УДК 62I.39I:5I2.2

О НЕКОТОРЫХ ЗАДАЧАХ ТАКСОНОМИИ (ГРУППИРОВКИ) ОБЪЕКТОВ

В.И. Котиков

В данной статье предложены методы оптимизации дисперсионного и дискриминантного критериев качества таксономии, а также рассмотрена постановка ряда других задач как задач таксономии. Приведены алгоритмы минимизации исходного описания в задачах таксономии.

§ I. Постановка задачи. Оптимизация дисперсионного критерия

Допустим, в ρ – мерном исходном пространстве признаков $X=\{X_1,\dots,X_\rho\}$ дано описание N объектов $\{a_1,\dots,a_N\}$. Необходимо разбить N объектов на K непересекающихся таксонов (групп, подмножеств) так, чтобы оптимизировать при этом некоторый критерий. Это обычная постановка задачи таксономии объектов [I,2].

Предположим, что вычислена матрица взаимных расстояний $\{\rho_{ij}\}$ между объектами в пространстве X. Тогда задача заключается в минимизации величины

$$\mathcal{D}=\sum_{\ell=1}^{K}\sum_{i=1}^{N_\ell}(\rho_{i0}(\ell))^2, \qquad (I)$$

где N_ℓ – число объектов в ℓ –м таксоне, а $\rho_{i0}(\ell)$ – – расстояние i – го объекта ℓ – го таксона до центра "тяжести" этого же таксона.

В работе [3] данная задача сведена к задаче квадратичного, а также целочисленного линейного программирования. Здесь же будет дано более простое решение.

Будем минимизировать несколько иной дисперсионный критерий:

$$\mathcal{D}^*=\sum_{\ell=1}^{K}\sum_{i=1}^{N_\ell}(\rho_{qi}(\ell))^2, \qquad (2)$$

где $\rho_{qi}(\ell)$ – расстояние q –го объекта ℓ –го таксона до i – го объекта того же таксона; причем q – й объект это такой объект ℓ – го таксона, который наиболее близок к центру "тяжести" того же таксона.

Очевидно, что если в каждом из K таксонов будет хотя бы один объект, достаточно близкий к центру "тяжести" своего таксона, то решения задачи по критерию $\mathcal{D}$ и критерию $\mathcal{D}^*$ будут близки.

Обозначим $\rho_{ij}^2=t_{ij}$. Введем целочисленные переменные (неизвестные):

$$x_{ij}=\{0,1\};\ i=1,\dots,N;\ j=1,\dots,N+1 \qquad (3)$$

Причем будем полагать "затраты", равными:

$$\left.\begin{aligned} t_{i(N+1)}&=0, \quad i=1,\dots,N;\\ t_{ij}&=\rho_{ij}^2; \quad i,j=1,\dots,N. \end{aligned}\right\} \qquad (4)$$

Столбец $(N+1)$ матрицы $\{x_{ij}\}$ является фиктивным и служит для контроля числа получающихся таксонов.

Определим функцию

$$Q^*=\sum_{i=1}^{N}\sum_{j=1}^{N+1}t_{ij}x_{ij} \qquad (5)$$

при ограничениях на неизвестные:

$$\sum_{i=1}^{N} z_{ij} = b_j \; ; \quad j = 1, \ldots, N+1 \; ; \qquad (6)$$

где $b_j = 1$; $j = 1, \ldots, N$ и $b_{N+1} = K$;

$$\sum_{j=1}^{N} z_{ij} - N z_{i(N+1)} \leq 0; \quad i = 1, \ldots, N \; . \qquad (7)$$

Число ненулевых строк в получившейся после решения матрице $\{z_{ij}\}$ будет равно числу таксонов. Это гарантируется неравенствами (7), а также равенством

$$\sum_{i=1}^{N} z_{i(N+1)} = b_{N+1} = K \; .$$

Равенства (6) гарантируют попадание каждого объекта только в один таксон.

Если q - я строка в матрице $\{z_{ij}\}$ окажется ненулевой и $\{z_{qv}=1, z_{qt}=1, z_{qc}=1\}$, то это означает, что объекты a_q, a_v, a_t и a_c входят в один таксон, к "центру" которого наиболее близким является объект a_q.

ЛЕММА. Объект a_q является наиболее близким к центру "тяжести" e - го таксона.

ДОКАЗАТЕЛЬСТВО. Дисперсия

$$\sigma = (1/N_e) \sum_{i=1}^{N_e} (x_i - \mu)^2$$

достигает минимума при

$$\mu = \bar{x} = (1/N_e) \sum_{i=1}^{N_e} x_i \; ;$$

это следует из уравнения $(\partial\sigma/\partial\mu) = 0$, то есть если $\mu = a_q$, то

$$\sum_{i=1}^{N_e} \rho_{iq}^2 \geq \sum_{i=1}^{N_e} \rho_{i\bar{x}}^2$$

Так как функция (5) подвергается минимизации, то утверждение леммы справедливо.

Таким образом, нами доказана следующая

ТЕОРЕМА I. Минимизация целевой функции (5) при ограничениях (3), (6), (7) и равенствах (4) эквивалентна минимизации критерия $\mathcal{D}^*$.

Данная задача есть задача целочисленного линейного программирования [4].

Для случая большого N сформулируем алгоритм, позволяющий достигать локальный минимум функции $\mathcal{D}^*$. Структура алгоритма ("процедуры γ") аналогична структуре многих итерационных методов таксономии [6].

Определим "процедуру γ" следующим образом: допустим, на i -м шаге процедуры осуществлено какое-то разбиение множества N объектов на K подмножеств и в каждом из этих подмножеств найден объект (эталон) a_q, сумма квадратов расстояний которого до всех остальных объектов данного подмножества минимальна. Тогда на $(i+1)$ -м шаге алгоритма происходит новое перераспределение всех N объектов по K таксонам: объект относится к тому подмножеству, к эталону которого a_q он наиболее близок. Для вновь полученных K подмножеств по вышеописанному правилу находятся новые эталоны a_q. Далее осуществляется $(i+2)$ - й шаг. Из самого алгоритма непосредственно вытекает

ТЕОРЕМА 2. Если $\mathcal{D}_i^*$ и $\mathcal{D}_{i+1}^*$ есть значения критерия качества таксономии $\mathcal{D}^*$ на i -м и $(i+1)$ -м шагах "процедуры γ", то справедливо соотношение $\mathcal{D}_{i+1}^* \leq \mathcal{D}_i^*$.

Хотя критерий $\mathcal{D}^*$ и "слабее" в статистическом смысле критерия $\mathcal{D}$, но имеет более широкую область применения, так как не требует, чтобы величины $\{\rho_{ij}\}$ обязательно удовлетворяли всем аксиомам метрического пространства: величина ρ_{ij} может просто численно выражать степень "различия" между объектами a_i и a_j. Кроме того, в некоторых задачах понятие центра "тяжести" (оценки математического ожидания) таксона физически просто не реализуемо.

Рассмотрим некоторые иные задачи, которые могут быть решены с помощью описанного метода.

Группировка признаков $\{X_1, \ldots, X_p\}$. Если здесь за объекты принять сами признаки, а в качестве расстояний между ними $\{\rho_{ij}\}$ взять $\{(1 - |r_{ij}|)\}$, где $|r_{ij}|$ - модуль коэффициента корреляций, то мы можем осу-

ществить разбиение всей совокупности исходных признаков X на K групп, в каждой из которых они между собой коррелируют достаточно сильно, а между группами слабо.

Минимизация описания объектов

Допустим, что каждый из ρ исходных признаков $\{X_1, \ldots, X_\rho\}$ имеет n_ℓ числовых градаций. Необходимо из имеющихся N градаций шкал ($\sum_{\ell=1}^{\rho} n_\ell = N$) выбрать те K градаций ($K < N$), при которых был бы минимум среднеквадратичного отклонения описания объектов в новом пространстве $\widetilde{X}$ от их описания в исходном пространстве X. Пусть всего M объектов. Если объект a_i имел по признаку X_j значение x_{ij}, то после удаления этой градации он примет значение $\widetilde{x}_{ij}$ ближайшей градации из неудаленных у X_j. Этого требует критерий минимума среднеквадратичного отклонения.

Если значение κ - й градации x_j^κ признака X_j имело m_κ из исходных объектов, а значение ℓ - й градации x_j^ℓ - m_ℓ объектов, то в качестве квадрата расстояния нужно взять:

$$\rho^2_{\kappa\ell} = m_\ell (x_j^\kappa - x_j^\ell)^2 \quad \text{и} \quad \rho^2_{\ell\kappa} = m_\kappa (x_j^\kappa - x_j^\ell)^2.$$

Расстояния между градациями различных признаков нужно положить равными ∞.

Справедливость подобной постановки вытекает из равенства

$$\sum_{i=1}^{M}\sum_{j=1}^{\rho} (x_{ij} - \widetilde{x}_{ij})^2 = \sum_{j=1}^{\rho}\sum_{i=1}^{M} (x_{ij} - \widetilde{x}_{ij})^2.$$

Ненулевые строки матрицы найденных неизвестных $\{z_{\kappa\ell}\}$ будут соответствовать тем K из исходных градаций пространства X, которые нужно оставить.

Квантование сигналов по уровню

Допустим, имеется M сигналов $y = f(t)$, заданных своими отсчетами по оси времени. Амплитудное значение при каждом отсчете времени может принимать одно из N значений. Нужно из N выбрать такие K градаций ($K < N$), чтобы среднеквадратичное отклонение семейства M сигналов от их исходных было

бы минимальным. Пусть число всех отсчетов по всем сигналам, имеющих значение градации y_κ, равно m_κ, а для градации y_ℓ - m_ℓ, соответственно. Тогда положим

$$\rho^2_{\kappa\ell} = m_\ell (y_\kappa - y_\ell)^2 \quad \text{и} \quad \rho^2_{\ell\kappa} = m_\kappa (y_\kappa - y_\ell)^2.$$

Аналогичным образом нетрудно убедиться, что решение этой задачи как задачи таксономии будет минимизировать вышеупомянутый критерий. Ненулевые строки матрицы $\{z_{\kappa\ell}\}$ будут соответствовать тем K из N исходных градаций уровня y, которые нужно оставить.

§ 2. Оптимизация дискриминантного критерия

Как и ранее, будем предполагать, что N объектов заданы своим описанием в ρ - мерном признаковом пространстве $\{X_1, \ldots, X_\rho\}$. Необходимо разбить N объектов на два класса A и B так, чтобы максимизировать значение критерия Фишера

$$F = \frac{(\bar{y}_a - \bar{y}_b)^2}{\sigma_a + \sigma_b} \qquad (8)$$

по наилучшему дискриминантному вектору $y = \alpha_1 X_1 + \ldots + \alpha_\rho X_\rho$. Величины $\bar{y}_a$ и $\bar{y}_b$ есть центры "тяжести" проекции выборки классов A и B на вектор y, а σ_a и σ_b - оценки соответствующих дисперсий по вектору y.

Допустим, имеется произвольный вектор $y = \alpha X$. Значения проекций N исходных объектов (точек) на y равны y_1, y_2, ..., y_N, соответственно.

Сначала рассмотрим одномерную задачу установления такого порога y^0 (границы между классами) на фиксированном векторе y, который бы доставлял $F = F_{max}$. Данную задачу будем решать методом "ветвей и границ" [4].

Относительно любого порога y^i расположение классов будем считать следующим:

$$y_\ell < y^i;\ y_\ell \in A \quad \text{и} \quad y_\kappa \geq y^i;\ y_\kappa \in B.$$

Допустим, для произвольных порогов y^i и y^j у нас вычислены следующие характеристики $\langle F^i, \bar{y}_a^i, \bar{y}_b^i, \sigma_a^i, \sigma_b^i, N_a^i \rangle$ и $\langle F^j, \bar{y}_a^j, \bar{y}_b^j, \sigma_a^j, \sigma_b^j, N_a^j \rangle$, соответственно. N_a^i - объем выборки

класса A при границе y^i на векторе Y, $N_б^i = N - N_a^i$. Пусть $F^i < F^j$ и $y^i < y^j$.

Вычислим верхнюю границу значения F_{ij} для порога на отрезке (y^i, y^j).

Между порогами y^j и y^i находится $(N_a^j - N_a^i)$ точек. Нетрудно убедиться в том, что величина F достигнет максимума на отрезке (y^i, y^j), если все $(N_a^j - N_a^i)$ этих точек будут иметь одну и ту же координату, почти совпадающую с y^i.

Тогда верхняя граница значения F_{ij} для порога на отрезке (y^i, y^j) вычисляется так:

$$\hat{F}_{ij} = (\hat{\mu}_a - \hat{\mu}_б)^2 / (\hat{\sigma}_a + \hat{\sigma}_б),$$

где:

$$\hat{\mu}_a = \frac{1}{N_a^j}(N_a^i \cdot \bar{y}_a^i + (N_a^j - N_a^i)\, y^i);$$

$$\hat{\mu}_б = \bar{y}_б^j; \quad \hat{\sigma}_a = \hat{\sigma}_{a\mu} + \hat{\sigma}_{ac}; \quad \hat{\sigma}_б = \sigma_б^j;$$

$$\hat{\sigma}_{a\mu} = (1/N_a^j)(N_a^i(\bar{y}_a^i - \mu_a)^2 + (N_a^j - N_a^i)(y^i - \mu_a)^2);$$

$$\hat{\sigma}_{ac} = (1/N_a^j)(\sigma_a^i N_a^i + 0);$$

$$\mu_a = (\bar{y}_a^i + y^i)/2 .$$

Если $(N_a^j - N_a^i) \leq 1$, то $\hat{F}_{ij} = \max(F^i, F^j)$. Координату нового порога y^q на отрезке $(y^i; y^j)$ можно выбирать пропорционально величинам F^i и F^j. Поскольку получено выражение для определения верхней границы значения критерия на отрезке и дано правило дробления отрезков, то возможно применение метода "ветвей и границ". В силу конечности выборки решение оптимизационной одномерной задачи может быть получено точное.

Перейдем теперь к решению задачи в многомерном пространстве.

Допустим, для произвольного вектора Y_i мы решили одномерную задачу таксономии, получив множества $\{A_i, B_i\}$ и значение критерия F_i. Для множеств A_i и B_i в пространстве X найдем дискриминантный вектор Y_{i+1}, максимизирующий критерий F. Вектор Y_{i+1} целесообразно находить методом, описанным в работе [5].

Для полученного вектора Y_{i+1} решим одномерную задачу таксономии и получим $F_{i+1} \longleftrightarrow \{A_{i+1}, B_{i+1}\}$.

Легко может быть доказана

ТЕОРЕМА 3. Если выполняется неравенство $Y_{i+1} \neq Y_i$, то справедливо соотношение $F_{i+1} > F_i$.

Таким образом, имеем сходимость к экстремуму функционала F.

На основе вышеописанной процедуры и процедуры работы [4] мы несложным образом можем осуществить таксономию на два класса N исходных объектов по вектору Y, ортогональному первому, и так далее до тех пор, пока значение качества таксономии F по вектору Y, ортогональному всем предыдущим, не станет достаточно плохим. Это можно использовать для общей таксономии N объектов на большее число классов.

§ 3. Минимизация исходной системы признаков

Необходимо из исходной системы r признаков $\{X_i\}$ выбрать такие m признаков $\{X_j\}$ $(m < r)$, в m - мерном пространстве которых качество группировки объектов на K таксонов было бы наилучшим. Это так называемая задача комбинированного типа "SX" [6]. Рассмотрим ее решение для двух частных случаев.

Дисперсионный критерий

Качество таксономии здесь будем измерять величиной критерия $\mathcal{D}^*$. Квадрат расстояния между i - м и j - м объектами, например, в евклидовой метрике измеряется таким образом:

$$\rho_{ij}^2 = \sum_{l=1}^{r} (\rho_{il}' - \rho_{jl}')^2 = \sum_{l=1}^{r} \rho_{ijl}^2 .$$

Если $t_{ij} = \rho_{ij}^2$, то $t_{ij} = \sum_{l=1}^{r} t_{ijl}$, где $t_{ijl} = \rho_{ijl}^2$.

Введем целочисленные переменные $x_{ijl} = \{0, 1\}$, где $i =$

$= 1,\dots, (N+1)$; $j = 1,\dots, (N+1)$; $\ell = 1,\dots, (\rho+1)$.

Неизвестные $z_{(N+1)j\ell}$ введены для контроля числа признаков $\{X_j\}$.

Теперь наша цель заключается в минимизации функционала

$$Q^* = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\ell=1}^{\rho} t_{ij\ell}\, z_{ij\ell} \qquad (9)$$

при ограничениях:

$$\left.\begin{aligned} &\sum_{i=1}^{N}\sum_{j=1}^{N+1} z_{ij\ell} - (N+K) z_{(N+1)\ell(\rho+1)} = 0; \\ &\ell = 1,\dots,\rho; \\ &\sum_{\ell=1}^{\rho} z_{(N+1)\ell(\rho+1)} = m; \end{aligned}\right\} \qquad (10)$$

$$\left.\begin{aligned} &\sum_{i=1}^{N}\sum_{\ell=1}^{\rho} z_{ij\ell} = b_j; \quad j = 1,\dots, N+1; \\ &b_j = m \ \text{ при } \ j = 1,\dots, N; \quad b_{(N+1)} = mK \end{aligned}\right\} \qquad (11)$$

$$\sum_{j=1}^{N}\sum_{\ell=1}^{\rho} z_{ij\ell} - N\sum_{\ell=1}^{\rho} z_{i(N+1)\ell} \le 0; \quad i = 1,\dots, N \qquad (12)$$

Трактовка переменных и ограничений аналогична задаче для критерия $\mathcal{D}^*$ второго параграфа.

Задача $<(9), (10), (11), (12)>$ есть задача целочисленного линейного программирования [3].

Для случая больших N и ρ может быть предложен такой алгоритм. Сначала задача таксономии решается в исходном ρ-мерном пространстве X. Заметим, что функционал (9) можно записать и в таком виде:

$$Q^* = \sum_{\ell=1}^{\rho} Q_\ell^*, \quad \text{где} \quad Q_\ell^* = \sum_{i=1}^{N}\sum_{j=1}^{N} t_{ij\ell}\, z_{ij\ell}$$

Далее, исходя из получившихся таксонов, для каждого из $\{X_\ell\}$ определим величину Q_ℓ^*. Так как величины Q_ℓ^* входят в Q^* чисто аддитивно, то "наилучшие" m признаков $\{X_j\}$ будут те, которые обладают меньшими величинами $\{Q_j^*\}$. Таким образом, здесь наилучшей подсистемой из m признаков будет та, которая вносит "наилучший" вклад в оптимизацию критерия качества

$Q^*(\mathcal{D}^*)$ таксономии во всем исходном признаковом пространстве.

Дискриминантный критерий

Допустим, в ρ-мерном пространстве $\{X_\ell\}$ на основе методики, изложенной в предыдущем параграфе, получены разбиения объектов на множества A и B по n ортогональным дискриминантным функциям $\{Y_1, \dots, Y_n\}$. Величина n выбирается такой, что качество таксономии F для $Y_{(n+1)}$ становится уже достаточно "плохим".

Этот факт можно трактовать как существование в исходном пространстве n возможных "различных" вариантов таксономии объектов на два класса A и B. Так как каждый Y_ℓ из полученных векторов $\{Y_1, \dots, Y_n\}$ обладает качеством таксономии по нему, равным F_ℓ, и является "наилучшим" собственным вектором некоторой матрицы $H_{(\ell)}$, то мы для выбора оптимальной m-мерной подсистемы $\{X_j\}$ можем воспользоваться алгоритмом, изложенным в работе [4].

В заключение автор выражает благодарность Н.Г. Загоруйко за обсуждение данной работы.

Литература

1. ДОРОФЕЮК А.А. Алгоритмы автоматической классификации (обзор). –"Автоматика и телемеханика", М., 1971, № 12.

2. ЛБКИНА В.Н., ЗАГОРУЙКО Н.Г. Количественные критерии качества таксономии и их использование в процессе принятия решений. –"Вычислительные системы", Новосибирск, "Наука", 1969, вып. 36.

3. КОТИКОВ В.И. Об одном методе таксономии множества объектов и параметров. –"Вычислительные системы", Новосибирск, 1971, вып. 46.

4. КОРБУТ А.А., ФИНКЕЛЬШТЕЙН Ю.Ю. Дискретное программирование. М., "Наука", 1969.

5. КОТИКОВ В.И. Оптимизация критерия Фишера-Уилкса и сокращение исходной системы описания в задачах распознавания образов. –Настоящий сборник, стр. 136–142.

6. ЗАГОРУЙКО Н.Г. Одновременный поиск эффективной системы признаков и наилучшего варианта таксономии (алгоритм " "). –"Вычислительные системы", Новосибирск, "Наука", 1969, вып. 36.

Поступила в ред.-изд.отд.
1. 1. 1972 г.