PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 384

Taksonomia 24
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronie internetowej Wydawnictwa
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 384 • 2015
Taksonomia 24 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu
e-mail: marcin.pelka@ue.wroc.pl

ADAPTACJA METODY *BAGGING* Z ZASTOSOWANIEM KLASYFIKACJI POJĘCIOWEJ DANYCH SYMBOLICZNYCH

Streszczenie: Podejście wielomodelowe może być z powodzeniem zastosowane w zagadnieniach dyskryminacyjnych i regresyjnych analizy danych. Podejście to może zostać także zastosowane w klasyfikacji. W odróżnieniu od obiektów w sensie klasycznym obiekty symboliczne mogą być opisywane także przez zmienne interwałowe, zmienne wielowariantowe, zmienne wielowariantowe z wagami, zmienne interwałowe z wagami oraz zmienne strukturalne. Celem artykułu jest zaproponowanie zastosowania metody *bagging*, z użyciem adaptacji propozycji Leischa [1999], w klasyfikacji wielomodelowej danych symbolicznych. W wyniku wykorzystania tego podejścia otrzymuje się wyniki w postaci klas opisywanych przez pojęcia. W części empirycznej przedstawiono wyniki badań symulacyjnych z wykorzystaniem rzeczywistych i sztucznych zbiorów danych symbolicznych.

Słowa kluczowe: klasyfikacja wielomodelowa, dane symboliczne, klasyfikacja pojęciowa.

DOI: 10.15611/pn.2015.384.24

1. Wstęp

Cechą odróżniającą obiekty symboliczne od klasycznych jest fakt, że obiekty symboliczne mogą być opisywane przez wiele różnych typów zmiennych. Oprócz zmiennych w ujęciu klasycznym (metrycznych lub niemetrycznych) mogą one być opisywane przez zmienne interwałowe, zmienne wielowariantowe i zmienne wielowariantowe z wagami, zmienne histogramowe, a także zmienne strukturalne (zob. np. [Bock, Diday 2000, s. 2-3]). Pozwala to z jednej strony na dokładniejszy opis obiektów, ale z drugiej strony utrudnia analizę skupień.

Podejście wielomodelowe polega na łączeniu wyników otrzymanych za pomocą wielu modeli celem otrzymania jednego, bardziej dokładnego modelu zagregowanego. Idea ta była z powodzeniem stosowana w rozwiązywaniu zagadnień z zakresu dyskryminacji i regresji (zob. np. [Gatnar 2008]). Niemniej idea podej-

ścia wielomodelowego może być z powodzeniem zastosowana także w zagadnieniu klasyfikacji danych symbolicznych. Podejście wielomodelowe w klasyfikacji oznacza łączenie (czyli agregację) wielu klasyfikacji (inaczej modeli) bazowych w jedną klasyfikację złożoną (por. [Fred, Jain 2005]).

Celem artykułu jest zaproponowanie zastosowania metody *bagging*, z użyciem adaptacji propozycji Leischa [1999], w klasyfikacji wielomodelowej danych symbolicznych. W wyniku wykorzystania tego podejścia otrzymuje się wyniki w postaci klas opisywanych przez pojęcia. W części empirycznej przedstawiono wyniki badań symulacyjnych z wykorzystaniem rzeczywistych i sztucznych zbiorów danych symbolicznych.

2. Dane symboliczne

Obiekty symboliczne mogą być opisywane przez następujące rodzaje zmiennych [Bock, Diday (red.) 2000, s. 2-3; Billard, Diday 2006, s. 7-30; Dudek 2013, s. 35-36]:

- zmienne nominalne,
- zmienne porządkowe,
- zmienne przedziałowe,
- zmienne ilorazowe,
- zmienne interwałowe – czyli przedziały liczbowe,
- zmienne wielowariantowe – czyli listy kategorii lub wartości,
- zmienne wielowariantowe z wagami – czyli listy kategorii z wagami,
- zmienne histogramowe – czyli listy wartości z wagami.

Szerzej o obiektach i zmiennych symbolicznych, sposobach otrzymywania zmiennych symbolicznych z baz danych, różnicach i podobieństwach między obiektami symbolicznymi a klasycznymi piszą m.in.: Bock, Diday (red.) [2000, s. 2-8], Dudek [2013, s. 42-43; 2004], Billard, Diday [2006, s. 7-66]; Noirhomme-Fraiture, Brito [2011]; Diday, Noirhomme-Fraiture [2008, s. 3-30].

3. Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym

W przypadku podejścia wielomodelowego w analizie skupień dla danych symbolicznych literatura przedmiotu pozwala rozróżnić trzy główne podejścia (zob. [de Carvalho i in. 2012; Pełka 2012a; Fred, Jain 2005; Ghaemi i in. 2009; Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999]):

1. Łączenie wyników wielu klasyfikacji bazowych[1].
2. Łączenie wielu macierzy odległości, z których każda jest traktowana jako od-

[1] Szerzej o tych metodach piszą m.in.: Ghaemi i in. [2009], Pełka [2012a] oraz Fred i Jain [2005].

rębny punkt widzenia na zbiór danych. Klasyfikacja polega tu na połączeniu (zagregowaniu) informacji z różnych macierzy odległości[2].

3. Adaptacja metody bagging na potrzeby podejścia wielomodelowego w klasyfikacji.

W klasyfikacji pojęciowej, w przeciwieństwie do klasyfikacji w rozumieniu tradycyjnym (gdzie klasa reprezentowana jest przez wektor średnich czy medoidę), klasa reprezentowana jest przez pojęcie.

„Pojęcie jest poznawczą reprezentacją skończonej liczby wspólnych cech, które w jednakowym stopniu przysługują wszystkim reprezentantom (desygnatom) danej klasy" (cyt. za [Gatnar 1998, s. 71]). Oznacza to, że obiekty przynależą do klasy nie na podstawie miar bliskości czy odległości, ale na podstawie pewnych wspólnych dla nich cech.

W wyniku zastosowania klasyfikacji pojęciowej otrzymuje się zwykle etykiety klas oraz pojęcia reprezentujące klasy. Dla niektórych metod klasyfikacji pojęciowej wynikiem klasyfikacji oprócz etykiet klas oraz pojęć reprezentujących te klasy są także reguły przynależności obiektów do klas (zob. np. [Gatnar 1998]).

W części empirycznej artykułu zastosowano algorytm klasyfikacji hierarchicznej/piramid P. Brity. Jest to metoda klasyfikacji pojęciowej, która pozwala na otrzymanie klas nierozłącznych (metoda piramid) albo klas rozłącznych (metoda hierarchiczna). Idea tych metod opiera się na podejściu zaproponowanym przez Didaya i Britę [1989], gdzie każda klasa reprezentowana jest przez syntetyczny obiekt symboliczny (pojęcie).

Konstrukcja dendrogramu klas, podobnie jak w przypadku klasycznej metody hierarchicznej, zakłada, że w pierwszym kroku klasyfikacji obiekty symboliczne tworzą klasy jednoelementowe i jednocześnie są pojęciami reprezentującymi te klasy. W następnych krokach spośród obiektów (klas) poszukuje się takich par P_i oraz P_j, aby klasa powstała w wyniku ich połączenia (AS_t) była kompletna. Spośród tych par obiektów wybiera się to połączenie dla którego współczynnik uogólnienia (*generality degree*) jest najmniejszy [Dudek 2013, s. 77-78; Billard, Diday 2006, s. 294; Diday, Noirhomme-Fraiture (red.) 2008, s. 163]:

$$G\left(AS_t\right) = \prod_{k=1}^{m} \frac{\mu\left(AS_{t,k}\right)}{\mu\left(AS_{\max,k}\right)},$$

gdzie: $AS_{t,k}$ – syntetyczny obiekt symboliczny (pojęcie) reprezentujące klasę, które powstało w t-tym kroku algorytmu; $AS_{max,k}$– syntetyczny obiekt symboliczny odpowiadający całemu analizowanemu zbiorowi danych; $k = 1,\ldots,m$ – numer zmiennej, $t = 1, \ldots, T$ – numer kolejny kroku w algorytmie.

Następnie redukuje się liczbę klas o jeden i ponownie poszukuje takiej pary obiektów (klas), aby klasa powstała w ich połączeniu była kompletna, a współczynnik uogólnienia najmniejszy.

[2] Szerzej piszą o tym de Carvalho i in. 2012.

W artykule Pełki [2014] zaprezentowano adaptację podejścia wielomodelowego z wykorzystaniem klasyfikacji pojęciowej danych symbolicznych jako klasyfikatora bazowego. Zastosowano przy tym ideę macierzy współwystąpień (*co-occurence matrix, co-association matrix*) jako sposobu łączenia wielu klasyfikacji. Wykorzystanie macierzy współwystąpień pozwoliło na otrzymanie w miarę stabilnych (w sensie skorygowanego indeksu Randa) wyników klasyfikacji, ale niestety wynik klasyfikacji zagregowanej nie był reprezentowany przez pojęcia. Rozwiązaniem tego problemu może być zastosowanie metody *bagging* do łączenia wyników klasyfikacji bazowych.

Metoda *bagging* jest jedną z bardziej znanych metod agregacji modeli bazowych w przypadku zagadnień dyskryminacyjnych lub regresyjnych por. ([Gatnar 2008, s. 140; Kuncheva 2004, s. 203]). Metoda ta wykorzystuje w swej konstrukcji architekturę równoległą modeli zagregowanych [Gatnar 2008, s. 68]. Metoda *bagging* polega na zbudowaniu M modeli bazowych na podstawie prób uczących $U_1,\ldots,U_M$ losowanych ze zwracaniem ze zbioru uczącego. Próby te nazywane są próbami bootstrapowymi [Gatnar 2008, s. 140]. W przypadku dyskryminacji stosuje się metodę głosowania większościowego natomiast w przypadku regresji wyniki są uśredniane [Gatnar 2008, s. 240; Kuncheva 2004, s. 204].

Zastosowanie metody *bagging* w klasyfikacji wymagało opracowania nieco innych rozwiązań niż proste głosowanie większościowe czy uśrednianie wyników. W literaturze przedmiotu zaproponowano trzy adaptacje metody *bagging* na potrzeby klasyfikacji [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999]:

1. Propozycja Leischa [1999], która zostanie zastosowana w części empirycznej artykułu:

- utworzenie kolejnych prób bootstrapowych (podprób),
- klasyfikacja podprób z zastosowaniem bazowej metody klasyfikacji (zwykle jest to algorytm iteracyno-optymalizacyjny) – w części empirycznej zastosowany zostanie tu algorytm hierarchiczny P. Brity,
- centra skupień z każdego podziału (w przypadku klasyfikacji pojęciowej będą to pojęcia reprezentujące klasy) przekształcane są w nowy zbiór danych, który poddawany jest klasyfikacji (zwykle stosowana jest tu jedna z metod hierarchicznych) – w części empirycznej zastosowana zostanie tu metoda hierarchiczna P. Brity,
- otrzymany dendrogram klas jest dzielony na klasy w celu otrzymania obserwacji podobnych do siebie,
- obserwacje z pierwotnego zbioru danych są przydzielane do tej klasy, której zalążek znajduje się najbliżej. W przypadku zastosowania klasyfikacji pojęciowej obiekty będą przydzielane do klas zgodnie z regułami tejże klasyfikacji.

2. Propozycja przedstawiona w pracy Dudoit i Fridlyand [2003],

- utworzenie prób bootstrapowych,
- zastosowanie algorytmu iteracyjno-optymalizacyjnego do oryginalnego zbioru danych i utworzonych podprób. W przypadku zastosowania klasyfikacji poję-

ciowej danych symbolicznych należałoby zastosować ten sam algorytm dla zbioru danych oraz podprób,
- dokonanie permutacji etykiet klas dla obiektów z prób bootstrapowych, tak aby zachodziła jak największa zgodność z etykietami dla obiektów z oryginalnego zbioru danych,
- zastosowanie głosowania majoryzacyjnego w celu określenia ostatecznych wyników klasyfikacji.

3. Propozycja Hornika [2005]:
- utworzenie prób bootstrapowych,
- zastosowanie klasycznego algorytmu klasyfikacyjnego dla każdej z nich. W przypadku klasyfikacji pojęciowej będzie to zastosowanie jej algorytmu dla każdej z tych podprób,
- uzyskanie ostatecznego podziału poprzez optymalizację funkcji:

$$\sum_{b=1}^{B} dist\left(c, c_b\right)^2 \Rightarrow \min_{c \in C},$$

gdzie: C – zbiór wszystkich możliwych klasyfikacji zagregowanych; $c_b \in \left(c_1, \ldots, c_B\right)$ – elementy klasyfikacji zagregowanej; *dist* – miara odległości euklidesowej (w przypadku danych symbolicznych musi to być jedna z miar odległości adekwatna dla tego typu danych).

4. Przykład empiryczny

Na potrzeby badań empirycznych przygotowano dwa sztuczne zbiory danych symbolicznych interwałowych. Sztuczne zbiory danych wygenerowano z zastosowaniem funkcji `cluster.Gen` z pakietu `clustersim` [Walesiak, Dudek 2014]. Zbiór danych I to 100 0biektów symbolicznych podzielonych na dwie klasy o wydłużonych kształtach, które są opisywane przez dwie zmienne symboliczne interwałowe. Obserwacje w tych klasach wylosowano z rozkładu normalnego o średnich (0, 0), (1, 5) oraz macierzy kowariancji $\sum\left(\sigma_{jj} = 1, \sigma_{jl} = -0,9\right)$. Zbiór danych II to 150 obiektów symbolicznych podzielonych na trzy klasy o wydłużonym kształcie, które są opisywane przez dwie zmienne symboliczne interwałowe. Obserwacje w tych klasach wylosowano z rozkładu normalnego o średnich (0, 0), (1,5, 7) (3, 14) oraz macierzy kowariancji $\sum\left(\sigma_{jj} = 1, \sigma_{jl} = -0,9\right)$.

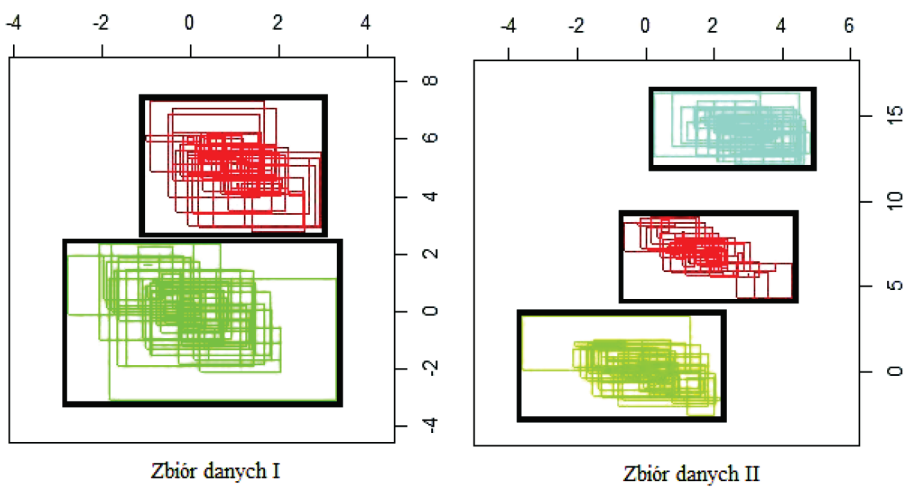
Następnie każdy ze zbiorów danych podzielono na 30 podzbiorów. W przypadku zbioru danych I losowano z niego 67 obiektów, a przypadku zbioru danych II losowano 100 obiektów. Do każdego z podzbiorów zastosowano metodę pojęciowej klasyfikacji hierarchicznej (z różnymi liczbami klas), otrzymując pojęcia

opisujące klasy. Pojęcia te utworzyły nowy zbiór danych, do którego zastosowano pojęciową klasyfikację hierarchiczną, otrzymując ostateczny podział na klasy (reprezentowane przez pojęcia). Obiekty, które nie znalazły się w żadnym z podzbiorów, przydzielono do klas zgodnie z regułami klasyfikacji pojęciowej. Pojęcia reprezentujące klasy dla zbioru danych I oraz II zawarto w tab. 1, są to pojęcia rozłączne, co oznacza, że opisują one wyłącznie obiekty zawarte w danej klasie, natomiast nie opisują obiektów z innych klas – obiekty te na rys. 2 zaznaczono grubszą linią w kolorze czarnym.

Tabela 1. Pojęcia reprezentujące klasy w sztucznych zbiorach danych

Nazwa zbioru danych	Klasy	Pojęcie (obiekt) reprezentujące klasę	
		zmienna v_1	zmienna v_2
Zbiór danych I	Klasa 1	[–1,09, 2,23]	[0,08, 2,94]
	Klasa 2	[–2,78, 3,28]	[–3,08, 2,35]
Zbiór danych II	Klasa 1	[–0,65, 0,44]	[4,29, 9,10]
	Klasa 2	[–3,61, 2,17]	[–2,58, 3,20]
	Klasa 3	[0,24, 4,82]	[12,15, 16,41]

Źródło: opracowanie własne.



Rys. 1. Sztuczne zbiory danych symbolicznych interwałowych – wyniki klasyfikacji
Źródło: opracowanie własne.

Dodatkowo w badaniu wykorzystano także jeden zbiór danych rzeczywistych interwałowych, który opisuje 28 modeli samochodów osobowych (obiektów symbolicznych drugiego rzędu) opisywanych przez dziesięć zmiennych symbolicznych interwałowych. Dane te opisują samochody osobowe z trzech różnych segmentów – A, B, C oraz D (por. [Pełka 2012a]).

Zbiór ten podzielono na 20 podzbiorów, z których każdy zawiera 19 obiektów. W wyniku zastosowania adaptacji metody *bagging* zgodnie ze zmodyfikowaną propozycją Leischa otrzymano dwie klasy. Pierwsza z nich reprezentuje samochody osobowe z segmentów A i B, są samochody najmniejsze, służące głównie do jazdy w mieście (np. Fiat 500), oraz samochody nieco większe, niezaliczane do segmentu C (np. Fiat Punto). Druga klasa zawiera samochody z segmentów C oraz D, są to samochody „klasy niższej średniej", zapewniające względny komfort jazdy dla czterech osób i przeciętnie duży bagażnik (np. Opel Astra) oraz samochody klasy „średniej", czyli względnie duże i wygodne auta rodzinne (np. Mazda 6). Podobne wyniki otrzymano, stosując klasyfikację pojęciową (algorytm klasyfikacji hierarchicznej P. Brity) do całego zbioru danych. Dodatkowo oceniono stabilność otrzymanej klasyfikacji z zastosowaniem skorygowanego indeksu Randa – otrzymano wartość 0,675487 co świadczy o relatywnie stabilnym podziale 28 obiektów na dwie klasy.

5. Podsumowanie

Adaptacja metody *bagging*, która jest modyfikacją propozycji Leischa z wykorzystaniem klasyfikacji pojęciowej, może z powodzeniem znaleźć zastosowanie w klasyfikacji danych symbolicznych dowolnego typu (dzięki zastosowaniu algorytmu hierarchicznego P. Brity).

Zaprezentowane podejście, podobnie jak podejście oparte na macierzy współwystąpień, pozwala na otrzymanie bardzo dobrych wyników (w sensie skorygowanego indeksu Randa). Przewagą metody *bagging* jest fakt otrzymywania wyników w postaci pojęć oraz zapewnienie lepszego zróżnicowania modeli bazowych dzięki adaptacji metody *bagging*.

Niewątpliwą zaletą proponowanego podejścia jest fakt, że w wyniku klasyfikacji, oprócz etykiet klas, otrzymujemy także opis klas w postaci pojęć (syntetycznych obiektów symbolicznych), co znacznie ułatwia zarówno interpretację klas, jak i ich opis.

Wadą proponowanego rozwiązania jest z pewnością złożoność i czasochłonność samego algorytmu. Kolejnym problemem jest niewielka liczba metod klasyfikacji pojęciowej, która może być zastosowana do danych symbolicznych dowolnego typu. Niewątpliwie istotnym ograniczeniem dla szerszego zastosowania proponowanego podejścia jest brak oprogramowania, które pozwalałoby na prowadzenie badań.

Celem dalszych prac będzie porównanie propozycji Leischa z innymi rozwiązaniami w zakresie metody *bagging* oraz porównanie tych metod z innymi metodami klasyfikacji wielomodelowej.

Literatura

Bock H.-H., Diday E. (red.), 2000, *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin-Heidelberg.

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

De Carvalho F.A.T., Lechevallier Y., de Melo F.M., 2012, *Partitioning hard clustering algorithms based on multiple dissimilarity matrices*, Pattern Recognition, 45(1), s. 447-464.

Diday E., Brito P., 1989, *Symbolic Cluster Analysis*, [w:] O. Opitz (red.), *Conceptual and Numerical Analysis of Data*, Springer-Verlag, Berlin-Heidelberg, s. 45-84.

Diday E., Noirhomme-Fraiture M., 2008, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester.

Dudek A., 2004, *Tworzenie obiektów symbolicznych z baz danych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1021, s. 107-114.

Dudek A., 2013, *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wyd. UE we Wrocławiu, Wrocław.

Dudoit S., Fridlyand J., 2003, *Bagging to improve the accuracy of a clustering procedure*, Bioinformatics, vol. 19, no. 9, s. 1090-1099.

Fred A.L.N., Jain A.K., 2005, *Combining multiple clustering using evidence accumulation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, s. 835-850.

Gatnar E., 1998, *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E., 2008, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Ghaemi R., Sulaiman N., Ibrahim H., Mustapha N., 2009, *A Survey: Clustering Ensemble Techniques*, [w:] *Proceedings of World Academy of Science*, Engineering and Technology, vol. 38, s. 636-645.

Hornik K., 2005, *A CLUE for CLUster Ensembles*, „Journal of Statistical Software", vol. 14, s. 65-72.

Ichino M., 1988, *General metrics for mixed features – the Cartesian space theory for pattern recognition,* [w:] *Proceedings of the 1988 IEEE International Conference on Systems*, Man and Cybernetics, vol. 1, International Academic Publishers, Beijing, s. 494-497.

Kuncheva L.I., 2004, *Combining Pattern Classifiers. Methods and Algorithms*, Wiley, New Jersey.

Leisch F., 1999, *Bagged clustering*, Adaptive Information Systems and Modeling in Economics and Management Science, Working Papers, SFB, 51.

Noirhomme-Fraiture M., Brito P., 2011, *Far beyond the classical data models: symbolic data analysis*, Statistical Analysis and Data Mining, vol. 4, issue 2, s. 157-170.

Pełka M., 2012a, *Ensemble approach for clustering of interval-valued symbolic data*, Statistics in Transition, vol. 13, no. 2, s. 335-342.

Pełka M., 2012b, *Skalowanie wielowymiarowe i klasyfikacja danych symbolicznych w ocenie pozycji produktów na rynku*, Marketing i Rynek nr 3/2012, s. 21-26.

Pełka M., 2014, *Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 327, s. 202-209.

Walesiak M., Dudek A., 2014, *The `clusterSim` package*, http://www.r-project.org.

THE ADAPTATION OF BAGGING WITH THE APPLICATION OF CONCEPTUAL CLUSTERING OF SYMBOLIC DATA

Summary: Ensemble learning can be succesfully applied in discrimination and regression tasks [Gatnar 2008]. However, the idea of combining results obtained from different models can be applied in clustering [Fred, Jain 2005]. Unlike classical data, symbolic objects can be described by interval-valued variables, multinominal variables, histogram variables and multinominal variables with weights. Symbolic variables can also present dependencies [Bock, Diday 2000, pp. 2-3]. The main aim of the paper is to present an application of bagging algorithm for clustering, according to proposal made by Leisch [1999]. Conceptual clustering for symbolic data will be used as the base model. The resulting clusters are described by concepts. In the empirical part of the article results obtained with the application of artificial and real data sets are presented.

Keywords: ensemble clustering, symbolic data, conceptual clustering.