ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Программная инженерия»

УДК 004.62, 004.8, 004.91, 515.1, 612.8

СОГЛАСОВАНО	УТВЕРЖДАЮ
Доцент департамент математики факультета экономических наук НИУ ВШЭ, кандидат физико-математических наук	Академический руководитель образовательной программы «Программная инженерия» старший преподаватель департамента программной инженерии
Е. В. Михайлец	Н.А. Павлочев
«__» _____ 2024 г.	«__» _____ 2024 г.

Отчет

по исследовательскому курсовому проекту

на тему “Нахождение скрытых функциональных состояний по ЭЭГ, основываясь на топологических признаках”

по направлению подготовки бакалавров 09.03.04 «Программная инженерия»

Выполнил
студент группы БПИ213
образовательной программы
09.03.04 «Программная инженерия»

Абрамов Александр Сергеевич


23 марта 2024

Москва 2024

РЕФЕРАТ

Отчёт 69 с., 1 кн., 13 рис., 5 табл., 41 источн., 1 прил.

ВРЕМЕННОЙ РЯД, ЭЭГ, ПОИСК ФУНКЦИОНАЛЬНЫХ СОСТОЯНИЙ, КЛАСТЕРИЗАЦИЯ, STATE-DETECTING ALGORITHM, ТОПОЛОГИЧЕСКИЙ АНАЛИЗ ДАННЫХ, ТОПОЛОГИЧЕСКИЕ ПРИЗНАКИ, АНАЛИЗ ИНФОРМАЦИОННОЙ ЦЕННОСТИ

Объектом исследования являются ЭЭГ процесса медитации по методу Tantric Guhyasamaja, для которых недавно разработанный алгоритм SDA показывает хорошее качество нахождения функциональных состояний по традиционным признакам.

Цель работы – оценить применимость другого подхода к извлечению признаков, основанного на топологическом анализе данных.

В рамках проекта были изучены и реализованы основные методы топологического анализа данных ЭЭГ, разработан алгоритм извлечения практически 20000 признаков, выбраны метрики оценки их качества, включая анализ информационной ценности, а также произведена оценка их применимости к решаемой задаче.

В результате работы алгоритм SDA был впервые применён к топологическим признакам, полученным по сигналу ЭЭГ, и показал возможность нахождения ответа, близкого к ранее известному. Тем не менее был замечен и ряд интересных различий, которые могут свидетельствовать как о потере информации топологическими признаками, так и об обнаружении новых, ранее неизвестных закономерностей.

На основании этого сделан вывод о целесообразности проведения дальнейших исследований с применением дополнительных автоматизированных методов оценки качества результатов и возможным привлечением экспертов в области нейрофизиологии для их оценки не только на основании вычисленных метрик качества, но и с учётом существующих знаний о мозговой активности человека во время изучаемого процесса.

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящем отчете о НИР применяют следующие термины, сокращения и обозначения с соответствующими определениями:

Временной ряд – набор данных, показывающий изменение значений параметров какого-либо процесса с течением времени

Гиперпараметр – параметр, управляющий поведением модели или процессом её обучения, задаваемый до начала обучения и не изменяющийся в процессе

Детерминированная система – система, для которой состояние в любой момент времени однозначно определяется начальными условиями и может быть предсказано

Кластер – набор (множество, группа) объектов, получаемый в результате кластеризации

Кластеризация – метод анализа данных, производящий разделение набора объектов на группы (кластеры) на основании сходства их признаков

Метрика качества кластеризации – численное значение, показывающее, насколько хороший результат был получен при применении алгоритма кластеризации

Метрическое пространство – множество (набор) точек некоторой размерности с корректно заданной на нём функцией расстояния

Многомерный временной ряд – временной ряд, описывающий динамику значений двух или более параметров; совокупность одномерных временных рядов

Одномерный временной ряд – временной ряд, описывающий изменение значения одного параметра с течением времени

Признак – численное значение, описывающее некоторый объект

Признаковое описание – набор (вектор) признаков некоторого объекта

Пространство признаков – векторное (линейное) пространство, в котором лежат векторы признаков

Случайная величина – числовая переменная, значения которой описывают изучаемый случайный процесс

ЭЭГ – (сокр.) электроэнцефалограмма

Coherence – (англ.) когерентность – один из традиционных для работы с временными рядами признаков, являющийся мерой синхронизации двух временных рядов

PLV – (сокр.) phase-locking value; (англ.) коэффициент фазовой синхронизации – один из традиционных для работы с временными рядами признаков, являющийся мерой синхронизации фаз двух сигналов

PSD – (сокр.) power spectral density; (англ.) спектральная плотность мощности – один из традиционных для работы с временными рядами признаков, описывающий распределение мощности сигнала по частотам

SDA – (сокр.) state-detecting algorithm; (англ.) алгоритм определения функциональных состояний

$||z|| = \sqrt{\sum_{i=1}^{n} z_i^2}$ – евклидова норма вектора z размерности n; то же, что $||z||_2$

$||z||_1 = \sum_{i=1}^{n} |z_i|$ – манхэттенская норма вектора z размерности n

$||z||_p = \sqrt[p]{\sum_{i=1}^{n} |z_i|^p}$ – p-норма вектора z размерности n

$|Z|$ – количество элементов во множестве Z

ВВЕДЕНИЕ

Традицонно, анализ ЭЭГ полагается на информацию о событиях, искусственно созданных во время эксперимента (внешнее взаимодействие с объектом, реакция на него, получение ответа и др). Тем не менее большую ценность для нейрофизиологии представляет задача анализа ЭЭГ непрерывных процессов, в ходе которых невозможно внешнее взаимодействие с объектом. Для решения этой задачи в [1] был разработан алгоритм SDA, способный выделять функциональные состояния по ЭЭГ в отсутствие какой-либо дополнительной информации о происходивших во время эксперимента событиях (в том числе, об их количестве) с помощью методов кластеризации данных. Алгоритм состоит из двух основных этапов:

1) Нахождение потенциальных границ функциональных состояний с помощью иерархического метода кластеризации с возможностью задания матрицы связности при различных значениях гиперпараметров;

2) Выбор лучших границ функциональных состояний путём применения подходящего алгоритма кластеризации к совокупностям полученных на первом этапе потенциальных ответов.

Для применения алгоритма требуется предварительно произвести очистку данных ЭЭГ, разделить их на эпохи по времени и представить соответствующие временные ряды в некотором пространстве признаков. Как показано в [1], алгоритм демонстрирует хорошее качество для ЭЭГ, полученных во время медитации буддистских монахов по методу Tantric Guhyasamaja при извлечении традиционных для работы с временными рядами признаковых описаний: PSD, PLV и Coherence.

Тем не менее в настоящее время стремительно набирает популярность другой подход, основанный на использовании алгебраической топологии и заключающийся в анализе пространственной структуры данных. Известно, что в некоторых задачах, связанных с временными рядами, в том числе в области физиологии, топологические признаки могут быть удобнее в использовании, позволяют достичь лучшего качества и относительно просты в интерпретации. В рамках исследования требуется проверить, подходят ли такие признаки для задачи анализа ЭЭГ непрерывных процессов, для которых отсутствует информация о событиях, происходивших во время её записи. Положительный результат откроет новую ветвь

исследований в области нейрофизиологии и позволит эффективнее и качественнее находить функциональные состояния по ЭЭГ.

Работа содержит один промежуточный отчёт первого этапа проекта.

1. НАБОР ДАННЫХ

Для проведения исследования были использованы данные, собранные и предобработанные авторами исходной статьи [1]. Краткое описание использованных методов приведено далее в настоящем отчёте.

1.1. Сбор данных

Guhyasamaja Tantra – традиционный процесс медитации, производимый в соответствии со строгими принципами, закреплёнными священными писаниями буддизма. Его основная часть состоит из 8 последовательных стадий, длительность которых может варьироваться в различных исполнениях.

Для получения данных был проведён эксперимент, в ходе которого были записаны ЭЭГ успешных процессов медитации трех тибетских буддистских монахов, практикующих Guhyasamaja Tantra в течение многих десятилетий. Регистрация сигналов производилась с помощью системы NVX-52 при частоте дискретизации 500 Гц с аналоговой полосовой фильтрацией от 0,1 до 200 Гц и режекторным фильтром на частоте 50 Гц для удаления артефактов, вызванных линией электропередач. В результате были получены записи длительностью 935, 2344 и 1302 секунды (далее, Объект – 1, Объект – 2 и Объект – 3 соответственно).

1.2. Предобработка данных

Предобработка данных производилась на языке Python с помощью библиотеки MNE [2]. Для очистки ЭЭГ от шумов и другой информации, не относящейся к интересующему процессу (дыхательная и мышечная активность, моргание глаз, биение сердца и др.), был применён ряд методов, включая полосовой фильтр на частотах 0.9 – 40 Гц и анализ независимых компонент. Для дальнейшей работы данные были разбиты на эпохи по времени длительностью 1 секунда, среди которых было удалено 10 – 15% наиболее отклоняющихся от среднего по спектральной плотности мощности. Для увеличения объема данных, доступных для объекта 1, разделение на эпохи производилось с наложением в 0,2 секунды; для других объектов наложение не применялось. В результате были получены массивы эпох длины 1046, 2019 и 1180 соответственно, где каждая эпоха является многомерным временным рядом, описанным матрицей размера 40 x 501. Каждый из 40 каналов соответствует одному из электродов, использованных во время эксперимента.

Для удобства анализа аналогично исходной статье электроды были условно разделены на 17 групп по частям головы, в которых они производили запись, как показано на рисунке 1: префронтальная (Fp1, Fp2, Fpz), левая фронтальная (F3, F7, FC3, FT7), средняя фронтальная (Fz, FCz), правая фронтальная (F4, F8, FC4, FT8), левая центральная (C3, CP3), центральная (Cz, CPz), правая центральная (C4, CP4), левая височная (T3, T5, TP7), правая височная (T4, T6, TP8), левая теменная (P3, P5), средняя теменная (Pz), правая теменная (P4, P6), левая затылочная (PO3, PO7, O1), средняя затылочная (POz, Oz), правая затылочная (PO4, PO8, O2) и ушные электроды A1(левый) и A2 (правый).

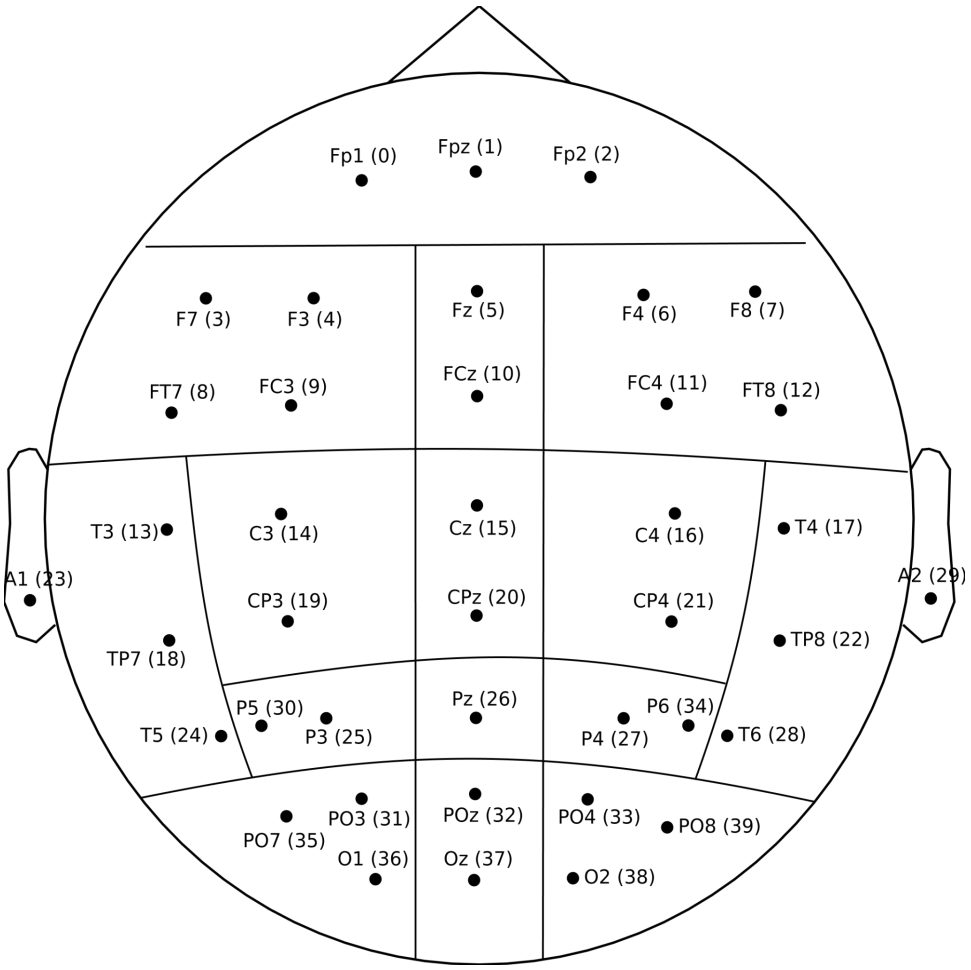

Рисунок 1 – Расположение электродов во время эксперимента и их разделение по группам

2. ПРИЗНАКОВОЕ ОПИСАНИЕ ДАННЫХ

2.1. Топологический анализ данных

Топологический анализ – это современный подход к анализу данных различного рода, основанный на понижении размерности без значительной потери информации путём изучения пространственных характеристик данных (формы, структуры и др.) методами алгебраической топологии. Главные инструменты такого подхода – устойчивые гомологии и диаграммы устойчивости, вычисляемые для метрических пространств и показывающие количество и стабильность многомерных “дырок” в них. Доказано [3], что незначительные изменения и деформации исходных данных не оказывают видимого влияния на результат, что позволяет использовать информацию, содержащуюся в диаграммах, для решения различных задач машинного обучения по следующему общему принципу:

1) Преобразование исходных данных (изображения, временного ряда, текста и др.) в метрическое пространство – набор точек с корректно определённой на нём функцией расстояния.

2) Построение последовательности вложенных симплициальных комплексов – топологических пространств, представляющих собой объединение простейших геометрических фигур различной размерности (отрезков, треугольников, тетраэдров и т.д.), называемых симплексами, образующихся в результате соединения точек, находящихся на расстоянии не более некоторого ε друг от друга. Общий алгоритм построения этой структуры заключается в наблюдении за появлением и исчезновением симплексов по мере увеличения значения параметра ε.

3) Вычисление устойчивой гомологии [4] полученной структуры по “времени” появления и исчезновения симплексов, построение и очистка от шума диаграммы устойчивости, в которой каждому симплексу соответствует одна точка с координатами по осям абсцисс и ординат, равными времени его появления и исчезновения соответственно.

4) Получение признакового описания исходных данных путём векторизации – вычисления различных топологических и статистических характеристик – диаграммы устойчивости в соответствии с основной теоремой устойчивых гомологий [5]. Полученные таким образом признаки могут использоваться в классических алгоритмах машинного обучения для решения различных задач.

В рамках настоящей работы вычисление признаков производилось несколькими способами (см. п. 2.16), после чего полученные векторы объединялись для получения итогового признакового описания.

Для упрощения реализации описанного процесса на языке Python были использованы библиотеки giotto-tda [6], numpy [7], scipy [8] и scikit-learn [9], а также matplotlib [10], pandas [11] и tqdm [12] для удобства визуального представления результатов и наблюдения за процессом работы алгоритмов.

2.2. Алгоритм получения эмбеддингов Такенса

2.2.1. Применимо к одномерным временным рядам

Одним из способов преобразования временных рядов в метрические пространства является алгоритм Такенса [13 – 14].

Пусть дан одномерный временной ряд $(X_1, X_2, ..., X_n)$, задана размерность d желаемого пространства и равномерно в интервале от 1 до n выбрана последовательность чисел $t_1, t_2, ..., t_q$ так, что разность между соседними элементами последовательности равна фиксированному значению s, называемому шагом (stride). Тогда координаты точки под номером $i \in [1; q]$ вычисляются по формуле (1).

$$Q_i = (X_{t_i}, X_{t_i + \tau}, X_{t_i + 2 \cdot \tau}, ..., X_{t_i + (d-1) \cdot \tau}), \quad (1)$$

где τ – фиксированная временная задержка между соседними координатами точки.

Описанный алгоритм реализован в библиотеке giotto-tda классом SingleTakensEmbedding, который в качестве параметров ожидает значения d, τ и s, на основе которых производится вычисление последовательности $t_1, t_2, ..., t_q$ так, чтобы последняя координата последней точки $(Q_{q,d})$ равнялась последнему значению исходного временного ряда (X_n).

2.2.2. Применимо к многомерным временным рядам

При работе с многомерными временными рядами описанный в п. 2.2.1 алгоритм применяют независимо к каждой компоненте, объединяя полученные результаты. Таким образом, каждая компонента описывается набором точек размерности d, а весь временной ряд – метрическим пространством размерности $d \cdot nChannels$, где $nChannels$ – количество компонент в ряду.

Этот подход реализован в библиотеке giotto-tda классом TakensEmbedding.

2.2.3. Принципы подбора гиперпараметров

Для подбора гиперпараметров описанного алгоритма (d, τ и s) помимо классических подходов (например, кросс-валидации) существуют и другие способы, основанные на оптимизации значений различных метрик, лучше отражающих специфическую структуру временных рядов.

Описанные далее алгоритмы реализованы в библиотеке giotto-tda функцией takens_embedding_optimal_parameters и классом SingleTakensEmbedding с указанием parameters_type = ‘search’.

2.2.3.1. Алгоритм подбора параметра τ

Показано [15], что для определения лучшего значения временной задержки между соседними координатами точек (τ) целесообразно оптимизировать величину коэффициента взаимной информации следующим образом:

1. На вход алгоритму поступает одномерный временной ряд (X_1, X_2, ..., X_n);
2. Определяются его минимальное (X_{min}) и максимальное (X_{max}) значения;
3. Отрезок [X_{min}; X_{max}] равномерно разделяется на некоторое достаточно большое число контейнеров (в настоящей работе использовалось значение 100 – стандартное в библиотеке giotto-tda);
4. Для каждого рассматриваемого значения τ (обычно, в интервале [1; max_τ]) вычисляется коэффициент взаимной информации по формуле (2);

$$MI(\tau) = -\sum_{i=1}^{K}\sum_{j=1}^{K} P_{ij}(\tau) \cdot ln\left(\frac{P_{ij}(\tau)}{P_i \cdot P_j}\right), \quad (2)$$

где K – количество контейнеров;

P_i, P_j – вероятности попадания случайного элемента временного ряда в контейнеры под номерами i и j соответственно;

$P_{ij}(\tau)$ – вероятность попадания случайного элемента временного ряда X_t в контейнер под номером i при условии попадания элемента $X_{t+\tau}$ в контейнер под номером j.

5. Значение τ, соответствующее первому минимуму величины коэффициента взаимной информации, то есть "добавляющее" наибольшее количество информации в данные, считается оптимальным и подаётся на выход алгоритма.

2.2.3.2. Алгоритм подбора размерности эмбеддингов

Алгоритм определения оптимального значения параметра d основан на предположении, что повышение размерности данных, описывающих детерминированную систему, происходит "гладко", то есть точки, находящиеся рядом в пространстве маленькой размерности, находятся рядом и в пространстве большей размерности. Для количественной оценки этой метрики и выбора лучшей размерности метрического пространства предлагается [16] оптимизировать количество ложно-соседних пар точек:

1. На вход алгоритму поступает одномерный временной ряд (X_1, X_2, ..., X_n) и полученное ранее методом, описанным в п. 2.2.3.1, оптимальное значение τ;

2. Выберем Q_i и Q_j – "соседние" точки после применения алгоритма Такенса с $d = d_0$. В рамках настоящего исследования и в реализации, приведенной библиотекой giotto-tda, для заданной точки Q_i соседними считаются две точки с наименьшим евклидовым расстоянием до Q_i.

3. Будем считать Q_i и Q_j ложно-соседними, если вычисляемая по формуле (3) величина R_{ij}, показывающая отношение модуля разности добавляемых координат к

расстоянию между точками, превосходит некоторое пороговое значение R_{th}, принятое за 10 в библиотеке giotto-tda и в рамках настоящего исследования.

$$R_{ij} = \frac{|X_{i+d_0\tau} - X_{j+d_0\tau}|}{||Q_i - Q_j||}, \quad (3)$$

4. Значение d, при котором количество пар ложно-соседних точек минимально, считается оптимальным и подаётся на выход алгоритма.

2.3. Скользящее окно

Скользящее окно – частный случай алгоритма получения эмбеддингов Такенса при $\tau = 1$, вычисляющий точку Q_i размерности d как вектор из d последовательных элементов временного ряда, начиная с t_i. Алгоритм реализован в библиотеке giotto-tda классом SlidingWindow.

2.4. Пирсоновское несходство

Пирсоновское несходство [17] является мерой “похожести” распределений двух случайных величин и может служить способом преобразования многомерных временных рядов в метрические пространства, пригодные для топологического анализа.

Пусть даны случайные величины $X = (X_1, X_2, ..., X_n)$ и $Y = (Y_1, Y_2, ..., Y_n)$ – одномерные временные ряды, являющиеся компонентами многомерного временного ряда (ЭЭГ) с $nChannels$ переменными. Тогда их Пирсоновским несходством называют величину D_{XY}, лежащую в интервале $[0;\ 1]$ и вычисляемую по формуле (4).

$$\mathrm{C}_{XY} = \frac{1}{N}\sum_{i=1}^{N}\left((X_i - \frac{1}{N}\sum_{j=1}^{N} X_j)(Y_i - \frac{1}{N}\sum_{j=1}^{N} Y_j)\right),$$

$$R_{XY} = \frac{C_{XY}}{\sqrt{C_{XX} \cdot C_{YY}}}, \tag{4}$$

$$D_{XY} = \frac{1 - R_{XY}}{2},$$

где C_{XY} – выборочная ковариация случайных величин X и Y;

R_{XY} – коэффициент выборочной корреляции Пирсона случайных величин X и Y.

Матрицу значений Пирсоновского несходства для всех способов выбора пары X и Y из данного многомерного временного ряда называют матрицей несходства, которая задаёт набор из $nChannels$ точек с расстояниями, равными величинам D_{XY}. Полученное таким образом метрическое пространство пригодно для топологического анализа корреляций между компонентами временного ряда.

Описанный алгоритм реализован классом PearsonDissimilarity библиотеки giotto-tda.

2.5. Построение диаграммы устойчивости

Пусть дано метрическое пространство (X, d). Для вычисления его устойчивой гомологии и диаграммы устойчивости требуется для всех положительных значений параметра ε построить симплициальный комплекс, гомотопически эквивалентный объединению шаров радиуса ε вокруг точек из множества X. Теорема о нерве [18] гарантирует, что такому свойству удовлетворяет комплекс Чеха [19], содержащий те и только те симплексы, для которых множество шаров радиуса ε вокруг вершин имеет непустое пересечение.

Тем не менее построение такой структуры вычислительно является достаточно сложной задачей, из-за чего на практике обычно применяется приближение комплекса Чеха – комплекс Вьеториса – Рипса [20], который в общем случае не удовлетворяет требуемому свойству, но достаточно хорошо подходит для решения большинства практических задач. Комплекс Вьеториса – Рипса определяется формулой (5) и содержит те и только те симплексы, для которых все попарные расстояния между вершинами не превышают ε.

Примеры комплексов Чеха и Вьеториса – Рипса для некоторого набора точек представлены на рисунке 2.

$$VR_{\varepsilon}(X,\ d) = \{\,[x_1,\ x_2,\ \ldots,\ x_n] \mid \forall\ i,\ j \in [1;\ n]^2\ d(x_i,\ x_j) \leq \varepsilon\,\}, \tag{5}$$

где x_i – точка в метрическом пространстве $(X,\ d)$;

$[x_1,\ x_2,\ \ldots,\ x_n]$ – симплекс с вершинами $x_1,\ x_2,\ \ldots,\ x_n$.

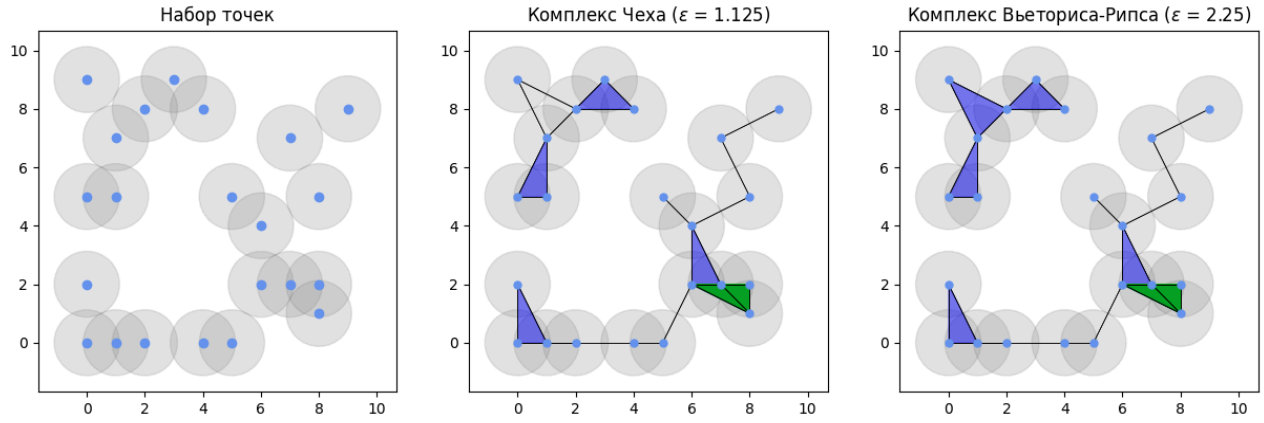

Рисунок 2 – Набор точек и комплексы Чеха и Вьеториса – Рипса для него

Для получения диаграммы устойчивости необходимо для каждого симплекса определить минимальное и максимальное значения ε, при которых он входит в соответствующий комплекс, и нанести точки на диаграмму, отложив “моменты” появления и исчезновения по осям абсцисс и ординат соответственно.

Известно большое количество алгоритмов построения комплексов Вьеториса – Рипса, устойчивых гомологий и диаграмм устойчивости [21 – 23]. Их строгие описания и конкретные реализации не являются существенными для настоящего исследования. В рамках работы была использована реализация, представленная классом VietorisRipsPersistence библиотеки giotto-tda, вычисляющая размерности, а также моменты появления и исчезновения симплексов для заданного метрического пространства.

2.6. Фильтрация диаграмм устойчивости

Особенный интерес для анализа представляют наиболее “долгоживущие” симплексы – те, для которых момент исчезновения значительно больше момента появления. На диаграмме устойчивости они соответствуют точкам, наиболее удалённым от диагонали. Остальные точки, в свою очередь, несут мало информации и с большой вероятностью описывают “шум” – закономерности, которые присутствуют в исходных данных, но не являются общими для изучаемых объектов.

Для повышения обобщающей способности алгоритмов машинного обучения, которые будут применяться к извлечённым признакам, целесообразно отфильтровать диаграммы – убрать точки, наиболее близкие к диагонали. В настоящей работе было принято решение удалять фиксированную долю точек (F) с наименьшем временем жизни.

2.7. Статистические характеристики

Помимо прочего, для векторизации отфильтрованных диаграмм устойчивости применялись традиционные статистические методы анализа распределений.

Пусть $L = (L_1, L_2, ..., L_n)$ – некоторая последовательность чисел, случайная величина. Тогда векторным описанием этой последовательности считался вектор, содержащий следующие значения:

1. Количество, сумма, среднее и стандартное отклонение элементов последовательности;
2. Максимальный элемент последовательности;
3. 25, 50 и 75 перцентили последовательности;
4. Манхэттенская (1) и евклидова (2) нормы L как вектора;
5. “Асимметрия” последовательности – отношение третьего центрального момента к кубу стандартного отклонения;
6. “Эксцесс” последовательности – отношение четвертого центрального момента к четвёртой степени стандартного отклонения.

При вычислении статистических характеристик функций производилось их преобразование в последовательности чисел путём выбора 100 значений равномерно на интересующем интервале.

В настоящем исследовании таким образом описывались последовательности времён жизни симплексов, последовательности средних значений между моментами их появления и исчезновения, кривые Бетти (см. 2.8) и др. Расчёты производились как для всей диаграммы, так и отдельно для симплексов каждой размерности.

Вычисления выполнялись с помощью библиотек numpy и scipy языка Python, реализующих эффективное нахождение описанных характеристик.

2.8. Числа и кривые Бетти

Пусть симплициальный комплекс K состоит из симплексов размерностей $(s_1, s_2, \ldots, s_n)$, где $s_i \in [a; b]$ – рассматриваются симплексы размерностей от a до b. Для нахождения числа Бетти [5] под номером k (B_k, $k \in [a; b]$) определим δ_k как отображение, сопоставляющее каждому симплексу размерности k его границу – набор симплексов размерности $(k-1)$. Тогда B_k вычисляется по формуле (6) как ранг – максимальное количество линейно-независимых векторов – k-ого гомологического пространства – факторпространства ядра δ_k по образу δ_{k+1}.

$$B_k = Rg\left(Ker(\delta_k) \,/\, Im(\delta_{k+1})\right) \tag{6}$$

Числа Бетти можно интерпретировать как максимальное количество разрезов поверхности, которые могут быть сделаны до её разделения на две части. Можно также доказать, что B_k равно количеству k-мерных “дыр” в исходном метрическом пространстве, что эквивалентно количеству симплексов размерности k.

Кривой Бетти под номером k называют функцию, сопоставляющую каждому значению ε величину k-го числа Бетти B_k в соответствующем симплициальном комплексе. Реализация алгоритма её вычисления представлена в библиотеке giotto-tda классом BettiCurve.

Тем не менее при анализе особенный интерес представляют не сами кривые Бетти, а закономерности изменения их значений, описываемые первой производной по ε. В рамках настоящей работы для векторизации диаграмм устойчивости и признакового описания данных использовались статистические характеристики (см. п. 2.7) первых производных кривых Бетти.

2.9. Энтропия устойчивости

Пусть дана диаграмма устойчивости D, содержащая n симплексов некоторой размерности, для каждого из которых известны моменты его появления и исчезновения – b_i и d_i соответственно. Тогда энтропия устойчивости [24] этой диаграммы вычисляется по формуле (7).

$$p_i = \frac{d_i - b_i}{\sum_{j=1}^{n} d_j - b_j},$$
$$E(D) = -\sum_{i=1}^{n} p_i \cdot ln(p_i), \tag{7}$$

Если диаграмма содержит симплексы нескольких различных размерностей, значения вычисляются независимо для каждой из них и объединяются в итоговый вектор. Полученные таким способом векторы пригодны для использования в классических методах машинного обучения.

Описанный алгоритм реализован в библиотеке giotto-tda классом PersistenceEntropy.

2.10. Ландшафт устойчивости

Уровнем k ландшафта устойчивости [25] диаграммы D, содержащей симплексы с временами появления и исчезновения $(b_1, b_2, ..., b_n)$ и $(d_1, d_2, ..., d_n)$ соответственно, называют функцию λ_k: $\mathbb{R} \to \mathbb{R}$, которая сопоставляет каждому значению ε величину k-ого максимального элемента во множестве $\{ \Lambda_i(\varepsilon) \mid i \in [1, n] \}$, где $\Lambda_i(\varepsilon)$ определяется формулой (8).

$$\Lambda_i(\varepsilon) = max(0,\ min(\varepsilon - b_i,\ d_i - \varepsilon)) \tag{8}$$

Алгоритм вычисления ландшафта устойчивости реализован в библиотеке giotto-tda классом PersistenceLandscape.

В качестве признакового описания диаграммы устойчивости применимы статистические характеристики (см. п. 2.7) одного или нескольких первых уровней её ландшафта. В рамках настоящей работы использовался только первый уровень.

2.11. Силуэт устойчивости

Взвешенным силуэтом [26] степени q диаграммы устойчивости D, содержащей симплексы фиксированной размерности с временами появления и исчезновения $(b_1,\ b_2,\ ...,\ b_n)$ и $(d_1,\ d_2,\ ...,\ d_n)$ соответственно, называют функцию $\phi(\varepsilon)$: $\mathbb{R} \to \mathbb{R}$, определяемую формулой (9).

$$w_i = \left|d_i - b_i\right|^q,$$
$$\phi(\varepsilon) = \frac{\sum_{i=1}^{n} w_i \cdot \Lambda_i(\varepsilon)}{\sum_{i=1}^{n} w_i}, \tag{9}$$

где $\Lambda_i(\varepsilon)$ – величина, вычисляемая по формуле (8).

Описанный алгоритм реализован в библиотеке giotto-tda классом Silhouette.

Аналогично ландшафту устойчивости, в качестве признакового описания диаграммы применимы статистические характеристики (см. п. 2.7) её силуэта.

2.12. Амплитуды диаграммы устойчивости

Пусть дана диаграмма устойчивости D, содержащая симплексы некоторой фиксированной размерности. Тогда амплитудой D называют число, являющееся мерой отличия D от “пустой” диаграммы Δ, “расстояние” между D и Δ.

При работе с диаграммами, содержащими симплексы нескольких размерностей, следует независимо проанализировать их подмножества, содержащие симплексы только одной размерности, после чего объединить результаты в вектор, p-норма которого (или сам вектор, если это приемлемо) и будет считаться амплитудой исходной диаграммы.

В рамках настоящей работы для получения признакового описания данных производилось вычисление набора амплитуд соответствующих диаграмм устойчивости по различным метрикам с помощью реализации, представленной классом Amplitude библиотеки giotto-tda.

2.12.1. Амплитуда по расстоянию Васерштейна

Пусть D_1 и D_2 – диаграммы устойчивости, содержащие симплексы некоторой фиксированной размерности. Тогда расстояние Васерштейна между ними [27] вычисляется по формуле (10) как нижняя грань p-нормы вектора максимальных отклонений элементов от их образов по всем возможным биективным отображениям γ между D_1 и D_2 со включенными диагоналями.

$$WD(D_1, D_2, p) = min_{\gamma: D_1 \cup \Delta \to D_2 \cup \Delta} \sqrt[p]{\sum_{x \in D_1 \cup \Delta} ||x - \gamma(x)||_\infty^p}, \quad (10)$$

где $\Delta = \{ (s, s) \mid s \in \mathbb{R} \}$ – “пустая” диаграмма устойчивости, содержащая все диагональные точки и только их;
γ – биективное отображение $D_1 \cup \Delta$ в $D_2 \cup \Delta$.

Если диаграмма D_2 пуста, то есть $D_2 = \Delta$, расстояние Васерштейна может быть вычислено по упрощённой формуле (11). Это значение и называется амплитудой диаграммы D_1 по расстоянию Васерштейна.

$$WD(D_1, \triangle, p) = A_w(D_1, p) = ||\frac{d-b}{2}||_p, \quad (11)$$

где d – вектор моментов исчезновения симплексов;

b – вектор моментов появления симплексов.

2.12.2. Амплитуда по расстоянию узкого горлышка

Расстоянием узкого горлышка между диаграммами устойчивости D_1 и D_2 и соответствующей амплитудой называют частный случай расстояния Васерштейна при $p \to \infty$.

2.12.3. Амплитуды по кривым Бетти, ландшафтам и силуэтам устойчивости

Пусть $f\colon U \to \mathbb{R}$ и $g\colon U \to \mathbb{R}$ – непрерывные вещественные функции, где $U \subseteq \mathbb{R}^n$. Таковыми являются, например, кривые Бетти (п. 2.8), ландшафты (п. 2.10) и силуэты (п. 2.11) устойчивости. Тогда L^p-расстоянием между ними называется величина, вычисляемая по формуле (12).

$$L^p(f, g) = \sqrt[p]{\int_U |f(x) - g(x)|^p dx}, \quad (12)$$

Если функции заданы дискретным набором значений в некоторых точках: $f \sim F(F_1, F_2, ..., F_n)$ и $g \sim G(G_1, G_2, ..., G_n)$, то L^p-расстояние между ними может быть оценено векторным расстоянием между F и G по формуле (13).

$$L^p(f, g) = l^p(F, G) = ||F - G||_p \quad (13)$$

Расстояниями под номером p между диаграммами устойчивости D_1 и D_2 по кривым Бетти, ландшафтам и силуэтам устойчивости называют L^p-расстояния между этими

функциями. Соответствующими амплитудами называют расстояния между заданной диаграммой устойчивости и “пустой” диаграммой.

2.13. Признаковое описание диаграмм устойчивости

С использованием алгоритмов, описанных в п. 2.6 – 2.12, был реализован класс FeatureCalculator, извлекающий признаковые описания из набора диаграмм устойчивости. Общая схема его работы представлена на рисунке 3.

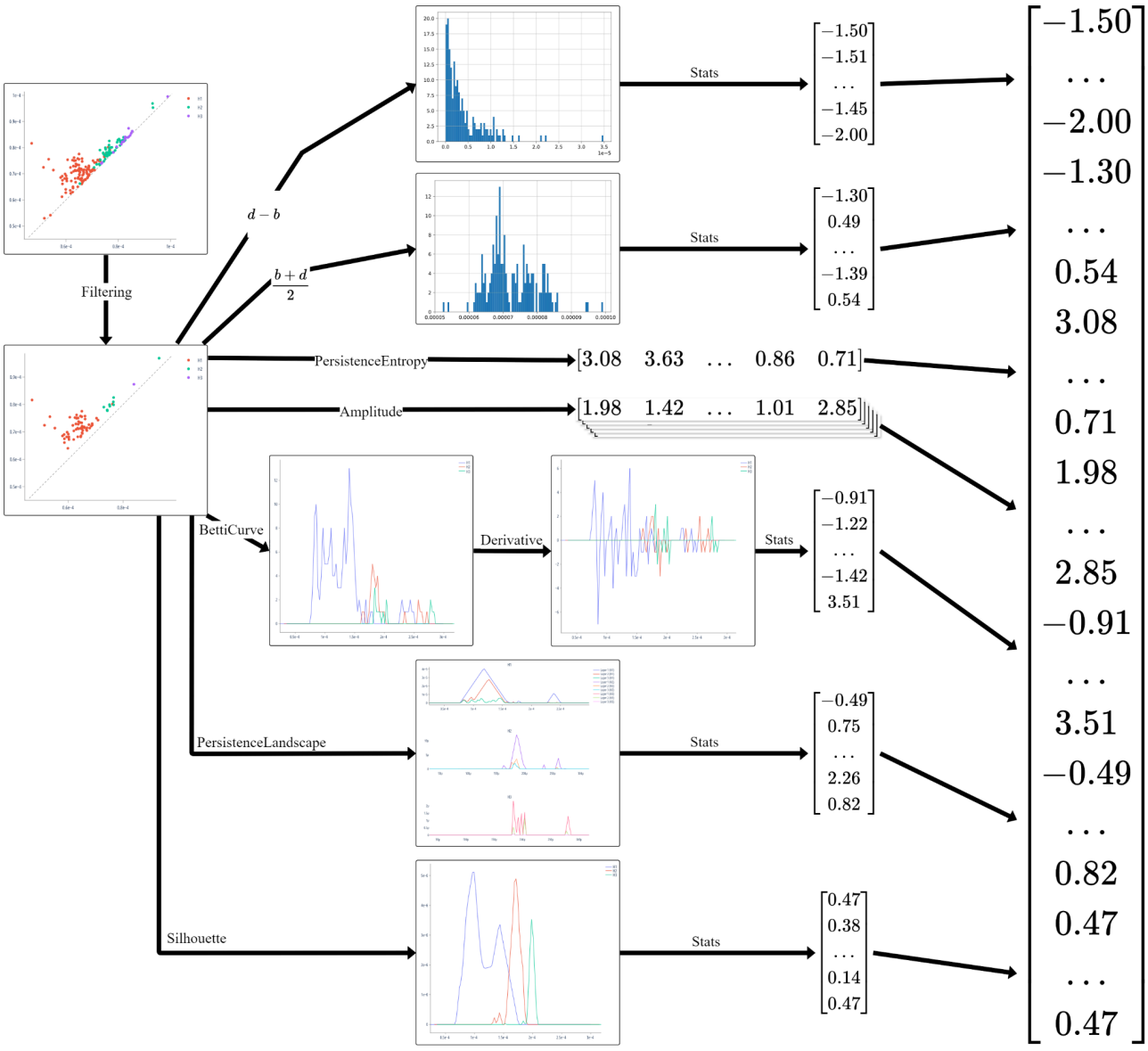

Рисунок 3 – Схема алгоритма векторизации диаграмм устойчивости классом FeatureCalculator

В ходе работы алгоритм производит фильтрацию полученных диаграмм – удаление F процентов наименее долгоживущих симплексов – и их векторизацию – вычисление различных статистических и топологических характеристик. В результате для каждой диаграммы определяются перечисленные далее значения, которые затем объединяются в итоговый вектор и подаются на выход алгоритма:

1. Статистические характеристики (п. 2.7) последовательности времён жизни симплексов для всей диаграммы и по каждой размерности независимо;
2. Статистические характеристики (п. 2.7) последовательности средних значений между моментами появления и исчезновения симплексов для всей диаграммы и по каждой размерности независимо;
3. Энтропия устойчивости (п. 2.9);
4. Набор амплитуд устойчивости (п. 2.12) и их векторные нормы первого и второго порядков:
    a. По расстоянию узкого горлышка;
    b. По расстоянию Васерштейна с $p = 1$ и $p = 2$;
    c. По кривым Бетти с $p = 1$ и $p = 2$;
    d. По первым и вторым слоям ландшафтов устойчивости с $p = 1$ и $p = 2$;
    e. По силуэтам устойчивости степеней 1 и 2 с $p = 1$ и $p = 2$;
5. Статистические характеристики (п. 2.7) первых производных кривых Бетти (п. 2.8);
6. Статистические характеристики (п. 2.7) первых уровней ландшафтов устойчивости (п. 2.10).
7. Статистические характеристики (п. 2.7) силуэтов устойчивости (п. 2.11) степеней 1 и 2.

2.14. Стандартизация признаков

Стандартизация – один из традиционных подходов, позволяющих привести известные признаки к единому масштабу, что способствует повышению качества алгоритмов машинного

обучения и позволяет избежать подавления одних признаков другими исключительно из-за величины их значений.

Пусть для каждого из n объектов вычислено по m признаков, представленных в виде матрицы A размера $n \times m$. Процессом стандартизации этих признаков называется приведение каждого столбца матрицы A к распределению с нулевым средним и единичным стандартным отклонением по формуле (14).

$$\mu = \frac{1}{n}\sum_{i=1}^{n} A_{ij}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (A_{ij} - \mu)^2} \tag{14}$$

$$A_j = \frac{A_j - \mu}{\sigma},$$

где A_{ij} – элемент матрицы A, записанный на пересечении строки под номером i и столбца под номером j;

μ – выборочное среднее столбца A_j;

σ – выборочное стандартное отклонение столбца A_j.

Описанный алгоритм реализован в библиотеке scikit-learn классом StandardScaler.

2.15. Метод главных компонент

Метод главных компонент [28] – один из способов понижения размерности данных с потерей как можно меньшего количества информации, основанный на линейном отображении пространства высокой размерности в пространство желаемой, меньшей размерности. Алгоритм заключается в нахождении проекции исходного признакового пространства на некоторое подпространство, заданное ортонормированным базисом, построенным так, чтобы каждый следующий вектор “объяснял” как можно большую часть дисперсии исходных данных, которая не объяснена уже найденными векторами.

Доказано (см. [29]), что таким базисом является набор собственных векторов ковариационной матрицы исходных признаков, а объясняемые ими доли дисперсии пропорциональны величинам соответствующих собственных значений. Вследствие этого, для нахождения главных компонент удобно применять сингулярное разложение исходной матрицы признаков.

Пусть A – исходная матрица размера $n \times m$, содержащая по m признаков для каждого из n объектов. Её сингулярным разложением называется представление в виде произведения трёх матриц: унитарной V размера $n \times n$, диагональной Σ размера $n \times m$ и ортогональной U^T размера $m \times m$. Можно проверить, что разложение верно, если Σ состоит из сингулярных чисел матрицы A (то есть из корней собственных чисел ковариационной матрицы $A^T A$), отсортированных по убыванию, U – из соответствующих собственных векторов матрицы $A^T A$, а столбцы V вычислены по формуле (15).

$$v_i = \frac{Au_i}{\sigma_i} \tag{15}$$

где v_i и u_i – i-е столбцы матриц V и U соответственно;

σ_i – i-е сингулярное значение матрицы A.

Легко заметить, что столбцы матрицы U составляют необходимый ортонормированный базис, проекция A' строк матрицы A на который может быть вычислена по формуле (16).

$$A' = AU = V\Sigma U^T U = V\Sigma \tag{16}$$

где V, Σ, U – соответствующие матрицы сингулярного разложения матрицы A.

При необходимости, “последние” столбцы Σ, содержащие наименьшие сингулярные значения, могут быть удалены или вовсе не вычислены, что приведёт к искомому понижению размерности признакового пространства с минимальной потерей информации. В рамках

настоящей работы вычислялись первые 65 главных компонент, объясняющие достаточную долю дисперсии без значительного влияния на время работы методов кластеризации.

Описанный алгоритм реализован классом PCA библиотеки scikit-learn.

2.16. Стратегии извлечения топологических признаков

В рамках настоящей работы извлечение топологических признаков производилось независимо тремя способами, показанными на рисунке 4, после чего результаты объединялись в итоговый вектор, производилась стандартизация полученных признаков и выделение главных компонент для понижения размерности, отбора и удаления низкоинформативных и коррелированных переменных. Таким образом, в результате каждая эпоха описывалась вектором относительно небольшой размерности.

Важно отметить, что, в отличие от исходной работы [1], данные, полученные на ушных электродах A1 и A2 (см. рис. 1) не были удалены из анализа, а также не производилось разбиение записей по частотным диапазонам и усреднение полученных признаков по частям мозга, где были расположены соответствующие электроды во время эксперимента. Первичная экспериментальная оценка показала, что эти операции не оказывают кардинального влияния на результат, но их выполнение могло бы усложнить анализ качества извлечённых признаков, что не соответствует цели работы – оценка применимости топологических признаков к задаче нахождения функциональных состояний по ЭЭГ, а не достижение высокого качества предсказаний.

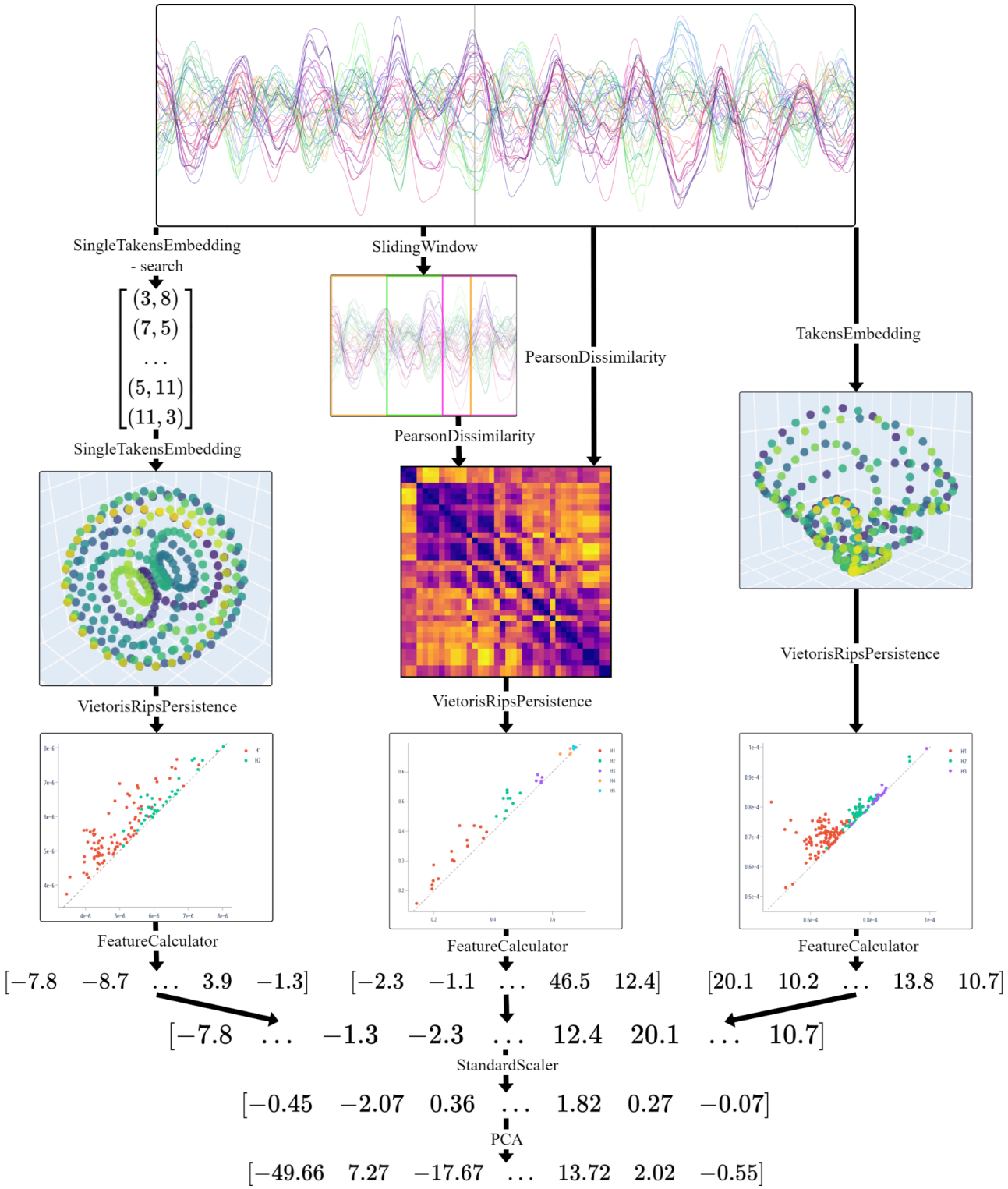

Рисунок 4 – Схема алгоритма извлечения топологических признаков

2.16.1. Анализ каждой переменной независимо

В первую очередь компоненты временного ряда обрабатывались независимо для анализа записей, полученных на каждом электроде, отдельно от других параметров системы следующим образом:

1. Эпохи ЭЭГ – многомерные временные ряды – разделялись на наборы одномерных временных рядов, каждый из которых соответствовал одному электроду во время эксперимента.

2. Для каждого полученного временного ряда производился поиск оптимальных гиперпараметров алгоритма извлечения эмбеддингов Такенса методом, описанным в п. 2.2.3. Таким образом, для каждой компоненты ЭЭГ был получен набор оптимальных для каждой эпохи гиперпараметров.

3. Для каждой компоненты выбиралось одно значение гиперпараметров путём “голосования большинства” – выбирались значения, которые оказались оптимальными для большинства эпох.

4. Каждая компонента каждой эпохи преобразовывалась в метрическое пространство с помощью алгоритма Такенса с полученными на шаге 3 гиперпараметрами (п. 2.2.1).

5. Для полученных метрических пространств производилось построение последовательностей комплексов Вьеториса – Рипса и диаграмм устойчивости (п. 2.5).

6. Каждая диаграмма устойчивости векторизовывалась классом FeatureCalculator, описанным в п. 2.13. Таким образом, для каждой компоненты каждой эпохи был получен один вектор признаков.

7. Для получения итогового признакового описания эпохи значения, соответствующие компонентам этой эпохи, объединялись в единый вектор.

Ожидается, что такой подход достаточно хорошо подходит для анализа отдельных компонент ЭЭГ и позволяет обнаружить большую часть значимых закономерностей. Тем не менее он не учитывает корреляции между компонентами временного ряда, которые могут содержать важную для решения поставленной задачи информацию.

2.16.2. Анализ корреляций между переменными

Другая стратегия извлечения топологических признаков основана на анализе информации, заключенной во взаимосвязях между компонентами ЭЭГ, и производилась по следующему алгоритму:

1. Вычисление Пирсоновского несходства (п. 2.4) для каждой эпохи, а также для её скользящих окон некоторого размера с фиксированным шагом. В результате получен набор матриц, описывающих метрические пространства, содержащие $nChannels = 40$ точек, с расстояниями, отражающими корреляции между соответствующими компонентами временных рядов.

2. Построение последовательностей комплексов Вьеториса – Рипса и диаграмм устойчивости (п. 2.5) для полученных метрических пространств.

3. Векторизация построенных диаграмм устойчивости с помощью класса FeatureCalculator (п. 2.13), в результате чего для каждой эпохи и всех её окон вычисляются векторы признаков.

4. Получение итогового признакового описания путём объединения векторов, соответствующих каждой эпохе.

2.16.3. Анализ ЭЭГ в целом

В заключение, производился топологических анализ временных рядов эпох без явного выделения анализируемых характеристик:

1. Эпохи преобразовывались в метрические пространства алгоритмом получения эмбеддингов Такенса в соответствии с п. 2.2.2 с подобранными экспериментально одинаковыми для всех компонент гиперпараметрами.

2. Для полученных метрических пространств производилось построение последовательностей комплексов Вьеториса – Рипса и диаграмм устойчивости в соответствии с алгоритмом, описанным п. 2.5.

3. Каждая диаграмма устойчивости векторизовывалась классом FeatureCalculator, описанным в п. 2.13. Таким образом, для каждой эпохи был получен один вектор признаков.

3. АЛГОРИТМ НАХОЖДЕНИЯ ФУНКЦИОНАЛЬНЫХ СОСТОЯНИЙ

3.1. Агломеративный метод иерархической кластеризации Уорда

Иерархическая кластеризация – класс алгоритмов кластеризации, основанный на создании дерева вложенных кластеров и примечательный возможностью задания матрицы связности, что необходимо при работе с данными ЭЭГ в связи с их непрерывной во времени структурой. Выделяют два типа таких методов: агломеративные, основанные на построении новых кластеров путем объединения существующих, и дивизионные, решающие задачу путем разделения существующих групп на более маленькие.

Экспериментально выявлено [30], что наилучшее качество для данных ЭЭГ показывает метод Уорда [31] – один из агломеративных алгоритмов иерархической кластеризации, принимающий решения о целесообразности объединения кластеров на основе расстояния Уорда с целью минимизации внутрикластерной дисперсии. Изначально каждый элемент входных данных образует отдельный кластер, после чего пары кластеров с минимальным расстоянием Уорда между ними объединяются в один, пока не будет получено необходимое разбиение или минимальное расстояние между парами кластеров не окажется достаточно велико.

Расстояние Уорда вычисляется по формуле (17) как прирост суммарного квадратичного отклонения элементов кластеров от их центров в результате объединения.

$$D(X,\ Y) = \sum_{z \in X \cup Y} ||z - \overline{X \cup Y}||^2 - \left(\sum_{x \in X} ||x - \overline{X}||^2 + \sum_{y \in Y} ||y - \overline{Y}||^2 \right), \quad (17)$$

где $X \cup Y$ – кластер, получаемый в результате объединения X и Y

$\overline{X}$, $\overline{Y}$, $\overline{X \cup Y}$ – центры множеств X, Y и $X \cup Y$ соответственно, вычисляемые по формуле (18).

$$\overline{X} = \frac{1}{|X|} \cdot \sum_{x \in X} x, \quad (18)$$

В рамках работы была использована реализация на языке Python, представленная классом AgglomerativeClustering библиотеки scikit-learn.

3.2. Метод k-средних

Метод k-средних [32] – один из классических алгоритмов кластеризации, стремящийся разделить данные на k групп так, чтобы суммарное квадратичное отклонение элементов от центров соответствующих кластеров было минимальным. В простейшей реализации k центров кластеров устанавливаются случайным образом, и каждое наблюдение относится в ближайший из них, после чего центры кластеров пересчитываются по формуле (18) с учетом полученного разбиения, и процесс повторяется до тех пор, пока происходит уменьшение внутрикластерных расстояний.

Хотя определение разбиения, соответствующего глобальному минимуму метрики, не гарантируется, алгоритм находит широкое применение в различных задачах анализа данных и его реализация, представленная классом KMeans библиотеки scikit-learn языка Python, была использована в настоящей работе.

3.3. State-Detecting Algorithm

Для нахождения функциональных состояний по признаковому описанию ЭЭГ был изучен и улучшен алгоритм SDA [1], комбинирующий два метода кластеризации для решения поставленной задачи с учётом непрерывности изучаемого процесса во времени.

На вход алгоритму подаётся признаковое описание ЭЭГ – матрица, сопоставляющая каждой эпохе вектор признаков некоторой размерности. Опционально перед запуском основной части алгоритма признаки могут быть дополнительно стандартизованы, что может положительно влиять на качество результата.

На первом этапе алгоритма производится нахождение потенциальных границ функциональных состояний с помощью агломеративного метода иерархической кластеризации Уорда (п. 3.1) с различными значениями количества искомых состояний ($n\_clusters$) и максимального времени между эпохами в матрице связности ($k\_neighbours$). Далее полученные границы сортируются и соседние состояния объединяются, если они слишком коротки (содержат не более len_min эпох) или недостаточно отличаются (расстояние Уорда между ними не превосходит $dist\_rate$ среднего значения для всех пар соседних

состояний). В результате для каждой тройки значений ($n\_clusters$, $k\_neighbours$, len_min) получается набор потенциальных границ функциональных состояний с достаточно большими различиями между ними.

На втором этапе алгоритма производится выбор кандидатов итоговых границ функциональных состояний на основе результатов первого этапа с помощью метода k-средних (п. 3.2). Для этого алгоритм кластеризации применяется к объединению наборов потенциальных границ, полученных на прошлом этапе при $n\_clusters \leq n\_cl$, $k\_neighbours \leq k\_nb\_max$ и фиксированном len_min, для выделения $n\_edge\_clusters$ состояний. Итоговые границы функциональных состояний вычисляются как средние, медианы и моды полученных кластеров. Таким образом, на втором этапе алгоритма для каждого набора значений (n_cl, k_nb_max, st_len, $n\_edge\_clusters$) определяется 3 варианта итогового результата.

Дальнейший выбор ответа производится экспертным решением с учётом метрик качества кластеризации, описанных далее.

Более подробное описание работы алгоритма и обоснование его корректности выходит за рамки настоящего исследования и представлено в [1].

4. МЕТРИКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ

В рамках исследования были использованы три способа оценки качества результата: внутренняя оценка, показывающая, насколько сильно полученные кластеры отличаются друг от друга, внешняя оценка, позволяющая сравнить полученные результаты между собой, а также анализ информационной ценности признаков, с помощью которого были оценены корреляции признаков как с правильными, так и с полученными наборами границ.

4.1. Внутренняя оценка

Так как для задачи нахождения функциональных состояния по ЭЭГ отсутствует “правильный ответ” (в ином случае исследование не имело бы смысла), в качестве основных способов оценки качества результата были использованы методы внутренней оценки, анализирующие сходства и различия полученных кластеров. Эти метрики показывают, насколько похожи объекты, которые были отнесены к одной группе, и насколько отличаются объекты, определенные в разные группы. Более того, вычисление метрик производилось не для всего набора данных, а отдельно для каждой пары смежных по времени кластеров с последующим усреднением значений, что лучше отражает непрерывную во времени структуру ЭЭГ (разные кластеры могут оказаться похожими, но их объединение невозможно, так как они не являются соседними во времени). В рамках исследования для консистентности результатов применялись те же метрики, что и в исходной статье [1].

4.1.1. Расстояние Уорда

Расстояние Уорда вычисляется по формуле (17) и показывает, насколько сильно увеличится сумма внутрикластерных расстояний при объединении двух кластеров. Лучшим результатам кластеризации соответствуют большие значения расстояния Уорда.

4.1.2. Центроидное расстояние

Центроидным расстоянием называется расстояние между центрами двух кластеров и вычисляется по формуле (19). Большое значение метрики соответствует удаленным кластерами и высокому качеству кластеризации. В настоящем исследовании при вычислении центроидного расстояния использовалась евклидова норма с $p = 2$.

$$d_c(X,\ Y) = ||\overline{X} - \overline{Y}||_p, \tag{19}$$

где $\overline{X}$, $\overline{Y}$ – центры множеств X и Y соответственно, вычисляемые по формуле (18).

4.1.3. Коэффициент силуэта

Коэффициент силуэта [32] вычисляется для одного объекта x по формуле (20) и показывает, насколько этот объект похож на другие объекты своего кластера (C_x) в сравнении с объектами других кластеров (C_y). Для получения единственного числа значения усредняются по всем исследуемым объектам. Возможные значения метрики лежат в интервале $[-1,\ 1]$, где положительные значения указывают на то, что объект отнесен к верному кластеру, который хорошо отделен от других кластеров.

$$\begin{gathered} a(x,\ C_x) = \frac{1}{|C_x| - 1} \cdot \sum_{y \in C_x} ||x - y||, \\ b(i,\ C_i) = min_{C_y \neq C_x} \frac{1}{|C_y|} \cdot \sum_{y \in C_y} ||x - y||, \\ s(x,\ C_x) = \frac{b - a}{max(a, b)}, \end{gathered} \tag{20}$$

Описанный алгоритм реализован функцией silhouette_score библиотеки scikit-learn.

4.1.4. Индекс Калински – Харабаса

Индекс Калински – Харабаса [34] вычисляется по формуле (21) как отношение сумм межкластерных отклонений (B) к суммам внутрикластерных отклонений (W), нормированных на количество их степеней свободы. Большие значения индекса Калински – Харабаса показывают, что кластеры разделены хорошо и расположены далеко друг от друга.

$$B = \sum_{i=1}^{k} |C_i| \cdot ||\overline{C_i} - \overline{C}||^2,$$

$$W = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \overline{C_i}||^2, \quad (21)$$

$$CH = \frac{B}{W} \cdot \frac{n-k}{k-1},$$

где n – количество объектов;

k – количество выделенных кластеров;

$\overline{C_i}$ – центр кластера C_i, вычисляемый по формуле (18);

$\overline{C}$ – центр множества всех точек, вычисляемый по формуле (18).

Описанный алгоритм реализован функцией calinski_harabasz_score библиотеки scikit-learn.

4.1.5. Индекс Дэвиса – Болдина

Индекс Дэвиса – Болдина [35] вычисляется по формуле (22) как среднее “сходство” – отношение между размерами кластеров и расстояниями между ними – пар наиболее близких кластеров. Таким образом, наилучшему качеству кластеризации соответствуют близкие к нулю значения метрики.

$$DB = \frac{1}{k} \cdot \sum_{i=1}^{k} max_{i \neq j} \frac{s_i + s_j}{d_{ij}}, \quad (22)$$

где k – количество кластеров;

d_{ij} – центроидное расстояние между кластерами i и j, вычисляемое по формуле (19);

s_i и s_j – диаметры кластеров i и j соответственно, вычисляемые по формуле (23).

$$s_i = \frac{1}{|C_i|} \cdot \sum_{x \in C_i} ||x - \overline{C_i}||, \quad (23)$$

где C_i – кластер под номером i;

$\overline{C_i}$ – центр множества C_i, вычисляемый по формуле (18).

Описанный алгоритм реализован функцией davies_bouldin_score библиотеки scikit-learn.

4.2. Внешняя оценка

При анализе применимости топологических признаков для нахождения функциональных состояний по ЭЭГ были также использованы и внешние методы оценки – метрики, основанные на сравнении результатов с некоторыми “правильными ответами”, в качестве которых были взяты результаты, полученные в [1] с использованием традиционных признаков.

4.2.1. Коэффициент взаимной информации

Пусть U и V – два распределения N точек по кластерам. Тогда коэффициент взаимной информации [36] вычисляется по формуле (24) и является мерой их согласованности – насколько одно распределение “зависит” от другого. Значения, близкие к единице, указывают на высокое совпадение распределений.

$$\begin{gathered} P_U(i) = \frac{|U_i|}{N}, \\ P_V(j) = \frac{|V_j|}{N}, \\ P(i, j) = \frac{|U_i \cap V_j|}{N}, \\ MI(U, V) = \sum_{i=1}^{C_U} \sum_{j=1}^{C_V} P(i, j) \cdot ln\left(\frac{P(i,j)}{P_U(i) \cdot P_V(j)}\right), \end{gathered} \tag{24}$$

где U_i – множество элементов, распределённых в кластер i в U;

V_j – множество элементов, распределённых в кластер j в V;

C_U и C_V – количества кластеров в распределениях U и V соответственно.

Тем не менее коэффициент взаимной информации как правило увеличивается с ростом количества кластеров, что не всегда коррелирует с повышением качества результата. Во избежание этого на практике применяют так называемые нормализованный и скорректированный коэффициенты взаимной информации [37]. В настоящей работе использовался скорректированный коэффициент, вычисляемый по формуле (25).

$$norm(U,\ V) = \ -\frac{1}{2}\left(\sum_{i=1}^{C_U} P_U(i)\ \cdot\ ln(P_U(i))\ +\ \sum_{i=1}^{C_V} P_V(i)\ \cdot\ ln(P_V(i))\right),$$
$$AMI(U,\ V) = \frac{MI(U,V) - E[MI(U,V)]}{norm(U,V) - E[MI(U,V)]}, \tag{25}$$

где C_U и C_V – количества кластеров в распределениях U и V соответственно;

$P_U(i)$, $P_V(j)$, $MI(U,\ V)$ – величины, вычисляемые по формуле (24);

$E[MI(U,\ V)]$ – математическое ожидание коэффициента взаимной информации.

Описанный алгоритм реализован функцией adjusted_mutual_info_score библиотеки scikit-learn.

4.2.2. Индекс Рэнда

Пусть U и V – распределения N точек по кластерам. Тогда индекс Рэнда [38] является мерой их подобия и вычисляется по формуле (26) как доля пар точек, отнесённых к одному или разным кластерам в обоих распределениях. Таким образом, значения, близкие к единице, указывают на высокое сходство результатов.

$$RI\ = \frac{SS + DS}{C_2^N}, \tag{26}$$

где SS – количество пар точек, определённых в один кластер и в U, и в V;

DS – количество пар точек, определённых в разные кластеры и в U, и в V;

C_2^N – общее количество различных пар точек в наборе данных.

Тем не менее индекс Рэнда, как и коэффициент взаимной информации, принимает хорошие значения для распределений с большим количеством кластеров, что не всегда соответствует высокому качеству. Для противодействия этому на практике применяют скорректированный индекс Рэнда [39], лежащий в интервале $[-0.5, 1]$ и вычисляемый по формуле (27).

$$ARI = \frac{RI - E[RI]}{1 - E[RI]}, \tag{27}$$

где RI – нескорректированный индекс Рэнда, вычисляемый по формуле (26);
$E[RI]$ – математическое ожидание индекса Рэнда.

Описанный алгоритм реализован функцией adjusted_rand_score библиотеки scikit-learn.

4.2.3. Индекс Фаулкса-Маллоуса

Индекс Фаулкса-Маллоуса [40] вычисляется по формуле (28) как среднее геометрическое попарных точности и полноты и примечателен тем, что, в отличие от коэффициента взаимной информации и индекса Рэнда, с увеличением числа кластеров его величина стремится к нулю. Таким образом, значения, близкие к единице, позволяют более уверенно утверждать о высокой схожести результатов.

$$FMI = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}, \tag{28}$$

где TP – количество пар точек, определённых в один кластер в обоих распределениях;
FP – количество пар точек, определённых в один кластер в “правильном” распределении, но в разные кластеры в анализируемом;
FN – количество пар точек, определённых в один кластер в анализируемом распределении, но в разные кластеры в “правильном”.

Описанный алгоритм реализован функцией fowlkes_mallows_score библиотеки scikit-learn.

4.3. Анализ информационной ценности признаков

Анализ информационной ценности [41] – один из типичных подходов к отбору признаков в задачах бинарной классификации, основанный на оценке влияния отдельных переменных на принадлежность объекта к положительному или отрицательному классам.

Пусть дано n объектов, для которых известны $x = (x_1, x_2, ..., x_n)$ – вектор категориальных признаков со значениями из множества $\{X_1, X_2, ..., X_k\}$ – и $y = (y_1, y_2, ..., y_n)$ – вектор значений целевой переменной (0 или 1). Назовём "весом доказательств" (WoE_i) значения X_i величину, вычисляемую по формуле (29) и показывающую соотношение долей положительных и отрицательных объектов при этом значении переменной.

$$\begin{gathered} percentEvents_i = numEvents_i \,/\, numEvents \\ percentNonEvents_i = numNonEvents_i \,/\, numNonEvents \\ WoE_i = \frac{percentEvents_i}{percentNonEvents_i}, \end{gathered} \tag{29}$$

где $numEvents_i$ и $numNonEvents_i$ – количества объектов, относящихся к положительному и отрицательному классам соответственно при значении признака X_i;

$numEvents$ и $numNonEvents$ – общие количества объектов, относящихся к положительному и отрицательному классам соответственно.

Тогда информационной ценностью (IV) признака x называют величину, вычисляемую по формуле (30) и показывающую, насколько распределение этого признака совпадает с распределением целевой переменной, т.е. насколько этот признак "важен" для правильной классификации объектов.

$$IV = \sum_{i=1}^{k} (percentEvents_i - percentNonEvents_i) \cdot WoE_i, \quad (30)$$

где $percentEvents_i$, $percentNonEvents_i$ и WoE_i – величины, вычисляемые по формуле (29).

При работе с количественными (не категориальными) признаками, производят преобразование значений к категориальным путём разбиения множества значений на заданное количество контейнеров равной длины. В рамках настоящей работы были принято решение использовать 10 контейнеров.

При решении задачи многоклассовой классификации информационную ценность вычисляют независимо для каждого класса – относительно целевой переменной вида "объект принадлежит к классу под номером i" – с последующим усреднением значений. А задача кластеризации, решаемая в настоящей работе, может быть сведена к многоклассовой классификации путём рассмотрения целевой переменной вида "объект принадлежит к кластеру под номером i".

Большие значения информационной ценности соответствуют высокой обобщающей способности признаков и их значительному влиянию на верное предсказание целевой переменной, в то время как значения, близкие к нулю, указывают на бесполезность признака для предсказаний.

5. РЕЗУЛЬТАТЫ

5.1. Выбор гиперпараметров

Подбор гиперпараметров производился преимущественно экспериментально с применением алгоритмов, описанных в п. 2, а также значений, использовавшихся в исходной работе [1]. Итоговые выбранные значения приведены в таблице 1.

Таблица 1 – Значения гиперпараметров

Стадия	Этап	Обозначение	Значение(я)	Описание
Извлечение топологических признаков – анализ каждой переменной независимо	Получение эмбеддингов Такенса	d	$[\,1;\ 20\,]$	Рассматриваемые размерности пространств
		τ	$[\,1;\ 20\,]$	Рассматриваемые значения временной задержки между координатами точек
		s	3	Шаг между первыми координатами точек
	Построение диаграммы устойчивости	n	$\{\,1,\ 2\,\}$	Размерности анализируемых симплексов
Извлечение топологических признаков – анализ корреляций между переменными	Скользящее окно	$size$	50	Размер скользящих окон
		$stride$	20	Шаг вычисления скользящих окон
	Построение диаграммы устойчивости	n	$[\,1;\ 5\,]$	Размерности анализируемых симплексов
Извлечение топологических признаков – анализ ЭЭГ в целом	Получение эмбеддингов Такенса	d	5	Размерность пространства точек
		τ	11	Временная задержка между координатами точек
		s	3	Шаг между первыми координатами точек
	Построение диаграммы устойчивости	n	$\{\,1,\ 2,\ 3\,\}$	Размерности анализируемых симплексов

Продолжение таблицы 1

Признаковое описание диаграмм устойчивости	Фильтрация	F	10%	Доля симплексов, удаляемых в результате фильтрации
	Извлечение признаков	n_bins	100	Количество рассматриваемых значений ε при дискретном вычислении топологических характеристик
Метод главных компонент		n_comp	65	Количество выделяемых главных компонент
SDA: этап 1		$n\_clusters$	[2, 20]	Количество искомых состояний
		$k\_neighbours$	[20, 50]	Максимальное время между эпохами в матрице связности
		len_min	{0, 20, 40, 60}	Минимальная длина найденных функциональных состояний
		$dist\_rate$	0.3	Коэффициент расстояния Уорда при объединении состояний
SDA: этап 2		n_cl	{ 10, 15, 20 }	Наибольшее значение $n\_clusters$ при объединении наборов границ
		k_nb_max	{35, 40, 45, 50}	Наиб. значение $k\_neighbours$ при объединении наборов границ
		$n\_edge\_clusters$	[2, 15]	Итоговое количество искомых состояний
Анализ информационной ценности		n_bins	10	Количество контейнеров при работе с колич. признаками

5.2. Получение результатов

Для получения результатов описанные алгоритмы применялись последовательно к записям ЭЭГ каждого объекта с сохранением промежуточных результатов на диск для дальнейшего анализа. Таким образом, для каждого объекта, помимо прочего, были получены следующие элементы:

1. Признаковое пространство размерности 8400, описывающее компоненты ЭЭГ независимо (п. 2.16.1);

2. Признаковое пространство размерности 10872, описывающее корреляции между компонентами ЭЭГ (п. 2.16.2);

3. Признаковое пространство размерности 291, описывающее ЭЭГ в целом (п. 2.16.3);

4. Итоговое признаковое пространство размерности 19563;

5. Признаковое описание размерности 65, полученное путём применения метода главных компонент к итоговому признаковому пространству (п. 4), а также доли дисперсии исходного признакового пространства, объясняемые каждой компонентой;

6. Два результата итоговых границ функциональных состояний: лучший по коэффициенту силуэта (внутренняя метрика, п. 4.1.3) и лучший по индексу Фаулка – Маллоуса (внешняя метрика, п. 4.2.3), для каждого из которых определены следующие величины:

   a. Значения всех метрик качества кластеризации (п. 4.1, п. 4.2);

   b. Информационные ценности (п. 4.3) каждого признака относительно полученного и “верного” ответов, а также гистограмма различий между ними;

   c. Изображение ЭЭГ с визуально выделенными границами функциональных состояний;

   d. Изображение нормализованных метрик качества кластеризации для всех найденных границ, а также диаграмма рассеяния потенциальных границ функциональных состояний, вычисленных на первом этапе алгоритма и использованных для получения анализируемого результата на втором.

В заключение для полноты анализа признаки, оказавшиеся наиболее информативными относительно лучшего ответа по коэффициенту силуэта, были объединены с традиционными признаками в соотношении 1:1 (по 765 признаков), после чего производилась их стандартизация, выделение 15 главных компонент, что обеспечивает объяснение около 70% дисперсии, и применение алгоритма SDA с вычислением всех метрик качества.

Исходный код использованных алгоритмов и результаты работы опубликованы в репозитории GitHub, доступном по адресу: https://github.com/TTPO100AJIEX/TopoEEG.

5.3. Объект – 1

Для первого объекта производился поиск 9 функциональных состояний. Суммарная доля дисперсии исходных данных, объяснённая главными компонентами, составила 0,5.

При использовании топологических признаков алгоритм определил границы $\{ 0, 52, 252, 280, 554, 682, 801, 842, 976, 1046 \}$ при выборе ответа по внутр. метрике и $\{ 0, 52, 210, 280, 486, 554, 682, 842, 976, 1046 \}$ – по внешней метрике против границ $\{ 0, 39, 282, 492, 560, 682, 784, 857, 976, 1046 \}$, полученных на традиционных признаках и $\{ 0, 186, 282, 490, 558, 682, 788, 857, 977, 1046 \}$ при использовании признаковых описаний вместе. Их сравнение представлено в таблице 2 и на рисунке 5.

Таблица 2 – Метрики качества результатов для объекта 1

	Ответ на традиционных признаках	По коэффициенту силуэта	По индексу Фаулкса – Маллоуса	Ответ при объединении признаков
Расстояние Уорда	24208	149649	145861	42752
Центроидное расстояние	21.38	52.64	48.35	28.08
Коэффициент силуэта	0.199	0.070	0.055	0.234
Индекс Калински – Харабаса	69.90	19.67	19.16	85.13
Индекс Дэвиса – Болдина	1.637	3.348	3.594	1.582
Коэфф. взаимной информации		0.869	0.878	0.920

Продолжение таблицы 2

Индекс Рэнда		0.791	0.795	0.832
Индекс Фаулкса-Маллоуса		0.823	0.824	0.857

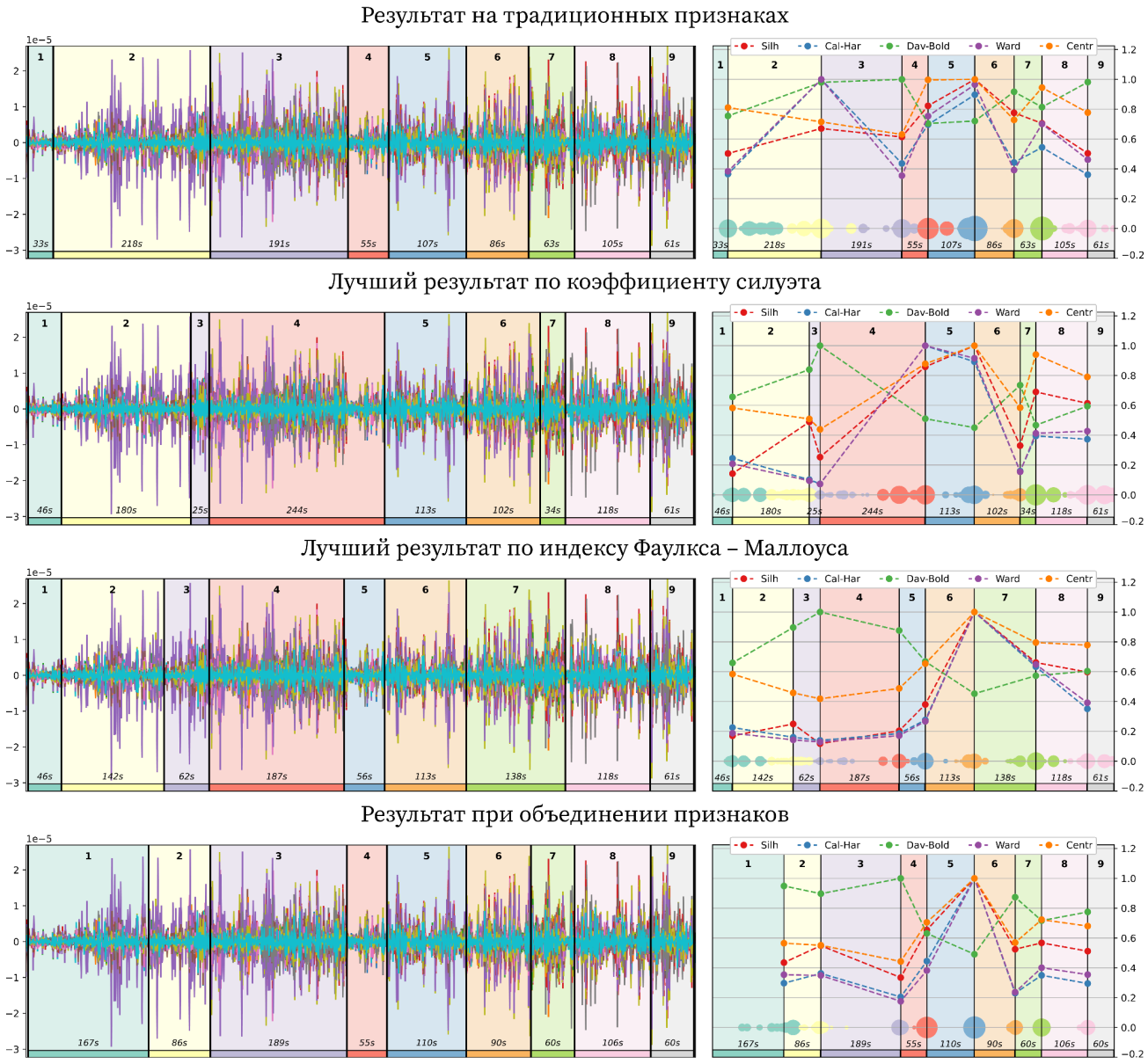

Рисунок 5 – Изображения ЭЭГ, диаграмм рассеяния потенциальных границ и нормированных метрик качества кластеризации всех найденных границ для объекта 1

Анализ информационной ценности показал, что топологические признаки имеют общие закономерности как с полученными, так и с “правильным” результатом кластеризации:

“лучшие” признаки имеют информационную ценность около 2,5 во всех случаях. Тем не менее многие признаки (более 13000) имеют и достаточно низкую ценность, не превышающую 0,25. Важно отметить, что отличия информационной ценности относительно различных ответов, как изображено на рисунке 6, невелики.

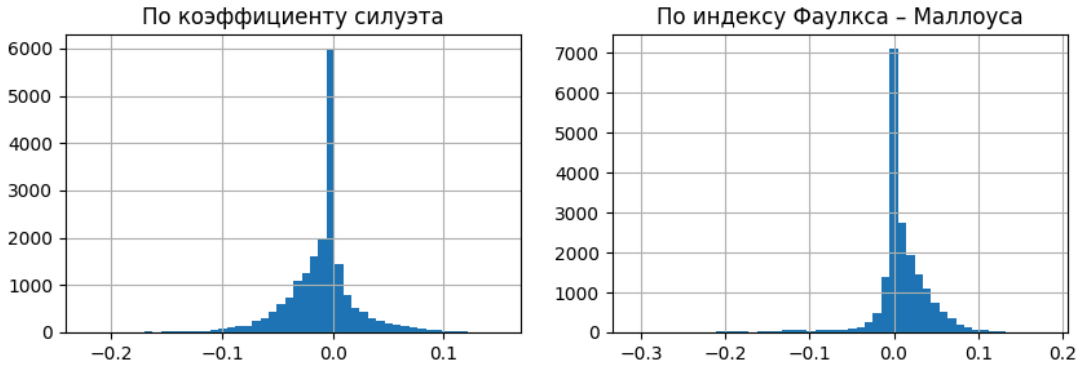

Рисунок 6 – Гистограммы разностей между информационной ценностью признаков относительно полученных ответов и относительно “правильного” ответа для объекта 1

Как показано на рисунке 7, наиболее информативными оказались амплитуды по кривыми Бетти с $p = 1$ и по расстоянию Васерштейна с $p = 1$ и $p = 2$, а также статистические характеристики последовательностей средних значений между моментами появления и исчезновения симплексов. При этом самые ценные признаки были получены при обработке 6, 7, 11 и 12 компонент временных рядов, соответствующих электродам F4, F8, FC4 и FT8 в правой лобной части мозга (рис. 1). По размерности анализируемых симплексов самыми ценными оказались признаки, описывающие все размерности сразу. Это говорит о том, что симплексы высших размерностей лишь дополняют признаковое пространство, а не дублируют уже найденные закономерности, хотя наибольший вклад и вносят симплексы размерностей 1 и 2. Наиболее важными статистическими характеристиками оказались векторные нормы с $p = 1$ и $p = 2$.

Наименее информативными, в свою очередь, оказались признаки, полученные путём анализа корреляций между переменными, а также путём векторизации диаграмм устойчивости, соответствующих ушному электроду A1 (компонент 23), электроду CP3 (комп. 19) в левой центральной части мозга, а также электродам T5 (24) и T6 (28) в левой и правой височных частях мозга соответственно. При этом наименьшая величина информационной ценности наблюдается у статистических характеристик кривых Бетти, а также ландшафтов и

силуэтов устойчивости. Интересно, что наименее важными оказались все перцентили (25, 50 и 75), а также эксцесс и асимметрия.

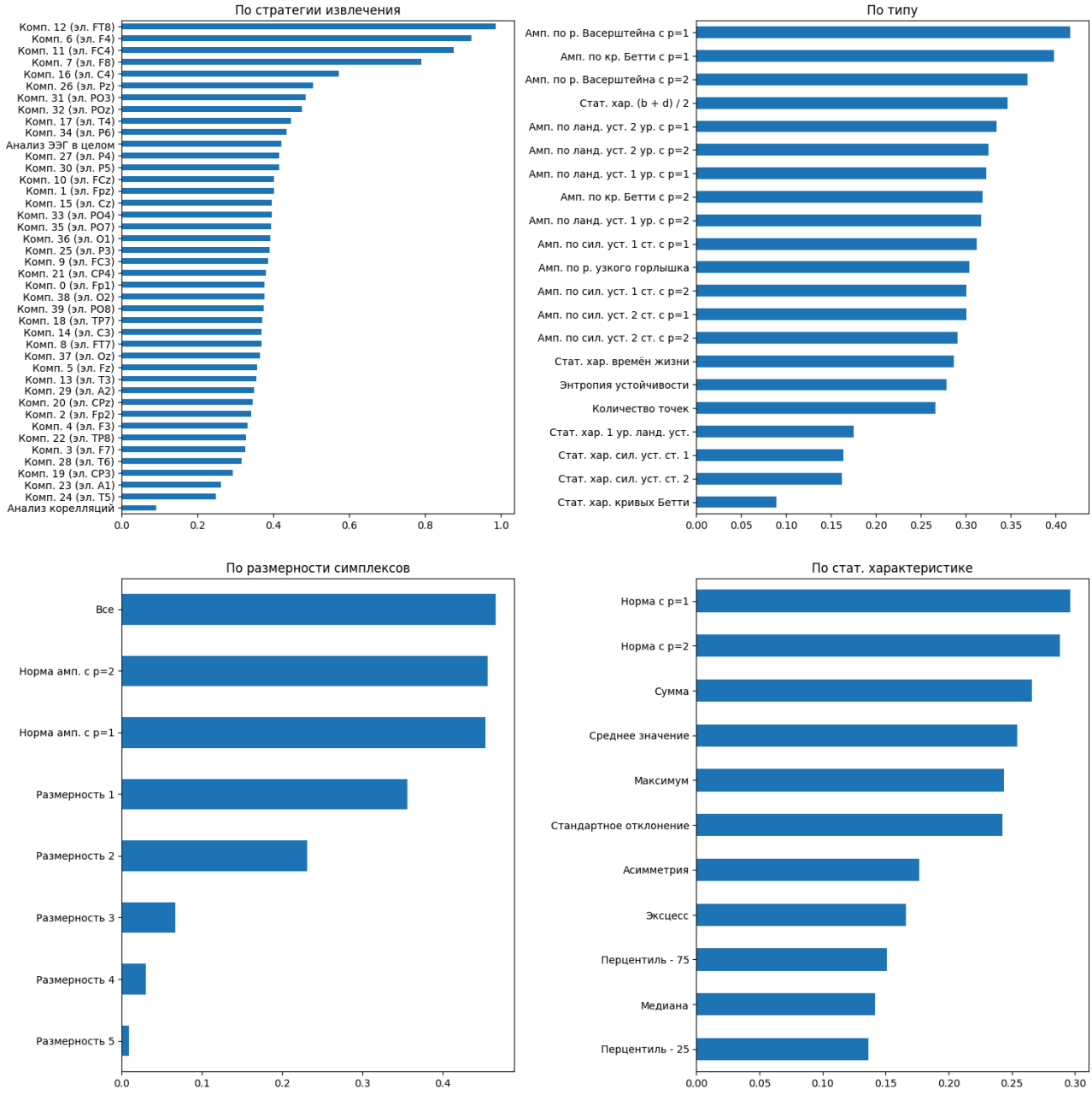

Рисунок 7 – Информационная ценность признаков для объекта 1 относительно лучшего результата по коэффициенту силуэта, агрегированная по параметрам их источника

5.4. Объект – 2

Для второго объекта производился поиск 8 функциональных состояний. Суммарная доля дисперсии исходных данных, объяснённая главными компонентами, составила 0,4.

При использовании топологических признаков алгоритм определил границы $\{0, 201, 497, 646, 994, 1171, 1746, 1995, 2017\}$ при выборе ответа по внутр. метрике и $\{0, 224, 543, 793, 1003, 1197, 1482, 1826, 2017\}$ – по внешней метрике против границ $\{0, 370, 526, 728, 1052, 1275, 1489, 1857, 2017\}$, полученных на традиционных признаках и $\{0, 346, 529, 843, 1109, 1276, 1493, 1795, 2017\}$ при использовании признаковых описаний вместе. Их сравнение представлено в таблице 3 и на рисунке 8.

Таблица 3 – Метрики качества результатов для объекта 2

	Ответ на традиционных признаках	Лучший ответ по коэффициенту силуэта	Лучший ответ по индексу Фаулкса – Маллоуса	Ответ при объединении признаков
Расстояние Уорда	18009	81393	86129	39687
Центроидное расстояние	12.38	28.61	23.99	17.27
Коэффициент силуэта	0.102	0.046	0.024	0.094
Индекс Калински – Харабаса	59.89	13.31	13.47	62.99
Индекс Дэвиса – Болдина	2.936	6.226	7.411	3.229
Коэфф. взаимной информации		0.693	0.793	0.850
Индекс Рэнда		0.474	0.668	0.762
Индекс Фаулкса-Маллоуса		0.557	0.713	0.795

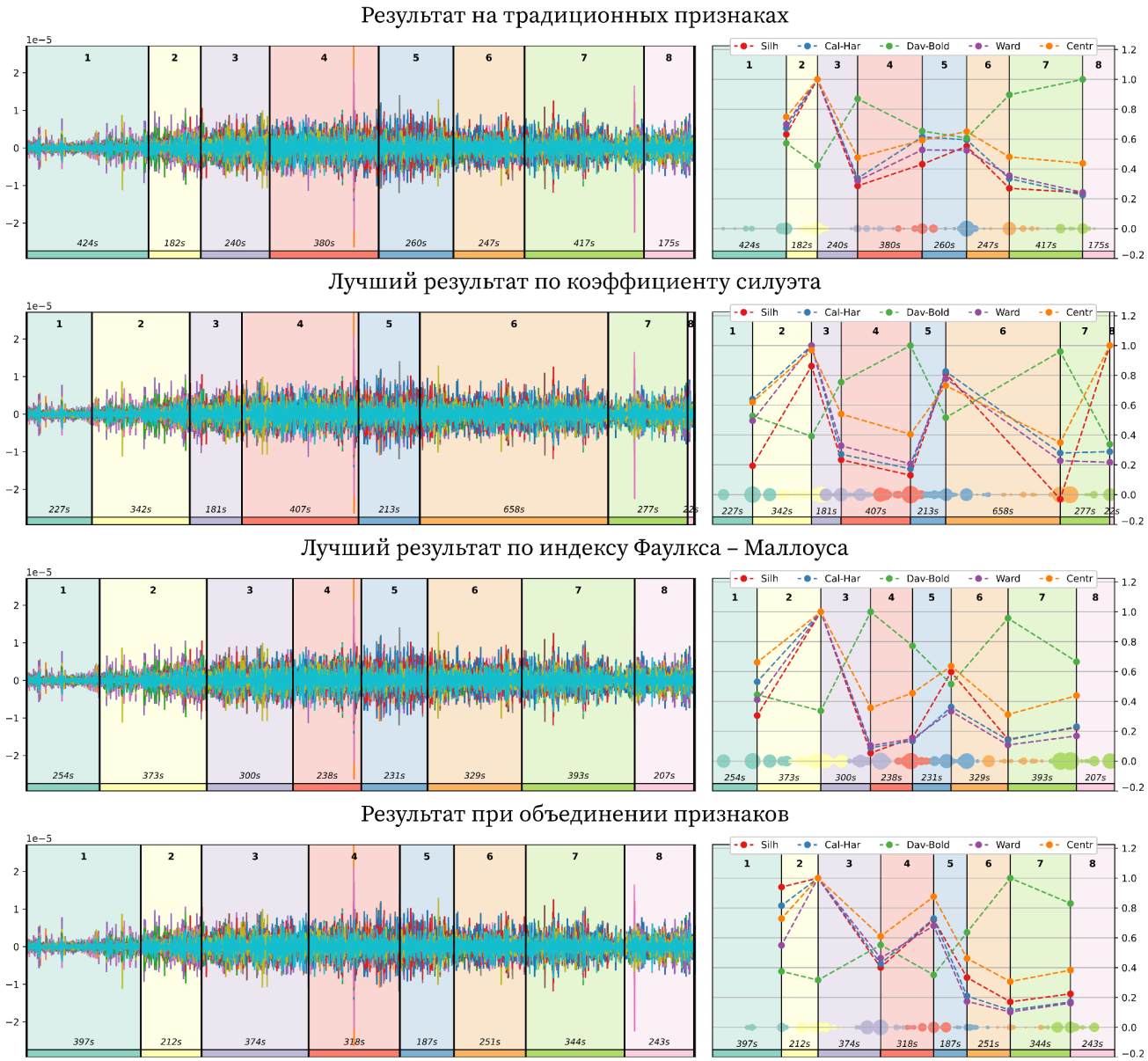

Рисунок 8 – Изображения ЭЭГ, диаграмм рассеяния потенциальных границ и нормированных метрик качества кластеризации всех найденных границ для объекта 2

Для второго объекта, как показано на рисунке 9, отличия информационной ценности относительно различных ответов оказались заметными, особенно для результата, выбранного по значению коэффициента силуэта, относительно которого наиболее информативные признаки имеют ценность около 1,2 при ценности относительно “правильного” ответа лишь немного превышающей 1,0.

Рисунок 9 – Гистограммы разностей между информационной ценностью признаков относительно полученных ответов и относительно “правильного” ответа для объекта 2

Наиболее информативными признаками (см. рис. 10) оказались амплитуды диаграмм устойчивости по их ландшафту и по расстоянию Васерштейна с $p = 2$, а также статистические характеристики последовательностей средних значений между моментами появления и исчезновения симплексов. Наибольшая ценность наблюдается у признаков, полученных при анализе 0, 2, 4 и 5 компонент временных рядов, соответствующих электродам Fp1, Fp2, F3 и Fz в префронтальной, а также левой и средней фронтальных частях мозга (рис. 1). Более того, высокие значения информационной ценности показали характеристики диаграмм устойчивости симплексов 1, 2 и 3 размерностей, полученных путём анализа ЭЭГ в целом (п. 2.16.3). Наиболее важными статистическими характеристиками, аналогично объекту 1, оказались нормы с $p = 1$ и $p = 2$, а также среднее значение и стандартное отклонение.

Наименее информативными оказались признаки, полученные путём анализа корреляций между переменными, а также при описании данных, собранных электродами T4 (комп. 17), CP3 (комп. 19), P3 (комп. 25) и O2 (комп. 38), находящимися в правой височной, левой центральной, левой теменной и правой затылочной частях мозга соответственно. При этом энтропия устойчивости, а также описания кривых Бетти и силуэтов показали наименьшую величину информационной ценности, а наименее важными статистическими характеристиками вновь оказались перцентили, эксцесс и асимметрия.

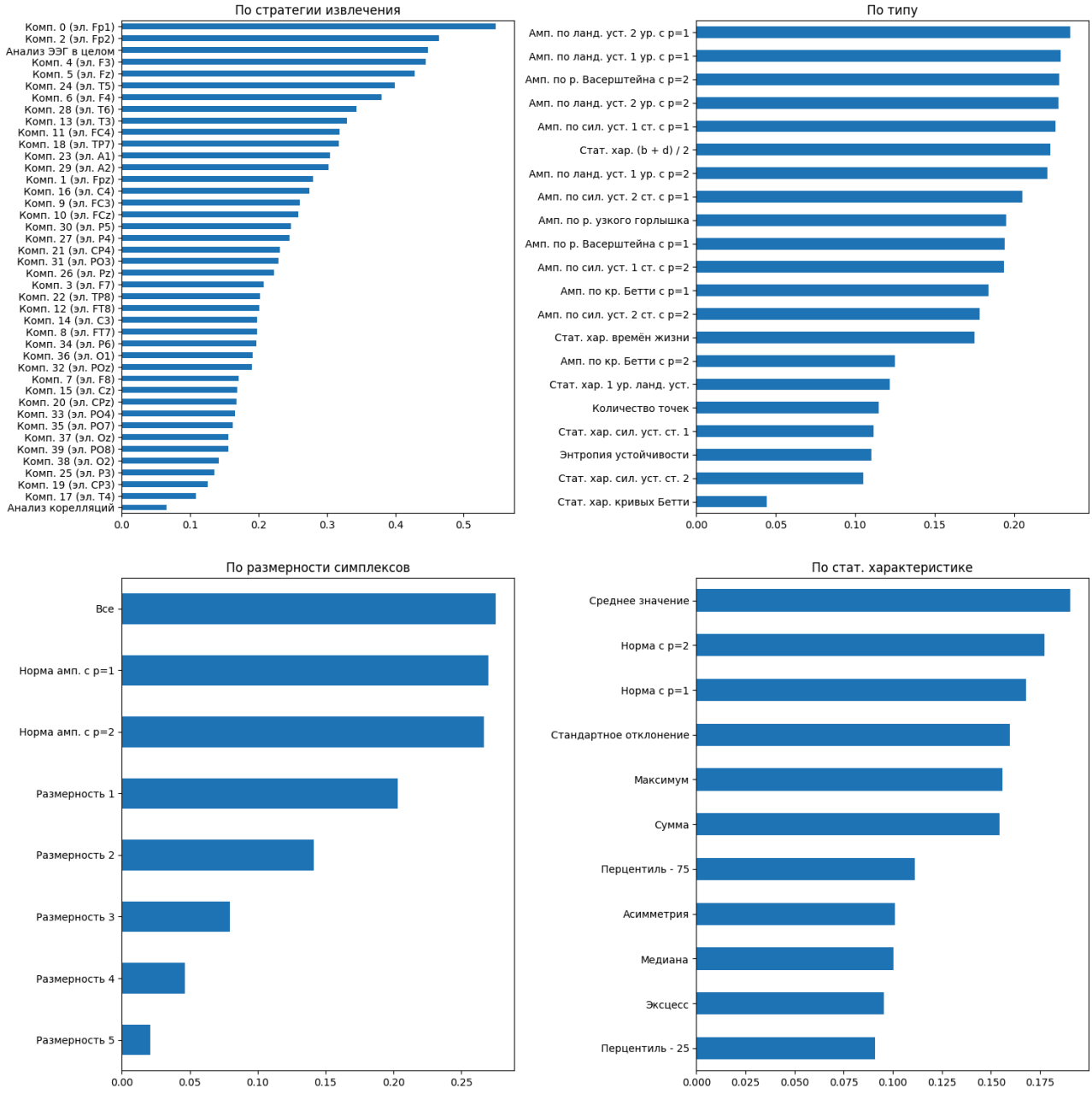

Рисунок 10 – Информационная ценность признаков для объекта 2 относительно лучшего результата по коэффициенту силуэта, агрегированная по параметрам их источника

Совокупно, влияния отдельных признаков на целевую переменную для второго объекта оказались значительно меньше, чем для первого объекта, а величина информационной ценности оказалась ниже 0,25 для более, чем 17000 признаков при максимальном значении лишь немного превышающем 1,25.

5.5. Объект – 3

Для третьего объекта производился поиск 10 функциональных состояний. Суммарная доля дисперсии исходных данных, объяснённая главными компонентами, составила 0,48.

При использовании топологических признаков при выборе ответа по внутр. метрике алгоритм определил границы { 0, 17, 195, 262, 466, 684, 859, 949, 1046, 1120, 1180 }, по внешней метрике – { 0, 153, 197, 262, 458, 684, 827, 949, 1046, 1120, 1180 } против границ { 0, 133, 175, 261, 458, 685, 783, 938, 1037, 1126, 1180 }, полученных на традиционных признаках, и { 0, 13, 132, 215, 460, 679, 846, 938, 1036, 1134, 1180 } при использовании признаковых описаний вместе. Их сравнение представлено в таблице 4 и на рисунке 11.

Таблица 4 – Метрики качества результатов для объекта 3

	Ответ на традиционных признаках	Лучший ответ по коэффициенту силуэта	Лучший ответ по индексу Фаулкса – Маллоуса	Ответ при объединении признаков
Расстояние Уорда	9973	97877	90143	38967
Центроидное расстояние	14.06	43.89	40.59	27.11
Коэффициент силуэта	0.105	0.059	0.048	0.191
Индекс Калински – Харабаса	27.34	13.44	12.34	64.20
Индекс Дэвиса – Болдина	2.735	4.404	4.401	2.072
Коэфф. взаимной информации		0.853	0.893	0.872
Индекс Рэнда		0.786	0.871	0.802
Индекс Фаулкса-Маллоуса		0.814	0.887	0.828

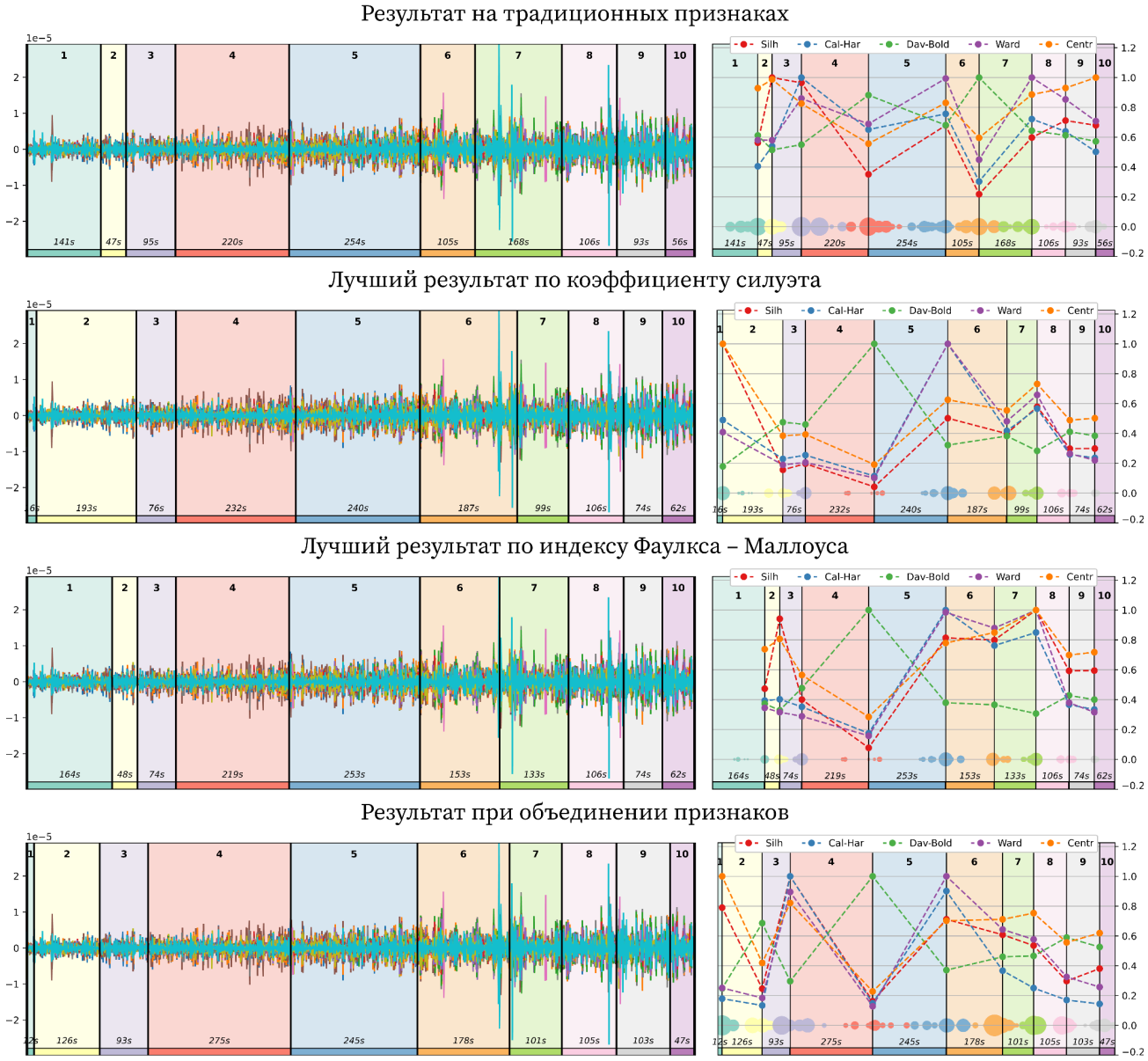

Рисунок 11 – Изображения ЭЭГ, диаграмм рассеяния потенциальных границ и нормированных метрик качества кластеризации всех найденных границ для объекта 3

Аналогично объекту 1, многие признаки показали высокое сходство информационной ценности относительно различных результатов. Тем не менее для отдельных признаков наблюдается значительная разница, достигающая 0,4 по абсолютной величине, как показано на рисунке 12.

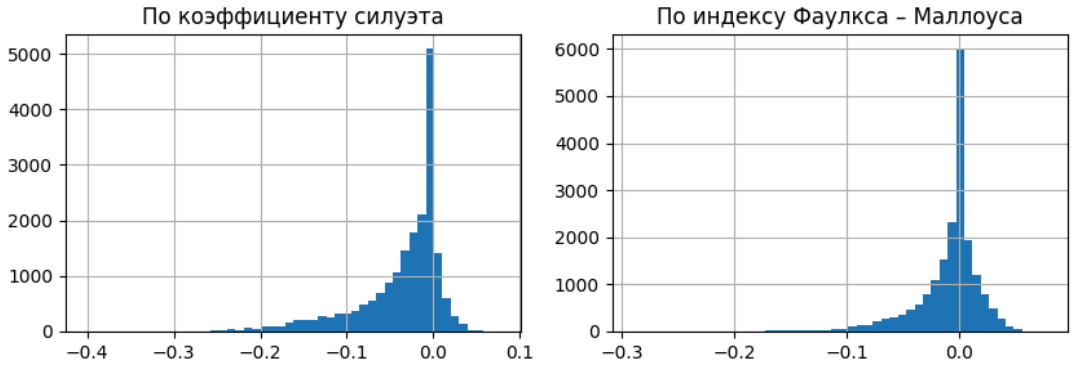

Рисунок 12 – Гистограммы разностей между информационной ценностью признаков относительно полученных ответов и относительно “правильного” ответа для объекта 3

Для третьего объекта, аналогично результатам для других объектов, наибольшую ценность (см. рис. 13) имеют статистические характеристики последовательностей средних значений между моментами появления и исчезновения симплексов, а также амплитуды по расстоянию Васерштейна. При этом наибольшая важность наблюдается у признаков, описывающие компоненты 0 (эл. Fp1), 3 (эл. F7), 4 (эл. F3) и 28 (эл. T6), соответствующие электродам, расположенным в префронтальной, левой фронтальной и правой височной областях головного мозга. Наиболее значимыми статистическими характеристиками также оказались векторные нормы порядков 1 и 2.

Наименее информативными, как и замечено для других объектов, являются признаки, полученные путём анализа корреляций между переменными, а также описывающие данные, соответствующие электродам Pz (комп. 26), PO3 (комп. 31), POz (комп. 32) и PO4 (комп. 33), находящиеся в средней теменной и затылочной частях мозга. Наименее ценными типами признаков оказались описания кривых Бетти и силуэтов устойчивости, а статистическими характеристиками – эксцесс, асимметрия и перцентили.

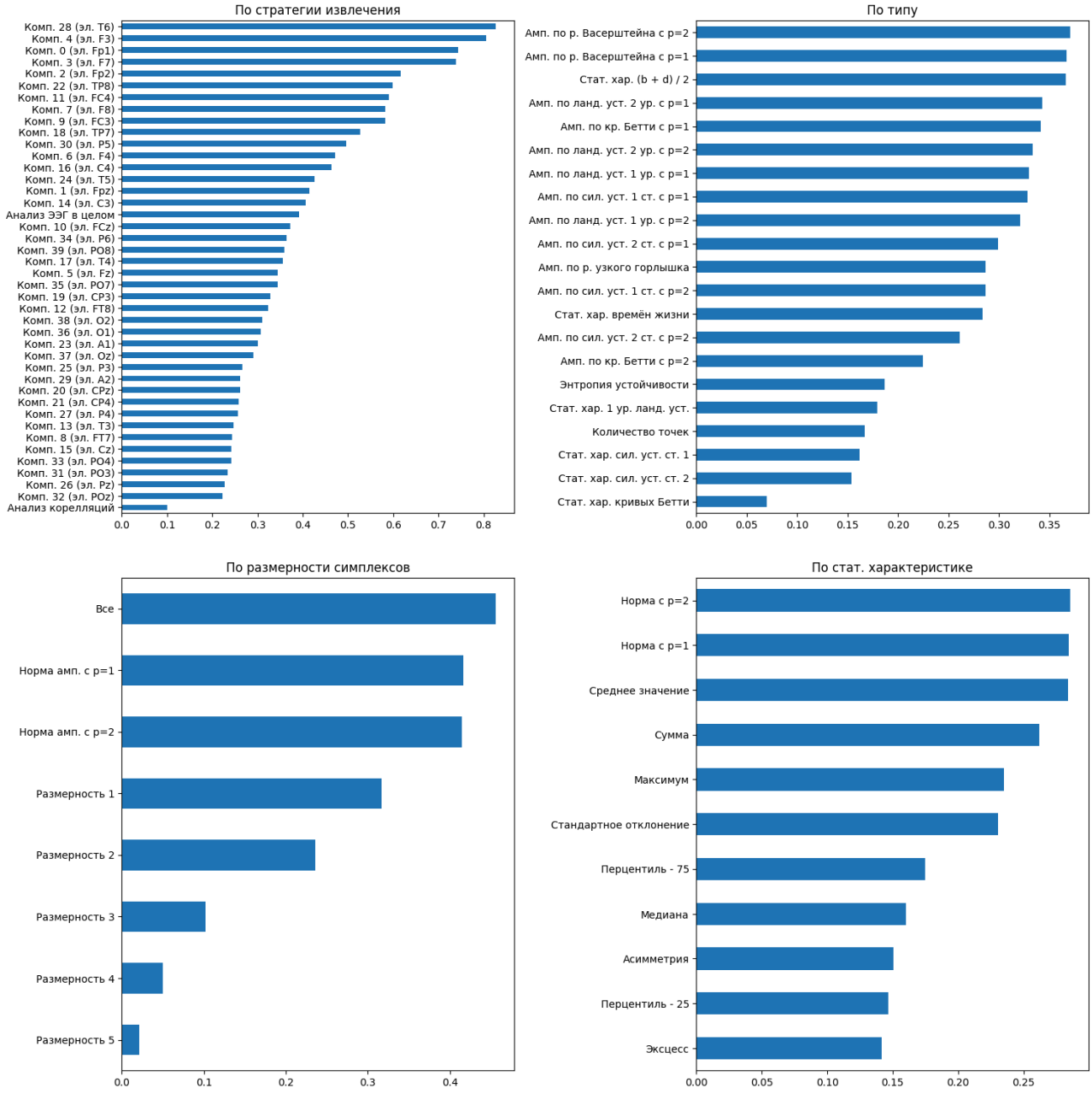

Рисунок 13 – Информационная ценность признаков для объекта 3 относительно лучшего результата по коэффициенту силуэта, агрегированная по параметрам их источника

6. ВЫВОД

Хотя полного совпадения результатов, полученных на основе топологических признаков, с результатами, полученными на традиционных признаках, достичь и не удалось, для всех трёх объектов явно прослеживаются корреляции полученных ответов с ранее известными, что подтверждается относительно неплохими значениями внешних метрик. Легко заметить, что границы, хорошо разделяющие соответствующие кластеры, были найдены во всех случаях как на традиционных и топологических признаках по-отдельности, так и при их совместном использовании. Тем не менее отдельные границы, разделяющие похожие состояния с низкими показателями межкластерных расстояний и других внутренних метрик качества, при использовании топологических признаков были смещены или не найдены вовсе. Так, например, в процессе подбора гиперпараметров и проведения экспериментов было замечено, что алгоритм плохо находит границу “784” для объекта 1, добавляя лишнюю границу в начало записи. А для объекта 2 сходство результата, выбранного по коэффициенту силуэта, с “правильным” оказалось гораздо ниже, чем для результата, выбранного по индексу Фаулкса-Маллоуса, что указывает на способность алгоритма получить похожий ответ, хотя он и не оказывается лучшим. Аналогичный результат прослеживается и для объекта 3. Это может указывать как на то, что топологические признаки теряют важные закономерности, присутствующие в исходных данных, так и на то, что они обнаруживают дополнительные значимые зависимости, не зафиксированные традиционными признаками. Для дальнейшего принятия решения потребуется использование более точных методов оценки результатов и обращение к экспертам в области нейрофизиологии.

Также важно заметить, что значения внутренних метрик качества кластеризации при использовании топологического подхода оказались в 2 – 3 раза хуже, чем на традиционных признаках. Тем не менее это не позволяет однозначно утверждать об их низком качестве. В ходе эксперимента производилось вычисление почти 20000 признаков и выделение 65 главных компонент, что значительно превышает размер традиционного признакового пространства – 15 главных компонент при 765 значениях. Из-за этого использованные для кластеризации главные компоненты, хотя и объясняют лишь 50% дисперсии всех признаков, описывают большое количество низкоуровневых закономерностей (“шума”), что

отрицательно сказывается на разделимости и стабильности получаемых границ и, как следствие, на величинах внутренних метрик качества.

Это подтверждается и результатами, полученными при объединении традиционного и топологического признаковых пространств, для которых ожидаемо заметно увеличиваются значения внешних метрик, но улучшаются и значения внутренних метрик, что говорит о повышении разделимости кластеров. Таким образом, лучшие топологические признаки не только обнаруживают новые закономерности, но и подтверждают известные, что может свидетельствовать о возможности их применения для получения хороших результатов.

Более того, для всех объектов многие признаки показывают высокое сходство значений информационной ценности относительно различных результатов, хотя в отдельных случаях и прослеживаются явные различия, которые свидетельствуют о том, что топологические признаки фиксируют закономерности в исходных данных, которые не были обнаружены традиционными признаками. Тем не менее для определения, насколько эти зависимости способствуют нахождению верных границ функциональных состояний (действительно ли они описывают общие закономерности, а не “шум”), требуется их дальнейшая более детальная оценка.

Анализ информационной ценности также выявил ряд закономерностей об “источниках” наиболее важных признаков:

1. Признаки, полученные по стратегии “анализ каждой переменной независимо” (п. 2.16.1) имеют наибольшую ценность для решения задачи, причём наиболее важными оказываются данные, собранные во фронтальной и правой частях мозга, а наименее связанными с результатом кластеризации – в левой и затылочной частях;

2. Признаки, полученные по стратегии “анализ корреляций между переменными” (п. 2.16.2) наименее информативны для определения границ функциональных состояний;

3. Наиболее ценными признаками являются статистические характеристики последовательностей средних значений между моментами появления и исчезновения симплексов, а также амплитуды по различным метрикам (особенно – по расстоянию Васерштейна и по ландшафтам устойчивости), тогда как наименее информативными для всех рассмотренных объектов оказались статистические характеристики кривых Бетти и силуэтов устойчивости;

4. Наибольшая информация заключена в симплексах размерностей 1 и 2, хотя симплексы больших размерностей лишь дополняют признаковое пространство и также могут фиксировать важные закономерности;

5. Наиболее ценные статистические характеристики – манхэттенская и евклидова нормы, а наименее важные – перцентили (25, 50 и 75), эксцесс и асимметрия последовательностей.

ЗАКЛЮЧЕНИЕ

В ходе работы над проектом были изучены, реализованы и применены различные алгоритмы топологического анализа данных ЭЭГ непрерывных процессов, а также существующие методы нахождения функциональных состояний по их признаковому описанию. Полученные результаты позволяют утверждать, что топологические признаки применимы к задаче нахождения функциональных состояний по ЭЭГ и не уступают, а может быть даже превосходят по качеству традиционные признаки.

Как показал проведённый эксперимент на записях ЭЭГ процесса медитации Guhyasamaja Tantra, при получении на вход данных, обработанных методами алгебраической топологии, алгоритм SDA способен находить границы, похожие на таковые, полученные с помощью традиционных признаков, при наличии информации о менее, чем 50% дисперсии всех извлечённых признаков. Тем не менее в отдельных случаях алгоритм не посчитал этот ответ оптимальным, отдав предпочтение результатам, менее похожим на ранее известные, что может свидетельствовать как о потере важной информации в процессе топологического анализа, так и об обнаружении новых закономерностей, ранее не выявленных существующими решениями.

Анализ информационной ценности топологических признаков также показал, что они имеют много общего с традиционными подходами и даже позволяют выявлять области головного мозга, данные из которых наиболее влияют на результат. Но было замечено, что прослеживается и ряд значимых различий в информационной ценности топологических признаков относительно различных ответов, что может свидетельствовать об обнаружении важных, ранее не выявленных закономерностей.

При использовании традиционных и топологических признаков совместно наблюдается увеличение качества результата в том числе по внутренним метрикам, что свидетельствует об улучшении разделимости найденных кластеров в сравнении с подходом, использующим только традиционные методы. Таким образом, топологические признаки не только позволяют найти хорошие границы функциональных состояний, но и информационно дополняют традиционное описание, подчеркивая известные закономерности и обнаруживая новые.

Проведение дальнейших исследований на тему применимости топологического анализа данных к нахождению функциональных состояний по ЭЭГ целесообразно в следующих направлениях:

1. Отбор наиболее информативных топологических признаков, увеличение объема и количества обрабатываемых данных: удаление признаков, с наибольшей вероятностью описывающих “шум”, а не общие закономерности, как следствие, понижение размерности итогового признакового пространства и повышение доли объясняемой им дисперсии, сбор новых записей ЭЭГ процессов медитации других монахов.

2. Анализ совместного поведения традиционных и топологических признаков: выявление общих закономерностей, исследование ключевых отличий и определение набора признаков, наилучшим образом дополняющих друг друга при решении поставленной задачи.

3. Более детальная оценка информационного вклада топологических признаков путём построения моделей многоклассовой классификации, предсказывающих номер кластера по признаковому описанию эпохи.

4. Привлечение экспертов в области нейрофизиологии для сравнения получаемых результатов не только на основании метрик качества кластеризации, но и с учётом существующих знаний о мозговой активности человека во время изучаемых процессов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. SDA: a data-driven algorithm that detects functional states applied to EEG of Guhyasamaja meditation [Электронный ресурс] / E. V. Mikhaylets, A. R. Razorenova, V. L. Chernyshev, N. V. Syrov, L. V. Yakovlev, J. A. Boytsova, E. V. Kokurina, Y. S. Zhironkina, S. V. Medvedev and A. Y. Kaplan. – Front. Neuroinform., 29 January 2024. – URL: https://doi.org/10.3389/fninf.2023.1301718. (дата обращения: 01.02.24).
2. MEG and EEG data analysis with MNE-Python [Электронный ресурс] / A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen and M. Hämäläinen. – Front. Neurosci., 26 December 2013. – URL: https://doi.org/10.3389/fnins.2013.00267. (дата обращения: 01.02.24).
3. Stability of persistence diagrams [Электронный ресурс] / D. Cohen-Steiner, H. Edelsbrunner and J. Harer. – Discrete & Computational Geometry: электрон. журн., vol. 37 (2007), pp. 103 – 120. – Springer, 12 December 2006. – URL: https://doi.org/10.1007/s00454-006-1276-5. (дата обращения: 01.02.24).
4. Topological Persistence and Simplification [Электронный ресурс] / H. Edelsbrunner, D. Letscher and A. Zomorodian. – Discrete & Computational Geometry: электрон. журн., vol. 28 (2002), pp. 511 – 533. – Springer, 01 November 2002. – URL: https://doi.org/10.1007/s00454-002-2885-2. (дата обращения: 01.02.24).
5. Computing persistent homology [Электронный ресурс] / A. Zomorodian, G. Carlsson. – Discrete & Computational Geometry: электрон. журн., vol. 33 (2005), pp. 249 – 274. – Springer, 19 November 2004. – URL: https://doi.org/10.1007/s00454-004-1146-y. (дата обращения: 01.02.24).
6. giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration [Электронный ресурс] / M. Rucco, F. Castiglione, E. Merelli, M. Pettini. – arXiv, 5 Mar 2021. – URL: https://doi.org/10.48550/arXiv.2004.02551. (дата обращения: 01.02.24).
7. Array Programming with NumPy [Электронный ресурс] / C.R. Harris, K.J. Millman, S.J. van der Walt et al. – arXiv, 18 Jun 2020. – URL: https://doi.org/10.48550/arXiv.2006.10256. (дата обращения: 16.03.24).
8. SciPy 1.0 – Fundamental Algorithms for Scientific Computing in Python [Электронный ресурс] / P. Virtanen, R. Gommers, T. E. Oliphant et al. – arXiv, 23 Jul 2019. – URL: https://doi.org/10.48550/arXiv.1907.10121. (дата обращения: 16.03.24).

9. Scikit-learn: Machine Learning in Python [Электронный ресурс] / F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay. – arXiv, 5 Jun 2018. – URL: https://doi.org/10.48550/arXiv.1201.0490. (дата обращения: 01.02.24).
10. Matplotlib: A 2D Graphics Environment [Электронный ресурс] / J. D. Hunter et al. – Computing in Science & Engineering, vol. 9, iss. 3, pp. 90 – 95. – IEEE, 18 June 2007. – URL: https://doi.org/10.1109/MCSE.2007.55. (дата обращения: 16.03.24).
11. pandas-dev/pandas: Pandas [Электронный ресурс] / The pandas development team. – Zenodo, Feb 2020. – URL: https://doi.org/10.5281/zenodo.3509134. (дата обращения: 16.03.24).
12. tqdm: A fast, Extensible Progress Bar for Python and CLI [Электронный ресурс] / C. da Costa-Luis, S. K. Larroque, K. Altendorf et al. – Zenodo, 2019. – URL: https://doi.org/10.5281/zenodo.595120. (дата обращения: 16.03.24).
13. Detecting strange attractors in turbulence [Электронный ресурс] / F. Takens. – Dynamical Systems and Turbulence, Warwick 1980. Lecture Notes in Mathematics, vol 898. – Berlin: Springer, 07 October 2006. – URL: https://doi.org/10.1007/BFb0091924. (дата обращения: 01.02.24).
14. Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis [Электронный ресурс] / J. Perea and J. Harer. – arXiv, 25 Nov 2013. – URL: https://doi.org/10.48550/arXiv.1307.6188. (дата обращения: 01.02.24).
15. Independent coordinates for strange attractors from mutual information [Электронный ресурс] / A. M. Fraser and H. L. Swinney. – Physical Review A: электрон. журн., vol. 33, iss. 2 (1986), pp. 1134 – 1140. – American Physical Society, 22 July 1985. – URL: https://doi.org/10.1103/PhysRevA.33.1134. (дата обращения: 16.03.24).
16. Determining embedding dimension for phase-space reconstruction using a geometrical construction [Электронный ресурс] / M. B. Kennel, R. Brown, and H. D. I. Abarbanel. – Physical Review A: электрон. журн., vol. 45, iss. 6 (1992), pp. 3403 – 3411. – American Physical Society, 24 April 1991. – URL: https://doi.org/10.1103/PhysRevA.45.3403. (дата обращения: 16.03.24).

17. Notes on regression and inheritance in the case of two parents [Электронный ресурс] / K. Pearson. – Proceedings of the Royal Society of London, 1 January 1895. – URL: https://doi.org/10.1098/rspl.1895.0041. (дата обращения: 16.03.24).
18. On the imbedding of systems of compacta in simplicial complexes [Электронный ресурс] / K. Borsuk. – Fundamenta Mathematicae: vol. 35.1 (1948), pp. 217 – 234. – EUDML. – URL: http://eudml.org/doc/213158. (дата обращения: 01.02.24).
19. Théorie générale de l'homologie dans un espace quelconque [Электронный ресурс] / E. Čech. – Fundamenta Mathematicae: vol. 19.1 (1932), pp. 149 – 183. – EUDML. – URL: http://eudml.org/doc/212569. (дата обращения: 01.02.24).
20. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen [Электронный ресурс] / L. Vietoris. – Mathematische Annalen: vol. 97 (1927), pp. 454 – 472. – Springer, December 1927. – URL: https://doi.org/10.1007/BF01447877. (дата обращения: 01.02.24).
21. Fast construction of the Vietoris – Rips complex [Электронный ресурс] / A. Zomorodian. – Computers & Graphics: vol. 34 (2010), pp. 263 – 271. – ScienceDirect, 20 March 2010. – URL: https://doi.org/10.1016/j.cag.2010.03.007. (дата обращения: 01.02.24).
22. A Note on the Simplex-Tree Construction of the Vietoris – Rips Complex [Электронный ресурс] / U. Bauer. – arXiv, 30 Jan 2023. – URL: https://doi.org/10.48550/arXiv.2301.07191. (дата обращения: 01.02.24).
23. Ripser: efficient computation of Vietoris – Rips persistence barcodes [Электронный ресурс] / A. Rieser. – arXiv, 26 Feb 2021. – URL: https://doi.org/10.48550/arXiv.1908.02518. (дата обращения: 01.02.24).
24. Characterisation of the idiotypic immune network through persistent entropy [Электронный ресурс] / G. Tauzin, U. Lupo, L. Tunstall, J. Burella Pérez, M. Caorsi, W. Reise, A. Medina – Mardones, A. Dassatti, K. Hess. – Springer Proceedings in Complexity. – Springer, Cham, 04 May 2016. – URL: https://doi.org/10.1007/978-3-319-29228-1_11. (дата обращения: 01.02.24).
25. Statistical topological data analysis using persistence landscapes [Электронный ресурс] / P. Bubenik. – arXiv, 27 Jul 2012. – URL: https://doi.org/10.48550/arXiv.1207.6437. (дата обращения: 16.03.24).

26. Stochastic Convergence of Persistence Landscapes and Silhouettes [Электронный ресурс] / F. Chazal, B. Terese Fasy, F. Lecci, A. Rinaldo, L. Wasserman. – arXiv, 2 Dec 2013. – URL: https://doi.org/10.48550/arXiv.1312.0308. (дата обращения: 16.03.24).
27. Geometry Helps to Compare Persistence Diagrams [Электронный ресурс] / M. Kerber, D. Morozov, A. Nigmetov. – arXiv, 10 Jun 2016. – URL: https://doi.org/10.48550/arXiv.1606.03357. (дата обращения: 16.03.24).
28. On lines and planes of closest fit to systems of points in space [Электронный ресурс] / K. Pearson. – The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science: ser. 6, vol. 2, iss. 11 (1901), pp. 559 – 572. – Taylor & Francis, 10 Jun 2010. – URL: https://doi.org/10.1080/14786440109462720. (дата обращения: 16.03.24).
29. The approximation of one matrix by another of lower rank [Электронный ресурс] / C. Eckart and G. Young. – Psychometrika: vol. 1 (1936), pp. 211 – 218. – Springer, September 1936. – URL: https://doi.org/10.1007/BF02288367. (дата обращения: 16.03.24).
30. An examination of the effect of six types of error perturbation on fifteen clustering algorithms [Электронный ресурс] / G. W. Milligan. – Psychometrika: vol. 45 (1980), pp. 325 – 342. – Springer, September 1980. – URL: https://doi.org/10.1007/BF02293907. (дата обращения: 01.02.24).
31. Hierarchical Grouping to Optimize an Objective Function [Электронный ресурс] / H. Joe, Jr. Ward. – Journal of the American Statistical Association: vol. 58 (1963), pp. 236 – 244. – Taylor & Francis, 10 Apr 2012. – URL: https://doi.org/10.1080/01621459.1963.10500845. (дата обращения: 01.02.24).
32. Steinhaus, H. Sur la division des corps materiels en parties / H. Steinhaus // Bulletin L'Académie Polonaise des Science – 1957 – vol. 4 – c. 801 – 804.
33. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis [Электронный ресурс] / P. J. Rousseeuw. – Journal of Computational and Applied Mathematics: vol. 20, pp. 53 – 65. – ScienceDirect, November 1987. – URL: https://doi.org/10.1016/0377-0427(87)90125-7. (дата обращения: 01.02.24).
34. A dendrite method for cluster analysis [Электронный ресурс] / T. Calinski and J. Harabasz. – Communications in Statistics: vol. 3 (1974), pp. 1 – 27. – Taylor & Francis, 27 June 2007. – URL: https://doi.org/10.1080/03610927408827101. (дата обращения: 01.02.24).

35. A Cluster Separation Measure [Электронный ресурс] / D. L. Davies and D. W. Bouldin. – IEEE Transactions on Pattern Analysis and Machine Intelligence: vol. PAMI-1, pp. 224 – 227. – IEEE, April 1979. – URL: https://doi.org/10.1109/TPAMI.1979.4766909. (дата обращения: 01.02.24).
36. A mathematical theory of communication [Электронный ресурс] / C. E. Shannon. – The Bell System Technical Journal: vol. 27, pp. 623 – 656. – Nokia Bell Labs, October 1948. – URL: https://doi.org/10.1002/j.1538-7305.1948.tb00917.x. (дата обращения: 01.02.24).
37. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance [Электронный ресурс] / N. X. Vinh, J. Epps, J. Bailey. – Journal of Machine Learning Research: vol. 11 (2010), pp. 2837 – 2854. – JMLR, October 2010. – URL: https://jmlr.org/papers/volume11/vinh10a/vinh10a.pdf. (дата обращения: 01.02.24).
38. Objective Criteria for the Evaluation of Clustering Methods [Электронный ресурс] / W. M. Rand. – Journal of the American Statistical Association: vol. 66 (1971), pp. 846 – 850. – Taylor & Francis, 05 Apr 2012. – URL: https://doi.org/10.1080/01621459.1971.10482356. (дата обращения: 01.02.24).
39. Comparing partitions [Электронный ресурс] / L. Hubert and P. Arabie. – Journal of Classification: vol. 2 (1985), pp. 193 – 218. – Springer, December 1985. – URL: https://doi.org/10.1007/BF01908075. (дата обращения: 01.02.24).
40. A Method for Comparing Two Hierarchical Clusterings [Электронный ресурс] / E. B. Fowkles and C. L. Mallows. – Journal of the American Statistical Association: vol. 78 (1983), pp. 553 – 569. – Taylor & Francis, 12 Mar 2012. – URL: https://doi.org/10.1080/01621459.1983.10478008. (дата обращения: 01.02.24).
41. Osteyee, D. B. [Электронный ресурс]: Information, Weight of Evidence. The Singularity Between Probability Measures and Signal Detection. / D. B. Osteyee, I. J.Good. – Lecture Notes in Mathematics: vol. 376. – Springer, 1974. – URL: https://doi.org/10.1007/BFb0064126. (дата обращения: 16.03.24).

ПРИЛОЖЕНИЕ А

Календарный план работ

Работы по проекту производились в соответствии с календарным планом, представленным таблицей А.1.

Таблица А.1 – Календарный план работ

Этап работ	Дата завершения
Изучение алгоритмов топологического анализа данных	15.11.2023
Изучение принципов работы с данными ЭЭГ и алгоритма SDA	15.12.2023
Реализация необходимых алгоритмов и получение промежуточных результатов	15.01.2024
Подготовка черновой версии промежуточного отчета о проекте	01.02.2024
Подготовка финальной версии промежуточного отчета о проекте и его загрузка в Smart LMS	15.02.2024
Завершение заключительного этапа проекта, получение итоговых результатов	01.03.2024
Подготовка черновой версии итогового отчета о проекте	18.03.2024
Подготовка финальной версии итогового отчета о проекте и его загрузка в Smart LMS	25.03.2024
Получение отчёта на Антиплагиат итогового отчёта и его загрузка в Smart LMS	26.03.2024
Подготовка презентации проекта к защите и ее загрузка в Smart LMS	07.04.2024
Защита проекта комиссии	16.04.2024