ЭМПИРИЧЕСКОЕ ПРЕДСКАЗАНИЕ И РАСПОЗНАВАНИЕ ОБРАЗОВ
(Вычислительные системы)
1978 год Выпуск 76

УДК 578.087.I:5I9.9

ТАКСОНОМИЯ В АНИЗОТРОПНОМ ПРОСТРАНСТВЕ

И.Г.Загоруйко

I. Процесс таксономии состоит в разбиении множества объектов $M = \{a_1, a_2, \ldots, a_L\}$ на некоторое число подмножеств (таксонов) $S = \{S_j\}$, $j = \overline{1,k}$. При этом из всех возможных вариантов разбиения выбирается тот, при котором некоторый функционал F качества таксономии достигает экстремального значения. Функционал F определяется так, чтобы выбранная по нему таксономия отвечала требованию минимума потерь $\rho(a_p^j, a_q^j)$, связанных с использованием одних объектов a_p^j таксона S_j вместо других объектов a_q^j того же таксона.

В распознавании образов, например, множество M_0 объектов обучающей выборки стремятся заменить "типичными", "эталонными" представителями каждого образа. Характеристиками этих объектов-заменителей потом пользуются при построении решающих функций, и если эталоны выбраны хорошо, то потери, возникающие из-за ошибок распознавания по этим эталонам, бывают относительно небольшими.

При таксономии объектов, не имевших предварительного разбиения на образы, т.е. в процессе первичной классификации, потери ρ связываются с возможностью замены всех объектов одного таксона одним определенным (или любым) объектом этого же таксона в качестве эталона в тех или иных будущих задачах распознавания.

Вопрос о том, будет ли хороший заменитель известных объектов таким же хорошим заменителем и всех других объектов генеральной совокупности, сводится к нерешенной пока проблеме о представительности обучающей выборки. Здесь мы не будем обсуждать этот вопрос, так же как и то, почему в реальных условиях таксономия, выполненная по тому или иному критерию, не связанному явно с конкретной задачей распознавания, т.е. таксономия "без суперцели", оказывается обычно такой же или почти такой же, как и таксономия "с суперцелью".

Ответ на этот вопрос надо было бы искать в виде закономерных связей между причинами, по которым разные исследователи, работая в одной и той же области, как бы находятся в одинаковых рамках, ограничивающих произвол (в одинаковых "фреймах"), в результате чего они выбирают почти одинаковые характеристики для описания объектов, задают похожие критерии таксономии объектов по этим характеристикам и ставят похожие задачи распознавания.

В данной работе мы коснемся лишь того, какие критерии используются в таксономии и как их применять при некоторых специфических свойствах пространства характеристик изучаемых объектов.

2. Если считать, что таксономия делается в предположении, что роль объекта заменителя a_p^j может переходить к любому из l_j объектов таксона S_j, то требование "наилучшего замещения" будет выглядеть так: суммарные потери R при всех вариантах выбора объекта-заменителя внутри каждого таксона S_j должны быть минимальными, т.е.

$$R_1 = \min \sum_{j=1}^{k} \sum_{p,q=1}^{l_j} \rho(a_p^j, a_q^j) \ldots \quad .$$

Если в качестве заменителя выбирается один определенный "центральный" объект a_0^j, то естественно потребовать, чтобы в таксоны S_j объединялись такие объекты a_q^j, при замене которых на объект a_0^j суммарные потери были минимальными:

$$R_2 = \min \sum_{j=1}^{k} \sum_{q=1}^{l_j} \rho(a_0^j, a_q^j) \ldots$$

Можно потребовать также, чтобы потери при замене самого далекого от центра таксона объекта $a_{q^*}^j$ на его "центральный" объект a_0^j были бы минимальными

$$R_3 = \min \sum_{j=1}^{k} \rho(a_0^j, a_{q^*}^j).$$

Иногда нужно делать "эстафетные" замены, т.е. иметь возможность переходить от объекта a_q^j к другому объекту a_p^j того же таксона S_j, делая любые промежуточные замены, но так, чтобы максимальные потери, которые могут возникнуть на любом шаге этой последовательности замен, были минимальными. Если последовательность переходов организована наилучшим образом, то качество разбиения будет зависеть от потерь ρ в максимальном звене t этой

последовательности в каждом таксоне, т.е. от ρ_k^j , и критерий разбиения на таксоны будет выглядеть так:

$$R_4 = \min \sum_{j=1}^{k} \rho_k^j \ .$$

При изменении числа таксонов k потери будут также меняться, причем рост k приводит к уменьшению R, но при этом будут возрастать потери Q, связанные с хранением и использованием информации о таксонах. Если стоимость потерь этого рода для одного j-го таксона равна c_j, то $Q = \sum_{j=1} c_j$ и наилучший вариант таксономии находится по критерию $F=\min(R+Q)$.

3. Если объекты множества M описаны n признаками в сильных шкалах (не слабее шкалы интервалов), то каждый объект можно представить точкой в n-мерном метрическом пространстве X_n и потери $\rho(a_p,a_q)$ можно вычислять как некоторую функцию f от расстояния r между точками a_p и a_q в этом пространстве.

Часто общие потери, возникающие из-за различия двух объектов по n свойствам, представляют собой сумму потерь из-за различия по каждому свойству в отдельности, так что потери $\rho(a_p,a_q)$ равны взвешенному расстоянию между точками a_p и a_q в n-мерном пространстве типа l_1:

$$\rho(a_p,a_q) = \sum_{i=1}^{n} \beta_i f(x_i^p - x_i^q),$$

где β_i – относительная стоимость (вес) единичных потерь по каждой координате.

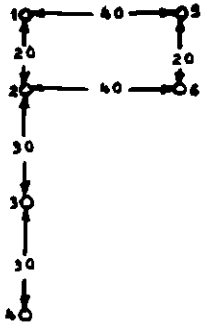

Рис. I

Существуют критерии, использующие не только меру близости между точками в таксонах, но также расстояние между таксонами, равномерность количества объектов в таксоне и т.п. [I].

Проиллюстрируем введенные выше критерии R_1–R_4 на примере. Пусть будет дано множество $M = \{a_1,\dots,a_6\}$ на плоскости (x_1, x_2), которое требуется разделить на 2 таксона (рис.I). По критерию R_1 нужно найти разбиение, при котором достигает минимума сумма ребер k полных подграфов, соединяющих все точки в каждом

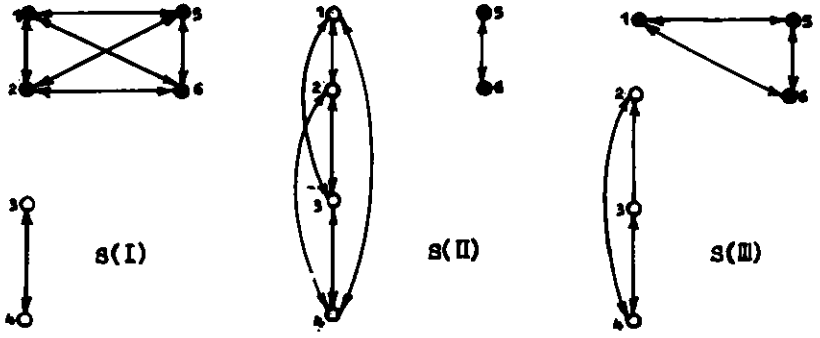

Рис. 2

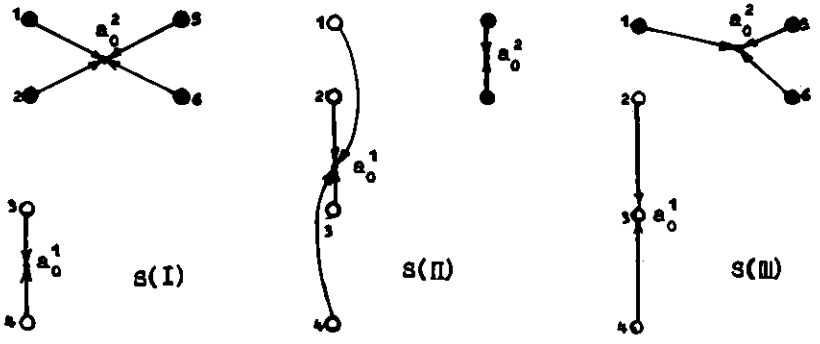

Рис. 3

таксоне. На рис.2 показаны три конкурирующих варианта таксономии: S (I), S (II) и S(III). По критерию R_1 предпочтение будет отдано варианту S (III).

Критерий R_2 эквивалентен минимуму суммы длин ребер k звездных графов, центральные вершины которых a_0^j лежат в центрах тяжести каждого из таксонов (рис.3). По этому критерию лучшей будет считаться таксономия S (II).

Использование критерия R_3 эквивалентно разбиению по минимуму суммы длин радиусов гиперсфер с центрами в точках a_0^j, равноудаленных от крайних точек таксона *)(см.рис.4). Этот критерий

*) В метрике l_2 критерий R_3 эквивалентен критерию, используемому алгоритмом "Форель" [2].

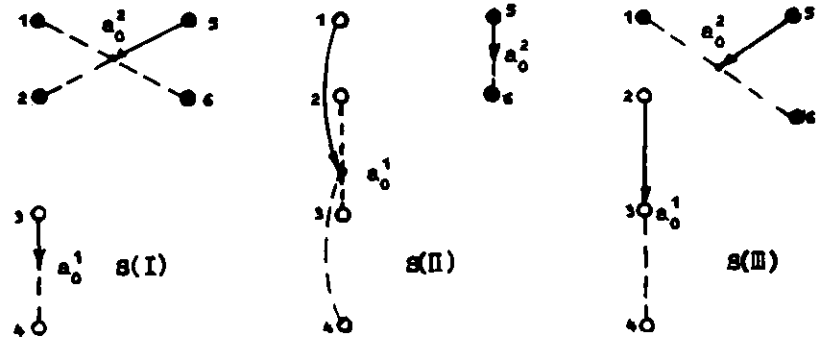

Рис. 4

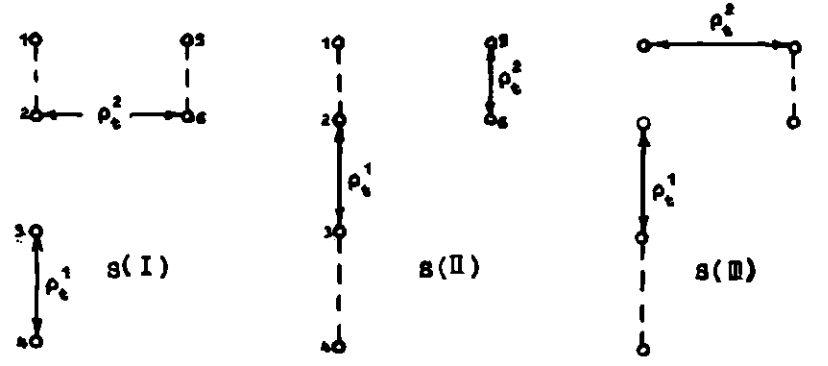

Рис. 5

предпочитает вариант s(I). При использовании критерия R_4 делается разбиение, которое минимизирует сумму самых длинных ребер кратчайшего незамкнутого пути, соединяющего точки в каждом таксоне (рис.5). По этому критерию лучшей считается таксономия s(II).

4. Вид и смысл требований R_1 R_2 R_3 и R_4 не изменяются и в анизотропном пространстве, т.е. когда $r(p,q) \neq r(q,p)$, однако результаты таксономии при этом будут другими. Пусть анизотропия пространства X_n задана коэффициентами $\alpha_i(p,q)$, $i = 1,\ldots, n$, показывающими, во сколько раз i-я компоненте потерь при замене p на q больше, чем при замене q на p, если $x_p^i > x_q^i$. Положим в нашем примере $\alpha_1(p,q) = 2$, $\alpha_2(p,q) = 10$. Тогда при замене точки a_5 на a_1 потери в 2 раза будут большими, чем при замене a_1 на a_5, т.е. $\rho(a_1,a_5) = 2\rho(a_5,a_1)$. Аналогич-

но $\rho(a_1,a_2) = 10\rho(a_2,a_1)$. Потери от замены объекта q на объект p будут равны

$$\rho(p,q) = \sum_{i=1}^{n} [\alpha_i(p,q)\ \beta_i f(x_{iq}^p - x_i^q)] .$$

В табл.I показаны величины потерь для приведенного выше примера при всех попарных заменах точек множества M друг на друга для значений $\beta_1=\beta_2=1$, $f_i(x_i^p - x_i^q) = |x_i^p - x_i^q|$, $\alpha_1 = 2$ и $\alpha_2 = 10$.

Таблица I

Потери при попарном замещении точек M
в анизотропном пространстве

Объекты		Объекты-заместители (a_p)					
		p = I	p = 2	p = 3	p = 4	p = 5	p = 6
Замещаемые объекты (a_q)	q = 1	0	200	500	800	40	240
	q = 2	20	0	300	600	60	40
	q = 3	50	30	0	300	90	70
	q = 4	80	60	30	0	I20	I00
	q = 5	80	280	580	880	0	200
	q = 6	I00	80	380	680	20	0
	q = 7	330	650	I790	3260	330	650

В этом пространстве по всем критериям $R_1 - R_4$ предпочтение отдается таксономии S(I). Величины потерь R_i для разных вариантов таксономии приведены в табл.2.

Таблица 2

Суммарные потери при разных вариантах таксономии

Критерии	$\rho(p,q) = \rho(q,p)$ $\alpha_1 = \alpha_2 = 1$			$\rho(p,q) \neq \rho(q,p)$ $\alpha_1 = 2,\ \alpha_2 = 10$		
	S(I)	S(II)	S(III)	S(I)	S(II)	S(III)
R_1	480	580	<u>450</u>	<u>I690</u>	3I90	I800
R_2	I50	<u>I20</u>	I43	<u>234</u>	267	285
R_3	<u>45</u>	50	60	<u>72</u>	I34	I20
R_4	70	<u>50</u>	70	<u>380</u>	500	500

5. При большом числе объектов L в пространстве большой размерности n поиск точного решения задачи таксономии связан с перебором и оценкой большого числа вариантов. Для сокращения перебора можно воспользоваться аналогами алгоритмов направленного перебора - "группировки" [2,3] для критериев R_1, R_2 и R_3 и "Краба" [1] для критерия R_4.

Введем понятия, аналогичные тем, которые используются в алгоритме "группировка".

В н у т р е н н и м и потерями D_q^j назовем потери, с которыми связано наличие элемента a_q в j-м таксоне. Для критерия R_1 потери D_q^j есть сумма длин ребер звездного графа, проходящих от вершины a_q ко всем другим вершинам j-го таксона. Для R_2 потери представляют собой длину ребра от вершины a_q^j к центральной вершине a_0^j j-го таксона. При использовании критерия R_3 величина D_q^j равна длине ребра от самой "периферийной" точки a_{q*}^j до центра j-го таксона a_0^j и $D_p^j = 0$ для всех других точек этого таксона.

В н е ш н и м и потерями Q_q^{jh} будем называть потери, связанные с заменой объекта a_q^j j-го таксона на соответствующие объекты h-го таксона. При работе с критерием R_1 потери Q_q^{jh} равны сумме длин ребер от вершины a_q^j до всех вершин h-го таксона, при критериях R_2 и R_3 потери Q_q^{jh} равны длине ребра от вершины a_q^j до центра h-го таксона.

О б щ и м весом B_p объекта a_p будем называть суммарную длину ребер от всех вершин множества M к вершине a_p. В табл. I вес B_p есть сумма элементов p-го столбца.

Для сокращения перебора процедура группировки начинается с выбора ядер групп. Я д р а м и k групп будем называть такие k объектов, которые максимально удалены друг от друга и имеют малый общий вес B_p. Если нужно найти два ядра, то следует для каждой пары объектов p и q вычислить величину

$$\Delta_{pq} = \frac{\rho_{pq} + \rho_{qp}}{B_p + B_q} = \frac{\rho^*_{pq}}{B^*_{pq}}$$

и выбрать объекты с наибольшим значением Δ_{pq}. Аналогично для трех ядер

$$\Delta_{pqs} = \frac{\rho_{pq} + \rho_{qp} + \rho_{ps} + \rho_{sp} + \rho_{qs} + \rho_{sq}}{B_p + B_q + B_s} = \frac{\rho^*_{pqs}}{B^*_{pqs}}.$$

В общем случае k ядер

$$\Delta_{a_1,a_2,\ldots,a_k} = \frac{\rho^*_{a_1,a_2,\ldots,a_k}}{B^*_{a_1,a_2,\ldots,a_k}} .$$

К одному из выделенных ядер присоединяется самый близкий элемент a_q, т.е. такой, для которого выполняется условие

$$D^j_q = \min_{\substack{q \in M/k \\ j=1\div k}} D^j_q .$$

Для остальных элементов процедура присоединения выполняется по этому же условию. После каждого очередного присоединения для всех объектов, входящих в состав уже образованных таксонов, производится вычисление величины $N_q = \frac{D^j_q}{D^h_q}$ для всех $j=1\div k$ и $h=1\div k$. Если $N_q > 1$, то это означает, что внутренние потери для элемента q больше внешних и его следует передать в таксон h. В работе [2] показано, что такой текущий контроль и перегруппировка гарантируют оптимальное решение задачи таксономии даже для случая неудачного (случайного) выбора первичных ядер групп.

При использовании критерия R_k для нахождения k таксонов нужно построить кратчайший незамкнутый путь [1] и изъять из него k - 1 самых длинных ребер.

Литература

1. ЁЛКИНА В.Н., ЗАГОРУЙКО Н.Г. Количественные критерии качества таксономии и их использование в процессе принятия решений. - В кн.: Вычислительные системы. Вып.36. Новосибирск, 1969, с. 29-46.

2. ЗАГОРУЙКО Н.Г., ЁЛКИНА В.Н. Алфавит с минимальной избыточностью. - В кн.: Вычислительные системы. Вып.28. Новосибирск, 1967, с.49-58.

3. ЗАГОРУЙКО Н.Г. Методы распознавания и их применение. М., "Сов.радио", 1972.

Поступила в ред.-изд.отд.
10 июля 1978 года