

Małgorzata Gliwa

Gimnazjum nr 20 w Katowicach

ZASTOSOWANIE ANALIZY SKUPIEŃ DO KLASYFIKACJI OBIEKTÓW SYMBOLICZNYCH

1. Wstęp

Symboliczna analiza danych (*symbolic data analysis*) uwzględnia głównie zmienne jakościowe zgromadzone w symbolicznej tabeli danych. Na podstawie tych danych tworzone są obiekty symboliczne, które następnie podlegają klasyfikacji. Metody analizy symbolicznej zaliczane są do metod nieklasycznych, a uzyskane wyniki są dość łatwe w interpretacji, ponieważ do opisu klas obiektów symbolicznych wykorzystuje się pojęcia naturalne.

W artykule zostanie przedstawiona analiza skupień (*divisive clustering*) dla symbolicznej macierzy danych oraz przykład jej zastosowania do klasyfikacji obiektów symbolicznych opracowany za pomocą programu Sodas. Omawiana metoda to hierarchiczna metoda podziału. Polega na tym, aby z rozpatrywanego zbioru obiektów wydzielić możliwie jednorodne podzbiory (skupienia), zapewniając jednocześnie jak największe zróżnicowanie tych podzbiorów między sobą. Proces podziału zbioru obiektów zatrzymywany jest po określonej przez użytkownika liczbie skupień, a jego rezultatem jest drzewo, w którym każdy węzeł reprezentuje skupienie.

2. Pomiar odległości między obiektami symbolicznymi

Rozważmy zbiór obiektów symbolicznych $\Omega = \{1, \dots, n\}$. Załóżmy, że obiekt symboliczny $k \in \Omega$ opisany jest przez następujący wektor zmiennych $Y(k) = (Y_1(k), \dots, Y_p(k))^T$. Zmienne te mogą mieć różne dziedziny, tj. zmienne,

których dziedziną jest zbiór liczb rzeczywistych, lub zmienne, których dziedziną jest skończony zbiór kategorii. Program Sodas wymusza założenie, aby do analizy nie uwzględniać zmiennych o różnych typach dziedzin. Nawiązując do powyższego założenia, należy stwierdzić, że miary odległości definiowane są w inny sposób dla macierzy zawierającej dane ilościowe i w inny dla macierzy zawierającej dane jakościowe.

Dla macierzy zawierającej dane ilościowe w pierwszym etapie liczone są odległości δ pomiędzy odpowiednimi wartościami cech obiektów symbolicznych k i l ze zbioru Ω . Jeśli przez $\xi_{kj} = Y_j(k)$ oznaczona zostanie wartość cechy j obiektu symbolicznego k , a przez $\xi_{lj} = Y_j(l)$, wartość cechy j obiektu symbolicznego l , to funkcja odległości δ_j wyraża się następującym wzorem:

$$\delta_j(\xi_{kj}, \xi_{lj}) = \max\{|\alpha_k^j - \alpha_l^j|, |\beta_k^j - \beta_l^j|\}. \quad (1)$$

Wzór (1) uwzględnia przypadek zmiennej przedziałowej, gdzie $\alpha_k^j, \alpha_l^j, \beta_k^j, \beta_l^j$ są odpowiednio początkiem i końcem przedziałów ξ_{kj} i ξ_{lj} . W sytuacji zmiennych skokowych, wzór (1) upraszcza się do postaci:

$$\delta_j(\xi_{kj}, \xi_{lj}) = |\xi_{kj} - \xi_{lj}|. \quad (2)$$

W kolejnym kroku obliczana jest miara odległości $d: \Omega \times \Omega \rightarrow R^+$ dla wszystkich obiektów symbolicznych ze zbioru Ω :

$$d(k, l) = \left(\sum_{j=1}^p \delta_j(\xi_{kj}, \xi_{lj})^2 \right)^{\frac{1}{2}}. \quad (3)$$

Miara wyrażona wzorem (3) może być znormalizowana w następujący sposób:

$$d'(k, l) = \left(\sum_{j=1}^p \left(\frac{1}{m_j} \delta_j(\xi_{kj}, \xi_{lj}) \right)^2 \right)^{\frac{1}{2}}, \quad (4)$$

gdzie wskaźnik m_j zdefiniowany może być, jako:

$$1) m_j^2 = \frac{1}{2n^2} \sum_{k=1}^n \sum_{l=1}^n \delta_j(\xi_{kj}, \xi_{lj})^2, \quad (5)$$

$$2) m_j - \text{różnica między największą a najmniejszą wartością z dziedziny } Y_j. \quad (6)$$

Oba te sposoby zostały zaproponowane przez M. Chavent [1997].

Z kolei w przypadku w macierzy zawierającej dane jakościowe przed obliczeniem miary odległości, należy zbudować tabelę częstości występowania odpowiednich wartości cech każdego obiektu symbolicznego ze zbioru Ω .

Tabela 1. Tabela częstości występowania wartości cech obiektu symbolicznego

	1	...	j	...	t	Σ
1	f_{11}		f_{1j}		f_{1t}	p
$\vdots$						
k	f_{k1}		f_{kj}		f_{kt}	p
$\vdots$						
n	f_{n1}		f_{nj}		f_{nt}	p
Σ	$f_{\cdot 1}$		$f_{\cdot j}$		$f_{\cdot t}$	$n \cdot p$

Źródło: opracowanie własne.

W tabeli 1 t oznacza liczbę kategorii wszystkich zmiennych, n liczbę obiektów symbolicznych, a p liczbę zmiennych.

Funkcja odległości d dla obiektów symbolicznych k i l ze zbioru Ω obliczana jest według wzoru:

$$d(k, l) = \left(\sum_{j=1}^t \frac{1}{p_{\cdot j}} \left(\frac{p_{kj}}{p_{k \cdot}} - \frac{p_{lj}}{p_{l \cdot}} \right)^2 \right)^{1/2}, \quad (7)$$

gdzie: t – liczba kategorii wszystkich zmiennych, $k = 1, \dots, n$, $l = 1, \dots, n$,

$$p_{kj} \stackrel{\text{def}}{=} \frac{f_{kj}}{np}, \quad p_{k \cdot} \stackrel{\text{def}}{=} \sum_{j=1}^t p_{kj}, \quad p_{\cdot j} \stackrel{\text{def}}{=} \sum_{k=1}^n p_{kj}.$$

3. Podział zbioru obiektów na skupienia

3.1. Kryterium jakości podziału

Po obliczeniu miar odległości między obiektami symbolicznymi ze zbioru Ω , budowana jest macierz odległości $D = (d_{kl})_{k, l \in \Omega}$. Na jej podstawie w każdym kolejnym etapie obliczana jest miara jakości podziału skupienia C_i .

Definicja 1. Niech $D = (d_{kl})_{k, l \in \Omega}$ będzie macierzą odległości obiektów symbolicznych ze zbioru Ω . Miara jakości podziału skupienia C_i wyraża się następującym wzorem:

$$Q(C_i) = \frac{1}{2 \cdot n_i} \sum_{k \in C_i} \sum_{l \in C_i} d^2(k, l), \quad (8)$$

gdzie: n_i oznacza liczbę obiektów należących do skupienia C_i .

W trakcie podziału zbioru Ω , po $m-1$ krokach otrzymujemy m skupień $C_1, \dots, C_m$. Miara jakości podziału dla tych skupień definiowana jest w następujący sposób.

Definicja 2. Niech $Z_m = (C_1, \dots, C_m)$ będzie zbiorem m skupień. Miara jakości podziału skupień wyraża się wzorem:

$$W(Z_m) = \sum_{i=1}^m Q(C_i). \quad (9)$$

3.2. Wybór skupienia do podziału

W każdym kolejnym etapie wybrane skupienie C_i dzielone jest na dwa skupienia C_{i1} i C_{i2} . Jeśli założymy, że $Z_m = (C_1, \dots, C_m)$ jest zbiorem m skupień, to powstaje pytanie, które skupienie wybrać, aby po m krokach otrzymać możliwie najmniejszą wartość funkcji $W(Z_{m+1})$.

$$W(Z_{m+1}) = W(Z_m) - [Q(C_i) - Q(C_{i1}) - Q(C_{i2})] \rightarrow \min. \quad (10)$$

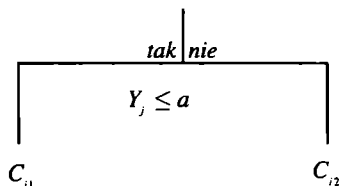
Minimalizacja funkcji oznaczonej wzorem (10) jest równoznaczna z wyborem takiego skupienia, aby różnica miar jakości podziału dla skupień C_i , C_{i1} oraz C_{i2} była maksymalna.

$$Q(C_i) - Q(C_{i1}) - Q(C_{i2}) \rightarrow \max. \quad (11)$$

W związku z tym, rozważone zostają wszystkie możliwe podziały i porównane wartości funkcji (11).

3.3. Podział skupienia

Podział wybranego skupienia C_i na C_{i1} i C_{i2} odbywa się zgodnie z funkcją logiczną „prawda/fałsz” i odpowiedzią „tak” lub „nie” na pytanie binarne: *Czy wartość cechy j obiektu symbolicznego jest mniejsza lub równa od wartości a .* Jeśli „tak”, to funkcja logiczna przyjmuje wartość „prawda” i obiekt symboliczny przydzielony zostaje do skupienia C_{i1} , jeśli „nie”, to funkcja przyjmuje wartość „fałsz”, a obiekt przydzielony zostaje do skupienia C_{i2} .



Rys. 1. Schemat przydziału obiektu do skupienia
Źródło: opracowanie własne.

Tabela 2. Definicje funkcji logicznych

Typ zmiennej	Definicja funkcji	Parametr c_j
Skokowa	$q_a = \begin{cases} \text{prawda, dla } c_j \leq a \\ \text{fałsz, dla } c_j > a \end{cases}$	c_j – mediana wartości
Przedziałowa		$c_j = \frac{\alpha'_j + \beta'_j}{2}$
Modalna	$q_a = \begin{cases} \text{prawda, dla } \sum_{x \leq a} \pi_k(x) \geq 0.5 \\ \text{fałsz, dla } \sum_{x \leq a} \pi_k(x) < 0.5 \end{cases}$	
Nominalna		

Źródło: opracowanie własne na podstawie [Bock 2000].

3.4. Algorytm DIV (*divisive clustering*)

Algorytm przydziału obiektów symbolicznych do skupień jest realizowany według następujących kroków:

- KROK 1. Ustal liczbę skupień (K).
- KROK 2. Utwórz jedno skupienie (klasę) C zawierające wszystkie obiekty.
- KROK 3. Porównaj wartości odpowiedniej funkcji jakości dla wszystkich możliwych podziałów skupienia.
- KROK 4. Wybierz najlepszy podział i skupienie C podziel na dwa skupienia C_1 i C_2 .
- KROK 5. Porównaj wartości funkcji jakości dla wszystkich możliwych podziałów skupień C_1 i C_2 .
- KROK 6. Wybierz najlepszy podział i dziel wybrane skupienie na kolejne dwa.
- KROK 7. Kroki 5 i 6 powtarzaj rekurencyjnie $K - 2$ razy.
- KROK 8. Zakończ algorytm po $K - 1$ iteracjach.

4. Przykład zastosowania metody DIV

4.1. Charakterystyka zbioru danych

Badanie przeprowadzono za pomocą programu Sodas wersji 1.2. Jako danych wejściowych użyto zbioru obiektów symbolicznych. Został on wygenerowany z relacyjnej bazy danych zbudowanej w programie Access za pomocą odpowiednich zapytań języka SQL. W tym celu zastosowano moduł DB2SO programu Sodas. Plik zawiera 28 obiektów symbolicznych odpowiadających kartom kredytowym. Obiekty opisane są przez 10 zmiennych, w tym 4 zmienne jakościowe i 6 ilościowych; a mianowicie:

- *konto* (czy klient zobowiązany jest mieć konto w danym banku),
- *ubezpieczenie* (czy bank oferuje pakiet ubezpieczeń w ramach posiadanej karty),

- *usługi dodatkowe* (rodzaje usług dodatkowych oferowanych w ramach posiadanej karty, przedstawione jako lista kategorii),
- *grace period* (długość okresu odroczenia rozpoczęcia spłaty kredytu),
- *nazwa banku*,
- *limit kredytu*,
- *minimalny dochód*,
- *opłata za kartę główną*,
- *maksymalne oprocentowanie*,
- *prowizja od wypłat z bankomatów własnych*.

Zgodnie z odpowiednimi założeniami omówionymi w rozdziale 2, podział zbioru obiektów odbywał się osobno dla danych ilościowych i osobno dla danych jakościowych. Dla zmiennych ilościowych miarę odległości między obiektami symbolicznymi obliczono na podstawie wzoru (3). Liczba klas ustalona została przez użytkownika przed rozpoczęciem wykonania algorytmu. Plik wynikowy, który generuje program Sodas, zawiera obiekty symboliczne, czyli nazwy kart kredytowych, przydzielone do odpowiednich skupień oraz binarne drzewo podziału. Ponadto wskazany jest stopień (w %) prawidłowo zaklasyfikowanych obiektów do klas. Poniżej (tab. 3 i 4) przedstawiono interpretację kilku wybranych skupień odpowiednio dla zmiennych ilościowych i jakościowych, opracowaną na podstawie informacji z bazy danych utworzonej przez autora.

4.2. Wyniki eksperymentu

Zmienne ilościowe

```

DRZEWO PODZIAŁU
-----
- Ng <-> tak and Nd <-> nie

          +---- Classe 1 (Ng=9)
          |
          !----4- [limit <= 2500.000000]
          |
          |      +---- Classe 5 (Nd=6)
          |      |
          |      !----2- [limit <= 7500.000000]
          |      |
          |      |      +---- Classe 3 (Nd=8)
          |      |      |
          |      |      !----1- [limit <= 12500.000000]
          |      |      |
          |      |      |      +---- Classe 2 (Ng=3)
          |      |      |      |
          |      |      |      !----3- [limit <= 25000.000000]
          |      |      |      |
          |      |      |      +---- Classe 4 (Nd=2)

```

Rys. 2. Binarne drzewo klasyfikacyjne wygenerowane dla zmiennych ilościowych
Źródło: opracowanie własne za pomocą programu Sodas v. 1.2.

Tabela 3. Przydział obiektów symbolicznych do skupień

Numer skupienia	Nazwa obiektu symbolicznego
1 ($n = 9$)	AmericanExpres, VisaEconomic, VisaClasic_11, VisaClasic_12, VisaClasic_17, MasterCardO-net, VisaCyfra+, BłkitnaKKVisaElektronic, BłkitnaKKMasterCardElectronic
2 ($n = 3$)	MasterCardSilver_20, VisaSilver, KartazLwicaMasterCard
3 ($n = 8$)	VisaCredit, VisaClasic_1, MasterCardSilver_7, VisaSilver_8, VisaSilver_14, SrebnnaKKVisa, SrebnnaKKMasterCardElectronic, KKVisaElectron
4 ($n = 2$)	VisaOrkiestra, VisaLot
5 ($n = 6$)	VisaElectron, VisaRegata, IdeaProfit, VisaClasic_15, VisaNowa, BursztynowaKartaKredytowa
98.440140%	

Źródło: opracowanie własne za pomocą programu Sodas v. 1.2.

Skupienie 3 zawiera 8 kart kredytowych, których właściciele zobowiązani są posiadać dochód netto w granicach 1400-1650 zł, otrzymują limit wynoszący 10 000 zł, a długość okresu odraczającego rozpoczęcie spłaty kredytu wynosi od 52 do 54 dni.

Skupienie 4 zawiera dwie karty kredytowe, których właściciele są zobowiązani posiadać dochód netto 2000 zł; w obu przypadkach długość okresu odraczającego rozpoczęcie spłaty kredytu jest w porównaniu z innymi kartami krótsza i wynosi 51 dni. Jednocześnie posiadacze tych kart mogą korzystać z limitu kredytu wynoszącego 30 000 zł (to jedyne karty z takim limitem). Cena za wydanie karty głównej jest, w porównaniu z innymi kartami, wysoka i wynosi odpowiednio 80 i 95 zł. Karty te oferowane są przez Pekao SA.

Zmienne jakościowe

DRZEWO PODZIAŁU

- Ng <-> tak and Nd <-> nie

```

          +---- Classe 1 (Ng=2)
          !
          !----4- [uslugi <= pakiet_podrozny]
          !
          !      +---- Classe 5 (Nd=16)
          !
          !----2- [konto = Nie]
          !
          !      +---- Classe 3 (Nd=3)
          !
          !----1- [ubezpz = Tak]
          !
          !      +---- Classe 2 (Ng=3)
          !
          !----3- [uslugi <= zdjecie]
          !
          !      +---- Classe 4 (Nd=4)

```

Rys. 3. Binarne drzewo klasyfikacyjne wygenerowane dla zmiennych jakościowych

Źródło: opracowanie własne za pomocą programu Sodas v. 1.2.

Tabela 4. Przydział obiektów symbolicznych do skupień

Numer skupienia	Nazwa obiektu symbolicznego
1 ($n = 2$)	VisaClasic_15, VisaNowa
2 ($n = 3$)	MasterCardOnet, VisaCyfra+, BursztynowaKartaKredytowa
3 ($n = 3$)	AmericanExpres, VisaEconomic, VisaClasic_11
4 ($n = 4$)	VisaOrkiestra, VisaLot, VisaRegata, KKVisaElectron
5 ($n = 16$)	VisaCredit, VisaElectron, VisaClasic_1, MasterCardSilver_7, VisaSilver_8, VisaClasic_12, IdeaProfit, VisaSilver_14, VisaClasic_17, MasterCardSilver_20, VisaSilver, SrebrnaKKVisa, SrebrnaKKMasterCardElectronic, BłkitnaKKVisaElektronic, BłkitnaKKMasterCardElectronic, KartazLwicaMasterCard
63.480779 %	

Źródło: opracowanie własne za pomocą programu Sodas v. 1.2.

Skupienie 3 zawiera trzy karty kredytowe, których właściciele są zobowiązani do posiadania konta w danym banku, a w ramach posiadanej karty klient otrzymuje usługi dodatkowe (pakiet podróży, assistance, zdjęcie na karcie) oraz pakiet ubezpieczeń. Są to karty oferowane przez bank Millennium.

Skupienie 4 zawiera cztery karty kredytowe, których właściciele są nie zobowiązani do posiadania konta w danym banku (oprócz KKVisaElectron), ale w ramach karty klient nie otrzymuje ani żadnych usług dodatkowych, ani pakietu ubezpieczeń. Trzy karty oferowane są przez Pekao SA, a jedna przez ING BŚ.

5. Wnioski

Zaprezentowana w artykule analiza skupień pozwala uwzględniać dane symboliczne, z których tworzone są obiekty symboliczne. Zaletą tej metody jest możliwość interpretacji otrzymanych wyników za pomocą naturalnych pojęć, natomiast wadą jest konieczność wykonywania podziału zbioru obiektów symbolicznych dla różnych typów dziedzin zmiennych.

Literatura

- Bock H.H., Diday E. (2000), *Analysis of Symbolic Data*, Springer-Verlag, Berlin, s. 294-311.
- Chavent M., Touati M. (2001), *A Divisive and Clustering Method*, Sodas-school, France.
- Chavent M. (1997), *Analyse de données symboliques. Une méthode divisive de classification*, praca doktorska, Université Paris IX Dauphine.
- Diday E. (2002), *An introduction to Symbolic Data Analysis and the Sodas software*, „The Electronic Journal of Symbolic Data Analysis”, lipiec, nr 1.

APPROACH DIVISIVE CLUSTERING INTO SYMBOLIC OBJECTS CLASSIFICATION

Summary

The article presents divisive clustering method for symbolic data array. This is a hierarchical clustering method. Its main aim is division of the set of symbolic objects for K separate clusters, inside which objects are as homogeneous as possible, however each cluster is as different as possible. K is the number of clusters given as input by the user. As a summary they present an example worked out by Sodas v. 1.2 programme.