

ПАВЕЛ ГАМБАРЯН

МАТЕМАТИЧЕСКИЙ МЕТОД В КЛАССИФИКАЦИИ

В основу всякой классификации кладется расположение объектов в упорядоченную систему на основании одного или многих признаков. Если для классификации используется один признак со своими градациями, то полученная система — линейная и искусственная (например, система Линнея).

В общем виде всякая классификация по многим признакам основывается на том, что разные объекты, наряду с различающими признаками, обладают также и общими. Абстрагируясь от качества отдельных признаков, математические методы таксономии дают определенные правила оценки близости (или различия) таксонов, в основу которых кладется число общих (совпадающих) признаков.

Простейший метод классификации — подсчет общего числа признаков каждой сравниваемой пары объектов. Этот метод прост и обладает малым выборочным варьированием. Но признаки объективно разнокачественные. Так, например, признак, свойственный одному виду — видовой, а свойственный всем видам рода — родовой. Из этого свойства признаков следует, что максимальную оценку следует давать самому частому признаку.

Классификация организмов — иерархическая. Объекты разбиваются на виды, виды объединяются в роды, роды — в семейства и т. д. Если 2 вида относятся к одному роду, то тем самым они относятся к одному семейству и т. д. Логически из этого вытекает, что максимальную оценку должен получить самый редкий признак. На этой основе Смирновым был предложен математический метод таксономии и аналогичные методы, дающие максимальную оценку совпадению объектов по самому редкому признаку. Если в классифицируемой группе ни один признак редковой формы не утрачивается, а происходит перераспределение старых признаков и возникновение новых без утраты старых, то объекты, обладающие самым редким признаком, действительно самые близкие. При утрате части признаков или замене их новыми самым редким признаком могут обладать и далекие таксоны.

Кроме того, давая максимальную оценку самому редкому признаку, мы автоматически даем минимальную оценку его антитезе. Трудно объяснить, почему для целей классификации наличие какого-нибудь признака надо оценивать иначе, чем отсутствие того же признака.

Мною [2, 3, 4] был предложен математический метод таксономии, который разрешает противоречие между объективной оценкой признака по его частоте и оценкой в связи с иерархической классификацией.

Кроме того, предложенный метод дает одинаковую оценку совпадению объектов по тезе и антитезе признака.

По упрощенному методу показатель сходства (при использовании гомологических признаков—родства) классифицируемых объектов предложено было оценивать величиной

$$T = \sum_{i=1}^m 4 p q_i, \quad (1)$$

где m число общих признаков и антитез у сравниваемых объектов, p —частота тезы, q —частота антитезы, $p+q=1$.

Для оценки сходства объектов было предложено использовать максимальные значения T , и поэтому ошибка T оценивалась по формуле

$$S_T = 4 \sqrt{\sum_{i=1}^m \left(p q \sqrt{\frac{2}{n-1}} \right)^2} \quad (2)$$

(в формуле 1 и 2 множитель 4 был введен для соответствия упрощенного метода основному и в дальнейшем изложении опускается).

При использовании не только максимальных значений T оказалось, что оценка S_T по предложенной формуле существенно занижена, и надо изменить предел суммирования в (2), учитывая ошибку несовпадающих признаков.

Правильная оценка

$$S_T = \sqrt{\sum_{i=1}^k \left(p q \sqrt{\frac{q}{n-1}} \right)^2}, \quad (3)$$

где k число используемых признаков.

Метод остается справедливым, если признаки некоррелированы между собой. Корреляция, автоматически возникающая между признаками в связи с их частотой, учитывается методом. Признаки с наименьшей частотой приведут к совпадению многих объектов по антитезам, что означает коррелированность этих признаков. Именно поэтому оценка самого редкого признака (с частой антитезой) минимальна. Но может существовать коррелированность между признаками и не автоматическая. Для оценки мы предлагаем определять величину, обозначаемую $T_{пр.}$, аналогичную показателю близости между объектами T . Для этого надо взять признаки как классифицируемые объекты, а классифицируемые объекты — как признаки. Существенным будет считаться такое значение $T_{пр.}$, вероятность случайного появления которого мала. Для этого воспользуемся критерием отбрасывания крайних вариантов [5].

$$T_{\max} = \frac{x_{\max} - \bar{x}}{S}; \quad T_{\min} = \frac{x_{\min} - \bar{x}}{S}; \quad (4)$$

подставляя вместо $\bar{X}$ значения $\bar{T}_{пр.} = \frac{\sum_{i=1}^k [np_i (np_i - 1) + nq_i (nq_i - 1)] pq_i}{n (n - 1)}$

и вместо S значение $S_{T_{пр.}} \cdot \sqrt{2}$.

Если $T_{пр.}$ существенно отличается от случайного, при суммировании вычесть из суммы весов обоих коррелированных признаков величину двух коррелированных pq для вычисления T надо внести поправку и

$$\frac{pq_1 + pq_2}{2} \cdot \left| \frac{T_{пр.} - \bar{T}_{пр.}}{T_{пр.}} \right|, \text{ где } \left| \frac{T_{пр.} - \bar{T}_{пр.}}{T} \right|$$

аналог коэффициента корреляции.

Для иллюстрации метода с помощью таблицы случайных чисел [1] смоделируем таксон из 20 объектов с 20 признаками. Исследуя 20 объектов, используем признаки, встречающиеся от 2 до 10 объектов (преобразовывая в нужном случае тезу в антитезу). В среднем признак будет встречаться у 3 объектов из 10. Я получил следующую модель таксона после ранжирования признаков и объектов по частоте (табл. 1).

Таблица 1

Модель таксона

		№ признака																											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	тез							
№ объекта	1	+	+	+		+			+	+	+			+				+											9
	2	+	+		+	+	+					+	+		+	+													8
	3	+	+				+					+	+		+		+												8
	4	+									+	+	+				+			+		+							8
	5	+	+	+		+	+										+		+	+									7
	6		+	+	+		+	+									+	+		+									7
	7		+	+		+	+		+	+			+		+		+			+									7
	8	+	+	+	+	+								+			+												6
	9			+			+	+				+					+	+		+									6
	10			+	+		+	+	+								+					+							6
	11	+	+				+	+	+		+												+						6
	12		+				+		+		+								+	+									6
	13		+				+				+				+		+		+	+									6
	14				+	+	+					+				+			+										5
	15				+			+	+			+				+				+									5
	16	+					+	+		+		+				+													5
	17		+	+				+					+			+													4
	18				+					+	+		+								+								4
	19	+			+	+						+	+																4
	20	+		+	+												+												4
тез		10,9		8	8	8	7	7	6	6	6	6	6	5	5	5	5	4	4	4	2								
pq		0,25		0,24			0,23		0,21					0,19				0,16			0,9								
Spq		0,08		0,08			0,08		0,07					0,06				0,05			0,3								

При подсчете значений T и $T_{пр.}$ для модели (табл. 1) получены следующие распределения (табл. 2).

Таблица 2
Распределение значений Т и Т_{пр.}

Т	n _i		Т _{пр.}	n _i
0,9	0	T=2,273 S=0,44	0,9	1
1,1	1		1,1	0
1,3	1		1,3	1
1,5	10		1,5	10
1,7	16	T _{пр.} =2,324 S=0,47	1,7	14
1,9	22		1,9	21
2,1	35		2,1	30
2,3	31		2,3	30
2,5	29		2,5	31
2,7	24		2,7	20
2,9	13		2,9	16
3,1	5		3,1	11
3,3	2		3,3	5
3,5	1		3,5	0
	190			190

По формуле (3) $S_T = 0,31$. При сравнении любых двух значений Т или Т_{пр.} ошибка разности, интересующая нас, будет равна

$$S_x = \sqrt{S_1^2 + S_2^2} \text{ или в нашем примере } \sqrt{0,31^2 + 0,31^2} = S_T \sqrt{2} = 0,44, \text{ как и в табл. 2.}$$

Распределение значений Т статистически несущественно отличается от нормального.

Все вышеприведенные рассуждения касаются и основного метода, расчеты по которому дают практически тот же результат, что и по упрощенному.

Ботанический институт
АН АрмССР

Поступило 26.II 1968 г.

ՊԱՂԵԼ Պ. ՂԱՄԲԱՐՅԱՆ

ԴԱՍԱԿԱՐԳՄԱՆ ՄԱԹԵՄԱՏԻԿԱԿԱՆ ՄԵԹՈԴԸ

Ա մ փ ո փ ո լ մ

Յուրաքանչյուր դասակարգման հիմքում դրված է առարկաների համընկնումը մեկ կամ շատ հատկանիշներով: Հեղինակը առաջարկել էր տաքսոնոմիայի մաթեմատիկական մեթոդ, որտեղ համընկնումը ըստ հատկանիշների գնահատելու համար վերցված է pq արժեքը, իսկ տաքսոնների մոտիկության գնահատման համար $T = \sum_{i=1}^m pq_i$ (4 բազմապատկիչը իջեցված է):

Т-ի ոչ միայն մաքսիմալ արժեքների օգտագործման ընթացքում պարզվեց, որ S_T -ն գնահատելու համար նախկինում առաջարկված բանաձևը տալիս է ցածր գնահատական: Երկրորդ գնահատումը՝ $S_T = \sqrt{\sum_{i=1}^k \left(pq \sqrt{\frac{2}{n-1}} \right)^2}$

օգտագործվող հատկանիշների թիվն է: Հարկավոր է հաշվի առնել ոչ միայն համընկնող, այլև չհամընկնող հատկանիշների սխալը:

Առաջարկված է նաև հատկանիշների միջև եղած հարաբերակցությունը գնահատելու համար որոշել $T_{\text{пр.}}$ -ն, հատկանիշների միջև եղած հարաբերակցության էության գնահատման համար օգտվել ծայրային տարրերակների ցուցման չափանիշից:

Եթե $T_{\text{пр.}}$ -ն լինի էական, ապա երկու հարաբերակցված pq -ն գումարելիս T -ն հաշվելու համար հարկավոր է ուղղում մտցնել և հանել

$$\frac{pq_1 + pq_2}{2} \cdot \left| \frac{T_{\text{пр.}} - \bar{T}_{\text{пр.}}}{\bar{T}_{\text{пр.}}} \right|$$

մեծությունը:

Մեթոդը պատկերացված է տաքսոնի վիճակագրական տիպարով:

Л И Т Е Р А Т У Р А

1. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики, М., 1965.
2. Гамбарян П. П. Изв. АН АрмССР (биолог. науки), т. 17, 12, 1964.
3. Гамбарян П. П. Изв. АН АрмССР (биолог. науки), т. 18, 8, 1965.
4. Гамбарян П. П. К систематике кавказских представителей рода *Pinus* L. (диссертация), Ереван, 1966.
5. Урбах В. Ю. Биометрические методы, М., 1964.