IV
Навчання та самонавчання розпізнаванню образів

Training and Selftraining Problems in Pattern Recognition

Розкладання функцій на динамічні складові методами навчання розпізнаванню образів

ВОЛОДИМИР ВАСИЛЬЄВ, ЮРІЙ ГОРІЛОВ

Інститут кібернетики НАН *252022 Київ, просп. Академіка Глушкова 40*

Владимир Васильев, Юрий Горилов. Разложение функций на динамические составляющие методами обучения распознаванию образов.

Показано, как разработанная специальная процедура обучения распознаванию образов может быть использована для восстановления дифференциального уравнения динамической системы по экспериментальным данным.

Часто для математичної формалізації законів функціонування динамічних систем застосовуються залежності, що включають як свої елементи похідні різних порядків у часі від деяких динамічних характеристик системи. При цьому за динамічні характеристики системи обирають як фазові, так і керувальні або контрольовані збурювальні траєкторії.

В [1] показано, що спеціяльну процедуру (альфа-процедура) навчання розпізнаванню образів (НРО) шляхом особливих перетворень можна використати як для розв'язування задачі відновлення багатовимірних функцій, так і для розв'язування задачі розкладання функції одного аргументу на детерміновані складові, як от на гармонічні складові.

Так само, як у [1], розглядається функція y(t), подана у формі скінченного ряду, наприклад, тригонометричного. У цій доповіді складовими функціями ряду є похідні різних порядків від функції y(t). Методами НРО за цих умов обираються такі комбінації похідних від функції y(t), які дозволяють досить добре (у межах коридору) апроксимувати функцію y(t).

Показано, що якщо динаміку системи можна описати лінійним диференціяльним рівнянням з правою частиною, що задовольняє умову Ліпшиця, то пропонований метод може використовуватися для відновлення диференційного рівняння системи за експериментальними даними.

Література

1. Васильев В.И. Теория редукции в проблемах экстраполяции // Проблемы управления и информатики, 1996, №№ 1—2, с. 239-251.

♦

Ефективність процедури розпізнавання, що ґрунтується на використанні роздільної гіперплощини

АНАТОЛІЙ ГУПАЛ, СЕРГІЙ ПАШКО

Науково-учбовий центр прикладної інформатики НАН

252207 Київ, просп. Академіка Глушкова 40
Тел.: (044) 266-4314
Електронна пошта: pashko@compuchair.icyb.kiev.ua

Анатолий Гупал, Сергей Пашко. Эффективность процедуры распознавания, основанной на использовании разделяющей гиперплоскости.

Даны оценка снизу сложности класса задач распознавания образов с признаками, удовлетворяющими условию независимости, и оценка сверху байесовской процедуры распознавания. Обе оценки совпадают с

точностью до абсолютной мультипликативной константы. В булевом случае доказана эквивалентность байесовского подхода и процедуры распознавания, основанной на использовании разделяющей гиперплоскости.

Розглянемо таку задачу розпізнавання. Нехай є скінченна множина B об'єктів b. Кожен об'єкт $b \in B$ ототожнюється з цілочисловим вектором $(x_1, x_2, \ldots x_n, f)$; тут n - натуральне число $(1 \le n < \infty)$; $x_j \in \{0,1,\ldots,g-1\}$, $j = 1,2,\ldots,n$; $f \in \{0,1,\ldots,h-1\}$; g, h - натуральні числа, $g \ge 2, h \ge 2$. Припустимо, на множині B заданий розподіл імовірностей P, який нам невідомий. Із множини B отримана навчальна вибірка V; структура та спосіб її отримання описані нижче. Нехай деякий об'єкт отриманий з множини B незалежно від вибірки V у відповідності з розподілом P, причому відомі тільки значення ознак $x_1, x_2, \ldots, x_n$. Вимагається за цими значеннями та за навчальною вибіркою V визначити значення цільової ознаки f.

Для вивчення ефективності процедури роздільної гіперплощини формалізуємо такі поняття, як клас задач, процедура розпізнавання, складність класу задач тощо.

Позначимо $x = (x_1, x_2, \ldots, x_n)$. Величини x, f, V є випадковими елементами. Нехай $\{\xi\}$ - множина значень, які приймає випадковий елемент ξ. Назвемо функцією розпізнавання функцію A, яка визначена на множині $\{x\}$ і набуває значень з множини $\{f\}$. Вважаємо, що процес розпізнавання об'єкта за відомими ознаками x здійснюється за допомогою функції A: в якості припустимого значення цільової ознаки f вибирається $A(x)$. Доцільно вибирати функцію A у такий спосіб, щоб якомога більшим було значення $P(A(x) = f)$. Оскільки число функцій A скінченне, то серед них існує така функція A^*, що для всіх A справедлива нерівність $P(A^*(x) = f) \ge P(A(x) = f)$. Назвемо похибкою функції A на розподілі P величину

$$\nu(A, P) = P(A^*(x) = f) - P(A(x) = f). \tag{1}$$

Процедурою розпізнавання Q назвемо однозначну функцію, що визначена на множині навчальних вибірок V і набуває значень з множини функцій розпізнавання A; процедура Q будує функцію A за V, тобто $A = Q(V)$. Похибкою процедури Q на розподілі P назвемо величину

$$\nu(Q, P) = \sum\nolimits_{W \in \{V\}} \nu(A, P) P_1(V = W); \tag{2}$$

тут $A = Q(W)$, $P_1(V = W)$ - імовірність події, яка полягає в тому, що навчальна вибірка V приймає значення W.

Підмножину об'єктів із B, у яких цільова ознака дорівнює i, назвемо i-м класом об'єктів. У навчальній вибірці V кількість об'єктів різних класів відома; більше того, на практиці ця кількість часто визначається заздалегідь. Тому будемо вважати, що вибірка V складається з $(h+1)$-ї частини, $V = (V_0, V_1, \ldots, V_h)$. У випадку, коли $0 < m_i < \infty$, $i \in \{0,1,\ldots,h-1\}$, частина V_i є цілочисловою матрицею розмірністю $m_i \times n$. Кожен рядок цієї матриці є спостережуваним значенням вектора $x = (x_1, x_2, \ldots, x_n)$, що описує об'єкт класу i, який вибраний з множини B у відповідності з розподілом імовірностей P при умові $f = i$.

Використовуючи матрицю V_i, легко обчислити частоти подій $\{x_j = s\}$, $s = 0,1,\ldots,g-1$, за умовою $f = i$. Іноді відомі точні значення умовних імовірностей $P(x_j = s | f = i)$. У цьому випадку зручно вважати, що $m_i = +\infty$, а V_i є матрицею розмірністю $n \times g$ з елементами $v_{ijs} = P(x_j = s | f = i)$, $j = 1,2,\ldots,n$; $s = 0,1,\ldots,g-1$. За умови $m_i = 0$ в якості V_i вибирається число -1.

Якщо $0 < m_h < \infty$, то остання частина V_h є цілочисловим вектором розмірністю m_h. Кожна компонента цього вектора є спостереженим значенням цільової ознаки f, що вибирається у відповідності з розподілом P, тобто ймовірність того, що j-та компонента вектора V_h приймає значення i, дорівнює $P(f = i)$. Якщо $m_h = \infty$, то V_h є вектором розмірністю h з компонентами $v_{hi} = P(f = i)$, $i = 0,1,\ldots,h-1$. За умови $m_h = 0$ в якості V_h вибирається число -1.

Всі рядки матриць V_i, $i = 0,1,\ldots,h-1$, і компоненти вектора V_h є незалежними випадковими елементами. Цілі невід'ємні числа $m_0, m_1, \ldots, m_h$ будемо вважати детермінованими. Позначимо $m^{(h)} = (m_0, m_1, \ldots, m_h)$. Вважаємо $m_0 \le m_1 \le \ldots \le m_{h-1}$; виконання цих нерівностей завжди можна забезпечити шляхом перенумерації класів об'єктів.

Назвемо класом задач $C = C(g, h, m^{(h)}, n)$ сукупність всіх можливих розподілів імовірностей P на множині B разом з величинами $g, h, m^{(h)}, n$. Множина задач та множина розподілів імовірностей даного класу знаходяться у взаємно-однозначній відповідності. Похибкою процедури розпізнавання Q на класі C назвемо число

$$\nu(Q, C) = \sup_{P \in C} \nu(Q, P). \tag{3}$$

Складністю класу задач C назвемо число

$$\mu(C) = \inf_{Q} \nu(Q, C). \tag{4}$$

Зрозуміло, що $0 \le \mu(C) \le \nu(Q, C) \le 1$, і бажано користуватися такими процедурами розпізнавання Q, для яких число $\nu(Q, C)$ мало відрізняється від числа $\mu(C)$.

Нехай $d = (d_1, d_2, \ldots, d_n)$ - цілочисловий вектор. Вважаємо, що розподіл імовірностей P з класу C при кожному d задовольняє умову $P(x_1 = d_1, \ldots, x_n = d_n | f = i) = \prod_{j=1}^{n} P(x_j = d_j | f = i)$, $i = 0,1,\ldots,h-1$, що означає незалежність ознак x_j для кожного класу об'єктів. Вважаємо також, що виконуються співвідношення $h \le 2g^n$; $P(f = i) > 0$, $i = 0,1,\ldots,h-1$; $0/0 = 0$.

Розглянемо випадкові величини $\xi(d, i)$, які залежать від d та i як від параметрів:

$$\xi(d, i) = q_i \prod_{j=1}^{n} p_j, \quad i = 0,1,\ldots,h-1; \tag{5}$$

тут $$q_i = \begin{cases} k_i / m_h, & 0 < m_h < \infty, \\ 0, & m_h = 0, \\ P(f = i), & m_h = \infty; \end{cases} \qquad p_j = \begin{cases} k_{jid_j} / m_i, & 0 < m_i < \infty, \\ 0, & m_i = 0, \\ P(x_j = d_j | f = i), & m_i = \infty; \end{cases}$$

k_{jid_j} - кількість значень, що дорівнюють d_j, j-ї ознаки в j-му стовпчику матриці V_i; k_i - кількість значень цільової ознаки, що дорівнюють i, серед компонент вектора V_h. Значення $A(d)$ виберемо рівним найменшому числу s із множини $\{0,1,\ldots,h-1\}$ такому, що:

$$\xi(d, s) \ge \xi(d, i), \quad i = 0,1,\ldots,h-1. \tag{6}$$

Процедуру розпізнавання (5)-(6) позначимо Q_1.

Величини $\xi(d, i) / \sum_{j=0}^{h-1} \xi(d, j)$ є наближеними значеннями ймовірностей $P(f = i | x_1 = d_1, \ldots, x_n = d_n)$, обчислених за допомогою теореми Байєса, тому процедуру розпізнавання Q_1 назвемо байєсівською.

Справедливі такі дві теореми.

Теорема 1. Існує абсолютна константа $a_0 < \infty$ така, що виконується нерівність $\nu(Q_1, C) \le \min(1, a_0 \sqrt{gn / m_0 + h / m_h})$.

Теорема 2. Існує абсолютна константа $a_1 > 0$ така, що виконується таке. Які б не були натуральні числа g, h, n і цілі числа $m_0, m_1, \ldots, m_h$, що задовольняють нерівностям $g \ge 2$, $2 \le h \le 2g^n$, $0 \le m_0 \le m_1 \le \ldots \le m_{h-1}$, $m_h \ge 0$, та процедура розпізнавання Q, у класі C існує такий розподіл імовірностей P, що виконується нерівність $\nu(Q, P) \ge a_1 \min(1, \sqrt{gn / m_0 + h / m_h})$.

Доведення теорем міститься в роботах [1,2]. З теореми 2 випливає, що складність $\mu(C)$ класу задач C не менша, ніж число $a_1 \min(1, \sqrt{gn / m_0 + h / m_h})$. Похибка $\nu(Q_1, C)$ відрізняється від

складності $\mu(C)$ класу задач C не більше, ніж в абсолютну константу разів, оскільки $\nu(Q_1, C) / \mu(C) \le \max(1, a_0) / a_1$. У такому розумінні байєсівська процедура розпізнавання Q_1 є субоптимальною. Таким чином, встановлена (з точністю до абсолютної мультиплікативної константи) складність класу задач C.

Перейдемо до розгляду процедури розділяючої гіперплощини за умов $g = h = 2$ (тобто коли всі ознаки приймають значення з множини $\{0,1\}$), $0 < m_0 \le m_1 \le \infty$, $0 < m_2 \le \infty$. Позначимо $C' = C(g = 2, h = 2, m^{(2)}, n)$, $d = (d_1, \ldots, d_n)$ - булів вектор.

Байєсівська процедура Q_1 на класі C' стає такою. Вектор d відноситься до класу об'єктів 0, якщо виконується нерівність

$$q_0 \prod_{j=1}^{n} p_{j0d_j} \ge q_1 \prod_{j=1}^{n} p_{j1d_j}, \tag{7}$$

де $q_i = \begin{cases} k_i / m_2, & m_2 < \infty, \\ P(f = i), & m_2 = \infty, \end{cases}$ $p_{jid_j} = \begin{cases} k_{jid_j} / m_i, & m_i < \infty, \\ P(x_j = d_j | f = i), & m_i = \infty, \end{cases}$ $i = 0,1;$ $j = 1,2,\ldots,n.$ Якщо (7) не виконується, то вектор d відноситься до класу об'єктів 1.

Процедура роздільної гіперплощини (позначимо її R) полягає в наступному. Якщо виконується нерівність $\sum_{j=1}^{n} \alpha_j d_j + \alpha_0 \ge 0$, то вектор d відноситься до класу об'єктів 0, інакше - до класу об'єктів 1; тут $\alpha_0, \alpha_1, \ldots, \alpha_n$ - дійсні числа.

Позначимо $J_i = \{j: 0 < p_{ji1} < 1\}$, $i = 0,1$; $I = \{i: 0 < q_i < 1\}$;

$t = \max\{\max_{i\in\{0,1\}} \max_{j\in J_i} \max_{s\in\{0,1\}} |\ln p_{jis}|, \max_{i\in I} |\ln q_i|\}$ (тут $\max_{k\in\varnothing} r_k = 0$), $t_0 = (n+1)t + 1$, $t_1 = (n+1)t_0 + 1$.

Запишемо нерівність процедури роздільної гіперплощини в наступному вигляді

$$\tau_0(q_0) + \sum_{j=1}^{n} [\tau_{0j}(p_{j01}) d_j + \tau_{0j}(p_{j00})(1 - d_j)] \ge \tau_1(q_1) + \sum_{j=1}^{n} [\tau_{1j}(p_{j11}) d_j + \tau_{1j}(p_{j10})(1 - d_j)], \tag{8}$$

де $\tau_i(z) = \tau_{ij}(z) = \begin{cases} \ln z, & z > 0, \\ -t_i, & z = 0, \end{cases}$ $i = 0,1;$ $j = 1,2,\ldots,n.$

Неважко довести, що нерівності (7) та (8) еквівалентні, тобто для будь-якого булівського вектора d обидві нерівності або одночасно виконуються, або одночасно не виконуються. Доведемо це у випадку, коли справедливі нерівності $0 < q_i < 1$, $0 < p_{ji1} < 1$, $i = 0,1$; $j = 1,2,\ldots,n$. Нерівність (8) приймає вигляд

$$\ln q_0 + \sum_{j=1}^{n} [(\ln p_{j01}) d_j + (\ln p_{j00})(1 - d_j)] \ge \ln q_1 + \sum_{j=1}^{n} [(\ln p_{j11}) d_j + (\ln p_{j10})(1 - d_j)],$$

або $\ln [q_0 \prod_{j=1}^{n} p_{j01}^{d_j} p_{j00}^{1-d_j}] \ge \ln [q_1 \prod_{j=1}^{n} p_{j11}^{d_j} p_{j10}^{1-d_j}]$. Але $p_{ji1}^{d_j} p_{ji0}^{1-d_j} = p_{jid_j}$; тому остання нерівність еквівалентна (7). У загальному випадку еквівалентність нерівностей (7) та (8) випливає з визначення чисел t_0, t_1.

Таким чином, $Q_1(V) \equiv R(V)$. Звідси випливає субоптимальність процедури R на класі C', а також те, що на класі C' процедури R і Q_1 мають однакову похибку $\nu(R, C') = \nu(Q_1, C') \le \min(1, a_2 \sqrt{n / m_0 + 1 / m_2})$, де $a_2 < \infty$ - абсолютна константа.

Можна довести, що за умов $g \ge 3, h \ge 2, n \ge 1, 0 \le m_i \le \infty, i = 0,1,2,$ виконується нерівність $\nu(R, C) \ge a_3$, де $a_3 > 0$ - абсолютна константа.

Література

1. Гупал А.М., Пашко С.В. Сложность задач классификации целочисленных объектов // Математические методы в компьютерных системах. - Киев: Ин-т кибернетики им. В.М. Глушкова НАН Украины, 1996, с. 4-22.
2. Сергиенко И.В., Гупал А.М., Пашко С.В. О сложности задач распознавания образов // Кибернетика, 1996, № 4, с. 70-88.

♦

Формування репрезентативної навчальної вибірки для систем контролю та діягностування

ВЛАДИСЛАВ МАРЧЕНКО, АНАТОЛІЙ КРАСНОПОЯСОВСЬКИЙ

Державний університет *244007 Суми, вул. Римського-Корсакова 2*
Тел.: (0542) 33-5055

Владислав Марченко, Анатолий Краснопоясовский. Формирование репрезентативной обучающей выборки для систем контроля и диагностирования.

Рассматривается определение в задачах распознавания образов минимальной длины обучающей выборки из условия получения приемлемых статистических погрешности и оперативности реализации алгоритма обучения. Получена зависимость статистической погрешности от числа испытаний и предложен алгоритм ее оценки методом динамических доверительных интервалов.

Методологічні та теоретичні основи оптимізації параметрів навчання. Навчальна вибірка має на практиці скінченну довжину, що обумовлює наявність статистичної похибки ε між імовірністю p_i та емпіричною частістю k_i/n находження значення і-го параметра об'єкту в полі допуску за n іспитів.

Верхня оцінка похибки ε в залежності від кількості іспитів n визначається за теоремою Муавра-Лапласа [1]:

$$P\left\{\left|\frac{k_i}{n}-p_i\right|\geq\varepsilon\right\}=P\left\{\left|\frac{k_i-np_i}{\sqrt{np_iq_i}}\right|-\frac{\varepsilon\sqrt{n}}{\sqrt{p_iq_i}}\geq 0\right\}=2\Phi\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{p_iq_i}}\right)\geq 2\Phi(-2\varepsilon\sqrt{n}) \quad , \tag{1}$$

де k_i - кількість подій, при яких маємо знаходження і-го параметру у полі допусків; $q_i=1-p_i$ - імовірність відсутності значення і-го параметру в полі допусків; Φ(...) - функція Лапласа.

Визначення мінімальної довжини навчальної вибірки зробимо за умови отримання допустимої статистичної похибки й оперативности роботи алгоритму визначення n_{min}. Ці вимоги є суперечливими, що обумовлює компромісний характер розв'язання цього завдання. Скористаємось методом динамічного довірчого інтервального оцінювання. Суть цього методу полягає в побудові після кожного випробування інтервалу, який включає саму імовірність p_i, що оцінюється, та знаходження і-го параметра в полі допусків з імовірністю 1-Q, де Q - рівень значення:

$$P\left\{\frac{k_i}{n}-\varepsilon_Q\leq p_i\leq\frac{k_i}{n}+\varepsilon_Q\right\}=1-Q \ . \tag{2}$$

Визначення верхньої оцінки похибки ε_Q в залежності від числа іспитів при заданому рівні значення Q здійснюється за співвідношенням:

$$2\Phi(-2\varepsilon_Q\sqrt{n})=Q \ . \tag{3}$$

Рівень значення може бути вибраний як будь-яке мале позитивне число (звичайно вибирають одне зі значень: 0.05; 0.01; 0.001).

З урахуванням властивості функції Лапласа Φ(x)=1-Φ(-x) перетворимо (3) до вигляду:

$$\Phi(2\varepsilon_Q\sqrt{n})=1-\frac{Q}{2} \ . \tag{4}$$

Оскільки при незмінному значенні функції Лапласа $\arg\left[\Phi(2\varepsilon_Q\sqrt{n})\right]=const$, то похибка ε_Q відповідно (4) змінюється в залежності від довжини навчальної вибірки n за гіперболічним законом.

Для Q=0.05 за таблицею функцій Лапласа, з урахуванням (4) для Φ(x)=0,975, знайдемо значення аргументу $x=2\varepsilon_Q\sqrt{n}=1.98$. Тоді

$$\varepsilon_Q=\frac{0.98}{\sqrt{n}}, \quad n>1. \tag{5}$$

На рис. 1а на основі (5) подано графік функції $\varepsilon_Q=f(n)$, де умовно виділено три області значень n, що відрізняються крутизною функції. Область I є забороненою областю для припинення іспитів, оскільки похибка перебільшує допустиму. Область III характеризується значними економічними затратами при малій швидкості зменшення похибки ε_Q. Область II є компромісною і, як видно на рисунку, охоплює інтервал приблизно з 40 до 80 іспитів. При різних значеннях Q графік функції $\varepsilon_Q=f(n)$ буде переміщуватися паралельно по вертикалі, не змінюючи свого вигляду.

Графічно довірчий інтервал можна створити, обчислюючи для кожного іспиту за виразом (5) значення похибки ε_Q й відкладаючи його зверху та знизу від графіка частости k_i/n. На рис. 1б - графік частости k_i/n , верхня $\underset{n}{Sup}\,\tilde{p}_i$ та нижня $\underset{n}{Inf}\,\tilde{p}_i$ межі довірчого інтервалу для оцінки імовірністи знаходження значення параметру, який контролюється, в полі допусків при змінюванні n та заданому рівні значення Q=0,05.

Для знаходження мінімального числа іспитів n_{min}, що гарантує прийнятну з практичних міркувань величину похибки й оперативність реалізації алгоритму обчислювання, необхідно знайти критерій припинення іспитів. Моментом припинення іспитів будемо вважати такий іспит n_{min}, при якому поточний довірчий інтервал накривається заданим інтервалом $[0.5\pm\Delta]$, де $|\Delta|<0.5$. Останній (правий) перетин заданого інтервалу з однією з меж довірчого інтервалу визначає це число n_{min}. Оскільки область довірчого інтервалу при збільшенні n, як показано на рис.1б, має вигляд коридору, що зменшується до частости k_i/n, то при $n>n_{min}$ можна гарантувати з імовірністю 1-Q, що похибка ε_Q не буде перебільшувати значення, отриманого при $n=n_{min}$.

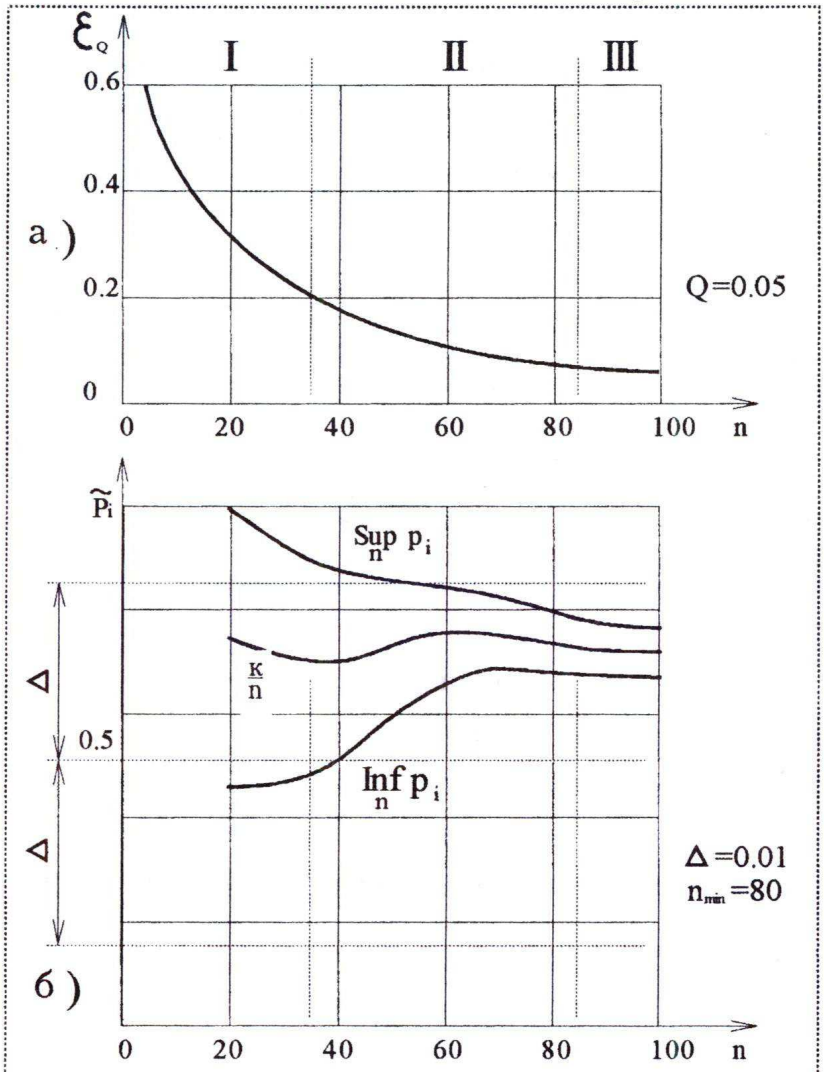


Рис. 1.
До визначення N_{min}: а) - залежність ε_Q від n; б) - динамічний довірчий інтервал

Зменшення статистичної похибки й часу реалізації алгоритму обчислювання є суперечливими. Тому вибрати n_{min} можна тільки на основі компромісного рішення. Так, аналіз рис.1 показує, що n_{min} краще вибирати у другій області при відсутності викидань довірчого інтервалу, що близькі до нуля або одиниці, при $n>n_{min}$.

Для багатьох практичних задач значення Δ вибирається з інтервалу [0,3;0,35].

В загальному випадку, треба будувати довірчі інтервали для усіх N контрольованих параметрів та вибирати n_{min} з умови:

$$n_{min}=\max_{N}{}_i(n_{min1},\dots n_{mini},\dots n_{minN}).$$

Однак при відповідному виборі системи допусків на контрольовані параметри та забезпеченні умов статистичної стійкости та однорідности проведення іспитів, можна вибирати n_{min} за довірчим інтервалом, побудованим для будь-якого одного параметру, що набагато знижує обчислювальну трудомісткість алгоритму.

Підготовчий етап. На підготовчому етапі виконуються такі кроки: вибір рівня значення; знаходження табличного значення аргумента функції Лапласа [2]; формування бінарної навчальної вибірки, яке здійснюється на практиці шляхом оцінки належності значень контрольованих параметрів своїм полям допусків.

Опис алгоритму визначення мінімальної довжини навчальної вибірки. Для ілюстрації методу розглянемо приклад визначення мінімальної довжини репрезентативної вибірки.

На рис. 2 наведена структурна схема алгоритму визначення n_{min} , що містить ітераційну процедуру пошуку оптимальної в розумінні Парето константи Δ.

Вхідними даними є: {X(I,J)} - масив навчальної вибірки, що належить до класу X_1^0; де J - число реалізацій (іспитів) $I=1,I_{MAX}$, I - число параметрів (ознак), $I=1,I_{MAX}$; H, SH - величини кроків зміни константи Δ на 0.1 і 0.01 відповідно; DL - константа Δ; ND - мінімально допустимий

іспит, котрий вибирається на межі областей I та II; C - таблична константа, що дорівнює половині значення аргумента функції Лапласа при заданому рівні значення.

Блок 4 обчислює при кожному іспиті суму одиниць - число подій, що відповідають знаходженню значення параметра у полі контрольних допусків. Блок 5 перевіряє виконання умови накриття поточним довірчим інтервалом заданого інтервалу [0.5 + DL].

Вихідними даними є: N_{MIN} - значення n_{min} для кожного стовпця масиву {X(I,J)}; NM - максимальне значення n_{min}.

Прикінцеві положення. Наведений алгоритм визначення репрезентативної навчальної вибірки реалізовано при розробці програмного забезпечення автоматизованої системи контролю та керування газоперекачувальною станцією ГПА-Ц-6.3А на алгоритмічній мові C++.

Висновки. Вперше запропоновано непараметричний метод визначення мінімальної довжини репрезентативної навчальної вибірки за умови отримання прийнятних статистичної похибки та оперативности реалізації алгоритму навчання системи контролю та діягностування.

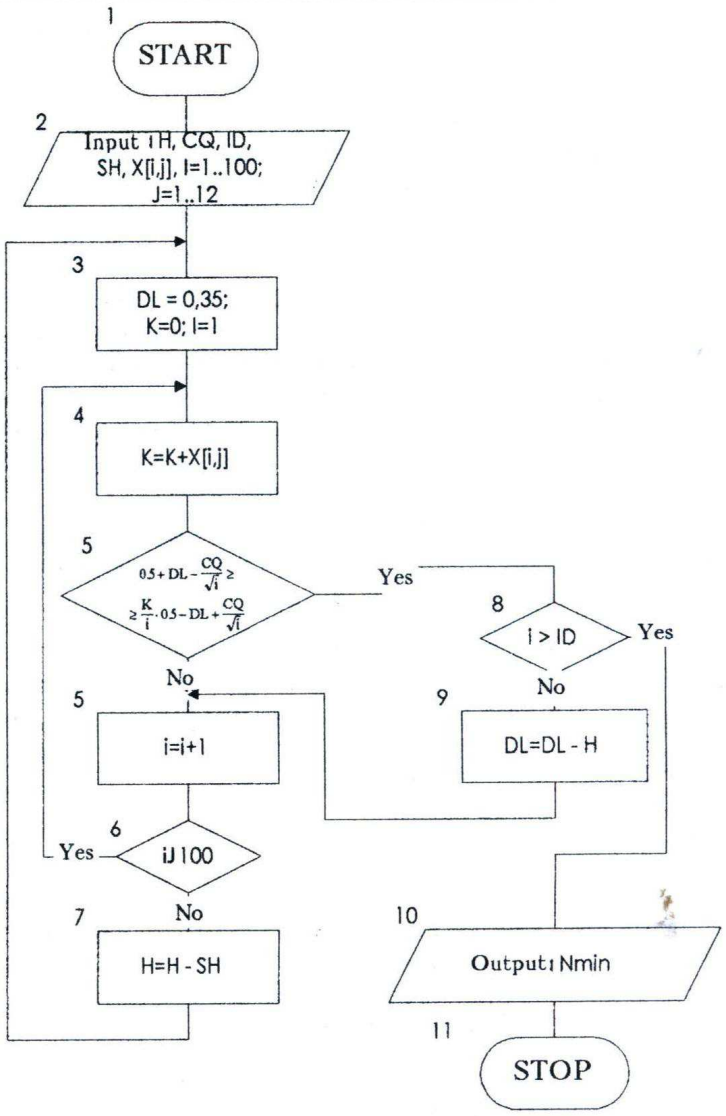


Рис. 2. Структурна схема алгоритму визначення n_{min}

Практична реалізація алгоритму визначення мінімальної довжини репрезентативної вибірки здійснюється в рамках алгоритму навчання системи, наприклад, за методом функціонально-статистичних випробувань як в умовах імітаційних і натурних випробувань, так і безпосередньо при функціонуванні об'єкту діягностування.

Література

1. Тутубалин В.П. Теория вероятностей. - Москва: МГУ, 1973, 272 с.
2. Большев, Смирнов. Таблицы математической статистики. - Москва: Наука.

♦

Метод достатніх наближень в проблемі пошуку емпіричних закономірностей

ВАЛЕРІЙ СУШКО

Інститут кібернетики НАН *252022 Київ, просп. Академіка Глушкова 40*

Тел.: (044) 266-4534

Валерий Сушко. Метод достаточных приближений в проблеме поиска эмпирических закономерностей.

В докладе рассматривается проблема поиска эмпирических закономерностей по выборкам данных ограниченного объема. Главные вопросы, которые возникают при решении этой проблемы, состоят в определении наиболее существенных характеристик или свойств изучаемого явления, а также в выборе правильного соотношения между "сложностью" класса функций, в котором ищется восстанавливающая функция, и объемом имеющихся эмпирических данных. В предлагаемом методе достаточных приближений

задача оценивания функции регрессии рассматривается как задача обучения распознаванию образов. Исследователю, выступающему в роли "учителя", предоставляется возможность при формировании образов учесть свои знания о природе изучаемого явления, а также формализовать требования к поведению восстанавливающей функции относительно имеющихся в выборке наблюдений значений зависимой переменной в виде уклонений, достаточных для требуемого приближения. Причем, уклонения могут быть заданы в виде произвольной функции или функций от переменных, описывающих решаемую задачу. Для решения задачи обучения распознаванию сформированных исследователем образов могут быть использованы любые алгоритмы обучения, гарантирующие качество и надежность работы на новых данных получаемых решений.

Задача пошуку емпіричних закономірностей полягає в тому, щоб знайти наближення до функціональної залежності, властивості якої відображені у вибірці прикладів. Найбільш розповсюдженою гіпотезою породження емпіричних даних є статистична гіпотеза, в основі якої наша віра в те, що емпіричні вибірки породжуються випадково і незалежно у відповідності з деякою об'єктивно існуючою ймовірнісною закономірністю. В цьому випадку такі відомі постановки як задача навчання розпізнаванню образів, оцінювання регресії, інтерпретації результатів непрямих експериментів можуть бути зведені до схеми мінімізації середнього ризику за даними експерименту.

Нехай вектор $Z \in \boldsymbol{Z}$ утворений парою X, y, де y - скалярна величина, яка може приймати значення із множини Y, а $X \in \boldsymbol{X}$ - n-вимірний вектор. На $\boldsymbol{Z}$ визначена ймовірнісна міра $P(Z)$ та задана множина вимірюваних відносно $P(Z)$ функцій $Q(Z, \alpha)$, які визначають величину втрат при появі вектора Z. Тут α - деякий параметр або вектор параметрів, котрий визначає конкретну функцію множини. Задача мінімізації середнього ризику полягає в пошуку на множині $Q(Z, \alpha)$ *фунції*, котра доставляє мінімум функціоналу

$$I(\alpha) = \int Q(Z,\alpha)\,dP(Z), \tag{1}$$

якщо ймовірнісна міра $P(Z)$ невідома, але задана випадкова і незалежна вибірка $Z_1, Z_2, \ldots, Z_l$, розподілена згідно $P(Z)$.

Якщо функція втрат задана у вигляді $Q(Z,\alpha) = (y - F(X,\alpha))^2$, то функціонал (1) може бути записаний у вигляді

$$I(\alpha) = \int (y - F(X,\alpha))^2 P(X,y)\,dX\,dy. \tag{2}$$

Мінімізацію функціонала (2), коли густина ймовірності невідома, але задана випадкова і незалежна вибірка пар

$$(X_1, y_1), (X_2, y_2), \ldots, (X_l, y_l) \tag{3}$$

довжиною l, прийнято називати задачею відновлення залежності за емпіричними даними. У випадку, коли величина y може приймати скінченне число значень, таку задачу називають задачею навчання розпізнаванню образів, а коли величина y може набувати яких завгодно числових значень із множини Y - задачею оцінювання регресії. Один із шляхів розв'язання задачі відновлення залежности за результатами спостережень, полягає в тому, що замість функціоналу середнього ризику (2) мінімізується функціонал емпіричного ризику

$$I_e(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(y_i - F(X_i,\alpha)), \tag{4}$$

побудований на основі вибірки (3). Але така заміна функціоналів в умовах обмеженого обсягу вибірки даних можлива не для будь-якого класу функцій втрат, а лише тоді, коли має місце рівномірне сходження середніх до їхніх математичних сподівань на всій множині функцій $Q(Z, \alpha)$ [1]. Тому, для того щоб отримати гарантований розвя'зок задачі мінімізації середнього ризику за емпіричними даними обмеженого обсягу, необхідно мати або абсолютну оцінку величини втрат, або ж оцінку відносної величини дисперсії. Для задач оцінювання регресії та інтерпретації результатів непрямих експериментів існування абсолютної оцінки є досить проблематичним. Такої оцінки може взагалі не існувати навіть для класу лінійних функцій, якщо на параметри α не накладено ніяких обмежень. Зовсім інша ситуація виникає при розв'язанні задачі навчання розпізнаванню образів, коли величина y може набувати значень 1 або 0 (випадок дихотомії) - тут абсолютна величина втрат ніколи не може бути більшою за одиницю. У зв'язку з чим виникла думка використати цю характерну особливість задачі навчання розпізнаванню образів в інших задачах пошуку емпіричних закономірностей.

В роботі [2], мабуть вперше, задача оцінювання регресії в класі лінійних функцій сформульована в термінах задачі навчання розпізнаванню двох образів. Перехід від задачі відновлення регресії до задачі навчання розпізнаванню образів здійснювався таким чином.

Нехай задана вибірка спостережень виду (3), за якою треба відновити функцію регресії в класі лінійних за параметрами функцій. При цьому, відновлювальна функція $F(X,\alpha_0)$ повинна бути такою, щоб для кожного об'єкта вибірки спостережень виконувалась така нерівність:

$$|y - F(X,\alpha_0)| \le 2\xi. \quad (5)$$

Тут ξ - невід'ємне число. Далі, кожному об'єктові вибірки (3) поставимо у відповідність два значення y: $y^{(1)} = y + \xi$ та $y^{(2)} = y - \xi$. Тоді вибірка (3) збільшиться вдвічі і може бути розділена підмножини V_1 і V_2, одна з яких містить елементи $(X, y^{(1)})$, а інша - елементи $(X, y^{(2)})$. Об'єкти підмножин V_1 і V_2 можна розглядати як об'єкти відповідно образів V_1^* і V_2^*. Якщо в процесі навчання вдасться безпомилково розділити ці образи за допомогою лінійної функції, то тим самим роздільна функція буде шуканою, яка оцінює функцію регресії і задовольняє умову (5).

Але більшість явищ матеріяльного світу мають нелінійний характер. А тому пошук функції, яка відновлює регресію в класі лінійних функцій може виявитись досить проблематичним і навіть практично недоцільним. При розв'язанні деяких задач вимога виконання умови (5) також є не зовсім доречною. Наведемо деякі приклади: 1) є достатнє число спостережень для побудови прогнозуючої моделі в умовах нормального функціювання об'єкту досліджень і всього декілька спостережень при виникненні екстремальних ситуацій (наприклад, аварійних). Треба побудувати загальну прогнозуючу модель об'єкта досліджень; 2) ціна похибки у прогнозі величини y в різних областях простору ознак має суттєве значення; 3) абсолютна похибка вимірювань величини y є змінною і залежить від умов проведення екперименту чи можливостей вимірювальної апаратури; 4) в певних точках простору ознак відновлювальна функція повинна набувати певних значень.

З огляду на приведені вище обставини пропонується такий підхід до розв'язання задачі відновлення регресії за вибірками емпіричних даних обмеженого обсягу. Досліднику пропонується формалізувати свої вимоги щодо точності прогнозу величини y та поведінки шуканої функції в певних областях і точках простору ознак. Це можна зробити в такий спосіб. Нехай величина ξ у виразі (5) є функцією y та X, тобто $\xi = f(y, X)$. Вибираючи певний тип функціонального зв'язку цих величин, дослідник має змогу у явному вигляді виписати умови, котрим повинна задовольняти функція, яка оцінює регресію. Далі, за вище приведеною схемою, формуються підмножини об'єктів V_1 і V_2 відповідно образів V_1^* і V_2^* та розв'зується задача навчання розпізнаванню образів. Величина досягнутого в процесі навчання ризику, тобто ймовірність неправильної класифікації за допомогою знайденої функції, котра розділяє образи V_1^* і V_2^* на навчальній вибірці довжиною $2l$ і водночас апроксимує регресію, може бути оцінена такою нерівністю [1]:

$$\int Q(Z,\alpha)dP(Z) \le \frac{1}{2l}\sum_{i=1}^{2l} Q(Z_i,\alpha) + \frac{h(\ln\frac{2l}{h}+1) - \ln\eta}{2l}\left(1 + \sqrt{1 + \frac{\frac{1}{2l}\sum_{i=1}^{2l} Q(Z,\alpha)}{h(\ln\frac{2l}{h}+1) - \ln\eta}}\right). \quad (6)$$

Ця нерівність з ймовірністю 1-η справедлива одночасно для всіх α на класі функцій скінченної ємності h. Якщо задати величину ймовірності неправильної класифікації, котру бажано досягти в процесі навчання, та припустити, що величина емпіричного ризику досягне нульового значення, то можна з урахуванням довжини навчальної вибірки визначити "складність" класу функцій, серед яких має рацію вести пошук функції, котра оцінює регресію і при цьому гарантовано задовольняє певні вимоги, які є достатніми для її практичного застосування. Зауважимо, що оцінка (6) одержана на основі стратегії мінімаксу втрат, яка передбачає випадок розв'язання задачі навчання розпізнаванню образів у найнесприятливіших умовах, а тому є надто завищеною. Цю обставину слід враховувати при розв'занні практичних задач.

Література

1. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. - Москва: Наука, 1979, 448 с.

2. Синтез пространств линейных зависимостей. // В.И. Васильев, Ю.И. Горелов, В.И. Сушко, В.В. Иголкин / Автоматика, 1990, № 4, с. 51-55.

♦

Оптимізація параметрів навчання за методом функціонально-статистичних випробувань

АНДРІЙ ЧЕРНИШ

Державний університет *244007 Суми, вул. Римського-Корсакова 2*
Тел.: (0542) 33-5055

Andrij Chernysh. Parameters' Optimization of Training's Process of Diagnostic System Using Method of Functional-Statistic Tests.

It's proved that system of the control admissions (SCA) has influence on a functional effectiveness of diagnosing's system (DS). Iterative algorithm of optimization SCA on attributes of recognition on information criterion of functional efficiency (CFE) is developed by unparametric method of automatic classification - method of functional-static tests (MFST) . The algorithm consists of two stages. Problem of algorithmic setup DS is solved at the first step. It is considered as a sequence of directed experiments during the training. At the second stage determination of the optimal control admissions in information sense on attributes recognition by the way iterative procedure of search extreme information CFE is executed, which is realized during training. The reduction scale is considered for use optimization of SCA in the general case .

Методологічні та теоретичні основи оптимізації параметрів навчання. При використанні системи розпізнавання (СР), що навчається, для розв'язання завдань технічної діягностики складних систем суттєвий вплив на достовірність розпізнавання має вибір системи контрольних допусків (СКД) на ознаки розпізнавання. Однак до теперішнього часу теоретичне розв'язання проблеми вибору СКД на ознаки розпізнавання все ще не отримано. Серед причин такого становища відзначимо дві: відсутність теорії оптимізації процесу навчання СР, що базується на загальному критерії функціональної ефективності (КФЕ); статистична невизначеність і неоднорідність вибірки значень ознак реального об'єкту діягностування (ОД), що обумовлює необхідність використання непараметричної математичної статистики.

В рамках непараметричного методу розпізнавання образів — методу функціонально-статистичних випробувань (МФСВ) розв'язання цієї проблеми досягається шляхом визначення інформаційної здатності СР в процесі її навчання в рамках алгоритму дискримінантного типологічного аналізу [1,2].

Основні завдання, які послідовно розв'язуються на етапі навчання в рамках МФСВ, такі: формування бінарної навчальної вибірки та еталонних векторів для заданого алфавіту класів розпізнавання; визначення мінімальної довжини репрезентативної навчальної вибірки; оцінка функціональної ефективности процесу навчання за інформаційним критерієм; оптимізація параметрів роздільних гіперповерхонь (РГП) для класів розпізнавання; оптимізація СКД за інформаційним КФЕ; оптимізація рівня селекції даних при формуванні еталонних векторів класів розпізнавання; прогнозування зміни технічного стану СР та моменту її перенавчання.

Центральне місце при оптимізації параметрів навчання СР в МФСВ займає обчислення інформаційного КФЕ, який має такий загальний вигляд:

$$E = \frac{H_0 - H(\gamma)}{H_0}; \qquad (1)$$

де H_0, $H(\gamma)$ - апріорна (безумовна) ентропія і умовна апостеріорна ентропія, що характеризує залишкову невизначеність рішення, що приймається.

При застосуванні відомої формули Байеса критерій (1) може бути виражений через точнісні характеристики СР [1].

Введемо для двохальтернативного рішення такі оцінки точнісних характеристик:

$$\alpha = \frac{K_1}{n^*}; \quad \beta = \frac{K_2}{n^*}; \quad D_1 = \frac{K_3}{n^*}; \quad D_2 = \frac{K_4}{n^*};$$

де α, β, D_1, D_2 - помилки першого та другого роду, перша та друга достовірність відповідно; K_1, K_3 - кількість явищ, що відбуваються при знаходженні виміряного значення ознаки розпізнавання (ОР) поза полем допусків і в полі допусків за умови знаходження істинного значення ознаки в полі допусків; K_2, K_4 - кількість явищ, що відбуваються при знаходженні виміряного значення ОР в полі допусків і поза полем допусків за умови знаходження істинного значення ознаки поза

полем допусків; n* - достатня кількість випробувань (спостережень), що визначає репрезентативність навчальної вибірки. Тоді критерій (1) для рівноймовірних явищ може мати вигляд:

$$E = 1 + \frac{1}{2}\left(\frac{K_1}{K_1+K_4}\log_2\frac{K_1}{K_1+K_4} + \frac{K_2}{K_2+K_3}\log_2\frac{K_2}{K_2+K_3} + \frac{K_3}{K_3+K_2}\log_2\frac{K_3}{K_3+K_2} + \frac{K_4}{K_4+K_1}\log_2\frac{K_4}{K_4+K_1}\right). \quad (2)$$

В загальному випадку оптимізація в інформаційному сенсі СКД потребує побудови наведеного поля допусків (ППД) [2]. Це дозволяє звести багатофакторне завдання експериментів до вибору спочатку на ППД оптимальних значень контрольних допусків за максимумом інформаційного КФЕ, а потім - до зворотнього відображення цих значень на поля допусків ознак ропізнавання.

Математичною моделлю поля допусків є відрізок числової прямої [a,b], a,b∈N, який називається шкалою. Як відомо, для двох шкал $[a_1,b_1]$, $[a_2,b_2]$ існує єдине лінійне відображення f: $[a_1,b_1] \to [a_2,b_2]$, яке називається перетворенням шкал. При цьому прямим відображенням є:

$$f(x) = \frac{b_2 - a_2}{b_1 - a_1}x + \frac{a_2 b_1 - a_1 b_2}{b_1 - a_1}; \quad (3)$$

а зворотнім — $$f^{-1}(x) = \frac{b_1 - a_1}{b_2 - a_2}x + \frac{a_1 b_2 - a_2 b_1}{b_2 - a_2}. \quad (4)$$

Підготовчий етап. На підготовчому етапі здійснюють: 1) градуювання шкал ОР та визначення ціни градації; 2) побудову ППД [2]; 3) вибір початкових значень нижнього та верхнього контрольних допусків на ППД; 4) вибір стратегії та напрямку пошуку екстремума КФЕ. На практиці достатньо обмежитись двома стратегіями: симетричною стратегією, при якій здійснюється покрокове симетричне відносно середини ППД зменшення або збільшення контрольного поля допусків до знаходження екстремуму КФЕ або асиметричною стратегією, при якій нижній та верхній допуски змінюються в одному напрямку з заданим кроком.

Опис алгоритму. Структурна схема алгоритму оптимизації контрольних допусків наведена на рис.1.

Вхідні дані: АН, АВ - нижній та верхній допуски ППД відповідно, АНК, АВК - початкові нижній та верхній допуски ППД відповідно, {A(L)}, {B(L)} - нижні та верхні експлуатаційні допуски на ОР відповідно.

Вихідні дані: CT[S] - тип обраної стратегії, значення інформаційного КФЕ для кожного кроку оптимізації, АНКО і АВКО - оптимальні значення нижнього та верхнього контрольних допусків на ППД відповідно, {АНКО(L)}, {АВКО(0)} - оптимальні значення нижнього та верхнього контрольних допусків на ОР.

Блок 9 обчислює значення АНК і АВК на кожному кроку оптимізції за формулами:

для першої стратегії:

АНК(S)=АНК(S-1)+K*S*H;

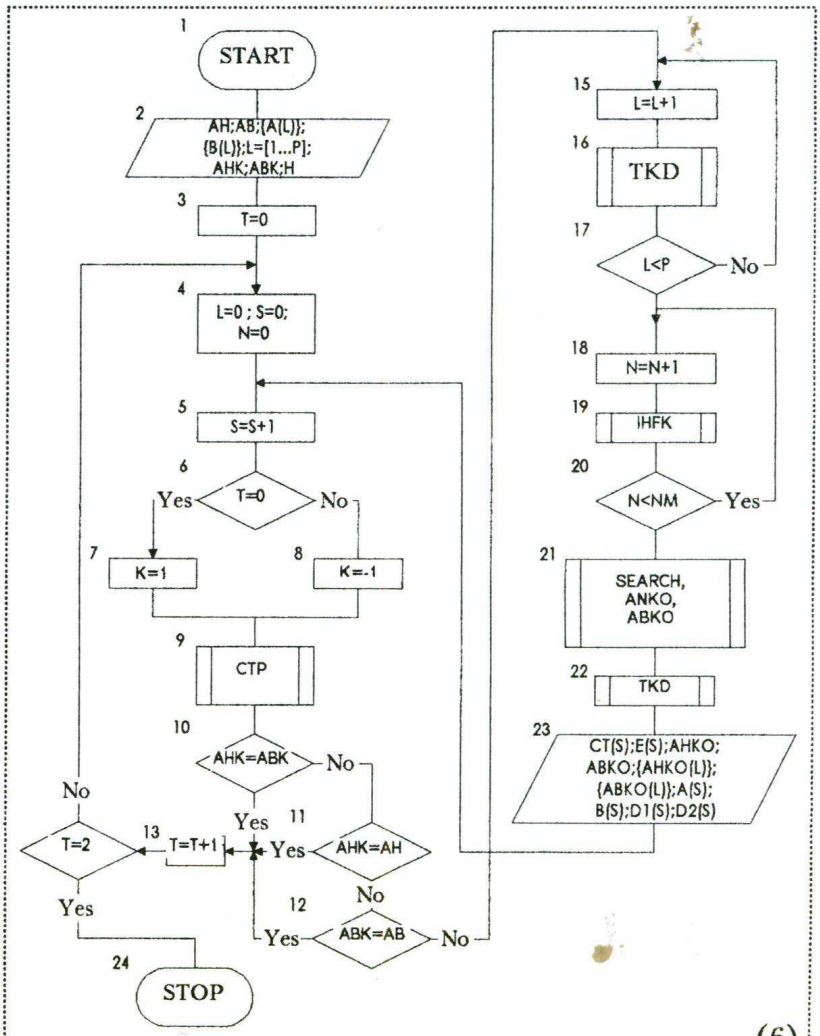
АВК(S)=АВК(S-1)-K*S*H ;

для другої стратегії:

АНК(S)=АНК(S-1)+K*S*H ;

АВК(S)=АВК(S-1)+K*S*H .

Блоки 10,11,12 проводять перевірку обмежень на застосування



(6)

Рис. 1. Структурна схема алгоритму оптимизації СКД.

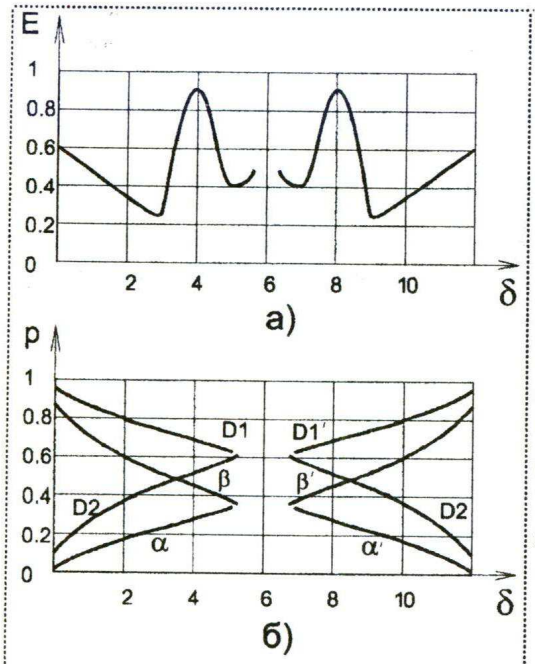


Рис. 2. До визначення оптимальної СКД на ОР:
а) залежність КФЕ від поля допусків;
б) залежність точнісних характеристик від поля допусків.

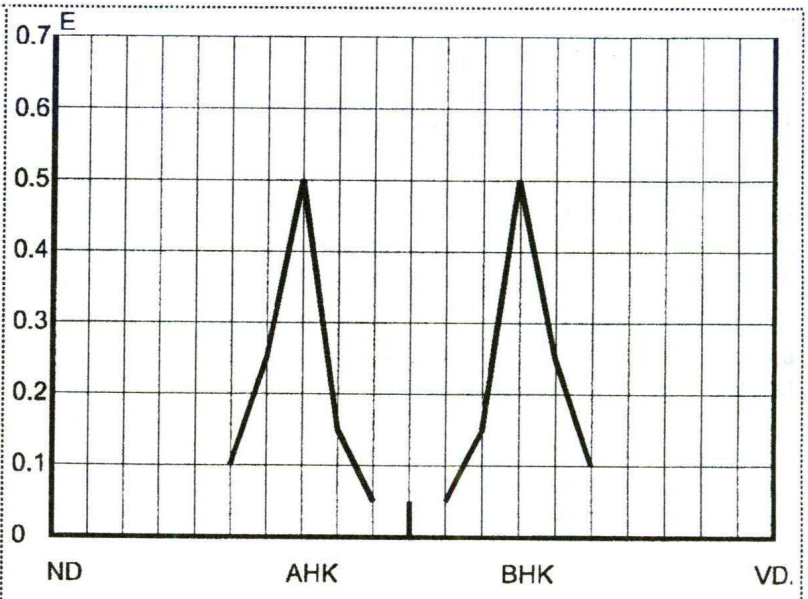


Рис. 3. Результати реалізації алгоритму.

стратегій. Таким чином блок 10 може бути виконаний, якщо обрана перша стратегія. При наявності обмежень блок 13 збільшує значення T на 1, що є командою для зміни напрямку стратегій на протилежний. Якщо у процесі виконання алгоритму T=2 (програма не знайшла екстремуму), блок 14 завершує роботу алгоритма. При відсутності обмежень (блоки 10,11,12) виконується підпрограма TKD, яка покроково обчислює експлуатаційні допуски шляхом зворотнього відображення AHK(S) і ABK(S) на поле допусків ОР за формулою (4). Потім виконується програма IHFK, яка обчислює значення КФЕ. Блок 21 знаходить екстремум критерію $E^{*}(S)$ та екстремальні значення AHKO и ABKO. Після зворотнього відображення координат AHKO и ABKO на поля допусків ОР (блок 22) йде видача результатів.

Прикінцеві положення. Наведений алгоритм реалізовано при розробленні програмного забезпечення автоматизованої системи контролю та керування газоперекачувальної станції ГПА-Ц-6.3А на алгоритмічній мові С++ для діягностування технічного стану авіадвигуна Д-336, який використовується як привід компресора. Для визначеня технічного стану ОД задіяно 11 аналогових та 9 сигнальних давачів. Діягностувалося шість класів стану ОД, у тому числі і виникнення помпажу двигуна.

На рис. 2 наведено пошук оптимальних контрольних допусків (мал.2а) та розподіл точнісних характеристик в ППД. На рис. 3 наведені оптимальні контрольні допуски ППД для першого класу розпізнавання, який фіксує працездатний стан ОД.

Висновки.

1. Доведено залежність СКД на ОР і точнісних характеристик СР від інформаційного КФЕ в його робочій області.

2. В загальному випадку для розв'язання задачі оптимізації СКД в інформаційному сенсі доцільною є побудова ППД з урахуваням закономірності впливу ОР на функціональну ефективність СР.

Література

1. Краснопоясовский А.С., Сергеев В.П., Проценко И.Г. Оценка функциональной эффективности автоматизированных систем контроля и управления // Электронная техника. Сер.9 Экономика и системы управления, 1988, Вып. 2, с. 30-36.
2. Краснопоясовский А.С., Калюжная С.А. О выборе обобщенной шкалы для входных переменных при многофакторном эксперименте // Автоматизированные системы управления. - Харьков, Харьк. авиац. ин-т, 1984, Вып. 5, с. 114-118.

♦♦♦