Prace Naukowe
Instytutu Cybernetyki Technicznej
Politechniki Wrocławskiej

Nr 18

Monografie 3

Włodzimierz Greblicki

Asymptotycznie optymalne algorytmy rozpoznawania i identyfikacji w warunkach probabilistycznych

Wrocław 1974

Prace Naukowe
Politechniki Wrocławskiej

Redaktor Naczelny
Marian Kloza

Redaktor Naukowy
Włodzimierz Greblicki

Opracowanie redakcyjne
Danuta Sowińska

Okładkę projektował
Jacek Sikorski

Korekta
Małgorzata Kmietowicz

Redakcja Wydawnictw Politechniki Wrocławskiej
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław

Nakład 450 + 65 egz. Ark. wyd. 6. Ark. druk. 6 5/8. Papier offset. kl. III, 70 g, B1
Zakład Graficzny Politechniki Wrocławskiej. Zam. 6021/74 — P-12 — Cena zł 21,—

Prace Naukowe Instytutu Cybernetyki Technicznej
Politechniki Wrocławskiej

Nr 18 Nr 18
Monografie Nr 3 1974

Identyfikacja, rozpoznawanie, uczenie, estymacja gęstości, funkcja decyzyjna, problem decyzyjny Bayesa

Włodzimierz GREBLICKI

Asymptotycznie optymalne algorytmy rozpoznawania i identyfikacji w warunkach probabilistycznych

Rozpoznawanie i identyfikację traktuje się w pracy jako szczególne przypadki ogólnego problemu decyzyjnego. Przyjmuje się przy tym całkowity brak jakiejkolwiek informacji o rozkładach prawdopodobieństwa. Sytuacja taka w rozpoznawaniu polega na nieznajomości rozkładu klas i gęstości w klasach, a w identyfikacji oznacza nieznajomość łącznego rozkładu wejścia i wyjścia obiektu. Na podstawie ciągu uczącego, tzn. obrazów i ich prawidłowych klasyfikacji oraz obserwacji wejścia i wyjścia obiektu można estymować nie znane rozkłady i z otrzymanych oszacowań korzystać w optymalnych regułach decyzyjnych zastępując nimi nieznane rozkłady. Wykazano, że takie algorytmy uczenia, przy stosowaniu nieparametrycznych metod estymacji gęstości prawdopodobieństwa, stają się coraz bliższe rozwiązaniom, które można byłoby wyznaczyć przy pełnej informacji, co oznacza, że są one asymptotycznie optymalne. Odpowiednia zgodność estymatora gęstości implikuje zbieżność reguły klasyfikującej do optymalnej reguły Bayesa oraz ryzyka do minimalnego ryzyka Bayesa. Omówiono procedury, które otrzymuje się przy stosowaniu różnych typów nieparametrycznych oszacowań gęstości. Dla kwadratowej funkcji strat uzyskiwane w ten sposób algorytmy identyfikacji są także asymptotycznie optymalne. Stosowanie estymatorów gęstości o różnych własnościach prowadzi do różnego typu zbieżności charakterystyk modelu do modelu optymalnego. Jako drugą zastosowano w identyfikacji metodę rozwinięć ortogonalnych i, przy pewnych założeniach, wykazano jej asymptotyczną optymalność.
Oceniono także jakość sterowania przy wykorzystywaniu charakterystyki modelu i pokazano, że dla odpowiednich procedur identyfikacji oszacowania sterowań wyznaczone na podstawie modelu są zbieżne do nieznanych sterowań zapewniających żądaną średnią lub ekstremalną wartość wyjścia obiektu.

WSTĘP

Zadania rozpoznawania i wyznaczania modelu obiektu w warunkach niepewności można traktować jako szczególne przypadki ogólnego problemu decyzyjnego. Jeśli posiada się pełną informację probabilistyczną, tzn. znajomość rozkładów w klasach i ich prawdopodobieństwa lub łączny rozkład wejścia i wyjścia obiektu, wyznaczenie optymalnych rozwiązań nie stwarza trudności i nie jest zbyt interesujące z punktu widzenia zastosowań. W sytuacjach praktycznych bowiem, dane o problemie są zazwyczaj niekompletne - nie zna się często rozkładów w klasach i rozkładów sygnałów w obiekcie. Wskazane jest wówczas prowadzenie eksperymentów, które pozwoliłyby uzupełniać dane wyjściowe. Eksperyment taki może polegać na obserwacji tzw. ciągu uczącego, tzn. ciągu obrazów i ich klas w rozpoznawaniu oraz wejścia i wyjścia obiektu w identyfikacji. Ponieważ stopień nieznajomości rozkładów może być różny, to stosuje się różne algorytmy uczenia rozpoznawania i identyfikacji. Jeśli wiadomo jakiej postaci są rozkłady, ale nie zna się pewnych parametrów, to uczenie polega na ich szacowaniu. Dalej posunięty brak danych o rozkładach, to całkowita ich nieznajomość. Typowe postępowanie polega wówczas na parametrycznym ustaleniu klasy rozważanych rozwiązań, tzn. charakterystyk modelu lub np. funkcji klasyfikujących. Ciąg uczący wykorzystuje się wtedy do estymacji parametrów charakteryzujących rozwiązania najlepsze w arbitralnie zazwyczaj ustalonych klasach.

Ograniczenie rozwiązań do określonej parametrycznie klasy prowadzi wprawdzie zwykle do prostych procedur uczenia, lecz nie zapewnia zbieżności (przy wzroście długości ciągu uczącego) otrzymywanych reguł rozpoznawania i modeli do optymalnych rozwiązań Bayesa, tzn. nie zapewnia asymptotycznej, bayesowskiej optymalności.

Praca ta poświęcona jest właśnie asymptotycznie bayesowskim algorytmom uczenia rozpoznawania i identyfikacji w sytuacji całkowitego braku danych o rozkładach charakteryzujących problem. Własności rozważanych procedur stają się pod różnymi względami coraz bliższe rozwiązaniom optymalnym, które można by było wyznaczyć przy znajomości odpowiednich rozkładów.

Własności takie ma znana metoda funkcji potencjalnych w zastosowaniu zarówno do uczenia rozpoznawania jak i identyfikacji. Większość prac, a jest ich stosunkowo niewiele, rozwiązujących przedstawiony problem dotyczy uczenia rozpoznawania. Główną ich ideą jest korzystanie z nieparametrycznych metod szacowania gęstości prawdopodobieństwa.

W pracy tej wykorzystano taki sposób uczenia rozpoznawania i udowodniono jego asymptotycznie optymalne własności. Odpowiednie twierdzenia podają warunki zapewniające zbieżność procedur ocenianą zarówno pod kątem zachowania się reguł rozpoznawania jak i ryzyka. W podobny sposób skonstruowano algorytmy identyfikacji, w których wykorzystano nieparametryczne estymatory gęstości prawdopodobieństwa w celu szacowania łącznej gęstości wejścia i wyjścia obiektu. Druga z przedstawionych metod identyfikacji to odpowiednie wykorzystanie rozwinięć ortogonalnych. Twierdzenia podane w pracy pozwoliły także sformułować nowe, asymptotycznie optymalne własności metody funkcji potencjalnych.

W rozdziale I przedstawiono ogólny, decyzyjny problem Bayesa i na jego tle rozważono zadania rozpoznawania i wyznaczania modelu. Następnie rozpatrzono sytuację braku danych i omówiono zagadnienie uczenia, a zwłaszcza uczenie rozpoznawania i uczenie w celu wyznaczenia modelu, czyli identyfikację. Przedstawiono także znane metody uczenia rozpoznawania i identyfikacji, ze szczególnym uwzględnieniem sytuacji całkowitego braku danych o rozkładach.

Rozdziały II i III dotyczą uczenia rozpoznawania. W pierwszym z nich podano ogólne twierdzenia o asymptotycznej optymalności procedur uczenia, w których wykorzystuje się nieparametryczne oszacowania gęstości. Wykazano odpowiednią zbieżność reguły rozpoznawania i ryzyka przy stosowaniu estymatorów o różnych własnościach, np. zgodnych, mocno zgodnych i zgodnych w sensie całkowym itp. Wykazano także ciekawą własność asymptotycznej optymalności algorytmu funkcji potencjalnych. W rozdziale III przedstawiono algorytmy uczenia rozpoznawania, które otrzymuje się dla różnych oszacowań gęstości prawdopodobieństwa, np. typu Parzena, Loftsgaardena i Quesenberry'ego itp. Podano także ich geometryczną interpretację. Na początku rozdziału zamieszczono ponadto krótkie omówienie znanych, nieparametrycznych oszacowań gęstości. Na zakończenie podano także przykład obliczeniowy.

Rozdział IV dotyczy identyfikacji. Podano w nim kilka twierdzeń o asymptotycznej optymalności procesów identyfikacji, przy kwadratowej funkcji strat, w których stosuje się nieparametryczne oszacowania gęstości, a następnie rozwinięto je dla oszacowań Parzena. Drugą z przedstawionych metod jest odpowiednio wykorzystana metoda rozwinięć ortogonalnych. Podano przykład numeryczny, a następnie wykazano interesujące własności sterowań wyznaczanych na podstawie modelu. Wykazano, że dla pewnych algorytmów identyfikacji takie oszacowania stają się coraz bliższe nieznanym sterowaniom, co pozwala te algorytmy uważać za asymptotycznie optymalne także pod względem sterowania.

W zakończeniu przedstawiono wreszcie nie rozwiązane i godne uwagi problemy związane z omawianymi w pracy zagadnieniami.

Autor dziękuje prof. drowi Zdzisławowi Bubnickiemu za dyskusje i cenne uwagi, prof. drowi Jerzemu Seidlerowi za przejrzenie rękopisu.

I. DECYZYJNE PROBLEMY ROZPOZNAWANIA I IDENTYFIKACJI

1.1. Decyzyjne problemy rozpoznawania i wyznaczania modelu

Przedstawimy teraz ogólny problem decyzyjny Bayesa, a następnie zagadnienia rozpoznawania i wyznaczania modelu obiektu jako jego szczególne przypadki. Niech Ω, $\mathcal{X}$, $\mathcal{D}$ będą odpowiednio przestrzeniami parametrów, obserwacji i decyzji, a G - apriorycznym rozkładem (dystrybuantą) parametrów, tzn. elementów przestrzeni Ω. Rozkład prawdopodobieństwa obserwacji, tzn. elementów przestrzeni $\mathcal{X}$ (na której określona jest σ - skończona miara μ) zależy od parametru ω, $\omega \in \Omega$. Gęstość tego rozkładu (względem μ), w sytuacji, gdy parametrem jest $\omega \in \Omega$ oznaczymy przez f_ω. Jeśli parametrem jest $\omega \in \Omega$, a decyzją $d \in \mathcal{D}$, to ponosi się stratę $L(d, \omega) \geq 0$. Dla reguły decyzyjnej ϕ, tzn. funkcji określonej na przestrzeni obserwacji $\mathcal{X}$ o wartościach w przestrzeni decyzji $\bar{\mathcal{D}}$ (mierzalnej μ) ryzyko warunkowe w sytuacji, gdy parametrem jest $\omega \in \Omega$, wyraża się wzorem

$$R(\phi, \omega) = \int_{\mathcal{X}} L(\phi(x), \omega) f_\omega(x)\, dx. \tag{1.1}$$

Jakość reguły decyzyjnej ϕ ocenia wartość oczekiwana straty, tzn. ryzyko

$$R(\phi) = \int_{\Omega} R(\phi, \omega)\, dG(\omega). \tag{1.2}$$

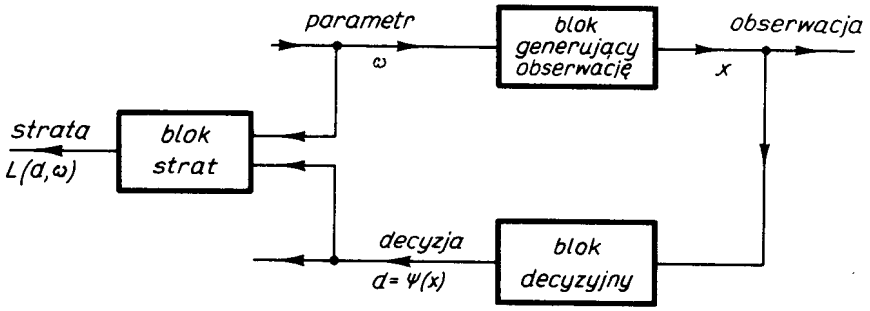
Każda reguła decyzyjna ϕ spełniająca dla prawie wszystkich (względem μ) $x \in \mathcal{X}$ równość

$$\Phi(\phi(x), x) = \min_{d \in \mathcal{D}} \Phi(d, x), \tag{1.3}$$

w której

$$\Phi(d, x) = \int_{\Omega} L(d, \omega) f_\omega(x)\, dG(\omega)$$

jest optymalną, bayesowską regułą decyzyjną, $R(\phi^*)$ natomiast - przy czym ϕ^* jest dowolną, optymalną, bayesowską regułą decyzyjną - jest minimalnym ryzykiem.

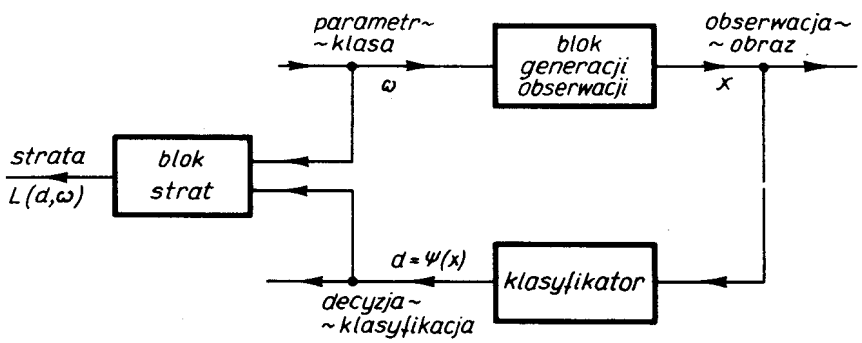


Rys. 1. Blokowa interpretacja decyzyjnego problemu Bayesa
Fig. 1. Block scheme for the Bayes decision problem

Przedstawiony problem decyzyjny można interpretować jak na rys.1. Własności bloku generacji obserwacji opisane są gęstością warunkową f_ω. W sytuacji, gdy na jego wejście zostanie podany parametr ω, wylosowany zgodnie z apriorycznym rozkładem G, to generuje on na wyjściu obserwację według gęstości f_ω. Blok decyzyjny o charakterystyce ϕ podejmuje, na podstawie obserwacji x, decyzję $d = \phi(x)$, co pociąga za sobą stratę $L(d, \omega)$ wyliczaną przez blok strat. Optymalny blok decyzyjny ma charakterystykę zapewniającą minimalną stratę średnią.

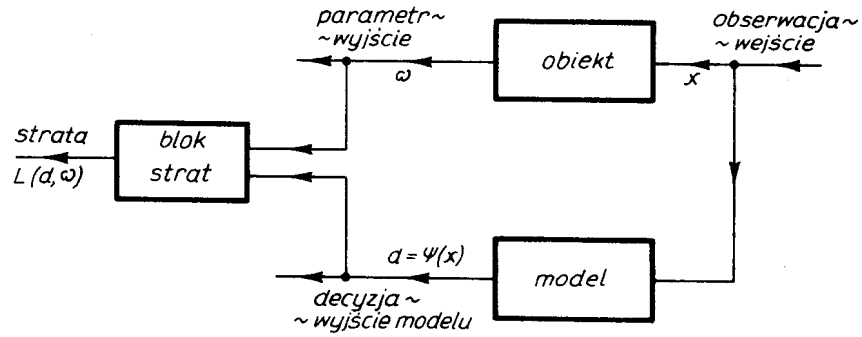
Jako szczególne przypadki przedstawionego powyżej decyzyjnego problemu Bayesa otrzymuje się zadania rozpoznawania i wyznaczania modelu. Omówimy teraz pierwsze z nich. W zadaniach rozpoznawania (klasyfikacji) skończoną przestrzeń parametrów $\Omega = \{\omega_1, \ldots, \omega_M\}$ nazywa się przestrzenią klas, a jej elementy klasami. Wygodnie przy tym przyjąć $\Omega = \{1, \ldots, M\}$. Apriorycznyrozkład klas określa prawdopodobieństwa pojawienia się każdej z nich. Przestrzeń obserwacji $\mathcal{X}$ jest p-wymiarową przestrzenią wektorową tzn. $\mathcal{X} = \mathcal{R}^p$ (μ jest miarą Lebesgue'a). Gęstości $f_1, \ldots, f_M$ nazywają się rozkładami w klasach, a obserwacja - obrazem. Przestrzenią decyzyjną jest $\mathcal{D} = \Omega = \{1, \ldots, M\}$, natomiast $L(d, \omega)$ jest stratą spowodowaną przez zaliczenie do klasy d obrazu należącego do klasy ω. Problem rozpoznawania można przedstawić jak na rys. 2. Blok generacji obserwacji scharakteryzowany jest teraz zestawem gęstości w klasach $f_1, \ldots, f_M$ i każdemu parametrowi, tzn. klasie przyporządkowuje, zgodnie z odpowiednim rozkładem, obserwację, tzn. obraz. Blok decyzyjny nazywa się klasyfikatorem

lub urządzeniem rozpoznającym i klasyfikuje obraz do jednej z M klas.



Rys. 2. Blokowa interpretacja decyzyjnego problemu rozpoznawania
Fig. 2. Block scheme for the pattern recognition problem

Schemat odpowiadający zadaniu wyznaczania modelu pokazano na rys. 3. Blok generacji obserwacji nazywa się teraz obiektem o wejściu x i wyjściu ω. Ponadto $\mathcal{X} = \mathcal{R}^p$(μ jest miarą Lebesgue'a) oraz $\mathcal{D} = \Omega = \mathcal{R}$. Urządzenie decyzyjne nazywa się modelem obiektu, a strata



Rys. 3. Blokowa interpretacja decyzyjnego problemu wyznaczania modelu
Fig. 3. Block scheme of decision problem for the model determination

ocenia odchylenie między wyjściami obiektu i modelu. Zarówno aprioryczny rozkład G wyjścia obiektu, jak i warunkowa gęstość f nie mają jednak wygodnej interpretacji technicznej i dlatego chętniej operuje się innymi rozkładami, np. łącznym rozkładem wejścia i wyjścia obiektu.

Przyjęty model decyzyjny sformułowano pod kątem problemów rozpatrywanych w tej pracy. Nie jest on uniwersalny i nie obejmuje oczywiście wszystkich sytuacji spotykanych w uczeniu rozpoznawania i identyfikacji. Jeśli na przykład odpowiednie rozkłady znane są z dokładnością do parametrów, to decyzje mogą dotyczyć nieznanych parametrów, a nie wyjścia obiektu jak to wyżej przyjęto.

1.2. Empiryczne problemy decyzyjne rozpoznawania i identyfikacji

W przedstawionych powyżej decyzyjnych problemach rozpoznawania i wyznaczania modelu optymalną regułę decyzyjną tzn. optymalny klasyfikator lub model wyznacza się na podstawie pełnej informacji probabilistycznej. Dla rozpoznawania wymaga to znajomości prawdopodobieństw wystąpienia poszczególnych klas i gęstości w klasach, przy wyznaczaniu modelu natomiast pełna informacja polega na znajomości łącznego rozkładu wejścia i wyjścia obiektu (co jest równoznaczne ze znajomością rozkładów G i f_ω). Zagadnienia takie nie są jednak szczególnie interesujące z praktycznego punktu widzenia, ponieważ w wielu technicznych zadaniach nie dysponuje się niestety pełną informacją probabilistyczną. W sytuacjach takich wskazane byłoby prowadzenie dodatkowych eksperymentów, które pomimo tych trudności pozwoliłyby zdobywać brakujące dane i coraz lepiej rozwiązywać odpowiedni problem decyzyjny.

W celu przedstawienia zagadnienia podejmowania decyzji na podstawie wcześniej wykonanych eksperymentów załóżmy, że zarówno rozkład apriorycznyy G, jak i gęstość f_ω, są całkowicie nie znane, a wspomniany eksperyment polega na wielokrotnym mierzeniu wejścia i wyjścia bloku generatora obserwacji (rys. 1), tzn. mierzeniu parametrów i odpowiadających im obserwacji. Jego wynikiem jest realizacja tzw. ciągu uczącego

$$(\Omega_1, X_1), \ldots, (\Omega_n, X_n),$$

tzn. ciągu niezależnych par zmiennych losowych o jednakowych rozkładach; rozkładem zmiennej losowej Ω_i jest G, natomiast rozkładem warunkowym zmiennej losowej X_i, gdy $\Omega_i = \omega$ jest f_ω. Niech teraz ψ_n będzie funkcją określoną na przestrzeni $(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X}$ o wartościach w przestrzeni decyzji $\mathcal{D}$, która każdej realizacji $(\omega_1, x_1), \ldots, (\omega_n, x_n)$ ciągu uczącego oraz elementowi $x \in \mathcal{X}$ przyporządkowuje decyzję

$$\psi_n(\omega_1, x_1, \ldots, \omega_n, x_n, x) \in \mathcal{D}.$$

Funkcję tę nazywa się empiryczną regułą decyzyjną, a ich ciąg $\{\psi_n\}$ algorytmem lub procedurą uczenia. Decyzja o obserwacji x za-

leży teraz od ciągu uczącego i jest natury losowej. Dlatego też wygodnie jest wprowadzić probabilistyczną (zrandomizowaną) regułę decyzyjną, która każdemu elementowi $x \in \mathcal{X}$ przyporządkowuje zmienną losową o wartościach w przestrzeni decyzji $\mathcal{D}$. Dla zrandomizowanej reguły decyzyjnej ϕ prawdopodobieństwo decyzji d, $d \in \mathcal{D}$, w sytuacji, gdy zaobserwowano $x \in \mathcal{X}$ jest równe $P\{\phi(x) = d\}$. Regułą taką może być również (deterministyczna) reguła decyzyjna wprowadzona w pkcie 1.1.

Zauważmy teraz, że funkcja ϕ_n wraz z ciągiem uczącym określają zrandomizowaną regułę decyzyjną, która decyzję d, gdy zaobserwowano x, podejmuje z prawdopodobieństwem

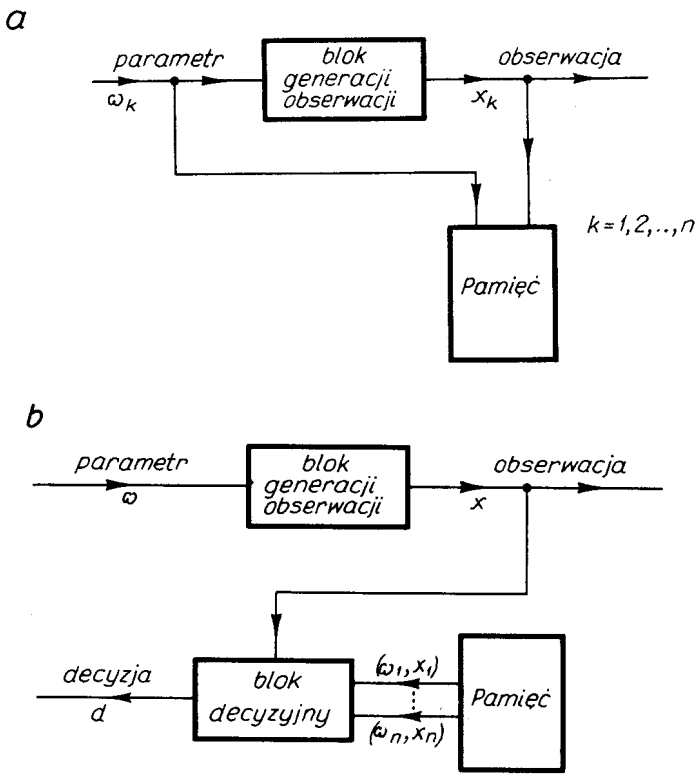
$$P\{\phi_n(\Omega_1, X_1, \ldots, \Omega_n, X_n, x) = d\}.$$

Symbolem ϕ_n będziemy oznaczać także zrandomizowaną regułę decyzyjną związaną z empiryczną regułą decyzyjną φ_n i będziemy nazywać (zrandomizowaną) regułą uczenia.

Problem podejmowania decyzji na podstawie uprzednich obserwacji $x_1, \ldots, x_n$ ($\omega_1, \ldots, \omega_n$ - nieznane) przy nieznajomości rozkładu apriorycznego G, czyli tzw. empiryczny problem Bayesa rozpatrywał Robbins [55], lecz zakładał znajomość rozkładu warunkowego f_ω. W zagadnieniu rozpatrywanym przez nas przyjmuje się dodatkowo nieznajomość gęstości f_ω, lecz za to dysponuje się także realizacjami parametru $\omega_1, \ldots, \omega_n$. Zatem w interesującym nas problemie, który można nazwać empirycznym zagadnieniem decyzyjnym Bayesa (rozkład apriorycznego parametru istnieje lecz jest nieznany i przed powzięciem decyzji estymuje się go) podjęcie decyzji poprzedzone jest uczeniem (nazywanym także uczeniem z trenerem) tzn. obserwacją ciągu uczącego, której wyniki są zapamiętywane. Dopiero po zakończeniu cyklu uczenia następuje proces podejmowania decyzji. Empiryczny problem decyzyjny można interpretować jak na rys. 4.

Blokową interpretację empirycznego problemu decyzyjnego rozpoznawania pokazano natomiast na rys. 5. W cyklu uczenia zapamiętuje się obrazy i ich klasy. Na ich podstawie podejmuje się następnie decyzję o rozpoznawanym obrazie. Charakterystyką klasyfikatora jest empiryczna reguła decyzyjna ϕ_n. Ciąg empirycznych reguł decyzyjnych nazywa się algorytmem lub procedurą uczenia rozpoznawania.

Uczenie w celu wyznaczenia modelu nazywa się identyfikacją obiektu. Na rysunku 6 przedstawiono cykle uczenia i podejmowania decyzji. Uczenie polega na wielokrotnym mierzeniu wejścia i wyjścia identyfikowanego obiektu. Empiryczną regułę decyzyjną nazywa się modelem, a ich ciąg algorytmem lub procedurą identyfikacji.



Rys. 4. Empiryczny problem decyzyjny Bayesa: a) uczenie, b) podejmowanie decyzji
Fig. 4. Empirical Bayes decision problem: a) learning, b) decision making

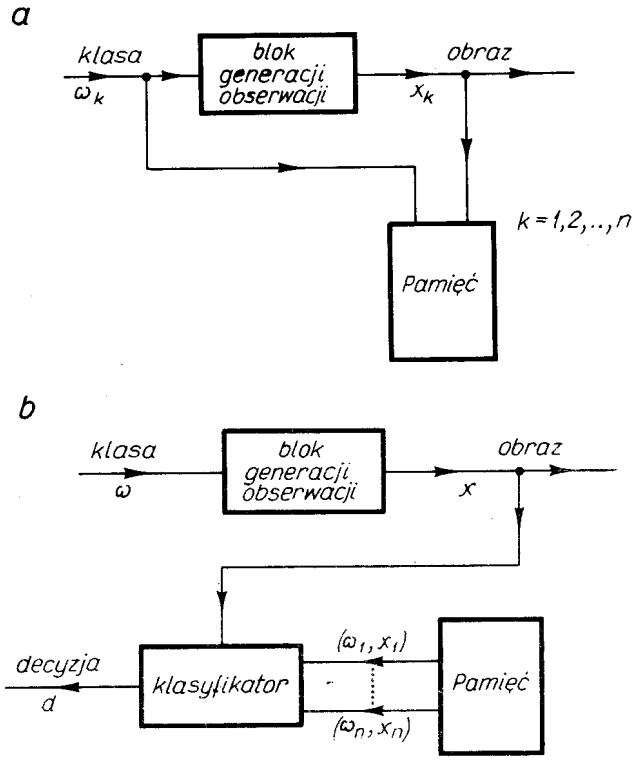
W przypadku rozpoznawania uczenie polega więc na obserwowaniu klas i obrazów, w identyfikacji natomiast - na obserwowaniu wejścia i wyjścia badanego obiektu.

Dla zrandomizowanej reguły decyzyjnej ψ_n ryzyko

$$R(\psi_n) = \int_{\Omega} R(\psi_n, \omega)\, dG(\omega),$$

gdzie $R(\psi_n, \omega)$ wyraża się wzorem (1.1), jest oczywiście zmienną losową. Proces uczenia będziemy oceniać badając ciąg zrandomizowanych reguł decyzyjnych $\{\psi_n\}$ lub odpowiadający mu ciąg ryzyk $\{R(\psi_n)\}$.

Interesować nas będą procedury uczenia, które przy wzroście długości ciągu uczącego coraz bardziej upodobniają się do optymalnych reguł Bayesa, tzn. procedury asymptotycznie optymalne w sensie Bayesa, czyli po prostu asymptotycznie optymalne. Podstawowym założeniem obowiązującym przy tym w całej pracy jest całkowita nieznajomość rozkładów f_ω i G. Asymptotyczną optymalność można różnie definiować, po-



Rys. 5. Empiryczny problem decyzyjny rozpoznawania: a) uczenie, b) rozpoznawanie
Fig. 5. Empirical pattern recognition decision problem: a) learning, b) decision making

nieważ różnie można oceniać zbieżność procedury $\{\phi_n\}$ do optymalnej reguły Bayesa (lub zbioru reguł optymalnych). Będziemy więc ją różnie rozumieli w zależności od tego czy np.

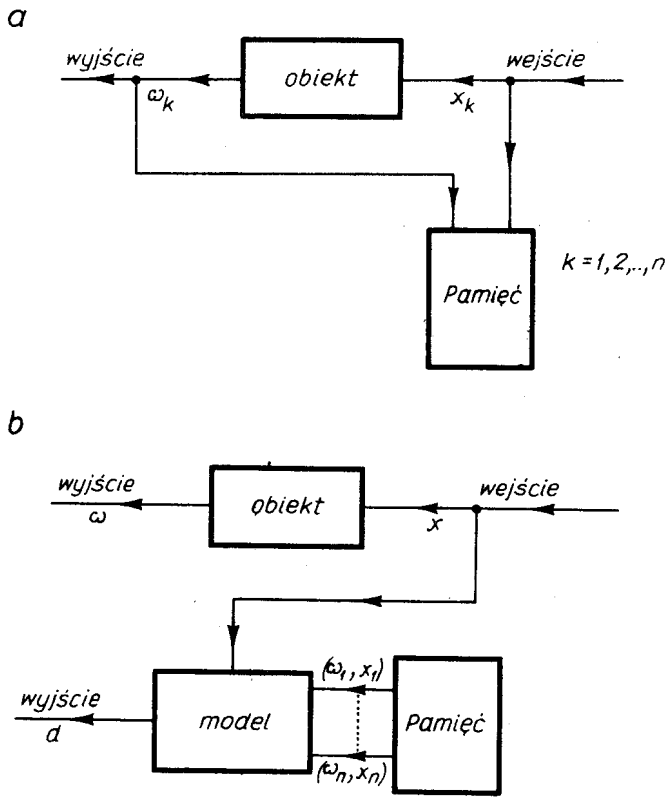
$$\phi_n(\Omega_1, X_1, \ldots, \Omega_n, X_n, x) \longrightarrow \phi^*(x)$$

według prawdopodobieństwa, z prawdopodobieństwem 1 lub w sensie średniokwadratowym, gdy $n \longrightarrow \infty$. Inny zastosowany sposób określenia asymptotycznej optymalności wykorzystuje zbieżności

$$R(\phi_n) \longrightarrow R(\phi^*)$$

według prawdopodobieństwa, z prawdopodobieństwem 1 lub w sensie średnim, gdy $n \longrightarrow \infty$.

Asymptotycznie optymalne algorytmy uczenia rozpoznawania i identyfikacji otrzymuje się w pracy poprzez estymację nieznanych rozkładów prawdopodobieństwa, a dokładniej dzięki wykorzystaniu nieparame-



Rys. 6. Decyzyjny problem identyfikacji: a) uczenie, b) podejmowanie decyzji
Fig. 6. Identification decision problem: a) learning, b) decision making

tycznych oszacowań gęstości. Drugim z zastosowanych sposobów jest odpowiednio użyta metoda rozwinięć ortogonalnych. Przedstawiono także wyniki jakie można otrzymać znaną metodą funkcji potencjalnych.

1.3. Algorytmy uczenia rozpoznawania i identyfikacji

W większości ze znanych metod uczenia rozpoznawania i identyfikacji, które można interpretować jako problemy decyzyjne typu przedstawionego w pktach 1.1 i 1.2 klasę rozważanych funkcji decyzyjnych ogranicza się do takich, które dają się przedstawić w formie parametrycznej. Za optymalną, w tak przyjętym zbiorze reguł decyzyjnych, oznaczmy go literą A, uważa się wówczas dowolną regułę decyzyjną ze zbioru A, która minimalizuje ryzyko (1.2). Problemy uczenia, które pojawiają się przy całkowitej nieznajomości rozkładów G i f_ω pole-

gają wówczas na znalezieniu procedur, które byłyby zbieżne do reguły decyzyjnej, najlepszej w przyjętej klasie A.

W przypadku identyfikacji zbiór A rozważanych modeli określa się z reguły jako liniową kombinację arbitralnie zazwyczaj ustalonego układu funkcji ortogonalnych $\varphi_1, \ldots, \varphi_N$, tzn.

$$\psi(x) = \sum_{i=1}^{N} \alpha_i \varphi_i(x).$$

Współczynniki $\alpha^*_1, \ldots, \alpha^*_N$ charakteryzujące najlepszy w takiej klasie model szacuje się następnie na podstawie ciągu uczącego, tzn. obserwacji wejścia i wyjścia identyfikowanego obiektu [9, 11, 12, 16, 18, 20, 37, 45, 46, 59, 80].

Analogiczny problem uczenia rozpoznawania, który polega na znalezieniu reguły decyzyjnej, najlepszej w ustalonej parametrycznie klasie, jest o wiele bardziej skomplikowany. W pracach [67, 79] podano algorytmy poszukiwania w jednowymiarowym problemie dychotomii takiego punktu rozdzielającego przestrzeń obrazów na dwa zbiory, który zapewnia najmniejsze prawdopodobieństwo błędnej klasyfikacji. Jednakże duże trudności związane z uczeniem w sytuacji, gdy jakość rozpoznawania oceniana jest według ryzyka powodują, że często wykorzystuje się inne wskaźniki jakości. Otrzymywane w ten sposób algorytmy nie są wprawdzie zbieżne do reguły decyzyjnej, najlepszej w ustalonej klasie, tzn. tej która minimalizuje ryzyko w przyjętym zbiorze A reguł decyzyjnych, ale odznaczają się dużą prostotą. Polegają one na aproksymacji odpowiednich, najczęściej bayesowskich funkcji klasyfikujących liniowymi kombinacjami arbitralnie ustalonego układu funkcji $\varphi_1, \ldots, \varphi_N$. Jako wskaźnik jakości aproksymacji przyjmuje się np.

$$\sum_{i=1}^{M} p_i \int \left(h_i(x) - \sum_{j=1}^{N} \alpha_{ij} \varphi_j(x)\right)^2 f_i(x)\, dx,$$

w którym p_i jest apriorycznym prawdopodobieństwem klasy i. Współczynniki α^*_{ij} najlepszej aproksymacji ustala się na podstawie ciągu uczącego. Jako przykłady funkcji h_i można podać

$$h_i(x) = \frac{p_i f_i(x)}{f(x)},$$

$i = 1, \ldots, M$, przy czym f jest bezwarunkową gęstością obrazów [16, 22].

Wyniki aproksymacji wykorzystuje się następnie w optymalnej regule klasyfikacyjnej Bayesa. W problemie dychotomii przyjmuje się też

np. [37, 54, 73]

$h_1(x) = (L(2,1)p_1f_1(x) - L(1,2)p_2f_2(x))/f(x)$, $\alpha^*_{2j} = 0$, $j = 1,\ldots,N$.

lub

$$h_i(x) = c_i, \quad i = 1,2 ,$$

przy czym c_1 i c_2 są odpowiednio wybranymi liczbami [52, 64, 83].

Na podstawie ciągu uczącego można także aproksymować gęstości w klasach, np. liniową kombinacją ustalonego układu funkcji [6, 7, 16, 19, 21] i wyniki wykorzystać następnie w optymalnej regule Bayesa [33, 65]. Niekiedy, między innymi z powodu trudności obliczeniowych, rezygnuje się z ryzyka jako oceny jakości rozpoznawania i przyjmuje inne typy wskaźników [5, 36, 40, 43, 48, 53, 64].

Istotną cechą powyższych metod jest zdecydowane uproszczenie problemów decyzyjnych do zagadnień poszukiwania rozwiązań najlepszych w arbitralnie wybranej klasie, określonej skończoną ilością parametrów. Spowodowana tym strata jakości modelu, czy też reguły rozpoznawania może być w wielu sytuacjach całkowicie dopuszczalna, niemniej jednak pojawia się zagadnienie znalezienia takich metod uczenia rozpoznawania i identyfikacji, które zapewniłyby zbieżność algorytmów do rozwiązań bayesowskich.

Heurystycznym sposobem uczenia rozpoznawania, w którym nie ogranicza się parametrycznie klasy rozwiązań, jest algorytm NN, tzn. najbliższy sąsiad [15]. Nie ma on jednak asymptotycznie optymalnych własności. Zapewnia je natomiast, choć przy dość skomplikowanych założeniach, metoda funkcji potencjalnych [8, 35] w zastosowaniu zarówno do rozpoznawania jak i identyfikacji. Można ją zapisać w postaci rekurencyjnego wzoru

$$t_{n+1}(x) = t_n(x) + r_n K(x, x_{n+1}) ,$$

w którym r_n jest odpowiednio wyliczanym współczynnikiem, a K tzw. funkcją potencjalną.

Algorytm ten, przy pewnych założeniach jest zbieżny do optymalnej decyzyjnej reguły Bayesa zarówno w zadaniu rozpoznawania jak i identyfikacji.

W pracach [10, 27, 29, 68, 70] przedstawiono pewne procedury uczenia rozpoznawania, w których zastosowano nieparametryczne metody estymacji gęstości prawdopodobieństwa i wykazano ich optymalne własności asymptotyczne. Podobne własności algorytmów identyfikacji otrzymano w pracach [30, 31, 34], w których wykorzystano także nieparametryczne oszacowania gęstości oraz metodę rozwinięć ortogonalnych.

Rozpatrywane w pracy modele decyzyjne i schematy uczenia nie wyczerpują oczywiście wszystkich zagadnień rozpoznawania i identyfika-

cji, z którymi można spotkać się przy tzw. niepełnej informacji probabilistycznej. Pomija się pośredni stopień między pełną informacją probabilistyczną [43, 50], a jej zupełnym brakiem, którego typowymi przykładami mogą być sytuacje powstające przy znajomości z dokładnością do parametrów rozkładów w klasach [1, 5, 14, 36, 39, 48, 50, 60, 61] lub łącznego rozkładu wejścia i wyjścia obiektu [5, 9, 45].

Inne zadania podobnego typu powstają przy znajomości z dokładnością do parametrów funkcji klasyfikujących lub powierzchni rozdzielających [2, 4, 22], albo też charakterystyki obiektu [2, 3, 9, 45, 46].

II. ASYMPTOTYCZNA OPTYMALNOŚĆ ALGORYTMÓW UCZENIA ROZPOZNAWANIA

2.1. Wstęp

W rozdziale tym zajmiemy się asymptotycznymi własnościami algorytmów uczenia rozpoznawania wykorzystujących nieparametryczne oszacowania gęstości prawdopodobieństwa. Wykażemy, że przy pewnych, bardzo ogólnych założeniach, algorytmy te, przy wzroście długości ciągu uczącego stają się coraz bliższe optymalnemu algorytmowi rozpoznawania Bayesa, tzn. takiemu, który można byłoby wyznaczyć przy znajomości rozkładów w klasach i ich prawdopodobieństw. Szczególną uwagę zwrócimy na asymptotyczne własności zrandomizowanej reguły rozpoznawania i odpowiadającego jej ryzyka. Omówimy także asymptotyczne własności znanego algorytmu wykorzystującego metodę funkcji potencjalnych.

Uczenie rozpoznawania z zastosowaniem nieparametrycznych estymatorów gęstości badali już Van Ryzin [68, 70], Wolverton i Wagner [81], Bubnicki [11], Glick [27] oraz autor [29, 32] i wykazali, przy mniej lub bardziej ogólnych założeniach, jego asymptotycznie bayesowskie własności. Wyniki te omówione zostaną szczegółowo w dalszych częściach pracy.

Jak wynika z wzoru (1.2), dla reguły klasyfikacyjnej ψ ryzyko wyraża się wzorem

$$R(\psi) = \sum_{j=1}^{M} p_j \int_{\mathcal{X}} L(\psi(x), j)\, f_j(x)\, dx, \tag{2.1}$$

w którym, jak wiadomo, p_j jest apriorycznym prawdopodobieństwem pojawienia się obrazu z klasy j, a f_j gęstością o tej klasie.

Reguła rozpoznawania ψ^*, która dla wszystkich $x \in \mathcal{X}$ spełnia równość

$$\sum_{j=1}^{M} L(\psi^*(x), j) p_j f_j(x) = \min_i \sum_{j=1}^{M} L(i,j)\, p_j f_j(x) \tag{2.2}$$

jest oczywiście optymalną regułą Bayesa.

Oznaczmy przez Φ^* zbiór wszystkich takich optymalnych reguł rozpoznawania. Niech ponadto Φ^*_x oznacza zbiór wszystkich klas, dla których spełniona jest równość (2.2). Będziemy uważać, że reguła ψ klasyfikuje obraz x optymalnie, jeśli $\psi(x) \in \Phi^*_x$. Jeżeli zatem optymalna klasyfikacja obrazu x jest jednoznaczna, to zbiór Φ^*_x jest jednoelementowy tzn. $\Phi^*_x = \{\psi^*(x)\}$, przy czym $\psi^* \in \Phi^*$.

Jak już zaznaczono w rozdziale I, nie znane są ani prawdopodobieństwa wystąpienia klas $p_1, \ldots, p_M$, ani gęstości obrazów w klasach $f_1, \ldots, f_M$. Jedyną informacją jest ciąg uczący. Oznacza to, że zaobserwowano zarówno obrazy jak i ich klasy. Nie znane prawdopodobieństwa p_i szacuje się ilorazami

$$p_{in} = \frac{n_i}{n}, \tag{2.3}$$

$i=1, \ldots, M$, przy czym n_i jest zaobserwowaną w ciągu uczącym ilością obrazów z klasy i, f_i natomiast - za pomocą nieparametrycznego estymatora gęstości prawdopodobieństwa f_{in}, który każdej realizacji ciągu uczącego $(\omega_1, x_1), \ldots, (\omega_n, x_n)$ i punktowi $x \in \mathcal{X}$ przyporządkowuje liczbę $f_{in}(\omega_1, x_1, \ldots, \omega_n, x_n, x) \stackrel{def}{=} f_{in}(x)$.

Niech teraz ψ_n będzie empiryczną regułą rozpoznawania, tzn. funkcją, która każdej realizacji ciągu uczącego i punktowi $x \in \mathcal{X}$ przyporządkowuje jedną z M klas. Ich ciąg $\{\psi_n\}$ nazywa się algorytmem uczenia rozpoznawania. W tej pracy będziemy rozpatrywać algorytmy uczenia rozpoznawania, które obraz x zaliczają do dowolnej z klas $i \in \Omega$ dla której

$$\min_i \sum_{j=1}^{M} L(i,j)\, p_{jn}\, f_{jn}(\omega_1, x_1, \ldots, \omega_n, x_n, x) = \sum_{j=1}^{M} L(i,j)\, p_{jn}\, f_{jn}(\omega_1, x_1, \ldots, \omega_n, x_n, x). \tag{2.4}$$

Niech T będzie klasą wszystkich takich algorytmów uczenia działających według wzoru (2.4). Empiryczna reguła rozpoznawania oraz ciąg uczący określają zrandomizowaną regułę (uczenia) rozpoznawania oznaczaną dalej, zgodnie z przyjętą zasadą, także symbolem ψ_n. Konsekwentnie będziemy, także dla wygody, symbolem T oznaczać odpowiedni zbiór wszystkich zrandomizowanych reguł uczenia rozpoznawania wyznaczony przez zbiór T empirycznych reguł i ciąg uczący. Zrandomizowana reguła ψ_n zalicza obraz x do klasy i z prawdopodobieństwem $P\{\psi_n(\Omega_1, X_1, \ldots, \Omega_n, X_n, x) = i\} \stackrel{def}{=} P\{\psi_n(x) = i\}$.

Problem, który teraz pojawia się jest następujący: jak estymować gęstości w klasach, aby zapewnić zbieżność ciągu zrandomizowanych reguł uczenia rozpoznawania z klasy T do optymalnej reguły Bayesa, a ryzyka do minimalnego ryzyka Bayesa? W dalszej części rozdziału zbadamy asymptotyczne własności algorytmów uczenia z klasy T, w sytuacji, gdy

$$f_{in}(x) \longrightarrow f_i(x) \qquad \text{lub} \qquad \int_{\mathcal{X}} (f_i(x) - f_{in}(x))^2 dx \longrightarrow 0$$

w odpowiednim sensie, gdy $n \longrightarrow \infty$.

Oznaczmy jeszcze przez $\rho(i, \Omega')$ odległość między decyzją i oraz zbiorem decyzji $\Omega' \subset \Omega$;

$$\rho(i, \Omega') = \min_{j \in \Omega'} |i-j|.$$

Niech dalej

$$g_i(x) \stackrel{\text{def}}{=} \sum_{j=1}^{M} L(i,j)\, p_j f_j(x), \qquad g_{in}(x) \stackrel{\text{def}}{=} \sum_{j=1}^{M} L(i,j)\, p_{jn} f_{jn}(x). \tag{2.5}$$

W sytuacjach nie budzących wątpliwości zamiast np. $\int_{\mathcal{X}} dx$ będziemy pisać $\int dx$.

2.2. Zbieżność ciągu zrandomizowanych reguł uczenia rozpoznawania

Omówimy zachowanie się ciągu zrandomizowanych reguł uczenia rozpoznawania z klasy T. Rozpatrzymy teraz dwa przypadki, gdy oszacowania są zbieżne do nieznanych gęstości prawdopodobieństwa albo według prawdopodobieństwa, albo z prawdopodobieństwem 1. Podane poniżej twierdzenia 2.1, 2.2 oraz 2.3 są uogólnieniem rezultatów przedstawionych we wcześniejszej pracy autora [29].

2.2.1. Zgodna estymacja gęstości prawdopodobieństwa

Przedstawione dalej twierdzenie podaje warunki, w których ciąg zrandomizowanych reguł uczenia rozpoznawania $\{\psi_n\} \in T$ jest, w ustalonym punkcie $x \in \mathcal{X}$, zbieżny według prawdopodobieństwa do optymalnej reguły Bayesa lub do zbioru optymalnych reguł Bayesa.

Twierdzenie 2.1

Jeśli estymator gęstości prawdopodobieństwa jest zgodny w punkcie $x \in \mathcal{X}$ tzn.

$$f_{in}(x) \xrightarrow{p} f_i(x),$$

gdy $n \to \infty$, dla $i=1, \ldots, M$, to dla dowolnego algorytmu uczenia $\{\psi_n\} \in T$

$$\rho(\psi_n(x), \Phi_x^*) \xrightarrow{p} 0 \tag{2.6}$$

gdy $n \to \infty$.

D o w ó d

Zauważmy najpierw, że dla dowolnych zdarzeń $A_1, \ldots, A_k$

$$P\left\{\bigcap_{i=1}^{k} A_i\right\} \geq \sum_{i=1}^{k} P\{A_i\} - (k-1) . \tag{2.7}$$

Z założenia i (2.3) wynika, że dla wszystkich $i=1, \ldots, M$ także

$$g_{in}(x) \xrightarrow{p} g_i(x) , \tag{2.8}$$

gdy $n \to \infty$.

Niech

$$\varepsilon = \frac{1}{3} \min_{i \notin \Phi_x^*} (g_i(x) - g_j(x)), \quad j \in \Phi_x^* . \tag{2.9}$$

Wybierzmy $\delta > 0$. Zbieżność (2.8) oznacza, że istnieje takie N, że dla $n > N$

$$P\{|g_i(x) - g_{in}(x)| < \varepsilon\} > 1 - \delta/2M \tag{2.10}$$

dla wszystkich $i=1, \ldots, M$. Stąd i z (2.9) mamy, dla dowolnych $j \in \Phi_x^*$ oraz $i \notin \Phi_x^*$

$$P\{g_{in}(x) > g_{jn}(x)\} \geq P\{|g_i(x) - g_{in}(x)| < \varepsilon \ , \ g_i(x) - g_j(x) \geq 3\varepsilon \ ,$$

$$|g_j(x) - g_{jn}(x)| < \varepsilon\} = P\{|g_i(x) - g_{in}(x)| < \varepsilon, |g_j(x) - g_{jn}(x)| < \varepsilon\}.$$

Na podstawie (2.7) i (2.10) wnioskujemy, że dla $n > N$

$$P\{g_{in}(x) > g_{jn}(x)\} > 1 - \delta/M .$$

Niech teraz $j \in \Phi_x^*$. Z powyższej nierówności wynika, że

$$P\{\rho(\psi_n(x), \Phi_x^*) = 0\} = P\{\psi_n(x) \in \Phi_x^*\} \geq P\left\{\bigcap_{i \notin \Phi_x^*} (g_{in}(x) > g_{jn}(x))\right\}.$$

Powtórnie wykorzystując (2.7) i (2.10) otrzymujemy dla $n > N$

$$P\{\rho(\psi_n(x), \Phi_x^*) = 0\} > 1 - \delta .$$

Ponieważ δ było dowolne, to wynika stąd teza, co kończy dowód. ■

Wykazaliśmy więc, że jeśli estymator gęstości prawdopodobieństwa jest zgodny w punkcie $x \in \mathcal{X}$, to odległość między zrandomizowaną regułą uczenia rozpoznawania i zbiorem optymalnych decyzji w tymże punkcie maleje do zera według prawdopodobieństwa, gdy długość ciągu uczącego wzrasta.

Teza (2.6) oznacza oczywiście, że

$$\lim_{n\to\infty} P\left\{\psi_n(x) \in \Phi_x^*\right\} = 1 ,$$

tzn., że prawdopodobieństwo optymalnego rozpoznania obrazu x zdąża do 1. Jeśli decyzja optymalna jest w punkcie x jednoznaczna, to z twierdzenia wynika, że

$$\lim_{n\to\infty} P\left\{\psi_n(x) = \phi^*(x)\right\} = 1, \qquad (2.11)$$

przy czym ϕ^* jest regułą optymalną. Zauważmy ponadto, że ponieważ odległość jest ograniczona, to także

$$\lim_{n\to\infty} E\ \rho^2(\psi_n(x), \Phi_x^*) = 0. \qquad (2.12)$$

Jeśli optymalna decyzja jest w punkcie x jednoznaczna, oznacza to , że

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

2.2.2. Mocno zgodna estymacja gęstości

Omówimy teraz zbieżność ciągu zrandomizowanych reguł uczenia rozpoznawania w sytuacji, gdy estymator gęstości prawdopodobieństwa jest mocno zgodny.

Twierdzenie 2.2

Jeśli estymator gęstości prawdopodobieństwa jest mocno zgodny w punkcie $x \in \mathcal{X}$, tzn.

$$f_{in}(x) \longrightarrow f_i(x) \qquad \text{z p. 1,}$$

gdy $n\to\infty$, dla wszystkich i=1, ..., M, to dla dowolnego algorytmu uczenia $\{\psi_n\} \in T$

$$\rho(\psi_n(x), \Phi_x^*) \longrightarrow 0 \qquad \text{z p. 1,} \qquad (2.13)$$

gdy $n\to\infty$.

D o w ó d

Zauważmy, że z założenia wynika, że także

$$g_{in}(x) \longrightarrow g_i(x) \qquad \text{z p. 1,} \qquad (2.14)$$

gdy $n\to\infty$, dla wszystkich i=1, ..., M.

Wybierzmy $\delta > 0$. Z (2.14) wynika, że istnieje N takie,że dla $n > N$

$$P\left\{|g_i(x) - g_{in}(x)| < \varepsilon \quad \text{dla wszystkich} \quad n > N\right\} > 1 - \delta/2M, \qquad (2.15)$$

przy czym ε określone jest wzorem (2.9). Dla dowolnych $j \in \Phi_x^*$ oraz $i \notin \Phi_x^*$ otrzymujemy na podstawie (2.9)

$$P\{g_{in}(x) > g_{jn}(x) \quad \text{dla wszystkich } n > N\} \geq$$

$$\geq P\{|g_i(x) - g_{in}(x)| < \varepsilon \ , \ g_i(x) - g_j(x) > 3\varepsilon \ , \ |g_j(x) - g_{jn}(x)| < \varepsilon$$

dla wszystkich $n > N\} = P\{|g_i(x) - g_{in}(x)| < \varepsilon \ , \ |g_j(x) - g_{jn}(x)| < \varepsilon$ dla wszystkich $n > N\}$. Stąd na podstawie (2.7) i (2.15)

$$P\{g_{in}(x) > g_{jn}(x) \quad \text{dla wszystkich } n > N\} > 1 - \delta/M.$$

Niech teraz $j \in \Phi^*_x$. Z powyższej nierówności i z (2.7) wynika, że

$$P\{\sup_{n > N} \rho(\psi_n(x), \Phi^*_x) = 0\} = P\{\psi_n(x) \in \Phi^*_x \quad \text{dla wszystkich } n > N\} \geq$$

$$\geq P\{\bigcap_{i \notin \Phi^*_x} (g_{in}(x) - g_{jn}(x) \quad \text{dla wszystkich } n > N)\} > 1 - \delta .$$

Ponieważ δ było dowolne, to dowodzi to tezy i kończy dowód. ■

Wykazaliśmy więc, że jeśli estymator gęstości jest w punkcie $x \in \mathcal{X}$ mocno zgodny, to odległość między zrandomizowaną regułą uczenia rozpoznawania i zbiorem optymalnych decyzji w tymże punkcie zdąża do zera z prawdopodobieństwem 1, gdy długość ciągu uczącego rośnie. Ponieważ zbieżność z prawdopodobieństwem 1 implikuje zbieżność według prawdopodobieństwa, te prawdziwe są wszystkie wnioski podane po twierdzeniu 2.1

2.3. Zbieżność ryzyka

Uczenie rozpoznawania można oceniać badając nie tylko regułę decyzyjną, lecz także jej skuteczność, tzn. ryzyko. Dla zrandomizowanej reguły uczenia z ciągu $\{\psi_n\} \in T$ ryzyko

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

jest oczywiście zmienną losową. Zbadamy teraz jego zbieżność do minimalnego ryzyka Bayesa

$$R^* \overset{\text{def}}{=} R(\psi^*) = \sum_{j=1}^{M} p_j \int L(\psi^*(x), j) \, f_j(x) \, dx \ ,$$

przy czym $\psi^* \in \Phi^*$.

2.3.1. Zgodna estymacja gęstości prawdopodobieństwa

Niech $\{\psi_n\}$ będzie dowolną procedurą uczenia rozpoznawania z klasy T. Zauważmy, że

$$\sum_{j=1}^{M} p_j \, f_j(x) \left[L(\psi_n(x), j) - L(\psi^*(x), j)\right] \leq L \, \rho(\psi_n(x), \Phi^*_x) \quad f(x) \ ,$$

przy czym $\psi^* \in \Phi^*$, $L = \max_{i,j} L(i,j)$;

zatem

$$0 \leq R(\psi_n) - R^* \leq L \int \rho(\psi_n(x), \Phi^*_x) f(x)\, dx . \qquad (2.16)$$

Z powyższej nierówności, twierdzeń 2.1 i 2.2 oraz twierdzenia Lebesgue'a o zbieżności wynika następujące twierdzenie:

Twierdzenie 2.3

Jeśli estymator gęstości jest (mocno) zgodny w prawie wszystkich (μ) punktach $x \in \mathcal{X}$, tzn. dla i=1, ..., M

$$f_{in}(x) \xrightarrow[z\ p.\ 1]{p} f_i(x) , \qquad (2.17)$$

gdy $n \to \infty$, w prawie wszystkich $x \in \mathcal{X}$, to dla dowolnego algorytmu uczenia rozpoznawania $\{\psi_n\} \in T$ odpowiednio

$$R(\psi_n) \xrightarrow[z\ p.\ 1]{p} R^* , \qquad (2.18)$$

gdy $n \to \infty$ oraz

$$\int \rho(\psi_n(x), \Phi^*_x) f(x)\, dx \xrightarrow[z\ p.\ 1]{p} 0 , \qquad (2.19)$$

gdy $n \to \infty$. W obydwu przypadkach także

$$\lim_{n \to \infty} E\, R(\psi_n) = R^* . \qquad (2.20)$$

Należy zaznaczyć, że podobny rezultat otrzymał Glick [27] choć w zupełnie inny sposób. Założył jednak dodatkowo, że dla wszystkich i=1, ..., M

$$\int f_{in}(x)\, dx \xrightarrow[z\ p.\ 1]{p} 1 ,$$

gdy $n \to \infty$.

Jako prosty wniosek z twierdzenia otrzymujemy, że jeśli optymalne decyzje są dla prawie wszystkich $x \in \mathcal{X}$ jednoznaczne, to dla dowolnej, optymalnej reguły $\psi^* \in \Phi^*$

$$\int (\psi^*(x) - \psi_n(x))^2 f(x)\, dx \xrightarrow[z\ p.\ 1]{p} 0 , \qquad (2.21)$$

gdy $n \to \infty$, a także

$$\lim_{n \to \infty} E \int (\psi^*(x) - \psi_n(x))^2 f(x)\, dx = 0. \qquad (2.22)$$

2.3.2. Całkowo zgodna estymacja gęstości

Twierdzenie podane poniżej dotyczy sytuacji, gdy estymator gęstości jest zgodny lub mocno zgodny w sensie całkowym. Jest ono rozszerzeniem rezultatów otrzymanych przez Wolvertona i Wagnera [81] na przypadek wielu klas i dowolnej funkcji strat.

Twierdzenie 2.4

Jeśli dla wszystkich $i=1, \ldots, M$

$$\int (f_i(x) - f_{in}(x))^2 \, dx \xrightarrow[\text{z p. 1}]{p} 0 , \tag{2.23}$$

gdy $n \to \infty$, to dla dowolnego algorytmu uczenia rozpoznawania $\{\psi_n\} \in T$

$$R(\psi_n) \xrightarrow[\text{z p. 1}]{p} R^*, \tag{2.24}$$

gdy $n \to \infty$. W obydwu przypadkach także

$$\lim_{n \to \infty} E\, R(\psi_n) = R^*. \tag{2.25}$$

D o w ó d

Dla dowodu wygodnie jest ryzyko zapisać w innej formie niż (2.1). Reguła ψ rozbija przestrzeń obrazów $\mathcal{X}$ na M rozłącznych zbiorów $A_1, \ldots, A_M$ takich, że

$$A_i \overset{\text{def}}{=} \{x : \psi(x) = i\}.$$

Ryzyko można więc wyrazić następująco:

$$R(\psi) = \sum_{i=1}^{M} \int_{A_i} \sum_{j=1}^{M} L(i,j)\, p_j\, f_j(x)\, dx. \tag{2.26}$$

Łatwo sprawdzić, że założenie (2.23) implikuje zbieżność

$$\int (g_i(x) - g_{in}(x))^2 \, dx \xrightarrow[\text{z p. 1}]{p} 0 , \tag{2.27}$$

gdy $n \to \infty$, dla $i=1, \ldots, M$.

Oznaczmy przez I_A funkcję charakterystyczną zbioru $A \subset \mathcal{X}$, a przez $\bar{A}$ jego dopełnienie.

Wybierzmy $\delta > 0$. Niech $B \subset \mathcal{X}$ będzie takim zbiorem, że $\mu(B) < \infty$ oraz

$$\int_{\bar{B}} g_i(x)\, dx < \delta/4\, M \tag{2.28}$$

dla $i=1, \ldots, M$. Dla dowolnej procedury $\{\psi_n\} \in T$ oraz ustalonej realizacji ciągu uczącego otrzymujemy na podstawie (2.26)

$$0 \leq R(\psi_n) - R^* = \sum_{i=1}^{M} \int_{A_{in}} g_i(x)\,dx - \sum_{i=1}^{M} \int_{A_i^*} g_i(x)\,dx ,$$

gdzie $A_{1n}, \ldots, A_{Mn}$ oraz $A_1^*, \ldots, A_M^*$ są rozbiciami odpowiadającymi ψ_n (przy ustalonej realizacji ciągu uczącego) oraz optymalnej regule ψ^*. Stąd

$$0 \leq R(\psi_n) - R^* = \sum_{i=1}^{M} \int_{B} g_i(x)\,(I_{A_{in}}(x) - I_{A_i^*}(x))\,dx +$$

$$+ \sum_{i=1}^{M} \int_{\overline{B}} g_i(x)\,(I_{A_{in}}(x) - I_{A_i^*}(x))\,dx .$$

Z własności ψ_n wynika, że

$$\sum_{i=1}^{M} \int_{B} g_{in}(x)\,(I_{A_i^*}(x) - I_{A_{in}}(x))\,dx \geq 0 .$$

Dodając powyższe dwa wyrażenia i wykorzystując (2.28) oraz nierówność Schwartza otrzymujemy

$$0 \leq R(\psi_n) - R^* \leq 2 \sum_{i=1}^{M} \int_{B} |g_i(x) - g_{in}(x)|\,dx + \delta/2 \leq$$

$$\leq 2(\mu(B))^{1/2} \sum_{i=1}^{M} \left(\int_{B} (g_i(x) - g_{in}(x))^2\,dx\right)^{1/2} + \delta/2 .$$

Ponieważ δ było dowolne, to wynika stąd (2.24), a w konsekwencji (2.25), co kończy dowód. ■

Warto jeszcze zaznaczyć, że twierdzenie to, podobnie jak i oryginalne twierdzenie Wolvertona i Wagnera, jest uogólnieniem wcześniejszych rezultatów Van Ryzina [68], który zakładał, że przestrzeń obrazów jest ograniczona.

2.4. Zależność między zbieżnościami reguły i ryzyka

Z nierówności (2.16) wynika, że odpowiednia zbieżność ciągu reguł uczenia implikuje zbieżność ryzyka. Interesujący jest problem, czy zbieżność ryzyka implikuje zbieżność reguły?

Załóżmy, że (co praktycznie nie stanowi ograniczenia)

$$L(i,i) < L(i \neq j, j) \tag{2.29}$$

dla wszystkich i oraz j, co oznacza, że strata dla klasyfikacji błędnych jest większa niż dla prawidłowych.

Twierdzenie 2.5

Jeśli funkcja strat spełnia nierówność (2.29), to

$$R(\psi_n) \xrightarrow[\text{z p. 1}]{p} R^*,$$

gdy $n \to \infty$, lub

$$\lim_{n\to\infty} E\, R(\psi_n) = R^*$$

wtedy i tylko wtedy, gdy odpowiednio

$$\int \rho(\psi_n(x), \Phi^*_x)\, f(x)\, dx \xrightarrow[\text{z p. 1}]{p} 0,$$

gdy $n \to \infty$, lub

$$\lim_{n\to\infty} E \int \rho(\psi_n(x), \Phi^*_x)\, f(x)\, dx = 0.$$

D o w ó d

Niech

$$\min_{\substack{i,j \\ i\neq j}} L(i,j) = 1.$$

Dla dowolnej reguły optymalnej $\psi^* \in \Phi^*$ zatem

$$0 \leq \frac{1}{M} \int \rho(\psi_n(x), \Phi^*_x)\, f(x)\, dx \leq R(\psi_n) - R^*.$$

Stąd i z nierówności (2.16) wynika teza, co kończy dowód. ■

Jeśli więc optymalne decyzje są dla prawie wszystkich $x \in \mathcal{X}$ jednoznaczne, to warunkami koniecznymi i wystarczającymi na to, aby zachodziły zbieżności

$$R(\psi_n) \xrightarrow[\text{z p. 1}]{p} R^*,$$

gdy $n \to \infty$, lub

$$\lim_{n\to\infty} E\, R(\psi_n) = R^*$$

są odpowiednio

$$\int (\psi^*(x) - \psi_n(x))^2 f(x)\, dx \xrightarrow[\text{z p.1}]{p} 0,$$

gdy $n \to \infty$ oraz

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

2.5. Metoda funkcji potencjalnych

Asymptotycznie optymalne własności procedury uczenia rozpoznawania zapewnia także metoda funkcji potencjalnych [2, 3, 4, 8, 35]. Dla omówienia jej załóżmy, że funkcja strat jest typu 0-1 (tzn. $L(i,j) = 1$ dla $i \neq j$ oraz $L(i,j) = 0$ dla $i=j$) i oznaczmy przez $L^2(f)$ przestrzeń funkcji określonych na $\mathcal{X}$ takich, że dla $t \in L^2(f)$

$$\int t^2(x)\, f(x)\, dx < \infty ,$$

gdzie, jak wiadomo, $f(x) = \sum_{i=1}^{M} p_i f_i(x)$. Przez funkcję potencjalną rozumie się

$$K(x,y) \sim \sum_{i=0}^{\infty} \lambda_i^2 \, \varphi_i(x) \, \varphi_i(y), \quad x,y \in X ,$$

gdzie $\{\varphi_i\}_{i=0}^{\infty}$ jest dowolnym układem funkcji w przestrzeni $L^2(f)$ oraz

$$\sum_{i=0}^{\infty} \lambda_i^2 < \infty .$$

Załóżmy, że prawdziwa jest tzw."podstawowa hipoteza" metody [2, 8] tzn.

$$t_j(x) \overset{\text{def}}{=} \frac{p_j f_j(x)}{f(x)} \sim \sum_{i=0}^{\infty} c_{ji} \, \varphi_i(x), \quad j=1, \ldots, M, \tag{2.30}$$

przy czym

$$\sum_{i=0}^{\infty} \left(\frac{c_{ji}}{\lambda_i}\right)^2 < \infty \quad , \; j=1, \ldots, M. \tag{2.31}$$

W takich sytuacjach można zastosować następujący algorytm rekurencyjny:

$$t_{j,n}(x) = t_{j,n-1}(x) + r_{jn} \, K(x,x_n) , \tag{2.32}$$

$n = 1,2, \ldots$, w którym współczynniki r_{jn} wyznacza się w opisany poniżej sposób.

Dla funkcji $t \in L^2(f)$ określmy nową funkcję $\bar{t}$ w następujący sposób:

$$\bar{t}(x) = \begin{cases} t(x) & \text{jeśli} \quad 0 \leq t(x) < 1 , \\ 0 & \text{jeśli} \quad t(x) < 0, \\ 1 & \text{jeśli} \quad t(x) > 1. \end{cases}$$

Kolejny, n-ty obraz ciągu uczącego, tzn. x_n rozpoznaje się z prawdopodobieństwem $\bar{t}_{j,n-1}(x_n)$ jako przychodzący z klasy j oraz z

prawdopodobieństwem $1 - \bar{t}_{j,n-1}(x_n)$ jako należący do jednej z pozostałych. Zachodzi wtedy jedna z trzech sytuacji. Albo rozpoznanie w klasie j (tzn. dla j-go z algorytmów (2.32) było prawidłowe - wówczas $r_{jn} = 0$, albo obraz należał do klasy j i został rozpoznany błędnie - wówczas $r_{jn} = \gamma_n$, albo wreszcie obraz błędnie zaliczono do klasy j - wówczas przyjmuje się $r_{jn} = -\gamma_n$.

Jeśli teraz

$$\gamma_n > 0, \quad \sum_{n=0}^{\infty} \gamma_n = \infty , \quad \sum_{n=0}^{\infty} \gamma_n^2 < \infty$$

oraz

$$t_{j,o}(x) \sim \sum_{i=0}^{\infty} c_{ji}^{o} \varphi_i(x) ,$$

przy czym

$$\sum_{i=0}^{\infty} \left(\frac{c_{ji}^{o}}{\lambda_i} \right)^2 < \infty ,$$

to

$$\int (t_j(x) - t_{j,n}(x))^2 f(x)\, dx \xrightarrow{P} 0 , \qquad (2.33)$$

gdy $n \to \infty$ [8].

Algorytm uczenia jest więc zbieżny w sensie (2.33), jeśli prawdziwa jest hipoteza (2.30) - (2.31). O zbieżności tej decyduje wybór funkcji potencjalnej, który określa klasę gęstości, względem której procedura jest zbieżna. Ustalenie jej jest więc najistotniejszym problemem w stosowaniu metody. Jeśli $\{\varphi_i\}_{i=0}^{\infty}$ jest pełnym układem ortonormalnym w przestrzeni $L^2(f)$, to

$$c_{ji} = \int t_j(x)\, \varphi_i(x)\, f(x)\, dx .$$

Warunek (2.31) narzuca więc ograniczenie na współczynniki rozwinięcia funkcji t_j. Algorytm funkcji potencjalnych jest więc zbieżny dla funkcji t_j z podprzestrzeni przestrzeni $L^2(f)$ określonej nierównością (2.31). Podprzestrzeń tę ustala się pośrednio przez wybór funkcji potencjalnej, tzn. współczynników $\lambda_1, \lambda_2, \ldots$. Warto dodać, że wykazano zbieżność procedury typu (2.33) także z prawdopodobieństwem 1 [35].

Twierdzenie 2.4 pozwala sformułować nowe, asymptotycznie optymalne własności metody funkcji potencjalnych. Zauważmy bowiem, że punktem wyjściowym dowodu jest założenie o zbieżności (2.27). Ponieważ dla funkcji strat typu 0-1

$$g_i(x) = f(x)\, (1 - t_i(x)) ,$$

położmy

$$g_{in}(x) \overset{\text{def}}{=} f(x)\,(1 - t_{in}(x)),$$

stąd

$$\int (g_i(x) - g_{in}(x))^2\,dx \leq \sup_x f(x) \int (t_i(x) - t_{in}(x))^2 f(x)\,dx .$$

Jeśli teraz gęstość bezwarunkowa f jest ograniczona, to z powyższej nierówności wynika, że zbieżność (2.33) implikuje (2.27), z której z kolei w myśl twierdzenia 2.4 wynika asymptotyczna optymalność w sensie średniego ryzyka algorytmu uczenia rozpoznawania według metody funkcji potencjalnych.

2.6. Uwagi

Twierdzenia podane w tym rozdziale pozwalają, przez zastosowanie nieparametrycznych metod estymacji gęstości prawdopodobieństwa, otrzymywać algorytmy uczenia rozpoznawania, których własności zbliżają się przy wzroście długości ciągu uczącego, do optymalnego algorytmu Bayesa. Asymptotyczną optymalność można przy tym określać w różny sposób, albo jako zbieżność ciągu zrandomizowanych reguł uczenia albo ryzyka.

Do badania zbieżności reguł wygodne jest wprowadzenie odległości ρ, oceniającej odległość między regułą decyzyjną, a zbiorem decyzji optymalnych. Asymptotyczną optymalność dla ciągu reguł można wówczas określić kilkoma sposobami, w zależności od zachodzenia odpowiedniej z następujących zbieżności:

$$\rho(\psi_n(x), \Phi^*_x) \xrightarrow[\text{z p. 1}]{p} 0, \qquad E\,\rho(\psi_n(x), \Phi^*_x) \longrightarrow 0,$$

gdy $n \to \infty$.

Zbieżność reguł można badać także w całej przestrzeni obrazów definiując odpowiednio asymptotyczną optymalność w zależności od zachodzenia następujących własności

$$\int \rho(\psi_n(x), \Phi^*_x) f(x)\,dx \xrightarrow[\text{z p. 1}]{p} 0, \qquad E \int (\psi_n(x), \Phi^*_x) f(x)\,dx \longrightarrow 0 ,$$

gdy $n \to \infty$.

Ostatnią z zastosowanych ocen uczenia rozpoznawania jest ryzyko. Korzystanie z niego, w celu badania asymptotycznych własności, prowadzi także do trzech definicji, ponieważ można rozważać, czy

$$R(\psi_n) \xrightarrow[\text{z p. 1}]{p} R^* \quad \text{lub} \quad E\,R(\psi_n) \longrightarrow R^* ,$$

gdy $n \to \infty$.

III. ALGORYTMY UCZENIA ROZPOZNAWANIA Z ZASTOSOWANIEM NIEPARAMETRYCZNYCH OSZACOWAŃ GĘSTOŚCI

3.1. Wstęp

Twierdzenia przedstawione w poprzednim rozdziale pozwalają konstruować asymptotycznie optymalne algorytmy uczenia rozpoznawania. Zaobserwowane obrazy $x_1, \ldots, x_n$ ciągu uczącego można mianowicie rozdzielić na M podciągów:

$$x_{11}, \ldots, x_{1n_1}$$
$$x_{M1}, \ldots, x_{Mn_M},$$

z których każdy zawiera obrazy z odpowiedniej klasy. Gęstości w poszczególnych klasach estymuje się następnie na podstawie obserwacji zawartych w tych podciągach. Każdy z M estymatorów gęstości stosuje zatem losowe liczby obserwacji. Jest jednak zupełnie oczywiste, że gdy długość ciągu uczącego rośnie, to oszacowania gęstości zachowują interesujące nas własności pomimo, że wyznaczane są na podstawie losowych liczb danych. Zastosowanie na przykład w celu oszacowania $f_i(x)$ estymatora zgodnego prowadzi do oszacowania wykorzystującego tylko część obserwacji, a mianowicie n_i obrazów ciągu uczącego. Jest ono jednak zbieżne według prawdopodobieństwa do $f_i(x)$, gdy n rośnie do nieskończoności.

3.2. Nieparametryczne oszacowania gęstości prawdopodobieństwa

Korzystanie z różnych metod oszacowania gęstości prawdopodobieństwa prowadzi do różnych algorytmów uczenia rozpoznawania. Znane, nieparametryczne estymatory gęstości można podzielić na kilka typów, a mianowicie na oszacowania:

a) typu Parzena, np. [13, 41, 51, 58],
b) typu Loftsgaardena i Quesenberry'ego [44, 47],
c) otrzymywane metodą szeregów ortogonalnych [42, 63],
d) inne, np. [10, 26, 56, 69, 74, 75, 78].

Omówienie każdego ze znanych typów estymacji, a następnie odpowiadającego mu algorytmu rozpoznawania znacznie poszerzyłoby pracę i dlatego skoncentrujemy się jedynie na tych spośród najbardziej znanych, które mają pożądane przez nas własności i wymagają możliwie słabych założeń o nie znanych gęstościach. Dlatego też omówimy trzy pierwsze typy estymacji. Przegląd nieparametrycznych metod szacowania gęstości można znaleźć w pracach [57, 76, 77].

3.2.1. Estymator typu Parzena

W celu estymacji gęstości prawdopodobieństwa można zastosować estymator podany przez Rosenblatta [58], rozwinięty przez Parzena [51] i uogólniony na przypadek wielowymiarowy przez Cacoullosa [13]. Aby uniknąć skomplikowanego zapisu załóżmy chwilowo, że ciąg uczący zawiera obrazy tylko z jednej klasy (tzn. $p_1=1$) charakteryzującej się gęstością prawdopodobieństwa f (dla prostoty pomijamy indeks). Estymator Parzena ma następującą postać:

$$f_n(x) = \frac{1}{n\,h^p(n)} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h(n)}\right), \tag{3.1}$$

przy czym $\{h(n)\}$ jest pewnym ciągiem liczb, a K odpowiednio wybraną (mierzalną) funkcją określoną na przestrzeni obrazów o wymiarze p.

Jeśli

$$h(n) > 0, \quad \lim_{n\to\infty} h(n) = 0, \quad \lim_{n\to\infty} n\,h^p(n) = \infty \tag{3.2}$$

oraz

$$\sup_x |K(x)| < \infty, \quad \int K(x)dx = 1, \quad \int |K(x)|dx < \infty, \quad \lim_{\|x\|\to\infty} \|x\|^p\,|K(x)| = 0, \tag{3.3}$$

to

$$\lim_{n\to\infty} E(f(x) - f_n(x))^2 = 0 \tag{3.4}$$

w punktach ciągłości gęstości f [13].

Jako przykłady funkcji K można podać

$$\left.\begin{aligned}
&\text{a) } K(x) = \begin{cases} C & \text{jeśli } \|x\| \le 1 \\ 0 & \text{jeśli } \|x\| > 1 \end{cases} \\
&\text{b) } (2\pi)^{-p/2} e^{-\frac{1}{2}\|x\|^2}, \quad \text{przy czym } \|x\|^2 = x^T x, \\
&\text{c) } 2^{-p} e^{-\|x\|}, \quad \text{przy czym } \|x\| = \sum_{i=1}^{p} |x^{(i)}|, \\
&\text{d) } \pi^{-p} \prod_{i=1}^{p} (1 + |x^{(i)}|)^{-2}.
\end{aligned}\right\} \tag{3.5}$$

Ciągiem liczbowym może być np.

$$h(n) = cn^{-\alpha}, \quad c > 0 , \tag{3.6}$$

przy czym

$$0 < \alpha < 1/p . \tag{3.7}$$

Jeśli ponadto f jest funkcją jednostajnie ciągłą, to [23]

$$\lim_{n\to\infty} \sup_x E(f(x) - f_n(x))^2 = 0 . \tag{3.8}$$

Jeżeli f jest jednostajnie ciągła oraz

$$0 < \alpha < 1/2p, \tag{3.9}$$

to dla funkcji K podanych w przykładach b), c), d)

$$\sup_x |f(x) - f_n(x)| \xrightarrow{p} 0 , \tag{3.10}$$

gdy $n\to\infty$ [13].

Przy dodatkowych, skomplikowanych założeniach o ciągu liczbowym $\{h(n)\}$ i funkcji K można otrzymać mocną zgodność [71], tzn. zbieżność

$$f_n(x) \longrightarrow f(x) \quad \text{z p.1},$$

gdy $n\to\infty$.

Przykładami mogą być ciągi o własnościach (3.6), (3.7) oraz funkcje podane wzorami (3.5). Jeśli gęstość f jest jednostajnie ciągła, to funkcje K podane wzorem (3.5), z wyjątkiem pierwszej, i ciąg typu (3.6) o własności (3.9) zapewniają, że

$$\sup_x |f(x) - f_n(x)| \longrightarrow 0 \quad \text{z p. 1}, \tag{3.11}$$

gdy $n\to\infty$.

Prostsze warunki mocnej zgodności dla przypadku skalarnego ($p=1$) można znaleźć w pracy [49].

Pewną modyfikacją jest estymator rekurencyjny

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^p(i)} K\left(\frac{x - x_i}{h(i)}\right) \tag{3.12}$$

zaproponowany przez Wolvertona i Wagnera [82].

Łatwo zauważyć, że

$$f_{n+1}(x) = \frac{n}{n+1} f_n(x) + \frac{1}{(n+1)h^p(n+1)} K\left(\frac{x - x_{n+1}}{h(n+1)}\right) .$$

Jeśli

$$1 \geq h(1) \geq h(2) \geq \ldots > 0, \quad \sum_{n=1}^{\infty} \frac{h(n)}{n} < \infty \ , \quad \sum_{n=1}^{\infty} \frac{1}{n^2 h^p(n)} < \infty$$

oraz

$$\int K(x)dx = 1, \quad \sup_x |K(x)| < \infty \ , \quad \int \|x\| |K(x)| \, dx < \infty \ ,$$

to dla gęstości f spełniającej warunek Lipschitza

$$\int (f(x) - f_n(x))^2 \, dx \longrightarrow 0 \quad \text{z p. 1} \ ,$$

gdy $n \longrightarrow \infty$.

Dalej podamy twierdzenie o zgodności rekurencyjnego estymatora (3.12). Cytujemy je za pracą autora [32].

Twierdzenie 3.1

Jeśli ciąg $\{h(n)\}$ i funkcja K spełniają założenia (3.2) i (3.3), to dla estymatora (3.12)

$$\lim_{n \to \infty} E(f(x) - f_n(x))^2 = 0$$

w punktach ciągłości gęstości f.

D o w ó d

Dowód jest odpowiednią modyfikacją dowodu Parzena [51].

Załóżmy chwilowo, że drugie z założeń (3.3) niekoniecznie jest spełnione, lecz jedynie

$$\int K(x) \, dx < \infty .$$

Zauważmy, że

$$I_n \overset{\text{def}}{=} E f_n(x) - f(x) \int K(x)dx = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^p(i)} \int_Y (f(x-y) - f(x)) K\left(\frac{y}{h(i)}\right) dy,$$

przy czym $y \in Y = \mathcal{R}^p$.

Niech $\delta > 0$. Podzielmy teraz przestrzeń Y na dwa zbiory, w których odpowiednio $\|y\| \leq \delta$ i $\|y\| > \delta$. Zatem

$$|I_n| \leq \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{h^p(i)} \int_{\|y\| \leq \delta} |f(x-y) - f(x)| \left| K\left(\frac{y}{h(i)}\right) \right| dy + \int_{\|y\| > \delta} \frac{f(x-y)}{\|y\|^p} \frac{\|y\|^p}{h^p(i)} \left| K\left(\frac{y}{h(i)}\right) \right| dy + \right.$$

$$\left. + f(x) \int_{\|y\| > \delta} \frac{1}{h^p(i)} \left| K\left(\frac{y}{h(i)}\right) \right| dy \right] \leq \sup_{\|y\| \leq \delta} |f(x-y) - f(x)| \int |K(x)| \, dx +$$

$$+\frac{1}{\delta^{p_n}}\sum_{i=1}^{n}\sup_{\|x\|>\delta/h(i)}\|x\|^{p}|K(x)| + f(x)\frac{1}{n}\sum_{i=1}^{n}\int_{\|x\|>\delta/h(i)}|K(x)|\,dx . \tag{3.13}$$

Zauważmy, że jeśli gęstość f jest ciągła w punkcie $x\in\mathcal{X}$, to dla dowolnego $\varepsilon > 0$ istnieje $\delta > 0$ takie, że

$$\sup_{\|y\|\leqslant\delta}|f(x-y) - f(x)|\int|K(x)|dx < \varepsilon/4 .$$

Z założenia (3.3) wynika ponadto, że dla ustalonych powyżej ε i δ istnieje H takie, że dla $h < H$

$$\frac{1}{\delta^{p}}\sup_{\|x\|>\delta/h}\|x\|^{p}|K(x)| < \varepsilon/4$$

oraz

$$f(x)\int_{\|x\|>\delta/h}|K(x)|\,dx < \varepsilon/4 .$$

Stąd, z założenia (3.2) oraz (3.13) wynika, że istnieje N takie, że dla $n > N$

$$|I_n| < \frac{3}{4}\varepsilon + \frac{c}{n} ,$$

przy czym

$$c = N\left(\frac{1}{\delta^{p}}\sup_{x}\|x\|^{p}|K(x)| + f(x)\int|K(x)|dx\right).$$

Dla $n > \max(N, 4c/\varepsilon)$ zatem

$$|I_n| < \varepsilon ,$$

skąd wynika, że jeśli $x\in\mathcal{X}$ jest punktem ciągłości gęstości f, to

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^{p}(i)}EK\left(\frac{x - X_i}{h(i)}\right) = f(x)\int K(x)dx ,$$

zatem, jeśli zgodnie z założeniem

$$\int K(x)dx = 1 ,$$

to

$$\lim_{n\to\infty}Ef_n(x) = f(x) .$$

Należy jeszcze wykazać, że

$$\lim_{n\to\infty} \operatorname{var}\left|f_n(x)\right| = 0 .$$

Zauważmy w tym celu, że

$$\operatorname{var}\left|f_n(x)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^{2p}(i)}\,EK^2\left(\frac{x-X_i}{h(i)}\right) \leq \frac{1}{n^2h^p(n)}\sum_{i=1}^{n}\frac{1}{h^p(i)}\,EK^2\left(\frac{x-X_i}{h(i)}\right).$$

Z wykazanej zbieżności i założeń wynika, że

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p(i)}\,EK^2\left(\frac{x-X_i}{h(i)}\right) = f(x)\int K^2(x)dx ,$$

co kończy dowód. ■

3.2.2. Estymator Loftsgaardena i Quesenberry'ego

Niech dla ustalonej normy

$$V = \mu(\{x: \|x\| \leq 1\}),$$

w której μ jest, jak wiadomo, miarą Lebesgue'a. Loftsgaarden i Quesenberry zastosowali następujący estymator [44]:

$$f_n(x) = \frac{k(n)}{n}\,\frac{1}{VR^p(k(n))} \tag{3.14}$$

w którym $\{k(n)\}$ jest pewnym ciągiem liczbowym, a $R(k)$ odległością między punktem x i k-tą najbliższą obserwacją i przy założeniu

$$k(n) > 0, \quad \lim_{n\to\infty} k(n) = \infty , \quad \lim_{n\to\infty}\frac{k(n)}{n} = 0 \tag{3.15}$$

wykazali jego zgodność w punktach ciągłości gęstości f.

Jeśli gęstość f jest jednostajnie ciągła, to w przypadku jednowymiarowym ($p=1$) dla pewnych ciągów $\{k(n)\}$ estymator ten jest także mocno zgodny [47].

3.2.3. Estymacja metodą szeregów ortogonalnych

Estymację przez rozwinięcie gęstości w nieskończony szereg ortogonalny można uważać za rozwinięcie znanych sposobów aproksymacji gęstości skończonym szeregiem funkcji ortogonalnych [6, 7, 17, 19, 21, 38].

Niech $\{\varphi_i\}_{i=0}^{\infty}$ będzie pełnym układem wspólnie ograniczonych funkcji ortonormalnych w przestrzeni L^2 funkcji mierzalnych określonych na przestrzeni $\mathcal{X}$ o całkowalnym kwadracie. Niech ponadto $f \in L^2$ oraz

$$f(x) \sim \sum_{j=0}^{\infty} a_j \varphi_j(x)$$

i

$$a_{jn} = \frac{1}{n} \sum_{k=1}^{n} \varphi_j(x_k)$$

oraz

$$f_n(x) = \sum_{j=0}^{q(n)} a_{jn} \varphi_j(x) . \tag{3.16}$$

Jeśli

$$\lim_{n\to\infty} q(n) = \infty , \quad \lim_{n\to\infty} \frac{q(n)}{n} = 0 ,$$

to [63]

$$\lim_{n\to\infty} E \int (f(x) - f_n(x))^2 dx = 0 . \tag{3.17}$$

Jako układ ortonormalny można wybrać np. układ Hermite'a jeśli $\mathcal{X} = \mathcal{R}$ [63] lub układ trygonometryczny w sytuacji, gdy $\mathcal{X} = [a, b]$, tzn. obrazy są ograniczone [42].

Jeśli gęstość f jest ciągła i ma ograniczone wahanie oraz

$$\lim_{n\to\infty} q(n) = \infty , \quad \lim_{n\to\infty} \frac{q^2(n)}{n} = 0 , \tag{3.18}$$

to dla układu Hermite'a

$$\lim_{n\to\infty} E(f(x) - f_n(x))^2 = 0 \tag{3.19}$$

we wszystkich punktach $x \in \mathcal{X}$ [63.]

Estymator jest więc zgodny we wszystkich punktach przestrzeni $\mathcal{X}$. Podobną własność można uzyskać także dla układu trygonometrycznego, bowiem [63]

$$E(f(x) - f_n(x))^2 \leqslant \left[f(x) - \sum_{i=0}^{q(n)} a_i \varphi_i(x) \right]^2 + c^2 \frac{q^2(n)}{n} .$$

Zauważmy, że jeśli

$$\frac{q(n+1)}{q(n)} \geqslant \mu > 1 , \tag{3.20}$$

to

$$\lim_{n\to\infty} \sum_{i=0}^{q(n)} a_i \varphi_i(x) = f(x)$$

zachodzi dla prawie wszystkich $x \in \mathcal{X}$ [62]. Własność ta jest oczywiście zachowana także wtedy, gdy ciąg $\{q(n)\}$ zawiera powtórzenia. W rezultacie, jeśli $\{q(n)\}$ jest monotonicznym ciągiem z powtórzeniami spełniającym (3.18) takim, że kolejne, różne wyrazy spełniają (3.20), to zbieżność (3.19) zachodzi dla prawie wszystkich $x \in \mathcal{X}$. Estymator (3.16) jest wówczas zgodny w prawie wszystkich punktach przestrzeni $\mathcal{X}$.

3.3. Algorytmy uczenia rozpoznawania

Omówimy teraz algorytmy uczenia rozpoznawania, które otrzymuje się przez zastosowanie metod estymacji gęstości przedstawionych w pkcie 3.2. Ich odpowiednie asymptotyczne optymalne własności wynikają z twierdzeń podanych w rozdziale II.

3.3.1. Zastosowanie estymatora Parzena

Stosując estymator Parzena do oszacowania gęstości f_i na podstawie obserwacji ciągu uczącego otrzymuje się

$$f_{in}(x) = \frac{1}{n_i h(n_i)} \sum_{k=1}^{n_i} K\left(\frac{x - x_{ik}}{h(n_i)}\right). \tag{3.21}$$

Obraz x zalicza się więc do którejkolwiek z klas, dla których wyrażenie

$$\sum_{j=1}^{M} L(i,j) \frac{1}{h^p(n_j)} \sum_{k=1}^{n_j} K\left(\frac{x - x_{jk}}{h(n_j)}\right) \tag{3.22}$$

osiąga minimum. Wybierając funkcję K i ciąg $\{h(n)\}$ zgodnie z (3.2) i (3.3) otrzymuje się pożądane własności asymptotyczne, tzn. odpowiednią zbieżność reguł uczenia i ryzyka do optymalnej reguły Bayesa i minimalnego ryzyka Bayesa. Jeśli dla przykładu funkcja K jest jedną z podanych wzorami (3.5), a ciąg $\{h(n)\}$ spełnia warunki (3.2) lub (3.6) i (3.7), to jak wynika z twierdzeń 2.1 i 2.2 oraz (mocnej) zgodności estymatora, zrandomizowana reguła uczenia jest zbieżna według prawdopodobieństwa (z prawdopodobieństwem 1) do optymalnej reguły Bayesa (lub zbioru optymalnych reguł) w tych punktach przestrzeni $\mathcal{X}$, w których wszystkie gęstości $f_1, \ldots, f_M$ są ciągłe. Jeśli ciągłość ta zachodzi dla prawie wszystkich $x \in \mathcal{X}$, to, jak wynika z twierdzenia 2.3 także

$$R(\psi_n) \xrightarrow[\text{z p. 1}]{p} R^*, \quad \text{gdy} \quad n \longrightarrow \infty$$

oraz

$$\lim_{n \to \infty} E\, R(\psi_n) = R^*.$$

Ostatnia z powyższych zbieżności oznacza, że algorytm uczenia jest asymptotycznie optymalny w sensie podanym przez Robbinsa [55] względem klasy gęstości prawie wszędzie ciągłych.

Załóżmy teraz, że funkcja strat jest zero-jedynkowa, tzn.

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

Obraz x zalicza się wówczas do klasy maksymalizującej

$$\frac{1}{h^p(n_i)} \sum_{k=1}^{n_i} K\left(\frac{x - x_{ik}}{h(n_i)}\right).$$

Jeśli jako K wybierze się pierwszą z funkcji podanych wzorem (3.5), to otrzyma się bardzo prosty algorytm, który obraz x zalicza do klasy zapewniającej maksimum ułamka

$$\frac{\text{liczba obrazów z klasy } i \text{ w kuli } S(x,h(n_i))}{h^p(n_i)} \tag{3.23}$$

przy czym

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

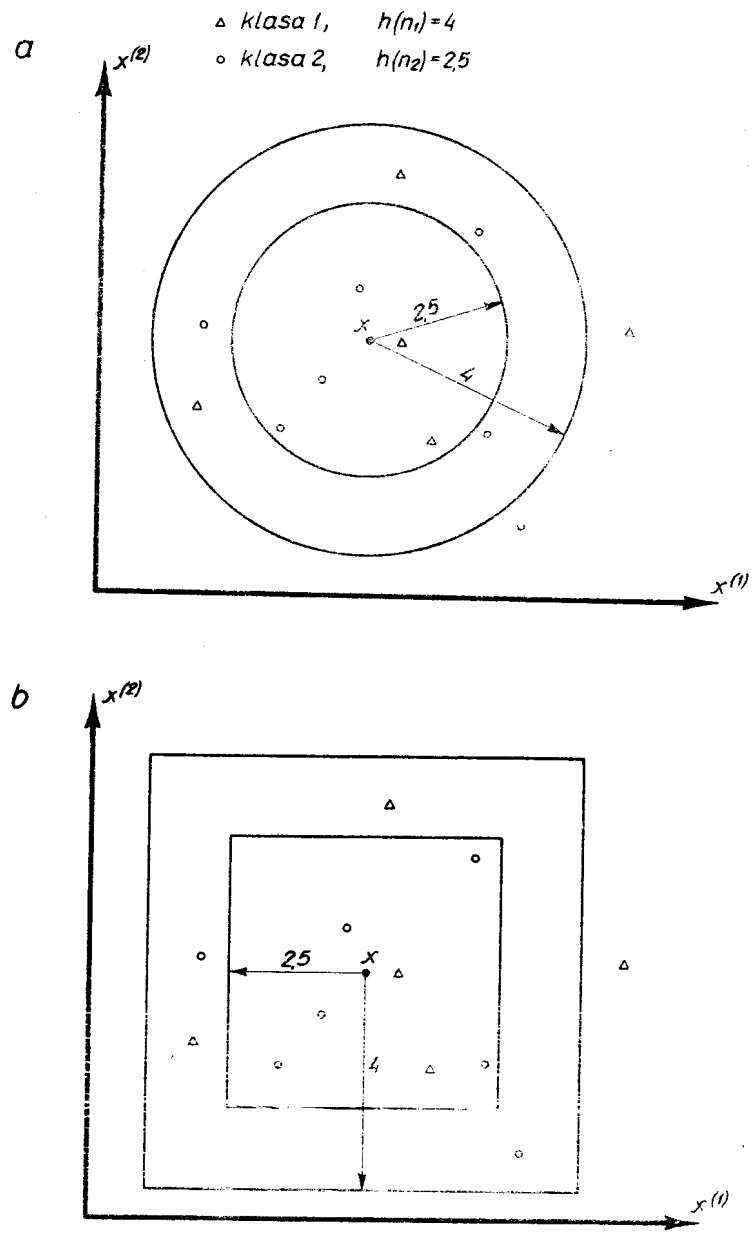
Działanie algorytmu, dla $p=2$ i problemu dychotomii, pokazano przykładowo na rys. 7. Ciągi uczące na rys. 7a i 7b są jednakowe, a jedyną różnicą są normy określające funkcję K. Dla obydwu obraz x zaliczony jest do klasy 2.

Charakterystyczną cechą algorytmu (3.23) jest zliczanie obrazów ciągu uczącego leżących w kulach, których promienie są różne dla różnych klas. Wprowadzając pewną modyfikację w sposobie estymacji gęstości prawdopodobieństwa można go jeszcze bardziej uprościć. Zamiast estymować gęstość f_i według wzoru (3.21) można wykorzystać oszacowanie nieco zmodyfikowane

$$F_{in}(x) = \frac{1}{n_i\, h^p(n)} \sum_{k=1}^{n_i} K\left(\frac{x - x_{ik}}{h(n)}\right). \tag{3.24}$$

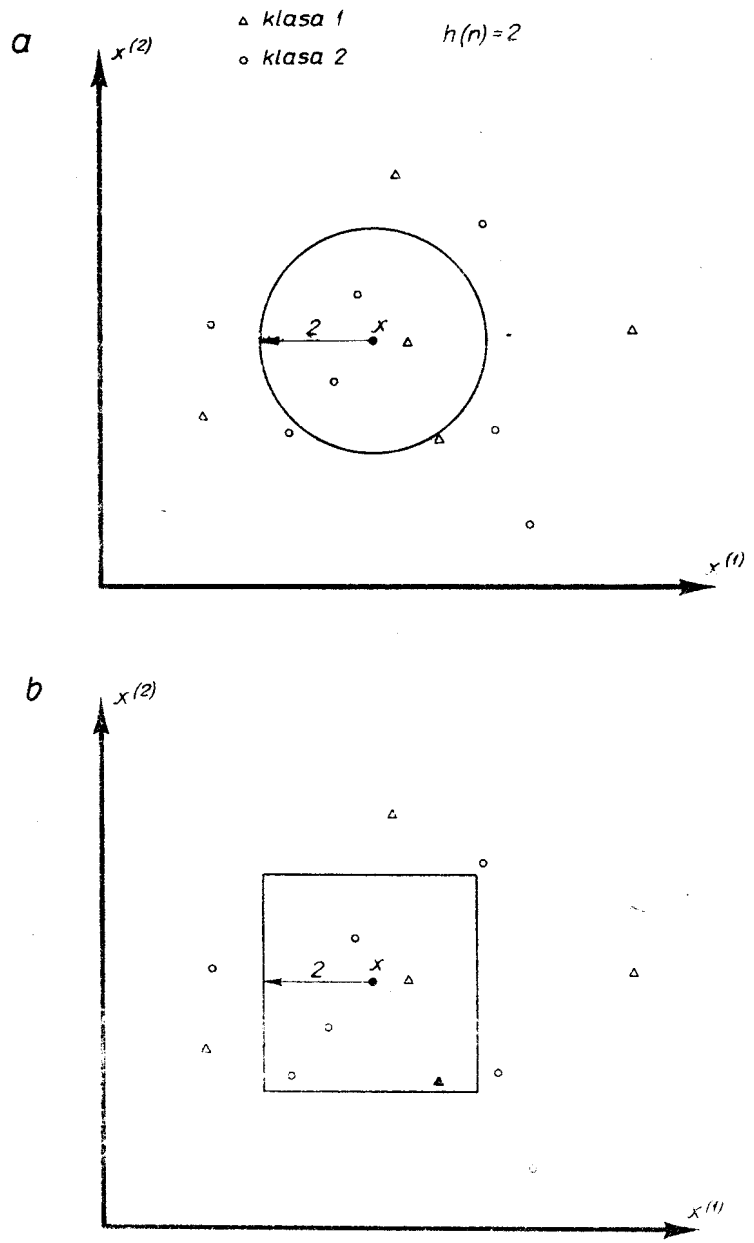
Jest oczywiste, że jeśli spełnione są założenia (3.2) i (3.3), to także i ten estymator jest (mocno) zgodny w punktach, w których estymator (3.21) jest (mocno) zgodny. Odpowiadający mu algorytm rozpoznawania zalicza obraz x do klasy, która minimalizuje

$$\sum_{j=1}^{M} L(i,j) \sum_{k=1}^{n_j} K\left(\frac{x - x_{jk}}{h(n)}\right). \tag{3.25}$$



Rys. 7. Rozpoznawanie z wykorzystaniem estymatora Parzena: a) $\|x\|^2 = x^T x$, b) $\|x\| = \max_i |x^{(i)}|$

Fig. 7. Recognizing with the Parzen estimator: a) $\|x\|^2 = x^T x$, b) $\|x\| = \max_i |x^{(i)}|$



Rys. 8. Zmodyfikowany algorytm wykorzystujący estymator Parzena: a) $\|x\|^2 = x^T x$, b) $\|x\| = \max_i |x^{(i)}|$

Fig. 8. Modified algorithm derived from the Parzen estimator: a) $\|x\|^2 = x^T x$, b) $\|x\| = \max_i |x^{(i)}|$

Dla funkcji strat typu 0-1 obraz x zalicza się więc do klasy, która maksymalizuje wyrażenie

$$\sum_{k=1}^{n_i} K\left(\frac{x - x_{ik}}{h(n)}\right). \tag{3.26}$$

Jeśli jako jądro K przyjąć pierwszą z funkcji podanych wzorami (3.5), to otrzymuje się algorytm, który obraz x rozpoznaje jako przychodzący z klasy, dla której

$$\text{liczba obrazów z klasy } i \text{ w kuli } S(x, h(n)) \tag{3.27}$$

jest największa.

Algorytm ten można nazwać algorytmem większości w kuli o promieniu $h(n)$. Jego działanie zilustrowano na rys. 8. Ciąg uczący jest taki sam jak na rys. 7, w podobny sposób wybrano także normy określające funkcje K. Dla obydwu obraz x zaliczany jest do klasy 2.

Należy jeszcze zaznaczyć, że zmodyfikowany algorytm uczenia jest także asymptotycznie optymalny według średniego ryzyka względem klasy gęstości prawdopodobieństwa ciągłych prawie wszędzie. Wybierając odpowiednio $\{h(n)\}$ i K można także zapewnić, że

$$R(\psi_n) \longrightarrow R^* \quad \text{z p. 1},$$

gdzie $n \to \infty$. Asymptotyczną optymalność, według średniego ryzyka, zmodyfikowanego algorytmu wykazał także Van Ryzin [70].

Stosowanie rekurencyjnej wersji (3.12) prowadzi do następującego oszacowania gęstości

$$f_{in}(x) = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{1}{h^p(k)} K\left(\frac{x - x_{ik}}{h(k)}\right). \tag{3.28}$$

Obraz x zalicza się więc do klasy, dla której

$$\sum_{j=1}^{M} L(i,j) \sum_{k=1}^{n_j} \frac{1}{h^p(k)} K\left(\frac{x - x_{jk}}{h(k)}\right) \tag{3.29}$$

osiąga najmniejszą wartość. Dla funkcji strat typu 0-1 obraz x zalicza się więc po prostu do klasy maksymalizującej

$$I_n(i,x) \overset{\text{def}}{=} \sum_{k=1}^{n_i} \frac{1}{h^p(k)} K\left(\frac{x - x_{ik}}{h(k)}\right).$$

Wyrażenie to można zapisać w formie rekurencyjnej

$$I_0(i,x) \equiv 0, \quad I_{n+1}(i,x) = I_n(i,x) + \xi_{n+1}(i) \frac{1}{h^p(n_i+1)} K\left(\frac{x - x_{n+1}}{h(n_i+1)}\right),$$

przy czym

$$\xi_n(i) = \begin{cases} 1 & \text{jeśli obraz } x_n \text{ należy do klasy } i \\ 0 & \text{w przeciwnym wypadku.} \end{cases}$$

Zauważmy teraz, że gdy obraz x, który należy rozpoznać, znany jest przed obserwowaniem ciągu uczącego, to przy stosowaniu powyższego algorytmu nie trzeba podczas uczenia pamiętać wszystkich obrazów ciągu uczącego, lecz jedynie M liczb $I_k(1,x),\ldots,I_k(M,x),k=1,\ldots,n$.

Asymptotycznie optymalne własności, np. według średniego ryzyka tego algorytmu uczenia rozpoznawania względem klasy gęstości prawie wszędzie ciągłych, wynikają z twierdzeń 2.1 i 2.3 oraz twierdzenia 3.1 o zgodności rekurencyjnego estymatora gęstości.

3.3.2. Zastosowanie estymatora Loftsgaardena i Quesenberry'ego

Stosowanie oszacowania Loftsgaardena i Quesenberry'ego prowadzi do algorytmu, który obraz x zalicza do klasy, która minimalizuje

$$\sum_{j=1}^{M} L(i,j) \frac{k(n_j)}{R_j^p(k(n_j))}, \tag{3.30}$$

przy czym $R_j(k(n_j))$ jest odległością między x, a $k(n_j)$ - tym najbliższym obrazem z klasy j. Dla różnych ciągów $\{k(n)\}$ otrzymuje się zgodnie z uwagami podanymi w pkcie 3.2.2, różne asymptotycznie optymalne własności uczenia rozpoznawania. Wybierając na przykład ciąg ten zgodnie z (3.14) uzyskuje się algorytm asymptotycznie optymalny według średniego ryzyka względem klasy gęstości ciągłych prawie wszędzie.

Jeśli funkcja strat jest typu 0-1, to obraz x zalicza się do klasy minimalizującej

$$\frac{1}{k(n_i)} R_i^p(k(n_i)) . \tag{3.31}$$

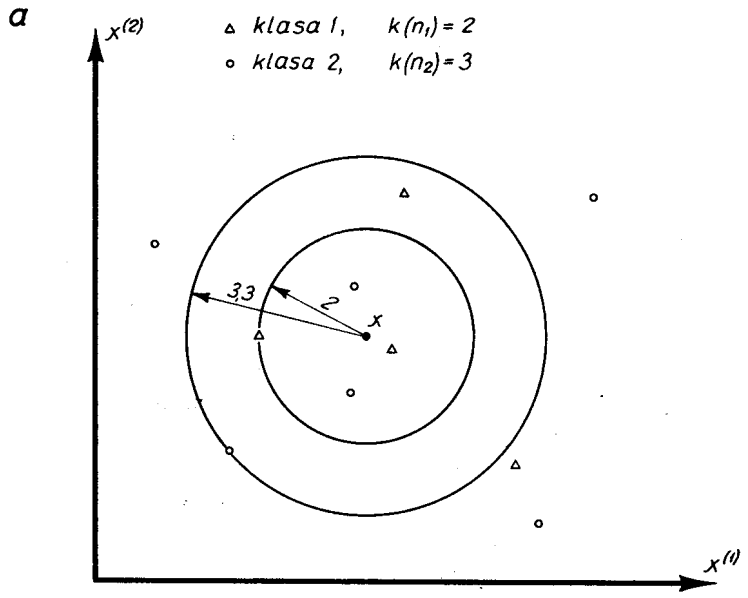
Działanie tego algorytmu dla M = 2 zilustrowane jest na rys. 9. Obraz x, w sytuacji a), zalicza się do klasy 1, ponieważ

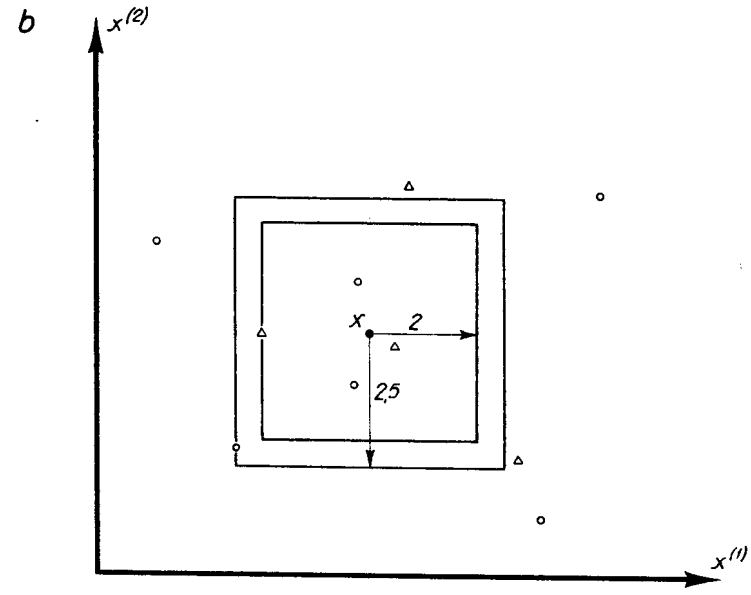
$$\frac{1}{k(n_1)} R_1^2(k(n_1)) = 0{,}56, \quad \frac{1}{k(n_2)} R_2^2(k(n_2)) = 1{,}20 ,$$

w sytuacji b) natomiast obraz ten zalicza się do klasy 2.

Algorytm ten można zmodyfikować, podobnie jak algorytm stosujący estymator Parzena. Zamiast szacować gęstość f_i według wzoru

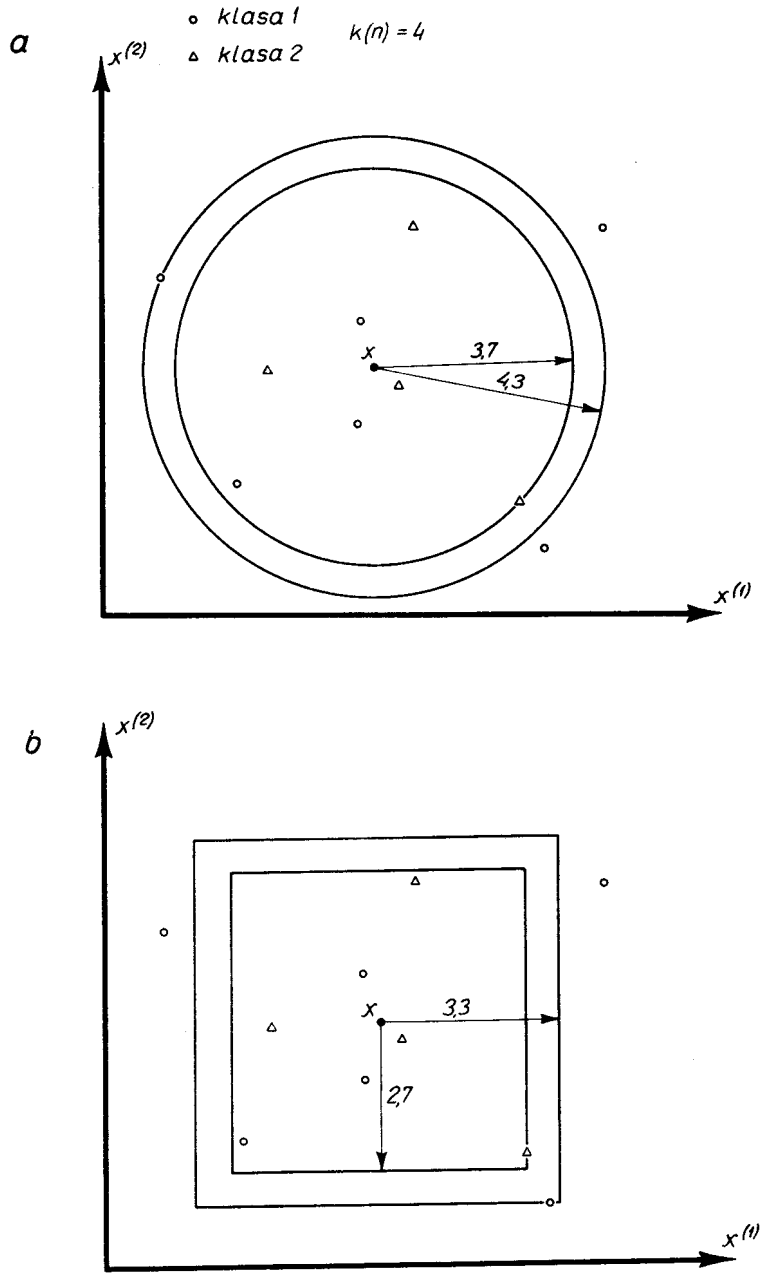
$$f_{in}(x) = \frac{k(n_i)}{n_i} \frac{1}{V R_i^p(k(n_i))} ,$$





Rys. 9. Rozpoznawanie z wykorzystaniem estymatora Loftsgaardena i Quesenberry'ego: a) $\|x\|^2 = x^T x$, b) $\|x\| = \max_i |x^{(i)}|$

Fig. 9. Recognizing with the Loftsgaarden-Quesenberry estimator: a) $\|x\|^2 = x^T x$, b) $\|x\| = \max_i |x^{(i)}|$



Rys. 10. Zmodyfikowany algorytm wykorzystujący estymator Loftsgaardena i Quesenberry ego: a) $\|x\|^2 = x^T x$, b) $\|x\| = \max_i |x^{(i)}|$

Fig. 10. Modified algorithm derived from the Loftsgaarden-Quesenberry estimator: a) $\|x\|^2 = x^T x$, b) $\|x\| = \max_i |x^{(i)}|$

co czyniono dotychczas, można estymować ją następująco:

$$\hat{f}_{in}(x) = \frac{k(n)}{n_i} \frac{1}{V R_i^p(k(n))} , \tag{3.32}$$

co, podobnie jak w przypadku estymatora Parzena, nie zmienia interesującej nas własności, tzn. (mocnej) zgodności w punktach ciągłości. Zmodyfikowany algorytm rozpoznawania zalicza teraz x do klasy, dla której

$$\sum_{j=1}^{M} L(i,j) \frac{1}{R_j^p(k(n))} \tag{3.33}$$

jest najmniejsze.

Jeśli funkcja strat jest typu 0-1, to obraz x klasyfikuje się do klasy, dla której po prostu

$$R_i(k(n)) \tag{3.34}$$

jest najmniejsze. Na rysunku 10 pokazano działanie tego algorytmu w problemie dychotomii. W obydwu sytuacjach obraz x zalicza się do klasy 2. Rozpoznawanie według otrzymanej reguły jest równoznaczne z zaliczaniem obrazu x do klasy, z której najwięcej obrazów znajduje się wśród 2 k(n)-1 obrazów ciągu uczącego najbliższych obrazowi x. Algorytm (3.34) zatem jest równoznaczny znanej regule k_n-NN [28] (tzn. k_n-ty najbliższy sąsiad). Z twierdzeń podanych w rozdziale II i własności estymatora wynikają odpowiednie własności asymptotyczne otrzymanej procedury uczenia. Jest ona zwłaszcza asymptotycznie optymalna w sensie średniego ryzyka względem klasy gęstości prawie wszędzie ciągłych.

3.3.3. Zastosowanie metody rozwinięć ortogonalnych

Zastosowanie estymatora (3.16) do oszacowania gęstości w klasach prowadzi do reguły, która obraz x zalicza do klasy minimalizującej

$$\sum_{j=1}^{M} L(i,j) \sum_{k=1}^{n_j} \sum_{m=1}^{q(n_j)} \varphi_m(x_{jk}) \varphi_m(x) . \tag{3.35}$$

Dla funkcji strat 0-1 reguła ta zalicza obraz x do klasy, która zapewnia maksimum wyrażenia

$$\sum_{k=1}^{n_i} \sum_{m=1}^{q(n_i)} \varphi_m(x_{ik}) \varphi_m(x) . \tag{3.36}$$

W przypadku, gdy $\mathcal{X} = \mathcal{R}$, to można wybrać ortonormalny układ Hermite'a, gdy obrazy są ograniczone można natomiast zastosować układ trygonometryczny. Asymptotycznie optymalne własności algorytmu w sensie średniego ryzyka względem klasy gęstości całkowalnych z kwadratem wynikają z (3.17) i twierdzenia 2.4.

Jeśli $\mathcal{X} = \mathcal{R}$, to dla układu Hermite'a, jak wiadomo

$$\varphi_n(x) = \frac{1}{\sqrt{2^n\, n!\, \sqrt{\pi}}}\; e^{-\frac{x^2}{2}} H_n(x), \quad n = 0, 1, 2, \ldots,$$

przy czym

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} .$$

Jeśli natomiast $\mathcal{X} = [a, b]$, to układem ortonormalnym może być

$$\frac{1}{\sqrt{b-a}}, \quad \sqrt{\frac{2}{b-a}} \cos 2\pi n \frac{x-a}{b-a}, \quad \sqrt{\frac{2}{b-a}} \sin 2\pi n \frac{x-a}{b-a}, \quad n=1, 2, \ldots .$$

Oszacowanie gęstości można także stosować w zmodyfikowanej formie ustalając liczbę wyrazów szeregu w zależności od n.Dla funkcji strat typu 0-1 obraz x zalicza się wówczas do klasy maksymalizującej

$$\sum_{k=1}^{n_i} \sum_{m=1}^{q(n)} \varphi_m(x_{ik})\; \varphi_m(x) .$$

3.4. Przykład

Dla zilustrowania rozważań o asymptotycznej optymalności metod uczenia rozpoznawania stosujących nieparametryczne oszacowania gęstości rozpatrzymy przykład, w którym wykorzystano estymator Parzena.Gęstości w klasach są następujące:

$$f_1 = N(0,1), \qquad f_2 = N(2,1)$$

oraz $p_1=p_2=1/2$; funkcja strat jest typu 0-1.

Jak łatwo sprawdzić $R^* = 0{,}159$. Gęstości w klasach estymowano następująco

$$f_{in}(x) = \frac{1}{n_i\, h(n_i)} \sum_{i=1}^{n_i} K\left(\frac{x - x_{ik}}{h(n_i)}\right), \quad i=1, 2,$$

przy czym jako K przyjęto

$$a)\quad K_1(x) = \begin{cases} 1/2 & \text{jeśli } |x| \leq 1 \\ 0 & \text{jeśli } |x| > 1, \end{cases}$$

b) $K_2(x) = \begin{cases} 2/3 - 4|x|/9 & \text{jeśli } |x| \leq 3/2 \\ 0 & \text{jeśli } |x| > 3/2, \end{cases}$

c) $K_3(x) = \frac{3/2}{(|x| + 1)^4}$.

Dla wszystkich tych funkcji $\int |x| K(x)dx = 1/2$. Ciąg $\{h(n)\}$ wybrano według wzoru

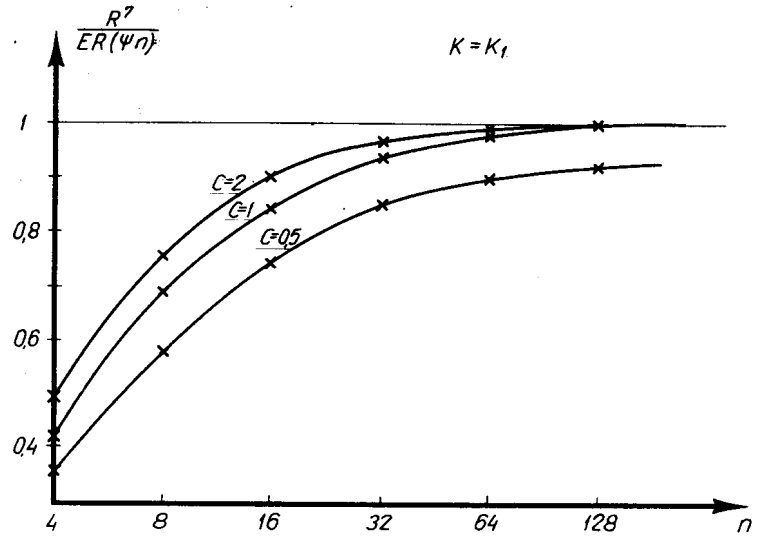
$$h(n) = cn^{-0,1},$$

przy czym jako c przyjęto 0,5, 1, 2.

Tabela 1

Jakość uczenia rozpoznawania
Quality of learning to recognize patterns

n	K_1			K_2			K_3		
	0,5	1	2	0,5	1	2	0,5	1	2
4	0,36	0,42	0,50	0,41	0,50	0,55	0,60	0,60	0,62
8	0,58	0,70	0,76	0,68	0,75	0,75	0,76	0,78	0,86
16	0,75	0,84	0,92	0,84	0,90	0,94	0,82	0,87	0,96
32	0,86	0,95	0,98	0,88	0,92	0,99	0,88	0,92	0,95
64	0,90	0,98	0,99	0,93	0,99	1,00	0,93	0,98	1,00
128	0,93	1,00	1,00	0,94	0,98	1,00	0,98	0,98	1,00



Rys. 11. Jakość procesu uczenia rozpoznawania
Fig. 11. Quality of learning for pattern recognition

Ryzyko średnie $ER(\psi_n)$ szacowano eksperymentalnie dla ciągów uczących o długościach 4, 8, 16, 32, 64, 128. Wyniki zestawiono w tabeli 1, podając w niej $R^*/ER(\psi_n)$. Zachowanie się tego ilorazu przy wzroście długości ciągu uczącego przedstawiono na rys. 11.

Bardzo dużą skuteczność algorytmu można uzyskać już dla ciągów uczących liczących kilkanaście obrazów. Jest ona dla wszystkich trzech funkcji K_1, K_2 i K_3 podobna. Zdecydowanie większy wpływ na wartość średniego ryzyka ma natomiast wybór stałej c w ciągu $\{h(n)\}$ (dla dużych n). Pokrywa się to z wnioskami wynikającymi z analizy [25] i badań eksperymentalnych [77] jakości estymatora gęstości.

3.5. Uwagi

Omówione procedury uczenia rozpoznawania stosują różne metody szacowania gęstości prawdopodobieństwa. Pewne ich elementy podlegające wyborowi można ustalić w zależności albo od liczby obrazów ciągu uczącego należących do poszczególnych klas, albo od długości ciągu uczącego. Ten drugi sposób prowadzi do tzw. zmodyfikowanych algorytmów. Szczególnie zmodyfikowany algorytm, w którym stosuje się estymator Loftsgaardena i Quesenberry'ego jest równoznaczny regule k_n-NN.

Liczba obliczeń niezbędnych do podjęcia decyzji o obrazie rośnie liniowo wraz z długością ciągu uczącego. Liniowo zależy też ona od wymiarowości problemu, tzn. od wymiaru przestrzeni obrazów.

W każdym z algorytmów można wybierać pewne elementy, np. ciąg $\{h(n)\}$, funkcję K, odległość, ciąg $\{k(n)\}$, układ ortonormalny lub też ciąg $\{q(n)\}$. Pewne sugestie co do tych ustaleń wynikają z prac Van Ryzina [68, 70], Wegmana [71], Epanecznikowa [25], Elkinsa [24], a nawet z ilustracyjnego przykładu 3.4. Problem właściwego wyboru wymienionych powyżej ciągów, funkcji itp. trudno jednak uważać za rozwiązany.

Interesujące byłoby, jak się wydaje, skonstruowanie adaptacyjnego estymatora, np. typu Parzena, w którym stałe c oraz α w ciągu $\{h(n)\}$ typu $h(n) = cn^{-\alpha}$, uzależnione byłyby od obrazów zaobserwowanych w ciągu uczącym.

IV. ASYMPTOTYCZNIE OPTYMALNE ALGORYTMY IDENTYFIKACJI

4.1. Wstęp

Korzystanie z nieparametrycznych metod estymacji gęstości prawdopodobieństwa w identyfikacji pozwala, podobnie jak w uczeniu rozpoznawania, uzyskać asymptotycznie optymalne rozwiązania. Badany obiekt ma p-wymiarowe wejście x oraz skalarne wyjście, które zwyczajowo będziemy oznaczać literą y (zamiast użytej w rozdziale I (rys.3) litery ω). $\mathcal{X}$ oraz $\mathcal{Y}$ są odpowiednio przestrzeniami wejść i wyjść, tzn. $\mathcal{X} = \mathcal{R}^p$, $\mathcal{Y} = \mathcal{Y} = \mathcal{R}$. Na skutek przypadkowych czynników działających na interesujący nas identyfikowany obiekt, nawet przy ustalonym wejściu, wyjście zmieniałoby się losowo. Niech f_x będzie gęstością wyjścia w sytuacji, gdy wejściem jest x, f zaś łączną gęstością wejścia i wyjścia, g natomiast - gęstością wejścia. Jako funkcję strat przyjmiemy

$$L(d,y) = (d-y)^2 ,$$

przy czym d jest wyjściem modelu.

Jest oczywiste, że dla wskaźnika jakości identyfikacji

$$\iint (y - \phi(x))^2 f(x,y)\, dx\, dy \qquad (4.1)$$

najlepszy model ma charakterystykę

$$\phi^*(x) = \int y\, f_x(y)\, dy . \qquad (4.2)$$

Identyfikacja polega na utworzeniu modelu na podstawie ciągu uczącego, tzn. ciągu zmiennych losowych na wejściu i wyjściu

$$(X_1, Y_1) , \ldots, (X_n, Y_n) .$$

Model ϕ_n wyznaczony dla ciągu uczącego o długości n przyporządkowuje każdej realizacji ciągu uczącego $(x_1, y_1), \ldots, (x_n, y_n)$ oraz wejściu x wyjście $\phi_n(x_1, y_1, \ldots, x_n, y_n, x)$. Zadanie polega więc na znalezieniu algorytmu identyfikacji, który zapewniłby odpowiednią zbieżność ciągu modeli $\{\phi_n\}$ do modelu optymalnego ϕ^*.

Na ten problem można spojrzeć nieco inaczej, pamiętając o późniejszym sterowaniu. Załóżmy bowiem, że polega ono na podaniu takiego wejścia, które zapewniłoby zadaną lub ekstremalną średnią wartość wyjścia, tzn. krzywej regresji $\int y\, f_x(y)\, dy$. Rozwiązania takich zadań można uzyskać, jak wykażemy później, estymując charakterystykę ϕ^* na podstawie ciągu uczącego.

Algorytmy zapewniające asymptotyczną optymalność procesu identyfikacji, tzn. zbieżność $\{\phi_n\}$ do ϕ^* podał autor w pracach [30,34] i [31]. W dwóch pierwszych korzystano z estymatora gęstości Parzena, w ostatniej natomiast ϕ^* estymowano szacując współczynniki odpowiedniego rozwinięcia w szereg ortonormalny.

Dalej podano twierdzenia pozwalające korzystać z różnych typów nieparametrycznych oszacowań gęstości prawdopodobieństwa, przy czym szczególną uwagę zwrócono na algorytmy, w których stosuje się oszacowania Parzena. Omówiono także procedurę identyfikacji stosującą rozwinięcia ortonormalne. Wykazano asymptotycznie optymalne własności otrzymanych algorytmów, tzn. zbieżność charakterystyk modelu do charakterystyki optymalnej.

Rozpatrzono przy tym dwie sytuacje, gdy rozkład wejścia jest znany lub nie znany. Zbadano także własności sterowań wyznaczanych na podstawie charakterystyki modelu.

4.2. Identyfikacja ze stosowaniem nieparametrycznych oszacowań gęstości

Poniżej podamy ogólne, asymptotycznie optymalne własności algorytmów identyfikacji wykorzystujących nieparametryczne metody estymacji gęstości prawdopodobieństwa. W sytuacji, gdy rozkład wejścia g jest znany, charakterystykę ϕ^* optymalnego modelu estymuje się wzorem

$$\phi_n(x) = \frac{1}{g(x)} \int y\, f_n(x,y)\, dy \,, \tag{4.3}$$

w którym f_n jest estymatorem gęstości f, gdy gęstość g natomiast nie jest znana - za pomocą wzoru

$$\bar{\phi}_n(x) = \frac{1}{g_n(x)} \int y\, f_n(x,y)\, dy \,, \tag{4.4}$$

w którym g_n jest estymatorem gęstości g.

4.2.1. Znany rozkład wejścia

Zauważmy, że z wzorów (4.2) i (4.3) wynika, że

$$|\phi^*(x) - \phi_n(x)| \leqslant \frac{1}{g(x)} \int |y|\, |f(x,y) - f_n(x,y)|\, dy \,.$$

Jeśli wyjście obiektu jest ograniczone, tzn. istnieje ograniczony zbiór $A \in Y$ taki, że

$$\int_A f_x(y)dy = 1 \quad \text{oraz} \quad f_n(x,y) = 0 \quad \text{dla} \quad y \notin A \tag{4.5}$$

dla wszystkich $x \in \mathcal{X}$, to wynika stąd, że

$$|\phi^*(x) - \phi_n(x)| \leq \frac{a}{g(x)} \sup_{x,y} |f(x,y) - f_n(x,y)| \tag{4.6}$$

przy czym $a = \int_A |y|dy$.

Ta nierówność pozwala sformułować twierdzenie:

Twierdzenie 4.1

Jeśli wyjście obiektu jest ograniczone oraz $g(x) > 0$ i

$$\sup_{x,y} |f(x,y) - f_n(x,y)| \xrightarrow[z\ p.\ 1]{p} 0 ,$$

gdy $n \to \infty$, to odpowiednio

$$\phi_n(x) \xrightarrow[z\ p.\ 1]{p} \phi^*(x) ,$$

gdy $n \to \infty$. W obydwu przypadkach ponadto (jeśli ϕ_n jest ograniczone)

$$\lim_{n \to \infty} E(\phi^*(x) - \phi_n(x)^2 = 0.$$

D o w ó d

Pierwsza część twierdzenia wynika bezpośrednio z nierówności (4.6). Ostatnia zbieżność jest także oczywista, ponieważ ϕ^* jest ograniczone ze względu na ograniczoność wyjścia, co kończy dowód. ■

Następne twierdzenie dotyczy innego oszacowania gęstości.

Twierdzenie 4.2

Jeśli wyjście obiektu jest ograniczone oraz $g(x) > 0$ i

$$\lim_{n \to \infty} \sup_{x,y} E(f(x,y) - f_n(x,y))^2 = 0 ,$$

to

$$\lim_{n \to \infty} E(\phi^*(x) - \phi_n(x))^2 = 0.$$

D o w ó d

Teza wynika z następującej nierówności:

$$E(\phi^*(x) - \phi_n(x))^2 \leq \frac{\mu(A)}{g^2(x)} \int_A y^2 dy \sup_{x,y} E(f(x,y) - f_n(x,y))^2 ,$$

co kończy dowód. ■

Twierdzenia 4.1 i 4.2 podają warunki punktowej zbieżności charakterystyki modelu, następne dotyczy natomiast zbieżności w sensie całkowym. Z (4.2) i (4.3) wynika, że

$$\int (\phi^*(x) - \phi_n(x))^2 g^2(x)\, dx \leq \int_A y^2\, dy \iint (f(x,y) - f_n(x,y))^2 dxdy.$$

Prawdziwe jest więc następujące twierdzenie:

Twierdzenie 4.3

Jeśli wyjście obiektu jest ograniczone oraz

$$\iint (f(x,y) - f_n(x,y))^2\, dx\, dy \xrightarrow[z\ p.\ 1]{p} 0 ,$$

gdy $n \to \infty$, lub

$$\lim_{n\to\infty} E \iint (f(x,y) - f_n(x,y))^2\, dx\, dy = 0 ,$$

to odpowiednio

$$\int (\phi^*(x) - \phi_n(x))^2 g^2(x)\, dx \xrightarrow[z\ p.\ 1]{p} 0 ,$$

gdy $n \to \infty$, lub

$$\lim_{n\to\infty} E \int (\phi^*(x) - \phi_n(x))^2 g^2(x)\, dx = 0 .$$

4.2.2. Rozkład wejścia nie znany

W sytuacji, gdy rozkład wejścia nie jest znany, algorytm identyfikacji ma postać (4.4). Z twierdzenia 4.1 wynika więc następne.

Twierdzenie 4.4

Jeżeli wyjście obiektu jest ograniczone i $g(x) > 0$ oraz odpowiednio

$$\sup_{x,y} |f(x,y) - f_n(x,y)| \xrightarrow[z\ p.\ 1]{p} 0 ,$$

gdy $n \to \infty$ i

$$g_n(x) \xrightarrow[z\ p.\ 1]{p} g(x) ,$$

gdy $n \to \infty$, to

$$\bar{\phi}_n(x) \xrightarrow[z\ p.\ 1]{p} \phi^*(x) ,$$

gdy $n \to \infty$.

4.3. Algorytmy identyfikacji stosujące estymator Parzena

Przedstawione twierdzenia pozwalają korzystać w procesie identyfikacji z różnych, nieparametrycznych oszacowań gęstości prawdopodobieństw omówionych w pkcie 4.2. Najwygodniejszy do stosowania jest estymator Parzena

$$f_n(x,y) = \frac{1}{n\, h^{p+1}(n)} \sum_{i=1}^{n} K_x\left(\frac{x - x_i}{h(n)}\right) K_y\left(\frac{y - y_i}{h(n)}\right),$$

w którym K_x i K_y są, jak wiadomo, odpowiednio wybranymi funkcjami.

Wówczas

$$\phi_n(x) = \frac{1}{n\, h^{p+1}(n)\, g(x)} \sum_{i=1}^{n} K_x\left(\frac{x - x_i}{h(n)}\right) \int y\, K_y\left(\frac{y - y_i}{h(n)}\right) dy .$$

W szczególności, jeśli $K_y(y) = K_y(-y)$, to ($K_x \overset{\text{def}}{=} K$),

$$\phi_n(x) = \frac{1}{n\, h^{p}(n)\, g(x)} \sum_{i=1}^{n} y_i\, K\left(\frac{x - x_i}{h(n)}\right) . \tag{4.7}$$

Jeśli gęstość g jest nie znana, to można ją także estymować stosując estymator Parzena, np. według wzoru

$$g_n(x) = \frac{1}{n\, h^{p}(n)} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h(n)}\right).$$

Charakterystyka modelu wyraża się wówczas wzorem

$$\bar{\phi}_n(x) = \frac{\sum_{i=1}^{n} y_i\, K\left(\frac{x - x_i}{h(n)}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h(n)}\right)} \tag{4.8}$$

Aby zapewnić zgodność procedur określonych wzorami (4.7) i (4.8) należy funkcję K i ciąg $\{h(n)\}$ wybrać w ten sposób, aby spełnione zostały odpowiednie założenia twierdzenia 4.1 lub 4.2. Jeśli na przykład f jest gęstością jednostajnie ciągłą i K oraz $\{h(n)\}$ wybierze się zgodnie z (3.2) i (3.3), to z (3.8) i twierdzenia 4.2 wynika, że

$$\lim_{n\to\infty} E\,(\phi^*(x) - \phi_n(x))^2 = 0$$

oraz

$$\phi_n(x) \xrightarrow{\ p\ } \phi^*(x) ,$$

gdy $n \to \infty$.

Wybierając K według (3.5) (z wyjątkiem przykładu a) można otrzymać słabą lub mocną jednostajną zgodność estymatora gęstości (patrz (3.10) i (3.11), a zatem i zgodność ciągu modeli, tzn. zbieżności

$$\psi_n(x) \xrightarrow[z\ p.\ 1]{p} \psi^*(x) ,$$

gdy $n \to \infty$ oraz

$$\bar{\psi}_n(x) \xrightarrow[z\ p.\ 1]{p} \psi^*(x) ,$$

gdy $n \to \infty$.

Jako przykłady algorytmów identyfikacji można podać

a)
$$\psi_n(x) = \frac{1}{(2\pi)^{p/2}\, n\, h^p(n)\, g(x)} \sum_{i=1}^{n} y_i\, e^{-\frac{1}{2}\|x-x_i\|^2}$$

$$\bar{\psi}_n(x) = \frac{\sum_{i=1}^{n} y_i\, e^{-\frac{1}{2}\|x-x_i\|^2}}{\sum_{i=1}^{n} e^{-\frac{1}{2}\|x-x_i\|^2}} ,$$

przy czym $\|x\|^2 = x^T x$,

b)
$$\psi_n(x) = \frac{1}{2^p\, n\, h^p(n)\, g(x)} \sum_{i=1}^{n} y_i\, e^{-\|x-x_i\|},$$

$$\bar{\psi}_n(x) = \frac{\sum_{i=1}^{n} y_i\, e^{-\|x-x_i\|}}{\sum_{i=1}^{n} e^{-\|x-x_i\|}} ,$$

przy czym $\|x\| = \sum_{i=1}^{p} |x^{(i)}|$.

Zbieżność algorytmów identyfikacji zapewniona jest jednak przy założeniu ograniczoności wyjścia obiektu. Okazuje się, że przy korzystaniu z oszacowania Parzena można je pominąć. Odpowiednią własność algorytmu identyfikacji podamy w formie twierdzenia. Oznaczmy jeszcze przez X i Y zmienne losowe na wejściu i wyjściu obiektu o łącznej gęstości f.

Twierdzenie 4.5

Jeżeli

$$EY^2<\infty \tag{4.9}$$

oraz

$$\sup_x |K(x)|<\infty,\ \int K(x)dx = 1,\ \int |K(x)dx<\infty,\ \lim_{\|x\|\to\infty} \|x\|^p |K(x)| = 0 \tag{4.10}$$

oraz

$$h(n) > 0,\quad \lim_{n\to\infty} h(n) = 0,\quad \lim_{n\to\infty} n\,h^{2p}(n) = \infty\ , \tag{4.11}$$

to

$$\lim_{n\to\infty} E\,(\phi^*(x) - \phi_n(x))^2 = 0 \tag{4.12}$$

i

$$\Phi_n(x) \xrightarrow{p} \phi^*(x)\ , \tag{4.13}$$

gdy $n\to\infty$, w punktach,w których iloczyn $g\phi^*$ jest funkcją ciągłą oraz $g(x) > 0$.

D o w ó d

Jest oczywiste, że wystarczy dowieść, że

$$\lim_{n\to\infty} E\left[\frac{1}{n\,h^p(n)} \sum_{i=1}^{n} Y_i\,K\left(\frac{x - X_i}{h(n)}\right) - \int y\,f(x,y)\,dy\right]^2 = 0\ .$$

Zauważmy w tym celu, że

$$E\left[\frac{1}{nh^p(n)} \sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h(n)}\right) - \int y\,f(x,y)dy\right]^2 = \operatorname{var}\left[\frac{1}{nh^p(n)} \sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h(n)}\right)\right] +$$

$$+\left[\frac{1}{nh^p(n)} E\left\{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h(n)}\right)\right\} - \int y\,f(x,y)dy\right]^2 \stackrel{def}{=} V_n + I_n^2\ .$$

Z nierówności

$$V_n = \frac{1}{nh^{2p}(n)} \operatorname{var}\left[YK\left(\frac{x - X}{h(n)}\right)\right] < \frac{1}{nh^{2p}(n)}\ EY^2 \sup_x K^2(x)$$

i założenia (4.11) wynika, że pierwszy składnik powyższej sumy dąży do zera, gdy $n\to\infty$. Dla wykazania, że również drugi składnik maleje do zera zauważmy, że

$$I_n = \frac{1}{h^p(n)} \int\int_{w\ y} y\,(f(x-w,y) - f(x,y))\ K\left(\frac{w}{h(n)}\right) dwdy\ ,$$

przy czym $w \in w = \mathcal{R}$.

Wybierzmy $\delta > 0$ i podzielmy przestrzeń w na dwa zbiory,w których odpowiednio $\|w\| \le \delta$ i $\|w\| > \delta$. Zatem

$$|I_n| \leq \sup_{\|w\|\leq \delta} \left| \int y\, f(x-w,y)dy - \int y\, f(x,y)dy \right| \int |K(x)|dx +$$
$$+ \int_{\|w\|>\delta} \int_{Y} \frac{1}{\|w\|^p} |y|\, f(x-w,y)\, \frac{\|w\|^p}{h^p(n)} \left| K\left(\frac{w}{h(n)}\right) \right| dwdy +$$
$$+ \left| \int y\, f(x,y)dy \right| \frac{1}{h^p(n)} \int_{\|x\|> \delta} \left| K\left(\frac{x}{h(n)}\right) \right| dx ,$$

stąd

$$|I_n| \leq \sup_{\|w\|\leq \delta} |\phi^*(x-w)\, g(x-w) - |\phi^*(x)\, g(x)| \int |K(x)|dx +$$
$$+ \frac{1}{\delta^p} \sup_{\|x\|> \delta/h(n)} \|x\|^p |K(x)| \sqrt{EY^2} + |g(x)\, \phi^*(x)| \int_{\|x\|>^{\circ}/h(n)} |K(x)|\, dx \qquad (4.14)$$

Z założeń wynika, że w punktach ciągłości $g\,\phi^*$, dla każdego $\varepsilon > 0$ istnieją $\delta > 0$ i $H > 0$ także, że dla $h(n) < H$, każdy ze składników sumy jest mniejszy od $\varepsilon/3$, co kończy dowód. ■

W identyfikacji można stosować także rekurencyjną wersję oszacowania Parzena otrzymując następujące algorytmy:

$$\phi_n'(x) = \frac{1}{n\, g(x)} \sum_{i=1}^{n} \frac{1}{h^p(i)}\, y_i\, K\left(\frac{x - x_i}{h(i)}\right), \qquad (4.15)$$

gdy gęstość wejścia jest znana oraz

$$\bar{\phi}_n'(x) = \frac{\sum_{i=1}^{n} \frac{1}{h^p(i)}\, y_i\, K\left(\frac{x - x_i}{h(i)}\right)}{\sum_{i=1}^{n} \frac{1}{h^p(i)}\, K\left(\frac{x - x_i}{h(i)}\right)}, \qquad (4.16)$$

gdy gęstość ta jest nie znana.

Wykażemy teraz zbieżność tych procedur identyfikacji.

Twierdzenie 4.6

Jeśli

$$EY^2 < \infty$$

oraz

$$\sup_x |K(x)| < \infty, \quad \int K(x)dx = 1, \quad \int |K(x)|dx < \infty, \quad \lim_{\|x\|\to\infty} \|x\|^p |K(x)| = 0 \qquad (4.17)$$

$$h(n) > 0, \quad \lim_{n\to\infty} h(n) = 0, \quad \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{h^{2p}(i)} = 0, \qquad (4.18)$$

to

$$\lim_{n\to\infty} E\,(\phi^*(x) - \phi_n'(x))^2 = 0 \tag{4.19}$$

oraz

$$\bar{\phi}_n'(x) \xrightarrow{p} \phi^*(x)\,, \tag{4.20}$$

gdy $n\to\infty$, w punktach, w których zarówno g jak i ϕ^* są funkcjami ciągłymi i $g(x) > 0$.

D o w ó d

Dla wykazania tez wystarczy dowieść, że

$$\lim_{n\to\infty} E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p(i)}\,Y_i\,K\left(\frac{x-X_i}{h(i)}\right) - \int y\,f(x,y)dy\right]^2 = 0.$$

Zauważmy, że

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p(i)}\,Y_i\,K\left(\frac{x-X_i}{h(i)}\right) - \int y\,f(x,y)dy\right]^2 =$$

$$= \operatorname{var}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p(i)}\,Y_i\,K\left(\frac{x-X_i}{h(i)}\right)\right] + I_n^2,$$

przy czym

$$I_n \overset{\text{def}}{=} E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p(i)}\,Y_i\,K\left(\frac{x-X_i}{h(i)}\right)\right] - \int y\,f(x,y)dy.$$

Ponieważ

$$\operatorname{var}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p(i)}\,Y_i\,K\left(\frac{x-X_i}{h(i)}\right)\right] \le \frac{1}{n^2}\,E\,Y^2 \sup_x |K(x)|^2 \sum_{i=1}^{n}\frac{1}{h^{2p}(i)}\,,$$

to z założeń wynika, że pierwszy ze składników sumy maleje do zera, gdy $n\to\infty$.

Dla drugiego składnika mamy (podobnie jak w twierdzeniu 4.5)

$$|I_n| \le \sup_{\|w\|\le\delta} |\phi^*(x-w)\,g(x-w) - \phi^*(x)\,g(x)| \int |K(x)|dx +$$

$$+ \frac{1}{\delta^p}\sqrt{EY^2}\,\frac{1}{n}\sum_{i=1}^{n}\sup_{\|x\|>\delta/h(i)}\|x\|^p|K(x)| + |\phi^*(x)g(x)|\,\frac{1}{n}\sum_{i=1}^{n}\int_{\|x\|>\delta/h(i)}|K(x)|\,dx, \tag{4.21}$$

gdzie δ jest dowolną liczbą dodatnią, skąd wynika teza (patrz np. (3.13) i dalej). ■

Zauważmy, że jeśli ciąg $\{h(n)\}$ jest monotoniczny, to ostatnie z założeń (4.18) można zastąpić poniższym

$$\lim_{n\to\infty} n\,h^{2p}(n) = \infty\,.$$

Algorytmy (4.15) i (4.16) można zapisać w wygodniejszej niekiedy formie rekurencyjnej:

$$\phi_0'(x) = 0, \quad \phi_n'(x) = \frac{n-1}{n}\,\phi_{n-1}'(x) + \frac{1}{nh^p(n)g(x)}\,y_n\,K\left(\frac{x-x_n}{h(n)}\right), \quad n=1,2,\ldots$$

oraz

$$L_n(x) = \frac{n-1}{n}\,L_{n-1}(x) + \frac{1}{nh^p(n)}\,y_n\,K\left(\frac{x-x_n}{h(n)}\right),$$

$$M_n(x) = \frac{n-1}{n}\,M_{n-1}(x) + \frac{1}{nh^p(n)}\,K\left(\frac{x-x_n}{h(n)}\right),$$

$$\Phi_0'(x) = 0, \quad \Phi_n'(x) = L_n(x) \,/\, M_n(x), \quad n = 1, 2, \ldots .$$

4.4. Rozwinięcie w szereg ortogonalny

W identyfikacji można stosować także oszacowania gęstości polegające na rozwijaniu jej w szereg ortogonalny. Zamiast gęstości f wygodniej jednak rozwijać w szereg funkcję R określoną wzorem

$$R(x) \overset{\text{def}}{=} \int y\,f(x,y)dy = g(x)\,\phi^*(x).$$

Niech $\{\varphi_i\}_{i=0}^{\infty}$ będzie pełnym układem funkcji ortonormalnych w przestrzeni L^2 funkcji całkowalnych w kwadracie określonych na przestrzeni $\mathcal{X}$. Załóżmy, że funkcje tego układu są wspólnie ograniczone, tzn.

$$\sup_{n,x} |\varphi_n(x)| \leqslant c < \infty . \tag{4.22}$$

Jeśli $EY^2 < \infty$, oraz $\sup_x g(x) < \infty$, to jak łatwo sprawdzić także $R \in L^2$. Funkcję R można więc rozłożyć w szereg

$$R(x) \sim \sum_{i=0}^{\infty} a_i\,\varphi_i(x),$$

w którym

$$a_i = \int R(x)\,\varphi_i(x)dx = \int y\,\varphi_i(x)\,f(x,y)\,dxdy .$$

Współczynniki rozwinięcia można estymować w naturalny sposób:

$$a_{in} = \frac{1}{n}\sum_{j=1}^{n} y_j\,\varphi_i(x_j) .$$

Estymator $R_n(x)$ nie znanego $R(x)$ określimy następująco:

$$R_n(x) = \sum_{i=0}^{N} a_{in}\,\varphi_i(x),$$

przy czym N jest pewną liczbą.

Jest oczywiste, że

$$\mathrm{E}\, a_{in} = a_i \,, \tag{4.23}$$

stąd i z założenia (4.22) wynika, że

$$\mathrm{E}(a_i - a_{in})^2 \leqslant \frac{1}{n}\, \mathrm{E}\left\{Y^2\, \varphi_i^2\,(X)\right\} \leqslant \frac{1}{n}\, c^2\, \mathrm{E}Y^2 \overset{\mathrm{def}}{=} c_1 \frac{1}{n}\,. \tag{4.24}$$

Korzystając z równości Parsevala otrzymujemy z kolei

$$\mathrm{E}\int (R(x) - R_n(x))^2 dx = \sum_{i=0}^{N} \mathrm{E}(a_i - a_{in})^2 + \sum_{i=N+1}^{\infty} a_i^2 \leqslant c_1 \frac{N}{n} + \sum_{i=N+1}^{\infty} a_i^2 .$$

Jeśli teraz N uzależni się od liczby obserwacji n w ten sposób, aby $\lim\limits_{n\to\infty} N(n) = \infty$, $\lim\limits_{n\to\infty} \frac{N(n)}{n} = 0$,

to z powyższej nierówności wynika, że wówczas

$$\lim_{n\to\infty} \mathrm{E}\int (R(x) - R_n(x))^2 dx = 0 \,,$$

tzn.

$$\lim_{n\to\infty} \mathrm{E}\int (\phi^*(x) - \phi_n(x))^2\, g^2(x)\, dx = 0 \,,$$

przy czym $\phi_n(x) = R_n(x)/g(x)$.

Wykazaliśmy więc następujące twierdzenie o asymptotycznej optymalności:

Twierdzenie 4.7

Jeśli $\{\varphi_i\}_{i=0}^{\infty}$ jest pełnym układem ortonormalnym w przestrzeni L^2 takim, że

$$\sup_{i,x} |\varphi_i(x)| < \infty$$

i

$$\mathrm{E}Y^2 < \infty \,, \quad \sup_x g(x) < \infty$$

oraz

$$\lim_{n\to\infty} N(n) = \infty \,, \quad \lim_{n\to\infty} \frac{N(n)}{n} = 0,$$

to

$$\lim_{n\to\infty} \mathrm{E}\int (\phi^*(x) - \phi_n(x))^2\, g^2(x) dx = 0 \,, \tag{4.25}$$

przy czym

$$\phi_n(x) = \frac{1}{g(x)} \sum_{i=0}^{N(n)} a_{in}\, \varphi_i(x) \,.$$

Wybierając odpowiednio układ ortonormalny i ciąg $\{N(n)\}$ można otrzymać punktowo zgodny estymator optymalnej charakterystyki ϕ^*. Oznaczmy w tym celu

$$\overline{R}^N(x) = \sum_{i=0}^{N} a_i \varphi_i(x).$$

Dla $N = N(n)$ zatem

$$E\,(R(x) - R_n(x))^2 = (R(x) - \overline{R}^N(x))^2 + E\,(\overline{R}^N(x) - R_n(x))^2 ,$$

ponieważ, jak wynika z (4.23) $ER_n(x) = \overline{R}^N(x)$. Na podstawie (4.24) i nierówności Schwartza mamy

$$E\,(\overline{R}^N(x) - R_n(x))^2 = E\Big(\sum_{i=0}^{N} (a_i - a_{in})\ \varphi_i(x)\Big)^2 \leqslant$$

$$\leqslant 2N \sum_{i=0}^{N} E\left\{(a_i - a_{in})^2\right\} \varphi_i^2(x) \leqslant c^2 c_1^2 \frac{2N^2}{n} \overset{\text{def}}{=} c_2 \frac{N^2}{n} .$$

Ostatecznie

$$E\,(R(x) - R_n(x))^2 \leqslant \Big(\sum_{i=0}^{N} a_i \varphi_i(x) - R(x)\Big)^2 + c_2 \frac{N^2}{n} .$$

Jeśli teraz układ ortonormalny i ciąg $\{N(n)\}$ wybiera się tak, aby spełnione były warunki

$$\lim_{n\to\infty} N(n) = \infty , \quad \lim_{n\to\infty} \frac{N^2(n)}{n} = 0 \tag{4.26}$$

oraz

$$\lim_{n\to\infty} \sum_{i=0}^{N} a_i \varphi_i(x) = R(x) , \tag{4.27}$$

to

$$\lim_{n\to\infty} E\,(R(x) - R_n(x))^2 = 0 ,$$

a zatem

$$\lim_{n\to\infty} E\,(\phi^*(x) - \phi_n(x))^2 = 0 ,$$

gdy $g(x) > 0$.

Udowodniliśmy więc następujące twierdzenie:

Twierdzenie 4.8

Jeśli spełnione są założenia twierdzenia 4.7 oraz (4.26) i (4.27), oraz $g(x) > 0$, to

$$\lim_{n\to\infty} E\left(\phi^*(x) - \phi_n(x)\right)^2 = 0 , \tag{4.28}$$

gdzie

$$\phi_n(x) = \frac{1}{g(x)} \sum_{i=0}^{N(n)} a_{in}\, \varphi_i(x).$$

Jeśli $\mathcal{X} = \mathcal{R}$, to można stosować układ ortonormalny Hermite'a (patrz pkt. 3.3.3) uzyskując zbieżność (4.25). Jeśli skalarne wejście jest ograniczone, np. $\mathcal{X} = [a,b]$, to można wybrać układ trygonometryczny.

Punktową zgodność estymatora optymalnego modelu można dla układu Hermite'a uzyskać jeśli R jest funkcją ciągłą o ograniczonym wahaniu. Wówczas zbieżność 4.27, a zatem i 4.28 zachodzi dla wszystkich $x \in \mathcal{X}$ [63]. Jeśli natomiast wejście jest ograniczone do odcinka [a,b] i ciąg $\{N(n)\}$ ma tę własność, że

$$\frac{N(n+1)}{N(n)} \geqslant \varkappa > 1 , \tag{4.29}$$

to dla układu trygonometrycznego zbieżność (4.27) zachodzi dla prawie wszystkich $x \in \mathcal{X}$ [62]. Jest oczywiste, że własność ta jest zachowana także wtedy, gdy dla pewnych n warunek (4.29) nie jest spełniony, a ciąg $\{N(n)\}$ ma powtórzenie. Liczbę powtórzeń należy tak dobierać, aby ciąg z powtórzeniami spełniał warunki (4.26). Przykładem ciągu spełniającego (4.29) może być 1, 2, 3, 5, 8, ... ($\varkappa = 1{,}5$). Ciągiem z powtórzeniami może być np. 1,1, 2,2,2, 3, ...Ich liczbę należy wybrać w ten sposób, aby spełniał on warunki (4.26).

W sytuacji, gdy gęstość g sygnału wejściowego nie jest znana, $\phi^*(x)$ można estymować następująco:

$$\bar{\phi}_n(x) = \frac{R_n(x)}{g_n(x)} ,$$

przy czym g_n jest estymatorem gęstości g.

Jest oczywiste, że jeśli $R_n(x)$ i $g_n(x)$ są estymatorami zgodnymi i $g(x) > 0$, to $\bar{\phi}_n(x)$ jest także zgodnym oszacowaniem. Gęstość prawdopodobieństwa g można estymować także metodą szeregów ortogonalnych tzn.

$$g_n(x) = \sum_{i=0}^{N(n)} b_{in}\, \varphi_i(x) ,$$

przy czym

$$b_{in} = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(x_j) ,$$

wówczas

$$\bar{\phi}_n(x) = \frac{\sum_{i=0}^{N(n)} a_{in} \varphi_i(x)}{\sum_{i=0}^{N(n)} b_{in} \varphi_i(x)} \qquad (4.30)$$

Przy nieograniczonym, skalarnym wejściu można stosować układ Hermite'a, przy ograniczonym natomiast układ trygonometryczny. Jeśli szeregiem ortogonalnym jest układ Hermite'a oraz ciąg $\{N(n)\}$ spełnia warunki (4.26), to (patrz 3.18))

$$\bar{\phi}_n(x) \xrightarrow{p} \phi^*(x), \quad \text{gdy} \quad n \to \infty$$

we wszystkich punktach $x \in \mathcal{X}$.

Podobną zbieżność dla prawie wszystkich $x \in \mathcal{X}$ można uzyskać także dla układu trygonometrycznego (patrz (3.20)).

4.5. Metoda funkcji potencjalnych

Zbieżność charakterystyki modelu do optymalnego modelu zapewnia także metoda funkcji potencjalnych. Własności obiektu przedstawimy teraz równaniem

$$y = \phi^*(x) + z .$$

Wejście x ma rozkład g, natomiast zakłócenie, w sytuacji, gdy na wejście podano x, ma gęstość warunkową $f_x(z + \phi^*(x))$, natomiast

$$\phi^*(x) = \int y \, f_x(y) \, dy .$$

Załóżmy ponadto, że dyspersja zakłóceń jest dla wszystkich wejść wspólnie ograniczona.

Niech teraz $L^2(g)$ będzie przestrzenią funkcji określonych na $\mathcal{X}$ takich, że dla $t \in L^2(g)$

$$\int t^2(x) \, g(x) \, dx < \infty$$

oraz $\{\varphi_i\}_{i=0}^{\infty}$ dowolnym układem funkcji w tej przestrzeni. Funkcję potencjalną, jak wiadomo, określa się następująco:

$$K(x,w) \sim \sum_{i=0}^{\infty} \lambda_i^2 \varphi_i(x) \varphi_i(w), \quad x,w \in \mathcal{X} ,$$

przy czym

$$\sum_{i=0}^{\infty} \lambda_i^2 < \infty .$$

Podstawowa hipoteza metody dotyczy przedstawienia ϕ^* i jest następująca:

$$\phi^*(x) \sim \sum_{j=0}^{\infty} c_j \varphi_j(x) , \tag{4.31}$$

przy czym

$$\sum_{j=0}^{\infty} \left(\frac{c_j}{\lambda_j}\right)^2 < \infty \tag{4.32}$$

Algorytm identyfikacji ma postać

$$\psi_n(x) = \psi_{n-1}(x) + \gamma_n (y_n - \psi_{n-1}(x_n)) K(x,x_n) . \tag{4.33}$$

Jeśli spełnione zostaną warunki

$$\gamma_n > 0, \quad \sum_{n=1}^{\infty} \gamma_n = \infty , \quad \sum_{n=1}^{\infty} \gamma_n^2 < \infty$$

oraz

$$\psi_0(x) \sim \sum_{j=0}^{\infty} c_j^0 \varphi_j(x) ,$$

przy czym

$$\sum_{j=0}^{\infty} \left(\frac{c_j^0}{\lambda_j}\right)^2 < \infty ,$$

to

$$\int \left[\phi^*(x) - \psi_n(x) \right]^2 g(x)\, dx \xrightarrow{p} 0 ,$$

gdy $n \to \infty$ [2, 8].

Podstawowym warunkiem zbieżności jest prawdziwość hipotezy. Podobnie jak w procesie uczenia rozpoznawania o zbieżności decyduje wybór funkcji potencjalnej, który powinien być taki, aby spełniona została hipoteza (4.32). Niestety, nie można podać żadnych konstruktywnych zaleceń co do wyboru funkcji potencjalnej. Jeśli $\{\varphi_i\}_{i=0}^{\infty}$ jest pełnym układem ortonormalnym w przestrzeni $L^2(g)$, to

$$c_j = \int \phi^*(x) \varphi_j(x) g(x)\, dx .$$

Warunek (4.32) zatem narzuca ograniczenie na współczynniki rozwinięcia optymalnej charakterystyki modelu ϕ^*. Powinna ona znajdować

się w odpowiedniej podprzestrzeni przestrzeni $L^2(g)$. O tym jaka to jest podprzestrzeń decydują współczynniki rozwinięcia funkcji potencjalnej względem układu ortonormalnego.

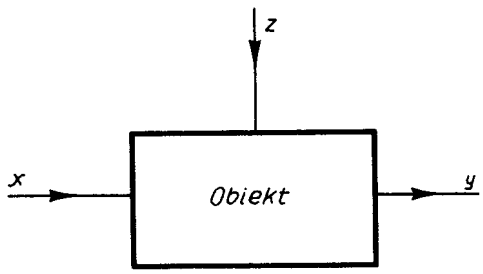
4.6. Przykład

Dla zilustrowania przedstawionej metody identyfikacji wykorzystującej estymator gęstości Parzena podamy przykład obliczeniowy, w którym przyjęto następujący obiekt (rys. 12):

$$y = \phi^*(x) + z ,$$

przy czym rozkład wejścia jest normalny N(0,1), rozkład zakłócenia jest niezależny od wejścia i także normalny N(0,1), natomiast

$$\phi^*(x) = \begin{cases} x & \text{jeśli} \quad x \geqslant 0 , \\ \frac{1}{4}x & \text{jeśli} \quad x < 0 . \end{cases}$$



Rys. 12. Obiekt identyfikacji
Fig. 12.The identified object

Jak łatwo sprawdzić dla kwadratowego wskaźnika jakości identyfikacji

$$\min_{\phi(x)} \iint (y - \phi(x))^2 f(x,y)\, dxdy = \iint (y - \phi^*(x))^2 f(x,y)\, dxdy = 1.$$

Algorytm identyfikacji jest następujący

$$\phi_n(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h(n)}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h(n)}\right)}$$

przy czym funkcję K przyjęto jak w przykładzie 3.4. Ciąg liczbowy wybrano następująco

$$h(n) = cn^{-0,1}$$

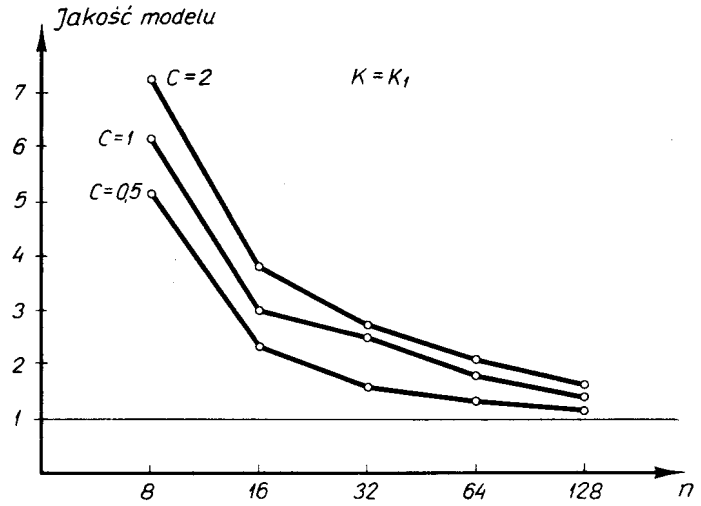
a jako c przyjęto 0,5, 1, 2. Średnie ryzyko oszacowano dla 8, 16, 32, 64 i 128 pomiarów wykonanych na obiekcie. Wyniki zestawiono w tabeli 2, w której podano średnie ryzyko dla różnych K, c oraz n i przedstawiono dla ilustracji na rys. 13.

Tabela 2

Jakość procesu identyfikacji
Quality of the identification process

n	K_1			K_2			K_3		
	0,5	1	2	0,5	1	2	0,5	1	2
8	5,01	6,03	7,22	5,32	6,31	7,51	4,52	6,01	7,00
16	2,03	3,00	3,78	2,37	3,20	4,05	2,02	3,00	4,04
32	1,52	2,98	2,65	1,70	2,99	2,80	1,40	2,81	2,49
64	1,27	1,69	2,02	1,40	1,73	2,07	1,21	1,60	2,07
128	1,16	1,23	1,57	1,20	1,23	1,60	1,10	1,18	1,50

Dobrą jakość modelu udaje się uzyskać już dla kilkunastu pomiarów. Podobnie jak w uczeniu rozpoznawania, niewiele zależy ona od jądra K, zdecydowany wpływ ma na nią natomiast stała c. Ponownie wyłania się więc zasygnalizowany już problem adaptacyjnego uzależnienia stałej c od wyników eksperymentu.



Rys. 13. Jakość procesu identyfikacji
Fig. 13. Quality of the identification process

4.7. Stosowanie modelu do sterowania

4.7.1. Przedstawienie problemu

Jak już wspomniano, cel sterowania obiektem może polegać na podaniu na wejście obiektu sygnału, dla którego $\int yf_x(y)\,dy$ osiąga pewną zadaną wartość lub też ekstremum. Dla wyznaczenia takich sterowań można posłużyć się charakterystyką modelu otrzymaną w procesie identyfikacji. Powstaje teraz następujący problem: czy przy coraz dłuższym prowadzeniu identyfikacji, tzn. przy wzroście n, sterowania wyznaczane na podstawie modelu stają się coraz bliższe nie znanym sterowaniom zapewniającym odpowiednie własności wyjścia obiektu? Wykażemy teraz, że przy właściwie prowadzonym procesie identyfikacji model nie tylko coraz lepiej opisuje własności obiektu (tzn. ϕ_n jest zbieżne do ϕ^*), lecz także sterowania wyznaczone na jego podstawie stają się coraz dokładniejsze. Można więc uważać, że taki proces identyfikacji posiada asymptotycznie optymalne własności także z punktu widzenia sterowania.

Oznaczmy przez ξ sygnał sterujący, dla którego charakterystyka ϕ^* osiąga wartość ekstremalną np. maksimum tzn.

$$\phi^*(\xi) = \max_x \phi^*(x) ,$$

przez η natomiast oznaczmy rozwiązanie równania

$$\phi^*(x) = \alpha ,$$

w którym α jest zadaną wartością. Niech teraz ξ_n i η_n będą estymatorami nie znanych ξ i η wyznaczonymi na podstawie charakterystyki modelu. Dla ξ_n charakterystyka ϕ_n osiąga maksimum, tzn.

$$\phi_n(\xi_n) = \max_x \phi_n(x) ,$$

η_n jest natomiast rozwiązaniem równania

$$\phi_n(x) = \alpha .$$

Wykażemy przy pewnych założeniach, że jeśli

$$\sup_x |\phi^*(x) - \phi_n(x)| \xrightarrow{p} 0 ,$$

gdy $n \to \infty$, to

$$\|\xi - \xi_n\| \xrightarrow{p} 0 \quad \text{i} \quad \|\eta - \eta_n\| \xrightarrow{p} 0,$$

gdy $n \to \infty$.

Można więc uważać, że proces identyfikacji jest wtedy asymptotycznie optymalny z punktu widzenia sterowania.

4.7.2. Estymacja sterowań

Załóżmy teraz, że charakterystyka ϕ^* jest jednostajnie ciągła. Niech ξ będzie jej jedynym punktem maksymalnym i ponadto niech dla każdego $\varepsilon > 0$

$$\sup_{\|x-\xi\|>\varepsilon} \phi^*(x) < \phi^*(\xi) . \tag{4.31}$$

Załóżmy także, że η jest jedynym rozwiązaniem równania $\phi^*(x) = \alpha$ oraz że dla każdego $\varepsilon > 0$

$$\inf_{\|x-\eta\|>\varepsilon} |\phi^*(x) - \alpha| > 0 . \tag{4.32}$$

Udowodnimy teraz twierdzenie, które wiąże zagadnienie identyfikacji ze sterowaniem.

Twierdzenie 4.9

Jeśli charakterystyka ϕ^* jest jednostajnie ciągła i spełnione są odpowiednio założenia (4.31) lub (4.32) oraz

$$\sup_x |\phi^*(x) - \phi_n(x)| \xrightarrow{p} 0 ,$$

gdy $n \to \infty$, to odpowiednio

$$\|\xi - \xi_n\| \xrightarrow{p} 0 \quad \text{lub} \quad \|\eta - \eta_n\| \xrightarrow{p} 0 , \tag{4.33}$$

gdy $n \to \infty$, oraz

$$\phi^*(\xi_n) \xrightarrow{p} \phi^*(\xi) \quad \text{lub} \quad \phi^*(\eta_n) \xrightarrow{p} \phi^*(\eta), \tag{4.34}$$

gdy $n \to \infty$.

D o w ó d

Wykażemy najpierw prawdziwość twierdzenia dla sterowania ekstremalnego. Z założeń wynika, że dla każdego $\varepsilon > 0$ istnieje $\delta > 0$ także, że z nierówności

$$|\phi^*(x) - \phi^*(\xi)| < \delta$$

wynika, że

$$\|x - \xi\| < \varepsilon .$$

Jest więc oczywiste, że zbieżność (4.34) implikuje (4.33). Aby dowieść ostatniej zauważmy, że

$$|\phi^*(\xi) - \phi^*(\xi_n)| \leq |\phi^*(\xi) - \phi_n(\xi_n)| + |\phi^*(\xi_n) - \phi_n(\xi_n)| \leq$$

$$\leq |\sup_x \phi^*(x) - \sup_x \phi_n(x)| + \sup_x |\phi^*(x) - \phi_n(x)| \leq$$

$$\leq 2 \sup_x |\phi^*(x) - \phi_n(x)| .$$

Z kolei dowiedziemy twierdzenia dla sterowania przy zadanym wyjściu. Zdefiniujemy w tym celu

$$\Phi^{x}(x) = \begin{cases} \phi^{*}(x) & \text{jeśli} \quad \phi^{*}(x) \leqslant \alpha , \\ -\phi^{*}(x) & \text{jeśli} \quad \phi^{*}(x) > \alpha \end{cases}$$

oraz

$$\Phi_n(x) = \begin{cases} \phi_n(x) & \text{jeśli} \quad \phi_n(x) \leqslant \alpha \\ -\phi_n(x) & \text{jeśli} \quad \phi_n(x) > \alpha . \end{cases}$$

Jest oczywiste, że η i η_n maksymalizują odpowiednio Φ^{*} i Φ_n. Z założenia wynika ponadto, że

$$\sup_x | \Phi^{*}(x) - \Phi_n(x)| \xrightarrow{p} 0 ,$$

gdy $n \to \infty$. Stąd i pierwszej części dowodu wynika ostatecznie teza, co kończy dowód. ■

4.7.3. Jednostajna zbieżność procesu identyfikacji

Wykażemy teraz, że przy pewnych założeniach algorytm identyfikacji stosujący estymator Parzena jest jednostajnie zbieżny według prawdopodobieństwa do charakterystyki optymalnego modelu. Interesować nas będzie przy tym zbieżność w dowolnym ograniczonym zbiorze A, dla którego

$$\inf_{x \in A} g(x) > 0 . \tag{4.35}$$

Załóżmy ponadto, że w zbiorze tym iloczyn $g\,\phi^{*}$ jest funkcją jednostajnie ciągłą. Aby wykazać jednostajną zbieżność w zbiorze A wystarczy więc udowodnić, że

$$\lim_{n \to \infty} \sup_{x \in A} | g(x)(E\,\phi_n(x) - \phi^{*}(x)) | = 0 \tag{4.36}$$

oraz

$$\lim_{n \to \infty} E \left\{ \sup_{x \in A} | g(x)(\phi_n(x) - E\,\phi_n(x)) |^2 \right\} = 0 , \tag{4.37}$$

ponieważ

$$\inf_{x \in A} g(x) \sup_{x \in A} | \phi^{*}(x) - \phi_n(x)| \leqslant \sup_{x \in A} g(x) | \phi^{*}(x) - \phi_n(x) | .$$

Zbieżność (4.36) wynika z (4.14) i jednostajnej ciągłości $g\,\phi^{*}$.

Niech

$$k(\omega) = \int K(x)\, e^{-j\omega^{T}x}\, dx$$

będzie absolutnie całkowalną transformatą Fouriera funkcji K, zatem

$$\phi_n(x)\, g(x) = (2\pi)^{-p} \int k\left(\omega h(n)\right) e^{j\omega^T x} \left[\frac{1}{nh^p(n)} \sum_{i=1}^{n} Y_i\, e^{j\omega^T X_i}\right] d\omega \quad ,$$

stąd

$$E\left\{\sup_x \left| g(x)\left(\phi_n(x) - E\,\phi_n(x)\right)\right|^2\right\} \leqslant (2\pi)^{-p} \frac{2}{nh^{4p}(n)} EY^2 \left[\int |k(\omega)|\, d\omega\right]^2,$$

bowiem

$$\sup_x \left| g(x)\left(\phi_n(x) - E\,\phi_n(x)\right)\right| \leqslant (2\pi)^{-p} \frac{1}{nh^{2p}(n)} \int |k(\omega)|\, d\omega \sum_{i=1}^{n} \left(|Y_i| + E|Y_i|\right).$$

Wykazaliśmy więc następujące twierdzenie.

Twierdzenie 4.10

Jeśli spełnione są założenia twierdzenia 4.5 oraz (4.35) i ponadto $g\phi^*$ jest jednostajnie ciągła w zbiorze A oraz k jest funkcją absolutnie całkowalną i

$$\lim_{n\to\infty} nh^{4p}(n) = \infty \quad ,$$

to

$$\sup_{x\in A} \left|\phi^*(x) - \phi_n(x)\right| \xrightarrow{p} 0 \quad ,$$

gdy $n\to\infty$, przy czym

$$\phi_n(x) = \frac{1}{nh^p(n)\, g(x)} \sum_{i=1}^{n} y_i\, K\left(\frac{x - x_i}{h(n)}\right) .$$

Analogicznie można wykazać jednostajną zbieżność dla zmodyfikowanego, rekurencyjnego algorytmu (4.14). Własność tę podamy także w formie twierdzenia.

Twierdzenie 4.11

Jeśli spełnione są założenia twierdzenia 4.6 oraz (4.35) i $g\phi^*$ jest jednostajnie ciągła w zbiorze A oraz k jest funkcją absolutnie całkowalną i

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^{4p}(i)} = 0 \quad ,$$

to

$$\sup_{x\in A} \left|\phi^*(x) - \phi_n'(x)\right| \xrightarrow{p} 0 \quad ,$$

gdy $n\to\infty$, gdzie

$$\phi_n'(x) = \frac{1}{n\, g(x)} \sum_{i=1}^{n} \frac{1}{h^p(i)}\, y_i\, K\left(\frac{x - x_i}{h(i)}\right).$$

D o w ó d

Zbieżność analogiczna do (4.36) wynika z jednostajnej ciągłości $g\,\phi^*$ oraz (4.21). Aby wykazać zbieżność odpowiadającą (4.37) zauważmy, że

$$\phi_n'(x)\, g(x) = (2\pi)^{-p} \int e^{j\omega^T x} \frac{1}{n} \sum_{n=1}^{n} \left(\frac{1}{h^p(i)} k(\omega h(i))\, Y_i\, e^{j\omega^T X_i} \right) d\omega ,$$

stąd

$$V_n \overset{\text{def}}{=} \sup_x |g(x)(\phi_n(x) - E\,\phi_n(x))| \leqslant$$

$$\leqslant (2\pi)^{-p} \frac{1}{n} \int |k(\omega)| d\omega \sum_{i=1}^{n} \frac{1}{h^{2p}(i)} (|Y_i| + E|Y_i|),$$

zatem

$$EV_n^2 \leqslant (2\pi)^{-2p}\, EY^2 \left[\int |k(\omega)|\, d\omega \right]^2 \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^{4p}(i)} ,$$

do kończy dowód. ■

Z powyższych twierdzeń oraz ogólnego twierdzenia 4.9 wynika, że jeśli nie znane sterowania leżą w zbiorze A (patrz (4.35)), to korzystanie z estymatora Parzena także w formie rekurencyjnej prowadzi do algorytmu identyfikacji asymptotycznie optymalnego ze względu na sterowanie. Należy zaznaczyć, że ze względu na założenie o absolutnej całkowalności transformaty Fouriera nie można stosować jako jądra K funkcji podanej w pierwszym z przykładów (3.5).

4.8. Uwagi

W przedstawionych algorytmach identyfikacji można stosować różne metody estymacji gęstości prawdopodobieństwa. Interesujące jest zwłaszcza korzystanie z oszacowania Parzena w oryginalnej a także i rekurencyjnej formie. Można wówczas otrzymać procedurę zbieżną jednostajnie według prawdopodobieństwa do krzywej regresji charakteryzującej optymalny model. Charakterystyką modelu można wtedy posługiwać się w celu znalezienia nie znanych sterowań zapewniających pożądane własności wyjścia obiektu.

Druga koncepcja polega na rozwinięciu odpowiedniej funkcji w nieskończony szereg ortonormalny. Wejście może być przy tym ograniczone lub nie. W tym drugim przypadku jako układ ortogonalny przyjmuje się wielomiany Hermite'a.

Liczby obliczeń rosną liniowo wraz z wymiarowością wejścia obiektu oraz długością ciągu uczącego, tzn. liczbą obserwacji wejścia i wyjścia obiektu. Stosowanie charakterystyki modelu do sterowania stwarza konieczność rozwiązywania odpowiedniego równania lub minimalizacji metodami iteracyjnymi. Wymiary tych zadań są równe liczbie wejść obiektu.

Jak wynika z podanego przykładu celowe byłoby opracowanie adaptacyjnych metod ustalania stałej c, przy korzystaniu z estymatora Parzena, na podstawie wyników eksperymentu, otrzymanych w procesie identyfikacji. Pewne sugestie co do jego wyboru przy dysponowaniu odpowiednią wstępną informacją podano w pracy [25].

ZAKOŃCZENIE

W pracy przedstawiono ogólne twierdzenia o asymptotycznej optymalności procesów uczenia w rozpoznawaniu i identyfikacji w sytuacji całkowitego braku danych probabilistycznych oraz podano algorytmy otrzymywane przy stosowaniu różnych typów nieparametrycznych oszacowań. Rozważano jedynie własności asymptotyczne procedur, nie analizowano ich natomiast przy ustalonych długościach ciągu uczącego. Podane przykłady wskazują, że istotny jest problem takiego wyboru pewnych elementów w algorytmach uczenia rozpoznawania i identyfikacji, który zapewniłby najlepszą dokładność. Badania eksperymentalne i analityczne wykazują, że dla przykładu, w oszacowaniu Parzena wybór ciągu $\{h(n)\}$ ma istotniejsze znaczenie niż jądra K. Jego optymizacja wymaga jednak praktycznie nieosiągalnych danych wstępnych. Pojawia się więc problem uzależnienia tego wyboru od ciągu uczącego. Za otwarte należy także uznać zagadnienie powiązania dokładności oszacowań gęstości z ryzykiem i oceny szybkości zbieżności ryzyka, pomimo znanych prób [70, 81]. Dla procesów identyfikacji powiązanie dokładności oszacowań gęstości z jakością modelu wydaje się łatwiejsze (patrz np. (4.6)), choć znaczne komplikacje powstają przy nieograniczonym wyjściu i nie znanym rozkładzie wejścia.

Istotny jest także aspekt obliczeniowy otrzymanych algorytmów. Liczby obliczeń niezbędnych do sklasyfikowania obrazu lub wyznaczenia charakterystyki modelu w ustalonym punkcie rosną liniowo ze wzrostem długości ciągu uczącego. Zależą one także liniowo od wymiarowości problemu, tzn. od wymiarów obrazów i wejścia obiektu. Wzrost wymiarowości powoduje także pogorszenie dokładności oszacowania gęstości [25], a zatem i jakości rozpoznawania i identyfikacji. Warto podkreślić, że liczby obliczeń wykonywanych przy korzystaniu z algorytmów przedstawionych w pracy i np. znanych algorytmów NN [15] i LI [10] są zbliżone. Reguła NN nie ma jednak asymptotycznych własności optymalnych.

Wspólną cechą rozważanych algorytmów jest konieczność pamiętania całego ciągu uczącego. Godną odnotowania jest więc próba ograniczenia pamięci, kosztem niewielkiej - jak się wydaje - straty dokładności przy stosowaniu estymatora Parzena [66].

Proces identyfikacji powiązano ze sterowaniem i wykazano, że oszacowania sterowań wyznaczone na podstawie modelu są zgodne. Ich wyznaczenie wymaga jednak stosowania odpowiednich procedur numerycznych w celu rozwiązania odpowiednich równań lub ekstremalizacji określonych funkcji.

Należy jeszcze zwrócić uwagę na wykazaną, bardzo ważną własność procedury uczenia rozpoznawania metodą funkcji potencjalnych, a mianowicie asymptotyczną optymalność w sensie średniego ryzyka.

LITERATURA

[1] A b r a m s o n N., B r a w e r m a n D., Learning to recognize patterns in a random environment, IRE Trans. IT, vol. IT-8, 1962, s. 58-63.

[2] A j z e r m a n M. A., B r a v e r m a n E. M., R o z o n o e r L. I., Metod potencjalnych funkcij v teorii obučenija mašin, Nauka, Moskva 1970.

[3] A j z e r m a n M. A., B r a v e r m a n E. M., R o z o n o e r L. I., Metod potencjalnych funkcij v zadače o vosstanovlenii charakteristiki funkcionalnogo preobrazovatelja po slučajno nabludaemym točkam, Avtomatika i Telemechanika, t. XXV, Nr 12, 1964, 1705-1714.

[4] A j z e r m a n M. A., B r a v e r m a n E. M., R o z o n o e r L. I., Verojatnostnaja zadača ob obučenii avtomatov rozpoznavaniju klassov i metod potencjalnych funkcij, Avtomatika i Telemechanika, t. XXV, Nr 9, 1964, 1307-1323.

[5] A n d e r s o n T., An introduction to multivariate statistical analysis, Wiley, New York, 1958.

[6] B l a y d o n C. C., Approximation of distribution and density functions, Proc. IEEE, Nr 2, 1967.

[7] B l a y d o n C. C., K a s h y a p R. L., W a g n e r T. J., Comments on the estimation of distribution functions, IEEE Trans. IT, vol. IT-16, 1970.

[8] B r a v e r m a n E. M., O metode potencjalnych funkcij, Avtomatika i Telemechanika, t. XXVI, Nr 12, 1965.

[9] B u b n i c k i Z., Identyfikacja obiektów sterowania, Warszawa 1974, PWN.

[10] B u b n i c k i Z., Least interval pattern classification and its application to control systems, IV IFAC Congress, Nr 21, Warszawa 1969.

[11] B u b n i c k i Z., System identification via estimation of probability distribution and moments, 2-nd IFAC Symposium on Identification, Prague 1970.

[12] Bubnicki Z., Zbieżność procesów automatycznej aproksymacji w układach dyskretnych,Zeszyty Naukowe Politechniki Wrocławskiej, Automatyka 5, Wrocław 1966.

[13] Cacoullos T., Estimation of multivariate density,Inst. Statist. Math., vol. 18, Nr 2, 1966, s. 179-189.

[14] Chien Y. T., Fu K. S., On bayesian learning and stochastic approximation, IEEE Trans. SSC, vol. SSC-3, Nr 1, 1967.

[15] Cover T. M., Hart P. E., Nearest neighbor pattern classification, IEEE Trans. IT, vol. IT-13, Nr 1, 1967, s. 21-27.

[16] Cypkin J. Z., Adaptacija, obučenie i samoobučenie v avtomatičeskich sistemach, Avtomatika i Telemechanika, t. XXVII, Nr 1, 1966.

[17] Cypkin J. Z., Ob algoritmach ocenki plotnosti raspredelenija i momentov po nabludenijam, Avtomatika i Telemechanika, t. XXVIII, Nr 7, 1967.

[18] Cypkin J. Z., O vosstanovlenii charakteristiki funkcionalnogo preobrazovatelja po slučajno nabludaemym točkam,Avtomatika i Telemechanika, t. XXVI, Nr 11, 1965.

[19] Cypkin J. Z., Primenenie metoda stochatičeskoj approksimacii k ocenke neizvestnoj plotnosti raspredelenija po nabludenijam, Avtomatika i Telemechanika, t. XXVII, Nr 3, 1966.

[20] Cypkin J. Z., Podstawy teorii układów uczących się, Warszawa 1973, WNT.

[21] Čencov N. N., Ocenka neizvestnoj plotnosti raspredelenija po nabludenijam, Doklady A. N. SSSR, t. 147, Nr 1, s. 45-48.

[22] Devjaternikov I. P., Propoj A. I., Cypkin J. Z., O rekurentnych algoritmach obučenija raspoznawaniju obrazov, Avtomatika i Telemechanika, t. XXVIII, Nr 2,1967, s. 122-132.

[23] Dobrovidov A. V., Ob odnom algoritme neparametričeskoj ocenki slučajnych mnogomernych signalov, Avtomatika i Telemechanika, t. XXXII, Nr 2, 1972, s. 88-99.

[24] Elkins T. A., Cubical and spherical estimation of multivariate probability density, J. Amer. Statist. Assoc.,vol.63, 1968, s. 1499-1513.

[25] Epanečnikov V. A., Neparametričeskaja ocenka mnogomernoj plotnosti verojatnosti, Teorija verojatnostej i ee primenenija, t. XIV, 1969, s. 156-162.

[26] Gessaman M. P., A consistent nonparametric multivariate density estimator based on statistically equivalent blocks, Ann. Math. Statist., vol. 41, Nr. 4, 1970, s. 1344-1346.

[41] K o n a k o v V. D., Neparametričeskaja ocenka plotnosti raspredelenija verojatnostej, Teorija verojatnostej i ee primenenija, t. XVII, Nr 2, 1972, s. 377-379.

[42] K r o n m a l R., T a r t e r M., The estimation of probability densities and cumulatives by Fourier series methods, J. Amer. Statist. Assoc., vol. 63, 1968, s. 925-952.

[43] K u l i k o w s k i J., Cybernetyczne układy rozpoznające, Warszawa 1972, PWN.

[44] L o f t s g a a r d e n D. O., Q u e s e n b e r r y C. P., A nonparametric estimation of multivariate density function, Ann. Math. Statist., vol. 36, Nr 3, 1965, s. 1049-1051.

[45] M a ń c z a k K., Metody identyfikacji wielowymiarowych obiektów sterowania, Warszawa 1971, WNT.

[46] M a ń c z a k K., Zastosowanie analizy regresyjnej do identyfikacji statycznych charakterystyk wielowymiarowych obiektów technologicznych, Archiwum Automatyki i Telemechaniki, Nr 1, 1966.

[47] M o o r e D. S., H e n r i c h o n E. G., Uniform consistency of some estimates of a density function, Ann. Math. Statist., vol. 40, Nr 4, 1968, s. 1499-1502.

[48] N a g y G., State of art in pattern recognition, Proc. IEEE, vol. 56, Nr 5, 1968, s. 836-862.

[49] N a d a r a j a E. A., O neparametričeskich ocenkach plotnosti verojatnosti i regressii, Teorija verojatnostej i ee primenenija, t. X, Nr 1, 1965, s. 199-203.

[50] N i l s o n N. J., Maszyny uczące się, Warszawa 1968, PWN.

[51] P a r z e n E., On estimation of a probability density function and mode, Ann. Math. Statist., vol. 33, 1962, s.1065-1076.

[52] P a t t e r s o n J. D., W a g n e r T. J., W o m a c k B. F., A mean-square performance criterion for adaptive pattern classification systems, IEEE Trans. AC, vol. AC-12, Nr 2, 1967, s. 195-197.

[53] P e t e r s o n D. W., M a t t s o n R. L., A method of finding linear discriminant functions for a class of performance criteria, IEEE Trans. IT, vol. IT-12, Nr 3, 1966, s. 380-387.

[54] P i t t J. M., W o m a c k B. F., Additional features of an adaptive multicategory pattern classification systems, IEEE Trans. SSC, vol. SSC-5, Nr 3, 1969, s. 183-191.

[55] R o b b i n s H., The empirical Bayes Approach to statistical decision problems, Ann. Math. Statist., vol. 35, 1964, s. 1-20.

[41] K o n a k o v V. D., Neparametričeskaja ocenka plotnosti raspredelenija verojatnostej, Teorija verojatnostej i ee primenenija, t. XVII, Nr 2, 1972, s. 377-379.

[42] K r o n m a l R., T a r t e r M., The estimation of probability densities and cumulatives by Fourier series methods, J. Amer. Statist. Assoc., vol. 63, 1968, s. 925-952.

[43] K u l i k o w s k i J., Cybernetyczne układy rozpoznające, Warszawa 1972, PWN.

[44] L o f t s g a a r d e n D. O., Q u e s e n b e r r y C. P., A nonparametric estimation of multivariate density function, Ann. Math. Statist., vol. 36, Nr 3, 1965, s. 1049-1051.

[45] M a ń c z a k K., Metody identyfikacji wielowymiarowych obiektów sterowania, Warszawa 1971, WNT.

[46] M a ń c z a k K., Zastosowanie analizy regresyjnej do identyfikacji statycznych charakterystyk wielowymiarowych obiektów technologicznych, Archiwum Automatyki i Telemechaniki, Nr 1, 1966.

[47] M o o r e D. S., H e n r i c h o n E. G., Uniform consistency of some estimates of a density function, Ann. Math. Statist., vol. 40, Nr 4, 1968, s. 1499-1502.

[48] N a g y G., State of art in pattern recognition, Proc. IEEE, vol. 56, Nr 5, 1968, s. 836-862.

[49] N a d a r a j a E. A., O neparametričeskich ocenkach plotnosti verojatnosti i regressii, Teorija verojatnostej i ee primenenija, t. X, Nr 1, 1965, s. 199-203.

[50] N i l s o n N. J., Maszyny uczące się, Warszawa 1968, PWN.

[51] P a r z e n E., On estimation of a probability density function and mode, Ann. Math. Statist., vol. 33, 1962, s.1065-1076.

[52] P a t t e r s o n J. D., W a g n e r T. J., W o m a c k B. F., A mean-square performance criterion for adaptive pattern classification systems, IEEE Trans. AC, vol. AC-12, Nr 2, 1967, s. 195-197.

[53] P e t e r s o n D. W., M a t t s o n R. L., A method of finding linear discriminant functions for a class of performance criteria, IEEE Trans. IT, vol. IT-12, Nr 3, 1966, s. 380-387.

[54] P i t t J. M., W o m a c k B. F., Additional features of an adaptive multicategory pattern classification systems, IEEE Trans. SSC, vol. SSC-5, Nr 3, 1969, s. 183-191.

[55] R o b b i n s H., The empirical Bayes Approach to statistical decision problems, Ann. Math. Statist., vol. 35, 1964, s. 1-20.

[56] R o b e r t s o n T., On estimating a density which is measurable with respect to a σ-lattice, Ann. Math. Statist., vol. 38, 1967, s. 482–493.

[57] R o s e n b l a t t M., Curve estimates, Ann. Math. Statist. vol. 42, 1971, s. 1815–1842.

[58] R o s e n b l a t t M., Remarks on some estimates of a density function, Ann. Math. Statist., vol. 27, 1956, s. 823–837.

[59] S a r i d i s G. N., N i c o l i c Z. J., F u K. S., Stochastic approximation algorithms for system identification, estimation and decomposition of mixtures, IEEE Trans. SSC, vol. SSC-5, Nr 1, 1969.

[60] S e i d l e r J., Optymizacja adaptacyjnych systemów informacyjnych, Warszawa-Wrocław 1971, PWN.

[61] S e i d l e r J., Statystyczna teoria odbioru sygnałów, Warszawa 1963, PWN.

[62] S i k o r s k i R., Funkcje rzeczywiste, Warszawa 1959, PWN.

[63] S c h w a r t z S. C., Estimation of probability density by an orthogonal series, Ann. Math. Statist. vol. 38, 1967, s. 1261–1265.

[64] S e b e s t y e n G., Decision-making processes in pattern recognition, Macmillan, New York 1962.

[65] S e b e s t y e n G., E d i e J., An algorithm for nonparametric pattern recognition, IEEE Trans. EC, vol. EC-15, 1966, s. 908–915.

[66] S p e c h t D. F., Generation of polynomial discriminant functions for pattern recognition, IEEE Trans. EC, vol. EC-16, 1967, s. 308–319.

[67] S t o l l e r D. S., Univariate two-population distribution-free discrimination, J. Amer. Statist. Assoc., vol. 49, 1954, s. 770–777.

[68] Van R y z i n J., Bayes risk consistency of classification procedures using density estimation, Sankhya, s. A, vol. 28, pts. 3–4, 1966, s. 161–170.

[69] Van R y z i n J., A histigram method of density estimation, (praca przedstawiona na IMS Meeting, April 8–9, 1970, Dallas, niepublikowana).

[70] Van R y z i n J., Non-parametric bayesian decision procedure for (pattern) classification with stochastic learning, Trans. of Fourth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Prague 1965.

[71] Van R y z i n J., On strong consistency of density estimates, Ann. Math. Statist., vol. 40, 1969, s. 1765–1772.

[72] W a g n e r T. J., Strong consistency of nonparametric estimate of a density function IEEE Trans. SMC, vol. SMC-3, 1973, s. 289-290.

[73] W a g n e r T. J., P i t t J. M., W o m a c k B. F., A comparison between pattern classification approaches. IEEE Trans. IT, vol. IT-13, 1967, s. 611-613.

[74] W a h b a G., A polynomal algorithm for density estimation, Ann. Math. Statist., vol. 42, 1971, s. 1870-1886.

[75] W e g m a n E. J., A note on estimating a unimodal density, Ann. Math. Statist., vol. 40, 1969, s. 1661-1667.

[76] W e g m a n E. J., Nonparametric probability density estimation; I. A summary of available methods, Technometrics, vol.14, Nr 3, 1972, s. 533-546.

[77] W e g m a n E. J., Nonparametric probability density estimation; II. A comparison of density estimation methods, J. Statist. Comput. Simul., vol 1, 1972, s. 225-245.

[78] W e i s s L., W o l f o w i t z J., Estimation of density function at a point, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, vol. 7, 1967, s. 327-335.

[79] W e s s e l G. N., S k l a n s k y J., Training a one-dimentional classifier to minimize the probability of error, IEEE Trans. SMC, vol. SMC-2, Nr 4, 1972, 535-541.

[80] W ę g r z y n S., Modele matematyczne i identyfikacja obiektów, [w:] Modele matematyczne i identyfikacja obiektów, Wrocław 1970, Ossolineum.

[81] W o l v e r t o n C. T., W a g n e r T. J., Asymptotically optimal discriminant functions for pattern classification, IEEE Trans. IT, vol. IT-15, Nr 2, 1969, s. 258-265.

[82] W o l v e r t o n C. T., W a g n e r T. J., Recursive estimates of probability densities, IEEE Trans. SSC, vol. SSC-5, Nr 3, 1969.

[83] Y a u S. S., S c h u m p e r t J. M., Design of pattern classifiers with the updating property using stochastic approximation techniques, IEEE Trans. C, vol. C-17, Nr 9, 1968, s. 861-872.

ASYMPTOTICALLY OPTIMAL PROBABILISTIC ALGORITHMS FOR PATTERN RECOGNITION AND IDENTIFICATION

The paper considers asymptotically Bayes optimal supervised learning procedures for pattern recognition, and identification derived from either nonparametric probability density estimators or orthonormal expantion. The Bayes optimal classification rule ϕ^* minimizes the risk

$$R(\phi) = \sum_{j \in \Omega} \int_{\mathcal{X}} L(\phi(x), j)\, p_j\, f_j(x)\, dx ,$$

where:

$\Omega = \{1, \ldots, M\}$ is a space of classes (populations),

$L(i,j)$ - the loss function,

p_i - prior class probability,

f_i - class density,

ϕ - a classification rule,

$x \in \mathcal{X}$ - ($\mathcal{X}$ - p-vector space of individuals with Lebesgue measure μ),

and classifies x as coming from any class of the set

$$\Phi^*_x \subset \Omega ,$$

where Φ^*_x is a set of all classes which minimize

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

p_i's and f_i's are not known but the learning sequence $(\omega_1, x_1), \ldots, (\omega_M, x_M)$, i.e. a sequence of independent and correctly classified individuals is given; ω_i being a class of a sample x_i. n_i

Probabilities p_i's are estimated by the rates $p_{in} = \frac{n_i}{n}$, where n_i is the number of samples from the class i. Probability densities are estimated nonparametrically.

An empirical classification rule ϕ_n assignes x to any class which minimizes

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

f_{in} is a nonparametric estimator of f_i. A distance between an action (classification) i and a set of actions $\Omega' \subset \Omega$ is denoted by

$$\rho(i, \Omega') = \min_{j \in \Omega'} |i-j| .$$

The papers prove that if

$$f_{in}(x) \xrightarrow[a.s.]{p} f_i(x) \qquad \text{as} \quad n \to \infty,$$

for every $i \in \Omega$ then

$$\varphi(\phi_n(x), \Phi_x^*) \xrightarrow[a.s.]{p} 0 \qquad \text{as} \quad n \to \infty.$$

It is clear that when the optimal classification rule ϕ^x is univocal for $x \in \mathcal{X}$ then

$$\phi_n(x) \xrightarrow[a.s.]{p} \phi^*(x) \qquad \text{as} \quad n \to \infty.$$

Moreover, if

$$f_{in}(x) \xrightarrow[a.s.]{p} f_i(x) \qquad \text{as} \quad n \to \infty$$

for almost all (μ) $x \in \mathcal{X}$
then

$$\int_{\mathcal{X}} \rho\left(\phi_n(x), \Phi_x^*\right) f(x)dx \xrightarrow[a.s.]{p} 0 \qquad \text{as} \quad n \to \infty,$$

where

$$f(x) = \sum_{i=1}^{M} p_i f_i(x)$$

and

$$R(\phi_n) \xrightarrow[a.s.]{p} R(\phi^*) \qquad \text{as} \quad n \to \infty.$$

Thus, when the optimal rule ϕ^* is univocal for in almost all $x \in \mathcal{X}$, then

$$\int_{\mathcal{X}} |\phi^*(x) - \phi_n(x)|^2 f(x)\, dx \xrightarrow[a.s.]{p} 0 \qquad \text{as} \quad n \to \infty.$$

The integrated consistency of the density estimation implies risk consistency.

In the next section pattern recognition algorithms obtainable when using Parzen, Loftsgaarden and Quesenberry or orthogonal series nonparametric probability density estimators are considered. Each of them is applied in two ways, e.g. for Parzen estimator:

$$f_{in}(x) = \frac{1}{n_i h^p(n_i)} \sum_{k=1}^{n_i} K\left(\frac{x - x_{ik}}{h(n_i)}\right)$$

or

$$f_{in}(x) = \frac{1}{n_i h^p(n)} \sum_{k=1}^{n_i} K\left(\frac{x - x_{ik}}{h(n)}\right),$$

where

x_{ij} - the j-th... sample from the class i,

$\{h(n)\}$ - a number sequence,

K - density type kernel.

Geometrical interpretation and numerical examples are given.

Analogous approach is applied to identification (for quadratic loss function) of an object with input x (x - p-vector) and scalar output y. On the basis of the learning sequence $(x_1,y_1),\ldots,(x_n,y_n)$, i.e. the sequence of independent input and output observations, the regression function

$$\phi^*(x) = \int y\, f_x(y)dy = \frac{1}{g(x)} \int y\, f(x,y)dy$$

is estimated;

f_x - the conditional output density when input is x,

g - the input density and $f(x,y) = g(x)\, f_x(y)$.

The estimators of ϕ^* take the form

$$\phi_n(x) = \frac{1}{g(x)} \int y\, f_n(x,y)dy ,$$

or

$$\phi_n'(x) = \frac{1}{g_n(x)} \int y\, f_n(x,y)dy$$

for g known and unknown, respectively (f_n and g_n are the corresponding non-parametric estimators of densities f and g).

For the Parzen estimation we have

$$\phi_n(x) = \frac{1}{nh^p(n)g(x)} \sum_{i=1}^{n} y_i\, K\left(\frac{x - x_i}{h(n)}\right)$$

and

$$\phi_n'(x) = \frac{\sum_{i=1}^{n} y_i\, K\left(\frac{x - x_i}{h(n)}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h(n)}\right)} .$$

It has been proved that if $g(x) > 0$ and

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

$$\lim_{n\to\infty} nh^{2p}(n) = \infty$$

then (under some assumptions)

$$\lim_{n\to\infty} E\,|\,\phi^*(x) - \phi_n(x)|^2 = 0$$

and

$$\phi_n'(x) \xrightarrow[z\ p.\ 1]{p} \phi^*(x) \qquad \text{as } n\to\infty$$

in all points of continuity of $g\,\phi^*$.

Moreover, if

$$\lim_{n\to\infty} n\,h^{4p}(n) = \infty$$

then (under additional assumptions)

$$\sup_x |\,\phi^*(x) - \phi_n(x)| \xrightarrow[z\ p.\ 1]{p} 0 \qquad \text{as } n\to\infty.$$

The above holds also for recursive estimators

$$\phi_n(x) = \frac{1}{ng(x)} \sum_{i=1}^{n} \frac{1}{h^p(i)}\, y_i\, K\left(\frac{x - x_i}{h(n)}\right)$$

and

$$\phi_n'(x) = \frac{\sum_{i=1}^{n} \frac{1}{h^p(i)}\, y_i\, K\left(\frac{x - x_i}{h(n)}\right)}{\sum_{i=1}^{n} \frac{1}{h^p(i)}\, K\left(\frac{x - x_i}{h(n)}\right)}.$$

The second methods is the orthonormal expantion. A function

$$\int y\, f(x,y)\, dy \in L^2$$

is expanded in orthonormal complete set $\{\varphi_i\}_{i=0}^{\infty}$ of commonly bounded functions. The regression function ϕ^* is estimated by

$$\phi_n(x) = \frac{1}{g(x)} \sum_{i=0}^{N(n)} a_{in}\, \varphi_i(x)$$

or

$$\phi_n'(x) = \frac{\sum_{i=0}^{N(n)} a_{in}\, \varphi_i(x)}{\sum_{i=0}^{N(n)} b_{in}\, \varphi_i(x)},$$

where

$$a_{in} = \frac{1}{n} \sum_{j=1}^{n} y_j \varphi_i(x_j)$$

and

$$b_{in} = \frac{1}{n} \sum_{j=1}^{n} \varphi_i(x_j) .$$

If

$$\lim_{n\to\infty} N(n) = \infty ,$$

$$\lim_{n\to\infty} \frac{N(n)}{n} = 0$$

then (under some assumptions)

$$\lim_{n\to\infty} E \int | \phi^*(x) - \phi_n(x)|^2 g^2(x)\, dx = 0 .$$

Under additional assumptions also

$$\lim_{n\to\infty} E | \phi^*(x) - \phi_n(x)|^2 = 0$$

and

$$\phi_n(x) \xrightarrow{p} \phi^*(x) \qquad \text{as } n\to\infty .$$

Finally the model is used for controlling. Input ξ which maximizes ϕ^*, and η which is a root of an equation $\phi^*(x) = \alpha$, (α - being given) are estimated by ξ_n and η_n respectively ξ_n maximizes ϕ_n and η_n is a root of an equation $\phi_n(x) = \alpha$. It is shown that if

$$\sup_x | \phi^*(x) - \phi_n(x)| \xrightarrow{p} 0 \qquad \text{as } n\to\infty$$

then

$$\| \xi - \xi_n \| \xrightarrow{p} 0$$

and

$$\phi^*(\xi_n) \xrightarrow{p} \phi^*(\xi) \qquad \text{as } n\to\infty$$

$$\| \eta - \eta_n \| \xrightarrow{p} 0$$

and

$$\phi^*(\eta_n) \xrightarrow{p} \phi^*(\eta) \qquad \text{as } n\to\infty.$$

The above convergencies make possible the construction of identification algorithms optimal asymptotically with respect to control either.

CONTENTS

SPIS RZECZY