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

ГЕНЕРАЦИЯ ДЕНДРОГРАММ ПО ЗАДАННЫМ ВРЕМЕННЫМ РЯДАМ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Плющева Игоря Андреевича

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

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

Саратов 2025

ВВЕДЕНИЕ

В настоящее время для изучения свойств сложных систем, получить математическое описание которых крайне затруднительно или невозможно, широко используется подход, заключающийся в исследовании таких систем посредством анализа производимых этой системой сигналов. Последовательность измерений таких сигналов, регистрируемых непрерывно или через некоторые интервалы времени называют наблюдаемой, или временным рядом. Его примерами являются, в частности, запись электрокардиограммы, запись колебаний земной коры, данные метеонаблюдений, данные стоимости акций, данные численности населения и т.п. Изучением данного математического объекта занимается подраздел теории динамических систем - анализ временных рядов.

Традиционно в анализе временных рядов выделяют две общие задачи:

1. Задача идентификации заключается в определении параметров породившей временной ряд динамической системы. Под параметрами понимают к примеру: корреляционную размерность, энтропию, размерность вложения и т.п.
2. Задача прогнозирования заключается в предсказании по имеющимся наблюдениям будущих характеристик динамической системы.

На данный момент для решения обеих задач разработан целый пласт математических методов и моделей, однако, даже несмотря на это анализ временных рядов является нетривиальной задачей, так как требуется выбрать методы и модели, а также подобрать адекватные данной ситуации параметры этих методов, в зависимости от природы и свойств системы породившей временной ряд, и к тому же не всегда приводящей к желаемому результату. Вместе с тем, в данное время получает развитие направление исследований, целью которых является разработка методов решения задач идентификации и прогнозирования для любого временного ряда в независимости от природы исследуемого явления. Одним из примеров исследований такого направления являются исследования С.Ф. Тимашева и его группы, используемый данной группой метод называется фликкер-шумовой спектроскопией.

Фликкер-шумовая спектроскопия - общий феноменологический подход к извлечению информации, содержащейся в сложных сигналах. Сущность ФШС-подхода состоит в придании информационной значимости корреля-

ционным взаимосвязям, которые реализуются в последовательности нерегулярностей сигналов — всплесков, скачков, изломов производных различных порядков как носителей информации на каждом пространственном уровне иерархической организации исследуемой эволюции. Данный метод анализа применяется во множестве исследований, объектами которых являются самые различные по природе системы, например: поиск электрических предвестников землетрясений, диагностирование функционального состояния сердечно-сосудистой системы по электрокардиографии, определение фоточувствительной эпилепсии и нейродегенеративных заболеваний по магнитоэнцефалографии.

Еще одним представителем исследований, пытающимся решить схожие задачи, является одна израильская группа ученых, их исследование «p-adic quantum potential» представляет собой приложение разработанного ими метода анализа временных рядов к классификации психических расстройств, а именно: на основании только лишь ЭЭГ записей у людей выявляется наличие или отсутствие когнитивных или психических расстройств. В этой работе каждой ЭЭГ записи сопоставляется показатель, называемый квантовым потенциалом, на основе значения которого можно осуществлять дифференциацию пациентов по различным группам.

Главными особенностями этого исследования являются:

— Во-первых, высокая, со слов исследователей, точность классификации.
— Во-вторых, применение нестандартных, с точки зрения анализа временных рядов методов, а именно методов из теории p-адического анализа и квантовой механики (методов принадлежащих двум различным математическим мирам).

Метод, используемый для вычисления так называемого квантового потенциала, содержит ряд последовательных математических преобразований исходных временных рядов в различные математические объекты. Ключевым, с теоретической точки зрения, объектом в данном ряде преобразований является дендрограмма, получаемая посредством агломеративной кластеризации. Под дендрограммой понимают бинарное дерево, описывающее вложенную группировку объектов, которая изменяется на различных уровнях иерархии, таким образом она показывает степень близости (схожести) отдельных объектов или кластеров, а также наглядно демонстрирует в гра-

фическом виде последовательность их объединения. В обсуждаемой работе ветви дендрограммы представляют собой события (Bohr's phenomena), а сама дендрограмма иерархические взаимосвязи между событиями.

Таким образом, обсуждаемое исследование, в силу оговоренных выше особенностей обуславливает **актуальность** новых исследований в заданном им направлении, так как по сути оно решает актуальные на данный момент задачи, в частности диагностирования психических и когнитивных расстройств, а в общем анализа временных рядов. Данная бакалаврская работа и является одним из новых исследований в этом направлении, посвященная промежуточному этапу метода p-адического квантового потенциала генерации дендрограмм.

Целью данной работы является генерация дендрограмм по заданным временным рядам, а также поиск оптимальных параметров генерации.

Заданная цель поставила следующие **задачи:**

1. Изучение теоретических основ кластерного анализа, в особенности тех, что связаны с методами иерархической кластеризации.
2. Поиск и выбор подходящего для последующего исследования набора данных.
3. Предварительная подготовка набора данных.
4. Разработка программы для создания дендрограмм на основании некоторого набора временных рядов.
5. Генерация дендрограмм по подготовленным временным рядам.
6. Сравнительный анализ различных методов создания дендрограмм.

Данная работа включает в себя введение, заключение, список использованных источников, 3 приложения и 2 главы:

1. Главу «теоретическая часть», содержащую 2 раздела:
   - Основы кластерного анализа;
   - Иерархическая кластеризация.
2. Главу «практическая часть», содержащую 4 раздела:
   - Поиск и выбор набора данных;
   - Подготовка набора данных;
   - Описание метода генерации дендрограмм по заданным временным рядам и его реализация;
   - Генерация дендрограмм и их анализ.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе приводятся основные теоретические положения кластерного анализа, а именно:

1. Вводятся основные понятия кластерного анализа: кластерный анализ, кластер, задача кластеризации, алгоритм кластеризации, центроид кластера, дисперсия кластера, радиус кластера.
2. Приводится теорема Клейнберга о невозможности универсального алгоритма кластеризации.
3. Рассматриваются общие цели кластеризации.
4. Приводятся примеры сфер применения кластерного анализа.
5. Обсуждается неоднозначность задачи и алгоритмов кластеризации.
6. Приводятся классификации алгоритмов кластерного анализа по разным критериям.
7. Описываются основные принципы работы нескольких популярных алгоритмов кластеризации.

Во втором разделе приводятся основные теоретические аспекты иерархической кластеризации:

1. Вводятся базовые понятия: иерархии, иерархической кластеризации, дендрограммы.
2. Приводится классификация алгоритмов иерархической кластеризации.
3. Приводятся примеры сфер применения иерархического кластерного анализа.
4. Описывается общая схема агломеративной кластеризации.
5. Рассматриваются используемые в агломеративной кластеризации методы нормализации, меры близости объектов и меры близости кластеров.
6. Приводятся применимые к иерархическим алгоритмам методы оценки качества кластеризации.
7. Рассматриваются преимущества и недостатки иерархической кластеризации.

В третьем разделе описывается процесс поиска и выбора набора данных.

В четвертом разделе описывается подготовка набора данных для последующего использования, а именно:

1. Первичный обзор набора данных.

2. Перезапись данных из файлов формата «HDF5» в форматы: «csv» и «fif».
3. Первичный анализ рассматриваемых данных.

В пятом разделе приводятся описания метода генерации дендрограмм и реализации данного метода.

В шестом разделе приводится описание генерации дендрограмм и сравнительного анализа дендрограмм посредством следующих подходов:

— Применение коэффициента кофенетической корреляцию.
— Применение коэффициента силуэта.
— Применение ручного графического анализа.
— Введение различных мер несходства между дендрограммами.
— Сравнение дендрограмм полученных при разных конфигурациях.

Также в данном разделе представлен общий анализ результатов, полученных в ходе сравнительного анализа.

ЗАКЛЮЧЕНИЕ

В ходе данной работы были изучены основы кластерного анализа и методов иерархической кластеризации. Помимо этого были произведены поиск и выбор НД, его обзор, подготовка и первичный анализ, была реализована программа для генерации дендрограмм по временным рядам, с использованием которой была произведена генерация дендрограмм по подготовленному НД, полученные дендрограммы были проанализированы с использованием некоторых методов оценки качества кластеризации. В результате были найдены оптимальные конфигурации генерации дендрограмм. Также была выявлено подтверждение зависимости результирующей дендрограммы от группы здоровья пациентов.

Основные источники информации:

[1] EEG p-adic quantum potential accurately identifies depression, schizophrenia and cognitive decline [Электронный ресурс].
URL: https://www.researchgate.net/publication/353728460_EEG_p-adic_quantum_potential_accurately_identifies_depression_schizophrenia_and_cognitive_decline
(дата обращения: 3.04.2025).

[2] B.S. Everitt. Cluster Analysis. — United Kingdom: John Wiley And Sons, 2011.

[3] Ш.У. Низаметдинов. Анализ данных. — Москва: МИФИ, 2006.

[4] Анализ временных рядов. Курс лекций [Электронный ресурс].
URL: https://chaos.phys.msu.ru/loskutov/PDF/Lectures_time_series_analysis.pdf
(дата обращения: 07.02.2025).

[5] The comparison of dendrograms by objective methods [Электронный ресурс].
URL: https://www.researchgate.net/publication/232128980_Sokal_RR_Rohlf_FJ_The_comparison_of_dendrograms_by_objective_methods_Taxon_11_33-40
(дата обращения: 25.02.2025).