Prace VIII Krajowej Konferencji Automatyki
Szczecin - wrzesień, 1980

Doc.dr hab.inż. Włodzimierz Greblicki
Instytut Cybernetyki Technicznej
Politechnika Wrocławska

ASYMPTOTYCZNIE OPTYMALNE UCZENIE ROZPOZNAWANIA

1. Wstęp

Wykorzystanie nieparametrycznych estymatorów gęstości prawdopodobieństwa prowadzi w pewnych sytuacjach do asymptotycznie optymalnych procedur uczenia rozpoznawania. Algorytmy, w których zastosowano oszacowania Rosenblatta-Parzena (RP) czy też Loftsgaardena-Quesenberry'ego (LQ) są asymptotycznie optymalne gdy np. wszystkie gęstości charakteryzujące poszczególne klasy są prawie wszędzie ciągłe, patrz Greblicki [3] lub całkowalne z kwadratem, Wolverton i Wagner [14]. Obecnie wykażemy, że procedury, w których stosuje się estymatory RP w wersji oryginalnej lub rekurencyjnej albo estymator LQ są asymptotycznie optymalne dla d o w o l n y c h gęstości w klasach. Pokażemy w tym celu, że wszystkie wymienione oszacowania są prawie wszędzie punktowo zgodne lub mocno zgodne.

2. Estymatory gęstości prawdopodobieństwa

Wykażemy teraz, że estymator RP, rekurencyjny estymator RP oraz estymator LQ są zgodne lub zgodne z p.1 prawie wszędzie. Niech $X_1,\ldots,X_n$ będzie ciągiem niezależnych obserwacji p-wymiarowej zmiennej losowej X o gęstości prawdopodobieństwa f. Rozważymy następujące oszacowania $f(x)$:

a/ estymator RP

$$\hat{f}_n(x) = n^{-1}h_n^{-p}\sum_{i=1}^{n} K(h_n^{-1}(x - X_i)) , \qquad (1)$$

b/ rekurencyjny estymator RP

$$\tilde{f}_n(x) = \sum_{i=1}^{n} h_i^{-ap}K(h_i^{-1}(x - X_i)) \Big/ \sum_{i=1}^{n} h_i^{p(1-a)} , \qquad (2)$$

c/ estymator LQ

$$\bar{f}_n(x) = k_n/VR^p(x,k_n) . \qquad (3)$$

W estymatorach (1) i (2) jądro K jest dowolną funkcją gęstości prawdopodobieństwa a $\{h_n\}$ jest ciągiem liczbowym. W oszacowaniu LQ $\{k_n\}$ jest ciągiem liczbowym, V objętością kuli jednostkowej w R^p, natomiast $R(k,x)$ jest odległością pomiędzy $x \in R^p$ i k-tą najbliższą obserwacją. Zauważmy, że dwa pierwsze oszacowania są - w odróżnieniu od trzeciego - sane gęstościami. Zauważmy ponadto, że dla a=1 oraz a=0 estymator (2) przyjmuje następujące postaci:

$$n^{-1}\sum_{i=1}^{n} h_i^{-p}K(h_i^{-1}(x - X_i)) , \qquad (2a)$$

$$\sum_{i=1}^{n} K(h_i^{-1}(x - X_i)) \Big/ \sum_{i=1}^{n} h_i^{-p} . \qquad (2b)$$

Estymator (2) był badany - lecz jedynie dla a=1 tzn. w postaci (2a)- przez Yamato [15].

Podamy teraz warunki, które zapewniają zgodność słabą i mocną wymienionych wyżej estymatorów w punktach Lebesgue'a gęstości f tzn. nie tylko w punktach jej ciągłości ale także w prawie wszystkich - w sensie miary Lebesgue'a - punktach $x \in R^p$. Twierdzenia 1-3 są rozwinięciem dotychczasowych prac, patrz Parzen [13] , Yamato [15] oraz Loftsgaarden ,Quesenberry [10] , w których badano zgodność jedynie w punktach ciągłości gęstości f. To, że wykazujemy ich zgodność w prawie wszystkich punktach pozwoli później udowodnić, że odpowiednie procedury uczenia rozpoznawania są asymptotycznie optymalne dla dowolnych gęstości w klasach.

Twierdzenie 1

Załóżmy, że dla nieujemnego i ograniczonego jądra K

$$\int K(x)dx = 1 \quad \text{oraz} \quad K(x) = O(\|x\|^{-p-t}) \tag{4}$$

dla dowolnego $t>0$. Jeśli

$$h_n \to 0 \quad \text{oraz} \quad nh_n^p \to \infty \quad \text{gdy } n \to \infty \ , \tag{5}$$

to estymator RP jest zgodny w punktach Lebesgue'a gęstości f. Jeśli ponadto

$$nh_n^p / \log n \to \infty \quad \text{gdy } n \to \infty \ , \tag{6}$$

to estymator ten jest zgodny z p.1 w tych samych punktach.

Uwaga 1

Przykładem jądra, które spełnia założenia twierdzenia może być np. gęstość rozkładu normalnego. Jeśli natomiast ciąg h_n jest typu n^{-b}, to warunek (5) jest spełniony dla $0 < b < 1/p$. Jest oczywiste, że zachodzi wtedy także (6) .

Twierdzenie 2

Załóżmy, że jądro K spełnia założenia Twierdzenia 1. Jeśli

$$h_n \to 0 \quad \text{oraz} \quad \sum_{i=1}^{n} h_i^{p(1-2a)} \Big/ \Big[\sum_{i=1}^{n} h_i^{p(1-a)} \Big]^2 \to 0 \quad \text{gdy } n \to \infty, \tag{7}$$

to rekurencyjny estymator RP jest zgodny w punktach Lebesgue'a gęstości f. Jeśli ponadto

$$\sum_{n=1}^{\infty} h_n^{p(1-2a)} \Big/ \Big[\sum_{i=1}^{n} h_i^{p(1-a)} \Big]^2 < \infty, \tag{8}$$

to estymator ten jest mocno zgodny w tych samych punktach.

Uwaga 2

Dla a=1 drugie z założeń (7) ma następującą postać:

$$n^{-2} \sum_{i=1}^{n} h_i^{-p} \to 0 \quad \text{gdy } n \to \infty,$$

natomiast (8) oznacza,że

$$\sum_{n=1}^{\infty} n^{-2} h_n^{-p} < \infty.$$

Dla a=0 obydwa założenia są spełnione gdy poprostu

$$\sum_{n=1}^{\infty} h_n^p = \infty .$$

Uwaga 3

Dla h_n typu n^{-b} założenia (7) są spełnione gdy $0 < b < 1/p$ (dla dowolnego a) . Spełniony jest wtedy także warunek (8) .

Twierdzenie 3

Estymator LQ jest zgodny w punktach Lebesgue'a gęstości f gdy

$$k_n \to \infty \quad \text{oraz} \quad k_n/n \to 0 \quad \text{gdy } n \to \infty \ . \tag{9}$$

Jeśli ponadto

$$k_n \log n/n \to 0 \quad \text{gdy } n \to \infty, \tag{10}$$

to estymator ten jest mocno zgodny w tych samych punktach.

3. Procedury uczenia rozpoznawania

Oznaczmy przez $p_1,\ldots,p_M$ oraz przez $f_1,\ldots,f_M$ prawdopodobieństwa a priori poszczególnych klas oraz p-wymiarowe gęstości prawdopodobieństwa w poszczególnych klasach. Jak wiadomo, dla funkcji strat typu 0-1 ryzyko jest równe prawdopodobieństwu błędnej klasyfikacji i jest najmniejsze - oznaczmy je przez R_0 - gdy każde $x \in R^p$ zalicza się do dowolnej klasy i, dla której $p_i f_i(x)$ osiąga największą wartość. Załóżmy, że ani prawdopodobieństwa a priori ani gęstości w klasach nie są znane. Będziemy je szacować na podstawie ciągu uczącego $(L_1,X_1),\ldots,(L_n,X_n)$ tzn. ciągu prawidłowo sklasyfikowanych obserwacji X_i; L_i jest klasą, z której pochodzi obserwacja X_i. Podzielimy w tam celu ciąg obserwacji na M podciągów

$$(X_1^1,\ldots,X_{N_1}^1),\ \ldots,\ (X_1^M,\ldots,X_{N_M}^M),$$

z których każdy składa się z obserwacji pochodzących z poszczególnych klas. Niech $p_{in}=N_i/n$ będzie estymatorem p_i. Przez $f_{in}\,x$ oznaczmy oszacowanie $f_i\,x$ wyznaczone na podstawie i-tego podciągu. Rozważymy regułę uczenia rozpoznawania, która każde $x \in R^p$ zalicza do dowolnej klasy i , która maksymalizuje $p_{in}f_{in}(x)$ i oznaczmy przez R_n odpowiadające jej ryzyko.

Stosując oszacowania (1) - (3) otrzymuje się zatem reguły o następujących funkcjach dyskryminacyjnych:

$$h_{N_i}^{-1}\sum_{j=1}^{N_i} K\left(h_{N_i}^{-1}(x-X_j^i)\right),\quad \sum_{j=1}^{N_i} h_j^{-a}K\left(h_j^{-1}(x-X_j^i)\right)/N_i\sum_{j=1}^{N_i}h_j^{p(1-a)},\quad k_{N_i}/R_i^p(x,k_n).$$

Z Twierdzenia 1 oraz z twierdzenia o asymptotycznej optymalności reguł uczenia rozpoznawania podanego w pracy Greblickiego [3] wynika

Twierdzenie 4

Jeśli jądro K spełnia założenia twierdzenia 1 oraz zachodzi (5), to dla procedury wykorzystującej estymator RP, $R_n \to R_0$ według prawdopodobieństwa gdy $n\to\infty$. Jeśli ponadto spełniony jest warunek (6) , to $R_n \to R_0$ z p.1 gdy $n\to\infty$. Obydwie zbieżności zachodzą dla dowolnych gęstości w klasach.

Jest oczywiste, że analogiczne twierdzenia można podać dla reguł stosujących rekurencyjny estymator RP lub estymator LQ. Ze względu na ograniczoność miejsca nie podamy ich tutaj. Należy jednak jeszcze raz podkreślić, że reguły o podanych powyżej funkcjach dyskryminacyjnych tzn. reguły wyprowadzone z rozpatrywanych w pracy estymatorów gęstości prawdopodobieństwa są asymptotycznie optymalne dla d o w o l n y c h gęstości w klasach. Rezultat ten stanowi istotne rozwinięcie dotychczasowych rezultatów. Dotychczas zakładano, że np. gęstości w klasach są całkowalne z kwadratem, Wolverton i Wagner [14], lub że są ciągłe prawie wszędzie, Greblicki [3,4] .

4. Uwagi końcowe

Przedstawione w pracy reguły uczenia rozpoznawania stają się jeszcze prostsze, i nie tracą przy tym swoich własności, gdy stosuje się je w tzw. zmodyfikowanej formie, patrz Greblicki [4].

Warto zaznaczyć, że asymptotyczną optymalność można uzyskać także w uczeniu z tzw. probabilistycznym trenerem tzn. takim, który może błędnie klasyfikować obserwacje ciągu uczącego, Greblicki [5]. Można także wykazać, że reguła uczenia może niekiedy klasyfikować lepiej niż trener.

Podejście nieco inne od przedstawionego w tej pracy polega na stosowaniu nieparametrycznych estymatorów funkcji regresji, co zrobili np. Devroye i Wagner [1]. Oszacowania ortogonalne stosował z podobnym skutkiem Greblicki [6] .

Jest oczywiste, że szybkość zbieżności procedur zależy od ciągów $\{h_n\}$ i $\{k_n\}$ oraz jądra K. Problemu ich wyboru nie będziemy tutaj poruszać, a zainteresowanych odsyłamy do prac Fukunagi i Hostetlera [2] oraz Greblickiego i Krzyżaka [7] .

5. Dodatek

Ze względu na szczupłość miejsca podamy jedynie szkice dowodów Twierdzeń 1-3. Zauważmy przede wszystkim, że dla jądra jak w Twierdzeniu 1

$$h^{-p}EK(h^{-1}(x-X)) \to f(x) \quad \text{gdy } h \to 0 \quad \text{i} \quad EK^2(h^{-1}(x-X)) \leq ch^p, \quad c > 0,$$

w punktach Lebesgue'a gęstości f, patrz Neri [12, str. 15] . Korzystając z powyższego można wykazać słabą zgodność estymatorów (1) i (2) . Dla wykazania mocnej zgodności pierwszego z nich wystarczy wziąć pod uwagę $P\{|\hat{f}_n(x)-E\hat{f}_n(x)|>t\}$ i skorzystać z nierówności Bernsteina, Hoeffding [8] , i lematu Borela-Cantellego. Przy drugim estymatorze wystarczy wziąć pod uwagę twierdzenie Kołmogorowa, Loéve [9, str. 250] . Twierdzenie 3 jest konsekwencją Twierdzenia 1 i Twierdzenia 1.1 w pracy Moore'a i Yackela [11] .

6. Literatura

1. Devroye L. P., Wagner T. J., On the L1 convergence of kernel estimators of regression functions with applications in classification, Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete, w druku.
2. Fukunaga K., Hostetler L. P., Optimization of k-nearest-neighbor density estimates, IEEE Trans. on Information Theory, May 1973.
3. Greblicki W., Asymptotically optimal pattern recognition procedures with density estimates, IEEE Trans. on Information Theory, March 1978.
4. Greblicki W., Pattern recognition procedures with nonparametric density estimates, IEEE Trans. on Systems, Man, and Cybernetics, September 1978.
5. Greblicki W., Learning to recognize patterns with a probabilistic teacher, Pattern Recognition, przyjęte do druku.
6. Greblicki W., Pattern recognition using nonparametric estimates of a density function, w przygotowaniu.
7. Greblicki W., Krzyżak A., Asymptotical properties of kernel estimates of a regression function, Journal of Statistical Planning and Inferrence, przyjęte do druku.
8. Hoeffding W., Probability inequalities for sums of bounded random variables, Journal of the American Statistical Association, March 1963.
9. Loève M., Probability Tneory, Springer-Verlag, 4-th Edition.
10. Loftsgaarden D. O., Quesenberry C. P., A nonparametric estimation of a multivariate density function, Annals of Mathematical Statistics, 1965.
11. Moore D. S., Yackel J. W., Consistency properties of nearest neighbor density function estimates, Annals of Statistics, 1977.
12. Neri U., Singular Integrals, Springer-Verlag 1971.
13. Parzen E., On the estimation of a probability density and the mode, Annals of Mathematical Statistics, 1962.
14. Wolverton C. T., Wagner T. J., Asymptotically optimal discriminant functions for pattern classification, IEEE Trans. on Information Theory, 1969.
15. Yamato H., Sequential estimation of a continuous density function and mode, Bulletin of Mathematical Statistics, 1971.