PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Ewa Witek
Akademia Ekonomiczna w Katowicach

METODA TAKSONOMII OPARTA NA MODELACH MIESZANYCH

1. Wstęp

McLachlan i Basford [1988] wykorzystali mieszankę rozkładów (kombinację wypukłą danego zbioru dystrybuant) do konstrukcji modeli mieszanych (*mixture models*). W modelach tych zakłada się, że obserwacje pochodzą z kilku populacji „mieszanych" w różnych proporcjach. Modele te znalazły swoje zastosowanie m.in. na gruncie taksonomii. Każda z populacji jest tu odpowiednikiem klasy lub sztucznej zmiennej modelu, których wagi wyrażają prawdopodobieństwo przynależności obiektów do poszczególnych klas. Każda z klas (sztucznych zmiennych modelu) posiada funkcję gęstości o nieznanych parametrach. W praktyce najczęściej wykorzystywane są funkcje gęstości rozkładu normalnego. Celem taksonomii jest więc nie tylko rozpoznanie liczby i struktury klas, ale także oszacowanie parametrów rozkładu każdej z tych klas.

Każdą kombinację liczby klas w poszczególnych modelach mieszanych Gaussa możemy przypisać innemu modelowi statystycznemu. Problem wyboru najlepszej z metod klasyfikacji (najlepszego modelu Gaussa) i optymalnej liczby klas sprowadza się do wyboru modelu statystycznego o najlepszej jakości dopasowania.

W artykule przedstawiona zostanie metoda taksonomii oparta na modelach (MBC – *model-based clustering),* wykorzystująca połączenie hierarchicznej metody aglomeracyjnej opartej na modelach (*model-based hierarchical clustering*), algorytmu EM oraz statystyki BIC. Hierarchiczne metody aglomeracyjne oparte na modelach dokonują podziału zbioru obiektów na klasy. Podział ten ma na celu zainicjowanie algorytmu EM, który daje lepsze wyniki wtedy, gdy znamy wartości startowe, kryterium informacyjne BIC zaś pozwala na wybór modelu o największej jakości dopasowania.

2. Model mieszany

Zakłada się, że każda wielowymiarowa obserwacja $[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$ jest realizacją zmiennej losowej o gęstości $f(\mathbf{y}_i|\Theta)$, zaś zmienne sztuczne modelu mieszanego identyfikowane są z poszczególnymi klasami, z których każda posiada swą własną funkcję gęstości.

$$\mathbf{y}_i \sim \sum_{k=1}^{G} \tau_k f_k(\mathbf{y}_i|\Theta_k), \tag{1}$$

gdzie: f_k – gęstość,
Θ_k – parametry k-tej klasy,
τ_k – wartość prawdopodobieństwa, że dana obserwacja należy do k-tej klasy
$(\tau_k \geq 0, \ \sum_{k=1}^{G} \tau_k = 1)$.

Problem polega na takim doborze parametrów Θ_k i podziale obiektów na klasy, by maksymalizować funkcję wiarygodności:

$$\alpha_{MIX} = (\Theta_1, \ldots\ldots, \Theta_G|\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{y}_i|\Theta_k). \tag{2}$$

Najczęściej za f_k przyjmujemy funkcję gęstości wielowymiarowego rozkładu normalnego o średniej μ_k, wyznaczającej środek klasy k i macierzy kowariancji $\mathbf{\Sigma}_k$ określonej wzorem:

$$\Phi_k(\mathbf{y}_i|\mu_k, \mathbf{\Sigma}_k) = \frac{\exp\{-\frac{1}{2}(\mathbf{y}_i - \mu_k)^T \sum_k^{-1} (\mathbf{y}_i - \mu_k)\}}{\sqrt{\det(2\pi \sum_k)}}. \tag{3}$$

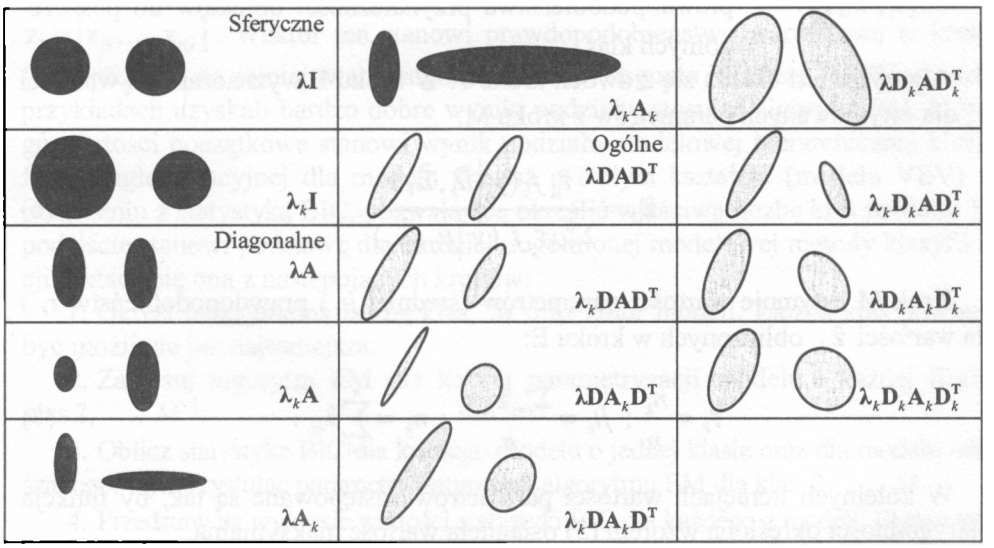
Banfield i Raftery [1993] przedstawili następujący zapis macierzy kowariancji $\mathbf{\Sigma}_k$:

$$\sum_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{\mathrm{T}}, \tag{4}$$

gdzie: $\mathbf{D}_k$ to macierz ortogonalna wartości własnych, $\mathbf{A}_k$ to macierz diagonalna, taka że $|\mathbf{A}_k| = 1$ i której elementy diagonalne to uporządkowane malejąco wartości własne macierzy $\mathbf{\Sigma}_k$, $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$. λ_k, $\mathbf{A}_k$, $\mathbf{D}_k$ traktowane są jako zbiory niezależnych parametrów, których wartości mogą być stałe lub różne dla poszczególnych klas. Gdy wartości parametrów poszczególnych macierzy przyjmują wartość stałą, klasy posiadają te same cechy geometryczne. Wartości parametrów macierzy $\mathbf{A}_k$ wyznaczają kształt, a wartości parametrów macierzy λ_k określają objętość klasy k. Klasy modelu mieszanego mogą więc posiadać te same lub różne kształty i objętości. Wartości parametrów macierzy $\mathbf{D}_k$ deter-

minują orientację klas modelu. Wyróżniamy trzy rodziny modeli: sferyczne, diagonalne i ogólne. W przypadku modeli modeli sferycznych (EII, VII) mówimy o braku orientacji klas. Klasy modeli diagonalnych (EEI, VEI, EVI, VVI) cechuje ta sama lub różna orientacja względem osi układu współrzędnych. Klasy modeli ogólnych (EEE, EEV, VEV, VVV) przyjmują tę samą lub różną orientację względem siebie. Klasy modeli sferycznych posiadają kształt sferyczny, a klasy modeli ogólnych kształt elipsoidalny. Stosowna parametryzacja macierzy kowariancji $\mathbf{\Sigma}_k$ określonej wzorem (4) pozwala zachować wzajemną spójność klas i umożliwia ich geometryczną interpretację.

Na rys. 1 przedstawiono różne możliwości parametryzacji modeli zawarte w pakiecie mclust programu R.



Rys. 1. Parametryzacja macierzy kowariancji $\mathbf{\Sigma}_k$ dostępna w pakiecie Mclust programu statystycznego R
Źródło: opracowanie własne na podstawie pracy [Fraley, Raftery 2006].

W pakiecie mclust programu R dostępnych jest 10 modeli mieszanych. Fraley i Raftery [2006, s. 7] zaproponowali następujące oznaczenia dla poszczególnych modeli: EII = $\lambda\mathbf{I}$, VII = $\lambda_k\mathbf{I}$, EEI = $\lambda\mathbf{A}$, VEI = $\lambda_k\mathbf{A}$, EVI = $\lambda\mathbf{A}_k$, VVI = $\lambda_k\mathbf{A}_k$, EEE = $\lambda\mathbf{DAD}^T$, EEV = $\lambda\mathbf{D}_k\mathbf{AD}_k^T$, VEV = $\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$, VVV = $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$. Wyróżnili oni dodatkowo dwa modele dla zmiennych jednowymiarowych – model E (dla klas o stałej objętości) i model V (dla klas o różnej objętości).

3. Algorytm EM

Algorytm EM jest skuteczną metodą iteracyjnego wyznaczania estymatorów największej wiarygodności parametrów modelu zawierającego zmienne nieobser-

wowalne [Dempster i in. 1977]. Zakładając, że dane kompletne modelu mieszanego określone są jako $\mathbf{x}_i = (\mathbf{y}_i, \mathbf{z}_i)$, gdzie $\mathbf{y}_i$ to dane obserwowalne, a $\mathbf{z}_i$ dane nieobserwowalne, *G*- wymiarowa wektorowa zmienna określająca przynależność do klas $\mathbf{z}_i = [\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iG}]$ przyjmuje wartość 1 (gdy $\mathbf{x}_i$ należy do *k*-tej klasy) lub 0 (w przeciwnym wypadku). Funkcja wiarygodności przybiera wtedy następującą postać:

$$l(\Theta_k, \tau_k, \mathbf{z}_{ik} | x = \sum_{i=1}^{n} \sum_{k=1}^{G} \mathbf{z}_{ik} \log[\tau_k f_k(\mathbf{y}_i | \Theta_k)], \tag{5}$$

gdzie: $f_k(\mathbf{y}_i | \Theta_k)$ – funkcja gęstości zmiennej $\mathbf{y}$,

$(\tau_1, \ldots, \tau_G)$ – prawdopodobieństwa przynależności obiektów do poszczególnych klas.

Algorytm EM składa się z dwóch kroków. W kroku E wyznaczane są wartości $\hat{\mathbf{z}}_{ik}$ dla estymowanych parametrów z kroku M:

$$\hat{\mathbf{z}}_{ik} = \frac{\hat{\tau}_{ik} f_k(y_i | \hat{\mu}_k, \hat{\Sigma}_k)}{\Sigma_{j=1}^{G} \hat{\tau}_j f_j(y_i | \hat{\mu}_j, \hat{\Sigma}_j)}. \tag{6}$$

Krok M estymuje wartości parametrów (średniej μ_k i prawdopodobieństw τ_k) dla wartości $\hat{\mathbf{z}}_{ik}$ obliczonych w kroku E:

$$\hat{\tau}_k = \frac{n_k}{n}; \ \hat{\mu}_k = \frac{\Sigma_{i=1}^{n} \hat{\mathbf{z}}_{ik} \mathbf{y}_i}{n_k}; \ n_k = \sum_{i=1}^{n} \hat{\mathbf{z}}_{ik}. \tag{7}$$

W kolejnych iteracjach wartości parametrów zastępowane są tak, by funkcja wiarygodności określona wzorem (5) osiągnęła wartość maksymalną.

4. Wybór modelu

W przypadku modeli szacowanych metodą największej wiarygodności statystyką pomiaru jakości dopasowania są kryteria informacyjne. Najbardziej popularnym kryterium, obok kryterium informacyjnego Akaike **AIC** (*Akaike information criterion*), jest bayesowskie kryterium informacyjne Schwarza **BIC** (*Bayesian information criterion*) dane wzorem:

$$BIC_k = 2 \log p(x | \hat{\Theta}_k, M_k) - v_k \log(n), \tag{8}$$

gdzie: $\log p(x | \hat{\Theta}_k, M_k)$ – logarytm funkcji wiarygodności dla oszacowanego wektora parametrów modelu M_k,

v_k – liczba parametrów modelu,

n – liczba obserwacji [Schwarz 1978, s. 461-464].

Pierwsza część równania (8) odpowiada za wybór modeli o najwyższej dobroci dopasowania, część druga zaś odrzuca modele z nadmierną liczbą parametrów. Statystyka BIC stosowana jest w celu porównania modeli o różnej parametryzacji oraz modeli o różnej liczbie klas. Im wyższa wartość statystyki BIC, tym lepsza jakość dopasowania danego modelu.

5. Metoda klasyfikacji oparta na modelach (*model-based strategy for clustering*)

Podział zbioru obiektów uzyskany za pomocą hierarchicznej klasyfikacji aglomeracyjnej można przekształcić w wektor przynależności do klas (selektor) $\mathbf{z}_i = [\mathbf{z}_{i1},\ldots,\mathbf{z}_{iG}]$. Wektor ten stanowi prawdopodobieństwo warunkowe w kroku 'M' i pozwala na zainicjowanie algorytmu EM. Dasgupta i Raftery [1998] w wielu przykładach uzyskali bardzo dobre wyniki podziału, stosując algorytm EM, którego wartości początkowe stanowi wynik podziału modelowej hierarchicznej klasyfikacji aglomeracyjnej dla modelu Gaussa o stałym kształcie (modelu VEV) w połączeniu z statystyką BIC, pozwalającą określić właściwą liczbę klas modelu. To podejście stanowi podstawę dla bardziej uogólnionej modelowej metody klasyfikacji. Składa się ona z następujących kroków:

1. Określ maksymalną liczbę klas, M oraz zbiór modeli. Liczba klas powinna być możliwie jak najmniejsza.
2. Zastosuj algorytm EM dla każdej parametryzacji modelu i każdej liczby klas 2, ..., M [1].
3. Oblicz statystykę BIC dla każdego modelu o jednej klasie oraz dla modelu mieszanego, wykorzystując parametry optymalne algorytmu EM dla klas 2, ..., M.
4. Przedstaw na wykresie wartości statystyki BIC dla każdego z modeli. Ekstremum lokalne pozwala wybrać właściwy model (jego parametryzację i liczbę klas).

6. Przykład empiryczny

Celem przeprowadzonego badania jest zastosowanie prezentowanej w artykule metody klasyfikacyjnej. Metoda klasyfikacji oparta na modelach mieszanych zostanie porównana z powszechnie stosowanymi metodami klasyfikacji, tj. metodą hierarchiczną oraz metodą *k*-średnich. W analizowanym przykładzie wykorzystano zbiór wyników diagnozy cukrzycy (*diabetes diagnosis*) [Reaven, Miller 1979] złożony z 145 obserwacji i 3 zmiennych objaśniających. Rozważane zmienne to: glukoza, insulina i sspg (stopień

[1] W metodzie taksonomii opartej na modelach mieszanek rozkładów normalnych za wartości początkowe algorytmu przyjmuje się wynik podziału hierarchicznej klasyfikacji aglomeracyjnej opartej na modelach. Hierarchiczną klasyfikację aglomeracyjną przeprowadza się dla nieograniczonego modelu Gaussa (modelu VVV) i liczby klas 2, ..., *M*. Model VVV odpowiada każdej parametryzacji modelu Gaussa, dlatego wykorzystywany jest do zainicjowania algorytmu EM.

odporności na insulinę). Założenia i treść przykładu zostały zaczerpnięte z pracy Fraleya i Raftery'ego [1998]. Zbiór dostępny jest w pakiecie mclust programu R.

Mimo że dane zawierają poprawnie sklasyfikowany zbiór obiektów, w badaniu pominięto znaną przynależność obiektów do poszczególnych klas. Informacja ta została jedynie wykorzystana w celach porównawczych poszczególnych metod klasyfikacji do obliczenia miar jakości klasyfikacji, tj. błędu klasyfikacji (e_v) i skorygowanej miary Randa (R_{HA}) [Hubert, Arabie 1985, s. 198].

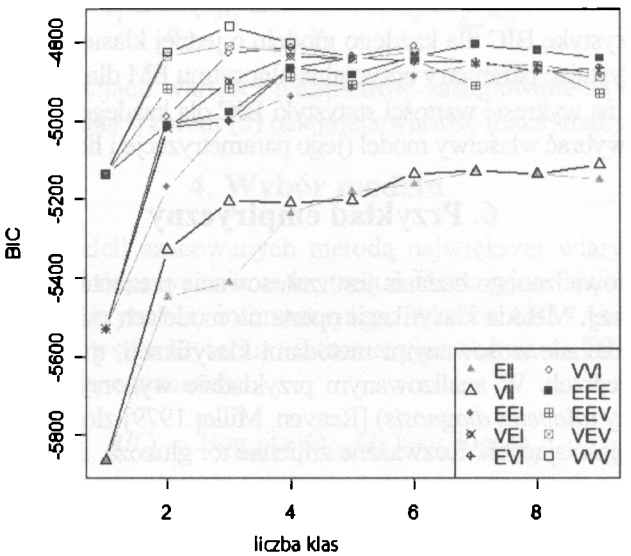
Dla omawianego zbioru dokonano klasyfikacji za pomocą hierarchicznej metody aglomeracyjnej (metody najbliższego sąsiada, metody Warda) oraz metody *k*-średnich (dla liczby klas $k = 3$ przyjętej *a priori*). W analizie skupień opartej na modelach liczba klas $k = 3$ została wyznaczona za pomocą statystyki BIC.

Tabela 1. Wynik klasyfikacji dla 3 klas uzyskany po zastosowaniu metody najbliższego sąsiada, Warda, *k*-średnich oraz metody taksonomii opartej na modelach (MBC)

Metoda klasyfikacji	Liczba obserwacji w klasach			L	R_{HA}	e_v
	klasa 1	klasa 2	klasa 3			
Najbliższego sąsiada	143	1	1	68	1,29%	46,20%
Warda	119	14	12	43	31,8%	29,65%
k-średnich	115	13	17	40	37,65%	27,59%
MBC	82	33	30	17	71,93%	11,72%

L – liczba obserwacji błędnie sklasyfikowanych, R_{HA} – skorygowana miara Randa, e_v – błąd klasyfikacji.

Źródło: obliczenia własne.



Rys. 2. Kryterium informacyjne BIC dla taksonomii opartej na modelach mieszanych dla zbioru wyników diagnozy cukrzycy (*diabetes diagnosis*)

Źródło: obliczenia własne.

Analiza tab. 1 pozwala na stwierdzenie, że zarówno metoda najbliższego sąsiada, metoda Warda, jak i metoda k-średnich nie dają satysfakcjonujących wyników podziału. W przypadku metody najbliższego sąsiada dwie z rozpoznanych klas zawierają obserwacje pojedyncze, więc prawie wszystkie obserwacje zostały przypisane do jednej klasy. Dla metody k-średnich obserwujemy zjawisko braku rozłączności klas. Ponadto oceny podobieństwa wyników klasyfikacji dla tych podziałów przyjmują bardzo niskie wartości. Z kolei dla metody taksonomii opartej na modelach otrzymano czterokrotnie niższą liczbę błędnie sklasyfikowanych obserwacji aniżeli w metodzie najbliższego sąsiada. Skorygowana miara Randa R_{HA} i błąd klasyfikacji dla tej metody dały zdecydowanie najlepsze wyniki.

Postępując zgodnie z zasadami metody klasyfikacyjnej opartej na modelach mieszanych zawartymi w części piątej artykułu, wartości statystyki BIC dla poszczególnych modeli przestawiono na rys. 2.

Pierwsze maksimum lokalne (a zarazem globalne) zostało osiągnięte dla nieograniczonego modelu Gaussa (modelu VVV) dla liczby klas $k = 3$. Model VVV, dla którego jakość dopasowania modelu wyrażona jest za pomocą statystyki BIC, dla trzech klas daje więc najlepszy możliwy podział analizowanego zbioru obserwacji. Należy również zauważyć, że na wykresie brakuje wartości BIC dla modelu VVV o liczbie klas $k > 4$. Wynikiem podziału modelowej hierarchicznej klasyfikacji aglomeracyjnej modelu VVV dla liczby klas $k > 4$ są skupienia o liczbie elementów równej 3. Ponieważ analizowane dane są trójwymiarowe, minimalna liczba obserwacji w każdej z klas wynosi 4. W przeciwnym wypadku wyznacznik macierzy kowariancji Σ_k będący mianownikiem we wzorze (3) wynosi 0, co powoduje, że nie można oszacować parametrów największej wiarygodności i zarazem obliczyć wartości statystyki BIC.

7. Podsumowanie

Przedstawione wyniki badania wykazują, że metoda klasyfikacji oparta na wielowymiarowych normalnych modelach mieszanych daje lepsze efekty aniżeli powszechnie stosowane metody w przypadku, gdy w zbiorze występują klasy o różnym rozmiarze oraz różnym kształcie. Zastosowana metoda pozwoliła na wybór odpowiedniego modelu (modelu o najlepszej parametryzacji), a zarazem stosownej liczby klas, której dobór dla nieparametrycznych metod klasyfikacji nadal nastręcza wiele trudności.

Literatura

Banfield J.D., Raftery A.E. (1993), *Model-based Gaussian and Non-gaussian Clustering*, „Biometrics", 49, s. 803-821.

Dasgupta A., Raftery A.E. (1998), *Detecting Features in Spatial Point Processes with Clutter via Model-based Clustering*, „Journal of the American Statistical Association", 93, s. 294-302.
Dempster A.P., Laird N.M., Rubin D.B. (1977), *Maximum Likelihood for Incomplete Data via the EM Algorithm (with Discussion)*, „Journal of the Royal Statistical Society", Ser. B, 39, s. 1-38.
Fraley C., Raftery A.E. (1998), *How Many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis*, „The Computer Journal", 41, s. 577-588.
Fraley C., Raftery A.E. (2002), *Model-based Clustering, Discriminant Analysis, and Density Estimation*, „Journal of the American Statistical Association", 97, s. 611-631.
Fraley C., Raftery A.E. (2006), *MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-based Clustering*, s. 1-50.
Hubert L.J., Arabie P. (1985), *Comparing Partitions*, „Journal of Classification", 1, s. 193-218.
McLachlan G.J., Basford K.E. (1988), *Mixture Models: Inference and Applications to Clustering*, G.J. Marcel Dekker, New York.
McLachlan G.J., Krishnan, T. (1997), *The EM Algorithm and Extensions*, Wiley, New York.
Reaven G.M., Miller R.G. (1979), *An Attempt to Define the Nature of Chemical Diabetes Using Multidimensional Analysis*, „Diabetologica", 16, s. 17-27.
Schwarz G. (1978), *Estimating the Dimension of a Model*, „The Annals of Statistics", 6, s. 461-464.

MODEL-BASED CLUSTERING

Summary

In model-based clustering approach, the data are viewed as coming from a mixture of probability distributions, each representing a different cluster. Models with varying geometric properties are obtained through Gaussian components with different parameterizations and cross-cluster constraints. Partitions are determined by the EM algorithm for maximum likelihood, with initial values from agglomerative hierarchical clustering. Models are compared on the basis of Bayesian Information Criterion (BIC). The problems of determining the number of clusters and the clustering method are at the same time solved by choosing the best model.