лированных помехах на входе и выходе // Пробл. передачи информ. 1986. Т.23, вып.2. С.61-74.

8. *Кушнир А.Ф.* Упрощенные асимптотически оптимальные оценки // Тез. докл. IV Междунар. конф. по теории вероятностей и математической статистике. Вильнюс, 1985. Т.II, С.94-96.

9. *Кушнир А.Ф., Лапшин В.М.* Параметрические методы анализа многомерных временных рядов. М.: Наука, 1986. 243с.

10. *Кушнир А.Ф.* Параметрические методы статистического анализа геофизических временных рядов: Дис. ... докт. физ.-мат. наук. 01.04.12. М.: ИФЗ АН СССР, 1989. 310с.

11. *Рабинер Л., Гоулд Б.* Теория и применение цифровой обработки сигналов. М.: Мир, 1978. 848с.

12. *Бриллинджер Д.* Временные ряды. Обработка данных и теория. М.: Мир, 1980. 536с.

13. *Андерсон Т.* Статистический анализ временных рядов. М.: Мир, 1976. 755с.

14. *Малютов М.Б.* Нижняя граница для средней длительности последовательно планируемого эксперимента // Изв. вузов. Математика. 1983. N 11. С.19-41.

15. *Nussbaum M.* An asymptotic minimax risk for estimation of a linear functional relationship//J. Multivar. Anal. 1984. Vol.14, N 3. P.300-314.

16. *Ибрагимов И.А., Хасьминский Р.З.* Асимптотическая теория оценивания. М.: Наука, 1979. 527с.

УДК 550.34

М.Н.Жижин

СИНТАКСИЧЕСКИЙ АНАЛИЗ ЗАПИСЕЙ СИЛЬНЫХ ДВИЖЕНИЙ. II. СЛУЧАЙ БЕСКОНЕЧНОГО АЛФАВИТА

M.N.Zhizhin

SYNTACTIC ANALYSIS OF STRONG MOTION DATA. II. THE CASE OF INFINITE GRAMMARS

In this work the Syntactic Pattern Recognition Scheme is constructed for the purpose of classification of multichannel seismic records in case of infinite cardinality of grammars, including the tools for verification of the classification reliability. It is applied for classification of three-component strong motion records from different geological regions.

Введение

Синтаксическое распознавание предполагает представление образов в виде предложений, составленных из символов некоторого алфавита с помощью набора грамматических правил (грамматики). Различные грамматики порождают образы-предложения различных классов. Обучение

состоит в нахождении соответствия между грамматиками и классами объектов. Собственно распознавание объекта сводится к определению (в стохастическом случае наиболее вероятной) грамматики, его породившей [1]. Если грамматические правила заранее неизвестны, то для каждого класса оценивается возможность того, что исследуемый объект порожден той же грамматикой, что и представители этого класса из материала обучения. Здесь часто применяется техника динамического программирования, использующая алгоритм Витерби [2].

Этот формализм, предполагающий конечность исходного алфавита, вполне естественен, например, при сравнении структуры генов [3]. При анализе временных рядов конечность алфавита не является необходимым требованием для адекватного описания структуры образа. С другой стороны, применение динамического программирования остается эффективным, если задача предполагает локальные нелинейные изменения скорости процесса (например, длительность фонем в распознавании речи). В результате возможна двоякая интерпретация алгоритма либо в терминах "динамического соотнесения по времени" (**dynamic time warping**), либо по-прежнему в терминах грамматического разбора, устойчивого при ошибках (**error detecting/correcting parsing**) [4], если опустить условие конечности алфавита.

Данная работа продолжает [5], где мы проводили синтаксическое распознавание волновых форм сильных движений (акселерограмм) из одного геологического региона с целью найти структурные различия в записях серий сейсмических событий. В работе строится алгоритм синтаксического распознавания многоканальных сейсмических записей для грамматик с бесконечным алфавитом, изложены методы его верификации, получены результаты распознавания волновых форм трехкомпонентных записей сильных движений, зарегистрированных в различных геологических регионах. Алгоритм [5] является частным случаем предлагаемой схемы, использующим кластеризацию фрагментов записей при параметризации вертикальных компонент.

Записи сильных движений в ближней зоне играют все большую роль в качестве источника информации для детального изучения очага землетрясения. В нулевом приближении акселерограмма содержит S-волны продолжительностью порядка очаговой длительности (5÷20 с) и моделируется как результат полосовой фильтрации белого шума. Параметры полосового фильтра определяются формой очагового спектра и высокочастотным затуханием в среде. Этого достаточно в рамках инженерной сейсмологии, но мало что добавляется к информации об очаге, получаемой при телесейсмических наблюдениях.

Представление о структуре очага как системе субочагов [6] (возможно, иерархической) позволяет рассматривать порожденные им сильные движения в первом приближении как суперпозицию "элементарных" волновых пакетов от этой системы субочагов и их отражений от рассеивателей на пути до регистрирующего прибора. Линейные размеры очага сравнимы с расстоянием до пункта регистрации, что приводит к необходимости рассматривать трехмерную картину распространения и рассеяния направленного сейсмического излучения.

Амплитудные характеристики "элементарного" волнового пакета определяются в основном величиной сброшенного напряжения и расстоянием до пункта регистрации. Спектральный состав зависит также от линейных размеров субочага, скорости распространения разрыва и характера высокочастотного затухания в среде [7]. Суперпозиция волновых пакетов определяется распределением субочагов на разломе, последовательностью их разрушения и расположением регистрирующего прибора относительно разлома. Заметим, что даже для субочагов с характерным размером 100÷800 м наблюдаются подобные суперпозиции [8], что может быть объяснено скачкообразным сбросом напряжений на отдельных неровностях в очаге.

Таким образом, сложность волновых форм сильных движений во многом определяется детальной структурой очага, что дает основание для интерпретации результатов сравнения структуры волновых форм в терминах различий очагов землетрясений. Формализм синтаксического распознавания образов представляется естественным на начальном этапе исследований. Получаемая в этом случае мера близости волновых форм является обобщением уже используемых в подобных задачах кросскорреляций записей [9,10] и сонограмм [11].

1. Алгоритмы SPARS

В этом разделе изложены алгоритмы, образующие Схему Синтаксического Распознавания Образов (SPARS - *Syntactic Pattern Recognition Scheme*). С помощью SPARS волновые формы сейсмических записей разбиваются на заранее известное число классов.

1.1. Фрагментация и параметризация сейсмических записей. В SPARS используются все три компоненты сейсмической записи, но обработка вертикальных и горизонтальных компонент проводится раздельно. Это связано с тем, что мы применяем SPARS в основном для анализа записей сильных движений, при изучении которых традиционно больше внимания уделяется горизонтальным компонентам записи. Изменения алгоритма, вычисляющего параметры горизонтальных компонент, на

случай вертикальных компонент или трехкомпонентной записи могут быть легко получены.

Пусть $b(t_i)$, $c(t_i)$ - две ортогональные компоненты сейсмической записи $\xi = \{(a(t_i),b(t_i),c(t_i)),\ i=1,2,\ldots,M\}$, где $a(t_i)$, $b(t_i)$, $c(t_i)$ - величины вертикального и двух горизонтальных ускорений, наблюденные в момент времени t_i. Интервал дискретизации $\Delta t=t_{i+1}-t_i$ фиксирован на протяжении всей записи. Для рассматриваемых ниже записей сильных движений возможны значения $\Delta t=0,005$ с, $\Delta t=0,01$ с, чаще всего $\Delta t=0.02$ с. Чтобы сделать исходные данные более однородными, в SPARS используются только первые N отсчетов $a(t_i)$, $i=1,2,\ldots,N$, где $N(\Delta t)\leq M$ выбирается в зависимости от приложений и является *свободным параметром* алгоритма.

Зафиксируем временной интервал $\delta t=k\Delta t$, где $k=k(\Delta t)$ целое и $\Delta t\ll\delta t\ll t_N - t_1$.

О п р е д е л е н и е 1. Мы называем *горизонтальным фрагментом* F_H с началом в момент времени t_j и длительностью δt следующий двумерный временной ряд:

$$F_H = F_H(t_j,\delta t)=\{(b(t_i),c(t_i)):\ t_i\in[t_j,t_j+\delta t)\}.$$

Длительность фрагментов записей δt является *свободным параметром* алгоритма. В дальнейшем используются значения $\delta t=1,28$ с и $\delta t = 2.56$ с. Например, в [5] запись ξ при общей длительности $t_N - t_1 = 20,48$ с и длительности фрагментов $\delta t=1,28$ с представлена в виде объединения не более 16 неперекрывающихся фрагментов:

$$\xi_H = \{(a(t_i)b(t_i)),\ i=1,2,\ldots,1024\} = \bigcup_{k=0}^{15} F_H(t_1 + k\delta t,\delta t). \qquad (1)$$

Каждый фрагмент записи содержит 64 отсчета при интервале дискретизации $\Delta t=0,02$ с.

Наряду с (1) возможно следующее представление записи в виде объединения множества перекрывающихся фрагментов:

$$\xi_H = \{(b(t_i),c(t_i)):\ i=1,2,\ldots,N\} \to \bigcup_{k=0}^{2N\Delta t/\delta t-2} F_H(t_1+ k\delta t/2,\delta t). \qquad (2)$$

Для каждого фрагмента $F_H = F_H(t_j,\delta t)$ вычисляются параметры $X_E(F_H)$, $X_Z(F_H)$ и $X_P(F_H)$, оценивающие энергию, доминирующую частоту и изменение поляризации сигнала на этом фрагменте. Функция

$$(\ln(ENG))(F_H)=\ln \sum_{t_i\in F_H} (b^2(t_i)+c^2(t_i))/ord(F_H) \qquad (3)$$

оценивает логарифм энергии, выделившейся на фрагменте и является

тривиальным обобщением аналогичной функции для вертикальных компонент [5].

Несколько сложнее обобщение на двумерный случай числа нуль-пересечений при оценке доминирующей частоты. Для этого мы определим методом наименьших квадратов каноническую ось 1 на горизонтальной плоскости и найдем число пересечений нуля, совершаемых проекцией колебания маятника записывающего прибора на эту ось.

Для каждого фрагмента $F_H(t_i,\delta t)$, являющегося конечным множеством точек на горизонтальной плоскости, найдем прямую 1, сумма квадратов проекций на которую для всех точек из F_H минимальна. Следуя [12], направление 1 можно определить с помощью собственного вектора ковариационной матрицы координат точек плоскости из F_H:

$$\mathcal{S} = \begin{pmatrix} (N-1)^{-1}\Sigma_{t_i\in F_H}(b(t_i)-b_{AV})^2, & (N-1)^{-1}\Sigma_{t_i\in F_H}(b(t_i)-b_{AV})(c(t_i)-c_{AV}) \\ (N-1)^{-1}\Sigma_{t_i\in F_H}(b(t_i)-b_{AV})(c(t_i)-c_{AV}), & (N-1)^{-1}\Sigma_{t_i\in F_H}(c(t_i)-c_{AV})^2 \end{pmatrix}, \quad (4)$$

имеющим наибольшее собственное значение. Здесь b_{AV} и c_{AV} обозначают средние значения соответствующей горизонтальной компоненты на фрагменте. Они задают новое начало координат. через которое должна проходить прямая 1. Если через b_{EIG} и c_{EIG} обозначить компоненты нормированного собственного вектора матрицы $\mathcal{S}$ с максимальным собственным значением. то проекция $h(t_i)$ горизонтального движения на ось 1 задается формулой

$$h(t_i)=Pr_1(b(t_i),c(t_i))=\langle(b(t_i)-b_{AV},c(t_i)-c_{AV}),(b_{EIG},c_{EIG})\rangle, \quad (5)$$

где $\langle x,y\rangle$ обозначает скалярное произведение.

Определим число нуль-пересечений следующей проекции колебаний на прямую 1 формулой

$$(ZCR)(F_H)=ord\left\{i:\ sgn(h(t_i))\neq sgn(h(t_{i+1})),\ t_i\in F_H\right\}. \quad (6)$$

Для интервалов дискретизации $\Delta t\neq 0.02$ с в формулу (6) вносятся соответствующие поправки.

В добавление к (3) и (6) для горизонтальных компонент мы рассматриваем также параметр. отражающий изменение поляризации сигнала в горизонтальной плоскости. Поляризация S-волн в ближней зоне землетрясения может быть использована для оценки параметров очага и даже для реконструкции истории распространения разрыва [13]. Для

каждого фрагмента $F_{H_k} = F_H(t+k\delta t/2, \delta t)$ вычислим ковариационную матрицу (4) и найдем ее собственные значения и векторы. Затем определим углы α_k и $(180^\circ - \alpha_k)$ между собственными векторами l_k и l_{k+1} двух соседних фрагментов сейсмической записи. Значение параметра теперь определим по формуле

$$(POL)(F_{H_k}) = \min(\alpha_k, 180^\circ - \alpha_k). \qquad (7)$$

Обозначим через $P(F)$ любую из функций, определенных формулами (3), (6), (7), и рассмотрим среднее $M(P) = \Sigma P(F)/L$ по всем фрагментам $F \in \xi$ всего множества исследуемых записей $\xi \in \mathfrak{N}$, так что $L = ord(\{F: F \in \xi, \xi \in \mathfrak{N}\})$ является общим числом всех фрагментов всех рассматриваемых записей. Для оценки дисперсии параметра P воспользуемся формулой:

$$D(P) = \Sigma(P(F) - M(P))^2/(L-1).$$

Окончательно, каждому фрагменту F_{H_k} (кроме последнего) поставим в соответствие значение трех параметров:

$$X_E(F_H) = [D(\ln(ENG))]^{-0,5}[(\ln(ENG))(F_H) - M(\ln(ENG))], \qquad (8)$$

$$X_Z(F_H) = (D(ZCR))^{-0,5}((ZCR)(F_H) - M(ZCR)), \qquad (9)$$

$$X_P(F_H) = (D(POL))^{-0,5}((POL)(F_H) - M(POL)). \qquad (10)$$

Последний фрагмент каждой записи используется только для вычисления параметра $X_P(F_H)$. Из определения параметров $X_E(F)$, $X_Z(F)$ и $X_P(F)$ следует, что

$$M(X_E) = M(X_Z) = M(X_P) = 0; \quad D(X_E) = D(X_Z) = D(X_P) = 1.$$

Таким образом, каждый фрагмент $F_H \in \xi_H$ представлен вектором $(X_E(F_H), X_Z(F_H), X_P(F_H))$, и каждая запись представлена матрицей параметров фрагментов

$$\xi_H \to \begin{pmatrix} X_E(F_{H_1}), \ldots, X_E(F_{H_{l-1}}) \\ X_Z(F_{H_1}), \ldots, X_Z(F_{H_{l-1}}) \\ X_P(F_{H_1}), \ldots, X_P(F_{H_{l-1}}) \end{pmatrix}, \qquad (11)$$

где l - число фрагментов записи ξ_H. Матрицу (11) мы называем *образом* записи ξ_H.

1.2. Синтаксическое расстояние между записями. При вычислении расстояния в SPARS исходными данными служат представления записей

в виде матриц параметров фрагментов (11). Обобщим это представление и рассмотрим случай, когда фрагмент $F_i \in \xi$ описан вектором из p параметров $X(F_i)$, а Ω - вектор весов параметров фрагментов, отражающий информативность каждого из них (*свободные параметры алгоритма*):

$$X(F_i)=\begin{pmatrix} X_1(F_i) \\ \vdots \\ X_r(F_i) \\ \vdots \\ X_p(F_i) \end{pmatrix}, \qquad \Omega=\begin{pmatrix} \omega_1 \\ \vdots \\ \omega_r \\ \vdots \\ \omega_p \end{pmatrix}, \tag{12}$$

Тогда каждая запись представима в виде 1×p матрицы

$$\xi \to mat_p(\xi)=(X(F_1),\ldots,X(F_l)), \tag{13}$$

где $l=l(\xi)$ - число фрагментов записи ξ.

Введем три вида *допустимых преобразований* матриц вида (12).

О п р е д е л е н и е 2. Удаление столбца матрицы определим формулой

$$T_D(X(F_j))\xi = (X(F_1),\ldots,X(F_{j-1}),X(F_{j+1}),\ldots,X(F_l)). \tag{14}$$

Это преобразование определено только для матриц $mat_p(\xi)$, j-й столбец которых равен $X(F_j)$. В этом случае $l(T_D\xi)=l(\xi)-1$.

О п р е д е л е н и е 3. Вставку столбца

$$a_j=\begin{pmatrix} a_1 \\ \vdots \\ a_r \\ \vdots \\ a_p \end{pmatrix} \tag{15}$$

определим формулой

$$T_I(X(F_j))\xi=(X(F_1),\ldots,X(F_j),a_j,X(F_{j+1}),\ldots,X(F_l)). \tag{16}$$

Это преобразование определено для матриц с числом столбцов $l(\xi)\geq j-1$. В этом случае

$$l(T_I\xi) = l(\xi)+1.$$

О п р е д е л е н и е 4. Замену столбца $X(F_j)$ столбцом a_j вида (15) определим формулой

$$T_S(X(F_j),a_j)=(X(F_1),\ldots,X(F_{j-1}),a_j,X(F_{j+1}),\ldots,X(F_l)). \qquad (17)$$

Это преобразование определено только для матриц $mat_p(\xi)$, j-й столбец которых равен $X(F_j)$. В этом случае

$$l(T_S\xi) = l(\xi).$$

Заметим, что возможна иная терминология при определении допустимых преобразований. Мы исходим из представления записи ξ в виде последовательности фрагментов F_j, каждый из которых, в свою очередь, представлен вектором параметров $X(F_j)$. Поэтому удаление j-го столбца из матрицы $mat_p(\xi)$ эквивалентно удалению фрагмента F_j из записи ξ, вставка столбца в матрицу эквивалентна вставке фрагмента в запись, и т.д.

Для преобразований (14), (16), (17) введем следующие веса, зависящие от векторов Ω, $X(F_j)$ и a_j:

$$\text{для } T_D(X(F_j)): \qquad w_D(X(F_j)) = \sum_{i=1}^{p} \omega_i^2 X_i(F_j)^2; \qquad (18)$$

$$\text{для } T_I(X(F_j)): \qquad w_I(a_j) = \sum_{i=1}^{p} \omega_i^2 a_i^2; \qquad (19)$$

$$\text{для } T_S(X(F_j),a_j): \qquad w_D(X(F_j),a_j) = \sum_{i=1}^{p} \omega_i^2 (X_i(F_j)-a_i)^2. \qquad (20)$$

О п р е д е л е н и е 5. Для пары записей $\xi,\tau \in \mathfrak{A}$ назовем *путем* $J(\xi\to\tau)$ с началом в ξ и концом в τ любую композицию из n преобразований:

$$J = J(\xi\to\tau) = T_1\circ T_2\circ\ldots\circ T_n, \qquad n>1, \qquad (21)$$

такую, что T_i - допустимые преобразования T_D, T_I или T_S, и $(T_1\circ T_2\circ\ldots\circ T_n)\xi = \tau$.

О п р е д е л е н и е 6. Назовем *длиной пути* $len(J(\xi\to\tau))$ сумму весов допустимых преобразований, входящих в композицию (21):

$$len(J(\xi\to\tau))=\Sigma w_i, \qquad (22)$$

где w_i - вес преобразования T_i, вычисленный по формулам (18)-(20).

Очевидно, что даже длина $len(J(\xi\to\xi))$ может быть не равна нулю. Например,

$$len(J(\xi\to\xi)) = len(T_D(X(F_1))\circ T_I(X(F_1))) = 2\sum_{i=1}^{p} \omega_i^2 X_i(F_1)^2 > 0,$$

если для некоторого i имеем $\omega_i X_i(F_1)\neq 0$.

О п р е д е л е н и е 7. *Синтаксическое расстояние* между записями ξ и η задается функцией

$$d(\xi,\tau) = \min \mathrm{len}(J(\xi\to\tau)), \qquad (23)$$

где минимум берется по всем путям с началом в ξ и концом в τ.

У т в е р ж д е н и е 1. Функция $d(\xi,\tau)$, определенная формулой (23), обладает всеми свойствами расстояния.

Из формул (21)-(23) видно, что $d(\xi,\tau)=d(\tau,\xi)$ и $d(\xi,\tau)=0 \Leftrightarrow \xi=\tau$. Условие

$$d(\xi,\eta)\le d(\xi,\tau)+d(\tau,\eta) \text{ для любых } \xi,\tau,\eta \in \mathfrak{U}$$

также выполнено. Чтобы доказать это, рассмотрим путь $J_{\xi\tau}$, дающий минимум (23) и, следовательно, определяющий расстояние $d(\xi,\tau)$ по формуле (23). Пусть $J_{\tau\eta}$ будет аналогичным путем из τ в η. Тогда композиция $J_{\xi\tau}\circ J_{\tau\eta} = J(\xi\to\eta) = J$ является путем из ξ в η и

$$d(\xi,\eta) \le \mathrm{len}(J_{\xi\tau}\circ J_{\tau\eta}) = \sum_{T_i\in J} w_i = \sum_{T_i\in J_{\xi\tau}} w_i + \sum_{T_i\in J_{\tau\eta}} w_i =$$

$$= \mathrm{len}(J_{\xi\tau}) + \mathrm{len}(J_{\tau\eta}) = d(\xi,\tau) + d(\tau,\eta).$$

1.3. Классификация записей и оценка ее качества. В результате предыдущих шагов SPARS исследуемые записи $\xi \in \mathfrak{U}$ были представлены в виде матриц параметров фрагментов, и было определено синтаксическое расстояние $d(\xi,\eta)$ между записями $\xi,\eta \in \mathfrak{U}$. В этом разделе мы изложим процедуру классификации записей на основе этих данных, а также два контрольных эксперимента: метод "складного ножа" и случайное перемешивание материала обучения.

1.3.1. Решающие правила. Предположим, что множество записей $\mathfrak{U}$ разбито на m+1 непересекающихся подмножеств $\mathfrak{U}_1, \mathfrak{U}_2, \ldots, \mathfrak{U}_{m+1}$, представляющих заданные классы сейсмических записей, включая подмножество $\mathfrak{U}_{m+1}$ записей, класс которых заранее неизвестен. Такими классами могут быть, например, записи, порожденные землетрясением и ядерным взрывом [14], записи, порожденные форшоком, главным толчком и афтершоком [5], или событиями из разных сейсмических регионов. Предположим также, что задано разбиение $\mathfrak{U} = \mathfrak{U}_L \cup \mathfrak{U}_T$ на *базу знаний* $\mathfrak{U}_L$ и *пробное множество* записей $\mathfrak{U}_T$; в общем случае $\mathfrak{U}_L \cap \mathfrak{U}_T \neq \emptyset$. Воспользуемся обобщением известного правила ближайшего соседа [15], чтобы для каждой записи $\xi\in\mathfrak{U}_T$ из пробного множества найти наиболее близкий класс $\mathfrak{U}_i$ или отнести ее к неопределенному классу $\mathfrak{U}_{m+1}$, если база знаний не полна. Эту процедуру назовем *решающим*

правилом. В результате мы получим дизъюнктное разбиение пробного множества $\mathfrak{U}_T = \coprod_{j=1}^{m+1} \mathfrak{C}_j$ на подмножества $\mathfrak{C}_j$, отвечающие заданным классам записей в базе знаний $\mathfrak{U}_L$.

О п р е д е л е н и е 8. Пусть известны расстояния от записи ξ до всех записей из множества $\mathfrak{U}_L$. Подмножество $\mathfrak{R}_K(\mathfrak{U}_L,\xi) \subseteq \mathfrak{U}_L$ назовем множеством *K-ближайших соседей* записи ξ, если $\xi \notin \mathfrak{R}_K(\mathfrak{U}_L,\xi)$, $\mathrm{ord}(\mathfrak{R}_K(\mathfrak{U}_L,\xi))=K$, $0<K\le r=\mathrm{ord}(\mathfrak{U}_L)$ и оно ближайшее к ξ среди всех K-элементных подмножеств множества $\mathfrak{U}_L$:

$$D(\xi,\mathfrak{R}_K(\mathfrak{U}_L,\xi)) := \frac{1}{K}\sum_{i=1}^{K} d(\xi,\eta_i) = \min_{\substack{\mathrm{ord}(G)=K \\ G\subseteq\mathfrak{U}_L}} D(\xi,G).$$

В общем случае выбор K-ближайших соседей неоднозначен, но на практике вероятность совпадений достаточно мала, и в случае неоднозначности допустим случайный выбор.

О п р е д е л е н и е 9. *Решающее правило K-ближайших соседей* относит запись ξ к классу, представленному в базе знаний $\mathfrak{U}_L$ подмножеством $\mathfrak{U}_i$, имеющим наибольшее число представителей среди K-ближайших соседей записи ξ :

$$\mathrm{ord}(\mathfrak{U}_j \cap \mathfrak{R}_K(\mathfrak{U},\xi)) < \mathrm{ord}(\mathfrak{U}_i \cap \mathfrak{R}_K(\mathfrak{U},\xi)) \quad \forall\, j\neq i.$$

В случае неоднозначности запись ξ относится к неопределенному классу. Число K является *свободным параметром* алгоритма.

О п р е д е л е н и е 10. *Решающее правило K-средних расстояний* относит запись ξ к классу, представленному в базе знаний $\mathfrak{U}_L$ подмножеством $\mathfrak{U}_i$, если среднее расстояние от K-ближайших соседей из $\mathfrak{U}_i \cap \mathfrak{U}_L$ до ξ минимально, т.е. для любого $j\neq i$:

$$D(\xi,\mathfrak{R}_K(\mathfrak{U}_i\cap\mathfrak{U}_L,\xi)) := \frac{1}{K}\sum_{\eta_1\in\mathfrak{R}_K(\mathfrak{U}_i\cap\mathfrak{U}_L,\xi)} d(\xi,\eta_1) < D(\xi,\mathfrak{R}_K(\mathfrak{U}_j\cap\mathfrak{U}_L,\xi)) := \frac{1}{K}\sum_{\eta_1\in\mathfrak{R}_K(\mathfrak{U}_j\cap\mathfrak{U}_L,\xi)} d(\xi,\eta_1).$$

В случае неоднозначности запись ξ относится к неопределенному классу. Число K является *свободным параметром* алгоритма.

Заметим, что при K=1 определения (9) и (10) приводят к одному и тому же правилу ближайшего соседа. При K≠1 правило K-средних расстояний значительно реже дает неопределенный ответ, чем правило K-ближайших соседей и представляется более устойчивым при зашумленном материале обучения.

1.3.2. Функция ошибок классификации используется в SPARS для оценки качества классификации и сравнения результатов при различных значениях свободных параметров. В общем случае, пусть задана

матрица потерь $\mathfrak{L} = \mathrm{mat}(l_{ij})$, $i,j=1,2,\ldots,m+1$, если при ошибке классификации запись из класса i отнесена в класс j, причем значения $i, j=m+1$ означают, что класс записи неопределен.

Обозначим через $I_{\mathfrak{G}}(\xi)$ индикаторную функцию множества $\mathfrak{G}$,

$$I_{\mathfrak{G}}(\xi)=\begin{cases}1, & \text{если } \xi\in\mathfrak{G};\\ 0, & \text{если } \xi\notin\mathfrak{G}.\end{cases}$$

О п р е д е л е н и е 11. Пусть заданы матрица потерь $\mathfrak{L} = \mathrm{mat}(l_{ij})$ и два дизъюнктные (возможно, совпадающие) разбиения пробного множества записей $\mathfrak{U}_T$:

i) $\mathfrak{U}_T = \bigcup_{i=1}^{m+1} \mathfrak{F}_i$: $\xi \in \mathfrak{F}_i$ означает, что *фактически* запись принадлежит i-му классу;

ii) $\mathfrak{U}_T = \bigcup_{j=1}^{m+1} \mathfrak{C}_j$: $\xi \in \mathfrak{C}_j$ означает, что *в результате* классификации запись отнесена в j-й класс.

Функция ошибок классификации F_{ERR} вычисляется по формуле:

$$F_{ERR}=\sum_{\xi\in\mathfrak{U}}\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} l_{ij} I_{\mathfrak{F}_i}(\xi) I_{\mathfrak{C}_j}(\xi). \qquad (24)$$

Часто матрица потерь имеет простой вид, а именно

$$l^*_{ij} = \begin{cases}0, & \text{если } i=j,\ i\le m+1,\ j\le m;\\ 1 & \text{в противном случае.}\end{cases}$$

В этом случае значением функции ошибок классификации будет число записей, для которых заранее известно, к какому классу они принадлежат, но в результате классификации попавших в другой (возможно, неопределенный) класс. Обозначим функцию ошибок с такой матрицей потерь F^*_{ERR}. Очевидно, $F^*_{ERR}/\mathrm{ord}(\mathfrak{U}_T\setminus\mathfrak{F}_{m+1})\times 100$ даст относительную ошибку классификации, которую удобно использовать для сравнения результатов распознавания при разных объемах исходных данных.

Если фактически принадлежность к определенному классу известна для малого множества записей и дизъюнктное разбиение $\mathfrak{U} = \mathfrak{U}_L \cup \mathfrak{U}_T$ делает базу знаний $\mathfrak{U}_L$ непредставительной, мы оцениваем значение функции ошибок, классифицируя каждую запись $\xi \in \mathfrak{U}$ с базой знаний $\mathfrak{U}_L = \mathfrak{U}\setminus\{\xi\}$ и используя все множество $\mathfrak{U}$ с полученным на нем таким образом дизъюнктным разбиением $\mathfrak{U} = \bigcup_{j=1}^{m+1} \mathfrak{C}_j$ в качестве пробного множества в формуле (24).

1.3.3. Метод "складного ножа" (**jackknife**) используется в SPARS для оценки устойчивости результатов классификации. При этом из базы знаний исключаются по одной все входящие в нее записи ($\mathfrak{U}'_L = \mathfrak{U}_L \setminus \{\eta\}$ для любой $\eta \in \mathfrak{U}_L$), и каждый раз проводится классификация пробного множества $\mathfrak{U}_T$. Если при исключении некоторой записи η_0 получится меньшее значение функции ошибок, то эта процедура повторяется уже для базы знаний без этой записи (т.е. $\mathfrak{U}''_L = (\mathfrak{U}_L \setminus \{\eta_0\}) \setminus \{\eta\}$ для любой $\eta \in \mathfrak{U}_L \setminus \{\eta_0\}$) и т.д., пока возможно уменьшение функции ошибок.

В результате применения метода "складного ножа" в базе знаний выделяется подмножество записей $\{\eta_0, \eta_1, \ldots, \eta_j\} \subseteq \mathfrak{U}_L$, последовательное исключение которых улучшает результаты классификации пробного множества $\mathfrak{U}_T$. Они могут быть подвергнуты дополнительному изучению для подтверждения необходимости их включения в базу знаний.

1.3.4. "Перемешивание" материала обучения. В этом контрольном эксперименте при классификации используется база знаний основного варианта, но со случайным разбиением на классы того же объема (т.е. $\mathfrak{U}_L = \bigcup_{i=1}^{m+1} \mathfrak{F}'_i$, где $\mathbf{ord}(\mathfrak{F}'_i) = \mathbf{ord}(\mathfrak{F}_i)$, но $\mathfrak{F}_i \neq \mathfrak{F}'_i$).

Случайное разбиение базы знаний $(\mathfrak{F}'_1, \mathfrak{F}'_2, \ldots, \mathfrak{F}'_{m+1})$ мы получаем из разбиения $(\mathfrak{F}_1, \mathfrak{F}_2, \ldots, \mathfrak{F}_{m+1})$ простым N-кратным перемешиванием. На первом шаге случайно выбираются две записи $\xi, \eta \in \mathfrak{U}_L$ и переставляются номера классов, к которым они принадлежат (если $\xi \in \mathfrak{F}_i$ и $\eta \in \mathfrak{F}_j$, то $\xi \in \mathfrak{F}'_j$ и $\eta \in \mathfrak{F}'_i$). На N-м шаге совершается перестановка пары $\xi_N, \eta_N \in \mathfrak{U}_L$. В приложениях N берется достаточно большим, чтобы можно было считать полученное в результате этой процедуры разбиение случайным. Далее проводится классификация записей так же, как в основном варианте.

Функция ошибок классификации усредняется по нескольким N-кратным перемешиваниям и среднее значение сравнивается со значением функции ошибок основного варианта.

2. Классификация волновых форм записей сильных движений из различных геологических регионов

В инженерной сейсмологии возникает проблема сравнения форм огибающих, длительностей, спектров реакций и других параметров акселерограмм из различных геологических регионов. Эти региональные различия в параметрах записей могут быть следствием как различий в механизмах землетрясений, скоростях распространения разрыва, характере распределения неоднородностей на разломе, так и различий в

геологических средах. На характер записей оказывают влияние также особенности регистрирующей аппаратуры и метода оцифровки и коррекции аналоговых записей. Мы полагаем, что в среднем они дают меньший эффект по сравнению с геологическими особенностями, так как мы рассматриваем откорректированные записи.

Применим изложенную выше схему SPARS для проверки гипотезы о том, что записи в ближней зоне калифорнийских и итальянских землетрясений имеют достаточно отличительных признаков для классификации их волновых форм в зависимости от региона.

2.1. Исходные наборы итальянских и калифорнийских акселерограмм. Мы отобрали 40 трехкомпонентных акселерограмм (20 из Италии и 20 из Калифорнии), записанных на расстоянии до 100 км от 16 неглубоких ($h \leq 25$ км) землетрясений (9 в Италии, 7 в Калифорнии) с магнитудами $M=3,5 \div 7$ (рис.1). Эти данные для каждой записи приведены в табл.1. За исключением землетрясения Ancona, сейсмические события в обоих регионах сравнимы по магнитуде. Исследуемые землетрясения в обоих регионах имеют различные механизмы, причем для выборки из Калифорнии преобладают горизонтальные подвижки, тогда как в Италии часто присутствует вертикальная компонента.

2.2. Постановка задачи и параметризация записей. Подтверждением гипотезы о различии волновых форм сильных движений мы бы считали положительный ответ на вопрос:

Возможно ли устойчиво классифицировать каждую из 40 анализируемых записей, используя в качестве материала обучения в SPARS параметры волновых форм оставшихся 39 записей и информацию об их геологическом регионе?

Для ответа на этот вопрос в соответствии с п.1.1 проведем фрагментацию и параметризацию акселерограмм из табл.1. Чтобы оценить устойчивость классификации относительно длительности фрагментов, параметризация проводилась для перекрывающихся фрагментов (2) с длительностью $\delta t=1,28$ с или $\delta t=2,56$ с. Общее число перекрывающихся фрагментов каждой записи для $\delta t=1,28$ с приведено в табл.1. Очевидно, для оценки общей длительности параметризованной записи следует воспользоваться формулой $T = (N_f/2+0,5)\delta t$, где N_f - число фрагментов, δt - длительность фрагмента.

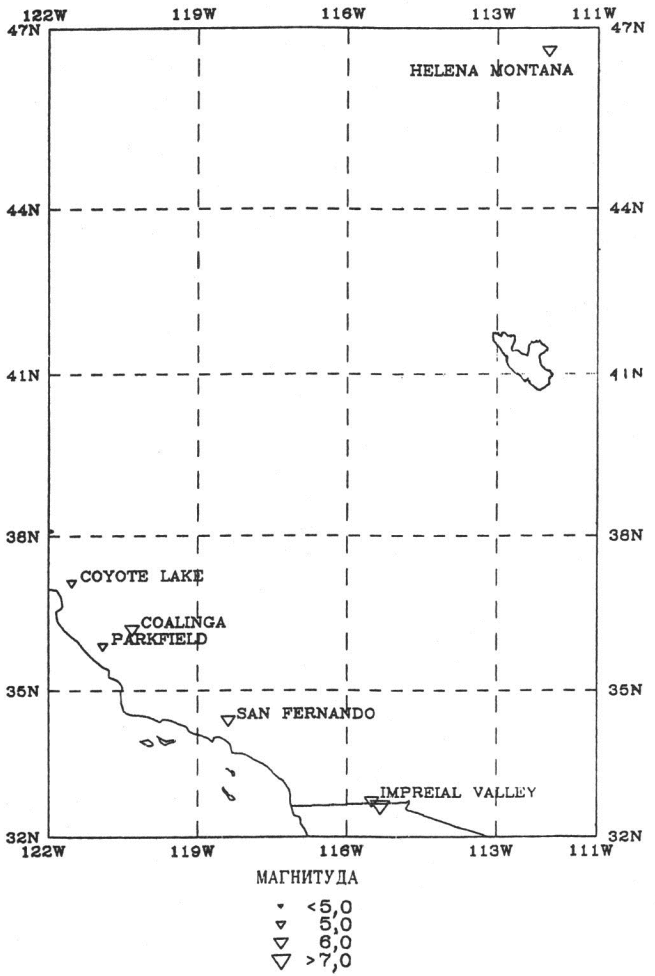
2.3. Классификация и оценка ее стабильности. Свободными параметрами алгоритма вычисления попарных расстояний между записями в SPARS согласно п.1.2. являются максимальное число фрагментов записи и веса параметров фрагментов. Варьируя свободные параметры, мы

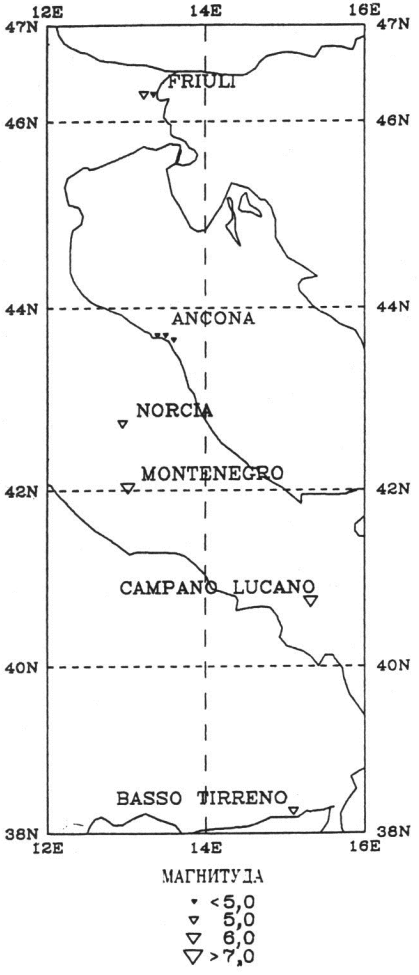
Таблица 1

Записи сильных движений калифорнийских и итальянских землетрясений, использованные при распознавании

Запись	Число фрагментов	Эпицентральное расстояние, км	M	Глубина очага, км	Механизм очага	Дата землетрясения	Землетрясение
			ИТАЛИЯ				
basmes1	37	35	5,8	24	-	15.04.78	BASSO TIRENO
basnas1	44	32	5,8	24	-	15.04.78	BASSO TIRENO
cambag1	122	28	6,5	15	NOR	23.11.80	CAMPANO LUCANO
camcal1	133	26	6,5	15	NOR	23.11.80	CAMPANO LUCANO
camstu1	109	37	6,5	15	NOR	23.11.80	CAMPANO LUCANO
if001	17	19	3,9	8	-	06.02.72	ANCONA
if016	18	16?	3,5	8	-	04.04.72	ANCONA
if019	8	20?	3,5	8	-	04.04.72	ANCONA
if022	25	9	4,7	3	-	14.06.72	ANCONA
if025	21	9	4,7	3	-	14.06.72	ANCONA
if142	74	17	6,8	4	D-S	15.04.79	MONTENEGRO
if148	73	6	6,8	4	D-S	15.04.79	MONTENEGRO
if151	73	58	6,8	4	D-S	15.04.79	MONTENEGRO
if160	70	51	6,8	4	D-S	15.04.79	MONTENEGRO
frifor1	16	31	6,2	8	REV	06.05.76	FRIULI
frimai1	17	27	6,2	8	REV	06.05.76	FRIULI
fritol1	55	23	6,2	8	REV	06.05.76	FRIULI
friufor1	22	15	5,5	6	REV	11.09.76	FRIULI
friusan1	15	14	5,5	6	REV	11.09.76	FRIULI
norbev1	36	37	5,8	-	NOR	19.09.79	NORCIA
			КАЛИФОРНИЯ				
coyote1	92	50	5,9	9	S-S	06.08.79	COYOTE LAKE
coyote28	41	12	5,9	9	S-S	06.08.79	COYOTE LAKE
coyote31	41	10	5,9	9	S-S	06.08.79	COYOTE LAKE
cf001	100	29	6,5	7	S-S	02.05.83	COALINGA
cf004	92	39	6,5	7	S-S	02.05.83	COALINGA
cf007	100	35	6,5	7	S-S	02.05.83	COALINGA
cf01	87	70	6,6	12	S-S	15.10.79	IMPERIAL VALLEY
cf04	88	28	6,6	12	S-S	15.10.79	IMPERIAL VALLEY
cf07	87	28	6,6	12	S-S	15.10.79	IMPERIAL VALLEY
h073	78	2	6,0	16	NOR	31.10.35	HELENA MONTANA
i001	83	7	7,1	16	S-S	19.05.40	IMPERIAL VALLEY
p103	39	55	5,5	10	S-S	28.06.66	PARKFIELD
p106	68	54	5,5	10	S-S	28.06.66	PARKFIELD
p109	46	61	5,5	10	S-S	28.06.66	PARKFIELD
p112	44	72	5,5	10	S-S	28.06.66	PARKFIELD
s151	80	43	6,5	8	REV	09.02.71	SAN FERNANDO
s169	127	36	6,5	8	REV	09.02.71	SAN FERNANDO
s202	56	34	6,5	8	REV	09.02.71	SAN FERNANDO
s316	153	35	6,5	8	REV	09.02.71	SAN FERNANDO
s328	151	31	6,5	8	REV	09.02.71	SAN FERNANDO

Примечание. NOR - normal; D-S - dip-slip; S-S - strike-slip; REV - reverse





Р и с. 1. Эпицентры калифорнийских (а) и итальянских (б) землетрясений, записи сильных движений которых использовались при распознавании

в каждом конкретном случае учитываем информативность параметров фрагментов записи и ее длительности.

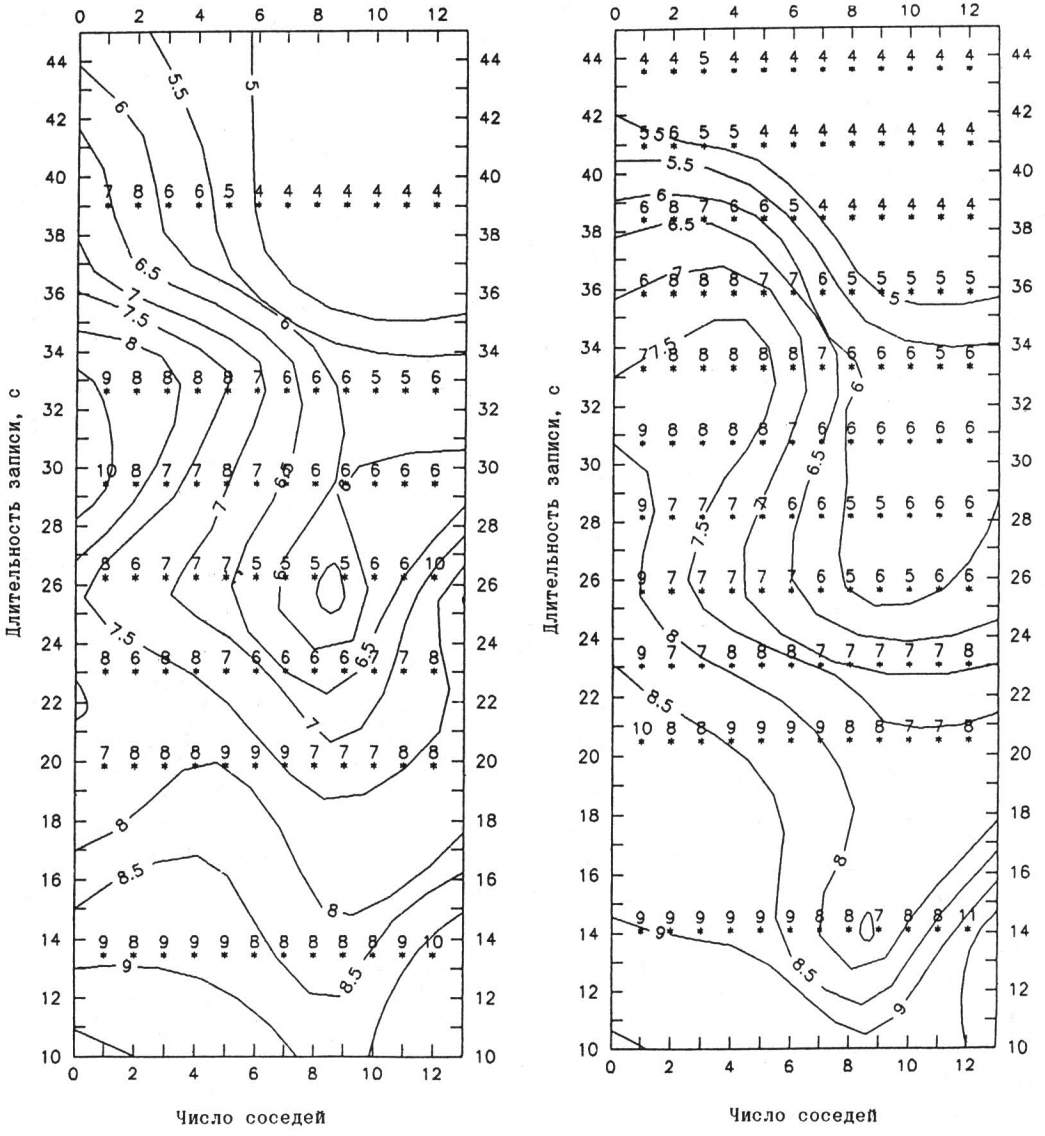
Как видно из табл.1, общие длительности исследуемых записей неоднородны, в среднем итальянские записи короче калифорнийских. Поэтому наибольшее внимание при оценке влияния свободных параметров на результаты классификации мы уделили именно общей длительности записи. Результаты вычислений попарных расстояний с разными порогами максимальной длительности записи при равных весах параметров фрагментов приведены на рис. 2.3. Графики строились отдельно для вертикальных и горизонтальных компонент и для длительностей фрагментов δt=1.28 с и δt=2.56 с. По оси X каждого графика отложено число соседей в процедуре классификации, по оси Y - максимальная длительность записи (в секундах). Поверхность значений функции ошибок F^*_{ERR}, зависящей от этих параметров, была сглажена и изображена на рис.2.3 изолиниями. Звездочками отмечены конкретные значения параметров, для которых проводились расчеты, и рядом указано полученное значение F^*_{ERR}. Рисунки 2.3 показывают, что увеличение максимальной длительности записей в общем уменьшает значения функции ошибок. Для вертикальных компонент поверхность значений F^*_{ERR} более регулярна, чем для горизонтальных.

На рис.4 раздельно для 20 калифорнийских и 20 итальянских акселерограмм приведены их длительности, отсортированные по возрастанию. На нем также показаны значения F^*_{ERR} (минимум по числу соседей 1÷12) для вертикальных и горизонтальных компонент при длительности фрагментов δt=2.56 с. Видно, что в среднем число ошибок распознавания вертикальных компонент ниже, чем горизонтальных, и длительность записи не является достаточным признаком для разделения двух этих множеств со сравнимыми значениями F^*_{ERR}.

Ошибки классификации для различного числа фрагментов в записях раздельно для вертикальных и горизонтальных компонент при длительности фрагмента 2.56 с сведены в табл.2. Она показывает, что наиболее часто ошибки возникают на записях землетрясения Campano Lucano (Италия) и землетрясения Helena (Montana). Заметим, что множества ошибок при классификации вертикальных и горизонтальных компонент различны.

Варьирование весов параметров фрагментов не привело к уменьшению значений F^*_{ERR} по сравнению со случаем равных весов.

Можно заключить, что применение SPARS в задаче распознавания записей сильных движений из разных геологических регионов на данном материале устойчиво приводит не более чем к 10÷15% ошибок в



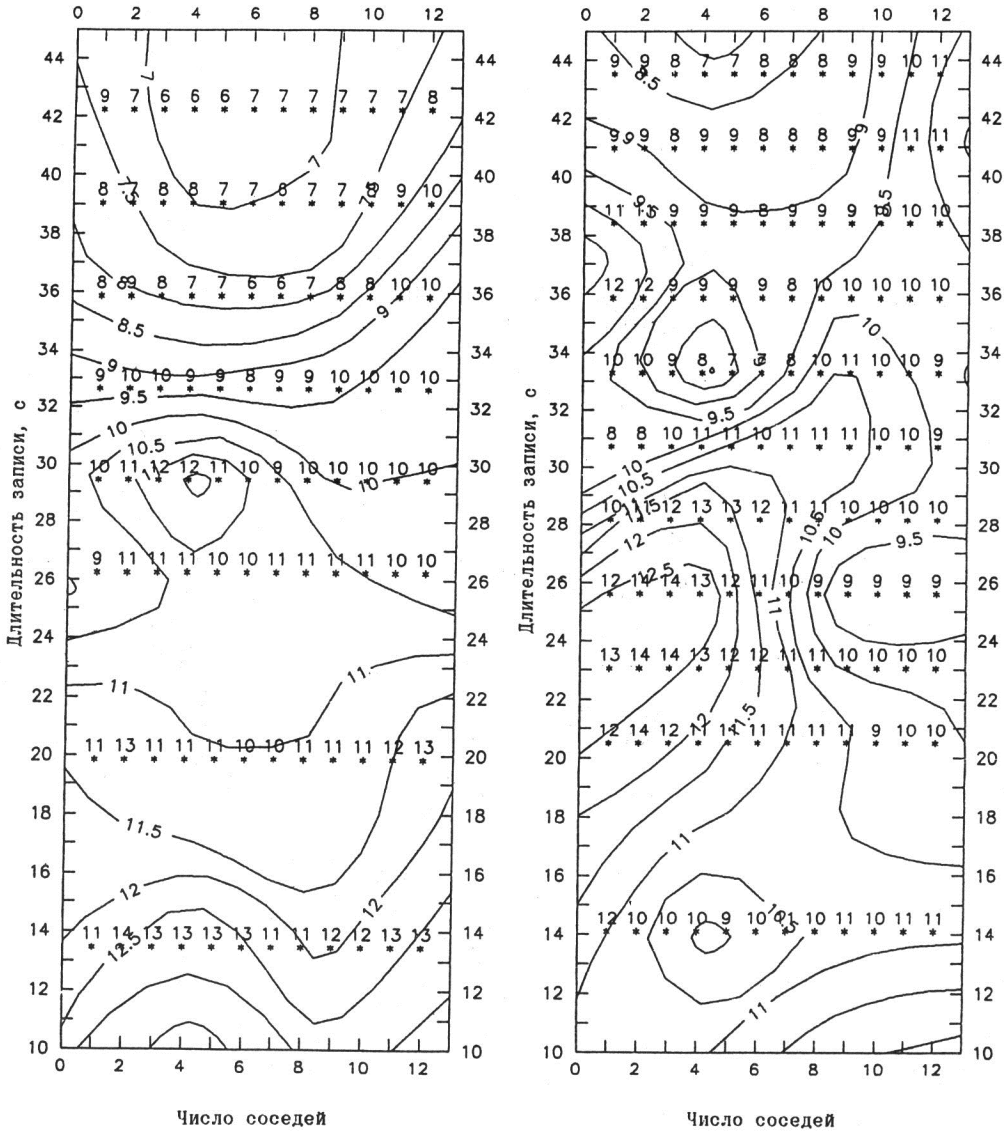
Р и с. 2. График значений функции ошибок для вертикальных компонент

Длительность фрагментов: а - δt=1,28 с. б - δt=2,56 с

Таблица 2

Ошибки классификации в зависимости от максимального числа фрагментов записи для горизонтальных и вертикальных компонент

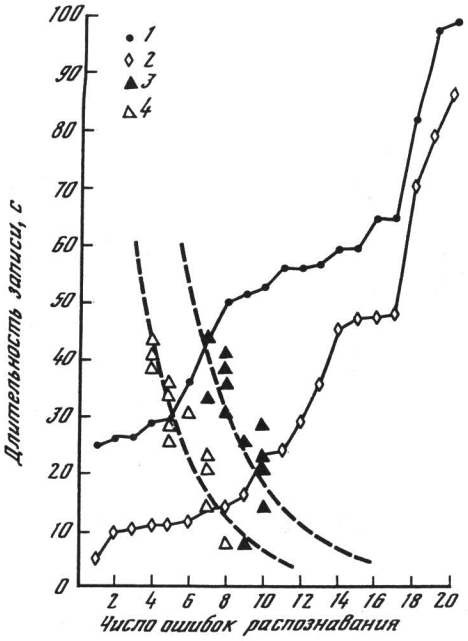
Запись	Максимальное число фрагментов													
	Горизонтальная компонента							Вертикальная компонента						
	10	15	19	23	27	31	35	10	15	19	23	27	31	35
basmes1			*											
basnas1				*				*	*	*	*			
cambag1	*	*	*	*	*	*	*	*	*	*	*	*	*	*
camcal1	*	*	*		*	*	*	*	*	*	*	*	*	
camstu1	*	*	*	*	*	*	*		*	*	*	*	*	*
if001														
if016														
if019														
if022								*						
if025														
if142	*	*	*		*									
if148		*	*	*	*	*								
if151														
if160														
frifor1														
frimai1														
fritol1	*	*	*	*	*	*	*	*	*					
friufor1														
friusan1														
norbev1														
coyote1	*													
coyote28														
coyote31														
cf001														
cf004														
cf007														
cf01														
cf04														
cf07	*			*										*
h073	*	*	*					*	*	*	*	*		
i001				*										
p103														
p106														
p109														
p112	*	*	*	*	*	*	*	*	*					
s151														
s169														
s202														
s316						*					*	*	*	
s328						*	*							



Р и с. 3. График значений функции ошибок для горизонтальных компонент
Длительность фрагментов: а - δt=1,28 с, б - δt=2,56 с

Р и с. 4. Минимальные значения функции ошибок для различных значений максимальной длительности записи

1, 2 - длительности калифорнийских и итальянских записей; 3, 4 - минимум функции ошибок для горизонтальных и вертикальных компонент



достаточно широком диапазоне значений свободных параметров (по сравнению с 50% ошибок при перемешивания материала обучения и тех же значениях параметров).

Результаты классификации для вертикальных компонент несколько лучше, чем для горизонтальных практически при любых значениях свободных параметров SPARS. Заметим, что аналогичные наблюдения верны и в случае задачи [5].

Автор благодарен ЕССЦ в Страсбурге за поддерджку этой работы, Ж.Бонину, Б.Мохамадьюну, Ж.Салантану, и особенно А.Д.Гвишиани, за полезные обсуждения и внимание к работе.

Литература

1. *Ту Дж., Гонсалес Р.* Принципы распознавания образов. М.: Мир, 1978. 324 с.

2. *Блейкут Р.* Быстрые алгоритмы цифровой обработки сигналов. М.: Мир, 1989. 448 с.

3. *Sellers P.H.* Pattern recognition in genetic sequences // Proc. Natl. Acad. Sci. USA. 1979. Vol.76. P.3041.

4. *Thomason M.G.* Structural methods in pattern analysis / Ed. by Devijer P.A. and Kittler J // Pattern recognition theory and applikations. NATO ASI Series, F30. 1987. 307 p.

5. *Гвишиани А.Д., Жижин М.Н., Иваненко Т.И.* Синтаксический анализ записей сильных движений // Компьютерный анализ геофизических полей. М.: Наука. 1990. С.235-253. (Вычисл. сейсмология; Вып. 23).

6. *Das S., Aki K.* Fault plane with barriers: a versatile earthquake model // J.Geophys.Res. 1977. Vol.82. P.5648-5670.

7. *Гусев А. А.* Модель очага землетрясения со множеством неровностей // Вулканология и сейсмология. 1988, N 1. С.41-55.

8. *Davis J.P., Sacks I.S., Linde A.T.* Source complexity of small earthquakes near Matsushiro, Japan // Tectonophysics. 1989. Vol.166. P.175-187.

9. *Peckmann J.C.,Kanamori H.* Waveforms and spectra of aftershocks of the 1978 Imperial Valley, California, earthquake: evidence of fault heterogeneity? // J.Geophys. Res. 1982. Vol.87. P. 10579-10597.

10. *Peckmann J.C., Thorbjarnardottir B.S.* Waveform analysis of two preshock-main shock-aftershock sequences in Utah // Bull. Seismol.Soc.Amer. 1990. Vol.80. P.519-550.

11. *Joswig M.* Pattern recognition for earthquake detection // Bull. Seismol.Soc.Amer. 1990. Vol.80. P.170-186.

12. *Diday E. et al.* Eléments d'analyse de données. Paris, Dunod: 1988. 256 p.

13. *Bernard P. and Zollo A.* Inversion of near-source S polarization for parameters of double-couple point sources // Bull. Seismol.Soc.Amer. 1989. Vol.79. P.1779-1809.

14. *Hsi-Ho Liu and K.S.Fu,* A syntactic approach to seismic pattern recognition // IEEE transactions on pattern analysis and machine intelligence, PAMI-4. 1982. P.136-140.

15. *Fu K.S.* Syntactic approach to pattern recognition / Ed. by Simon J.C. // Spoken language generation and underestanding. D. Reidel Publishing Co., 1980. P.221-251.

УДК 550.34.012

К.В.Кислов. Ю.А.Колесников. А.Ю.Марченков. Ю.О.Старовойт

СЕЙСМИЧЕСКИЙ МИКРОБАРОГРАФ

K.V.Kislov, Yu.A.Kolesnikov, A.Yu.Marchenkov, Yu.O.Starovoit

SEISMIC MICROBAROGRAPH

Design and operational principles of a high resolution microbarograph with capacity transducer are described. The microbarograph can be used in investigation of seismic noise caused by local fluctuationof of atmospheric pressure in problems of pressure noise compensation.

Выделение сейсмических сигналов в длиннопериодной области спектра во многом осложнено сейсмической помехой барического происхождения. Микрофлуктуации атмосферного давления в диапазоне периодов