PODSTAWY STEROWANIA – TOM 2 – ZESZYT 3 – 1972

Nieparametryczna estymacja w uczeniu rozpoznawania

WŁODZIMIERZ GREBLICKI

(Maszynopis wpłynął 25 kwietnia 1972)

1. Wstęp

Uczenie rozpoznawania polega na gromadzeniu informacji statystycznej o rozkładach w poszczególnych klasach i odpowiednich prawdopodobieństwach. Regułę decyzyjną otrzymuje się na podstawie prawidłowo rozpoznanych obiektów tzn. ciągu uczącego. W pracy tej rozpatrzymy własności reguł decyzyjnych, które otrzymuje się przez nieparametryczną estymację nieznanych gęstości prawdopodobieństwa i wykażemy, że gdy długość ciągu uczącego rośnie, reguła decyzyjna zdąża do reguły Bayesa, a ryzyko do ryzyka Bayesa. Otrzymane w ten sposób algorytmy uczenia rozpoznawania są też bezwzględnie, asymptotycznie optymalne [5].

2. Wprowadzenie

Obiekty należące do M klas rozpoznaje się na podstawie pomiarów, które tworzą wektor x w przestrzeni wektorowej X. Prawdopodobieństwo pojawienia się obiektu z klasy i wynosi p_i, a gęstość prawdopodobieństwa w tej klasie $f_i(x)$.

$L(i,j)$ jest stratą poniesioną przez zaliczenie obiektu z klasy j do klasy i. Ryzyko jest więc równe

$$R[\psi(x)] = \sum_{j=1}^{M} p_j \int_x L(\psi(x), j) f_j(x)\, dx\,,$$

przy czym $\psi(x)$ jest regułą decyzyjną. Wiadomo, że optymalna reguła decyzyjna $\psi^*(x)$ zalicza obiekt o pomiarze x do klasy i, dla której ryzyko warunkowe

$$\sum_{j=1}^{M} L(i,j) p_j f_j(x) \tag{1}$$

jest najmniejsze. Dla uniknięcia niejednoznaczności założymy, że ryzyko to osiąga minimum dokładnie dla jednego indeksu dla prawie wszystkich $x \in X$.

Rozpoznawanie prowadzi się w sytuacji, gdy ani prawdopodobieństwa, ani rozkłady w poszczególnych klasach nie są znane. Jedyną informacją jest ciąg uczący

$$(x_1, j_1), (x_2, j_2), \ldots, (x_n, j_n),$$

gdzie j_k jest numerem klasy, do której należy obiekt o pomiarze x_k. Zarówno kolejne pomiary, jak i numery klas są stochastycznie niezależne.

Ciąg uczący wykorzystuje się do estymacji prawdopodobieństw p_i i gęstości $f_i(x)$. Prawdopodobieństwa p_i szacuje się ułamkami

$$p_{in} = \frac{n_i}{n}, \tag{2}$$

przy czym n_i jest ilością obiektów z klasy i, natomiast gęstości $f_i(x)$ ocenia się przy pomocy nieparametrycznych estymatorów $f_{in}(x)$ [1, 2, 3]. Przykładem algorytmu uczenia otrzymanego w ten sposób jest LI, w którym na podstawie ciągu uczącego konstruuje się odpowiednie gęstości empiryczne [5]. W tej pracy rozważymy ogólne własności algorytmów uczenia, w których wykorzystuje się estymatory zbieżne do $f_i(x)$ według prawdopodobieństwa i z prawdopodobieństwem 1.

Reguła decyzyjna $\psi_n(x)$ otrzymana na podstawie ciągu uczącego klasyfikuje obiekt x do klasy i, dla której

$$\sum_{j=1}^{M} L(i,j) p_{jn} f_{jn}(x) \tag{3}$$

osiąga minimum.

3. Zbieżność reguły decyzyjnej i ryzyka

Jak już wspomniano, rozpatrzymy dwa typy nieparametrycznych estymatorów gęstości prawdopodobieństwa. Twierdzenie podane poniżej dotyczy przypadku, gdy estymator jest zbieżny według prawdopodobieństwa.

TWIERDZENIE 1

Jeśli dla pewnego $x \in X$ oraz wszystkich i

$$f_{in}(x) \to f_i(x) \text{ według prawdopodobieństwa, gdy } n \to \infty, \tag{4}$$

to

$$\lim_{n\to\infty} P\{\psi_n(x) = \psi^*(x)\} = 1 \tag{5}$$

oraz

$$\lim_{n\to\infty} E|\psi_n(x) - \psi^*(x)|^2 = 0. \tag{6}$$

Jeżeli (4) spełnione jest dla prawie wszystkich $x \in X$, to

$$\lim_{n\to\infty} E\,|R[\psi_n(x)]-R[\psi^*(x)]|^2 = 0 \tag{7}$$

i

$$\lim_{n\to\infty} E\int_{\Omega} |\psi_n(x)-\psi^*(x)|^2 dx = 0 \tag{8}$$

dla każdego ograniczonego zbioru $\Omega \subset X$.

Dowód: Zauważmy najpierw, że

$$P\{A \cap B\} \geqslant P\{A\}+P\{B\}-1\,. \tag{9}$$

Ustalmy teraz $x \in X$ i załóżmy, że (4) spełnione jest w tym punkcie dla wszystkich i oraz (1) osiąga minimum dokładnie dla jednego wskaźnika i_0. Wynika stąd, że istnieje $\varepsilon > 0$ takie, że

$$\sum_{j=1}^{M} L(i,j)p_j f_j(x) - \sum_{j=1}^{M} L(i_0,j)p_j f_j(x) > 3\varepsilon \tag{10}$$

dla wszystkich $i \neq i_0$.

Z (4) wynika z kolei, że dla dowolnie małego $\delta > 0$ istnieje N_1 takie, że

$$P\left\{\max_i |f_{in}(x)-f_i(x)| < \frac{\varepsilon}{2L}\right\} > 1-\frac{\delta}{4} \tag{11}$$

dla $n > N_1$, przy czym

$$L = \max_{i,j} L(i,j)\,. \tag{12}$$

Jest oczywiste, że istnieje N_2 takie, że

$$P\left\{\max_i |p_{in}-p_i| < \frac{\varepsilon}{2LfM}\right\} > 1-\frac{\delta}{4} \tag{13}$$

dla $n > N_2$, gdzie

$$f = \max_i f_i(x)\,.$$

Na podstawie (2), (9), (11) i (13) wnioskujemy, że dla $n > N \overset{\text{df}}{=} \max(N_1, N_2)$ (dla uproszczenia zapisu użyto oznaczeń f_i oraz f_{in} zamiast $f_i(x)$ oraz $f_{in}(x)$)

$$P\left\{\left|\sum_{j=1}^{M} L(i,j)p_{jn}f_{jn} - \sum_{j=1}^{M} L(i,j)p_j f_j\right| < \varepsilon\right\} =$$

$$= P\left\{\left|\sum_{j=1}^{M} L(i,j)p_{jn}(f_{jn}-f_j) + \sum_{j=1}^{M} L(i,j)(p_{jn}-p_j)f_j\right| < \varepsilon\right\} \geqslant$$

$$\geqslant P\left\{\left|\sum_{j=1}^{M} L(i,j)p_{jn}(f_{jn}-f_j)\right| < \frac{\varepsilon}{2}\,,\ \left|\sum_{j=1}^{M} L(i,j)(p_{jn}-p_j)f_j\right| < \frac{\varepsilon}{2}\right\} \geqslant$$

$$\geqslant P\left\{L\max_i|f_{in}-f_i|<\frac{\varepsilon}{2},\ ML\max_i|p_{in}-p_i|<\frac{\varepsilon}{2}\right\}=$$

$$=P\left\{\max_i|f_{in}-f_i|<\frac{\varepsilon}{2L},\ \max_i|p_{in}-p_i|<\frac{\varepsilon}{2MLf}\right\}\geqslant$$

$$\geqslant\left(1-\frac{\delta}{4}\right)+\left(1-\frac{\delta}{4}\right)-1=1-\frac{\delta}{2}.$$

Dla $n>N$, na podstawie (9), (10) i ostatniej nierówności znajdujemy, że

$$P\left\{\sum_{j=1}^{M}L(i,j)p_{jn}f_{jn}-\sum_{j=1}^{M}L(i_0,j)p_{jn}f_{jn}>\varepsilon\right\}=$$

$$=P\left\{\left[\sum_{j=1}^{M}L(i,j)p_{jn}f_{jn}-\sum_{j=1}^{M}L(i,j)p_jf_j\right]+\left[\sum_{j=1}^{M}L(i,j)p_jf_j-\right.\right.$$

$$\left.\left.-\sum_{j=1}^{M}L(i_0,j)p_jf_j\right]+\left[\sum_{j=1}^{M}L(i_0,j)p_jf_j-\sum_{j=1}^{M}L(i_0,j)p_{jn}f_{jn}\right]>\varepsilon\right\}\geqslant$$

$$\geqslant P\left\{\left|\sum_{j=1}^{M}L(i,j)p_{jn}f_{jn}-\sum_{j=1}^{M}L(i,j)p_jf_j\right|<\varepsilon,\ \sum_{j=1}^{M}L(i,j)p_jf_j-\right.$$

$$\left.-\sum_{j=1}^{M}L(i_0,j)p_jf_j>3\varepsilon,\ \left|\sum_{j=1}^{M}L(i_0,j)p_jf_j-\sum_{j=1}^{M}L(i_0,j)p_{jn}f_{jn}\right|<\varepsilon\right\}\geqslant$$

$$\geqslant\left(1-\frac{\delta}{2}\right)+\left(1-\frac{\delta}{2}\right)-1=1-\delta$$

dla wszystkich $i\neq i_0$.

Wykazaliśmy więc, że dla dowolnie małego $\delta>0$ istnieje N takie, że

$$P\left\{\sum_{j=1}^{M}L(i,j)p_{jn}f_{jn}(x)-\sum_{j=1}^{M}L(i_0,j)p_{jn}f_{jn}(x)>\varepsilon\right\}>1-\delta$$

dla wszystkich $i\neq i_0$ tzn., że

$$P\{\psi_n(x)=\psi^*(x)\}>1-\delta\,.$$

Udowodniliśmy zatem (5).

Ponieważ $0<\psi(x)\leqslant M$, to

$$E|\psi_n(x)-\psi^*(x)|^2\leqslant M^2P\{\psi_n(x)\neq\psi^*(x)\}\,. \tag{14}$$

Stąd

$$\lim_{n\to\infty}E|\psi_n(x)-\psi^*(x)|^2\leqslant M^2\lim_{n\to\infty}P\{\psi_n(x)\neq\psi^*(x)\}=0\,,$$

co oznacza prawdziwość (6).

W celu udowodnienia (7) skorzystamy z (14). Na podstawie nierówności Schwartza otrzymuje się

$$E|R[\psi_n(x)]-R[\psi^*(x)]|^2 = E\Big|\sum_{j=1}^{M} p_j \int_x [L(\psi_n(x),j)-L(\psi^*(x),j)]f_j(x)\,dx\Big|^2 \leqslant$$

$$\leqslant M\sum_{j=1}^{M}\int_x E|L(\psi_n(x),j)-L(\psi^*(x),j)|^2 f_j(x)\,dx \leqslant$$

$$\leqslant ML^2\sum_{j=1}^{M}\int_x P\{\psi_n(x)\neq\psi^*(x)\}f_j(x)\,dx\,. \tag{15}$$

Z (5), nierówności $|P\{\psi_n(x)\neq\psi^*(x)\}|\leqslant 1$ i twierdzenia o przechodzeniu do granicy pod znakiem całki wynika natomiast, że [6]

$$\lim_{n\to\infty}\int_x P\{\psi_n(x)\neq\psi^*(x)\}f_j(x)\,dx = \int_x \lim_{n\to\infty} P\{\psi_n(x)\neq\psi^*(x)\}f_j(x)\,dx = 0\,,$$

ponieważ (5) jest spełnione prawie wszędzie. Zatem przechodząc do granicy po prawej stronie nierówności (15) znajduje się (7).

Ze wspomnianego powyżej twierdzenia granicznego i nierówności $E|\psi_n(x)-\psi^*(x)|^2\leqslant M^2$ otrzymuje się z kolei (8), co kończy dowód.

Zauważmy teraz, że jeśli istnieje ograniczony zbiór $\Omega\subset X$ taki, że $P\{x\in\Omega\}=1$, co oznacza, że pomiary obiektów należą do zbioru Ω, to udowodniona powyżej własność (8) oznacza, że

$$\lim_{n\to\infty} E\int_x |\psi_n(x)-\psi^*(x)|^2 dx = 0\,.$$

Reguła decyzyjna jest wówczas zbieżna do reguły Bayesa także w średnim, całkowo kwadratowym sensie.

Udowodniliśmy więc, że jeśli estymator gęstości prawdopodobieństwa jest zgodny, to prawdopodobieństwo rozpoznawania zgodnie z regułą Bayesa zdąża w procesie uczenia do jedności. Reguła decyzyjna jest ponadto zbieżna do reguły Bayesa także według średniej drugiego rzędu, a ryzyko do ryzyka Bayesa w tym samym sensie.

Interesujący jest przypadek, gdy $L(i,j)$ jest zero-jedynkową funkcją strat, tzn. gdy

$$L(i,j) = \begin{cases} 0 & \text{jeśli } i=j \\ 1 & \text{jeśli } i\neq j \end{cases}.$$

Ryzyko Bayesa jest wówczas, jak wiadomo, najmniejszym prawdopodobieństwem błędnej klasyfikacji. Jest oczywiste, że $R[\psi_n(x)]$ jest natomiast prawdopodobieństwem błędnej klasyfikacji dla reguły $\psi_n(x)$. Zatem z (7) wynika, że

$$P\{\text{błędnej klasyfikacji}\}\to P\{\text{błędnej klasyfikacji dla reguły Bayesa}\}$$

gdy $n\to\infty$.

Rozpatrzymy teraz drugi przypadek, gdy estymator gęstości prawdopodobieństwa jest zbieżny z prawdopodobieństwem 1.

TWIERDZENIE 2

Jeśli dla pewnego $x \in X$ oraz wszystkich i

$$f_{in}(x) \to f_i(x) \text{ z prawdopodobieństwem 1, gdy } n \to \infty, \tag{16}$$

to

$$P\{\lim_{n\to\infty} \psi_n(x) = \psi^*(x)\} = 1 \tag{17}$$

oraz

$$\lim_{n\to\infty} E|\psi_n(x) - \psi^*(x)|^2 = 0 \tag{18}$$

Ponadto, jeśli (16) spełnione jest dla prawie wszystkich $x \in X$, to

$$P\{\lim_{n\to\infty} R[\psi_n(x)] = R[\psi^*(x)]\} = 1\,, \tag{19}$$

$$\lim_{n\to\infty} E|\psi_n(x) - \psi^*(x)|^2 = 0 \tag{20}$$

oraz

$$\lim_{n\to\infty} E \int_{\Omega} |\psi_n(x) - \psi^*(x)|^2 dx = 0 \tag{21}$$

dla każdego ograniczonego zbioru $\Omega \subset X$.

Dowód: Jest zupełnie oczywiste, że z (16) wynika zbieżność (4). Zatem z twierdzenia 1 i założeń wynika prawdziwość relacji (18), (20) i (21).

Udowodnimy teraz (17). Ustalmy w tym celu $x \in X$ i załóżmy, że (16) jest spełnione w tym punkcie dla wszystkich i, oraz że (1) zachodzi dla jednego wskaźnika i_0. Wynika stąd, że istnieje $\varepsilon > 0$ takie, że

$$\sum_{j=1}^{M} L(i,j) p_j f_j(x) - \sum_{j=1}^{M} L(i_0,j) p_j f_j(x) > 3\varepsilon\,. \tag{22}$$

Na podstawie (2) i (16) otrzymujemy natomiast

$$v_{in} \overset{\text{df}}{=} \sum_{j=1}^{M} L(i,j) p_{jn} f_{jn}(x) - \sum_{j=1}^{M} L(i,j) p_j f_j(x) \to 0$$

z prawdopodobieństwem 1, gdy $n \to \infty$, skąd wynika, że istnieje N takie, że

$$|v_{in}| < \varepsilon$$

z prawdopodobieństwem 1 dla wszystkich i oraz $n > N$. Zatem dla $n > N$

$$\sum_{j=1}^{M} L(i,j) p_{jn} f_{jn} - \sum_{j=1}^{M} L(i_0,j) p_{jn} f_{jn} = \Big[\sum_{j=1}^{M} L(i,j) p_{jn} f_{jn} -$$
$$- \sum_{j=1}^{M} L(i,j) p_j f_j\Big] + \Big[\sum_{j=1}^{M} L(i,j) p_j f_j - \sum_{j=1}^{M} L(i_0,j) p_j f_j\Big] +$$
$$+ \Big[\sum_{j=1}^{M} L(i_0,j) p_j f_j - \sum_{j=1}^{M} L(i_0,j) p_{jn} f_{jn}\Big] > v_{in} + 3\varepsilon + v_{i_0 n} > \varepsilon$$

z prawdopodobieństwem 1, skąd wynika (17).

Z udowodnionej już relacji (17) wynika, że

$$P\{\lim_{n\to\infty} L(\psi_n(x), j) = L(\psi^*(x), j)\} = 1 \tag{23}$$

dla wszystkich j. Zatem

$$|R[\psi_n(x)] - R[\psi^*(x)]| = \sum_{j=1}^{M} p_j \left| \int_x [L(\psi_n(x), j) - L(\psi^*(x), j)] f_j(x)\, dx \right| \leqslant$$

$$\leqslant M \sum_{j=1}^{M} \int_x |L(\psi_n(x), j) - L(\psi^*(x), j)| f_j(x)\, dx .$$

Stąd, z nierówności $\sum_{j=1}^{M} |L(\psi_n(x), j) - L(\psi^*(x), j)| \leqslant M^2$ i faktu, że (17) spełnione jest prawie wszędzie oraz z twierdzenia o przechodzeniu do granicy pod znakiem całki wynika (19).

4. Algorytmy uczenia

Stosowanie różnych metod estymacji nieznanych gęstości prawdopodobieństwa prowadzi do różnych algorytmów uczenia rozpoznawania. Rozpatrzymy teraz kilka nieparametrycznych sposobów estymacji i otrzymane przez ich stosowanie reguły decyzyjne.

Przykład 1

Parzen wykazał, że jeśli

$$\sup_y \|K(y)\| < \infty, \quad \sup_y \|yK(y\| < \infty, \quad \int_y K(y)\, dy = 1, \quad \int_y K^2(y)\, dy < \infty$$

oraz

$$\lim_{n\to\infty} h(n) = 0, \quad \lim_{n\to\infty} nh(n) = \infty ,$$

to

$$f_n(y) = \frac{1}{nh^p(n)} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h(n)}\right),$$

(gdzie $y_1, y_2, \ldots, y_n$ są niezależnymi obserwacjami zmiennej losowej o gęstości $f(y)$) zdąża do $f(y)$ według prawdopodobieństwa w punktach ciągłości funkcji $f(y)$, tzn. prawie wszędzie [1].

Stosowanie tego estymatora prowadzi do reguły decyzyjnej, która zalicza obiekt x do klasy, dla której wyrażenie

$$\sum_{j=1}^{M} \frac{1}{h^p(n_j)} L(i, j) \sum_{l=1}^{n_j} K\left(\frac{x - x_{jl}}{h(n_j)}\right) \tag{24}$$

osiąga minimum, przy czym x_{jl} jest kolejnym obiektem z klasy j.

Dla zero-jedynkowej funkcji strat i

$$K(x) = \begin{cases} \text{const} & \text{jeśli } \|x\| \leqslant 1 \\ 0 & \text{jeśli } \|x\| > 1 \end{cases}$$

x zalicza się do klasy, dla której ułamek

$$\frac{\text{ilość obiektów z klasy } i \text{ leżących w kuli } S(h(n_i)x)}{h(n_i)}$$

jest największy, przy czym

$$S(h, \bar{x}) = \{x : \|x - \bar{x}\| \leqslant h\} .$$

Przykład 2

Dla estymacji nieznanej gęstości prawdopodobieństwa Loftsgaarden i Quesenberry zaproponowali ciąg funkcji [2]

$$f_n(y) = \frac{k(n)-1}{n} \frac{p\Gamma\left(\frac{p}{2}\right)}{2\sqrt{\pi}R^p(k(n))},$$

gdzie $R(k)$ jest euklidesową odległością pomiędzy y, a k-ym najbliższym pomiarem w ciągu obserwacji $y_1, y_2, \ldots, y_n$. Estymator ten jest zbieżny do $f(x)$ według prawdopodobieństwa w punktach ciągłości $f(y)$, jeśli

$$k(n) > 0, \quad \lim_{n\to\infty} k(n) = \infty, \quad \lim_{n\to\infty} n^{-1}k(n) = 0 .$$

Stosując powyższą metodę estymacji do oceny gęstości w poszczególnych klasach otrzymujemy regułę decyzyjną, która zalicza x do klasy, dla której wyrażenie

$$\sum_{j=1}^{M} [k(n_j)-1] L|(i,j) R_j^{-p}(k(n_j))$$

jest najmniejsze, przy czym $R_j(k)$ jest euklidesową odległością pomiędzy x, a k-ym najbliższym pomiarem z klasy j.

Dla zero-jedynkowej funkcji strat otrzymuje się natomiast algorytm, który klasyfikuje x do klasy, dla której wyrażenie

$$[k(n_i)-1]^{-1} R_i(k(n_i)) .$$

jest najmniejsze. Jest to więc algorytm typu zmodyfikowany $k_n - NN$ [4].

Przykład 3

Jeśli $f(y)$ jest funkcją jednostajnie ciągłą oraz

$$p = 1, \quad K(y) \geqslant 0, \quad \int_{-\infty}^{\infty} K(y)dy = 1$$

i $h(n)$ jest ciągiem takim, że

$$\lim_{n\to\infty} h(n) = 0, \qquad \lim_{n\to\infty} \sum_{k=1}^{n} e^{-\gamma k h^2(k)} < \infty$$

dla każdego $\gamma > 0$, to zbieżność (16) zachodzi dla wszystkich y [3]. Reguła decyzyjna (24) jest wówczas zbieżna do reguły Bayesa z prawdopodobieństwem 1. Jako $h(n)$ można wybrać np. $\frac{1}{\sqrt{n}}$.

5. Zakończenie

W pracy rozważono ogólne własności algorytmów uczenia, w których wykorzystuje się nieparametryczne metody estymacji nieznanych gęstości prawdopodobieństwa w poszczególnych klasach. Otrzymane reguły decyzyjne są zbieżne do optymalnej reguły Bayesa, a ryzyko do ryzyka Bayesa. Z relacji (7) i (19) wynika ponadto, że

$$\lim_{n\to\infty} ER[\psi_n(x)] = R[\psi^*(x)],$$

co oznacza, że omówione algorytmy uczenia są bezwzględnie, asymptotycznie optymalne [5]

Nonparametric Estimation to Recognize Patterns

The paper deals with learning to recognize patterns. Decision rules which can be obtained by nonparametric estimation of unknown density function are investigated. Two cases are considered: the estimator converges in probability to the true density and with probability equal one. It is proved that the decision rule tends to the Bayes rule and the risk to the Bayes risk when the number of correctly classified samples increases.

Непараметрические оценки в обучении распознаванию

В работе исследуются алгоритмы распознавания объектов, которые можно получить в результате непараметрической оценки плотности распределения. Рассматриваются случаи, когда оценка сходится по вероятности и с вероятностью 1. Показано, что алгоритм распознавания сходится к Байесовому алгоритму, а риск к Байесовому риску.

Literatura

[1] E. Parzen, *On Estimation of Probability Density Function and Mode*, Ann. Math. Statist., vol. 33, 1962.

[2] D. O. Loftsgaarden, C. P. Quesenberry, *A Nonparametric Estimate of Multivariate Density Function*, Ann. Math. Statist. vol. 36, 1965.

[3] E. A. Nadaraya, *O neparametriczeskich ocenkach plotnosti werojatnosti i regressii*, Teoria werojatnostej i jejo primenenija, tom X, wyp. 1, 1965.
[4] T. M. Cover, P. E. Hart, *Nearest Neigbor Pattern Classification*, IEEE Trans. IT, nr 1, 1967.
[5] Z. Bubnicki, *Least Interval Pattern Recognition and its Application to Control Systems*, IV Congress of IFAC, nr 21.
[6] R. Sikorski, *Funkcje rzeczywiste*, Warszawa 1958.

DR INŻ. W. GREBLICKI
POLITECHNIKA WROCŁAWSKA, INSTYTUT CYBERNETYKI TECHNICZNEJ
ZAKŁAD PROCESÓW STEROWANIA, WROCŁAW, POLSKA