnia kilka sytuacji, w których można stosować dane symboliczne do analizy danych. Wśród nich są:

– przetwarzanie danych pochodzących z dużych baz danych,
– przetwarzanie danych mających charakter jakościowy (kategorialny),
– przetwarzanie danych połączonych silną logiczną zależnością między zmiennymi.

Wychodząc z typowego dla podejścia symbolicznego założenia, że same wartości liczbowe nie wystarczą do pełnego przedstawienia modelowanej w procesie analizy rzeczywistości, można wyróżnić pięć typów zmiennych symbolicznych, których łączne stosowanie do opisu danych powinno dawać dużo pełniejsze odwzorowanie świata rzeczywistego. Są to:

- Zmienne o realizacjach w postaci pojedynczych wartości liczbowych.
- Zmienne o realizacjach w postaci łańcuchów tekstowych.
- Zmienne o realizacjach w postaci przedziałów liczbowych rozłącznych lub nierozłącznych. Zmienne tego typu noszą nazwę zmiennych symbolicznych interwałowych. Zmienne symboliczne o realizacjach w postaci listy kategorii. Zmienne tego typu nazywane są zmiennymi wielowariantowymi lub wielokategorialnymi.
- Zmienne symboliczne o realizacjach w postaci listy wartości z wagami. Noszą one nazwę zmiennych wielowariantowych z wagami lub zmiennych wielokategorialnych z wagami.

3. Pakiet symbolicDA

Pakiet symbolicDA zawiera implementacje najważniejszych algorytmów analizy danych symbolicznych w środowisku statystycznym **R**. Składa się on z 17 funkcji głównych (oraz funkcji pomocniczych, które ze względu na ograniczenia wielkości artykułu nie będą opisywane). Cztery z nich (`SO2Simple`, `simple2SO`, `generate.SO`, `parse.SO`) to funkcje użytkowe przeznaczone do zmiany formatu danych pomiędzy danymi pełnymi (zawierającymi obiekty opisane zmiennymi symbolicznymi dowolnego typu) oraz danymi w postaci uproszczonej (opisane tylko zmien-

nymi o realizacjach w postaci przedziałów liczbowych), do generowania zbiorów danych symbolicznych o określonej strukturze klas oraz do ładowania zbiorów danych symbolicznych z plików zewnętrznych w standardzie XML[3].

Funkcje `zoomStar`, `plot3dInterval` i `plot.decisionTree.SDA` służą do graficznej prezentacji danych symbolicznych i wyników działania poszczególnych metod.

Funkcja `dist.SDA` oblicza odległości pomiędzy obiektami symbolicznymi. Umożliwia obliczenie odległości Ichino-Yaguchiego [Ichino, Yaguchi 1994] dowolnego typu, miar de Carvalho opartych na funkcji porównującej, miar de Carvalho [de Carvalho, Souza 1998] opartych na pojęciu potencjału opisowego obiektu, miar odległości dla zmiennych symbolicznych probabilistycznych.

Funkcja `cluster.Description.SDA` służy do tworzenia charakterystyk skupień powstałych w wyniku zastosowania metod klasyfikacji dla danych symbolicznych.

Pozostałe funkcje są implementacjami najważniejszych algorytmów podejścia symbolicznego. Są to odpowiednio:

`DClust` Agorytm dynamicznej analizy skupień oparty na macierzach odległości [Verde, Lechevalier, Chavent 2003].

`SClust` Algorytm dynamicznej analizy skupień oparty na tablicy danych symbolicznych [Verde 2004].

`decisionTree.SDA` Drzewo klasyfikacyjne dla obiektów symbolicznych oparte na procedurze optymalnego podziału [Bock, Diday (red.) 2000, s. 244-265].

`HINoV.SDA` Modyfikacja algorytmu Heuristic Identification of Noisy Variables [Carmone, Kara, Maxwell 1999] − dla danych symbolicznych. Metoda służy do selekcji zmiennych zakłócających w I etapie procedury klasyfikacyjnej) [Carmone, Kara, Maxwell 1999].

`IchinoFS.SDA` Metoda Ichino selekcji zmiennych [Ichino 1994].

`kernel.SDA` Jądrowa analiza dyskryminacyjna dla danych symbolicznych [Bock, Diday (red.) 2000, s. 244-265].

`kohonen.SDA` Samoorganizujące się mapy Kohonena dla danych symbolicznych interwałowych [El Golli, Conan-Guez, Rossi 2004].

Funkcje pakietu symbolicDA zostaną wykorzystane w przykładowych analizach przedstawionych w dalszej części artykułu.

4. Klasyfikacja województw pod względem struktury bezrobocia

W badaniu wykorzystano zbiór danych symbolicznych, składający się z 16 obiektów symbolicznych odpowiadających województwom i zawierających informację

[3] Do przygotowywania tego typu plików może posłużyć SDAEditor (http://keii.ae.jgora.pl/sdaeditor) lub program SODAS (http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm).

o strukturze osób pozostających bez pracy pod koniec 2007 r., opisanych zmiennymi: *płeć* – procentowy udział kobiet i mężczyzn w ogólnej liczbie bezrobotnych, *wiek* – procentowy udział osób w wieku 18-24; 25-34; 35-44; 45-54; 55-59; 60-65 w ogólnej liczbie bezrobotnych, *wykształcenie* – procentowy udział osób z różnymi typami wykształcenia w ogólnej liczbie bezrobotnych, *staż* – procentowy udział osób mających odpowiednio 0-1; 1-5; 5-10; 10-20; 20-30; 30 i więcej lat stażu pracy w ogólnej liczbie bezrobotnych, *bez_pracy* − procentowy udział osób pozostających bez pracy odpowiednio 1-3; 3-6; 6-12; 12-18; 18-24; 24 i więcej miesięcy w ogólnej liczbie bezrobotnych.

Dla danych tych zastosowano metodę dynamicznej klasyfikacji obiektów symbolicznych SCLUST za pomocą następującego ciągu poleceń języka **R**.

```
library(symbolicDA);library(clusterSim)
sdt<-parse.SO("bezrobocie")
klasy<-SClust(sdt,3)
index.G2(dist.SDA(sdt,type="L_2"),klasy)
index.G3(dist.SDA(sdt,type="L_2"),klasy)
```

Rys. 1. Klasyfikacja województw Polski według struktury bezrobotnych (poszczególne klasy oznaczone są różnymi kolorami)

Źródło: opracowanie własne.

Dla podanej klasyfikacji indeks Bakera-Huberta wynosi 0,7268, a indeks Huberta-Levine − 0,0677, co świadczy o czytelnej strukturze klas. Na rysunku 1 przedstawiono strukturę otrzymanych skupień na mapie Polski. Należy przy tym podkreślić, że w przeciwieństwie do większości tego typu analiz dla danych w postaci klasycznej macierzy danych numerycznych, klasyfikacja ta dotyczy *struktury* osób bezrobotnych, a nie ilościowego porównania.

5. Przykład zastosowania drzew klasyfikacyjnych dla obiektów symbolicznych w ocenie wiarygodności kredytobiorców

W badaniu wykorzystano zbiór danych symbolicznych zawierających informacje o tysiącu kredytobiorców niemieckich banków. Dane opisane są następującymi zmiennymi symbolicznymi: *kredyt* − zmienna nominalna przyjmująca wartości 1 (spłacany z problemami) i 2 (spłacany planowo), *okres* − zmienna symboliczna interwałowa, *poprzednie* − informacja o poprzednio zaciągniętych kredytach, zmienna nominalna, *cel* − zmienna nominalna, *wysokosc* − zmienna symboliczna interwałowa, *oszczednosci* zmienna symboliczna interwałowa, zatrudnienie − zmienna symboliczna interwałowa, *procentDochodow* − zmienna symboliczna interwałowa, *plec* − zmienna nominalna, *zyranci* − zmienna wielokategorialna, *wlasnosc* − zmienna nominalna, *wiek* − zmienna symboliczna interwałow, *inneKredyty* − zmienna nominalna, *lokum* − zmienna nominalna, *poprzednieKredyty* − zmienna nominalna, *zajecie* − zmienna nominalna, *obcokrajowiec* − zmienna nominalna.

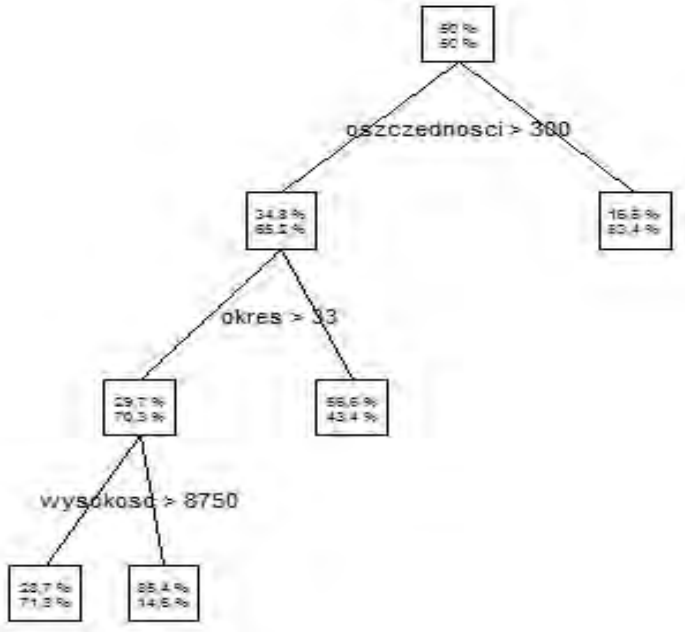
Zbiór wejściowy został podzielony na zbiór uczący o liczebności 928 obiektów i zbiór testowy o liczebności 72 obiekty.

Następujący ciąg poleceń języka R tworzy drzewo klasyfikacyjne dla danych symbolicznych, oparte na procedurze optymalnego podziału.

```
testSet<-c(650:690,720:750)
dt<-decisionTree.SDA(sdt,"kredyt~.",testSet)
.plot.decisionTree.SDA(dt)predicted<-dt$prediction
variables<-as.matrix(sdt$variables)
indivN<-as.matrix(sdt$indivN)
categorialVariable<-
as.numeric(variables[variables[,"label"]=="kredyt","num"])
classes<-
as.numeric(indivN[indivN[,"variable"]==categorialVariable,
"value"])
actualClasses<-classes[testSet]
detailsListNom<-as.matrix(sdt$detailsListNom)
cat<-detailsListNom[detailsListNom[,"details_no"]==variables[
categorialVariable,"details"],"label"]
print(paste("Przydział do klas poprzez predykcję z
wykorzystaniem drzewa klasyfikacyjnego
kategorie:",paste(cat[predicted], collapse=";")))
print(paste("Błąd klasyfikacji:",100*(1-
sum(apply(cbind(actualClasses,
predicted),1,function(x){if(x[1]==x[2]){1}else{0}}))/length(tes
tSet) ),"procent"))
```

W wyniku zastosowania algorytmu otrzymano drzewo o dziewięciu węzłach (w tym pięciu węzłach końcowych).

Na rysunku 2 przedstawiono drzewo klasyfikacyjne otrzymane na podstawie zbioru uczącego przy założeniu, że zmienną dyskryminującą jest zmienna *kredyt*.



Rys. 2. Drzewo klasyfikacyjne określające przynależność do klas kredytobiorców

Źródło: opracowanie własne.

Algorytm zakwalifikował poprawnie 66 kredytobiorców, a niepoprawnie 6. Otrzymany błąd klasyfikacji wyniósł 8,3%.

6. Podsumowanie

W artykule przedstawiono podstawy metodologiczne podejścia symbolicznego w analizie danych oraz program symbolicDA, służący do analizy danych symbolicznych w środowisku statystycznym R. Program ten powstał w ramach projektu badawczego MNiSW N N111 105234 „Obiekty symboliczne w wielowymiarowej analizie statystycznej" i jest dostępny na licencji GPL 2.0 pod adresem htpp://keii.ae.jgora.pl /symbolicDA. Po przejściu odpowiednich testów zgodności będzie również umieszczony w repozytorium pakietów środowiska R (http://cran.r-project.org/).

Mimo że analiza danych symbolicznych rozwija się już od ponad 20 lat, w literaturze przedmiotu zaobserwować można brak równowagi między podstawami teo-

retycznymi a sferą zastosowań. O ile ta pierwsza sfera rozwija się dość szybko i cały czas pojawiają się nowe propozycje metod, a istniejące algorytmy są ciągle udoskonalane, o tyle w obszarze ich zastosowań praktycznych istnieją dość istotne luki. Być może wynika to z braku standaryzowanego oprogramowania komputerowego do tego typu analizy danych. Autor wyraża nadzieje, że zaproponowane oprogramowanie pomoże zmienić ten stan rzeczy, przyczyniając się do popularyzacji tego podejścia i owocując ciekawymi praktycznymi przykładami zastosowań.

Literatura

Berry M.J.A, Linoff G.S., *Data Mining Techniques*: *For Marketing, Sales, and Customer Support*, John Willey & Sons, New York 1997.

Billard L., Diday E. (red.), *Symbolic Data Analysis, Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester 2006.

Bock H.H., Diday E. (red.), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin 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" 1999, 36, s. 501-509.

de Carvalho F.A.T., Souza R., *Statistical Proximity Functions of Boolean Symbolic Objects Based on Histograms*, [w:] *Advances in Data Science and Classification*, Springer-Verlag, Heidelberg 1998, s. 391-396.

Diday E., *An introduction to symbolic data analysis and the SODAS software*, J.S.D.A., International e-Journal, 2002.

Diday E., Noirhomme-Fraiture M. (red.), *Symbolic Data Analysis with SODAS Software*, John Wiley & Sons, Chichester 2008.

El Golli A., Conan-Guez B., Rossi F., *Self organizing map and symbolic data*, „Journal of Symbolic Data Analysis" 2004, vol. 2.

Gatnar E., *Modelowanie jakościowe zjawisk ekonomicznych*, praca doktorska (rękopis) [1993].

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

Ichino, M., Yaguchi H., *Generalized Minkowski metrics for mixed feature-type data analysis*, „IEEE Transactions on Systems, Man, and Cybernetics" 1994, vol. 24, no. 4, s. 698-707.

Jajuga K., *Statystyczna analiza wielowymiarowa*, PWN, Warszawa 1993.

Maddala G.S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, 1986.

Manski C., McFadden D., (red.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge 1981.

Mardia K.V., Kent J.T., Bibby J.M., *Multivariate Analysis*, Academic Press, London 1979.

Marshall C., Rossman G.B., *Designing Qualitative Research*, SAGE Publicatons, Thousand Oks − London − New Delhi 2009.

Neil J., *Qualitative versus Quantitative Research*: URL: http://wilderdom.com/research/ualitativeer susQuantitativeResearch.html (4.10.2010).

Nikodemska-Wołowik A.M., *Jakościowe badania marketingowe*, PWE, Warszawa 1999.

Olson D.L., Delen D., *Advanced Data Mining Techniques*, Springer-Verlag, Berlin − Heidelberg 2008.

Silverman D., *Prowadzenie badań jakościowych*, Wyd. Naukowe PWN, Warszawa 2007.

Silverman D., *Interpretacja badań jakościowych*, Wyd. Naukowe PWN, Warszawa 2008.

Smith W.E., Su A.G. (2008), *Translations from Organon of Aristotle*, Google Books (http://books.ooglecom/books?id=MEpl lLqYC&printsec=frontcover&dq=aristotle+organon&hl=pl).

Verde R., *Clustering Methods in Symbolic Data Analysis, Classification*, [w:] D. Banks, L. House, P. Arabie, W. Gaul (red.), *Clustering and Data Mining*, Springer-Verlag, Berlin 2004, s. 299-318.

Verde R., Lechevalier Y., Chavent M, *Symbolic clustering interpretation and visualization*, „The Electronic Journal of Symbolic Data Analysis" 2003, vol. 1, no. 1.

Wolcott H.F., *Writing up qualitative research*, Sage Publications, Thousand Oaks − London − New Delhi 2010.

ANALYSIS OF SYMBOLIC DATA IN R ENVIRONMENT. METHODOLOGICAL BACKGROUND AND EXAMPLES OF USAGE

Summary: Symbolic data analysis is a branch of a multivariate statistical analysis dealing with large data sets (mainly gained from computer data bases) aggregated into symbolic objects which may include the data in the form of numbers, text strings, intervals, sets of categories and weighted sets of categories. The paper describes symbolicDA software package developed by the author, dedicated to the analysis of symbolic data in popular statistical **R** environment (http://www.r-project.org/). It is divided in two main parts. The first part is a trial defining and placing a symbolic approach in statistical research with special concern on economic applications. The second part presents procedures and functions of package along with some empirical usage examples.