ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

В.Э. Борзых, А.В. Быков

АГРЕГИРОВАНИЕ СТРАТИГРАФИЧЕСКОЙ ИНФОРМАЦИИ С ПОМОЩЬЮ ТЕОРИИ НЕЧЕТКИХ МНОЖЕСТВ

Создание различных геологических моделей широко используется при разработке месторождений нефти и газа. Однако применение стратиграфической информации в этих моделях осложняется ее неопределенностью и зачастую противоречивостью. С помощью аппарата теории нечетких множеств возможно и на основе такой информации получать приемлемые результаты. В статье описан алгоритм получения различных характеристик для стратиграфических подразделений из исходных геологических данных и разработанное на его основе программное обеспечение.

Стратиграфия, нефть, газ, модели, нечеткие множества.

При разработке нефтяных и газовых месторождений обращаются к созданию различных геологических моделей. Однако стратиграфическую информацию при этом используют нечасто из-за большой ее сложности и даже противоречивости (это свойственно всем геологическим данным, но стратиграфические в этом плане выделяются). Теория нечетких множеств позволяет манипулировать такой информацией и получать приемлемые результаты.

Будем использовать следующую информацию:

- данные о пробуренных скважинах (названия, координаты, альтитуды, траектории стволов и др.);
- результаты геофизических исследований скважин;
- данные исследований керна;
- разбивки траекторий скважин на стратиграфические подразделения (полученные специалистами на основе других данных);
- метаданные (исполнители — организации, авторы; время получения; использовавшиеся методики и инструменты; ссылки на исходные данные и др.).

В результате можно получить следующие характеристики для стратиграфических подразделений:

- статистические параметры (средневзвешенное значение, доверительный интервал, медиана и др.) для любого геофизического метода, по которому имеется информация;
- все результаты исследования керна; для признаков, измеряемых номинальной шкалой (тип породы, найденные остатки организмов и др.),— частоты встречаемости; для непрерывных данных (пористость, проницаемость, продуктивность и др.) — получение всех статистических характеристик;
- границы распространения подразделения.

Также можно получить обобщающую информацию по всем стратонам:

- все встречающиеся подразделения для заданной территории (стратоны должны быть одного ранга и находиться на одной шкале);
- районы распространения различных подразделений одного возраста (например, границы расположения свит, содержащихся в одном горизонте).

Рассмотрим общий алгоритм. Сначала происходит загрузка данных с исправлением ошибок, которые возможно исправить. Затем по различным критериям определяется качество исходной информации (критерии качества можно изменить в дальнейшем, они не определяются только на этапе загрузки). На основании загруженных данных строится матрица принадлежности *U* разбивок скважин стратиграфическим подразделениям по дополненному для нашего случая алгоритму байесовской классификации, при этом непрерывные переменные приводятся к дискретному виду. После этого можно получить различные характеристики.

При неудовлетворительных результатах, а также при появлении новых исходных данных необходимо вернуться в начало, отредактировать исходные данные и/или пересмотреть их критерии качества. Общий алгоритм работы системы показан на рис. 1. Рассмотрим его более подробно.

Рис. 1. Общий алгоритм

Качество исходных данных играет очень важную роль в любой информационной системе. Для обеспечения достоверного результата необходимо провести предварительную обработку исходных данных, которая позволит получить качественные данные для последующего анализа [6]. Рассмотрим особенности исходных геологических данных и методы повышения их качества.

Можно выделить следующие параметры, влияющие на качество данных:

- существует несколько видов исходных данных, для каждого из которых сложились свои форматы представления, зачастую плохо согласующиеся между собой;
- данные поступают из различных организаций, форматы представления в которых могут различаться;
- большинство исходного материала получено в 1960–1970-е гг. в период массового бурения скважин, когда использовались бумажные технологии, переход от которых не был системным (выполнялся разными организациями в течение длительного периода времени, часто без сохранения преемственности, особенно в 1990-е гг.) и породил большое количество некачественных данных;
- во время массового бурения скважин в 1960–1970-е гг. качеству получаемых данных не всегда уделялось должное внимание, что особенно характерно для эксплуатационных скважин.

В связи с тем что данные поступают из различных источников, необходимо провести интеграцию данных совместно с «очисткой». Полученные после «очистки» и интеграции качественные данные можно будет использовать для анализа.

При обработке геологической информации данные, полученные из различных источников, могут различаться по степени достоверности, точности. Рассмотрим примеры разного качества исходных данных.

Одной из главных проблем является противоречивость данных друг другу. Положение усугубляется тем, что даже эксперты не могут прийти к общей точке зрения относительно их достоверности. Такая ситуация сложилась из-за наличия небольшого количества достоверных данных (которые к тому же основаны главным образом на косвенных измерениях), когда геологи вынуждены строить гипотезы, опираясь на свой опыт и интуицию, что приводит к различной интерпретации одних и тех же исходных данных и разделению специалистов на школы и направления. Кроме того, в рамках одной школы у разных авторов могут быть расхождения; одни и те же авторы со временем меняют свое мнение. В качестве примера можно привести разногласия «новосибирской» и «тюменской» школ по разбивкам Тюменской сверхглубокой скважины СГ-6 [6, 9]: расхождения в отметках границ одного и того же подразделения достигают более 400 м, а некоторые авторы неоднократно высказывали различные мнения.

Все это не позволяет непосредственно совместно использовать информацию от этих школ, можно лишь рассматривать данные каждой из них отдельно (желательно одного автора). К тому же необходимо использовать только последние данные по каждому объекту, чтобы избегать противоречий, возникающих из-за изменения авторами мнений с течением времени.

Рассмотрим поисково-разведочные и эксплуатационные скважины. Как показывает практика, достоверность данных, полученных с первых, выше. Часто даже данные с эксплуатационных скважин не берутся в расчет из-за их низкого качества. Основными причинами этого являются разное их предназначение, что влияет на характер траекторий скважин, степень их изученности и качество отбора материала.

Основным преимуществом эксплуатационных скважин является их распространенность. Данные с разведочных скважин хоть и более точные, но эти скважины пробурены сравнительно редко.

Рассмотрим процесс проставления экспертами разбивок. При хороших данных геофизических исследований скважин (ГИС) можно сопоставить несколько графиков и найти границу стратиграфических подразделений. Однако стенки скважины могут быть заглинизированы, границы пород в этом месте могут быть нечеткими. Эксперты вынуждены проводить границы условно, основываясь на данных с соседних скважин и своем опыте. Таким образом, все разбивки можно поделить на точные и условные.

Можно привести еще ряд примеров, когда выделяются группы данных разного качества. Существуют признаки, позволяющие судить о качестве исходных данных и их совместимости между собой.

Для каждого признака определяется значение достоверности данных. Например, оно может лежать в диапазоне от 0 до 1 (информация со значением 0 полностью игнорируется как недостоверная, а информация со значением 1 будет иметь наибольший вес при вычислениях). Если данные содержат несколько признаков, то общее значение получается из произведения значений всех признаков.

Следует отметить, что значение достоверности данных каждого признака может меняться в зависимости от задачи. При наличии нескольких школ целесообразно рассчитать статистические характеристики для каждой школы отдельно. Полученные таким образом характеристики качества исходной информации будут использоваться в дальнейшем при вычислениях.

Итак, необходимо построить несколько моделей, каждая из которых в дальнейшем может уточняться, и использовать их совместно. Сравнивая результаты разных моделей, можно найти области максимальной неопределен-

ности (там, где модели максимально расходятся и требуют более детального изучения, вычисления статистических данных расхождения моделей).

Основная задача состоит в следующем. Для каждого набора характеристик необходимо определить степени принадлежности всем подразделениям. Кроме того, следует использовать характеристики качества исходных данных. Представим это в формальном виде.

Дано множество объектов

$$I = \{i_1, i_2, ..., i_j, ..., i_n\},$$

где i_j — информация по участку скважины; $j = 1...n$.

Каждый объект характеризуется множеством переменных и их критериями качества:

$$i_j = \{x_1, x_2, ..., x_h, ..., x_m, k_1, k_2, ...k_h, ...k_m\},$$

где x_h — независимые переменные, значения которых известны; k_h — критерий качества x_h. Каждая переменная x_h может принимать значения из некоторого множества R_h, которое может быть конечным или бесконечным.

Критерии качества должны соответствовать следующему условию:

$$k_h \in [0, 1].$$

На основании этих переменных и их критериев качества выбирается значение матрицы принадлежности U. Значение u_{nm} показывает степень принадлежности участка скважины n подразделению m, причем должно выполняться следующее условие:

$$u_{ij} \in [0, 1].$$

Следует отметить, что здесь рассматриваются подразделения с малой мощностью, на которые обычно разбивают траектории скважин. Степени принадлежности стратонов, стоящих выше в иерархии, получаются из объединения степеней принадлежности входящих в них подразделений.

Рассмотрим основные алгоритмы, используемые для решения подобных задач. Это алгоритмы классификации или кластеризации. В случае классификации имеется обучающее множество объектов, для которых известны значения переменных и класс. Необходимо каждый поступающий новый объект по известным значениям переменных отнести к одному из классов. Кластеризация отличается тем, что обучающее множество отсутствует, необходимо разбить объекты по известным значениям переменных на кластеры (группы) похожих объектов.

Для данной задачи классы объектов (к какому относятся стратону) известны, однако эта информация является дополнительной, а за основу нового разделения на классы берутся свойства пород. В результате строится новая матрица принадлежности, которая может по некоторым параметрам существенно отличаться от первоначального разделения объектов.

Рассмотрим алгоритмы кластеризации [3, 4, 11]. Выделяют иерархические и неиерархические алгоритмы. Среди иерархических нет дающих результаты, которые можно привести к нечетким переменным, поэтому далее мы их касаться не будем. Среди неиерархических наиболее популярными являются метод нечеткой самоорганизации c-means и его обобщение в виде алгоритма Густафсона — Кесселя, с помощью которых можно найти множества с ядром в центре кластера. Однако существуют и алгоритмы другого вида, без использования центра кластера. Хотя такие алгоритмы используют нечеткие отношения и носят более универсальный характер, их применение проблематично из-за того, что они дают в результате множество разбиения, а не степени принадлежности.

Существует большое количество алгоритмов классификации [1, 5, 10], однако большинство из них дают дискретные результаты. Степени принадлежности дают только алгоритмы, создающие правила классификации на основе статистики. Это относительно простой 1R-алгоритм и алгоритм байесовской классификации, который успешно использовался для определения функции принадлежности [8]. Одним из недостатков этих алгоритмов является невозможность прямой обработки непрерывных переменных, их необходимо преобразовывать в дискретные, что может привести к потере закономерностей.

Таким образом, имеются алгоритмы классификации и кластеризации (будем использовать алгоритмы Густафсона — Кесселя и байесовской классификации, как наиболее совершенные в своих группах), необходимо найти среди них оптимальный. Дополним их для решения данной задачи.

В случае классификации исходное множество данных интервалов скважин, на основе которого первоначально создаются правила, прогоняется через эти правила для определения степени принадлежности стратонам. В этом случае необходимо следить, чтобы не происходило переобучения [10]. Под переобучением будем понимать слишком точное соответствие правил конкретному обучающему множеству (т.е. исходным разбивкам), когда теряется способность к обобщению. В нашем случае это приведет к тому, что результаты будут зависеть больше от разбивок конкретных авторов, а не от непосредственных свойств подразделений. В худшем случае мы получим исходные данные.

С другой стороны, должны быть выработаны такие правила, чтобы классификация приводила к результатам, достаточно близким к исходным разбивкам скважин.

В случае кластеризации появляется другая проблема. Так как алгоритм самостоятельно делит все множество информации по разбивкам на группы, получившиеся группы могут даже приблизительно не совпадать с данными, полученными от авторов разбивок. Это связано еще с тем, что в своей работе геологи полагаются и на плохо формализуемые характеристики, которые невозможно учесть при автоматической обработке.

Если же данные совпадают, то это говорит о хорошем выборе характеристик для разделения стратонов, высокой степени использования всей имеющейся информации в автоматическом режиме и о высоком качестве имеющихся разбивок. Однако, учитывая часто субъективный характер проведения разбивок, добиться такого совпадения при кластеризации чрезвычайно сложно, поэтому остановимся на использовании байесовской классификации для создания матрицы принадлежности.

При вычислении характеристик стратиграфических подразделений необходимо использовать максимальное количество доступной информации. Однако в этом случае на границах подразделений (где одни стратоны плавно перетекают в другие) будут учитываться и свойства пород соседних подразделений. Чтобы уменьшить влияние граничных районов, вводятся веса, которые уменьшаются от центра к периферии. В этом случае целесообразно в качестве весов использовать степени соответствия стратонам.

Используя матрицу принадлежности U, можно получить различные характеристики для стратиграфических подразделений. После этого необходимо проанализировать результаты. Если они не устраивают, следует либо загрузить дополнительные данные (если они появились), либо пересмотреть критерии качества (например, отдать предпочтение сторонникам другой школы).

Описанные алгоритмы работы со стратиграфическими данными позволили разработать уникальное программное обеспечение, предназначенное для извлечения дополнительной информации о стратонах. Опишем его основную структуру.

Исходные данные могут быть получены из различных источников. Это как локальные, так и серверные СУБД, файлы геологических форматов. Причем в нескольких СУБД необходимая нам информация может быть расположена в полях с разными именами и типами данных.

Предлагается следующий выход из сложившейся ситуации. Используя универсальную среду доступа к СУБД (DBE) и дополнив ее средствами работы с геологическими файлами, пользователю необходимо поставить в соответствие некоторым из всех доступных атрибутов внешних данных внутренние атрибуты системы (например, атрибуту «name» из таблицы «well» удаленной СУБД поставить в соответствие атрибут «имя скважины» локальных данных). После этого все данные с таким атрибутом автоматически будут перенесены в локальное хранилище системы, проверены на наличие ошибок и подготовлены к работе.

Одним из наиболее важных факторов разрабатываемого программного обеспечения является скорость обработки исходных данных. Хотя количество исходной информации сравнительно невелико — например, в Ханты-Мансийском автономном округе около 15 тыс. скважин, до 70 разбивок каждая [2], работать с ней при помощи стандартных инструментов доступа к реляционным СУБД не всегда приемлемо из-за больших временных затрат, обусловливаемых в первую очередь удаленной СУБД.

Безусловно, исходная информация должна храниться в этих СУБД, где имеются надежные системы хранения, резервного копирования, обработки и извлечения данных. Однако количество данных, необходимых для этой задачи, и объем оперативной памяти современных компьютеров позволяет хранить всю рабочую информацию на одном компьютере, без постоянного обращения к серверу СУБД за новой порцией данных. Такая структура обеспечивает значительное сокращение времени доступа к данным, что позволяет более тщательно обрабатывать исходную информацию, выявлять больше зависимостей, в том числе неявных, и оперативно отображать результаты.

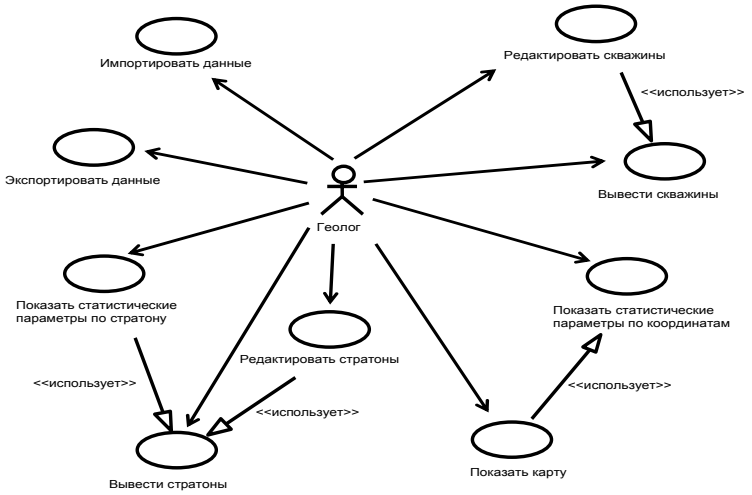

Рис. 2. Диаграмма вариантов использования

Стоит отметить, что исходную информацию, а часто и результаты работы все же следует хранить при помощи реляционных СУБД, т.е. должны быть предусмотрены средства импорта и экспорта данных в них, а также работа с форматами представления данных других программных продуктов.

В качестве языка описания модели было решено использовать UML (Unified Modeling Language), де-факто являющийся стандартным языком моделирования в процессе создания информационных систем. Была разработана модель, содержащая диаграмму вариантов использования, диаграммы последовательности, классов, компонентов и размещения.

В качестве действующего лица информационной системы будет выступать специалист, работающий с геологической информацией (геолог).

Для определения вариантов использования будем использовать функциональные требования к системе. Это загрузка данных в систему, вывод их, а также результатов их анализа. Оформим их в качестве прецедентов: «Загрузить данные», «Выгрузить данные», «Редактировать стратоны», «Редактировать скважины», «Вывести стратоны», «Вывести скважины», «Показать статистические параметры по стратону», «Показать статистические параметры по координатам», «Показать карту». Так как некоторые варианты использования могут быть инициированы не только геологом, но и другими прецедентами, отобразим это на диаграмме (рис. 2).

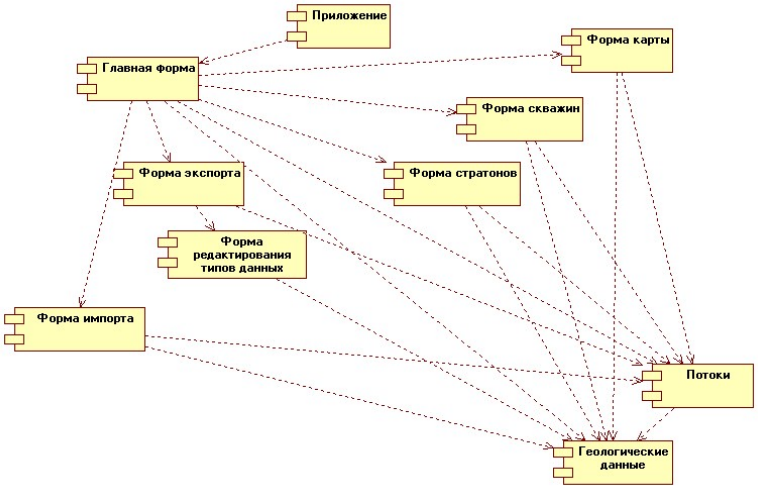

Рис. 3. Диаграмма компонент

Опишем базовые сценарии вариантов использования.

«Импортировать данные» — извлечь из СУБД или из файлов геологических форматов информацию и поместить ее в локальное хранилище для дальнейшего использования.

«Экспортировать данные» — записать информацию из локального хранилища во внешние СУБД или в файлы геологических форматов.

«Редактировать стратоны» — добавить, удалить или изменить информацию о стратонах и их взаимодействии между собой и скважинами.

«Редактировать скважины» — добавить, удалить или изменить информацию о скважинах.

«Вывести стратоны» — вывести доступную информацию о стратонах.

«Вывести скважины» — вывести доступную информацию о скважинах.

«Показать статистические параметры по стратону» — рассчитать и вывести статистические параметры по конкретному стратону.

«Показать статистические параметры по координатам» — рассчитать и вывести статистические характеристики по конкретным координатам на основе данных с окрестностей.

«Показать карту» — рассчитать для каждой точки конкретные статистические характеристики и вывести их в виде карты.

На диаграмме компонент показано расположение физических модулей кода: отображены компоненты исходного кода («Главная форма», «Форма экспорта», «Форма импорта», «Форма редактирования типов данных», «Форма стратонов», «Форма скважин», «Форма карт», «Потоки» и «Геологические данные»), а также исполняемое приложение. Связи показывают зависимости компонентов, т.е. какие из них должны быть скомпилированы раньше других (рис. 3).

Спроектированная модель позволила непосредственно реализовать приложение, использующее методы обработки стратиграфической информации.

ЛИТЕРАТУРА

1. *Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.* Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004. 336 с.

2. *Болотова И.Л., Гришкевич В.Ф., Колосова Л.А. и др.* Опыт информационно-технологического обеспечения решения геологических задач стратиграфическими методами // Вестн. недропользователя ХМАО. 2003. № 12.

3. *Дуда Р., Харт П.* Распознавание образов и анализ сцен. М.: Мир, 1976. 511 с.

4. *Дюран Б., Оделл П.* Кластерный анализ. М.: Статистика, 1977. 342 с.

5. *Елисеева И.И., Рукавишников В.О.* Группировка, корреляция, распознавание образов. М.: Статистика, 1977. 144 с.

6. *Киричкова А.И., Чижова В.А., Сташкова Э.К. и др.* Стратиграфия в нефтяной геологии: методология исследований и актуальные проблемы // Нефтегазовая геология: Теория и практика: Электрон. науч. журн. 2007. № 2. 32 с.

7. *Коровкин С.Д., Левенец И.А., Ратманова И.Д. и др.* Решение проблемы комплексного оперативного анализа информации хранилищ данных // СУБД. 1997. № 5–6. С. 47–51.

8. *Осис Я.Я.* Распознавание неисправностей сложных объектов диагностики с использованием теории размытых множеств // Кибернетика и диагностика. Рига: РПИ, 1968. С. 13–18.

9. *Филиппович Ю.В.* Некоторые аспекты стратиграфического расчленения мезозоя Западной Сибири // Вестн. недропользователя ХМАО. 2002. № 8.

10. *Чубукова И.А.* Data Mining. М.: БИНОМ: Лаборатория знаний, 2008. 382 с.

11. *Watada J., Tanaka H. and Asai K.* A heuristic method of hierarchal clustering for fuzzy intransitive relations // Fuzzy Sets and Possibility Theory / Ed. by R.R. Yager. N.Y.: Pergamon Press, 1982. P. 148–166.

V.E. Borzykh, A.V. Bykov

The aggregation of stratigraphic information using theory of fuzzy sets

The creation of different geological models is extensively used under the development of oil and gas fields. Nevertheless, the use of stratigraphic data in these models being complicated by its ambiguity and, often, inconsistency. Using the apparatus of theory of fuzzy sets allows to obtain the acceptable results even from such information. The article gives a description of the algorithm in order to obtain different characteristics for stratigraphic subdivisions using the initial geological data as well as software developed on that basis.

Stratigraphy, oil, gas, models, fuzzy sets.