УДК 001.891:65.011.56

Повхан І.Ф.
ДВНЗ «Ужгородський національний університет»

ПРОБЛЕМА ФУНКЦІОНАЛЬНОЇ ОЦІНКИ НАВЧАЛЬНОЇ ВИБІРКИ В ЗАДАЧАХ РОЗПІЗНАВАННЯ ДИСКРЕТНИХ ОБ'ЄКТІВ

Статтю присвячено актуальній проблемі оцінки навчальної вибірки на основі функціоналів. Правильна оцінка вибірки дозволяє побудувати ефективне правило класифікації, що забезпечує просте та повне розпізнавання дискретних об'єктів.

Ключові слова: *розпізнавання дискретних об'єктів, важливість аргументів, функція розпізнавання.*

Постановка проблеми. У процесі розв'язання практичних завдань із розпізнавання дискретних об'єктів часто доводиться стикатися з такими моментами:

– кодування вихідної інформації;
– перевірка та оцінка навчальної вибірки;
– знаходження важливості ознак і груп ознак;
– мінімізація вихідного опису;
– визначення якості кодування вихідної інформації;
– побудова множини алгоритмів розпізнавання і знаходження коректного алгоритму для певної задачі.

Указані завдання доводиться вирішувати в рамках деяких обмежень: час; пам'ять; ефективність та інше.

Аналіз останніх досліджень і публікацій. Дослідженням порушеної проблеми займалися такі науковці, як І.Ф. Повхан, Ю.А. Василенко, Е.Ю. Василенко та інші.

Постановка завдання. Метою статті є знайти шляхи розв'язання проблеми функціональної оцінки навчальної вибірки в завданнях із розпізнавання дискретних об'єктів.

Виклад основного матеріалу дослідження. *Кодування вихідної інформації.* Якщо об'єкти із множини G є деякими точками n-вимірного простору, то завжди, за умови вихідної різниці між класами, можемо розбити n-вимірний простір, обмежений можливими значеннями ознак, гіперплощинами, на n-мірні гіперпаралелепіпеди так, щоб точки з різних класів не попадали в один і той самий паралелепіпед. Після цього кожному гіперпаралелепіпеду в n-мірному просторі можливо поставити у відповідність двійковий код. Ці двійкові коди, залежно від того, куди попадають об'єкти із множини G, і визначають набір бінарних ознак завдання розпізнавання.

Поняття навчальної вибірки (далі – НВ), зокрема проблема її функціональної оцінки, як відомо, є одним із найважливіших у теорії розпізнавання об'єктів. На його основі запропоновано різноманітні алгоритми розпізнавання, а також визначено відносну важливість ознак у дискретному наборі [1].

У статті розглядаються деякі функціонали, що розраховуються безпосередньо за НВ, які дозволяють вирішувати доволі широке коло завдань стосовно розпізнавання об'єктів та технічної діагностики. Як показано далі, такі функціонали відносяться до загальних та більш широких понять, ніж поняття НВ.

Нехай НВ задана в такому вигляді:

$$(x_1, f_R(x_1)), \ldots, (x_M, f_R(x_M)), \tag{1}$$

де $x_i \in G$ (G – деяка множина);

$f_R(x_i) \in \{0, 1, 2, \ldots, k-1\}, (i = 1, 2, \ldots, M)$, а, відповідно, $f_R(x_i) = l, (0 \le l \le k-1)$ означає, що $x_i \in H_l, H_l \subset G$. Тут f_R – деяка гранична функція, яка задає розбиття R множини G, що складається з підмножин (образів) $H_0, H_1, H_2, \ldots, H_{k-1}$.

Таким чином, НВ – це сукупність (точніше послідовність) деяких наборів, причому кожний набір – це комплекс значень певних ознак і функцій у цьому наборі. Сукупність значень ознак – це зображення, а значення функції відносить це зображення до відповідного об'єкта [2].

Визначимо наступний функціонал за НВ з огляду на (1):

$$W(x_i) = \sum_{j=0}^{k_i-1} \frac{b_j}{h} \rho_j, \quad \rho_j = \max_{0 \le m \le k} \frac{q_j^m}{b_j}, \tag{2}$$

де b_j – кількість наборів НВ, у яких $x_i = j, (j = 0, \ldots, k_i - 1)$;

q_j^m – кількість наборів НВ, у яких $x_i = j$, а $f_R(x) = m, (j = 0, 1, \ldots, k_i - 1), (m = 0, 1, \ldots, k-1)$;

h – кількість усіх наборів НВ.

При $k_1 = k_2 =, \ldots, k_n = k = 2$ (бінарний випадок) функціонал (2) набуває більш простого вигляду:

$$W(x_i) = \frac{1}{h}(\max_m q_0^m + \max_m q_1^m), \tag{3}$$

де q_0^m, q_1^m, h має той самий сенс, що і в (2).

Функціонал (2) отримаємо з таких міркувань (3). Величину $\frac{q_j^m}{b_i}$ можливо інтерпретувати як імовірність того, що функція f_R з (1) набуває значення $m, (0 \le m \le k-1)$ за умов, що значення ознаки x_i дорівнює $j, (0 \le j \le k_i - 1)$.

Величина ρ_j є максимальною з указаних імовірностей. Варто зазначити, що величина ρ_j – це та інформація, яку можливо отримати про значення функції $f_R(x)$, знаючи, що в наборі x значення ознаки x_i дорівнює j.

Величина $W(x_i)$ визначається функціоналом (2), характеризує ту інформацію, яку можливо отримати про функцію $f_R(x)$, знаючи значення ознаки x_i в наборі x.

Відтак ознака x_i, *(1≤i≤n)*, для якої вказана інформація є найбільшою, є найважливішою стосовно $f_R(x)$, тобто функції, яка розбиває НВ на об'єкти.

Узагальненням функціоналу (2) (він залежить від однієї функції (ознаки) $f_i(x_1,...,x_n) = x_i$) є наступний функціонал, який залежить від групи ознак (функцій):

$$W(x_1,...,x_\gamma) = \sum_{\Delta \in Q} \frac{B_\Delta}{h} \delta_\Delta , \qquad (4)$$

$$\delta_\Delta = \max_{0 \le m \le k-1} \frac{B_\Delta^m}{B_\Delta} ,$$

де $\Delta = t_1 t_2 .... t_\gamma, (0 \le t_j \le k_j - 1), (j = 1,2,....,\gamma)$ – довільний набір ознак;

B_Δ – кількість усіх наборів НВ, для яких виконується співвідношення $x_j = t_j, (j = 1,2,...,\gamma)$;

B_Δ^m– кількість усіх наборів НВ, для яких виконується співвідношення $x_j = t_j, (j = 1,2,...,\gamma)$ та $f_R(\mathrm{x}_1, ..., \ \mathrm{x_n}) = m, (0 \le m \le k-1)$;

Q – множина всіх наборів ознак $\mathrm{x}_1, ..., \mathrm{x}_\gamma$.

Обґрунтування функціоналу (4) аналогічне обґрунтуванню функціоналу (2). Функціонал (4) визначає важливість групи ознак $x_1, ..., x_\gamma$ стосовно $f_R(x)$.

Практично для обчислення функціоналу (4) зручніше використовувати таку формулу:

$$W(x_1, ..., x_\gamma) = \frac{1}{h} \sum_{\Delta \in Q} \max B_\Delta^m, (m = 0,1) . \qquad (5)$$

Таким чином, загалом

$$W(x_{i_1}, ..., x_{i_\gamma}), (1 \le \gamma \le n) . \qquad (6)$$

При $\gamma = 1$ визначає важливість ознаки $x_i, (i = 1,2,...,n)$ стосовно $f_R(x)$, а у випадку $1 < \gamma \le n$ – важливість групи ознак $x_{i_1}, ..., x_{i_\gamma}$ щодо $f_R(x)$.

Очевидно, що $\frac{1}{k} \le W(x_{i_1}, ..., x_{i_\gamma}) \le 1, (1 \le \gamma \le n)$.

Функціонал (6) розраховується безпосередньо за НВ, він є загальним та ширшим поняттям порівняно з поняттям тесту.

Дійсно, якщо числова величина вказаного функціоналу дорівнює одиниці, то група ознак, від яких він залежить, є тестом [2; 3]. Ця особливість функціоналу (6) дає можливість запропонувати на його основі алгоритм знаходження всіх тупикових тестів НВ. Таким чином, поняття тесту можливо розглядати як деякий частковий випадок функціоналу (6).

Водночас при $W < 1$ функціонал (6) уможливлює безпосередньо за НВ визначити важливість ознак (однієї або групи) стосовно функції, яка задає розбиття НВ на об'єкти [4].

Загальна схема алгоритму. Наведемо один із можливих машинних алгоритмів знаходження важливості ознак і груп ознак за допомогою функціоналу (5) і його можливе використання для вирішення широкого кола завдань із розпізнавання дискретних об'єктів.

Особливістю згаданого алгоритму є випадковий вибір ознак та груп ознак для оцінки їх інформативності і тому подібне.

Опишемо призначення деяких масивів і змінних:

N – кількість ознак;

M – кількість об'єктів у G;

$N2$ – кількість класів;

$HB[1..N,1..M]$ – масив для зберігання НВ у пам'яті ЕОМ;

$KOB[1..N2]$ – масив для зберігання порядкових номерів останніх об'єктів кожного класу;

$HH[1..MX]$ – масив для зберігання поліномів за модулем 2, необхідних для формування випадкових бітових рядків B, одиниці яких задають номери ознак. Для останніх надалі знаходиться їх значення за допомогою функціоналу (5);

MX – кількість поліномів (кожний поліном ступеня N задається у вигляді бітового рядку).

Масиви $GRP[1..N], KOB2[1..N2], IOB[1..M]$ – допоміжні масиви для роботи поточного алгоритму.

Змінній W надається значення груп ознак, номери яких зберігаються в масиві $GRP[1..N]$.

Опис алгоритму.

Крок 1. Ввести вихідні значення: M, N, $N2$, $HB[1..N,1..M]$, $KOB[1..N2]$, $HH[1..MX]$, $IJB = 0$, $M2 = 2 * N$, N5=64.

Крок 2. Формування випадкового бітового рядку $B.IJB = IJB + 1$.

Крок 3. $\mathrm{GRP} = 0$, $\mathrm{KI} = 0$, $\mathrm{IWJ} = 0$, $\mathrm{N4} = 0$, $\mathrm{N3} = \mathrm{N5{-}N} + 1$.

Крок 4. $\mathrm{I} = \mathrm{N3{-}1}$.

Крок 5. $\mathrm{I} = \mathrm{I} + 1$, $\mathrm{N4} = \mathrm{N4} + 1$.

Крок 6. Якщо 1-ий біт рядка В дорівнює I, то $\mathrm{KI} = \mathrm{KI} + 1$, $\mathrm{GRP[KI]} = \mathrm{N4}$; якщо $\mathrm{I} < \mathrm{N3} + \mathrm{N}$, то перейти на Крок 5.

Крок 7. $L2 = 0$, $IOB = 0$, $I = 0$.

Крок 8. $I = I + 1$. Якщо $IOB[I] = 1$, то перейти на Крок 17.

Крок 9. $I1 = 1$, $KOB2 = 0$, $J = 0$.

Крок 10. $J = J + 1$; якщо $J = KOB[I1]$, то $I1 = I + 1$; якщо $IOB[J] = 1$, то перейти на Крок 14.

Крок 11. $K = K + 1$.

Якщо $HB[J,GRP[K]] > HB[I,GRP[K]]$, то перейти на Крок 14.

Крок 12. Якщо $K < KI$, то перейти на Крок 14.

Крок 13. $IOB[J] = 1$, $KOB[I1] = KOB2[I1] + 1$.

Крок 14. Якщо $J < M$, то перейти на Крок 10.

Крок 15. Знайти максимальне значення в масиві $KOB2$.

Крок 16. $IWJ = IWJ + IW$.

Крок 17. Якщо $I < M$, то перейти на Крок 8.

Крок 18. $W = IWJ/M$. Роздрукувати значення GRP і W.

Крок 19. Якщо $IBJ < M2$, то перейти на Крок 2.

Крок 20. Закінчити роботу

Зазначимо, що за незначної модифікації описаний алгоритм можливо використовувати для широкого кола завдань із розпізнавання об'єктів (знаходження всіх тестів у G, тестів фіксованої довжини, наперед заданої кількості тестів, деяких тестів за обмежений час, розв'язання завдань із геологічного прогнозування, визначення якості кодування інформації, знаходження ознак для розпізнавання дискретних об'єктів, вирішення деяких соціологічних завдань та інше).

Висновки. Перерахуємо тепер основні завдання, пов'язані з обробкою НВ і які можливо вирішити на основі застосування функціоналу (6), а також визначимо деякі його особливості.

Як уже зазначалося вище, на основі формули (6) запропоновано метод знаходження всіх тестів НВ (зокрема, тупикових та мінімальних).

Функціонал (6) дозволяє безпосередньо за НВ вирахувати значення ознаки (групи ознак) стосовно функції розбиття $f_R(x)$ НВ на об'єкти.

Формула (6) дозволяє легко знаходити значення ознак (або групи ознак) в будь-якому випадку (багатозначному) стосовно функції, яка задає розбиття НВ.

Формула (6) дає можливість досить легко знаходити тести (тупикові або мінімальні), а також значення окремих ознак (групи ознак) у випадку, коли НВ складається із двох і більше образів.

Чисельна величина $W(x_{i_1},....,x_{i_\gamma}),(1 \le \gamma \le n)$, як бачимо, є частиною об'єктів НВ, які правильно розпізнаються групою ознак $x_{i_1},....,x_{i_\gamma}$.

У деяких випадках дуже важко (неможливо) знайти тест (наприклад, прогнозування погоди). Очевидно, у такому разі слід спробувати знайти таку сукупність ознак, яка б дозволяла найкращим чином відрізняти одні об'єкти від інших. Указану групу ознак, на відміну від тесту, будемо називати квазітестом. Формула (6) дозволяє знаходити квазітести безпосередньо за заданою НВ.

Функціонал (6) легко узагальнити на випадок, якщо об'єкти НВ є наборами дійсних чисел, тобто складаються з нескінченних множин. У такому разі можливо ввести поняття тесту, аналогічне наведеному, а також тупикового (мінімального) тесту. Однак під ознакою тут розуміється вже не функція вибору аргументу (як тестових методів), а більш загальне функціональне перетворення об'єктів (наборів дійсних чисел) НВ.

Формула (6) дозволяє ефективно знаходити групи ознак (або одну ознаку), які розпізнають образи НВ з деякою заданою наперед точністю. Використовуючи її, можемо завжди знайти найважливішу групу ознак (стосовно функції, яка задає розбиття НВ) за їх фіксованої кількості у групі.

Якщо в НВ не існує тесту (крім тривіального, тобто такого, який містить всі ознаки), то функціонал (6) дозволяє знайти мінімальну групу ознак із максимальним значенням стосовно функції, яка задає розбиття НВ на об'єкти.

Значення ознак, а також тести (тупикові та мінімальні) за допомогою (6) знаходяться безпосередньо за НВ (тобто без додаткових перетворень НВ).

Особливо важливою якістю функціоналу (6) є те, що обчислення значення ознак (групи ознак) та знаходження тестів за (6) здійснюється одночасно, паралельно, що дозволяє виділити групу ознак, для яких $W(x_{i_1},....,x_{i_\gamma}) < 1$, як деякий квазітест, що розпізнає об'єкти НВ із фіксованою точністю, рівною W. У відомих тестових методах ці завдання (визначення значення і тестовості ознак) вирішуються послідовно, що збільшує машинний час виконання вказаних процедур.

Значення окремих ознак (групи ознак), яке розраховується за (6), як уже було сказано вище, чисельно співпадає із частиною об'єктів, що правильно розпізнаються (за допомогою ознак) – фактично це інтерпретація функціоналу (6) стосовно НВ.

На основі функціоналів (2), (4) можемо запропонувати ефективний метод (для багатьох видів НВ) знаходження всіх тупикових тестів НВ.

Функціонал (6) можливо застосовувати для знаходження контрольних та діагностичних тестів (у дискретних та аналогових пристроях) під час пошуку технічних несправностей.

Можемо навести приклади таких НВ, для яких на основі побудови всіх тупикових тестів неможливо визначити значення ознак. Функціонал (6) дозволяє це зробити.

Функціонал (6) можливо застосовувати для обробки НВ, яка складається з одного об'єкта (однієї таблиці), тобто знаходити тести (тупикові, мінімальні), визначати значення ознак стосовно рядків таблиці та інше. Наприклад, таке завдання досить часто зустрічається в геологічних дослідженнях.

Функціонал (6) можливо застосовувати до НВ, які не зведені до допустимого вигляду.

Для обчислення функціоналу (6) не обов'язково вводити НВ у пам'ять ЕОМ. Є можливість (за допомогою деяких фіксованих алгоритмів) обчислювати (2) та (4) без занесення в пам'ять машини.

Функціонал (6) має сенс також для всіх ознак $x_1,....,x_n$. Дійсно, нехай G – довільна множина, а $f_R(x)$ – деяка функція, яка задана на цій множині. Нехай $x_1,.....,x_2$ є деякими ознаками, тобто функціями вигляду $G \to \{0,1,.....,k_i - 1\}, (i = 1,2,.....,n)$. Систему ознак $x_1,....,x_n$ будемо називати повною щодо $f_R(x)$, якщо виконується така умова: якщо $f_R(x) \uparrow f_R(y)$, де $x, y \in G$, то набір ознак $x_1,....,x_n$ відрізняється від набору ознак $y_1,....,y_n$. Очевидно, що величина $W(x_1,....,x_n)$ тоді і тільки тоді дорівнює одиниці (тобто група ознак $x_1,....,x_n$ є тестом), коли група ознак $x_1,....,x_n$ є повною стосовно функції f_R. Для НВ допустимого вигляду $W(x_1,....,x_n) = 1$.

Якщо ж група ознак $x_1,....,x_n$ неповна, то величина $W(x_1,....,x_n)$ строго менша одиниці. Таким чином, величина $W(x_1,....,x_n)$ слугує оцінкою повноти групи ознак $x_1,....,x_n$ стосовно функції $f_R(x)$.

На основі функціоналу (6) можливо запропонувати звичайну класифікацію НВ зі скінчених множин: для кожного n, $(n = 1,2,...)$ буде лише 2^n різних типів НВ, відмінних один від одного хоч одним тупиковим тестом.

Функціонал (6) можливо застосовувати для розв'язання ігрових завдань (наприклад, гра «Морський бій»), а також для знаходження мінімальних диз'юнктивних нормальних форм, тобто для вирішення завдань, які зводяться до знаходження тупикових тестів.

На основі застосування функціоналу (6) можемо запропонувати різні алгоритми розпізнавання об'єктів, які описуються дискретними наборами $x_1,....,x_n$, $(x_i \in \{0,1,....,k_i - 1\})$ (або наборами дійсних чисел).

На основі фунціоналу (6) можливо знайти максимальне число тупикових тестів для кожного n. Це число буде рівне $\max(C_n^1,....,C_n^{n-1})$ за деяких обмежень щодо значення функціоналу (6).

Функціонал (2) можливо використовувати для вибору розташування аргументів у граф-схемах нульових та багатозначних функцій за мінімізації.

В оцінці повноти всіх ознак $x_1,....,x_n$ інтерес викликає така величина: $B(x_1,....,x_n) = \min_{\Delta \in Q} \delta_\Delta$, де $\varnothing Q$ мають той самий сенс, що і в (6). Ця оцінка є більш жорсткою, ніж $W(x_1,....,x_n)$, і тому її більш доцільно застосовувати тоді, коли необхідно отримати значний ефект у розпізнаванні об'єктів $x, (x \in G)$, тобто коли потрібно максимально обмежити ризик неправильного розпізнавання.

Аналогічно функціоналу (2) можемо ввести функціонал $W^`(x_i) = \mathrm{W}(x_1,....,x_{i-1},x_{i+1},...,x_n)$. При цьому найважливішою ознакою стосовно функції, яка задає розбиття НВ на об'єкти, вважається та ознака, для якої ця величина мінімальна.

Список літератури:

1. Повхан І.Ф. Визначення поняття ознаки в теорії розпізнавання образів. Науково технічний журнал «Штучний Інтелект». 2002. № 4. С. 512–517.
2. Повхан І.Ф. Групова та індивідуальна оцінка важливості бульових аргументів. Вісник НТУ «ХПІ». 2011. № 53. С. 57–64.
3. Повхан І.Ф., Василенко Ю.А., Василенко Е.Ю. Концептуальна основа систем розпізнавання образів на основі методу розгалуженого вибору ознак. European Journal of Enterprise Technologies. 2004. № 7(1). С. 13–15.
4. Повхан І.Ф., Василенко Е.Ю., Василенко Ю.А. Метод розгалуженого вибору ознак у математичному конструюванні багаторівневих систем розпізнавання образів. Штучний інтелект. 2003. № 7. С. 246–249.

ПРОБЛЕМА ФУНКЦИОНАЛЬНОЙ ОЦЕНКИ ОБУЧАЮЩЕЙ ВЫБОРКИ В ЗАДАЧАХ РАСПОЗНАВАНИЯ ДИСКРЕТНЫХ ОБЪЕКТОВ

Статья посвящена актуальной проблеме оценки обучающей выборки на основе функционалов. Правильная оценка выборки позволяет построить эффективное правило классификации, обеспечивает простое и полное распознавание дискретных объектов.

Ключевые слова: *распознавание дискретных объектов, важность аргументов, функция распознавания.*

THE PROBLEM OF FUNCTIONAL ASSESSMENT OF THE TRAINING SET IN THE TASKS OF RECOGNITION OF DISCRETE OBJECTS

The work is devoted to the actual problem of evaluating a training sample based on functional. The correct estimate of the sample allows you to build an effective classification rule, provides a simple and complete recognition of discrete objects.

Key words: *recognition of discrete objects, importance of arguments, recognition function.*