Доклады Академии наук СССР
1972. Том 206, № 6

УДК 519.95. *КИБЕРНЕТИКА И ТЕОРИЯ РЕГУЛИРОВАНИЯ*

Ю. Л. ВАСИЛЬЕВ, А. Н. ДМИТРИЕВ

СПЕКТРАЛЬНЫЙ ПОДХОД К СРАВНЕНИЮ ОБЪЕКТОВ, ОХАРАКТЕРИЗОВАННЫХ НАБОРОМ ПРИЗНАКОВ

(Представлено академиком С. Л. Соболевым 23 II 1972)

1. Здесь рассматривается вопрос о заключительном этапе процедур, называемых диагностическими, распознаванием образов и т. п., а именно: вопрос об уточнении диагностики на основе той или иной числовой меры. Предлагаемый подход вводит на объектах и характеризующих их признаках числовую меру и дает способ ее вычисления, основанный на итерациях. Следует отметить, что эта мера в некотором смысле отражает естественное соотношение между сравниваемыми объектами (п. 3) и оправдывается при практических применениях (п. 6). Способ вычисления отличается малой трудоемкостью и делает доступными задачи с весьма большим числом признаков (пп. 3, 4, 6). На обрабатываемые таблицы налагается некоторое естественное структурное ограничение (связность, п. 4).

2. Пусть имеется h объектов $W_1, \ldots, W_h$, охарактеризованных двузначными признаками $P_1, \ldots, P_l$. Пусть $T = (t_{ij})$ — таблица размера $h \times l$, которая составлена из единиц и нулей и в которой i-я строка $(t_{i1}, \ldots, t_{il})$ отвечает объекту W_i, $i = 1, \ldots, h$, j-й столбец $(t_{1j}, \ldots, t_{hj})$ отвечает признаку P_j, $j = 1, \ldots, l$; таблица T отражает выраженность признаков у объектов: если для i-го объекта и j-го признака она превышает некоторый уровень, то $t_{ij} = 1$, а если меньше, то $t_{ij} = 0$. Условимся, что $h \geqslant 2$, $l \geqslant 2$ и что в таблице T нет строк и нет столбцов, составленных сплошь из нулей.

Если признаки k-значны, $k > 2$, то t_{ij} может принимать соответствующее число значений из отрезка $[0, 1]$. Ограничимся рассмотрением случая $k = 2$, так как переход на случай $k > 2$ будет очевиден.

Таблица описывает объекты в связи с тем или иным фактором X, и задача состоит в оценке последнего по картине, даваемой таблицей T. Предполагается, что признаки $P_1, \ldots, P_l$ существенны для X, что набор их достаточно полон и что объекты $W_1, \ldots, W_h$ родственны по отношению к X. Пусть T — геологическая таблица, X — запасы месторождений, объекты — месторождения одного и того же полезного ископаемого, которые родственны по признакам и сравнимы по запасам, но оценка последних требует уточнения.

Нас интересуют случаи, когда число объектов сравнительно невелико (10—20), число признаков велико (десятки, сотни), влияния признаков находятся в сложном переплетении. Поскольку статистический подход встречает здесь серьезные препятствия, возникли другие подходы. Сначала они проверяются на всесторонне изученных примерах, затем распространяются не менее изученные ситуации вплоть до решения задач прогноза. Хорошо согласующиеся с практикой оценки дает основанный на аппарате тупиковых тестов ([1]) известный тестовый подход ([2–4]), на который мы будем ориентироваться при сопоставлении результатов *.

* Напомним, что для таблицы T тестовый вес j-го столбца (j-го признака) $c_j = (N_j / N)$, где N_j — число тупиковых тестов, содержащих j-й столбец, $j = 1, \ldots, l$, N — число всех тупиковых тестов; тестовый вес i-й строки (i-го объекта) $r_i = t_{i1}c_1 + \ldots + t_{il}c_l$, $i = 1, \ldots, h$.

3. Вектор $\mathbf{a} = (a_1, \ldots, a_m)$ называется положительным, если $a_1, \ldots, a_m > 0$. Норма вектора $\mathbf{a} = (a_1, \ldots, a_m)$ — число $|\mathbf{a}|$, равное $\max\limits_{1 \leqslant i \leqslant m} |a_i|$; вектор $\mathbf{a}$ называется нормальным, если $|\mathbf{a}| = 1$.

Числовую меру для объектов и признаков естественно задавать в виде положительных нормированных векторов $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_h)$ и $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_l)$. Ниже мы определим их как нагрузку строк и нагрузку столбцов таблицы T. По ним предлагается судить о проявленности фактора X соответственно в объектах $W_1, \ldots, W_h$, а также о степени влияния на нее признаков $P_1, \ldots, P_l$.

Определение нагрузки подсказывается анализом грубой оценки объектов и признаков, при которой им приписывается вес, равный количеству единиц в соответствующих строках и столбцах таблицы T. При этом

	P_1	P_2	P_3	P_4	Вес
W_1	1	1	0	0	2
W_2	0	1	1	0	2
W_3	0	0	1	1	2
Вес	1	2	2	1	

Рис. 1. Грубая оценка

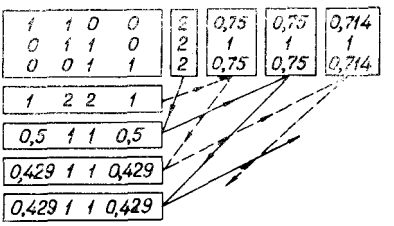

Рис. 2. Итерационный процесс («качели»)

строки и столбцы выступают как бы порознь. В примере на рис. 1 строки получают равные веса, однако они различны по отношению к столбцам: во второй строке обе единицы отвечают признакам, представленным единицами и в других строках, а в первой и третьей строках имеется лишь по одной такого рода единице. Содержательно это может означать, что вторая строка отвечает более сильному проявлению фактора X.

Определение нагрузки отражает итог описанных ниже пересчетов весов строк с учетом весов столбцов, и наоборот. При каждом пересчете вес строки W_i, $i = 1, \ldots, h$, определяется через веса столбцов, найденные на предшествующем шаге как сумма весов тех столбцов, по которым в строке W_i стоят единицы; вес столбца P_j, $j = 1, \ldots, l$, аналогично определяется через найденные на предшествующем шаге веса строк; затем полученные два набора чисел нормируются, начинается новый пересчет и т. д. Процесс сходится к некоторым предельным векторам $\boldsymbol{\omega}$ и $\boldsymbol{\pi}$, которые примем в качестве нагрузок. Для таблицы на рис. 1 $\boldsymbol{\omega} \approx \approx (0{,}707;\ 1;\ 0{,}707)$, $\boldsymbol{\pi} \approx (0{,}414;\ 1;\ 1;\ 0{,}414)$. По сравнению с грубой оценкой нагрузки $\boldsymbol{\omega}$ выделяет вторую строку. На рис. 2 представлены результаты трех первых пересчетов.

4. Исходя из таблицы T и пары векторов $\mathbf{w} = (w_1, \ldots, w_h)$, $\mathbf{p} = (p_1, \ldots, p_l)$, определим пару векторов $'\mathbf{w} = ('w_1, \ldots, 'w_h)$, $'\mathbf{p} = ('p_1, \ldots, 'p_l)$, а также нормирующие множители α, β и пару векторов $\mathbf{w}'$ и $\mathbf{p}'$:

$$'\mathbf{w} = T(\mathbf{p}), \text{ т. е. } 'w_i = t_{i1}p_1 + \ldots + t_{il}p_l, \ \mathbf{w}' = \alpha \cdot {}'\mathbf{w}, \qquad (1)$$

$$i = 1, \ldots, h, \quad |\mathbf{w}'| = 1$$

$$'\mathbf{p} = T^*(\mathbf{p}), \text{ т. е. } 'p_j = t_{1j}w_1 + \ldots + t_{hj}w_h, \ \mathbf{p}' = \beta \cdot {}'\mathbf{p}, \qquad (2)$$

$$j = 1, \ldots, l, \quad |\mathbf{p}'| = 1$$

Формулы (1), (2) дают формальное определение упомянутых пересчетов.

Определения. Пару нормированных векторов $\varphi = (\varphi_1, \ldots, \varphi_h)$, $\psi = (\psi_1, \ldots, \psi_l)$ назовем фракционирующей парой таблицы, а сами векторы — фракционирующими векторами, если $\varphi' = \varphi$ и $\psi' = \psi$.

Фракционирующую пару векторов ω, π назовем нагрузкой таблицы T, а сами векторы — нагрузкой строк и нагрузкой столбцов, если эти векторы положительны.

Таблицу T, нагрузка которой существует и единственна, назовем измеримой.

Таблицу T назовем разрозненной, если среди всех ее h строк найдутся h_1 таких строк, каждая из которых ортогональна* каждой из прочих $h - h_1$ строк ($h_1 < h$). Иными словами, совокупность объектов распадается на две такие группы по h_1 и $h - h_1$ объектов, что каждый из признаков обязательно представлен нулями во всех объектах какой-то одной из этих двух групп (грубо говоря, эти две группы «не имеют ничего общего»). Перестановками строк и столбцов любую разрозненную таблицу можно привести к виду $\begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix}$. Любая разрозненная таблица неизмерима **.

Таблицу T, не являющуюся разрозненной, назовем связной.

Теорема 1. *Любая связная таблица измерима.*

Теорема 2 (о качелях). *Нагрузка строк ω и нагрузка столбцов π измеримой таблицы T являются пределами соответственно последовательностей векторов*

$$\omega^0, \omega^1, \ldots, \omega^n, \ldots, \quad \text{где } \omega^0 = (\overbrace{1, \ldots, 1}^{h \text{ единиц}}), \quad \omega^n = (\omega^{n-1})', \tag{3}$$

$$\pi^0, \pi^1, \ldots, \pi^n, \ldots, \quad \text{где } \pi^0 (\underbrace{1, \ldots, 1}_{l \text{ единиц}}), \quad \pi^n = (\pi^{n-1})'. \tag{4}$$

5. Доказательство теоремы 1 сводит ее к известной теореме Фробениуса ([5]) о спектре неотрицательной неразложимой матрицы Ф. Роль матрицы Ф играют матрицы TT^* и T^*T, объединяемые матрицей $K^2 = \begin{pmatrix} TT^* & 0 \\ 0 & T^*T \end{pmatrix}$, где $K = \begin{pmatrix} 0 & T \\ T^* & 0 \end{pmatrix}$.

Последняя связана с (3), (4) и служит посредником между Ф и T. Сокращения с.в. и с.з. означают собственный вектор и собственное значение. Для векторов $\mathbf{a} = (a_1, \ldots, a_h)$ и $\mathbf{b} = (b_1, \ldots, b_l)$ обозначим через $\mathbf{a} \vee \mathbf{b}$ вектор $(a_1, \ldots, a_h, b_1, \ldots, b_l)$. Аккомодацией назовем преобразование вектора $\mathbf{a} \vee \mathbf{b}$ в пару нормированных векторов $v_a \cdot \mathbf{a}$, $v_b \cdot \mathbf{b}$ ($|v_a \cdot \mathbf{a}| = |v_b \cdot \mathbf{b}| = 1$). Пусть X — множество нормированных с.в. матрицы K и пусть Y — множество фракционирующих пар таблицы T.

(А) Аккомодация взаимно однозначно отображает X на Y; при этом положительные собственные векторы переходят в нагрузки, и обратно. С.з. матриц TT^* и T^*T совпадают (различиями в кратности с.з. 0 пренебрегаем).

Б) Если $K(\mathbf{a} \vee \mathbf{b}) = \lambda(\mathbf{a} \vee \mathbf{b})$, то $\|\mathbf{a}\| = \|\mathbf{b}\|$ (норма эвклидова) и $TT^*(\mathbf{a}) = \lambda^2 a$, $T^*T(\mathbf{b}) = \lambda^2 b$; имеет место и обратное.

Пусть Ф — любая из матриц TT^* и T^*T. Матрица Ф неотрицательна; так как таблица T связная, то Ф неразложима. Согласно теореме Фробениуса, матрица Ф имеет такое с.з. $\lambda_{\max}$ и отвечающий ему с.в. $\mathbf{u}$, что а) $\lambda_{\max} > 0$, $\lambda_{\max}$ больше всех других с.з., $\lambda_{\max}$ является простым с.з.; б) с.в. $\mathbf{u}$ положителен. Из (б), (Б) и (А) следует существование нагрузки таблицы T; из простоты с.з. $\lambda_{\max}$, (Б), ортогональности с.в. матрицы K и из (А) следует единственность нагрузки.

---

* Два вектора $(x_1, \ldots, x_l)$ и $(y_1, \ldots, y_l)$ ортогональны, если $x_1 y_1 + \ldots + x_l y_l = 0$.

** Можно доказать, что доля разрозненных таблиц среди всех $(h \times l)$- таблиц T стремится к нулю при $h \leqslant l \leqslant 2^{h-1}$ и $h \to \infty$.

Доказательство теоремы 2 сводит ее к обычному итерационному построению ([6]) с.в. **u** матрицы Ф, отвечающего с.з. λ_{max}.

6. Примеры решения прогнозно-поисковых задач из геологии.

Пример 1. При сравнительном изучении гигантских месторождений нефти ([7]) требуется из общей совокупности (около 350) характеристических признаков выделить группу признаков, имеющих поисковое значение. Для этого каждая из выделенных групп испытывается на спо-

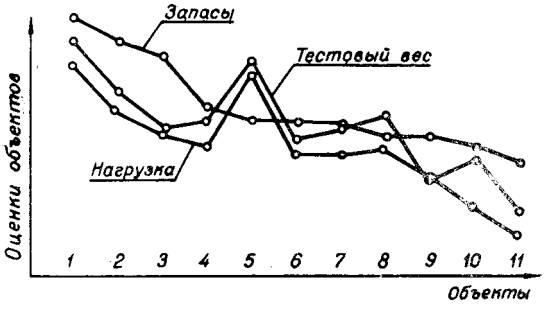

Рис. 3. Группа месторождений Аравийской платформы

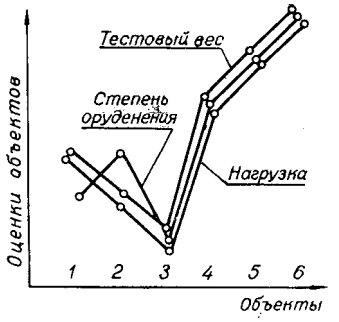

Рис. 4. Дифференцированные траповые интрузии

собность фракционировать месторождения по масштабу запасов. Одна из таких групп, именуемая «структурная ловушка» ($l = 28$), нами взята для сопоставления фракционирующих возможностей тестового и спектрального подходов (рис. 3).

Пример 2. При изучении дифференцированных траповых интрузий, перспективных на медно-никелевое оруденение, рассматривались интрузии норильского типа ([8]). Упорядоченность последних по степени оруденения, а также фракционирование их по результатам обработки таблиц, содержащих данные о петрохимических признаках ($l = 32$), тестовым и спектральным подходами представлены на рис. 4.

Сравнительное изучение тестового и спектрального подходов было проведено для большого набора геологических таблиц. Как и в приведенных примерах, имеет место близость оценок объектов по обоим подходам. Время обработки таблиц: для одних и тех же таблиц требовалось 30—90 мин. на БЭСМ-6 при тестовом подходе и не более 1 мин. на М-20, М-220 при спектральном подходе.

Наряду с близостью нагрузки и тестового веса для объектов наблюдаются существенные различия между нагрузкой и тестовым весом для признаков.

Институт математики

Институт геологии и геофизики
Сибирского отделения Академии наук СССР
Новосибирск

Поступило
17 II 1972

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] И. А. Чегис, С. В. Яблонский, Тр. Матем. инст. им. В. А. Стеклова АН СССР, **51**, 270 (1958). [2] А. Н. Дмитриев, Ю. И. Журавлев, Ф. П. Кренделев, Сборн. Дискретный анализ, в. 7, 3 (1966). [3] А. Н. Дмитриев, Ю. И. Журавлев, Ф. П. Кренделев, Геология и геофизика, № 5, 50 (1968). [4] С. В. Яблонский, Н. Г. Демидова и др., Геология рудных месторожд., № 2, 3 (1971). [5] Ф. Р. Гантмахер, Теория матриц, М., 1967. [6] Д. К. Фаддеев, В. Н. Фаддеева, Вычислительные методы линейной алгебры, М., 1960. [7] В. С. Вышемирский, А. Н. Дмитриев, А. А. Трофимук, Мировой нефтяной конгресс, М., 1971, [8] А. Н. Дмитриев, В. В. Золотухин, Ю. Р. Васильев. Сов. геол., № 12, 64 (1968).