ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Программная инженерия»

УДК 004.62, 004.8, 004.91, 515.1, 612.8

Отчет об исследовательском проекте

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

(промежуточный, этап 1)

Выполнен студентом:
Группы БПИ213 Абрамовым Александром Сергеевичем
ФИО студента

Проверен руководителем проекта:

Михайлец Екатерина Викторовна, кандидат физико-математических наук
ФИО, научная степень (если есть)

Приглашённый преподаватель, доцент
Должность

НИУ ВШЭ, Департамент математики
факультета экономических наук
Место работы (организация или департамент НИУ ВШЭ)

Москва 2024

РЕФЕРАТ

Отчёт 34 с., 1 кн., 5 рис., 4 табл., 27 источн., 0 прил.

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

Объектом исследования являются ЭЭГ процесса медитации по методу Tantric Guhyasamaja. Недавно разработанный алгоритм SDA показывает хорошее качество нахождения функциональных состояний по традиционным признакам, извлечённым из этого сигнала.

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

В ходе промежуточного этапа исследования были изучены и реализованы основные алгоритмы извлечения топологических признаков и анализа качества результата, а также произведён их запуск для получения и анализа первого ответа.

В результате работы алгоритм SDA был впервые применён к топологическим признакам, полученным по сигналу ЭЭГ, и показал приемлемое для первого запуска качество, определив большую часть границ функциональных состояний с достаточной точностью в сравнении с ранее полученным результатом.

На основании этого сделан вывод о целесообразности продолжения исследования. В ходе заключительного этапа работы будет произведён статистический анализ полученных признаков и произведён подбор гиперпараметров для достижения хорошего качества результата.

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

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 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 динамику значений одного параметра

Признак - численное значение, описывающее некоторый объект

Признаковое описание - набор (вектор) признаков некоторого объекта

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

Стандартизация - процесс приведения набора данных к заданным среднему и дисперсии

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

$|z|$ - Евклидова норма вектора z.

$|Z|$ - количество элементов во множестве Z.

ВВЕДЕНИЕ

Традицонно, анализ ЭЭГ полагается на информацию о событиях, искусственно созданных во время эксперимента (предъявление объекту исследования некоторой информации, реакция на неё, 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 отсутствии 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 сходства при различных значениях гиперпараметров;

2) Выбор лучших границ функциональных состояний путём применения подходящего алгоритма кластеризации к совокупности 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 получено в [1], алгоритм показывает 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 теории 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 удобнее в использовании, позволяют достичь лучшего качества и 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.

В ходе первого этапа проекта был изучен, реализован и оптимизирован алгоритм SDA и методы оценки качества выводимого им ответа, что позволило воспроизвести полученные в [1] результаты, а также был построен алгоритм извлечения топологических признаков, получен и проанализирован первый результат.

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. 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.

2. ПРИЗНАКОВОЕ ОПИСАНИЕ ДАННЫХ

2.1. Топологический анализ данных

Топологический анализ данных - 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 наборов точек и показывающие количество и устойчивость 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 (изображений, временных рядов, текстов и др.) в набор точек некоторым образом. В частности, для работы с временными рядами применим метод получения эмбеддингов Такенса [4-5], который был использован в настоящей работе и будет подробно рассмотрен далее.

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

3) Вычисление устойчивой гомологии [6] 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 по диаграмме устойчивости в соответствии с основной теоремой устойчивых гомологий [7] с помощью, например, построения последовательностей чисел Бетти [7] или вычисления энтропии

устойчивости [8]. Полученные таким образом признаки могут использоваться в классических алгоритмах машинного обучения для решения различных задач.

Для реализации описанного процесса была использована библиотека giotto-tda [9] на языке Python. В рамках промежуточного этапа работы извлечение топологических признаков производилось по алгоритму, проиллюстрированному рисунком 1. На каждом этапе производилось соответственно преобразование временных рядов в наборы точек с помощью алгоритма получения эмбеддингов Такенса (TakensEmbedding), построение комплексов Вьеториса - Рипса для вычисления устойчивых гомологий и диаграмм устойчивости (VietorisRipsPersistence), масштабирование и фильтрация диаграмм для удаления шума (Scaler и Filtering) и получение признаковых описаний по диаграммам с помощью энтропии устойчивости (PersistenceEntropy).

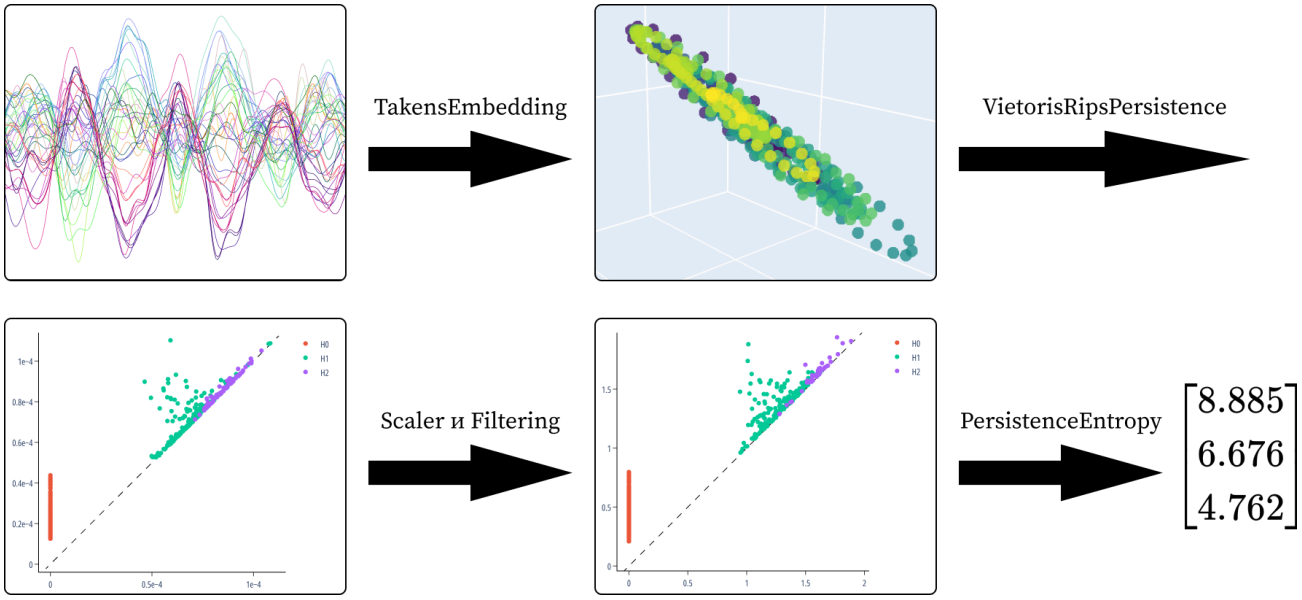

Рисунок 1 - Схема алгоритма извлечения топологических признаков

2.2. Алгоритм получения эмбеддингов Такенса

Для преобразования временных рядов, полученных в результате разделения ЭЭГ на эпохи по времени, в наборы точек был использован алгоритм Такенса [4-5].

Пусть дан одномерный временной ряд $(X_0, X_1, X_2, ..., X_n)$, задана размерность d пространства желаемого набора точек и равномерно в интервале от 0 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).

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

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 $t_1,\ t_2,\ \ldots,\ t_q$ так, чтобы последняя координата последней точки ($P_{q,\,d}$) равнялась последнему значению исходного временного ряда (X_n).

При работе с многомерными временными рядами описанный алгоритм применяют независимо к каждой компоненте ряда, объединяя полученные результаты. Этот подход реализован в библиотеке giotto-tda классом TakensEmbedding.

2.3. 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 [10] гарантирует, что такому свойству удовлетворяет комплекс Чеха [11], 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 [12], который в общем случае не удовлетворяет требуемому свойству, но достаточно хорошо подходит для решения большинства практических задач. Комплекс Вьеториса-Рипса определяется формулой (2) 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.

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

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

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

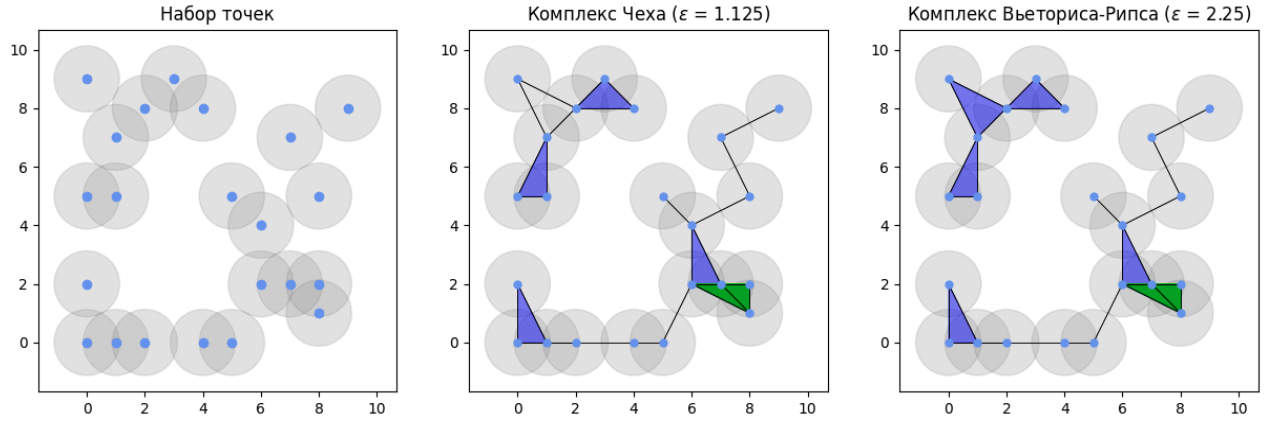

Рисунок 2 - Набор точек и комплексы Чеха и Вьеториса-Рипса для него

Далее 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 [13-15]. 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.

2.4. Масштабирование и фильтрация

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 диаграмму - удалить точки, наиболее близкие к диагонали. Для унификации величины фильтрации следует предварительно привести значения к единому масштабу. Описанные процессы были реализованы с помощью классов Scaler и Filtering библиотеки giotto-tda соответственно.

2.5. Энтропия устойчивости

В заключение необходимо преобразовать диаграммы устойчивости в векторы признаков, для чего в рамках промежуточного этапа работы производилось вычисление энтропии устойчивости [8].

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 для этой диаграммы вычисляется по формуле (3).

$$p_i = \frac{d_i - b_i}{\sum_{j=1}^{n} d_j - b_j},$$
$$E(D) = -\sum_{i=1}^{n} p_i \cdot ln(p_i), \tag{3}$$

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.

3. АЛГОРИТМ НАХОЖДЕНИЯ ФУНКЦИОНАЛЬНЫХ СОСТОЯНИЙ

3.1. 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 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 [16] с различными значениями количества искомых состояний ($n\_clusters$) и максимального времени между эпохами в матрице сходства ($k\_neighbours$). 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 [17]. 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].

3.2. Агломеративный метод иерархической кластеризации Уорда

Иерархическая кластеризация - класс алгоритмов кластеризации, основанный на создании дерева вложенных кластеров и примечательный возможностью задания матрицы сходства, 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 [18], что наилучшее качество для данных ЭЭГ показывает метод Уорда [16] - 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 (4) как прирост суммарного квадратичного отклонения элементов кластеров от их центров в результате объединения.

$$D(X, Y) = \sum_{z \in X \cup Y} |z - \overline{X \cup Y}|^2 - \left(\sum_{x \in X} |x - \overline{X}|^2 + \sum_{y \in Y} |y - \overline{Y}|^2 \right), \quad (4)$$

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 (5).

$$\overline{X} = \frac{1}{|X|} \cdot \sum_{x \in X} x, \tag{5}$$

В рамках работы была использована реализация на языке Python, представленная классом AgglomerativeClustering библиотеки scikit-learn [19].

3.3. Метод k-средних

Метод k-средних [17] - 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 (5) 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 применения в различных задачах анализа данных и его реализация, представленная классом KMeans библиотеки scikit-learn языка Python, была использована в настоящей работе.

4. МЕТРИКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ

В рамках исследования были использованы два способа оценки качества результата: внутренняя оценка, показывающая, насколько сильно полученные кластеры отличаются друг от друга, и внешняя оценка, позволяющая сравнить полученные результаты между собой.

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 оригинальной статье [1].

4.1.1. Расстояние Уорда

Расстояние Уорда вычисляется по формуле (4) 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 (6). Большое значение метрики соответствует удаленным кластерами и высокому качеству кластеризации.

$$d_c(X, Y) = |\overline{X} - \overline{Y}|_p, \tag{6}$$

где $\overline{X}$, $\overline{Y}$ - центры множеств X и Y соответственно, вычисляемые по формуле (5).

$|z|_p$ - норма вектора z под номером p, в настоящем исследовании использовалась Евклидова норма ($p = 2$).

4.1.3. Коэффициент силуэта

Коэффициент силуэта [20] вычисляется для одного объекта x по формуле (7) 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. 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.

$$\begin{gathered} a(x, C_x) = \frac{1}{|C_x| - 1} \cdot \sum_{y \in C_x} |x - y|, \\ b(i, C_i) = min_{C_y \neq C_x} \frac{1}{|C_y|} \cdot \sum_{y \in C_y} |x - y|, \\ s(x, C_x) = \frac{b - a}{max(a, b)}, \end{gathered} \tag{7}$$

Описанный алгоритм реализован функцией silhouette_score библиотеки scikit-learn.

4.1.4. Индекс Калински-Харабаса

Индекс Калински-Харабаса [21] вычисляется по формуле (8) 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.

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

$$W = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \overline{C_i}|^2, \quad (8)$$

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

где n - количество объектов;

k - количество выделенных кластеров

$\overline{C_i}$ - центр кластера C_i, вычисляемый по формуле (5);

$\overline{C}$ - центр множества всех точек, вычисляемый по формуле (5).

Описанный алгоритм реализован функцией calinski_harabasz_score библиотеки scikit-learn.

4.1.5. Индекс Дэвиса-Болдина

Индекс Дэвиса-Болдина [22] вычисляется по формуле (9) 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.

$$DB = \frac{1}{k} \cdot \sum_{i=1}^{k} max_{i \neq j} \frac{s_i + s_j}{d_{ij}}, \quad (9)$$

где k - количество кластеров;

d_{ij} - центроидное расстояние между кластерами i и j, вычисляемое по формуле (6);

s_i и s_j - диаметры кластеров i и j соответственно, вычисляемые по формуле (10).

$$s_i = \frac{1}{|C_i|} \cdot \sum_{x \in C_i} |x - \overline{C_i}|, \quad (10)$$

где C_i - кластер под номером i;

$\overline{C_i}$ - центр множества C_i, вычисляемый по формуле (5).

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 использованы 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 [23] вычисляется по формуле (11) и является мерой их согласованности - насколько одно распределение “зависит” от другого. Значения, близкие к единице, указывают на высокое совпадение распределений.

$$\begin{gathered} P_U(i) = \frac{|U_i|}{N}, \\ P_V(j) = \frac{|V_j|}{N}, \\ P(i, j) = \frac{|U_i \cap V_j|}{N}, \\ MI(U, V) = \sum_{i=1}^{C_U} \sum_{j=1}^{C_V} P(i, j) \cdot ln\left(\frac{P(i,j)}{P_U(i) \cdot P_V(j)}\right), \end{gathered} \tag{11}$$

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 [24]. В настоящей работе использовался скорректированный коэффициент, вычисляемый по формуле (12).

$$norm(U,\ V) = \ -\frac{1}{2}\left(\sum_{i=1}^{C_U} P_U(i)\ \cdot\ ln(P_U(i))\ +\ \sum_{i=1}^{C_V} P_V(i)\ \cdot\ ln(P_V(i))\right),$$
$$AMI(U,\ V) = \frac{MI(U,V) - E[MI(U,V)]}{norm(U,V) - E[MI(U,V)]}, \tag{12}$$

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

$P_U(i)$, $P_V(j)$, $MI(U,\ V)$ - вычисляемые по формуле (11) величины;

$E[MI(U,\ V)]$ - 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 [25] является мерой их подобия и вычисляется по формуле (13) как доля пар точек, отнесённых к одному или разным кластерам в обоих распределениях. Таким образом, значения, близкие к единице, указывают на высокое сходство результатов.

$$RI\ = \frac{SS + DS}{C_2^N}, \tag{13}$$

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 [26], лежащий в интервале $[-0.5, 1]$, вычисляемый по формуле (14).

$$ARI = \frac{RI - E[RI]}{1 - E[RI]}, \tag{14}$$

где RI - нескорректированный индекс Рэнда, вычисляемый по формуле (13);
$E[RI]$ - математическое ожидание индекса Рэнда.

Описанный алгоритм реализован функцией adjusted_rand_score библиотеки scikit-learn.

4.2.3. Индекс Фаулкса-Маллоуса

Индекс Фаулкса-Маллоуса [27] вычисляется по формуле (15) 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.

$$FMI = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}, \tag{15}$$

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.

5. ПРОМЕЖУТОЧНЫЕ РЕЗУЛЬТАТЫ

5.1. Выбор гиперпараметров

Подбор гиперпараметров в рамках промежуточного этапа проекта не производился. При запуске алгоритма преимущественно применялись стандартные значения, а также значения, использовавшиеся в [1], в соответствии с таблицей 1:

Таблица 1 - Значения гиперпараметров

Этап	Обозначение	Значение(я)	Описание
Получение эмбеддингов Такенса	d	10	Размерность пространства точек
	τ	1	Временная задержка между координатами точек
	s	1	Шаг между первыми координатами точек
Построение комплекса Вьеториса-Рипса	n	$\{0, 1, 2\}$	Размерности искомых симплексов
Фильтрация	ε	0.01	Минимальная допустимая разница между моментам появления и исчезновения симплекса
SDA: этап 1	$n\_clusters$	$[2, 20]$	Количество искомых состояний
	$k\_neighbours$	$[20, 50]$	Максимальное время между эпохами в матрице сходства
	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

5.2. Объект - 1

Для первого объекта производился поиск 9 функциональных состояний. При использовании топологических признаков алгоритм определил границы {0, 105, 252, 345, 490, 566, 670, 855, 975, 1046} против “правильных” {0, 39, 282, 492, 560, 682, 784, 857, 976, 1046}. Их сравнение представлено в таблице 2 и на рисунке 3.

Таблица 2 - Анализ полученного результата для объекта 1

	Расстояние Уорда	Центроид. расстояние	Коэфф. силуэта	Индекс Калински-Харабаса	Индекс Дэвиса-Болдина
“Правильный”ответ	24208	21.4	0.199	69.9	1.64
Полученный ответ	90	1.15	0.126	48.8	2.58
Коэфф. взаимной информации		0.821			
Индекс Рэнда		0.662			
Индекс Файлкса-Маллоуса		0.709			

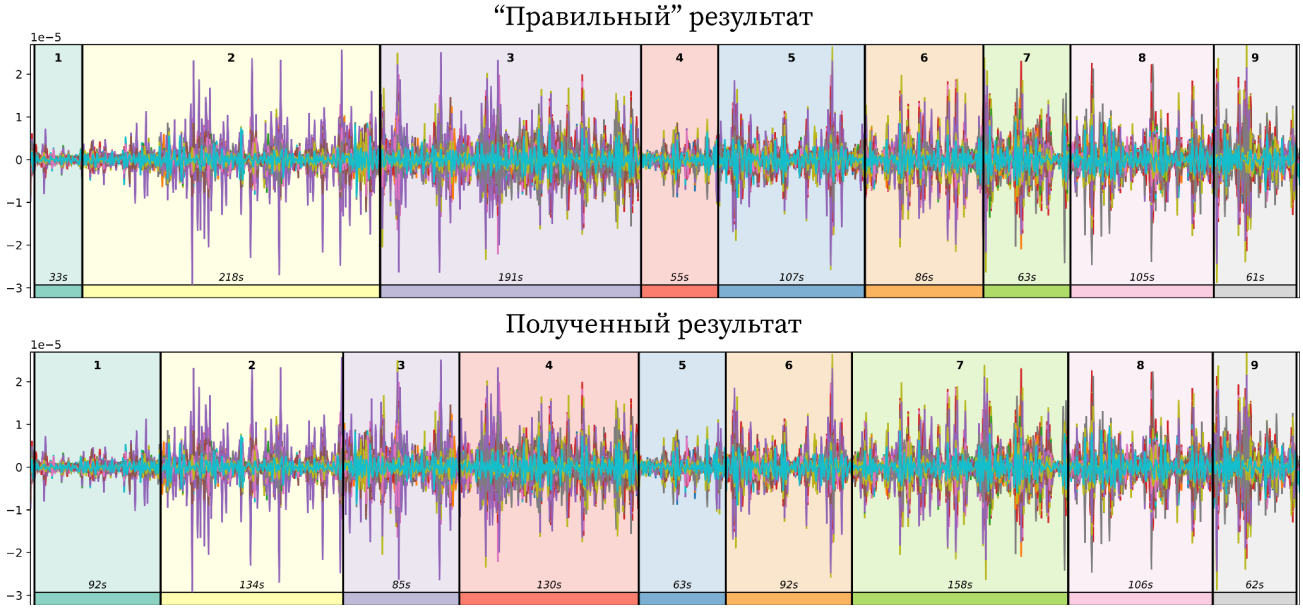

Рисунок 3 - Сравнение результатов для объекта 1

5.3. Объект - 2

Для второго объекта производился поиск 8 функциональных состояний. При использовании топологических признаков алгоритм определил границы {0, 420, 505, 722, 1069, 1272, 1463, 1723, 2017} против “правильных” {0, 370, 526, 728, 1052, 1275, 1489, 1857, 2017}. Их сравнение представлено в таблице 3 и на рисунке 4.

Таблица 3 - Анализ полученного результата для объекта 2

	Расстояние Уорда	Центроид. расстояние	Коэфф. силуэта	Индекс Калински-Харабаса	Индекс Дэвиса-Болдина
“Правильный”ответ	18009	12.4	0.102	59.9	2.94
Полученный ответ	103.6	0.96	0.067	40.1	3.87
Коэфф. взаимной информации	0.852				
Индекс Рэнда	0.782				
Индекс Файлкса-Маллоуса	0.813				

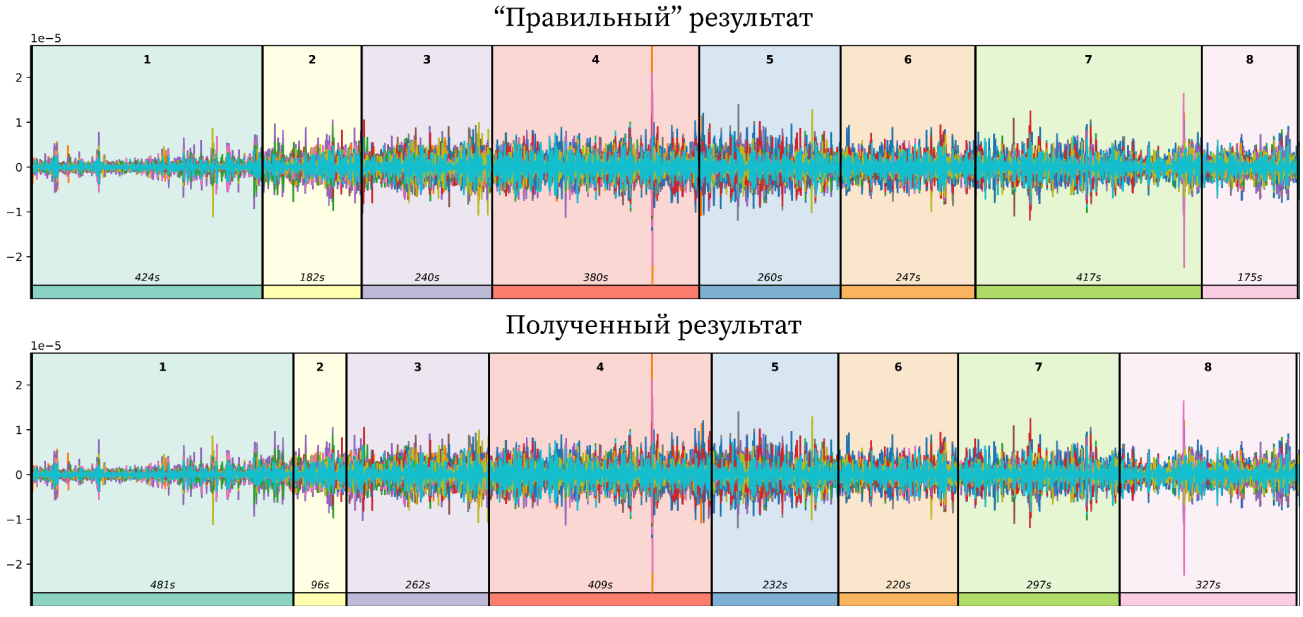

Рисунок 4 - Сравнение результатов для объекта 2

5.4. Объект - 3

Для третьего объекта производился поиск 10 функциональных состояний. При использовании топологических признаков алгоритм определил границы {0, 13, 199, 260, 498, 621, 735, 911, 1038, 1109, 1180} против “правильных” {0, 133, 175, 261, 458, 685, 783, 938, 1037, 1126, 1180}. Их сравнение представлено в таблице 4 и на рисунке 5.

Таблица 4 - Анализ полученного результата для объекта 3

	Расстояние Уорда	Центроид. расстояние	Коэфф. силуэта	Индекс Калински-Харабаса	Индекс Дэвиса-Болдина
“Правильный”ответ	9973	14.1	0.105	27.3	2.74
Полученный ответ	60.1	1.1	0.113	29.2	2.53
Коэфф. взаимной информации	0.784				
Индекс Рэнда	0.635				
Индекс Файлкса-Маллоуса	0.681				

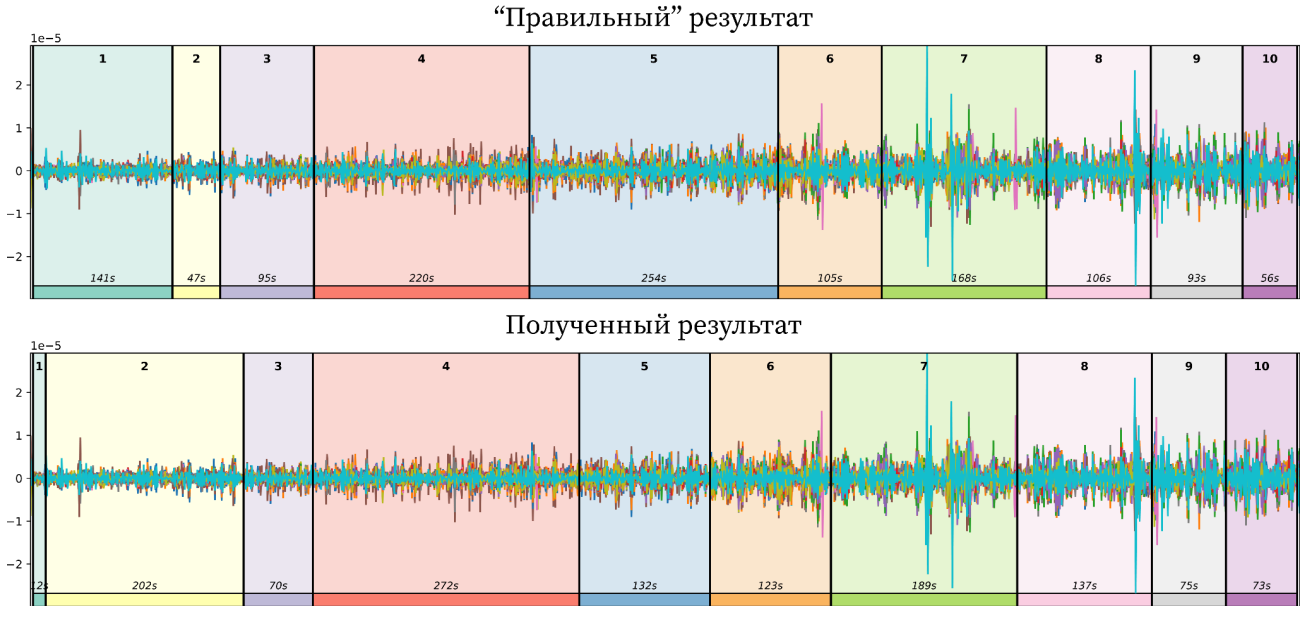

Рисунок 5 - Сравнение результатов для объекта 3

5.5. Вывод

Для всех трёх объектов явно прослеживаются сходства результатов, полученных с помощью топологических признаков, с результатами, полученными на основе традиционных признаков, что подтверждается относительно неплохими значениями внешних метрик. Хотя отдельные состояния, особенно в начале медитации, и не были найдены, многие границы были определены с неплохой точностью, учитывая, что подбор гиперпараметров для экспериментов не производился. Более того, значения внутренних метрик, не зависящих от абсолютных величин признаков (коэффициент силуэта, индексы Калински-Харабаса и Дэвиса-Болдина), для результатов на основе топологического подхода также уступают значениям для исходного результата лишь незначительно, а для объекта 3 и вовсе превосходят их.

ЗАКЛЮЧЕНИЕ

В рамках промежуточного этапа проекта были изучены и реализованы основные методы для решения задачи нахождения функциональных состояний по ЭЭГ непрерывных процессов с использованием подходов топологического анализа данных, а также были воспроизведены существующие результаты решения данной задачи. Как показал проведённый эксперимент на данных процесса медитации Guhyasamaja Tantra, даже без подбора гиперпараметров алгоритм способен находить границы состояний с неплохой точностью, особенно в конце исследуемого процесса. Такой результат позволяет предполагать, что топологические признаки действительно применимы при решении поставленной задачи и позволяют достичь точности, не уступающей или даже превосходящей таковую при использовании традиционных для работы с временными рядами признаков.

Дальнейшие исследования будут направлены на подбор гиперпараметров и улучшение алгоритмов признакового описания ЭЭГ для достижения более хорошего результата. При этом могут быть применены различные способы статистического анализа получаемых признаков, алгоритмы анализа их информационной ценности и метод главных компонент. При необходимости также может быть доработана реализация алгоритмов для повышения удобства подбора гиперпараметров с помощью кросс-валидации.

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. 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).
5. 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).
6. 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).
7. 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).
8. 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).

9. 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).
10. 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).
11. 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).
12. 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).
13. 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).
14. 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).
15. 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).
16. 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).
17. Steinhaus, H. Sur la division des corps materiels en parties / H. Steinhaus // Bulletin L'Académie Polonaise des Science - 1957 - vol. 4 - c. 801 - 804.

18. 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).
19. 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).
20. 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).
21. 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).
22. 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).
23. 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).
24. 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).
25. 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).

26. 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).
27. 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).