

ЭЛН 62-5: 07:621.391:519.2

КОЛИЧЕСТВЕННЫЕ КРИТЕРИИ КАЧЕСТВА ТАКСОНОМИИ
И ИХ ИСПОЛЬЗОВАНИЕ В ПРОЦЕССЕ ПРИНЯТИЯ РЕШЕНИЙ

В.Н. Елкина, Н.Г. Загоруико

§ I. Критерий качества таксономии

Процедура разбиения некоторого множества Z исходных данных в пространстве признаков X на K непересекающихся подмножеств S делается для получения более краткого описания множества Z за счет частичной потери информации об индивидуальных свойствах представителей этого множества. Если точно известно, в каких условиях будет использоваться этот более краткий алфавит S , то можно, пользуясь алгоритмом таксономии [1], получить группировку, оптимальную для этих условий (для этой "суперцели"). Когда "суперцель" неизвестна, тогда группировка делается "на все случаи жизни".

Информация об индивидуальных свойствах группируемых объектов (т.е. об элементах множества Z) при этом заменяется описанием средних свойств группы объектов. Чем меньше мы теряем при такой перекодировке, тем при прочих равных условиях выше качество таксономии.

Естественно стремиться к тому, чтобы качество таксономии F было как можно лучшим, а результат таксономии описывался бы по возможности наиболее просто (кратко). Если затраты, связанные с потерями качества F обозначить через λ , а затраты на описание результатов - через λ_2 , то задача таксономии сводится

к поиску такого варианта разбиения множества Z на K таксонов S , чтобы суммарные потери^{*} $N = N_1 + N_2$ были минимальными. Немного упростим эту задачу: сначала будем стремиться получить вариант таксономии, наилучший с точки зрения критерия качества F , а потом попытаемся найти наипростейшее описание полученных таксонов.

Цель данной работы состояла в определении критерия оценки качества таксономии при отсутствии "суперцели".

Если Z есть некоторый массив в двумерном пространстве (точки на плоскости), то таксономия этого множества, сделанная разными людьми "вручную", "на глаз", обычно дает один и тот же результат. Естественно потребовать, чтобы критерии автоматической таксономии совпадали с критериями, которыми пользуется человек при выполнении таксономии "вручную".

Для определения человеческих критериев качества таксономии группе испытуемых ("экспертов"^{**}) предъявлялись различные варианты расположения L точек массива Z на плоскости и предлагалось объединить эти точки в K таксонов. В ходе экспериментов выяснилось следующее:

1. Одни случаи эксперты называли "легкими", другие "трудными".

Максимально простым является случай, когда требуемое число таксонов (K) равно числу точек (L), а максимально сложным - случай (при $K < L$) равномерного распределения точек по плоскости. Реальные задачи находятся в промежутке между этими крайностями.

2. Результаты разных экспертов в простых случаях одинаковы, а в сложных - не всегда одинаковы.

3. Эксперты руководствуются некоторыми критериями "близости" (ρ) точек внутри таксона и "удаленности" (α) таксонов друг от друга.

4. Кроме этих критериев, эксперты учитывают также локальный характер распределения точек или "одинаковость распределения" (λ) точек внутри таксона.

* По аналогии с однопродуктовой моделью в математической экономике.

** Группа "экспертов" состояла из 10 человек различных специальностей (математики, инженеры, лингвисты), имеющих среднее и высшее образование.

5. При прочих равных условиях эксперты предпочитают вариант таксономии с одинаковым числом точек в каждом таксоне (h).

Были предприняты попытки установления способа оценки этих параметров - ρ, α, λ, h - и вида функции качества таксономии

$$F = f(\rho, \alpha, \lambda, h)$$

Правильность различных предположений проверялась на примерах, решаемых параллельно экспертами и на ЭВМ.

Выяснилось, что таксономические параметры удобно оценивать с помощью кратчайшего незамкнутого пути (КНП), соединяющего все L точек множества Z в связный неориентированный граф без петель, с минимальной суммарной длиной ребер [2,3] (см. рис. 1 и 2).

Обозначим через α - длину участка пути между двумя соседними точками, а через β - минимальный из участков, непосредственно примыкающих к рассматриваемому участку α .

Тогда мерой (ρ_j^*) "близости" g_j точек внутри j -го таксона можно считать величину

$$\rho_j^* = \frac{1}{g_j - 1} \sum_{i=1}^{g_j-1} \alpha_i$$

а мерой средней "близости" внутри K таксонов -

$$\rho = \frac{1}{K} \sum_{j=1}^K \rho_j^*$$

Чем меньше ρ , тем выше качество таксономии.

"Удаленность" K таксонов друг от друга можно измерять величиной

$$\alpha = \frac{1}{K-1} \sum_{j=1}^{K-1} z_j$$

где z_j - расстояние между таксонами, т.е. длина участка цепи, по которому проходит граница между таксонами.

В качестве меры "одинаковости распределения" точек внутри таксона оказалось возможным использовать величину

$$\lambda = \frac{1}{K-1} \sum_{j=1}^{K-1} \frac{h_j}{z_j}$$

Лучшим вариантом таксономии отвечают меньшие значения λ .

"Одинаковость" числа точек в таксонах можно считать величиной

$$h = K \sqrt[K]{\frac{L}{K}}$$

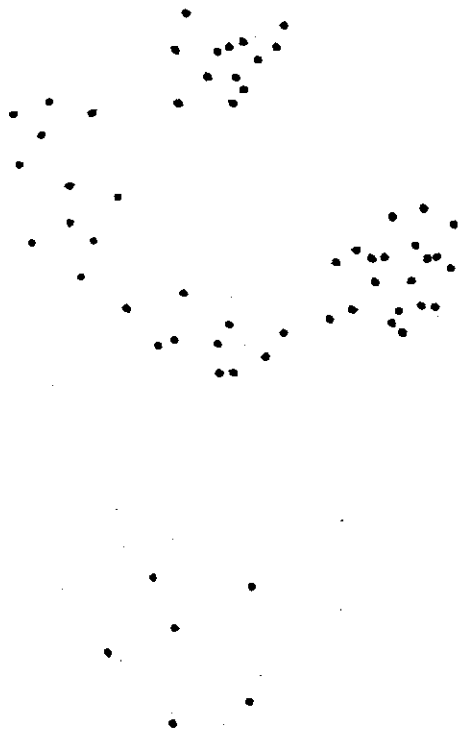


FIG. 1.

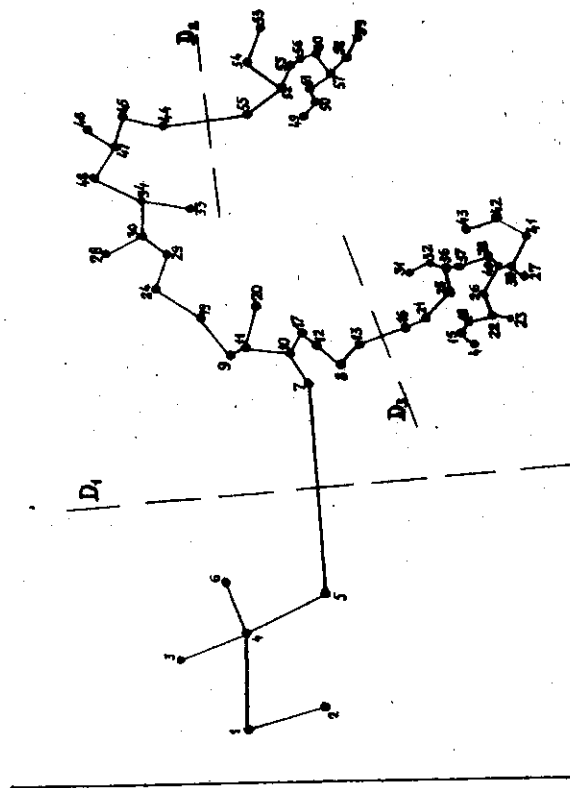


FIG. 2

Здесь K^k - нормирующий множитель, делающий величину h не зависящей от числа таксонов K и общего числа точек L . При одинаковом числе точек в каждом таксоне, т.е. при $L_j = L/K$ величина h достигает максимального значения, равного единице.

Количественной мерой общего качества таксономии в итоге выбора величина

$$F = L_n \frac{d^a h^c}{(1+\rho^m)(\lambda^{b+1})}$$

Точное значение коэффициентов a, b, c и m , характеризующих относительную значимость соответствующих параметров, устано-вить пока не удалось. Однако оказалось, что при $a=b=c=m=1$ функция F может служить хорошим критерием качества таксономии. Варианты решения, получившие большие значения F , признавались также лучшими всеми или подавляющим большинством экспертов.

§ 2. Алгоритмы таксономии

Функцию F можно использовать не только для сравнения качества результатов, полученных тем или иным способом, но также и для построения алгоритмов, которые давали бы сразу наилучший с точки зрения этого критерия F , вариант таксономии.

Одним из таких алгоритмов является алгоритм "Краб". В соответствии с этим алгоритмом для разделения множества Z , изображенного на рис. 1, на K таксонов нужно выполнить следующие операции:

1. Соединить все точки множества между собой кратчайшим незамкнутым путем (см. рис. 1). Эта цепочка соединений будет иметь $L-1$ участков.

2. Предполагая, что $K-1$ границ между K таксонами могут пройти по любому из $L-1$ участков, вычислить для каждого из 2^{L-1} вариантов величины ρ, α, λ, h и оценку качества F .

3. Выбрать тот вариант, которому соответствует F_{max} .

4. Построить упрощенный вариант описания точек, попавших в один таксон. Для этого можно воспользоваться, например, алгоритмом "дробящихся эталонов" [4], который позволяет описать таксон небольшим числом простых фигур, покрывающих точки только этого таксона.

В итоге будет получено наилучшее и наиболее простое решение задачи таксономии конкретного материала. Так, например, было получено решение задачи таксономии множества Z (рис. 1) на 4 таксона, показанное на рис. 3. Расположение границ между таксонами, найденное на ЭВМ и экспертами, было одинаковым. Однако легко видеть, что при большом числе исходных точек L этот алгоритм становится очень громоздким. Можно воспользоваться таким методом сокращения перебора, как метод СПА [5], но можно попытаться найти и более простые реализации описанного выше "идеального" алгоритма. Такие упрощения даются, естественно, ценой некоторого риска получения не самого лучшего варианта таксономии.

Рассмотрим два алгоритма таксономии, близких к оптимальным.

а) Алгоритм "Краб-1".

1) Вначале, так же как и в предыдущем случае, все точки соединяются между собой кратчайшим незамкнутым путем.

2) Для каждого участка пути вычисляется величина

$$\lambda_i = \frac{\beta_i}{\alpha_i}$$

3) Так как граница обычно проходит по участку более длиннее, чем соседние с ним, то из дальнейшего рассмотрения безболезненно можно взять m участков, для которых $\lambda_i < 1$ (в более общем случае - $\lambda_i < \epsilon$). В итоге остается L_1 участков- претендентов на проведение границ.

4) Для каждого из 2^{L_1-1} вариантов размещения границ вычисляются величины ρ, α, λ, h и по ним оценка качества F .

5) Выбирается тот вариант таксономии, которому соответствует F_{max} .

6) Строится наиболее простое описание областей, включающих в себя точки только одного таксона.

В примере, представленном на рис. 1, как по предыдущему полному алгоритму, так и по алгоритму "Краб-1" было получено одно и то же решение.

б) Другой разновидностью упрощения полного алгоритма является алгоритм "Краб-П". Здесь сокращение объема вычислений достигается за счет предварительной группировки точек массива Z на Q мелких простых таксонов, причем $L > Q > K$. (Алгоритм такой группировки "Форэль-1" описан в работе [6]). Далее с найденными таксонами можно обращаться, как с точками, объеди-

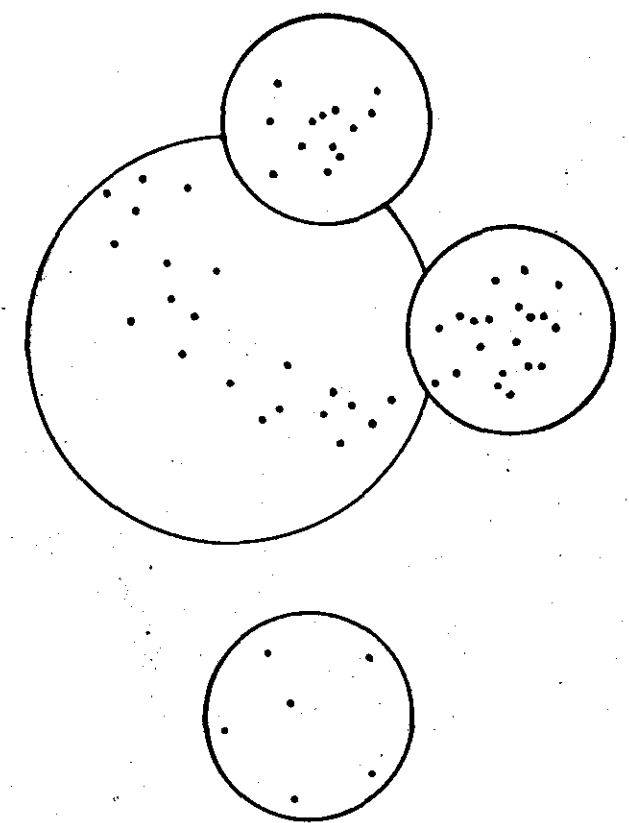


Рис. 3

няя их в цепочку. Расстояние между двумя таксонами S_i, S_j будем измерять как расстояние между множествами в метрическом пространстве:

$$d_{ij} = \inf d(q_m, q_l)$$

где q_m и q_l - произвольные точки из i -го и j -го таксонов.

После построения кратчайшего незамкнутого пути между Q таксонами множества Z , выполняются те же операции, что и в описанном выше "полном" алгоритме или в алгоритме "Краб-1".

И этот алгоритм дал решение того же примера (рис.1), совпадающее с оптимальным.

§ 3. Выбор числа таксонов

Так как критерий качества таксономии F нормирован по отношению к числу точек Z и числу таксонов K , его можно использовать для выбора наиболее естественного (с точки зрения "экспертов") числа таксонов.

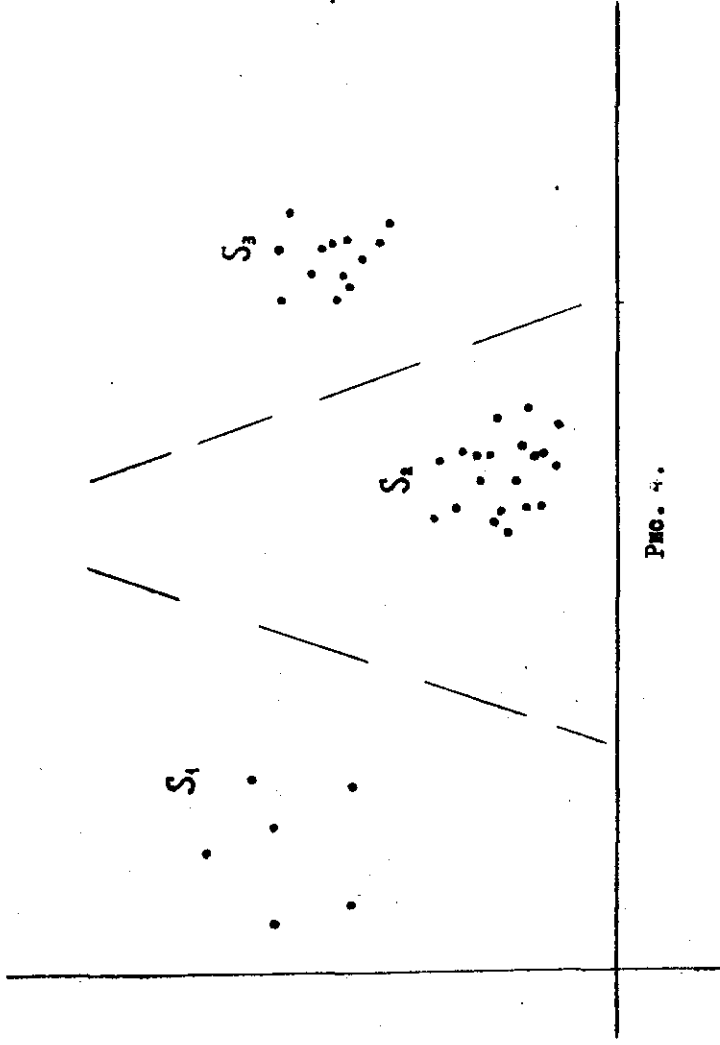
В табл. 1 приводятся значения величин α, ρ, λ, h и F для различных вариантов группировки массива Z (рис.1) на 2, 3, и 4 таксона. Максимальным значениям F соответствует варианты таксономии, предпочтительные, по мнению "экспертов".

Наибольшее значение $F = 0,53$ соответствует разбиению Z на два таксона - варианту, единодушно признанному экспертами наилучшим.

Функция F позволяет сравнить между собой качество таксономии различных множеств. Так, качество таксономии множества Z_1 , представленного на рис.4, оценивается величиной $F_1 = 1,440$, а Z_2 (рис.5) - величиной $F_2 = 0,243$.

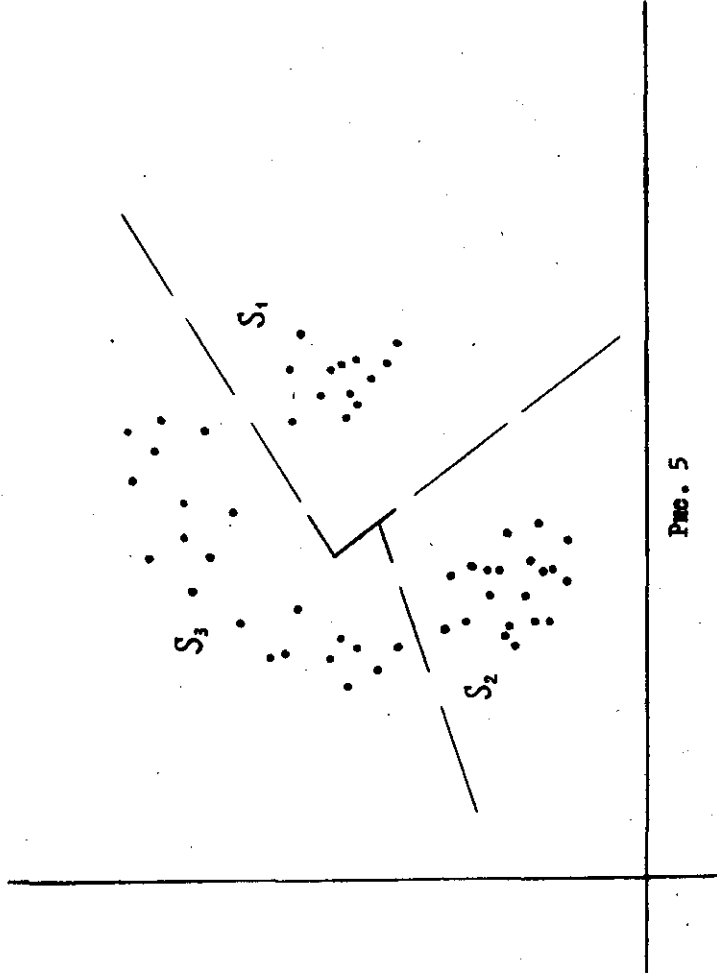
§ 4. Таксономические решающие функции

Критерий качества таксономии F можно использовать также и в качестве решающей функции в процессе распознавания принадлежности реализации q к тому или иному образу.



Pmo. 4.

• • • • •



Pmo. 5

Один из алгоритмов, использующих таксономические критерии в качестве решающей функции (алгоритм "ТФ-1"), состоит в следующем:

1. Для каждого образа строится кратчайший незамкнутый путь, соединяющий все точки его обучающей последовательности.

2. Исходя из предположения, что точка Q относится к образу i , строится кратчайший путь от точки Q до ближайшей точки i -го образа.

3. Строится участки α кратчайшего незамкнутого пути, соединяющего между собой все K образов (так же, как соединяются таксоны S_m в алгоритме "Крб-2", описанном в § 2).

4. Вычисляются параметры ρ_i , α_i , λ_i , h_i и функция качества F .

5. Процедуры по п.п. 2, 3, и 4 повторяют K раз, относя Q поочередно ко всем K образам.

6. Реализация X считается принадлежащей j -му образу, если $F_j = \sup F_i$, $i = 1 \div K$.

Если разрешается реализации Q не только относить к одному из K образов, но и считать её представителем $(K+1)$ -го образа (такое решение иногда называют "отказом" или "распознаванием с порогом"), то в алгоритме "ТФ-1" после п.5 делаются следующие процедуры:

6. Точка Q выделяется в $K+1$ образ. Для этого варианта снова строится кратчайший путь между образами и вычисляется оценка качества F .

7. Реализация X считается принадлежащей к j -му образу, если $F_j = \sup F_i$, где $i = 1 \div (K+1)$. Если количество точек обучающей последовательности ($L_{об} = \sum_{i=1}^K l_{(об)}i$) велико,

то может оказаться, что рациональнее хранить в памяти не все эти точки или соединяющие их кратчайшие пути, а описание поверхностей, отделяющих области принятия решения в пользу того или иного образа. Между образами i и j эта поверхность проходит по точкам Q , отношение которых к i -му и j -му образам дает одинаковые значения функции качества: $F_i = F_j$.

Второй алгоритм ("ТФ-2") удобно использовать в тех случаях, когда контрольные реализации предъявляются для распознавания не поштучно, а сразу группой из L контрольных штук. Тогда следует:

1. Объединить все $L_{об}$ точек обучающей ($Z_{об}$) и $L_{конт}$ точек контрольной ($Z_{конт}$) последовательностей в одно общее множество.

2. С помощью алгоритма "Крб" делается таксономия множества Z на K таксонов (K - число распознаваемых образов).

Объем вычислений можно существенно сократить, если границы искать только на участках пути, соединяющего точки из $Z_{об}$, принадлежащие разным образам.

3. Проверить распределение обучающих реализаций по таксонам. Если в таксоне имеются обучающие реализации только одного (i -го) образа, то и все контрольные реализации этого таксона также следует отнести к i -му образу.

4. Если в таксоне имеются представители K , разных образов, то следует точки этого таксона подвергнуть таксономии на K таксон.

5. Процедуры пп.4 и 4 повторяются до тех пор, пока ни в одном таксоне не будет точек из $Z_{об}$, принадлежащих разным образам.

6. Если выделится таксон, содержащий только контрольные точки, то эти точки следует считать либо представителями $(K+1)$ -го образа (распознавание "с порогом"), либо представителями того образа, присоединение к которому (по кратчайшему пути) даст наибольшее значение функции качества ρ (распознавание "без порога").

На рис. 6 показан пример, показывающий преимущество этого алгоритма перед любым алгоритмом, использующим в процессе построения решающей функции (α) только информации об обучающей последовательности (отмеченной крестиками). Штриховкой покрыты области ошибочных решений, получаемых с помощью функции α . Таксономические алгоритмы позволяют получить решение, совпадающее для дан -

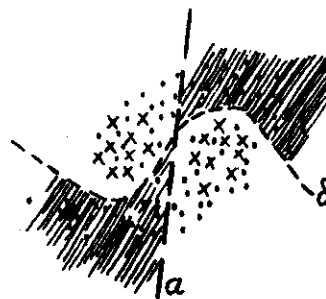


Рис. 6

ного примера с решением по оптимальной решающей функции δ .

Ясно, что границу α построить проще, чем найти решение по таксономическим критериям. Однако обычно групповое предъявление контрольного материала встречается в таких задачах (геология, социология и т.д.), при решении которых увеличение времени на принятие более точного решения вполне оправдано.

§ 5. Дополнительные замечания

Результаты таксономии в сильной степени зависят от метрики выборочного пространства. Один из способов обеспечения сопоставимости результатов решения разных задач состоит в нормализации масштабов измерения всех признаков. В настоящее время наиболее часто используется нормировка исходных данных по величине среднеквадратичного отклонения σ_j :

$$\sigma_j = \sqrt{\frac{1}{L} \sum_{i=1}^L (x_{ij} - \bar{x}_j)^2},$$

где \bar{x}_j - математическое ожидание значения j -го признака, а x_{ij} - значение j -го признака i -ой точки массива исходных данных.

Координаты точек в нормированном пространстве определяются по формуле

$$y_{ij} = \frac{x_{ij}}{\sigma_j}; \quad i=1, \dots, L; \quad j=1, \dots, n,$$

где L - число точек массива Z ,

n - размерность признакового пространства.

Если известны веса ("относительная значимость") β_j признаков, то нормирующий коэффициент равен $\frac{\beta_j}{\sigma_j}$.

Дальнейшие исследования критериев качества таксономии должны быть направлены на получение критерия, реагирующего не только на такие параметры, как описанные выше ρ , α , λ и h , но также и на форму (\mathcal{F}) получаемых таксонов: опыт показывает, что эксперты учитывают некоторый "фактор формы".

Таблица I

K	Граничные участки	ρ	α	λ	l_1	l_2	l_3	h	F
$K=2$	(1,2)	19,00	7,13	0,518	2	58	-	0,129	-1,61
	(2,3)	17,00	7,16	0,580	5	55	-	0,305	-0,91
	(3,4)	42,11	6,73	0,152	6	54	-	0,360	0,53
	(4,5)	9,49	7,29	0,473	20	40	-	0,890	-0,37
	(4,6)	9,22	7,30	0,343	25	35	-	0,972	-0,22
	(6,8)	10,77	7,27	0,650	18	42	-	0,840	-0,42
	(7,8)	17,12	7,16	0,456	13	47	-	0,680	-0,02
	(1,2) (2,3)	18,00	6,96	0,549	2	3	55	0,041	-2,80
	" (3,4)	30,56	6,51	0,335	2	4	54	0,054	-1,81
	" (4,5)	14,25	7,09	0,496	2	20	38	0,190	-1,85
	" (4,6)	14,11	7,09	0,431	2	33	25	0,206	-0,99
	" (6,8)	14,89	7,06	0,584	2	40	18	0,18	-1,53
	" (7,8)	18,89	6,92	0,487	2	45	13	0,146	-1,14
	(2,3) (3,4)	29,56	6,55	0,366	5	1	54	0,034	-2,33
" (4,5)	13,25	7,12	0,527	5	20	35	0,228	-1,68	
" (4,6)	13,11	7,13	0,462	5	30	25	0,469	-0,67	
" (6,8)	13,89	7,10	0,615	5	37	18	0,416	-0,81	
" (7,8)	17,06	6,99	0,518	5	42	13	0,342	-0,73	
(3,4) (4,5)	25,80	6,68	0,313	6	20	34	0,51	0,27	
" (4,6)	25,66	6,69	0,248	6	29	25	0,543	0,38	
" (6,8)	26,44	6,66	0,401	6	36	18	0,486	0,18	
" (7,8)	29,61	6,54	0,304	6	41	13	0,400	0,18	
(4,5) (4,6)	9,35	7,26	0,408	20	15	25	0,937	-0,28	
" (6,8)	10,13	7,23	0,562	20	22	18	0,990	-0,25	
(4,5) (7,8)	13,30	7,18	0,465	20	27	13			

Граничные участки	ρ	α	λ	l_1	l_2	l_3	l_4	h	F
(4,5) (7,8)	13,30	7,18	0,465	20	27	13		0,878	-0,02
(4,6) (6,8)	20,00	6,89	0,497	35	7	18		0,551	-0,07
(7,8)	13,17	7,13	0,399	35	12	13		0,683	-0,24
(6,8) (7,8)	13,94	7,10	0,553	13	42	5		0,341	-0,97
(1,2) (2,3)(3,4)	26,04	6,33	0,417	2	3	1	54	0,006	-5,08
(1,2) (2,3)(4,5)	15,16	6,89	0,524	2	3	20	35	0,083	-2,26
" (4,6)	15,07	6,92	0,480	2	3	30	25	0,089	-2,17
" (6,8)	15,59	6,89	0,583	2	3	37	18	0,079	-2,29
" (7,8)	17,71	6,77	0,518	2	3	42	13	0,065	-2,37
(1,2)(3,4)(4,5)	23,53	6,46	0,381	2	4	20	34	0,107	-1,41
(1,2)(3,4)(4,6)	23,44	6,47	0,334	2	4	29	25	0,115	-1,31
" (6,8)	23,96	6,44	0,440	2	4	36	18	0,102	-1,48
" (7,8)	26,08	6,33	0,375	2	4	41	13	0,084	-1,53
(1,2) (4,5)(4,6)	12,57	7,05	0,445	2	20	13	25	0,257	-1,31
" (6,8)	13,09	7,02	0,547	2	20	20	18	0,284	-1,08
" (7,8)	15,20	6,91	0,482	2	20	25	13	0,257	-1,10
(1,2) (4,6)(6,8)	13,00	7,03	0,504	2	33	7	18	0,164	-1,74
(1,2) (4,6)(7,8)	15,11	6,91	0,439	2	33	12	13	0,203	-1,31
(1,2) (6,8)(7,8)	15,63	6,89	0,541	2	40	5	13	0,103	-2,01
(2,3) (3,4)(4,5)	22,87	6,50	0,402	5	1	20	34	0,067	-1,91
" (4,6)	22,78	6,50	0,358	5	1	29	25	0,072	-1,81
" (6,8)	23,29	6,47	0,461	5	1	36	18	0,064	-1,99
" (7,8)	22,08	6,54	0,376	5	1	41	13	0,053	-2,16
(2,3)(4,5) (4,6)	11,90	7,09	0,465	5	20	10	25	0,494	-0,70
" (6,8)	12,42	7,06	0,587	5	20	17	18	0,604	-0,53
" (7,8)	14,54	6,94	0,503	5	20	22	13	0,565	-0,38

Граничные участки	ρ	α	λ	l_1	l_2	l_3	l_4	h	F
(2,3)(4,6)(6,8)	12,33	7,06	0,524	5	30	7	18	0,320	-0,98
" (7,8)	14,45	6,95	0,480	5	30	12	13	0,462	-0,43
(2,3)(6,8)(7,8)	14,96	6,92	0,562	5	37	5	13	0,238	-1,25
(3,4)(4,5)(4,6)	20,27	6,64	0,323	6	20	9	25	0,533	-0,07
" (6,8)	20,79	6,60	0,428	6	20	16	18	0,683	-0,27
" (7,8)	22,91	6,50	0,360	6	20	21	13	0,647	0,38
(3,4)(4,6)(6,8)	20,70	6,61	0,382	6	29	7	18	0,361	-0,17
" (7,8)	22,82	6,50	0,317	6	29	12	13	0,536	-0,21
(3,4)(6,8)(7,8)	23,33	6,77	0,419	6	36	5	13	0,277	-0,49
(4,5)(4,6)(6,8)	9,83	7,20	0,489	20	15	7	18	0,746	-0,51
" (7,8)	11,94	7,08	0,424	20	15	12	13	0,924	-0,04
(4,5)(6,8)(7,8)	12,46	7,06	0,526	20	22	5	13	0,564	-0,56
(4,6)(6,8)(7,8)	12,37	7,06	0,483	35	7	5	13	0,315	-1,12

Л и т е р а т у р а

1. Н.Г. ЗАГОРУЙКО, В.Н. ЕЛКИНА. Алфавит с минимальной избыточностью. - Вычислительные системы, Новосибирск, Изд-во "Наука", Сибирское отделение, 1967, вып.22
2. O. BORUVKA, On a minimal problem. Prace Moravské Přírodovědecké společnosti, v. 3, (1966).
3. В.Н. ЕЛКИНА. Программа построения кратчайшего незамкнутого пути. Отчет ИМ СО АН СССР, 1968.
4. Н.Г. ЗАГОРУЙКО. Структура проблемы распознавания слуховых образов и методы её решения. Сб. "Распознавание слуховых образов", Изд-во "Наука", Новосибирск, 1966.
5. Г.С. Дюв. Выбор эффективной системы зависимых признаков. Вычислительные системы, Новосибирск, Изд-во "Наука", Сибирское отделение, 1965, вып. 19.
6. В.Н. ЕЛКИНА, Н.Г. ЗАГОРУЙКО. Об алфавите объектов распознавания. - Вычислительные системы, Новосибирск, Изд-во "Наука", Сибирское отделение, 1966, вып.22.

Поступила в редакцию
5.I.1969г.