Записки Ленинградского ордена Ленина, ордена Октябрьской Революции
и ордена Трудового Красного Знамени горного института им. Г.В.Плеханова,
1985, т. 103

УДК 550.4:518.54

ИСПОЛЬЗОВАНИЕ ПАКЕТА ПОДПРОГРАММ ДЛЯ РЕШЕНИЯ ГИДРОГЕОХИМИЧЕСКИХ КЛАССИФИКАЦИОННЫХ ЗАДАЧ

М.Б.ШТОКАЛЕНКО

Обязательным этапом обработки данных гидрогеохимического опробования является систематизация наблюдений. Систематизация проводится с целью выделения однородных гидрохимических полей [5], определения фоновых и аномальных содержаний компонентов, что позволяет установить общность условий формирования состава опробованных вод.

Для решения задачи используются формальные методы классификации многомерных данных [2, 4, 6]. Условия применения существующих реализаций этих методов имеют ряд принципиальных и технических ограничений [I]. Предлагается использовать для решения гидрогеохимических классификационных задач пакет подпрограмм КЛАСС (I424428.00900-0I), разработанный во ВНИИ разведочной геофизики НПО «Рудгеофизика" в качестве одной из компонент автоматизированной системы обработки материалов рудной геофизики (АСОМ РГ) для ЕС и СМ ЭВМ. Программное обеспечение пакета мобильно [I0]. Все подпрограммы могут эксплуатироваться на любых типах ЭВМ, имеющих транслятор с Фортрана.

Пакет подпрограмм КЛАСС предназначен для классификации произвольных объектов, представленных набором значений признаков. Пакет ориентирован на обработку данных поисковых гидрогеохимических съемок, но может быть использован в любой другой области знаний для упорядочения многомерных наблюдений.

В состав пакета включены три подпрограммы, реализующие различные алгоритмы таксономии (TAXON, CLSTR, CENTR), две подпрограммы сокращения числа выделенных классов (LIMIT, JOINT), подпрограммы-функции классифицирования (IDDD, IDDC), подпрограммы подготовки карты-распечатки результатов классификации (MAP), подпрограммы оценки статистических характеристик заданного класса (ESTIM, SEQNC) и др. Здесь использована терминология, предложенная в работе [I]. При определении состава пакета предпочтение отдано безэталонным методам, поскольку последние не навязывают готовой схемы классификации, предполагая тем самым более деликатное обращение с исходным материалом.

В качестве меры подобия в алгоритмах таксономии, сокращения числа классов и классифицирования IDDD используется квадрат евклидова расстояния между точками, изображающими объекты в нормированном пространстве признаков. Предусмотрена возможность замены меры подобия путем предвари-

тельной нормировки значений признаков каждого объекта на величину модуля соответствующего вектора признаков. В этом случае упомянутое расстояние становится равным удвоенному синусу половины угла между векторами.

Подпрограммы TAXON и CLSTR автоматически нормируют и квантуют исходные данные следующим образом: размах значений каждого признака делится на заданное пользователем число уровней квантования, что дает шаг квантования признака. Затем разность исходного значения признака и минимального в выборке делится на шаг квантования. Таким образом, все признаки выражаются в сопоставимых единицах.

Для избирательного изменения разрешающей способности метода предлагается добавлять в выборку фиктивный объект с аномально высокими значениями признаков, влияние которых требуется уменьшить. Тогда шаг квантования таких признаков увеличится.

Подпрограммы CENTR, JOINT, IDDD осуществляют нормировку путем умножения значений признаков на заданные веса. Последние входят в список формальных параметров подпрограмм, т.е. могут подбираться пользователем.

Подпрограмма TAXON реализует оригинальный эвристический алгоритм таксономии в предположении информационной равноценности признаков. Алгоритм не использует никаких априорных предположений о распределении вероятности по значениям признаков. В пространстве признаков выделенным классам соответствуют связные области произвольной конфигурации. Каждая такая область содержит одну точку локального максимума плотности распределения вероятности. Построение классов ведется от точек максимумов по правилу классификации, основанному на принципе ближайшего соседа [II]. Границы классов проводятся по гиперповерхностям относительных понижений плотности вероятности. Предложенный метод представляет собой многомерный вариант построения вариационных кривых или гистограмм распределения.

Алгоритм не использует ковариационной или корреляционной матриц исходной выборки. Определение коэффициента корреляции признаков в неоднородной выборке не вполне корректно, поскольку полученные значения не отражают возможного наличия подмножеств с высокой положительной или отрицательной корреляцией. В силу сказанного, отказ от использования объединенной ковариационной матрицы обеспечивает алгоритму TAXON преимущество в разрешающей способности по сравнению с методом главных компонент [4]. Последний часто применяют для безэталонной классификации многомерных объектов: изображенные точками на плоскости главных компонент объекты можно легко расклассифицировать вручную, графически [8].

Предлагаемый метод не требует специального определения критерия однородности выделяемых классов. Необходимость такого определения вносит дополнительную субъективность и в без того искусственные классификационные процедуры.

Результаты работы подпрограммы TAXON не зависят от последовательности объектов в выборке. Такая зависимость является распространенным недостатком алгоритмов таксономии [I].

Подпрограмма CLSTR предназначена для решения задачи таксономии при наличии главного по информативности признака. Выполняется сортировка объектов по совокупности признаков при заданной последовательности изменения веса признаков. Использована идея алгоритма сортировки одномерных объектов BUBBLESORT [3]. Сортировка обеспечивает отсутствие пересекающихся связей в иерархической структуре роста сходства объектов, которая выстраивается подпрограммой по мере отбрасывания малоинформативных признаков. В качестве веса, определяющего приоритет признака, используется абсолютная величина коэффициента корреляции с главным признаком. Подпрограмма выявляет отношения между классами, способствуя тем самым их интерпретации.

Третья подпрограмма таксономии – CENTR – реализует модификацию алгоритма k-внутригрупповых средних [II], ориентированного на выделение однородных классов. Мерой однородности служит величина дисперсии главного признака в каждом классе.

Подпрограммы сокращения до заданного числа выделенных классов – LIMIT и JOINT – служат для расформирования малочисленных классов по правилу ближайшего соседа с наложением ограничения на степень близости [I]. Подпрограмма LIMIT в качестве возможных соседей перебирает все объекты сохраняемых классов. Тем самым, при соблюдении ограничения продолжается процесс построения в пространстве признаков областей произвольной конфигурации. Подпрограмма JOINT рассматривает в качестве соседей центры выделенных классов. Расформировываются классы, имеющие наиболее близких соседей, превосходящих их по численности. Подпрограмма JOINT ориентирована на выделение однородных классов, которым соответствуют в пространстве признаков многомерные эллипсоиды.

Подпрограммы IDDD и IDDC выполняют классифицирование объекта по правилу ближайшего соседа с наложением ограничений на меру близости. В качестве меры IDDD использует квадрат евклидова расстояния, IDDC – коэффициент корреляции между векторами признаков.

Во всех подпрограммах пакета отсутствуют обращения к внешним устройствам ввода и вывода. Этим достигается мобильность программного обеспечения, его независимость от форматов представления исходных данных и от состава оборудования ЭВМ. Мобильность позволяет без изменений включать подпрограммы в состав программных комплексов.

Размещение исходных данных и результатов счета в оперативной памяти ЭВМ дает большой выигрыш в скорости вычислений по сравнению с вариантами реализации алгоритмов, использующими внешние носители. Хранение исходных данных и промежуточных результатов на внешних носителях – магнитных дисках и лентах – приводит к достаточно сложной специальной задаче организации страничной памяти [3].

Для обработки в рамках пакета больших массивов, превышающих допустимые размеры раздела оперативной памяти ЭВМ, предлагается следующая методика: решение задачи таксономии для части выборки и затем классифицирование оставшихся объектов с наложением ограничения на меру подобия.

Для множества объектов, отсеянных в силу наложенного ограничения, вновь решается задача таксономии. Таким образом вся большая выборка оказывается разбитой на классы.

Анализ сравнительных возможностей различных типов современной цифровой вычислительной техники позволяет ожидать высокой эффективности обработки больших массивов многомерных данных при размещении последних в оперативной памяти мультипроцессора ПС-2000. Мультипроцессор представляет собой параллельную вычислительную систему, состоящую из однотипных процессорных элементов, обрабатывающих по общему алгоритму несколько потоков данных. Оперативная память мультипроцессора позволяет разместить выборку из 16 тыс. объектов при числе признаков, не превышающем 64. Специальное исследование алгоритмов нормировки и классификации многомерных данных показало возможность их эффективной реализации на мультипроцессоре. Информационные графы, описывающие алгоритмы нормировки, обладают параллелизмом первого рода, т.е. полной независимостью потоков данных. В информационных графах, описывающих алгоритмы классификации, наряду с параллельными ветвями появляются сходящиеся, например, при суммировании квадратов разностей значений признаков. Такой, неполный, параллелизм называют параллелизмом второго рода [9].

Целесообразность постановки классификационных задач на высокопроизводительных параллельных ЭВМ диктуется значительными объемами обработки материалов гидрохимических съемок, спецификой алгоритмов классификационных задач, содержащих значительную долю операций, которые можно выполнять одновременно над различными данными. В настоящее время параллельные вычислительные системы используются в геологии для решения задач, требующих получения нескольких вариантов решения. Неоднозначность результатов характерна и для задачи систематизации многомерных наблюдений. Применение мультипроцессора позволит быстро изучить множество возможных классификаций и выбрать наилучшие с точки зрения содержательной интерпретации.

Для всех подпрограмм пакета КЛАСС предусмотрена возможность отсутствия исходных данных по отдельным признакам. Отсутствующие значения заменяются условным числом по выбору пользователя. Мера подобия объектов вычисляется с целью принятия решения о принадлежности объектов к одному классу. Решение принимается на основе сравнения значений меры подобия для разных объектов. Каждый из признаков может повлиять на это решение в ту или иную сторону. При отсутствии данных хотя бы по одному из сопоставляемых значений признака подпрограммы пакета относят неизвестное влияние в пользу близости объектов, т.е. разность значений признака считается нулевой. Тем самым обеспечиваются одинаковые условия оценки меры подобия для всех объектов. Искажение размерности пространства признаков формально не влияет на оценку меры подобия, поскольку все алгоритмы используют нормированные значения признаков. Вопрос о допустимой доле условных чисел нуждается в специальном исследовании.

Подпрограммы ESTIM и SEQNC обеспечивают возможность оценки статистических характеристик выделенных классов: вычисляются средние значе-

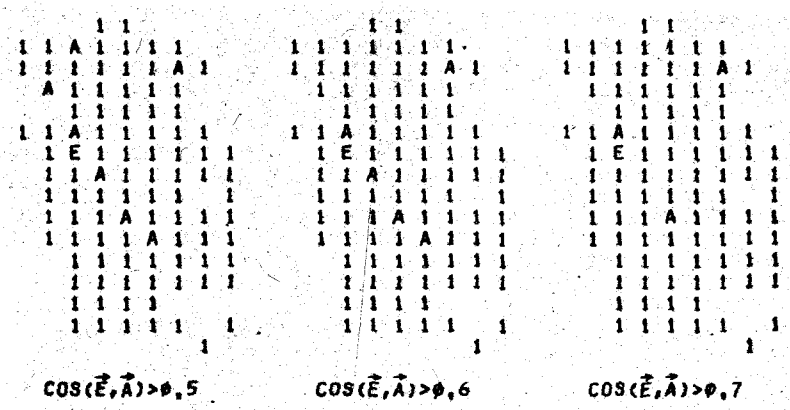
ния признаков, их стандартные отклонения, корреляционная матрица и ряд подвижности компонентов по Ю.К.Буркову [7].

Пакет КЛАСС использован для разбраковки комплексных десятикомпонентных аномалий, выделенных при поисках скрытых рудных месторождений. Район поисковых работ полностью перекрыт четвертичными отложениями мощностью более 6 м. Геологическая подзадача разбраковки аномалий заключалась в установлении положения предполагаемого фронта высокотемпературной гидротермальной деятельности. Наличие такого фронта представляет собой поисковую предпосылку.

На площади 10 х 20 км было отобрано около 6 тыс.проб. Эмиссионным спектральным анализом в пробах определены содержания цинка, никеля, кобальта, меди, серебра, свинца, молибдена и других элементов. В соответствии с требованиями инструкции [5] были установлены нижние аномальные значения содержаний элементов, что позволило выявить 107 аномалий – объектов классификации, представленных десятью признаками – содержаниями элементов, нормированными на стандартные отклонения.

Подпрограммы TAXON и LIMIT выделили три уникальных объекта – участки опробования со среднеаномальными значениями признаков. Эти аномалии были использованы в качестве эталонов при классифицировании объектов. Мерой подобия служил косинус угла, образованного векторами признаков. Для аномалий-аналогов значение меры подобия было ограничено снизу числами 0,5; 0,6 и 0,7. Тем самым критерий подобия последовательно ужесточался, что позволило уточнить степень сходства с эталоном для каждого аналога. Условные числа, заменяющие отсутствующие данные, не применялись.

В результате работы подпрограммы IDDC по пяти смежным участкам прослежена линейная зона аномалий-аналогов протяженностью более 8 км (см. рисунок). С известными геологическими и ландшафтными структурами зо-



Карты результатов разбраковки 10-компонентных аномалий
E – аномалия-эталон; A – аномалия-аналог, выделенная при заданном критерии подобия; 1 – аномалия, не обнаружившая сходства с эталоном при заданном критерии подобия, фон

на не коррелирует. Зона была интерпретирована как фронт высокотемпературной гидротермальной деятельности.

Ранее те же данные обрабатывались по схеме факторного анализа, и упомянутая линейная зона не была обнаружена. Таким образом, применение методов формальной классификации позволило получить новые сведения об исследуемой территории на основе обработки данных геохимического опробования.

ЛИТЕРАТУРА

1. Бабич В.В. Типизация задач распознавания. – В кн.: Распознавание образов в задачах качественного прогноза рудных месторождений. Новосибирск, Наука, 1980.

2. Гавришин А.И., Жогло В.Г. Результаты машинного анализа хорошо изученной задачи. – В кн.: Применение математических методов и ЭВМ в геологии. Новочеркасск, изд. Новочеркасского политехн. ин-та, 1980.

3. Гудман С., Хидетниеми С. Введение в разработку и анализ алгоритмов. М., Мир, 1974.

4. Девис Дж. Статистика и анализ геологических данных. М., Мир, 1977.

5. Инструкция по геохимическим методам поисков рудных месторождений / Мингео СССР. М., Недра, 1983.

6. Коротков А.И. Гидрогеохимический анализ при региональных геологических и гидрохимических исследованиях. Л., Недра, 1983.

7. Методические указания по геологической съемке масштаба 1:50000. Гидрохимические исследования. Л., Недра, 1970. Вып. 10.

8. Мураджанов С.А., Эленбоген А.М., Гараев С.Г., Таптугян Д.Т. Применение математических методов для оценки перспективности геофизических аномалий. – Сов. геология, 1982, № 6.

9. Прангишвили И.В., Виленкин С.Я., Медведев И.Л. Параллельные вычислительные системы с общим управлением. М., Энергоатомиздат, 1983.

10. Программы по математическому обеспечению обработки и интерпретации геолого-геофизических материалов на ЭВМ. Л., изд. НПО „Рудгеофизика", 1982.

11. Ту Дж., Гонсалес Р. Принципы распознавания образов. М., Мир, 1978.