PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

**Małgorzata Gliwa**

Akademia Ekonomiczna w Katowicach

# METODA PIRAMID W KLASYFIKACJI OBIEKTÓW SYMBOLICZNYCH

## 1. Wstęp

W praktycznych zastosowaniach metod klasyfikacji szczególne znaczenie mają hierarchiczne metody aglomeracyjne, w których początkowo każdy obiekt traktowany jest jako klasa jednoelementowa. Następnie klasy te łączy się aż do chwili, gdy wszystkie obiekty znajdą się w jednym zbiorze. Hierarchiczne metody aglomeracyjne generują zbiór klas rozłącznych, a ich graficzną prezentacją jest dendrogram. Zdarza się jednak, że istnieje konieczność przydzielenia tego samego obiektu do więcej niż jednej klasy (np. klasyfikacja słów ze względu na ich znaczenie, klasyfikacja kart bankowych ze względu na ich funkcjonalność).

W taksonomii symbolicznej istnieje grupa metod tworzących skupienia nierozłączne. Wśród tych najczęściej stosuje się metodę piramid. Nazwa metody pochodzi od jej graficznej prezentacji, czyli tzw. piramidy.

Celem artykułu jest przedstawienie własności metody piramid oraz porównanie ich z własnościami hierarchicznych metod aglomeracyjnych. W pracy przedstawiony również zostanie przykład zastosowania algorytmu metody do klasyfikacji obiektów symbolicznych z rzeczywistego zbioru danych.

## 2. Charakterystyka metody piramid

Metoda piramid jest rozszerzeniem hierarchicznych metod aglomeracyjnych [Brito, Diday 1990]. Proces grupowania odbywa się bezpośrednio dla zbioru obiektów symbolicznych. Początkowo każdy obiekt stanowi klasę jednoelementową, a piramidę buduje się przez stopniowe łączenie obiektów w klasy posiadające wspólne elementy. Metoda piramid odwołuje się do macierzy niepodobieństw Robinsona i teorii krat [Brito, Diday 1990].

Tabela 1. Porównanie własności metody piramid z własnościami hierarchicznych metod aglomeracyjnych

| Metoda piramid | Metody aglomeracyjne |
|---|---|
| Tworzy nierozłączne skupienia | Tworzą rozłączne skupienia |
| Kryterium łączenia klas (kryteria „symboliczne”):<br>$d \to \min$ (zob. formuła 4)<br>minimalny stopień ogólności lub minimalny wzrost stopnia ogólności | Metody grupowania zależą od sposobu określenia odległości między klasami, np.:<br>najmniejsza odległość spośród wszystkich odległości między obiektami należącymi do łączonych klas (metoda najbliższego sąsiedztwa)<br>największa odległość spośród wszystkich odległości między obiektami należącymi do łączonych klas (metoda najdalszego sąsiedztwa) |
| Każde skupienie ma najwyżej dwóch poprzedników | Każde skupienie ma najwyżej jednego poprzednika |
| Klasa $B \in P$ nazywana jest poprzednikiem klasy $A \in P$, jeżeli $A \subset B$, oraz nie istnieje taka klasa $C \in P$, że $A \subset C \subset B$ | |
| Połączone mogą być takie dwie klasy, które nie były wcześniej dwukrotnie łączone | Połączone mogą być takie dwie klasy, które nie były wcześniej łączone |
| Zawsze pozwala odkryć pojęcia (charakterystyki klas), ponieważ proces grupowania odbywa się na podstawie zbioru obiektów | Nie zawsze pozwalają w łatwy sposób odkryć pojęcia, ponieważ proces grupowania odbywa się na podstawie macierzy odległości |
| Graficzną prezentacją jest piramida<br>Piramida to skończona rodzina<br>$P = \{A, B, ...\}$ niepustych podzbiorów<br>$A, B, ... \subseteq E = \{s_1, ..., s_n\}$, takich że:<br>$E \in P$,<br>$\forall_{s_i} \{s_i\} \in P, \ i = 1, ..., n$,<br>$\forall_{A,B \in P} A \cap B = \phi \ \vee \ A \cap B \in P$,<br>w zbiorze $E$ istnieje porządek liniowy $\leq$ taki, że $A \in P$ jest przedziałem ze względu na $\leq$:<br>$A = [\alpha, \beta] = \{s_i \mid s_i \in E, \ \alpha \leq s_i \leq \beta\}$. | Graficzną prezentacją jest dendrogram<br>Dendrogram to skończona rodzina<br>$P = \{A, B, ...\}$ niepustych podzbiorów<br>$A, B, ... \subseteq E = \{s_1, ..., s_n\}$, takich że:<br>$E \in P$,<br>$\forall_{s_i} \{s_i\} \in P, \ i = 1, ..., n$,<br>$\forall_{A,B \in P} A \cap B = \phi \ \vee \ A \subseteq B \ \vee \ B \subseteq A$. |

Źródło: opracowanie własne.

W metodzie piramid definiuje się pojęcia: zupełnego obiektu symbolicznego, zakresu obiektu symbolicznego oraz stopnia ogólności obiektu symbolicznego [Diday, Bertrand 1986].

*Definicja 1.* Zakres obiektu symbolicznego $s = [y_1 \in V_1] \wedge [y_2 \in V_2] \wedge ... \wedge [y_p \in V_p]$ to zbiór wszystkich obiektów reprezentowanych przez dany obiekt ($|s|_E = \{\omega \in E : y_j(\omega) \in V_j, j = 1,...p\}$, $E = \{s_1,...,s_n\}$).

*Definicja 2.* Obiekt symboliczny $s$ nazywany jest zupełnym, jeżeli jego opis (charakterystyka) jest taki sam jak jego zakres.

*Przykład*. Rozważmy zbiór obiektów symbolicznych $E = \{s_1, s_2, s_3\}$ scharakteryzowanych przez zmienne płeć, wiek oraz waga:
$s_1 = K \wedge [20,30] \wedge [45,55]$, $s_2 = M \wedge [30,45] \wedge [60,80]$, $s_3 = K \wedge [30,40] \wedge [50,60]$, oraz obiekt symboliczny $s = K \wedge [20,40] \wedge [40,60]$.

Wówczas $|s|_E = \{s_1, s_3\}$, a obiekt $s$ jest zupełny.

*Definicja 3.* Stopień ogólności obiektu symbolicznego $s$ to funkcja:

$$G : E \to [0,1],$$

określona:

a) dla obiektów opisanych przez zmienne w postaci listy kategorii lub przedziałów w następujący sposób [Brito, Diday 1990; Brito 1994]:

$$G(s) = \prod_{j=1}^{p} \frac{\overline{\overline{V_j}}}{\overline{\overline{O_j}}}, \tag{1}$$

gdzie: $V_j \subseteq O_j$, $j = 1,...,p$, $V_j$ oznacza zbiór wartości $j$-tej zmiennej opisującej dany obiekt $s$, $O_j$ oznacza dziedzinę $j$-tej zmiennej, $\overline{\overline{V_j}}$, $\overline{\overline{O_j}}$ to odpowiednio:

- długość przedziału, jeżeli obiekty opisane były przez zmienne wyrażone w postaci przedziałów [Brito 2000],
- liczba kategorii, jeżeli obiekty opisane były przez zmienne w postaci listy kategorii [Brito 2000];

b) dla obiektów opisanych przez zmienne z wagami $(p_1, p_2,..., p_k)$ wyróżnia się [Brito 2002]:

- uogólnienie obiektu symbolicznego s wyrażone przez maksimum:

$$G(s) = \prod_{j=1}^{p} \frac{1}{\sqrt{k_j}} \sum_{i=1}^{k_j} \sqrt{p_{ij}}, \tag{2}$$

- uogólnienie obiektu symbolicznego s wyrażone przez minimum:

$$G(s)=\prod_{j=1}^{p}\frac{1}{\sqrt{k_j\left(k_j-1\right)}}\sum_{i=1}^{k_j}\sqrt{1-p_{ij}}\ . \tag{3}$$

W dalszej części artykułu przedstawiony zostanie algorytm klasyfikacji obiektów symbolicznych ze zbioru $E=\{s_1,...,s_n\}$ z wykorzystaniem metody piramid, który składa się z następujących etapów [Brito, Diday 1990]:

1) utwórz $n$ rozłącznych jednoelementowych klas $C_1,...,C_n$;

2) dla $i\neq j$ połącz klasy $C_i, C_j$. Sprawdź, czy ich suma $C_i\cup C_j$ zawiera zakres pewnego zupełnego obiektu symbolicznego $s$ oraz czy:

$$d=\overline{\overline{C_i\cup C_j-|s|_E}} \tag{4}$$

jest najmniejsze:

a) jeśli $d=0$, to utwórz klasę $C=C_i\cup C_j$ oraz znajdź reprezentujący ją obiekt symboliczny $s$,

b) w przeciwnym przypadku utwórz klasę zawierającą zakres takiego zupełnego obiektu symbolicznego $s$, dla którego $d$ było najmniejsze;

3) jeśli wszystkie obiekty stanowią jedną klasę lub dla każdego zupełnego obiektu symbolicznego $s$ istnieje $A\in C$ takie, że $|s|_E\subset A$, to zakończ algorytm.

Brito i Diday [1990] uważają, że w wyniku zastosowania powyższego algorytmu zbiór klas $C_1,...,C_n$ może nie zostać wybrany jednoznacznie. Oznacza to, że w etapie (2) algorytmu należy wprowadzić dodatkowe kryterium oparte na pojęciu stopnia ogólności obiektu sformułowanym w definicji 3. Z powyższego wynika, że w klasy łączy się te obiekty, dla których jednocześnie wartość $d$ oraz stopień ogólności $G(s)$ osiągają minimum. Wartość $G(s)$ wyznacza wysokość piramid [Brito, Diday 1990; Brito 1994].

W wyniku zastosowania algorytmu metody piramid dla zbioru obiektów symbolicznych każda klasa reprezentowana jest przez zupełny obiekt symboliczny [Diday, Bertrand 1986], co pozwala na uzyskanie w łatwy sposób charakterystyk klas. Należy jednak zauważyć, że w rezultacie można również otrzymać taki zbiór $C$, że $E\notin C$, co w konsekwencji spowoduje, że warunek 1 definicji piramidy nie jest spełniony (zob. tab. 1). Taka piramida nazywana jest wówczas niekompletną [Brito, Diday 1990].

## 3. Porównanie metody piramid z hierarchicznymi metodami aglomeracyjnymi

Metoda piramid, jak już wspomniano wcześniej, jest uogólnieniem metod hierarchicznych. W poniższym punkcie zostaną porównane własności metody piramid z własnościami hierarchicznych metod aglomeracyjnych.

Metoda piramid wykorzystuje zalety hierarchicznych metod aglomeracyjnych, do których m.in. zaliczyć można to, że działają według jednej procedury aglomeracyjnej, a wyniki klasyfikacji można przedstawić graficznie w formie dendrogramu wskazującego na kolejność połączeń między klasami. Różnice w procedurach poszczególnych metod aglomeracyjnych wynikają z odmienności definiowania odległości międzyklasowej [Gatnar, Walesiak 2004].

## 4. Przykład

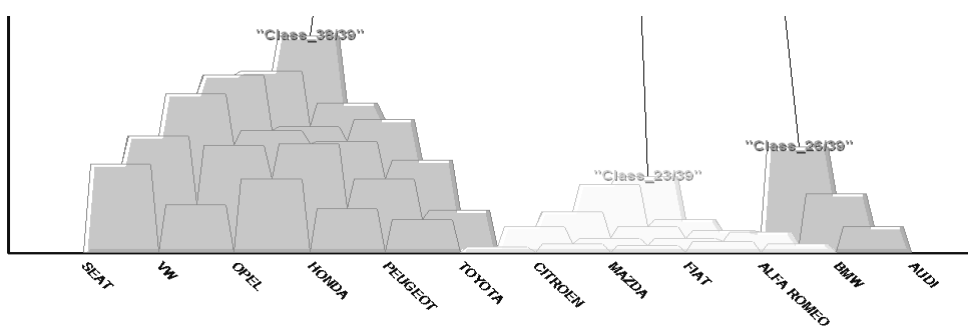
Celem empirycznej części artykułu jest klasyfikacja obiektów symbolicznych metodą piramid oraz porównanie jej z klasyfikacją obiektów przy wykorzystaniu hierarchicznych metod aglomeracyjnych.

W przykładzie wykorzystano dane producentów samochodów klasy średniej zamieszczone na stronie internetowej www.motofakty.pl. Na ich podstawie utworzono 12 obiektów symbolicznych (marki samochodów), które zostały scharakteryzowane przez 5 zmiennych przedziałowych:

- cenę samochodu (zł)[1],
- pojemność silnika ($cm^3$),
- moc silnika (KM),
- przyspieszenie do 100 km/h ( $s$ ),
- spalanie paliwa w cyklu miejskim ( $l$ ).

Do obliczeń wykorzystano moduł HIPYR programu SODAS ver. 2,5, moduł DI (*distances martix*) programu SODAS ver. 1,2 oraz pakiet `cluster` programu R.

Po przeprowadzeniu grupowania obiektów symbolicznych metodą piramid otrzymano klasy obiektów przedstawione na rys. 1.




Rys. 1. Piramida otrzymana dla zbioru obiektów symbolicznych
Źródło: opracowanie własne za pomocą programu SODAS ver. 2,5.

Piramida z rys. 1 przedstawia kolejne połączenia poszczególnych klas obiektów symbolicznych. Liczba klas, którą przyjęto za ostateczną, została ustalona na

[1] Ceny samochodów pochodzą z okresu od 5.02.2008 r. do 22.04. 2008 r.

podstawie analizy raportu otrzymanego z programu SODAS ver. 2,5. Sugerował on podział zbioru obiektów na 3 klasy:

```
THE MOST IMPORTANT CLASSES:
THE CLASS - "C_23/39„, THE CLASS -"C_26/39„,THE CLASS -
"C_38/39".
```

W klasie 1 znajdują się następujące obiekty symboliczne: {Seat, VW, Opel, Honda, Peugeot, Toyota, Citroen}. Klasa ta jest reprezentowana przez następujący zupełny obiekt symboliczny:

```
[cena = [43700, 115250]] ^ [pojemnosc = [1248, 2231]] ^
[moc = [75, 240]] ^ [przysp/100 = [6.7, 17]] ^ [pali-
wo_miasto = [4.9, 9.5]].
```

Klasa 2 zawiera następujące obiekty symboliczne: {Toyota, Citroen, Mazda, Fiat, Alfa Romeo, BMW} i jest reprezentowana przez następujący zupełny obiekt symboliczny:

```
[cena = [115000, 119000]] ^ [pojemnosc = [1349, 2231]]
^ [moc = [84,180]] ^ [przysp/100 = [8.1, 13]] ^ [pali-
wo_miasto = [7.6, 12.1]].
```

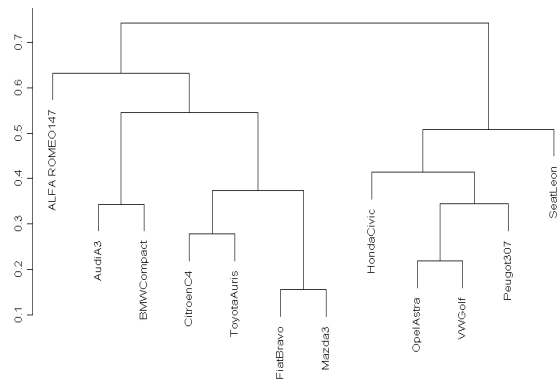
W klasie 3 znalazły się obiekty: {Fiat, Alfa Romeo, BMW, Audi}. Klasę tę charakteryzuje zupełny obiekt symboliczny zapisany w poniższej formie:

```
[cena = [119000, 130435]] ^ [pojemnosc = [1360, 1995]]
^ [moc = [90,200]] ^ [przysp/100 = [6.6, 12.4]] ^ [pa-
liwo_miasto = [6.5, 12.1]].
```

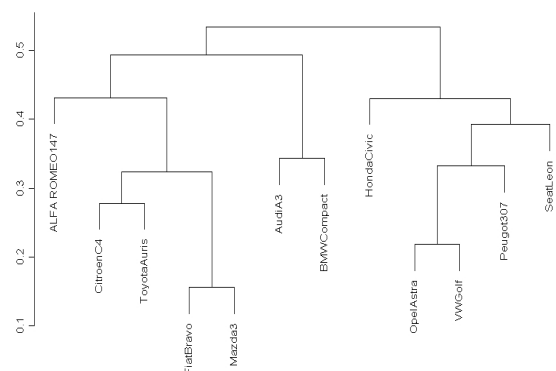
W wypadku analizowanych samochodów można stwierdzić, że w klasie 1 znajdują się samochody najtańsze i najbardziej ekonomiczne pod względem spalania paliwa w cyklu miejskim. Klasa 2 to samochody ze „średniej" półki cenowej. Jednak pod względem podstawowych parametrów charakteryzujących osiągi silnika, czyli mocy i przyspieszenia, samochody z klasy 2 osiągają gorsze wyniki w porównaniu z samochodami z pozostałych klas. Klasa 3 to samochody najdroższe, a zarazem o dość niskiej pojemności silnika w porównaniu z samochodami z pozostałych klas.

Poszczególne klasy zawierają obiekty wspólne, co w tym przypadku wydaje się naturalne, bowiem podjęcie decyzji o zakupie samochodu to proces poprzedzony analizą wielu parametrów. Jedne z nich mogą być wspólne, a inne różne dla poszczególnych marek samochodów.

Następnie za pomocą modułu DI programu SODAS ver. 1,2 wygenerowano macierz odległości. Do wyznaczenia odległości zastosowano znormalizowaną odległość Ichino-Yaguchiego [Ichino, Yaguchi 1994]. Jest to miara odległości stosowana dla obiektów symbolicznych. Na podstawie powstałej macierzy przeprowadzono grupowanie obiektów symbolicznych za pomocą dwóch wybranych metod aglomeracyjnych: metody najdalszego sąsiedztwa oraz metody średniej klasowej. Otrzymano dendrogramy przedstawione na rys. 2 i 3.




Rys. 2. Dendrogram otrzymany dla metody
Źródło: opracowanie własne z wykorzystaniem programu **R**.




Rys. 3. Dendrogram otrzymany dla metody najdalszego sąsiedztwa średniej klasowej
Źródło: opracowanie własne z wykorzystaniem programu **R**.

Tabela 2. Klasy obiektów symbolicznych otrzymane w wyniku zastosowania metody najdalszego sąsiedztwa i średniej klasowej

| | | |
|---|---|---|
| Klasa 1 | Seat, Peugeot, VW, Opel, Honda | Seat, Peugeot, VW, Opel, Honda |
| Klasa 2 | Audi, BMW, Citroen, Toyota, Fiat, Mazda | Audi, BMW |
| Klasa 3 | Alfa Romeo | Alfa Romeo, Citroen, Toyota, Fiat, Mazda |

Źródło: opracowanie własne na podstawie obliczeń wykonanych w programie **R**.

Analizując tab. 2, przekonujemy się, że otrzymane struktury klas, zwłaszcza klas 2 i 3, są inne niż te otrzymane w wyniku zastosowania metody piramid.

## 5. Podsumowanie

Metoda piramid realizuje proces grupowania obiektów w sposób podobny do metod aglomeracyjnych. Jednakże w trakcie tego grupowania nie trzeba wyznaczać

kolejnych macierzy podobieństw. Metoda piramid wykorzystuje „symboliczne" kryteria łączenia klas, które wyznacza się bezpośrednio dla zbioru obiektów symbolicznych. Metoda piramid tworzy zbiór klas nierozłącznych. Można ją przyrównać do procesu kategoryzacji występującego u ludzi w czasie myślenia i rozpoznawania [Gatnar 1998]. Każda powstała klasa jest nowym obiektem symbolicznym, a zatem można ją łatwo zinterpretować.

Niestety, dostępne oprogramowanie metod analizy danych symbolicznych stanowi problem dla stosowania metody piramid dla obiektów symbolicznych opisanych przez zmienne strukturalne.

## Literatura

Brito P. (1994), *Use of pyramids in symbolic data analysis*, [w:] E. Diday (red.), *New approaches in classification and data analysis*, Springer-Verlag, Berlin-Heidelberg, s. 378-386.

Brito P. (2000), *Hierarchical and pyramidal clustering with complete symbolic objects*, [w:] H.-H. Bock, E. Diday (red.), *Analysis of symbolic data*, Springer Verlag, Berlin-Heidelberg, s. 312-324.

Brito P. (2002), *Hierarchical and pyramidal clustering for symbolic data*, „Journal of the Japanese Society of Computational Statistics", vol. 15 (2), s. 231-244.

Brito P., Diday E. (1990), *Pyramidal representation of symbolic objects*, [w:] M. Schader, W. Gaul (red.), *Knowledge, data and computer-assisted decisions*, NATO ASI Series, Springer Verlag, Berlin-Heidelberg-New York, s. 3-16.

Diday E., Bertrand P. (1986), *An extension of hierachical clustering: the pyramidal presentation*, [w:] E.S. Geslema, L.N. Kanal (red.), *Pattern recognition in practice II*, Elsevier Science, Amsterdam.

Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa s. 122-131.

Gatnar E., Walesiak M. (2004) (red.), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław, s. 316-350.

Ichino M., Yaguchi H. (1994), *Generalized Minkowski metrics for mixed feature type data analysis*, „IEEE Transactions on Systems, Man and Cybernetics" vol. 24 (4), s. 698-708.

## PYRAMIDAL CLUSTERING IN THE CLASSIFICATION OF SYMBOLIC OBJECTS

### Summary

The aim of this article is a presentation of properties of the pyramidal clustering which is used to classify symbolic objects. The properties of the pyramidal clustering were compared to the properties of hierarchical agglomerative clustering. The example of using this method to classify the symbolic objects from real dataset was also presented. Calculation was made with the use of SODAS and R programmes.