МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра дискретной математики и информационных технологий

УМЕНЬШЕНИЕ РАЗМЕРНОСТИ ВРЕМЕННЫХ РЯДОВ ЭЭГ ДЛЯ ВЫЯВЛЕНИЯ ЗАВИСИМОСТЕЙ С ПОМОЩЬЮ UMAP

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Ешмакова Артема алексеевича

Научный руководитель
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2025

ВВЕДЕНИЕ

Электроэнцефалография является одним из ключевых методов изучения функциональной активности головного мозга, широко применяемым для диагностики неврологических и психоневрологических расстройств. Высокая размерность и сложность ЭЭГ-данных создают значительные трудности при их анализе, что требует применения современных методов обработки, таких как снижение размерности. В последние годы алгоритм UMAP зарекомендовал себя как эффективный инструмент для анализа сложных многомерных данных, обеспечивая сохранение как локальной, так и глобальной структуры.

Актуальность данной работы обусловлена необходимостью эффективной работы с ЭЭГ-сигналами и их анализом, особенно в контексте диагностики психоневрологических заболеваний.

Целью исследования является оценка алгоритма UMAP для выявления значимых различий в ЭЭГ-данных между различными группами пациентов, а также какие подходы к предварительной обработке способствуют улучшению результатов.

Задачи:

- Изучить методы анализа временных рядов и ЭЭГ-сигналов.
- Познакомиться с работой метода UMAP
- Разработать и применить алгоритм UMAP для обработки предварительно подготовленных ЭЭГ-данных.
- Провести сравнительный анализ результатов классификации с использованием UMAP и без него для различных датчиков и наборов данных
- Оценить влияние UMAP на точность классификации психоневрологических состояний на основе ЭЭГ-данных.

Материалы исследования Методологические основы исследования временных рядов ЭЭГ и методов снижения размерности представлены в работах: теоретические аспекты анализа временных рядов рассмотрены в трудах Д.Б. Егорова и С.Д. Захарова [1], где особое внимание уделено проблемам нестационарности и методам спектрального анализа. Физиологические основы ЭЭГ и современные подходы к обработке сигналов изложены в исследованиях M. Cohen [2], а также в работах M. Teplan [3], посвящённых стандартизации измерений и устранению артефактов. Методы снижения размерности, включая UMAP, детально проанализированы L. McInnes и J. Healy [4, 5], где представлены ма-

тематические основы алгоритма и сравнение с t-SNE. Применение машинного обучения для анализа ЭЭГ освещено в работах Y. Zhang [6] и N.D. Truong [7], где продемонстрирована эффективность нейросетевых моделей (CNN, LSTM) для классификации паттернов.

В работе использованы ЭЭГ-данные, полученные с помощью 19-канальной системы ЭЭГ (частота дискретизации 500 Гц), включающие записи от пациентов с различными психоневрологическими расстройствами (болезнь Альцгеймера, депрессия, шизофрения, когнитивное расстройство) и контрольной группы.

Структура работы

Магистерская диссертация состоит из четырёх глав:

1. Методы анализа временных рядов – обзор классических и современных подходов.
2. ЭЭГ и методы её анализа – физиологические основы, артефакты, предобработка.
3. Методы снижения размерности – сравнение PCA, t-SNE и UMAP.
4. Применение UMAP для анализа ЭЭГ – обработка данных, понижение размерности визуализация, классификация.

Научная новизна Проведена комплексная оценка применения алгоритма UMAP для снижения размерности ЭЭГ-данных. Разработан подход к предобработке данных, включающий фильтрацию и извлечение признаков, оптимизированный для использования с UMAP. Установлено влияние различных параметров UMAP на точность классификации, а также проведено сравнение эффективности классификации до и после применения UMAP на полном наборе данных и данных отдельных датчиков.

Научная и практическая значимость

Результаты работы вносят вклад в развитие методов анализа ЭЭГ-сигналов, предлагая эффективный подход к снижению размерности с сохранением диагностически значимой информации. Применение UMAP с использованием меток позволяет выявлять скрытые паттерны в ЭЭГ-данных, одноко без использования меток алгоритм не способен выявить закономерности в данных. Полученные выводы могут быть использованы для улучшения диагностических моделей в клинической практике, а также для дальнейших исследований в области нейроинформатики и машинного обучения.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе «Методы анализа временных рядов» представлен обзор методов анализа временных рядов, включая классические и современные подходы, применяемые для обработки данных. Рассмотрены основные характеристики временных рядов, их типы, а также ключевые требования к их корректному использованию: сопоставимость уровней данных, полнота наборов данных, высокая точность измерений и учет периодических изменений.

Описаны основные проблемы, возникающие при анализе временных рядов:

- Нестационарность, когда статистические свойства ряда меняются со временем, что усложняет прогнозирование.
- Ошибки прогноза, связанные с различиями между реальными и прогнозируемыми значениями.
- Выбор модели, зависящий от задачи, объема данных и вычислительных ресурсов.

Для решения этих проблем применяются методы предварительной обработки, такие как тесты на стационарность, разности и декомпозиция рядов, что особенно важно для ЭЭГ-данных, демонстрирующих нестационарность из-за физиологических процессов.

В работе рассмотрены следующие методы анализа временных рядов: экспоненциальное сглаживание (взвешенное среднее с экспоненциально убывающими весами), регрессионный анализ (для выявления линейных и нелинейных трендов), автокорреляционная функция (выявление сезонности и паттернов), ARIMA/SARIMA (статистические модели для стационарных/нестационарных рядов), скользящее среднее (простое выявление трендов), спектральный анализ (преобразование Фурье для выделения ритмов мозга), вейвлет-анализ (одновременный анализ во временной и частотной областях), LSTM-сети (анализ временных зависимостей), кластеризация (K-means, DBSCAN), методы снижения размерности (PCA, t-SNE, UMAP), классификация временных рядов, извлечение признаков (статистических и частотных характеристик), динамическое временное выравнивание (DTW) для сравнения паттернов, а также функции авто- и кросскорреляции для оценки ритмичности и функциональной связности в ЭЭГ-данных.

Временные ряды являются мощным инструментом анализа данных в био-

медицине, включая ЭЭГ. Классические методы остаются актуальными за счет интерпретируемости, тогда как современные подходы позволяют эффективно обрабатывать сложные и высокоразмерные данные. Выбор метода зависит от свойств данных и целей исследования, что подчеркивает важность комплексного подхода к анализу ЭЭГ-сигналов.

Второй раздел «ЭЭГ и методы её анализа» посвящен изучению ЭЭГ сигналов, особенностям и методам обработки. ЭЭГ-сигналы представляют собой записи электрической активности коры головного мозга, получаемые с помощью электродов, размещенных на поверхности головы. Основные характеристики сигналов включают низкую амплитуду, частотный диапазон и наличие специфических ритмов: дельта (0,5–4 Гц), тета (4–8 Гц), альфа (8–13 Гц), бета (13–30 Гц) и гамма (>30 Гц). ЭЭГ-сигналы обладают высокой чувствительностью к внешним и внутренним помехам, что делает их обработку сложной задачей.

Ключевые проблемы анализа ЭЭГ-сигналов:

- Шумы и артефакты: физиологические (движения глаз, мышечные сокращения, ЭКГ) и технические (помехи от сети).
- Нестационарность: изменение статистических свойств сигнала во времени из-за динамики физиологических процессов.
- Высокая размерность: многоканальные записи (до 40 каналов) создают объем данных, усложняя их интерпретацию.

Для решения этих проблем применяются методы предобработки:

1. Фильтрация: использование полосовых фильтров (0,5–40 Гц) для удаления частотных шумов, а также вейвлет-фильтрации для сохранения временной структуры сигнала.
2. Независимый компонентный анализ: выделение независимых источников сигналов для устранения артефактов, таких как глазные движения или сердечные сокращения.
3. Нормализация и стандартизация: приведение данных к единому масштабу для обеспечения совместимости при анализе.
4. Извлечение признаков: вычисление характеристик, включая:
    - Статистические: среднее, дисперсия, асимметрия, эксцесс.
    - Частотные: мощность ритмов, доминантная частота,
    - спектральная энтропия.
    - Нелинейные: энтропия Шеннона, корреляционная размерность, по-

казатель Херста.

Рассмотрены основные методы анализа ЭЭГ-сигналов:

- Классические подходы: анализ мощности ритмов, когерентности между каналами, автокорреляции для выявления паттернов, связанных с физиологическими состояниями.
- Машинное обучение: использование алгоритмов, таких как SVM, Random Forest, CatBoost и k-NN, для классификации психоневрологических состояний, включая болезнь Альцгеймера, депрессию, шизофрению и легкое когнитивное расстройство.
- Нейросетевые методы: применение сверточных нейронных сетей для анализа пространственных паттернов и рекуррентных сетей для учета временных зависимостей в ЭЭГ-данных.
- Анализ функциональной связности: изучение корреляций и фазовой синхронизации между каналами для оценки взаимодействия регионов мозга.

В третьем разделе «Методы понижения размерности» представлен обзор методов снижения размерности, применяемых для анализа высокоразмерных данных, таких как ЭЭГ-сигналы. Снижение размерности необходимо для упрощения данных, сохранения их структуры, улучшения визуализации и повышения эффективности алгоритмов машинного обучения. Рассмотрены основные проблемы высокоразмерных данных: вычислительная сложность, эффект «проклятия размерности» и сложность интерпретации.

Описаны следующие методы снижения размерности:

1. Анализ главных компонент (PCA): Линейный метод, преобразующий данные в новое пространство с помощью ортогональных компонент, максимизирующих дисперсию. Преимущества: простота, интерпретируемость, высокая скорость. Недостатки: ограниченная способность улавливать нелинейные зависимости, потеря локальной структуры данных.
2. t-SNE (t-distributed Stochastic Neighbor Embedding): Нелинейный метод, оптимизирующий сохранение локальной структуры данных для визуализации. Преимущества: высокое качество визуализации кластеров. Недостатки: высокая вычислительная сложность, чувствительность к параметрам (perplexity), слабое сохранение глобальной структуры.
3. UMAP (Uniform Manifold Approximation and Projection): Нелинейный метод, основанный на топологических принципах, сохраняющий как ло-

кальную, так и глобальную структуру данных. Преимущества: высокая скорость, гибкость настройки (параметры n_neighbors, min_dist), применимость к большим наборам данных. Недостатки: необходимость тщательной настройки гиперпараметров, сложность интерпретации.

Методы снижения размерности, такие как PCA, t-SNE и UMAP, решают проблему обработки высокоразмерных данных, улучшая визуализацию и классификацию. UMAP выделяется благодаря скорости, гибкости и способности сохранять как локальную, так и глобальную структуру данных.

Четвертый раздел «Особенности ЭЭГ-сигналов и методы их обработки» посвящен практическому применению алгоритма UMAP для снижения размерности ЭЭГ данных, их анализа, классификации психоневрологических состояний и визуализации результатов. Раздел включает детальное описание подготовки данных, настройки гиперпараметров UMAP, попарного анализа состояний, сравнения результатов классификации с использованием различных алгоритмов машинного обучения, а также оценки влияния UMAP на качество диагностики.

Для исследования зависимостей между ЭЭГ-сигналами и психоневрологическими заболеваниями использовался датасет, содержащий данные ЭЭГ пациентов с депрессией (28 человек), шизофренией (42 человека), когнитивными нарушениями (34 пациента), болезнью Альцгеймера (49 пациентов) и контрольной группой(96 пациенов). Данные были собраны в период с 2010 по 2019 годы в Медицинском центре Рабина в Израиле.

Для исследования использовались два типа датасетов: исходные сырые сигналы ЭЭГ и набор признаков, включающий спектральные характеристики, корреляции между датчиками и статистические метрики. Предобработка данных включала:

1. Удаление артефактов и выбросов. Артефакты представлены искажением сигнала, которые не связаны с работой мозга, например, движения глаз или напряжение мышц, выбросы - участки сигнала, который превышает нормальный диапазон значений. Мы использовали модифицированный Z-score метод, устойчивый к асимметричным распределениям. Пороговое значение threshold=3.5 было выбрано на основе анализа чувствительности к амплитудным артефактам (например, движения глаз).
2. Нормализация сигналов. Применен RobustScaler, который масштабирует

данные по межквартильному диапазону (5–95 перцентили), что минимизирует влияние выбросов:

3. Фильтрация частот. Для выделения значимых ритмов ЭЭГ использован FIR-фильтр с плавными переходными зонами:
4. Коррекция артефактов методом независимых компонент ICA.
5. Сглаживание сигналов с помощью фильтра Савицкого-Голея. Реализует сохранение фазовых характеристик, эффективное подавление высокочастотного шума и минимальное искажение полезных компонент

Сравнение датасетов показало, что обработанные данные обеспечивают более устойчивые результаты.

Алгоритм UMAP применялся в двух режимах: с обучением на метках заболеваний и без них. В первом случае наблюдалось четкое разделение кластеров, соответствующих разным диагнозам, что подтверждает способность UMAP сохранять классовые структуры при наличии информации о болезнях. Например, для пар «депрессия – контроль» и «шизофрения – болезнь Альцгеймера» проекции демонстрировали минимальное перекрытие кластеров. Во втором случае (без меток) разделение было менее выраженным, что свидетельствует о высокой вариабельности ЭЭГ-сигналов внутри групп.

Особый интерес представлял попарный анализ заболеваний по отдельным датчикам. Несмотря на то, что ни один из сенсоров не обеспечивал абсолютной разделимости классов, некоторые каналы показывали точность классификации до 81.6%. Но при этом другие - около 38%.

Для оценки практической пользы UMAP проведена серия экспериментов с классификацией. Использовались модели CatBoost, Random Forest и SVM на двух типах данных: исходных признаках и признаках, преобразованных UMAP в пространство сниженной размерности (n_components=5). Результаты показали неоднозначный эффект:

- В 40% случаев применение UMAP улучшало точность
- В 41.6% случаев наблюдалось ухудшение
- Наилучшие результаты достигнуты при классификации на полном наборе признаков всех датчиков (точность CatBoost: 52%), однако снижение размерности ухудшило этот показатель на 16%.

Дополнительная проверка на стороннем наборе данныхподтвердила ограничения UMAP: без обчения с метками или дополнительной обработки класте-

ры заболеваний значительно перекрывались. Это подчеркивает необходимость интеграции UMAP с более сложными методами и нейросетевыми архитектурами.

Эксперименты показали, что при использовании информации о диагнозах UMAP позволяет эффективно выявлять скрытые структуры и зависимости в данных. Но при использовании обучения без меток umap показывал плохие результаты, что означает, что алгоритм не способен найти зависимотси внутри данных электроэнцефалографии. В целом, алгоритм UMAP является эффективным инструментом для визуализации сложных многомерных биомедицинских данных, но только при использовании информации о метках болезней

ЗАКЛЮЧЕНИЕ

В данной магистерской работе произведена задача понижения размерности временных рядов ЭЭГ с целью вывления скрытых структур и различий между разными группами психоневрологических заболеваний. Была проведеная работа по изучению методов анализа временных рядов, в особенности ЭЭГ данных. Были проанализированы различные методы анализа временных рядов. Основное внимание было уделено алгоритму UMAP, показавшему высокую эффективность в задачах визуализации многомерных биомедицинских данных.

В работе был разработан алгоритм предобработки ЭЭГ-данных, включающий методы фильтрации, удаления артефактов и нормализации, что позволило улучшить качество последующего анализа. При этом применение UMAP может как улучшить, так и ухудшить точность в зависимости от конкретной пары заболеваний и используемых признаков. Было проведено сравнение работы UMAP в режиме обучения с метками и без меток. Эксперименты показали, что из-за высокой вариативности ЭЭГ, определённые паттерны позволяют различать состояния пациентов только при использовании информации о диагнозах. Однако без учёта меток классов структура данных оказывается существенно менее выраженной.

Проведённое исследование вносит вклад в развитие методов компьютерной обработки ЭЭГ и открывает новые возможности для создания интеллектуальных систем поддержки медицинских решений. Результаты работы подчёркивают важность комплексного подхода, сочетающего современные алгоритмы машинного обучения с глубоким пониманием нейрофизиологических основ сигналов ЭЭГ. Перспективными направления являются интеграция UMAP с глубоким обучением для более эффективного выделения диагностически значимых признаков и разработка интерактивных инструментов визуализации на основе UMAP для использования в клинической практике.

Основные источники информации:

1 Егоров, Д. Б. Современные методы анализа и прогнозирования временных рядов и их применение в медицине [Текст] / Егоров, Д. Б., Захаров, С. Д. и Егорова, А. О. // Врач и информационные технологии. — 2020. — № 1. — С. 21–26

2 Cohen, M. X. Where does EEG come from and what does it mean? [Text] // Trends in neurosciences. — 2017. —Vol. 40, no. 4.—P. 208–218.

3 Teplan, Michal. Fundamentals of EEG measurement [Text] / Teplan, Michal [et al.] // Measurement science review. — 2002. — Vol. 2, no. 2. —P. 1–11.

4 McInnes, Leland. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Text]. — 2020. —1802.03426.

5 McInnes, Leland. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Text] / McInnes, Leland and Healy, John // ArXiv. — 2018. — Vol. abs/1802.03426.

6 Ahybrid deep learning approach for emotion classification using EEG signals linked to PTSD[Text]/Zhang,Y., Wang,Y., Wang,L., Wang,Y., andWang,H.// Frontiers in Computational Neuroscience. — 2022. — Vol. 16. —P. 1019776.

7 A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals [Text] / Truong, N. D., Nguyen, A. D., Kuhlmann, L., Bonyadi, M. R., Yang, J., Ippolito, S., and Kavehei, O. // Computer Methods and Programs in Biomedicine. —2018.—Vol. 161.—P. 1–13.