Prace Naukowe Instytutu Cybernetyki Technicznej
Nr 1 Politechniki Wrocławskiej Nr 1

Studia i Materiały Nr 1 1972

Włodzimierz GREBLICKI

O pewnej metodzie uczenia w nieparametrycznym zadaniu rozpoznawania obrazów

W pracy podano pewien sposób uczenia rozpoznawania na podstawie znajomości ciągu uczącego. Polega on na wykorzystaniu procesu aproksymacji stochastycznej do ustalenia przybliżeń nieznanych rozkładów prawdopodobieństwa w poszczególnych klasach. Otrzymane przybliżenia stosuje się następnie w optymalnej regule bayesowskiej. Przeprowadzono także analizę dokładności oszacowania rozkładów.

WSTĘP

W zależności od stopnia znajomości procesu pojawiania się obrazów istnieją różne sposoby konstrukcji algorytmu rozpoznawania. Jeśli znana jest ilość klas, do których one należą oraz prawdopodobieństwa pojawienia się obrazu z każdej klasy i rozkłady w tych klasach, to wyznaczenie najlepszej reguły rozpoznającej nie przedstawia większych trudności. Problem znacznie się komplikuje, jeśli rozkłady i prawdopodobieństwa nie są znane. Właściwe rozpoznawania poprzedza się wówczas często tzw. cyklem uczenia, w którym obrazy są rozpoznawane przez urządzenie klasyfikujące bezbłędnie, zwane trenerem. Na podstawie tych prawidłowych rozpoznań zdobywa się pewne informacje o pojawiających się obrazach. Istnieją różne sposoby gromadzenia i wykorzystania tej informacji po cyklu uczenia, np. metoda najbliższego średniego obrazu NM, najbliższego sąsiada NN, czy też najmniejszego

przedziału LI [1]. W tej pracy podano jeszcze jeden sposób uczenia, który polega na stosowaniu metody aproksymacji stochastycznej do ustalania empirycznych przybliżeń nieznanych rozkładów na podstawie znajomości ciągu uczącego i zastosowaniu ich w tzw. optymalnej bayesowskiej regule rozpoznawania.

PRZEDSTAWIENIE PROBLEMU

Będziemy zakładać, że obiekty, które należy rozpoznać należą do M klas. Rozpoznawanie prowadzi się na podstawie pomiarów k cech tych obiektów. Wyniki pomiarów (obrazy), które oznaczymy przez x można uważać więc za elementy k-wymiarowej przestrzeni X. Będziemy dla uproszczenia pisać, że obraz x należy do klasy i, jeśli obiekt, u którego wykonano ten pomiar, należy do tej klasy.

Obiekty, a zatem i obrazy pojawiają się w sposób przypadkowy. Prawdopodobieństwo wystąpienia obrazu z klasy i oznaczymy przez p_i a gęstość prawdopodobieństwa obrazów w tej klasie przez $f_i(x)$. Jakość rozpoznawania ocenia funkcję strat $L(i,j)$, gdzie i jest numerem klasy, do której obraz zaliczono, a j numerem klasy, do której on należy. Załóżmy dalej, że funkcja strat przyjmuje wartość najmniejszą przy prawidłowym rozpoznaniu, tzn. że

$$L(i,j) \geqslant L(i,j=i).$$

Jak wiadomo, rozpoznawanie polega na przyporządkowaniu każdemu wektorowi $x \in X$ numeru klasy, do której będzie on zaliczany. Ustalenie algorytmu jest więc równoznaczne z określeniem na przestrzeni X funkcji $\Psi(x)$, która każdemu elementowi x przyporządkowuje jedną z liczb $1,2,\ldots,M$, tj. numer klasy, do której obraz x zostaje sklasyfikowany. Najlepszym algorytmem rozpoznawania $\Psi^*(x)$ jest taki, który minimalizuje ryzyko średnie

$$R\left[\Psi(x)\right] = \int_X \sum_{j=1}^{M} L(\Psi(x),j)p_j f_j(x)\,dx. \qquad (1)$$

Zatem

$$\Psi^*(x) = i,$$

jeśli

$$\sum_{j=1}^{M} L(i,j)p_j f_j(x) = \min_l \sum_{j=1}^{M} L(l,j)p_j f_j(x). \qquad (2)$$

Powyższy, bayesowski model rozpoznawania wymaga dużej informacji apriorycznej o procesie pojawiania się obrazów. Często, zarówno prawdopodobieństwa p_j, jak i rozkłady $f_j(x)$ *) nie są znane, co uniemożliwia stosowanie reguły (2). Załóżymy obecnie, że prawdopodobieństwa p_j i rozkłady $f_j(x)$ istnieją, ale nie są znane, dany jest natomiast tzw. ciąg uczący, tzn. ciąg prawidłowo rozpoznanych $\bar{n}$ obrazów. Oznaczymy go przez

$$(x_1,d_1),\ldots,(x_n,d_n),\ldots,(x_{\bar{n}},d_{\bar{n}}),$$

gdzie x_n jest kolejnym obrazem, a d_n numerem klasy, do której on należy. Inaczej mówiąc, danych jest $\bar{n}$ poprawnie rozpoznanych realizacji procesu pojawiania się obrazów. Ciąg ten jest podstawą do ustalenia empirycznych przybliżeń rozkładów $f_j(x)$ i prawdopodobieństw p_j. Po zakończeniu uczenia rozpoznawanie odbywa się według algorytmu (2), ale przy zastosowaniu empirycznych rozkładów.

USTALANIE EMPIRYCZNYCH PRZYBLIŻEŃ ROZKŁADÓW

Dla ustalenia empirycznych przybliżeń nieznanych rozkładów $f_j(x)$ zastosujemy metodę zaproponowaną przez Cypkina [2], która polega na rozwiązaniu pewnego układu równań regresji za pomocą procedury aproksymacji stochastycznej Robbinsa-Monro.

Rozkłady $f_j(x)$ będą aproksymowane liniowymi kombinacjami funkcji $\varphi_1(x),\ldots,\varphi_N(x)$. Nie tracąc na ogólności załóżymy, że funkcje te tworzą układ ortonormalny tzn. że

$$\int_X \varphi_i(x)\ \varphi_j(x)dx = \begin{cases} 0 & \text{jeśli } i \neq j \\ 1 & \text{jeśli } i = j. \end{cases}$$

Załóżymy dalej, że są one ograniczone, tzn. że

$$a_k \leqslant \varphi_k(x) \leqslant A_k \tag{3}$$

$k = 1,\ldots,N$. Przybliżenie rozkładu $f_j(x)$ wyraża się wzorem

$$\bar{f}_j(x) \overset{df}{=} \sum_{k=1}^{N} c_{jk}\,\varphi_k(x). \tag{4}$$

Jego dokładność oceniana jest zależnością

$$Q_j = \int_X \left[f_j(x) - \sum_{k=1}^{N} c_{jk}\,\varphi_k(x)\right]^2 dx. \tag{5}$$

*) Nie jest także znana postać funkcyjna gęstości $f_j(x)$ (problem nieparametryczny).

Przyrównując pochodne $\frac{\delta Q_j}{\delta c_{jk}}$ do zera otrzymuje się

$$\int_X \varphi_k(x) f_j(x)dx - c_{jk} = 0 \tag{6}$$

$j = 1,\ldots,M$; $k = 1,\ldots,N$. Współczynniki c_{jk}^o minimalizujące wyrażenie (5) są zatem równe

$$c_{jk}^o = \int_X \varphi_k(x) f_j(x)dx. \tag{7}$$

Najlepsza w sensie (5) liniowa kombinacja (4) wyraża się więc wzorem

$$\overline{f}_j^o(x) = \sum_{k=1}^{N} c_{jk}^o \varphi_k(x). \tag{8}$$

Do wyznaczenia współczynników c_{jk}^o wykorzystany jest cykl uczenia. Otrzymamy je stosując do rozwiązania NM równań regresji (6) NM niezależnych procesów aproksymacji stochastycznej Robbinsa-Monro

$$c_{jk}^{(n+1)} = c_{jk}^{(n)} - \gamma_{jk}^{(n)} \xi_j(x_n)\left[\varphi_k(x_n) - c_{jk}^{(n)}\right]. \tag{9}$$

$n = 1,2,\ldots$, przy czym x_n jest kolejnym obrazem w serii uczącej, a $\xi_j(x_n)$ jest funkcją daną wzorem

$$\xi_j(x_n) = \begin{cases} 1 & \text{jeśli } d_n = j \\ 0 & \text{jeśli } d_n \neq j, \end{cases}$$

$c_{jk}^{(n)}$ jest kolejnym oszacowaniem rozwiązania c_{jk}^o równania (6), a $\gamma_{jk}^{(n)}$ pewną liczbą. Łatwo jest sprawdzić, że równania regresji (6) spełniają nierówności podane przez Gładyszewa [4]

$$E\left\{(\varphi_k(x) - c_{jk})^2/c_{jk};\ \xi_j(x) = 1\right\} \leqslant 2c_{jk}^2 + \varphi_{jk}^2,$$

przy czym

$$\varphi_{jk}^2 = \int_X \varphi_k^2(x)\ f_j(x)dx = E\left\{\varphi_k^2(x)/\ \xi_j(x)=1\right\} \leqslant \max(a_k^2, A_k^2) < \infty \quad ,$$

oraz

$$\inf(c_{jk} - \varphi_{jk}^2)^2 > 0$$

$$\varepsilon < |c - \varphi_{jk}^2| < \frac{1}{\varepsilon}$$

dla każdego $\varepsilon > 0$. Jeśli zatem ciąg $\gamma_{jk}^{(n)}$ wybierze się tak, aby spełnione były warunki

$$\gamma_{jk}^{(n)} > 0, \quad \sum_{n=1}^{\infty} \gamma_{jk}^{(n)} = \infty, \quad \sum_{n=1}^{\infty} \left[\gamma_{jk}^{(n)}\right]^2 < \infty, \tag{10}$$

to, jak wynika z twierdzenia o zbieżności procesu Robbinsa-Monro podanego przez Gładyszewa [4]

$$c_{jk}^{(n)} \longrightarrow c_{jk}^{o}$$

dla każdego $c_{jk}^{(1)}$ z prawdopodobieństwem 1 i według średniej 2. rzędu, tzn.

$$P\left\{\lim_{n\to\infty} c_{jk}^{(n)} = c_{jk}^{o}\right\} = 1 \quad \text{i} \quad E(c_{jk}^{(n)} - c_{jk}^{o})^2 \to 0. \tag{11}$$

Po zakończeniu procesu uczenia otrzymuje się następujące przybliżenia $\bar{f}_j^{(\bar{n})}(x)$ nieznanych rozkładów $f_j(x)$:

$$\bar{f}_j^{(\bar{n})}(x) = \sum_{k=1}^{N} c_{jk}^{(\bar{n})} \varphi_k(x).$$

Relacja (11) oznacza, że także powyższe przybliżenia są zbieżne do najlepszego $f_j^o(x)$ dla każdego x, tzn.

$$\bar{f}_j^{(\bar{n})}(x) \longrightarrow \sum_{k=1}^{N} c_{jk}^{o} \varphi_k(x)$$

z prawdopodobieństwem 1 i według średniej 2. rzędu.

W celu wyznaczenia prawdopodobieństw p_j zastosujemy algorytm

$$\bar{p}_j^{(n)} = \frac{1}{n} \sum_{m=1}^{n} \zeta_j(x_m),$$

Zatem

$$\bar{p}_j^{(n)} \longrightarrow p_j$$

z prawdopodobieństwem 1.

ALGORYTM ROZPOZNAWANIA

Po zakończeniu trwającego $\bar{n}$ chwil cyklu uczenia urządzenie rozpoznające stosuje algorytm (2) z tym, że zamiast nieznanych prawdopodobieństw p_j i rozkładów $f_j(x)$ stosuje ustalone empirycznie ich przybliżenia $\bar{p}_j(\bar{n})$ i $\bar{f}_j^{(\bar{n})}(x)$. Algorytm rozpoznawania przeprowadza się następująco

$$\Psi(x) = i$$

jeśli

$$\sum_{j=1}^{M} L(i,j)\bar{p}_j^{(\bar{n})}\bar{f}_j^{(\bar{n})}(x) = \max_{l} \sum_{j=1}^{M} L(l,j)\bar{p}_j^{(\bar{n})}\bar{f}_j^{(\bar{n})}(x).$$

OCENA PROCESU UCZENIA

Przedstawione własności asymptotyczne procesu Robbinsa-Monro, zastosowanego do ustalania przybliżeń rozkładów, są jednak niewystarczające do oceny dokładności aproksymacji po wykonaniu skończonej ilości kroków procedury (9), tzn. po zakończeniu cyklu uczenia. Celowe jest zatem dokonanie oceny jakości oszacowania po każdym kroku, co umożliwiłoby określenie długości serii uczącej, niezbędnej do osiągnięcia wymaganej dokładności.

Oznaczmy przez $Q_j^{(n)}$ dokładność n-go przybliżenia zgodnie ze wzorem (5)

$$Q_j^{(n)} = E\left\{\int_X \left[f_j(x) - \sum_{k=1}^{N} c_{jk}^{(n)} \varphi_k(x)\right]^2 dx\right\} = Q_j^o + \sum_{k=1}^{N} E\left\{\left[c_{jk}^o - c_{jk}^{(n)}\right]^2\right\}, \quad (12)$$

przy czym Q_j^o jest dokładnością najlepszej aproksymacji, tzn.

$$Q_j^o = \int_X \left[f_j(x) - \sum_{k=1}^{N} c_{jk}^o \varphi_k(x)\right]^2 dx. \quad (13)$$

Z zależności (11) wynika, że

$$Q_j^{(n)} \longrightarrow Q_j^o .$$

Szybkość tej zbieżności zależy oczywiście od ciągów $\gamma_{jk}^{(n)}$ oraz dokładności pierwszego oszacowania liczb c_{jk}^o, którą można określić w podany poniżej sposób.

Zauważmy, że z nierówności (3) i wzoru (7) otrzymuje się

$$\left| c - c_{jk}^o \right| = \left| c - \int_X \varphi_k(x) f_j(x) dx \right| \leqslant \int_X \left| c - \varphi_k(x) \right| f_j(x) dx \leqslant \frac{1}{2} \left| A_k - a_k \right|$$

dla

$$c = \frac{1}{2} (A_k + a_k).$$

Jeśli teraz $c_{jk}^{(1)}$ wybierze się zgodnie ze wzorem

$$c_{jk}^{(1)} = \frac{1}{2} (A_k + a_k) , \quad (14)$$

to

$$(c_{jk}^{(1)} - c_{jk}^o)^2 \leqslant \frac{1}{4} (A_k - a_k)^2 . \quad (15)$$

Wśród wszystkich ciągów $\gamma_{jk}^{(1)},\ldots,\gamma_{jk}^{(n)}$ istnieje oczywiście ciąg minimalizujący

$$E(c_{jk}^{(n)} - c_{jk}^{o})^2.$$

Ponieważ procesy aproksymacji stochastycznej są dla wszystkich c_{jk} niezależne, to z wzoru (12) wynika, że te same ciągi minimalizują interesującą nas dokładność aproksymacji $Q_j^{(n)}$. Jak wykazał Dvoretzky [3], ciąg taki wyraża się zależnością

$$\gamma_{jk}^{(n)} = \frac{b_k}{\varphi_{jk}^2 + n b_k}, \tag{16}$$

przy czym

$$b_k = \frac{1}{4}(A_k - a_k)^2$$

jest dokładnością pierwszego oszacowania daną nierównością (15). Dla tego ciągu otrzymuje się

$$E(c_{jk}^{(n)} - c_{jk}^{o})^2 \leqslant \frac{b_k \varphi_{jk}^2}{\varphi_{jk}^2 + (n-1)b_k}. \tag{17}$$

Ciąg (16) jest optymalny w tym sensie, że dla każdego innego ciągu, dla którego istnieje pierwsze oszacowanie spełniające nierówność (15), nie zachodzi relacja (16).

Do stosowania optymalnego ciągu (16) konieczna jest znajomość φ_{jk}^2. W przypadku, gdy przestrzeń obrazów X jest jednowymiarowa i funkcje $\Phi_1(x),\ldots,\Phi_N(x)$, wybrano ortonormalizując układ

$$\Phi_k(x) = \begin{cases} t^{k-1} & \text{dla } -T \leqslant t \leqslant T \\ 0 & \text{dla } t < -T, \text{lub } t > T, \end{cases}$$

przy czym $k = 1,\ldots,N$, to przy znajomości momentów rozkładów w poszczególnych klasach można korzystać ze wzoru (16). Otrzymuje się wtedy

$$Q_j^{(n)} \leqslant Q_j^{o} + \sum_{k=o}^{N} \frac{b_k E\{\varphi_k^2(x)/\xi_j(x) = 1\}}{E\{\varphi_k^2(x)/\xi_j(x)=1\} + (n-1)b_k}. \tag{18}$$

Jeśli jednak nie posiada się takiej informacji apriorycznej, lub układ ortonormalny wybrano inaczej, to stosowanie ciągu (16) jest niemożliwe. Zauważmy jednak, że z wzoru (3) wynika, że

$$\varphi_{jk}^2 = \int_X \varphi_k^2(x) f_j(x)\,dx \leqslant \max_x \varphi_k^2(x) = \max(A_k^2, a_k^2) \stackrel{df}{=} \bar{a}_k .$$

Jeśli teraz zastosować

$$\gamma_{jk}^{(n)} = \frac{b_k \bar{a}_k}{\bar{a}_k + n b_k} \tag{19}$$

to, jak łatwo sprawdzić

$$E(c_{jk}^{(n)} - c_{jk}^{o})^2 \leqslant \frac{\bar{a}_k b_k}{\bar{a}_k + (n-1) b_k}$$

i

$$Q_j^{(n)} \leqslant Q_j^o + \sum_{k=1}^{N} \frac{\bar{a}_k b_k}{\bar{a}_k + (n-1) b_k} . \tag{20}$$

Zauważmy na koniec, że ciąg $\gamma_{jk}^{(n)}$ dany wzorem (19) nie zależy od wskaźnika j. Zatem w procesach (9) aproksymacji stochastycznej współczynniki $\gamma_{jk}^{(n)}$ nie zależą od klasy j, dla której ustalana jest gęstość, lecz jedynie od numeru k ustalającego pozycję współczynnika c_{jk} w liniowej kombinacji (4).

ZAKOŃCZENIE

Podaną zasadę uczenia rozpoznawania można stosować przy zupełnym braku informacji w procesie pojawiania się obrazów. Na podstawie nierówności (20) można ocenić dokładność przybliżenia nieznanych rozkładów. Przy niewielkiej informacji i odpowiednim wyborze układu funkcji ortonormalnych można nawet optymalizować proces.

Interesujący jest także problem wyboru funkcji ortonormalnych $\Phi_1(x), \ldots, \Phi_N(x)$. Załóżmy, że w nierówności (3)

$$a_k = -A_k .$$

Jeśli teraz, zgodnie z wzorem (14), wybrać $c_{jk}^{(1)} = 0$, to

$$(c_{jk}^{(1)} - c_{jk}^{o})^2 \leqslant \frac{A_k^2}{4} \stackrel{df}{=} b_k .$$

Z nierówności (20) wynika, że jeśli zastosuje się ciąg $\gamma_{jk}^{(n)}$ zgodnie z (19), to

$$E(c_{jk}^{(n)} - c_{jk}^{o})^2 \leqslant \frac{A_k^2}{4n} ,$$

tzn.

$$Q_j^{(n)} \leqslant Q_j^o + \frac{1}{4n} \sum_{k=1}^{N} A_k^2 .$$

Układ funkcji ortonormalnych należy więc wybrać tak, aby osiągnąć możliwie małe liczby A_k, bowiem jak widać z ostatniej nierówności, zapewnia to lepszą zbieżność.

LITERATURA

[1] B u b n i c k i Z., Least interval pattern recognition and its application in control systems, Materiały IV Kongresu IFAC, sekcja nr 21, Warszawa 1969.

[2] C y p k i n J. Z., Primenenije metoda stochastičeskoj approksimacji neizvestnoj plotnosti raspredelenija po nabliudenijam, Avtomatika i Telemechanika, 3, 1966.

[3] D v o r e t z k y A., On stochastic approximation, Proc. III Berkeley Symp. Math. Statistics and Probability, 1, 1956.

[4] G l a d y š e v E., O stochastičeskoj approksimacji, Teorija verojatnostej i ejo primenenija, X, 1965.

ON A CERTAIN METHOD OF LEARNING PATTERN RECOGNITION IN NONPARAMETRIC PROBLEM

In the paper a certain method of learning to recognize pattern is given. A stochastic approximation to estimate the unknown distribution densities is applied. The obtained estimations in the Bayes decision rule is used. Analysis of accuracy of the method is carried out.

О НЕКОТОРОМ МЕТОДЕ ОБУЧЕНИЯ В НЕПАРАМЕТРИЧЕСКОЙ ЗАДАЧЕ РАСПОЗНАВАНИЯ ОБРАЗОВ

В статье предлагается некоторый новый подход к обучению разпознаванию образов. К определению эмпирических распределений вероятности применяется метод стохастическрй аппроксимации. Полученные оценки распределений применяются в Баесовам законе распознавания.