МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
Национальный исследовательский ядерный университет «МИФИ»

Факультет Кибернетики и информационной безопасности

Кафедра кибернетики (№ 22)

Направление подготовки 09.03.02 Информационные системы и технологии

Расширенное содержание пояснительной записки

к учебно-исследовательской работе студента на тему:

Разработка алгоритма классификации когнитивных состояний по данным фМРТ на основе анализа межиндивидуальных корреляций

Группа Б14-506

Студент _____ (подпись) Шедько А. Ю. (ФИО)

Руководитель _____ (подпись) Трофимов А. Г. (ФИО)

Научный консультант _____ (подпись) - (ФИО)

Оценка руководителя _____ (0-5 баллов) Оценка консультанта _____ (0-5 баллов)

Москва 2017

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Национальный исследовательский ядерный университет «МИФИ»

Факультет кибернетики и информационной безопасности

КАФЕДРА КИБЕРНЕТИКИ

Задание на УИР

Студенту гр. Б14-506 (группа) Шедько Андрею Юрьевичу (фио)

ТЕМА УИР

Разработка алгоритма классификации когнитивных состояний по данным фМРТ на основе анализа межиндивидуальных корреляций

ЗАДАНИЕ

№ п/п	Содержание работы	Форма отчетности	Срок исполнения	Отметка о выполнении Дата, подпись рук.
1.	**Аналитическая часть**			
1.1.	Изучение и анализ подходов к классификации когнитивных состояний по данным фМРТ (статическим и динамическим) применительно к задачам медицинской диагностики	Пункт ПЗ	1.03.17	
1.2.	Сравнительный анализ методов классификации многомерных данных (линейный дискриминантный анализ, метод опорных векторов, нейросетевые методы) для выбора подходящего набора алгоритмов.	подраздел ПЗ	8.03.17	
1.3.	Сравнительный анализ программных средств визуализации трехмерных данных фМРТ и исследование возможности их использования.	Текст ПЗ	8.03.17	
1.4.	*Оформление расширенного содержания пояснительной записки (РСПЗ)*	Текст РСПЗ	27.03.17	
2.	**Теоретическая часть**			
2.1.	Формальная постановка задачи классификации сигналов фМРТ.	подраздел ПЗ	5.03.17	
2.2.	Выбор и разработка показателей точности классификации когнитивных состояний по фМРТ.	Формулы, Выражения	10.03.17	
2.3.	Разработка алгоритма выявления значимых для классификации зон головного мозга на основе анализа межиндивидуальных корреляций.	подраздел ПЗ	14.03.17	
2.4.	Формальное описание алгоритма классификации когнитивных состояний по фМРТ.	рабочие материалы	20.03.17	
2.5.	Формальное описание схемы применения алгоритма для классификации когнитивных состояний в режиме реального времени.	Текст ПЗ	20.03.17	
3.	**Инженерная часть**			
3.1.	Проектирование программного пакета выполняющего классификацию когнитивных состояний по данным фМРТ на основе анализа межиндивидуальных корреляций	Текст ПЗ	1.04.17	
3.2.	Результаты проектирования оформить с помощью UML диаграммы модели.	UML диаграмма	1.04.17	

4.	**Технологическая и практическая часть**			
4.1.	Реализация программных модулей для экспериментальных исследований алгоритма классификации когнитивных состояний по фМРТ. с использованием программных сред MATLAB и Scipy.	Исполняемые файлы, исходный текст, подключаемый модуль для ЯП	21.03.17	
4.2.	Описание типов когнитивных состояний и исходных данных для проведения экспериментальных исследований разработанного алгоритма.	Текст ПЗ	15.03.17	
4.3.	Составление плана экспериментальных исследований разработанного алгоритма.	План эксперимента	1.04.17	
4.4.	Исследование точности классификации при различных способах оценки межиндивидуальных корреляций с использованием программных сред MATLAB и Scipy.	Схемы, графики, исходные тексты	10.04.17	
4.5.	Исследование показателей точности классификации, выявление наименее и наиболее разделимых когнитивных состояний и соответствующих зон головного мозга с использованием программных сред MATLAB и Scipy.	Схемы, графики	10.04.17	
5.	*Оформление пояснительной записки (ПЗ) и иллюстративного материала для доклада*.	Текст ПЗ, презентация	15.05.17	

ЛИТЕРАТУРА

[1]	Дьяконов В. П. MATLAB. Полный самоучитель. – М.// ДМК Пресс, 2012. – 768 с.: ил.
[2]	*Pajula Juha, Kauppi Jukka-Pekka, Tohka Jussi*. Inter-Subject Correlation in fMRI: Method Validation against Stimulus-Model Based Analysis // PLOS ONE. — 2012. — 08. — Vol. 7, no. 8. — Pp. 1–13.
[3]	*Pereira Francisco, Mitchell Tom, Botvinick Matthew*. Machine learning classifiers and fMRI: A tutorial overview // NeuroImage. — 2009. — Vol. 45, no. 1, Supplement 1. — Pp. S199 – S209. — Mathematics in Brain Imaging. http://www.sciencedirect.com/science/article/pii/S1053811908012263.
[4]	*Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome.* The elements of statistical learning: data mining, inference and prediction – 2 edition – Springer, 2009.
[5]	ГОСТ Р 7.0.53-2007 Система стандартов по информации, библиотечному и издательскому делу. Издания. Международный стандартный книжный номер. Использование и издательское оформление. — М.: Стандартинформ, 2007. — 5 с.
[6]	*Буч Г., Рамбо Д., Джекобсон А.* Язык UML. Руководство пользователя: Пер. с англ. М.// ДМК, 2007
[7]	*Kauppi J. P. et al.* Clustering inter-subject correlation matrices in functional magnetic resonance imaging //Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on. – IEEE, 2010. – С. 1-6.
[8]	*Ivezić Ž. et al.* Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. //Princeton University Press, 2014.
[9]	*Pajula Juha*. Inter-Subject Correlation Analysis for Functional Magnetic Resonance Imaging: Properties and Validation. Tampere University of Technology. Publication. — Tampere University of Technology, 2016. — 4. — Awarding institution:Tampere University of Technology.

Дата выдачи задания: Руководитель _____(_____)
(подпись) (фио)

« » 2017г. Студент _____(_____)
(подпись) (фио)

Реферат

Пояснительная записка содержит страниц (из них XX страниц приложений). Количество использованных источников – XX. Количество приложений – X.

Ключевые слова: Межиндивидуальная корреляция, Машинное обучение, классификация, фМРТ, Кластеризация.

Целью данной работы является описание применения Межиндивидуальной корреляции для кластеризации признаков при анализе неестественных стимулов в фМРТ.

В первой главе проводится обзор и анализ

Во второй главе описываются использованные и разработанные/модифицированные методы-/модели/алгоритмы ….

В третьей главе приводится описание программной реализации и экспериментальной проверки ….

В приложении А приведены исходные тексты некоторых программ

Содержание

Введение

В настоящее время актуальны проблемы анализа многомерных данных, особенно в медицинских приложениях. Данная работа рассматривает новый подход задаче понижения размерности: матрицу Межиндивидуальных корреляций. Акттивно публикуются в этой области: Juha Pajula из университета Тампере (в 2016 году защитившего диссертацию по данной теме [1]), Jussi Tohka, Jukka-Pekka Kauppi, Юрия Хассона, впервые описавшего данный метод. Первое упоминание применения метода для задачи кластеризации данных фМРТ можно найти в статье Юрия Хассона и других в 2004[2]. Однако в этой работе рассматривались естественные стимулы (просмотр фильмов) что не соотносится с доступными авторам данными (вербальные и пространственные задачи). Из-за разреженности данных без модификации методы предыдущих исследований не применимы без модификаций.

Таким образом, получим задачу данной работы — использование метода межиндивидуальной корреляции для кластеризации данных в задаче понижения размерности. Также проводится сравнение нового метода с традиционными подходами к данной задаче, не использующими множество испытуемых (T-статистика, Обобщённая линейная модель (GLM)).

Новизна работы состоит в применении метода ISC для кластеризации в условиях неестественных стимулов.

В первой главе подробно рассматриваются теоретические аспекты задачи понижения размерности, задачи классификации (Метод опорных векторов (SVM), нейронные сети, Линейный дискриминантный анализ (ЛДА)) и специфических для проблемной области (фМРТ) подходов к анализу данных. Также описываются программные средства визуализации трёхмерных данных с примерами их использования. (`nilearn.plotting`[3], `matplotlib3d`[4], `NIFTI`, `MITK`[5])

Во второй главе описаны используемые в работе алгоритмы, а именно: кластеризация на основе ISC, формирование вектора признаков, вычисление показателей точности классификации, классификация в режиме реального времени.

В третьей главе рассматриваются программные аспекты реализации алгоритмов описанных в предыдущей главе.

В заключительной главе описывается характер экспериментальных данных и количественные показатели точности работы системы. Также проводится исследование эффективности различных показателей точности классификации применительно к конкретным экспериментальным данным.

1. Анализ проблематики задач классификации когнитивных состояний

Аннотация. *В первой главе подробно рассматриваются теоретические аспекты задачи понижения размерности, задачи классификации (Метод опорных векторов (SVM), нейронные сети, Линейный дискриминантный анализ (ЛДА)) и специфических для проблемной области (фМРТ) подходов к анализу данных. Также описываются программные средства визуализации трёхмерных данных с примерами их использования. (*`nilearn.plotting`*[3],* `matplotlib3d`*[4],* `NIFTI`*,* `MITK`*[5])*

1.1 Изучение и анализ подходов к классификации когнитивных состояний по данным фМРТ (статическим и динамическим) применительно к задачам медицинской диагностики

Аннотация. *Для каждого образца объекта или события с известным классом y рассматривается набор наблюдений x (называемых ещё признаками, переменными или измерениями). Набор таких образцов называется обучающей выборкой (или набором обучения, обучением). Задачи классификации состоит в том, чтобы построить хороший прогноз класса y для всякого так же распределённого объекта (не обязательно содержащегося в обучающей выборке), имея только наблюдения x.*

В роли объектов выступают пациенты. Признаки характеризуют результаты обследований, симптомы заболевания и применявшиеся методы лечения. Примеры бинарных признаков: пол, наличие головной боли, слабости. Порядковый признак — тяжесть состояния (удовлетворительное, средней тяжести, тяжёлое, крайне тяжёлое). Количественные признаки — возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата. Признаковое описание пациента является, по сути дела, формализованной историей болезни. Накопив достаточное количество прецедентов в электронном виде, можно решать различные задачи:

- классифицировать вид заболевания (дифференциальная диагностика);
- определять наиболее целесообразный способ лечения;
- предсказывать длительность и исход заболевания;
- оценивать риск осложнений;

– находить синдромы — наиболее характерные для данного заболевания совокупности симптомов.

Ценность такого рода систем в том, что они способны мгновенно анализировать и обобщать огромное количество прецедентов — возможность, недоступная специалисту-врачу.

1.2 Сравнительный анализ методов классификации многомерных данных

Аннотация. *Рассмотрим такие методы как: Метод опорных векторов (SVM), Линейный дискриминантный анализ (ЛДА), Логистическая Регрессия*

Вначале дадим общее определение *задачи классификации (обучения с учителем)*.

Существует неизвестная целевая зависимость — отображение $y^* : X \to Y$, значения которой известны только на объектах конечной обучающей выборки $\{(x_i, y_i)|i \in \overline{1, P}\}$, где P— количество примеров, $x_i \in X, y_i \in Y$, X– пространство входных признаков, чаще всего действительное векторное пространство ($\mathbb{R}^k$), Y– конечное множество классов. Часто множество Y является 2 элементным, в этом случае классификация называется *бинарной*. Требуется построить алгоритм $\alpha : X \to Y$, который для каждого $x \in X$ построить хороший прогноз класса y.

Говорят также, что алгоритм должен обладать способностью к обобщению эмпирических фактов, или выводить общее знание (закономерность, зависимость) из частных фактов (наблюдений, прецедентов).

Данная постановка является обобщением классических задач аппроксимации функций. В классической аппроксимации объектами являются действительные числа или векторы. В реальных прикладных задачах входные данные об объектах могут быть неполными, неточными, неоднородными, нечисловыми. Эти особенности приводят к большому разнообразию методов обучения с учителем. Далее приводятся описания некоторых методов бинарной классификации данных. Более подробные описания и обобщения на случай многоклассовой классификации приводятся на веб-сайте `machinelearning.ru` [6]

1.2.1 Логистическая регрессия

Пусть объекты описываются n числовыми признаками $f_j : X \to \mathbb{R}, \ j = 1, \ldots, n$. Тогда пространство признаковых описаний объектов есть $X = \mathbb{R}^n$. Пусть Y — конечное множество номеров (имён, меток) классов. Пусть задана обучающая выборка пар «объект, ответ» $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Случай двух классов Положим $Y = \{-1, +1\}$. В логистической регрессии строится линейный алгоритм классификации $a : X \to Y$ вида $a(x, w) = \text{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = \text{sign}\langle x, w\rangle$, где w_j — вес j-го признака, w_0 — порог принятия решения, $w = (w_0, w_1, \ldots, w_n)$ — вектор весов, $\langle x, w\rangle$ — скалярное произведение признакового описания объекта на вектор весов.

Предполагается, что искусственно введён «константный» нулевой признак: $f_0(x) = -1$. Задача обучения линейного классификатора заключается в том, чтобы по выборке X^m настроить вектор весов w. В логистической регрессии для этого решается задача минимизации эмпирического риска с функцией потерь специального вида:

$$Q(w) = \sum_{i=1}^{m} \ln\left(1 + \exp(-y_i\langle x_i, w\rangle)\right) \to \min_w. \tag{1.1}$$

После того, как решение w найдено, становится возможным не только вычислять классификацию a(x) = sign$\langle x, w\rangle$ для произвольного объекта x, но и оценивать апостериорные вероятности его принадлежности классам:

$$\mathbb{P}\{y|x\} = \sigma\left(y\langle x, w\rangle\right), \quad y \in Y, \tag{1.2}$$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ — сигмоидная функция. Во многих приложениях апостериорные вероятности необходимы для оценивания рисков, связанных с возможными ошибками классификации.

1.2.2 ЛДА

Линейный дискриминантный анализ (ЛДА), а также связанный с ним *линейный дискриминант Фишера* — методы статистики и машинного обучения, применяемые для нахождения линейных комбинаций признаков, наилучшим образом разделяющих два или более класса объектов или событий. Полученная комбинация может быть использована в качестве линейного классификатора или для сокращения размерности пространства признаков перед последующей классификацией.

Рассмотрим этот метод для случая 2 классов:

При ЛДА предполагается, что функции совместной плотности распределения вероятностей $p(\vec{x}|y = 1)$ и $p(\vec{x}|y = 0)$ - нормальны. В этих предположениях оптимальное байесовское решение – относить точки ко второму классу если отношение правдоподобия ниже некоторого порогового значения T:

$$(\vec{x} - \vec{\mu}_0)^T\Sigma_{y=0}^{-1}(\vec{x} - \vec{\mu}_0) + \ln|\Sigma_{y=0}| - (\vec{x} - \vec{\mu}_1)^T\Sigma_{y=1}^{-1}(\vec{x} - \vec{\mu}_1) - \ln|\Sigma_{y=0}| < T$$

Если не делается никаких дальнейших предположений, полученную задачу классификации называют квадратичным дискриминантным анализом (*англ. quadratic discriminant analysis, QDA*). В ЛДА делается дополнительное предположение о *гомоскедастичности* (т.е. предполагается, что ковариационные матрицы равны, $\Sigma_{y=0} = \Sigma_{y=1} = \Sigma$) и считается, что ковариационные матрицы имеют полный ранг. При этих предположениях задача упрощается и сводится к сравнению скалярного произведения с пороговым значением

$$\vec{\omega} \cdot \vec{x} < c$$

для некоторой константы c, где

$$\vec{\omega} = \Sigma^{-1}(\vec{\mu_1} - \vec{\mu_0}).$$

Это означает, что вероятность принадлежности нового наблюдения x к классу y зависит исключительно от линейной комбинации известных наблюдений.

1.2.3 SVM

Что предпринимать, если данные не гомоскедастичны? Рассмотрим метод опорных векторов, для чего вначале дадим определение метода.

Метод опорных векторов (*англ. SVM, support vector machine*) — набор схожих алгоритмов обучения с учителем, использующихся для задач классификации и регрессионного анализа. SVM в чистом виде – линейный классификатор.

Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Рассмотрим задачу нахождения наилучшего в некотором смысле разделения множества векторов на два класса с помощью линейной решающей функции. Пусть имеется множество прецедентов (Ξ, Y), где $\Xi = \{\vec{x}_1, ..., \vec{x}_N\}$ — обучающая выборка, а $Y = (y_1, ..., y_N)$ — множество меток двух классов ω_1 и ω_2. Требуется по обучающей выборке построить линейную решающую функции, т.е. такую линейную функцию $f(\vec{x})$, которая удовлетворяла бы условию

$$f(\vec{x}_i) > 0 \vec{x}_i \in \omega_1, f(\vec{x}_i) < 0 \vec{x}_i \in \omega_2.$$

Без ограничения общности можно считать, что метки классов равны

$$y_i = \{\ 1,\ \vec{x}_i \in \varpi_1,\ -1,\ \vec{x}_i \in \varpi_2.$$

Тогда поставленную выше задачу можно переформулировать следующим образом. Требуется найти линейную решающую функцию $f(\vec{x})$, которая бы удовлетворяла условию

$$y_i f(\vec{x}_i) > 0 \text{ для всех } \vec{x}_i \in \Xi \tag{1.3}$$

Умножая, если нужно, функцию f на некоторое положительное число, нетрудно видеть, что система неравенств (1.3) равносильна системе

$$y_i f(\vec{x}_i) > 1 \vec{x}_i \in \Xi$$

Кроме того, так как $f(\vec{x})$ — линейная функция, то последняя система неравенств примет вид (1.4)

$$y_i((\vec{w}, \vec{x}_i) + b) \geq 1, \quad i = 1, ..., N, \tag{1.4}$$

где $\vec{w}$ — вектор весовых коэффициентов, b — некоторое число. Тогда разделяющей два класса гиперплоскостью будет $(\vec{w}, \vec{x}) + b = 0$. Нетрудно видеть, что и все гиперплоскости вида $(\vec{w}, \vec{x}) + b' = 0$, где $b' \in (b - 1, b + 1)$, также будут разделяющими. Расстояние между граничными гиперплоскостями $(\vec{w}, \vec{x}) + b - 1 = 0$ и $(\vec{w}, \vec{x}) + b + 1 = 0$ равно $\frac{2}{\|\vec{w}\|}$. Действительно, $\left(\frac{\vec{w}}{\|\vec{w}\|}, \vec{x}\right) + \frac{b-1}{\|\vec{w}\|} = 0$ и $\left(\frac{\vec{w}}{\|\vec{w}\|}, \vec{x}\right) + \frac{b+1}{\|\vec{w}\|} = 0$ — нормальные уравнения этих гиперплоскостей. Тогда $p_1 = \frac{b-1}{\|\vec{w}\|} p_2 = \frac{b+1}{\|\vec{w}\|}$ — расстояния от этих гиперплоскостей до начала координат и $\frac{2}{\|\vec{w}\|}$ — расстояние между гиперплоскостями. На самих граничных плоскостях может находиться некоторое число обучающих векторов. Эти векторы называются опорными.

Для надежного разделения классов необходимо, чтобы расстояние между разделяющими гиперплоскостями было как можно большим, т.е. $\|\vec{w}\|$ была как можно меньше. Таким образом, ставится задача нахождения минимума квадратичного функционала $0.5(\vec{w}, \vec{w})$ (коэффициент 0.5 вводится для удобства дифференцирования) в выпуклом многограннике, задаваемым системой неравенств (2). В выпуклом множестве квадратичный функционал всегда имеет единственный минимум (если это множество не пусто). Из теоремы Куна — Таккера следует, что решение этой оптимизационной задачи равносильно поиску седловой точки лагранжиана

$$\mathcal{L}(\vec{w}, b, \vec{\lambda}) = 0.5(\vec{w}, \vec{w}) - \sum_{i=1}^{N} \lambda_i (y_i((\vec{w}, \vec{x}_i) + b) - 1) \rightarrow \min_{\vec{w}, b} \max_{\vec{\lambda}}$$

в ортанте по множителям Лагранжа $\lambda_i \geqslant 0 \ \ (i = 1, ..., N)$, при условии, что $\lambda_i \left(y_i((\vec{w}, \vec{x}_i) + b) - 1\right) = 0, \quad i = 1, ..., N$. Последнее условие равносильно тому, что

$$\lambda_i = 0 y_i((\vec{w}, \vec{x}_i) + b) - 1 = 0, \quad i = 1, ..., N \tag{1.5}$$

Из необходимых условий существования седловой точки (полагая $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{in})$) имеем

$$\begin{cases} 0 = \dfrac{\partial L}{\partial w_j} = w_j - \displaystyle\sum_{i=1}^{N} \lambda_i y_i x_{ij}, \quad j = 1, ..., n, \\ 0 = \dfrac{\partial L}{\partial b} = \displaystyle\sum_{i=1}^{N} \lambda_i y_i. \end{cases}$$

Откуда следует, что вектор $\vec{w}$ следует искать в виде

$$\vec{w} = \sum_{i=1}^{N} \lambda_i y_i \vec{x}_i, \tag{1.6}$$

причем

$$\sum_{i=1}^{N} \lambda_i y_i = 0. \tag{1.7}$$

В силу (3) в сумму (4) с ненулевыми коэффициентами λ_i входят только те векторы, для которых $y_i((\vec{w}, \vec{x}_i) + b) - 1 = 0$. Такие векторы называют опорными, так как это именно те векторы, через которые будут проходить граничные гиперплоскости, разделяющие классы. Для найденного весового вектора $\vec{w}$ смещение b можно вычислить как $b = {y_s}^{-1} - (\vec{w}, \vec{x}_s)$ для любого опорного вектора $\vec{x}_s$. Найдем значения множителей Лагранжа, как критических точек лагранжиана. Для этого подставим (4) и (5) в лагранжиан, получим

$$\mathcal{L}(\vec{w}, b, \vec{\lambda}) = 0.5(\vec{w}, \vec{w}) - \sum_{i=1}^{N} \lambda_i (y_i((\vec{w}, \vec{x}_i) + b) - 1) = 0.5(\vec{w}, \vec{w}) - \left((\vec{w}, \vec{w}) - \sum_{i=1}^{N} \lambda_i \right) = \sum_{i=1}^{N} \lambda_i - 0.5(\vec{w}, \vec{w}) =$$

Таким образом, задача сводится к нахождению критических точек функции

$$\Phi(\vec{\lambda}) = \sum_{i=1}^{N} \lambda_i - 0.5 \left\| \sum_{i=1}^{N} \lambda_i y_i \vec{x}_i \right\|^2 . \tag{1.8}$$

Так как эта функция представляет собой разность линейной и квадратичной функций, причем квадратичная функция отрицательно определена, то требуется найти наибольшее значение функции $\Phi(\vec{\lambda})$ при условии $\sum_{i=1}^{N} \lambda_i y_i = 0$ в области $\lambda_i \geq 0 \quad (i = 1, ..., N)$. В теории оптимизации существует множество алгоритмов решения этой задачи (например, градиентные методы, метод покоординатного спуска и т.д.).

В 1992 году в работе Бернарда Бозера (Boser B.), Изабелл Гийон (Guyon I.) и Владимира Вапника был предложен способ адаптации машины опорных векторов для *нелинейного разделения классов*.

1.3 Сравнительный анализ программных средств анализа и визуализации трехмерных данных фМРТ и исследование возможности их использования

Аннотация. *В разделе описаны различные программные компоненты для визуализации нейроданных.*

Nilearn

Данная библиотека предоставляет с лёгкостью использовать продвинутые техники машинного обучения, распознавания образов и статистики на «нейроданных» для таких задач как MVPA (мно-

говоксельный анализ закономерностей, *англ. Mutli-Voxel Pattern Analysis*), декодирование, предиктивное моделирование и других.

`Nilearn` может быть использован для анализа данных фМРТ в состоянии покоя и в случае выполнения испытуемым задач. Данная библиотека создана на основе библиотеки `SciKit-Learn` для языка `python` в которой уже реализована значительная часть алгоритмов описанных выше.

Analyze

`Analyze` – ППП, разработанный в *Mayo Clinic* компанией Biomedical Imaging Resource (BIR) для многомерных отображения, обработки и измерения медицинских изображений различного типа. Это коммерческая программа, импользуемая для изучения томорамм, результатов фМРТ, компьютерной томографии, позитрон-эмиссионной томографии (PET).

MITK

`Medical Imaging Interaction Toolkit (MITK)` – свободная система с открытым исходным кодом для разработки интерактивного ПО для обработки медицинских изображений. Внутри себя, MITK содержит Insight Toolkit (ITK), Visualization Toolkit (VTK) и набор инструментов для разработки приложений. Разработана в *German Cancer Research Center Division of Medical and Biological Informatics*

1.3.1 SPM12

1.4 Выводы и постановка задачи курсового проекта

Это всегда последний пункт. Здесь, по-первых, приводятся, попунктно, основные вывода из проделанного анализа. Например:

1. Выполнен сравнительный анализ таких-то формальных систем с точки зрения применимости к решению задачи классификации. Из-за доступности и легкости их применения решено провести сравнение их успешности для этой задачи
2. Были проанализированы варианты программных архитектур на основе систем. С учетом требований к поддержке больших объемов данных и высоких требований к потенциалу модернизируемости, была выбрана за основу такая-то архитектура.
3. Сравнительный анализ таких-то библиотек показал, что библиотека X проще в использовании, но менее производительна, в то время как библиотека Y обеспечивает высокую производительность, но и требует значительных трудозатрат для использования. В связи с такими-то соображениями были принято решение использовать такую-то библиотеку.

2. Алгоритм классификации когнитивных состояний по данным фМРТ на основе метода межиндувидуальных корреляций

2.1 Формальная постановка задачи.

Аннотация. *Суть алгоритма: посмотреть какие воксели действуют схожим образом для каждого типа стимулов. Для этого применим метод Межиндивидуальных корреляций.*

Основные Определения и Описание данных

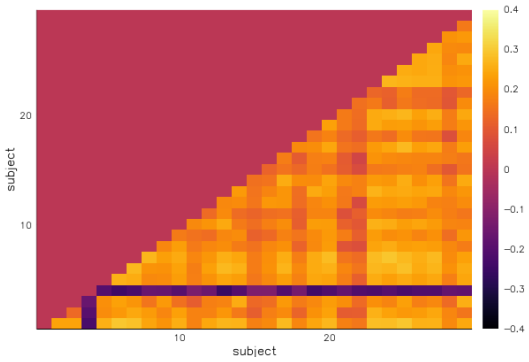
В работе используется несколько форматов представления данных:

- `NIfTI` — файлы для результатов сегментации и удобного взаимодействия со средствами визуализации. Формат является адаптацией ранее использовавшегося формата `ANALYZE™` и обладает множеством преимуществ перед последним. Например поддержкой различных типов данных (точность от 8 до 128 бит) ,хранением метаданных вместе с данными об объёме, возможностью хранить 4-d последовательности (x, y, z, t) и проч. [7]. В данной работе подразумевается что данные разбиты покадрово и находятся в пронумерованных директориях соответствующих испытуемым.
- `HDF5` — файлы использующиеся для хранения результатов промежуточных вычислений и «кэширования» данных для более удобной работы с ними.

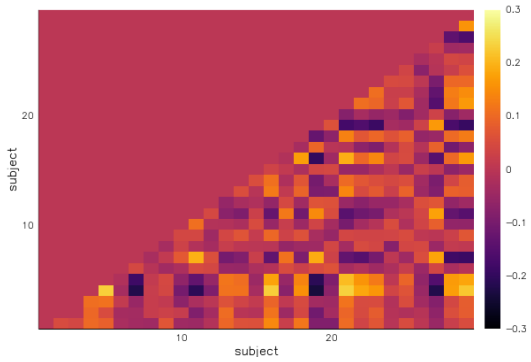
2.2 Алгоритм определения информативных вокселей фМРТ

Аннотация. *В данном разделе описывается алгоритм определения информативных вокселей. Применяются следующие подходы:*

- *Метод межиндивидуальных корреляций (ISC), состоящий в построении корреляционной матрицы для каждого вокселя для всех пар пациентов. Используется коэффициент корреляции Пирсона.*
- *Метод выделения t- статистики*
- *Обобщённая линейная модель.*



(а) Информативного вокселя.



(б) Неинформативного вокселя

Рис. 2.1 – Корреляционная матрица

Метод межиндивидуальных корреляций (ISC)

2.1

Метод t-статистик

В основе метода лежит предположение о том, что среднее значение информативных вокселей отличается во временных рядах, соответствующих различным стимулам. Напомним процедуру двухвыборочного t-критерия для независимых выборок[1]: Пусть имеются две независимые выборки объемами n_1 , n_2 нормально распределенных случайных величин X_1, X_2. Необходимо проверить по выборочным данным нулевую гипотезу равенства математических ожиданий этих случайных величин $H_0 : \ M_1 = M_2$

Рассмотрим разность выборочных средних $\Delta = \overline{X}_1 - \overline{X}_2$. Очевидно, если нулевая гипотеза выполнена $\mathbb{M}(\Delta) = M_1 - M_2 = 0$. Дисперсия этой разности равна исходя из независимости выборок: $V(\Delta) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$. Тогда используя несмещенную оценку дисперсии $s^2 = \dfrac{\sum_{t=1}^{n}(X_t - \overline{X})^2}{n-1}$ получаем несмещенную оценку дисперсии разности выборочных средних: $s_\Delta^2 = \dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}$. Следовательно, t-статистика для проверки нулевой гипотезы равна

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Эта статистика при справедливости нулевой гипотезы имеет распределение $t(df)$, где

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

[1]Выборки считаются независимыми так как в промежутках между стимулами пациент находится в состоянии покоя.

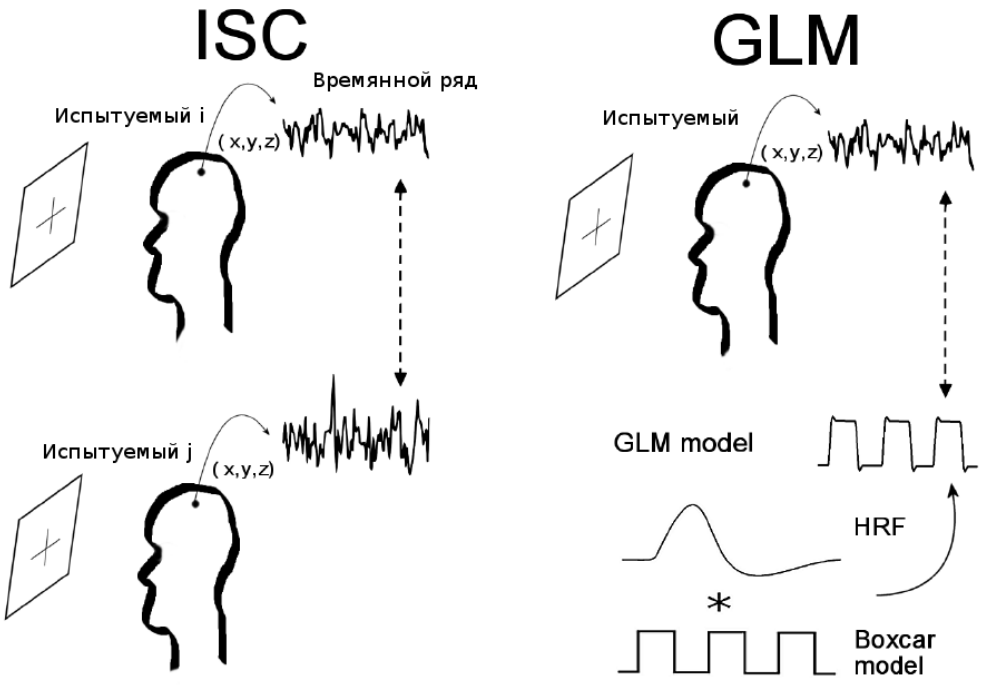


Рис. 2.2 – Сравнение принципов работы ISC и GLM

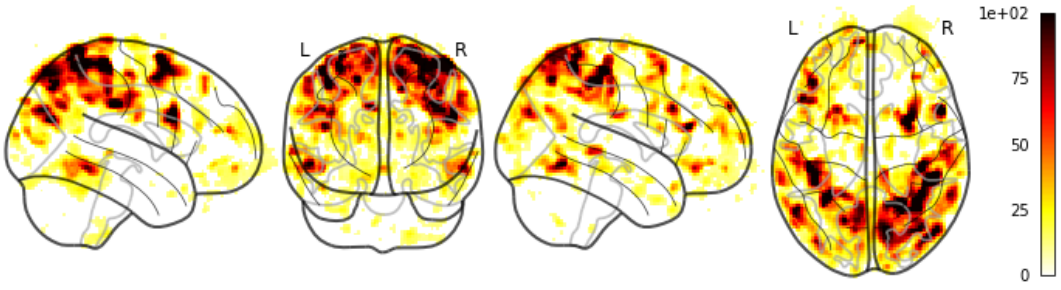


Рис. 2.3 – Сегментация с помощью *t*-статистики ($-log_{10}(P)$)

2.3 Алгоритм формирования вектора характерных признаков сигналов фМРТ для классификации

Аннотация. *Вектор признаков для классификатора формируется на основании результатов предыдущего шага с использованием метода главных компонент. Также убираются из рассмотрения признаки, не меняющиеся в зависимости от класса.*

Полученные в ходу предыдущего шага $\mathbb{S}$ данные фильтруются с помощью экспериментально выбранных пороговых значений, мотивированных количеством признаков получаемых в ходу анализа. Функция принадлежности вокселя (i, j, k) к множеству вокселей, значимых для классификации $(\mathcal{F})$, имеет вид:

$$I((A_{i,j,k} > 0) \wedge (\mathbb{S}_{i,j,k} > a) \wedge (\mathbb{CV} > b) \wedge (\mathbb{M}. > 100))$$

2.4 Показатели точности классификации

Аннотация. *В качестве показателей точности используются:*

- *Чувствительность (SEN) и специфичность (SPE)*
- $f1$ *мера*

Опишем формулы вычисления этих показателей:

$$SEN = TP/P = TP/(TP + FN)$$
$$SPE = TN/N = TN/(TN + FP)$$
$$Precision = TP/(TP + FP)$$
$$Recall = TP/(TP + FN)$$
$$f1 - measure = \sqrt{Precision \cdot Recall}$$

2.5 Формальное описание схемы применения алгоритма для классификации когнитивных состояний в режиме реального времени.

Аннотация. *В этой секции описывается применение алгоритма классификации к данным, поступающим с фМРТ-аппарата. Суть раздела: отображение фМРТ–снимку или небольшому набору снимков класса когнитивного состояния. (*$\mathrm{f}(\mathbb{N}^3 \to \mathbb{R}) \to Y$*, где* Y*–конечный набор классов когнитивных состояний).*

2.6 Выводы

Был разработан алгоритм сегментации на основе коэффициента корреляции Пирсона, осуществляющий выделение значимых для классификации вокселей изображения при определённых допущениях о предварительной обработке данных и характере эксперимента. Он представляет собой упрощённый вариант алгоритма ISC, где выбор временного интервала осуществляется априорно, как выбор матрицы регрессоров в обобщённой линейной модели.

Также описан алгоритм сегментации с помощью двухвыборочного t-критерия. Он отличается от стандартного подхода дифференциацией вокселей на основе P-значения.

Для решения задачи сегментации также применен стандартный метод обобщённой линейной модели (GLM) в качестве контрольного. Используется его реализация в программном пакете SPM12.

Необходимо перечислить, какие теоретические результаты были получены с указанием степени новизны. Например: «Была разработана такая-то модель. Она представляет собой адаптированную версию модели X, в которой уравнение Z заменено на уравнение Z'». Ещefindmax(res) пример: «Была предложена такая-то архитектура, она отличается от типовой в том-то и том-то. Это позволяет избежать таких-то проблем.». При этом следует заниматься «высасыванием из пальца»: «Поставленная задача является типовой; для ее решения применены стандартные средства (перечислить, какие).».

3. Разработка программной системы для классификации сигналов фМРТ

3.1 Проектирование программного пакета выполняющего классификацию когнитивных состояний по данным фМРТ на основе анализа межиндивидуальных корреляций

Для проведения сравнительного анализа различных методов сегментации и классификации было решено использовать гибкую архитектуру типа ”Data pipeline”. Учитывая значительные затраты времени и ресурсов на использования метода ISC было принято решение реализовать алгоритм на языке Julia[8]

3.2 Программная реализация системы классификации

Аннотация. *Здесь будет описана реализация*

- *алгоритма быстрой загрузки примеров для обучения и кластеризации.*

В этом разделе обосновывается выбор инструментальных средств; одним из критериев выбора могут быть какие-либо требования к разрабатываемой системе, и если этих требований много, они могут быть выделены в отдельный раздел, или же в приложение. Этот пункт не пишется, если в аналитической главе был раздел, посвященный сравнительному анализу и выбору инструментальных средств.

3.3 Состав и структура реализованного программного обеспечения

Аннотация. *Разработанное приложение является подключаемой библиотекой для использования в среде ”интерактивных тетрадей” `jupyter`. В состав библиотеки входят:*

- *Модуль параллельной загрузки/выгрузки примеров/изображений*
- *Модуль кластеризаци, содержащий алогритмы, описанные в части 2.*
- *Модуль классификации, предоставляющий на выбор несколько классификаторов и их входные параметры.*
- *Модуль визуализации для удобного представления данных для последующегог анализа специалистом*

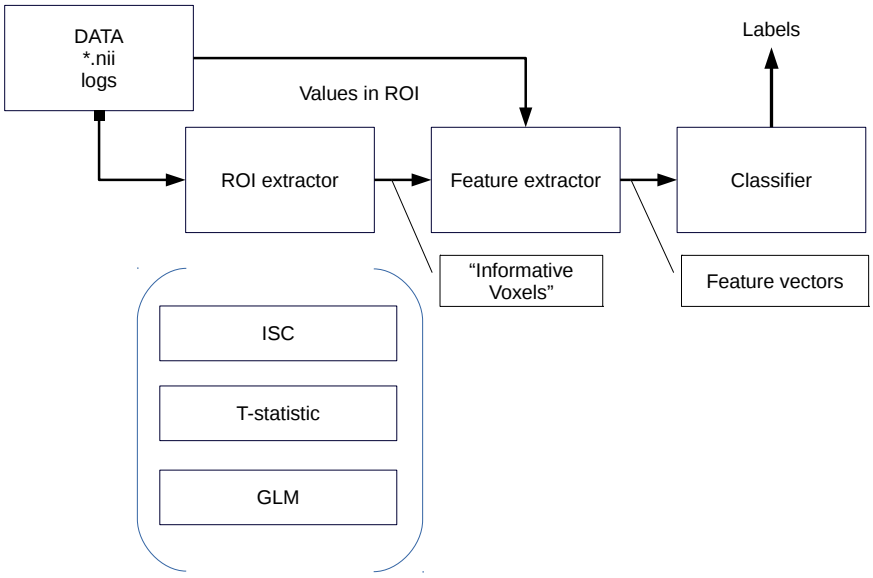


Рис. 3.1 – Архитектура разработанной системы

3.4 Основные сценарии работы пользователя

Аннотация. *Подразумевается следующий сценарий работы: пользователь подключатся к серверу интерактивных рабочих тетрадей декларативно описывает свои действия. Сценарий подразумевает загрузку дополнительных обучающих/тестовых выборок для проверки корректности работы алгоритмов.*

3.5 Сравнение реализованного программного обеспечения с существующими аналогами

Аннотация. *Автор не нашёл аналогов данного приложения по причине низкой востребованности.*

В сравнении должно быть отражено, чем полученное ПО выгодно (и невыгодно) отличается от прочих ближайших аналогов. Практика показывает, что аналоги есть всегда. А если нет аналогов, значит есть частичные решения, которые реализуют какие-то части функционала вашей системы. Тут тоже может быть относительно много таблиц и графиков.

3.6 Выводы

Следует перечислить, какие практические результаты были получены, а именно: какое программное или иное обеспечение было создано. В число результатов могут входить, например, методики тестирования, тестовые примеры (для проверки корректности/оценки характеристик тех

или иных алгоритмов) и др. По каждому результату следует сделать вывод, насколько он отличается от известных промышленных аналогов и исследовательских прототипов.

4. Экспериментальные исследования алгоритма классификации сигналов фМРТ

4.1 Описание исходных данных

S задач

примеров

испытуемых

временные отсчёты

Воксели и Мозги во времени

На рисунке 4.2 представлен результат работы набора команд представленного в приложении А.1

4.2 Составление плана экспериментальных исследований разработанного алгоритма

Что хотим

Параметры точности

Вопросы, ответы на которые хотим получить

При каком числе вокселей лучше точность

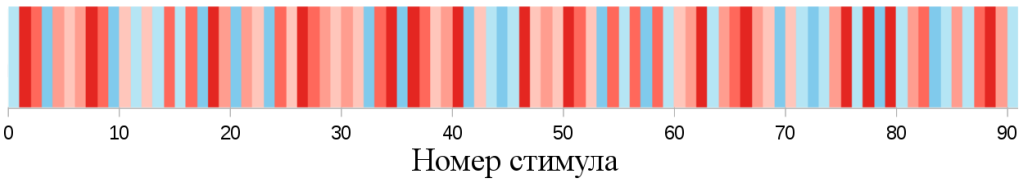


Рис. 4.1 – Дизайн эксперимента.

Оттенки красного обозначают стимулы типа **V**, синего — стимулы типа **S**

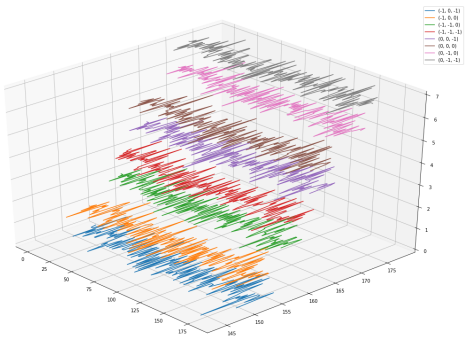


Рис. 4.2 – Пример активности локальной окрестности вокселя в течение $\approx$ 47 сек

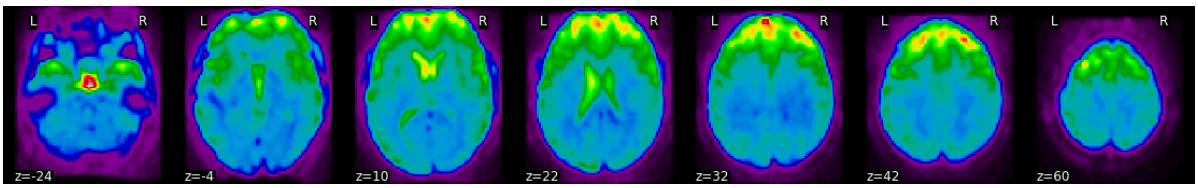


Рис. 4.3 – Разрезы мозга по z-координате

4.3 Исследование точности классификации при различных способах оценки межиндивидуальных корреляций

4.4 Исследование показателей точности классификации, выявление наименее и наиболее разделимых когнитивных состояний и соответствующих зон головного мозга

Графики, таблицы

ROC, AUC

4000

Бинарные признаки — объединение всех одинакового типа (S*, V*)в2 группы

Полученный результат находится на уровне state-of-the-art классификатора для разреженных данных по версии [9].

Валидация

В качестве подтверждения значимости шага сегментации при помощи ISC воспользуемся выбором случайных вокселей и их окрестностей в качестве признаков для классификатора:

Таблица 4.1 – Результаты многоклассовой классификации

Признак	Точность	Полнота	$f1$-мера	Количество
S1	0.77	0.69	0.73	90
S2	0.69	0.80	0.74	87
V1	0.72	0.67	0.70	85
V2	0.67	0.63	0.65	82
V3	0.73	0.65	0.69	84
V4	0.59	0.68	0.63	81
Среднее / Всё	0.69	0.69	0.69	509

Таблица 4.2 – Результаты многоклассовой для метода выделения t-статистики классификации

Признак	Точность	Полнота	$f1$-мера	Количество
S1	0.65	0.70	0.67	88
S2	0.61	0.68	0.64	80
V1	0.70	0.66	0.68	88
V2	0.55	0.65	0.60	71
V3	0.78	0.70	0.74	89
V4	0.72	0.62	0.67	92
Среднее / Всё	0.68	0.67	0.67	508

Таблица 4.3 – Результаты классификации для бинарных признаков

Признак	Точность	Полнота	$f1$-мера	Количество
S*	0.95	0.82	0.88	170
V*	0.91	0.98	0.94	339
Среднее / Всё	0.92	0.92	0.92	509

Таблица 4.4 – Результаты классификации для бинарных признаков при случайном выборе значимых вокселей

Признак	Точность	Полнота	$f1$-мера	Количество
S*	0.78	0.83	0.80	168
V*	0.91	0.89	0.90	341
Среднее / Всё	0.87	0.87	0.87	509

Таблица 4.5 – Результаты классификации для бинарных признаков при выборе значимых вокселей с помощью метода t-статистики

Признак	Точность	Полнота	$f1$-мера	Количество
S*	0.95	0.85	0.90	168
V*	0.93	0.98	0.95	340
Среднее / Всё	0.94	0.94	0.93	508

Таблица 4.6 – Результаты классификации для бинарных признаков при выборе значимых вокселей с помощью метода GLM

Признак	Точность	Полнота	$f1$-мера	Количество
S*	0.85	0.88	0.87	146
V*	0.95	0.94	0.94	362
Среднее / Всё	0.92	0.92	0.92	508

Заключение

В заключении в тезисной форме необходимо отразить результаты работы:

– аналитические (что изучено/проанализировано);

– теоретические;

– инженерные (что спроектировано);

– практические (что реализовано/внедрено).

Примерная формула такая: по каждому указанному пункту приводится по 3-5 результатов, каждый результат излагается в объеме до 5 фраз или предложений.

Также есть смысл привести предполагаемые направления для будущей работы.

Направления будущей работы

– Изучение применимости нелинейных классификаторов для задач Диагностики.

– Использование математических методов (online statistics, Tensor-Train Decomposition[10] для оптимизации времени работы и повышения точности классификации в задачах анализа фМРТ.

Список литературы

1. Pajula Juha. Inter-Subject Correlation Analysis for Functional Magnetic Resonance Imaging: Properties and Validation. Tampere University of Technology. Publication. — Tampere University of Technology, 2016. — 4. — ISBN: 978-952-15-3721-9. — Awarding institution:Tampere University of Technology.

2. Intersubject Synchronization of Cortical Activity During Natural Vision / Uri Hasson, Yuval Nir, Ifat Levy et al. // Science. — 2004. — Vol. 303, no. 5664. — P. 1634–1640. — http://science.sciencemag.org/content/303/5664/1634.full.pdf.

3. Machine learning for neuroimaging with scikit-learn / Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg et al. // Frontiers in Neuroinformatics. — 2014. — Vol. 8. — P. 14. — Access mode: `http://journal.frontiersin.org/article/10.3389/fninf.2014.00014`.

4. Hunter J. D. Matplotlib: A 2D graphics environment // Computing In Science & Engineering. — 2007. — Vol. 9, no. 3. — P. 90–95.

5. The Medical Imaging Interaction Toolkit (MITK)–a toolkit facilitating the creation of interactive software by extending VTK and ITK / Ivo Wolf, Marcus Vetter, Ingmar Wegner et al. // Proc. of SPIE Vol. — Vol. 5367. — 2004. — P. 17.

6. Классификация как раздел Машинного Обучения. — 2016. — Access mode: `http://www.machinelearning.ru/wiki/index.php?title=Классификация`.

7. A (sort of) new image data format standard: Nifti-1 / Robert W Cox, John Ashburner, Hester Breman et al. // Neuroimage. — 2004. — Vol. 22. — P. e1440.

8. Julia: A fresh approach to numerical computing / Jeff Bezanson, Alan Edelman, Stefan Karpinski, Viral B Shah // SIAM Review. — 2017. — Vol. 59, no. 1. — P. 65–98.

9. Zhang Lingsong, Lin Xihong. Some considerations of classification for high dimension low-sample size data // Statistical methods in medical research. — 2013. — Vol. 22, no. 5. — P. 537–550.

10. Oseledets Ivan V. Tensor-train decomposition // SIAM Journal on Scientific Computing. — 2011. — Vol. 33, no. 5. — P. 2295–2317.

Приложение А. Исходные тексты программ

```
def ind(i,l):
    return (i % l -1, (i % (l**2) // l)-1,
            i % (l**3) // l**2 -1)
def geti(x,i):
    return list(map(lambda el: el[i], x))
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
for (i,j) in zip(list(np.argsort([np.mean(geti(x,i))
for i in range(len(r)**3)])),range(len(r)**3)):
    ax.plot(range(183),geti(x,i),j, label =
    "{}".format(ind(i,len(r))));
```

Листинг А.1 – Код для иллюстрации