УДК 004.021

В.М. Ігнатенко, А.Ю. Кравченко, В.М. Чумаков

СПОСОБИ СТАТИСТИЧНОЇ КЛАСИФІКАЦІЇ ДИХОТОМІЧНИХ ГІПОТЕЗ

Анотація: Визначення основних напрямків у розвитку теорії класифікації дихотомічних гіпотез, виділення основних недоліків сучасної теорії класифікації та способів їх подолання.

Ключові слова: кластер, множина об'єктів, об'єднання кластерів.

Вступ

В сучасному суспільстві існує досить велика кількість неупорядкованої інформації, а сучасні пошукові системи намагаються класифікувати та кластеризувати безліч розрізненої інформацій в Інтернеті, при умові, що в наш час немає ідеальних алгоритмів класифікації, а це, в свою чергу, погіршує можливості пошуку та аналізу інформації. Одна з основних проблем класифікації - перевірка "реальності"кластера, його об'єктивного існування незалежно від розрахунків дослідника.

Підходи до класифікації

Окремий випадок перевірки "реальності" кластера – перевірка обґрунтованості об'єднання двох кластерів, які ми розглядаємо як дві множини об'єктів, а саме, множини $\{a_1, a_2, \ldots, a_k\}$ і $\{b_1, b_2, \ldots, b_m\}$. Нехай є дві сукупності мір близькості: одна – міри близькості між об'єктами, що лежать усередині одного кластера, тобто $d(a_i, a_j)$, $1 \leqslant i < j \leqslant k$, $d(b_\alpha, b_\beta)$, $1 \leqslant \alpha < \beta \leqslant m$, і інша – міри близькості між об'єктами, що лежать у різних кластерах, тобто $d(a_i, b_\alpha)$, $1 \leqslant i \leqslant k, 1 \leqslant \alpha \leqslant m$. Ці дві сукупності мір близькості пропонується розглядати як незалежні вибірки й перевіряти гіпотезу про збіг їхніх функцій розподілу. Якщо гіпотеза не відкидається, об'єднання кластерів вважається обґрунтованим, у протилежному випадку - поєднувати не можна, алгоритм припиняє роботу.

У розглянутому підході є дві неточності [4]. По-перше, міри близькості не є незалежними випадковими величинами. По-друге, не враховується, що поєднуються не заздалегідь фіксовані кластери, а отримані в результаті роботи деякого алгоритму, і їхній склад (кількість елементів) виявляється випадковими. Від першої із цих неточностей можна частково позбутися. Справедливо наступне твердження.

Теорема 1. Нехай $a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_m$ *– незалежні однаково розподілені випадкові величини (зі значеннями в довільному просторі). Нехай випадкова величина* $d(a_1, a_2)$ *має всі моменти. Тоді при* $k, m \to \infty$ *розподіл статистики.*

$$\frac{8\sqrt{3U} - 3(k+m)(k+m-1)(k(k+1)+m(m+1))}{2(k+m)\sqrt{km(k^2+m^2)}},$$

 ISSN 1560-8956

де U – сума рангів елементів першої вибірки в об'єднаній вибірці; перша вибірка складена із внутрикластерних відстаней (мір близькості) $d(a_i, a_j)$, $1 \leqslant i < j \leqslant k$, і $d(b_\alpha, b_\beta)$, $1 \leqslant \alpha < \beta \leqslant m$, а друга – з міжкластерних відстаней $d \quad d(a_i, b_\alpha)$, $1 \leqslant i \leqslant k$, $1 \leqslant \alpha \leqslant m$ сходиться до стандартного нормального розподілу з математичним очікуванням 0 і дисперсією 1.

Тоді можна вирахувати величину U, і якщо вона занадто мала, то статистична гіпотеза однорідності відхиляється (на заданому рівні значимості), і можливість об'єднання відкидається.

Перейдемо до етапу застосування прогностичних правил, коли класи, в які вносимо об'єкт, уже виділені.

Прогностичне правило – це алгоритм, що дозволяє по характеристиках матеріалу прогнозувати його властивості. Якщо прогноз дихотомічний ("є" або "ні"), то правило є алгоритмом діагностики, при якому матеріал відноситься до одного з двох класів. Звісно,що кожне правило має свої набори ознак, і тут виникає проблема відбору ознак [2].

Природно відбирати лише найбільш "надійні" прогнози. Для доданння точного змісту терміну "надійний" необхідно мати спосіб порівняння алгоритмів діагностики по прогностичній "силі".

Результати обробки даних за допомогою деякого алгоритму діагностики описуються частками: правильної діагностики в першому класі κ; правильної діагностики в другому класі λ; частками класів в об'єднаній сукупності $\pi_i, i = 1..2; \pi_1 + \pi_2 = 1$

Величини k, π, π_1, π_2 визначаються ретроспективно.

Нерідко як показник якості алгоритму діагностики (прогностичної "сили") використовують частку правильної діагностики [1] $\mu = \pi_1 k + \pi_2 \lambda$

Однак показник μ визначається, зокрема, через характеристики π_1 і π_2, частково задані дослідником (наприклад, на них впливає тактика відбору зразків для вивчення).

Для виявлення інформативного набору ознак доцільно використовувати метод перерахування на модель лінійного дискримінантного аналізу, відповідно до якого статистичною оцінкою прогностичної "сили" є

$$\delta* = \Phi(d*/2), d* = \Phi^{-1}(k) + \Phi^{-1}(\lambda),$$

де $\Phi(x)$ – функція стандартного нормального розподілу ймовірностей з математичним очікуванням 0 і дисперсією 1, а $\Phi^{-1}(y)$ – зворотна їй функція.

Якщо класи описуються вибірками з багатомірних нормальних сукупностей з однаковими матрицями коваріацій, а для класифікації застосовується класичний лінійний дискримінантний аналіз Р.Фішера, то величина $d*$ являє собою реальну статистичну оцінку так званої відстані Махаланобиса [4] між розглянутими двома сукупностями, незалежно від граничного значення, що визначає конкретне вирішальне правило. У загальному випадку показник $\delta*$ уводиться як евристичний.

Теорема 2. *Нехай* $m, n \to \infty$. *Тоді для всіх* x

$$P\left(\frac{\delta* - \delta}{A(k,\lambda)} < x\right) \to \Phi(x),$$

де δ – справжня "прогностична сила" алгоритму діагностики; $\delta*$ – її емпірична оцінка,

$$A^2(k,\lambda) = \frac{1}{4}\left\{\left[\frac{\phi(d*/2)}{\phi(\Phi^{-1}(k))}\right]^2 \frac{k(1-k)}{m} + \left[\frac{\phi(d*/2)}{\phi(\Phi^{-1}(\lambda))}\right]^2 \frac{\lambda(1-\lambda)}{n}\right\};$$

де $\phi(x) = \Phi'(x)$ – щільність стандартного нормального розподілу ймовірностей з математичним очікуванням 0 і дисперсією 1.

За допомогою теореми 2 по k і λ звичайним чином визначають довірчі границі для "прогностичної сили" δ.

Припустимо, що класифікація складається в обчисленні деякого прогностичного індексу y і порівнянні його із заданим порогом c; об'єкт відносять до першого класу, якщо $y \leqslant c$, до другого, якщо $y > c$. Візьмемо два значення порога c_1 і c_2. Якщо перерахування на модель лінійного дискримінантного аналізу обґрунтоване, то "прогностичні сили" для обох правил збігаються: $\delta(c_1) = \delta(c_2)$. Цю статистичну гіпотезу можна перевірити.

Нехай k_1 – частка об'єктів першого класу, для яких $y \leqslant c_1$, а k_2 – частка об'єктів першого класу, для яких $c_1 < y \leqslant c_2$. Аналогічно нехай λ_2 – частка об'єктів другого класу, для яких $c_1 < y \leqslant c_2$, а λ_3 – частка об'єктів другого класу, для яких $y < c_2$. Тоді можна вирахувати дві оцінки тієї самої відстані Махаланобиса, що мають наступний вигляд [3]:

$$d*(c_1) = \Phi^{-1}(k_1) + \Phi^{-1}(\lambda_2 + \lambda_3) d*(c_2) = \Phi^{-1}(k_1 + k_2) + \Phi^{-1}(\lambda_3).$$

Теорема 3. *Якщо справжні прогностичні сили двох правил діагностики збігаються,* $\delta(c_1) = \delta(c_2)$ *топриm* $\to \infty, n \to \infty$ *при всіх* x

$$P\left(\frac{d*(c_1) - d*(c_2)}{B} < x\right) \to \Phi(x),$$

де

$$B_2 = \frac{1}{m}T(k_1;k_2) + \frac{1}{n}T(\lambda_3;\lambda_2);$$

$$T(x;y) = \frac{x(1-x)}{\phi^2(\Phi^{-1}(x))} + \frac{(x+y)(1-x-y)}{\phi^2(\Phi^{-1}(x+y))} - \frac{2x(1-x-y)}{\phi(\Phi^{-1}(x))\phi(\Phi^{-1}(x+y))}.$$

З теореми 3 випливає метод перевірки розглянутої гіпотези:

при виконанні нерівності $\left|\frac{d*(c_1)-d*(c_2)}{B}\right| \leqslant \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ вона приймається на рівні значимості, що асимптотично дорівнює α, у противному випадку - відкидається.

Алгоритм використання теорем:

1. Отримавши нові данні розподіляємо кожен елемент в кожен окремий кластер.

2. Користуючись *теоремою 1* об'єднаємо найбільш бажані для об'єднання (ті у яких розподіл по теоремі найближчий до нормального) кластери.
3. Для визначення найбільш бажаної пари кластерів використовують *теорему 2*. Для цього, поступово зменшуючи поріг c, виявляють лише ту пару кластерів, яка при найменшому c задовольняє *теоремі 2*.
4. Після визначення найбільш "підходящої" пари кластерів перевіряємо можливість такого об'єднання за допомогою *тереми 3*.
5. Якщо на заданому рівні значимості гіпотеза про можливість об'єднання двох кластерів відкидається, то необхідно повторити алгоритм використання теорем з пункту 2 і 3, при цьому ігноруючи можливість об'єднання пари відкинутих кластерів.
6. Алгоритм повторюється доти, доки за пунктом 4 не будуть перебрані та відкинуті усі можливі комбінації пар кластерів.

Особливості алгоритму:

1. Необхідно попередньо визначити емпіричну оцінку прогностичної сили, що є досить нелегким завданням.
2. Рівень значимості, за допомогою якого приймаються чи відкидається гіпотеза про об'єднання двох кластерів, задається користувачем.

Висновки

У даній статті було вказано на недоліки існуючих підходів класифікації і визначено можливості їх подолання, що дозволило сформувати алгоритм контролю "правильності" статистичної класифікації. Крім того у статті подані теореми, які є теоретичною основою для досягнення запропонованим алгоритмом найбільш тонного результату класифікації. Проте слід розуміти, що для кожного конкретного випадку найбільш ефективним може виявитися метод класифікації, найменш ефективний у інших випадках. Тому визначити універсальний метод класифікації неможливо.

Література

1. Себер Дж. Линейный регрессионный анализ. - М.: Мир, 1980. - 456 с.
2. Крамер Г. Математические методы статистики. - М.: Мир, 1975. - 648 с.
3. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. – 736 с.
4. Орлов А.И. Некоторые вероятностные вопросы теории классификации. – В сб.: Прикладная статистика. Ученые записки по статистике, т.45. - М.: Наука, 1983. – С.166-179.

Отримано 14.11.2011 р.