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

Ewa Witek
Akademia Ekonomiczna w Katowicach

PROBLEM DOBORU ZMIENNYCH W PODEJŚCIU MODELOWYM W ANALIZIE SKUPIEŃ

1. Wstęp

W analizie skupień coraz większe znaczenie ma podejście modelowe (*model-based clustering*), w którym obserwacje pochodzą z mieszanki rozkładów, a rozkłady składowe mieszanek identyfikowane są z klasami. Dobór zmiennych w podejściu modelowym dokonywany jest zazwyczaj w sposób intuicyjny lub za pomocą strategii zachłannej. Wybór optymalnych zmiennych za pomocą strategii zachłannej w podejściu modelowym polega na tym, że w każdym kroku „szuka się" zmiennej, która w największym stopniu poprawia jakość klasyfikacji mierzoną za pomocą kryterium informacyjnego BIC. W rezultacie wybierany jest model o najlepszych własnościach i optymalnej liczbie klas (każde dwa konkurujące modele postrzegane są jako dwa zbiory zmiennych).

W artykule dokonane zostanie porównanie wyników klasyfikacji w przypadku wykorzystania wszystkich zmiennych oraz po dokonaniu ich wstępnej selekcji za pomocą strategii przeszukiwania zachłannego (*greedy search*) oraz metody *HINoV* [Carmone, Kara, Maxwell 1999, s. 508].

2. Model mieszanek

Zakłada się, że wielowymiarowe obserwacje $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^{\mathrm{T}}$ pochodzą z jednej z s $(s = 1, \ldots u)$ podpopulacji. Liczba podpopulacji, tj. rozkładów składowych mieszanki (*components of mixture model*), rozumiana jest tutaj jako liczba klas. Wartość funkcji gęstości dla obserwacji $\mathbf{x}_i$ jest równa:

$$f(\mathbf{x}_i \,|\, \boldsymbol{\Theta}) = \sum_{s=1}^{u} \tau_s f_s(\mathbf{x_i} \,|\, \boldsymbol{\Theta}_s), \tag{1}$$

gdzie: f_s – funkcja gęstości klasy P_s (s-tego rozkładu składowego mieszanki),
$\boldsymbol{\Theta}_s$ – wektor parametrów rozkładu w klasie P_s,
τ_s – wektor prawdopodobieństw *a priori* – wyznacza wartość prawdopodobieństwa, że dana obserwacja należy do klasy P_s ($\tau_s \geq 0, \sum_{s=1}^{u} \tau_s = 1$).

Najczęściej za f_s przyjmujemy funkcję gęstości wielowymiarowego rozkładu normalnego o wektorze średnich rozkładu $\boldsymbol{\mu}_s = [\mu_1, \ldots, \mu_m]^T$ i macierzy kowariancji $\boldsymbol{\Sigma}_s$. Nałożenie ograniczeń na macierz kowariancji poprzez jej dekompozycję według wartości własnych przyczyniło się do powstania modeli mieszanek Gaussa o różnych cechach geometrycznych. Cechy geometryczne 10 modeli dostępnych w pakiecie `mclust`, sposób estymacji parametrów modeli oraz wyboru modelu optymalnego można znaleźć m.in. w pracach [Fraley 2002; Witek 2008, s. 199-206].

3. Selekcja zmiennych w analizie skupień opartej na mieszankach rozkładów

Selekcja zmiennych w analizie skupień opartej na modelach mieszanek dokonywana jest na podstawie modelu statystycznego. Zbiór zmiennych Y w każdym kroku algorytmu dzielony jest na trzy rozdzielne podzbiory:

- $\mathrm{Y}^{(1)}$ zbiór zmiennych mających moc dyskryminacyjną,
- $\mathrm{Y}^{(2)}$ zmienna(e) kandydująca(e) do dołączenia do zbioru $\mathrm{Y}^{(1)}$ lub usunięcia z niego,
- $\mathrm{Y}^{(3)}$ zbiór pozostałych zmiennych.

Decyzja o dołączeniu zmiennej do zbioru zmiennych o mocy dyskryminacyjnej $\mathrm{Y}^{(1)}$ lub usunięciu jej z tego zbioru jest podejmowana na podstawie dwóch konkurujących modeli:

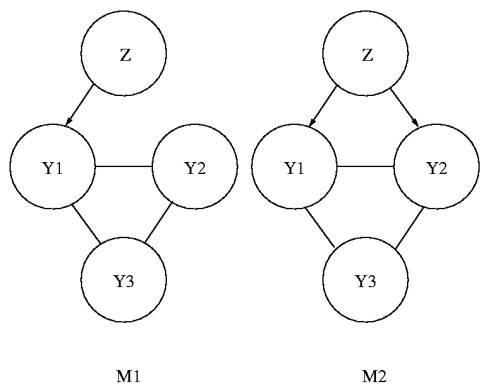
$$M_1: \; p(\mathrm{Y}|\mathbf{z}) = p(\mathrm{Y}^{(1)}, \mathrm{Y}^{(2)}, \mathrm{Y}^{(3)}|\mathbf{z}) = p(\mathrm{Y}^{(3)}|\mathrm{Y}^{(2)}, \mathrm{Y}^{(1)})\, p(\mathrm{Y}^{(2)}|\mathrm{Y}^{(1)})\, p(\mathrm{Y}^{(1)}|\mathbf{z}), \quad (2)$$

$$M_2: \; p(\mathrm{Y}|\mathbf{z}) = p(\mathrm{Y}^{(1)}, \mathrm{Y}^{(2)}, \mathrm{Y}^{(3)}|\mathbf{z}) = p(\mathrm{Y}^{(3)}|\mathrm{Y}^{(2)}, \mathrm{Y}^{(1)})\, p(Y^{(2)}, \mathrm{Y}^{(1)}|\mathbf{z}), \quad (3)$$

gdzie: $\mathbf{z}$ to (nieobserwowalny) wektor przynależności obiektów do klas. W modelu M_1 zakłada się, że dla danego zbioru $\mathrm{Y}^{(1)}$ $\mathrm{Y}^{(2)}$ jest warunkowo niezależny od przynależności obiektów do klas (określonej za pomocą wektora $\mathbf{z}$), co oznacza, że $\mathrm{Y}^{(2)}$ nie ma wpływu na strukturę klas. W modelu M_2 zakłada się, że przy danym zbiorze $\mathrm{Y}^{(1)}$ zmienne zbioru $\mathrm{Y}^{(2)}$ mają zdolność dyskryminacji analizowanego zbioru obiektów [Raftery, Dean 2006].

Zaletą jest to, że zmienne nieposiadające mocy dyskryminacyjnej nie muszą być niezależne od zmiennych klasyfikujących. Jeżeli natomiast założenie o niezależności $p(Y^{(2)}|Y^{(1)}) = p(Y^{(2)})$ zostało spełnione dla modelu M_1, to możliwe jest uwzględnie-

nie zmiennych skorelowanych ze zmiennymi o mocy dyskryminacyjnej. Zmienne te jednak same nie mają zdolności dyskryminacji. Zakłada się, że zmienne zbioru $Y^{(3)}$ są warunkowo niezależne od zmiennych o mocy dyskryminacyjnej $Y^{(1)}$ i zbioru $Y^{(2)}$. Różnica założeń dla zbiorów $Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$ w modelach M_1 i M_2 została przedstawiona na rys. 1, na którym strzałki oznaczają zależność pomiędzy $\mathbf{z}$ i $Y^{(1)}$, $Y^{(2)}$.



Rys. 1. Graficzna prezentacja modeli wykorzystywanych w metodzie doboru zmiennych w analizie skupień opartej na mieszankach rozkładów
Źródło: [Raftery, Dean 2005, s. 169].

Modele M_1 i M_2 są porównywane za pomocą przybliżenia czynnika Bayesa *(Bayes factor*). Czynnik Bayesa B_{12} dla modelu M_1 w zależności od M_2 jest określony jako:

$$B_{12} = \frac{p(Y|M_1)}{p(Y|M_2)}, \tag{4}$$

gdzie: $p(\mathrm{Y}|M_k) = \int p(\mathrm{Y}|\boldsymbol{\Theta}_k, M_k) p(\boldsymbol{\Theta}_k|M_k)\, d\boldsymbol{\Theta}_k$ to brzegowa gęstość wektora obserwacji w modelu M_k (*integrated likelihood for model* M_k), w której $\boldsymbol{\Theta}_k$ jest wektorem parametrów modelu M_k, a $p(\boldsymbol{\Theta}_k|M_k)$ jest prawdopodobieństwem *a priori* [Kass, Raftery 1995].

Zakładając, że prawdopodobieństwa *a priori* $\boldsymbol{\Theta}_\mathrm{k}$ są takie same dla modeli M_1 i M_2, tj. $p(Y^{(3)}|Y^{(2)}, Y^{(1)}, M_2) = p(Y^{(3)}|Y^{(2)}, Y^{(1)}, M_1)$, otrzymuje się:

$$B_{12} = \frac{p(Y^{(2)}, Y^{(1)}|M_1)\, p(Y^{(1)}|M_1)}{p(Y^{(2)}, Y^{(1)}|M_2)}. \tag{5}$$

Wzór (5) jest dużym uproszczeniem przez pominięcie wielowymiarowego zbioru $Y^{(3)}$.

W praktyce wykorzystywane jest przybliżenie czynnika Bayesa dane wzorem:

$$2\log(B_{12}) \approx BIC(M_1) - BIC(M_2). \tag{6}$$

Podwojony logarytm czynnika Bayesa jest w przybliżeniu równy różnicy wartości kryteriów BIC obserwacji [Schwarz 1978, s. 461- 464] dla dwóch porównywanych ze sobą modeli. Bayesowskie kryterium informacyjne Schwarza BIC (*Bayesian Information Criterion*) jest statystyką pomiaru jakości dopasowania w przypadku modeli szacowanych metodą największej wiarygodności [Witek 2008, s. 202].

Najczęściej rozważa się przypadek, w którym do zbioru $\mathrm{Y}^{(2)}$ należy tylko jedna zmienna i $p(Y^{(2)}\left|Y^{(1)}, M_1\right.)$ oznacza funkcję regresji. Wartość kryterium informacyjnego Schwarza można przybliżyć w następujący sposób:

$$2\log p(Y^{(2)}\left|Y^{(1)}, M_1\right.) \approx BIC_{reg} = -n\log(2\pi) - n\log(S/n) - n - (\dim(Y^{(1)}) + 2)\log(n), \tag{7}$$

gdzie: S to suma kwadratów reszt dla funkcji regresji o zmiennej objaśnianej należącej do zbioru $\mathrm{Y}^{(2)}$ i zmiennych objaśniających ze zbioru $\mathrm{Y}^{(1)}$, $\dim(Y^{(1)})$ zaś określa wymiar zbioru $Y^{(1)}$.

4. Metoda doboru zmiennych w analizie skupień wykorzystująca algorytm przeszukiwania zachłannego

Algorytm przeszukiwania zachłannego (*greedy search algorithm*) w analizie skupień opartej na mieszankach rozkładów polega na tym, że w każdym kroku przeszukiwany jest cały zbiór zmiennych tak, by wybrać zmienną, która w największym stopniu poprawia jakość klasyfikacji mierzoną za pomocą kryterium BIC. Następnie dokonywana jest ocena, czy któraś ze zmiennych może zostać usunięta. W każdym kroku wybierana jest najlepsza kombinacja liczby klas i modelu o różnych cechach geometrycznych. Algorytm zatrzymuje się w przypadku, gdy nie następuje żadna poprawa wartości kryterium BIC [Raftery, Dean 2006].

Zdolność dyskryminacji zbioru obiektów dla każdej zmiennej obliczana jest jako różnica wartości kryteriów BIC dla różnych modeli mieszanek (o różnych cechach geometrycznych) o różnej liczbie klas:

$$BIC_{diff}(Y^{(2)}) = BIC_{tak}(Y^{(2)}) - BIC_{nie}(Y^{(2)}), \tag{8}$$

$$BIC_{tak}(Y^{(2)}) = \max_{2\le s\le u} BIC(Y^{(2)}), \tag{9}$$

$$BIC_{tak}(Y^{(2)}) = BIC(p(Y^{(2)} \mid Y^{(1)}) - BIC(p(\mathrm{Y}^{(1)} \mid \mathbf{z})). \tag{10}$$

Kroki algorytmu:

1. Wybiera się zmienną o najwyższej mocy dyskryminacyjnej (rozpatrywany jest jednowymiarowy zbiór obserwacji), tj. zmienną o najwyższej wartości różnicy

kryteriów $BIC_{diff}(Y^{(2)})$ dla modeli jednowymiarowych E i V zgodnie ze wzorem (8). Model E to model, którego klasy cechują się taką samą objętością, natomiast w modelu V klasy mają różną objętość [Witek 2008, s. 201].

2. Wybiera się drugą zmienną o najwyższej mocy dyskryminacyjnej, tj. o najwyższej wartości $BIC_{diff}(Y^{(2)})$ (pierwsza wybrana zmienna pozostaje bez zmian).

3. Wybiera się kolejną zmienną o najwyższej mocy dyskryminacyjnej dla zbioru wielowymiarowego (wcześniej wybrane zmienne pozostają bez zmian). Zmienna ta jest akceptowana, jeżeli wartość różnicy $BIC_{diff}(Y^{(2)})$ jest wyższa dla zbioru zmiennych o mocy dyskryminacyjnej zawierającego analizowaną zmienną – $BIC_{tak}(Y^{(2)})$ w porównaniu ze zbiorem bez tej zmiennej – $BIC_{nie}(Y^{(2)})$, tzn. gdy $BIC_{diff}(Y^{(2)}) > 0$.

4. Proponuje się zmienną, którą można usunąć z wybranego zbioru zmiennych klasyfikujących, tj. zmienną o najniższej wartości $BIC_{diff}(Y^{(2)})$. Zmienna usuwana jest ze zbioru wówczas, gdy jej wartość BIC dla zbioru zmiennych o zdolności dyskryminacyjnej zawierających analizowaną zmienną jest mniejsza w porównaniu ze zbiorem zmiennych niezawierającym tej zmiennej (gdy $BIC_{diff}(Y^{(2)}) < 0$).

5. Powtarza się kroki 3 i 4, tak by jakość podziału mierzona kryterium BIC osiągnęła najwyższą wartość.

5. Przykłady empiryczne

Wybór zmiennych w pakiecie `clustvarsel` programu R

Za pomocą dwuwymiarowej zmiennej losowej o rozkładzie normalnym dla dwóch skupień wygenerowano po 200 obserwacji. Przyjęto następujące wektory wartości oczekiwanych dla skupień (1; 0,5), (0,5; 1) oraz macierze kowariancji $\mathbf{\Sigma}_1 = \begin{bmatrix} 1 & 0,5 \\ 0,5 & 1 \end{bmatrix}$; $\mathbf{\Sigma}_2 = \begin{bmatrix} 1,5 & -0,7 \\ -0,7 & 1,5 \end{bmatrix}$. Do analizy wprowadzono dodatkowo 6 zmiennych zakłócających (zmienne o nr 1-6) istniejącą w układzie dwuwymiarowym strukturę klas (tzw. *noisy variables*). Do wygenerowania danych wykorzystano funkcje `rnorm` i `mvnorm` pakietu `stats`. Wynik selekcji zmiennych w analizie skupień opartej na modelach z wykorzystaniem strategii zachłannej, uzyskany w pakiecie `clustvarsel` programu **R,** przedstawiono na rys. 2.

```
Variable proposed BIC of new clustering variables set
BIC difference
[1,] "8" "-805,765654963164" "35,5242020507347"
[2,] "7" "-1445,15007066406" "97,190339663286"
[3,] "2" "-1445,15007066406" "-8,71587638824872"
[4,] "7" "-1445,15007066406" "97,190339663286"
```

```
Type of step Decision
[1,] "Add" "Accepted"
[2,] "Add" "Accepted"
[3,] "Add" "Rejected"
[4,] "Remove" "Rejected"
```

Rys. 2. Wynik selekcji w pakiecie `clustvarsel`
Źródło: opracowanie własne.

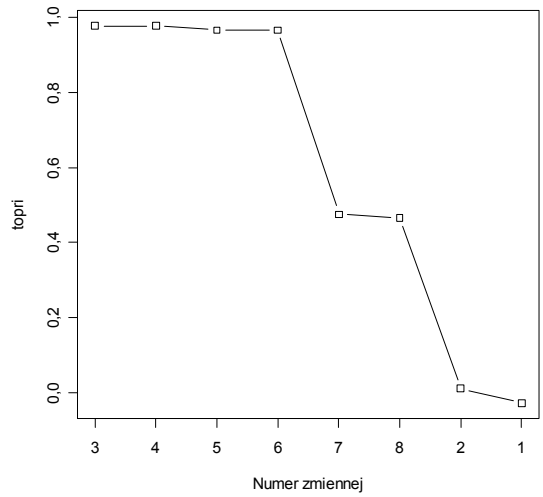
Jako pierwsze do zbioru zmiennych o mocy dyskryminacyjnej zostały wybrane zmienne nr 8 i 7 (wartość różnicy kryteriów BIC była dodatnia, dlatego zmienne zostały wybrane). Kolejną zmienną „kandydującą" do zbioru $Y^{(1)}$ była zmienna nr 2. Ponieważ wartość różnicy kryteriów BIC była ujemna, zmienna ta nie została dołączona do zbioru $Y^{(1)}$. W kolejnym kroku rozważano możliwość usunięcia zmiennej nr 7 – wartość różnicy kryteriów BIC była dodatnia, dlatego zmienna nie została usunięta ze zbioru $Y^{(1)}$. W wyniku selekcji do zbioru o mocy dyskryminacyjnej wybrano dwie zmienne o nr 7 i 8.

Dokonując podziału za pomocą analizy skupień opartej na modelach, przekonano się, że dla wszystkich zmiennych najlepszym modelem okazał się model VVV o 1 klasie. Przeprowadzając wstępną selekcję zmiennych za pomocą strategii zachłannej, stwierdzono, że liczba klas będąca wynikiem taksonomii opartej na mieszankach rozkładów wyniosła dokładnie 2. Model VVV to model dla zbiorów wielowymiarowych, którego klasy cechują się różnym kształtem, objętością i orientacją [Witek 2008, s. 201]. Raftery i Dean [2006, s. 172-176] w swej pracy pokazują, że metoda ta daje również bardzo dobre wyniki na zbiorach benchmarkowych, tj. `crabs`, `iris`, `texture`.

Analiza porównawcza z metodą *HINoV*

Znajdująca szczególne zastosowanie w analizie skupień opartej na modelach mieszanek metoda doboru zmiennych, wykorzystująca strategię przeszukiwania zachłannego, porównana została z heurystyczną procedurą doboru zmiennych *HINoV* [Carmone, Kara, Maxwell 1999, s. 504]. W tym celu wykorzystano funkcję `HINoV.Mod` pakietu `clusterSim` o następujących parametrach: liczba klas – 2, miara odległości – kwadrat odległości euklidesowej `d4`, metoda klasyfikacji – `kmeans`. Wynik selekcji dla wygenerowanego zbioru 8 zmiennych, jaki otrzymano po zastosowaniu metody *HINoV*, przedstawiono na rys. 3.

Na podstawie wykresu osypiska należy wybrać 4 zmienne, tj. zmienne o numerach 3, 4, 5, 6. Między zmiennymi 6 i 7 występuje znaczny spadek wartości skorygowanych współczynników Randa (`topri`). Zmienne o numerach: 1, 2, 7, 8, zakłócające w układzie dwuwymiarowym strukturę klas zostają usunięte. Warto podkreślić, że w przeciwieństwie do metody selekcji opartej na modelach w metodzie *HINoV* przyjęto *a priori* liczbę klas równą 2. Pomimo to, metoda ta nie dała wyników poprawnych



Rys. 3. Wynik metody *HINoV* – wykres osypiska
Źródło: opracowanie własne.

(zmienne o mocy dyskryminacyjnej to zmienne o nr 7 i 8). Dokładny opis metody *HINoV* i jej zastosowania w programie **R** można znaleźć w pracy [Walesiak 2005].

Aby dokonać analizy porównawczej obu metod selekcji zmiennych, za pomocą pakietów `clusterSim` (funkcja `cluster.Gen`) oraz `stats` (funkcje `norm` i `mvnorm`) wygenerowano kilkanaście zbiorów o 1-7 zmiennych o mocy dyskryminacyjnej, 1-7 zmiennych zakłócających. Analizowano zbiory 200-elementowe o liczbie klas równej 2-5. Wyniki badań, których rezultaty ze względu na ograniczenia objętościowe artykułu nie zostały zamieszczone, wskazują na to, że obie metody dokonują prawidłowej selekcji zmiennych w przypadku, gdy zmienne sztuczne wygenerowano z rozkładu jednostajnego lub normalnego o różnych parametrach. W przeciwieństwie do metody wykorzystującej strategię zachłanną heurystyczna metoda *HINoV* zawodzi, gdy przynajmniej 2 zmienne zakłócające generowane są z rozkładu normalnego o tych samych parametrach. Metoda selekcji oparta na modelach daje również lepsze rezultaty w przypadku, gdy w zbiorze jest tylko jedna zmienna z określoną strukturą klas, a inne są zmiennymi zakłócającymi (metoda *HINoV* zawodzi w tej sytuacji).

6. Podsumowanie

Metoda selekcji zmiennych wykorzystująca strategię zachłanną w analizie skupień opartej na modelach pozwala nie tylko na wybór zmiennych o największej mocy dyskryminacyjnej, ale również na wybór optymalnej liczby klas. W porównaniu z heurystyczną metodą *HINoV* metoda ta dała wyniki lepsze na analizowanych zbiorach

sztucznych. Główną zaletą metody jest to, że wybór każdej ze zmiennych i liczby klas opiera się na wnioskowaniu statystycznym (liczba klas w metodzie *HINoV* ustalana jest w sposób arbitralny). Ograniczeniem zaprezentowanej metody jest niemożność dokonywania selekcji dla zmiennych niemetrycznych (nominalnych i porządkowych).

Literatura

Carmone F.J., Kara A., Maxwell S. (1999), *HINoV: a new method to improve market segment definition by identifying noisy variables*, „Journal of Marketing Research", November, no 36, s. 501-509.

Fraley C., Raftery A.E. (2002), *Model-based clustering, discriminant analysis, and density estimation*, „Journal of the American Statistical Association" no 97, s. 611-631.

Hubert L.J., Arabie P. (1985), *Comparing partitions*, „Journal of Classification" no 1, s. 193-218.

Kass R.E., Raftery A.E. (1995), *Bayes factors*, „Journal of the American Statistical Association" no 90, s. 773-795.

Raftery A.E., Dean N. (2006), *Variable selection for model-based clustering*, „Journal of the American Statistical Association" no 101, s. 168-178.

Schwarz G. (1978), *Estimating the dimension of a model*, „The Annals of Statistics" no 6, s. 461-464.

Walesiak M. (2005), *Wybór zmiennych w zagadnieniu klasyfikacji – podejścia, problemy, metody*, Plenarne posiedzenie Komitetu Statystyki i Ekonometrii PAN, 15 marca, Wrocław.

Witek E. (2008), *Metoda taksonomii oparta na modelach mieszanych*, [w:] Taksonomia 15, red. K. Jajuga, M. Walesiak, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 1207, UE, Wrocław, s. 199-206.

VARIABLE SELECTION FOR MODEL-BASED CLUSTERING

Summary

We propose a variable selection for model-based clustering using greedy search algorithm. At each stage it searches for the variable to add that most improves the clustering as measured by BIC. As a result the best combination of number of groups and clustering model is chosen (two nested subsets of variables are recast as a model comparison problem). We compare variable selection using greedy search algorithm with a *HINoV* method of variable selection.