ПАЛЕОНТОЛОГИЧЕСКИЙ ЖУРНАЛ
1975 № 2

МЕТОДИКА ИССЛЕДОВАНИЙ

УДК 56.07

И. А. ВАНЧУРОВ

МАТЕМАТИЧЕСКИЙ МЕТОД СРАВНЕНИЯ ПАЛЕОНТОЛОГИЧЕСКИХ ОБЪЕКТОВ

Задача сравнения объектов слагается из двух отдельных проблем: выбора меры сходства и выбора признаков. В качестве меры близости палеонтологических объектов в общем случае могут рекомендоваться коэффициенты, удовлетворяющие определенным требованиям, в частности: требованию изменения в конечных пределах и требованию инвариантности относительно порядка значений признаков. Выбор признаков осуществляется на основании системы априорных данных о связях объектов. В целом метод позволяет выявить систему сходства объектов и определяющую ее систему признаков.

Сравнение объектов — одна из самых важных и распространенных задач в палеонтологии. Под объектами при этом могут подразумеваться отдельные окаменелости, совокупности образцов, представляющих остатки популяции, таксономические единицы, подразделения геологической временно́й шкалы и т. д. С одной стороны, каждый изучаемый объект необходимо сопоставить с другими из той же серии, для того чтобы определить его сходство с ними, с другой — нужно выявить его индивидуальные отличия, чтобы показать самостоятельную ценность.

Поставленная задача состоит из двух отдельных проблем: первая заключается в выборе меры близости объектов, вторая — в определении набора признаков, по которым следует устанавливать сходство. Рассмотрим каждую из проблем отдельно.

ВЫБОР МЕРЫ

Характеристикой близости двух объектов может служить число, которое назовем коэффициентом сходства, мерой близости или условным расстоянием между объектами. Для определения коэффициента сходства можно рекомендовать различные математические выражения. Но прежде чем останавливаться на каком-либо из них, определим требования, которым должно удовлетворять такое выражение.

Сначала рассмотрим наиболее простую ситуацию, когда объекты охарактеризованы только качественными признаками [1]. В этом случае к коэффициенту сходства предъявим следующие требования: 1) изменение в конечных пределах, 2) достижение верхнего предела при полном несовпадении значений признаков, нижнего — при полном их сходстве, 3) инвариантность относительно значений признаков (0 и 1). Необходимость выполнения первых двух пунктов очевидна. Требование инвариантности относительно значений признаков объясняется самой природой последних, невозможностью обосновать при кодировании предпочтительность одного из значений.

Для сравнения объектов по качественным признакам предложено большое число коэффициентов. Приведем некоторые из них, ни в коей мере не претендуя на полный обзор этой обширной темы.

[1] Качественными признаками будем называть те, которые могут принимать только два альтернативных значения.

Коэффициент Оцука:

$$z=\frac{C}{\sqrt{N_1N_2}}, \quad 0\leqslant z\leqslant 1 \tag{1}$$

Коэффициент Жаккара [2]:

$$z=\frac{C}{N_1+N_2-C}, \quad 0\leqslant z\leqslant 1 \tag{2}$$

Коэффициент Дайса:

$$z=\frac{2C}{N_1+N_2}, \quad 0\leqslant z\leqslant 1 \tag{3}$$

Коэффициент Фейджера:

$$z=\frac{C}{\sqrt{N_1+N_2}}-\frac{1}{2\sqrt{N_2}}, \quad -\frac{1}{2}\leqslant z<1 \tag{4}$$

Формула Престона:

$$\left(\frac{N_1}{N_1+N_2-C}\right)^{1/z}+\left(\frac{N_2}{N_1+N_2-C}\right)^{1/z}=1, \quad 0<z\leqslant 1 \tag{5}$$

Расстояние Хемминга [3]:

$$z=\sqrt{N_1+N_2-2C}, \quad 0\leqslant z\leqslant\sqrt{m} \tag{6}$$

Формула Экмона:

$$z=\frac{N_1+N_2-2C}{C}, \quad 0\leqslant z\leqslant\infty \tag{7}$$

Коэффициент Стугрена — Радулеску:

$$z=\frac{N_1+N_2-3C}{N_1+N_2-C}, \quad -1\leqslant z\leqslant 1, \tag{8}$$

где z — коэффициент сходства, C — число совпадающих единичных значений признаков у двух объектов, N_1 — сумма значений признаков первого объекта, N_2 — сумма значений признаков второго объекта, m — число признаков.

Во всех перечисленных коэффициентах в той или иной степени достаточно хорошо учитывается различие объектов, однако сходство отражается не полностью или вообще не отражается (расстояние Хемминга). Признаки, значения которых в обоих сравниваемых объектах равны нулю, не фиксируются ни в одном случае. Но поскольку совпадение нулевых значений в объектах указывает на их сходство (так же как и совпадение единичных значений), то в коэффициенте сходства должно учитываться все множество значений признаков. Этим объясняется рекомендация И. Н. Печерской и Ю. Н. Печерского (1972), предложивших под числом C понимать число совпавших единиц и пулей. В принятых нами обозначениях это число (обозначим его C) равно:

$$\hat{C}=m-N_1-N_2+2C. \tag{9}$$

Тогда, например, коэффициент Жаккара изменится следующим образом:

$$z=\frac{m-N_1-N_2+2C}{2(N_1+N_2-C)-m}. \tag{10}$$

---

[2] Коэффициент Жаккара представляет собой отношение числа единичных элементов в пересечении двух множеств к числу в их объединении и может быть соответственно выражен как $z=\frac{N(A\cap B)}{N(A\cup B)}$.

[3] Этот коэффициент выражает расстояние между вершинами m-мерного единичного гиперкуба. Подкоренное выражение представляет собой число несовпадений значений признаков в сравниваемых объектах.

Однако и последний вариант коэффициента сходства, как и все предыдущие, не удовлетворяет третьему требованию.

Одним из коэффициентов сходства объектов по качественным признакам может служить расстояние Хемминга, деленное на общее количество признаков:

$$z=\left(\frac{N_1+N_2-2C}{m}\right)^{1/2}, \quad 0\leqslant z\leqslant 1. \tag{11}$$

Легко убедиться, что это выражение удовлетворяет всем трем требованиям, предъявленным выше к коэффициенту сходства.

Конечно, выражение (11) не единственное из тех, которые обладают обусловленными свойствами. Рассмотрим еще один коэффициент сходства, для чего воспользуемся следующим выражением:

$$z=1-r, \tag{12}$$

где r — коэффициент корреляции между двумя объектами по совокупности (m) признаков. В принятых выше обозначениях для качественных признаков он соответствует следующему:

$$r_{1,2}=\frac{mC-N_1N_2}{\sqrt{N_1N_2(m-N_1)(m-N_2)}}. \tag{13}$$

Нетрудно убедиться, что предлагаемый коэффициент различия удовлетворяет всем трем предъявленным ранее требованиям: 1) поскольку $1\leqslant r\leqslant 1$, то $0\leqslant z\leqslant 2$; 2) при полном различии объектов $C=0$, $N_1+N_2=m$, $z=2$, при полном сходстве объектов $N_1=N_2=C$, $z=0$; 3) заменим значения признаков у сравниваемых объектов на альтернативные и рассчитаем новый коэффициент сходства $\hat{z}$ с параметрами $\hat{N}_1$, $\hat{N}_2$ и $\hat{C}$, причем $N_1=m-N_1$, $N_2=m-N_2$, $\hat{C}=m-N_1-N_2+C$. Откуда:

$$\hat{z}=1-\frac{m(m-N_1-N_2+C)-(m-N_1)(m-N_2)}{\sqrt{N_1N_2(m-N_1)(m-N_2)}}=$$

$$=1-\frac{mC-N_1N_2}{\sqrt{N_1N_2(m-N_1)(m-N_2)}}=1-r=z. \tag{14}$$

Таким образом, коэффициент (12) удовлетворяет всем трем перечисленным выше требованиям.

В некоторых задачах, когда число признаков, описывающих объекты, конечно и известны все их значения, в качестве характеристики сходства объектов можно использовать расстояние Хемминга. Такой задачей может, например, быть сравнение окаменелостей по данным качественного химического анализа. Совокупностью признаков в этом случае может служить перечень химических элементов, наличие или отсутствие которых кодируется 0 и 1. Если же количество признаков не имеет естественных границ, то расстояние Хемминга бесконечно велико, а его оценки, полученные при искусственном ограничении числа признаков, не представляют практической ценности. Коэффициент (11) избавлен от этого недостатка, однако ниже при рассмотрении обработки значений количественных признаков будет показано, что из-за некоторых задач коэффициенту (12) следует отдать предпочтение [4].

Перейдем к рассмотрению коэффициента сходства для порядковых признаков [5]. В этом случае к нему предъявляются те же требования, что и для качественных признаков, за исключением третьего пункта, который

[4] Возможно, имеются задачи, в которых соблюдение перечисленных выше требований не обязательно. В этих случаях применение коэффициентов (1—8) может быть вполне оправдано. В данной статье частные ситуации не рассматриваются.

[5] Порядковыми признаками назовем такие, которые могут принимать несколько (более двух) упорядоченных значений.

заменяется на следующий: 3) инвариантность относительно альтернативного порядка значений признаков. Это требование обусловлено тем, что при кодировании начало условной шкалы выбирается произвольно от любого крайнего значения признака.

Покажем, что коэффициент сходства (12) удовлетворяет и этому требованию.

Пусть r_{XY} — коэффициент корреляции m порядковых признаков объектов X и Y, а x_i и y_i — соответствующие значения i-го признака. Тогда значения признаков с альтернативным порядком $(p-x_i)$ и $(p-y_i)$, где p — максимальное значение принятой условной шкалы плюс единица. Вычислим коэффициент корреляции r_{XY} для новых значений признаков:

$$\hat{r}_{XY}=\frac{m\sum_{i=1}^{m}(p-x_i)(p-y_i)-\sum_{i=1}^{m}(p-x_i)\sum_{i=1}^{m}(p-y_i)}{\sqrt{\left\{m\sum_{i=1}^{m}(p-x_i)^2-\left[\sum_{i-1}^{m}(p-x)\right]^2\right\}\left\{m\sum_{i=1}^{m}(p-y_i)^2-\left[\sum_{i=1}^{m}(p-y_i)\right]^2\right\}}}.$$

После элементарных преобразований получаем:

$$\hat{r}_{XY}=\frac{m\sum_{i=1}^{m}x_iy_i-\sum_{i=1}^{m}x_i\sum_{i=1}^{m}y_i}{\sqrt{\left[m\sum_{i=1}^{m}x_i^2-\left(\sum_{i=1}^{m}x_i\right)^2\right]\left[m\sum_{i=1}^{m}y_i^2-\left(\sum_{i=1}^{m}y_i\right)^2\right]}}=r \qquad (15)$$

что и требовалось доказать.

Использование коэффициента корреляции в качестве характеристики сходства объектов по совокупности количественных признаков широко распространено. В литературе оно фигурирует обычно под названием корреляционного Q-анализа или Q-метода (Миллер и Кан, 1965; Крамбейн и Грейбилл, 1967). Коэффициент $z=1-r$ и в этом случае удовлетворяет первым двум требованиям, предъявляемым к нему при рассмотрении сходства по качественным и порядковым признакам, а последний пункт требований, связанный с инвариантностью порядка значений признаков, естественным образом отпадает и заменяется на следующий: 3) одинаковое изменение масштаба измерения всех признаков в одном из сравниваемых объектов не должно влиять на величину расстояния между ними.

Это требование объясняется тем, что выбор масштабных единиц для измерений количественных признаков осуществляется произвольно, а не определяется сущностью объекта. Кроме того, измеряя значения признаков, которые увеличиваются пропорционально увеличению индивидуального возраста организма, мы задаемся целью сравнить их соотношения, а не абсолютные величины.

Покажем, что коэффициент корреляции удовлетворяет этому требованию. Пусть x_i и y_i значения m — признаков в объектах X и Y, а r_{XY} — коэффициент корреляции между X и Y. Изменим значения признака в объекте X в p раз и вычислим новый коэффициент корреляции $r_{pX,Y}$:

$$r_{pX,Y}=\frac{m\sum_{i=1}^{m}px_iy_i-\sum_{i=1}^{m}px_i\sum_{i=1}^{m}y_i}{\sqrt{\left[m\sum_{i=1}^{m}(px_i)^2-\left(\sum_{i=1}^{m}px_i\right)^2\right]\left[m\sum_{i=1}^{m}y_i^2-\left(\sum_{i=1}^{m}y_i\right)^2\right]}}. \qquad (16)$$

Очевидно, что величина p в этом выражении сокращается, т. е. $r_{p\,X,\,Y} = r_{X,\,Y}$.

Если число качественных признаков, описывающих сравниваемые свойства объектов, конечно и все их значения фиксированы, то в качестве коэффициента сходства можно использовать эвклидово расстояние:

$$z = \left[\sum_{i=1}^{m} (x_i - y_i)^2\right]^{1/2}, \qquad (17)$$

где x_i и y_i — значения i-го признака в объектах X и Y, m — число признаков.

Такая задача может возникнуть, например, при сравнении формы скелетов организмов, когда известны все параметры, полностью описывающие эти формы. В противном случае эвклидово расстояние не служит мерой близости объектов — вывод, аналогичный тому, который сделан при сравнении по качественным признакам при помощи расстояния Хемминга.

Теперь, когда выбрана одна и та же мера сходства объектов для признаков всех трех классов, рассмотрим вопрос о сравнении объектов по комплексу всех признаков. Поскольку любые значения качественных и порядковых признаков можно рассматривать как частные случаи от значений количественных признаков, такой вопрос решается положительно. Однако признаки при этом приобретают различные, ничем не оправданные веса. Вес признака становится тем больше, чем больше его значение отличается от среднего значения признаков в каждом объекте. Таким образом, вес количественных признаков зависит от масштаба измерения их значений, а вес порядковых признаков — от дробности выбранной условной шкалы. Чтобы избежать искусственного введения веса признаков, нормируем их значения через средние:

$$a_{ij}^{*} = \frac{a_{ij} n}{\sum_{j=1} a_{ij}},$$

где a_{ij} — исходное значение i-го признака, a_{ij}^{*} — его нормированное значение, n — количество объектов.

Такой подход не обеспечивает, конечно, установления естественных весов признаков. Однако в палеонтологических задачах нет возможности априори определять числовые соотношения весов признаков, так как они зависят не только от самой природы объектов, но и от набора сравниваемых объектов и могут существенно измениться с изменением этого набора. Известные рекомендации по «вычислению» весовых коэффициентов признаков не обоснованы и содержат те или иные ошибки. Если бы можно было достаточно надежно обосновать величины весовых коэффициентов признаков в палеонтологических объектах, как предполагают многие исследователи, то задача по сравнению последних значительно упростилась бы. Но поскольку веса признаков неопределимы, приходится искать обходной путь решения задачи. Такой путь имеется. Он основан на выборе определенного набора признаков, т. е. выбора признакового пространства, и требует привлечения дополнительной информации о сходстве объектов.

ВЫБОР ПРИЗНАКОВ

Как было показано выше, значение признака в объекте зависит от нескольких факторов. Влияние этих факторов в различных признаках неравноценно. Тогда задача сравнения объектов может быть сформулирована следующим образом: из всего множества фиксированных признаков необходимо выбрать подмножество таких, у которых значения зависят от выбранного нами фактора в большей степени, чем от других факторов.

Очевидно, что сходство, рассчитанное по выбранному подмножеству признаков, будет прежде всего зависеть от ведущего фактора. Веса признаков при этом не обязательно должны соответствовать естественным весам, однако независимо от абсолютного значения вычисленных расстояний между признаками изменение веса признаков не нарушает их упорядоченности и, что самое главное, минимальный коэффициент сходства остается минимальным.

Требуемое подмножество признаков выбирается из всего фиксированного множества на основании дополнительной априорной информации о связях объекта. Обычно в палеонтологических задачах такая информация довольно значительна. Она состоит из системы достоверных и запрещенных связей. Достоверная связь — это связь между двумя объектами, о которой заранее известно, что она более тесна, более существенна, чем связь одного из этих объектов со всеми остальными, фигурирующими в задаче. Примером достоверной связи в задаче генетического сходства видов может служить уверенность в происхождении одного конкретного вида от другого конкретного вида. Основанием для такой уверенности могут служить любые соображения, зависящие от фиксированных признаков или не зависящие от них. Поэтому в дальнейшем эти основания условимся называть общими соображениями. При построении филогенетического древа группы организмов отдельные ветви этого древа могут не вызывать сомнений и разногласий специалистов — это и есть достоверные априорные связи. Запрещенными связями назовем те, которые специалистами на основании общих соображений уверенно отвергаются. В предыдущем примере запрещенными связями между видами будут те, относительно которых есть уверенность, что ни один из двух сравниваемых видов не является ближайшим предком или потомком второго. Если закодировать значения априорных сведений о связях, то можно составить матрицу априорной информации о связях. В задачах о генетическом сходстве такая матрица будет треугольной, в некоторых других случаях может быть и квадратной.

Априорная информация используется при отборе признаков следующим образом: находится такой набор признаков, по которому минимальные расстояния между объектами соответствуют достоверным связям и не соответствуют запрещенным связям. Тогда задача о взаимосвязях объектов может рассматриваться как задача интерполяции известной (априорной) информации о кратчайших расстояниях между объектами на основании комплекса признаков. С этих позиций надежность результатов в значительной степени определяется количеством априорной информации о связях и количеством интерполируемых на ее основании данных. Рекомендуемая методика решения задачи о связях объектов может привести к трем типам ответов:

1. Ни один набор признаков не отвечает априорным заданным условиям. Интерпретация этой ситуации такова — выбранный за основу сходства фактор не доминирует ни в одном из сочетаний признаков, описывающих объекты. Описание объектов не соответствует поставленной задаче.

2. Единственное сочетание признаков удовлетворяет априорно заданным связям. Задача решена.

3. Имеется несколько сочетаний признаков, удовлетворяющих априорным сведениям о связях. Такое решение может получиться при слабой обеспеченности априорными связями или при высокой корреляции отдельных групп признаков во всех объектах. В первом случае совокупность прогнозируемых связей может резко различаться, во втором — она однозначна или очень близка к этому (если значения признаков находятся в функциональной линейной зависимости, то не играет роли, какой из них оставлен для определения сходства объектов). Во всех случаях предпочтение отдастся системе связей, в основу которой положены сочe-

тания из максимального количества признаков, так как большая представительность выборки признаков, описывающих объекты, снижает в общем случае вероятность случайной ошибки решения задачи.

Полученной в результате решения задачи матрице минимальных расстояний между объектами можно поставить в соответствие перенумерованный граф дерево, в котором вершинам соответствуют объекты, а ребрам — связи между ними. Такая интерпретация чрезвычайно наглядна и близка к привычным палеонтологическим схемам. Например, в результате решения задачи о генетическом сходстве видов получается граф дерево с корнем (или совокупность деревьев), которое палеонтологами именуется филогенетическим древом (рис. 1).

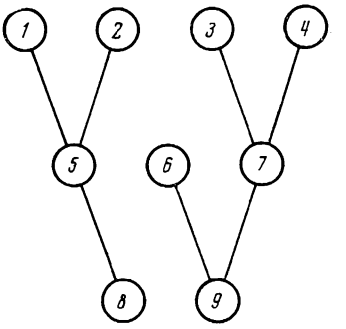

Рис. 1. Граф дерево, показывающий вычисленную связь между девятью объектами

Изложенный метод легко реализуется на ЭВМ среднего класса, если количество признаков не превышает 15. При большем числе признаков практические трудности, связанные с большим объемом вычислений, непреодолимы, так как сумма всех сочетаний из n равна 2^n. При $n=20$ $2^{20}>10^6$, т. е., если на анализ одного сочетания признаков затрачивается одна минута (обычно больше), то для анализа всех сочетаний потребуется около двух лет беспрерывного счета. Прямого решения задачи, которое позволило бы непосредственно определить необходимое сочетание признаков без «слепого» перебора их сочетаний, не найдено. Поэтому предлагается серия последовательных однотипных операций по сокращению набора признаков, содержащая не более n циклов. В случае независимости значений признаков удовлетворительно было бы следующее решение: устанавливаются веса признаков, показывающие их роль в проявлении несоответствий кратчайших расстояний между объектами по исходным данным и априорной информацией о связях; признаки ранжируются по этим весам, затем отбрасываются один за другим до того момента, когда несоответствие исчезнет. Такая задача имеет решение, но мы не будем его рассматривать, так как признаки в организмах обычно коррелированы. А следствием корреляции может быть существенное различие между совместным влиянием на сходство объектов нескольких признаков и суммой их индивидуальных влияний. Отбросив один самый «вредный» для решения задачи признак, можно настолько изменить ситуацию, что установленный ранее порядок признаков по «вредности» не будет соответствовать действительному. Поэтому, определив наиболее «вредный» признак и изъяв его из исходных данных, предлагается анализировать полученную ситуацию заново, т. е. повторить цикл предыдущих расчетов. Можно придумать такую искусственную модель (задача довольно сложная), когда изложенный метод не позволит найти существующее решение задачи, но вероятность случайного создания такой модели чрезвычайно мала и ею следует пренебречь. Практически этот метод направленного поиска необходимого сочетания признаков приводит к тем же результатам, что и перебор, но расчеты резко сокращаются (максимальное количество циклов меньше числа признаков).

Детальный алгоритм решения задачи приведен в одной из моих работ (Ванчуров, 1973) и запрограммирован для ЭВМ типа БЭСМ-4 В. С. Самариным (ВНИГНИ) под кодовым названием «Дендрограф». Программа «Дендрограф» моделирует работу палеонтолога по анализу сходства объектов, не внося в нее никаких принципиально отличных от общепринятых операций над фактическим материалом. В то же время эта программа позволяет произвести сопоставление до 120 объектов по 150 признакам за

время, не превышающее 3—4 часов,— задача, непосильная для исследователя, не использующего вычислительную технику. Принципиальная отличительная черта описанного выше алгоритма от других алгоритмов на эту тему — обоснованный априорной информацией выбор признаков, использующихся для установления сходства объектов, из их общего анализируемого числа.

ЛИТЕРАТУРА

Ванчуров И. А. 1973. О решении классификационных задач в палеонтологии математическими методами. Тр. Всес. н.-и. геологоразв. нефт. ин-та, вып. 135, стр. 34—48.

Крамбейн У. и Грейбилл Ф. 1967. Статистические модели в геологии. «Мир», стр. 1—397.

Миллер Р. и Кан Дж. 1965. Статистический анализ в геологических науках. «Мир», стр. 1—481.

Печерская И. Н. и Печерский Ю. Н. 1972. Применение методов теории графов для решения задач классификации палеонтологических объектов. В сб.: Цифровое кодирование систематических признаков древних организмов. «Наука», стр. 133—139.

Всесоюзный научно-исследовательский
геологоразведочный нефтяной институт
Москва

Статья поступила в редакцию
25 II 1974