УДК 62-5:007:621.391:519.2

ОБ ОДНОМ МЕТОДЕ ТАКСОНОМИИ МНОЖЕСТВ ОБЪЕКТОВ И ПАРАМЕТРОВ

В.И.Котиков

§ I. Постановка задачи

Пусть дано M множеств $\{L_1, \dots, L_M\}$. Каждое множество L_i содержит N_i объектов, где $N_i > 1$ и $\sum_{i=1}^{M} N_i = N_0$. Каждый объект x_α, в свою очередь, представляет собой определенную точку в исходном p-мерном пространстве его описания $x_\alpha = \{x_{1\alpha}, \dots, x_{p\alpha}\}$. Множества $\{L_i\}$ не содержат общих объектов.

Необходимо M множеств $\{L_i\}$ разбить на K таксонов (классов, групп), причем, каждое из множеств L_i должно входить только в один из K таксонов. $1 < K < M$. Эти "укрупненные" множества (таксоны) будем обозначать $\{T_1, \dots, T_K\}$.

§ 2. Критерий качества таксономии

Определим сначала "таксономичность" (возможность объединения) - t_{ij} для любой пары множеств $\{L_i, L_j\}$.

При определении t_{ij} мы будем исходить из следующих двух интуитивно понятных требований:

а) множества L_i и L_j должны быть достаточно плохо "распознаваемы" ("отличимы") друг от друга (в этом случае величина t_{ij} будет достаточно большой);

б) множества L_i и L_j должны быть достаточно "похожими" друг на друга.

Как будет видно из дальнейшего, требования а) и б) не являются тавтологией друг друга.

Распознаваемость γ_{ij} множеств L_i и L_j мы будем определять величиной критерия Фишера $((\mu_i - \mu_j)^2/(\sigma_i + \sigma_j))$ по вектору

$y = \sum_{i=1}^{p} d_i X_i$ в исходном пространстве $X = \{X_1, \ldots, X_p\}$, который максимизирует данный критерий. Величины μ и σ – есть соответствующие оценки математического ожидания и дисперсии множеств. Если мы на основании множеств L_i и L_j составим матрицу ковариации $\| V_{kq} \|$ ($k, q = 1, \ldots, p$), где

$$V_{kq} = \frac{(\mu_{ki} - \mu_{kj})(\mu_{qi} - \mu_{qj})}{\sigma_{kqi} + \sigma_{kqj}},$$

то нетрудно доказать, что вектор y, максимизирующий критерий Фишера, есть вектор, соответствующий максимальному собственному числу λ_{max} матрицы $\| V_{kq} \|$. При этом

$$r_{ij} = \lambda_{max}(V) = \left(\frac{(\mu_i - \mu_j)^2}{\sigma_i + \sigma_j} \right)_{max}.$$

Сходство (похожесть) C_{ij} множеств L_i и L_j определим так. Допустим, что $\lambda_{min}(L_i)$, $\lambda_{min}(L_j)$ и $\lambda_{min}(L_i \cup L_j)$ – минимальные собственные числа матриц ковариаций, полученных по объектам множеств L_i, L_j и $(L_i \cup L_j)$ соответственно. Можно доказать, что $\lambda_{min}(L)$ – есть остаточная дисперсия при аппроксимации точек множества L линейным регрессионным полиномом по методу наименьших квадратов в модели X.

Пусть $\lambda_{min}(L_i, L_j) = \min\{\lambda_{min}(L_i), \lambda_{min}(L_j)\}$, тогда величина $C_{ij} = \lambda_{min}(L_i, L_j) / \lambda_{min}(L_i \cup L_j)$ будет определять степень ухудшения описания в пространстве X множеств L_i и L_j при их объединении. Если множества L_i и L_j "похожи" друг на друга, то величина C_{ij} будет достаточно большой.

Введем также и весовой коэффициент $(N_i + N_j)/N_0$, характеризующий степень доверия к полученным оценкам r_{ij} и C_{ij}.

Таким образом, величина t_{ij} определяется так:

$$t_{ij} = \frac{N_0 \cdot \lambda_{min}(L_i, L_j)}{(N_i + N_j) \cdot \lambda_{max}(V) \cdot \lambda_{min}(L_i \cup L_j)}.$$

Очевидно, что $t_{ij} = t_{ji}$.

На основании обучающей выборки определяется матрица $\| t_{ij} \|$; $i, j = 1, \ldots, M$. Пусть $t_{ii} \neq 0$.

Определять качество таксономии F в целом можно с помощью, например, одного из следующих двух различных критериев.

1. Квазиминимаксный критерий.

$$F_1 = \sum_{k=1}^{K} \sum_{i=1}^{M} \max(t_{ij}),$$

где: $L_i \subset T_\ell \wedge L_j \subset T_\ell$; $i \neq j$; $i, j = 1, \ldots, M$.

2. Квазибайесов (усредненный) критерий.

$$F_2 = \sum_{k=1}^{K} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} t_{ij},$$

где $L_i \subset T_\ell \wedge L_j \subset T_\ell$ и $i \neq j$.

§ 3. Алгоритмы решения задачи

Алгоритм 1.

1. Пусть $T_i = L_i$; $i = 1, \ldots, M$.
2. Множества L_i и L_j, не входящие ещё в один таксон, для которых величина t_{ij} максимальна, объединяются в один таксон T_{ij}.
3. Если два таксона T_{ij} и $T_{j\ell}$ имеют хотя бы один общий индекс, то они объединяются в один таксон $T_{ij\ell}$.
4. Если число таксонов больше K, то переходим к пункту 2.

Очевидно, что алгоритм 1 максимизирует критерий F_1.

Алгоритм 2.

Введем матрицу неизвестных $\| z_{\ell ij} \|$ ($\ell = 1, \ldots, K$; $i, j = 1, \ldots, M$). Целевая функция:

$$F_2 = \sum_{\ell=1}^{K} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} t_{ij} z_{\ell ij}; \qquad z_{\ell ij} = \{0, 1\}.$$

Ограничения:

$$\sum_{i=1}^{M} z_{\ell ii} \geq 1 ; \quad \ell = 1, \ldots, K ;$$

$$\sum_{\ell=1}^{K} z_{\ell ii} = 1 ; \quad i = 1, \ldots, M ;$$

$$z_{\ell ii} - \frac{1}{M-K+2} \left(\sum_{j=i}^{M} z_{\ell ij} + \sum_{j=1}^{i} z_{\ell ji} \right) \geq 0 ; \quad \ell = 1, \ldots, K ; \; i = 1, \ldots, M .$$

Это задача целочисленного линейного программирования, решение которой максимизирует критерий F_2 .

§ 4. Таксономия параметров

Часто возникает задача таксономии (группировки) исходных параметров (признаков) $\{X_1, \ldots, X_p\}$ на K непересекающихся таксонов (групп). В основу, как и в других работах, может быть взята, например, матрица коэффициентов корреляций $\| r_{ij} \|$ ($i, j = 1, \ldots, p$) между признаками $\{X_i\}$. Положив ($M = p$, $L_i = X_i$, $t_{ij} = |r_{ij}|$), мы тем самым, свели данную задачу к предыдущей. В этом случае можно пользоваться также критериями F_1 или F_2 и соответствующими алгоритмами.