Записки Ленинградского ордена Ленина, ордена Октябрьской Революции
и ордена Трудового Красного Знамени горного института им. Г.В.Плеханова,
т.XCII, 1982 г.

УДК 550.834.05

АЛГОРИТМЫ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ПРИ РЕШЕНИИ ЗАДАЧИ КЛАССИФИКАЦИИ ДОННЫХ ОТЛОЖЕНИЙ АКУСТИЧЕСКИМИ МЕТОДАМИ

А.И.СВЕЧНИКОВ

Разделение донных отложений по гранулометрическому составу на валунные, галечные, песчано-гравийные, песчаные, глинистые и илистые, определение принадлежности идентифицируемых осадков к определенному литологическому типу на основании результатов акустических исследований, выполняемых при картировании донных отложений на акватории, сводится к решению задачи восстановления совокупности свойств объекта по его реакции на внешнее воздействие (акустическое излучение). Подобные задачи относятся к обратным задачам геофизики.

Рассмотрим некоторые особенности постановки задачи классификации донных отложений акустическими методами. Во-первых, изучаемый объект является геологической средой, физико-механические свойства которой зависят от целого ряда неконтролируемых факторов, что проявляется в неоднородности и случайной изменчивости характеристик среды, во-вторых, акустические измерения, на основании которых определяются свойства изучаемого объекта, проводится в среде, обладающей некоторой неопределенностью характеристик. Кроме того, момент измерений (контакт акустического луча с исследуемым объектом) приводит к третьему типу неопределенности – к геометрической неопределенности процесса измерения. В этих условиях однозначное решение получить трудно.

Проанализируем возможности разделения донных отложений по результатам акустических измерений на статистически обособленные литологические классы. Успех классификации будет определяться следующими основными факторами:

1) существованием разделения донных отложений по физико-механическим свойствам на обособленные статистически не связанные либо слабо связанные классы;

2) наличием устойчивых корреляционных зависимостей между акустическими характеристиками объекта, получаемыми при акустических измерениях, и физико-механическими свойствами объекта.

Анализ физико-механических свойств донных осадков [1, 4] указывает на возможность их разделения на отдельные литологические группы в соответствии с принятой геологической классификацией на основании акустических характеристик.

Экспериментальные работы, проведенные с целью выявления статистических закономерностей, связывающих разнородные литологические типы донных осадков и определенные изменения в эхо-сигнале, доказали существование устойчивых связей между физико-механическими свойствами донных осадков и их реакцией на акустическое излучение, проявляющееся в изменении характеристик эхо-сигнала (амплитудного значения огибающей, коэффициентов формы, спектральных составляющих).

В процессе сейсмоакустических измерений присутствуют факторы (статистический шум), которые нивелируют статистические распределения акустических характеристик донных осадков. Эти факторы обычно связаны с анизотропией свойств, обусловленной геологическим генезисом осадка, геометрической неопределенностью контакта акустического луча с поверхностью осадка, неоднородностью акустических характеристик водного слоя.

С позиций статистической теории задача определения литологического типа осадка на основании сейсмоакустических измерений сводится к отысканию таких признаков эхо-сигнала, которые при минимальном уровне статистических шумов сводят к минимуму риск при принятии решения о принадлежности выделенного компактного признакового класса к конкретному литологическому типу донных отложений. Поставленная задача связана с проблемой классификации сигналов и относится к классу задач распознавания образов.

В решении задач распознавания образов в последнее время наметилось два основных подхода: при первом полагают, что вся априорная информация о классах сигналов в точности известна, при втором считаются с тем, что исходных сведений недостаточно для непосредственного применения методов теории статистических решений [3]. В настоящее время, когда предпринимаются первые шаги в создании цифровых методик обработки, второй подход, по-видимому, более эффективен в силу своей оперативности и гибкости при небольшом объеме информации.

С учетом этих факторов структурная схема обработки сейсмоакустической информации может быть разделена на два этапа: предварительную обработку, служащую для вычисления признаковых характеристик, отражающих интересующие свойства классифицируемого объекта и окончательную итоговую обработку, предназначенную для классификации исследуемого объекта на основании вычисленных характеристик и оценки качества классификации, т.е. для принятия решения о принадлежности исследуемого осадка к определенному литологическому классу.

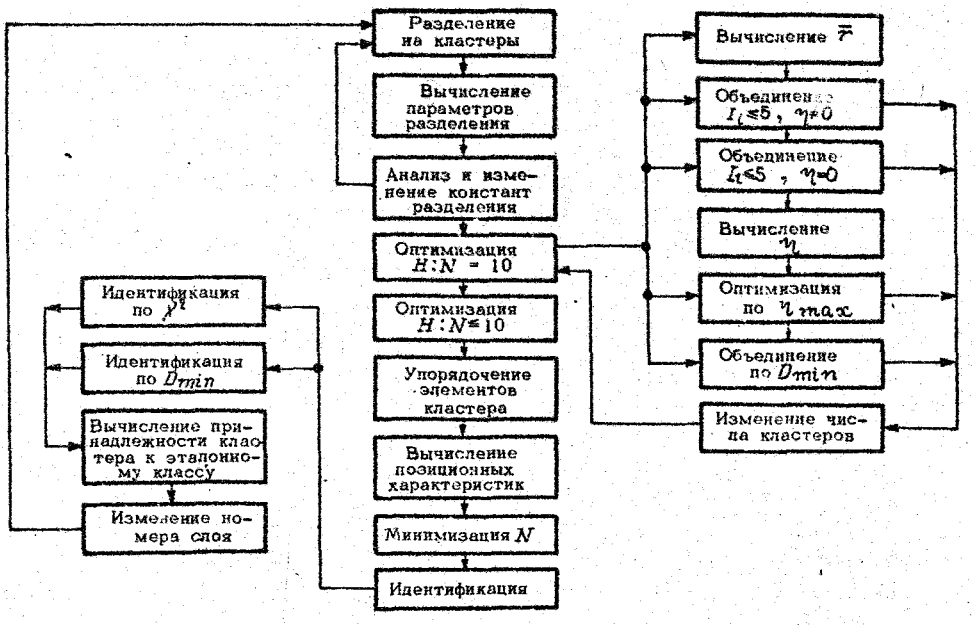
Этап предварительной обработки включает в себя вычисление признаковых характеристик эхо-сигнала и определение акустических характеристик. Акустические характеристики являются признаковыми характеристиками (признаковыми описаниями) стратифицируемых слоев донных осадков и формируются на основе полученных признаковых характеристик эхо-сигналов. На

этапе итоговой обработки исходной информацией являются акустические характеристики, но в качестве дополняющего набора признаковых описаний могут использоваться и отдельные признаковые характеристики эхо-сигналов, например длительность эхо-сигнала, характеристика спада амплитуды в эхо-сигнале. Применение акустических характеристик на этапе окончательной обработки обусловлено в первую очередь их независимостью от характеристик зондирующего акустического импульса.

При итоговой обработке сформированное пространство признаков по профилю проверяется на однородность с помощью принципа послойной обработки признаковой информации. Если пространство признаков неоднородно, то оно разделяется на области, для которых признаки, принадлежащие одной и той же области обладают максимальным сходством; соответственно признаки, принадлежащие разным областям, должны обладать минимальным сходством [2]. Эта операция выполняется с использованием методов кластерного анализа. Полученное на первом этапе разделение на кластеры проверяется на оптимальность и на этапе окончательной обработки предъявляется на идентификацию с использованием эталонных массивов признаковых описаний отдельных литологических типов донных осадков.

Граф статистической обработки в соответствии с изложенным состоит из отдельных процедур, основными из которых являются процедура разделения на кластеры, процедура оптимизации полученного кластерного разделения и процедура идентификации (см. рисунок).

Процедура разделения на кластеры состоит из отдельных блоков, которые выполняют операции разделения на кластеры, вычисления параметров разделения и изменения констант разделения.



Граф обработки

Блок разделения на кластеры выполняет признаковый анализ с фиксированным окном по всему массиву характеристик. Параметры используемого окна контролируются блоком изменения констант разделения. Начальный размер окна по всем признакам принимается равным $\sigma = 3t$, где t – погрешность измерений.

Основная цель процедуры перебора по признакам – выделение признаков объектов, имеющих наибольшую вероятность появления, определяемую по числу объектов, совпавших с принятым за эталонный.

Одновременно с вычислением вероятности встречи объекта вновь принятый за основу объект проверяется на возможную связь с предыдущими объектами. Если анализируемый объект уже встречался на предыдущих этапах анализа, т.е. входил в состав организованных групп, то в дальнейшем он рассматривается как зависимый объект, который не в состоянии организовать собственную группу. В противном случае организуется новая группа с элементами a_{l_i}, где l и i – индексы номера соответственно кластера и выборки в кластере. Процедура выделения обособленных групп производится по всему признаковому массиву, что в итоге приводит к разделению признакового массива на L групп – кластеров, причем

$$\left[a_{l_1}\right] \bar{\in} \left[a_{\substack{l_i \\ l,I_l}}\right] \quad i \neq 1 .$$

Блок вычисления параметров разделения на кластеры предназначен для определения констант G, $\bar{\eta}$, $\bar{I}$, характеризующих качество разделения (здесь G – константа, определяющая число несвязанных кластеров в признаковом массиве; $\bar{\eta}$ – коэффициент связи между кластерами; $\bar{I}$ – математическое ожидание числа элементов в кластере).

Константы

$$\bar{\eta} = 1 - \frac{j_m}{\sum I_l}; \qquad \bar{I} = j_m \frac{1}{N} p_k,$$

где j_m – максимальное число признаковых выборок; I_l – число выборок в кластере с номером l; N – максимальное число выделяемых литологических классов; p_k – вероятность образования кластера с размером $\bar{I}$.

В случае равновероятного появления каждого литологического класса на исследуемом профиле и при $N = 10$ и $p_k = 1$ константа $\bar{I} = 0{,}10 j_m$.

Блок изменения констант разделения определяет номер кластера, содержащего максимальное число элементов I_{max}, анализирует параметры разделения и на основании анализа принимает одно из следующих решений:

1) если между кластерами сильная связь ($\bar{\eta} \geq 0{,}3$), то необходимо повторить этап разделения на кластеры с новым значением окна $\sigma = \sigma - \Delta\sigma$;

2) если $\bar{\eta} = 0$ и $L \geq 10$, т.е. все кластеры обособлены, то вырабатывается решение на объединение $\sigma = \sigma + \Delta\sigma$;

3) если $G \neq 0$, $I_{max} \geq \bar{I}$, то считается, что достигнуто удовлетворительное разделение;

4) если $G \neq 0$; $I_{max} \leq \bar{I}$, т.е. все кластеры связаны и малопредставительны, то вырабатывается решение на объединение $\sigma = \sigma + \Delta\sigma$;

5) если $L \neq 1$; $L \leq 10$, то считается, что достигнуто удовлетворительное разделение;

6) если $L = I$; $\eta = 0$, т.е. разделение на кластеры не наступило, то принимается решение на разделение $\sigma = \sigma - \Delta\sigma$.

При появлении контролирующих признаков: I) $\sigma = 0$ – уменьшение размера окна до значения t; 2) $\sigma = 6t$; 3) режим объединения меняется на режим разделения или наоборот – блок изменения констант разделения прекращает свою работу и передает управление на выполнение процедуры оптимизации разделения.

В основе процедуры оптимизации лежит гипотеза H о существующем разделении признакового пространства, полученного при измерении акустических характеристик по интерпретируемому разрезу донных отложений, на N самостоятельных кластеров.

На первом этапе оптимизации осуществляется выполнение гипотезы $H{:}N = 10$. Для второго этапа оптимизации на основании позиционного критерия размещения объектов по профилю проверяется и уточняется гипотеза $H{:}N \leq 10$.

Первый этап оптимизации включает в себя выполнение следующих основных операций:

I) вычисление среднего расстояния $\bar{r}$ между кластерами;

2) объединение кластеров с параметрами $I_l \leq 5$; $\eta \neq 0$;

3) объединение кластеров с параметрами $I_l \leq 5$; $\eta = 0$;

4) объединение по минимальному расстоянию между кластерами;

5) вычисление коэффициента связи между кластерами η;

6) объединение по максимальному значению коэффициента связи между кластерами.

Блок вычисления среднего расстояния между кластерами

$$\bar{r} = \frac{\sum^{L-1}\sqrt{\sum^{k}\left[X_{jnk}(1) - X_{jnk}(l+1)\right]^2} + \sum^{L-2}\sqrt{\sum^{k}\left[X_{jnk}(2) - X_{jnk}(l+2)\right]^2} + \dots}{S},$$

где $X_{jnk}(l)$ – характеристика эхо-сигнала, образующая кластер с номером l; S – число независимых расстояний между кластерами, $S = (L-1)+(L-2)+\dots+1$, является вспомогательным и служит для обеспечения работы блока объединения по параметрам $I_l \leq 5$; $\eta = 0$.

Блок объединения по $I_l \leq 5$; $\eta = 0$ предназначен для уплотнения кластеров за счет уменьшения их общего числа. Процесс уплотнения кластеров включает в себя две процедуры объединения. Первая обеспечивает объединение кластеров, для которых межкластерное расстояние $r_l \leq \bar{r}$. После объединения кластеров с параметрами $I_l \leq 5$; $\eta = 0$ по критерию $r_l \leq \bar{r}$ осуществляется объединение оставшихся кластеров на основании позиционного критерия максимальной близости. При этом считается, что кластеры с характеристикой $I_l \leq 5$ являются малопредставительными и приводят к излишней детализации на этапе интерпретации.

Блок объединения по параметрам $I_l \leq 5$; $\eta \neq 0$ минимизирует общее число кластеров с размером $I_l \leq 5$ на основании максимального значения коэффициента связи η, причем присоединение выполняется к кластеру с большим значением I_l.

Блок вычисления коэффициента связи $Q_{l\eta}$ (коэффициент связи кластера с номером l с кластерами из совокупности $[L-1]$, причем размер массива по η определяется параметром I_l) выполняет вспомогательную функцию подготовки блока данных для процедуры оптимизации по характеру связи.

Коэффициент связи T_l представляет собой одномерный массив чисел, который определяет представительность (полноту) связей кластера с номером l в пространстве кластеров. Значение коэффициента связи $a_{l\eta}$ вычисляется по формуле $a_{l\eta} = \phi / I_l$, где ϕ – параметр силы связи кластера, который определяет число элементов a_{l_i} кластера с номером l, совпавшее с элементами кластера из массива $M_{l\eta}$ (здесь массив номеров кластеров из совокупности $[L-1]$, с которыми связан кластер с номером l).

Блок оптимизации по коэффициенту связи выполняет операцию объединения сильносвязанных кластеров (критерий $a_{l_r} = max[a_{l_r}]$), причем процедура оптимизации заканчивается либо при выполнении гипотезы $H: N = 10$, либо после завершения объединения всех сильносвязанных кластеров с коэффициентами связи $a_{l_r} \geq 0{,}1$.

Процедура оптимизации по минимальному расстоянию между кластерами является завершающей операцией первого этапа оптимизации кластерного разделения.

Блок оптимизации находит минимальное нормированное расстояние D_l между кластерами и выполняет объединение кластеров по критерию $D_l = min[D_l]$:

$$[D_l]_L^L = \sum_1^{L-1}\sum_1^{k_m}\left\{\frac{X_{jnk}(1)-X_{jnk}(l+1)}{0{,}5[X_{jnk}(1)-X_{jnk}(l+1)]}\right\}^2 + \sum^{L-2}\sum^{k_m}\left\{\frac{X_{jnk}(2)-X_{jnk}(l+2)}{0{,}5[X_{jnk}(2)-X_{jnk}(l+2)]}\right\}^2 + \dots,$$

где k_m – максимальное число признаков для одного эхо-сигнала.

Блок изменения числа кластеров является блоком принятия управляющих решений. В качестве управляющего решения для выяснения приоритета кластера при объединении используется параметр, определяющий представительность или размер кластера. Приоритет организации кластера остается за тем кластером, значение I_l для которого выше.

Основным критерием принятия решения о том, что наступило оптимальное разделение на кластеры, является выполнение принятой гипотезы $H: N = 10$.

На втором этапе оптимизации осуществляется проверка гипотезы $H: N \leq 10$ и определение значения N.

В состав процедуры оптимизации второго этапа входят три основные операции: упорядочение элементов (ооъектов) кластера, вычисление позиционных характеристик кластера и минимизация общего числа кластеров.

Операция упорядочения предназначена для исключения повторяющихся элементов в кластере и размещения их в соответствии с позиционным номером по критерию $a_{l,i-1} < a_{l,i} < a_{l,i+1}$.

Для проверки гипотезы $H: N \leq 10$ осуществляется вычисление позиционных характеристик кластера: среднего значения элемента a_{l_i} для позиционного класса $b_{l_c} = \sum_{i=A}^{B} a_{l_i} / f_c$, числа элементов в позиционном классе $f_c = B - A + 1$ (здесь B и A – границы позиционного класса), числа позиционных классов в кластере c_l (здесь c – текущий номер позиционного класса в кластере).

Граничные условия:

$$\left(a_{l_i}\Big|_{i=A+1}^{B} - a_{l_i}\Big|_{i=A}^{B-1}\right) \leq 0; \qquad \left(a_{l_i}\Big|_{i=A} - a_{l_i}\Big|_{i=A-1}\right) > 0;$$

$$\left(a_{l_i}\Big|_{i=B+1} - a_{l_i}\Big|_{i=B}\right) > 0; \qquad f_c > F,$$

где F – минимально допустимое число элементов в позиционном классе;
θ – шаг позиционного разброса в классе.

На основании вычисленных позиционных характеристик блок минимизации проверяет выполнение гипотезы $H: N < 10$, вычисляет значение N и производит объединение кластеров с близкими позиционными характеристиками. При этом кластер с номером l считается близким к кластеру с номером $(l+\xi)$ если выполняется условие

$$\left(b_{l_c}\Big|_{l}^{L-1} - b_{l_c}\Big|_{l=l+\xi}^{L}\right) \leq \theta$$

для значений в интервале $[1, C_l]$, если $C_l < C_{l+\xi}$, или в интервале $[1, C_{l+\xi}]$, если $C_l > C_{l+\xi}$, причем параметр ξ при фиксированном значении l изменяется от 1 до $(L-l)$.

После завершения процедуры оптимизации сформированные кластеры сопоставляются с эталонными литологическими классами.

Процедура идентификации состоит из трех основных операций: идентификации по критерию χ^2, идентификации по D_{min} и операции вычисления принадлежности кластера к эталонному литологическому классу.

Выбор той или иной операции идентификации определяется размером кластера I_l. При $I_l > 50$ выполняется идентификация по критерию χ^2, в противном случае по критерию минимального расстояния между кластером и литологическим классом D_{min}.

Блок идентификации кластера по критерию χ^2 служит для вычисления соответствия закона эмпирического распределения признаков в идентифицируемом кластере с теоретическим распределением эталонного массива литологического класса.

В качестве принадлежности анализируемого класса к эталонному классу используется критерий χ^2 (критерий Пирсона). Принятие решения о принадлежности анализируемого кластера к эталонному объекту осуществляется по минимальному значению величины

$$\chi^2 = \sum_{m=1}^{M} \frac{(\eta_m - \eta_{m\,теор})^2}{\eta_{m\,теор}},$$

где M – число областей разбиения, на которые делится все пространство объектов для анализируемого кластера; m – индекс номера области; η_m и $\eta_{m\,теор}$ – эмпирическая и теоретическая вероятность распределения объектов для эталонного класса в области с индексом m.

Если объем выборки равен n, то

$$\chi^2 = \sum_{m=1}^{M} \frac{(\eta_m - np_m)^2}{np_m},$$

где p_m – теоретическая плотность вероятности попадания объекта в область m.

Принцип работы блока идентификации по χ^2 заключается в следующем. Все пространство признаков разбивается на M участков так, чтобы в каждую область попало k объектов, затем производится вычисление пределов изменения параметра в фиксированной области, по которой определяется

число значений параметра эталонного массива, попавшего в эти пределы, и, наконец, вычисляется значение критерия χ_p для каждого из эталонных литологических классов.

В результате выполнения алгоритма идентификации по критерию χ^2 получается одномерный массив $[\chi_p]$, минимизация которого определяет принадлежность анализируемого кластера к эталонному литологическому классу.

Блок идентификации по D_{min} выполняет процедуру идентификации для кластеров с размером $I_l \leq 50$. Процедура идентификации выполняется в следующей последовательности: вычисляется величина математического ожидания $\overline{p_{jnk}(l)}$ для анализируемого кластера, затем расстояние между каждым эталонным классом и анализируемым кластером, после чего формируется одномерный массив расстояний χ_p.

Блок вычисления принадлежности кластера к эталонному литологическому классу предназначен для принятия решений об отнесении анализируемого кластера к конкретному литологическому классу и вычисления параметров, определяющих качество идентификации.

В результате работы блока вычисляется первый минимум $\chi_{p\,min}$ и определяется номер класса эталонного массива, к которому этот минимум относится, далее рассчитывается второй минимум и вычисляется переметр, определяющий качество классификации $\mu_l = \chi_{p\,min1}/\chi_{p\,min2}$, где $\chi_{p\,min1}$ и $\chi_{p\,min2}$ – первое и второе наименьшие значения критерия χ^2.

ЛИТЕРАТУРА

1. Белл Д.Л., Портер У.Дж. Возможности дистанционной классификации осадков по характеристикам отраженных дном акустических сигналов. – В кн.: Акустика морских осадков / Под ред. Л.Хэмптона. М., Мир, 1977.

2. Дюран Б., Оделл П. Кластерный анализ. М., Статистика, 1977.

3. Миленький А.В. Классификация сигналов в условиях неопределенности. М., Сов.радио, 1975.

4. Смит Д.Т. Акустические и механические свойства морских осадков. – В кн. Акустика морских осадков / Под ред. Л.Хэмптона. М., Мир, 1977.