А. Н. Дмитриев, Ф. П. Кренделев, А. А. Бишаев,
Е. А. Смертин, Т. И. Штатнова

ОРГАНИЗАЦИЯ И ОБРАБОТКА ГЕОЛОГИЧЕСКОЙ ИНФОРМАЦИИ С ПОМОЩЬЮ ЭВМ НА ОСНОВЕ ПОСТРОЕНИЯ ТУПИКОВЫХ ТЕСТОВ

В данной статье излагаются результаты определенного подхода к классификации геологических объектов и явлений, получившие частичное отражение в ранее публиковавшихся работах (Кренделев, Дмитриев, Журавлев, 1967; Дмитриев, Журавлев, Кренделев, 1966; Дмитриев, Журавлев, Кренделев, 1968; Дмитриев, Васильев, Золотухин, 1968; Дмитриев, Золотухин, Васильев, 1968; Трофимчук и др., 1969; Нестеренко и др., 1969; Модников и др., 1969; Кренделев, Дмитриев, 1969; Константинов, Дмитриев, 1970; Вышемирский и др., 1971), и сделана попытка обобщения этих работ с целью создания одной из возможных моделей целостной обработки геологической информации.

Предлагаемую статью следует рассматривать и как своего рода методическое пособие, поскольку в ней дается свод рекомендаций теоретического и прикладного характера по обобщению геологических описаний. Для обработки больших массивов геологической информации и с целью расширения практических возможностей геолога при истолковании результатов решения конкретных задач на ЭВМ в процессе исследований был найден характеризуемый в данной статье ряд алгоритмов программ и приемов интерпретации. Таким образом, основным содержанием работы является описание новых средств обработки геологической информации. В отличии от получивших широкое распространение процедур, связанных с применением математической статистики, эти средства построены на базе дискретного анализа и используют процедуры распознавания и классификации. При этом характеризуются процедуры, как существенно зависящие от обработки данных на ЭВМ, так и те, результаты которых могут быть получены при ручной обработке информации. Ссылки в квадратных скобках относятся к номерам уравнений, вписанных в текст пособия.

ЛОГИЧЕСКИЕ МЕТОДЫ В ПРЕДАЛГОРИТМИЧЕСКОЙ ОБРАБОТКЕ ИНФОРМАЦИИ

Изучение любого геологического объекта – минерал, рудное тело, месторождение, рудное поле или провинция – начинается с качественного и (или) количественного описания некоторого числа

Таблица 1

Основная таблица характеристик объектов

T_1

Объект	x_1	x_2	. . .	x_i	. . .	x_n	x_{n+1}
S_1	t_{11}	t_{21}	. . .	t_{i1}	. . .	t_{n1}	$t_{n+1,1}$
S_2	t_{12}	t_{22}	. . .	t_{i2}	. . .	t_{n2}	$t_{(n+1),2}$
.	.	.		.			.
.	.	.	. . .	.	. . .	. . .	.
.	.	.		.			.
S_j	t_{1j}	t_{2j}	. . .	t_{ij}	. . .	t_{nj}	$t_{(n+1),j}$
.	.	.		.		.	.
.	.	.	. . .	.	. . .	.	.
.	.	.		.		.	.
S_m	t_{1m}	t_{2m}	. . .	t_{im}	. . .	t_{nm}	$t_{(n+1),m}$

свойств, представляемых в виде набора признаков, характеризующих данный объект. Таким образом, каждый объект может быть охарактеризован строкой признаков $(x_1, x_2, \ldots, x_n)$, называемых характеристиками, где каждый признак имеет установленные для него значения.

Описание совокупности объектов представляется в виде матрицы, в которой строкам соответствуют объекты, а столбцам – признаки (способы кодировки значений признаков будут рассмотрены позже). В общем виде матрица (табл. 1 и 2) имеет следующий вид. В таблицах (T_1) и (T_2) t_{ij} – значения функции x_i на объекте S_j.

В геологических исследованиях табличное задание информации выражается компактной совокупностью функционально связанных и логически сцепленных признаков, характеризующих группу объектов или явлений. Этот факт и специфика геологических постановок задач в ряде случаев исключают применение мощных численных методов решения и приводят к поиску нового аппарата математического исследования геологических объектов. Большинство геологических исследований сводится к поиску меры аналогии исследуемого объекта с уже известными. Многие из геологических задач можно определить как задачи классификации, т.е. определения места данного объекта среди множества родственных ему объектов, для которых известно одно или несколько из интересующих исследователя свойств. Одновременно эту задачу можно считать и диагностической, т.е. задачей, которая сводится к тому, чтобы по наименьшему числу наиболее существенных свойств определить значение исследуемого объекта по заданному свойству. Такова общая направленность геологических задач, подлежащих решению предлагаемым методом.

В дальнейшем в соответствии с принятой терминологией таблица называется таблицей обучения или таблицей эталонов $(T_э)$, если для всех ее строк известны значения целевого признака x_{n+1}; и таблицей проб, или таблицей экзаменуемых объектов, если для всех ее строк значения целевого признака x_{n+1} неизвестны.

Табличное описание совокупности объектов или явлений позволяет решать многие задачи, из которых здесь рассматриваются следующие:

1. Оценка существенности признаков таблицы эталонов и каждой из ее строк; минимизация числа признаков.
2. Нахождение места исследуемого объекта S_γ в ряду объектов обучающей последовательности, описываемых таблицей $T(m \times n)$.
3. Сравнительное изучение двух или более таблиц и классификация объектов.
4. Нахождение места исследуемого объекта, представленного строкой S_γ, в ряду двух или более таблиц обучения, т.е. диагностика.
5. Сравнительное изучение нескольких пятистрочных таблиц, в которых число признаков произвольно, т.е. таблиц $T(m \times n)$, в которой $m = 5$, n – произвольно.
6. Разбиение объектов таблицы $T(m \times n)$ на подтаблицы, т.е. ранжировка таблицы по системам признаков и группам объектов (задачи таксономии).

Способы решения этих задач с помощью алгоритмов и программ, разработанных на базе построения тупиковых тестов, приведены в табл. 3. Эти же задачи исследовались средствами теории распозна-

Таблица 2

Основная таблица характеристик объектов в двоичных символах

T_2

Объект	x_1	. . .	x_i	. . .	x_n	x_{n+1}
S_1	1	. . .	0	. . .	0	β_1
.	.					.
.	.	. . .	. . .	. . .	. . .	.
.	.					.
S_j	0	. . .	1	. . .	1	β_j
.	.					.
.	.	. . .	. . .	. . .	. . .	.
.	.					.
S_m	1	. . .	0	. . .	1	β_m

Таблица 3

Перечень алгоритмов и программ

№ п/п	Группа алгоритмов	Программа ЭВМ	Вычисляемые параметры	Предназначение
	Вычисление тестовых параметров	П–1 (М–20, М–220)	K; K_i; $r(T)$; $P_{(i)}$; $I(S)$	Оценка существенности строк и столбцов таблиц решения
	таблиц двоичных	П–2 (БЭСМ–6)	K, $P_{(i)}$	
1	символов	П–3 (БЭСМ–6)	K^*; $P^*_{(i)}$	Сравнительное изучение таблиц
		П–4 (БЭСМ–6)	$P_{(i)}$; $\delta^{\text{абс}}$; K	Нахождение места пробы среди объектов обучающей последовательности
		П–5 (БЭСМ–6)	K^*	Нахождение места пробы в ряду таблиц обучения
2	Вычисление тестовых параметров по составу таблиц	П–6 (М–220, БЭСМ–4)	K; K_i; $P_{(i)}$; $I(S)$	Обработка пятистрочных таблиц произвольной длины

Примечание. Программисты: Т.Л.Слуцкая (П–1–П–5) и Е.А.Смертин – П–6.

вания образов на базе статистики. Однако логическая природа разработанных алгоритмов позволяет исследователю получать удовлетворительные результаты в решении задач, которые по тем или иным причинам не удовлетворяют статистическим требованиям. Излагаемый метод предполагает предварительную организацию больших массивов информации и обработку таких массивов с помощью ЭВМ. При этом информация может предварительно протоколироваться в описаниях (отчеты, статьи, монографии, документация), а точность исследований и их изложение соответствуют природе и особенностям фиксирования характеристических свойств геологических объектов.

При организации массивов информации и подготовке ее к алгоритмической обработке предполагается, что специфика геологического исследования и фиксирования результатов исследования такова, что:

а) варианты описания одного и того же геологического объекта разными исследователями в основном не противоречат друг другу.

б) описание родственных объектов (например, единого генетического типа месторождений) сопоставимо по большому числу характеристических признаков, отражающих это родство;

в) каждый геологический объект или их группа описываются набором признаков, фиксирующих характеристические особенности именно этих объектов.

Метод неприменим, если таблицей описываются объекты различной природы (например, интрузивное тело и пачка ленточных глин), несопоставимые по основным характеристическим признакам.

Перечисленные требования учитываются при выборе средств, необходимых для решения геологических задач, сформулированных в терминах информационных задач.

Логические основы метода

В практике геологических исследований обнаруживается следующая последовательность операций изучения геологических объектов:

1) наблюдение и фиксирование свойств геологических объектов (геологические тела, процессы);

2) систематизация фиксированных свойств объектов и составление описаний объектов (месторождений, интрузий, свит и т.д.);

3) обобщение полученных описаний объектов (составление классификаций, схем, карт);

4) принятие решения на ориентацию дальнейших исследований.

В связи с тем, что число и разнообразие форм наблюдения и фиксирования результатов непрерывно растет, происходит интенсивное накопление описаний разнообразных фактов и явлений геологического характера. Это накопление описательного материала в проблемах поисково-разведочного характера почти всюду перешло за грань возможностей его осмысливания без применения ЭВМ, что снижает вероятность правильного обобщения материалов и решения.

Предлагаемый метод уменьшает трудности, связанные с обобщением большого количества логических сообщений. При этом предполагается, что каждое геологическое описание может быть подразделено на характеристические признаки (логической или количественной природы), по которым можно составить однозначное представление об объекте, заданном описанием. Для описания, предварительно разделенного на отдельные признаки материала, выбирается код, который соответствует математическим и содержательным требованиям. Закодированные описания обрабатываются на ЭВМ по определенным формализованным правилам, и таким путем можно уменьшить трудности, связанные с большим количеством логических вариантов, которые необходимо рассмотреть при обобщении материала.

Особенностью геологических описаний является то, что большинство характеристических признаков представлены логическими, качественными понятиями. При сравнительном изучении объектов обнаруживается, что характеристические признаки или их совокупности обладают следующими свойствами:

а) при совпадении значений признаков эти признаки отождествляют исследуемые объекты;

б) при несовпадении значений признаков эти признаки различают объекты.

Например, в ряду минералов магнетит–пирротин–пирит признак "магнитность" будет отождествляющим для первой пары и различающим для пар: пирротин–пирит, магнетит–пирит. Указанные свойства признаков широко используются в наборе формализованных процедур, связанных с обработкой геологической информации. Излагаемый метод в своей основе позволяет различать или отождествлять объекты между собой как по отдельному признаку, так и по совместным наборам этих признаков. При этом сравнительное изучение геологических объектов производится путем отыскивания по фиксированной системе признаков всех возможных минимальных описаний, позволяющих распознавать объекты между собой. Такое изучение осуществляется на базе логико–дискретного анализа с привлечением некоторых эвристических процедур, неизбежно возникающих в случае, когда полная разработка математического аппарата отсутствует.

Общая схема обработки информации

Большинство задач геологии сводится к сравнению исследуемого объекта (рудного тела, пласта, залежи, рудного поля, района, провинции) с уже известными и оценке перспектив этого объекта. Поэтому изучение даже одного объекта сопровождается сбором информации по совокупности объектов, которые можно считать эталонами. Сравнение исследуемых районов с эталонными заканчивается представлением информации в виде, удобном для принятия решения. В такой постановке задачи ее решением является обнаружение недостающих признаков и (или) определение запасов полезного ископаемого.

Общая схема исследований следующая. Изучается некоторая проблемная ситуация, которая и является целью и контролирует постановку задачи. Возможны два случая: 1) проблема решается в общем виде; 2) проблема решается по частям.

При решении задачи по частям необходимо уточнить цель. Например, общая задача "увеличить вдвое промышленные запасы золота" может иметь частные задачи "поиск коренных месторождений золота" или "поиск россыпных месторождений золота". Такая частная задача может быть еще более уточнена. Например, "поиск россыпных месторождений золота с запасами на менее 10 т в каждом" или "поиск россыпных месторождений золота с запасами, обеспечивающими добычу его со стоимостью не выше n копеек за грамм".

От указания цели задачи зависит дальнейшая схема учета сбора и обработки информации. В эту схему входят: 1) сбор информации по эталонным объектам (материал обучения) и ее подготовка к обработке на ЭВМ; 2) сбор информации по исследуемому (экзаменуемому) объекту (или объектам); 3) обработка информации материала обучения и объектов экзамена; 4) интерпретация (т.е. геологическое осмысливание полученных результатов); 5) принятие решения о направлении съемочных и (или) поисково–разведочных работ или их прекращении.

Неудача решения вызывает необходимость пересмотра целеуказания, метода и материала исследования, схемы интепретации. В таком случае необходимо попытаться найти другие средства решения задачи.

Геологическую информацию условно можно разделить на три главных вида[1]: а) измерительная (график, число, анализ, кривая и т.п.) б) логико–описательная (да–нет; выше–ниже; больше–меньше; сечет–пересекается и т.п.); в) графическая (карта, план, рисунок, контур и т.д.).

Средства обработки информации должны учитывать специфику геологической информации и изучения совокупности объектов. Поскольку подавляющая часть геологических сообщений представляется логическими формулировками, одной из главных особенностей алгоритмов обработки такой информации должно быть умение работать с логическими данными. В таком случае может возникнуть в соответствии с решаемой задачей проблема преобразования измерительной и графической информации в логическую, т.е. проблема представления всей информации в виде, удобном для работы с автоматическими средствами обработки информации по данному алгоритму.

В практической деятельности все этапы подготовки информации, ее обработка и интерпретация тесно переплетаются, что отражено на схеме (рис. 1).

В данной статье рассматриваются виды работ, связанные с этапом предварительной обработки информации и ее интерпретации (рис. 1, блоки I и III). Все операции формализованной обработки (рис. 1, блок II) вынесены в приложения.

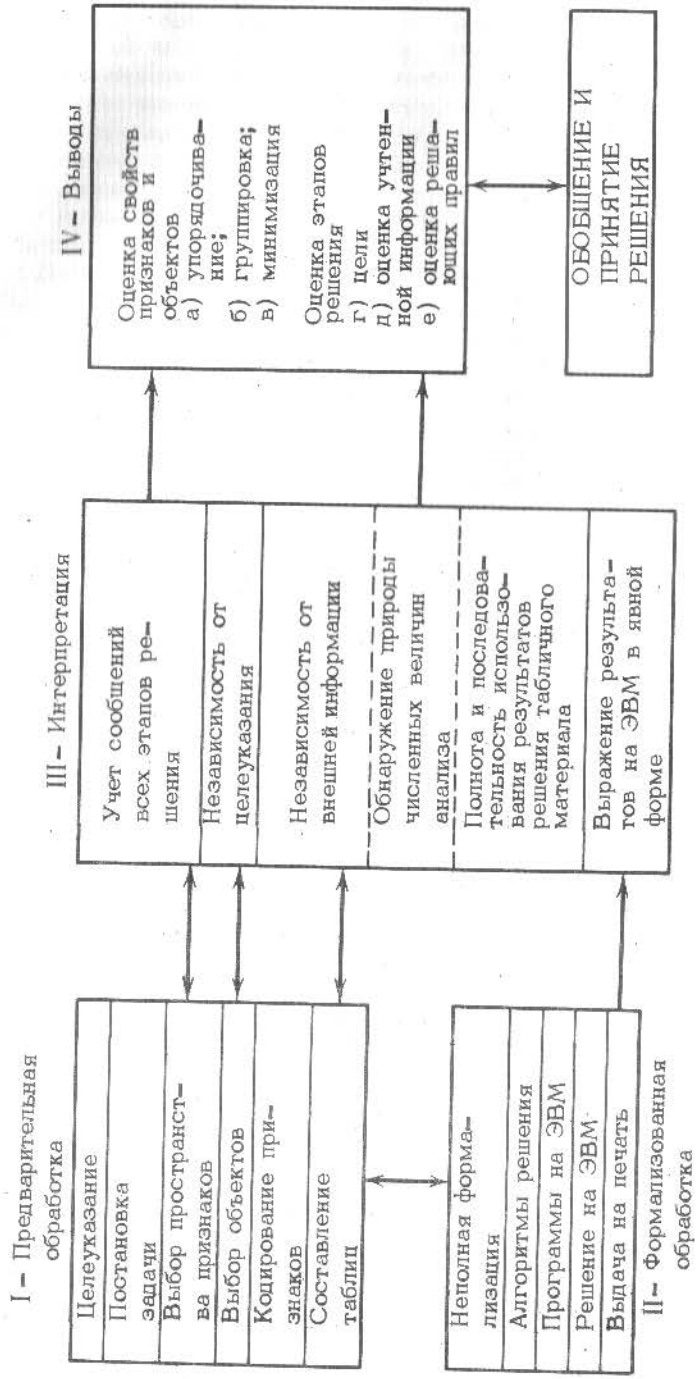
Правильность интерпретации полученных выводов и принятия решения постоянно контролируется:

а) геологической логикой, соответствием полученных выводов установленным закономерностям;

б) включением в объекты экзамена проб с известным значением целевого признака и нахождением его места в обучающей последовательности;

---

[1] Это подразделение вызвано скорее не природой самих сообщений а нашим методом их употребления. Машины недалекого будущего смогут автоматически перерабатывать сообщения всех подразделений в едином цикле. Поэтому мы пользуемся очень общей условной схемой.



Р и с. 1. Общая схема обработки геологической информации на ЭВМ на базе дискретного анализа

в) практикой поисково-съемочных и разведочных работ.

Качество интерпретации в максимальной степени зависит от того, насколько наглядно, удобно для принятия решения представлена обработанная информация.

Постановка задачи

Под постановкой задачи понимается исследование структуры и функционирования отношений между частями целеуказания и информацией, направленной на реализацию данной цели. Развернутая геологическая постановка задачи включает в себя:

- формулирование цели;
- указание на средства ее достижения;
- указания сферы данных, имеющих отношение к целеуказанию.

Пример постановки задачи. Цель – выделение на территории СССР площадей, перспективных на обнаружение месторождений нефти (Дмитриев, 1970).

Средства – сравнительное изучение месторождений в других странах и некоторых районов на территории СССР по самым общим признакам.

Сфера информации – геолого-геофизические характеристики бассейнов (данные картирования, бурения, геофизических съемок, сконцентрированные в фондах, печатных трудах, экономических сводках, устных сообщениях и т.д.).

Таким образом, уже на этапе постановки задачи происходит общее указание на целевой признак. Это позволяет исключить из рассмотрения все объекты и признаки, не имеющие отношения к сформулированной цели. Постановка задачи ограничивает произвол в выборе исследуемых объектов и характеризующих их признаков.

Выбор пространства признаков

Выбор пространства начинается с указания целеобразующих признаков, которые ограничивают сферу сбора информации. Для каждого объекта обучения (например, месторождения, строго соответствующего сформулированной цели) составляет описание, каждый признак которого регистрируется под своим номером.

Полная совокупность признаков всесторонне характеризует объект (полнота и всесторонность понимаются в смысле полноты набора существенных для данного объекта признаков). Выбор признаков контролируется следующими правилами (Волков и др., 1958; Нестеренко и др., 1969; Кренделев и Дмитриев, 1969; Дмитриев, 1970).

1. Каждый признак, заданный логически или количественно, описывает факт, наблюдение, а не его истолкование.

2. Признаки группируются в соответствии с содержательным родством (например, признаки структуры, возраста, состава; региональные и локальные и т.д.).

Рис. 2. Схема деления учтенной и обработанной информации

3. Признаки в своей полной совокупности должны характеризовать объект с подробностью, которая может быть получена на сходной стадии изученности (например, с помощью геологической съемки или разведки, или эксплуатации, т.е. признаки отбираются в соответствии с предполагаемым видом деятельности над объектом экзамена).

4. При прогнозной оценке новых регионов для характеристики объектов обучения отдается предпочтение тщательному отбору общих признаков ("добуровая охарактеризованность") независимо от их природы (геофизические, геологические или геохимические).

5. Не допускается предпочтения фактов; набор признаков не должен составляться в пределах одной какой-либо гипотезы[1].

6. Признаки, интерпретируемые как генетические, задаются в виде однозначно наблюдаемого факта, который потенциально содержит "признак генезиса".

Пример: При решении упоминавшейся выше задачи прогноза месторождений нефти на территории СССР совокупность признаков разбилась на группы так, как показано на рис. 2.

Перечисленные правила не являются исчерпывающими и могут дополняться в соответствии с задачей и опытом геолога, решающего конкретную задачу.

---

[1] Выводы о недопустимости предпочтения одних фактов другим, как и о недопустимости смещения информации, характеризующей объекты разной степени изученности, был сделан на опыте решения задач по золотоносным конгломератам (Трофимук и др., 1969).

Выбор объектов исследования

Из сказанного выше следует, что выбор объектов зависит от целеуказания и определяется им. Целеуказание определяет не только род объекта (полиметаллические руды, угли, нефть – газ, ртуть и другие полезные ископаемые), но и его масштабы (рудное тело, месторождение, район, провинция и т.д.).

Геологические объекты можно разделить на три группы:

1) изученные объекты (эталоны), заданные полными описаниями;

2) изучаемые объекты (пробы), заданные неполными описаниями;

3) подлежащие изучению, для которых описаний еще не имеется.

Машинная обработка информации осуществляется практически только для первых двух групп, причем обычно задача заключается в том, чтобы распознать неполно описанный объект (пробу) и сопоставить его с эталонными. Правила выбора эталонов и проб различаются.

При выборе эталонов учитываются следующие правила.

1. В каждом эталоне значение целевого признака строго соответствует фиксированному целеуказанию (пример: указание вида полезного ископаемого и его запасов).

2. Каждый эталон имеет полный перечень и описание характеристических признаков.

3. Для совокупности эталонов, образующих матрицу $T(m \times n)$, имеется логическая структура их охарактеризованности, т.е. имеется некоторый набор отождествляющих признаков, называемых далее критерием общности. Например, при целеуказании "поиск оруденения типа докембрийских металлоносных конгломератов" в группу эталонов могут включаться только такие месторождения, рудные тела которых локализуются в конгломератах и только докембрийского возраста;

4. Для совокупности эталонов информация может быть разбита на категории, соответствующие этапам геологической деятельности (геологическая съемка, поиск, разведка, эксплуатация).

5. Совокупность эталонов содержит такие признаки, которыми могут быть охарактеризованы и совокупности проб.

При выборе проб главным условием является: а) наличие набора отождествляющих признаков (см. пункт 3 для эталонов), определяющих сходство общих признаков, описывающих как эталоны, так и пробы; б) охарактеризованность проб должна быть не меньше какого-либо фиксированного "порога изученности", т.е. должна заполняться прочерками в определенной пропорции с 1 и 0; в) признаки, которыми характеризуются пробы, имеются и для эталонов.

Выбор эталонов и проб должен обеспечиваться добротным фактическим материалом и гарантировать непредвзятость выбора объектов и признаков (Нестеренко и др., 1969; Дмитриев, 1968; Модников и др., 1969; Дмитриев, 1970).

Кодирование признаков

Сложность кодирования признаков связана с различием в природе геологических сообщений (логические, измерительные, графические). Кодирование – наиболее ответственная процедура, так как от нее зависят результаты решения и возможности их интерпретации. Кодирование должно осуществляться с помощью специалиста-геолога достаточно высокой квалификации.

Способы кодирования зависят от природы информации.

Логические сведения во всех случаях кодируются по принципу: 1 – i–й признак обладает свойством x_i; 0 – i–й признак не обладает свойством x_i; (–) – не известно, обладает ли i–й признак свойством x_i.

Для количественной информации могут применяться различные виды кода:

Первый. 1 – значение x_i i–го признака больше среднего значения; 0 – значение x_i i–го признака меньше среднего значения;

Второй. Для каждого признака строится график (или вычисляется функция) распределения его значений по всем эталонам; затем вычисляется

$$z_i = (x_i - x_{\min})/x_{\max} - x_{\min}. \qquad (1)$$

С помощью (1) все значения x_i i–го признака нормируются к единице. Возможны два случая:

$Z_i >$ 0,5 – ставится 1,

$Z_i <$ 0,5 – ставится 0.

В случае, если $z_i = 0{,}5$, значения 1 или 0 выбираются произвольно или из практических соображений.

Третий. Возможно кодирование по максимуму (или медиане) кривой распределения значений x_i, тогда:

$z_i > \epsilon$ – ставится 1,

$z_i < \epsilon$ – ставится 0,

где ϵ – значение нормируемых величин x_i в максимуме кривой распределений i–го признака[1]. Возможны кодирования по другим критериям. Это будет зависеть от точности измерительной информации и величины выборки (т.е. от объема матрицы $T(m \times n)$).

Представление информации в табличной форме

После выбора кода всех признаков сообщения о признаках и объектах сводятся в таблицу типа T_2, заполненную единицами, нулями и прочерками (табл. 4).

[1] В случае, если $z_i = \epsilon$, выбор значения 1 или 0 выбирается произвольно или из практических соображений.

Среди таблиц выделяются два главных класса:

$T_э$ – таблица эталонов; в ней все строки представлены объектами обучения (эталонами);

T_n – таблица проб (экзаменуемых объектов); в ней все строки представлены пробами.

По характеру заполнения столбцов таблицы T_4 (Дмитриев, Журавлев, Кренделев, 1966; Дмитриев, 1968; Волков и др., 1968; Нестеренко и др., 1968) все признаки могут подразделяться следующим образом:

1. Унарные (взятые по отношению к самим себе и не содержащие незаполненных строк):

а) сквозные или отождествляющие (x_1, x_2), где все строки заполнены одинаковыми символами;

б) различающие (x_3, x_5), где не все строки заполнены одинаковыми символами.

Таблица 4

Объект	Признаки							
	1	2	3	4	5	6	7	8
S_1	1	0	1	0	1	0	1	1
S_2	1	0	1	0	1	1	0	0
S_3	1	0	1	1	0	1	0	1
S_4	1	0	0	1	0	1	0	0
S_5	1	0	0	1	0	0	1	1
S_6	1	0	0	1	1	0	1	1

Таблица 4 (окончание)

Объект	Признаки						Запасы
	9	10	11	12	...i	...n	$n+1$
S_1	1	1	1	–	... 1	... 1	β_1
S_2	0	0	–	–	... 0	... 0	β_2
S_3	1	–	0	–	... 0	... 0	β_3
S_4	0	1	–	–	... 1	... 1	β_4
S_5	1	–	1	–	... 1	... 1	β_5
S_6	1	–	–	–	... 0	... 1	β_6

2. Бинарные (рассматривается отношение двух признаков друг к другу, причем предполагается, что все строки заполнены):

а) симметричные ($x_6 x_7$), где в каждой строке одного из столбцов стоит символ, противоположный символу, стоящему в соответствующей строке другого признака;

б) тождественные (x_8, x_9), где символы, стоящие в соответствующих строках рассматриваемых признаков, одинаковы.

3. Неполно изученные (x_{10}, x_{11}), где для части строк значения признака неизвестны.

4. Неизученные (x_{12}), где для всех строк значения признака неизвестны.

Существуют задачи, в которых объекты разбиваются на классы и каждый класс описывается отдельной таблицей. Тогда количество таблиц соответствует числу классов M, а таблица T_l фиксирует исходную информацию, характеризующую l –й класс объектов, где l –1,2, ..., M. Таблица T_l в содержательном смысле аналогична S_j, но характеризует класс объектов, а не один объект.

Допуск проб к обработке осуществляется не только после проверки на сходство признаков, обобщающих эталоны, но и после оценки степени изученности проб в общем перечне характеристических признаков (см. ниже).

Таблица T, получившаяся по заполнении всех строк и столбцов в соответствии с принятым кодом, называется исходной. Ясно, что исходная T может быть преобразована в соответствии с целями ее обработки. Из таблицы должны быть исключены все столбцы, заполненные прочерками, как неизученные, а также столбцы, заполненные нулями или единицами, как составляющие критерий общности исследуемого типа объектов. Кроме того, из таблицы должны быть удалены строки, в которых все признаки обозначены одинаковыми символами, так как в этом случае объекты становятся эквивалентными (см. выше). В таблице – только одна из таких строк.

Практика показывает, что этап подготовки геологической информации к обработке на ЭВМ является наиболее трудоемким и на него приходится до 80% рабочего времени, затрачиваемого на решение задачи. Этот этап предполагает сотрудничество геолога и математика и заканчивается получением материала в табличной форме, доступной для автоматических средств обработки информации.

Результатом предалгоритмической обработки являются три документа:

1) кривая или таблица значений целевого признака (пример – рис. 2);

2) кодовое описание признаков, т.е. перечень номеров признаков и их содержательная (или количественная) характеристика;

3) исходная таблица описания изучаемых объектов.

НЕПОЛНОСТЬЮ ФОРМАЛИЗОВАННАЯ ОБРАБОТКА ГЕОЛОГИЧЕСКОЙ ИНФОРМАЦИИ, ЗАДАННОЙ В ТАБЛИЧНОЙ ФОРМЕ

Общая структура неполно формализованной обработки геологической информации иллюстрируется схемой (см. рис. 1, блок 2).

Укажем, что везде, где это возможно, употреблены простейшие варианты и способы изложения математического материала. В связи с этим в работе значительно ослаблена строгость математического изложения и приведены только необходимые для понимания метода определения и терминология со ссылкой на соответствующую литературу, где дано более обстоятельное изложение.

Общие положения и определения

Пусть задано конечное множество $T = \{S_j = \{t_{ij}\},\ j = 1, \ldots, m;\ i = 1, \ldots, n\}$, на элементах которого определены две системы признаков (предикатов): $X = < x_1, x_2, \ldots, x_n >$ и $X = < x_{n+1}, x_{n+2}, \ldots, x_{n+k} >$. Признаки $x_{n+1}, x_{n+2}, \ldots, x_{n+k}$ назовем целеобразующими (основными или целевыми предикатами), а признаки $x_1, x_2, \ldots, x_n$ – характеристическими признаками (далее просто признаками или предикатами).

Элементы множества $S_j = \{t_{ij}\}$ называются эталонными описаниями объекта S_j, если для них известны значения основного предиката x_{n+1}; а коэффициент изученности [12] $I(T) = 1$; и пробами S_γ (или описаниями объектов экзамена), если значения основного предиката x_{n+1} неизвестны, а $I(T) \leqslant 1$. В дальнейшем эталонные и пробные описания будут называться просто эталонами и пробами. Отметим также, что решение задач прогнозно-поискового характера связано с выяснением мер близости проб и эталонов по признакам $X = < x_1, \ldots, x_n >$. Установленные с помощью формальных процедур меры близости между пробами и эталонами являются диагностическими величинами проб. Нередко также возникает задача сравнительного изучения классов объектов (таблице соответствует класс) между собой и другие задачи подобного характера. Для решения этих задач, помимо алгоритмов, требуются некоторые предварительно организующие информацию правила с нежесткой конструкцией.

Совокупность признаков X, значения которых t_{ij} заданы в алфавите $\{0, 1, -\}$

$$t_{ij} = \begin{cases} 0 - \text{объект } S_j \text{ обладает признаком } x_i, \\ 1 - \text{объект } S_j \text{ не обладает признаком } x_i, \\ (-) - \text{значение признака } x_i \text{ для } S_j \text{ неизвестно,} \end{cases}$$

в реальных таблицах подразделяется на две главные группы – отождествляющую и различающую группы признаков.

Признак x_i над объектами множества $\{S_j,\ j = 1,2,\ldots, m\}$ называется отождествляющим, если для всех описаний элементов таблицы T искомого признака, не содержащих знак –, значение x_i вез-

де одинаково; признак x_i называется положительно отождествляющим, если для всех объектов из T, не содержащих знака –, значение $x_i(S_j) = 1$, и отрицательным, если $x_i(S_j) = 0$.

Совокупность отождествляющих признаков образует критерий общности класса M таблицы $T(m \times n)$. Критерий общности класса выступает в роли меры сродства объектов в классе. Чем выше процент отождествляющих признаков в $X = < x_1, x_2, \ldots, x_n >$, тем сильнее сжаты объекты исследования в класс. Критерий общности организует объекты в класс, и процедура диагностики проб должна начинаться с выявления соответствия исследуемой пробы критерию общности класса. Нередко отождествляющие признаки фигурируют в качестве целеобразующих.

Признак x_i над объектами множества $\{S_j, j = 1, \ldots, m\}$ называется различающим, если для всех описаний элементов T искомого признака, не содержащих –, существуют такие i_1, i_2, что значение признака

$$x_i(S_{j_1}) = 1, \text{ а } x_i(S_{j_2}) = 0.$$

Различающий признак x_i называется пропорциональным в T, если число S_j, таких, что $x_i(S_j) = 1$, равно числу S_j, таких, что $x_i(S_j) = 0$.

Если таблица T предварительно подразделена на T_1 и T_2 и для T_1 все значения $x_i(S_j) = 1$, а для T_2 все значения $x_i(S_j) = 0$ (или наоборот), то такой признак называется максимально различающим для T_1 и T_2. Различающий признак, принимающий с вероятностью 0,9 и более значение $x_i(S_j) = 1$, называется сходящимся к положительному, а при значении $x_i(S_j) = 0$ – к отрицательному отождествляющему признаку. Большое количество таких признаков нежелательно в таблицах, подвергаемых непосредственной обработке на ЭВМ. Роль различающих признаков сводится к ранжированию объектов внутри исследуемого класса и к распознаванию места пробы внутри класса, к которому она отнесена процедурой распознавания.

Основные алгоритмы обработки информации, заданной в виде одной таблицы

В отличие от изложения алгоритмов, данных в математических справках к программам решения, приводимое ниже описание нацелено не столько на формальное построение, сколько на показ идейного источника и понятийной базы, из которых строятся алгоритмы.

Идеи алгоритмов решения задач исследуемого профиля восходят к работе (Смертин, Дмитриев, 1970), а их понятийная база в применении к геологическим задачам изложена в работах (Дмитриев, Журавлев Кренделев, 1968; Волков и др., 1968).

Алгоритмический этап решения задач начинается с момента окончательного построения исходных таблиц, преобразование которых заканчивается построением допустимых таблиц.

Определение 1. Таблица T, состоящая из значений признаков в алфавите $\{1, 0, -\}$, называется допустимой, если:

- в наборе столбцов таблицы $X = \{x_1, \ldots, x_n\}$ не содержится отождествляющих признаков;
- все строки таблицы T различны;
- каждая строка таблицы T удовлетворяет требованиям критерия общности для T;

Отметим, что две строки S_{j_1} и S_{j_2} таблицы T различны, если имеется столбец за номером i – такой, что

$$t_{ij_1}, t_{ij_2} \in \{0,1\} \text{ и } t_{ij_1} \neq t_{ij_2}. \quad (1)$$

Допустимые таблицы согласно определению 1 концентрируют в себе все характеристики объектов, которые в выбранном алфавите значений признаков позволяют исследовать различия объектов и (или) их сходство. Основополагающими мерами объектов, подвергаемых исследованию, будут меры оценки столбцов и строк допустимых таблиц.

Такие оценки вычисляются на базе построения всех тупиковых тестов для допустимых таблиц и называются информационными весами строк и столбцов. Изложим наиболее простым способом понятия, лежащие в основе алгоритмов решения.

Определение 2. Набор столбцов с номерами $i_1, i_2, \ldots, i_l$ допустимой таблицы называется тестом таблицы T, если после удаления из T всех столбцов, за исключением столбцов с номерами $i_1, i_2, \ldots, i_l$, получается таблица T', все строки которой попарно различны

Пример 1.

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

T_5

1	2	5
1	0	0
0	1	0
1	1	0
1	1	1

T'_5

Набор столбцов T'_5 (1, 2, 5) является тестом таблицы T_5. Действительно, убрав из T_5 столбцы 3 и 4, получили T'_5, все строки которой попарно различны. Заметим, что сама таблица T_5 также является тестом.

Таким образом, с точки зрения своей способности различать все строки T_5 является избыточной. Указанное рассмотрение позволяет сформулировать следующее определение.

Определение 3. Тест, составленный из столбцов с номерами $i_1, \ldots, i_l$, называется тупиковым, если из него нельзя удалить ни одного столбца без того, чтобы он перестал быть тестом. Из определения 2 следует, что если столбцы образуют тупиковый тест, то удаление из таблицы, составленной из этих столбцов, любого столб-

ца приведет к появлению хотя бы двух тождественных, совпадающих по всем значениям строк. Так, таблица T_5' в примере 1 образует тупиковый тест таблицы T_5, поскольку удаление из нее любого столбца приводит к появлению тождественных пар строк:

1–1		1–2		1–3	
1	2	1	5	2	5
1	0	1	0	0	0
0	1	0	0	1	0
1	1	1	0	1	0
1	1	1	1	1	1

Пусть $t_1, t_2, \ldots, t_k$ – все тупиковые тесты допустимой таблицы T. Возьмем столбец с номером i, соответствующий различающему признаку, и выделим все тупиковые тесты, в которые входит этот столбец. Полученную величину обозначим через K_i.

Определение 4. Величину $P_{(i)}$, определяемую равенством

$$P_{(i)} = K_i/K, \qquad (2)$$

назовем информационным весом i-го различающего признака.

В величине $P_{(i)}$ условно принимается эквивалентность тупиковых тестов различной длины. Вообще говоря, тупиковые тесты различной длины диагностически неравноценны. Выявление распределения длин тупиковых тестов дает дополнительные сведения о характере исследуемых объектов. Основное диагностическое значение несет число тупиковых тестов данной длины, но не сама длина. Число минимальных и максимальных тестов, как правило, составляет меньше 5% от общего числа тестов. Таким образом, при решении задач практического характера $P_{(i)}$ можно принимать без поправок на вхождение тестов различной длины. Подробно этот вопрос кратко освещен в работе (Дмитриев и Смертин, 1970).

Пример 2

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

Приведенная таблица имеет три тупиковых теста: <1, 2, 5>, <2, 4, 5>, <1, 2, 3, 4>, тогда $P_{(1)} = 2/3$; $P_{(2)} = 1$; $P_{(3)} = 1/3$; $P_{(4)} = 2/3$ и $P_{(5)} = 2/3$. Столбец 2 вошел во все тупиковые тесты.

Содержание понятия "информационный вес" можно пояснить следующим образом. Описание выбранной совокупности объектов всеми исходными различающими признаками является избыточным по отношению к элементарной процедуре – различать все строки таблицы. После удаления некоторых признаков описание сохраняется основное свойство – оно различает все строки таблицы. Последовательно удаляя столбцы, мы приходим к несжимаемому (неизбыточному) описанию, которое при дальнейшем сжатии теряет свойство различать строки допустимой T. Такие несжимаемые описания или тупиковые тесты являются как бы корнями, основами остальных описаний.

Естественно считать, что чем в большее число таких основных (коренных) описаний входит признак, тем он существеннее при описании строк данной таблицы. Из двух различающих столбцов, входящих в допустимую таблицу, более существенным для характеристики различий строк таблицы является тот, для которого информационный вес больше. Значения $P_{(i)}$ колеблются в пределах от нуля до единицы.

На базе значений $P_{(i)}$ оцениваются меры важности строк (объектов). Основными из этих мер являются: информационные веса строк, определяемые по равенству

$$I(S_j) = \sum_{i=1}^{n} t_{ij} P_{(i)}, \qquad (3')$$

в случае, если T не содержит прочерков;

$$I(S_j) = \sum_{t_{ij}=1} P_{(i)} + \frac{1}{2} \sum_{t_{ij}=\text{—}} P_{(i)}, \qquad (3'')$$

если T содержит прочерки; или в общем виде:

$$I(S_j) = \sum_{i=1}^{n} P_{(i)} |t_{ij}|, \text{ где}$$

$$|t_{ij}| = \begin{cases} 0 \text{ при } t_{ij} = 0, \\ 1 \quad " \quad t_{ij} = 1, \\ \frac{1}{2} \quad " \quad t_{ij} = (-) \end{cases} \quad \text{для } j = 1,2,\ldots,m. \qquad (3)$$

Во избежание влияния ориентации в ряде конкретных задач (2, 3, 20) были введены величины, не зависящие от ориентации кода. Одна из таких величин – взвешенный информационный вес строк, определяющийся суммой только тех значений $P_{(i)}$ в строке S_j, которые соответствуют большей условной вероятности вхождения 0 или 1 для данного столбца в целом. Взвешенный информационный вес строки определяется выражением

$$I(S_j) = L - \sum_{i=1}^{n} ([2|t_{ij}| - \mathcal{P}_i|] - [|t_{ij}| - \mathcal{P}_i|])P_{(i)} =$$

$$= L - \sum_{i=1}^{n} P_{(i)} (|t_{ij}| - [\mathcal{P}_i]), \qquad (4)$$

где L – средняя длина тупикового теста данной таблицы (см. ниже, уравнение (13)), а $\mathcal{P}_i$ – пропорциональность столбца (см. ниже, уравнение (8)).

Информационный вес признака (2), информационный вес строки (3) и взвешенный информационный вес строки (4) являются основополагающими величинами в задачах, связанных с обработкой логической информации.

$$[P_i] = \begin{cases} 1 \text{ при } P_i \geqslant 0{,}5, \\ 0 \text{ при } P_i \leqslant 0{,}5. \end{cases}$$

В случае $P_i = 0{,}5$ выбор значения производится случайным образом или по практическим мотивам.

Основные алгоритмы обработки информации, заданной в виде нескольких таблиц

В случае, когда требуется одновременное исследование двух и более таблиц, используется понятие тестора (Дмитриев, Журавлев, Кренделев, 1968; Волков и др., 1968).

Определение 5. Набор столбцов $(i_1, i_2, \ldots, i_l)$ в T называется тестором для T_1 и T_2 если после удаления из T_1 и T_2 столбцов, не вошедших в набор $(i_1, i_2, \ldots, i_l)$, в таблицах T_1 и T_2 не будет общих строк. Тестор, согласно определению, может обладать избыточностью (по "затрате" столбцов на один тестор), поэтому естественно сформулировать следующее определение.

Определение 6. Тестор t^* для T_1 и T_2 называется тупиковым тестором, если после удаления из него какого-либо столбца он перестает быть тестором для T_1 и T_2.

Отметим, что в определении тупиковости тестора содержится указание на качество столбцов, из которых составлен данный тестор. Чем меньше столбцов затрачивается на построение тупикового тестора, тем бо́льшая различающая способность столбцов в распознавании строк, принадлежащих T_1 и T_2.

Пример 3. Пусть T_6 разбита на T_6^1 и T_6^2 таким образом, что все строки T_6^1 отличны от строк T_6^2

$$T_6 = \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \qquad T_6^1 = \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

$$T_6^2 = \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

Очевидно, что набор столбцов (x_2, x_3) является тестором для T_6^1 и T_6^2. Удалив из T_6 столбцы x_1, x_4, x_5 и переходя к T_6^{1-1} и T_6^{2-1}, получим

$$T_6^{1-1} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{и} \quad T_6^{2-1} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

Для того чтобы выделить группу признаков (столбцов), которыми охарактеризованы объекты (строки таблицы) по их свойству максимально разделять строки таблиц $T_1, T_2, \ldots,$ нам и потребуется величина следующего содержания.

Пусть K^* – число всех тупиковых тесторов таблицы T, составленной из T_1 и T_2, в которые вошел столбец за номером i.

Определение 7. Число

$$P^*_{(i)} = K^*_i / K^*, \quad i = 1,2,\ldots,n, \qquad [5]$$

называется тесторным информационным весом признака x_i.

Пример 4. Пусть имеются две таблицы.

$$T_7 = \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array} \quad \text{и} \quad T_8 = \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Для них тупиковыми тесторами будут

$$T_7^1 = \begin{array}{c} \begin{array}{ccc} 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \end{array} \qquad T_7^2 = \begin{array}{c} \begin{array}{ccc} 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \end{array}$$

$$T_8^1 = \begin{array}{c} \begin{array}{ccc} 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \end{array} \qquad T_8^2 = \begin{array}{c} \begin{array}{ccc} 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

Отсюда можно вычислить:

$$P^*_{(1)} = P^*_{(3)} = 1/2; \quad P^*_{(2)} = P^*_{(4)} = 1, \quad P^*_{(5)} = 0,$$

поскольку x_5 не входит ни в один набор.

На базе перечисленных процедур и количественных величин построены алгоритмы голосования по тестам и тесторам (Дмитриев, Журавлев, Кренделев, 1966), которые прямо примыкают к распространенным способам распознавания образов.

НЕКОТОРЫЕ ПУТИ ИСПОЛЬЗОВАНИЯ ТЕСТОВЫХ ПАРАМЕТРОВ БИНАРНЫХ ТАБЛИЦ ПОСЛЕ ИХ ОБРАБОТКИ НА ЭВМ

Параметрами бинарных таблиц назовем все вводимые числовые величины, которыми характеризуются или могут быть охарактеризованы таблицы. Параметры таблиц подразделяются на априорные и апостериорные (Дмитриев, Журавлев и Кренделев, 1966). Априорными параметрами называются первичные числовые величины, которыми таблицы характеризуются еще до их обработки на ЭВМ, апостериорными – числовые или иные величины, которые характеризуются таблицы после их обработки на ЭВМ.

А. Априорные параметры (первичные или заданные). Следует отметить, что уже при организации информации по решаемой проблеме обнаруживаются изъяны в качестве и количестве информации. Табличное задание сообщений в свою очередь является не только формой обобщенного и разового задания информации, но служит источником информации еще до обработки таблицы на ЭВМ. Действительно упорядоченная согласно целеуказанию и выраженная таблицей бинарных символов информация представляет собой своего рода информационную карту, вмещающую много полезных сообщений. Для облегчения получения таких сообщений можно пользоваться априорными параметрами таблицы.

После вынесения всех отождествляющих столбцов (признаков) таблицы и образования критерия общности (класса объектов) вычисляются параметры для одной допустимой таблицы $T(m \times n)$, заполненной символами из алфавита $\{0, 1, -\}$. Среди этих параметров отметим следующие.

Относительная ширина h таблицы T помогает оценивать тестовую трудоемкость таблицы и выражается соотношением

$$h = n/m. \tag{6}$$

Отношение

$$Q = q/(n-q), \tag{7}$$

где q – число признаков, образующих критерий общности, выражает степень сжатости изучаемых объектов в класс.

Степень пропорциональности i-го признака, контролирующего ориентацию кода, в таблице $T(m \times n)$, заполненной символами алфавита $\{1, 0, -\}$, может быть вычислена по соотношению

$$\mathcal{P}_i = \frac{1}{m} \sum_{j=1}^{m} |t_{ij}|. \tag{8}$$

Необходимо соблюдать некоторые пороговые значения $\mathcal{P}_i$, из которых рекомендуются те его значения, которые принадлежат интервалу $[0{,}2;\ 0{,}8]$. Пропорциональность таблицы в целом определяется из соотношения

$$\mathcal{P} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{P}_i. \tag{9}$$

Выражением

$$d_{j,j'} = \frac{1}{n} \sum_{i=1}^{n} \big| |t_{ij}| - |t_{ij'}| \big| \tag{10}$$

характеризуется исходное различие между любой парой строк на априорных параметрах, поскольку согласно требованиям к построению допустимых таблиц, в T не должно быть одинаковой пары строк. Для более четкого выделения различающей особенности строк потребуем, чтобы $d_{j,j'} \geqslant 0{,}2$.

Для характеристики исходной различимости T воспользуемся выражением

$$d = \frac{1}{C_m^2} \sum_{(j,\, j')} d_{j,j'}, \tag{11}$$

где C_m^2 – это число всех пар строк (j, j'), таких, что $1 \leqslant j \leqslant j' \leqslant m$ (знак $\sum_{(j,\, j')}$ означает суммирование по таким парам). Параметры для нескольких таблиц вводятся в соответствии с характером параметров для одной таблицы.

Наличие прочерков в таблице позволяет оценить степень изученности признаков объектов и(или) классов объектов. Для грубой оценки коэффициента изученности объектов одной таблицы произведем переобозначение, заменив прочерк на 0,5. Тогда описание всех S_j будет представлено как $\{S_1, S_2, \ldots, S_m\}$, где каждая строка задана некоторым n-мерным вектором $X = (x_1, x_2, \ldots, x_n)$, где $x_i \in \{0;\ 1;\ 0{,}5\}$ для всех $i = 1, 2, \ldots, n$. Сформулируем общее правило количественной оценки состояния изученности объектов по данному перечню признаков при условии, что полная изученность объектов оценивается единицей, а полная неизученность – нулем, в предположении, что все признаки одинаково существенны для характеристики исследуемых объектов. Это правило можно выразить в виде следующей эмпирически построенной числовой зависимости (4):

$$I(T) = \sum_{i=1}^{n} 4\nu_i (x_i - 0{,}5)^2, \tag{12a}$$

которая каждой таблице T ставит в соответствие число из отрезка $(0; 1)$, где ν_i – доля признака x_i в изученности, при этом

$$\sum_{i=1}^{n} \nu_i = 1.$$

В конкретных случаях возможно предварительное фракционирование объектов по коэффициенту изученности. Полагая $\epsilon = 0{,}5$, получим:

$I(T) > \epsilon$ – объект допускается к обработке, (12б)

$I(T) \leqslant \epsilon$ – объект направляется на доисследование.

Б. Апостериорные параметры. В результате вычислений на ЭВМ получаются основополагающие величины $P_{(i)}$; $I(S_j)$; $P^*_{(i)}$ и некоторые другие, однако при свертывании информации часть полезных сведений теряется, но может оказаться полезной при анализе тонких структур таблиц.

Для вычисления трех основополагающих величин необходимо знать общее число тупиковых тестов. Однако из самого определения (6, 7, 21) ясно, что тупиковые тесты имеют различную длину. Введем следующие параметры: K^l – число тупиковых тестов длины l в таблице T, $l = 1,2,\ldots, m-1$; K^l_i – число тупиковых тестов длины l, в которые входит столбец x_i, $i = 1,2,\ldots, n$; $l = 1,2,\ldots, m-1$; K – число тупиковых тестов таблицы T; очевидно, что

$$K = \sum_{l=1}^{m-1} K^l.$$

K_i – число тупиковх тестов T с участием i-го столбца; очевидно, что

$$K_i = \sum_{l=1}^{m-1} K^l_i.$$

Определим L – среднюю длину тупикового теста таблицы T

$$L = \frac{\sum\limits_{t\in\tau(t)} l(t)}{K} = \frac{\sum\limits_{l=1}^{m-1} lK^l}{K} = \frac{\sum\limits_{i=1}^{n} K_i}{K}. \qquad (13)$$

Величина L требуется для установления центра тяжести таблицы по отношению к минимальной и максимальной длинам тестов. Сдвиг L к минимуму длин тестов свидетельствует о большой различающейся способности столбцов (признаков), а в сторону максимума длин – об отождествляющей способности столбцов. Это указание контролирует качество учтенной информации и ее соответствие целеуказанию. Для i-го столбца таблицы определим величину l_i – среднюю длину тупиковых тестов, в которые входит столбец x_i, по формуле

$$l_i = \frac{\sum\limits_{i=1}^{m-1} lK^l_i}{K_l}.$$

При исследовании свойств нескольких таблиц (классов объектов) для каждой таблицы находятся значения пропорциональностей признаков по соотношению (8). Строку, составленную из этих значений, назовем среднетипичным представителем таблицы (класса) $(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n)$, тогда

$$I^*(S) = \sum_{i=1}^{n} P_{(i)} \mathcal{P}_i. \qquad (14)$$

называется информационным весом таблицы (класса). Эта величина имеет тот же смысл, что и $I(S)$ для строки, но относится к таблице в целом. Используется $I^*(S)$ для сравнительного изучения классов.

Перечисленные параметры не исчерпывают полученные результаты решения, но достаточны для работы интерпретатора. Отметим, что в задачах, прямо связанных с распознаванием образов, используется различающая мера столбца

$$R_{(i)} = \frac{n(m-l_i)}{(m-1)L} P_{(i)} \qquad (15)$$

и различающая мера между парой строк $S_j, S_{j'}$ с использованием апостериорных параметров, определяемая из соотношения

$$a_{j,j'} = \sum_{i=1}^{n} R_{(i)} |t_{ij} - t_{ij}| \Big/ \sum_{i=1}^{n} R_{(i)}, \qquad (16)$$

причем ясно, что $1 \geqslant a_{j,\,j'} \geqslant 0$. Аналогично (16) можно записать выражение для оценки расстояния между объектами по значениям целевого признака

$$\Delta_{j,j'} = \frac{|x^j_{n+1} - x^{j'}_{n+1}|}{x^{\max}_{n+1}}, \qquad (17)$$

причем значение признака x_{n+1} изменяется в широких пределах (например, исследуются сравнительные особенности гигантских и мелких месторождений), а так как нас интересуют при малых значениях x_{n+1} небольшие различия и при больших значениях x_{n+1} – большие различия, то лучше воспользоваться вместо уравнения [17] уравнением вида

$$\Delta'_{j,j'} = \frac{|\ln x^j_{n+1} - \ln x^j_{n+1}|}{\ln x^{\max}_{n+1}}. \qquad (18)$$

После нахождения коэффициентов различия между объектами можно перейти к оценке общей информативности данной системы признаков, собранной в соответствии с целеуказанием.

Почти во всех случаях для задач, связанных с прогнозом и поиском месторождений, на конечном этапе решения требуется выявить характер взаимосвязи между значениями $a_{j,j'}$ и $\Delta_{j,j'}$.

Если взаимосвязь обнаруживается, т.е.

$$a + \epsilon = f(\Delta), \tag{19}$$

то функция, удовлетворяющая уравнению (19), будет отражать вид этой взаимосвязи. В конкретном случае это может быть любая из следующих функций:

$$c,\ x^n,\ e^x,\ \ln x.$$

Указание вида функциональной зависимости дает возможность оценить информативность данной системы признаков для распределения объектов по целевому назначению.

Для случая, когда найдены информационные веса строк $I(S)$, уравнение (19) запишется как

$$I(S) \pm \epsilon = f(x_{n+1}). \tag{20}$$

Вполне очевидно, что для каждой строки S_j значение ϵ_j будет иметь конкретный размер – такой, что

$$I(S_j) \pm \epsilon_j = f(x^j_{n+1}). \tag{21}$$

Тогда, воспользовавшись статистическими критериями, запишем:

$$\bar{\sigma} = \frac{1}{m}\sum_{j=1}^{m} |\epsilon_j| \quad \text{или} \quad \sigma = \frac{1}{m}\sum_{j=1}^{m} \frac{|\epsilon_j|}{x^j_{n+1}}, \tag{22}$$

где $\bar{\sigma}$ – среднее отклонение объекта по диагностической величине от величины значения по x_{n+1} и σ – коэффициент отклонения. Соотношения вида (19) – (22) были использованы в ряде конкретных задач, связанных с прогнозированием полезных ископаемых (Соловьев, 1968; Модников и др., 1969). Величина отклонений, как правило, зависит от характера и точности характеристических признаков, которые отбирает профессионал-геолог. По степени отклонений можно сформировать пространство признаков, более всего относящееся к величине запасов.

Обработка результатов голосования по тестам и тесторам

В задачах диагностики объектов на тестовой основе в постановке распознавание образов используется алгоритм голосования по тестам и тесторам (Чегис и Яблонский, 1958; Дмитриев и Смертин, 1969). Здесь мы укажем на простейшие формы использования результатов голосования.

А. Голосование по тестам. Пусть K – количество всех тупиковых тестов, построенных для таблицы T и пробы S_γ. Через K_j обозначим количество тупиковых тестов, в которых проба совпадает с объектом S_j из эталонной таблицы T.

Поскольку из понятия тупикового теста вытекает, что $K \geqslant \sum_{j=1}^{m} K_j$, то коэффициент принадлежности пробы к классу (таблице T) можно вычислить как

$$\delta = \sum_{j=1}^{m} K_j / K, \tag{23}$$

причем $1 \geqslant \delta \geqslant 0$.

Коэффициент отнесения пробы S_γ к конкретному эталонному объекту S_j в T вычисляется по соотношению

$$\delta_j = K_j / \sum_{j=1}^{m} K_j. \tag{24}$$

Этот коэффициент определяет силу отнесения пробы к объекту S_j по сравнению с другими объектами. Нахождение места пробы в ряду эталонов состоит в отыскании $\max \delta_j$ (чем больше проба тяготеет к одному из эталонов, тем ближе $\max \delta_j$ к 1).

Неопределенность в отнесении пробы к конкретному эталону возрастает с уменьшением $\max \delta_j$. В этом случае $\max \delta_j \to \frac{1}{m}$. Отметим, что для интерпретации необходимо учитывать величины δ и δ_j одновременно, т.е. об абсолютной величине отнесения пробы к конкретному объекту S_j можно судить по величине

$$\delta^{\text{абс}}_j = \delta\delta_j = K_j / K. \tag{25}$$

Б. Голосование по тесторам. Пусть K^* – количество всех тупиковых тесторов, а m – число классов объектов, т.е. число таблиц T_j, $K_j(J)$ – количество голосов, поданных за объект S_j в классе T_j, где $j = 1,2,\ldots,m(J)$; $J = 1,2,\ldots,m$.

Для тесторов по аналогии с тестами вытекает, что

$$\sum_{J=1}^{m} K_j^{\max}(J) \leqslant K^*. \tag{26}$$

Принадлежность пробы к группе таблиц T_J (классов m) можно установить, используя величину

$$\delta^* = \sum_{J=1}^{m} K_j^{\max}(J) / K^*. \tag{27}$$

Таблица 5

Основные тестовые апостериорные параметры, вычисляемые с помощью программ П-1 – П-6

Вычисляемый параметр	Номер формулы	Природа и значение
		Исследование столбцов (признаков)
$P(i)$	(2)	Информационный вес признака – относительная важность признака для количественного сравнения всех эталонных строек (объектов) в таблице[1] при изучении различных строк
$P^*_{(i)}$	(5)	Различающий тесторный вес столбца указывает на интенсивность различия нескольких сравниваемых таблиц по данному признаку
$R(i)$	(15)	Нормированный различающий вес признака, в который, кроме $P_{(i)}$, входят $n/m-1$ – нормирующий множитель таблицы, L – коэффициент экономичности различия всех столбцов таблицы, $(m - l_i)$ – мера экономичности различия данного столбца. Используется как величина, уточняющая $P_{(i)}$ при ранжировании столбцов и строк таблицы
		Исследование строк (объектов)
$I(S_j)$	(3)	Информационный вес строки оценивает степень приближения количественной оценки строки к величине средней длины тупикового теста данной таблицы; указывает существенность строки по количеству информации в связи с целевым признаком; служит для упорядочивания строк таблицы
$a_{j,j}$	(16)	Относительная существенность пары строк и их взаимное распределение по мере различия независимо от направленности кодирования
$I_{(p)}(S_j)$	(4)	Взвешенные информационные веса строк оценивают близость каждой строки таблицы к наиболее типичной (усредненной) строке и используются для определения степени компактности объектов в классе (родства)

[1] В настоящей таблице основополагающий для интерпретации параметр подчёркнут.

Таблица 5 (окончание)

Вычисляемый параметр	Номер формулы	Природа и значение
$\Delta_{j,j'}$	(17)	Относительное различие между любой парой строк по целевому признаку
$I^*(S)$	(14)	Величина, аналогичная $I(S_j)$, для исследования свойств нескольких таблиц (классов объектов) при их сравнительном изучении
σ	(22)	Мера отличия распределения объектов по существенности данной системой признаков от распределения этих же объектов по целевому признаку. Служит для обнаружения качества целевой информативности учтенной системы признаков
		Диагностика объектов
$I(S_\gamma)$	(3)	Информационный вес строки (пробы) позволяет найти место пробы среди объектов, заданных таблицей эталонов (одного класса)
$\underline{I^*(S)}$	(14)	Место пробы среди объектов нескольких классов (нескольких таблиц, подвергнутых сравнительному изучению)
δ	(23)	Мера принадлежности пробы к таблице в целом, находимая процедурой тестового голосования; служит оценкой "погружения" экзаменуемого объекта в класс, к которому объект принадлежит
$\delta^{абс}_j$	(25)	Мера принадлежности пробы к конкретному объекту (строке) класса (таблицы); обнаруживается ближайший аналог среди изученных объектов класса
δ^*	(27)	Меры, аналогичные δ и $\delta^{абс}$, но вычисляемые с помощью процедур тесторного голосования. Величины (23), (25), (27), (29) являются основными диагностическими величинами в задачи распознавания образцов
$\delta^{*абс}_j$	(29)	

Тогда коэффициент отнесения пробы к одному из классов найдем, как

$$\delta^*_{j(J)} = K_j(J) / \sum_{J=1}^{m} K_j^{\max}(J), \qquad (28)$$

δ^*_J определяет величину отнесения к классу и аналогична δ_j при отнесении пробы к объекту в тестовом подходе, поэтому для нее справедливы все выводы, полученные для δ_j, с той разницей, что роль объекта выполняет таблица T_J, т.е. класс.

С учетом предыдущего

$$\delta^{*абс}_{j(J)} = K_j(J)/K^*. \qquad (29)$$

Очевидно, что δ_J или $\delta^{абс}_J$ указывает на номер класса, к которому относится проба. Индекс j в $K_j^{\max}(J)$ дает номер объекта, к которому проба тяготеет в данном классе.

Для того чтобы судить о принадлежности пробы к классу J, можно воспользоваться соотношением

$$\overline{\delta^*_J} = \frac{1}{m(J)} \sum_{j=1}^{m(J)} \frac{K_j(J)}{K^*}. \qquad (30)$$

О мере неоднородности объектов в классе по отношению к пробе можно судить по величине

$$(\delta^{абс}_J - \delta^*_J)/\delta^{*абс}_J. \qquad (31)$$

Указанные способы обработки результатов в значительной мере исчерпывают возможности использования исходных данных и результатов выдачи машинного счета по разработанным программам.

Совокупность тестовых апостериорных параметров, сведенных в табл. 5, поступает в распоряжение геолога, который интерпретирует количественные результаты решения в соответствии с целью и спецификой данной задачи.

ГЕОЛОГИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ РЕШЕНИЯ ЗАДАЧ

Введение в геологическую практику новых методов побуждает авторов сделать ряд замечаний о значении и характере применения предлагаемых методов. Ограниченный объем статьи не позволяет подробно рассмотреть этот вопрос, поэтому мы выделим для рассмотрения лишь несколько существенных моментов.

Геологическое описание того или иного объекта или явления можно считать моделью того объекта или явления в том смысле, в каком термин модель используется в естествознании. Однако одно важное свойство позволяет выделить его среди всех остальных моделей геологического объекта или явления. А именно, от других моделей описание отличается степенью присутствия в нем нашего понимания ненаблюдаемых непосредственно процессов или особенностей структуры, характеризующих конкретный описываемый объект или явление. Конечно, сам выбор описываемых свойств связан с объектом, например, через его общерегиональные характеристики, характеристики, относящие его к тому или иному типу месторождений и т.д. Однако все они относятся к классу наблюдаемых величин, а не являются следствием тех или иных теоретических построений, статистической обработкой данных и т.п. Именно это свойство описаний делает их основным материалом для предлагаемых в настоящей статье процедур. Указанное ограничение в выборе модели изучаемого объекта или явления проистекает из самой природы распознавания, имеющего дело непосредственно с наборами реально наблюдаемых величин, или, как принято говорить в теории распознавания, с изображениями. Для надежного распознавания и классификации изучаемого материала необходимо иметь дело именно с изображениями предметов и явлений, а не с теми или иными их моделями, построенными на основании общетеоретических соображений и интуиции интерпретации.

Следует отметить, что предлагаемый набор средств обработки геологической информации (модель) не является образцом (моделью) для составления тех или иных геологический описаний; эти модели относятся; первая – к описанию средств исследования, а вторая – к сфере моделирования объектов исследования.

Интерпретация – наиболее ответственный этап работы, так как от качества интерпретации зависит принятие решения, а следовательно, и направление, сроки, стоимость и эффективность поисково-съемочных и геолого-разведочных работ. Интерпретация есть связующее звено между всеми этапами сбора, учета, предалгоритмической и машинной обработки информации и принятием решения. Интерпретация существенно облегчается после машинной обработки информации и зависит главным образом от того, насколько наглядно будут представлены результаты вычислений.

Поэтому важнейшим условием принятия правильного решения является приведение результатов расчетов на ЭВМ в явную форму, т.е. в форму выразительную и привычную в геологической практике. Результаты вычислений на ЭВМ по разработанным шести программам выдаются на печать в виде последовательности цифр, выражающих основные величины из табл. 6.

Эти результаты относятся к трем основным параметрам: $P_{(i)}$, $I(S)$ и $I^*(S)$ (в таблице подчеркнуты), а все остальные параметры являются вспомогательными, производными от основных. Способы их интерпретации ясны из сказанного выше.

Таблица 6 (фрагмент)

Оценки признаков, характеризующих докембрийские металлоносные конгломераты

Номер признака	Содержательный смысл	Оценивающие данные: Учет признаков	Информационный вес $P_{(i)}$	Частота встречаемости	Ранг значимости
	Геологический возраст				
x_1	Докембрий	–[1]	–	–	–
x_2	Нижний протерозой	A[2]	0,354	0,65	III
x_3	Средний протерозой	A	0,447	0,82	II
x_4	Верхний протерозой	A	0,219	0,54	–
	Структура региона				
x_5	Второй структурный этаж	–	–	–	–
x_8	Платформенный чехол на протерозойском основании	A	0,323	0,59	III
	Характер рудовмещающей толщи				
x_9	Ритмично-слоистые отложения передовых прогибов с олигомиктовыми базальными толщами	A	0,473	0,87	II
	Рудоносная толща перекрывается платформенными отложениями	A	0,215	0,39	IV
	Тектоническое строение рудного поля				
	Крупные пологие синклинальные складки, осложненные складками второго порядка	A	0,214	0,39	IV
	Литология вмещающей толщи				
	На базальном горизонте залегают силлы и силлообразные дайки диабазов, порфиритов или эти породы секут рудную толщу в рудных полях	B[3]	0	0	VII
	Магматические проявления				
	Граниты только в подстилающей толще	A	0,528	0,97	I
	В структуре, вмещающей продуктивную толщу, известны кимберлитовые трубки или в перекрывающих отложениях выявлена алмазоносность	A	0,343	0,03	III

[1] – прочерк означает сквозные признаки или неопределенные.

[2] A – пространственно-временные поисковые признаки.

[3] B – вещественные поисковые признаки.

Информационный вес признака $P_{(i)}$ и способы его интерпретации

Величина функции $P_{(i)}$ выдается на печать в виде следующей записи: порядковый номер признака, его информационный вес. Система команд предусматривает два типа выдачи на печать:

а) последовательную выдачу номеров признаков и затем их информационного веса;

б) информационные веса упорядочены по убыванию (или возрастанию) затем идут номера признаков.

По существу, после расчетов на ЭВМ кодовая таблица признаков дополняется графами, оценивающими каждый из признаков.

Пример 5. Кодовая таблица и информационный вес признаков в обучающей выборке "Докембрийские металлоносные конгломераты" (Волков и др., 1968).

Однако табличная форма недостаточно наглядна. Значимость признаков более отчетливо проявляется на графиках упорядочивания признаков по величине P_i. На рис. 4 приведены кривые для примера 5, использованного в табл. 6. Из рис. 4 следует, что различающие признаки по величине $P_{(i)}$ подразделяются на ранги, обозначенные римскими цифрами.

Наиболее важными в этом примере оказались признаки x_{65}, x_9, x_{69} (рис. 4, *а*) и x_{33}, x_{37}, x_{25} (рис. 4, б), относимые к I рангу. Менее важными можно считать признаки II, III, IV рангов и так далее. Признаки x_{43}, x_{21}, x_{51}, x_{59} оказываются несущественными.

Содержательный смысл наиболее информативных признаков в расматриваемом примере следующий:

x_{65} – гранитоиды выявлены только в подстилающей толще;

x_9 – ритмично-слоистые отложения передового прогиба содержат олигомиктовые горизонты в базальных толщах;

x_{66} – основные породы развиваются сингенетично рудовмещающей толще и (или) после нее;

x_{33} – в основании толщи есть горизонт грубообломочных конгломератов;

x_{37} – в подстилающей толще известны основные эффузивы или интрузии;

x_{25} – в рудовмещающей толще встречаются прослойки карбонатных пород.

Как видим, максимальный информационный вес присущ признакам, описывающим крупные в региональном плане события, выявляемые наиболее дешевыми видами поисково-разведочных работ. Все эти признаки представляются существенными в обычных геологических исследованиях и согласуются с общепринятыми представлениями (Трофимчук и др., 1969; Нестеренко и др., 1969). Наименее существенными оказываются признаки, описывающие детали строения толщ, текстурные особенности пород и рудных тел. Иными сло-

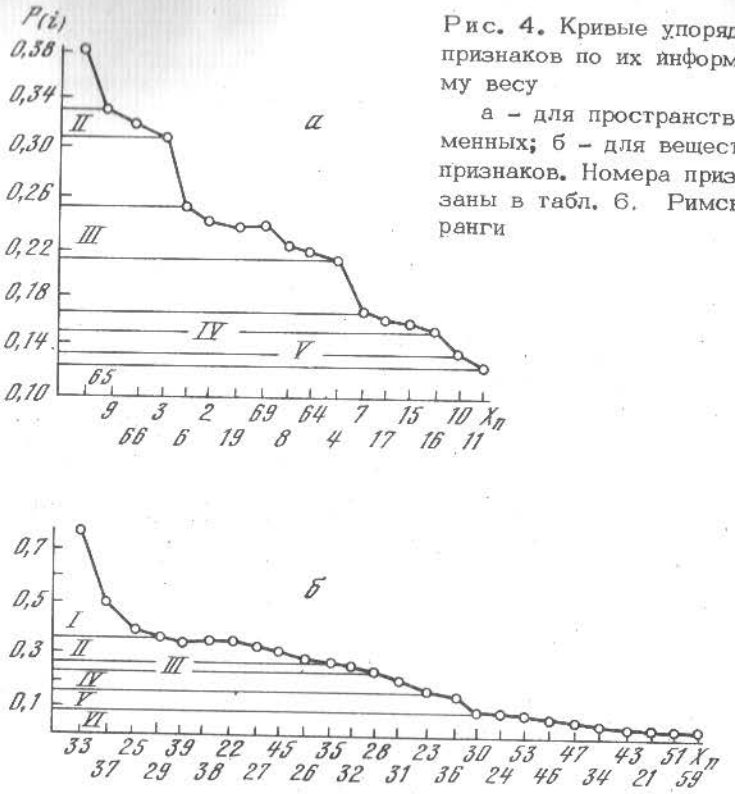


Рис. 4. Кривые упорядочивания признаков по их информационному весу

а – для пространственно-временных; б – для вещественных признаков. Номера признаков указаны в табл. 6. Римские цифры – ранги

вами, в процессе поисково-разведочных работ наиболее дорогостоящими являются признаки с наименьшим информационным весом.

Решение конкретных задач (Волков и др., 1968; Дмитриев, Золотухин и Васильев, 1968; Соловьев, 1968; Нестеренко и др., 1969; Модников и др., 1969; Кренделев, Дмитриев, 1969; Дмитриев, 1970) убеждает в том, что ранжирование признаков на группы является общей закономерностью описания геологических объектов. Такое ранжирование отражает степень изученности объектов.

Наиболее существенные для оценки параметров районов признаки выявляются на стадии геологических съемок, затем следуют признаки, выявляемые поисково-разведочными работами, и, наконец, признаки, выявляющиеся на стадии детальной разведки и эксплуатации, они несут наименьшую информационную нагрузку.

Поскольку признаки ранжируются, можно утверждать, что нельзя выделить какую-то одну главную причину рудообразования, раз-

работать универсальную классификацию, в основе которой лежит один признак или его модификация. Характер рудных концентраций определяется сочетанием признаков, входящих в высокие ранги на кривых упорядочивания признаков по величине $P_{(i)}$.

Информационный вес строки I (S) и способы интерпретации этой величины

Интерпретация величины $I(S)$ в общем подобна интерпретации $P_{(i)}$. На печать выдаются порядковый номер объекта (строки) и величина $I(S)$. В зависимости от цели и здесь можно на выдаче упорядочить объекты по номерам или величине $I(S)$. На основе цифровых данных строится график упорядочивания объектов по $I(S)$ (рис. 5, *а*).

Пример 5 (тот же, что и в предыдущем параграфе). Упорядочиваются по информационному весу семь известных в мире эталонных месторождений типа докембрийских конгломератов (рис. 5, *а*). Запасы руд известны только на двух из них – Витватерсранд (*1*) и Блайнд-Ривер (*2*). Однако ясно, что по запасам месторождения могут быть разделены на три группы.

1. Витватерсранд (*1*) с уникальными по масштабу запасами уран-золотоносных с торием руд;
2. Блайнд-Ривер (*2*), Тарква (*6*), Жакобина (*3*) – средние по запасам урановых (*2*) золотых (*6* и *3*) руд, с непромышленными содержаниями второго элемента.
3. Мунана (*4*), Австралия (*5*) и Эно-Коли (*7*) – мелкие месторождения урана в древних конгломератах.

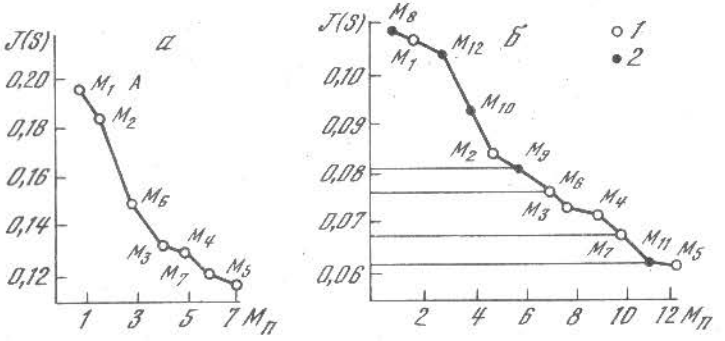


Рис. 5. Упорядочивание месторождений по информационному весу соответствующих им строк

а – на материале обучения; б – общая кривая эталонов (1) и проб (2). Номера объектов указаны в тексте

На рис. 5, *а*, построенном на материале обучения, месторождения четко разбиваются на те же три группы. Кривая отражает не только запасы месторождений, но и вещественный состав руд, поскольку для первой группы характерны три элемента: Au, U, Th; для второй – два из них; для третьей – только один.

Рис. 5 подтверждает правильность отнесения всех месторождений к единой формации и наличие взаимопереходов между сульфидными конгломератами и чисто магнетитовыми.

Строится график упорядочивания эталонных объектов и проб (рис. 5, *б*).

В качестве проб участвуют следующие регионы:

8 – Район X–0; *9* – Район X–1; *10* – X–2; *11* – X–3; *12* – X–4.

Для всех проб критерий общности выдерживается, но значение целевого предиката неизвестно. Упорядочивание по $I(S)$ (в рамках учтенной и обработанной информации) показывает, что наиболее перспективным оказывается район X–0.

При полевых исследованиях особое внимание обращалось на поиски признаков с высоким информационным весом и были найдены рудопроявления витватерсрандского типа, что подтверждает перспективность района и правильность метода.

Для контроля в пробы включен неизвестный объект X–1, в котором в рудных телах присутствуют и сульфиды, и магнетит. На рис. 5, *б* этот район попадает во второй ранг, т.е. в группу смешанных руд среднего размера, что соответствует истине.

Этот пример относится к качественной разбраковке месторождений на крупные классы без точной оценки масштабов.

Пример 6. Оценивается масштаб редкометального оруденения, локализованного в вулканических аппаратах (Модников и др., 1969). В таблицу обучения входило десять месторождений и шесть рудопроявлений с запасами, подсчитанными по категориям $A+Б+C$. Для краткости изложения целевой график запасов совмещен с кривой информационных весов для тех же объектов (рис. 6). Совершенно очевидно, что кривая информационных весов достаточно хорошо коррелирует с логарифмом запасов, что доказывает наличие зависимости между геологическими факторами, определяющими условия локализации оруденения и вулканических аппаратах, и его масштабами. Нетрудно вычислить коэффициент корреляции между логарифмом запасов и величиной $I(S)$.

Вычислив $I(S)$ для пробы, можно определить и запасы исследуемого объекта. В данном примере коэффициент корреляции очень высок (0,81–0,93).

Практика решения задач с достаточно точно определенными запасами показывает, что чем выше категория запасов, тем выше упомянутый коэффициент корреляции, и наоборот: для слабо изученных объектов (категории запасов C_1, $C_1 + C_2$, C_2) коэффициент корреляции снижается (до 0,63). И это вполне закономерно, так как степень изученности объектов выше – и, следовательно, количество прочерков в матрицах на объектах обучения меньшее.

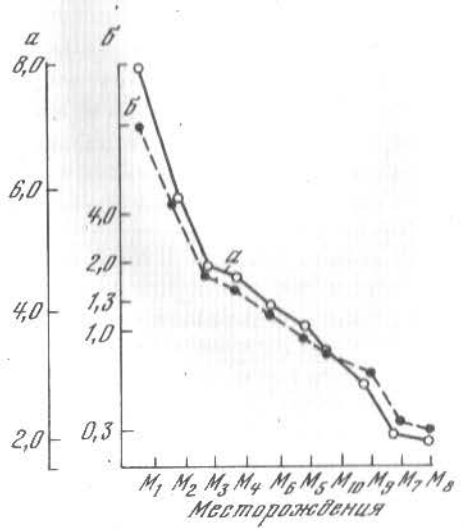


Рис. 6. Кривые упорядочивания объектов на материале обучения

а – по величине информационного веса строки; б – по логарифму запасов руд

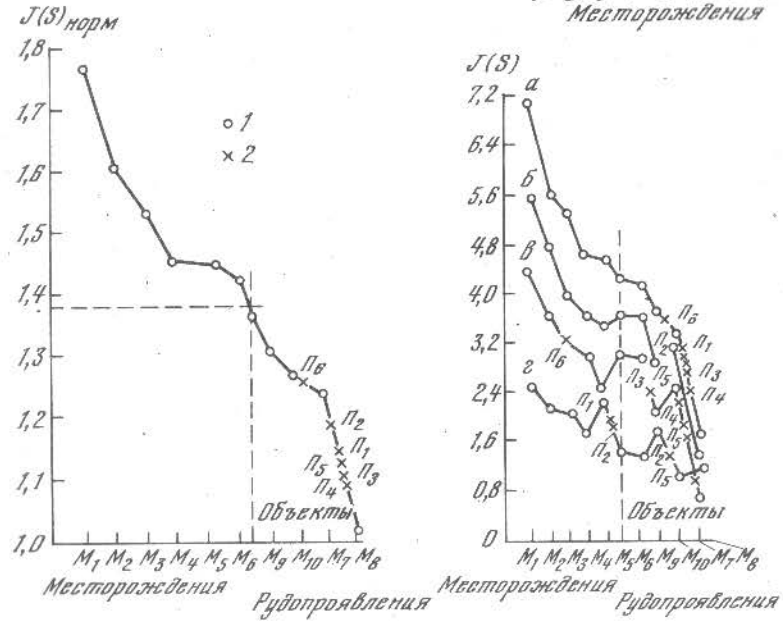


Рис. 7. Упорядочивание объектов, локализованных в вулканических аппаратах по величине $I(S)$

1 – объекты обучения; 2 – пробы

Рис. 8. Упорядочивание эталонов и проб по признакам разных групп

а – по совокупности всех значимых признаков; б – по значениям легко выявляемых признаков; в – по признакам, отражающим только региональные факторы (для объектов $П_3$, $П_4$ и $П_6$ величины $I(S) = 0,9$)

Упорядочивание эталонных объектов и исследуемых проб в данном примере (рис. 7) показывает отличное совпадение графиков и дает оценку изучаемым месторождениям. Как видно из рис. 7, объекты контроля попадают в разряд рудопроявлений. Для того чтобы эти рудопроявления перевести в разряд месторождений, необходимо для каждого из них выявить недостающие признаки с высоким информационным весом.

В данном примере производился анализ информативности региональных и локальных признаков (рис. 8), показывающий, что благоприятные региональные критерии являются необходимым, но недостаточным условием локализации оруденения в пределах региона. Только сочетание признаков всех групп, обладающих высоким информационным весом, делает прогноз надежным и не дает ошибок на материале обучения.

Диагностика классов объектов с разбиением таблицы на подтаблицы

Существуют задачи, в которых объекты разнесены на близкие между собой классы, не имеющие резких различий по целевому предикату. Кроме того, каждый класс объектов изучен с разной степенью детальности и характеризуется признаками, исследуемыми разными методами. Требуется дать сравнительную оценку объектов и классов, выразить меру сходства и различия между ними в количественном виде. Имеется объект – проба, которую по некоторому неполному перечню признаков следует отнести к какому-либо из уже выделенных классов.

В таком виде задача допускает разбиение общей таблицы на подтаблицы либо по классам объектов, либо по группам признаков, либо по обоим показателям одновременно.

Пример 7. В одном из районов Забайкалья собрана информация по 38 россыпям, которые подразделяются по крупности на четыре класса: а), б), в), г) рудопроявления. Признаки (их 181) качественно делятся на четыре группы: I – пространственно-временные; II – геологические; III – геоморфологические; IV – вещественные. Таким образом, получена общая таблица охарактеризованности $T(m \times n)$, где $m = 38$, $n = 181$. Разбиение исходной таблицы на подтаблицы показано в табл. 7. Индексы Ia, $Iб$,..., $IVв$, $IVг$ соответствуют номерам 16 подтаблиц, охарактеризованных в бинарных символах с алфавитом {0, 1}.

Вначале для каждой подтаблицы получаются информационные веса $P_{(i)}$ характеристических признаков (столбцов). При их упорядочивании по $P_{(i)}$ в пределах каждой подтаблицы выявляется разбиение признаков на ранги (рис. 9). Вычисляются также различающие веса признаков $R_{(i)}$, описывающие существенность каждого признака при обнаружении различия между классами по заданным порогам запасов. Суммарное значение различающих весов для всех четырех групп признаков может быть использовано как показатель расстояния между

Таблица 7

Классификация признаков золотоносных рассыпей Забайкалья

Классы россыпей	Группы признаков				Число строк (объектов)
	пространственно-временны́е	геологические	геоморфологические	вещественные	
	I	II	III	IV	
(а)	Ia	IIa	$IIIa$	IVa	8
(б)	$Iб$	$IIб$	$IIIб$	$IVб$	9
(в)	$Iв$	$IIв$	$IIIв$	$IVв$	11
Рудопроявления (г)	$Iг$	$IIг$	$IIIг$	$IVг$	10
Число столбцов	43	45	40	53	38 181

классами. В данной задаче межклассовое сравнение обнаружило (по сумме $R_{(i)}$ всех групп признаков) различные "расстояния по запасам" между классами. В частности, расстояние (коэффициент различия между классами; Кренделев, Дмитриев, 1969) между классами $а$ и $г$ в 2,5 раза больше, чем расстояние между классами $б$ и $в$. Иначе говоря, различие между классами $б$ и $в$ по вычисленным параметрам оказалось несущественным, и в действительности граница между этими классами условна, так как объекты, разграничивающие классы, имеют слабое различие.

Следует подчеркнуть, что вычисленные расстояния тесно коррелируют с разницей классов в запасах. Этот факт хорошо иллюстрирует кривые рис. 10.

Для того чтобы выяснить, с какой точностью отдельные группы признаков упорядочивают объекты по запасам, на рис. 11 приведены суммарные кривые $\overline{I(S)}$ для каждой из групп признаков. Ради большей компактности результатов существенность групп признаков в ранжировке россыпей по запасам оценена коэффициентами ранговой корреляции (ρ – коэффициент Спирмена). Эти показатели (табл. 8) подтверждают качественную картину ранжировки классов (рис. 11). Из этого рисунка видно, что наилучшая корреляция имеется между запасами объектов данного класса и суммарным информационным весом строк $\overline{I(S)}$. Следовательно, чтобы спрогнозировать запасы россыпи, относящейся к этому классу, необходима полная охарактеризованность объекта. Это соответствует требованию о всесторонней логической охарактеризованности объектов для исследования данным методом.

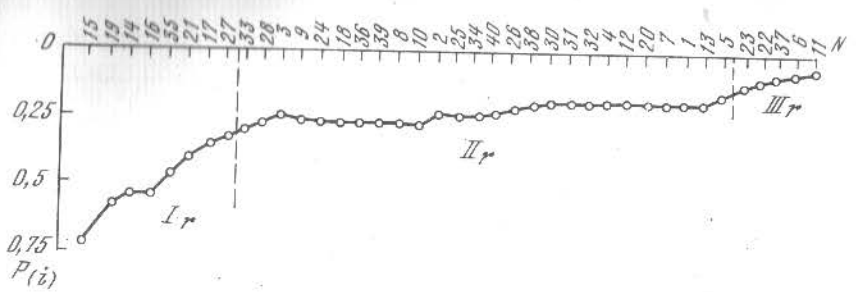


Рис. 9. Кривая упорядочивания и ранжировки признаков III группы (геоморфологические) в классе а по величине P_i

15 – продольный уклон долины меньше (больше) 0,020; 19 – длина долины больше (меньше) 10 км; 14 – приуроченность россыпей к долинам со ср. (0,005–0,020) продольн. уклон.; 16 – приуроченность россыпей к долинам (не) I–II порядков; 35 – то же с правой асимметрией; 21 – то же от 30 до 90 км; 17 – то же к долинам III–IV порядков; 27 – продольный уклон долины с россыпью близок к среднему уклону; 33 – асимметричность долин; 28 – то же аномально пологий; 3 – приуроченность россыпей к долинам Ю3–СВ простирания; 9 – то же с Ю3 течением; 24 – то же близким по длине к средней; 18 – то же V–VI порядков; 36 – густота речной сети не очень высокая (к 0,7); 39 – асимметричность речных бассейнов; 8 – приуроченность россыпей к долинам с Ю3 течением; 10 – то же с ЮВ течением; 2 – то же к субширотным долинам; 25 – то же к аномально коротким долинам; 34 – принимающая долина и долина с россыпью близкого порядка; 40 – приуроченность золотоносной долины к бассейну с правой асимметрией; 26 – приуроченность россыпей к аномально крутым долинам; 38 – густота речной сети средняя (0,5–0,7); 30 – приуроченность россыпей к долинам со средним (150–250 м) врезом; 31 – то же к существенно (более 250 м) врезанным долинам; 32 – то же к глубоко (более 400 м) врезанным долинам; 4 – то же к долинам СЗ–ЮВ простирания; 12 – то же к долинам с СВ течением; 21 – то же к ср. (10–30 км) по длине долинам; 7 – приуроченность к долинам с восточным течением; 1 – приуроченность россыпей к субмеридиональным долинам; 13 – продольный уклон долины менее 0,005; 5 – приуроченность к долинам с южным течением; 20 – приуроченность к аномально длинным долинам; 22 – то же к очень длинным долинам; 37 – густота речной сети низкая; 6 – приуроченность к долинам с северным течением; 11 – приуроченность к долинам с СВ течением

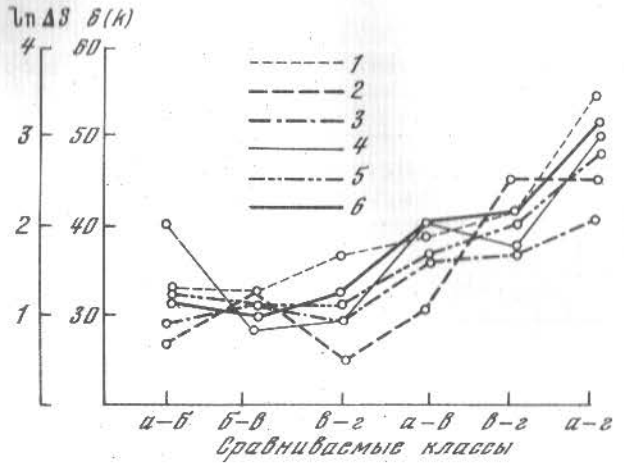


Рис. 10. Кривые зависимости разности запасов россыпей разных классов с коэффициентом различия между классами

1–4 – информационные веса строк соответствующих групп признаков; 5 – суммирующие для всех групп признаков; 6 – логарифм запасов

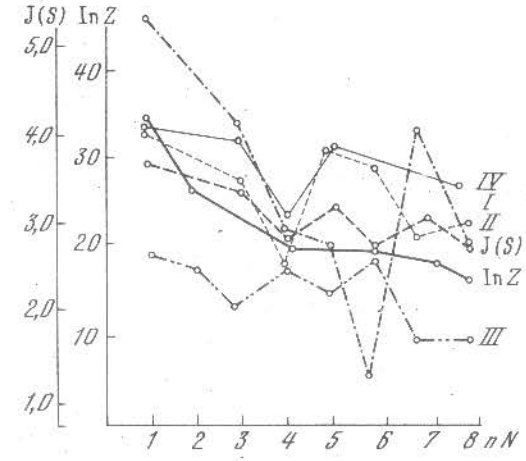


Рис. 11. Кривые упорядочивания россыпей по суммарным кривым $I(S)$ для каждой группы признаков и по логарифму запасов (класс а)

Римские цифры – группы признаков, N – объекты

Таблица 8

Значение различных групп признаков упорядочения золотоносных россыпей по запасам

S_j Порядок россыпей класса a по запасам	Упорядочение объектов по $I(S)$ в группах				Упорядочение по средней
	1	2	3	4	
1	5	1	6	1	1
2	6	2	1	3	2
3	3	3	4	5	3
4	2	7	2	2	5
5	6	4	5	6	7
6	7	5	3	7	4
7	1	8	8	8	8
8	4	6	7	4	8
Коэффициент Спирмена ρ	−0,480	+0,755	+0,357	+0,573	+0,847

Характерно, что, как и при межклассовом сравнении, внутри класса a обнаруживается та же последовательность существенности групп признаков при упорядочивании объектов по запасам. Неодинаковая способность различных групп признаков ранжировать объекты по запасам имеет большое значение в построении диагностических схем. В частности, для объектов обследованного района наиболее существенными группами признаков являются геологические и вещественные. Геоморфологическая группа признаков слабо упорядочивает объекты по запасам. Это явление может быть связано не только с внутренней природой признаков слабодиагностирующей группы, но и с их недостаточно строгой (в профессиональном смысле) подборкой.

Таким образом, использование межклассовых и внутриклассовых тестовых параметров позволяет ставить и решать вопросы прямого производственного и научного характера.

Некоторые рекомендации для интерпретации результатов решения

Автоматическая обработка данных не заменяет геолога, а только подготавливает данные и приводит их к виду, удобному для использования на этапе принятия решений. Очевидно, что после обработки информации на ЭВМ задача интерпретации значительно облегчается. Эффективность интерпретации также возрастает по мере совершенствования навыков геолога в информационной работе.

Сложность и разнообразие задач, возникающих в производственной и исследовательской работе геологов, вызывают необходимость в формулировании указаний, организующих логику интерпретации. Приведем наиболее общие указания.

1. Необходимо быть объективными при выборе данных и не пренебрегать сведениями, которые не соответствуют взгляду интерпретатора на исследуемый объект или явление. Избирательность возможна, но опасность ошибки уменьшается, если все расчетные данные доведены до явной формы, т.е. в таблицу или на график вынесены все расчетные величины, полученные в ходе решения задачи.

2. Необходимо избегать способов, посредством которых искусственно совмещают расчетные данные с результатами, ожидаемыми согласно целеуказанию. Независимость рекомендаций от ожидаемых результатов иногда помогает обнаружить факты, непредусмотренные при постановке задачи. Не исключено, что полученные данные дадут основания для изменения формулировки задачи и ревизии целеуказания.

3. Интерпретации подвергаются только те сведения, которые включались в исходную таблицу. Привлечение внешней информации недопустимо. Если появляются неучтенные при постановке задачи признаки, их необходимо включить в исходную таблицу, заново произвести все расчеты и только после этого перейти к интерпретации. При этом возможны два случая:

а) новые сообщения подтверждают полученные результаты решения;

б) новые сообщения противоречат полученным результатам решения.

В этом случае необходимо пересмотреть не только пространство признаков, но и постановку задачи.

4. Необходимо полно и четко выражать результаты решения в явной форме. Такими формами могут быть итоговые таблицы численных результатов решения с указанием тех величин, которые составляют базу для выводов и принятия решения. Изображение количественных результатов, помимо табличной формы, можно представлять в виде графиков, суммирующих кривых, гистограмм и в других приемлемых для данного случая формах.

5. Следует соблюдать полноту (по индукции) и последовательность использования результатов решения с тем, чтобы не производить избирательной интерпретации нужного места задачи. В случае неполного и непоследовательного анализа результатов решения производится разрезание полученных результатов без учета информационного центра тяжести. Необходимый для интерпретатора участок результатов подвергается более детальному обследованию после выяснения общей задачи.

После окончания интерпретации наступает заключительный этап в решении задачи, состоящей в том, чтобы сформулировать и обобщить данные для принятия практического решения, для реальной гео-

логической деятельности. На этом этапе основное заключение делается геологом, который по своему усмотрению может привлекать дополнительную информацию для согласования с ней основных результатов решения. Проводятся дополнительные операции по выбору способов экспериментальной проверки, намечаются первоочередные задачи для деятельности в области съемочных, поисковых, разведочных работ.

ВЫВОДЫ

1. Обработка геологической информации дискретными методами делится на следующие основные этапы: подготовка геологической информации к обработке на ЭВМ (предалгоритмический этап) обработка информации на ЭВМ с помощью одного из рассмотренных в статье алгоритмов (тестов, тесторов, голосования по тестам, голосования по тесторам), геологической интерпретации результатов решения задач.

2. Подготовка геологической информации заключается в постановке задачи, выборе объекта исследования, кодировании признаков, представлений информации в табличной форме. В результате первого этапа работы должны быть получены график или таблица значений целевого признака, кодовое описание признаков, т.е. перечень номеров признаков и их содержательная (или количественная) характеристика, исходная таблица описания изучаемых объектов.

3. Обработка геологической информации на ЭВМ заключается в использовании одного из алгоритмов, определения информационных весов признаков с помощью нахождения тестов, тесторов, голосования по тестам или голосования по тесторам. Выбор алгоритма зависит от того, в каком виде представлена исходная геологическая информация: от количества исходных таблиц, от числа эталонных объектов в каждом классе, от неоднородности объектов. В результате обработки информации на ЭВМ в распоряжение геолога поступают набор параметров, имеющих определенное содержательное значение. К их числу относятся информационные веса признаков, различающие тесторные веса признаков, нормированные различающие веса признаков, информационные веса строк, относительная существенность каждой пары строк, взвешенные информационные веса строк, относительные различия между парами строк по целевому признаку, мера отличия полученного распределения объектов от их распределения по целевому признаку, информационные веса контрольных объектов (проб) и меры их принадлежности к классу и к конкретным объектам.

4. Геологическая интерпретация результатов решения задачи на ЭВМ заключается в правильном содержательном анализе полученных параметров, их представлении в форме, удобной для геологической практики. При этом важным является ранжирование признаков по их информационным весам на группы, отражающие степень изученности геологических объектов.

5. Проведенные исследования позволяют утверждать, что нельзя выделить какую-то одну главную причину рудообразования, разработать универсальную классификацию, в основе которой лежит один признак или его модификация. Характер рудных концентраций определяется сочетанием признаков, имеющих наибольшие информационные веса.

ЛИТЕРАТУРА

Бугаец А.Н., Дворниченко Г.К., Мацак А.П., Серова Л.Л. Алгоритмы и программы решения геологических задач на ЭВМ "Минск-2", и "БЭСМ-3М". Алма-Ата, КазНИИ минерального сырья, вып. 2, 1969.

Волков П.П., Дмитриев А.Н., Нагаева Г.М., Пантелеев В.И. Дифференциальная диагностика шизофрении и органических заболеваний мозга логико-дискретным методом. – В сб.: "Проблемы моделирования психической деятельности". Новосибирск, "Наука", 1968.

Вышемирский В.С., Дмитриев А.Н., Трофимук А.А. В сб.: "Геология и математика". Новосибирск, "Наука", 1967.

Дмитриев А.Н. Некоторые табличные числа. – В сб.: "Дискретный анализ", вып. 12. Новосибирск, "Наука", 1968.

Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П. О математических принципах классификации предметов и явлений. – В сб.: "Дискретный анализ", вып. 7. Новосибирск, "Наука", 1966.

Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П. Об одном принципе классификации и прогноза геологических объектов и явлений. – Геология и геофизика, 1968, № 5.

Дмитриев А.Н., Васильев Ю.Р., Золотухин В.В. Логико-математическая обработка информации при выявлении перспективности сульфидного оруденения в некоторых трапповых интрузиях Севера Сибирской платформы. – Геология и геофизика, 1968, № 7.

Дмитриев А.Н., Золотухин В.В., Васильев Ю.Р. Опыт применения дискретной математической обработки информации по дифференцированным трапповым интрузиям Северо-Запада Сибирской платформы. – Сов. геология, 1968, № 12.

Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П. Логический способ построения многомерных классификаций в геологии. – В сб.: "Применение математических методов в геологии". Алма-Ата, Изд-во АН КазССР, 1968.

Дмитриев А.Н., Смертин Е.А. Связь тестовых параметров таблиц с повторяемостью столбцов. Всесоюзная конференция по проблемам теоретической кибернетики. Тезисы докладов. Новосибирск, "Наука", 1969.

Дмитриев А.Н. Использование длин тупиковых тестов при обработке таблиц. – В сб.: "Дискретный анализ", вып. 17, Новосибирск, "Наука", 1970.

Дмитриев А.Н., Смертин Е.А. Алгоритмы вычисления тестовых параметров бинарных таблиц в задачах распознавания. – В сб.: "Алгоритмы и программы решения геологических задач", вып. 3. Алма-Ата, Изд-во АН КазССР, 1970.

Константинов Р.М., Дмитриев А.Н. Использование математических методов для анализа геологических факторов, влияющих на масштабы оруденения (на примере месторождений касситерит-сульфидной формации). – Геология рудн. месторождений, 1970, № 2.

Кренделев Ф.П., Дмитриев А.Н., Журавлев Ю.И. Сравнение геологического строения зарубежных месторождений докембрийского конгломерата с помощью дискретной математики. – ДАН СССР, 1967, т. 173, № 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. – В кн.: "Проблема металлоносности древних конгломератов на территории СССР". М., "Наука", 1969.

Модников И.С., Еремеев А.Н., Писаревский В.И. и др. Оценка масштаба редкометального оруденения, локализованного и вулканических аппаратах (с помощью ЭВМ). – Сов. геология, 1969, № 11.

Нестеренко Г.В., Дмитриев А.Н., Кренделев Ф.П. и др. Проблемы геологии россыпей (на примере Восточного Забайкалья). – "Труды II–III Всесоюзного совещения по геологии россыпей". Магадан, 1969.

Слуцкая Т.Л. Алгоритмы вычисления информационных весов признаков. – В сб.: "Дискретный анализ", вып. 12. Новосибирск, "Наука", 1968.

Смертин Е.А., Дмитриев А.Н. Дополнение к алгоритму распознавания "голосованием" по тестам и тесторам. – В сб.: "Алгоритмы и программы", вып. 3. Алма-Ата, Изд-во АН КазССР, 1970.

Соловьев Н.А. Об одном свойстве таблиц с тупиковыми тестами одинаковой длины. – В сб.: "Дискретный анализ", вып. 12. Новосибирск, "Наука", 1968.

Трофимчук А.А., Вышемирский В.С., Дмитриев А.Н. и др. О сравнительном изучении гигантских месторождений нефти с использованием логико-дискретного анализа. – Геология нефти и газа, 1969, № 6.

Чегис И.А., Яблонский С.В. Логические способы контроля работы электрических схем. – Труды Мат. ин-та им. В.А. Стеклова", т. 51. М., Изд-во АН СССР, 1958.