АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
ГЕОЛОГИЯ И ГЕОФИЗИКА

№ 5 1968

А. Н. ДМИТРИЕВ,
Ю. И. ЖУРАВЛЕВ, Ф. П. КРЕНДЕЛЕВ

ОБ ОДНОМ ПРИНЦИПЕ КЛАССИФИКАЦИИ И ПРОГНОЗА ГЕОЛОГИЧЕСКИХ ОБЪЕКТОВ И ЯВЛЕНИЙ

Геологический поиск сводится, по существу, к оценке меры аналогии исследуемого района с ранее изученным, причем в большинстве случаев такое сравнение ведется по совокупности признаков, важность каждого из которых неизвестна. В геологической практике такая оценка во многом зависит от интуиции и опыта исследователя и в ряде случаев приводит к положительным результатам.

Ниже дается попытка построения системы правил, позволяющих: а) отобрать и классифицировать сведения; б) выявить степень существенности признаков, упорядочить их по степени существенности и сформировать поисковый критерий; в) построить классификацию ранее изученных месторождений по комплексу признаков; г) найти в построенной классификации место изучаемого района и оценить перспективность этого района, т. е. найти меру аналогии с изученными районами.

Работа иллюстрируется примерами по изучению золотоносных месторождений типа докембрийских конгломератов и никеленосных интрузий северо-запада Сибирской платформы.

Авторы выражают признательность А. А. Трофимуку и А. А. Ляпунову за внимание к данной работе.

Правила отбора информации

Формирование цели. Первым этапом обработки информации, по-видимому, является выбор и формулирование цели обработки [1]. Например, целью может быть: поиск любых месторождений золота. Такая цель может быть уточнена указанием масштаба месторождения, генетическим типом месторождений и т. д. Примером уточненной цели является следующая: поиск месторождений золота с запасом свыше 1000 *т* докембрийского возраста в обломочных породах.

Второй пример уточнения цели: исследование трапповых интрузий → исследование никеленосных трапповых интрузий → исследование никеленосных трапповых интрузий северо-запада Сибирской платформы.

Отбор эталонов. На втором этапе отбираются ранее исследованные объекты, для которых выполнены требования цели. Например, цели — сравнительное изучение комплексных древних месторождений золота в обломочных породах с запасами золотых руд свыше 100—1000 *т* соответствуют семь зарубежных объектов (эталонов): Витватерсранд

(ЮАР), Блайнд-Ривер (Канада), Жакобина (Бразилия), Мунана (Габон), Тарква (Гана), Австралия, Эно-Коли (Финляндия).

Для неуточненной цели — поиск любых промышленных месторождений золота, число эталонных объектов возросло бы, по крайней мере, в сотни раз, а суммарное количество признаков практически необозримо. Задача такого объема вряд ли может быть решена даже с помощью современных ЭВМ. Данное обстоятельство указывает на необходимость максимального уточнения цели и внимательного отбора эталонных объектов.

Отбор признаков. а) Мобилизуется вся информация по эталонным объектам, сконцентрированная в описаниях этих объектов. На основе этой информации конструируется по возможности полная система признаков. Каждый из признаков должен быть строго выполнен, строго невыполнен или неизвестен для каждого эталонного объекта. Если существует неопределенность в утверждении «признак выполнен» («признак невыполнен»), следует считать признак неизвестным для данного объекта. При отборе информации следует избегать предпочтения отдельных гипотез и крайне осторожно отбрасывать якобы несущественную информацию, поскольку в задачу входит определение степени ее существенности. Вся информация должна отображать лишь наблюденные факты и величины, но не выводы или мнения.

б) Различными видами геологической деятельности могут выявляться разные группы признаков. Обычно выделяются четыре основных вида деятельности. В соответствии с этим, получаемые признаки также можно разбить на четыре группы, а именно: получаемые при геологическом картировании (группа A_1); при поиске (A_2); при разведке (A_3); и при эксплуатации (A_4).

Очевидно, количество видов деятельности и, соответственно, групп признаков, может быть детализировано или сокращено.

в) Как правило, хорошо изученные эталоны имеют сведения о выполнимости признаков всех групп. Для исследуемого района это может не выполняться. Так как в основе дальнейших рассмотрений будет лежать определение меры близости эталонов и исследуемого района, в перечне признаков эталонных объектов следует оставить лишь те, которые соответствуют видам деятельности, приведенным в исследуемом районе.

В задачах геологического поиска такими видами чаще всего являются геокартирование и поиск.

Построение таблицы. Вся имеющаяся признаковая информация об эталонах (кроме признаков, характеризующих цель) сводится в таблицу. Строкам таблицы соответствуют эталоны, столбцам — признаки.

Таблица 1

Эталоны \ Признаки	x_1	x_2	...	x_j	...	x_n
M_1	1	0	.	—		0
M_2	0	1	.	1	.	0
⋮	.	.	.	.	.	.
M_i	0	1	.	.	.	1
⋮	.	.	.	.	.	.
M_n	1	—	.	0	.	1

Если признак j выполнен, невыполнен или неизвестен для i-го эталона, на пересечении i-й строки и j-го столбца таблицы ставится 1, 0, или прочерк (—) соответственно (табл. 1). Построенную таблицу будем называть в дальнейшем основной.

Следует отметить, что для удобства работы: а) желательно формировать список признаков для каждого эталона отдельно и лишь затем сводить полученные признаки в одну таблицу; б) операцию удаления

Таблица 2

Основная кодовая таблица по никеленосным трапповым интрузивам северо-запада Сибирской платформы

№ эталона	Группа	Объекты	Признаки																																Сумма номеров занятых мест по таблице	Общий порядок объектов по значимости
			1	4	6	7	13	14	18	20	21	23	24	25	26	27	28	2	5	9	10	11	12	15	16	17	19	22	29	30	31	3	8		
1	Первая	Норильск I	1	0	1	0	1	1	1	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	14	3
2		Талнах	1	0	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	4	1
3		Норильск II	1	1	1	0	1	0	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	8	2
4		г. Черная	1	0	1	0	0	1	0	1	1	0	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	18	4
5		Имангда	1	0	1	0	1	0	1	0	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	23	6
6		Накохоз	0	1	0	1	0	0	0	1	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	21	5
7		г. Зуб	1	1	1	1	0	0	0	0	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	24	7
8	Вторая	Хюкта	1	1	0	1	0	0	1	0	1	0	0	1	0	0	1	1	0	1	0	1	1	1	0	1	0	0	1	1	1	1	1	41	8
9		Второй порог	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	1	0	1	1	1	1	1	1	0	0	0	1	1	1	55	15
10		г. Озерная	1	1	0	0	1	0	1	0	0	1	1	0	0	1	0	0	1	1	0	1	1	1	0	1	0	0	0	0	1	1	1	45	9
11		г. Подкова	0	1	0	0	0	0	1	0	0	1	1	0	0	1	0	0	1	1	0	1	1	1	0	1	1	0	0	0	1	1	1	52	13
12		Рудничный камень	0	1	1	1	0	0	0	0	0	1	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0	0	0	1	1	1	50	11
13		Нижняя I	1	1	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	1	1	1	0	1	1	0	0	0	1	1	0	53	14
14		Нижняя II	1	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	1	1	1	0	1	1	0	0	1	1	1	1	58	17
15		Кулюмбинская	0	1	0	1	0	1	1	0	1	0	1	0	0	1	1	0	1	1	0	1	1	1	0	0	0	0	0	0	1	1	1	48	10
16		Галельская	0	1	0	1	0	0	0	0	1	1	1	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	1	1	1	63	18
17		Моронга	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	1	1	1	0	1	1	1	1	0	1	0	1	0	1	0	0	51	12
18		Н. Фокинская	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0	1	0	1	1	1	0	0	56	16

признаков, соответствующих видам деятельности, не проведенным для исследуемого объекта, обычно удобнее выполнять после построения основной таблицы.

Понятие формального типа

В основной таблице выделяем все признаки, которые выполнены или невыполнены одновременно для всех эталонов (строк). Иными словами, столбцы, которые для всех строк имеют значение единица или нуль, называются отождествляющими. Остальные признаки будем называть различающими [1, 2, 4]. Например, для табл. 2, построенной по никеленосным трапповым интрузиям Северо-Запада Сибирской платформы, признаки X_{11}, X_{15}, X_{31} являются отождествляющими [3].

Остальные 28 признаков являются различающими.

Содержательный смысл признаков кодовой таблицы приведен в табл. 3.

Таблица 3

СОДЕРЖАТЕЛЬНЫЙ СМЫСЛ КОДОВЫХ ПРИЗНАКОВ

I. Петрохимические признаки

X_1 — магнезиальность исходного расплава: $0>9\%$ $MgO>1$*.
X_2 — магнезиальность пикритового горизонта: $0>16{,}5\%$ $MgO>1$.
X_7 — содержание щелочей в исходном расплаве: $0>2{,}9\%$ $(Na_2O+K_2O)>1$.

X_8 — содержание щелочей в нижних горизонтх: $0>1{,}9\%$ $(Na_2O+K_2O)>1$.

X_{16} — степень дифференциации: $0>21$ К.Ф.>1

$$\left[\text{КФ} = \frac{Fe_2O_3 + FeO}{Fe_2O_3 + FeMgO} \cdot 100\right]$$

II. Минералогические признаки

X_3 — магнезиальность оливином нижних горизонтов (пикриты и др.): $0>25\%$ $Fa>1$.

X_4 — магнезиальность оливинов верхних горизонтов: $0>35\%$ $Fa>1$.
X_5 — наличие ортопироксенов в нижних дифференциатах интрузий: 1 — да, 0 — нет.

X_6 — основность плагиоклазов нижних горизонтов: $0>76\%$ $Au>1$.
X_9 — наличие биотита: $0>(\sim 0{,}1\%)>1$.
X_{10} — наличие хромита: 1 — да, 0 — нет.
X_{11} — наличие неправильных выделений магнетита и титаномагнетита: 1 — да, 0 — нет.
X_{12} — наличие вкрапленности сульфидов: $0>(\sim 1\%)>1$.
X_{13} — наличие вкрапленности сульфидных оруденений в верхних горизонтах: 1 — да, 0 — нет.
X_{14} — наличие жильного оруденения: 1 — да, 0 — нет.

III. Строение интрузий

X_{15} — наличие дифференцированности: 1 — да, 0 — нет.
X_{17} — » нижних такситовых габбро-долеритов: 1 — да, 0 — нет.
X_{18} — » верхних такситовых габбро-долеритов: 1 — да, 0 — нет.
X_{19} — » пикритовых габбро-долеритов: 1 — да, 0 — нет.
X_{20} — » кислых гибридных пород в крыле интрузива: 1 — да, 0 — нет.
X_{22} — форма интрузии: 1 — хоналит, 0 — нет.

---

* Для признаков X_{1-4}, X_{6-8} цифровые данные получены делением сумм средневзвешенных значений для каждого объекта. Для X_{16} число 21 К. Ф. получено от деления суммированных разностей крайних значений К. Ф. для каждого объекта на число объектов.

IV. Геологическая позиция и тектоника

X_{21} — интенсивность метаморфизма пород в экзо- и эндоконтактах: 1 — велика, 0 — нет.
X_{29} — присутствие дифференцированных лав во вмещающих толщах: 1 — есть, 0 — нет.
X_{30} — наличие ангидридов, гипсовых слоев в подстилающих толщах: 1 — есть, 0 — нет.
X_{31} — связь интрузий с зонами глубинных разломов: 1 — да, 0 — нет.

Характер пород кровли

X_{23} — лавы, туфы, цитрузивные породы: 1 — есть, 0 — нет.
X_{24} — наличие песчано-сланцевой толщи: 1 — есть, 0 — нет.
X_{25} — » карбонатно-глинистых пород и мергелей: 1 — есть, 0 — нет.

Характер пород подошвы

X_{26} — лавы, туфы, интрузии: 1 — есть, 0 — нет.
X_{27} — наличие грубообломочного материала, песчаников, сланцев: 1 — есть, 0 — нет.
X_{28} — » карбоната, мергелей, глин: 1 — есть, 0 — нет.

Выделим в основной таблице все отождествляющие признаки. Совокупность объектов, для которых все отождествляющие признаки, выполненные на эталонах, хотя бы на 20%, выполнены (или невыполнены) и для исследуемых объектов, назовем формальным типом по данной группе признаков.

В дальнейшем мы будем вести обработку информации по исследуемому району только в том случае, если он принадлежит к тому же типу, что и эталоны.

После формирования типа все отождествляющие признаки, (т. е. столбцы, сплошь заполненные единицами или нулями) из таблицы удаляются.

В ряде случаев понятие формального типа совпадает с понятием совокупности объектов, объединенных по генетической классификации.

Заметим, что образование формального типа производится на основе группы признаков, что является сходным с идеей классификации геологических объектов по комплексному основанию. Классификации с многопризнаковыми основаниями изучались, например, в работах [2, 6].

По-видимому, чем более геологически важные признаки содержатся в совокупности отождествляющих, тем больше аналогия исследуемых объектов в формальном типе.

Информационный вес признака

Таблицу, которая получается из основной, если удалить все столбцы, соответствующие отождествляющим признакам, назовем допустимой.

Введем классификацию различающих признаков по величине, которую в дальнейшем будем называть информационным весом признака.

Рассмотрим алфавит из трех символов: 0, 1, «—». Будем говорить, что символы 0 и 1 являются различными, символы 0 и «—», 1 и «—» — совпадающими.

Две строки $S_1 = (\alpha_1 \alpha_2, \ldots, \alpha_n), = (\beta_1 \beta_2, \ldots, \beta_n)$, составленные из символов алфавита {0, 1, —} назовем различными, если существует хотя бы один номер i, $1 \leqslant l \leqslant n$, такой, что символы α_i и β_i различны.

Строки, не являющиеся различными, назовем совпадающими.

Так, строки (0—0—1) и (01101) являются различными, так как третьи по порядку символы разнятся между собой. Строки (0—0—1) и (00011) являются совпадающими.

Определение 1. Набор столбцов допустимой таблицы T с номерами $(i_1, i_2, \ldots, i_e)$ называется тестом таблицы T, если после удаления из T всех столбцов, за исключением столбцов с номерами $i_1, i_2, \ldots, i_e$ получается таблица, все строки которой различны.

Пример:

1	2	3	4
0	0	0	—
0	—	1	1
1	1	0	1

Таблица T

5	1	3	5
1	0	0	1
0	0	1	0
0	1	0	0

Таблица T′

В таблице T столбцы 1, 3, 5 образуют тест. Действительно, после вычеркивания из T второго и четвертого столбцов, получается таблица T′, все строки которой различны.

Определение 2. Тест, составленный из столбцов с номерами $i_1, i_2, \ldots, i_e$, называется тупиковым, если из него нельзя удалить ни одного столбца без того, чтобы он перестал быть тестом

Смысл определения 2 состоит в том, что если столбцы $i_1, i_2, \ldots, i_e$ образуют тупиковый тест, то удаление из таблицы, составленной из этих столбцов, любого столбца приводит к появлению хотя бы двух совпадающих строк.

Пример. Таблица T′ не образует тупиковый тест таблицы T. T составлена из столбцов с номерами 1, 3, 5. Удаляя из T′ соответственно столбцы 1, 3, 5, получим T_1', T_3', T_5':

0	1	0
1	0	0
0	0	1

Таблица T_1'

1	0	0
0	0	1
0	1	0

Таблица T_3'

0	0
0	1
1	0

Таблица T_5'

В таблицах T_1', T_3', T_5' все строки различны. Каждый из наборов (1,3), (1,5), (3,5) столбцов образует тупиковый тест таблицы T.

Действительно, если, например, удалить из T_1' любой столбец, то в оставшемся получим две совпадающие строки.

Заметим, что число различных тупиковых тестов (минимальных кодов таблицы) может быть весьма велико. Алгоритмы построения всех тупиковых тестов изложены в [2].

Пусть $T_1, T_2, \ldots, T_k$ — тупиковые тесты допустимой таблицы T. Возьмем столбец с номером i, соответствующий различающему признаку, и выделим все тупиковые тесты, в которые входит этот столбец. Пусть такими тестами являются $T_1^i, T_2^i, \ldots, T_{k(i)}^i$.

Положим $m(i) = \dfrac{k(i)}{k}$, $N = \sum\limits_{i=1}^{n} \dfrac{k(i)}{k}$.

Определение 3. Величину $P(i)$, определяемую равенством $P_i = \dfrac{k(i)}{k \cdot N}$, назовем информационным весом i-го признака.

Величина $P(i)$ с точностью до нормирующего множителя N есть отношение числа тупиковых тестов, в которые входит столбец, соответствующий i-му признаку, к общему числу тупиковых тестов исходной допустимой таблицы.

Пример:

1	2	3	4	5
1	0	0	1	1
1	1	0	1	1
1	0	1	1	0
1	1	1	0	0
0	1	1	0	1
0	1	1	1	0

Таблица T_2

Приведенная здесь таблица имеет три тупиковых теста: $\langle 1, 2, 5 \rangle$, $\langle 2, 4, 5 \rangle$, $\langle 1, 2, 3, 4 \rangle$ и соответственно $P(1) = \frac{1 \cdot 3}{2 \cdot 7} = \frac{3}{14}$, $P_2 = \frac{3}{7}$, $P_3 = \frac{1}{7}$, $P_4 = \frac{3}{14}$, $P_5 = \frac{3}{14}$, $N = \frac{7}{3}$.

Нормирующий множитель $\frac{1}{N}$ введен здесь для удобства вычислений. Поэтому за информационный вес можно принять число $P(i)$, где $P(i) = \frac{k(i)}{k}$.

Содержание понятия «информационный вес» может быть пояснено следующим образом. Описание выбранной совокупности эталонов всеми исходными различающими признаками обычно является избыточным. После удаления некоторых признаков описание сохраняет основное свойство — оно различает все объекты (эталоны) из заданного класса. Последовательно удаляя столбцы, мы приходим к несжимаемому (абсолютно неизбыточному описанию), которое при дальнейшем сжатии теряет свойство различения объектов данного класса. Такие несжимаемые описания (тупиковые тесты) являются как бы корнями, основами остальных описаний.

Естественно считать, что чем в большее число таких основных (корневых) описаний входит признак, тем он существеннее при описании данного класса объектов.

Мы приходим таким образом к формулировке принципа классификации признаков.

Из двух различающих признаков, входящих в допустимую таблицу, более существенным для характеристик данного класса объектов (месторождений, интрузивов, геологических районов и т. п.) является тот, для которого информационный вес больше.

Данный принцип, по-видимому, имеет широкую область применений, но вряд ли является всеобъемлющим. Он был проверен на задаче классификации признаков, описывающих комплексные месторождения урана и золота в древних конгломератах [4, 5], трапповые интрузии северо-запада Сибирской платформы [3], и некоторых других задач [4].

В этих задачах сформированный выше принцип оказался правомерным. По-видимому, при построении других классификаций следует предварительно проверить корректность принципа на совокупности хорошо изученных эталонов.

Пример. При обработке таблицы, описывающей зарубежные комплексные месторождения золота в древних конгломератах [4, 5] оказалось, что наибольший информационный вес имеет признак X_9 («наличие конгломератов в основании продуктивной толщи»), а наименьший информационный вес — признаки X_{10} и X_{11} (соответственно трансгрессивная и регрессивная серия осадков, участвующих в строении рудовмещающей толщи), что вполне соответствует общепринятым геологическим представлениям.

Для ряда признаков их место в классификации по информационным весам оказалось неожиданным, но в дальнейшем было подтверждено дополнительными исследованиями. Так, например, высоким информационным весом обладает группа признаков, описывающих магматические проявления в подстилающей толще. Иными словами, оказывается важным, на чем залегают комплексные месторождения типа древних конгломератов. Это кажется неожиданным с точки зрения оса-

дочного генезиса оруденения этих месторождений. При дальнейших исследованиях геологическая существенность этой группы признаков нашла свое объяснение (см. [4, 5]).

Положим информационный вес каждого из отождествляющих признаков равным единице. В таблице встречаются такие признаки, столбцы которых заполнены только единицами и прочерками или нулями и прочерками (частично-отождествляющие признаки).

Пусть такой признак x_i из общего числа m эталонов для l эталонов неизвестен (закодирован прочерком), а для всех остальных $m-l$ эталонов выполнен (или невыполнен). Тогда информационным весом признака x_i назовем величину $P_{(i)}$, где $P_{(i)} = \frac{m-l}{m}$.

Информационные веса отождествляющих признаков и частично-отождествляющих признаков вводятся для удобства некоторых вычислений. Они не имеют того содержательного смысла, который присущ информационным весам различающих признаков.

Формирование поискового критерия

Пусть вычислены информационные веса всех различающих признаков и сами признаки $x_1, x_2, \ldots, x_{n-1}, x_n$ упорядочены по убыванию информационного веса, т. е. $P_1 \geqslant P_2 \geqslant \ldots \geqslant P_{n-1} \geqslant P_n$.

При исследовании конкретных таблиц Т, описывающих классы различных эталонов (геологических объектов), замечено следующее.

По значениям $P(i)$ признаки распадаются на группы (рис. 1), причем колебание величины $P(i)$ при переходе к соседней группе (рангу) значительно больше, чем колебание между соседними признаками в одной группе.

В различных задачах колебание информационного веса между смежными группами выражено с разной степенью четкости, но всегда сохраняется качественная картина типа, изображенная на рис. 1. Число групп также широко варьирует.

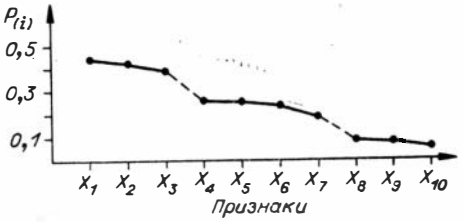

Рис. 1. Принципиальная схема упорядочивания признаков информационным весом $P(i)$ (без нормирования).

В дальнейшем мы будем упорядочивать группы по убыванию информационного веса: признаки первой группы имеют наибольший информационный вес, признаки второй группы имеют вес, меньший, чем признаки первой группы, но больший, чем признаки третьей группы, и т. д.

Назовем поисковым критерием l-го ранга совокупность всех признаков, входящих в группы с номерами 1, 1, . . . , l.

Наибольший практический интерес представляют критерии первого и второго ранга, составленные из признаков первой и второй группы.

При реальном поиске на объектах, для которых l значения отождествляющих признаков совпадают с эталонными значениями, следует, по-видимому, в первую очередь изучить те объекты, для которых выполнены (максимально сближены в пространстве) признаки первой и второй группы. Если эти признаки для изучаемого объекта не выявлены, следует провести работу по определению значений этих признаков.

Пример. После того, как была проведена классификация признаков, описывающих зарубежные комплексные месторождения золота в

древних конгломератах, было замечено, что в одном из районов развития докембрия СССР все признаки первой группы, кроме одного, X_9 (см. стр. 56), выполнены, а этот признак не был исследован. Полевыми исследованиями в районах, где выполнялись остальные признаки, недостающий признак был обнаружен. Одновременно в выбранных районах были найдены сульфидные конгломераты докембрийского возраста.

Информационный вес строки

Пусть задана строка S (объект, эталон), составленная из n символов, принимающих значения 0, 1, «—». Пусть также информационные веса признаков равны соответственно $P_1, P_2, \ldots, P_n$. Выделим в строке S все места $i_1, i_2, \ldots, i_e$, на которых стоят единицы, и все места $j_1, j_2, \ldots, j_r$, на которых стоят нули.

Определение 4. Величина $I(S) = \sum_{k=1}^{l} P(i_k) + \frac{1}{2} \sum_{k=1}^{r} P(j_k)$ называется информационным весом строки S.

Пусть в основной таблице в j столбце из общего числа m символов, n_1^j раз встречается прочерк, n_2^j раз — 0, n_3^j раз — 1. Обозначим через $q(j)$ величину — $\frac{n_3^j}{n_2^j + n_3^j}$.

Определение 4′. Величина $\overline{I(S)} = \sum_{k=1}^{l} P(i_k) + \sum_{k=1}^{r} q(j_k) \cdot P(j_k)$ называется информативностью строки S [1].

Так как каждая строка S соответствует некоторому объекту, величины $I(S)$ и $\overline{I(S)}$ мы будем называть также информационным весом, или информативностью объекта, который характеризует строка S.

На ряде примеров установлено, что информационный вес, или информативность, могут служить хорошей основой для построения классификации объектов. Оказывается также, что на основе исследования данных величин в ряде случаев может быть сделан геологический прогноз, основанный на формальном вычислении близости прогнозируемого объекта к ранее изученным.

Пример 1. При сравнительном изучении комплексных месторождений золота в древних конгломератах [2, 4] упорядочение эталонных месторождений по убыванию информационного веса (информативности) был получен следующий график (рис. 2).

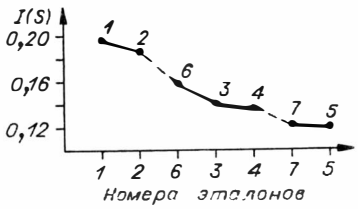

Рис. 2. Упорядочивание эталонных месторождений типа докембрийских металлоносных конгломератов по информационному весу строк:

1 — Витватерсранд; *2* — Блайнд-Ривер, *6* — Австралия, *3* — Жакобина, *4* — Мунана, *7* — Эно-Коли (Финляндия), *5* — Тарква. Для удобства значение информационного веса (по ординате) взято с нормирующим множителем.

Как видно из графика, наибольший информационный вес получили месторождения с максимальными запасами, содержательно классифицируемые, как месторождения сульфидного типа [5]. Минимальным информационным весом обладают месторождения с небольшими запасами сырья, относящиеся к магнетит-гематитовому типу. Таким образом, формальная классификация месторождений по величине их информационного веса совпадает с содержательной классификацией по запасам сырья и типам. Следует отметить, что трехэлементным

месторождениям соответствует больший информационный вес, двуэлементным и одноэлементным — меньший вес. Мы видим, что информационный вес формально разделил месторождения с разным качественным составом руд.

По тем же правилам была проведена оценка возможностей пяти районов развития докембрия в СССР. Эти районы характеризовались теми же признаками, что и эталонные месторождения. Ясно, что в сравниваемых районах отсутствовали целевые признаки. Упорядочивание сравниваемых районов (номера от 8 до 12) с эталонами по $I(S)$ показано на графике (рис. 3). Вполне очевидно, что наиболее перспективным оказывается район, обозначенный на графике восьмым номером, поскольку именно этот район получил информационный вес, близкий к таковому для Витватерсранда.

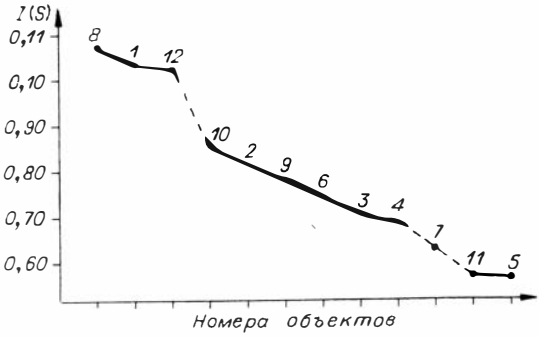

Рис. 3. Упорядочивание эталонных и сравниваемых районов по I (S). Номера эталонных месторождений см. на рис. 2, сравниваемых районов — в тексте (с нормирующим множителем).

Проверка метода заключалась в том, чтобы в заданном районе, в данном случае в восьмом, найти рудопроявления изучаемого типа. При полевых исследованиях в этом районе были найдены рудопроявления изучаемого типа, что наилучшим образом подтверждает правильность предлагаемого метода классификаций и выражения меры аналогии по совокупности признаков.

Пример 2. Как известно из литературы, посвященной рудоносным трапповым интрузиям Сибирской платформы [3], наиболее неясным моментом является связь сульфидного оруденения с той или иной трапповой интрузией. Эмпирически давно установлена связь сульфидного оруденения с дифференцированными трапповыми интрузиями.

Впоследствии эта связь была уточнена. Интрузии с сульфидным оруденением на Сибирской платформе дифференцированы, как правило, по норильскому типу. Но масштабы оруденения могут сильно колебаться внутри группы интрузий этого типа. Более того, на основании геологопоисковых работ стало ясно, что не все дифференцированные интрузии этого типа рудоносны. Выяснение причин разномасштабности вплоть до полного отсутствия оруденения в однотипных интрузиях представляет большой теоретический и практический интерес.

В соответствии с правилами, сформулированными в первом разделе, была задана общая таблица объемом (18×31) (см. табл. 2). В таблице отсутствовали названия эталонов и признаков. Эти названия были раскрыты составителями (В. В. Золотухиным и Ю. И. Васильевым) после проведения формального решения.

На первом этапе решения требовалось: 1) установить, разложимы ли эталоны на несколько типов или это объекты одного типа; 2) указать наиболее важные признаки, описывающие данные объекты; 3) наметить комплексы наиболее важных признаков, объединенные общностью геологического содержания.

В процессе решения вычислялись информационные веса признаков и эталонов, после чего эталоны были упорядочены по убыванию ин-

формационного веса. Оказалось, что эталоны разбились на две четкие группы (рис. 4). Первая группа состоит из семи объектов, вторая — из 11. Эталоны первой группы получили информационные веса, значительно большие, чем эталоны второй группы. После указания названий эталонов и проверки геологической существенности каждого из них оказалось, что в первую группу вошли все наиболее важные объекты, во вторую — менее важные. Среди первой группы эталон № 2 набрал наибольший информационный вес и выступает как наиболее существенный с формальной точки зрения объект. И действительно, интрузия Талнах является наиболее существенной и в геологическом смысле.

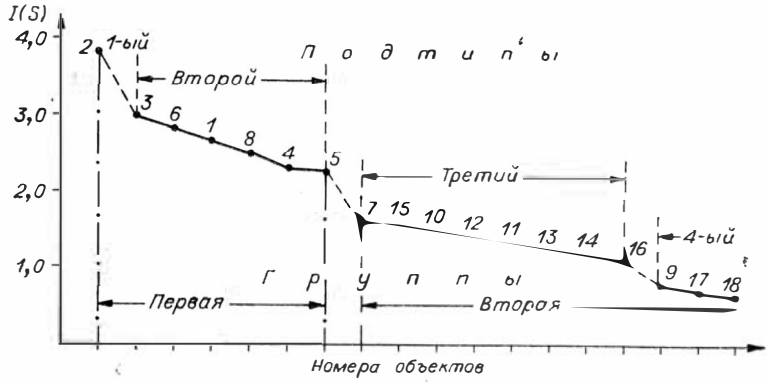

Рис. 4. Упорядочивание никеленосных трапповых интрузий северо-запада Сибирской платформы по информационным весам строк. Номера эталонов см. табл. 2 (без нормирования).

Если провести более полный анализ графика (рис. 4), то можно сказать, что представленные эталоны разделились на 4 подтипа (по два в каждой группе). Первый подтип представлен одним объектом 2. Этот подтип можно назвать наиболее существенным. Второй подтип представлен шестью объектами — 3, 6, 1, 8, 4, 5, третий восемью — 7, 15, 10, 12, 11, 13, 14, 16, четвертый тремя эталонами — 9, 17, 18. Подробное разбиение полностью соответствует общим геологическим представлениям об исследуемом типе, но было выполнено формально.

Признаки с наибольшим информационным весом оказались для типа в целом (18 эталонов) X_{28}, X_{20}, X_{14} (см. табл. 3).

В дальнейшем рассматривались две подтаблицы (верхняя и нижняя части табл. 2), образованные объектами первой (эталоны 2, 3, 6, 1, 8, 4, 5) и второй групп (эталоны 7, 15, 10, 12, 11, 13, 14, 16, 9, 17, 18). Для каждой из групп было проведено упорядочение эталонов по убыванию информационного веса. При этом оказалось, что порядок, построенный по общей таблице, в основном совпадает с порядками, построенными по подтаблицам для отдельных групп.

Было также проведено следующее: с помощью информационных весов признаков, вычисленных по общей таблице для групп, найдены информационные веса объектов второй группы и по этим весам упорядочены эталоны этой группы. Затем первая и вторая группы поменялись местами. Оказалось, что порядки, построенные таким образом, в основном совпали с порядками, построенными по общей таблице. В первой группе не произошло никаких существенных перестановок, во второй — поменялись местами последние по значимости объекты (см. табл. 2).

Соответствующие вычисления проводились и на смешанных таблицах, составленных из представителей (но не всех) первого и второго типов. При всех таких перестановках сохранился на 90% порядок, построенный по общей таблице, причем перемещения в основном происходили среди объектов второй группы.

Для общей таблицы и для каждой из подтаблиц были найдены информационные веса признаков, список которых приведен в табл. 3. При этом оказалось, что признаки, имевшие большой информационный вес на подтаблице первой группы, получили малый информационный вес на подтаблице второй группы и наоборот. Таким образом, были обнаружены комплекты признаков, существенные для каждой группы объектов в отдельности.

После выполнения всех расчетов оказалось, что 28 признаков, включенных в общую таблицу (табл. 2), разделились на три комплекта. Первый комплект — признаки, существенные для объединения типов (X_{23}, X_{18}, X_7, X_{24}, X_1, X_{29}, X_{30}, X_{26}, X_9, X_{12}, X_{17}, X_{22}, X_6, X_8); второй комплект — признаки, существенные для первой и несущественные для второй группы (X_{14}, X_{20}, X_{28}, X_{25}, X_4, X_{13}, X_{21}); третий комплект — признаки, существенные для второй и несущественные для первой группы (X_{27}, X_{19}, X_3, X_{16}, X_2, X_5, X_{10}).

Вся совокупность признаков (табл. 3) распалась по содержанию на 4 группы:

а) петрохимическую (X_1, X_2, X_7, X_8, X_{16}); б) минералогическую (X_4, X_3, X_5, X_6, X_{9-14}); в) группу признаков X_{15}, X_{17-20}, X_{22} (ими задано строение интрузий; г) геолого-тектоническую (X_{21}, X_{23}, X_{24}, X_{25} — X_{31}).

После выполнения всех вычислений оказалось возможным упорядочить данные группы по информационным весам. При этом оказалось, что группы упорядочиваются по убыванию веса таким образом: строение интрузий — петрохимия — минералогия, геологическая обстановка — породы кровли. Следует отметить, что точность упорядочения содержательных групп признаков зависит от точности содержательной классификации признаков по группам.

Сравнительное изучение интрузий северо-запада Сибирской платформы изложенным методом показывает, что формирование интрузий первой группы обусловлено многими факторами, особенно теми, которые отображают строение самих интрузий. Исходя из этого, рассмотрение генезиса оруденений невозможно оторвать от генезиса такситовых горизонтов, наиболее обогащенных сульфидами [3].

По-видимому, для того чтобы в сульфидах появился никель и, следовательно, интрузия обрела значительность, она должна содержать горизонт пикритовых габбро-долеритов, с условием, что в них будет не меньше 16,5% MgO, а железистость оливина не будет превышать 25% Fa компонента.

Некоторые алгоритмы распознавания образов

Используя введенное в работе понятие информационного веса признака и ряда понятий, которые будут приведены ниже, можно построить серию алгоритмов для задачи распознавания образов. Очевидно, эта задача имеет прямое отношение к задаче прогноза месторождений полезных ископаемых.

Рассмотрим ее в следующей постановке.

Даны две таблицы, составленные из нулей и единиц, причем строки первой таблицы не встречаются во второй и наоборот. Строки первой

таблицы T_1 мы будем называть эталонами первого класса, строки второй таблицы T_2 — эталонами второго класса.

Рассматривается произвольная строка той же длины n, что и строки таблиц.

Требуется отнести эту строку к одному из двух классов эталонов.

Объединим таблицы T_1 и T_2 в одну таблицу T.

Определение 5 [2]. Набор $(i_1, \ldots, i_t)$ столбцов таблицы T назовем тестором для T_1, T_2, если после удаления из T всех столбцов, не вошедших в совокупность $(i_1, \ldots, i_t)$, все строки из T_1 будут различаться со всеми строками из T_2.

Определение 6. Тестор для T_1, T_2 называется тупиковым, если после удаления из него любого столбца он перестает быть тестором для T_1, T_2.

Пример:

Пусть

$$T = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \quad T_1 = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \end{array} \quad T_2 = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}.$$

Набор столбцов с номерами 2, 3 является тупиковым тестором для T_1, T_2. Действительно, после удаления из T столбцов с номерами 1, 4, 5 таблица T_1 переходит в T_1' таблица T_2 — в T_2' и

$$T_1' = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}, \qquad T_2' = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

Все строки таблицы T_1' отличны от всех строк таблицы T_2'. Поэтому набор (2, 3) является тестором для T_1, T_2. Тупиковость (2, 3) очевидна.

Пусть k — число тупиковых тесторов относительно (T_1, T_2), k_i — число тупиковых тесторов, в которые вошел i-й столбец (признак). Положим $S(i) = \frac{k_i}{k}$, $i = 1, 2, \ldots, n$, $N = \sum_{i=1}^{n} S(i)$.

Определение 7. Величина $R(i)$, где $R(i) = \frac{k_i}{k \cdot N}$, называется разделяющим весом i-го столбца (признака), $i = 1, 2, \ldots, n$.

Алгоритмы построения всех тупиковых тесторов даны в [2]. Рассмотрим две операции $A \sim B$, $A \bigcirc B$:

B \ A	01
0	10
1	01

$A \sim B$

B \ A	0 1
0	0 1
1	1 0

$A \bigcirc B$

Пусть задана таблица T, разделенная на две подтаблицы T_1 и T_2, составленная из строк длины n, причем T_1 имеет k строк и T_2 — l строк.

Рассмотрим произвольную строку S длины n, составленную из единиц и нулей. Сопоставим этой строке и таблицам T_1 и T_2 следующие величины:

I. $R_1, \ldots, R_2, R_n$ — разделяющие веса признаков.

II. $P_1(1), P_1(2), \ldots, P_1(n)$ — информационные веса признаков по таблице T_1.

III. $P_2(1), P_2(2), \ldots, P_2(n)$ — информационные веса признаков по таблице T_2.

IV. Пусть α_{ij}, β_{ij} $(i=1, 2, \ldots, n, \quad j=1, 2, \ldots, k)$ — элементы таблиц T_1 и T_2 соответственно и $S = (\gamma_1, \ldots, \gamma_n)$.

Положим:

$$1^\circ. \quad \rho_1 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} R(i)(\alpha_{ij} \odot \gamma_i)}{k}, \quad \rho_2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{l} R(i)(\beta_{ij} \odot \gamma_i)}{l}.$$

$$2^\circ. \quad \bar{\rho}_1 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} P_1(i)(\alpha_{ij} \sim \gamma_i)}{k}, \quad \bar{\rho}_2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{l} P_2(i)(\beta_{ij} \sim \gamma_i)}{l}.$$

$$3^\circ. \quad (\rho_{\min})_1 = \min_j \sum_{i=1}^{n} R(i)(\alpha_{ij} \odot \gamma_i), \quad (\rho_{\min})_2 = \min_j \sum_{i=1}^{n} P_2(i)(\beta_{ij} \odot \gamma_i).$$

$$4^\circ. \quad (\bar{\rho}_{\min})_1 = \min_j \sum_{i=1}^{n} P_1(i)(\alpha_{ij} \odot \gamma_i), \quad (\bar{\rho}_{\min})_2 = \min_j \sum_{i=1}^{n} P_2(i)(\beta_{ij} \odot \gamma_i).$$

$$5^\circ. \quad N_1 = \rho_1 \cdot (\rho_{\min})_1, \quad N_2 = \rho_2 \cdot (\rho_{\min})_2.$$

$$5^{\circ\prime}. \quad N_1' = \frac{k}{2^n} \cdot N_1, \quad N_2' = \frac{l}{2^n} \cdot N_2.$$

$$6^\circ. \quad \bar{N}_1 = \frac{\bar{\rho}_1}{(\bar{\rho}_{\min})_1}, \quad \bar{N}_2 = \frac{\bar{\rho}_2}{(\bar{\rho}_{\min})_2}.$$

$$6^{\circ\prime}. \quad \bar{N}_1' = \frac{k}{2^n} \cdot \bar{N}_1, \quad \bar{N}_2' = \frac{l}{2^n} \cdot N_2.$$

Соответственно пунктам $1^\circ - 6^\circ$(IV) формируется шесть алгоритмов распознавания.

1. Для таблицы T вычисляются величины $R(i)$; экзаменуемая строка S относится к первому классу, если $\rho_1 < \rho_2$, и ко второму, если $\rho_1 > \rho_2$. При $\rho_1 = \rho_2$ распознавание строки S не проводится.

Соответствующий алгоритм распознавания не зависит от кодировки признаков (замены признака на противоположный), «почти всегда» устойчив по отношению к малым изменениям таблицы T, но обладает двумя существенными недостатками. Во-первых, нетрудно указать таблицы T_1 и T_2, такие, что, предъявляя для опознания одну из строк этих таблиц, мы придем к выводу, что ее следует отнести не к тому классу, в котором она содержится (ошибка на исходном материале). Во-вторых, при $n \to \infty$ для таблиц, имеющих один тупиковый тестор $(1, 2, \ldots, n)$, стремится к единице.

Но для таких таблиц все $R(i)$ равны $1/n$ и метрика 1° вырождается просто в среднее расстояние по Хэмингу от строки S до соответствующего множества. Однако если высота таблицы мала по сравнению с $2^{n/2}$, последнее обстоятельство не имеет места.

2. Вычисляются $\bar{\rho}_1$ и $\bar{\rho}_2$. Если $\bar{\rho}_1 > \bar{\rho}_2$, эталон S относится к первому классу, если $\bar{\rho}_1 < \bar{\rho}_2$ — ко второму. При $\bar{\rho}_1 = \bar{\rho}_2$ опознавание не производится.

Достоинства алгоритма 2 те же, что и 1. Недостаток — возможная ошибка при классификации исходного материала. Следует отметить, что этот недостаток присущ многим алгоритмам с усредненной метрикой.

3. Вычисляются $(\rho_{\min})_1$ и $(\rho_{\min})_2$. Если $(\rho_{\min})_1 > (\rho_{\min})_2$, то S относится ко второму классу, если $(\rho_{\min})_1 < (\rho_{\min})_2$ — к первому. При $(\rho_{\min})_1 = (\rho_{\min})_2$ опознавание S не происходит.

Недостатком 3 является неустойчивость. Малое изменение исходной таблицы может существенно изменить прогноз. Ошибки на исходном материале алгоритм не допускает.

4. Строится аналогично 3, но с величинами $(\bar{\rho}_{\min})_1$, $(\bar{\rho}_{\min})_2$.

5. Если $N_1 < N_2$, эталон S относится к первому классу, при $N_1 > N_2$ — ко второму. При $N_1 = N_2$ возникает определенность. Алгоритм достаточно устойчив, ошибка на исходном материале не возникает.

6. Если $\bar{N}_1 < \bar{N}_2$, эталон S относится ко второму классу, если $\bar{N}_1 > \bar{N}_2$ — к первому. При $\bar{N}_1 = \bar{N}_2$ прогноз не делается.

7. Алгоритмы 1—6 предназначены для работы при близких k и l. Если одна из исходных таблиц существенно меньше, необходимо вводить множители $\frac{k}{2^n}$ и $\frac{l}{2^n}$ (см. 5°, 6°).

8. Можно рассматривать и более сложные матрицы. Например, вместо N_1, N_2, $\bar{N}_1$, $\bar{N}_2$ (см. 5°, 6° — IV) можно рассматривать

$$N_1 = (\rho_1)^\alpha (\rho_{\min})_1^\beta, \quad N_2 = (\rho_2)^\alpha (\rho_{\min})_2^\beta,$$

$$N_1 = \frac{(\bar{\rho}_1)^\alpha}{(\bar{\rho}_{\min})^\beta}, \quad N_2 = \frac{(\bar{\rho})^\alpha}{(\rho_{\min})_2^\beta}.$$

Выбор параметров α и β может быть связан с процессом обучения, алгоритмы не зависят от кодировки, устойчивы, не допускают ошибок на исходном материале.

ЛИТЕРАТУРА

1. Дмитриев А. Н. Некоторые табличные числа. Сб. Дискретный анализ, № 12, 1968.
2. Дмитриев А. Н., Журавлев Ю. И., Кренделев Ф. П. О математических принципах классификации предметов и явлений. Сб. Дискретный анализ, № 7, 1966.
3. Дмитриев А. Н., Золотухин В. В., Васильев Ю. И. Опыт применения дискретной математической обработки информации по дифференцированным траповым интрузиям северо-запада Сибирской платформы. Сов. геол. (в печати).
4. Кренделев Ф. П., Дмитриев А. Н., Журавлев Ю. И. Сравнение геологического строения зарубежных месторождений докембрийских конгломератов с помощью дискретной математики, ДАН СССР 1967, т. 173, № 5.
5. Кренделев Ф. П., Дмитриев А. Н. Применение дискретной математики для выбора районов и направления поисково-разведочных работ с целью выявления крупных месторождений типа Витватерсранд. Сб. Пробл. металлоносности древних конгломератов на территории СССР, 1968.
6. Родионов Д. А. Статистические методы разграничения геологических объектов по комплексу признаков. «Недра», 1967.

Институт геологии и геофизики
Сибирского отделения АН СССР,
Институт математики
Сибирского отделения АН СССР,
Новосибирск

Статья поступила в редакцию
5 апреля 1967 г.