



O ZASTOSOWANIU STATYSTYCZNYCH METOD ROZPOZNAWANIA OBRAZÓW DO WSPOMAGANIA PROCESÓW PODEJMOWANIA DECYZJI W DIAGNOSTYCE MEDYCZNEJ

Małgorzata Misztal

Uniwersytet Łódzki, Wydział Ekonomiczno–Socjologiczny, Katedra Metod Statystycznych

Uwagi wstępne

Działalność człowieka to nieustanny proces podejmowania decyzji. Z każdą decyzją związana jest jednak możliwość popełnienia błędu, a dodatkowo, wybór danej decyzji ze zbioru decyzji dopuszczalnych nie przesądza w sposób jednoznaczny o wyniku lub skutku podjęcia decyzji.

Podjęcie decyzji wymaga zwykle rozważnego zgłębienia wszystkich możliwych sposobów działania, a następnie wyboru jednego z nich. Coraz większa złożoność otaczających nas zjawisk sprawia, że niezbędne staje się poszukiwanie metod wspomagających procesy podejmowania decyzji w warunkach mnogości informacji i niepewności.

W celu efektywnego rozwiązywania praktycznych problemów wymagających przechowywania i przetwarzania dużej ilości danych opisanych w przestrzeniach wielowymiarowych zaproponować można metody rozpoznawania obrazów.

Obraz definiowany będzie jako ilościowy opis obiektu, zdarzenia lub zjawiska.

Ogólnie zadanie teorii rozpoznawania obrazów polega na określaniu przynależności rozmaitego typu obiektów do pewnych klas w sytuacji braku apriorycznej informacji co do reguł przynależności, a jedyną dostępną informację stanowi zwykle tzw. ciąg uczący, złożony z obiektów, których prawidłową klasyfikację znamy (tzw. rozpoznawanie z nauczycielem).

Dokładniej rozpoznawanie obrazów można zdefiniować jako wieloetapowy proces przetwarzania informacji, podczas którego relatywnie duża ilość danych wejściowych zostaje przetworzona na mniejszą ilość danych użytecznych, zakończony klasyfikacją, czyli przypisaniem obiektowi numeru klasy (por. [1]).

Wśród metod rozpoznawania obrazów wyróżnić można (por. np. [5], [9]):

- ◆ metody fizjologiczne (biocybernetyczne), w których dąży się do naśladowania procesów umysłowych przez tworzenie modeli systemu nerwowego;



- ♦ metody programowe (algorytmiczne), w których głównym celem jest tworzenie formalnych metod opisu zadania rozpoznawania i związanych z nimi algorytmów możliwych do komputerowej realizacji. Wyróżnia się tutaj rozpoznawanie strukturalne i rozpoznawanie teoriodecyzyjne.

Prezentowane w artykule teoriodecyzyjne metody rozpoznawania wymagają przyjęcia założenia, że rozpoznawany obiekt, scharakteryzowany wartościami p cech, może być rozpatrywany jako punkt $\mathbf{x}=(x_1, \dots, x_p)^T$ p -wymiarowej przestrzeni \mathbf{X} ($\mathbf{X} \subseteq \mathbb{R}_n$) i traktowany jako realizacja wektora losowego \mathbf{X} o funkcji gęstości $f_i(\mathbf{x})$, $i \in \mathbf{K}$, gdzie $\mathbf{K}=\{1, \dots, k\}$ - jest zbiorem numerów klas. Decyzja zaliczająca obiekt do klasy wynika z transformacji zaobserwowanych wartości za pomocą pewnego algorytmu, zwanego algorytmem rozpoznawania.

Algorytmem rozpoznawania ψ (algorytmem klasyfikacji, regułą decyzyjną) nazywamy przepis, według którego odbywa się przyporządkowanie rozpoznawanemu obiektowi $\mathbf{x} \in \mathbf{X}$ numeru klasy $i \in \mathbf{K}$: $\psi(\mathbf{x}) = i$.

Innymi słowy, mamy tu do czynienia z odwzorowaniem przestrzeni cech w zbiór numerów klas: $\psi: \mathbf{X} \rightarrow \mathbf{K}$ bądź też z generowaniem rozkładu przestrzeni cech na rozłączne obszary decyzyjne: $R_i = \{\mathbf{x} \in \mathbf{X}: \psi(\mathbf{x}) = i\}$, $i \in \mathbf{K}$. Obszary decyzyjne R_i w pełni opisują konkretny algorytm rozpoznawania - obiekt dany wektorem cech \mathbf{x} zaliczany jest do klasy i , jeśli \mathbf{x} należy do obszaru decyzyjnego R_i .

Kolejne cechy algorytmu rozpoznawania to jednoznaczność i kompletność - rozpoznawany jest każdy obiekt (bo przestrzeń \mathbf{X} jest zbiorem wszystkich możliwych wartości cech) i zaliczany jest on do jednej i tylko jednej klasy ze zbioru \mathbf{K} .

W rozpoznawaniu teoriodecyzyjnym do opisu analizowanej sytuacji wykorzystuje się modele probabilistyczne i statystyczne, ze względu na ich szczególną przydatność do wykrywania niepewnych i niejednoznacznych związków między klasami i ilościowymi charakterystykami obiektów.

Wybrane metody tworzenia algorytmów rozpoznawania

Wśród metod tworzenia algorytmów rozpoznawania wyróżniamy podejście oparte na modelu probabilistycznym oraz podejście oparte na modelu statystycznym.

W przypadku modelu probabilistycznego zakłada się, że dla każdego rozpoznawanego obiektu \mathbf{x} znane jest prawdopodobieństwo *a priori* q_i zdarzenia, że pochodzi on z klasy o numerze i ; $i \in \mathbf{K}$; a także znane są warunkowe gęstości rozkładów cech w poszczególnych klasach:

$$f(\mathbf{x}/i) = f_i(\mathbf{x}) \quad \mathbf{x} \in \mathbf{X}. \quad (1)$$



W takiej sytuacji możliwe jest obliczenie wskaźnika jakości rozpoznawania oraz, poprzez rozwiązanie odpowiedniego problemu optymalizacyjnego, wyznaczenie reguły decyzyjnej minimalizującej ten wskaźnik. W zadaniach rozpoznawania opartych na modelach probabilistycznych wykorzystuje się np. klasyfikację bayesowską lub regułę minimaxową (por. np. [7, 8, 9, 12]).

W praktycznych zastosowaniach metod rozpoznawania obrazów korzysta się zwykle ze źródła informacji, jakim jest pewien zbiór obiektów, zwany zbiorem uczącym. Dla każdego obiektu z tego zbioru (czyli obiektu uczącego) znany jest wektor wartości cech oraz numer klasy, do której należy. Mamy więc:

$$U = \{ (\mathbf{x}_1, i_1), (\mathbf{x}_2, i_2), \dots, (\mathbf{x}_N, i_N) \}. \quad (2)$$

Podzbiór zbioru U złożony z obiektów uczących należących do i -tej klasy oznaczamy:

$$U_i = \{ \mathbf{x}_{i,l} \in \mathbf{X}, l=1, 2, \dots, N_i \}, \quad i \in \mathbf{K}, \quad (3)$$

i zakładamy, że jego elementy pochodzą z populacji o warunkowej gęstości $f_i(\mathbf{x})$. Oczywiście: $U = \{ U_1, U_2, \dots, U_k \}$ oraz $N = \sum N_i$.

Zatem podstawą konstrukcji reguł decyzyjnych ze zbiorem uczącym jest model statystyczny. Wobec tego rozważyć można dwie sytuacje:

- ◆ znamy z założenia postać funkcyjną warunkowych gęstości w klasach, a nie znamy ich parametrów – dokonujemy więc ich estymacji na podstawie zbioru uczącego;
- ◆ brak jest jakichkolwiek założeń co do postaci funkcyjnej warunkowych gęstości w klasach – dokonujemy więc estymacji funkcji gęstości za pomocą metod nieparametrycznych.

W grupie algorytmów rozpoznawania opartych na parametrycznym modelu statystycznym najczęściej wykorzystywane są te metody, w których przyjmuje się założenie o normalności rozkładów cech obiektów w klasach. Wymienić tu można m. in. algorytm rozpoznawania wykorzystujący odległość Mahalanobisa oraz algorytmy wykorzystujące estymatory liniowych i kwadratowych funkcji klasyfikacyjnych.

Algorytm rozpoznawania oparty na odległościach Mahalanobisa zapisać można w następujący sposób:

$$\psi(\mathbf{x}) = i, \quad \text{gdy} \quad {}^M D_i^2(\mathbf{x}) = \min_{g \in \mathbf{K}} \{ {}^M D_g^2(\mathbf{x}) \}, \quad i \in \mathbf{K} \quad (4)$$

gdzie:

$${}^M D_i^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i), \quad i \in \mathbf{K} \quad (5)$$

przy czym $\bar{\mathbf{x}}_i$ i \mathbf{S} są zwykłymi estymatorami wektora średnich w i -tej klasie i macierzy kowariancji.

Reguła klasyfikacyjna jest następująca: wybieramy jako rozpoznanie tę klasę, najbliższą której (w myśl odległości Mahalanobisa) znajduje się rozpoznawany obiekt.

Algorytmy rozpoznawania wykorzystujące zwykle obciążone, zwykle nieobciążone, bayesowskie i quasi-bayesowskie estymatory liniowych funkcji klasyfikacyjnych można zapisać następująco (por. [7]):

$$\psi(\mathbf{x}) = i, \text{ gdy } {}_{(j)}\hat{e}_i(\mathbf{x}) = \max_{g \in \mathbf{K}} {}_{(j)}\hat{e}_g(\mathbf{x}); \quad j=1, \dots, 4; \quad i \in \mathbf{K} \quad (6)$$

gdzie:

$${}_{(1)}\hat{e}_i(\mathbf{x}) = -\frac{1}{2}d_i^2(\mathbf{x}) + \ln q_i \quad (7)$$

$${}_{(2)}\hat{e}_i(\mathbf{x}) = -\frac{1}{2} \frac{N-k-p-1}{N-k} d_i^2(\mathbf{x}) + \frac{1}{2} \frac{p}{N_i} + \ln q_i \quad (8)$$

$${}_{(3)}\hat{e}_i(\mathbf{x}) = -\frac{1}{2}d_i^2(\mathbf{x}) - \frac{1}{2} \frac{p}{N_i} + \ln q_i \quad (9)$$

$${}_{(4)}\hat{e}_i(\mathbf{x}) = -\frac{N-k+1}{2} \ln \left[1 + N_i(N_i+1)^{-1}(N-k)^{-1}d_i^2(\mathbf{x}) \right] + \frac{p}{2} \ln \frac{N_i}{N_i+1} + \ln q_i \quad (10)$$

przy czym:

$$d_i^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i) \quad (11)$$

$$g_i = \left[\frac{N_i}{\pi(N-k)(N_i+1)} \right]^{p/2} \frac{\Gamma((N-k+1)/2)}{\Gamma((N-k-p+1)/2) |\mathbf{S}|^{1/2}} \quad (12)$$

a $\bar{\mathbf{x}}_i$ i \mathbf{S} są zwykłymi estymatorami wektora średnich w i -tej klasie i macierzy kowariancji.

Reguła klasyfikacyjna jako rozpoznanie wybiera tę klasę, dla której funkcja klasyfikacyjna przyjmuje największą wartość.

Algorytmy rozpoznawania wykorzystujące zwykle obciążone, zwykle nieobciążone, bayesowskie i quasi-bayesowskie estymatory kwadratowych funkcji klasyfikacyjnych można przedstawić w następujący sposób (por. [7]):

$$\psi(\mathbf{x}) = i, \text{ gdy } {}_{(j)}\hat{u}_i(\mathbf{x}) = \max_{g \in \mathbf{K}} {}_{(j)}\hat{u}_g(\mathbf{x}); \quad j=1, \dots, 4; \quad i \in \mathbf{K}; \quad (13)$$

gdzie:

$${}_{(1)}\hat{u}_i(\mathbf{x}) = -\frac{1}{2}D_i^2(\mathbf{x}) - \frac{1}{2} \ln |\mathbf{S}_i| + \ln q_i \quad (14)$$

$$\begin{aligned} {}_{(2)}\hat{u}_i(\mathbf{x}) = & -\frac{1}{2} \frac{N_i-p-2}{N_i-1} D_i^2(\mathbf{x}) - \frac{1}{2} \ln |\mathbf{S}_i| + \frac{1}{2} \frac{p}{N_i} + \\ & + \frac{1}{2} \sum_{n=1}^p \boldsymbol{\Psi} \left(\frac{1}{2}(N_i-n) \right) - \frac{p}{2} \ln \left(\frac{N_i-1}{2} \right) + \ln q_i \end{aligned} \quad (15)$$



$${}_{(3)}\hat{u}_i(\mathbf{x}) = -\frac{1}{2}D_i^2(\mathbf{x}) - \frac{1}{2}\ln|\mathbf{S}_i| - \frac{1}{2}\frac{p}{N_i} + \frac{1}{2}\sum_{n=1}^p \boldsymbol{\psi}\left(\frac{1}{2}(N_i - n)\right) - \frac{p}{2}\ln(N_i - 1) + \ln q_i \quad (16)$$

$${}_{(4)}\hat{u}_i(\mathbf{x}) = -\frac{N_i}{2}\ln\left[1 + N_i(N_i^2 - 1)^{-1}D_i^2(\mathbf{x})\right] + \ln(c_i q_i) \quad (17)$$

przy czym:

$$D_i^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{S}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) \quad (18)$$

$$\boldsymbol{\psi}(x) = \frac{d \ln \Gamma(x)}{dx} \quad (19)$$

$$c_i = \left[\frac{N_i}{\pi(N_i^2 - 1)} \right]^{\frac{p}{2}} \frac{\Gamma(N_i/2)}{\Gamma((N_i - p)/2) |\mathbf{S}_i|^{1/2}} \quad (20)$$

a $\bar{\mathbf{x}}_i$ i \mathbf{S}_i są zwykłymi estymatorami wektora średnich i macierzy kowariancji w i -tej klasie.

Reguła klasyfikacyjna wybiera jako rozpoznanie tę klasę, dla której funkcja klasyfikacyjna przyjmuje największą wartość.

Wśród metod rozpoznawania opartych na nieparametrycznym modelu statystycznym wyróżnić można m.in. algorytmy oparte na estymatorze Parzena z gaussowską funkcją jądra czy algorytmy minimalnoodległościowe.

Algorytm rozpoznawania oparty na estymatorze Parzena z gaussowską funkcją jądra zapiszemy w sposób następujący (por. np. [7], [9]):

$$\psi(\mathbf{x})=i, \text{ gdy } \frac{1}{h^p(N_i)} \sum_{s=1}^{N_i} K\left[\frac{\mathbf{x} - \mathbf{x}_s}{h(N_i)}\right] = \max_{g \in \mathbf{K}} \frac{1}{h^p(N_g)} \sum_{s=1}^{N_g} K\left[\frac{\mathbf{x} - \mathbf{x}_s}{h(N_g)}\right] \quad (21)$$

gdzie:

$$K(\mathbf{y}) = (2\pi)^{-\frac{p}{2}} \exp\left[-\frac{\|\mathbf{y}\|^2}{2}\right] \quad (22)$$

Spośród algorytmów bazujących na pojęciach sąsiedztwa i odległości wymienić warto algorytm najbliższego sąsiada, algorytm α najbliższych sąsiadów oraz algorytm DB oparty na odległościach.

Reguła klasyfikacyjna najbliższego sąsiada (ang. *Nearest Neighbour* - NN) wskazuje jako rozpoznanie tę klasę, do której należy obiekt najbliższy w myśl przyjętej miary odległości d rozpoznawanemu obiektowi \mathbf{x} , co zapisujemy (por. np. [9, 13]):

$$\psi(\mathbf{x}) = i; i \in \mathbf{K}, \text{ gdy } d(\mathbf{x}; \mathbf{x}_{i,l_i}) = \min_{g \in \mathbf{K}} d(\mathbf{x}; \mathbf{x}_{g,l_g}) \quad l_i = 1, \dots, N_i \quad l_g = 1, \dots, N_g \quad (23)$$



gdzie $d(*)$ jest miarą odległości, np.:

◆ Euklidesa:

$$d(\mathbf{x}_m; \mathbf{x}_n) = \left[\sum_{r=1}^p |x_{mr} - x_{nr}|^2 \right]^{\frac{1}{2}} \quad (24)$$

◆ Canberra:

$$d(\mathbf{x}_m; \mathbf{x}_n) = \sum_{r=1}^p \frac{|x_{mr} - x_{nr}|}{|x_{mr} + x_{nr}|} \quad (25)$$

◆ GDM Walesiaka ([14]):

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^m a_{ikj} b_{kij} + \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq i, k}}^n a_{ilj} b_{klj}}{2 \left[\left(\sum_{j=1}^m a_{ikj}^2 + \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq i, k}}^n a_{ilj}^2 \right) \left(\sum_{j=1}^m b_{kij}^2 + \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq i, k}}^n b_{klj}^2 \right) \right]^{\frac{1}{2}}} \quad (26)$$

przy czym dla zmiennych mierzonych na skali ilorazowej i (lub) przedziałowej stosowane jest podstawienie:

$$\begin{aligned} a_{ipj} &= x_{ij} - x_{pj} \quad \text{dla } p = k, l \\ b_{krj} &= x_{kj} - x_{rj} \quad \text{dla } r = i, l \end{aligned} \quad (27)$$

a dla zmiennych mierzonych na skali porządkowej:

$$a_{ipj}(b_{krj}) = \begin{cases} 1 \Leftrightarrow x_{ij} > x_{pj} & (x_{kj} > x_{rj}); \\ 0 \Leftrightarrow x_{ij} = x_{pj} & (x_{kj} = x_{rj}); \\ -1 \Leftrightarrow x_{ij} < x_{pj} & (x_{kj} < x_{rj}) \end{cases} \quad \text{dla } p = k, l; r = i, l \quad (28)$$

Reguła klasyfikacyjna α najbliższych sąsiadów (ang. α - *Nearest Neighbours* - α -NN) wskazuje jako rozpoznanie tę klasę, która jest najliczniej reprezentowana wśród α najbliższych rozpoznawanemu obiektowi \mathbf{x} obiektów z ciągu uczącego, co zapisujemy:

$$\psi(\mathbf{x}) = i; i \in \mathbf{K}, \quad \text{gdy } \alpha_i = \max_{g \in \mathbf{K}} \alpha_g \quad (29)$$

Algorytm DB (ang. *Distance -based*) oparty na odległościach (por. [3]) zapisać można następująco:

$$\psi(\mathbf{x}) = i; i \in \mathbf{K}, \quad \text{gdy } {}^{DB}D_i(\mathbf{x}) = \min_{g \in \mathbf{K}} \{{}^{DB}D_g(\mathbf{x})\} \quad (30)$$



gdzie ${}^{DB}D_i(\mathbf{x})$ jest funkcją klasyfikującą postaci:

$${}^{DB}D_i(\mathbf{x}) = \frac{1}{N_i} \sum_{m=1}^{N_i} d(\mathbf{x}; \mathbf{x}_m) - \frac{1}{2N_i^2} \sum_{m=1}^{N_i} \sum_{n=1}^{N_i} d(\mathbf{x}_m; \mathbf{x}_n) \quad (31)$$

a d^* jest odległością między obiektami. Jako rozpoznanie wybieramy tę klasę, dla której funkcja klasyfikująca przyjmuje wartość najmniejszą.

Przedstawione algorytmy tworzenia reguł decyzyjnych, oparte na modelu statystycznym, określić można mianem klasycznych, bazują one bowiem na rozwiązaniach analizy dyskryminacji, metod decyzji statystycznych, teorii estymacji (zarówno parametrycznej, jak i nieparametrycznej), bayesowskiej teorii decyzji czy metod optymalizacyjnych.

Ocenić jakość reguły klasyfikacyjnej wymaga wykorzystania zbioru testowego, złożonego z M obiektów (\mathbf{x}_l) wraz z ich prawidłowymi klasyfikacjami (i_l) :

$$T_M = \{ (\mathbf{x}_1, i_1), (\mathbf{x}_2, i_2), \dots, (\mathbf{x}_M, i_M) \} \quad l=1, \dots, M. \quad (32)$$

Jakość algorytmu ψ określa się poprzez oszacowanie prawdopodobieństwa błędnej klasyfikacji:

$$\hat{P}_e(\psi) = \frac{1}{M} \sum_{l=1}^M I\{\psi(\mathbf{x}_l) \neq i_l\} \quad (33)$$

gdzie $I\{A\}$ jest funkcją wskaźnikową postaci:

$$I\{A\} = \begin{cases} 1, & \text{w przypadku zajścia zdarzenia } A \\ 0, & \text{w przeciwnym wypadku} \end{cases} \quad (34)$$

W praktycznych zadaniach rozpoznawania rzadko zachodzi możliwość wykorzystania zbioru testowego. Wobec tego do oceny jakości algorytmu rozpoznawania wykorzystuje się takie metody jak: metoda resubstytucji, metoda wydzielenia, metoda usuwania, metoda rotacji sprawdzania krzyżowego (por. np. [9]).

W rozważanym dalej przykładzie do oceny dokładności klasyfikacji wykorzystano metodę usuwania (ang. *leave-one-out*), polegającą na tym, że na podstawie zbioru U_{N-1} konstruowana jest reguła klasyfikacyjna, a brakujący element traktowany jest jako jednoelementowy zbiór testujący T_1 . Taką procedurę powtarza się N razy, zmieniając eliminowany obiekt ze zbioru uczącego. Zatem:

$$\hat{P}_e(\psi_{N-1}) = \frac{1}{N} \sum_{l=1}^N I\{\psi_{N-1}(\mathbf{x}_l) \neq i_l\}. \quad (35)$$

Alternatywę dla klasycznych metod rozpoznawania obrazów stanowić mogą nieklasyczne metody określania reguł przynależności obiektów do klas. Szczególną uwagę zwrócić tu należy na metodę rekurencyjnego podziału, której graficzną prezentacją jest drzewo decyzyjne.



Metoda rekurencyjnego podziału polega na stopniowym podziale p-wymiarowej przestrzeni cech na rozłączne podzbiory, aż do uzyskania ich homogeniczności ze względu na wyróżnioną cechę. W wyniku rekurencyjnego podziału zbiór uczący U zostaje podzielony na M rozłącznych podzbiorów U_1, U_2, \dots, U_M , zgodnie z następującą procedurą ([4]):

1. Dla danego zbioru obiektów sprawdzić, czy jest on jednorodny ze względu na wartości zmiennej zależnej lub spełnione jest inne przyjęte kryterium stopu. Jeśli tak – zakończyć postępowanie.
2. Jeśli nie – rozważyć wszystkie możliwe podziały zbioru U na rozłączne podzbiory U_1, U_2, \dots, U_M , w oparciu o wartości kolejno wybieranych zmiennych objaśniających.
3. Ocenić jakość każdego z podziałów zgodnie z przyjętym kryterium i wybrać najlepszy z nich.
4. Podzielić zbiór obiektów w wybrany sposób.
5. Kroki 1-4 wykonać rekurencyjnie dla każdego podzbioru U_1, U_2, \dots, U_M .

Procedurę podziału kończymy, jeżeli zostało osiągnięte założone kryterium stopu – zwykle jednorodność podzbiorów U_1, U_2, \dots, U_M lub określona, minimalna liczebność podzbiorów. Proces rekurencyjnego podziału zbioru U można przedstawić graficznie w postaci drzewa klasyfikacyjnego.

Wśród algorytmów tworzących drzewa klasyfikacyjne wymienić można np. algorytm CART – ang. *Classification and Regression Trees* (por. [2]), algorytm QUEST – ang. *Quick Unbiased Efficient Statistical Trees* (por. [10]), algorytm CRUISE – ang. *Classification Rule with Unbiased Interaction Selection and Estimation* (por. [6]).

Zwrócić należy uwagę na fakt, że procedury tworzenia drzew klasyfikacyjnych nie mają wymagań co do rozkładu badanych zmiennych i są odporne na obserwacje nietypowe.

Drzewa klasyfikacyjne nie stawiają warunków dotyczących skali pomiaru badanych zmiennych, a także umożliwiają klasyfikację obrazów opisanych wektorem cech z wartościami brakującymi. Uzyskane w wyniku analizy drzew klasyfikacyjnych reguły decyzyjne są proste w interpretacji, a klasyfikacja obiektów ciągu testowego nie wymaga zwykle pomiaru wszystkich cech objaśniających, co zmniejsza koszty prowadzonych analiz.

Przestawione, wybrane algorytmy rozpoznawania ze zbiorem uczącym znajdują zastosowanie w wielu konkretnych problemach badawczych z różnych dziedzin nauki, a dokładniej mówiąc – wszędzie tam, gdzie mamy do czynienia ze zbiorem wielowymiarowych obserwacji z pewnej próby, o których wiemy dokładnie, z jakich populacji (klas) pochodzą. Jedną z takich dziedzin nauki jest diagnostyka medyczna.

Reguły klasyfikacyjne w diagnostyce medycznej

Zakwalifikowanie pacjenta z chorobą wieńcową do leczenia operacyjnego jest przykładem decyzji podejmowanej w warunkach niepewności. Za ryzyko operacyjne przyjmuje się w takim przypadku prawdopodobieństwo wystąpienia mniej lub bardziej niebezpiecznych



powikłań, wynikających z bardzo różnych przyczyn, a zaistniałych jeszcze przed, podczas lub po zakończeniu operacji.

Niech rozpoznawanymi obiektami będą pacjenci Kliniki Kardiologii UM w Łodzi poddani operacji wszczepienia by-passów (CABG) w związku z chorobą wieńcową.

Obiekty należą do dwóch klas:

- ◆ klasa 1 – grupa niskiego ryzyka operacyjnego ($N_1=96$ osób);
- ◆ klasa 2 – grupa wysokiego ryzyka operacyjnego ($N_2=96$ osób).

Zestaw cech diagnostycznych, uznanych za przedoperacyjne czynniki ryzyka, przedstawia się następująco (dla uproszczenia obliczeń wykorzystano tylko zmienne mierzone na skali co najmniej porządkowej):

1. Wiek w latach;
2. BSA – wskaźnik powierzchni ciała;
3. RRs – ciśnienie skurczowe (w mmHg);
4. RRd – ciśnienie rozkurczowe (w mmHg);
5. EF% – frakcja wyrzutowa lewej komory serca (w %);
6. AspAt – aminotransferaza asparagianowa (w U/L);
7. Poziom kreatyniny (w mg/dL).

Prawdopodobieństwa błędnej klasyfikacji szacowano metodą *leave-one-out*. Do obliczeń wykorzystano:

- ◆ Pakiet *STATISTICA PL* – moduły: *Analiza dyskryminacyjna*, *Estymacja nieliniowa i Drzewa klasyfikacyjne*.
- ◆ Autorskie programy napisane w *STATISTICA Basic*, realizujące algorytmy najbliższego sąsiada, α -najbliższych sąsiadów, dyskryminacji DB z miarami odległości Euklidesa i Canberra oraz algorytm wykorzystujący liniowe i kwadratowe funkcje klasyfikacyjne z uwzględnieniem metody *leave-one-out* szacowania prawdopodobieństwa błędnych klasyfikacji.
- ◆ Udostępnione w Internecie przez autorów programy tworzące drzewa klasyfikacyjne: algorytmy – QUEST (<http://www.stat.wisc.edu/~loh/quest.html>) i CRUISE (<http://www.wpi.edu/~hkim/cruise/>).
- ◆ Program komputerowy GDM for Windows udostępniany wraz z książką Walesiaka [14].

Uzyskane wyniki przedstawia tablica 1 oraz rysunki 1 i 2. W przypadku algorytmów minimalnoodległościowych podano najlepsze otrzymane rezultaty. Dodatkowo przedstawione zostały również wyniki klasyfikacji uzyskane za pomocą metody regresji logistycznej, często stosowanej w diagnostyce medycznej.



Tablica 1. Błędne klasyfikacje dla zbioru pacjentów poddanych CABG

Algorytm rozpoznawania	Odsetek błędnych klasyfikacji [%] (metoda <i>leave-one-out</i>)		
	Niskie ryzyko operacyjne	Wysokie ryzyko operacyjne	Ogółem
Algorytm najbliższego sąsiada z miarą odległości GDM	34/96 (35,42%)	30/96 (31,25%)	33,33%
Algorytm 11 najbliższych sąsiadów z miarą odległości GDM	18/96 (18,75%)	23/96 (23,96%)	21,35%
Algorytm DB z miarą odległości Canberra	17/96 (17,71%)	19/96 (19,79%)	18,75%
Liniowe funkcje klasyfikacyjne (niezależnie od typu estymatora)	21/96 (21,88%)	19/96 (19,79%)	20,83%
Kwadratowe funkcje klasyfikacyjne (estymator zwykły)	36/96 (37,50%)	12/96 (12,50%)	25,00%
Algorytm wykorzystujący odległość Mahalanobisa	21/96 (21,88%)	19/96 (19,79%)	20,83%
CART – reguła stopu 1-SE	11/96 (11,46%)	16/96 (16,67%)	14,06%
CRUISE – reguła stopu 0-SE	11/96 (11,46%)	11/96 (11,46%)	11,46%
Regresja logistyczna	20/96 (20,83%)	20/96 (20,83%)	20,83%

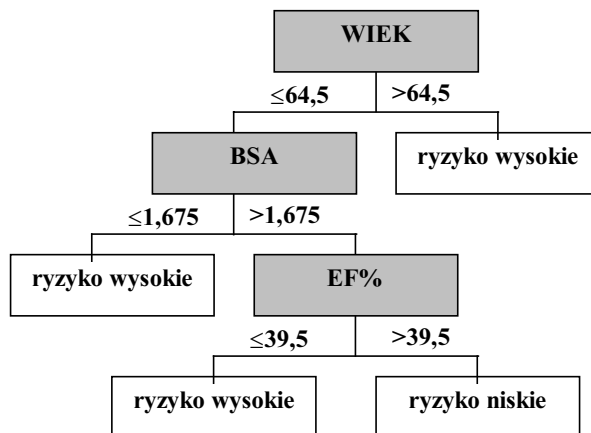
Źródło: obliczenia własne

Jak łatwo zauważyć, zdecydowanie najgorsze wyniki dostajemy dla algorytmu najbliższego sąsiada, gdzie co trzeci pacjent zostaje nieprawidłowo zaklasyfikowany.

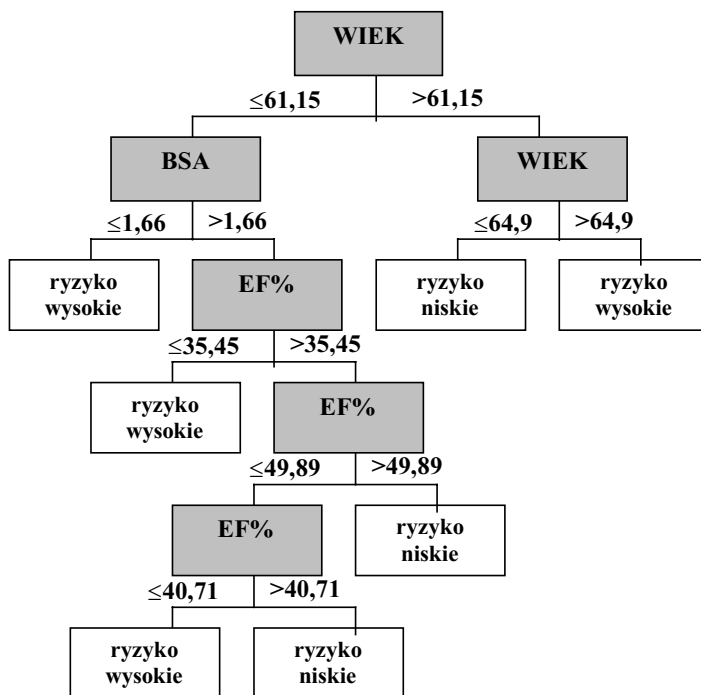
Z algorytmów bazujących na pojęciach sąsiedztwa i odległości najlepsze wyniki daje algorytm BD oparty na odległościach z miarą odległości Canberra. Odsetek błędnych klasyfikacji w tym przypadku wynosi 18,75%.

Liniowe funkcje klasyfikacyjne (niezależnie od typu estymatora), algorytm wykorzystujący odległość Mahalanobisa oraz metoda regresji logistycznej dają identyczne wyniki – 20,83% niepoprawnych zaklasyfikowań. Gorsze rezultaty daje reguła decyzyjna oparta na wartościach kwadratowych funkcji klasyfikacyjnych (estymator zwykły). 25% wszystkich pacjentów zostaje źle zdiagnozowanych. Zauważmy przy tym, że algorytm ten błędnie rozpoznaje pacjentów z grupy niskiego ryzyka – odsetek błędów wynosi 37,5%. Pacjenci z grupy wysokiego ryzyka są w większości prawidłowo rozpoznawani.

Najlepsze wyniki dają algorytmy tworzące drzewa klasyfikacyjne (por. rys. 1 i rys. 2).



Rys. 1. Drzewo klasyfikacyjne – algorytm CART; źródło: opracowanie własne



Rys. 2. Drzewo klasyfikacyjne – algorytm CRUISE; źródło: opracowanie własne

Drzewo klasyfikacyjne uzyskane w wyniku zastosowania algorytmu CART ma 4 węzły końcowe. Łatwo zauważyć, że do podziału w węzłach wykorzystano tylko trzy z siedmiu analizowanych czynników ryzyka: wiek pacjenta, wskaźnik powierzchni ciała oraz wielkość frakcji wyrzutowej lewej komory serca.



Otrzymane w wyniku zastosowania algorytmu CART reguły klasyfikacyjne można łatwo zapisać. Np. pacjentów z grupy wysokiego ryzyka można opisać jako osoby w wieku powyżej 64,5 lat lub osoby o niskim wskaźniku powierzchni ciała (nie wyższym niż 1,675) lub osoby z niską frakcją wyrzutową (co najwyżej 39,5%).

Odsetek błędnych klasyfikacji ogółem dla algorytmu CART wynosi 14,06%. Nieco gorzej jest rozpoznawana grupa pacjentów wysokiego ryzyka operacyjnego – 16,7% przy 11,5% błędnych klasyfikacji dla osób z grupy niskiego ryzyka.

Drzewo klasyfikacyjne powstałe w wyniku zastosowania algorytmu CRUISE jest nieco bardziej rozbudowane. Liczba węzłów końcowych jest równa 7, ale do podziału w węzłach wykorzystane są tylko trzy czynniki ryzyka: wiek, BSA i EF%.

Odsetek błędnych rozpoznań wynosi 11,46%. Reguły klasyfikacyjne są podobne do uzyskanych dla algorytmu CART. Pacjenci z grupy wysokiego ryzyka to osoby w wieku powyżej 64,86 lat lub o wskaźniku powierzchni ciała równym co najwyżej 1,66, lub o frakcji wyrzutowej lewej komory nie wyższej niż 40,71%.

Uwagi końcowe

Przedstawiony przykład zastosowania wybranych algorytmów rozpoznawania i uzyskane wyniki klasyfikacji wskazują, że metody te można z powodzeniem wykorzystać do wspomagania procesu podejmowania decyzji w diagnostyce medycznej. Oczywiście każda z omawianych metod tworzenia reguł decyzyjnych ma pewne wady i zalety.

W przypadku metod minimalnoodległościowych problemem może być wybór odpowiedniej miary odległości. W zasadzie nie ma reguły wskazującej najlepszą miarę. Wybór miary odległości odbywać się może tylko na drodze eksperymentalnej – z kilku sprawdzonych miar wybieramy tę, dla której dostajemy niższe odsetki błędnych klasyfikacji. Dodatkowym problemem jest tutaj wybór miary odległości dla obiektów opisanych zestawem cech mieszanych.

Zastosowanie minimalnoodległościowych algorytmów rozpoznawania wymaga od badacza przechowywania całego ciągu uczącego, bowiem klasyfikacja każdego nowego obiektu wymaga obliczenia jego odległości od wszystkich obiektów ze zbioru uczącego. Może to znacznie wydłużyć czas obliczeń.

Użyteczną metodą klasyfikacji w praktycznych zastosowaniach są liniowe funkcje klasyfikacyjne oraz metoda regresji logistycznej. Wiąże się to z dostępnością tych metod w pakietach statystycznych. Pamiętać jednak należy, że liniowe i kwadratowe funkcje klasyfikacyjne, algorytm oparty na odległościach Mahalanobisa oraz regresję logistyczną można stosować w przypadku, gdy spełnione są założenia o wielowymiarowej normalności rozkładów cech obiektów w klasach.

Wykorzystanie do analizy danych metod, dla których nie są spełnione wszystkie założenia, prowadzić może do mało wiarygodnych, a nawet błędnych wyników. Stąd też wynika potrzeba poszukiwania metod optymalnych w warunkach prowadzonych badań

empirycznych, w których najistotniejszą własnością jest odstępstwo od klasycznych założeń (np. normalności rozkładu, sposobu pomiaru cech itp.).

Szczególnie użyteczne zdają się być algorytmy tworzące drzewa klasyfikacyjne, które nie mają wymagań co do rozkładu i skali pomiaru badanych zmiennych i są odporne na obserwacje nietypowe. Uzyskane w wyniku analizy drzew klasyfikacyjnych reguły decyzyjne są proste w interpretacji, a ich graficzna prezentacja ułatwia proces podejmowania decyzji. Klasyfikacja obiektów ciągu testowego nie wymaga zwykle pomiaru wszystkich cech objaśniających, co zmniejsza koszty prowadzonych analiz. Podstawowe algorytmy budowy drzew klasyfikacyjnych (CART, QUEST) są dostępne w pakiecie *STATISTICA*.

Literatura

1. Bobrowski L. (1987), *Dyskryminacja symetryczna w rozpoznawaniu obrazów. Teoria, algorytmy, zastosowania w komputerowym wspomaganii diagnostyki medycznej*, Ossolineum, Wrocław.
2. Breiman L., Friedman J., Olshen R., Stone C. (1984), *Classification and Regression Trees*, CRC Press, London.
3. Cuadras C. M. (1989), *Distance Analysis in Discrimination and Classification Using Both Continuous and Categorical Variables*, (w:) *Statistical Data Analysis and Inference*, (Dodge ed.), Elsevier Science Publishers B. V., North Holland, s. 459-473.
4. Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa.
5. Jajuga K. (1990), *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa.
6. Kim H., Loh W.-Y. (2001), Classification Trees With Unbiased Multiway Splits, *Journal of the American Statistical Association* 96, s. 598-604.
7. Krzyśko M. (1990), *Analiza dyskryminacyjna*, WNT, Warszawa.
8. Krzyśko M. (1997), *Statystyka matematyczna, część II*, Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza, Poznań.
9. Kurzyński M. (1997), *Rozpoznawanie obiektów. Metody statystyczne*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.
10. Loh W.-Y., Shih Y.-S. (1997), Split Selection Methods for Classification Trees, *Statistica Sinica* 7, s. 815-840.
11. Misztal M. (2001), *Statystyczne metody rozpoznawania obrazów i ich zastosowania*, rozprawa doktorska, maszynopis, Łódź.
12. Rao R. C. (1982), *Modele liniowe statystyki matematycznej*, PWN, Warszawa.
13. Tadeusiewicz R., Flasiński M. (1991), *Rozpoznawanie obrazów*, PWN, Warszawa.
14. Walesiak M. (2002), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.