STUDIA INFORMATICA 2003
Volume 24 Number 1 (52)

Katarzyna STĄPOR, Alina MOMOT
Politechnika Śląska, Instytut Informatyki
Magdalena TROJNAR
Wojewódzka Przychodnia Okulistyczna, Katowice

ASYMPTOTYCZNA OPTYMALNOŚĆ W ALGORYTMACH UCZENIA ROZPOZNAWANIA OBRAZÓW

Streszczenie. W artykule przedstawiono decyzyjny problem Bayesa w sytuacji braku danych o rozkładach prawdopodobieństwa, szczegółowo opisano algorytm uczenia rozpoznawania oparty na nieparametrycznej metodzie Loftsgaardena/Quesenberry'ego oszacowania gęstości prawdopodobieństwa i wykazano jego asymptotyczną optymalność. Pokazano również zastosowanie tego algorytmu dla rozwiązania praktycznego zagadnienia klasyfikacji osób chorych na jaskrę na podstawie obrazów dna oka.

Słowa kluczowe: rozpoznawanie obrazów, decyzyjny problem Bayesa.

ASYMPTOTIC OPTIMALITY IN PATTERN RECOGNITION LEARNING ALGORITHMS

Summary. The article presents the Bayes' decision problem in which probability distributions of features are unknown, describes the pattern recognition learning algorithm based on nonparametric Loftsgaarden/Quesenberry method of estimating probability density function and proves its asymptotic optimality. The application of the algorithm to solve practical problem of digital fundus eye images classification into normal and glaucomatous ones is also shown.

Keywords: image recognition, Bayes' decision problem.

1. Model statystyczny zadania rozpoznawania

Zadanie rozpoznawania obiektu można przedstawić jako szczególny przypadek ogólnego problemu decyzyjnego Bayesa. Statystycznym problemem podejmowania decyzji jest gra (Θ, D, L), gdzie poszczególne symbole oznaczają:

Θ - zbiór możliwych stanów natury, zwany **przestrzenią parametrów,**

D - **zbiór decyzji** dostępnych statystykowi,

$L(d,\theta)$ - funkcja rzeczywista określona na produkcie kartezjańskim $D\times\Theta$, zwana **funkcją straty.**

Każda gra jest interpretowana następująco. Natura wybiera stan $\theta\in\Theta$, natomiast statystyk, który nie jest poinformowany o wyborze dokonanym przez naturę, wybiera decyzję $d\in D$. W konsekwencji tych dwóch wyborów statystyk ponosi stratę o wartości $L(d,\theta)$, w wyniku podjęcia decyzji $d\in D$, gdy θ jest „prawdziwym stanem natury". Gra jest powiązana z eksperymentem, będącym procesem zbierania informacji przez statystyka. Statystyk, nim podejmie decyzję, może obserwować wartości skalarne lub wektorowe zmiennej losowej X, której rozkład zależy od prawdziwego stanu natury θ. **Przestrzeń obserwacji** (zbiór wszystkich możliwych wyników prób n-elementowych), oznaczana przez E, jest skończenie wymiarową przestrzenią euklidesową, a rozkłady prawdopodobieństwa zmiennej losowej X są określone na podzbiorach borelowskich $B \subset E$. Dla każdego $\theta\in\Theta$ istnieje miara prawdopodobieństwa p_θ określona na B oraz odpowiednia dystrybuanta $F_\theta(x)$ (lub gęstość prawdopodobieństwa $f_\theta(x)$), która przedstawia rozkład zmiennej losowej X, gdy θ jest prawdziwą wartością parametru. Jeśli X jest wektorem p-wymiarowym, to należy traktować X jako skrócony zapis $(X_1, ..., X_p)$, a $F_\theta(x)$ jako taki zapis dla dystrybuanty wielowymiarowej $F_\theta(x_1, ..., x_p)$.

Regułą decyzyjną (algorytmem rozpoznawania) nazywamy funkcję:

$$\Psi : E \to D, \tag{1}$$

która odwzorowuje przestrzeń obserwacji E w przestrzeń decyzji D.

Na podstawie wyniku eksperymentu $X=x$ (x jest obserwowaną wartością zmiennej X) statystyk podejmuje decyzję $d=\Psi(x)\in D$.

Wartość oczekiwaną zmiennej losowej $L(\Psi(x),\theta)$, wyznaczoną przy założeniu, że θ jest prawdziwym stanem natury, nazywamy funkcją ryzyka reguły decyzyjnej Ψ:

$$R(\Psi,\theta) = E_\theta L(\Psi(x),\theta) = \int_E L(\Psi(x),\theta)dp_\theta(x) = \int_E L(\Psi(x),\theta)f_\theta(x)dx\,. \tag{2}$$

Niech Ξ jest wyróżnionym, najmniejszym σ-ciałem podzbiorów zbioru Θ, zawierającym jako swoje elementy wszystkie jednoelementowe podzbiory zbioru Θ, względem którego są mierzalne funkcje $R(d,.)$ dla każdego $d\in D$. Miara probabilistyczna p, określona na przestrzeni mierzalnej (Θ,Ξ), nazywa się rozkładem a priori parametru θ. Zbiór wszystkich rozkładów

a priori na (Θ,Ξ) będziemy oznaczać Θ^*. Każdy rozkład $p\in\Theta^*$ pozwala wprowadzić porządek w klasie reguł decyzyjnych D.

Ryzykiem bayesowskim reguły decyzyjnej $\Psi\in D$ względem rozkładu a priori $p\in\Theta^*$ nazywamy całkę:

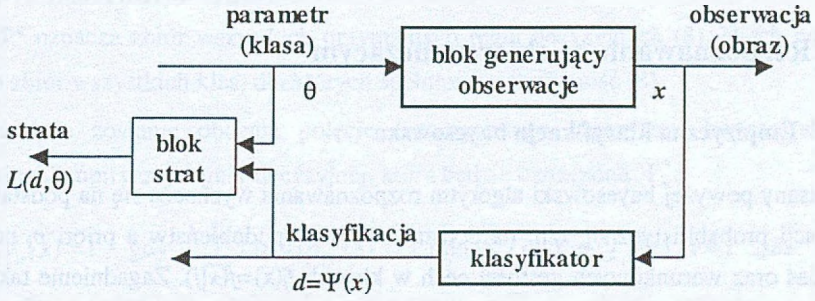
$$r(\Psi,p)=E_p(R(\Psi,\theta))=\int_{\Theta}R(\Psi,\theta)dp(\theta). \tag{3}$$

Reguła decyzyjna $\Psi^*\in D$ jest **regułą bayesowską** ze względu na rozkład a priori p, jeżeli

$$r(\Psi^*,p)=\inf_{d\in D} r(\Psi,p), \tag{4}$$

$r(\Psi^*,p)$ nazywamy **minimalnym ryzykiem bayesowskim** względem rozkładu a priori p.

Jako szczególny przypadek przedstawionego wyżej decyzyjnego problemu Bayesa zostanie przedstawione teraz **zadanie rozpoznawania (klasyfikacji).** W zadaniach rozpoznawania skończoną przestrzeń parametrów $\Theta=\{\theta_1,\ ...,\ \theta_M\}$ nazywa się przestrzenią klas, a jej elementy **klasami**. Wygodnie przy tym przyjąć $\Theta=\{1,\ ...,\ M\}$. Obserwacja nazywa się **obrazem**. Przestrzenią decyzyjną jest $D=\Theta=\{1,\ ...,\ M\}$, $L(d=\Psi(x),\ \theta)$ jest stratą spowodowaną przez zaliczenie do klasy $d=\Psi(x)$ obrazu należącego do klasy θ. Problem rozpoznawania można przedstawić tak jak na rys. 1.



Rys. 1. Blokowa interpretacja decyzyjnego problemu rozpoznawania
Fig. 1. Block scheme of the pattern recognition decision problem

Wektor wartości cech $x=(x_1,\ ...,\ x_p)$ opisujących rozpoznawany obiekt oraz wynik rozpoznawania - numer klasy j do której on należy stanowią realizację pary zmiennych losowych (θ,X). Zmienna losowa θ jest typu dyskretnego i przyjmuje wartości ze zbioru numerów klas $\Theta=(1,...,\ M\}$. Rozkład tej zmiennej losowej jest scharakteryzowany prawdopodobieństwami a priori klas p_j (tj. prawdopodobieństwami pojawiania się klas):

$$\Pr(\theta=j)=p_j,\quad j\in\Theta. \tag{5}$$

Zmienna losowa X to p-wymiarowa zmienna losowa o ciągłym charakterze. Rozkład tej zmiennej losowej dla każdej wartości $j\in\Theta$ jest określony poprzez funkcję gęstości prawdopodobieństwa $f_j(x)$ – tzw. warunkową gęstość prawdopodobieństwa cech w klasie j:

$$f(x \mid j)=f_j(x). \tag{6}$$

Niech teraz funkcja straty ma postać:

$$L(d_i,\theta_j)=\begin{cases}L(i,j) & i\neq j\\ 0 & i=j\end{cases} \qquad i,j=1,\ldots,M$$

Bayesowski algorytm rozpoznawania, czyli algorytm z wykorzystaniem bayesowskich, a więc optymalnych reguł decyzyjnych, zwany również **regułą minimalizacji ryzyka średniego,** ma następującą postać [5]:

$$\Psi^*(x)=i \quad \text{gdy} \quad \sum_{j=1}^{M} L(i,j)p_j f_j(x)=\min_{1\le i\le M}\sum_{j=1}^{M} L(i,j)p_j f_j(x) \qquad i=1,\ldots,M\,. \tag{7}$$

W przypadku $L(i,j)=1$ (tak zwana **prosta funkcja straty**) otrzymujemy następującą regułę decyzyjną:

$$\Psi^*(x)=i \quad \text{gdy} \quad p_i f_i(x)=\max_{1\le j\le M} p_j f_j(x) \qquad i=1,\ldots,M\,. \tag{8}$$

Reguła ta zwana jest **regułą największego prawdopodobieństwa a posteriori.** Sens tej reguły jest intuicyjnie oczywisty: rozpatrywany obiekt należy zaliczyć do tej klasy, która dla zaobserwowanych wartości cech x jest najbardziej prawdopodobna.

2. Rozpoznawanie ze zbiorem uczącym

2.1. Empiryczna klasyfikacja bayesowska

Opisany powyżej bayesowski algorytm rozpoznawania wyznacza się na podstawie pełnej informacji probabilistycznej, tzn. na podstawie prawdopodobieństw a priori p_j poszczególnych klas oraz warunkowych gęstości cech w klasach $f_j(x)=f(x|j)$. Zagadnienie takie nie jest jednak szczególnie interesujące z praktycznego punktu widzenia, ponieważ w wielu technicznych zadaniach nie dysponuje się niestety pełną informacją probabilistyczną. Brak tej wstępnej informacji rekompensuje się przez n-elementowy zbiór obiektów uczących, tzw. **ciąg uczący**:

$$V_n=\{(\theta_1,X_1),\ldots,(\theta_n,X_n)\}, \tag{9}$$

tzn. ciąg n niezależnych obserwacji pary zmiennych losowych (θ,X), które są wynikiem wykonanego eksperymentu, polegającego na wielokrotnym mierzeniu wejścia i wyjścia bloku generatora obserwacji (rys. 1).

Empiryczną regułą decyzyjną (empirycznym algorytmem rozpoznawania), czyli regułą opartą na zbiorze uczącym, nazywamy funkcję, która odwzorowuje przestrzeń $(\Theta\times E)^n\times E$ w przestrzeń decyzji D (zbiór klas), lub inaczej, która każdej realizacji ciągu uczącego V_n oraz elementu $x\in E$ przyporządkowuje decyzję $i\in D$:

$$\Psi_n(V_n, x) = \Psi_n(\{(\theta_1, x_1), ..., (\theta_n, x_n)\}, x) = i\,. \tag{10}$$

Prawdopodobieństwo a priori p_{in} klas ze zbioru uczącego szacuje się poprzez udział poszczególnych klas w zbiorze uczącym ilorazem:

$$p_{in} = \frac{n_i}{n} \quad i = 1, ..., M. \tag{11}$$

gdzie n_i jest zaobserwowaną w ciągu uczącym liczbą obrazów z klasy i, oraz oczywiście zachodzi $n = \sum_{j=1}^{M} n_j$. Jest to nieobciążony i zgodny estymator prawdopodobieństwa [2].

Jeśli chodzi o empiryczne szacowanie warunkowych gęstości w klasach, to wyróżnia się 2 przypadki: **parametryczny** i **nieparametryczny**. W pierwszym rozkłady prawdopodobieństwa w klasach znane są z dokładnością do skończonej i znanej liczby parametrów. W przypadku nieparametrycznym, który oznacza brak jakichkolwiek założeń co do postaci funkcyjnej warunkowych gęstości cech w klasach, f_i szacuje się za pomocą nieparametrycznego estymatora gęstości prawdopodobieństwa f_{in}, który każdej realizacji ciągu uczącego $V_n = \{(\theta_1, x_1), ..., (\theta_n, x_n)\}$ i punktowi $x \in E$ przyporządkowuje liczbę:

$$f_{in}(\theta_1, x_1, ..., \theta_n, x_n, x) \overset{df}{=} f_{in}(x)\,.$$

Niech Φ^* oznacza zbiór wszystkich optymalnych reguł decyzyjnych (8). Niech ponadto Φ_x^* oznacza zbiór wszystkich klas, dla których spełniona jest równość (8).

Wprowadzone zostanie obecnie pojęcie **empirycznego algorytmu bayesowskiego**, tzn. optymalnej, empirycznej reguły decyzyjnej, która będzie oznaczona Ψ_n^*:

$$\Psi_n^*(V_n, x) = i \quad \text{gdy} \quad \sum_{j=1}^{M} L(i,j)\, p_{jn} f_{jn}(x) = \min_i \sum_{j=1}^{M} L(i,j)\, p_{jn} f_{jn}(x)\,,\ i = 1, ..., M \tag{12}$$

lub dla przypadku prostej funkcji strat:

$$\Psi_n^*(V_n, x) = i \quad \text{gdy} \quad p_{in} f_{in}(x) = \max_{1 \le j \le M} p_{jn} f_{jn}(x)\,,\ i = 1, ..., M\,. \tag{13}$$

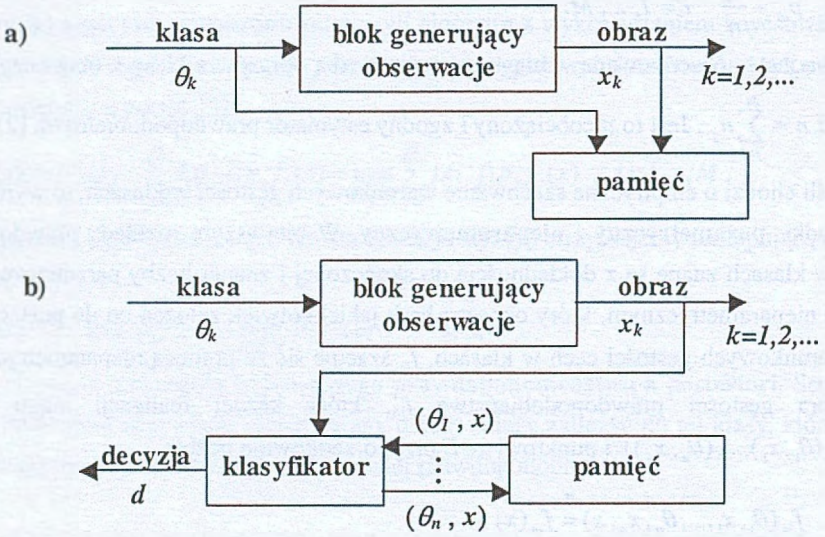
Niech T będzie klasą wszystkich empirycznych reguł rozpoznawania działających według (13). Problem podejmowania decyzji na podstawie uprzednich obserwacji $x_1, ..., x_n$, przy nieznajomości rozkładów, tzw. **empiryczny problem Bayesa** przedstawia rys. 2.

Podjęcie decyzji poprzedzone jest tzw. uczeniem z nauczycielem, tzn. obserwacją ciągu uczącego, której wyniki są zapamiętywane. Dopiero po zakończeniu cyklu uczenia następuje proces podejmowania decyzji.

W sytuacji gdy prawdopodobieństwa klas i rozkłady w klasach nie są znane, pojawia się tzw. zadanie uczenia rozpoznawania.

Algorytmem uczenia rozpoznawania (regułą uczenia rozpoznawania) nazywa się ciąg empirycznych reguł decyzyjnych [2]:

$$\{\Psi_n\} = \{\Psi_n(x, V_n)\}. \tag{14}$$



Rys. 2. Empiryczny problem decyzyjny Bayesa: a) uczenie, b) podejmowanie decyzji
Fig. 2. Empirical pattern recognition decision problem: a) learning, b) decision making

2.2. Asymptotyczna optymalność algorytmów uczenia rozpoznawania

Algorytm uczenia rozpoznawania $\{\Psi_n\}$ nazywa się **asymptotycznie optymalnym**, jeśli:

$$\Psi_n((\theta_1, x_1), \ldots, (\theta_n, x_n), x) \xrightarrow{\Pr} \Psi^*(x), \quad \text{gdy } n \to \infty.$$

Przedstawione poniżej twierdzenie podaje warunki, w których ciąg reguł decyzyjnych $\{\Psi_n\} \in T$ jest w ustalonym punkcie $x \in E$ zbieżny według prawdopodobieństwa do optymalnej reguły Bayesa lub do zbioru optymalnych reguł Bayesa, czyli jest asymptotycznie optymalny.

Twierdzenie 1

Jeżeli estymator gęstości f_{jn} jest zgodny w punkcie $x \in E$ do funkcji gęstości f_j, tzn. $f_{jn} \xrightarrow{\Pr} f_j$, gdy $n \to \infty$, dla $j \in \{1,2,\ldots,M\}$, to dla dowolnego ciągu $\{\Psi_n\} \in T$: $\lim_{n\to\infty} \Pr\left(\Psi_n(x) \in \Phi_x^*\right) = 1$.

Dowód

Na przykład w [2].

3. Metody nieparametryczne

3.1. Estymator Loftsgaardena/Quesenberry'ego

Niech $x_1, ..., x_p$ są niezależnymi obserwacjami p-wymiarowej zmiennej losowej $X=(X_1, ..., X_p)$ o ciągłej dystrybuancie $F(x_1, ..., x_p)$. Obserwacja x_i zmiennej X to $(x_{1i}, ..., x_{pi})$. Niech $k(n)$ będzie niemalejącym ciągiem liczb całkowitych dodatnich, spełniającym następujące warunki:

$$\lim_{n\to\infty} k(n) = \infty, \quad \lim_{n\to\infty} k(n)/n = 0. \tag{15}$$

Dla wyznaczenia estymatora funkcji $f(x_1, ..., x_p)$ w punkcie $z=(z_1, ..., z_p)$ należy utworzyć ciąg obszarów $S_{r_{k(n)},z}$, z których każdy jest hiperkulą o środku w punkcie z i promieniu $r_{k(n)}$. Promień $r_{k(n)}$ hiperkuli równy jest odległości punktu z, środka hiperkuli, od $k(n)$-tej najbliższej próbki x_i (obiektu ciągu uczącego), $V_{r_{k(n)},z}$ oznacza objętość hiperkuli. Tak więc każda hiperkula obejmuje wyspecyfikowaną liczbę $k(n)$ próbek. Następująca funkcja jest estymatorem gęstości f dla rozkładu zmiennej losowej elementów próby [2]:

$$f_n(z) = \frac{k(n)-1}{nV_{r_{k(n)},z}} = \{(k(n)-1)/n\}\{p\Gamma(p/2)/2r_{k(n)}^p\pi^{p/2}\}, \tag{16}$$

(gdzie Γ oznacza funkcję gamma). Jako ciąg liczbowy $k(n)$ można przyjąć np. $k(n)=(cn^{\beta})$, przy czym $c > 0$, $0< \beta<1$. Podane zostaną teraz pomocnicze lematy wykorzystywane w dowodzie zgodności estymatora (16).

Lemat 1

Dla dowolnych zdarzeń $A_1, A_2, ..., A_k$ prawdziwa jest nierówność:

$$\Pr\left(\bigcap_{i=1}^{k} A_i\right) \geq \sum_{i=1}^{k} \Pr(A_i) - k + 1 .$$

Lemat 2

Niech $\{X_n\}$ oraz $\{Y_n\}$ będą dowolnymi ciągami zmiennych losowych. Jeżeli $X_n \xrightarrow{\Pr} X$ i $Y_n \xrightarrow{\Pr} Y$, gdy $n\to\infty$, oraz istnieje takie M, że $\Pr(|X| \leq M) = 1$ i $\Pr(|Y| \leq M) = 1$, to $X_nY_n \xrightarrow{\Pr} XY$, gdy $n\to\infty$.

Lemat 3

Dla dowolnej nieujemnej, parzystej i niemalejącej dla $x>0$ funkcji f oraz dowolnej zmiennej losowej X zachodzi nierówność:

$$\Pr(|X| \geq \varepsilon) \leq \frac{Ef(X)}{f(\varepsilon)}.$$

Twierdzenie 2

Estymator gęstości f_n określony wzorem (16) jest zgodny w punktach ciągłości gęstości f, tzn. $f_n \xrightarrow{\Pr} f$, gdy $n \to \infty$.

Dowód

W pierwszej części dowodu zostanie pokazane, że zmienna losowa $U_{k(n)} = \Pr\left(S_{r_{k(n)},z}\right)$ ma rozkład beta $\mathrm{B}\left(k(n), n-k(n)+1\right)$.

Niech $Y = |X - z|$, gdzie z jest punktem ciągłości f, w którym sprawdzana jest zgodność estymatora, natomiast X zmienną losową o funkcji gęstości f. Można zauważyć, że z określenia promienia $r_{k(n)}$ wynika, że jest on $k(n)$-tą statystyką pozycyjną n-elementowej próby z rozkładu zmiennej losowej Y, tzn. $r_{k(n)} = Y_{k(n),n}$. Zatem:

$$\Pr\left(S_{r_{k(n)},z}\right) = \Pr\left(\left\{X \in S_{r_{k(n)},z}\right\}\right) = \Pr\left(|X - z| \le r_{k(n)}\right) = \Pr\left(|X - z| \le Y_{k(n),n}\right) = F_Y\left(Y_{k(n),n}\right).$$

Korzystając z faktu, że dla n-elementowej próby z rozkładu zmiennej losowej Y k-ta statystyka pozycyjna ma rozkład o dystrybuancie [1]:

$$F_{Y_{k,n}}(x) = \Pr\left(Y_{k,n} \le x\right) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F_Y(x)} t^{k-1}(1-t)^{n-k}\,dt$$

otrzymuje się:

$$F_U(u) = \Pr\left(U_{k(n)} \le u\right) = \Pr\left(\Pr\left(S_{r_{k(n)},z}\right) \le u\right) =$$
$$= \Pr\left(F_Y\left(Y_{k(n),n}\right) \le u\right) = \Pr\left(F_Y^{-1}F_Y\left(Y_{k(n),n}\right) \le F_Y^{-1}(u)\right),$$

gdzie $F_Y^{-1}(u)$ jest uogólnioną dystrybuantą odwrotną określoną wzorem $F_Y^{-1}(u) = \inf\left\{x:\ F_Y(x) \ge u\right\}$ dla $u \in (0,1)$. Natomiast:

$$\Pr\left(F_Y^{-1}F_Y\left(Y_{k(n),n}\right) \le F_Y^{-1}(u)\right) = \Pr\left(Y_{k(n),n} \le F_Y^{-1}(u)\right),$$

gdyż wartości przyjmowane przez zmienną losową $Y_{k(n),n}$ zawierają się w zbiorze wartości zmiennej losowej Y (wykorzystuje się tu bowiem nierówność $F^{-1}(F(x)) \le x$ spełnioną dla dowolnej dystrybuanty F).

Korzystając z przytoczonego wyżej wzoru opisującego postać dystrybuanty statystyki pozycyjnej otrzymujemy:

$$\Pr\left(Y_{k(n),n} \le F_Y^{-1}(u)\right) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F_Y\left(F_Y^{-1}(u)\right)} t^{k-1}(1-t)^{n-k}\,dt =$$
$$= \frac{n!}{(k-1)!(n-k)!} \int_0^{u} t^{k-1}(1-t)^{n-k}\,dt,$$

gdzie ostatnia równość jest konsekwencją faktu, że dla ciągłej dystrybuanty F_Y zachodzi równość $F_Y(F_Y^{-1}(u)) = u$ (jest to szczególny przypadek nierówności $F(F^{-1}(t)) \geq t$ spełnionej dla dowolnej dystrybuanty), co kończy pierwszą część dowodu.

W dalszej części dowodu zostanie pokazane, że $\frac{U_{k(n)}}{V_{r_{k(n)},z}} \xrightarrow{\Pr} f(z)$, gdy $n \to \infty$. Można stwierdzić, że $U_{k(n)} \xrightarrow{\Pr} 0$, gdy $n \to \infty$. Jest to konsekwencją zastosowania lematu 3 dla funkcji $f(x) = |x|$ oraz faktu, że zmienna losowa o rozkładzie $B(\alpha, \beta)$ przyjmuje tylko nieujemne wartości i jej wartość oczekiwana wynosi $\frac{\alpha}{\alpha+\beta}$ [1], tzn. dla dowolnego $\varepsilon > 0$:

$$\Pr\left(|U_{k(n)}| \geq \varepsilon\right) \leq \frac{\mathrm{E}|U_{k(n)}|}{|\varepsilon|} = \frac{\mathrm{E}U_{k(n)}}{\varepsilon} = \frac{k(n)}{\varepsilon(n+1)}.$$

Korzystając ponadto z założenia, że $\lim_{n\to\infty} \frac{k(n)}{n} = 0$, otrzymuje się:

$$\lim_{n\to\infty} \Pr\left(|U_{k(n)}| \geq \varepsilon\right) \leq \lim_{n\to\infty} \frac{k(n)}{\varepsilon(n+1)} = 0 \Leftrightarrow U_{k(n)} \xrightarrow{\Pr} 0, \text{ gdy } n \to \infty.$$

Ponieważ z jest punktem ciągłości funkcji gęstości f rozkładu zmiennej losowej X, można stwierdzić również, że:

$$U_{k(n)} \xrightarrow{\Pr} 0 \Leftrightarrow V_{r_{k(n)},z} \xrightarrow{\Pr} 0 \Leftrightarrow r_{k(n)} \xrightarrow{\Pr} 0, \text{ gdy } n \to \infty.$$

Pierwsza równoważność jest prawdziwa w przypadku, gdy nie istnieje hiperkula wokół punktu z, dla której gęstość f jest tożsamościowo równa zero. Jednak to założenie nie jest istotne, gdyż w przeciwnym przypadku $f_n(z) = \frac{k(n)-1}{nV_{r_{k(n)},z}} \leq \frac{k(n)-1}{nV} \xrightarrow{n\to\infty} 0 = f(z)$, gdzie V jest objętością tej hiperkuli.

Z faktu, że gęstość f jest ciągła w punkcie z, wynika również (na podstawie definicji):

$$f(z) = \lim_{r\to 0} \frac{\Pr(S_{r,z})}{V_{r,z}} \Leftrightarrow \forall \varepsilon > 0 \ \ \exists R \ \ \forall r \ \ r < R \Rightarrow \left| f(z) - \frac{\Pr(S_{r,z})}{V_{r,z}} \right| < \varepsilon.$$

Jeżeli więc $r_{k(n)} \xrightarrow{\Pr} 0$, to dla wybranych powyżej ε i R otrzymuje się:

$$\forall \delta > 0 \ \ \exists N \ \ \forall n \ \ n > N \Rightarrow \Pr(r_{k(n)} < R) > 1 - \delta,$$

zatem zgodnie z powyższymi założeniami:

$$\Pr\left(\left| f(z) - \frac{\Pr(S_{r_{k(n)},z})}{V_{r_{k(n)},z}} \right| < \varepsilon \right) > 1 - \delta,$$

czyli

$$\lim_{n\to\infty}\Pr\left(\left|f(z)-\frac{\Pr\left(S_{r_{k(n)},z}\right)}{V_{r_{k(n)},z}}\right|<\varepsilon\right)=1\Leftrightarrow\frac{U_{k(n)}}{V_{r_{k(n)},z}}\xrightarrow{\Pr}f(z),\text{ gdy }n\to\infty.$$

Można zauważyć, że $\frac{U_{k(n)}}{V_{r_{k(n)},z}}\xrightarrow{\Pr}f(z)\Leftrightarrow\frac{nU_{k(n)}}{k(n)-1}\cdot\frac{k(n)-1}{nV_{r_{k(n)},z}}\xrightarrow{\Pr}f(z)$, gdy $n\to\infty$. Zatem, gdy zostanie wykazane, że $\frac{nU_{k(n)}}{k(n)-1}\xrightarrow{\Pr}1$, gdy $n\to\infty$, będzie to równoznaczne ze stwierdzeniem, że $\frac{k(n)-1}{nV_{r_{k(n)},z}}\xrightarrow{\Pr}f(z)$, gdy $n\to\infty$, co stanowi tezę twierdzenia. Tak więc ostatnia część dowodu będzie pokazywała, że $\frac{nU_{k(n)}}{k(n)-1}\xrightarrow{\Pr}1$, gdy $n\to\infty$.

Korzystając z lematu 1 oraz 3 dla funkcji $f(x)=x^2$ oraz z faktu, że dla zmiennej losowej Z o rozkładzie $\mathrm{B}(\alpha,\beta)$ jej drugi moment wynosi $\mathrm{E}Z^2=\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$ [1], otrzymujemy dla dowolnego $\varepsilon>0$:

$$\Pr\left(\left|\frac{nU_{k(n)}}{k(n)-1}-1\right|\geq\varepsilon\right)\leq\frac{1}{\varepsilon^2}\mathrm{E}\left(\frac{nU_{k(n)}}{k(n)-1}-1\right)^2=$$

$$=\frac{1}{\varepsilon^2}\left(\left(\frac{n}{k(n)-1}\right)^2\mathrm{E}U_{k(n)}{}^2-\frac{2n}{k(n)-1}\mathrm{E}U_{k(n)}+1\right)=$$

$$=\frac{1}{\varepsilon^2}\left(\left(\frac{n}{k(n)-1}\right)^2\frac{k(n)\left(k(n)+1\right)}{(n+1)(n+2)}-\frac{2n}{k(n)-1}\frac{k(n)}{(n+1)}+1\right)=$$

$$=\frac{1}{\varepsilon^2}\left(\frac{n^2}{(n+1)(n+2)}\cdot\frac{k(n)\left(k(n)+1\right)}{\left(k(n)-1\right)^2}-\frac{2n}{(n+1)}\cdot\frac{k(n)}{k(n)-1}+1\right).$$

Korzystając z założenia, że $\lim_{n\to\infty}k(n)=+\infty$, mamy:

$$\lim_{n\to\infty}\Pr\left(\left|\frac{nU_{k(n)}}{k(n)-1}-1\right|\geq\varepsilon\right)\leq$$

$$\leq\frac{1}{\varepsilon^2}\lim_{n\to\infty}\left(\frac{n^2}{(n+1)(n+2)}\cdot\frac{k(n)\left(k(n)+1\right)}{\left(k(n)-1\right)^2}-\frac{2n}{(n+1)}\cdot\frac{k(n)}{k(n)-1}+1\right)=0,$$

co oznacza, że $\frac{nU_{k(n)}}{k(n)-1}\xrightarrow{\Pr}1$, gdy $n\to\infty$ i kończy dowód.□

Asymptotycznie optymalny algorytm uczenia gwarantuje, że dla wzrastającej liczby obiektów ciągu uczącego prawdopodobieństwo tego, że jego ryzyko będzie dostatecznie bliskie ryzyku bayesowskiemu, staje się dowolnie bliskie *1*. Intuicyjnie oznacza to, że dla algo-

rytmu uczenia prawdopodobieństwo faktu, że nauczy się on na podstawie licznego zbioru uczącego rozpoznawać tak dobrze jak algorytm bayesowski, jest bliskie *1*.

3.2. Algorytm rozpoznawania

Stosowanie oszacowania Loftsgaardena/Quesenberry'ego prowadzi do reguły decyzyjnej zwanej **algorytmem k-tego najbliższego sąsiada**. Dla prostej funkcji strat empiryczny algorytm bayesowski, po uwzględnieniu oszacowania gęstości (18) oraz prawdopodobieństwa a priori klas (11), ma postać:

$$p_i f_i(x) = p_i f_{in}(z) = p_i \frac{k(n_i)-1}{n_i V_{r_{k(n_i)},z}} = \frac{n_i}{n} \frac{k(n_i)-1}{n_i V_{r_{k(n_i)},z}} = \frac{k(n_i)-1}{n V_{r_{k(n_i)},z}}. \tag{17}$$

Ponieważ:

$$V_{r_{k(n)},z} = \frac{2r_{k(n)}^p \pi^{p/2}}{p\Gamma(\frac{p}{2})}, \tag{18}$$

należy maksymalizować wyrażenie:

$$\frac{k(n_i)-1}{n\frac{2r_{k(n_i)}^p \pi^{p/2}}{p\Gamma(\frac{p}{2})}} = \frac{p\Gamma(\frac{p}{2})k(n_i)-1}{2r_{k(n_i)}^p \pi^{p/2} n}. \tag{19}$$

Ze względu na to, że $\frac{p\Gamma(\frac{p}{2})}{2\pi^{p/2}n}$ nie zależy od i otrzymuje się funkcję klasyfikującą postaci:

$$\frac{k(n_i)-1}{r_{k(n_i)}^p}. \tag{20}$$

Zatem bayesowska reguła decyzyjna ma ostatecznie następującą postać:

$$\Psi_n^*(V_n, x) = i, \text{ gdy } \frac{k(n_i)-1}{r_{k(n_i)}^p} = \max_{j\in\Theta} \frac{k(n_j)-1}{r_{k(n_j)}^p}. \tag{21}$$

Przykład 1

Działanie reguły decyzyjnej (21) zostanie zilustrowane na przykładzie, który został zaczerpnięty z wykonanego w ramach pracy eksperymentu opisanego w następnym punkcie. Dla uczenia klasyfikatora wybrano 5 próbek zdrowych (klasa 0) oraz 5 chorych (klasa 1), z których każda charakteryzowana jest przez wektor cech złożony z dwóch współczynników kształtu: stosunek pole/obwód oraz współczynnik Haralicka. W tabeli 1 zamieszczono wartości wektorów cech dla tychże próbek, a także wektora cech dla próbki, na podstawie której zilustrowana zostanie procedura klasyfikacji nieparametrycznej. Próbka ta pochodzi z klasy chorych. Natomiast rys. 3 pokazuje rozmieszczenie próbek w przestrzeni cech.

Obliczone według wzoru (20) wartości funkcji klasyfikujących odpowiednio: dla klasy zdrowych(klasa 0) i chorych(klasa 1) wynoszą:

$p_0 f_0(x) = 1.406781,$

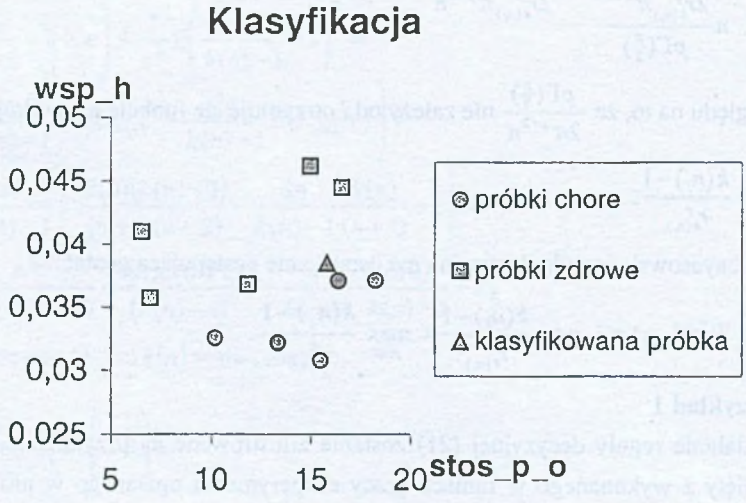
$p_1 f_1(x) = 2.860297.$

Zatem, przy zastosowaniu reguły decyzyjnej (21) próbka została poprawnie zaklasyfikowana jako chora.

Tabela 1

Wartości wektora cech dla próbek z klas zdrowych i chorych

Próbki chore		Próbki zdrowe	
stosunek pole/obwód	współczynnik Haralicka	stosunek pole/obwód	współczynnik Haralicka
13.33472	0.032191	14.9403	0.046176
18.23172	0.037139	6.52437	0.040996
15.46161	0.030787	6.950128	0.03576
16.37466	0.037113	11.84396	0.036835
10.19214	0.032582	16.52475	0.044499
15.78338	0.038519		

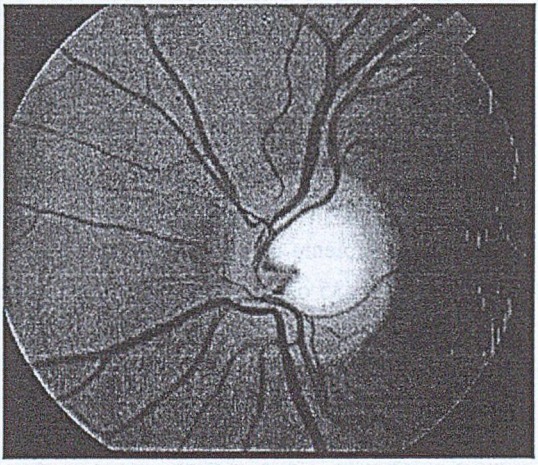


Rys. 3. Próbki uczące z dwóch klas: zdrowych i chorych wraz z próbką do klasyfikacji
Fig. 3. Learning samples from two classes: healthy and ill with sample for classifying

4. Opis eksperymentu

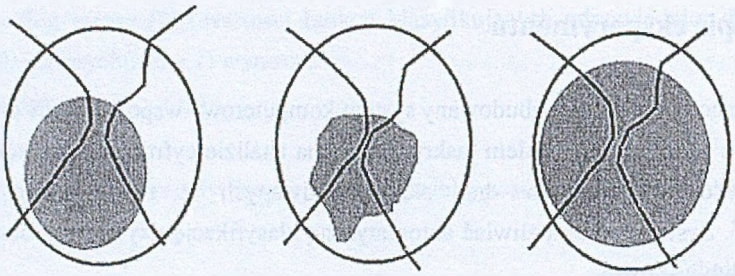
W ramach pracy został zbudowany system komputerowy wspomagający diagnostykę okulistyczną u osób z podejrzeniem jaskry, oparty na analizie cyfrowych obrazów tarczy nerwu wzrokowego na obrazach dna oka uzyskiwanych z funduskamery typu Canon CF-60Uvi. System ma umożliwiać automatyczną klasyfikację uzyskanych obrazów jako prawidłowe lub jaskrowe.

Prawidłowa tarcza nerwu wzrokowego (np. rys. 4) ma: 1) pierścień nerwowo-siatkówkowy o różowym zabarwieniu położony na obwodzie tarczy, 2) zagłębienie fizjologiczne, tj. obszar całkowicie pozbawiony włókien nerwowych, żółtawy (na rysunku jest to obszar najjaśniejszy), położony centralnie.



Rys. 4. Obraz prawidłowej tarczy nerwu wzrokowego na obrazie dna oka
Fig. 4. The image of correct disc of optical nerve on the image of eye's fundus

Jaskra to grupa schorzeń charakteryzujących się postępującą neuropatią nerwu wzrokowego, która prowadzi do powstania ubytków w polu widzenia, kończących się ostatecznie ślepotą. Jaskrowe zmiany w wyglądzie tarczy nerwu wzrokowego obejmują różnorodne zmiany kształtu pierścienia nerwowo-siatkówkowego (a co za tym idzie - zagłębienia) wskutek uszkodzeń włókien nerwowych. Przykładowe zmiany kształtu to: zlokalizowane, wklęsłe ubytki, rozległe zwężenia i zaniki pierścienia nerwowo-siatkówkowego, pionowa owalizacja zagłębienia (np. rys. 5).



Rys. 5. Przykłady zmian jaskrowych w wyglądzie tarczy nerwu wzrokowego
Fig. 5. Examples of glaucoma changes in disc of optical nerve

Zatem opis kształtu zagłębienia za pomocą odpowiednio dobranych współczynników kształtu może pozwolić na odróżnienie zdrowych i chorych (jaskrowych) tarcz.

Pozyskane obrazy dna oka, po przetwarzaniu wstępnym, które obejmowało odszumianie oraz operacje poprawy kontrastu obrazu, zostały poddane algorytmom segmentacji obszarowej, w wyniku której uzyskano obszary: tarczy nerwu wzrokowego oraz zagłębienia.

Dane z tych dwóch automatycznie rozpoznawanych obszarów tarczy są przekształcane w ilościowe współczynniki charakteryzujące ich kształty. Spośród istniejących w literaturze wytypowano do badań następujące współczynniki kształtu [3]: cyrkularności, Malinowkiej, Blaira-Blissa, Danielssona, Haralicka, Lp1, stosunek pole/obwód.

Obliczone współczynniki dla zagłębienia na obrazach dna oka w grupie osób zdrowych (28 osób) i w grupie chorych na jaskrę (35 osób) zostały poddane analizie statystycznej [10], na którą składała się:

1. analiza zgodności rozkładów (test Manna-Whitneya),
2. analiza normalności rozkładów (test Shapiro-Wilka),
3. analiza niezależności cech (test Spearmana).

W wyniku tej analizy dokonano wyboru następujących dwóch współczynników spośród wymienionych wyżej:

stosunek pole/obwód:

$$stos\_p\_o = \frac{S}{P} \tag{22}$$

oraz współczynnik Haralicka:

$$wsp\_h = \sqrt{\frac{(\sum d^2)}{n\sum d^2 - 1}}, \tag{23}$$

gdzie S oznacza pole obiektu, P – obwód obiektu, d - odległość pikseli konturu obiektu od jego środka ciężkości, n jest liczbą punktów konturu.

Współczynniki te utworzyły wektor cech zastosowany następnie do opisanej, nieparametrycznej procedury klasyfikacyjnej. Po realizacji procedury uczenia klasyfikatorów dokonano

testowania obu klasyfikatorów metodą usuwania [7]. Wyniki testów powyższą metodą są następujące (0 oznacza klasę zdrowych, 1 zaś klasę chorych):

$P(0|0) = 97{,}7\%$ (prawdopodobieństwo poprawnej klasyfikacji próbki zdrowej),

$P(1|1) = 98.8\%$ (prawdopodobieństwo poprawnej klasyfikacji próbki chorej),

$P(1|0) = 2.3\%$ (prawdopodobieństwo błędnej klasyfikacji próbki zdrowej),

$P(0|1) = 1.2\%$ (prawdopodobieństwo błędnej klasyfikacji próbki chorej).

Wnioski

1. Klasyfikator nieparametryczny jest bardziej uniwersalny w porównaniu z klasyfikatorem parametrycznym, gdyż nie jest wymagana w przypadku jego konstrukcji znajomość postaci funkcyjnej rozkładu cech w klasach – sytuacja najczęściej występująca w praktyce. Natomiast wyniki przeprowadzonego eksperymentu wskazują na jego dużą skuteczność.
2. Wyraźnie większa wartość prawdopodobieństwa błędnej klasyfikacji próbki zdrowej w stosunku do wartości prawdopodobieństwa błędnej klasyfikacji próbki chorej (prawie dwukrotnie większa) nie stanowi dużego zagrożenia, gdyż sytuacja zaklasyfikowania osoby zdrowej jako chorej nie pociąga za sobą tak negatywnych skutków jak sytuacja odwrotna.
3. Zastosowanie bardziej precyzyjnych metod segmentacji obszarowej prawdopodobnie pozwoliłoby uzyskać nieco lepsze rezultaty.

LITERATURA

1. Bartoszewicz J.: Wykłady ze statystyki matematycznej. PWN, Warszawa, 1989.
2. Devroye L.i in.: A probabilistic theory of pattern recognition. Springer Verlag, N.Y., 1998.
3. Duda R., Hart P.: Pattern classification and scene analysis. John Wiley&Sons, N.Y., 1973.
4. Fisz M.: Rachunek prawdopodobieństwa i statystyka matematyczna. PWN, Warszawa, 1969.
5. Krzyśko M.: Statystyka matematyczna. Statystyczne funkcje decyzyjne. UAM, Poznań, 1998.
6. Kanski J. i in.: Glaucoma. A color manual of diagnosis and treatment. Butterworth-Heinemann, 1996.
7. Kurzyński M.: Rozpoznawanie obiektów: metody statystyczne. Wyd. Pol. Wrocławskiej, Wrocław, 1997.
8. Loftsgaarden D. O., Quesenberry C. P.: A nonparametric estimate of a multivariate density function. Annals of Mathem. Statistics, Vol. 36, s. 11049-1051, 1965.

9. Rao C. R.: Modele liniowe statystyki matematycznej. PWN, Warszawa, 1982.
10. Zieliński R., Zieliński W.: Podręczne tablice statystyczne. PWN, Warszawa, 1987.
11. Żurada J. i in.: Sztuczne sieci neuronowe. PWN, Warszawa, 1998.

Recenzent: Dr Irena Wistuba

Wpłynęło do Redakcji 2 października 2002 r.

Abstract

In this article the general Bayes' decision problem is presented. It is examined the case which is most frequently met in practice, namely the lack of data concerning probability distributions and the pattern recognition learning problems are discussed. In the presented case the unknown probability density function is replaced by its estimate and it is given the theorem (Theorem 1) in the paper, which guarantees that if the estimate is consistent, the recognition rule based on the estimate is also consistent, i.e. convergent in probability.

In the paper the pattern recognition learning algorithm based on the nonparametric Loftsgaarden/Quesenberry method for probability density function estimation is described and its asymptotic optimality (Theorem 3) is proved. In the last part of the article the application of the presented algorithm in solving practical problem of digital fundus eye images classification into normal and glaucomatous ones is given. On the basis of statistical tests performed on data calculated on images obtained from funduscamera and appropriately processed, the feature vector is selected and next the recognition rules are created. The results of testing the classification rule by "leave-one-out" method are given.

Adresy

Katarzyna STĄPOR: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-101 Gliwice, Polska.
Alina MOMOT: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-101 Gliwice, Polska.
Magdalena TROJNAR: Wojewódzka Przychodnia Okulistyczna, ul. Powstańców 31, 40-129 Katowice, Polska.