УДК 519.25
ББК 22.17

ПРО ПОСЛІДОВНИЙ МЕТОД КЛАСИФІКАЦІЇ

М. М. Осипчук
Прикарпатський національний університет імені Василя Стефаника; 76018, Івано-Франківськ, вул. Шевченка, 57; e-mail: myosyp@ukr.net

Розглядається задача оцінки ймовірностей належності об'єкта до двох сукупностей за відомими його незалежними характеристиками. Пропонується методика класифікації таких об'єктів.

Ключові слова: *прогнозування, класифікація, дискримінантний аналіз, статистика.*

Постановка завдання. Проблема полягає у необхідності розробки методу визначення ймовірності належності досліджуваного об'єкта до однієї з двох сукупностей та побудові ефективного методу визначення сукупності, до якої слід віднести об'єкт. Інформація про об'єкт містить значення певної кількості незалежних характеристик. Кожна характеристика може приймати скінченну кількість дискретних значень.

Завдання прогнозування належності досліджуваного об'єкта до однієї з кількох сукупностей є стандартним завданням дискримінантного аналізу. За наявності інформації про значення певної кількості характеристик всіх членів навчальних вибірок із заданих генеральних сукупностей існує можливість побудови дискримінантних функцій (лінійних), за значеннями яких приймається рішення про віднесення нового об'єкта до цих сукупностей. Крім того, можна знайти ймовірності (апостеріорні) належності цього об'єкта до кожної з них. Ситуація ускладнюється, якщо замість інформації про значення характеристик об'єктів даних сукупностей маємо тільки частоти, з якими ці значення спостерігалися. Вивченню саме цього питання присвячена запропонована робота.

Точкова оцінка ймовірності належності об'єкта до однієї з двох сукупностей**.** Для побудови оцінки ймовірності належності об'єкта до кожної із сукупностей розглянемо навчаючу вибірку (об'єму n) характеристик об'єктів аналогічних досліджуваному, поділену на дві частини (вибірки з двох сукупностей A і B) об'ємів n_1 та n_2 ($n_1+n_2=n$). Самі сукупності A і B вважаємо достатньо великими. Кожне вибіркове значення є вектором $\left(x_{1i}^{(j)}, x_{2i}^{(j)}, ..., x_{hi}^{(j)}\right)$, $i=\overline{1,n_j}$, $j=1,2$, $h\in N$. Значення $x_{li}^{(j)}$ належить до деякої k_l-елементної множини $\mathbf{M}_l$. Характеристики ξ_l, вибірковими значеннями яких є $x_{li}^{(j)}$, вважаємо незалежними.

Нехай апріорні ймовірності належності об'єкта до сукупностей A і B відповідно дорівнюють p_1 і p_2, а умовні ймовірності – $\mathbf{P}\left(\xi_l=\hat{x}_{lm}/A\right)=q_{lm}^{(1)}$, $\mathbf{P}\left(\xi_l=\hat{x}_{lm}/B\right)=q_{lm}^{(2)}$, де $\hat{x}_{lm}$ – m-тий елемент мно-

жини $\mathbf{M}_l$. За формулою Байєса ймовірність того, що об'єкт зі значенням $\hat{x}_{lm}$ характеристики ξ_l належить до сукупності A, дорівнює

$$p_{lm}^{(1)} = \mathbf{P}\left(A / \xi_l = \hat{x}_{lm}\right) = \frac{p_1 q_{lm}^{(1)}}{p_1 q_{lm}^{(1)} + p_2 q_{lm}^{(2)}}, \tag{1}$$

а до сукупності B – $p_{lm}^{(2)} = 1 - p_{lm}^{(1)}$. Враховуючи незалежність характеристик ξ_l, ймовірність належності об'єкта зі значеннями $\xi_1 = \hat{x}_{1m_1}$, $\xi_2 = \hat{x}_{2m_2}$, ..., $\xi_l = \hat{x}_{lm_l}$ до сукупності A можна знайти за рекурентною формулою

$$p_l^{(1)} = \mathbf{P}\left(A / \xi_1 = \hat{x}_{1m_1}, ..., \xi_l = \hat{x}_{lm_l}\right) = \frac{p_{l-1}^{(1)} q_{lm_l}^{(1)}}{p_{l-1}^{(1)} q_{lm_l}^{(1)} + p_{l-1}^{(2)} q_{lm_l}^{(2)}}, \tag{2}$$

$$p_l^{(2)} = 1 - p_l^{(1)}, \tag{3}$$

$$p_0^{(1)} = p_1, \; p_0^{(2)} = p_2. \tag{4}$$

Замінивши ймовірності, що входять до формул (2) – (4), на їх оцінки, побудовані за навчаючими вибірками, знайдемо оцінки ймовірностей $p_h^{(1)}$ та $p_h^{(2)} = 1 - p_h^{(1)}$. Оцінками ймовірностей p_1 та p_2 можуть бути, відповідно $\hat{p}_1 = 0,5$, $\hat{p}_2 = 0,5$, якщо із сукупностей A і B було одержано незалежні вибірки, об'єми яких ніяк не відображають об'єми цих сукупностей, чи $\hat{p}_1 = \frac{n_1}{n}$, $\hat{p}_2 = \frac{n_2}{n}$, якщо вибірки одержані розбиттям вибірки із об'єднаної сукупності, і їх об'єми пропорційні об'ємам сукупностей.

Оцінити ймовірності $q_{lm}^{(j)}$ можна, знайшовши надійні інтервали для них. Для спрощення викладок введемо такі позначення: p – оцінювана ймовірність, n – об'єм вибірки, ν – відносна частота, з якою у вибірці зустрічається значення, ймовірність якого оцінюється. Враховуючи асимптотичну нормальність величини $\frac{n\nu - np}{\sqrt{np(1-p)}}$, можна знайти наближений надійний інтервал для p із заданим рівнем надійності:

$$\left(\frac{2n\nu + \varepsilon^2 - \varepsilon\sqrt{4n\nu(1-\nu) + \varepsilon^2}}{2(n+\varepsilon^2)}; \frac{2n\nu + \varepsilon^2 + \varepsilon\sqrt{4n\nu(1-\nu) + \varepsilon^2}}{2(n+\varepsilon^2)} \right). \tag{5}$$

Тут ε є розв'язком рівняння $\mathbf{P}\left(\left| \frac{n\nu - np}{\sqrt{np(1-p)}} \right| < \varepsilon \right) = \gamma$, а γ – рівень надійності. За достатньо великого n можна взяти $\varepsilon = u_{\frac{\gamma+1}{2}}$ – квантиль порядку $\frac{\gamma+1}{2}$ стандартного нормального розподілу.

Вибір оцінок $\hat{q}_{lm}^{(j)}$ ймовірностей $q_{lm}^{(j)}$ з інтервалів типу (5) потрібно здійснювати з врахуванням того, що $\sum_{m=1}^{k_l} \hat{q}_{lm}^{(j)} = 1$. Один з варіантів такого вибору додає до середини інтервалу (5) поправку

$$\Delta_l^{(j)} = \frac{(2-k_l)\varepsilon^2}{2k_l(n_j+\varepsilon^2)}, \tag{6}$$

яка рівномірно розподіляє між оцінками $\widehat{q}_{lm}^{(j)}$ відмінність від 1 суми середин згаданих інтервалів. Таким чином оцінками ймовірностей $q_{lm}^{(j)}$ є статистики

$$\widehat{q}_{lm}^{(j)} = \frac{n_j k_l \nu_{lm}^{(j)} + \varepsilon^2}{k_l(n_j+\varepsilon^2)}, \tag{7}$$

де $\nu_{lm}^{(j)}$ – відносна частота значення $\widehat{x}_{lm}$ характеристики ξ_l в j-тій навчаючій вибірці. Зауважимо, що для бінарних характеристик ($k_l = 2$) поправка (6) дорівнює нулю і відповідні оцінки є серединами інтервалів (5).

З властивостей оцінки (7) слід відмітити такі.

1. Оцінка (7) належить інтервалу (5).

Дійсно, це підтверджують очевидні нерівності

$$\left|\Delta_l^{(j)}\right| = \frac{\left(1-\frac{2}{k_l}\right)\varepsilon^2}{2(n_j+\varepsilon^2)} < \frac{\varepsilon^2}{2(n_j+\varepsilon^2)} \le \frac{\varepsilon\sqrt{4n_j\nu_{lm}^{(j)}(1-\nu_{lm}^{(j)})+\varepsilon^2}}{2(n_j+\varepsilon^2)}.$$

2. Оцінка (7) асимптотично незміщена.

Для доведення, врахувавши, що $\mathbf{M}\,\nu_{lm}^{(j)} = q_{lm}^{(j)}$, знайдемо

$$\mathbf{M}\widehat{q}_{lm}^{(j)} = \frac{n_j k_l \mathbf{M}\,\nu_{lm}^{(j)} + \varepsilon^2}{k_l(n_j+\varepsilon^2)} = \frac{n_j k_l q_{lm}^{(j)} + \varepsilon^2}{k_l(n_j+\varepsilon^2)} \to q_{lm}^{(j)}, \quad n_j \to \infty.$$

3. Дисперсія оцінки (7) еквівалентна дисперсії відносної частоти $\nu_{lm}^{(j)}$ і прямує до нуля при $n_j \to \infty$. Це випливає з рівності

$$\mathbf{D}\widehat{q}_{lm}^{(j)} = \frac{n_j^2 \mathbf{D}\,\nu_{lm}^{(j)}}{(n_j+\varepsilon^2)^2} = \frac{n_j^2}{(n_j+\varepsilon^2)^2} \cdot \frac{q_{lm}^{(j)}(1-q_{lm}^{(j)})}{n}.$$

Надійний інтервал для ймовірності належності об’єкта до однієї з двох сукупностей**.** Такий інтервал може бути знайдений з допомогою відомих надійних інтервалів для умовних ймовірностей значень кожної характеристики в розглянутих сукупностях. Нехай $q_{lm}^{(j)} \in \left(\alpha_{lm}^{(j)}; \beta_{lm}^{(j)}\right)$ з ймовірністю $\gamma_{lm}^{(j)}$. Досліджуючи екстремуми функції

$$f(x,y,z) = \frac{xy}{xy+(1-x)z},$$

можемо стверджувати, що врахування перших l характеристик приводить до такого надійного інтервалу $\left(p_l^-; p_l^+\right)$ для ймовірності належності об’єкта до першої ($j=1$) із сукупностей , для якого мають місце рівності

$$p_l^- = \frac{p_{l-1}^- \alpha_{lm_l}^{(1)}}{p_{l-1}^- \alpha_{lm_l}^{(1)} + (1-p_{l-1}^-)\beta_{lm}^{(2)}}; \tag{8}$$

$$p_l^+ = \frac{p_{l-1}^+ \beta_{lm_l}^{(1)}}{p_{l-1}^+ \beta_{lm_l}^{(1)} + (1 - p_{l-1}^+)\alpha_{lm}^{(2)}}; \tag{9}$$

$$p_0^- = p_0^+ = \hat{p}_1. \tag{10}$$

Тут ми вважаємо, що досліджуваний об'єкт має значення l-тої характеристики, рівне $\hat{x}_{lm_l}$. Ітеруючи рівності (8), (9) з врахуванням (10), одержимо, що врахування всіх h характеристик дає надійний інтервал $(p_j^-; p_j^+)$ з межами:

– для ймовірності належності до сукупності A

$$p_1^- = \frac{\prod_{l=1}^{h} \alpha_{lm_l}^{(1)}}{\prod_{l=1}^{h} \alpha_{lm_l}^{(1)} + \prod_{l=1}^{h} \beta_{lm_l}^{(2)} \frac{\bar{p}_2}{\bar{p}_1}}, \; p_1^+ = \frac{\prod_{l=1}^{h} \beta_{lm_l}^{(1)}}{\prod_{l=1}^{h} \beta_{lm_l}^{(1)} + \prod_{l=1}^{h} \alpha_{lm_l}^{(2)} \frac{\bar{p}_2}{\bar{p}_1}}; \tag{11}$$

– для ймовірності належності до сукупності B

$$p_2^- = \frac{\prod_{l=1}^{h} \alpha_{lm_l}^{(2)}}{\prod_{l=1}^{h} \alpha_{lm_l}^{(2)} + \prod_{l=1}^{h} \beta_{lm_l}^{(1)} \frac{\bar{p}_1}{\bar{p}_2}}, \; p_2^+ = \frac{\prod_{l=1}^{h} \beta_{lm_l}^{(2)}}{\prod_{l=1}^{h} \beta_{lm_l}^{(2)} + \prod_{l=1}^{h} \alpha_{lm_l}^{(1)} \frac{\bar{p}_1}{\bar{p}_2}}. \tag{12}$$

Надійність цих інтервалів становить $\gamma = \prod_{l=1}^{h} \gamma_{lm_l}^{(1)} \gamma_{lm_l}^{(2)}$. Якщо ж вибирати надійності інтервалів для ймовірностей $q_{lm_l}^{(j)}$ однаковими, то вони повинні бути рівними $\sqrt[2h]{\gamma}$ для забезпечення надійності γ.

Метод дискримінації**.** Побудова дискримінуючого правила повинна враховувати наведені вище викладки та умови проведення дослідження. Найбільш очевидним і водночас досить вимогливим щодо величини об'єму навчаючої вибірки є наступне. Досліджуваний об'єкт слід віднести до сукупності A, якщо $p_2^+ < p_1^-$, або – до сукупності B, якщо $p_1^+ < p_2^-$. Інші варіанти є випадками невизначеності. Такий підхід вимагає вузьких надійних інтервалів з межами (11), (12), що зменшує ймовірність невизначеності. Цього можна досягти за умови достатньо великих об'ємів навчаючих вибірок.

Інший варіант дискримінуючого правила простіший для застосування, хоча виглядає дещо необґрунтованим. Згідно цього правила спостереження відноситься до сукупності A, якщо ймовірність з (2) $p_h^{(1)} > 0,75$, та до сукупності B, якщо $p_h^{(1)} < 0,25$. В інших випадках маємо невизначеність. Надійність такого методу може бути оцінена з використанням частоти правильних дискримінацій за екзаменуючою вибіркою. Вибір меж 0,25 та 0,75 є безперечно умовним, хоча досвід робіт [1], [2], вказує на можливість побудови досить якісного методу дискримінації саме з такими межами.

На завершення ще раз звернемо увагу на основне обмеження щодо застосування описаного методу. Характеристики, що застосовуються для класифікації, повинні бути незалежними в сукупності.

Література

1. Матейко Г.Б. Прогнозування ускладнень вагітності і пологів у жінок з герпесвірусною інфекцією / Г.Б. Матейко, М.М. Осипчук, Б.М. Дикий // Галицький лікарський вісник. – 2006. – Т.13. – № 4. – С. 103-105.
2. Матейко Г.Б. Антенатальне прогнозування внутрішньоутробного інфікування плода у вагітних жінок з герпетичною і цитомегаловірусною інфекціями / Г.Б. Матейко, М.М. Осипчук // Архів клінічної медицини. – 2006. – № 2(10). – С. 45-47.

Стаття поступила в редакційну колегію 02.11.2009 р.
Рекомендовано до друку чл.-кореспондентом НАН України,
професором ***Портенком М.О.***

ABOUT A CONSECUTIVE DISCRIMINATORY ANALYSIS

M. M. Osypchuk
Precarpathian National University named by Vasil Stefanic,
76000, Ivano-Frankivs'k, Shevchenko street, 57;
e-mail: myosyp@ukr.net

A task is examined about the evaluation of probabilities of that an object belongs each of two aggregates. We consider known his independent descriptions. The method of classification of such objects is built.

Keywords: *prognostication, classification, discriminant analysis, statistics.*