PRACE VII KRAJOWEJ KONFERENCJI AUTOMATYKI
Rzeszów, 15 – 17 września 1977 r., tom I

Włodzimierz Greblicki, doc. dr hab.
Instytut Cybernetyki Technicznej, Politechnika Wrocławska, Wrocław

ROZPOZNAWANIE OBIEKTÓW NA PODSTAWIE OSZACOWAŃ GĘSTOŚCI PRAWDOPODOBIEŃSTWA

I. Wstęp

Rozważania nasze rozpoczniemy od przedstawienia zagadnienia rozpoznawania jako decyzyjnego problemu Bayesa. Dla prostoty założymy, że rozpoznawane obiekty należą do dwóch klas 1 i 2 oraz przez p_1 i p_2 oznaczymy aprioryczne prawdopodobieństwa każdej z nich. Decyzję o przynależności obiektu podejmuje się na podstawie wektora obserwacji $x \in X = R^k$. Gęstości w poszczególnych klasach oznaczymy przez $f_1(x)$ i $f_2(x)$. Jakość reguły decyzyjnej $\psi(x)$ oceniana jest przez ryzyko

$$R(\psi) = \sum_{j=1}^{2} p_j \int L(\psi(x), j) \, f_j(x) \, dx,$$

gdzie $L(i,j)$ jest funkcją strat, o której zakładamy, że $L(i,j)=1-\delta_{ij}$ gdzie $\delta_{i,j}$ - symbol Kroneckera. Jak wiadomo, optymalna reguła decyzyjna $\psi_0(x)$ klasyfikuje x do klasy, dla której

$$p_i f_i(x)$$

osiąga największą wartość. W całej pracy zakładamy, że aprioryczne prawdopodobieństwa p_1 i p_2 nie są znane. Prowadzi to do tzw. empirycznego problemu decyzyjnego Bayesa, patrz [8]. Zakładamy także, że nieznane są rozkłady w klasach $f_1(x)$ i $f_2(x)$.

II. Uczenie rozpoznawania jako empiryczny problem Bayesa.

Podstawą konstruowania procedur rozpoznawania jest ciąg uczący

$$(w_1,x_1), \ldots, (w_n,x_n),$$

tzn. ciąg niezależnych par (klasa, obserwacja). Dla wygody, obserwacje tego ciągu podzielimy na dwa podciągi

$$x_{11}, \ldots, x_{1n_1},$$

$$x_{21}, \ldots, x_{2n_2},$$

$n_1+n_2=n$, obserwacji pochodzących odpowiednio z klasy 1 i 2. Prawdopodobieństwa aprioryczne szacuje się następująco

$$\hat{p}_{in} = n_i/n, \quad i = 1,2.$$

Oznaczymy przez $\hat{f}_{1n}(x)$ i $\hat{f}_{2n}(x)$ oszacowania nieznanych gęstości. Załóżmy, że empiryczna regu[ł]a decyzyjna, patrz [8] , tzn. reguła uczenia rozpoznawania $\psi_n(x)$ zalicza x do klasy, dla której

$$\hat{p}_{in}\hat{f}_{in}(x)$$

osiąga największą wartość. Ciąg reguł uczenia rozpoznawania $\{\psi_n(x)\}$ nazywa się procedurą uczenia rozpoznawania.

Jest oczywiste, że podstawowym wymaganiem nałożonym na procedurę uczenia jest żądanie zbieżności ciągu zmiennych losowych $\{R(\psi_n)\}$ do minimalnego ryzyka Bayesa $R(\psi_o)$ tzn. żądanie asymptotycznej optymalności. Jest naturalne, że fakt zachodzenia tej własności zależy od własności oszacowań gęstości w poszczególnych klasach. Warunki asymptotycznej optymalności procedur uczenia rozpoznawania omawianych w tej pracy podają następujące twierdzenia, patrz [3] i [4] :

Twierdzenie 1

Jeśli oszacowania gęstości są zgodne prawie wszędzie, tzn.

$$\hat{f}_{in}(x) \to f_i(x) , \quad i = 1,2, \tag{1}$$

według prawdopodobieństwa, gdy $n \to \infty$, dla prawie wszystkich (według miary Lebesgue'a) $x \in X$, to

$$R(\psi_n) \to R(\psi_o)$$

według prawdopodobieństwa, gdy $n \to \infty$, oraz

$$\lim_{n\to\infty} ER(\psi_n) = R(\psi_o) .$$

Twierdzenie 2

Jeśli oszacowania gęstości są mocno zgodne prawie wszędzie, tzn.

$$\hat{f}_{in}(x) \to f_i(x) , \quad i = 1, 2, \tag{2}$$

z prawdopodobieństwem 1, gdy $n \to \infty$, dla prawie wszystkich (według miary Lebesgue'a) $x \in X$, to

$$R(\psi_n) \to R(\psi_o)$$

z prawdopodobieństwem 1, gdy $n \to \infty$, oraz

$$\lim_{n\to\infty} ER(\psi_n) = R(\psi_o) .$$

Rozpatrzymy teraz dwa podstawowe przypadki. W pierwszym, tzw. parametrycznym, rozkłady w klasach zależą od nieznanych parametrów, a w drugim, tzw. nieparametrycznym, rozkłady te są zupełnie nieznane.

III. Procedury parametryczne

3.3. Oszacowania najbardziej wiarogodne

Załóżmy, że rozkłady w klasach zależą od parametrów λ_1 i λ_2 w następujący sposób: $f_i(x; \lambda_i)$ oraz, że rzeczywiste wartości tych parametrów λ_{10} i λ_{20}, nie są znane.

Nieznane parametry można szacować maksymalizując funkcje wiarogodności

$$L_i(\lambda_i) = \prod_{j=1}^{n_i} f_i(x_{ij}; \lambda_i) \quad .$$

Jako oszacowania gęstości w klasach można następnie przyjąć

$$\hat{f}_{in}(x) = f_i(x; \lambda_{in}) ,$$

gdzie λ_{in} są najbardziej wiarogodnymi oszacowaniami nieznanych parametrów.

Jak wiadomo, przy dość ogólnych założeniach oszacowania te są mocno zgodne, patrz np. [2] . Wynika stąd, że przy założeniu ciągłości funkcji $f_i(x; \lambda_i)$, $i = 1,2$, także oszacowania gęstości są mocno zgodne dla wszystkich $x \in X$. Oznacza to, że spełnione jest założenie (2) , a zatem, zgodnie z twierdzeniem 2, procedura uczenia rozpoznawania jest asymptotycznie optymalna.

Jeśli ponadto oszacowania nieznanych parametrów są dostateczne, patrz [7] , to procedury te są bardzo proste pod względem obliczeniowym.

Przykład

Załóżmy , że rozkłady w klasach są normalne o nieznanych średnich m_1 i m_2 oraz znanych dyspersjach σ_1 i σ_2. Metodą największej wiarogodności otrzymuje się następujące oszacowania wartości średnich

$$\hat{m}_{in} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} ,$$

które są dostateczne i, jak to wynika z mocnego prawa wielkich liczb, są mocno zgodne. Jest więc oczywiste, że

$$\hat{f}_{in}(x) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(x-\hat{m}_{in})^2}{2\sigma_i^2}\right) \to \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(x-m_i)^2}{2\sigma_i^2}\right)$$

z prawdopodobieństwem 1, gdy $n \to \infty$, dla wszystkich $x \in X$. Stąd i z twierdzenia 2 wynika asymptotyczna optymalność procedury uczenia rozpoznawania.

3.2. Oszacowania Bayesa

Załóżmy, że rozkłady w klasach zależą od parametrów. Niech ponadto $f(\lambda_1)$ i $f(\lambda_2)$ będą apriorycznymi rozkładami tych parametrów. Jako oszacowania gęstości można przyjąć

$$\hat{f}_{in}(x) = f_i(x/x_{i1}, \ldots, x_{in_i}) =$$

$$= \int f_i(x/\lambda_i)\; f(\lambda_i/x_{i1}, \ldots, x_{in_i})\, d\lambda_i , \qquad (3)$$

przy czym rozkład a posteriori jest równy

$$f(\lambda_i/x_{i1}, \ldots, x_{in_i}) = \frac{f(\lambda_i) \prod_{j=1}^{n_i} f_i(x_{ij}/\lambda_i)}{\int [\text{licznik}]\; d\lambda_i} .$$

Jak wiadomo, przy odpowiednich założeniach rozkłady a posteriori koncentrują się wokół nieznanych parametrów, patrz [2] , a przy odpowiednich założeniach o ciągłości, estymatory gęstości są mocno zgodne dla wszystkich $x \in X$, tzn., że procedura uczenia rozpoznawania jest asymptotycznie optymalna.

Jeśli rozkłady aprioryczne reprodukują się, patrz [1] , [2] , [9], to procedury uczenia rozpoznawania są bardzo oszczędne pod względem obliczeniowym.

Przykład

Załóżmy, że rozkład w klasie i jest normalny o nieznanej średniej m_i i znanej dyspersji S_i. Rozkład apriorycznej średniej jest także normalny o średniej μ_i i dyspersji σ_i. Rozkład a posteriori średniej jest także normalny, patrz [2] , o wartości oczekiwanej

$$\frac{S_i \mu_i + n_i \sigma_i \bar{x}_i}{S_i + n_i \sigma_i} ,$$

gdzie

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} ,$$

i dyspersji

$$\frac{\sigma_i S_i}{S_i + n_i \sigma_i} .$$

Estymator wartości średniej jest więc mocno zgodny, a zatem i oszacowanie (3) jest mocno zgodne we wszystkich punktach $x \in X$.

Inne przykłady procedur wykorzystujących rozkłady reprodukujące się można znaleźć w pracach [5] i [9] .

IV. Procedury nieparametryczne

Jeżeli gęstości prawdopodobieństwa w poszczególnych klasach są całkowicie nieznane, to do ich estymacji można stosować tzw. oszacowania nieparametryczne, których bogaty przegląd można znaleźć np. w pracy [10] . Dla przykładu omówimy procedurę wykorzystującą estymator Parzena [6]

$$\hat{f}_{in}(x) = \frac{1}{n_i \, h^k(n_i)} \sum_{j=1}^{n_i} K\left(\frac{x - x_{ij}}{h(n_i)} \right) ,$$

gdzie $\{h(n)\}$ jest ciągiem liczbowym, a funkcja $K(.)$ - jądrem oszacowania. Dla pewnych ciągów np. $n^{-1/2k}$ i pewnych jąder np. $2\pi^{-k/2} \exp(-\|x\|^2/2)$, gdzie $\|x\|^2 = x^T x$, estymator ten jest zgodny w punktach ciągłości gęstości $f_i(x)$. Jeśli zatem gęstości w klasach są ciągłe dla prawie wszystkich (według miary Lebesgue'a) $x \in X$, to,

zgodnie z twierdzeniem 1, omawiana procedura rozpoznawania jest asymptotycznie optymalna.

Zauważmy, że dla podanych powyżej ciągu liczbowego i jądra obserwację x zalicza się do klasy, dla której wyrażenie

$$n_i \sum_{j=1}^{n_i} \exp\left(-\sqrt[2k]{n_i}\,\|x-x_{ij}\|^2/2\right)$$

osiąga największą wartość. Stosując ten estymator w nieco zmodyfikowanej formie otrzymuje się procedurę o identycznych własnościach, która przyporządkowuje x do klasy, dla której

$$\sum_{j=1}^{n_i} \exp\left(-\sqrt[2k]{n}\,\|x-x_{ij}\|^2/2\right)$$

jest największe, patrz [4] .

Inne, nieparametryczne procedury uczenia rozpoznawania i dokładne omówienie ich własności można znaleźć w monografii [4] .

L I T E R A T U R A

[1] N. Abramson, D. Braverman, "Learning to recognize patterns in a random environment", IRE Trans. on Information Theory, vol. IT-8, 1962, s. 58 - 63.

[2] M.M. DeGroot, "Optimal statistical decisions", Mc Graw-Hill, 1970.

[3] W. Greblicki, "Nieparametryczna estymacja w uczeniu rozpoznawania", Podstawy Sterowania, t. 2, z. 3, 1972, s. 221 - 230.

[4] W. Greblicki, "Asymptotycznie optymalne algorytmy rozpoznawania i identyfikacji w warunkach probabilistycznych", Prace Naukowe Instytutu Cybernetyki Technicznej Politechniki Wrocławskiej Nr 18, Seria: Monografie Nr 3, Wrocław 1974.

[5] D.G. Keehn, "A note on learning for gaussian properties", IEEE Trans. on Information Theory, January 1965, s. 126 - 132.

[6] E. Parzen, "On estimation of a probability density function and mode", Ann. Math. Statist., vol. 33, 1962, s. 1065 - 1076.

[7] C. R. Rao, "Linear statistical inference and its applications", Wiley, 1965.

[8] H. Robbins, "The empirical Bayes approach to statistical decision problems", Ann. Math. Statist., vol. 35, 1964, s. 1 - 20.

[9] J. Spragins, "A note on the iterative application of Bayes rule", IEEE Trans. on Information Theory, October 1965, s. 544 - 549.

[10] E.J. Wegman, "Nonparametric probability density estimation; I. A summary of available methods", Technometrics, vol. 14, No 3, 1972, s. 533 - 546.