МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. Ломоносова

На правах рукописи

Попова Елена Александровна

Методы и программные средства для обработки данных электроэнцефалографии

Специальность 05.13.11 - Математическое и программное обеспечение
вычислительных машин, комплексов
и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва 2009

Работа выполнена на кафедре автоматизации систем вычислительных комплексов факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.

Научный руководитель: доктор физико-математических наук,
член-корреспондент РАН,
профессор Королев Лев Николаевич.

Официальные оппоненты: доктор физико-математических наук,
профессор Жданов Александр Аркадьевич;

доктор биологических наук,
кандидат физико-математических наук,
профессор Фролов Александр Алексеевич.

Ведущая организация: Институт прикладной математики им. М.В.Келдыша
Российской академии наук.

Защита состоится 6 марта 2009г. в 11 час. 00 мин. на заседании диссертационного совета Д 501.001.44 в Московском государственом университете имени М.В.Ломоносова по адресу: 119991,ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, ВМиК, ауд. 685.

С диссертацией можно ознакомиться в библиотеке факультета ВМиК МГУ. С текстом автореферата можно ознакомиться на официальном сайте ВМиК МГУ http://cs.msu.ru в разделе "Наука" – "Работа диссертационных советов" – "Д.501.001.44".

Автореферат разослан "_____" февраля 2009 г.

Ученый секретарь
диссертационного совета
профессор _____ Н.П.Трифонов

Общая характеристика работы

Актуальность темы. Исследованию проблем, связанных с функционированием человеческого мозга, всегда уделялось большое внимание со стороны специалистов из разных научных областей. Важная роль в решении возникающих задач отводится методам автоматизированной обработки сигналов, измеряемых при изучении человеческого мозга. Одно из важных направлений в изучении человеческого мозга связано с исследованиями его электрической активности, регистрируемой в виде сигнала электроэнцефалограммы (ЭЭГ). Ключевое место в этой проблеме занимает задача локализации источников. Определение зон активности открывает новые возможности для проведения диагностирования и лечения заболеваний, исследования реакций мозга на внешние воздействия, исследования проблем восприятия человеком внешнего мира.

Определение зон активности коры головного мозга при реакциях на различные внешние воздействия широко используется при создании интерфейса мозг-компьютер (Brain-Computer Interface). В нейрофизиологических исследованиях на основе ЭЭГ проводятся важнейшие эксперименты по выявлению реакции мозга на поступающие образы. Получая оценку функционального состояния работы мозга, нейрофизиологи выявляют принципы и механизмы внутреннего взаимодействия компонент мозга. Практическое внедрение новых программных систем, на основе новых информационных технологий позволяет дать дополнительные новые инструменты для понимания нейрофизиологами проблем восприятия человеком внешнего мира.

Принципиальную роль в этих исследованиях играет автоматизированный анализ данных ЭЭГ.

Нейронная система мозга представляет собой очень сложный объект, содержащий приблизительно 10^{14} нейронов [1], которые в совокупности отражают функциональное состояние мозговой активности. Судить о реакции мозга мы можем по измерениям потенциала на поверхности головы, которые являются многомодовыми, т.е. отражают реакцию мозга на разные сенсорные рецепторы – зрительные, слуховые, ментальные, двигательные и др. В полученном экспериментальном сигнале содержится несколько ритмов, которые отвечают за разный тип активности, и результат их взаимодействия. Главной особенностью экспериментальных данных, изучаемых в диссертации, является то, что они получены внешними по отношению к объекту диагностиками. Необходимо определить внутренние источники этих сигналов, чтобы понять модели, лежащие в основе этих данных.

Причиной развития параллельных алгоритмов в задачах локализации источников служат масштабы разрабатываемых моделей человеческого мозга. От общих многослойных моделей исследователи переходят к моделированию максимально приближенным к кон-

[1] Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. Sarvas J. // Phys.Med.Biol. 1987. **32**. p.11-22.

кретному испытуемому анатомических структур мозга и описанию более сложных процессов электрической активности, происходящих на микро-уровне. В связи с этим число элементов в рассматриваемой структуре возрастает в несколько раз, что требует увеличения вычислительных ресурсов для микро-моделирования. Создание новых поколений многопроцессорных вычислительных систем, содержащих сотни тысяч процессорных элементов, позволяет ставить новые задачи по прямому моделированию работы мозга.

Все это приводит к необходимости разработки новых алгоритмов и программных комплексов для обработки данных измерений, связанных с активностью и строением мозга. Основное, что предлагается в диссертации - это метод, основанный на использовании ансамблей деревьев решений и программный комплекс, построенный на базе этого метода для поиска источников анализируемых сигналов электрической активности головного мозга.

Цель диссертации.

Разработка алгоритмов и программных средств, реализующих автоматизированное построение пространственно-временных карт активности мозга по данным ЭЭГ.

Научная новизна и практическая ценность.

1. Разработан новый метод и создана программная система автоматизированного анализа сигнала ЭЭГ. Впервые показано, что задача локализации нейронных источников по сигналам ЭЭГ может быть поставлена как задача классификации данных и эффективно решена с помощью деревьев решений. Создан программный комплекс, позволяющий устойчиво выделять в трехмерном пространстве электрически активные зоны мозга, отвечающие за регистрируемый пространственно-временной сигнал ЭЭГ.

2. Впервые разработаны алгоритмы и программы параллельного анализа данных ЭЭГ большой размерности с помощью построения комитета классификаторов на высокопроизводительных вычислительных системах. Разработана методика параллельного обучения классификаторов. Показано, что сформулированная задача является масштабируемой по числу процессоров.

3. Предложены новые методы и алгоритмы автоматизированной оценки функционального состояния мозговой деятельности, которые могут служить источником дополнительной информации при анализе нейрофизиологических данных.

   Проведена практическая демонстрация эффективности разработанных программных средств при анализе данных ЭЭГ, полученных в нейрофизиологических экспериментах, связанных с обработкой мозгом зрительных стимулов.

Апробация работы.

Основные результаты работы докладывались на:

- 23-й международной конференции по статистической физике STATPHYS23, секция стастистические методы в биомедицине (Италия, Генуя, Июль 2007)
- 14-й международной конференции по параллельным вычислениям ParCo2007 (Германия, Аахен, Сентябрь 2007)
- семинаре Группы Изучения Мозга человека (биологический факультет МГУ)
- семинаре Группы Нейрофизиологии Когнитивных Процессов Института Высшей Нервной Деятельности и Нейрофизиологии РАН
- научной конференции "Тихоновские Чтения 2007" (Россия, Москва, МГУ, ф-т ВМиК)
- семинаре кафедры автоматизации систем вычислительных комплексов по руководством зав. Кафедры чл.-корр. РАН Л.Н. Королева

Публикации. Основные результаты диссертации опубликованы в работах [1-8].

Структура и объем диссертации. Диссертация состоит из введения, трех глав, заключения, списка литературы и приложения, содержащего рисунки. Текст изложен на 112 страницах, диссертация содержит 27 рисунков. Список литературы включает 114 наименований.

Содержание работы

Введение посвящено краткому описанию проблем и результатов, относящихся к теме диссертации.

Приведен анализ существующих математических подходов автоматизированной обработки экспериментальных данных, связанных с деятельностью мозга. Рассмотрены распространенные программные комплексы по обработке электрических сигналов мозга, отмечены их недостатки и преимущества.

Первая глава посвящена обоснованию нового подхода для анализа сигналов ЭЭГ на основе ансамблей случайных деревьев решений.

Главной особенностью предложенного метода является сведение задачи локализации нейронных источников внутри мозга к задаче классификации признаков источников с использованием комитета голосующих классификаторов, обученных на входном сигнале ЭЭГ.

В §1.1 описывается дипольная модель нейронных источников, создающих регистрируемый электрический потенциал на поверхности головы в технологии ЭЭГ. Приводится

математическая постановка задачи поиска активных нейронных источников внутри мозга, ответственных за суммарный электрический потенциал на поверхности головы.

В известной дипольной модели мозг рассматривается как объемный трехмерный проводник. Источниками электрической активности являются электролитические токи внутри нервных клеток коры головного мозга, которые заменяются в модели локализованными электрическими диполями. Ограничения на область распространения электрического поля вводятся уравнением поверхности головы.

Диполь в свободном пространстве задается пространственным положением $\mathbf{r}^{(p)}$ и вектором плотности тока (момента, $\nu^{(p)}$). Используя известные законы распространения электрического тока в ограниченном пространстве, можно вычислить потенциал электрического поля, создаваемый заданным набором диполей (прямая задача ЭЭГ). Задача поиска расположения активных нейронных источников, ответственных за регистрируемый на поверхности головы потенциал, называется обратной задачей ЭЭГ или задачей локализации активных нейронных источников.

В диссертации обратная задача формулируется следующим образом. Пусть даны экспериментальные сигналы ЭЭГ, представляющие запись электрического потенциала $U_{\exp}$ на поверхности головы в течение определенного времени. Задан набор данных – электрические потенциалы $U_{\exp}(\vartheta_i, \varphi_i, t^n)$ в ряде точек $(\vartheta_i, \varphi_i; i = 1, 2...I)$ на поверхности головы, измеренные в моменты времени $(t^n; n = 1, 2...T)$. Этот потенциал моделируется суммарным потенциалом дипольных источников $U_{model} = \sum_p U^{(p)}_{model}(r^{(p)}, \vartheta^{(p)}, \varphi^{(p)}, \nu^{(p)}_r, \nu^{(p)}_\vartheta, \nu^{(p)}_\varphi | \vartheta_i, \varphi_i, t^n)$, где $p = 1, 2, ..., P$ – номер источника, $\mathbf{r}^{(p)} = (r^{(p)}, \vartheta^{(p)}, \varphi^{(p)})$ – координаты источника в сферической системе (локализация), $\nu^{(p)} = (\nu^{(p)}_r, \nu^{(p)}_\vartheta, \nu^{(p)}_\varphi)$ – три проекции силы (момента) диполя, (ϑ_i, φ_i) – координаты i-го электрода измерения потенциала на поверхности головы в сферической системе. Точность аппроксимации экспериментальных данных модельным потенциалом от p диполей оценивается функционалом ошибки по потенциалу:

$$\varepsilon^p(t^n) = \sum_i [(U_{exp}(\vartheta_i, \varphi_i, t^n) - U^{(p)}_{model}(r^{(p)}, \vartheta^{(p)}, \varphi^{(p)}, \nu^{(p)}_r, \nu^{(p)}_\vartheta, \nu^{(p)}_\varphi | \vartheta_i, \varphi_i, t^n)]^2 \Delta S, \quad (1)$$

где ΔS – элемент площади поверхности. Задача определения источника, создающего измеряемый потенциал, состоит в нахождении такого полного набора параметров p диполей

$$\{(r^{(p)}, \vartheta^{(p)}, \varphi^{(p)}, \nu^{(p)}_r, \nu^{(p)}_\vartheta, \nu^{(p)}_\varphi)\}_{p=1...P}, \quad (2)$$

при котором ошибка (1) минимальна: $\varepsilon^p(t^n) \to min$. Задача является нелинейной и при

нахождении нескольких источников в каждый момент времени - переопределенной (параметров источника больше, чем условий совпадения потенциала), поэтому необходимо введение дополнительных ограничений на множество решений (например, априорные предположения о силе диполя в разные моменты времени).

В §1.2 на основе принятой дипольной модели предлагается подход, в котором задача локализации источников сводится к задаче классификации параметров источников.

Основная идея предлагаемого метода состоит в решении задачи минимизации функционала ошибки методом классификации источников на основе специально сформированного обучающего множества этих источников. Каждый образ в множестве – это набор диполей, заданных в сферической системе координат. Полное анализируемое множество состоит из множества различных комбинаций параметров диполей, располагаемых некоторым способом внутри пространственной области головы.

Каждый набор из n диполей характеризуется $6n$ признаками и меткой класса, к которому он принадлежит,

$$\mathbf{X}^p = \{x_1, x_2, ..., x_i, ..., x_M | C_k\}^p, \tag{3}$$

где p – номер примера (набора диполей), k – метка класса, к которому относится пример, x_i^p – i-й признак для набора диполей с номером p. Для каждого p-го примера вычисляется ошибка по потенциалу RRE (Residual Ralative Error)

$$\varepsilon^p = \sum_i \sum_j [(U_{exp}(\vartheta_i, \varphi_j) - V_K(\vartheta_i, \varphi_j)) - w^{(p)}(\mathbf{x^P} \mid \vartheta_i, \varphi_j)]^2 \sin\vartheta_i h_{\vartheta i} h_{\varphi j}.$$

Задается уровень допустимой ошибки ε_{th}, и все множество диполей разделяется на два класса: ниже порога, когда $\varepsilon^p < \varepsilon_{th}$ (первый класс — C_1) и выше порога (второй класс — C_2). Каждому набору диполей ставится в соответствие метка одного из классов. Таким образом формируется обучающее множество.

Целью классификации данных является нахождение параметров дипольных источников из обучающего множества, относящихся к классу C_1, т.е. имеющих допустимый уровень ошибки (1). Предложенный метод решения минимизационный задачи является новым альтернативным подходом решения обратной задачи.

В §1.3 предлагается алгоритм построения классификатора для определения зон электрической активности.

В параграфе разрабатывается модифицированный метод комитета случайных деревьев классификации Random Forest [2] для решения задачи классификации активных источников, основанный на эвристике задачи локализации. Подход комбинирования простых классификаторов является эффективным методом решения задачи бинарной классификации

[2]Breiman L. Random Forests//Machine Learning.2001.**45**P.5-32.

с непересекающимися классами.

В параграфе описан алгоритм построения множества простых классификаторов, который состоит в следующем. Для обучающего множества (3) строится P классификаторов по алгоритму:

1. Формируется P случайных независимых векторов размерности $N, \{\vartheta_i\}_{i=1}^{P}$ с одинаковым распределением.
2. Каждому вектору $\vartheta_k^l, l = 1..P, k = 1..N$ ставится в соответствие подмножество исходного обучающего множества $D_i = \{(\vec{x}, C)_i^j\}_{i=1,j=l}^{N}$.
3. Каждое подмножество D_i разбивается на два: $D_i = \{(\vec{x}, C_i)_1^{2/3N} \cup oob = (\vec{x}, C_j)_{2/3N}^{N}\}$. Первое будет использоваться для обучения классификатора, второе – для вычисления ошибки классификации.
4. Используя первое подмножество примеров каждого D_i, обучается набор P классификаторов со случайным параметром m: $\{h_k(x, \vartheta_k, m)\}_{k=1}^{P}$.
5. Выполняется процедура голосования классификаторов, определяются итоговые области активности.
6. Вычисляется значение ошибки классификации, используя второе подмножество примеров каждого D_i: $OOB_{error} = \sum_{(\vec{x},j)\in X,} \frac{\sum_{h_k: x\in oob(h_k)} I(h_k \neq j)/N_k}{N}$, где N_k – общее число примеров: $(\vec{x}, j) \in oob$, N – число примеров обучающего множества, $I(\cdot)$ – функция-индикатор.

В качестве простого классификатора в работе используется дерево решений. Дерево решений осуществляет кусочно-постоянную аппроксимацию исходных данных, что делает метод устойчивым при обработке сильно зашумленных данных. Иерархическая структура дерева представляет признаки исходного множества в порядке "значимости". Обучение дерева без отсечения ветвей уменьшает вычислительную сложность алгоритма.

Для формирования комитета классификаторов в параграфе определяется процедура голосования деревьев классификации. В работе предлагается использовать эвристику задачи локализации, которая состоит в следующем.

В каждом дереве рассматриваются все пути из корневой вершины v_{root} в листовую вершину с меткой класса C_1. Через $V = \{v_1, v_2, ..., v_N\}$ обозначается множество терминальных узлов дерева, через $B = \{br_1, br_2, ..., br_B\}$ – множество ветвей дерева, принадлежащих каждому из таких путей. Каждому пути i в дереве t, соединяющему корневую вершину и лист с классом C_1, ставится в соответствие множество вида

$$\mathrm{Reg}_i = \{x_i, x_{th}, rt\}$$

где x_i - признак узла v_i, x_{th} - пороговое значение узла v_i из множества V, rt – отношение

$\{<,>\}$, определенное ветвью br_i из множества B. Множество будет определять область в исходном пространстве признаков.

Рассматривается множество областей, выделенных всеми деревьями комитета: $Reg_{all} = \{Reg_{1_1}, Reg_{1_2}, ..., Reg_{2_1}, Reg_{2_2}, ..., Reg_{P_1}, Reg_{P_2}, ...\}$, где P – число деревьев. Строится отображение множества Reg_{all} на сферическую систему координат. Для этого выделяется два подмножества: первое из них состоит из признаков, отвечающих за пространственное положение диполей $R_{pos} = \{(r^p, \vartheta^p, \varphi^p)\}_{p=1}^{L}$, второе – из моментов диполей $R_{mom} = \{(\nu_r^p, \nu_\vartheta^p, \nu_\varphi^p)\}_{p=1}^{L}$. Отображение формируется для множества положений источников R_{pos} в сферической системе, определяющее множество областей. Тогда пересечение этих областей $Z_p = (\bigcap_{i=1}^{P}(R_{pos}^p), nt)$ будет соответствовать областям класса C_1, а nt – числу деревьев, участвующих в голосовании (пересечении). Для того, чтобы выделить наиболее значимую область решающим пересечением определяется такое пересечение, в котором участвовало бы наибольшее число деревьев. Моменты активных зон определяются элементами из множества R_{mom}.

Таким образом, множество активных областей соответствует числу дипольных источников при построении исходного обучающего множества. Данный способ локализации позволяет исключить из рассмотрения области, возникающие из-за шума в исходных данных.

В §1.4 исследуется сходимость и определение основных параметрических зависимостей предложенного метода.

Сходимость и точность разработанного алгоритма решения обратной задачи ЭЭГ была исследована путем решения модельной обратной задачи ЭЭГ и сравнения результатов работы разработанного алгоритма с аналитическим решением.

В результате были выявлены основные параметры алгоритма, влияющие на его сходимость и точность локализации: порог по потенциалу, число деревьев в ансамбле, число признаков для вычисления узла дерева.

Вторая глава посвящена разработке программного комплекса для реализации предложенных в диссертации методов обработки экспериментальных данных ЭЭГ. В главе на основе разработанного метода и требований, предъявляемых к программному комплексу, определена структура комплекса и рассмотрена программная реализация основных модулей.

В §2.1 на основе анализа функциональности существующих программных систем локализации дипольных источников, учете специфики входных экспериментальных данных и предложенных в первой главе алгоритмов определяются требования к разрабатываемому программному комплексу для обработки данных ЭЭГ. Сформулированные в диссертации требования определяют функциональные возможности комплекса и требования к производительности разрабатываемого комплекса программ и масштабируемости предлагаемых параллельных алгоритмов. В параграфе также обсуждаются существующие способы реа-

лизации программных систем для обработки данных ЭЭГ.

В §2.2 описана структура разработанного программного комплекса для обработки данных ЭЭГ.

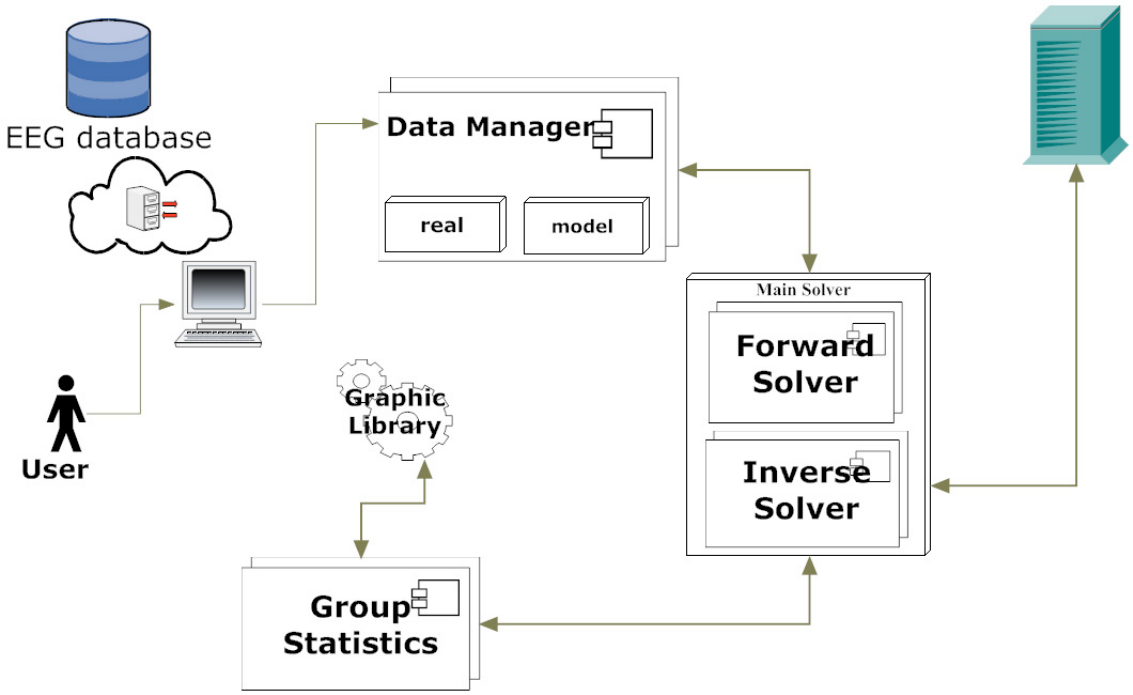

Рис. 1: Структура программного комплекса для обработки сигнала ЭЭГ.

Модули комплекса обеспечивают обработку входных сигналов, реализуют разработанные в диссертации алгоритмы решения прямой и обратной задач, выполняют сбор и анализ результатов обработки, поддерживают визуализацию и интерпретацию полученных результатов. Структура разработанного комплекса приведена на рис. 1.

Загрузка данных сигналов ЭЭГ, их предварительная обработка и представление во внутреннем формате осуществляется модулем *Data Manager*. Параметрами модуля являются спецификация формата входных данных согласно принятой системе регистрации сигнала ЭЭГ. В настоящее время разработанный комплекс поддерживает работу семи форматов входных данных.

Основными модулями комплекса являются модули, реализующие решение прямой задачи (Forward Solver) и выполняющие построение ансамбля деревьев решений с возможностью параллельного исполнения на многопроцессорных вычислительных (Inverse Solver). Для визуализации и интерпретации результатов работы алгоритмов реализован модуль мониторинга основных параметрических зависимостей и модуль визуализации активности на реальных структурах мозга (Graphics Library). Отдельный модуль реализцет алгоритм построения усредненных карт активности на основе кластерного анализа (Group Statistics).

Программная реализация предложенных алгоритмов выполнена на языке C++. Параллельная реализация модуля Inverse Solver выполнена с использованием технологий па-

раллельного программирования MPI и OpenMP. Реализация интерпретации результатов локализации использует библиотеку трехмерных моделей мозга BrainVisa, представленную в виде классов на языке C++.

В параграфе описаны базовые классы основных модулей программного комплекса, реализующие алгоритмы: построения ансамбля деревьев решений, сведения задачи локализации к задаче классификации, построения пространственно-временных карт активности.

В §2.3 проводится временная оценка алгоритма решения задачи локализации. Целью данного анализа является выявление наиболее "узких" мест алгоритма и их дальнейшая параллельная реализация. Под "узким" местом будем понимать наиболее вычислительноемкие части алгоритма.

В параграфе анализируется зависимость времени решения прямой задачи ЭЭГ от параметров задачи: размерности обучающего множества и способа задания геометрии поверхности головы человека. Представлены зависимости роста времени решения задачи от размерности входного сигнала ЭЭГ и числа сигналов, используемых при усреднении результатов.

Показано, что вычислительные затраты,необходимые для построения ансамбля деревьев решений, занимают около 90% от общего времени решения задачи. Голосование деревьев и определение результирующей зоны требуют около 7% общего времени решения. Исходя из полученных результатов, для реализации параллельной обработки предлагается разделение исходной задачи построения множества классификаторов на независимые подзадачи построения одного классификатора.

В параграфе проводится анализ последовательного алгоритма построения одного дерева решений. В результате выделены те стадии построения дерева решений, которые требуют наибольших временных затрат.

В §2.4 описан параллельный алгоритм построения ансамблей деревьев решений.

Предлагается способ “быстрого построения” множества деревьев-классификаторов, каждое из которых определяет зону активности в зависимости от количества доступных ресурсов вычислительной системы. Группа процессоров, на которых выполняется алгоритм, разбивается на блоки в соответствие с числом деревьев решений в ансамбле. Независимые подзадачи алгоритма распределяются на процессоры этих блоков.

Пусть множество свободных процессоров P_{free} и N_t – число деревьев решений, которые нужно построить для формирования ансамбля. Тогда группа процессоров P_{free} разбивается на блоки по $P_{\text{free}}/N_t \text{mod} N_b$, где N_b – число процессоров в каждом блоке. Обучение каждого дерева выполняется на одном блоке процессоров. Исходное множество примеров B распределяется между блоками по $B/(P_{\text{free}}/N_t \text{mod} N_b)$ примеров на блок. Множество примеров для каждого блока разбивается на равные части между всеми процессорами блока.

Предположим, что число процессоров в блоке есть степень двойки и подсистема связи процессоров друг с другом представляет собой топологию гиперкуба. На практике это условие реализуется путем использования виртуальных топологий MPI.

Шаги параллельного алгоритма:

1. База данных обучающих примеров дерева решений равномерно распределяется между P процессорами. Таким образом, если N – число обучающих примеров, то у каждого процессора в локальной памяти будет N/P примеров.
2. Изначально все процессоры обрабатывают один узел дерева N_0 и работают вместе для вычисления точки разбиения.

   - Каждый процессор для каждого атрибута, находящегося у него в памяти, вычисляет распределение классов и записывает его в свою копию хэш-таблицы.
   - Каждый процессор обменивается данными о распределении классов с другими процессорами посредством записи в глобальную редукционную переменную (хэш-таблицу).
   - Одновременно процессоры вычисляют значение энтропии и индекс Гини для каждого атрибута узла и выбирают наилучший атрибут для разбиения на подмножестве,а затем и на всем обучающем множестве посредством записи в редукционную переменную.
   - Создаются потомки для рассматриваемого узла и проводится разбиение тренировочного множества на подмножества-потомки.

3. С увеличением глубины дерева объем собираемой статистики на каждом уровне увеличивается. Вследствие этого на каком-то из уровней временная стоимость обмена информацией между процессорами становится чрезмерно большой. Тогда в этом случае группа процессоров, работающих в блоке над каждым узлом, разбивается на две подгруппы, выполняющие построение поддеревьев параллельно.
4. Шаги алгоритма, начиная с п.1, повторяются для созданных подгрупп процессоров, которые могут работать независимо.
5. Если число процессоров в группе становится меньше двух, то эта группа объединяется с группой процессоров, которая имеет такое же число процессоров.

Ключевым моментом в данном алгоритме является критерий разбиения группы работающих процессоров на подгруппы. Выполнение разбиения на каждом уровне дерева решений (асинхронный подход) или отказ от проведения разбиения (синхронный подход), могут приводить к увеличению накладных расходов, связанных с передачей данных большого объема. В параграфе подробно рассмотрены оба случая и предложен гибридный

метод, основанный на комбинировании синхронного и асинхронного способов построения дерева решений.

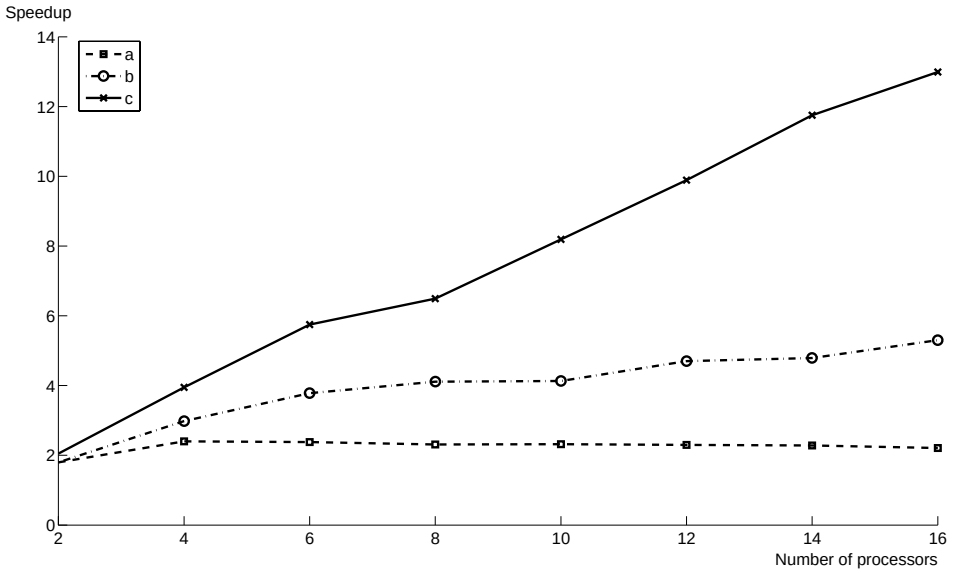

Рис. 2: Ускорение времени работы алгоритма построения одного члена ансамбля: a) при отсутствии разбиения группы процессоров на подгруппы на всех этапах b) при разбиении на каждом уровне построения дерева, c) при разбиении с использованием предложенного критерия.

В п. 2.4.1 проведен анализ эффективности MPI-реализации предложенного параллельного алгоритма. Реализация параллельного алгоритма выполнена на вычислительной системе IBM pSeries 690 Regatta. В параграфе представлены результаты ускорения параллельной программы построения полного ансамбля деревьев решений для обучающих множеств различной размерности, исследована зависимость среднего времени выполнения задачи построения одного дерева решений от порога для разбиения процессоров на подгруппы. Результаты эксперимента представлены на рис.2.

В п. 2.4.2 рассматриваются возможности параллельной реализации алгоритма локализации источников на многопроцессорных архитектурах с общей памятью, в том числе, многоядерных процессорах. Многопоточная реализация алгоритмов позволяет сократить время обработки данных ЭЭГ непосредственно в режиме их регистрации. В параграфе предлагается следующая схема параллельной многопоточной реализации алгоритма. Алгоритм разбивается на два блока. Первый блок состоит в независимом обучении каждого дерева решений в исходном ансамбле и построении множества "активных" путей для построенного дерева. Второй блок реализует параллельное выполнения кластеризации локализованных активных источников для группы сигналов. Декомпозиция алгоритма осуществляется по временным окнам усреднения. Многопоточная реализация описанного алгоритма выполнена с помощью технологии параллельного программирования OpenMP.

Третья глава посвящена разработке алгоритмов построения пространственно-

усредненных карт активности нейронных дипольных источников и анализу результатов применения разработанного комплекса программ для обработки экспериментальных данных при исследовании нейрофизиологических проблем восприятия.

В §3.1 описывается алгоритм построения пространственно-усредненных карт активности нейронных дипольных источников. Алгоритм основан на предложенном в работе методе кластерного анализа временных последовательностей параметров положений источников, использующем эвристику задачи локализации.

Входными данными алгоритма построения карт активности является множество временных последовательностей найденных параметров дипольных источников. В параграфе сформулирован общий алгоритм построения пространственных карт активности для группы сигналов ЭЭГ в фиксированном временном окне τ_0. Пусть дана группа N файлов с записью ЭЭГ сигналов одинаковой длины , и T - число измерений потенциала на поверхности головы, сделанное за время τ_0. Предположим, что в каждый временной отсчет были найдены параметры активного источника-диполя. Обозначим через $I = \{I_1, I_2, ..., I_N\}$ множество последовательностей вида

$$I_i = \{(r_t, \vartheta_t, \varphi_t, v_{r_t}, v_{\vartheta_t}, v_{\varphi_t})^t\}, \tag{4}$$

где i - номер файла ЭЭГ в рассматриваемой группе, $t = 1..T$, T – число анализируемых временных отсчетов, $(r_t, \vartheta_t, \varphi_t, v_{r_t}, v_{\vartheta_t}, v_{\varphi_t})^t$ - параметры найденного диполя для временного отсчета t. Тогда множеству I поставим в соответствие новое множество $I^K = \left\{I_1^{K_1}, I_2^{K_2}, ..., I_N^{K_N}\right\}$, где $K_1...K_N$ - некоторые числа, которые могут принимать целые значения в интервале $[1..T]$, а $I_i^{K_i} = \{(r_j, \vartheta_j, \varphi_j, v_{r_j}, v_{\vartheta_j}, v_{\varphi_j})^j, C_j, P_j\}^K$ - упорядоченное по всем C_j множество, т.е. $C_j > C_{j+1}$, $j = 1...U_i^{K_i}$, где $U_i^{K_i}$- количество всех неповторяющихся возможных положений $(r_i, \vartheta_i, \varphi_i)$ диполя для всех временных моментов множества $I_i^{K_i}$, $(r_j, \vartheta_j, \varphi_j, v_{r_j}, v_{\vartheta_j})^j$ - параметры найденного диполя для временных отсчетов интервала $[P_j - C_j]$, C_j - число временных отсчетов, начиная с P_j, для которых найденные параметры диполя не меняли значений.

В параграфе описывается разработка алгоритма кластеризации источников множества I, основанная на специфике задачи локализации источников, которая заключается в следующих утверждениях:

1. Положения центроидов и порядок их рассмотрения заданы априори множеством I^K.
2. Центроид кластера не может менять свое положение (исключается из множества кластеров).
3. Число кластеров неизвестно и итеративно определяется минимизацией критерия функционала качества, где степени принадлежности i-го объекта j-му кластеру задается отношением $w_{ij} = (\sum_{g=1}^{l} (\frac{\|x_i - v_j\|}{\|x_i - v_g\|})^{2/m-1})^{-1}$, где l – число кластеров, $m > 1$.

4. Все точки, имеющие степени принадлежности всем кластерам, которые отличаются на заданную пороговую величину, исключаются из рассмотрения.

В §3.2 описана проблематика нейрофизиологических исследований, для которых применялись предложенные методы и разработанный программный комплекс.

Разработанный комплекс программ был применен для анализа различных стадий эксперимента зрительного восприятия человеком стимулов и последующего их влияния на мозговую деятельность для выявления новой информации о мозговой деятельности при визуальном восприятии лиц и связи электрических характеристик при обработке сигнала ЭЭГ.

В §3.3 приводится протокол нейрофизиологического эксперимента, характеристики исследуемых экспериментальных данных и цели математической обработки экспериментальных данных.

Целью обработки описанных экспериментальных данных было исследование предложенного метода локализации источников и построение усредненных карт активности для его применения в качестве способа оценки функционального состояния мозговой активности.

В §3.5 представлены результаты обработки экспериментальных данных. Выполнено сравнение областей возникновения источников активности и их смещения для четырех стадий эксперимента. Показано, что с помощью построения пространственно-временных карт активности дипольных источников, можно различить стадии эксперимента. Это свидетельствует о том, что данный алгоритм может быть использован для оценки функционального состояния мозговой деятельности.

С помощью кластерного анализа областей возникновения и их перемещения в мозге испытуемых на различных стадиях эксперимента выявлено наличие близких изменяющихся пространственно-временных активных структур.

Показано, что созданные методы и программные средства анализа полученных результатов обработки экспериментальных данных на реальной геометрии мозга позволяют интерпретировать положения найденных источников на мозговых структурах, а также позволяют решить проблему соотношения сигнал - шум и отфильтровывать случайные скачки дипольных источников, вызванные артефактными событиями.

В §3.6 проведено сравнение разработанных алгоритмов локализации источников с существующими подходами решения обратных задач ЭЭГ.

Сравнительный анализ предложенного алгоритма с другими методами показал, что он входит в число наиболее точных и устойчивых методов. Большое преимущество алгоритма было обнаружено при обработке сильно зашумленных модельных данных. Точность лока-

лизации алгоритма соответствовала методам adapR(RAP)-MUSIC [3], BK Beam [4] и MEM2 [5] для "нескрытых" положений модельного источника и в некоторых случаях превосходила алгоритмы при локализации "скрытых" зон. Алгоритм также определял изменения вектора силы модельного источника, и имел наименьшее число ложных локализаций среди пяти рассмотренных алгоритмов. В результате исследования можно сделать заключение, что предложенные алгоритмы локализации источников могут давать новую информацию о зонах активности и обрабатывать экспериментальный сигнал ЭЭГ, содержащий значительное число артефактов.

В **заключении** перечислены основные результаты диссертации.

Основные результаты работы

1. Предложен новый метод автоматизированного анализа сигнала ЭЭГ, основанный на ансамбле деревьев решений, который позволяет устойчиво выделять электрически активные зоны мозга, отвечающие за регистрируемый пространственно-временной сигнал ЭЭГ. Предложен алгоритм параллельного обучения ансамблей классификаторов, позволяющий значительно сократить временные затраты на обработку сигнала.

2. Разработан новый метод построения усредненных пространственно-временных карт активности нейронных источников, основанный на кластерном анализе результатов локализации. Метод позволяет находить общие закономерности активности для группы сигналов ЭЭГ и отфильтровывать случайные скачки дипольных источников, вызванные артефактными событиями.

3. На основе предложенного подхода, разработанных методов и алгоритмов реализован программный комплекс для анализа сигналов ЭЭГ. Эффективность предложенного подхода продемонстрирована на примере использования комплекса для исследования нейрофизиологических проблем восприятия зрительной информации.

Публикации

Основные результаты диссертации опубликованы в работах:

1. Попова Е.А. Локализация нейронных источников электрической активности мозга с помощью метода Random Forest // *Программные Системы и Инструменты : Тематический сборник ф-та ВМиК МГУ.* — 2005. — N6. — С. 106-122.

[3] Efficient localization of synchronous EEG source activities using a modified RAP-MUSIC algorithm, Hesheng Liu, Schimp P.H. Biomedical Engineering, IEEE Transactions on Vol. 53, Is.4, pp. 652-661, 2006

[4] Electromagnetic brain mapping Baillet S, Mosher JC, Leahy RD. IEEE Signal Processing Magazine, vol. 18, pp. 14-30, 2001

[5] Three-dimensional EEG source imaging via maximum entropy method Khosla, D.; Singh, M.; Rice, D. Nuclear Science Symposium and Medical Imaging Conference Record, 2000, 2000 IEEE Volume 3, Issue , 21-28 Oct 2000 Page(s):1515 - 1519 vol.3

2. Попова Е.А. Разработка адаптивной системы машинного обучения основе метода деревьев решений // *Программные Системы и Инструменты : Тематический сборник ф-та ВМиК МГУ.* — 2005. — N6. — С. 17-28.

3. Попова Е.А. Обзор методов построения ансамблей классификаторов // *Программные Системы и Инструменты : Тематический сборник ф-та ВМиК МГУ.* — 2006. — N7. — С. 4-12.

4. Попова Е.А. Анализ электрической активности человеческого мозга на основе ансамблей деревьев решений // *Вестн. Моск. Ун-та Сер. 15 Вычисл. матем. и киберн.* — 2008. — N3. — С. 46-55.

5. Popova E. Ensemble of decision trees for neuronal source localization of the brain // *Abstract Book of the XXIII IUPAP International Conference on Statistical Physics*, Genova, Italy, 2007.

6. Popova E. Parallel ensemble of decision trees for neuronal source localization of the brain // *Abstract Book of the International Conference ParCo2007*, Aachen, Germany, 2007.

7. Попова Е.А. Метод параллельного построения комитета деревьев решений для обработки сигналов электроэнцефалографии // *Вестн. Моск. Ун-та Сер. 15 Вычисл. матем. и киберн.* — 2009. — N1.— С. 43-49 (в печати).

8. Попова Е.А. Анализ масштабируемости и производительности параллельного алгоритма построения ансамблей деревьев решений для задачи локализации нейронных источников // *Программные Системы и Инструменты : Тематический сборник ф-та ВМиК МГУ.* — 2008. — N8. — С. 15-25.