

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ ЭНТРОПИИ ПРЕФИКСНЫХ КОДОВ

(автореферат бакалаврской работы)

студента 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Губарева Алексея Александровича

Научный руководитель
доцент, кандидат ф.-м. наук

подпись, дата

Л.Б. Тяпаев

Зав. кафедрой
кандидат социологических наук, доцент

подпись, дата

И.Г. Малинский

Саратов 2025

ВВЕДЕНИЕ

Актуальность **темы** **выпускной** **квалификационной**
работы "Кластеризация на основе энтропии префиксных кодов" обусловлена стремительным ростом объемов данных в различных сферах деятельности, включая социально-экономические исследования. Современные методы анализа данных требуют новых подходов, способных эффективно выявлять скрытые закономерности и группировать объекты на основе их структурных характеристик. В этом контексте использование энтропии префиксных кодов, интерпретируемых через дендрограммы, представляет собой перспективное направление, позволяющее оценивать степень упорядоченности данных и устойчивость кластеров. Данный метод особенно актуален для анализа многомерных временных рядов, таких как социально-экономические показатели стран ОЭСР, где традиционные методы кластеризации зачастую оказываются недостаточно информативными.

Степень проработанности проблемы в научной литературе охватывает широкий спектр исследований, посвященных методам кластеризации и теории информации. Традиционные подходы, такие как k-means и иерархическая кластеризация, хорошо изучены, однако их ограничения, включая чувствительность к начальным параметрам и сложность интерпретации результатов, стимулируют поиск новых решений¹. В работах Тяпаева, Анашина и Давыдова показана эффективность использования энтропии префиксных кодов для анализа сложных структур данных, что открывает новые возможности для социально-экономических исследований². Тем не менее, применение этого метода для кластеризации стран ОЭСР остается малоизученной областью, что определяет научную новизну данного исследования.

¹ Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2009. 745 p.

² 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 её приложения». М.: ИПМ им. М.В. Келдыша РАН, 2022. С. 207–210.

Целью исследования является разработка и применение метода кластеризации социально-экономических показателей стран ОЭСР на основе анализа энтропии префиксных кодов, интерпретируемых через дендрограммы, для выявления устойчивых групп стран со схожими характеристиками и оценки внутренней согласованности кластеров.

Задачи исследования:

1. Провести обзор современных методов кластерного анализа с акцентом на подходы, использующие теорию информации и префиксные коды.
2. Разработать методологию представления социально-экономических данных в виде дендрограмм и их преобразования в максимальные префиксные коды с вычислением энтропии, математического ожидания и дисперсии длин кодовых последовательностей.
3. Реализовать программный инструмент для автоматизированного анализа данных стран ОЭСР, включая предобработку, построение дендрограмм и расчет информационных метрик.
4. Провести кластеризацию стран на основе полученных метрик, проанализировать устойчивость и интерпретируемость результатов, а также сравнить эффективность предложенного метода с традиционными подходами.

Объектом исследования являются социально-экономические показатели стран ОЭСР, представленные в виде временных рядов, включающие такие параметры, как валовой внутренний продукт, численность населения, уровень выбросов углекислого газа, продолжительность жизни, расходы на образование и другие ключевые индикаторы, характеризующие экономическое и социальное развитие государств.

Предметом исследования является метод анализа иерархических структур данных на основе энтропии префиксных кодов, включающий преобразование социально-экономических показателей в дендрограммы, вычисление информационных метрик и их применение для кластеризации стран, а также оценку сравнительной эффективности данного подхода.

Структура ВКР: состоит из введения, четырех глав, заключения и списка использованных источников.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе «Теоретические основы» рассматриваются ключевые концепции, связанные с префиксными кодами, дендрограммами и их применением в кластерном анализе социально-экономических данных. Основное внимание уделяется математическим ожиданиям, энтропии и дисперсии префиксных кодов как мерам для оценки устойчивости и структуры кластеров.

Префиксные коды и дендрограммы

Префиксные коды определяются как класс кодов, где ни одно кодовое слово не является началом другого, что обеспечивает однозначность декодирования. Они тесно связаны с дендрограммами через механизм кодирования ветвей дерева, где каждая ветвь от корня к терминальному узлу может быть представлена уникальной бинарной последовательностью. Это свойство позволяет использовать дендрограммы не только для визуализации иерархической кластеризации, но и для формального анализа структуры данных через префиксные коды. Важным аспектом является выполнение неравенства Крафта-Макмиллана, которое отражает полноту кода и отсутствие избыточности в структуре дерева³.

Математическое ожидание, энтропия и дисперсия префиксных кодов

Математическое ожидание длины кода (E) характеризует среднюю сложность иерархической структуры данных. Для дендрограммы оно вычисляется по формуле:

$$E = \sum_{i=1}^r i \cdot \frac{N_i}{2^i},$$

где N_i – количество кодовых слов длины i , а r – общее число кодовых слов.

³ Иорданский М.А., Смышляева О.В. Кодирование комбинаторных объектов: учебное пособие. Н. Новгород: Мининский университет, 2017. 128 с.

Энтропия (H) служит мерой неопределенности и рассчитывается как:

$$H = - \sum_{i=1}^r \frac{N_i}{2^i} \log_2 \left(\frac{N_i}{2^i} \right),$$

Она отражает степень упорядоченности распределения длин кодовых последовательностей.

Дисперсия (Var) показывает изменчивость структуры дендрограммы и вычисляется по формуле:

$$Var = \sum_{i=1}^r \frac{N_i}{2^i} (i - E)^2,$$

Эти метрики позволяют количественно оценить устойчивость и внутреннюю согласованность кластеров⁴.

Методы кластеризации

Иерархическая кластеризация, основанная на построении дендрограмм, является одним из ключевых методов анализа данных. В отличие от плоских методов, таких как k-means, иерархический подход сохраняет информацию о вложенности кластеров и позволяет анализировать данные на различных уровнях детализации⁵. Особое внимание уделяется методам агломерации (например, среднему связыванию), которые менее чувствительны к выбросам и обеспечивают более стабильные результаты⁶. Предлагаемый метод кластеризации на основе энтропии префиксных кодов позволяет анализировать структуру данных через информационные характеристики, минуя этап вычисления попарных расстояний между объектами.

Энтропия как мера кластеризации

⁴ Тяпаев Л. Б., Анашин В. С., Давыдов В. В. О методе обработки большой совокупности аналоговых сигналов с целью выделения характеристических признаков источников сигналов // Дискретные модели в теории управляющих систем : Труды XI международной конференции / Отв. ред. С. А. Ложкин, Д. С. Романов, В. В. Подымов. Москва : МАКС Пресс, 2023, С. 110–113. EDN: XTMWWX

⁵ Kaufman L., Rousseeuw P.J. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken: Wiley, 1990. 342 p.

⁶ Everitt B.S., Landau S., Leese M., Stahl D. Cluster Analysis. 5th ed. Chichester: Wiley, 2011. 346 p.

Энтропия префиксных кодов служит индикатором структурных особенностей данных, позволяя выявлять скрытые закономерности в сложных многомерных наборах⁷. Её применение особенно эффективно для социально-экономических показателей, где традиционные методы могут давать поверхностные результаты. Например, высокая энтропия может указывать на сложную, слабо структурированную группировку данных, тогда как низкие значения соответствуют четко выраженным кластерам. Сочетание энтропийного анализа с методами машинного обучения открывает новые возможности для повышения точности и интерпретируемости результатов кластеризации⁸.

Во второй главе «Методика исследования» рассматриваются ключевые аспекты организации и проведения исследования, включая описание данных стран ОЭСР и подготовку инструментария для их анализа. В главе подробно освещаются социально-экономические показатели, используемые для кластеризации, такие как ВВП, численность населения, уровень выбросов CO₂, продолжительность жизни и другие индикаторы, характеризующие развитие стран-членов Организации экономического сотрудничества и развития (ОЭСР)⁹. Особое внимание уделяется комплексному характеру данных ОЭСР, их стандартизированной методологии сбора и высокой степени сопоставимости между странами, что обеспечивает надежность результатов анализа.

Далее описывается разработанный программный комплекс на языке Python, предназначенный для обработки, нормализации и кластеризации данных. Архитектура решения построена по модульному принципу, где каждый компонент выполняет строго определенные функции: загрузка данных, предварительная обработка (включая заполнение пропущенных значений и нормализацию), кластерный анализ с использованием метрики Хеллингера и

⁷ Chen J., Li K., Rong H. Entropy-based Clustering for Complex Data Sets // IEEE Transactions on Knowledge and Data Engineering. 2022. Vol. 34. № 3. P. 1124–1137.

⁸ Zhang Y., Li P., Liu S. Machine Learning with Entropy Regularization // Journal of Machine Learning Research. 2023. Vol. 24. P. 1–35.

⁹ Organisation for Economic Co-operation and Development (OECD). Official Documents. 2023. [Электронный ресурс] / URL: <https://www.oecd.org>. (дата обращения 09.04.2025)

визуализация результатов. Подчеркивается важность этапа нормализации данных, особенно для разнородных показателей, где применяется метод Min-Max, преобразующий значения в унифицированный диапазон [0, 1] для устранения искажений при вычислении расстояний¹⁰.

Особую научную ценность представляет модуль работы с префиксными кодами, который преобразует дендрограммы в систему бинарных кодов. Этот модуль автоматически генерирует уникальные префиксные коды для каждого признака, вычисляет их ключевые характеристики (математическое ожидание, энтропию, дисперсию) и выполняет сравнительный анализ структур кодов между странами. Методологическая основа системы объединяет традиционные подходы машинного обучения с авторскими разработками в области анализа иерархических структур, что позволяет количественно оценивать сложность кластерных структур и их устойчивость.

В заключении главы обсуждаются практические аспекты реализации инструментария, включая использование библиотек `pandas`, `scikit-learn` и `matplotlib` для эффективной работы с данными и визуализации результатов. Подчеркивается, что разработанный комплекс обеспечивает полный цикл обработки данных — от загрузки до кластеризации и интерпретации результатов, что делает его ценным инструментом для социально-экономических исследований¹¹. Результаты анализа могут быть использованы для разработки экономических стратегий, прогнозирования тенденций и принятия управленческих решений, способствуя развитию международного сотрудничества и устойчивого развития¹².

В третьей главе «Практическая реализация» представлен детальный анализ процесса кластеризации социально-экономических показателей стран

¹⁰ Milligan G.W., Cooper M.C. A Study of Standardization of Variables in Cluster Analysis // *Journal of Classification*. 1988. Vol. 5. P. 181–204.

¹¹ Hyndman R.J., Wang E., Laptev N. Dimension Reduction for Clustering Time Series Using Global Characteristics // *Computational Statistics*. 2015. Vol. 30. P. 1–20.

¹² Organisation for Economic Co-operation and Development (OECD). Policy Frameworks. 2023. [Электронный ресурс] / URL: <https://www.oecd.org>. (дата обращения 09.04.2025)

ОЭСР на основе энтропии префиксных кодов. Глава начинается с предобработки данных, где описываются ключевые этапы подготовки информации для анализа: отбор значимых показателей, обработка пропущенных значений с использованием методов линейной интерполяции и заполнения крайними значениями, а также структурирование данных по странам и временным периодам. Особое внимание уделяется нормализации данных с применением метода Min-Max, что позволяет устранить проблему разнородности шкал измерений и обеспечить корректность последующего анализа.

Далее рассматривается методика построения дендрограмм, где для количественной оценки сходства между временными рядами используется расстояние Хеллингера. Этот подход демонстрирует высокую эффективность при работе с распределениями и неотрицательными векторами, что особенно важно для анализа социально-экономических показателей. На основе матрицы попарных расстояний выполняется иерархическая кластеризация с применением метода среднего связывания, который обеспечивает устойчивость результатов и учитывает структуру всего кластера. Полученные дендрограммы преобразуются в систему бинарных префиксных кодов, где каждый узел дерева кодируется последовательностью битов, что позволяет формализовать иерархическую структуру данных.

В части, посвященной вычислению математического ожидания, энтропии и дисперсии префиксных кодов, подробно описываются формулы и методы расчета этих характеристик. Математическое ожидание длины кода отражает среднюю сложность кластерной иерархии, энтропия служит мерой неопределенности в распределении длин ветвей, а дисперсия количественно оценивает изменчивость структуры дендрограммы. Эти метрики позволяют выявлять устойчивые кластеры и анализировать их внутреннюю согласованность.

Заключительная часть главы включает визуализацию результатов, где представлены столбчатые диаграммы распределения характеристик по странам, а также трехмерные графики, отображающие взаимосвязь математического

ожидания, энтропии и дисперсии. Визуализация помогает интерпретировать полученные данные и выявлять закономерности в социально-экономическом развитии стран. На основе анализа результатов выделены три ключевых кластера: сбалансированные развитые экономики, динамичные переходные экономики и контрастные специализированные экономики, каждый из которых обладает уникальными структурными особенностями. Практическая значимость исследования подчеркивается возможностью применения результатов для разработки экономических стратегий и принятия управленческих решений.

В четвёртой главе «Результаты и анализ» представлены итоги проведённого исследования, включая интерпретацию полученных данных и их значение для понимания социально-экономических различий между странами ОЭСР. На основе анализа математического ожидания, энтропии и дисперсии префиксных кодов выявлены три устойчивых кластера стран, каждый из которых обладает уникальными характеристиками.

Первый кластер, включающий такие страны, как Норвегия, Италия и Дания, отличается низкой энтропией, умеренным математическим ожиданием и минимальной дисперсией, что свидетельствует о сбалансированности и стабильности их экономических систем. Эти страны демонстрируют чёткую дифференциацию социально-экономических показателей и высокую степень внутренней упорядоченности.

Второй кластер, к которому относятся Израиль, Мексика, Чили и другие динамично развивающиеся экономики, характеризуется высокой энтропией, что отражает переходный характер их структур. Умеренная дисперсия указывает на наличие как развитых, так и отстающих секторов, что связано с внутренней неоднородностью и процессами трансформации.

Третий кластер, включающий Люксембург, Швейцарию и Австралию, выделяется высокой дисперсией и сложной кластерной иерархией, что говорит о выраженных контрастах между секторами экономики. Средняя энтропия в этом кластере отражает частичную упорядоченность при сохранении значительных внутренних диспропорций.

Практическая значимость исследования подчёркивается возможностью использования полученных результатов для разработки экономических стратегий, прогнозирования тенденций и принятия обоснованных управленческих решений. Методология, основанная на энтропии префиксных кодов, доказала свою эффективность в анализе сложных многомерных данных и может быть адаптирована для исследований в других областях. В заключении главы отмечается, что предложенный подход сочетает теоретическую строгость с практической применимостью, открывая новые перспективы для изучения социально-экономических процессов.

ЗАКЛЮЧЕНИЕ

В данной ВКР был разработан и применен метод кластеризации социально-экономических показателей стран ОЭСР на основе анализа энтропии префиксных кодов, интерпретируемых через дендрограммы. Исследование продемонстрировало, что использование математического ожидания, энтропии и дисперсии префиксных кодов позволяет эффективно выявлять устойчивые группы стран со схожими характеристиками, а также оценивать внутреннюю согласованность кластеров. Этот подход сочетает теоретическую строгость с практической применимостью, предлагая новый взгляд на анализ многомерных данных.

Результаты исследования подтвердили гипотезу о том, что энтропия префиксных кодов может служить надежной мерой кластеризации, отражая степень упорядоченности и устойчивости социально-экономических структур. Анализ выявил три ключевых кластера стран: сбалансированные развитые экономики, динамичные переходные экономики и контрастные специализированные экономики. Каждый из этих кластеров обладает уникальными характеристиками, что позволяет глубже понять закономерности их развития и взаимосвязи между различными показателями.

Практическая значимость работы заключается в возможности применения полученных результатов для разработки экономических стратегий, прогнозирования тенденций и принятия управленческих решений. Методология,

основанная на энтропии префиксных кодов, может быть адаптирована для анализа других типов данных, что открывает новые перспективы для исследований в области экономики, социологии и смежных дисциплин.

В заключение можно отметить, что предложенный метод кластеризации на основе энтропии префиксных кодов представляет собой ценный инструмент для анализа сложных многомерных данных. Его применение позволяет не только выявлять скрытые закономерности, но и оценивать устойчивость кластерных структур, что делает его особенно актуальным для современных социально-экономических исследований. Дальнейшее развитие этого направления может включать адаптацию метода для работы с другими типами данных, а также его интеграцию с современными алгоритмами машинного обучения для повышения точности и интерпретируемости результатов.