Eugeniusz Gatnar
Uniwersytet Ekonomiczny w Katowicach

ANALIZA DYSKRYMINACYJNA – STAN AKTUALNY I KIERUNKI ROZWOJU

Wprowadzenie

W literaturze statystycznej pojęcie analizy dyskryminacyjnej jako zadania polegającego na znalezieniu charakterystyki klas pojawiło się już w latach 30. XX w., gdy Fisher (1936) próbował w tym celu zastosować liniowe modele regresji. Prowadził on eksperymenty na najbardziej znanym w wielowymiarowej analizie statystycznej zbiorze danych, zawierającym charakterystykę trzech gatunków kosaćca (IRYS) za pomocą czterech zmiennych.

W Polsce jednym z pierwszych badaczy zajmującym się metodami analizy dyskryminacyjnej był prof. Józef Kolonko, który w swojej pracy (Kolonko, 1980) stosował pojęcie analizy dyskryminacyjnej w nieco innym, znacznie szerszym znaczeniu. W przedstawionym przez niego ujęciu, związanym z podejściem cybernetycznym obecnym w latach 70. XX w. w naukowej literaturze rosyjskiej, dyskryminacja zbioru obserwacji to jego podział („rozbicie"), który optymalizuje wartość pewnego funkcjonału stanowiącego kryterium. Analiza dyskryminacyjna dzieli się wobec tego na klasyfikację z wzorcem oraz klasyfikację bez wzorca (taksonomię).

Obecnie w polskiej literaturze statystycznej panuje zgodny pogląd, że analizą dyskryminacyjną jest zbiór metod prowadzących do znalezienia reguły klasyfikacyjnej, charakterystyki klas lub funkcji rozdzielających klasy, na podstawie zbioru uczącego, tj. zawierającego obiekty o znanej przynależności do klas. Jest to więc jedynie klasyfikacja z nauczycielem.

W niniejszym artykule, który ma charakter przeglądowy, zagadnienie dyskryminacji zostało zdefiniowane jako problem znalezienia charakterystyki klas, poprzez identyfikację sposobów ich odseparowania. W pierwszym punkcie zostanie sformułowane ogólne zadanie dyskryminacji, natomiast w drugim – klasyczne podejście zaproponowane przez Fishera, które prowadzi do powstania liniowych funkcji dyskryminacyjnych. Część trzecia zawiera omówienie założeń, których spełnienie pozwala na zastosowanie modeli liniowych. W przeciwnym przypadku można stosować kwadratowe funkcje dyskryminacyjne. W czwartym punkcie artykułu

przedstawiono najbardziej popularną nieparametryczną metodę dyskryminacji wykorzystującą drzewa klasyfikacyjne. Na końcu pracy znajduje się omówienie aktualnych kierunków rozwoju metod dyskryminacji.

1. Zadanie dyskryminacji

Dyskryminacja to decyzja o przydzieleniu obiektu do klasy, która jest dokonywana na podstawie znajomości rozkładów zmiennych w klasach oraz prawdopodobieństw *a priori*. Jeżeli zbiór $\{C_1, \ldots, C_J\}$ zawiera wartości zmiennej objaśnianej Y (nazwy klas), to prawdopodobieństwo *a priori* dla klasy C_j ($j = 1, \ldots, J$) to $p(C_j)$.

Ważne jest także założenie o losowości wektora zmiennych $\mathbf{X} = (X_1, \ldots, X_L)$ i o tym, że jego warunkowe rozkłady prawdopodobieństw w klasach $g(\mathbf{X} \mid C_j)$ są znane. Jeżeli oba założenia są spełnione, to można wyznaczyć prawdopodobieństwo, że klasyfikowana obserwacja $[\mathbf{x}_i, y_i]$ należy do klasy C_j, na podstawie wzoru Bayesa:

$$p(C_j \mid \mathbf{x}_i) = \frac{g(\mathbf{x}_i \mid C_j)\, p(C_j)}{\sum_{k=1}^{J} g(\mathbf{x}_i \mid C_k)\, p(C_k)}. \tag{1}$$

Prawdopodobieństwo $p(C_j \mid \mathbf{x}_i)$ jest nazywane prawdopodobieństwem *a posteriori*, gdyż decyzja o przydzieleniu obserwacji $[\mathbf{x}_i, y_i]$ do klasy C_j podejmowana jest już po jej zaobserwowaniu. Mianownik (1) nie wpływa na wartość $p(C_j \mid \mathbf{x}_i)$, zatem zależy ona jedynie od rozkładu warunkowego $g(\mathbf{x}_i \mid C_j)$ ważonego prawdopodobieństwami *a priori*.

Model dyskryminacyjny (1) może błędnie zaklasyfikować obserwację $[\mathbf{x}_i, y_i]$, z czym związana jest pewna strata, którą wyraża przyjęta arbitralnie funkcja straty (*loss function*) $L(Y, f(\mathbf{X}))$, najczęściej funkcja zero-jedynkowa:

$$L\big(Y, f(\mathbf{X})\big) = \begin{cases} 1 & gdy \quad Y \neq f(\mathbf{X}) \\ 0 & gdy \quad Y = f(\mathbf{X}) \end{cases} \tag{2}$$

lub entropia krzyżowa (*cross entropy*):

$$L\big(Y, f(\mathbf{X})\big) = -2\sum_{j=1}^{J} (Y = C_j) \log p(C_j \mid \mathbf{X}). \tag{3}$$

Błąd klasyfikacji to spodziewana wielkość funkcji straty $e(f) = E\left[L(Y, f(\mathbf{X})\right]$.

Jeżeli obserwacja $[\mathbf{x}_i, y_i]$ została błędnie zaklasyfikowana, to wartość błędu związanego z taką decyzją wynosi:

$$e(\mathbf{x}_i) = \sum_{j=1}^{J} L(y_i, f(\mathbf{x}_i)) p(C_j|\mathbf{x}_i). \tag{4}$$

Wyrażenie (4) dla funkcji straty (2) osiąga minimum, gdy:

$$\hat{f}(\mathbf{x}_i) = \arg\min_{j=1,\ldots,J} \left\{1 - p(C_j|\mathbf{x}_i)\right\}. \tag{5}$$

Oznacza to, innymi słowy, że błąd predykcji (4) jest najmniejszy, gdy obserwacja $[\mathbf{x}_i, y_i]$ zostaje przydzielona do klasy, dla której prawdopodobieństwo *a posteriori* jest największe:

$$\hat{f}(\mathbf{x}_i) = C_j \;\; gdy \;\; p(C_j|\mathbf{x}_i) = \max_{k=1,\ldots,J} \left\{p(C_k|\mathbf{x}_i)\right\}. \tag{6}$$

Reguła (6) nazywana jest bayesowską regułą klasyfikacji (*Bayes rule*), a model (5) to optymalny klasyfikator bayesowski (*Bayes classifier*). Z kolei jego błąd klasyfikacji $e^{Bayes} = 1 - \max_j \left\{p(C_j|\mathbf{x}_i) \cdot p(C_j)\right\}$ nazywany jest błędem bayesowskim (*Bayes error*).

Bayesowska reguła klasyfikacji jest prosta, lecz wymaga znajomości rozkładów warunkowych w klasach $g(\mathbf{X}|C_j)$ oraz prawdopodobieństw *a priori*. W praktyce rozkłady te są estymowane na podstawie próby uczącej. Na przykład, najczęściej stosuje się frakcje obiektów należących do klasy C_j, tj. $p(C_j) = n_j / N$, gdzie $\sum_{j=1}^{J} p(C_j) = 1$. Można też przyjąć, że są równe, tj. $p(C_1) = \ldots = p(C_J)$, oszacować na podstawie zbioru testowego lub ustalić subiektywnie.

Zgodnie z regułą Bayesa, aby poprawnie sklasyfikować obserwację $[\mathbf{x}_i, y_i]$, należy znaleźć maksimum wyrażenia $g(\mathbf{x}_i|C_j)p(C_j)$. Wyniki klasyfikacji jednak nie zmienią się, jeśli na funkcję dyskryminacyjną nałożymy odwzorowanie monotonicznie rosnące, otrzymując np. funkcje: $f_j(\mathbf{X}) = g(\mathbf{X}|C_j)$, $f_j(\mathbf{X}) = g(\mathbf{X}|C_j)p(C_j)$ lub $f_j(\mathbf{X}) = \ln(g(\mathbf{X}|C_j)) + \ln(p(C_j))$. W tym ostatnim przypadku szukamy klasy C_j, dla której funkcja:

$$f_j(\mathbf{X}) = -\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_j) + \ln(p(C_j)) \tag{7}$$

osiąga maksimum, gdzie $\boldsymbol{\mu}_j$ jest środkiem ciężkości klasy C_j, zaś $\boldsymbol{\Sigma}$ macierzą wariancji i kowariancji. Warto zauważyć, że funkcja dyskryminacyjna (7) jest funkcją liniową.

W podobny sposób można skonstruować funkcje dyskryminacyjne, które separują poszczególne klasy. Biorąc pod uwagę np. klasy C_j oraz C_k, można zbudować funkcję:

$$f_{jk}(\mathbf{X}) = \ln\left(\frac{p(C_j \mid \mathbf{X})}{p(C_k \mid \mathbf{X})}\right) + \ln\left(\frac{p(C_j)}{p(C_k)}\right), \tag{8}$$

która wyznacza równanie hiperpłaszczyzny:

$$\ln\left(\frac{p(C_j)}{p(C_k)}\right) - \frac{1}{2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_k) + (\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}^{-1}\mathbf{X} = 0\,. \tag{9}$$

Widać tutaj związek z liniowymi funkcjami dyskryminacyjnymi, zaproponowanymi przez Fishera, które zostaną omówione w następnej części artykułu.

W literaturze mowa jest także o naiwnym klasyfikatorze bayesowskim (*naive Bayes*), który opiera się na założeniu, że w klasie C_j zmienne $X_1,...,X_L$ są niezależne:

$$f_j(\mathbf{X}) = \prod_{l=1}^{L} f_{jl}(X_l)\,, \tag{10}$$

gdzie funkcje gęstości f_{lj} są przybliżane za pomocą rozkładu normalnego. Uzyskane za jego pomocą wyniki klasyfikacji są często bardziej dokładne niż w przypadku innych modeli, ponieważ obciążenie estymatorów funkcji gęstości w klasach nie przekłada się na obciążenie prawdopodobieństw *a posteriori*.

2. Liniowe modele dyskryminacyjne Fishera

Jak już wspomniano, zagadnienie dyskryminacji zostało po raz pierwszy sformułowane przez Fishera (1936), a zaproponowana przez niego metoda opiera się na redukcji wymiaru przestrzeni cech. Pierwotna przestrzeń $\mathbf{X}^L$ zostaje zredukowana do przestrzeni zmiennych kanonicznych (*canonical variables*) $\mathbf{Z}^{J-1} = Z_1 \times ... \times Z_{J-1}$.

Należy podkreślić, że wywodzące się od Fishera pojęcie liniowej analizy dyskryminacyjnej (Linear Discriminant Analysis – LDA) używane jest często w węższym znaczeniu i określa jedynie jego podejście. Fisher rozwiązał pro-

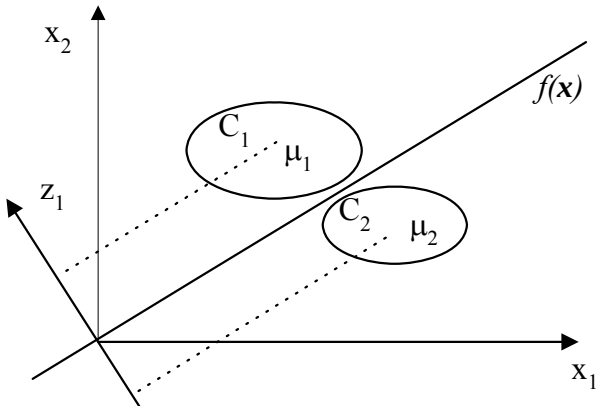
blem dla dwóch klas ($J = 2$), zaś uogólnienie dla $J > 2$ klas podali Rao (1948) i Bryan (1951).

Zmiennymi kanonicznymi są:

$$Z_j = \sum_{l=1}^{L} a_{jl} X_l , \tag{11}$$

które powodują, że w tej nowej przestrzeni klasy są optymalnie odseparowane, tj. ich środki ciężkości leżą jak najdalej od siebie.

Na rys. 1 pokazano dwie klasy C_1, C_2 w dwuwymiarowej przestrzeni cech X_1, X_2 oraz zmienną kanoniczną Z_1. Prosta $f(\mathbf{X})$, która je oddziela, jest ortogonalna względem Z_1.



Rys. 1. Separacja klas w przestrzeni dwóch zmiennych

Źródło: Gatnar (2008).

Poszukiwane jest więc maksimum funkcji:

$$Q = \frac{\left|\mathbf{a}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\right|^2}{\mathbf{a}\mathbf{W}\mathbf{a}^T}, \tag{12}$$

gdzie licznik jest kwadratem odległości środków ciężkości klas wzdłuż kierunku $\boldsymbol{a}$, zaś $\mathbf{W}$ jest macierzą wariancji i kowariancji wewnątrzklasowej:

$$\mathbf{W} = \frac{1}{N-J} \sum_{j=1}^{J} \sum_{i=1}^{N_j} (\mathbf{x}_i - \boldsymbol{\mu}_j)^2 . \tag{13}$$

Maksimum (12) jest osiągane dla wektora $\hat{\mathbf{a}}$ o kierunku $\mathbf{W}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Hiperpłaszczyzna separująca obie klasy jest prostopadła do kierunku $\hat{\mathbf{a}}$ i przechodzi przez środek odcinka łączącego środki ciężkości klas. Ma ona postać:

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathbf{W}^{-1}\left(\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) = 0\,. \tag{14}$$

Można pokazać, że licznik funkcji (12) jest równy $\mathbf{aBa}^T$, a zatem kryterium to sprowadza się do znalezienia takiego wektora $\hat{\mathbf{a}}$, aby iloraz wariancji:

$$F = \frac{\mathbf{aBa}^T}{\mathbf{aWa}^T} \tag{15}$$

był jak największy, gdzie $\mathbf{B}$ jest macierzą wariancji i kowariancji międzyklasowej, wyznaczoną na podstawie próby:

$$\mathbf{B} = \frac{1}{J-1}\sum_{j=1}^{J} N_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T\,. \tag{16}$$

Wektor $\hat{\mathbf{a}}$ maksymalizujący kryterium (14) jest wektorem własnym macierzy $\mathbf{W}^{-1}\mathbf{B}$ odpowiadającym największej wartości własnej tej macierzy.

Zmienne kanoniczne są konstruowane w sposób sekwencyjny. Wektor $\hat{\mathbf{a}}$ wyznacza kierunek, wzdłuż którego klasy są najlepiej odseparowane, a kombinacja liniowa $\mathbf{Z}_1 = \hat{\mathbf{a}}_1\mathbf{X}^T$ jest pierwszą zmienną kanoniczną. Jeśli jakość klasyfikacji nie jest zadowalająca, można wyznaczyć kolejne wektory $\hat{\mathbf{a}}_2, \ldots, \hat{\mathbf{a}}_{J-1}$, które spełniają warunek wzajemnej ortogonalności i są wektorami własnymi macierzy $\mathbf{W}^{-1}\mathbf{B}$, odpowiadającymi uporządkowanym malejąco wartościom własnym.

Drugie podejście związane jest z metodą preceptronową Rosenblatta (1958), która ma charakter iteracyjny. Metoda ta minimalizuje wartość kryterium:

$$Q(\mathbf{a}) = -\sum_{\mathbf{x}_i \in E} \mathbf{a}\mathbf{x}_i^T\,, \tag{17}$$

gdzie $E = \{\mathbf{x}_i : \mathbf{a}\mathbf{x}_i^T < 0\}$ jest zbiorem obserwacji błędnie sklasyfikowanych. Ponieważ funkcja (17) jest ciągła, do jej minimalizacji można wykorzystać klasyczną gradientową metodę spadku (*gradient descent*). Gradient funkcji (17) wynosi:

$$\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} = -\sum_{\mathbf{x}_i \in E} \mathbf{x}_i\,. \tag{18}$$

W pierwszym kroku wybierane są losowo początkowe wartości wektora parametrów $\mathbf{a}^{(0)}$, a w następnych krokach następuje jego modyfikacja:

$$\mathbf{a}^{(r+1)} = \mathbf{a}^{(r)} + \lambda_r \sum_{\mathbf{x}_i \in E} \mathbf{x}_i \tag{19}$$

w kierunku przeciwnym do gradientu. Parametr λ_r określa długość kroku i nazywany jest także współczynnikiem uczenia. Inna często stosowana odmiana reguły (19) wykorzystuje pojedynczą, błędnie sklasyfikowaną obserwację:

$$\mathbf{a}^{(r+1)} = \mathbf{a}^{(r)} + \lambda_r \mathbf{x}_i . \tag{20}$$

Opisane sposoby modyfikacji nie zmieniają wartości wektora wag dla obserwacji poprawnie sklasyfikowanych.

Metoda perceptronowa jest zbieżna w przypadku liniowej separowalności klas. Naturalnym kryterium stopu jest więc wartość $Q(\mathbf{a}) = 0$. Wynik poszukiwania zależy od wyboru wartości początkowej wektora wag, dlatego procedurę optymalizacji można przeprowadzić kilka razy, dla różnych losowo wybranych wektorów wag początkowych $\mathbf{a}^{(0)}$, a następnie wybrać rozwiązanie najlepsze.

3. Kwadratowe funkcje dyskryminacyjne

Jak już zaznaczono, podejście regresyjne w analizie dyskryminacyjnej polega na tym, że w każdej klasie C_j budowany jest osobny model:

$$f_j(\mathbf{X}) = \alpha_{j0} + \sum_{l=1}^{L} \alpha_{jl} X_l \tag{21}$$

dla $j = 1, \ldots, J$, a położenie hiperpłaszczyzny separującej klasy, np. C_j, C_k, wyznacza się za pomocą równania:

$$\hat{f}_j(\mathbf{x}) = \hat{f}_k(\mathbf{x}) . \tag{22}$$

Jeżeli chodzi o etap klasyfikacji, to obserwacja $[\mathbf{x}_i, y_i]$ jest przydzielana do tej klasy, dla której wartość teoretyczna funkcji (21) jest największa:

$$\hat{y}_i = \arg\max_j \hat{f}_j(\mathbf{x}_i) . \tag{23}$$

Można także szukać prawdopodobieństw *a posteriori* w klasach, tj. $p(C_j \mid \mathbf{X})$. Jeżeli $g_j(\mathbf{X})$ jest funkcją gęstości w klasie C_j, to ze wzoru Bayesa wynika, że:

$$p(C_j \mid \mathbf{X}) = \frac{g_j(\mathbf{X}) p(C_j)}{\sum_{k=1}^{J} g_k(\mathbf{X}) p(C_k)} . \tag{24}$$

Zakładając, że $g_j(\mathbf{X})$ jest funkcją gęstości wielowymiarowego rozkładu normalnego:

$$g_j(\mathbf{X}) = \frac{1}{(2\pi)^{L/2}\left|\mathbf{\Sigma}_j\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T \mathbf{\Sigma}_j^{-1}(\mathbf{X}-\boldsymbol{\mu}_j)}, \tag{25}$$

gdzie $\boldsymbol{\mu}_j = [\mu_{j1}, \ldots, \mu_{jL}]$ jest środkiem ciężkości klasy C_j, tj. $\mu_{jl} = \frac{1}{N_j}\sum_{i=1}^{N_j} x_{il}$.

Wtedy, w przypadku dwóch klas C_j, C_k okazuje się, że logarytm ilorazu wiarygodności ma postać:

$$\log\frac{p(C_j \mid \mathbf{X})}{p(C_k \mid \mathbf{X})} = \log\frac{p(C_j)}{p(C_k)} - \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k) + \mathbf{X}^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k). \tag{26}$$

Wartość logarytmu (26) wyznaczono przy założeniu, że macierze wariancji i kowariancji w obu klasach są równe, tj. $\mathbf{\Sigma}_j = \mathbf{\Sigma}_k$.

W przypadku gdy wszystkie macierze wariancji i kowariancji w klasach są równe:

$$\mathbf{\Sigma} = \mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_J, \tag{27}$$

to obserwacje w każdej klasie tworzą hipersferyczne skupienia tej samej wielkości i otrzymujemy dla każdej klasy liniową funkcję dyskryminacyjną (LDA):

$$f_j(\mathbf{X}) = \mathbf{X}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_j + \log p(C_j). \tag{28}$$

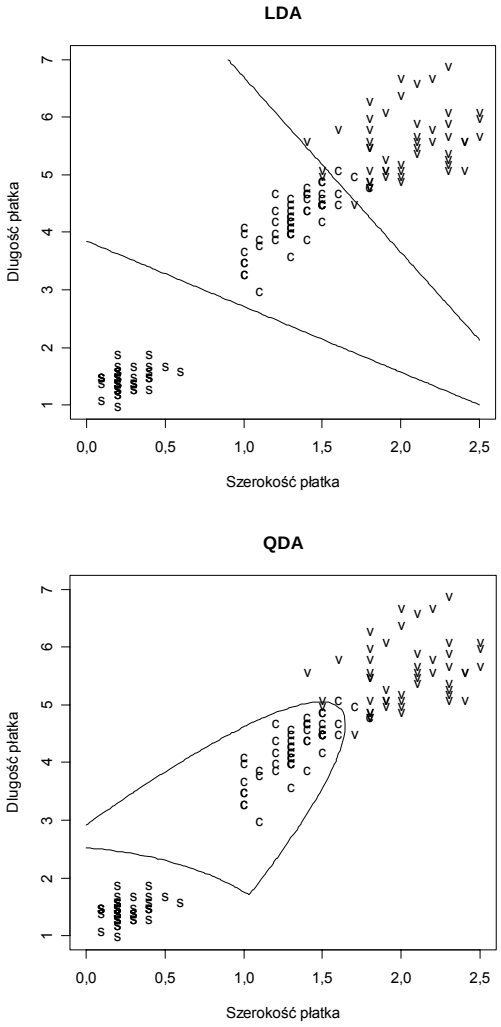
Jeżeli zmienne $\mathbf{X} = [X_1, \ldots, X_L]$ są niezależne i macierz $\mathbf{\Sigma}$ jest diagonalna, to można udowodnić, że hiperpłaszczyzna rozdzielająca dwie klasy C_j i C_k jest prostopadła do odcinka łączącego środki ciężkości obu tych klas. Ponadto, jeśli prawdopodobieństwa *a priori* dla tych klas są jednakowe, hiperpłaszczyzna przechodzi przez środek tego odcinka. W przeciwnym razie hiperpłaszczyzna jest „przesunięta" w stronę środka ciężkości tej klasy, która ma mniejsze prawdopodobieństwo *a priori* $p(C_j)$. Mówiąc obrazowo, obserwacja $\mathbf{x}_i$ ze zbioru rozpoznawanego będzie przydzielona do klasy C_j, której środek ciężkości $\boldsymbol{\mu}_j$ leży najbliżej w sensie odległości euklidesowej.

Jeżeli macierz wariancji i kowariancji $\mathbf{\Sigma}$ nie jest diagonalna, to wynik klasyfikacji obserwacji $\mathbf{x}_i$ ze zbioru rozpoznawanego zależy od odległości Mahalanobisa do środka ciężkości najbliższej klasy.

Gdy warunek (27) nie jest spełniony, to otrzymujemy kwadratowe funkcje dyskryminacyjne (QDA) w klasach:

$$f_j(\mathbf{X}) = -\frac{1}{2}\log\left|\mathbf{\Sigma}_j\right| - \frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T \mathbf{\Sigma}_j^{-1}(\mathbf{X}-\boldsymbol{\mu}_j) + \log p(C_j). \quad (29)$$

Na rys. 2 pokazano położenie liniowych i kwadratowych funkcji dyskryminacyjnych dla zbioru IRYS w przestrzeni dwuwymiarowej ($L = 2$). Wybrano dwie zmienne o największej zdolności dyskryminacyjnej, tj. długość płatka (*dp*) i szerokość płatka (*sp*).



Rys. 2. Liniowe i kwadratowe funkcje dyskryminacyjne dla zbioru IRYS

Źródło: Ibid.

Friedman (1989) zaproponował rozwiązanie kompromisowe pomiędzy liniowymi i kwadratowymi funkcjami dyskryminacyjnymi. Metoda ta nazywana jest regularyzowaną analizą dyskryminacyjną (*regularized discriminant analysis*) oraz polega na przekształceniu macierzy wariancji i kowariancji:

$$\boldsymbol{\Sigma}_j(\delta) = \delta \cdot \boldsymbol{\Sigma}_j + (1-\delta)\boldsymbol{\Sigma}. \tag{30}$$

Parametr regularyzacji $\delta \in [0,1]$ pozwala zbudować model dyskryminacyjny w postaci pośredniej pomiędzy liniową ($\delta = 1$) i kwadratową ($\delta = 0$). W praktyce parametr δ ustalany jest eksperymentalnie na podstawie zbioru testowego lub w wyniku sprawdzania krzyżowego.

Dwuparametrową rodzinę macierzy kowariancji można otrzymać, wstawiając do wzoru (30) zamiast $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}_y = \gamma \boldsymbol{\Sigma} + (1-\gamma)\sigma^2 \mathbf{I}, \tag{31}$$

gdzie $\boldsymbol{I}$ jest macierzą jednostkową, $\gamma \in [0,1]$ jest parametrem regularyzacji, zaś $\sigma^2 \boldsymbol{I}$ to diagonalna macierz wariancji i kowariancji wyznaczona na podstawie próby.

4. Drzewa klasyfikacyjne

W opozycji do klasycznych, parametrycznych metod dyskryminacji, powstały metody nieparametryczne, niewymagające spełnienia przedstawionych w poprzedniej części artykułu wymagań. Należą do nich m.in. metoda *K*-najbliższych sąsiadów i metoda drzew klasyfikacyjnych.

Ta ostatnia polega na sekwencyjnym podziale *L*-wymiarowej przestrzeni zmiennych $\mathbf{X}^L$ na podprzestrzenie R_k (segmenty), aż do chwili, gdy zmienna zależna Y osiągnie w każdej z nich minimalny poziom zróżnicowania (mierzony za pomocą odpowiedniej funkcji straty). Metoda ta nazywana jest metodą rekurencyjnego podziału (*recursive partitioning*) i była stosowana w statystyce już przez Morgana i Sonquista (1963). Jej wykorzystanie w analizie dyskryminacyjnej i regresji przedstawili Breiman i in. (1984), proponując algorytm CART. W języku polskim wyczerpującą monografią poświęconą zagadnieniom budowy modeli w postaci drzew klasyfikacyjnych i regresyjnych jest praca Gatnara (2001).

Przebieg procedury rekurencyjnego podziału najlepiej reprezentuje drzewo, tj. graf spójny i bez cykli; stąd nazwa metody – drzewa klasyfikacyjne* (*classification trees*). W ramach omawianej metody model jest tworzony nie globalnie,

* W istocie prawidłowa nazwa w języku polskim powinna brzmieć: drzewa dyskryminacyjne.

lecz poprzez złożenie modeli lokalnych o najprostszej postaci (tj. stałej), budowanych w każdym z K rozłącznych segmentów, na jakie dzielona jest wielowymiarowa przestrzeń zmiennych:

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} \alpha_k I(\mathbf{x}_i \in R_k), \tag{32}$$

gdzie R_k ($k = 1, \ldots, K$) to podprzestrzenie (segmenty) przestrzeni $\mathbf{X}^L$, α_k – parametry modelu, zaś I jest funkcją wskaźnikową.

Każdy z obszarów R_k jest definiowany poprzez jego granice w przestrzeni $\mathbf{X}^L$, które dla zmiennych metrycznych $X_1, \ldots, X_L$, można przedstawić jako:

$$I(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} I\left(v_{kl}^{(d)} \le x_{il} \le v_{kl}^{(g)}\right), \tag{33}$$

gdzie wartości $v_{kl}^{(d)}$ oraz $v_{kl}^{(g)}$ oznaczają odpowiednio jego górną i dolną granicę w *l*-tym wymiarze przestrzeni.

Gdy zmienne $X_1, \ldots, X_L$ mają charakter niemetryczny, to podprzestrzeń R_k można zdefiniować jako:

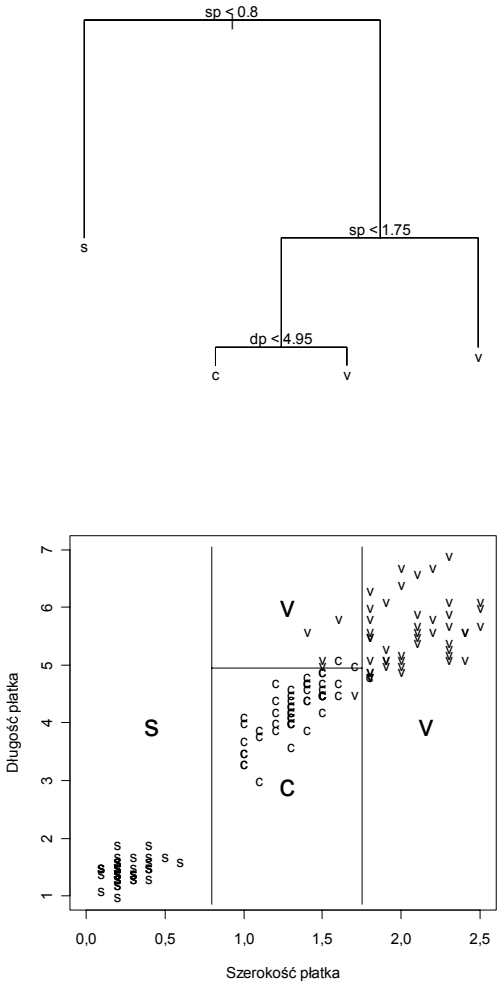
$$I(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} I\left(x_{il} \in B_{kl}\right), \tag{34}$$

gdzie B_{kl} to podzbiór zbioru kategorii zmiennej X_l, tj. $B_{kl} \subseteq V_l$.

Jeżeli zmienna zależna Y w modelu (32) jest zmienną nominalną, to model ten nazywany jest dyskryminacyjnym i reprezentuje go drzewo klasyfikacyjne. Wtedy parametry α_k modelu (32) są wyznaczane jako:

$$\alpha_k = \arg\max_j p(C_j \mid \mathbf{x}_i \in R_k). \tag{35}$$

Model w postaci drzewa klasyfikacyjnego dla zbioru IRYS, wykorzystujący dwie zmienne objaśniające: długość płatka *(dp)* i szerokość płatka *(sp)*, oraz odpowiadający mu podział przestrzeni dwuwymiarowej na 4 segmenty, zostały pokazane na rys. 1. Jak widać, segment oznaczony literą *S* jest homogeniczny, ponieważ należą do niego wszystkie kwiaty z gatunku *Setosa*. Jego granice wyznacza formuła: $0 < sp < 0{,}8$. Z kolei segment oznaczony literą *C* zawiera większość kwiatów z gatunku *Versicolor*, a jego granicami w pierwszym wymiarze jest $0{,}8 < sp < 1{,}75$ oraz w drugim – $0 < dp < 4{,}95$.



Rys. 3. Drzewo klasyfikacyjne oraz podział przestrzeni zmiennych dla zbioru IRYS

Źródło: Ibid.

Do oceny stopnia zróżnicowania podprzestrzeni R_k można wykorzystać jedną z następujących miar:

– błąd klasyfikacji (*misclassification error*):

$$Q(R_k) = 1 - \arg\max_j p(C_j \mid R_k), \tag{36}$$

– wskaźnik Giniego (*Gini index*):

$$Q(R_k) = 1 - \sum_{j=1}^{J} \left(p(C_j \mid R_k) \right)^2 , \tag{37}$$

– entropia:

$$Q(R_k) = -\sum_{j=1}^{J} p(C_j \mid R_k) \log_2 p(C_j \mid R_k) . \tag{38}$$

Omówienie własności przedstawionych powyżej miar oraz charakterystyka innych, nieco mniej znanych, znajduje się w pracy Gatnara (2001).

Powyższe miary homogeniczności są wykorzystywane do kontroli procesu podziału przestrzeni zmiennych. Stosowana jest w tym celu strategia wspinaczki (*hill climbing*), pozwalająca dokonać podziału, który jest optymalny w sensie lokalnym. Nie gwarantuje to oczywiście osiągnięcia rozwiązania optymalnego w sensie globalnym.

W każdym kroku ocena jakości podziału podprzestrzeni R na segmenty $R_1, \ldots, R_K$ odbywa się za pomocą kryterium:

$$\Delta Q(R) = Q(R) - \sum_{k=1}^{K} Q(R_k)\, p(k) , \tag{39}$$

gdzie $p(k)$ oznacza frakcję obserwacji w segmencie R_k. Kryterium (39) podlega maksymalizacji, tj. szukany jest taki podział, który zapewni jak największą jednorodność uzyskanych podprzestrzeni, czyli osiągnięcie minimum przez $Q(R_k)$ dla $k = 1, \ldots, K$.

Breiman i in. (1984) wykorzystali w swojej pracy do oceny homogeniczności segmentów wskaźnik Giniego (37). Ma on jednak pewną wadę, ponieważ osiąga maksimum również wtedy, gdy segmenty R_k zawierają jednakową liczbę obserwacji. Z kolei Quinlan (1993) w swoim algorytmie C4.5 stosuje entropię (38), której główna wada polega na tym, że preferuje ona taki podział, który generuje maksymalną liczbę segmentów R_k. Aby tego uniknąć, można zastosować normalizację, uzyskując tzw. względny przyrost informacji (*gain ratio*):

$$\Delta Q^*(R) = \frac{\Delta Q(R)}{-\sum_{k=1}^{K} p(k) \log_2 p(k)} . \tag{40}$$

Podział przestrzeni $\mathbf{X}^L$ na podprzestrzenie odbywa się za pomocą hiperpłaszczyzn równoległych do osi (gdy zmienne $X_1, \ldots, X_L$ są zmiennymi metrycznymi). Równanie takiej hiperpłaszczyzny ma wtedy postać $X_l = c$, gdzie zarówno wybór zmiennej X_l, jak i wartości c kontroluje miara (39).

Aby wyznaczyć stałą c, należy obliczyć wartość kryterium (39) dla wszystkich możliwych wariantów podziału zbioru wartości $V_l = \{v_{l1}, \ldots, v_{lT}\}$ zmiennej X_l:

$$c = \frac{v_{lt} + v_{lt+1}}{2}. \tag{41}$$

Zawsze uzyskuje się w ten sposób dwa zbiory obserwacji: $\{\mathbf{x}_i : x_{il} \le c\}$ oraz $\{\mathbf{x}_i : x_{il} > c\}$. Inaczej mówiąc, dokonywana jest dyskretyzacja zmiennej X_l, której rezultatem jest powstanie drzewa binarnego, w którym z każdego węzła wychodzą dwie krawędzie.

W procesie budowy modelu w postaci drzewa klasyfikacyjnego najpierw każda zmienna metryczna poddawana jest dyskretyzacji*, a następnie wybierana jest ta spośród nich, dla której kryterium (39) osiąga maksimum.

Jeżeli zmienna X_l ma charakter niemetryczny, to zbiór jej kategorii $V_l = \{v_{l1}, \ldots, v_{lT}\}$ jest dzielony na dwa podzbiory (w przypadku drzewa binarnego), tak aby wartość kryterium (39) była jak największa (takich podziałów jest 2^T dla zmiennych porządkowych oraz $2^{T-1} - 1$ dla zmiennych nominalnych). Najczęściej punktem wyjścia jest podział V_l na T podzbiorów $\{v_{l1}\}, \ldots, \{v_{lT}\}$, a następnie te podzbiory są stopniowo łączone. W metodzie CHAID, którą zaproponował Kaas (1980) tym procesem łączenia steruje statystyka χ^2.

W przypadku modeli w postaci drzew klasyfikacyjnych pojawia się problem wyboru takiej postaci modelu, by jego błąd predykcji był jak najmniejszy. Spośród metod wykorzystywanych w celu wyeliminowania tego zjawiska i zmniejszenia stopnia złożoności modelu, najczęściej** stosuje się tzw. przycinanie krawędzi (*pruning*). Zabieg ten powoduje redukcję wielkości drzewa poprzez usunięcie niektórych jego fragmentów, co może oznaczać, że z modelu zostaną wyeliminowane niektóre zmienne.

Breiman i in. (1984) zaproponowali pewną formę regularyzacji, która pozwala uzyskać kompromis pomiędzy złożonością modelu i jego jakością w postaci kryterium:

$$S_\lambda(D) = Q(D) + \lambda \cdot K, \tag{42}$$

---

* W pracy Gatnara (2001) omówiono także metody podziału zbioru wartości zmiennej X_l na trzy i więcej przedziałów (*multiway split*), w rezultacie czego powstają drzewa niebinarne. To zagadnienie jest jednak jeszcze bardziej złożone.

** Gatnar (2001) omawia także rzadziej stosowaną metodę skracania krawędzi drzewa (ang. *shrinking*), które są proporcjonalne do stopnia homogeniczności w węzłach.

które podlega minimalizacji. W powyższej formule $Q(D) = \sum_{k=1}^{K} Q(R_k)\,p(k)$ to miara jakości modelu D w postaci drzewa, K oznacza liczbę liści i jest oceną złożoności modelu, zaś λ to tzw. parametr złożoności ($\lambda \geq 0$). Duże wartości parametru λ oznaczają podział na niewiele segmentów (małe drzewa), zaś małe wartości – drzewa bardziej rozbudowane, o dużej liczbie liści. W przypadku gdy $\lambda = 0$, powstaje drzewo maksymalne (pełne) D_0.

5. Kierunki rozwoju analizy dyskryminacyjnej

Wzrost możliwości przetwarzania dużych zbiorów danych przez współczesne komputery oraz dostępność zaawansowanego oprogramowania statystycznego powoduje, że spośród metod analizy dyskryminacyjnej najszybciej rozwijają się metody nieparametryczne, wykorzystywane w systemach *business intelligence* i *data mining*.

Należą do nich np.: metoda K-najbliższych sąsiadów oraz drzewa klasyfikacyjne. Ważną zaletą tej ostatniej klasy modeli jest możliwość klasyfikacji danych niepełnych, tj. zawierających obserwacje, dla których nie można określić wartości pewnych zmiennych, np. gdy są one trudne do zmierzenia. Metoda ta jest również odporna na występowanie wartości nietypowych. Ponadto w modelu dyskryminacyjnym zmiennymi objaśniającymi mogą być zmienne mierzone zarówno na skalach mocnych, jak i na skalach słabych, bez konieczności dokonywania ich transformacji.

Od dłuższego czasu utrzymuje się zainteresowanie także takimi nieklasycznymi metodami dyskryminacji, jak: sieci neuronowe (*neural networks*) oraz metoda wektorów nośnych SVM (*Suport Vector Machines*). Sieć neuronowa to model złożony z wielu modeli liniowych znajdujących się w poszczególnych warstwach sieci, które przetwarzają dane wejściowe i korygują („uczą się"), w czasie tysięcy powtórzeń, parametry poszczególnych modeli składowych (Rosenblatt, 1958).

Z kolei metoda SVM, zaproponowana przez Vapnika (1995), polega na transformacji obserwacji z pierwotnej przestrzeni zmiennych objaśniających w przestrzeń o wiele większym wymiarze, w której klasy są łatwiej separowalne. Obserwacje definiujące położenie funkcji oddzielających klasy nazywane są wektorami nośnymi.

Warto także wspomnieć o metodach łączenia modeli dyskryminacyjnych, które powodują znaczące zwiększenie dokładności klasyfikacji. Zalety tego podejścia, nazywanego wielomodelowym, przedstawił wyczerpująco w swojej monografii Gatnar (2008).

Duży wysiłek badawczy jest wkładany obecnie w poszukiwanie metod dyskryminacji dla dużych i wielowymiarowych zbiorów danych. Należą do nich przede wszystkim zbiory danych o genotypach (*gene microarray data*).

Bibliografia

Breiman L., Friedman J., Olshen R., Stone C. (1984): *Classification and Regression Trees*. CRC Press, London.

Bryan J.G. (1951): *The Generalized Discriminant Function: Mathematical Foundation and Computational Routine*. Harvard Education Review, 21, s. 90-95.

Duda R. O., Hart P. E., Storck G. E. (2001): *Pattern Classification*. John Wiley & Sons, New York.

Fisher L.A. (1936): *The Use of Multiple Measurements in Taxonomic Problems*. Annals of Eugenics, t. 7, s. 179-188.

Friedman J. H. (1989): *Regularized Discriminant Analysis*. „Journal of the American Statistical Association", 84, s. 165-175.

Gatnar E. (1998): *Symboliczne metody klasyfikacji danych*. Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E. (2001): *Nieparametryczna metoda dyskryminacji i regresji*. Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E. (2008): *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*. Wydawnictwo Naukowe PWN, Warszawa.

Hastie T., Tibshirani R., Friedman J. (2001): *The Elements of Statistical Learning*. Springer Series in Statistics, Springer, Berlin.

Huberty G. J. (1995): *Applied Discriminant Analysis*. John Wiley & Sons, New York.

Jajuga K. (1990): *Statystyczna teoria rozpoznawania obrazów*. Wydawnictwo Naukowe PWN, Warszawa.

Jajuga K. (1993): *Statystyczna analiza wielowymiarowa*. Państwowe Wydawnictwo Naukowe, Warszawa.

Kaas G. V. (1980): *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. „Applied Statistics", 29, s. 119-127.

Kolonko J. (1980): *Analiza dyskryminacyjna w badaniach ekonomicznych*. PWN, Warszawa.

McLachlan G.J. (1992): *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, New York.

Morgan J.N., Sonquist J.A. (1963): *Problems in the Analysis of Survey Data: A Proposal*. „Journal of the American Statistical Association", 58, s. 417-434.

Nilsson N.J. (1965): *Learning Machines: Foundations of Trainable Pattern-Classifying Systems*. McGraw-Hill.

Quinlan J.R. (1983): *Learning Efficient Classification Procedures and their Application to Chess and Games*. W: R. Michalski, J. Carbonell, T. Mitchell (eds.): *Machine Learning. An Artificial Intelligence Approach*. Tioga, Palo Alto, s. 126-142.

Quinlan J.R. (1986): *Induction of decision trees*, Machine Learning, 1, s. 81-106.

Quinlan J.R. (1993): *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Rao C. (1948): *The Utilisation of Multiple Measurements in Problems of Biological Classification*. „Journal of the Royal Statistical Society B", 10, s. 159-203.

Ripley B.D. (1996): *Pattern Recognition and Neural Networks*. Cambridge University Press. Cambridge.

Rosenblatt F. (1958): *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*. „Psychological Review", 65(6), s. 386-408.

Titterington D.M., Murray G.D., Murray L.S., Spiegelhalter D.J., Skene A.M., Habbema J.D., Gelpke G.J. (1981): *Comparison of Discriminant Techniques Applied to Complex Data Sets of Head Injured Patients*. „Journal of the Royal Statistical Society, Series A", 144, s. 145-175.

Vapnik V. (1995): *The Nature of Statistical Learning Theory*. Springer, Berlin.

Vapnik V. (1998): *Statistical Learning Theory*. John Wiley and Sons, New York.

Wernecke K.-D. (1992): *A Coupling Procedure for Discrimination of Mixed Data*. „Biometrics", 48, s. 497-506.

DISCRIMINANT ANALYSIS – STATE OF THE ART AND FUTURE DEVELOPMENTS

Summary

The aim of the discriminant analysis is to partition the multivariate feature space into subspaces in order to separate observations belonging to different classes. In other words, its task is to find a model that can give class descriptions on the basis of a set containing previously classified observations. Then the model is applied to classify new ones with a minimum error. Founded in 1936 by Fisher, the discriminant analysis had become an important part of multivariate statistical analysis. It has many applications and is an obligatory procedure in many available data mining systems.

In Poland prof. Józef Kolonko has been one of the pioneering statisticians interested in discriminant analysis. He published his book on discriminant analysis in 1980, based on cybernetics. Therefore the analysis had a broader meaning, including both supervised and unsupervised classification.

Nowadays, in statistical literature, discriminant analysis is considered only as a set of methods that can discover class descriptions on the basis of the training set, i.e. supervised classification.

This article is devoted to review the existing approaches and new developments in discriminant analysis starting from its roots, i.e. Fisher's approach based on regression analysis, and concluding with classification trees as a nonparametric discriminant analysis techmique. We also mention new approaches recently proposed in statistical literature, such as: neural nets, K-nearest neighbors and support vector machines.