

ПАВЕЛ П. ГАМБАРЯН

СРАВНЕНИЕ 10 ТИПИЧНЫХ МЕТОДОВ ЧИСЛОВОЙ ТАКСОНОМИИ

Изучение сравнительной эффективности разных методов числовой таксономии (ЧТ) в биологии затрудняется отсутствием эталонных таксонов, таких бесспорных, как таблица Менделеева для классификация элементов. Сходство биологических объектов объясняется их общим происхождением, а сходство химических элементов нельзя объяснить тем же. Для сравнения эффективности разных методов мною [5] предложено использовать отношение межгрупповых дисперсий оценок сходства к внутригрупповым. Чем больше это отношение, тем эффективнее метод. При модели со случайным распределением признаков чем больше отношение дисперсий, тем метод хуже. Наконец, чем проще метод в вычислительном отношении, тем он лучше.

Задача сравнения эффективности разных методов ЧТ очень сложна [1]. Моделирование таксонов, адекватных реальным, невозможно, а сравнение разных методов на реальных объектах чрезвычайно трудно, да и ответ неопределен. Именно поэтому упрощение проблемы и использование дисперсионного отношения может быть оправдано. К сожалению, отношение дисперсий нельзя применить для оценки существенности классификации, так как модель со случайным распределением признаков при группировке по максимальным, особенно по максимальным средним арифметическим оценкам сходства, дает существенное отношение дисперсий. Распределение оценок сходства (без оценок сходства объектов самих с собой) при большом числе объектов и признаков должно быть приблизительно нормальным, и для оценки существенности классификации можно пользоваться средним квадратическим.

Мы используем простейшие оценки сходства, равные сумме весов совпадающих признаков, кроме метода Смирнова [6]. Вычисления проводились с четырьмя значащими цифрами, результаты приведены с двумя значащими цифрами.

В ЧТ оценка признака связана с его частотой. В работе исследовано 10 методов (табл. 1). 1—с равной оценкой признаков [8], 2—с оценкой признака его частотой, 3—оценка признака частотой его антитезы, 4—оценка признака логарифмом частоты тезы $\eta p = -p \log_2 p$, 5—оценка признака логарифмом частоты антитезы $\eta q = -b \log_2 q$, 6—по методу Смирнова [6], 7—с оценкой признака величиной $H(x) = \eta p + \eta q$, т. е. количеством информации [2], 8—частотой несовпадений pq , 9—так как величина $H(x)$ коррелирует с дисперсией биномиального признака, использована удобная величина $2-H(x)$, которая тем больше, чем меньше дисперсия; 10—так как pq —это дисперсия биномиального признака, использована удобная величина, обратно пропорциональная дисперсии, $0,5-pq$. Эти

Таблица 1

Отношение межгрупповых дисперсий оценок сходства к внутригрупповым, полученных для десяти моделей десятью различными способами

№ метода	Оценка совпадения двух ОТЕ		Отношения межгрупповых дисперсий к внутригрупповым дисперсиям оценок сходства									
			№ модели									
	по тезе	антитезе	1	2	3	4	5	6	7	8	9	10
1	I	I	8	12	11	17	17	20	16	7,3	40	23
2	p	q	8	12	11	17	19	20	17	7,3	27	16
3	q	p	8	11	13	17	15	20	16	7,3	48	32
4	rp	rq	10	17	16	29	22	33	26	7,3	40	30
5	rq	rp	10	17	16	29	25	33	27	7,3	50	30
6	q/p	p/q	6	8	7	9	9	12	8	8,5	40	26
7	H(x)	H(x)	10	17	17	29	25	36	28	7,3	48	27
8	pq	pq	11	19	19	33	28	39	32	7	50	32
9	2-H(x)	2-H(x)	7	9	8	11	12	14	11	7,1	32	25
10	0,5-rq	0,5-rq	7	9	8	11	11	12	11	5,2	34	24

десять методов далеко не исчерпывают всех возможных оценок признака в связи с его частотой, однако охватывают все тенденции таковых оценок.

Исследовались следующие модели таксонов:

- 1 4 ОТЕ, разделенные на 2 группы по 2 ОТЕ в группе,
- 2 5 ОТЕ, разделенные на 2 группы по 2 и 3 ОТЕ в группе,
- 3 5 ОТЕ, разделенные на 3 группы по 1, 2 и 2 ОТЕ в группе,
- 4 6 ОТЕ, разделенные на 2 группы по 3 ОТЕ в группе,
- 5 6 ОТЕ, разделенные на 2 группы по 2 и 4 ОТЕ в группе,
- 6 6 ОТЕ, разделенные на 3 группы по 2 ОТЕ в группе,
- 7 6 ОТЕ, разделенные на 3 группы по 1, 2 и 3 ОТЕ в группе.

Все перечисленные модели с одной специфической тезой или антитезой на группу и с одной специфической тезой на каждую ОТЕ. 8-я модель из десяти ОТЕ, анализируемая по 20 случайно распределенным признакам. Тезой обозначалась четная цифра из таблицы случайных чисел, антитезой—нечетная. 9-я модель из 10-ти ОТЕ, разделенных на 2 группы по пять ОТЕ, исследуемая по 20-ти признакам. 10-я модель из 10-ти ОТЕ, разделенная на 3 группы по 2,3 и 5 ОТЕ в группе, анализируемая по 20-ти признакам. Для двух последних моделей первая ОТЕ группы имеет случайный набор признаков, а все остальные ОТЕ группы имеют с первой по 10 общих признаков, каких именно—определялось с помощью таблицы случайных чисел. Остальные признаки распределены случайно. Чтобы не попадались дублирующие признаки, строилась модель с 30-ью признаками, из коих 15 общих с первой ОТЕ группы, затем дублирующие признаки объединялись и оставлялось 20 признаков. Для трех последних моделей в таблице приведены средние арифметические отношений из 10 разных проб.

Наилучшим методом для данных моделей оказался метод с оценкой признака величиной pq , дающий малое отношение дисперсий в модели со случайным распределением признаков и большое отношение во всех других исследованных случаях. В вычислительном отношении этот метод

наименее трудоемкий после метода с равной оценкой всех признаков.

Наиболее сложный в вычислительном отношении метод Смирнова [6] дает максимальное отношение дисперсий в модели со случайным распределением признаков и относительно малое отношение дисперсий в остальных исследованных случаях. Шмидт [7] при сравнении двух методов отдает предпочтение методу Смирнова как математически более обоснованному. Надо признать, что, хотя сравнение методов говорит не в пользу метода Смирнова, все методы с оценкой признака в зависимости от частоты разработаны под значительным влиянием его идей.

Институт ботаники
АН АрмССР

Поступило 25.VI 1970 г.

ՊԱՎԵԼ Պ. ՂԱՄԲԱՐՅԱՆ

ԹՎԱՅԻՆ ՏԱՔՍՈՆՈՄԻԱՅԻ 10 ՏԻՊԻԿ ՄԵԹՈԴՆԵՐԻ ՀԱՄԵՄԱՏՈՒԹՅՈՒՆԸ

Ա մ փ ո փ ո լ մ

Ուսումնասիրված մեթոդները միմյանցից տարբերվում են հատկանիշի տարբեր գնահատմամբ՝ կախված հաճախականությունից: Համեմատության համար օգտագործվել է նմանության գնահատականների միջխմբային դիսպերսիաների հարաբերությունը ներխմբայինի նկատմամբ: Օգտագործվել են տարրոնների 10 տարբեր մոդելներ: Բոլոր մոդելներում դիսպերսիաների ամենամեծ հարաբերությունը, բացի հատկանիշների պատահական դասավորության մոդելից, տվեց ԲՊ մեծության հատկանիշի գնահատմամբ մեթոդը:

Л И Т Е Р А Т У Р А

1. Бейли Н. Математика в биологии и медицине. М., 1970.
2. Гамбарян П. П. Изв. АН АрмССР, сер. биол., 17, 12, 1964.
3. Гамбарян П. П. Изв. АН АрмССР, сер. биол., 18, 8, 1965.
4. Гамбарян П. П. Биологический журнал Армении, 21, 10, 1968.
5. Гамбарян П. П. Журнал общей биологии, 31, 4, 1970.
6. Смирнов Е. С. Таксономический анализ. М., 1969.
7. Шмидт В. М. Ботанический журнал, 53, 3, 1970.
8. Sokal R. R. and Sheath P. H. A. Principles of Numerical taxonomy. S. Franc. ez London, 1963.