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

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК 515.1

Отчет об исследовательском проекте на тему:

Анализ топологических признаков в ЭЭГ-сигнале при изучении ASMR

Выполнил студент:

группы #БПМИ213, 3 курса Зиманов Темирхан

Принял руководитель проекта:

Кантонистова Елена Олеговна
Доцент, к.ф.-м.н.
Факультет компьютерных наук НИУ ВШЭ

Москва 2024

Содержание

Аннотация

В данной курсовой работе рассматривается анализ топологических признаков в ЭЭГ-сигналах при восприятии ASMR (Autonomous Sensory Meridian Response) с использованием методов топологического анализа данных (TDA) и теории устойчивых гомологий. Обнаружение специфических топологических признаков в этих сигналах может способствовать лучшему пониманию механизмов возникновения ASMR.

Ключевые слова

ASMR, ЭЭГ, топологический анализ данных, устойчивые гомологии, топологические признаки.

1 Введение

1.1 Некоторые топологические термины

В данном разделе представлены основные определения и понятия, лежащие в основе топологического анализа данных, необходимые для понимания дальнейшего изложения.

Топологическое пространство — это множество точек, для которого определены понятия близости и непрерывности с помощью системы открытых множеств. Формально, пара (X, τ), где X — множество, а τ — семейство подмножеств X, называется топологией, если выполнены следующие условия:

- Пустое множество и X принадлежат τ;
- Любое объединение множеств из τ принадлежит τ;
- Пересечение любого конечного числа множеств из τ принадлежит τ.

Гомеоморфизм — это непрерывное взаимно однозначное отображение между двумя топологическими пространствами, обратное к которому также непрерывно. Гомеоморфные пространства считаются топологически эквивалентными, поскольку они имеют одинаковую топологическую структуру.

Инварианты в топологии — это свойства топологических пространств, которые остаются неизменными при топологических преобразованиях, таких как гомеоморфизмы. Инварианты играют ключевую роль в классификации топологических пространств, позволяя установить их эквивалентность или различие. Они могут быть как числовыми (например, число компонент связности), так и более сложными структурами (например, группы гомологий).

Примеры инвариантов:

1. *Число компонент связности* — один из самых простых топологических инвариантов, указывающий на количество отдельных "кусков" , из которых состоит пространство. Например, два отдельных круга на плоскости имеют две компоненты связности, в то время как одиночный круг имеет одну компоненту связности.

2. *Эйлерова характеристика* — числовой инвариант, который можно вычислить для многих топологических пространств, включая поверхности. Для простых двумерных фигур эта характеристика вычисляется как $V - E + F$, где V — количество вершин, E — количество ребер, и F — количество граней. Например, для сферы (как и для тетраэдра, куба и других многогранников, гомеоморфных сфере) эйлерова характеристика равна 2.

3. *Фундаментальная группа* — инвариант, отражающий пути, которые можно проложить в пространстве, возвращаясь в исходную точку. Она показывает, какие петли в пространстве можно стянуть в точку (тривиальные), а какие нет. Например, для простого круга фундаментальная группа тривиальна, поскольку любую петлю можно стянуть в точку, в то время как для тора фундаментальная группа не тривиальна из-за наличия "дыр".

Инварианты позволяют классифицировать топологические пространства, определять их свойства и строить мосты между различными областями математики. Они являются мощным инструментом в топологическом анализе данных, позволяя выявлять фундаментальные структурные особенности в сложных наборах данных.

Гомотопия между двумя непрерывными функциями $f, g : X \to Y$, где X и Y являются топологическими пространствами, — это непрерывное отображение $H : X \times [0,1] \to Y$ такое, что для каждого $x \in X$ выполняются условия $H(x,0) = f(x)$ и $H(x,1) = g(x)$. Параметр $t \in [0,1]$ в функции $H(x,t)$ можно интерпретировать как "время" , при этом H показывает, как точка x "перемещается" из $f(x)$ в $g(x)$ по мере изменения t от 0 до 1. Если такое отображение H существует, функции f и g называются гомотопически эквивалентными или просто гомотопными.

Гомотопическая эквивалентность. Два топологических пространства X и Y называются гомотопически эквивалентными, если существуют непрерывные отображения $f : X \to Y$ и $g : Y \to X$ такие, что композиция $f \circ g$ гомотопна тождественному отображению id_Y на Y, и композиция $g \circ f$ гомотопна тождественному отображению id_X на X. Это означает, что каждое пространство может быть "деформировано" в другое с помощью этих отображений, и эта деформация обратима.

Гомотопическая эквивалентность является важным понятием в алгебраической топологии, поскольку она позволяет классифицировать топологические пространства с точки зрения их "формы" , не учитывая точную геометрию. Пространства, гомотопически эквивалентные точке, называются стягиваемыми и считаются "простейшими" в топологическом смысле.

Симплициальный комплекс определяется как множество V с выделенным набором его конечных подмножеств S, удовлетворяющих следующему условию: если $X \in S$ и $Y \subset X$, то $Y \in S$. При этом элементы множества V называются вершинами комплекса, а элементы множества S — его симплексами.

Симплициальные комплексы являются мощным инструментом в топологии и топологическом анализе данных, поскольку они позволяют моделировать пространства со сложной структурой с помощью относительно простых геометрических блоков.

Цепь, цикл, граница, гомология:

- **Цепь** в симплициальном комплексе определяется как формальная линейная комбинация его симплексов с коэффициентами из некоторого поля, обычно вещественных или целых чисел. Для симплициального комплекса K, цепь C может быть представлена как $C = \sum a_i \sigma_i$, где a_i — коэффициенты, а σ_i — симплексы комплекса K.

- **Граница симплекса** σ определяется через его вершины. Если $\sigma = [v_0, v_1, \ldots, v_k]$ является k-симплексом, то его граница $\partial\sigma$ вычисляется как $\partial\sigma = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$, где $\hat{v}_i$ означает, что вершина v_i исключается из списка.

- **Цикл** — это такая цепь Z, что её граница равна нулю, т.е., $\partial Z = 0$. Это означает, что сумма границ всех симплексов в цикле взаимно уничтожается, не оставляя "края".

- **Гомология** — это метод классификации циклов симплициального комплекса по их "дырам" , с точностью до границ. Гомологические классы определяются как эквивалентные классы циклов, где два цикла считаются эквивалентными, если их разность является границей некоторой другой цепи в комплексе. Группы гомологий $H_k(K)$ для симплициального комплекса K собирают все такие классы эквивалентности для циклов размерности k.

Устойчивые гомологии являются развитием идеи гомологий в контексте топологического анализа данных. Этот подход позволяет анализировать топологическую структуру данных, учитывая изменения этой структуры на различных масштабах.

Для определения устойчивых гомологий, в первую очередь, необходимо ввести понятие **фильтрации**. Фильтрация комплекса — это последовательность вложенных друг в друга симплициальных комплексов $K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K$, которая начинается с пустого комплекса и завершается полным комплексом. Каждый комплекс K_i в этой последовательности представляет собой структуру данных на определённом "уровне разрешения где каждый уровень добавляет новые симплексы или связи.

Устойчивые гомологии рассматриваются как последовательности гомологических групп этих комплексов, выявляя, какие топологические особенности (например, дыры различной размерности) сохраняются по мере развития фильтрации. Важным аспектом устойчивых гомологий является их способность квантифицировать устойчивость этих особенностей: некоторые "дыры" могут появляться и исчезать при небольшом изменении масштаба, в то время как другие остаются "устойчивыми" на протяжении значительных интервалов фильтрации.

Для визуализации устойчивости топологических особенностей используется **диаграмма устойчивости** или **диаграмма персистенции**. Эта диаграмма отображает, на каком уровне фильтрации каждая особенность появляется и когда она исчезает. Оси диаграммы обычно представляют собой уровни фильтрации; точка на диаграмме (b, d) означает, что некоторая особенность появляется на уровне b и исчезает на уровне d. Чем длиннее интервал между b и d, тем более "устойчивой" считается соответствующая топологическая особенность.

Математическое выражение устойчивых гомологий может быть описано следующим образом: для каждого k гомологической размерности, строится последовательность гомологических групп $H_k(K_0), H_k(K_1), \ldots, H_k(K_n)$. Гомологические изменения от одного комплекса к другому отслеживаются с помощью гомоморфизмов, которые индуцированы вложениями $K_i \hookrightarrow K_{i+1}$. Устойчивые гомологии фокусируются на изучении ядра и образа этих гомоморфизмов для идентификации и классификации устойчивых особенностей в данных.

Таким образом, устойчивые гомологии предоставляют мощный инструмент для анализа сложных наборов данных, позволяя выявлять и количественно оценивать структурные и топологические особенности данных на различных уровнях детализации.

1.2 Топологический анализ данных

Топологический анализ данных (TDA) — это метод, применяемый для изучения формы данных на основе их топологических свойств. Согласно учебному пособию [5], процесс TDA можно разбить на несколько ключевых этапов:

- **Преобразование данных в облако точек:** Первым шагом в TDA является представление исходных данных в виде облака точек в многомерном пространстве. Это позволяет использовать геометрическую и топологическую информацию о данных.

- **Конструкция комплекса Вьеториса — Рипса:** Далее, на основе облака точек строится симплициальный комплекс Вьеториса — Рипса. Этот метод включает выбор параметра масштабирования ϵ, который определяет, какие точки будут соединены. Если расстояние между двумя точками меньше ϵ, они соединяются ребром. Набор точек, все попарные расстояния между которыми меньше ϵ, образует симплекс. Таким образом, комплекс Вьеториса — Рипса представляет собой коллекцию симплексов, которые кодируют информацию о пространственных отношениях между точками в облаке.

- **Вычисление устойчивых гомологий:** Затем, используя полученный комплекс, вы-

числяются гомологии — топологические инварианты, которые описывают структуры, такие как дыры различной размерности в данных. Устойчивые гомологии изучаются с помощью так называемых диаграмм устойчивости, которые показывают, как топологические особенности данных (например, циклы или пустоты) появляются и исчезают при изменении параметра ϵ.

- **Анализ и интерпретация:** Финальный этап заключается в анализе и интерпретации полученных топологических инвариантов.

Таким образом, TDA предоставляет уникальный инструмент для изучения нелинейных и сложных структур в данных, которые трудно выявить с помощью традиционных статистических методов.

1.3 Электроэнцефалография

Электроэнцефалография (ЭЭГ) представляет собой неинвазивный метод исследования, который позволяет регистрировать электрическую активность мозга. Этот метод основан на измерении напряжений, возникающих в результате работы нейронов мозга, с помощью электродов, размещаемых на поверхности кожи головы. ЭЭГ позволяет получить ценную информацию о функциональном состоянии мозга, его реакции на различные стимулы и в процессе выполнения разнообразных задач.

ЭЭГ широко используется в клинической практике и исследованиях благодаря своей способности предоставлять динамическую картину электрической активности мозга во времени. Это делает ЭЭГ ценным инструментом для изучения состояний сознания, сна, эпилепсии, а также для исследования нейронных основ поведения и восприятия. В контексте исследований ASMR, ЭЭГ предоставляет уникальную возможность наблюдать изменения в мозговой активности, вызванные специфическими стимулами, что может способствовать лучшему пониманию механизмов этого явления.

1.4 Автономная сенсорная меридиональная реакция

Автономная сенсорная меридиональная реакция (Autonomous Sensory Meridian Response, ASMR) является явлением, при котором определённые звуковые, визуальные или тактильные стимулы вызывают приятные ощущения покалывания или мурашки, распространяющиеся по коже головы, шеи и других частей тела. Хотя ASMR стал объектом повышенного

интереса как для широкой общественности, так и для научного сообщества в последние годы, механизмы, лежащие в основе этого явления, до конца не изучены.

Исследования ASMR с использованием электроэнцефалографии (ЭЭГ) представляют собой новое и перспективное направление. ЭЭГ позволяет наблюдать изменения в мозговой активности в реальном времени, предоставляя уникальную возможность изучить, как различные ASMR стимулы воздействуют на разные области мозга. Исследования показали, что ASMR может влиять на амплитуду и частоту мозговых волн [4], что указывает на потенциальное изменение состояния релаксации и внимания у индивидов.

1.5 Цель проекта

Целью данной работы является анализ топологических признаков в ЭЭГ-сигналах при восприятии ASMR с использованием методов топологического анализа данных и устойчивых гомологий. Конкретной задачей является разработка классификатора, который сможет определять наличие ASMR-реакции у человека на основе анализа двухсекундных фрагментов ЭЭГ.

2 Обзор литературы

Существующие исследования подчеркивают значимость и многообещающий потенциал применения TDA и ЭЭГ для глубокого понимания ASMR и его влияния на мозговую активность. Однако детальное изучение топологических признаков в ЭЭГ-сигналах при восприятии ASMR остается малоизученным направлением (авторы не нашли подобных исследований в интернете), что подчеркивает новизну и актуальность данного исследования.

3 Обработка данных

3.1 Анализ имеющейся размеченной выборки фрагментов ЭЭГ

Мы использовали данные, представленные в статье [4]. В рамках исследования 25 участников просматривали видео, вызывающие ASMR, и сообщали о своих ощущениях, которые варьировались от отсутствия изменений (Baseline) до повышенной релаксации (Relaxed) и самого ASMR. Для каждого участника было собрано от 300 до 500 размеченных двухсекундных ЭЭГ-фрагментов с использованием 64 электродов при частоте съема 512 Гц. Суммарно получилось 10759 фрагментов. На Рисунке 3.1 представлены примеры двухсекундных

фрагментов, записанных первым электродом, демонстрирующие типичные сигналы, полученные в ходе эксперимента.

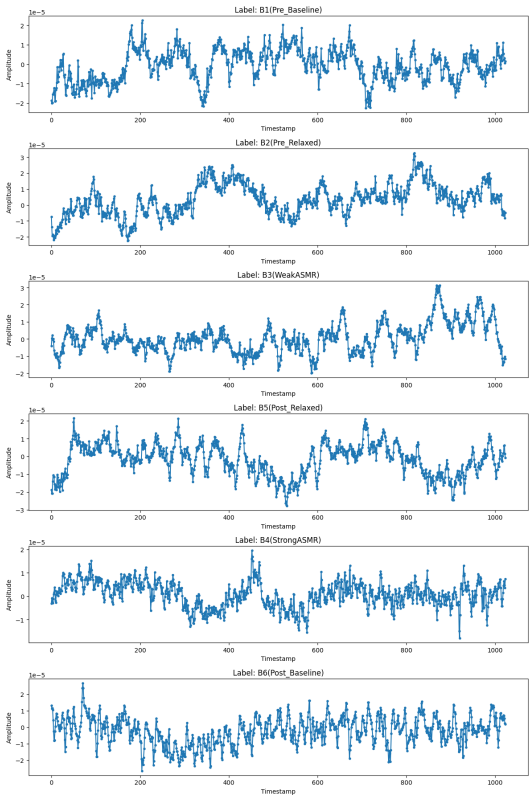

Рис. 3.1: Примеры двухсекундных фрагментов, записанных первым электродом.

Распределение меток по участникам показано на Рисунке 3.2, демонстрирующем неравномерность распределения как между различными метками, так и среди участников.

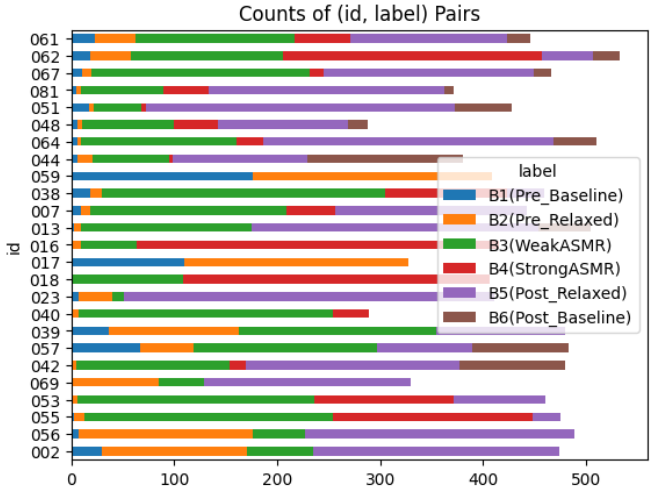

Рис. 3.2: Распределение меток по участникам.

Для упрощения анализа данных мы решили различать только две категории состояний: наличие ASMR-реакции (WeakASMR/StrongASMR) и её отсутствие (все остальные

метки). Распределение этих упрощенных меток среди участников иллюстрируется на Рисунке 3.3.

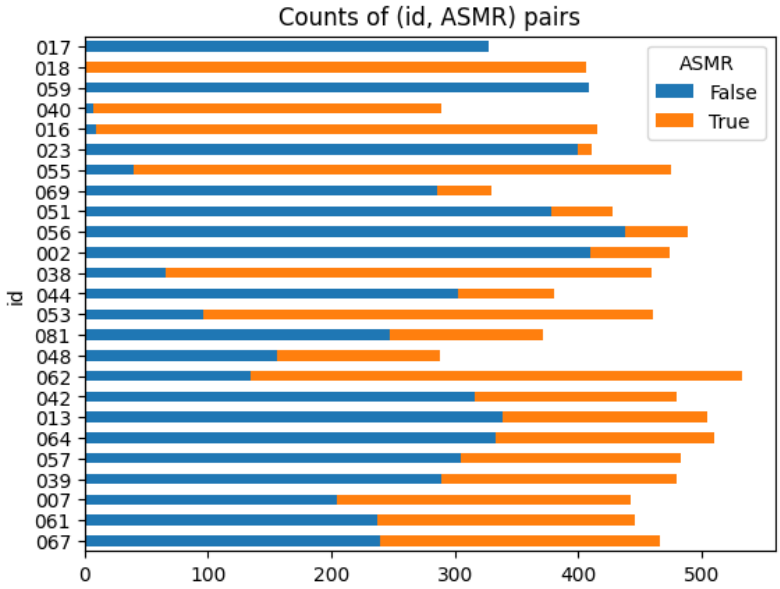

Рис. 3.3: Распределение упрощенных меток по участникам.

3.2 Преобразование ЭЭГ сигналов в облака точек

Для преобразования ЭЭГ сигналов в облака точек мы использовали вложения Такенса. Вложения Такенса — это методика, предназначенная для реконструкции динамического пространства состояний из временных рядов данных. Этот метод широко применяется в областях анализа временных рядов и топологического анализа данных, особенно когда исходные данные о структуре фазового пространства не доступны.

Вложения Такенса работают на принципе задержек, создавая векторное пространство из временного ряда путём взятия отсчетов с фиксированными временными интервалами. Каждый вектор в новом пространстве составляется из значений временного ряда, взятых через равные промежутки времени, что позволяет воссоздать динамику исходной системы.

Для временного ряда $x(t)$, вложение Такенса определяется как множество векторов вида:

$$y(t) = (x(t), x(t+\tau), x(t+2\tau), \ldots, x(t+(d-1)\tau))$$

где:

- τ — временная задержка,
- d — размерность вложения.

Временная задержка τ и размерность вложения d являются ключевыми параметрами, которые необходимо настроить. Их выбор влияет на качество реконструкции динамики системы:

- τ должно быть достаточным, чтобы компоненты вектора не были чрезмерно коррелированными, но и не слишком великим, чтобы избежать потери важной информации о динамике.
- d выбирается достаточно большим, чтобы обеспечить, что реконструированное пространство имеет вложенное измерение, которое способно улавливать всю динамику системы.

Для определения оптимальных параметров размерности и задержки времени при вложении Такенса ЭЭГ сигналов, был разработан эксперимент, использующий следующий методический подход:

1 Исходные ЭЭГ данные считываются и обрабатываются для извлечения двухсекундных интервалов с помощью библиотеки MNE-Python [2]. Эти интервалы сгруппированы по идентификаторам участников и меткам состояния. Из каждой группы случайно выбираются по два интервала с фиксированным сидом, чтобы обеспечить воспроизводимость эксперимента.

2 Каждый интервал ассоциируется с конкретным каналом ЭЭГ. Для этого используется равномерное распределение интервалов по всем каналам, доступным в данных.

3 Для каждого интервала и канала применяется вложение Такенса с помощью класса `SingleTakensEmbedding` библиотеки Giotto-tda [1]. Эксперименты проводятся с максимальными значениями размерности вложения до 10 и временной задержкой до 90, с шагом в один тик данных. Вложение настраивается так, чтобы исследовать пространство возможных параметров (`parameters_type="search"`).

4 Для каждого интервала вычисляются оптимальные значения размерности и временной задержки, на которых вложение достигает наилучших результатов. Эти значения регистрируются вместе с идентификатором участника, меткой состояния и каналом ЭЭГ.

Такой подход позволяет систематически оценить, как различные параметры вложения Такенса влияют на качество представления ЭЭГ данных в топологическом пространстве. На

Рисунке 3.4 демонстрируются 10 наиболее часто встречающихся пар (размерность, временная задержка) среди результатов эксперимента.

	dimension	time_delay	Count
0	6	15	178
1	6	12	173
2	6	20	158
3	6	18	155
4	6	22	152
5	6	9	152
6	6	19	148
7	6	11	148
8	6	8	145
9	6	10	144

Рис. 3.4: 10 наиболее часто встречающихся пар подбираемых параметров.

На основании результатов эксперимента было решено установить размерность вложения на уровне 6 и временную задержку на уровне 15. Рисунок 3.5 демонстрирует примеры облаков точек, созданных из ЭЭГ-фрагментов первого канала. На левой части изображения показано облако точек для случая отсутствия ASMR, в то время как на правой части — для случая присутствия ASMR.

Рис. 3.5: Примеры облаков точек, полученных из фрагментов ЭЭГ сигналов для первого канала. Слева ASMR отсутствует, справа — присутствует.

Тем не менее, как указывается в статье [3], алгоритмы вычисления устойчивых гомологий требуют значительных вычислительных ресурсов, работая за время, кубически зависящее от количества симплексов. Чтобы сократить время вычислений, мы увеличили шаг вложения Такенса с 1 до 10. Это позволило существенно уменьшить размерность облаков точек. Примеры более маленьких облаков точек представлены на Рисунке 3.6, где также

показано различие между случаями отсутствия и наличия ASMR.

Рис. 3.6: Примеры уменьшенных облаков точек с шагом 10. Слева ASMR отсутствует, справа — присутствует.

Таким образом, мы преобразовали исходные ЭЭГ сигналы в облака точек, тем самым подготовив их к TDA.

4 Извлечение топологических признаков

Извлечение топологических признаков включает в себя несколько шагов, начиная с вычисления устойчивых гомологий облаков точек и заканчивая вычислением и анализом топологических инвариантов.

4.1 Конструкция симплициальных комплексов

Первым шагом в извлечении топологических признаков является построение симплициальных комплексов из облаков точек, полученных из ЭЭГ сигналов. Для этого применяется метод Вьеториса — Рипса, который позволяет создавать симплексы на основе евклидова расстояния между точками облака. Параметр масштаба в этом методе определяет, как близко должны находиться точки друг к другу, чтобы быть соединенными в симплекс.

4.2 Вычисление устойчивых гомологий

После построения симплициальных комплексов следующим этапом является вычисление устойчивых гомологий. Этот процесс позволяет идентифицировать и классифицировать топологические особенности (например, дыры различной размерности), которые сохраняются на разных уровнях фильтрации данных. Устойчивые гомологии предоставляют ценную информацию о том, какие структурные особенности данных являются значимыми. На практике вычисляются гомологии размерностей 0, 1 и 2.

4.3 Визуализация диаграмм устойчивости

Для визуализации полученных топологических данных используются диаграммы устойчивости. Эти диаграммы представляют собой графическое изображение, на котором каждая точка отражает моменты появления и исчезновения топологических особенностей при изменении параметра фильтрации. Диаграммы устойчивости позволяют легко оценить, какие особенности являются устойчивыми и значимыми, а какие — временными или шумовыми. На Рисунке 4.1 отображены по 4 диаграммы устойчивости для каждой из 6 меток.

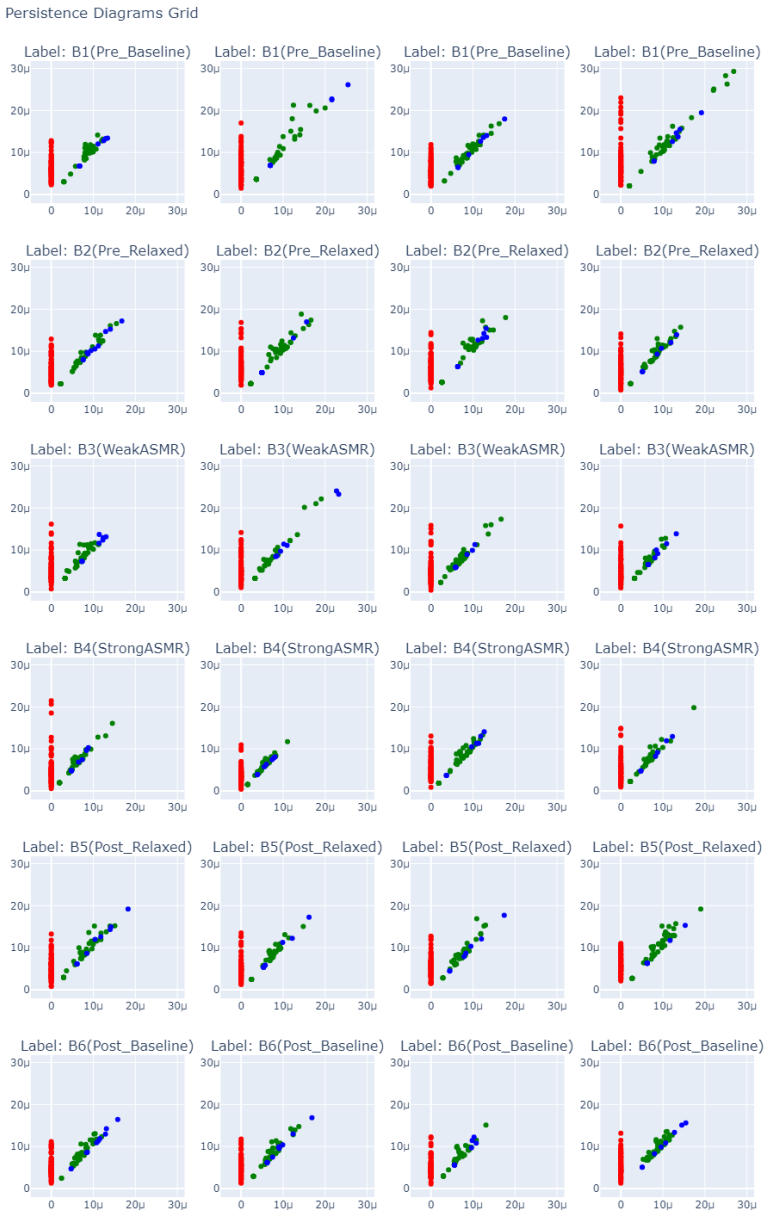

Рис. 4.1: Диаграммы устойчивости для четырех случайных облаков точек, соответствующих первому каналу, для каждой из шести имеющихся меток. Ось абсцисс соответствует времени рождения, а ось ординат — времени смерти гомологий. Точки разных цветов соответствуют гомологиям разных размерностей (красные — 0, зеленые — 1, синие — 2).

4.4 Квантификация топологических признаков

Для использования топологических данных в дальнейшем анализе необходимо квантифицировать извлеченные особенности. Это достигается за счет преобразования диаграмм устойчивости в векторные признаки, которые могут быть использованы в машинном обучении и статистическом анализе. Примеры таких признаков включают амплитуды устойчивых гомологий, энтропии персистенций и числа Бетти, которые отражают количество особенностей каждой размерности.

Эти признаки предоставляют конкретные числовые значения для каждой рассматриваемой размерности, которые отражают топологическую сложность и структурную уникальность исследуемых ЭЭГ сигналов. Используя эти данные, можно попробовать разрабатывать алгоритмы для классификации состояний ASMR, а также для других прикладных исследований в нейронауке.

4.4.1 Амплитуды устойчивых гомологий

Амплитуды устойчивых гомологий являются одним из ключевых числовых показателей в топологическом анализе данных. Они измеряют "вес"или "размер"каждой топологической особенности, улавливаемой на диаграмме устойчивости. Каждая точка на такой диаграмме представляет собой особенность, а амплитуда этой точки описывает диапазон масштабов, на которых данная особенность наблюдается. Большие значения амплитуд указывают на более значимые и устойчивые особенности в структуре данных. Эти значения могут быть использованы для оценки схожести или различия топологических структур в разных наборах данных.

4.4.2 Энтропии устойчивостей

Энтропия устойчивости — это мера, описывающая сложность и разнообразие топологических особенностей, сохраняющихся на различных масштабах. Она вычисляется на основе распределения длительностей жизни топологических особенностей, отраженных на диаграммах устойчивости. Высокие значения энтропии указывают на наличие большого количества особенностей с различной продолжительностью жизни, что может свидетельствовать о сложной топологической структуре данных. Эта метрика полезна для сравнения структурных и динамических различий между разными группами данных.

4.4.3 Числа Бетти

Числа Бетти представляют собой один из фундаментальных топологических инвариантов, которые используются для описания количества особенностей в различных измерениях топологических пространств. Например, число Бетти B_0 описывает количество компонент связности, B_1 — количество "дыр" или петель, а B_2 и выше — количество полостей и других многомерных "дыр". Числа Бетти могут быть рассчитаны на основе диаграмм устойчивости и предоставляют важную информацию о топологической устойчивости и изменчивости данных.

4.5 Применение в исследовании

В итоге, объектом в нашей задаче классификации будет вектор, состоящий из идентификатора человека, названия канала (электрода), и девяти числовых признаков: по одному из описанных выше признаков на каждую размерность (0, 1 и 2). Первые два признака и числа Бетти принимают строковые или целочисленные значения. К ним мы применили `OneHotEncoder` из библиотеки scikit-learn. Остальные признаки принимают действительные значения. К этим признакам мы применили `StandardScaler`.

5 Разбиение данных на обучающую и валидационную выборки

5.1 Подход к разбиению

Для обеспечения эффективности и надежности результатов, получаемых моделью машинного обучения, крайне важно правильно разбить исходные данные на обучающую и валидационную выборки. Особенное внимание следует уделить тому, что данные асимметричны как по меткам, так и по участникам, что показано на Рисунке 3.3. Это означает, что количество данных, относящихся к различным меткам и участникам, неодинаково, что может привести к смещению в модели.

5.2 Стратегия выбора

Для минимизации смещения в данных и повышения обобщающей способности модели была принята стратегия сбалансированного выбора участников и интервалов по каждой

метке. Основным заданием данного подхода является обеспечение равномерного распределения каждой метки и представительства каждого участника в обучающей и валидационной выборках.

Мы определили два ключевых параметра, которые играют важную роль в этой стратегии:

- **Number of People (количество участников)**: Этот параметр определяет, сколько участников будет включено в исследование. Выбор определенного количества участников позволяет нам управлять масштабом данных и гарантировать, что данные будут взяты из разнообразной выборки популяции.

- **Intervals per Person per Class (количество интервалов на каждую метку у каждого участника)**: Этот параметр обеспечивает, что для каждого участника и для каждой метки будет выбрано одинаковое количество интервалов. Такой подход позволяет сбалансировать влияние каждой метки и каждого участника на обучение и валидацию модели, что критически важно для предотвращения переобучения на доминирующих классах или данных отдельных участников.

Важно отметить, что выборки для обучения и валидации формируются таким образом, чтобы быть не только сбалансированными по классам и участникам, но и иметь одинаковый размер. Это обеспечивает равные условия для оценки модели на различных этапах обучения и валидации, улучшая тем самым надежность оценки её эффективности.

Эта стратегия позволяет систематически подходить к выбору и разделению данных, учитывая их асимметричность и разнообразие. Таким образом, модель будет тренироваться и проверяться на более репрезентативной выборке, что способствует повышению точности и надежности получаемых результатов.

6 Результаты экспериментов

6.1 Влияние увеличения размера обучающей выборки

В одном из экспериментов мы исследовали влияние увеличения размера обучающей выборки на эффективность классификации с использованием решающего дерева. Результаты показаны на Рисунке 6.1. Из таблицы видно, что с увеличением объема обучающей выборки классификатор не перестает переобучаться, в то время как показатели на валидационной

выборке остаются на уровне ROC AUC = 0.5, что свидетельствует об отсутствии улучшения обобщающей способности модели.

	Number of People	Intervals per Person per Class	Training Samples	Classifier	Training ROC AUC	Validation ROC AUC
0	1	10	1280	DecisionTree	1.0	0.514062
1	1	20	2560	DecisionTree	1.0	0.511328
2	1	30	3840	DecisionTree	1.0	0.511198
3	1	40	5120	DecisionTree	1.0	0.509766
4	1	50	6400	DecisionTree	1.0	0.496406
5	1	100	12800	DecisionTree	1.0	0.497500
6	2	100	25600	DecisionTree	1.0	0.508867
7	3	100	38400	DecisionTree	1.0	0.502526
8	10	50	64000	DecisionTree	1.0	0.505141

Рис. 6.1: Влияние размера обучающей выборки на эффективность решающего дерева. Переобучение модели видно по результатам на тренировочной выборке. На валидационной выборке решающее дерево работает не лучше случайного гадания.

6.2 Сравнение классификаторов

В другом эксперименте было проведено сравнение шести различных классификаторов на обучающих выборках трех разных размеров. Результаты представлены на Рисунке 6.2. Наблюдается различное поведение классификаторов: некоторые из них показывают признаки переобучения, в то время как другие сохраняют стабильность. Несмотря на это, результаты на валидационной выборке для всех классификаторов остаются низкими с ROC AUC = 0.5, что указывает на сложности в данных или потенциальные недостатки в моделях, не позволяющие эффективно различать классы.

Результаты, полученные в ходе экспериментов, указывают на необходимость дальнейшего анализа для улучшения обобщающей способности моделей, возможно, через более продвинутую предобработку данных, выбор признаков или использование нейронных сетей.

7 Заключение

В рамках данного курсового проекта мы столкнулись с рядом проблем при попытке обучить эффективный классификатор для определения реакции ASMR по ЭЭГ-сигналам. Несмотря на использование различных подходов и моделей, результаты остались на уровне

	Number of People	Intervals per Person per Class	Training Samples	Classifier	Training ROC AUC	Validation ROC AUC
0	1	100	12800	DecisionTree	1.000000	0.497500
1	1	100	12800	LogisticRegression	0.554624	0.499197
2	1	100	12800	KNeighbors	0.750994	0.502715
3	1	100	12800	RandomForest	1.000000	0.504728
4	1	100	12800	GradientBoosting	0.665015	0.499542
5	1	100	12800	AdaBoost	0.534707	0.495745
6	3	100	38400	DecisionTree	1.000000	0.502526
7	3	100	38400	LogisticRegression	0.534488	0.505452
8	3	100	38400	KNeighbors	0.751235	0.509321
9	3	100	38400	RandomForest	1.000000	0.512389
10	3	100	38400	GradientBoosting	0.596488	0.511487
11	3	100	38400	AdaBoost	0.527861	0.502229
12	10	50	64000	DecisionTree	1.000000	0.505141
13	10	50	64000	LogisticRegression	0.532609	0.489024
14	10	50	64000	KNeighbors	0.753570	0.504133
15	10	50	64000	RandomForest	1.000000	0.511272
16	10	50	64000	GradientBoosting	0.586270	0.511031
17	10	50	64000	AdaBoost	0.530619	0.493143

Рис. 6.2: Сравнение различных классификаторов на обучающих выборках разного размера. Все модели показывают схожие результаты на валидационной выборке, что свидетельствует об общей трудности задачи классификации в данном контексте.

случайного угадывания, что указывает на возможные причины, требующие дополнительного анализа.

7.1 Возможные причины неудачи

- Одной из ключевых проблем является субъективность в разметке данных участниками. У каждого участника могут быть собственные представления о том, что такое ASMR, что приводит к неоднородности меток и затрудняет обучение классификатора. Возможно, потребуется более строгий контроль за процессом разметки или использование профессионально подготовленных оценщиков.

- Использование топологических признаков, таких как числа Бетти, может приводить к потере значимой информации. Возможно, стоит рассмотреть работу непосредственно с диаграммами устойчивости или использование более комплексных признаков.

7.2 Предложения по улучшению

- Объединение информации со всех 64 каналов ЭЭГ в один объект. Интеграция данных со всех доступных каналов ЭЭГ в единую структуру данных может значительно улучшить точность моделирования. Каждый канал собирает уникальную информацию о различных регионах мозга. Объединение этой информации позволит создать более полное представление о мозговой активности в ответ на ASMR стимулы. Это, в свою очередь, может улучшить обучение классификаторов, позволяя им учитывать комплексные взаимосвязи между различными частями мозга.

- Применение нейронных сетей непосредственно к диаграммам устойчивости. Использование сверточных нейронных сетей (CNN), которые могут адаптировать и выявлять сложные паттерны в таких данных, потенциально может улучшить классификацию.

Проект подчеркивает сложность задачи классификации физиологических состояний, таких как ASMR, основываясь на ЭЭГ-данных. Неудача в достижении высокой точности классификации подсказывает, что для успеха в этом направлении потребуется глубокий мультидисциплинарный подход, объединяющий нейронауку, машинное обучение и, возможно, психологию. Также это подчеркивает значимость тщательной подготовки данных и выбора методов их анализа в зависимости от специфики задачи.

Список литературы

[1] *Giotto-ai Giotto TDA*. Библиотека для топологического анализа данных. URL: https://giotto-ai.github.io/gtda-docs/0.5.1/library.html (дата обр. 09.02.2024).

[2] *MNE: MEG and EEG software*. Инструментарий для анализа данных магнито- и электроэнцефалографии. URL: https://mne.tools/stable/index.html (дата обр. 09.02.2024).

[3] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod и Heather A. Harrington. "A roadmap for the computation of persistent homology". В: *EPJ Data Science* 6 (1 2017), с. 17. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-017-0109-5.

[4] Thomas R. Swart, Michael J. Banissy, Thomas P. Hein, Ricardo Bruña, Ernesto Pereda и Joydeep Bhattacharya. "ASMR amplifies low frequency and reduces high frequency oscillations". В: *Cortex* 149 (2022), с. 85—100. ISSN: 0010-9452. DOI: https://doi.org/10.1016/j.cortex.2022.01.004.

[5] Антон Айзенберг. *Методичка по симплициальным комплексам и гомологиям*. Курс по топологическому анализу данных. URL: https://github.com/raxtemur/TDA_course_at_FCS/blob/main/Supplementary/%D0%90%D0%B9%D0%B7%D0%B5%D0%BD%D0%B1%D0%B5%D1%80%D0%B3_%D0%9C%D0%B5%D1%82%D0%BE%D0%B4%D0%B8%D1%87%D0%BA%D0%B0_%D0%BF%D0%BE_%D0%B3%D0%BE%D0%BC%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F%D0%BC.pdf (дата обр. 09.02.2024).