

Małgorzata Wójcik

Akademia Ekonomiczna w Katowicach

ZASTOSOWANIE SYMBOLICZNEJ ANALIZY CZYNNIKOWEJ DO KLASYFIKACJI OBIEKTÓW SYMBOLICZNYCH

1. Wstęp

Gwałtowny rozwój technologii komputerowej znacznie ułatwia gromadzenie danych, co prowadzi do powstawania coraz większych baz danych. Informacje zawarte w bazach danych często mają charakter zmiennych symbolicznych, których nie można przetworzyć za pomocą klasycznych metod. W związku z tym coraz bardziej popularna staje się symboliczna analiza danych (*symbolic data analysis*). Uwzględnia ona głównie zmienne jakościowe zgromadzone w symbolicznej tabeli danych. Na podstawie tych danych są tworzone obiekty symboliczne, które następnie podlegają klasyfikacji. Metody analizy symbolicznej są zaliczane do metod nienumerycznych, a uzyskane wyniki są dość łatwe w interpretacji, ponieważ do opisu klas obiektów lub obiektów symbolicznych wykorzystuje się pojęcia naturalne.

W artykule zaprezentowano podstawy symbolicznej analizy czynnikowej (*factorial discriminant analysis*) oraz przykład jej zastosowania do klasyfikacji obiektów symbolicznych.

2. Główne cele

Do głównych celów symbolicznej analizy czynnikowej należą:

- przedstawienie na płaszczyźnie czynnikowej klas obiektów symbolicznych jak najlepiej od siebie oddzielonych,
- definiowanie geometrycznej reguły klasyfikacyjnej,
- ocena reguły klasyfikacyjnej.

W metodzie tej punktem wyjścia jest zbiór danych symbolicznych zgromadzonych w symbolicznej tabeli danych, z których następnie są tworzone obiekty symboliczne jako koniunkcje wartości poszczególnych cech obiektu (deskryptorów). Na przykład jedną z kart kredytowych, oferowaną przez mBank można opisać następująco:

$$\{ \text{VisaCredit} \} = [\text{oprocentowanie}=27\%] \wedge [\text{grace_period}=53] \wedge \\ \wedge [\text{karta_glowna}=70] \wedge [\text{konto} = \text{nie}] \wedge [\text{limit}=10000] \wedge \\ \wedge [\text{prowizja}=3\% \text{ min } 5\text{zł}] \wedge [\text{uslugi}=\text{tak}].$$

Następnie dokonuje się numerycznej transformacji deskryptorów obiektów symbolicznych, aby ostatecznie przedstawić symboliczną interpretację otrzymanych rezultatów.

Procedury symbolicznej analizy czynnikowej, dostępne w pakiecie SODAS 1.200, dopuszczają następujące typy deskryptorów obiektów symbolicznych:

- numeryczne (tj. typu rzeczywistego lub przedziałowego),
- nominalne lub multinominalne,
- modalne.

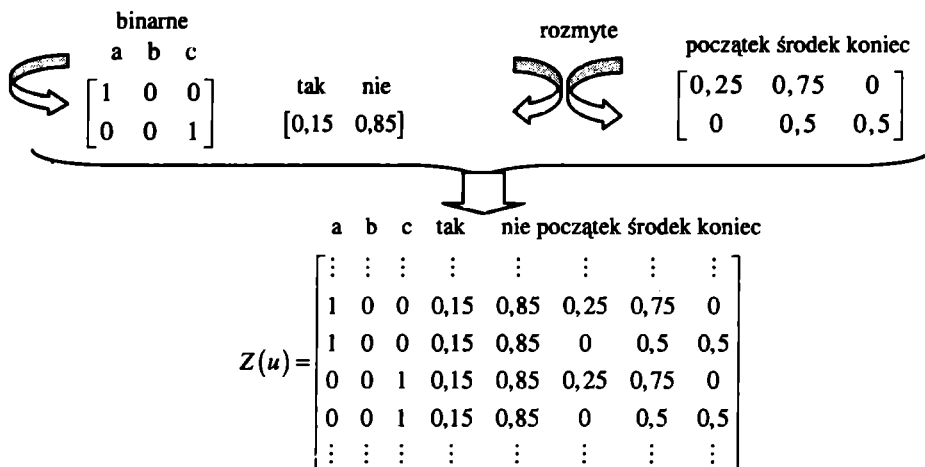
Jako dane wejściowe rozważany jest zbiór uczący E oraz zbiór testowy E^+ obiektów symbolicznych. Obiekty symboliczne zbioru testowego są tworzone z tej samej populacji, co obiekty zbioru uczącego. Na podstawie zbioru uczącego są budowane reguły klasyfikacji geometrycznej, które dostępne są w pakiecie SODAS, na podstawie zbioru testowego dokonywana jest zaś ocena tych reguł.

3. Kodowanie deskryptorów obiektów symbolicznych

Każdy deskryptor obiektu symbolicznego musi zostać zakodowany. Dokonuje się tego za pomocą kodowania binarnego lub rozmytego, a wyniki są gromadzone w globalnej macierzy kodującej $Z_{N \times K}$. Wybór funkcji kodującej zależy od rodzaju deskryptorów obiektu symbolicznego.

Przykład kodowania deskryptorów obiektu symbolicznego u

Niech obiekt symboliczny u opisany będzie przez zmienne $X_1 = \{a, c\}$, $X_2 = \{tak(0,15), nie(0,85)\}$, $X_3 = [55, 70]$ odpowiednio o dziedzinach $D_{X_1} = \{a, b, c\}$, $D_{X_2} = \{tak, nie\}$, $D_{X_3} = [40, 80]$. Wówczas kodowanie dla obiektu symbolicznego u można przedstawić następująco:



Nazwy kolumn „początek”, „środek”, „koniec” w macierzy $Z(u)$ to odpowiednio zakodowane końce oraz wartość z wnętrza przedziału $X_3 = [55, 70]$.

Liczbę kolumn globalnej macierzy kodującej Z można obliczyć po skorzystaniu z następującego wzoru :

$$K = 3 \cdot g + \sum_{j=1}^q k_j, \quad (1)$$

gdzie : $g \leq p$ wyraża liczbę zmiennych ilościowych (rzeczywistych lub przedziałowych),

$q = p - g$ wyraża liczbę zmiennych nominalnych,

k_j jest liczbą wszystkich kategorii zmiennej nominalnej $Y_j (j = 1, \dots, q)$.

Całkowita liczba wierszy macierzy Z zależy od liczby zmiennych typu przedziałowego oraz od liczby kategorii zmiennej nominalnej. Zauważyć należy, iż zmienne modalne nie powodują wzrostu liczby wierszy macierzy Z , ponieważ są one kodowane w ten sposób, że jeden wiersz (zakodowana zmienna modalna zawiera odpowiednie wagi) odpowiada dokładnie jednemu obiektowi symbolicznemu. Jeśli przez $h \leq g$ oznaczona zostanie liczba zmiennych typu przedziałowego, przez $l = q + (g - h)$ – liczba zmiennych zarówno nominalnych wielowartościowych, jak i zmiennych rzeczywistych, przez $k_{uj} = |Y_j(u)|$ – liczba kategorii zmiennej nominalnej Y_j dla obiektu symbolicznego u , to liczba wierszy macierzy Z wyraża się następującym wzorem :

$$N = \sum_{i=1}^n \left(2^h \cdot \prod_{v=1}^l k_{u_v} \right). \quad (2)$$

4. Dobór predyktorów dyskryminacyjnych

Po zakodowaniu deskryptorów obiektów symbolicznych należy dobrać najlepsze predyktory dyskryminacyjne. Zatem niech $G_{N \times n}$ będzie macierzą wskaźników identyfikujących różne obiekty ze zbioru E , tzn. $g_{wu} = 1$ lub 0 , jeśli odpowiednio kolumna w w macierzy Z zawiera bądź nie element z . Optymalną ilość p -deskryptorów z K -kategorii otrzymuje się jako rozwiązanie następującego równania :

$$(G'Z\Sigma^{-1}ZG)w_{\alpha} = \mu_{\alpha}w_{\alpha}, \quad (3)$$

przy założeniu: $w_{\alpha}w_{\alpha} = 1$ oraz $w_{\alpha}w_{\alpha'} = 0$, dla $\alpha \neq \alpha'$,

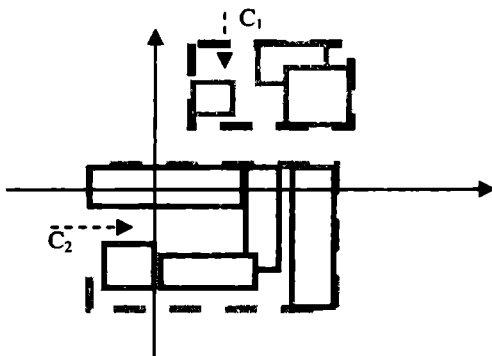
gdzie μ_{α} i w_{α} są odpowiednio α -tą wartością własną i wektorem własnym macierzy $G'Z\Sigma^{-1}ZG$ we wzorze 3, Σ^{-1} jest odpowiednią macierzą diagonalną. Zakładając, że liczba K wszystkich kategorii jest mniejsza od n , to liczba wszystkich nietrywialnych wartości własnych jest równa: $K - p + 1$. W ten sposób współrzędne wierzchołków obiektów symbolicznych można wyznaczyć ze wzoru :

$$\Phi_{\alpha} = Zw_{\alpha}, \quad \alpha = 1, \dots, K - p + 1. \quad (4)$$

Tak otrzymane zmienne czynnikowe mogą być użyte jako nowe predyktory.

5. Wybór podprzestrzeni klas obiektów i budowa reguł klasyfikacyjnych

W dalszej części należy wskazać taką podprzestrzeń, w której można geometrycznie przedstawić m optymalnie rozdzielonych klas $C_1, \dots, C_m$, tak aby rozproszenie obiektów wewnątrz klasy było jak najmniejsze (rys. 1).



Rys.1. Przykładowe klasy obiektów symbolicznych
Źródło: opracowanie własne.

Jeśli przez C oznaczona będzie macierz wskaźników, które identyfikują obiekt symboliczny z klasą, np. :

$$C = \left. \begin{matrix} 1 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ N \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix} \end{matrix} \right\} \begin{matrix} \text{wierzchołki tych obiektów} \\ \text{symbolicznych należą do} \\ \text{klasy 1,} \end{matrix}$$

to równanie charakterystyczne symbolicznej analizy czynnikowej ma następującą postać:

$$(\Phi' \Phi)^{-1} \Phi' C (C' C)^{-1} C' \Phi w_\alpha = \mu_\alpha w_\alpha, \quad (5)$$

gdzie: $\Phi = [\Phi_1 | \dots | \Phi_\alpha] =$ nowe predyktory.

Przydziału nowych obiektów symbolicznych do *a priori* zdefiniowanych klas dokonuje się po skorzystaniu z jednej z dwóch reguł klasyfikacji geometrycznej.

Pierwsza z tych reguł jest oparta na mierze niepodobieństwa De Carvalho. W celu sklasyfikowania obiektu symbolicznego $v \in E^+$, należy obliczyć jego niepodobieństwo z każdym obiektem symbolicznym $u \in E$, za pomocą wzoru (6):

$$d(u, v) = \sqrt[r]{\sum_{\alpha=1}^s [\kappa_\alpha \varphi(S_{v\alpha} S_{u\alpha})]^r}, \quad (6)$$

gdzie: $\kappa_\alpha > 0$ to α -ta wartość własna,

$r \geq 1$ jest odpowiednio dobranym wykładnikiem (najczęściej równym 2), ponadto:

$$\varphi(S_{v\alpha} S_{u\alpha}) = \frac{\mu(S_{v\alpha} \oplus S_{u\alpha}) - \mu(S_{v\alpha} \cap S_{u\alpha}) + \gamma(2\mu(S_{v\alpha} \cap S_{u\alpha}) - \mu(S_{v\alpha}) - \mu(S_{u\alpha}))}{\mu(S_{v\alpha} \oplus S_{u\alpha})},$$

gdzie $0 \leq \gamma \leq 1$.

Miara d wyraża odległość obiektu v od każdego obiektu u z klasy C_j . Aby jednak znaleźć odległość obiektu V od klasy $C_i \subseteq E$, należy obliczyć średnią odległość v na podstawie wzoru (7):

$$\bar{d}(v, C_i) = \frac{1}{n_i} \sum_{u \in C_i} d(u, v), \quad (7)$$

gdzie: $n_i = |C_i|$.

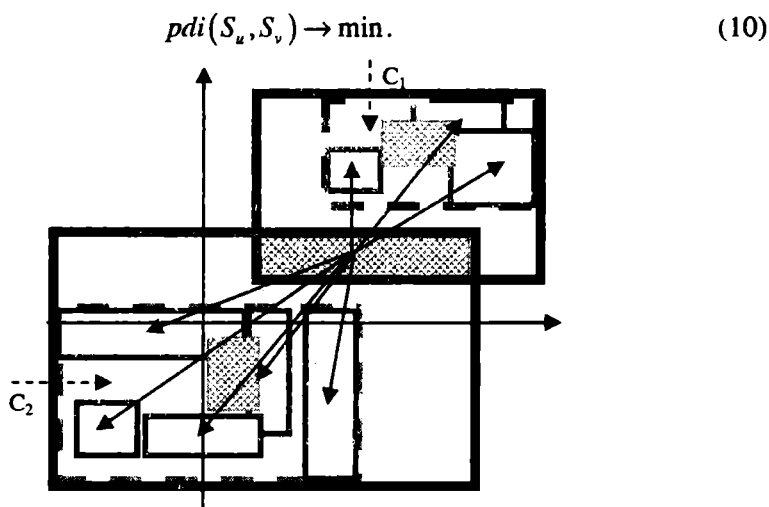
Ostatecznie obiekt symboliczny v zostanie przydzielony do tej klasy $\Pi_i \subseteq E^+$, dla której średnia odległość $\bar{d}(v, C_i)$ jest minimalna (por. rys. 2):

$$\bar{d}(v, C_i) \rightarrow \min. \quad (8)$$

Druga z reguł klasyfikacji geometrycznej jest oparta na pojęciu minimalnego inkrementacyjnego potencjału opisowego, związanego *de facto* ze wzrostem objętości klasy:

$$pdi(S_u, S_v) = \frac{\prod_{\alpha=1}^s \mu(S_{u\alpha} \oplus S_{v\alpha}) - \prod_{\alpha=1}^s \mu(S_{v\alpha})}{\prod_{\alpha=1}^s \mu(S_{u\alpha})}. \quad (9)$$

Zgodnie z tą regułą obiekt symboliczny v zostanie przydzielony do tej klasy obiektów symbolicznych, dla których wartość $pdi(S_u, S_v)$ jest minimalna (rys. 2), dla każdego $u \in E$:



Rys. 2. Schemat przydziału nowego obiektu symbolicznego do klasy
Źródło: opracowanie własne.

W ostatniej fazie następuje ocena jakości klasyfikacji przez ostateczne przydzielenie obiektów zbioru testowego do klas, a następnie obliczenie współczynnika poprawności klasyfikacji.

6. Przykład

Do zaprezentowania przykładu wykorzystano metodę FDA (*factorial discriminant analysis*), dostępną w programie SODAS 1.200. Do obliczeń wykorzystano plik testowy TESTSDT.SDS dołączony do programu. Zawiera on 720 obiektów symbolicznych, charakteryzowanych przez 12 zmiennych. Obiekty przydzielono do 4 klas:

- specjaliści,
- technicy,
- urzędnicy,
- robotnicy.

Tabela 1 zawiera odległości poszczególnych obiektów symbolicznych od klas, obliczone na podstawie miary De Carvalho za pomocą wzorów (6) oraz (7). Autorka za wartość parametru r przyjęła 2, a $\gamma = \frac{1}{2}$.

Ze względu na to, że liczba obiektów symbolicznych wynosi 720, wielkość tab. 1 i 2 ograniczono i zawierają one jedynie kilkadziesiąt początkowych obiektów.

Tabela 1. Tabela odległości obiektów od klas

Metoda przydziału obiektu do klasy na podstawie najmniejszej odległości obliczonej za pomocą miary De Carvalho				
test/klasy	specjaliści	technicy	urzędnicy	robotnicy
„rdg p8 e100_249”	1	0,880	0,855	0
„rdh p8 e100_249”	1	0,845	0,815	0
„rdi p8 e100_249”	1	0,852	0,832	0
„rdj p8 e100_249”	1	0,849	0,831	0
„rdk p8 e100_249”	1	0,892	0,867	0
„rdl p8 e100_249”	1	0,872	0,856	0
„rdn p8 e100_249”	1	0,854	0,846	0
„rda p9 e100_249”	1	0,850	0,858	0
„rdd p9 e100_249”	1	0,879	0,857	0
„rde p9 e100_249”	1	0,890	0,861	0
„rdg p9 e100_249”	1	0,872	0,850	0
„rdh p9 e100_249”	1	0,869	0,857	0
„rdj p9 e100_249”	1	0,872	0,857	0
„rdn p9 e100_249”	1	0,875	0,857	0
„rg p9 e100_249”	1	0,892	0,865	0
„rh p9 e100_249”	1	0,862	0,839	0
„ri p9 e100_249”	1	0,873	0,848	0
„rk p9 e100_249”	1	0,893	0,814	0
„rj p4 e10_99”	1	0,890	0	0,885
„rdg p4 e10_99”	1	0,857	0	0,854
„rf p8 e100_249”	1	0,850	0	0,825

Źródło: opracowanie własne.

Tabela 2. Tabela klasyfikacji obiektów do klas

Test/klasy	Specjaliści	Technicy	Urzędnicy	Robotnicy
„rdg p8 e100_249”	0	0	0	1
„rdh p8 e100_249”	0	0	0	1
„rdi p8 e100_249”	0	0	0	1
„rdj p8 e100_249”	0	0	0	1
„rdk p8 e100_249”	0	0	0	1
„rdl p8 e100_249”	0	0	0	1
„rdn p8 e100_249”	0	0	0	1
„rda p9 e100_249”	0	0	0	1
„rdd p9 e100_249”	0	0	0	1
„rde p9 e100_249”	0	0	0	1
„rdg p9 e100_249”	0	0	0	1
„rdh p9 e100_249”	0	0	0	1
„rdj p9 e100_249”	0	0	0	1
„rdn p9 e100_249”	0	0	0	1
„rg p9 e100_249”	0	0	0	1
„rh p9 e100_249”	0	0	0	1
„ri p9 e100_249”	0	0	0	1
„rk p9 e100_249”	0	0	0	1
„rj p4 e10_99”	0	0	1	0
„rdg p4 e10_99”	0	0	1	0
„rf p4 e10_99”	0	0	1	0
„rk p4 e10_99”	0	0	1	0

Źródło: opracowanie własne.

Tabela 3. Liczba obiektów należących do klas

Klasy/klasy	Specjaliści	Technicy	Urzednicy	Robotnicy
Specjaliści	153	0	0	0
Technicy	0	95	0	0
Urzednicy	0	0	176	0
Robotnicy	0	0	0	296
Współczynnik poprawności klasyfikacji: 100%				

Źródło: opracowanie własne.

Zgodnie ze wzorem (8), obiekt symboliczny v powinien zostać przydzielony do tej klasy, dla której jego odległość od klasy jest minimalna. Z tabeli 1 wynika, że najmniejsza odległość obiektu symbolicznego *rdg p8 e100_249* jest dla klasy *robotnicy* i wynosi 0, zatem obiekt ten powinien być przydzielony do klasy *robotnicy*. Tabela 2, przedstawiająca rezultaty klasyfikacji poszczególnych obiektów do klas, potwierdza powyższe rozważania.

Tabela 3 przedstawia sumaryczne zestawienie obiektów należących do odpowiednich klas oraz współczynnik poprawności klasyfikacji. Do klasy *specjaliści* sklasyfikowano 153 obiekty symboliczne, do klasy *technicy* – 95, do klasy *urzednicy* – 176, a do klasy *robotnicy* – 296; współczynnik poprawności klasyfikacji, z zastosowaniem miary De Carvalho z parametrami $r = 2$ i $\gamma = \frac{1}{2}$, wynosi 100%.

7. Podsumowanie

Metoda analizy danych zaprezentowana w tym artykule pozwala na korzystanie z danych symbolicznych, z których są tworzone obiekty symboliczne. Obiekty te można np. uzyskać z relacyjnych baz danych Access, po zastosowaniu modułu DB2SO programu SODAS.

Literatura

- Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, red. H.H. Bock, E. Diday, Springer Verlag 2000.
- Diday E., *An Introduction to Symbolic Data Analysis and The Sodas Software*, „The Electronic Journal of Symbolic Data Analysis”, lipiec 2002.
- Dudek A., *Obiekty symboliczne w klasyfikacji danych*, Prace Naukowe Akademii Ekonomicznej nr 1022, AE, Wrocław 2004, s. 284-290.
- Gatnar E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa 1998.
- Miles B., Huberman A., *Analiza danych jakościowych*, Wydawnictwo Trans Humana, Warszawa 2000.

APPROACH SYMBOLIC FACTORIAL DISCRIMINANT ANALYSIS INTO SYMBOLIC OBJECTS CLASSIFICATION

Summary

In the article the bases of Factorial Discriminant Analysis are introduced. The method allows to discover dependency among various symbolic objects characteristic. The main objectives are defining a geometrical classification rules and visualizing on a factorial plane the classes of symbolic objects, separated from each other in the best way. To sum up the example based on SODAS programme was introduced.