УДК 519.23(045)

Тупіцина Н. М.
Національний авіаційний університет, Київ

СТАТИСТИЧНИЙ АЛГОРИТМ ПОДІЛУ СУБ'ЄКТІВ НА КЛАСИ ЗА ПРОФЕСІЙНОЮ ПРИДАТНІСТЮ

В статті аналізується алгоритм, заснований на модифікації послідовного статистичного аналізу відношення ймовірностей. Для цілей визначення професійної придатності цей алгоритм повинен бути ще більш ефективним, так як психологічні ознаки є слабо статистично залежними, а за цих умов послідовний аналіз відношень ймовірностей є оптимальною процедурою для класифікації на два класи.

Нехай інформація про психологічні особливості людини міститься в n-мірному векторі ν ($\nu_1, \nu_2, ..., \nu_n$). Кожне з ν_i (i = 1, 2 ,..., n) –число, отримане за допомогою тієї або іншої методики (серед них можуть бути певним чином закодовані і якісні характеристики людини). Надалі компоненти ν будуть називатися ознаками. Вибір ознак зазвичай проводиться з урахуванням психологічних вимог до професійної придатності. Пропонований алгоритм дозволяє відкинути ті з використовуваних ознак, які виявляються неінформативними для даного конкретного завдання визначення професійної придатності.

Передбачається, що групам осіб, з одного боку, придатних (група «А»), а з іншого боку, непридатних (група «В») до розглянутої діяльності відповідають два класи векторів $\{\nu_A\}$ і $\{\nu_B\}$, які можуть сильно перетинатися, але статистично різні. У подальшому завжди будемо вважати, що $\{\nu_A\}$ – клас векторів, що характеризують придатних до даної діяльності суб'єктів.

З математичної точки зору завдання визначення професійної придатності полягає у віднесенні з певною ймовірністю помилки вектора (ν1, ν2, ..., νη) до одного з двох класів – «А» або «В».

Є багато різних методів вирішення цієї задачі. У всіх методах необхідний етап «навчання»: статистичний аналіз вже наявного досвіду. Для цілей визначення професійної придатності вони не набули великого поширення – одні з-за крайньої громіздкості та складності застосування навіть за допомогою обчислювальних машин, інші тому, що виявилися не дуже ефективними.

Успіх класифікації за багатьма ознаками в задачах діагностики залежить від інформативності цих ознак і способу інтеграції інформації. Цей спосіб інтеграції має бути:

1) простим в обчислювальному відношенні і доступним при використанні;

2) малочутливим до відсутності якої-небудь ознаки;

3) у якійсь мірі інваріантним до зсуву розподілів ознак (останнє істотно в силу необхідності рахуватися з різними методичними умовами отримання однієї і тієї ж ознаки).

Цим вимогам у значній мірі задовольняє алгоритм, заснований на модифікації послідовного статистичного аналізу відношення ймовірностей.

Алгоритм

Алгоритм складається з двох етапів: першого – етапу навчання, під час якого накопичується інформація про ознаки на підставі вже наявного досвіду і оцінюється інформативність обраних ознак, і другого - етапу класифікації, на якому виноситься рішення про придатність суб'єкта до певної діяльності.

Навчання. Передбачається, що на підставі попереднього досвіду можна виділити групи суб'єктів «А» і «В», які відображають наше розуміння придатності (або непридатності) до даної роботи і є певними еталонами для подальшого прогнозування придатності. Далі передбачається, що є якийсь набір ознак $\nu_1, \nu_2 ,..., \nu_n$, істотність яких для визначення професійної придатності можна і не знати. Тепер можна побудувати множини векторів $\{\nu_A\}$ і $\{\nu_B\}$, що відповідно характеризують групи суб'єктів «А» і «В». Процес навчання складається з одержання оцінки дискретних одновимірних розподілів ймовірностей ознак $\nu_1, \nu_2 ,..., \nu_n$ для класу «А»:

$$f_A^1(\nu_1), f_A^2(\nu_1), ..., f_A^n(\nu_1),$$

для класу «В»:

$$f_B^1(\nu_1), f_B^2(\nu_1), ..., f_B^n(\nu_1).$$

Передбачається, що $\nu_1, \nu_2 ,..., \nu_n$ слабо залежні. Якщо, проте, цього немає, то для збільшення ефективності процедури в розгляд вводяться складні ознаки – синдроми, визначення яких можна отримати на підставі досвіду і теоретичних

міркувань або ж використовуючи відповідний математичний апарат. Побудова одновимірних розподілів істотно полегшує процес навчання, а в разі слабкої залежності втрати інформації при цьому невеликі.

Якщо класи «А» і «В» численні, то можна отримати досить хорошу оцінку необхідних ймовірностей:

$$\{f_A(v_1)\} u \{f_B(v_1)\} \quad (i=1,2,\ldots,n).$$

У тих же випадках, коли чисельності класів «А» і «В» невеликі, доводиться вдаватися до грубого квантуванню ознак на 2-3-4 градації. Практична перевірка показує, що за наявності в групі 25-30 чоловік і відповідному квантуванні можна отримати задовільні результати.

Отримані в результаті обстеження даного контингенту осіб показники можуть мати різну цінність для цілей прогнозування професійної придатності. Тому наступним етапом «навчання» є оцінка інформативності ознак.

Ознака буде тим більш інформативною, чим більше різниця між його розподілами у представників класу «А» і «В». Оцінка інформативності ознаки ν, може виражатися величиною P_j – ймовірністю того, що розподіли

$$f_A^j(\nu_j) \quad \text{і} \quad f_B^j(\nu_j)$$

різні. Це досягається за допомогою обчислення χ_2.

Інтуїтивно ясно, що ймовірність Р може бути хорошою мірою інформативності ознаки ν при даній конкретній класифікації. Необхідно відзначити, що ознаки, інформативні в одному випадку, можуть виявитися зовсім не інформативними для розв'язання задачі профвідбору інших фахівців.

$$\chi_j^2 = N_A^{(j)} \cdot N_B^{(j)} \cdot \sum_{i=1}^{S_j} \left[\frac{1}{A_i^{(j)} + B_i^{(j)}} \left(\frac{A_i^{(j)}}{N_A^{(j)}} - \frac{B_i^{(j)}}{N_B^{(j)}} \right) \right].$$

Обчислення χ_2 проводилося за формулою:

де $N_A^{(j)}$ і $N_B^{(j)}$ – загальне число осіб відповідно в класах «А» і «В», дані яких використовувалися при побудові розподілів для *j*-ї ознаки; $A_i^{(j)}$ і $B_i^{(j)}$ – частоти появи індивідів у *i*-градації *j*-ї ознаки для порівнюваних класів; S – число градацій для j-ї ознаки. Вірогідність *Pj* визначалися за таблицями Л. Большова і Н. Смирнова. Оцінка інформативності може бути також отримана і за допомогою відстані Кульбака. У прийнятих тут позначеннях і дещо зміненій формі ця відстань має вигляд:

$$I_j = \left| I_j^A \right| + \left| I_j^B \right|,$$

де

$$I_j^A = \sum_{i=1}^{s_j} \left(\frac{A_i^j}{N_A^j} \lg \frac{A_i^j \cdot N_B^j}{B_i^j \cdot N_A^j} \right) \text{ і } I_j^B = \sum_{i=1}^{s_j} \left(\frac{B_i^j}{N_B^j} \lg \frac{A_i^j \cdot N_B^j}{B_i^j \cdot N_A^j} \right).$$

Цей захід має ряд переваг, особливо при теоретичних досліджень. Для практики становить інтерес можливість виміру значимості ознак ν_1 (j = 1, 2 ,..., n) окремо для винесення рішення про належність ν до $\{\nu_A\}$ або $\{\nu_B\}$ (відповідно складові I_j^A і I_j^B).

Використовуючи ту чи іншу міру, ознаки доцільно розташувати за їх зменшуваною інформативністю, а ті з них, які неінформативні (Р занадто велике або І – мале), використовувати не треба. Якщо виявиться, що інформативних ознак залишилося мало, то необхідно ввести нові ознаки. Процес «навчання» можна вважати закінченим, коли оцінки розподілів

$$f_j^A(v_j) \text{ і } f_j^B(v_j), \quad (j = 1, 2, \ldots, n),$$

досить надійні, ознаки впорядковані за їх інформативністю і їх досить багато.

Класифікація (вирішальне правило). При класифікації можна допустити дві помилки. Суб'єкт з класу «А» може бути помилково віднесений до класу «В» і, навпаки, суб'єкт з класу «В» може бути помилково зарахований до класу «А». Першу з вказаних помилок класифікації будемо позначати через α, а другу через β.

Ймовірності помилок а і β визначаються до проведення класифікації. При виборі цих ймовірностей повинна бути врахована важливість тієї чи іншої помилки класифікації, а також реальна ситуація, що виникла при вирішенні даного конкретного завдання.

Нехай при обстеженні суб'єкта S були отримані ознаки

$$v_1^0, v_2^0, \ldots, V_n^0,$$

(вони наведені тут в порядку зменшення їх інформативності). Нехай на підставі здорового глузду обрані допустимі ймовірності помилок α і β. Розглянемо відношення ймовірностей, відповідних першій ознаці:

$$\frac{f_B^1(v_1^0)}{f_A^1(v_1^0)}.$$

Якщо це відношення буде менше ніж: $\frac{\alpha}{1-\beta}$, то це буде означати, що отримане значення ознаки настільки найімовірніше для класу «А», що можна з обраним рівнем надійності (α, β) стверджувати, що дана особа відноситься до класу «А» (придатна до даної професійної діяльності).

Якщо це відношення $> \frac{1-\alpha}{\beta}$, то з тим же рівнем надійності приймається рішення про непридатність до розглянутої діяльності. Якщо

$$\frac{\alpha}{1-\beta} < \frac{f_B^1(v_1^0)}{f_A^1(v_1^0)} < \frac{1-\alpha}{\beta}$$

то інформація, укладена в ознаці, недостатня для віднесення до класів «А» і «В» і розглядається наступна ознака v_2^0.

Якщо

$$\frac{f_B^1(v_1^0)\cdot f_B^2(v_2^0)}{f_A^1(v_1^0)\cdot f_A^2(v_2^0)} < \frac{\alpha}{1-\beta},$$

то виноситься рішення про віднесення індивіда до класу «А» якщо

$$\frac{f_B^1(v_1^0)\cdot f_B^2(v_2^0)}{f_A^1(v_1^0)\cdot f_A^2(v_2^0)} > \frac{1-\alpha}{\beta},$$

то в клас «В». Коли ж

$$\frac{\alpha}{1-\beta} < \frac{f_B^1(v_1^0)\cdot f_B^2(v_2^0)}{f_A^1(v_1^0)\cdot f_A^2(v_2^0)} < \frac{1-\alpha}{\beta},$$

то розглядається значення третьої ознаки v_3^0 і т.д.

Якщо, перебравши всі ознаки, які не вдається віднести суб'єкта до того чи іншого класу з даним рівнем надійності, тобто розглянуте відношення не виходить за межі необхідних рубежів, то це означає, що наявні результати обстеження не дозволяють зробити прогноз з обраним рівнем надійності. У цих випадках можна знизити цей рівень і таким чином зробити прогноз або звернутися за додатковою інформацією.

При відсутності додаткової інформації для мінімізації ймовірності помилки доцільно побудувати два розподіли відношення правдоподібності за всіма ознаками відповідно для груп «А» і «В» і на основі цих розподілів вибрати один поріг. Особливості розподілу зазвичай такі, що цим порогом рідко буває 1.

Як відомо, в схемах послідовного статистичного аналізу. процедури обґрунтовуються для однорідного випадку, коли

$$f_A^1(v_1) = f_A^2(v_2) = \ldots = f_A^n(v_n) \quad \text{і}$$

$$f_B^1(v_1) = f_B^2(v_2) = \ldots = f_B^n(v_n).$$

Проте неважко показати, що залежність порогів від імовірності помилок α і β переноситься і на випадок неоднакових розподілів.

Практично зручно мати справу не з відношеннями ймовірностей, а з логарифмом цього відношення. Тоді всі обчислення зводяться до послідовного додавання.

Отже, визначення належності векторів ν (ν_1, ν_2 ,..., ν_n) до множини $\{\nu_A\}$ або $\{\nu_B\}$ здійснюється наступним чином. Послідовно обчислюються величини L_1 L_2 ,..., L_k, де:

$$L_k = \sum_{j=1}^{k} R_j, \text{ а } R_j = \lg \frac{f_B^j(v_j)}{f_A^j(v_j)}.$$

Кожне обчислене L_k порівнюється з порогами

$$\frac{\alpha}{1-\beta} \quad \text{и} \quad \frac{1-\alpha}{\beta}.$$

Якщо при деякому k<n

$$\frac{\alpha}{1-\beta} < L_k < \frac{1-\alpha}{\beta}.$$

То обчислюється L_{k+1}. Якщо ж

$$L_k > \frac{\alpha}{1-\beta}.$$

То $v \subset \{v_B\}$; якщо ж

$$L_k > \frac{1-\alpha}{\beta}.$$

Вибір порога

У послідовній статистичної процедурі відношення ймовірностей передбачаються два пороги

$$\frac{\alpha}{1-\beta} \quad \text{и} \quad \frac{1-\alpha}{\beta}$$

де α, β – помилки класифікації, які призначаються заздалегідь Проста залежність порогів від ймовірностей помилок класифікації дозволяє вибирати потрібний поріг, основуючись на кон'юнктурі

Необхідність вибору невеликого числа осіб з великих контингентів робить можливим визначити a = b порядку 0,001 або навіть 0,0001 З іншого боку, при обмеженій кількості осіб природно вибрати α = β = 0,05 або навіть 0,10

Якщо виявиться, що помилка пропустити хорошого фахівця і, навпаки, помилка прийому малопридатного нерівноцінні, то є можливість враховувати це, вибираючи різні ймовірності α і β. Таким чином, вибір порогів є дуже гнучким і враховує реальну обстановку, а також ціну можливих помилок.

Науковий керівник – Лещинський О.Л.,
канд. фіз.-мат. наук, доц.