МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

АНАЛИЗ СОЦИОЛОГИЧЕСКИХ ДАННЫХ С ПОМОЩЬЮ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ В ВИДЕ ДЕНДРОГРАММ

(автореферат бакалаврской работы)

Студента 4 курса 451 группы
направления 09.03.03 – «Прикладная информатика»
профиль «Прикладная информатика в социологии»
Социологического факультета
Вайцуля Александра Николаевича

Научный руководитель
канд.физ.-мат.наук, доцент _____ Л. Б. Тяпаев
подпись, дата

Зав. кафедрой
канд.социол.наук, доцент _____ И. Г. Малинский
подпись, дата

Саратов 2023

ВВЕДЕНИЕ

Актуальность темы. Анализ данных в настоящее время представляется сложной задачей, связанной с неограниченным и постоянным ростом их количества. Согласно статистическим прогнозам на период с 2020 по 2025 год автора Petroc Taylor "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025"[1], объем создаваемых, собираемых, копируемых и потребляемых данных во всем мире стремительно увеличивается. В 2020 году объем достиг отметки в 64,2 зеттабайта, и по прогнозам, в следующие пять лет, до 2025 года, он вырастет более чем на 180 зеттабайт. Увеличение объема данных в 2020 году оказалось выше ожидаемого, причиной этому стал возросший спрос в потреблении информации, связанный с пандемией COVID-19. Большее количество людей работало и училось дома, а также активнее пользовалось домашними развлечениями. Объем хранилища также растет, но лишь небольшой процент вновь созданных данных сохраняется. Всего два процента данных, произведенных и потребленных в 2020 году, будут сохранены в 2021 году. В связи с сильным ростом объема данных, база емкости хранилища будет постепенно увеличиваться, с ежегодными темпами роста на 19,2 процента в период с 2020 по 2025 год. В 2020 году базовая емкость хранилища составила 6,7 зеттабайта. С учетом стремительного роста данных, традиционные инструменты уже не могут удовлетворить потребности в их обработке и хранении. Для работы с большими данными применяются различные методы понижения размерности, включая как пространство объектов, так и пространство признаков. Для сокращения размерности пространства объектов и выявления соответствующих структур применяется кластерный анализ. Кластерный анализ представляет собой широкий набор методов и алгоритмов, используемых для группировки данных. Один из таких методов – иерархический анализ.

[1] Taylor, P. "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025". Statista. [Электронный ресурс]: URL: https://www.statista.com/statistics/871513/worldwide-data-created/ (дата обращения 11.06. 2023). Загл. с экрана. Яз. англ.

Степень разработанности проблемы. Методы кластеризации используются в качестве статистического инструментария в самых разнообразных научных направлениях. К примеру, многомерная классификация данных находит широкое применение в медицинских исследованиях и психологии. Так, В. А. Альбахели в исследовании "Сегментация магнитно-резонансных изображений на основе кластерного анализа"[2] проводит кластерный анализ работы медицинской техники с целью повышения качества диагностики заболеваний с помощью МРТ. В работе В. П. Пономарева и И.Ю. "Белоглазовой исследование показателей крови больных" проведено на основе кластерного и факторного видов анализа. Применение кластерного анализа для обработки данных психологических исследования показано в работе Савченко Т. Н. "Применение методов кластерного анализа для обработки данных психологических исследований"[3]. Автор рассматривает теоретико-методические, а также прикладные вопросы применения этого вида анализа, предлагает варианты развития методики классификации и пути совершенствования алгоритмов анализа данных, реализуемых в современных пакетах прикладных программ. В социально-экономических исследованиях сфера применения многомерной классификации данных также широка. Так, кластерный анализ часто применяется в маркетинговых исследованиях. С его помощью чаще всего проводят сегментацию рынка. Однако есть авторы, которые применяют этот вид анализа в других направлениях маркетинговой деятельности. К примеру, О.Н. Ломидзе в работе "Кластерный анализ в социологических исследованиях"[4] показывает возможности кластеризации для выбора наиболее эффективного способа управления персоналом. В работе

[2] Альбахели В.А. "Сегментация магнитно-резонансных изображений на основе кластерного анализа" [Электронный ресурс]: URL: https://www.elibrary.ru/item.asp?id=24332502 (дата обращения 22.12.2022). Загл. с экрана. Яз. рус.

[3] Савченко Т.Н., Применение методов кластерного анализа для обработки данных психологических исследований [Электронный ресурс]: URL: https://lib.ipran.ru/upload/papers/15101448.pdf (дата обращения 22.12.2022). Загл. с экрана. Яз. рус.

[4] Ломидзе О.Н. Кластерный анализ в социологических исследованиях // Ученые записки Российского государственного социального университета. 2011. № 9 (97). Ч. 1. С. 38-42.

Колышкина Т. Б. "Восприятие концепта “красота” различными группами участников рекламной коммуникации"[5] показаны особенности восприятия населением концепта “красота” в рекламной продукции. Чаще всего кластерный анализ применяется для сегментации территориальных образований по набору социально-экономических индикаторов. К примеру, в исследовании Дегтярева "Исследование дифференциального социального развития сельских территорий"[6] проводится сегментация районов Оренбургской области на основе показателей, характеризующих уровень социального развития сельских территорий. Авторы И.Л. Фрумина и Е.В. Цветкова методом кластерного анализа исследуют проблемы аграрной экономики и проводят типологизацию муниципальных районов Челябинской области[7]. В исследовании Богорсукова Н. Я., Халафяна А. А. "Применение кластерного при изучении динамики численности населения районов Краснодарского Края"[8] авторы применяют кластерный анализ для оценки динамики численности населения Краснодарского края, выделяя группы районов со схожими характеристиками протекания демографических процессов.

Иерархическая агломеративная кластеризация часто применяется в качестве промежуточного этапа исследования. В работах Анашина В. С., Тяпаева Л. Б., Давыдова В. В. "Классификация психических заболеваний на основе дендрограмм ЭЭГ головного мозга и их характеристик"[9] и Oded Shor, Amir Glik,

---

[5] Колышкина Т.Б., Шустина И.В. Восприятие концепта «красота» различными группами участников рекламной коммуникации // Вестник Томского государственного университета. Филология. 2015. № 6 (38). С. 46-54.

[6] Дегтярева Т.Д., Чулкова Е.А. и Торбина Е.С., Исследование дифференциации социального развития сельских территорий // Известия Оренбургского государственного аграрного университета. 2015. № 5. С. 212-216.

[7] Фрумина И.Л. и Цветкова Е.В. Исследование некоторых проблем аграрной экономики методом кластерного анализа // Известия Челябинского научного центра УРО РАН. 2007. № 4. С. 93-97

[8] Богорсукова Н.Я., Халафян А.А., Ракачев В.Н. Применение кластерного анализа при изучении динамики численности населения районов Краснодарского края // Вестник Северокавказского федерального университета. 2014. №2 (41). С. 142-146.

[9] Анашин В.С., Тяпаев Л.Б., Давыдов В.В. "Классификация психических заболеваний на основе дендрограмм ЭЭГ головного мозга и их характеристик". // Труды XIV международного научного семинара «Дискретная математика и ее приложения» имени академика О.Б.

Amit Yaniv-Rosenfeld, Avi Valevski, Abraham Weizman, Andrei Khrennikov, Felix Benninger "EEG p-adic quantum potential accurately identifies depression, schizophrenia and cognitive decline"[10] иерархическая кластеризация используется для построения дендрограмм, которые затем рассматриваются как бинарные деревья. Это связано с тем, что каждая ветвь дендрограммы кодируется 2-адическими числами. Дальнейший анализ основан на различных подходах. В обеих работах анализ строится на основе построенных дендрограмм, однако в одной из них применяется теория кодирования с использованием префиксных кодов, а в другой – p-адический квантовый потенциал. Подробнее эти исследования представлены в главе 3.

Объект исследования: массивы данных.

Предмет исследования: кластеризация разнородных данных с помощью иерархического кластерного анализа.

Цель исследования: демонстрация работы алгоритма иерархической агломеративной кластеризации на разных наборах данных.

Задачи:

1) Реализовать несколько алгоритмов кластеризации с целью выбора оптимального для данного набора данных;

2) Оценка каждого алгоритма кластеризации, используя коэффициенты качества;

3) Выбор оптимального алгоритма с дальнейшей интерпретацией каждого из кластеров;

4) Демонстрация алгоритма кластеризации в виде промежуточного этапа анализа временных рядов ЭЭГ мозга.

Лупанова (20-25 июня 2022 г., Москва). М.: ИПМ им. М.В.Келдыша РАН, 2022, С.207–210. URL: https://doi.org/10.20948/dms-2022-64

[10] Shor O., Glik A., Yaniv-Rosenfeld A., Valevski A., Weizman A., Khrennikov A., и др. "EEG p-adic quantum potential accurately identifies depression, schizophrenia and cognitive decline". PLoS ONE, 16(8): e0255529, 2021. URL: https://doi.org/10.1371/journal.pone.0255529

Структура выпускной квалификационной работы представлена введением, тремя главами, заключением, списком использованных источников и приложением.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе "Методы анализа данных" вводятся все теоретические основы, которые используются во втором разделе. К ним относится: формальная задача кластеризации, описание несколько алгоритмов кластеризации (K-means, спектральная кластеризация, алгомеративная иерархическая кластеризация), методы анализа качества модели.

Во второй главе "Сегментация клиентов рынка" производится предобработка данных и реализуются 3 алгоритма кластеризации: агломеративная иерархическая кластеризация, K-means, Spectral Clustering. Опираясь на результаты визуализации графиков зависимости коэффициентов SSE и Silhouette от количества кластеров, выбирается приблизительное число кластеров, которые возможно достаточно точно интерпретировать, а также берется в рассмотрение распределение числа объектов в соответствии с числом кластеров. По результатам анализа лучше интерпретируются три кластера.

После всех перечисленных этапов, выбирается наиболее оптимальный алгоритм кластеризации для представленных данных. Алгоритм спектральной кластеризации проигрывает в качестве, поэтому не рассматривается как итоговый алгоритм для интерпретации. Остальные два алгоритма показывают хорошие результаты качества. Разница, согласно коэффициенту силуэта, составляет 0.03. Побеждает алгоритма K-means, хоть и не с большим разрывом в качестве. В итоге интерпретация кластеров основывается на результате работы K-means.

Интерпретация

Кластер 0. Данный кластер характеризуется семьями с детьми, а также в некоторой мере семьями без детей. Родители обладают образованием высшего или незаконченного высшего уровня. В семье преимущественно не более трех детей, в основном в возрасте до подросткового. Средний доход составляет 37 000

долларов. Средний возраст родителей составляет 45 лет. Кластер также характеризуется небольшим объемом покупок. Клиенты из этой группы зарегистрированы на сайте относительно недавно - в среднем 34 месяца. В данной группе среднее количество покупок со скидкой. Наблюдается низкое количество покупок через веб-ресурсы, каталог и в магазине, однако чаще других клиенты посещают сайт. В данном кластере также наблюдается низкое количество покупок различных типов продуктов.

Возможно, клиенты данного кластера посещают сайт магазина из интереса, не являясь активными покупателями.

Кластер 1. Данный кластер характеризует людей без партнера или с партнером, но без детей. Уровень образования в основном высшее или незаконченное высшее. Средний доход составляет 79 000 долларов. Средний возраст клиентов в данной группе составляет 47 лет. Эта группа людей совершает в среднем наибольшее количество покупок – 1250. Клиенты из данного кластера зарегистрированы на сайте более 36 месяцев. Количество покупок со скидкой в этой группе незначительное. Клиенты данного кластера осуществляют среднее количество покупок через веб-ресурсы, но больше всего они ориентированы на покупки с использованием каталога или в магазине. В данной группе наблюдается наибольшее количество покупок вина, фруктов, мяса, рыбы, сладостей, а также немного больше покупок золота по сравнению с другими продуктами.

Эти клиенты представляют потенциальных покупателей с высоким доходом и большим объемом покупок.

Кластер 2. Данный кластер представлен семьями с детьми. Уровень образования в основном высшее. В этих семьях имеется 1-2 ребенка, преимущественно подросткового возраста. Средний доход составляет 60 000. Средний возраст родителей - 50 лет. Среднее количество покупок в данной группе - 750. Клиенты данного кластера зарегистрированы на сайте уже 38 месяцев. В данной группе наблюдается наибольшее число покупок со скидкой. Большинство покупок было совершено в магазине, на веб-ресурсах, а также

имеется среднее количество покупок по каталогу. По количеству покупок разных типов продуктов данная группа находится в среднем.

Потенциальные покупатели из данного кластера обладают средним доходом и средним объемом покупок. Возможно, для них больший интерес представляют товары для детей.

В третьей главе "Анализ временных рядов с помощью дендрограмм на примере ЭЭГ мозга" представляется международная система 10-20, которая является стандартом для проведения и расчета данных ЭЭГ мозга. Далее рассматривается два научных исследования, в которых промежуточным этапом является построение дендрограмм на основе ЭЭГ записей. Задача исследований состоит в том, что необходимо на основе дендрограмм найти числовую характеристику, которая смогла бы классифицировать психологические заболевания. В первом Анашина В. С., Тяпаева Л. Б., Давыдова В. В. "Классификация психических заболеваний на основе дендрограмм ЭЭГ головного мозга и их характеристик" ученые опираются на теорию кодирования и рассматривают дендрограмму с точки зрения префиксных кодов. На основе агрегирования данных по всем пациентам и временным окнам вычисляются максимальные, минимальные и средние показатели математического ожидания, дисперсии и энтропии. В результате приводится таблица 1.

Таблица 1 – Агрегированные результаты по всем коэффициентам и группам людей

Файл	Параметр	Мат. ожидание	Энтропия	Дисперсия
alz	Среднее	3.5741733	1.8168757	1.6849957
	Максимум	4.0937500	2.2526204	4.0898438
	Минимум	2.4453125	1.1292237	0.3974609
controls	Среднее	3.5953878	1.7979498	1.6737500
	Максимум	4.1875000	2.2995481	4.3085938
	Минимум	2.3896484	0.6962123	0.1523438
dep	Среднее	3.6573661	1.7742883	1.4166489
	Максимум	4.1875000	2.2995481	3.9940796
	Минимум	2.8359375	0.6962123	0.1523438
mci	Среднее	3.6290458	1.8588847	1.4953717

	Максимум	4.0937500	2.2653195	4.0082397
	Минимум	2.4140625	1.1862781	0.3974609
shiz	Среднее	3.6320219	1.8641116	1.5150029
	Максимум	4.0937500	2.2653195	4.1875000
	Минимум	2.7968750	1.2741306	0.3974609

Анализ вычисленных параметров префиксных кодов выявил следующую особенность: средние значения математического ожидания, энтропии и дисперсии для группы дендрограмм пациентов, принадлежащих одному ментальному кластеру, являются уникальными.

Во втором исследовании "EEG p-adic quantum potential accurately identifies depression, schizophrenia and cognitive decline" авторов Oded Shor, Amir Glik, Amit Yaniv-Rosenfeld, Avi Valevski, Abraham Weizman, Andrei Khrennikov, Felix Benninger ученые опираются на результаты вычисления p-адического квантового потенциала, который тоже следует из преобразования дендрограмм. На основе изменчивости квантового потенциала и среднего значения строятся несколько моделей классификации, которые позволяют сравнить контрольные группы здоровых людей и больных, а также сравнить группы больных попарно. Результаты представлены на рисунках 1, 2.

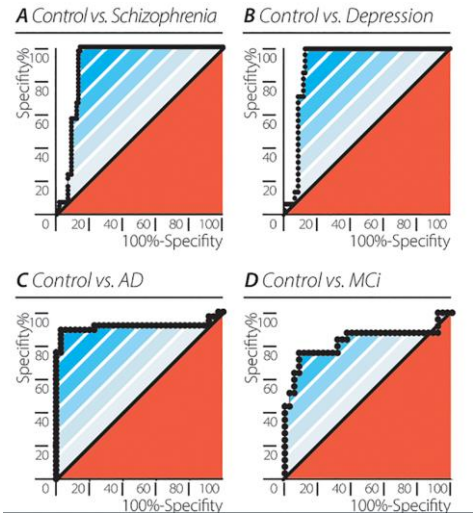

Рисунок 1 –Точность оценки квантового потенциала, основанного на ЭЭГ, показатель среднего значения и изменчивости (qpmvs), в дифференциации нейропсихиатрических групп пациентов от здоровых контрольных групп

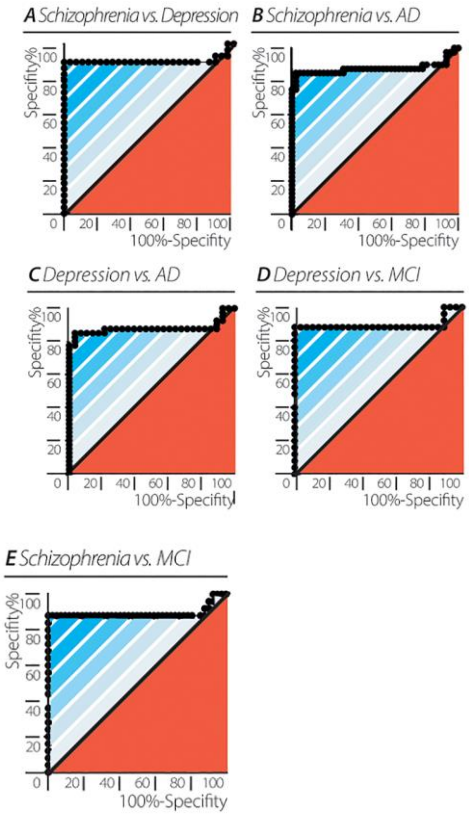

Рисунок 2 – Дифференциация между группами нейропсихиатрических пациентов с помощью показателя среднего значения и изменчивости квантового потенциала ЭЭГ (qpmvs)

В итоге парный анализ ROC-кривых контрольной группы здоровых лиц и группы пациентов с депрессией, шизофренией, болезнью Альцгеймера и легким когнитивным нарушением (MCI) показал чрезвычайно высокие значения площади под кривой (AUC), что указывает на то, что $qpmvs$ может быть полезным инструментом для диагностики наличия нейропсихиатрических или нейрокогнитивных заболеваний.

ЗАКЛЮЧЕНИЕ

В данной выпускной квалификационной работе был проведен анализ социологических данных с использованием трех различных методов кластеризации: K-means, агломеративной иерархической кластеризации и спектральной кластеризации. Результаты исследования показали, что метод K-means демонстрирует наилучшую производительность в данной задаче, выделяя более четкие и интерпретируемые кластеры социологических данных. Коэффициенты силуэта, использованные для оценки качества кластеризации,

отличались всего на 0.03 для трех кластеров, что свидетельствует о схожести результатов двух методов.

Тем не менее, стоит отметить, что иерархическая кластеризация также могла быть использована для интерпретации и анализа данных социологических исследований. Оба метода показали сопоставимые результаты, и возможность использования иерархической кластеризации для этой задачи остается актуальной.

Кроме того, в работе была продемонстрирована применимость иерархической кластеризации для анализа временных рядов ЭЭГ мозга. Данный пример исследования подтверждает, что иерархическая кластеризация может использоваться не только для данных количественного типа, но и для временных рядов. Два метода, основанные на префиксных кодах дендрограмм и их свойствах, а также на использовании p-адического квантового потенциала, были применены для анализа временных рядов ЭЭГ мозга, и они предоставили ценную информацию для предварительного определения ментального класса психических заболеваний.

Таким образом, данная работа подтверждает, что иерархическая кластеризация является мощным инструментом для анализа различных типов данных, включая как количественные данные, так и временные ряды. Ее использование позволяет выявлять скрытые структуры, группы схожих объектов и паттерны, что может быть полезным в различных областях, включая социологию и нейрофизиологию.