АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
ГЕОЛОГИЯ И ГЕОФИЗИКА

№ 11 1976

УДК 550:51+553.98

А. Н. ДМИТРИЕВ, В. О. КРАСАВЧИКОВ

ПРОЦЕДУРЫ МАТЕМАТИЧЕСКОЙ ОБРАБОТКИ ОПИСАНИЙ НЕФТЯНЫХ МЕСТОРОЖДЕНИЙ

Излагаются процедуры математической обработки описаний месторождений углеводородов. Цель обработки информации — решение задач прогнозно-поискового профиля. Описываемые алгоритмы — элементы системы последовательного распознавания. Их изложение и краткая содержательная трактовка даются с учетом возможного применения и для рудных ископаемых.

В работе излагается ряд процедур математической обработки описаний геологических объектов. Эти процедуры являются составной частью системы последовательного распознавания, созданной в процессе сравнительного изучения описаний нефтяных месторождений и предназначенной для решения задач прогнозно-поискового профиля. Последовательное распознавание проводится за несколько шагов, поэтапно, причем результаты его на предыдущем шаге прямо или косвенно учитываются на последующем шаге. Используемый на последующем шаге алгоритм распознавания может совпадать с алгоритмом предыдущего шага, но при этом изменяются объекты обучения, либо наоборот, для той же обучающей выборки может быть использован другой алгоритм распознавания. Кроме того, на одном шаге может быть параллельно получено несколько результатов распознавания для одного и того же исследуемого объекта (пробы) в связи с использованием различных наборов характеристических признаков.

При построении указанной выше системы выбор или разработка новых математических процедур осуществлялись с учетом целеуказания, природы сообщений, подлежащих обработке, и объема информации. Особенности геологических постановок задач [15, 16] и требуемая детальность распознавания обусловили применение комплекса алгоритмов. В этом комплексе, проверенном на широком круге задач нефтяного прогноза и поиска, выделено три функциональных блока.

1. Блок первоначального распознавания — грубая, установочная сортировка объектов. Алгоритм распознавания блока основан на методе согласованных оценок объектов и признаков — строк и столбцов таблиц описаний (программа «Качели П2») [3, 6, 9].

2. Блок уточнения распознавания — уточнение распознавания и его детализация. Используются аглоритмы метода суммарного учета мер приуроченности и согласования (программа «Каскад П6») [2, 14].

3. Блок поиска аналога — для каждого описания пробы осуществляется поиск ближайшего аналога среди описаний эталонов. Используется алгоритм голосования по тесторам, входящий в тестовый подход (программа «Тестор П5») [7, 11—13].

В соответствии со спецификой каждого блока и процедур общая задача, определяемая основным целеуказанием, в каждом конкретном случае подвергалась соответствующей перепостановке.

1. Метод согласованных оценок в распознавании образов

Описываемый алгоритм и реализующая его программа [6, 9] основаны на приложении метода согласованных оценок («Качелей») [3] к общей схеме распознавания на базе понятия типичной строки («реплики») [4].

1.1. Общая схема распознавания

Пусть задано два класса объектов, чьими эталонами являются $s^1_1, \ldots, s^1_m$ и $s^2_1, \ldots, s^2_{m_2}$ соответственно. Описания эталонов бинарными характеристическими признаками $x_1, \ldots, x_n$ образуют таблицы T_1 и T_2.* Строка $(\tilde{t}_1, \ldots, \tilde{t}^k_n) = \tilde{s}^k$, где $k=1, 2$, называется типичной для T_k, если

$$\tilde{t}^k_j = \begin{cases} 1 \text{ при } \sum\limits_{i=1}^{m_k} t^k_{ij} \geqslant \dfrac{m_k}{2} \\ 0 \text{ при } \sum\limits_{i=1}^{m_k} t^k_{ij} < \dfrac{m_k}{2}, \end{cases}$$

где t^k_{ij} — элемент таблицы T_k. При $\sum\limits_{i=1}^{m_k} t^k_{ij} = \dfrac{m_k}{2}$ значение $\tilde{t}^k_j$ может определяться также из неформальных, практических соображений или случайным путем. Рассмотрим следующую схему распознавания на базе понятия «реплика». Пусть $P^k = (P^k_1, \ldots, P^k_n)$ — произвольный неотрицательный вектор нагрузок столбцов таблицы T_k, причем $\sum\limits_{j=1}^{n} P^k_j = 1$. Для произвольной строки $s = (t_1, \ldots, t_n)$ положим $r(P^k, s) = \sum\limits_{j=1}^{n} |t_j + \tilde{t}^k_j - 1| \times$ $\times P^k_j$ и $R(P^1, P^2, s) = \dfrac{r(P^1, s)}{r(P^2, s)}$ (при $r(P^2, s) \neq 0$). Величина $r(P^k, s)$ представляет собой взвешенное число совпадений строки s с типичной строкой таблицы T_k, и ее можно принять в качестве меры близости s к классу, представленному эталонами $s^k_1, \ldots, s^k_{m_k}$. Величина $R(P^1, P^2, s)$ оценивает тяготение строки s к одному из указанных двух классов при заданных мерах близости. Если $R(P^1, P^2, s) > 1$, то s тяготеет к первому классу, если $R(P^1, P^2, s) < 1$, то ко второму. При $R(P^1, P^2, s) = 1$ мы ничего не можем сказать о принадлежности s. Естественно сформулировать такое решающее правило для диагностики испытуемых объектов s, описаниями которых являются строки $(t_1, \ldots, t_n)$:

а) при $R(P^1, P^2, s) \geqslant 1+\varepsilon$ s относится к первому классу;

б) при $R(P^1, P^2, s) \leqslant 1-\varepsilon$ s относится ко второму классу;

в) при $1-\varepsilon_2 < R(P^1, P^2, s) < 1+\varepsilon$ s не распознается.

Здесь пороги $\varepsilon_1, \varepsilon_2 > 0$ определяют «решительность» алгоритма распознавания.

Наиболее простой процедурой подобного рода является, по-видимому, следующая: положим $P = P^k_{II} = (1/n, \ldots, 1/n)$, т. е. все признаки считаются в равной степени существенными для диагностики. Тогда $r(P^k_{II}, s) = r_{II}(s) = \dfrac{1}{n} \sum\limits_{j=1}^{n} |t_j + \tilde{t}^k_j - 1|$, т. е. равняется числу совпадений

* Где на пересечении i-той строки и j-того столбца стоит значение признака x_j на объекте s^k_i, к=1,2 : единица, если x_j выполняется, и ноль в противном случае.

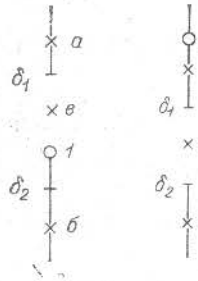

Рис. 1.

строки s с типичной строкой $\tilde{s}^k$, поделенному на n, а $R(P^1_{II}, P^2_{II}, s)=R_{II}(s)$ — отношение числа совпадений с типичной строкой первого класса к числу совпадений с типичной строкой второго класса.

Для того чтобы указанное решающее правило можно было применять к решению той или иной конкретной задачи диагноза, необходимо убедиться в его способности распознавать объекты обучения $s^1_1, \ldots, s^1_{m_1}$, $s^2_1, \ldots, s^2_{m_2}$, т. е. проверить, что если s^k_i — строка, соответствующая объекту s^k_i где $i=1, \ldots, m_k$, то

$$(*)\begin{cases} R(P^1, P^2, s^1_i) > 1 & i=1, \ldots, m_1 \\ R(P^1, P^2, s^2_i) < 1 & i=1, \ldots, m_2. \end{cases}$$

Такое распознавание объектов обучения будем называть устойчивым *. Однако в ряде случаев условия (*) могут не выполняться, тем не менее может быть сформулирована приемлемая с содержательной точки зрения процедура распознавания. Такая ситуация имеет место, если выполняются условия (**) $R(P^1, P^2, s^1_i) > R(P^1, P^2, s^2_{i'})$ для всех $i=1, \ldots, m_1$, $i'=1, \ldots, m_2$. Положим $\delta_1 = \min\limits_{j=1,\ldots,m_1} R(P^1, P^2, s^1_i)$, $\delta_2 = \max\limits_{i'=1,\ldots,m_2} R(P^1, P^2, s^2_{i'})$, тогда (**) означает, что $\delta_1 > \delta_2$. Сформулируем решающее правило:

а) при $R(P^1, P^2, s) \geqslant \delta_1$ s относится к первому классу;

б) при $R(P^1, P^2, s) \leqslant \delta_2$ s относится ко второму классу;

в) при $\delta_2 < R(P^1, P^2, s) < \delta_1$ s не распознается.

На рис. 1 схематически показаны случаи а, б, в. Распознавание по этому правилу при невыполнении условий (*) называется неустойчивым. Устойчивое распознавание представляется более приемлемым.

1.2. Меры характеристичности и типичности

Сформулируем основные требования к выбору нагрузок P^1, P^2. Обозначим через E^N, $N=1, 2 \ldots$ N-мерный единичный куб. Пусть T — таблица описаний объектов $s_1, \ldots, s_m$ бинарными признаками $x_1, \ldots, x_n$. Произвольному m-мерному вектору-столбцу $x=(\alpha_1, \ldots \alpha_m)'$, где „′“ — знак транспонирования, сопоставим его типизированный вариант $\tilde{x}=(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_m)'$, где $\tilde{\alpha}_i=\alpha_i$ при $\sum\limits_{i=1}^{m} \alpha_i \geqslant m/2$ и $\tilde{\alpha}_i=1-\alpha_i$ при $\sum\limits_{i=1}^{m} \alpha_i < m/2$**.

Таблице T сопоставим ее типизированный вариант $\tilde{T}$ [10]. Столбцами таблицы $\tilde{T}$ являются векторы $\tilde{x}_j$ — типизированные варианты столбцов x_j таблицы T. Для произвольных векторов $a=(a_1, \ldots, a_N)$, $b=(b_1, \ldots, b_N)$, $a \leqslant b$ означает, что $a_1 \leqslant b_1, \ldots, a_N \leqslant b_N$***. Пусть $\tilde{T}$ составлена из описаний объектов $s_1, \ldots, s_m$ бинарными признаками $x_1, \ldots, x_n$.

Определение 1. Неотрицательная функция $\mu(x)$, заданная на E^m, называется мерой характеристичности признака, если для любых $x', x'' \in E^m$ $\tilde{x}' < \tilde{x}''$ влечет $\mu(x') < \mu(x'')$.

Пусть $\tilde{s}^* = (\tilde{t}^*_1, \ldots, \tilde{t}^*_n)$ — „реплика“ T, где $\tilde{t}_j$ представляет собой типичное значение признака x_j, $j=1, \ldots, n$.

---

* Подобная проверка не исключает и обычного внешнего экзамена процедуры распознавания.

** Таким образом, при $x=x_j$, $j=1, \ldots, n$ в первом случае типичным значением x_j для объектов $S_1, \ldots, S_m$ является единица, а во втором ноль. Кроме того, при $x=x_j$ и $\sum\limits_{i=1}^{m} \alpha_i = m/2$ $\tilde{x}_j$ может определяться и иным путем (см. 1.1), но так, чтобы это не приводило к недоразумениям.

*** $a<b$ равносильно $a \leqslant b$ и $a \neq b$.

Для произвольного вектора $s \in E^n$ положим $\tilde{s} = (\tilde{t}_1, \ldots, \tilde{t}_n)$, где $\tilde{t}_j = |t_j + \tilde{t}_j^* - 1|$

Определение 2. Неотрицательная функция $\eta(s)$, заданная на E^n, называется мерой типичности объекта связанной с $\tilde{s}^*$, если для любых s', $s'' \in E^n$ $\tilde{s}' < \tilde{s}''$ влечет $\eta(s') < \eta(s'')$. Если $\alpha = (t_1, \ldots, t_m)'$ — столбец значений признака x на объектах $s_1, \ldots, s_m$, то число $\mu(\alpha)$ называется мерой характеристичности признака x для объектов $s_1, \ldots, s_m$. Положим $\mu(x) = \mu(\alpha)$. Для объекта s число $\eta(s)$ называется мерой типичности s в классе, представленном эталонами $s_1, \ldots, s_m$.

Если $\mu(x)$ — мера характеристичности, то (3*) $\eta(s) = a \sum_{j=1}^{n} \tilde{t}_j \mu(x_j)$ будет мерой типичности; и наоборот, если $\eta(s)$ — мера типичности, то (4*) $\mu(x) = b \sum_{i=1}^{m} \tilde{\alpha}_i \eta(s_i)$ — мера характеристичности, где a, b — произвольные положительные числа, $\mu(x_j) > 0$, $j = 1, \ldots, n$, $\eta(s_i) > 0$, $i = 1, \ldots, m$.

При определении векторов нагрузок P^1, P^2 естественно потребовать, чтобы числа P_j^1, P_j^2 были мерами характеристичности признаков x_j для объектов $s_1^1, \ldots, s_{m_1}^1$ и $s_1^2, \ldots, s_{m_2}^2$ соответственно. Тогда $r(P^k, s) = \eta^k(s)$ будет мерой типичности объекта s в указанном выше смысле.

1.3. Конкретизация мер характеристичности и типичности

Простейшим примером меры характеристичности является частота, с которой признак принимает свое типичное значение. Простейшим примером меры типичности будет число совпадений строки s с репликой таблицы Т. Однако при таком способе оценки мера типичности объекта не зависит от того, насколько характеристичны свойственные ему признаки, а мера характеристичности признака не зависит от типичности тех объектов, на которых он принимает свое типичное значение. Для установления взаимосвязи между η и μ, потребуем, чтобы для них выполнялись одновременно соотношения (3*) и (4*). В этом случае мерой характеристичности признака x_j является (с точностью до постоянного множителя) сумма мер типичности тех объектов, где он принимает свое типичное значение. И, наоборот, мерой типичности объекта s является сумма мер характеристичности тех признаков, которые на объекте s принимают свое типичное значение. При этом мера типичности, порождаемая по формуле (3*) функцией $\mu(x_j)$, совпадает (с точностью до постоянного множителя) с $\eta(s)$, а мера характеристичности, порождаемая $\eta(s)$, — с $\mu(x_j)$. Иначе говоря, меры характеристичности и типичности являются взаимосогласованными. Полагая для $i = 1, \ldots, m$; $j = 1, \ldots, n$; $\eta(s_i) = \tilde{\omega}_i$; $\mu(x_j) = \tilde{\pi}_j$; $1/a = c$, $1/b = d$, запишем соотношения (3*)—(4*) для строк и столбцов таблицы Т:

$$c\tilde{\omega}_i = \sum_{j=1}^{n} \tilde{t}_{ij} \cdot \tilde{\pi}_j, \tag{3*}$$

$$d\tilde{\pi}_j = \sum_{i=1}^{m} \tilde{t}_{ij} \cdot \tilde{\omega}_i, \tag{4*}$$

где c и d — произвольные положительные числа. Приходим к основным уравнениям метода согласованных оценок [3] для таблицы $\tilde{T}$. Решение этой системы — «качельные» нагрузки столбцов и строк $\tilde{T}$ — определяют взаимосогласованные меры характеристичности признаков и типичности объектов. Для этого, кроме требования измеримости $\tilde{T}$ [3], сле-

дует предположить, что $\widetilde{T}$ не имеет нулевой строки. Практически любая геологическая таблица удовлетворяет этому требованию.

Пусть $\widetilde{\pi}^k = (\widetilde{\pi}^k_1, \ldots, \widetilde{\pi}^k_n)$ $\widetilde{\omega}^k = (\widetilde{\omega}^k_1, \ldots, \widetilde{\omega}^k_{m_k})$ — «качельные» нагрузки столбцов и строк таблицы $\widetilde{T}_k$, $\mathscr{P}^k_j$ — частота принятия признаком x_j своего типичного значения. Справедлива формула (6*) $\widetilde{\pi}^k_j = \gamma^k \left[\mathscr{P}^k_j + \sum_{i=1}^{m_k} \times \right.$ $\left. \times \left(\frac{\widetilde{\omega}^k_i}{\gamma^k} - \frac{1}{m_k} \right) \widetilde{t}^k_{ij} \right] = \gamma^k \left(\mathscr{P}^k_j + \sum_{i=1}^{m_k} \sigma^k_i \widetilde{t}^k_{ij} \right)$, где $\gamma^k = \sum_{i=1}^{m_k} \widetilde{\omega}^k_i$, $\sigma^k_i = \frac{\widetilde{\omega}^k_i}{\gamma^k} - \frac{1}{m_k}$ причем $\sigma^k_i \geqslant 0$ означает, что $\omega^k_i \geqslant \gamma^k / m_k$, где γ^k / m_k — среднее арифметическое чисел $\widetilde{\omega}^k_1, \ldots, \widetilde{\omega}^k_{m_k}$. Отметим также, что $\widetilde{\omega}^k_i = r(\widetilde{\pi}^k, s^k_i)$.

Использование $\widetilde{\omega}^k_i$ как меры типичности объекта s^k_i в группе объектов $s^k_1, \ldots, s^k_{m_k}$ проводилось на ряде геологических примеров и дало результаты, хорошо согласующиеся с содержательной трактовкой типичности геологических объектов. Таким образом, можно сказать, что при $\sigma^k_i > 0$ «типичность» объекта s^k_i в группе $s^k_1, \ldots, s^k_{m_k}$ выше, а при $\sigma^k_i < 0$ — ниже средней. Если $\mathscr{P}^k_j$ оценивает частоту встречаемости типичного значения, то $\sum_{i=1}^{m_k} \sigma^k_i \widetilde{t}^k_{ij}$ оценивает «качество» выполнения этого значения. Например, если для x_j, $x_{j'}$ $\mathscr{P}^k_j = \mathscr{P}^k_{j'}$, то неравенство $\widetilde{\pi}^k_{j'} > \widetilde{\pi}^k_j$ означает, что объекты s^k_i, для которых $\widetilde{t}_{ij'} = 1$, в среднем более типичны, чем объекты s^k_i, для которых $t_{ij} = 1$.

Все вышесказанное позволяет утверждать, что с содержательной точки зрения величина $\widetilde{\pi}^k_i$ может быть истолкована как мера характеристичности признака x_j для объектов $s^k_1, \ldots, s^k_{m_k}$ совмещающая в себе как частотную, так и «качественную» оценку. Поэтому естественно ожидать, что если мы используем полученные нагрузки столбцов $\widetilde{\pi}^k_j$ в описанной выше диагностической схеме, положив $\mathrm{P}^k_{\mathrm{I}} = \left(\frac{\widetilde{\pi}^k}{\sum_{j=1}^{n} \widetilde{\pi}^k_j}, \ldots \right.$ $\left. \ldots, \frac{\widetilde{\pi}^k_n}{\sum_{j=1}^{n} \widetilde{\pi}^k_j} \right)$, $r^k_{\mathrm{I}} = r(\mathrm{P}^k_{\mathrm{I}}, s)$, $R_{\mathrm{I}}(s) = \frac{r(\mathrm{P}^1_{\mathrm{I}}, s)}{r(\mathrm{P}^2_{\mathrm{I}}, s)}$, то полученный алгоритм распознавания окажется применимым для широкого круга геологических задач. Опыт его применения для решения ряда геологических задач показал следующее: во всех случаях результаты распознавания по R_{I} были не хуже, чем по R_{II}. Отмечены случаи неразделения классов по R_{II} (т. е. $\delta_2 > \delta_1$) при их разделении по R_{I}, а также устойчивого распознавания по R_{I} при неустойчивом распознавании по R_{II}, хотя даже и в таких ситуациях общая картина распределения значений R_{I} и R_{II} примерно одинаковая.

2. Метод суммарного учета мер приуроченности и согласования в распознавании образов

Среди большого числа косвенных признаков, которыми оперирует геолог при решении прогнозных задач, особого внимания заслуживают признаки, которые выполняются на продуктивных (в принятых терминах нефтегеологии) объектах гораздо чаще, чем на «пустых» объектах.

Выявление и суммарный учет таких признаков, зачастую производимый на интуитивном уровне, играют значительную роль в традиционном геологическом прогнозировании.

Аналогичным образом может производиться априорная оценка важности перспективного на определенный вид сырья участка на стадии его добуровой охарактеризованности. При этом в первую очередь привлекаются те признаки, которые, как показывает профессиональный опыт, в большей степени коррелируют с целевым признаком (например, масштабом запасов) объектов.

Настоящий метод реализует один из возможных подходов к математизации этих процедур поиска и оценки перспектив месторождений полезных ископаемых. В этом подходе вводятся меры приуроченности (для задач распознавания) и меры согласования, или связи, между бинарным косвенным признаком и числовым целевым (для задач упорядочения). Вводимые меры определяются аксиоматическим путем. Следует отметить, что до введения аксиоматических определений были построены конкретные примеры этих мер [2] совместно с В. В. Бабичем и Г. С. Федоссевым. Ими же разработана методика применения данных мер при распознавании и упорядочении [2].

2.1. Меры приуроченности

Рассмотрим задачу распознавания для случая двух классов. Целевой признак x_{n+1} предполагается бинарным: равным единице для объектов первого класса и нулю — для второго. Задано пространство характеристических признаков $X=\{x_1, \ldots, x_n\}$ и выборки объектов первого и второго классов $s_1=\{s_1^1,\ldots, s_{m_1}^1\}$, $S_2=\{s_1^2,\ldots, s_{m_2}^2\}$, $S=S_1\cup S_2$. Для простоты изложения будем отождествлять признаки x_j с отвечающими им столбцами значений на объектах из S: $x_j=(x_j(s_1^1),\ldots,x_j(s_{m_1}^1), x_j(s_1^2), \ldots, x_j(s_{m_2}^2))'=(t_{1j}^1,\ldots,t_{m_1j}^1, t_{1j}^2,\ldots,t_{m_2j}^2)'$. Для произвольных столбцов из $E^{m_1+m_2}$ также будем использовать обозначение x_j, предполагая, что если столбец $x_j\notin X$, $x_j\neq x_{n+1}$, то $j>n+1$.

Определение 3. На множестве $E^{m_1+m_2}$ введем отношение частичного порядка $\overset{+}{\leqslant}$ следующим образом: $x_j \overset{+}{\leqslant} x_k$, если $t_{ij}^1\leqslant t_{ik}^1$, $i=1,\ldots,$ $,\ldots, m_1$ и $t_{ij}^2\geqslant t_{ik}^2$, $i=1,\ldots, m_2$ *.

Отметим, что максимальным элементом множества $E^{m_1+m_2}$ по отношению $\overset{+}{\leqslant}$ будет столбец, совпадающий со столбцом значений целевого признака x_{n+1} на обучающей выборке: $\underbrace{(1,1,\ldots,}_{m_1}\underbrace{,0,0\ldots 0)'}_{m_2}$. Минимальным же будет столбец, совпадающий с $\bar{x}_{n+1}$: $\underbrace{(0,\ldots,0,}_{m_1}\underbrace{\quad}_{m_2}\underbrace{1,\ldots,1)'}_{m_2}$. Здесь $\bar{x}_j$ означает булево отрицание признака x_j: $\bar{x}_j(s)=1-x_j(s)$ для любого объекта s из области определения x_j.

Пример 1. $m_1=2$, $m_2=2$. На рис. 2 линиями соединены соседние по отношению $\overset{+}{\leqslant}$ вершины E^4, причем если x_j и x_k соединены чертой и x_j лежит ниже x_k, то $x_j \overset{+}{\leqslant} x_k$.

Определение 4. Функция $g(x)$, заданная на $E^{m_1+m_2}$ называется оценкой приуроченности, если для любого $x_j\in E^{m_1+m_2}$

1) $|g(x_j)|\leqslant 1$,
2) $g(\bar{x}_j)=-g(x_j)$,
3) $g(x_j)=1$ в том и только в том случае, когда x_j совпадает с x_{n+1} и $g(x_j)=-1$ в том и только в том случае, когда x_j совпадает с $\bar{x}_{n+1}$,
4) если x_j — нулевой или единичный столбец, то $g(x_j)=0$,
5) если $x_j \overset{+}{\leqslant} x_k$, то $g(x_j)\leqslant g(x_k)$.

* Отношение aRb называется отношением частичного порядка, если 1) aRa, 2) aRb и bRa влечет $a=b$, 3) aRb и bRc влечет aRc для любых a, b, c из области его определения.

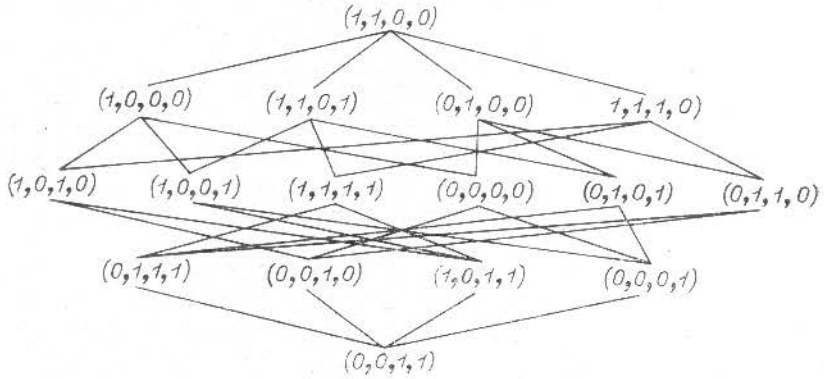

Рис. 2.

Определение 5. Функция $|g(x)|$, где $g(x)$ — оценка приуроченности, называется мерой приуроченности.

Наиболее естественным образом интуитивному содержанию понятия приуроченности отвечает величина разности частот выполнения оцениваемого признака x_j на объектах первого класса (месторождений или иных «важных» с позиции решаемой задачи геологических образований) и объектах второго класса, менее важных относительно поставленной цели. Обозначим ее P_j^y. Таким образом

$$(*)\quad P_j^y = \frac{\sum_{i=1}^{m_1} t_{ij}^1}{m_1} - \frac{\sum_{i=1}^{m_2} t_{ij}^2}{m_2} = \mathscr{P}_j^{I} - \mathscr{P}_j^{II} = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{l=1}^{m_2} (t_{ij}^1 - t_{lj}^2) =$$

$$= \sum_{\substack{\text{по всем} \\ \text{парам } (i,l)}} (t_{ij}^1 - t_{lj}^2) / m_1 m_2.$$

Условия 1—5 определения 4 выполняются для P_j^y очевидным образом.

При оценке признака величиной P_j^y все объекты внутри первого класса, как и все объекты внутри второго, рассматриваются равноправными. Однако предположение о равноправности приемлемо далеко не всегда. Неравноправность объектов обучения может быть связана как с различиями по их важности (запасам и т. п.), так и с различиями по их типичности для своих классов. Уточним величину P_j^y на тот случай, когда для объектов обоих классов можно определить неотрицательную функцию важности объекта $f(s)$ (например, запасы), такую, что $\min\limits_{s \in S_1} f(s) > \max\limits_{s \in S_2} f(s)$. Для этого введем

$$\varphi^1[x_j, f] = \varphi_j^1 = \frac{1}{m_1 m_2 (f^{cpI} - f^{cpII})} \sum_{i=1}^{m_1} \sum_{l=1}^{m_2} (t_{ij}^1 - t_{lj}^2)(f(s_i^1) - f(s_l^2)),$$

где f^{cpI}, f^{cpII} — средние значения f в S_1 и S_2. Нетрудно проверить, что φ_j^1 — оценка приуроченности. В случае, когда $f(s_i^1) = C_1$, $i = 1, \ldots, m_1$, $f(s_l^2) = C_2$, $i = 1, \ldots, m_2$, где $C_1 > C_2 \geqslant 0$ — постоянные, $\varphi_j^1 = P_j^y$. Если $x_j, x_{j'} \in E^{m_1+m_2}$ и $x_j + x_{j'} = x_r$, причем $x_r \in E^{m_1+m_2}$, то $\varphi^1[x_r, f] = \varphi^1[x_j, f] + \varphi^1[x_{j'}, f]$. Кроме того, φ_j^1 инвариантна относительно линейных преобразований функции f вида $y = ax + b$, где $a \geqslant 0$, оставляющих f неотрицательной. Помимо упомянутых выше, в [2] приведены и другие конкретные примеры оценок приуроченности.

Пример 2.

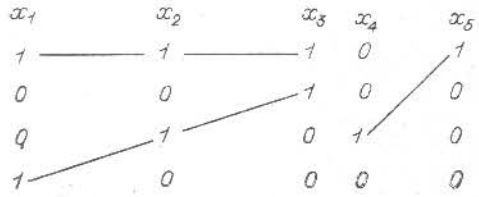

Рис. 3.

Задано m объектов $S=\{s_1, \ldots, s_m\}$, относящихся к одному и тому же классу. Задана неотрицательная функция важности f, которая является количественным выражением целевого признака x_{n+1}, $f(s_1)\geqslant f(s_2)\geqslant \ldots \geqslant f(s_m)$, где $f(s_1)>f(s_m)$. Как и выше, признаки x_j отождествляются с отвечающими им векторами $(x_j(s_1), x_j(s_2), \ldots, x_j(s_m))'=(t_{1j}, \ldots, t_{mj})'$. Обозначение x_j при $j>n+1$ соответствует векторам из E^m, не входящим в X. Обозначим через $\mathscr{P}_j$ частоту выполнения на множестве $\{s_1, \ldots, s_m\}$ $\mathscr{P}_j=\sum_{i=1}^{m} t_{ij}/m$. Пусть в столбце x_j имеется c единиц, стоящих в строках с номерами $i_1(j)<i_2(j)<\ldots<i_c(j)$, где $c=m\cdot\mathscr{P}_j$. На множестве E^m введем отношение частичного порядка $\overset{\circ}{\leqslant}$ следующим образом: $x_j\overset{\circ}{\leqslant}x_h$, если $\mathscr{P}_j=\mathscr{P}_h$ и $i_l(j)\geqslant i_l(k)$ для всех $l=1, \ldots, c$.

Пример 2. Здесь $x_1\leqslant x_2\leqslant x_3$, $x_1\leqslant x_5$. Соответствующие единицы сравниваемых наборов соединены линиями (рис. 3).

Определение 6. Функция $g(x)$ на E^m, удовлетворяющая сформулированным ниже условиям 1—4, называется оценкой согласования.

1. $g(\bar{x}_j)=-g(x_j)$.

2. Если $x_j=(\overbrace{0,0, \ldots, 0}^{m})'$ или $x_j=(\overbrace{1,1, \ldots, 1}^{m})'$, то $g(x_j)=0$.

3. Если $x_j\overset{\circ}{\leqslant}x_k$, то $g(x_j)\leqslant g(x_k)$.

4. Если $x_j=(\overbrace{1, 1, 1, \ldots, 1}^{r}, \overbrace{0, \ldots, 0}^{p})'$, $1\leqslant r<m$, то $g(x)>0$.

Функция $|g(x_j)|$, где $g(x_j)$ — оценка согласования, заданная на E^m, называется мерой согласования.

Пусть x_j, $x_k\in E^m$ и $x_j+x_k\in E^m$. Функция g, заданная на E^m и принимающая вещественные значения, называется аддитивной, если для любых таких x_j, x_h, $g(x_j+x_k)=g(x_j)+g(x_k)$. Положим $x^{(i)}=(\underbrace{0,0, \ldots 0, 1, 0}_{i}, \ldots)$

Тогда $g(x_j)=\sum_{i=1}^{m} t_{ij}\cdot g(x^{(ji)})$.

Характерным примером оценки согласования является хорошо известный выборочный коэффициент корреляции $r_{\xi\eta}$ (см., например, [8]), подсчитанный для выборок $f=\{f(s_1), f(s_2), \ldots, f(s_m)\}$ и $x_j=\{x_j(s_1), \ldots, x_j(s_m)\}$ обычным путем *. Однако такая оценка, вообще говоря, не будет аддитивной. Формулируемое ниже утверждение показывает, что аддитивные оценки согласования могут быть получены из аддитивных оценок приуроченности (P_j^{∇}, φ_j^1 и др.). Разобъем S на пары попарно не пересекающихся классов: $S=S_1^{(k)}\cup S_2^{(k)}$, где $S_1^{(k)}=\{s_1, \ldots, s_k\}$, $S_2^{(k)}=\{s_{k+1}, \ldots, s_m\}$, $k=1, \ldots, m-1$.

Теорема 1: Если $g^k(x_j)$ — аддитивные оценки приуроченности для пар $S_1^{(k)}, S_2^{(k)}$, такие, что при $i\leqslant l$ $g^k(x^{(i)})\geqslant g^k(x^{(l)})$ и $\alpha_1, \ldots, \alpha_{m-1}$ — произвольные неотрицательные числа, не все равные нулю, то $g(f, x_j)=\sum_{i=1}^{m-1}\times$ $\times\alpha_k g^k(x_j)$ — аддитивная оценка согласования. Справедливость этого утверждения вытекает непосредственно из определений 4 и 6.

* Если x_j —нулевой или единичный столбец, по определению полагаем $r_{fx_j}=0$.

Пример 3. Оценки P_j^{ν} и φ_j^1 очевидным образом удовлетворяют условию теоремы 1. Поэтому $\sum_{k=1}^{m-1}\sum_{i=1}^{k}\sum_{l=k+1}^{m}\frac{(t_{ij}-t_{lj})}{k(m-k)}=\sum_{k=1}^{m-1}P_j^{\nu(k)}$ и

$$\sum_{k=1}^{m-1}\sum_{i=1}^{k}\sum_{l=k+1}^{m}\frac{(t_{ij}-t_{lj})(f(s_i)-f(s_l))}{k(m-k)(f^{\text{ср I}(k)}-f^{\text{ср II}(k)})}=\sum_{k=1}^{m-1}\varphi_j^{1(k)}=\varphi_j^{*},$$

(где $f^{\text{ср I}(k)}$, $f^{\text{ср II}(k)}$ — средние значения $f(s)$ в $S_1^{(k)}$ и $S_2^{(k)}$ соответственно) являются аддитивными оценками согласования.

В работе [2] приводятся и другие примеры оценок согласования.

2.3. Алгоритмы распознавания и упорядочения

Согласно [2], опишем алгоритмы распознавания и упорядочения, основанные на мерах приуроченности и согласования, а также укажем способы минимизации пространства признаков. Пусть функция $\theta(x)$ определена для любого вещественного числа x следующим образом: $\theta(x)=1$ при $x>0$ и $\theta(x)=0$ при $x\leqslant 0$.

Алгоритм распознавания. Пусть для распознавания проб отобраны признаки $x_{j_1},\ldots,x_{j_l}$. Для объекта $s=(t_1,\ \ldots,\ t_n)$ положим $\mathcal{J}^l[s]=\sum_{r=1}^{l}(t_{j_r}|g(x_{j_r})|\theta(g(x_{j_r}))+(1-t_{j_r})|g(x_{j_r})|\theta(g(\bar{x}_{j_r}))$ где $g(x_{j_r})\neq 0$, $g(x_j)$ — оценка приуроченности. Пусть

$$a^l=\min_{i=1,\ldots,m_1}\mathcal{J}^l[s_i^1],\ b^l=\max_{r=1,\ldots,m_2}\mathcal{J}^l[s_i^2].$$ Тогда:

а) при $\mathcal{J}^l[s]\geqslant\max(a^l,\ b^l)+\varepsilon_1$ s относится к первому классу;

б) $\mathcal{J}^l[s]\leqslant\min(a^l,\ b^l)-\varepsilon_2$ s относится ко второму классу;

в) при $\min(a^l,\ b^l)-\varepsilon_2<\mathcal{J}^l[s]<\max(a^l,\ b^l)+\varepsilon_1$ s не распознается.

Минимизация пространства признаков для алгоритма распознавания. Пусть $x_{j_1},\ldots,x_{j_n}$ — упорядочение признаков исходного признакового пространства X по убыванию меры приуроченности $|g(x_j)|$. Для $l=1,\ \ldots,\ n$ положим $\rho^l=a^l-b^l$. Пусть $\bar{l}$ — наименьшее, удовлетворяющее соотношению $\bar{\rho}^l=\max\rho^l$. Пространство X минимизируется до набора $x_{j_1},\ \ldots,\ x_{j_{\bar{l}}}$.

Алгоритм упорядочения. Пусть для упорядочения объектов отобраны признаки $x_{j_1},\ \ldots,\ x_{j_l}$. Для объекта $s=(t_1,\ \ldots,\ t_n)$ положим

$$\mathcal{J}^l[s]=\sum_{r=1}^{l}(t_{j_r}|g(x_{j_r})|\theta(g(x_{j_r}))+(1-t_{j_r})|g(x_{j_r})|\theta(g(\bar{x}_{j_r}))),$$

где $g(x_j)$ — оценка согласования. Пусть упорядоченность объектов $s_1,\ \ldots,\ s_m$ по убыванию величины $\mathcal{J}^l[s_i]$ «совпадает или сильно коррелирована» [5] с их упорядоченностью по степени проявления целевого признака. Пусть s — проба. Вычисляя для нее $\mathcal{J}^l[s]$, найдем пару соседних объектов s_k, s_{k+1}, такую, что $\mathcal{J}^l[s_{k+1}]\leqslant\mathcal{J}^l[s]\leqslant\mathcal{J}^l[s_k]$, т. е. найдем место пробы в упорядоченном ряду эталонных объектов*. Кроме того, можно упорядочить заданную последовательность проб $s_{m+1},\ \ldots,\ s_q$ по степени проявления целевого признака.

Отметим, что приведенные алгоритмы упорядочения и распознавания инвариантны относительно инверсии признаков исходного признакового пространства X. Для реализации этих алгоритмов можно воспользоваться программами [1, 2, 14] и другими, описанными в [2]. Там же приводятся процедуры оптимального бинарного кодирования.

---

* Если указанные выше упорядоченности не вполне совпадают, то можно найти наиболее близкие к s по величине $\mathcal{J}^l[s_i]$ эталоны.

2.3. Алгоритм голосования по тупиковым тесторам

Пусть $T_1, \ldots, T_\mu, \ldots, T_M$ — бинарные таблицы, составленные из описаний объектов $S_1^\mu, \ldots, S_{m_\mu}^\mu$, $\mu = 1, \ldots, M$ бинарными признаками $x_1, \ldots, x_n$, где $T_\mu = (t_{ij}^\mu)_{m_\mu \times n}$. Объекты $s_1^\mu, \ldots, s_{m_\mu}^\mu$ — эталоны класса $\mathcal{J}_\mu$, $\mu = 1, \ldots, M$, причем при $\mu_1 \neq \mu_2$ $\mathcal{J}_{\mu_1} \cap \mathcal{J}_{\mu_2} = \varnothing$. Согласно [5], тупиковым тестором для $T_1, \ldots, T_M$ называется набор столбцов (признаков) $t = (x_{j_1}, \ldots, x_{j_l})$, обладающий следующими свойствами:

1) если $\mu_1 \neq \mu_2$, то для любых $s_i^{\mu_1}$, $s_k^{\mu_2}$, $i = 1, \ldots, m_{\mu_1}$, $k = 1, \ldots, m_{\mu_2}$ $(t_{ij_1}^{\mu_1}, t_{ij_2}^{\mu_1}, \ldots, t_{ij_l}^{\mu_1}) \neq (t_{kj_1}^{\mu_2}, t_{kj_2}^{\mu_2}, \ldots, t_{kj_l}^{\mu_2})$ при любых $\mu_1, \mu_2 = 1, \ldots, M$.

2) Никакой поднабор $t' \subset t$ указанным выше свойством не обладает.

В работах [7, 11, 13] рассмотрена общая схема распознавания, использующая процедуру голосования по тесторам. Опишем то конкретное решающее правило, которое использовано в настоящей работе. Ему соответствует программа для БЭСМ-6, приведенная в работе [12].

Определение 7. Для данной строки $s = (t_1, \ldots, t_n)$ тестор $t = (x_{j_1}, \ldots, x_{j_l})$ голосует за строку (объект) s_i^μ, где $i = 1, \ldots, m_\mu$, $\mu = 1, \ldots, M$, если $(t_{j_1}, \ldots, t_{j_l}) = (t_{ij_1}^\mu, \ldots, t_{ij_l}^\mu)$.

Обозначим через $\Gamma(ss_i^\mu)$ общее число голосов, поданных за s_i^μ для данной строки s, деленное на общее число тесторов. Пусть s — описание объекта, подлежащего распознаванию (пробы). Вычислим $\Gamma_\mu(s) = \max_{i=1,\ldots,m_\mu} \Gamma(s, s_i^\mu) = \Gamma(s, s_{i_0}^\mu)$. Пусть $\bar{\mu}$ таково, что $\Gamma_{\bar{\mu}}(s) = \max \Gamma_\mu(s)$ и числа $\varepsilon_1, \varepsilon_2 > 0$. Тогда, если $\Gamma_{\bar{\mu}}(s) \geqslant \varepsilon_1$ и $\Gamma_{\bar{\mu}}(s) - \max_{\mu \neq \bar{\mu}} \Gamma_\mu(s) \geqslant \varepsilon_2$, то пробе s отвечает эталон $s^{\bar{\mu}}$. В противном случае s не имеет аналога среди эталонов. Здесь числа ε_1, ε_2 — пороги, определяющие «решительность» алгоритма распознавания. В нашем случае ε_1 принималось равным 0,5, а ε_2 — 0,01.

Заключение

Изложенные математические средства составляют процедурное и программное обеспечение схемы последовательного распознавания, применявшейся для решения задачи сравнительного изучения определенной целевой совокупности нефтяных месторождений. Приведенные алгоритмы распознавания и упорядочения опробованы на широком круге геологических примеров, связанных с задачами прогнозно-поискового профиля. Авторы стремились описать эти алгоритмы таким образом, чтобы их можно было изъять из контекста указанной выше задачи сравнительного изучения нефтяных месторождений и применять для решения других геологических задач, сводимых к проблеме распознавания образов.

ЛИТЕРАТУРА

1. Бабич В. В. Программа ПЗ «Подсчет строчечных нагрузок по заданной информативности для бинарных таблиц», В кн. Логико-математич. обработка геологич. информации. Новосибирск, 1975.
2. Бабич В. В., Красавчиков В. О., Федосеев Г. С. Программы метода суммарного учета мер приуроченности и согласования. В кн. Логико-математич. обработка геологич. информации. Новосибирск, 1975.
3. Васильев Ю. Л., Дмитриев А. Н. Спектральный подход к сравнению объектов, охарактеризованных набором признаков. ДАН СССР, 1972, т. 206, № 6.

4. Волков П. П., Дмитриев А. Н. и др. Дифференциальная диагностика заболеваний головного мозга логико-дискретным методом. В кн. Пробл. моделир. психич. деятельности. Новосибирск, изд. фил. ЦИТП, 1968, вып. 2.
5. Дмитриев А. Н., Журавлев Ю. И., Кренделев Ф. П. О математических принципах классификации предметов и явлений. В кн. Дискретный анализ. Новосибирск, 1966, вып. 7.
6. Дмитриев А. Н., Красавчиков В. О. Программы метода согласованных оценок. В кн. Логико-математич. обработка геологич. информации. Новосибирск, 1975.
7. Дмитриев А. Н., Смертин Е. А. и др. Программы «тестовых оценок» (тесты, тесторы). В кн. Логико-математич. обработка геологич. информации. Новосибирск, 1975.
8. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М., 1973.
9. Кандыба В. Н. Программа П2 «Расчет коэффициентов». В кн. Логико-математич. обработка геологич. информации. Новосибирск, 1975.
10. Красавчиков В. О. Модификация тестового подхода к анализу таблиц описаний на основе понятия пакета. В кн. Дискретный анализ. Новосибирск, 1974, вып. 26.
11. Мацак А. П. Проблемы обучения по малым выборкам при геологическом прогнозировании. Изв. АН КазССР, серия физ.-мат., 1969, № 5.
12. Слуцкая Т. Л. Программа П5 «Диагностика объектов голосованием по тестерам». В кн. Логико-математич. обработка геологич. информации. Новосибирск, 1975.
13. Смертин Е. А., Дмитриев А. Н. Дополнение к алгоритму распознавания голосованием по тестам и тесторам. В кн. Алгоритмы и программы решения геологич. задач на ЭЦВМ «Минск-2» и «БЭСМ-3М». Алма-Ата, изд. КазИМС, 1970, вып. 3.
14. Соколов А. Д. Программа П5 «Оптимальное бинарное кодирование признаков, подсчет строчечных нагрузок и минимизация». В кн. Логико-математич. обработка геологич. информации. Новосибирск, 1975.
15. Трофимук А. А., Васильев Ю. Л., Вышемирский В. С., Дмитриев А. Н. Сравнительное изучение гигантских месторождений нефти спектральным методом. В кн. Применение математич. методов и ЭВМ для решения прогнозных задач нефтяной геологии. Новосибирск, изд. СНИИГГиМС, 1973.
16. Трофимук А. А., Вышемирский В. С., Дмитриев А. Н. и др. Распознавание образов гигантских нефтяных месторождений. В кн. Пробл. нефтеносности Сибири. Новосибирск, «Наука», 1971.

ИГиГ СО АН СССР
Новосибирск

Поступила в редакцию
4 марта 1976 г.

A. N. Dmitriev, V. O. Krasavchikov

MATHEMATICAL PROCESSING APPLIED TO DESCRIPTION OF THE OIL FIELDS

Mathematical processing applied to description of the oil fields associated with hydrocarbon pools is reported. The purpose of processing is solution of problems of prospecting and search of oil pools. The described algorithms are the elements of the system of successive pattern recognition. The methods and their treatment are reported in view of their application also to ore deposit prospecting.