УДК 550.34.012

МЕТОДИКА ПОСТАНОВКИ И РЕШЕНИЯ СЛАБОФОРМАЛИЗОВАННЫХ ГЕОЛОГИЧЕСКИХ ЗАДАЧ

Черемисина Евгения Наумовна[1], Костылева Татьяна Владимировна[2], Кирпичева Елена Юрьевна[3]

[1]Доктор технических наук, профессор, заведующий отделением;
ФГБУ «Всероссийский научно-исследовательский геологический нефтяной институт»;
Россия, 105118, Москва, Шоссе Энтузиастов, 36;
e-mail: e.cheremisina@geosys.ru.

[2]Начальник участка;
ФГБУ «Всероссийский научно-исследовательский геологический нефтяной институт»;
Россия, 105118, Москва, Шоссе Энтузиастов, 36;
e-mail: e.cheremisina@geosys.ru.

[3]Кандидат технических наук, доцент;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: kirphel@uni-dubna.ru.

Работа посвящена описанию методики постановки и решения геологических задач, в том числе слабо формализованных. Подробному рассмотрению основных этапов процесса постановки и решения задач. Отображено описание анализа данных, классификации формализованных задач по эталонам, или без них. Изложены постановки задач упорядочения и минимизации.

Ключевые слова: формулирование геологической задачи, формализация задач, выбор способа решения задач, формализация модели, поисковые признаки, алгоритм К-средних, Голотипные алгоритмы, алгоритм «Иерархическая таксономия».

Для цитирования:
Черемисина Е. Н., Костылева Т. В., Кирпичева Е. Ю. Методика постановки и решения слабоформализованных геологических задач // Системный анализ в науке и образовании: сетевое научное издание. 2024. № 1. С 1- 3. EDN : TKEVKO. URL : https://sanse.ru/index.php/sanse/article/view/602.

METHODOLOGY FOR SETTING AND SOLVING UNDERFORMALIZED GEOLOGICAL PROBLEMS

Cheremisina Evgeniya N.[1], Kostyleva Tatiana V.[2], Kirpicheva Elena Yu.[3]

[1]Grand PhD in Technical Sciences, professor, head of department;
FSBI "All-Russian Research Geological Petroleum Institute";
36 Entuziastov Highway, Moscow, 105118, Russia;
e-mail: e.cheremisina@geosys.ru.

[2]Head of the section;
FSBI "All-Russian Research Geological Petroleum Institute", Department of Geoinformatics;
36 Entuziastov Highway, Moscow, 105118, Russia;
e-mail: *tkostyleva@geosys.ru.*

[3]PhD in Technical Sciences, associate professor;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: kirphel@uni-dubna.ru.

The paper describes the methodology for setting and solving geological problems, including underformalized ones, and discusses in detail the main stages of the process of setting and solving problems. A description of data analysis, classification of formalized tasks according to standards, or without them, is displayed. Statements of ordering and minimization problems are presented.

Keywords: formulation of a geological problem, formalization of problems, choice of method for solving problems, formalization of a model, search features, K-means algorithm, Holotype algorithms, "Hierarchical taxonomy" algorithm.

For citation:
Cheremisina E., Kostyleva T., Kirpicheva E. Methodology for setting and solving underformalized geological problems. *System analysis in science and education*, 2024;(1):1-13 (In Russ). EDN : TKEVKO. Available from: https://sanse.ru/index.php/sanse/article/view/602.

Введение

Все науки делятся на *описательные*, которые описывают некоторое явления, законы и следствия, и *предсказательные*, которые на основании выявленных закономерностей предсказывают те или иные явления или объекты. Каждая наука в своем историческом развитии движется от статуса описательной к статусу предсказательной [1].

Все предсказательные науки изучают ***модели*** природы и общества, а не изучают саму природу или общество. Точность предсказаний зависит от адекватности и точности представленной модели. Если модель объекта исследования охватывает достаточно обширный круг процессов и явлений, рассматриваемых данной наукой, наука называется *формализованной*. В современной физике фигурируют 2-3 модели, описывающие явления и процессы, поэтому и является примером формализованной науки. Если модель должна создаваться отдельно для каждого объекта, явления или процесса, то такая наука называется *слабоформализованной*. К таким наукам, в частности, относится геология, экология, медицина и другие.

В формализованной науке для решения задачи необходимо собрать информацию, подобрать подходящую модель и перевести проблему на язык математики с дальнейшим решением в этой области. В слабоформализованных науках получить решение задачи, как правило, намного сложнее, так как не существует модели рассматриваемого объекта или явления, поэтому для решения таких задач необходимо разработать схему *постановки и решения*.

1. Схема постановки и решения задач

Схема постановки и решения задачи разбивается на этапы формулирования, формализации, выбора способа решения, решения, анализа и интерпретации результатов (см. рис. 1).

Формулирование геологической задачи включает указание цели, представлений о геологической (генетической) и выводимой из нее геолого-прогнозной модели объекта исследований, исходных данных, результата и формы его представления, требований к результату.

Формализация задачи состоит в переводе на формальный язык описания цели и объектов исследования, типизации и формализации прогнозной модели с указанием способов вычисления свойств, формировании цифровой модели данных для решения задачи, формализации требований к результату, проверке согласованности требуемого результата с целью.

Процесс выбора способа решения задачи включает все этапы анализа данных (анализ с целью их коррекции и оценки качества, анализ шкал, распределений, взаимосвязи и преобразование данных, анализ с целью выбора метода решения) и корректировки информации, а также определение и выполнение алгоритма решения задачи, обеспечивающего получение требуемого результата.

Решение задачи в соответствии с выбранным ранее способом.

Рис. 1. Схема постановки и решения геологических задач

Формальный анализ и интерпретация результатов состоят в проверке согласования результатов с целью исследования, сформулированными требованиями к результату и принятием решения об использовании результатов либо об уточнении модельных представлений и формулировки задачи [3].

Формулирование геологической задачи. Предметную, в частности, Геологическую задачу формулирует специалист-геолог (геофизик, геохимик) на языке предметной области. Формулирование геологической задачи включает указание цели, представлений о геологической модели объекта исследований (поиска), исходных данных, результата и формы его представления, требований к результату, под которыми понимается следующее.

Цель – предвосхищение в мышлении результата деятельности. Нет смысла говорить о наличии задачи, если четко не сформулирована цель исследования.

2. Моделирование объектов

Модель – образ (в том числе условный или мысленный) какого-либо объекта или системы объектов, используемый в определенных условиях в качестве их «заместителя». В естественных науках модель какой-либо системы – её описание на языке некоторой научной теории. Модель является упрощенным образом оригинала. Моделирование может быть как материальным, так и идеальным. Идеальное моделирование может происходить как на уровне самых общих не до конца фиксированных представлений, так и на уровне достаточно детализированных знаковых систем.

Предварительным этапом создания геологической модели является формулирование (выбор) закономерностей формирования геологического объекта. Затем модель соотносится с исходными данными, имеющимися у исследователя, для составления геолого-прогнозной модели с определением соответствующего ей пространства критериев и признаков.

Важным элементом модели является определение объекта поиска; этот объект должен быть выбран в соответствии с целью и масштабом исследования, и определен в терминах исходных данных. Объекты должны быть описаны свойствами, которые отвечают модельным представлениям и масштабу, и могут быть получены из исходных материалов. Объекты и области поиска должны быть взаимосвязаны (в рудных провинциях следует выделять рудные районы, в рудных районах – рудные узлы, и т.п.).

На практике геологическая модель прогнозного объекта формулируется лишь на уровне *представлений о модели* разной степени детальности, связанных с определением закономерностей размещения

полезных ископаемых, под которыми понимаются устойчивые, статистически доказанные связи месторождений с определенными геологическими образованиями, природными процессами, геофизическими полями, геоморфологическими и другими характеристиками территории. Эти различные факторы и поисковые признаки, а также их совокупности и производные, связь месторождений с которыми может быть пространственной, временной, генетической [2].

Факторы характеризуют геологические тела и структуры, создавшие их процессы, время и обстановки, являющиеся причинами или условиями образования, размещения и сохранения месторождений.

Региональные факторы определяют объект предыдущего масштаба и должны ограничивать область поиска в масштабе исследований.

Локальные факторы определяют положение, морфологию и ресурсы объекта прогноза.

Поисковые признаки – геологические образования и характеристики территории, являющиеся следствием процессов, сопутствующих формированию, изменениям и разрушению концентраций полезных минералов, и указывающие на возможное наличие объекта в определенном месте. Они соответствуют или предположительно отвечают отдельным объектам либо телам более крупного масштаба (например, месторождение).

К прямым признакам следует относить непосредственные находки прогнозируемого полезного ископаемого.

Исходные данные. Основным исходным материалом являются:

- Государственные геологические карты, карты полезных ископаемых и закономерностей их размещения, а также специальные минерагенические карты.
- Геофизическая, геохимическая и дистанционная основы геологической карты соответствующего масштаба.
- Сейсмические профили и кубы.
- Скважинные данные.
- Структурные карты.
- Представления о моделях промышленных объектов прогнозируемого полезного ископаемого.

Дополнительно могут быть использованы различные карты, результаты геохимической, геофизической съемок, анализов. Исходная информация может быть представлена в самой различной форме: карты, графики, результаты тех или иных анализов, экспертные оценки и т.п. Таким образом, форма представления исходной информации не имеет существенного значения. Однако очень важна связь между данными и модельными представлениями: модель должна быть сформулирована в терминах данных. Должно быть указано, каким образом данные связаны с моделью (факторами и признаками), а через модель – с целью.

Наиболее уязвимым местом карт закономерностей размещения полезных ископаемых, создаваемых на основе минерагенических факторов и поисковых признаков, является слабая увязка слоев изображения между собой.

При моделировании необходимо избегать большого количества факторов, создающих шум при механическом наложении. Каждый этап вполне достаточно характеризуется одним – двумя факторами, которые тесно логически увязываются между собой и создают общий образ прогнозного объекта.

Таким образом, важнейшей на первом этапе является задача выбора генетической модели наиболее отвечающей прогнозируемому типу формирования месторождения.

Важно подчеркнуть, что при наличии альтернативных генетических моделей для разработки прогнозно-поисковой модели факторы рассматриваются и анализируются в строгом соответствии с одной из них. Сочетания факторов разного идеологического содержания при создании прогнозно-поисковой модели недопустимо [4].

Результат должен быть конкретным и соответствовать цели. В нем формулируется утверждение об объектах поиска.

Требования к результату должны вытекать из модельных представлений. Они могут относиться к более крупным областям, чем объекты поиска.

В качестве наиболее часто применяемых требований выступают прямые поисковые признаки, не участвующие в решении задачи, а служащие для подтверждения и оценки полученных результатов.

От критерия оценки результата зависит выбор подхода к решению задачи. Отсутствие критерия оценки результата, приводит к случайному выбору метода решения задачи и к последующей не интерпретируемости результата. Таким образом, требования к результату должны вытекать из модельных представлений и должны быть верифицируемы, т.е. каждый конкретный результат может быть проверен на соответствие требованиям. При отсутствии сформулированных требовании к результату задача может быть решена. Однако в этом случае качество результата оценить невозможно, в частности, у содержательной задачи может оказаться множество решений.

Заметим, что цель, исходные данные и представления о результате обычно имеются у геолога, решающего прогнозную задачу. Как уже указывалось выше, без требований к результату задача может быть решена, хотя и не очень хорошо. Но решить задачу без модельных представлений об объекте поиска невозможно и формулирование их является наиболее сложным пунктом для геолога-постановщика.

3. Формализация задачи

Формализация задачи состоит в переводе на формальный язык описания цели и объекта исследования, типизации и формализации модельных представлений, определении способов вычисления свойств и построении цифровой модели (матрицы) данных для решения задачи, формировании требований к результату.

Формализация цели. Как следует из опыта автоматизированного решения задач прогноза, основными формальными задачами, решаемыми на отдельных этапах и стадиях геологоразведочного процесса по комплексу геологических, геофизических, геохимических характеристик, являются следующие:

- *выделение* объектов поиска в области поиска (районирование территорий);
- *разделение* объектов на перспективные и неперспективные по степени их похожести на эталонные объекты (разбраковка территории);
- *упорядочение* (ранжирование) всех или наиболее перспективных объектов по значимости с целью определения очередности их дальнейшего изучения (оценки).
- Кроме того, часто решаются вспомогательные задачи:
- сокращение признакового пространства, то есть избавление от признаков, которые не имеют серьезного значения для решаемой основной задачи;
- переописание, то есть описание объектов более крупных чем первоначальные при изменении формулировки задачи.

В качестве *объектов прогноза* в зависимости от целей и масштабов работ могут выступать различные по рангу продуктивные объекты. Описание информации, приходящей в виде растровых данных дистанционного зондирования, векторной картографической информации, числовых геофизических полей и относящейся к различным объектам привязки, должно приводиться к единому объекту исследования. Естественно разделить территорию на площадки (видимо, квадратной формы), размеры которых отвечают минимальному размеру объекта прогноза, а затем постараться привязать всю имеющуюся у нас информацию к этим площадкам, то есть для данной конкретной площадки некоторые конкретные значения должны быть получены по всем слоям информации. Ясно, что выбранный *объект привязки информации* должен отвечать цели исследования, а при возможности и о наличии полезного ископаемого, его запасах и т.п. Таким образом, максимальный размер прогнозной ячейки ограничен принципом непропуска объекта прогноза, то есть не должен превышать половины поперечника объекта прогноза, а минимальный - точностью проведения границ на исходной карте, то есть не должен быть меньше 1 мм в масштабе карты. Если эти условия несовместны, это значит, что данная задача не может быть решена при выбранном масштабе исследований. Обычно ячейка имеет размеры 2 – 10 мм в масштабе карты.

Целесообразно разбить процесс решения на этапы, а это в свою очередь приводит к распадению реализации цели на цепочку формальных задач. Например, если задача сформулирована как задача

прогноза площадей, перспективных на продуктивные объекты определенного типа, то естественно разбить ее на следующие подзадачи:

- ограничение области поиска,
- выделение в ней перспективных участков,
- разделение перспективных участков по ожидаемым типам.

Формализация модели. Формализация модели состоит в определении соответствия геолого-прогнозной модели исходным данным, описании критериев и признаков для решения задачи и получения их цифровой модели (матрицы), а также выборе способов их расчета. При этом можно выделить несколько типов формальных моделей.

1. Критериальная модель задается группой критериев, наличие каждого из которых благоприятствует (реже препятствует) достижению на данном объекте максимума целевой функции. Под целевой функцией здесь, например, можно понимать наличие и ожидаемые запасы полезного ископаемого на исследуемой территории и т.п. Наличие такой модели знаменует либо достаточную теоретическую проработанность вопроса об объекте исследований, либо большую статистику, позволяющую сделать такие выводы.

В качестве критериев выступают региональные и локальные минерагенические факторы. Критерии можно подразделить на:

- условия, ограничивающие область поиска;
- критерии, характеризующие совокупности объектов поиска;
- критерии, характеризующие объекты поиска (большинство из которых в дальнейшем составляет основу для формирования пространства свойств для решения задачи).

2. Аналоговая модель, в которых постулируется принцип аналогий и представлено достаточное количество объектов (эталонов), отвечающих различным классам. Здесь особенно важно иметь эталонных представителей не только для оптимальных, а для всех классов объектов. Например, должны быть известны не только месторождения, но и пустые участки территории.

При выборе эталонных объектов (материала обучения) необходимо учитывать следующее:

- в качестве эталонов могут выступать только экспериментально изученные (разбуренные) объекты;
- эталонные объекты должны быть одного ранга и соответствовать рангу объектов прогноза (при выделении районов нельзя использовать в качестве эталонов известные месторождения);
- представительность эталонов определяется их разнообразием (с точки зрения присущих особенностей), а не количеством однотипных;
- в ситуациях, когда задаются эталоны разных классов (например, перспективные и неперспективные), важно, чтобы степень их изученности была примерно одинакова.

3. Промежуточное положение занимает *критериально-аналоговая* модель, соединяющая в себе наличие критериев и использование принципа аналогий. В принципе от критериальной модели всегда можно перейти к критериально-аналоговой, организовав образы на основе создания идеальных по значениям критериев объектов разных классов. При этом необходимо помнить, что искусственно порожденные объекты могут не только не иметь аналогов в действительности, но и быть теоретически запрещенными, так как не любые наборы критериев оказываются совместными. Например, если критериями являются параметры вещественного состава, то в таком искусственном объекте различных компонент может оказаться в сумме более 100%.

Формализация исходных данных состоит в определении способов вычисления свойств, описывающих объекты поиска, и построении цифровой модели (матрицы) данных для решения задачи. При этом всегда должно осмысливать, каким образом каждое свойство соотносится с модельными представлениями, и на основании этого определять способы его вычисления на основе исходной информации.

Построение цифровой модели признакового пространства для решения задач состоит в расчете в выбранных ячейках сети формальных прогнозных характеристик по картам, данным полей, точкам наблюдения и т.д., а также производных характеристик, получаемых различными методами.

При наличии геологической информации, представленной в виде карт, для *формирования прогнозных характеристик* используются различные функции. Цифровая модель карты может содержать слои, состоящие из точечных, линейных и площадных объектов, для каждого из которых рассчитывается та или иная функция, выбирая которую, необходимо четко представлять, какой именно фактор она формализует и каким образом это согласуется с модельными представлениями [5].

При обработке информации потенциальных полей могут быть рассчитаны производные прогнозные характеристики методами трансформации полей, разделения полей на региональную и локальную составляющие, расчета корреляционных характеристик полей, вычисления статистик в скользящем окне.

Для геохимических данных широко применяются арифметические преобразования, позволяющие рассчитывать любые аддитивные, мультипликативные и другие, более сложные производные характеристики.

Формализация требований к результату обычно также приводит к расчету тех или иных характеристик, главным образом, с целью формализации прямых поисковых признаков, по которым может быть получено подтверждение результатов решения задачи (см. рис. 2).

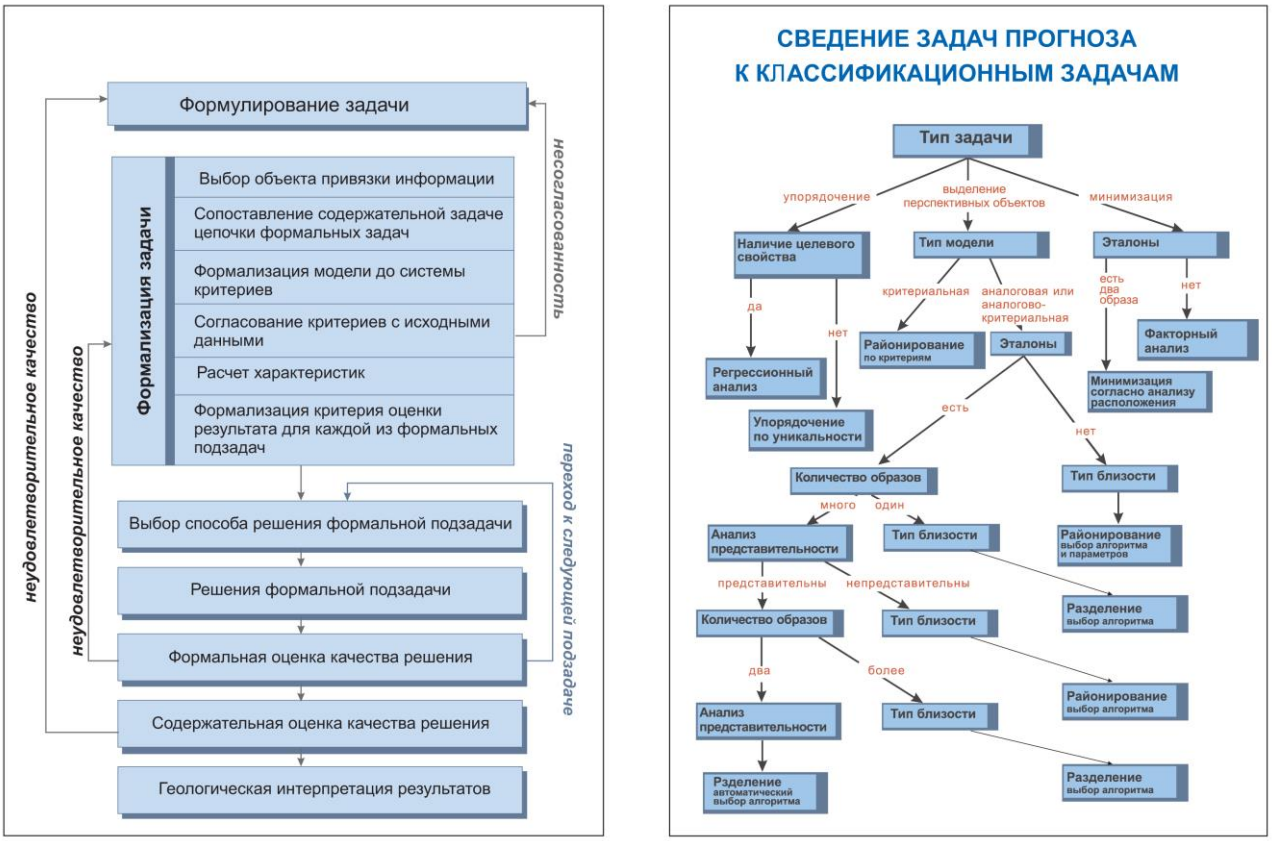


Рис. 2. Технология постановки и решения геологических задач

4. Выбор способа решения задачи

Этап выбора способа решения задачи можно подразделить на две стадии. Первая из них заключается в анализе и трансформации имеющейся (исходной и производной) информации, а вторая - собственно в выборе способа решения задачи (формальной подзадачи), базирующемся как на этом анализе, так и на самой формулировке задачи. Обе стадии тесно увязаны друг с другом.

Анализ данных содержит следующие процедуры:

- проверка вариабельности свойства,
- изменение масштаба свойства,
- выявление аномальных значений,

- выявление взаимосвязи свойств,
- проверка представительности эталонов.

Проверка вариабельности производится только для количественных свойств и направлена на проверку правильности установки шкалы. Если в каком-либо свойстве оказывается достаточно мало различных значений, то это свойство не может быть признано количественным и должно быть переведено в ранговую или номинальную шкалу.

Изменение масштаба свойства производится для приведения его к приближенно симметричному распределению. Это связано с тем, что все алгоритмы, базирующиеся на статистических методах, разработаны в предположении нормальности распределения. Без этого предположения их можно считать эвристическими. Но опыт показывает, что и статистические и эвристические алгоритмы не дают разумных результатов, если распределение свойств сильно отличается от симметричного, то есть имеет ярко выраженную асимметрию. Для того, чтобы понять необходимость изменения масштаба, строится гистограмма распределения соответствующего свойства. Если центр распределения свойства смещен влево (реже вправо), то изменение масштаба необходимо. Изменение масштаба производится применением к свойству какой-либо монотонной корректирующей функции. Для исправления правостороннего скоса рекомендуется применять экспоненту или степенную функцию при разных показателях степени. Для исправления левостороннего скоса положительно определенного свойства обычно применяется логарифм или корни различной степени. Понятно, что логарифм или корень можно применять только к положительным свойствам. Если же свойство не положительно, то его следует сдвинуть вправо, а затем применить функции, о которых говорилось выше.

Определение аномальных значений должно производиться только после приведения свойства к приблизительно симметричному виду, так как иначе аномальными будут оказываться «хвосты» ассиметричных распределений. Аномальные значения не должны препятствовать решению формальной задачи, но их наличие во многих случаях может насторожить исследователя и заставить его провести содержательный анализ материала. Особенно надо обращать внимание на объекты, для которых значения многих свойств оказываются аномальными, а также на те, у которых значение одного из свойств аномально с очень большим отлетом (7 или 10 стандартных отклонений).

Далее с помощью *вычисления коэффициентов корреляции* проводится анализ взаимосвязи свойств. Наличие сильной взаимосвязи при некоторых постановках задачи (задача разделения или районирования) должно побудить исследователя исключить какое-то из них из рассмотрения. Для других же постановок задач (основанных на методах регрессионного или факторного анализа) наличие сильно связанных между собой свойств является скорее правилом, чем исключением. Заметим, что значение коэффициента корреляции 1 или -1 с необходимостью означает линейную связь между свойствами. Однако, высокая корреляция равнозначна зависимости между свойствами только в случае, когда они распределены нормально. Например, если значения двух свойств очень малы, но имеется один аномальный объект, у которого значения обоих свойств достаточно велики, то легко убедиться, что коэффициент корреляции между ними будет близок к 1, хотя никакой связи между свойствами нет. И наоборот, равный 0 коэффициент корреляции означает отсутствие связи только для нормальных свойств. В самом деле, возьмем равномерно распределенное на отрезке [-1,1] свойство ν. Легко убедиться, что коэффициент корреляции между ν и ν^2будет равен 0, и в то же время говорить об отсутствии связи между этими двумя свойствами не приходится.

В заключении необходимо задать веса свойств**,** выбранных для решения задачи. Можно не делать этого, но тогда «по умолчанию» все признаки считаются равнозначными.

5. Методы решения формальной задачи

Будем считать, что решается одна из трех задач:

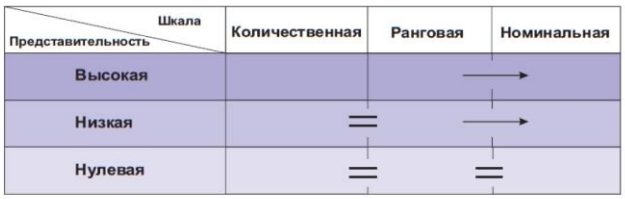
- выделить на данной территории перспективные объекты;
- при наличии целевого свойства, определенного только на части территории, необходимо оценить его на остальной территории (задача упорядочения);
- минимизировать признаковое пространство.

Начнем с первой, наиболее частой задачи. В этом случае необходимо проанализировать модель.

Если при постановке задачи удается сформулировать *критериальную* модель, то по критериям необходимо провести районирование территории по мере сходства с «идеально плохим» в смысле этих критериев объектом.

Далее будем рассматривать *аналоговую или критериально-аналоговую модели*. Здесь можно выделить несколько ситуаций (см. рис. 3):

- Имеются эталоны разных классов, и они адекватно представляют всю выборку (представительны).
- Имеются эталоны, и они непредставительны.
- Имеются эталоны только продуктивного (первого) класса.
- Эталонов нет.



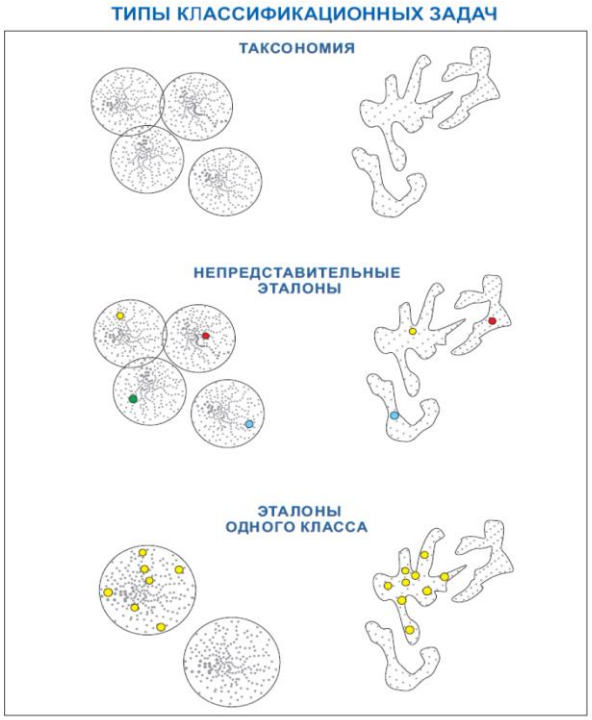


Рис. 3. Классификация постановок геологических задач

Классификация без эталонов. При отсутствии эталонов следует решать задачу районирования, и определить параметры, с которыми эта задача будет решаться. Такими параметрами являются количество групп, на которые будет разбита выборка и порядок малости, то есть количество объектов, которые должны содержаться в группе для того, чтобы она участвовала в подсчете количества групп. Основанием для выбора метода классификации является принцип организации однородных групп. Таких принципов усматривается два:

- между любыми двумя объектами, входящими в одну группу, должно быть большое сходство;
- для любых двух объектов, входящих в одну группу, должна найтись соединяющая их цепочка объектов из той же группы, такая, что сходство между любыми двумя соседями из этой цепочки очень велико.

Выбор принципа должен базироваться на формализации модельных представлений.

Понятно, что первый принцип приводит к построению шарообразных групп, а второй - к построению вытянутых «колбасообразных» групп. Второму принципу соответствуют алгоритмы, построенные на понятии «голотипа».

Для районирования на «шарообразные» группы используется широко известный алгоритм *К-средних* (Мак-Куин). Здесь порог разбиения итерационно подбирается для разбиения на нужное число групп. Для этого в качестве первого центра выбирается произвольный объект. Для следующего объекта

проверяется его мера сходства с уже имеющимися центрами. Если для какого-то из них она превышает порог, то объект присоединяется к группе этого центра, а центр корректируется так, чтобы он находился в центре тяжести группы. Если же эта мера сходства для всех центров оказывается меньше порога, то соответствующий объект становится центром новой группы. Понятно, что если центров оказывается больше, чем нужно, то надо увеличить порог, а если меньше, то уменьшить. Перебрав все объекты, нужно начать перебирать их сначала, пока все центра не «устаканятся» и не перестанут заметно двигаться. Таким образом, будет получено результирующее разбиение.

Недостатком этого метода является несогласованность разбиений, полученных для разного числа групп (разных порогов). То есть разбиение на 6 групп может быть непохожим на разбиение на 5 групп. Избежать данной проблемы поможет алгоритм *«Иерархическая таксономия».* В нем первоначально строится разбиение на большое число групп, например, на 100, а затем близкие группы объединяются между собой. Понятно, что в этом случае разбиение на 5 групп получается из разбиения на 6 групп объединением двух самых близких между собой групп. В системе присутствует алгоритм «иерархической таксономии», базирующийся на нейронных сетях Кохонена и имеющий множество режимов, соответствующих различным модификациям алгоритма.

Голотипные алгоритмы базируются на вычисление скелета, то есть кратчайшего подграфа, соединяющего все точки (объекты) в признаковом пространстве. В этом случае реализован соответствующий алгоритм, в котором строится связный граф, по ребрам которого можно совершить путь между любыми двумя вершинами графа. Кроме того этот граф не содержит петель, то есть нельзя совершить какой-либо круговой путь по графу, не проходя дважды одно и то же ребро. Можно доказать, что построенный граф имеет минимальную суммарную длину ребер среди всех связных графов, имеющих вершинами все объекты. Поэтому построенный граф не зависит от вершины, с которой начато его построение.

Разбиение на компоненты (таксономия). Понятно, что, если в кратчайшем графе разрезать К ребер с минимальной мерой сходства, то он распадется на К+1 компоненту. Тем самым решена задача таксономии.

При разбиении некоторые из компонент могут получиться слишком малочисленными. При необходимости процесс разрезания можно продолжить до тех пор, пока не получится К полноценных компонент. Правда при проведении этого процесса может оказаться, что все множество вершин разбилось на малочисленные компоненты. Тогда следует сделать вывод, что получить К полноценных компонент в этом случае невозможно.

Классификация в случае непредставленных эталонов. Проверка представительности эталонов проводится только для задачи разделения, когда имеется эталонный материал нескольких классов, а все остальные объекты следует разделить по принадлежности к этим классам. Проверка представляет из себя сравнение на гистограмме областей значения эталонного материала (по всем классам) и той части выборки, принадлежность которой к классам неизвестна. Если область значений эталонного материала существенно меньше или вообще лежит в стороне от области значений другой части выборки по некоторому свойству, то эталонный материал следует признать непредставительным для этого свойства.

При наличии непредставительных эталонов различных классов нет возможности решать задачу разделения и поэтому следует решать задачу районирования (частичной таксономии). Группу, полученную в результате районирования, в которые попали какие-либо эталонные объекты, следует относить к тому же классу, что и эти объекты. Группы, в которые не попало ни одного эталонного объекта, отнести к какому-либо классу затруднительно. Метод решения задачи выбирается пользователем в зависимости от особенностей алгоритмов. В системе задача частичной таксономии может быть решена как методом «К средних», так и голотипным алгоритмом.

Частичная таксономия, голотипный подход. Реализован алгоритм, позволяющий разбить все множество вершин на возможно более крупные компоненты так, чтобы в одну компоненту не попадали эталоны различных классов. Далее компоненты, содержащие эталоны, приписываются соответствующему классу.

Частичная таксономия, подход «К-средних». Итерационно подбирается такой порог, чтобы эталоны разных классов не попали в одну компоненту, но чтобы этих компонент было минимальное число. Понятно, что компонент окажется не меньше, чем эталонных классов.

Классификация в случае наличия эталонов только одного класса. При наличии эталонов только одного (продуктивного) класса также необходимо решать задачу разделения. Метод решения выбирается как и при непредставительных эталонах. Группы, содержащие эталоны, относятся к продуктивному классу, а остальные к непродуктивному. Общим принципом является формирование таких групп, чтобы все эталонные объекты попали в одну группу и, с другой стороны, чтобы эта группа была минимальной.

Для «шарообразных» групп эта задача решается итерационно.

При голотипном подходе строится, как и в предыдущих случаях, скелет, который разрезается в соответствии с указанным выше принципом, то есть все эталонные объекты окажутся в одной группе, а все остальные группы следует признать непродуктивными.

Классификация в случае нескольких представительных классов эталонов. В случае представительных эталонов следует решать задачу разделения. Для двух эталонных классов существует алгоритм, позволяющий определить три способа расположения классов в признаковом пространстве:

- классы далеки;
- классы близки;
- классы перемешаны.

Каждому из этих случаев соответствует класс алгоритмов распознавания: первому и второму случаю алгоритм разделения классов поверхностью, для этого реализован метод построения линейной дискриминантного функции; третий случай показывает, что сами классы не компактны и, следовательно, либо неправильно выбрано признаковое пространство и следует вернуться к этому этапу, либо необходимо применять алгоритм, работающий на некомпактных классах, например алгоритм «голотип», работающий аналогично тому, что описано выше в разделе, посвященном частичной таксономии [2].

Если же количество эталонных классов более 2, то рекомендуется провести разделение в несколько приемов, объединив исходные классы в две группы, проведя разделение, как указывалось выше, а затем делить внутри полученного разбиения. Разделение классов на группы следует проводить по содержательным соображениям. Другим вариантом является выбор способа решения задачи пользователем вручную, на основании сведений о попарном расположении классов эталонов в признаковом пространстве.

Задача упорядочения может иметь две различные формулировки. Во-первых, можно считать, что на некотором подмножестве объектов задано целевое свойство и необходимо вычислить это свойство на всем множестве объектов. Во втором случае нужно упорядочить объекты по степени их уникальности.

Первая задача решается с помощью регрессионных методов. В самом деле, можно попытаться выразить целевое свойство, заданное на подмножестве объектов через набор свойств (косвенных) которые заданы на всех объектах, а затем по полученной формуле вычислить целевое свойство для всех объектов. Обычно решается линейная задача, то есть целевое свойство выражается через косвенные по линейной формуле. Но можно с тем же успехом решить и полиномиальную задачу: для этого первоначально нужно рассчитать необходимые степени и произведения косвенных свойств, а затем решать линейную задачу от всех исходных и насчитанных косвенных свойств. Имеет смысл считать, что регрессионная задача успешно решена, если рассчитываемый при этом коэффициент множественной корреляции достаточно велик (он изменяется от 0 до 1).

Рассмотрим теперь задачу упорядочения без целевого свойства. Ее постановка может быть сформулирована следующим образом. Необходимо выделить на территории хотя бы один продуктивный объект. При этом постулируется, что продуктивных объектов мало и они не похожи друг на друга и на непродуктивные объекты, а непродуктивных объектов много и они похожи друг на друга. Из постановки задачи следует, что упорядочение должно производиться согласно уникальности каждого из объектов в пространстве признаков. Пользователь здесь должен задать два параметра: максимальное количество объектов, которые могут быть похожи на данный уникальный объект (понятно, что все они также будут уникальными), и общее число уникальных объектов. Эта задача решается с помощью построения скелета. Здесь последовательно рассекаются ребра, с минимальной мерой сходства между

соединяемыми ими объектами. Каждый раз подсчитывается количество объектов, принадлежащих малым компонентам. Для этого на каждом шаге запоминается имеющееся количество объектов в малых компонентах. Если вновь разрезаемое ребро лежало в малой компоненте, то количество не меняется. Если же оно было в крупной компоненте, то изменение происходит только в том случае, если от крупной компоненты отрезается мелкая или если крупная компонента разрезается на две мелкие. Как только оно превосходит порог, заданный пользователем, разрезание прекращается. Результатом решения этой задачи будет множество перспективных объектов, проранжированных по степени их уникальности. Можно надеяться, что среди них находится и продуктивный объект.

Задачи минимизации признакового пространства носят вспомогательный характер, решаются обычно при наличии избыточного набора признаков, и существуют в двух постановках - при наличии эталонного столбца, содержащего информацию об эталонах двух классов, и в его отсутствии. В первом случае методом исключения отыскивается такое подпространство, в котором разделение на два класса происходит не хуже, чем на первоначальном признаковом пространстве. Во втором случае задача решается методами факторного анализа.

Начнем с *факторного анализа*. Его основной посылкой является предположение, что в реальности все объекты лежат не в исходном m-мерном признаковом пространстве, а в каком-то его подпространстве меньшей размерности, и нашей задачей является отыскание этого подпространства. Для этого анализируется матрица корреляции или матрица ковариации. Поскольку это симметричные матрицы, то можно доказать, что существует такой базис в признаковом пространстве, в котором эти матрицы имеют диагональный вид. Такой базис называется базисом собственных векторов, а положительные числа, стоящие на диагонали матрицы – собственными числами. Оказывается, что собственные числа – дисперсии по соответствующим осям, а общая дисперсия выборки равна их сумме. Если собственные вектора упорядочены по мере убывания дисперсий, то, в случае избыточного признакового пространства, доля дисперсии, приходящаяся на последние оси очень невелика и ей можно пренебречь. Таким образом, первый этап факторного анализа состоит в определении такого количества собственных векторов (главных компонент), которые обеспечат нас достаточным процентом общей изменчивости выборки и переходом в соответствующее подпространство.

Первый этап факторного анализа весьма ясен и прозрачен, но, к сожалению, собственные вектора очень сложно интерпретировать. Поэтому проводится второй этап, состоящий в отыскании в полученном подпространстве такого базиса (факторов), который бы было интерпретировать легче. В идеале это должен быть такой базис, который бы выражался через исходные свойства с помощью коэффициентов 1, -1, 0. Тогда можно было бы сказать, что свойства, входящие с ненулевыми коэффициентами в какой-то фактор, образуют ассоциацию. Однако, в реальности добиться близости к таким коэффициентам обычно не удается. Многие свойства оказываются размазанными между несколькими факторами, а объяснить полученные ассоциации на содержательном уровне не представляется возможным. При этом необходимо отметить, что проценты изменчивости, который был у каждой главной компоненты (собственного вектора) на первом этапе ни в коем случае не распространяется на факторы. Они не являются независимыми и только все вместе отвечают за сохраненный процент изменчивости. Сложность интерпретации полученных факторов является препятствием для использования методов факторного анализа.

Перейдем теперь к сокращению признакового пространства при наличии эталонов двух классов. Как отмечалось ранее можно выделить три способа расположения классов в признаковом пространстве:

- классы далеки;
- классы близки;
- классы перемешаны.

Нашей целью будет сокращение размерности признакового пространства, которое не ухудшает (или незначительно ухудшает) расположение эталонных классов. Если классы перемешаны, то ничего хуже уже нельзя придумать, и мы этот случай рассматривать не будем. В двух других случаях важным параметром является расстояние (мера сходства) между центрами классов. Для сокращения размерности признакового пространства циклически выбрасывается по одному из свойств и определяется не ухудшилось ли качественно ситуация с классами, то есть не стали ли они из далеких близкими или из близких перемешанными. Если качественного ухудшения не произошло, то вычисляется расстояние между центрами и выбрасывается тот признак, при удалении которого расстояние между центрами

окажется наибольшим. Далее отыскивается следующий кандидат на удаление. Процедура завершается, если удаление любого признака приводит к качественному ухудшению ситуации.

Заключение

Предложенная в статье общая схема постановки и решения геологических задач положена в основу разработки блока постановки и решения геолого-прогнозных задач в ГИС INTEGRO. Рассмотренные в статье формальные задачи позволяют представить геологическую задачу в виде последовательности этих формальных задач и обеспечить ее решение с помощью описанных в статье алгоритмов. В перспективе предполагается использование нейронных сетей для процессов решения этих задач в ситуации наличия больших данных. Создание баз знаний, формализующих опыт предметных специалистов решения геологических задач, обеспечит применение методов ИИ для получения новых эффективных решений этих задач.

Список источников

1. Черемисина Е. Н., Миловидова А. А. Современные проблемы системного анализа и управления : электронный курс. Свидетельство о регистрации базы данных RU 2019620401, 13.03.2019.
2. ВНИГНИ – 65. Люди, результаты и перспективы / Под редакцией А.И. Варламова, В.И. Петерсилье. – Москва : Всероссийский научно-исследовательский геологический нефтяной институт, 2018. – 520 с. – EDN : EFFLFI.
3. Черемисина Е. Н., Любимова А. В., Малинина С. С. Системный подход к постановке и решению прикладных задач на базе отечественного программно-технологического комплекса ГИС INTEGRO // Системный анализ и информационные технологии САИТ-2019 : Труды Восьмой международной конференции, Иркутск, 08–14 июля 2019 года. – Иркутск: Федеральный исследовательский центр "Информатика и управление" Российской академии наук, 2019. – С. 363-369. – DOI 10.14357/SAIT2019047. – EDN OSLNJB.
4. Финкельштейн М. Я., Черемисина Е. Н., Деев К. В. Обработка геолого-геофизической информации при решении геологических задач на базе ГИС ИНТЕГРО // ГеоЕвразия 2018. Современные методы изучения и освоения недр Евразии : Труды Международной геолого-геофизической конференции, Москва, 05–08 февраля 2018 года. – Москва: ООО "ПолиПРЕСС", 2018. – С. 824-828. – EDN : XMUCKT.
5. Черемисина Е. Н., Никитин А. А. Количественные критерии системного анализа для принятия решений в проблемных ситуациях геолого-геофизических исследований. // Геоинформатика. – 2014. – № 2. – С. 20-28. – EDN : SMXFJR.