Санкт–Петербургский государственный университет

ДАШКОВА Олеся Вячеславовна

Выпускная квалификационная работа

Статистический анализ с применением методов кластерного анализа для исследования социально - экономических данных

Уровень образования: бакалавриат
Направление 01.03.02 «Прикладная математика и информатика»
Основная образовательная программа СВ.5005.2020 «Прикладная математика, фундаментальная информатика и программирование»

Научный руководитель:
профессор, кафедра математической теории игр и статистических решений,
д.т.н. Буре Владимир Мансурович

Рецензент:
старший преподаватель, кафедра диагностики функциональных систем,
Орехов Андрей Владимирович

Санкт-Петербург
2024 г.

Содержание

Введение

Классификация субъектов Российской Федерации является важным инструментом для изучения, систематизации и прогнозирования социально-экономического развития административно-территориальных образований страны.

Исследуя результаты, полученные экспертами в рассматриваемой предметной области, появляется возможность определить типовые регионы, разработать целенаправленные стратегии развития для каждой соответствующей группы и обеспечить более обоснованное и эффективное управление региональным развитием [1].

Кроме того, анализ социально-экономических показателей сформированных типовых групп субъектов Федерации может наглядно продемонстрировать сложившуюся диспропорцию в развитии экономики регионов, которая часто возникает вследствие неэффективного распределения доступных экономических и социальных ресурсов [2].

Для математического обоснования экспертной классификации регионов России можно использовать типологизацию на основе одного из современных методов машинного обучения, а именно — кластеризации (следует упомянуть, что в случае классификации известно число групп и свойства этих групп, а в задаче типологизации не известно ни то ни другое).

Данная процедура хорошо подходит для анализа данных, так как позволяет группировать объекты на основе сходства выбранных показателей. Применение данной методики к субъектам Российской Федерации даёт возможность выделить группы регионов, имеющие схожие характеристики в рассматриваемой предметной области.

При осуществлении типологизации регионов необходимо уделить особое внимание качественному конструированию признакового пространства, которое должно максимально точно описывать свойства рассматриваемых субъектов Федерации в выбранной области исследования (уровень жизни, экономика, социально-экономический аспект, миграция и т.д.). Не менее важным аспектом является выбор алгоритма кластеризации. Он осуществляется на основе состава исходных данных и на определённых критериях успеха

исследования.

Существует множество методик типологизации регионов, каждая из которых по-своему уникальна и определяется доступными наборами объективных исходных данных и конкретной целью проводимого исследования. Результат, при этом, оценивается либо статистическими методами, используя определённые метрики, основанные на внутрикластерной схожести и межкластерном различии, либо путём интерпретации качества полученных групп, основываясь на экспертной оценке.

Для того, чтобы приблизиться к классификации, полученной экспертами, необходимо исследовать наиболее популярные алгоритмы кластеризации, реализовать машинный эксперимент, осуществив подбор параметров.

Постановка задачи

В качестве научного задела для выполнения данной работы была использована статья с результатами кластеризации регионов России по экономическим и инновационным показателям [3]. В качестве формализованных экспертных знаний был использован рейтинг инновационного развития субъектов Российской Федерации [4].

Используя перечисленные документы как информационную базу исходных данных, необходимо сформировать набор данных и провести предобработку.

Кроме того, необходимо рассмотреть наиболее популярные алгоритмы кластеризации, метрики для оценки результатов и провести машинный эксперимент с имеющимися данными. Выполнить следующий анализ: какой из методов и наборов параметров даёт разбиение, ниаболее приближенное к полученному экспертами. Затем описать полученные результаты и выявленные недостатки тех или иных методов.

В качестве завершающего этапа работы отдельно рассмотреть эффективность иерархических методов кластеризации. При этом уделить внимание проблеме отсутствия чёткого критерия для определения оптимального числа кластеров. В целом, данная работа должна заключаться в реализации методов анализа данных и исследовании результатов вычислительных экспериментов.

Обзор литературы

Существует достаточное количество публикаций и литературы, в которых рассматриваются различные варианты классификации и типологизации административно-территориальных образований Российской Федерации и проводятся исследования социально-экономических показателей сгруппированных регионов. На основе полученной информации можно сделать несколько существенных выводов.

В настоящее время инновационное развитие страны в целом и отдельных регионов в частности является одной из наиболее приоритетных задач государства. Во многих работах экспертов-экономистов для классификации регионов в качестве одного из существенных факторов используются именно показатели, характеризующие научно-технический прогресс [5]. В частности, инновационные характеристики помимо ВРП являются весьма существенными для оценки стратегической конкурентоспособности регионов [6].

Одним из показателей качества экономической политики государства в современных условиях является выравнивание диспропорций в развитии экономики регионов.

Кластерный анализ весьма широко используется для выявления проблемных регионов и исследования факторов, оказывающих на них негативное влияние [7]. Это, в свою очередь, служит сигналом о необходимости принятия мер по стимулированию развития отстающих регионов за счёт исправления некоторых существенных показателей.

В частности, исследование вопроса кластеризации субъектов РФ по показателям, относящимся к инновационной деятельности описано в статьях [8], [9].

Глава 1. Разделительная и плотностная кластеризация

1.1 Введение в задачу кластеризации

Под кластеризацией понимается процесс разделения конечного числа объектов на группы (кластеры) на основе известных об объектах данных так, чтобы каждый из них принадлежал только одному кластеру. В качестве объектов выступают абсолютно любые сущности, которые могут отличаться друг от друга на основе своих характеристик или свойств. Такое множество элементов кластеризации представимо в виде n точек d-мерного пространства (n объектов кластеризации по d признакам). Различные типы кластеризации можно описать с помощью иерархии 1, предложенной в публикации об алгоритмах кластерного анализа [10].

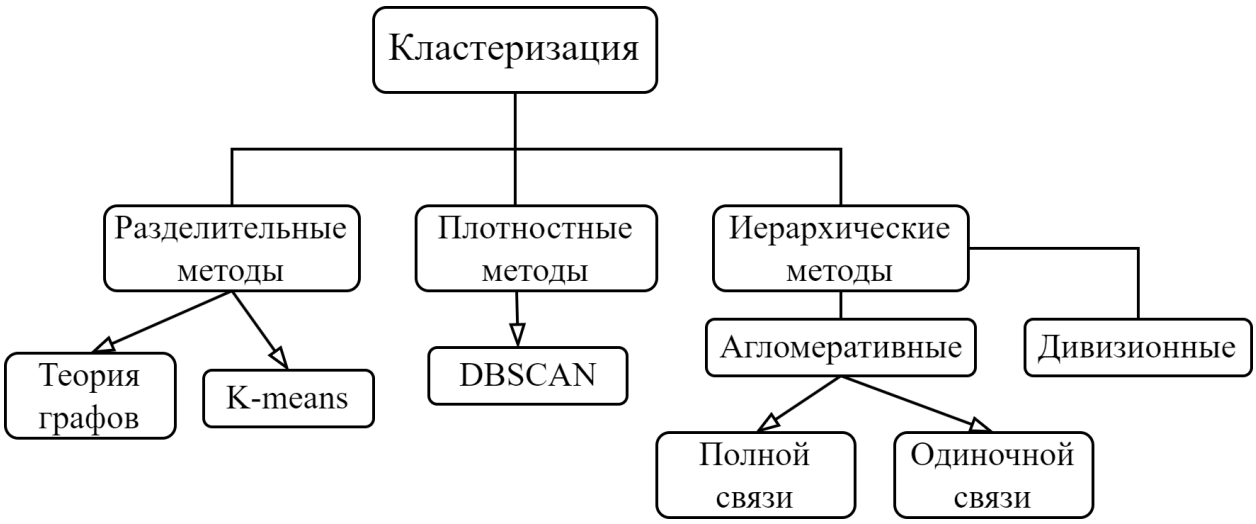

Рис. 1: Иерархия методов кластеризации

Методы кластеризации разделяются на иерархические, разделительные и основанные на плотности распределения данных [10]. Первые производят серию вложенных разбиений (будут подробнее рассмотрены во втором разделе). Разделительные методы с каждой итерацией улучшают разбиение исходного набора данных. Алгоритм плотностной кластеризации основан на оценке плотности распределения элементов внутри кластеров.

1.2 Обзор методов

Сначала рассмотрим разделительные методы. Преимущественно, они используются при работе с большим объемом данных. Всё потому, что в данном случае получение одного разбиения менее затратно с точки зрения памяти и времени, чем реализация и хранение целой серии разбиений. Хорошо работает с изолированными и компактными кластерами. На практике метод применяется несколько раз с разными начальными состояниями и выбирается наилучший результат.

Наиболее популярным представителем является алгоритм k-means (k-средних). В его основе лежит оптимизация критерия квадратичной ошибки.

$$e^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \|x_j^{(i)} - c_i\|^2$$

Действие алгоритма начинается со случайного выбора k экземпляров среди всего множества объектов. Они в дальнейшем отождествляются с геометрическими центрами каждого кластера (центроидами). Оставшиеся объекты соотносятся с ближайшими к ним кластерами. Затем перевычисляются геометрические центры кластеров и обновляется принадлежность элементов. Процесс завершается в момент, когда на очередной итерации суммарное квадратичное отклонение точек кластеров от соответствующих центроидов перестает изменяться.

Алгоритм имеет два существенных недостатка: необходимость указывать заранее количество кластеров и сильная чувствительность к начальному выбору центроидов.

Существует улучшенная версия алгоритма k-средних, которая помогает избежать проблем со сходимостью и получить более устойчивые результаты. Метод был предложен в статье «k-means++: The Advantages of Careful Seeding» Дэвидом Артуром с соавторами в 2007 году. [11] В этой работе они представили улучшенную стратегию по выбору начальных центроидов. Так, объекты, находящиеся дальше от существующих центров кластеров, более вероятно станут новыми центрами. Такая стратегия позволяет разместить исходные центроиды более разнообразно и эффективно, что способствует повышению

вероятности сходимости алгоритма кластеризации к глобальному минимуму и улучшает качество получаемых результатов.

Методы, основанные на плотности позволяют обнаруживать кластеры произвольной формы и обрабатывать выбросы. Один из самых известных плотностных алгоритмов кластеризации — DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [12]. Алгоритм выделяет кластеры путем определения областей высокой плотности точек данных. Он использует два основных параметра — радиус (eps), определяющий окрестность точек, и минимальное количество точек ($MinPts$), что определяет минимальное количество точек, необходимое для образования кластера.

Алгоритм начинается с рассмотрения некоторого из объектов и его окрестности радиуса eps. Если в этой окрестности присутствует $MinPts$ точек, то объект считается «ядром» или же элементом кластера. Так, с каждой итерацией, кластеры увеличиваются, добавляя все точки в окрестности ядра, и повторяя процесс для каждой новой точки.

Хотя DBSCAN является мощным и универсальным алгоритмом кластеризации, у него также есть недостатки. Если в данных присутствуют кластеры с различной плотностью точек, то может возникнуть проблема объединения их в один кластер или разделения на несколько кластеров в зависимости от параметров. Алгоритм не только чувствителен к входным параметрам (неправильный их выбор может привести к нежелательным результатам), но самое неприятное то, что эти параметры задаются эвристически, а эвристические методы основаны не на строгих математических формулах, а на правдоподобных рассуждениях.

1.3 Меры оценки качества кластеризации

Результаты кластеризации могут быть оценены с помощью мер оценки качества, которые подразделяются на две группы. Внешние (external) - использующие размеченные данные. Внутренние (internal) - используют только информацию о структуре данных и сформированных кластерах, они основываются на таких понятиях, как компактность и отделимость кластеров [13]. В данной работе будут использоваться только внутренние меры. Рассмотрим

наиболее популярные из них.

Метод Силуэта (Silhouette). В основе идеи метода лежит вычисление коэффициентов, которые присваиваются каждому объекту в кластере и образуют так называемый силуэт кластера. Коэффициенты изменяются в диапазоне от -1 до 1. Значения, близкие к 1, указывают на то, что объект является похожим на другие объекты в кластере и не похожим на объекты из других кластеров. Если большинство объектов имеют значения коэффициентов близкими к 1, можно утверждать, что кластерная структура хорошо выражена, и количество кластеров соответствует естественной группировке данных. Напротив, если в силуэте кластера много низких и отрицательных значений, это говорит о том, что кластерная структура плохо соответствует естественным группам данных, т.е. кластеров слишком много или слишком мало.

$$Sil = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{|c_j|} \frac{b(x_i, c_j) - a(x_i, c_j)}{\max(a(x_i, c_j), b(x_i, c_j))}$$

Индекс силуэта рассчитывается для каждого объекта как разность между средним расстоянием до объектов ближайшего кластера и средним расстоянием до объектов внутри того же кластера, деленная на максимальное из этих двух значений. Для оценки качества кластеризации можно усреднить значения индекса силуэта по всем объектам.

Индекс Дэвиса-Болдуина (Davies–Bouldin Index). Оптимальное значение индекса стремится к нулю, что указывает на хорошее разделение кластеров. Чем выше значение индекса, тем хуже качество кластеризации.

$$DB = \frac{1}{N} \sum_{c_j \in C} \max_{c_k \in C \setminus c_j} \left(\frac{S(c_j) + S(c_k)}{\|c_j - c_k\|} \right)$$

Показатель индекса вычисляется путем рассмотрения каждого кластера и сравнения его соседей. Более конкретно, для каждого кластера вычисляется отношение между суммой внутрикластерных расстояний (мера разброса внутри кластера) и расстоянием между центроидами текущего кластера и его наиболее близкого кластера. Затем берется среднее по всем кластерам, чтобы получить итоговое значение индекса.

Индекс Калинского-Харабаса (Calinski–Harabasz). Вычисляется путем сравнения оценки дисперсии между кластерами с оценкой дисперсии внутри кластеров. Более конкретно, данная метрика использует отношение межкластерной дисперсии к внутрикластерной дисперсии, где более высокое значение индекса указывает на более четкое разделение кластеров.

$$CH = \frac{N-K}{K-1} \cdot \frac{\sum_{c_j \in C} |c_j| \cdot \|\overline{c_j} - \overline{X}\|}{\sum_{c_j \in C} \sum_{x_i \in c_j} \|x_i - \overline{c_j}\|}$$

При анализе кластеризации данных рекомендуется использовать несколько метрик вместе, чтобы получить более полное представление о качестве кластеризации и выборе наилучших параметров модели.

1.4 Исходные данные

В качестве объектов кластеризации были рассмотрены субъекты Российской Федерации (82 штуки). Каждый из них описывается шестью признаками: три признака описывают экономическое развитие субъекта и три признака - инновационное развитие. Информационной базой для выбора признаков стала статья Заварухина о кластеризации регионов РФ [8]. Данные собраны с сайта Росстат.

В качестве признаков, указывающих на уровень экономического развития региона, выступили:

1. Валовый региональный продукт (ВРП) на душу населения в текущих (основных) ценах. Представляет собой суммарную добавленную стоимость продуктов и услуг, произведённых предприятиями того или иного региона, и определяется, как разница между выпуском и промежуточным потреблением [14];

2. Инвестиции в основной капитал на душу населения в фактически действовавших ценах. Предоставляют информацию о том, какие средства вложены в расширение и улучшение производственной базы, инфраструктуры и других средств производства в конкретном регионе [15];

3. Объем основных фондов на душу населения. Отражает накопленную

стоимость производственных активов, необходимых для осуществления производственной деятельности [16].

Для вычисления среднедушевого значения использована среднегодовая численность постоянного населения за соответствующие года.

Использованные признаки, олицетворяющие инновационную составляющую региона:

1. Коэффициент изобретательной активности без учёта полезных моделей. Число отечественных патентных заявок на изобретения, поданных в России, в расчете на 10 тыс. человек населения [17]. Под “изобретением” понимается техническое решение в некоторой области, относящееся к продукту. “Полезная модель” - техническое решение, относящееся к устройству. Разница в менее строгих условиях патентоспособности [18].

2. Внутренние затраты на научные исследования и разработки в расчете на одного научного работника. Отражают уровень финансовых ресурсов, которые компании, организации или государственные учреждения готовы инвестировать в инновационные процессы [19]. Для вычисления признака используется значение численности персонала, занятого научными исследованиями и разработками.

3. Объем отгруженных инновационных товаров, работ и услуг на душу населения. Показатель определяется отношением объема инновационных товаров, работ, услуг к общему объему отгруженных товаров, выполненных работ и услуг [20].

Для успешного выполнения задач классификации и кластеризации необходимо представить исходные данные в числовом формате и привести их к единой шкале измерений.

Признаки уже представлены в количественной шкале, но измеряются в разном масштабе. Преобразование данных в набор, сохраняющий статистические характеристики, но имеющий неопределенные минимальные и максимальные значения называется стандартизацией данных [21]. Существует несколько методов стандартизации, например Z-преобразование.

$$x_{ij} = \frac{x_{ij} - \mu}{\sigma}$$

μ - среднее, σ^2 - дисперсия. Используем его для стандартизации исходных данных о субъектах.

Также рассмотрим разбиение, полученное в статье Заварухина [8] методом k-means:

1. Средними по РФ уровнями экономического и инновационного развития: Белгородская, Липецкая, Смоленская, Архангельская, Вологодская, Ленинградская, Мурманская, Челябинская, Иркутская, Томская и др. (26 регионов)

2. Экономически слабые территории с низкими показателями инновационной деятельности: Брянская, Владимирская, Воронежская, Рязанская, Саратовская и т.д.(47 регионов)

3. Лидерами в научном и инновационном развитии: г. Москва, г. Санкт-Петербург, Московская область, Республика Татарстан. (4 региона)

4. Высокий ВРП и низкая инновационная активность: Тюменская область, Республика Саха (Якутия), Магаданская область, Сахалинская область и Чукотка. (5 регионов)

1.5 Применение рассмотренных методов

Реализация всех разбиений осуществлялась с использованием языка Python и библиотеки «scikit-learn». Рассмотрим метод «KMeans». У него есть достаточно большое количество параметров. Укажем только те, которые были изменены: число кластеров ($n\_clusters$) указываем равное 4, параметр $init$ - использовано значение «random» для произвольного выбора центроидов. Параметр n_init отвечает за то, сколько раз будут пересчитываться центроиды для выбора наилучшего варианта с точки зрения инерции. Данные параметры ($init$, n_init) помогают предотвратить проблему того, что элемент может оказаться изолированным от других и образовать кластер из одного элемента.

В данных о субъектах РФ безусловно присутствуют подобные элементы (Чукотский АО, Еврейская АО) - можно заметить это при исследовании ключевых характеристик данных. Следовательно данный параметр должен быть ненулевым, выбрано значение 5. Также задано максимальное количество итераций ($max\_iter = 20$) - данных не слишком много и кластеров тоже.

В качестве результата буду приводить размеры каждого из полученных кластеров: (n_1, n_2, n_3, n_4) и, в случае необходимости, какие субъекты вошли в каждый из них.

Провожу кластеризацию 100 раз для определения самого часто реализуемого разбиения. Им оказывается разбиение на кластеры размерами $(3, 4, 13, 62)$. Можем обратить внимание, что оно достаточно сильно похоже на разбиение представленное нам в статье. Но, нельзя однозначно утверждать, что полученная чаще всего кластеризация будет эталонной. «K-means» - это итеративный алгоритм, который зависит от начальных условий и случайного выбора центров кластеров. Даже при одних и тех же данных и числе кластеров результаты могут варьироваться в зависимости от случайности начальной инициализации.

Попробуем использовать метрики для нахождения подходящего разбиения. Лучшее с точки зрения метрик $silhouette, DB$ - $(1, 1, 13, 62)$. Лучшее по метрике CH: $(1, 4, 14, 63)$. Как мы видим, в данном случае наилучшей кластеризацией оказывается та, которая выделяет кластеры с одним элементом. Такой результат нас не устраивает.

После применения алгоритма выбора оптимального числа кластеров — метода «силуэта», количество кластеров для наилучших по метрикам моделей предположительно должно быть равно 2 или 3. В таком случае, для $k = 3$, получаем разбиение $(1, 5, 76)$, что также является результатом, не схожим с эталонным разбиением.

После применения улучшенного метода «k-means++» получаем частое разбиение - $(1, 4, 14, 63)$. Наилучшая кластеризация по метрикам так же выделяет по одному элементу в отдельный кластер: $(1, 3, 7, 71)$, $(1, 3, 10, 68)$, $(1, 4, 12, 65)$.

Применяя метод «DBSCAN», возникает проблема подбора входных параметров ($eps, MinPts$) - сложно выбрать их так, чтобы общее количество

кластеров равнялось 4. Тем не менее полученный результат: $(2, 3, 28, 49)$. Но, ни один из кластеров не совпадает с эталонным (не выделяет главную тройку субъектов Москва, Санкт-Петербург и Московская обл.).

Глава 2. Иерархическая кластеризация

2.1 Описание метода

Перейдём к реализации метода иерархической кластеризации. Он представляет из себя типологизацию объектов в иерархическую структуру. В качестве результата выступает дендограмма разбиения — древовидная схема, на которой отображена получившаяся структура данных [22], [23].

Преимуществами являются простота в понимании результатов, возможность в визуализации иерархической структуры данных. Недостатком являются вычислительные затраты в случае большого объема данных. Существует два алгоритма иерархической кластеризации: агломеративный и дивизионный.

Агломеративный подход заключается в том, что в начале алгоритма каждый объект является отдельным кластером; затем, пары кластеров объединяются по мере продвижения иерархии, пока не останется один общий кластер.

Основная проблема реализации любого алгоритма агломеративной кластеризации заключается в определении момента остановки, который задает количество кластеров.

Агломеративный алгоритм основан на построении матрицы расстояний между кластерами (обозначим их, как U и V). Существует несколько методов агломеративной кластеризации.

1. Одиночной связи (single linkage): в качестве расстояния между кластерами берётся минимальное расстояние среди всех попарных расстояний элементов двух кластеров.

$$d(U, V) = \min_{x_1 \in U,\, x_2 \in V} \{d(x_1, x_2)\}$$

2. Полной связи (complete linkage): в качестве расстояния между кластерами используется максимальное среди всех попарных расстояний между элементами двух кластеров.

$$d(U,V) = \max_{x_1 \in U,\, x_2 \in V} \{d(x_1, x_2)\}$$

3. Средней связи (average linkage): в качестве расстояния между кластерами используется среднее арифметическое среди всех попарных расстояний элементов двух кластеров.

$$d(U,V) = \frac{1}{|U| + |V|} \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} d(x_i, x_j)$$

4. Центроидный метод (centroid method): в качестве расстояния между кластерами используется расстояние между центроидами двух кластеров.

$$d(U,V) = d(c_U, c_V)$$

5. Метод Варда (Ward's method): расстояние между двумя кластерами высчитывается по формуле:

$$d(U,V) = SST_{U \cup V} - (SST_U + SST_V)$$

$$SST_U = \sum_{i=1}^{|U|} \|x_i - c\|; \quad SST_V = \sum_{i=1}^{|V|} \|x_i - c\|;$$
$$SST_{U \cup V} = \sum_{i=1}^{|U \cup V|} \|x_i - c\|$$

Для рассмотренных методов расстояние между кластерами на каждом шаге процесса можно вычислить при помощи формулы Ланса - Уильямса [24].

Результат с одним кластером не представляет практической ценности. Следовательно, имеет смысл остановить алгоритм на некоторой итерации, чтобы получить оптимальное количество кластеров. Для этого может быть использован критерий остановки агломеративного процесса, представленный в статье Орехова А.В. [25].

2.2 Марковский момент остановки агломеративного процесса кластеризации

После завершения агломеративного алгоритма кластеризации мы имеем множество минимальных расстояний $F_1, F_2, F_3, ..., F_{n-1}$, при этом для иерархических агломеративных алгоритмов (за исключением центроидного), числовые значения F_i монотонно возрастают: $0 \leq F_1 \leq F_2 \leq ... \leq F_{n-1}$ [26], [27].

При объединении «близких» кластеров, численные значения элементов множества минимальных расстояний возрастают медленно и их монотонное изменение почти линейно (если в случае центроидного метода $F_{i-1} > F_i$, то F_i заменяется на F_{i-1}). Так продолжается до тех пор, пока объединяемые кластеры достаточно близки. В момент, когда происходит объединение уже сформированных кластеров значение минимального расстояния резко возрастает. В этом случае для определения момента завершения процесса агломеративной кластеризации можно применить «метод локтя» и, соответственно, на рассматриваемом шаге абсцисса точки изгиба графика последовательности минимальных расстояний (обозначим её буквой k) совпадёт с предпочтительным количеством кластеров. Основная концепция аналитического обобщения «метода локтя» заключается в построении квадратичных форм аппроксимационно-оценочных критериев.

Обозначим набор $y_1, y_2, ..., y_k$ - множество тренда, полученное преобразованием $y_i = F_i + q \cdot i$, где q - коэффициент тренда, i - итерация в агломеративном процессе кластеризации. Рассматривается квадратическая погрешность линейной аппроксимации по m узлам:

$$\delta_l^2(m) = \sum_{i=0}^{m-1} (a \cdot i + b - y_i)^2$$

И неполная параболическая аппроксимация по m узлам:

$$\delta_q^2(m) = \sum_{i=0}^{m-1} (a \cdot i^2 + d - y_i)^2$$

Положим, что m = 4. Обозначим рассматриваемые на некоторой итерации узлы: $\hat{y_0}, \hat{y_1}, \hat{y_2}, \hat{y_3}$

Тогда, в качестве критерия используем разность двух погрешностей:

$$\delta^2(4_0) = \delta_l^2(4_0) - \delta_q^2(4_0) = \frac{1}{245}(19\hat{y_1}^2 - 11\hat{y_2}^2 + 41\hat{y_3}^2 + 12\hat{y_1}\hat{y_2} - 64\hat{y_1}\hat{y_3} - 46\hat{y_2}\hat{y_3})$$

Соответственно, характер возрастания числовой последовательности y_n изменится с линейного на параболический на k-ой итерации, если для узлов y_k, y_{k-1}, y_{k-2}, y_{k-3} справедливо неравенство $\delta > 0$, а для узлов y_{k-1}, y_{k-2}, y_{k-3}, y_{k-4} справедливо неравенство $\delta \leq 0$.

2.3 Применение методов к данным

Метод «AgglomerativeClustering», представленный в библиотеке «scikit-learn», в качестве параметров может принимать число кластеров и тип связи. В рассмотренной статье было получено 4 кластера. Будем предполагать, что это и будет оптимальным числом. В результате его применения получаем метки кластеров. В случае, когда параметр не указан - результат будет представлен в виде двух кластеров, что не представляет ценности для исследователя.

Результат после применения метода одиночной связи: кластеры размерами $(1, 1, 2, 78)$. Посмотрим на получившуюся дендограмму разбиения 2.

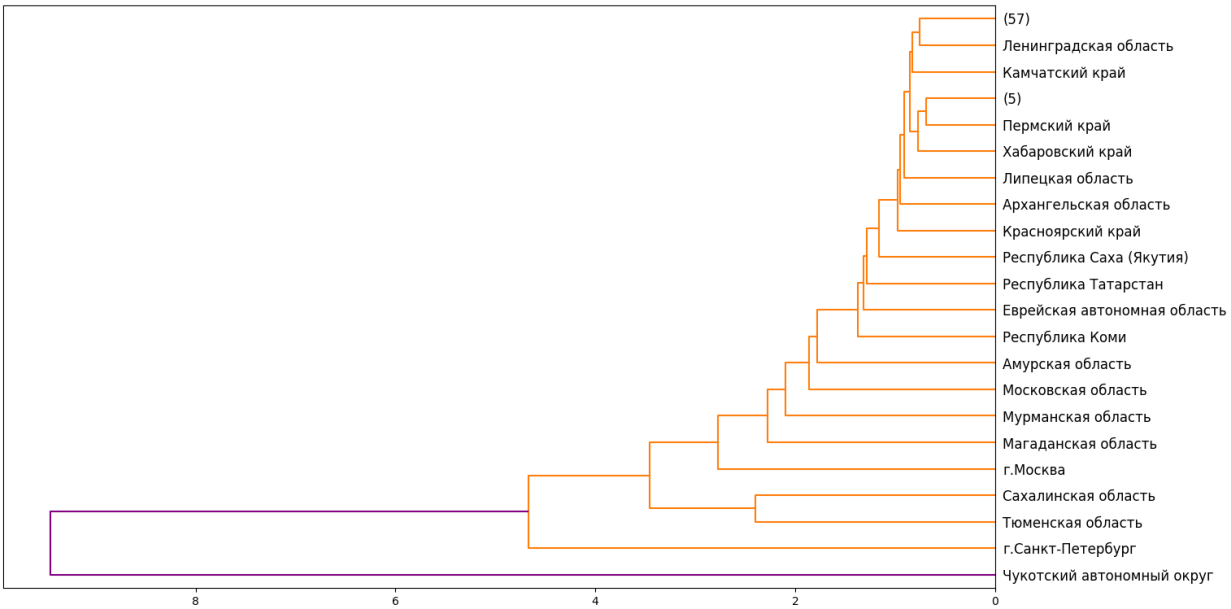

Рис. 2: Дендограмма (одиночная связь)

Можно заметить, что Санкт-Петербург и Чукотский АО оказались единственными элементами в двух кластерах, что указывает на так-называемый «цепной эффект» (chaining effect), которому подвержен алгоритм на основе одиночной связи [28]. Критерий поиска расстояния между кластерами является локальным. Так, объекты, которые сами по себе не обладают большим сходством, оказываются помещенными в один кластер. Такие последовательные объединения представляют из себя «цепочку», которую мы наблюдаем на дендограмме.

Рассмотренный эффект может быть вызван недостаточной чувствительностью кластеризационного алгоритма к отдельным объектам или шуму в данных.

Проанализируем результаты, полученные после применения метода полной связи. Получаем результат $(1, 2, 3, 76)$. Рассмотрим подробнее получившуюся дендограмму 3.

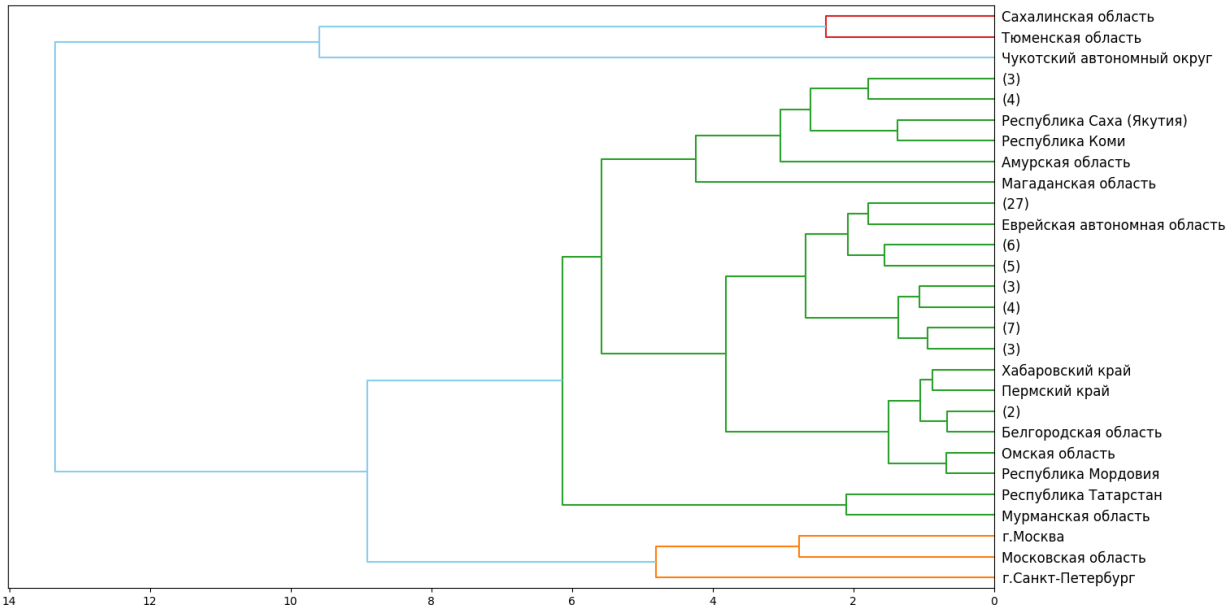

Рис. 3: Дендограмма (полная связь)

Визуальный анализ даёт понять, что были сформированы два эталонных кластера: (Сахалин, Тюмень, Чукотка), (Москва, Московская обл., Санкт-Петербург). Оставшееся большинство субъектов объединено в один большой кластер.

В отличие от рассмотренного ранее метода одиночной связи, кластеры получились более компактные и связанные. Более того, благодаря данному

способу не возникает «цепной эффект». Но, в данном случае слишком много объектов оказалось в одном кластере. Судя по дендограмме можно разбить его на три подкластера. В таком случае в одном кластере окажутся два объекта: Татарстан и Мурманская обл., что не является нормальным с точки зрения размера кластера.

Метод средней связи не является столь популярным, как те, что рассмотрены выше. Но тем не менее является наиболее предпочтительным, так как менее всего подвержен влиянию зашумленных данных [27]. На данных о субъектах получаем $(1, 3, 1, 77)$.

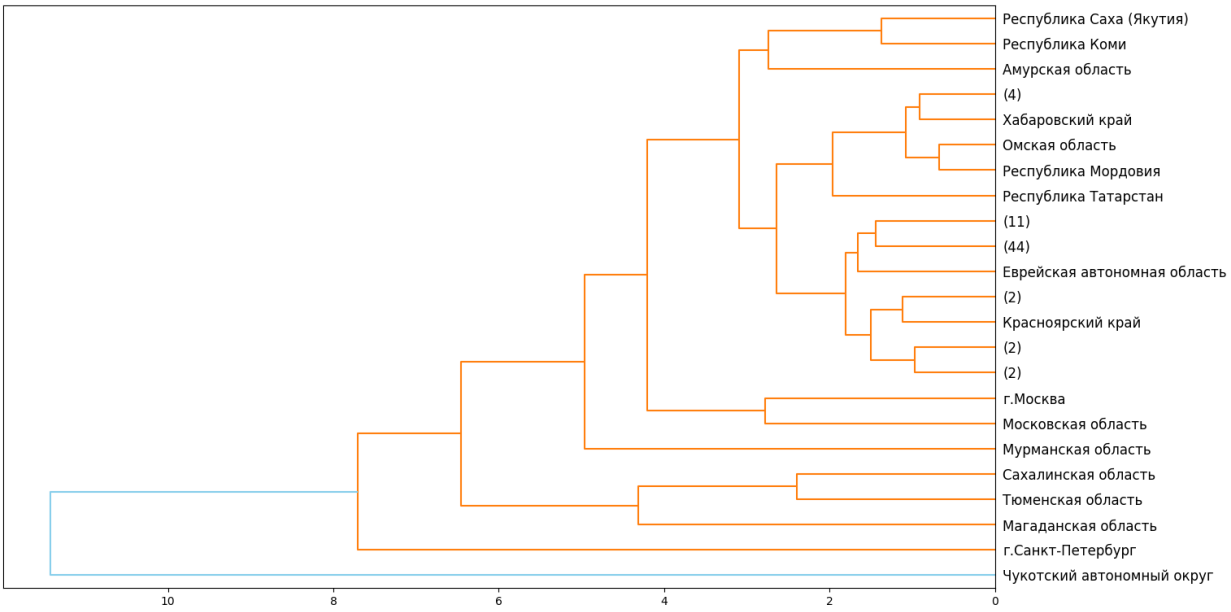

Рис. 4: Дендограмма (средняя связь)

Визуально анализируя дендограмму 4, замечаем, что Москва и Московская область были объединены в один кластер в середине агломеративного процесса. Большая часть субъектов по началу хоть и была поделена на два кластера, но в конце всё же был сформирован один большой кластер.

Центроидный метод и метод Варда. Для данных типов связи реализован способ нахождения числа кластеров. Рассмотрим центроидный метод. Так как в бибилиотеке «scikit-learn» нет соответствующего значения параматра в методе «AgglomerativeClustering», метод был запрограммирован отдельно на языке Python с использованием библиотеки pandas. Начнём с построения последовательности минимальных расстояний $F_1, F_2, ..., F_{81}$. Будем использовать критерий построений по четырём точкам.

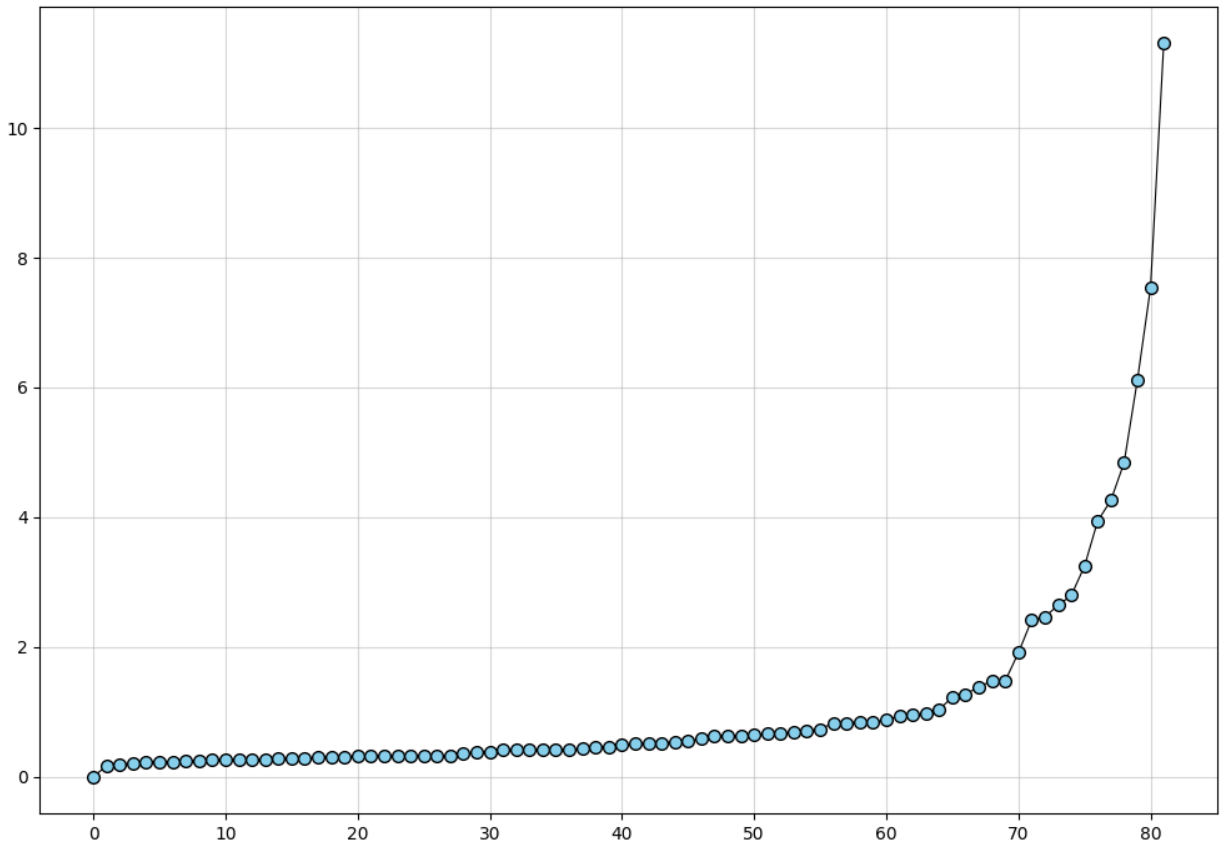

Рис. 5: График последовательности минимальных расстояний (центроидный метод)

Визуально анализируя график 5, можно заметить две области, на каждой из которых заметен переход с линейной аппроксимации на параболическую. Исследуем промежутки для различных значений параметров q. В зависимости от промежутка, будет получено различное число кластерных групп: Q_1, Q_2, ..., Q_{e-1}, Q_e (при $q \in Q_e$ все объекты объединены в один кластер). Наиболее предпочтительным является Q_{e-2}. Таким промежутком в данном случае является $[0.5, 0.73]$. Количество кластеров при этом будет равно 4. Рассмотрим дендограмму такого разбиения 6.

В результате получаем набор $(1, 1, 3, 77)$. В отдельные кластеры были выделены Санкт-Петербург и Чукотский АО; в одном кластере оказались Тюменская, Магаданская и Сахалинская области — есть совпадения с исходным разбиением. Остальные же элементы были сформированы в один большой кластер.

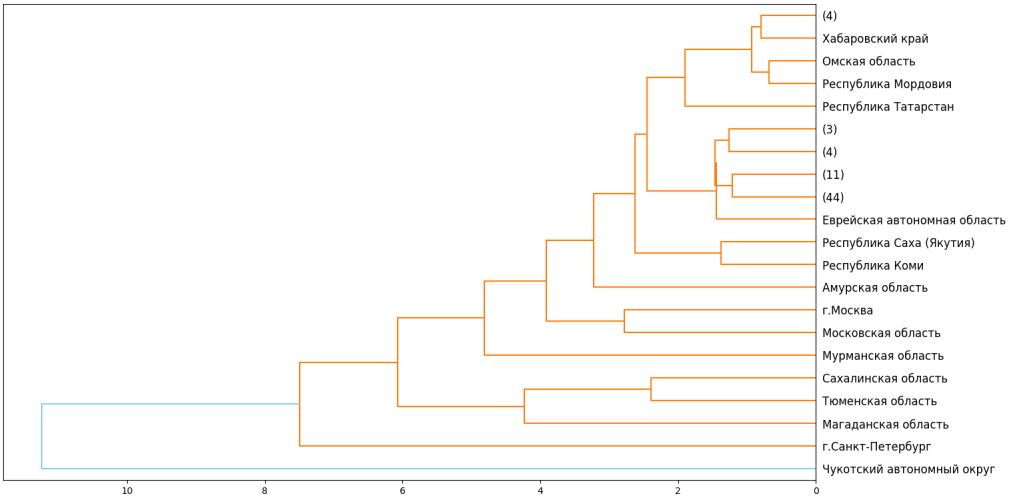

Рис. 6: Дендограмма (центроидный метод)

Применим метод Варда. Строим график последовательности минимальных расстояний 7. Выполнение расчётов занимает достаточно много времени. При $q \in [0.5, 1.2]$ получен результат в 8 кластеров. При $q \in [1.3, 3]$ получим 3 кластера. При $q \in [4, 25]$ формируются два кластера. Соответственно промежутком Q_{e-2} считается тот, при котором получено количество кластеров равное 3.

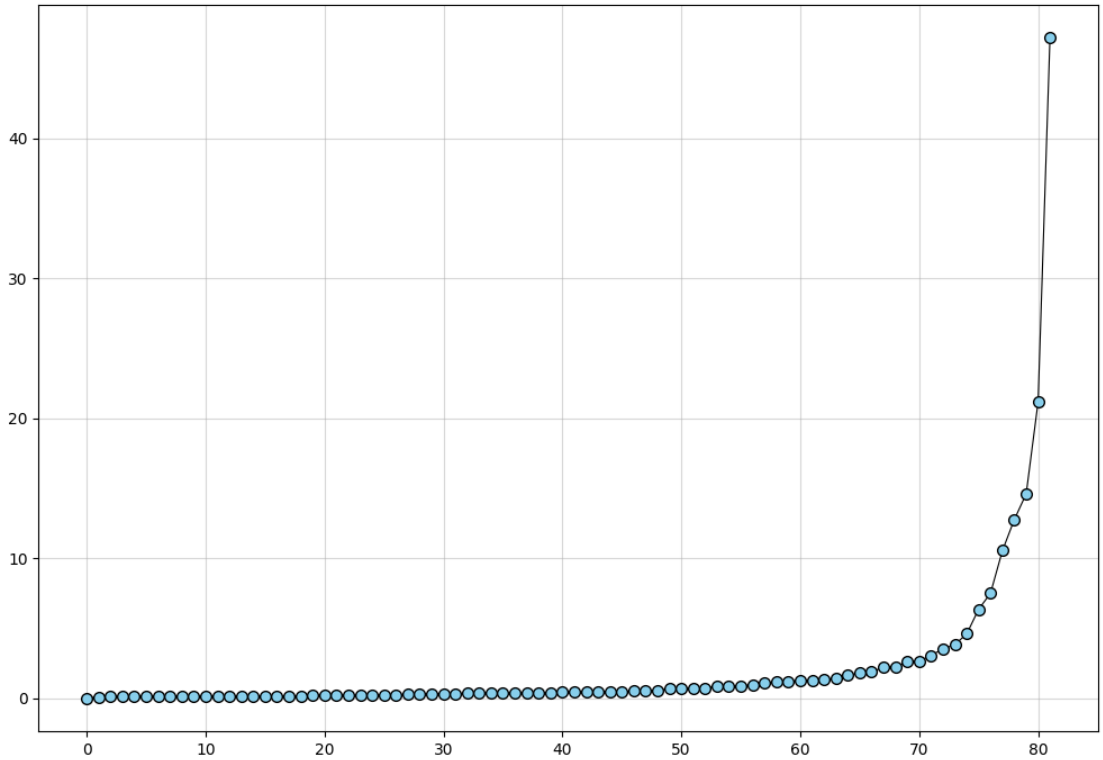

Рис. 7: График последовательности минимальных расстояний (метод Варда)

Построив дендограмму 8, наглядно видим полученное разбиение. По ней можно заметить 5 сформированных кластеров. Так, совмещая результаты, полученные визуальным анализом дендограммы и программной реализацией метода, стоит обратить внимание, что в случае $k = 8$ присутствуют совпадающие кластеры. Объединяя два кластера: (Чукотский АО) и (Тюменская обл., Сахалинская обл.), получим итоговое число кластеров, равное 7.

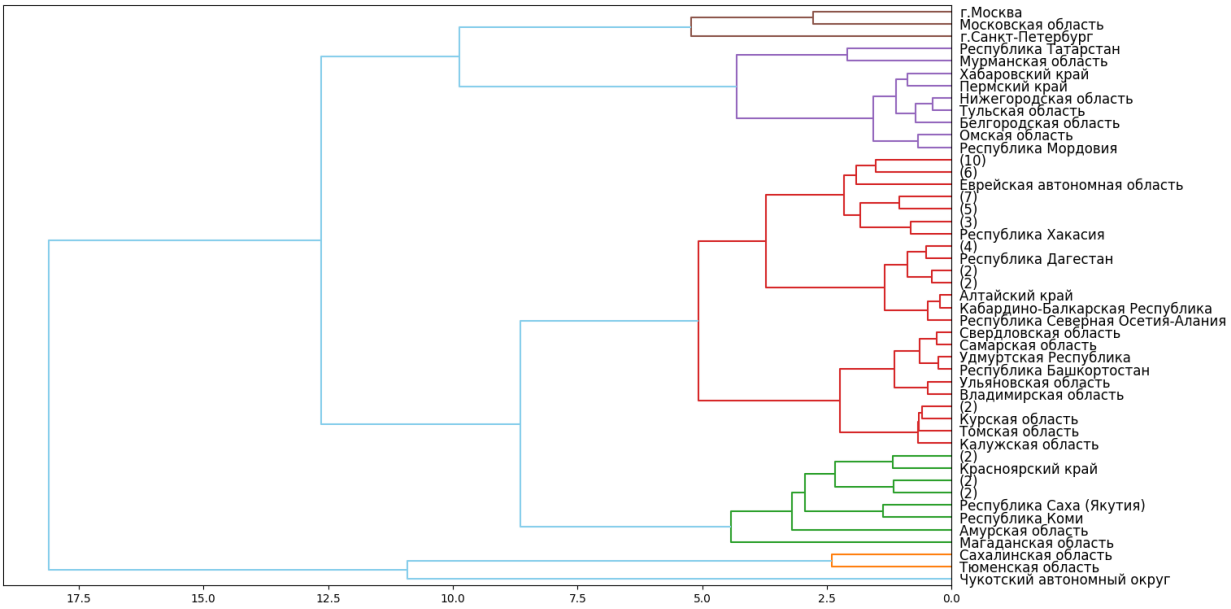

Рис. 8: Дендограмма (метод Варда)

Получившиеся группы представлены на Таблице 1. Проведя статистический анализ данных в каждом кластере, построив распределение средних значений по признакам для каждого кластера, приходим к следующим результатам:

1. Три субъекта, вошедшие в первый кластер имеют достаточно высокие экономические показатели и наивысший результат по значению внутренних затрат на исследования и разработки. При этом коэффициент изобретательной активности (КИА) является самым низким среди всех.

2. Во втором кластере оказались регионы с наименьшими показателями по всем признакам.

3. Третий кластер объединил субъекты с низкими значениями экономических признаков, но вторым по величине показателем КИА.

4. Сформирована группа из трёх основных субъектов: Москва, Московская обл. и Санкт-Петербург. Она имеет наивысшее значение КИА и второе по величине значение признака объема отгруженных инновационных товаров, работ и услуг.

5. Субъекты, попавшие в пятый кластер имеют значения ниже среднего по всем признакам в целом.

6. Шестой кластер представляет 9 субъектов с наивысшим значением объема отгруженных инновационных товаров, работ и услуг на душу населения и средними значениями всех остальных признаков.

7. В седьмом кластере оказались субъекты имеющие значения экономических показателей выше среднего, а инновационных показателей ниже среднего.

Таблица 1: Результаты кластеризации (метод Варда)

	Кол-во	**Вошедшие субъекты**
1 кластер	3	Чукотский АО, Тюменская обл., Сахалинская обл.
2 кластер	18	респ. Северная Осетия-Алания, Кабардино-Балкарская респ., Алтайский край, респ. Ингушетия, Чеченская респ., респ. Дагестан, Карачаево-Черкесская респ., респ. Тыва, Забайкальский край, респ. Калмыкия, респ. Крым, респ. Адыгея, респ. Бурятия, Курганская область, респ. Алтай, респ. Хакасия, Ивановская обл., Псковская обл.
3 кластер	11	Курская обл., Томская обл., Калужская обл., Воронежская обл., Новосибирская обл., Владимирская обл., Ульяновская обл., респ. Башкортостан, Удмуртская респ., Самарская обл., Свердловская обл.
4 кластер	3	г. Санкт-Петербург, Московская обл., г. Москва
5 кластер	27	Еврейская АО, респ. Карелия, Приморский край, Калининградская обл., Астраханская обл., г.Севастополь, Новгородская обл., Краснодарский край, Кировская обл., респ. Марий Эл, Костромская обл., Чувашская респ., Орловская обл., Тамбовская обл., Волгоградская обл., Саратовская обл., Ростовская обл., Пензенская обл., Рязанская обл., Тверская обл., Ярославская обл., Оренбургская обл., Челябинская обл., Брянская обл., Ставропольский край, Смоленская обл., Кемеровская обл.-Кузбасс
6 кластер	9	Мурманская обл., респ. Татарстан, респ. Мордовия, Омская обл., Белгородская обл., Тульская обл., Нижегородская обл., Пермский край, Хабаровский край
7 кластер	11	Красноярский край, Липецкая обл., Архангельская обл., Вологодская обл., Иркутская обл., Ленинградская обл., Камчатский край, Магаданская обл., Амурская обл., респ. Коми, респ. Саха (Якутия)

Выводы

В ходе работы был исследован набор экономических и инновационных показателей субъектов РФ. Полученные данные были предобработаны и в дальнейшем использовались в качестве тестовых. Была рассмотрена иерархия типов кластеризации, основные меры оценки качества, проведено сравнение различных методов кластеризации. Наиболее популярные методы были применены к данным о субъектах Российской Федерации, описан результат работы каждого из них и возникающие проблемы. Вероятнее всего, в статье, используемой в качестве информационной базы, разбиение было выбрано как самое частое после многократного повторения алгоритма k-средних. Разбиения, имеющие наилучшую оценку по метрикам не подошли, так как выделяют в отдельные кластеры выбросы либо концентрируют в одном кластере большую часть элементов.

Во второй главе были описаны и рассмотрены все типы иерархической кластеризации в зависимости от способа нахождения расстояния между кластерами. Также запрограммирован алгоритм для остановки агломеративного процесса и применен к центроидному методу и методу Варда. В результате визуального анализа дендограмм и результатов, полученных программно, наилучшим образом показал себя метод Варда, не выделяя в отдельные кластеры небольшое количество элементов и формируя некоторые кластеры, совпадающие с результатами, полученными экспертами-экономистами.

Заключение

Результаты данной работы будут полезны экспертам и разработчикам программных средств типологизации административно-территориальных образований РФ или же других схожих по структуре объектов обследования, так как здесь рассмотрены универсальные, наиболее популярные методы кластерного анализа, а также выделены их достоинства и недостатки.

Программная реализация метода остановки агломеративного процесса будет полезна в задачах, обрабатывающих большие объемы исходных данных, особенно в случае невозможности визуального анализа дендограмм.

В дальнейшем можно продолжить начатое исследование, дополнив состав исходных данных актуализированными социально-экономическими характеристиками развития регионов, и провести кластеризацию временных рядов.

Список литературы

[1] Виолин С.И. Типологизация регионов как основа для проведения дифференцированной государственной региональной политики [Электронный ресурс] // Региональная экономика и управление: электронный научный журнал. 2018. № 2 (54). URL: https://eee-region.ru/article/5406/

[2] Алтунина В.В., Анучина Д.А. Классификация регионов Российской Федерации в контексте пространственной поляризации // Экономика, предпринимательство и право. 2022. Том 12. № 5. с. 1453-1474. doi: 10.18334/epp.12.5.114641

[3] Заварухин В.П., Чинаева Т.И., Чурилова Э.Ю. Регионы России: результаты кластеризации на основе экономических и инновационных показателей // Статистика и Экономика. 2022. № 19(5). с. 35-47. https://doi.org/10.21686/2500-3925-2022-5-35-47

[4] Абашкин В.Л., Абдрахманова Г.И., Бредихин С.В. и др.; под ред. Гохберга Л.М. Рейтинг инновационного развития субъектов Российской Федерации. Вып. 8 // Нац. исслед. ун-т «Высшая школа экономики». 2023. с. 260. 80 экз. ISBN 978-5-7598-3000-9 (в обл.)

[5] Мыслякова Ю.Г. Разработка типологии регионов по их предрасположенности к научно-технологическому развитию // Экономика и управление. 2021. № 27 (10). с. 775-785. https://doi.org/10.35854/1998-1627-2021-10-775-785

[6] Тяпушова Е.В., Шеховцева Л.С. Исследование инновационного развития и типология регионов на основе интегральной оценки их конкурентоспособности // Известия Уральского государственного экономического университета. 2011. № 2 (34). с. 83-91.

[7] Орлова И.В., Филонова Е.С. Кластерный анализ регионов ЦФО по социально-экономическим и демографическим показателям // Статистика и Экономика. 2015. с. 111-115. https://doi.org/10.21686/2500-3925-2015-5-136-142

[8] Шматко А.Д., Губин С.В. Кластерный анализ инновационного потенциала субъектов РФ // Управленческое консультирование. 2020. № 3. с. 61-72. doi: 10.22394/1726-1139-2020-3-61-72

[9] Шамрай-Курбатова Л.В., Леденева М.В. Кластерный анализ субъектов РФ по уровню инновационной активности // Бизнес. Образование. Право. 2021. № 1 (54). с. 88-97. doi: 10.25683/VOLBI.2021.54.174

[10] Jain A.K., Murty M.N., Flynn P.J. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, № 3

[11] Arthur D., Vassilvitskii S. K-means++: the advantages of careful seeding // Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms SODA '07, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2007), p. 1027-1035

[12] Ester M., Kriegel H., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) / Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. — AAAI Press, 1996. p. 226–231

[13] Сивоголовко Е. В. Методы оценки качества четкой кластеризации // Компьютерные инструменты в образовании, № 4 2011.

[14] Понятия и определения (ВРП) [Электронный ресурс] // Федеральная служба государственной статистики
URL: https://rosstat.gov.ru/statistics/accounts (Дата обращения 24.01.2024)

[15] Понятия и определения (Инвестиции в основной капитал) [Электронный ресурс] // Федеральная служба государственной статистики
URL: https://rosstat.gov.ru/investment_nonfinancial (Дата обращения 24.01.2024)

[16] Понятия и определения (Основные фонды) [Электронный ресурс] // Федеральная служба государственной статистики
URL: https://rosstat.gov.ru/folder/14304 (Дата обращения 24.01.2024)

[17] Коэффициент изобретательской активности в регионах российской федерации // Аналитические исследования сферы интеллектуальной собственности, 2021.

[18] Изобретения и полезные модели [Электронный ресурс] // Федеральный институт промышленной собственности URL: https://www.fips.ru/to-applicants/inventions/

[19] Методология (Внутренние затраты на научные исследования и разработки) [Электронный ресурс] // Федеральная служба государственной статистики URL: https://rosstat.gov.ru/free_doc/new_site/business/nauka/mnayka7.htm

[20] Методологические пояснения. (Объем отгруженных товаров собственного производства, выполненных работ и услуг) [Электронный ресурс] // Федеральная служба государственной статистики
URL: https://rosstat.gov.ru/bgd/regl/b08_48/IssWWW.exe/Stg/metod.htm

[21] Старовойтов, В. В. Нормализация данных в машинном обучении // Информатика. 2021. Т. 18, № 3. с. 83–96. https://doi.org/10.37661/1816-0301-2021-18-3- 83-96

[22] Вирт Н. Алгоритмы и структуры данных. СПб.: Невский диалект, 2001. – 352 с.

[23] Вставская Е. Структуры данных: деревья, [Электронный ресурс] URL: https://prog-cpp.ru/data-tree/

[24] Lance G.N., Williams W.T.; A General Theory of Classificatory Sorting Strategies // 1. Hierarchical Systems, The Computer Journal, Volume 9, Issue 4, 1 February 1967, p. 373-380

[25] Орехов А.В. Марковский момент остановки агломеративного процесса кластеризации в евклидовом пространстве // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2019. Т. 15. Вып. 1. с. 76–92. https://doi.org/10.21638/11702/spbu10.2019.106

[26] Lance G.N., Williams W.T. A general theory of classificatory sorting strategies, 1. Hierarchical Systems // The Computer Journal, 1967, Vol. 9, p. 373–380.

[27] Milligan G.W. Ultrametric hierarchical clustering algorithms // Psychometrika, 1979, 44, p. 343-346

[28] Manning C.D., Raghavan P., Schütze H., Introduction to Information Retrieval // Cambridge University Press. 2008

[29] Jarman A.M. Hierarchical Cluster Analysis: Comparison of Single linkage, Complete linkage, Average linkage and Centroid Linkage Method // 2020 doi: 10.13140/RG.2.2.11388.90240

Приложения

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

class Hierarchical_clustering():
    def __init__(self, data, method='centroid'):
        self.data = data
        self.method = method
        self.transformed_data = None
        self.labels = None
        self.list_of_distances = None
        self.predicted_labels = []

    def find_labels(self, num_of_the_column):
        '''Saving the labels column'''
        self.labels = self.data[self.data.columns[num_of_the_column]]
        self.data = self.data.drop(self.data.columns[num_of_the_column], axis=1)

    def preprocess(self, scaler = StandardScaler):
        '''Preprocessing the data'''
        transformer = scaler().fit(self.data)
        transformed_data = transformer.transform(self.data)
        self.transformed_data = pd.DataFrame(transformed_data)
        self.transformed_data.columns = self.data.columns

    def Markov_moment(self, q):
        '''Get the appropriate clustering'''

        def delta_2(li):
            '''Approximation error'''
            new_li = np.array(li) - li[0]
            res = (1/245) * (19 * new_li[1] ** 2 - 11 * new_li[2] ** 2 + 41 * new_li[3]
    ** 2 + 12*new_li[1]*new_li[2]-64*new_li[1]*new_li[3]-46*new_li[2]*new_li[3])
            return res
        def delta(li, q):
            '''Convert to trend set'''
            res = []
            for i in range(len(li)):
                res.append(li[i] + q*i)
            return res
        flag = True
        list_of_distances = [0]
```

```
    df_start = self.transformed_data.copy()
    elements = df_start.index
    for _ in range(self.data.shape[0] - 1):
        distance_df = pd.DataFrame(np.zeros((len(elements), len(elements))))
        distance_df.index = elements
        distance_df.columns = elements
        for i in range(distance_df.shape[0]):
            obj_1 = [int(f) for f in str(distance_df.index[i]).split()]
            for j in range(distance_df.shape[0]):
                obj_2 = [int(f) for f in str(distance_df.index[j]).split()]
                # count the distance between clusters
                if obj_1 == obj_2:
                    distance_df[distance_df.index[i]][distance_df.index[j]] = 0
                else:
                    centroid_1 = sum(list(map(lambda x: df_start.iloc[x, :] / len(
obj_1), obj_1)))
                    centroid_2 = sum(list(map(lambda x: df_start.iloc[x, :] / len(
obj_2), obj_2)))
                    if self.method == 'ward':
                        distance_df[distance_df.index[i]][distance_df.index[j]] = (
len(obj_1 * len(obj_2))) / (len(obj_1) + len(obj_2)) * np.linalg.norm(centroid_1 -
centroid_2)
                    else:
                        distance_df[distance_df.index[i]][distance_df.index[j]] = np.
linalg.norm(centroid_1 - centroid_2)
        # finding the minimum distance
        x_coord = 0
        y_coord = 0
        min_distance = 10 ** 5 + 1
        for i in range(distance_df.shape[0]):
            for j in range(distance_df.shape[1]):
                if distance_df.iloc[i, j] < min_distance and distance_df.iloc[i, j]
!= 0:
                    min_distance = distance_df.iloc[i, j]
                    x_coord = i
                    y_coord = j
        if self.method == 'centroid' and min_distance < list_of_distances[-1]:
            min_distance = list_of_distances[-1]
        list_of_distances.append(min_distance)
        trend = delta(list_of_distances, q)
        if len(trend) >= 5:
            if delta_2(trend[len(trend) - 5:len(trend) - 1]) <= 0 and delta_2(trend[
len(trend) - 4:len(trend)]) > 0 and flag:
                #print('the nature of the increase has changed')
                res_elements = list(map(lambda x: str(x).split(), elements))
                for i in range(len(self.labels)):
```

```
                        for j in range(len(res_elements)):
                            if i in [int(elem) for elem in res_elements[j]]:
                                self.predicted_labels.append(j)
                    flag = False
            new_index = f'{distance_df.index[x_coord]} {distance_df.index[y_coord]}'
            # remove rows and columns in the similarity matrix
            distance_df = distance_df.drop(distance_df.columns[[x_coord, y_coord]], axis
    =0)
            distance_df = distance_df.drop(distance_df.columns[[x_coord, y_coord]], axis
    =1)
            # add a new cluster to the elements list
            elements = list(distance_df.index)
            elements.append(new_index)
        self.list_of_distances = list_of_distances

    def draw_sequence_of_distances(self):
        '''Display the sequence of the minimum distances'''
        x = np.arange(0, len(self.list_of_distances), 1)
        y = np.array(self.list_of_distances)
        fig = plt.figure(figsize=(10, 7))
        plt.plot(x,y, linewidth = 0.8, color = 'black', marker='o', ms= 7,
    markerfacecolor='skyblue')
        plt.grid(True, alpha = 0.5)
        plt.show()

    def print_clustering(self):
        for i in np.unique(self.predicted_labels):
            print(f'{i} cluster: {list(self.labels[self.predicted_labels == i])}')

'''Upload data'''
data_2020 = pd.read_excel('dataset_path', sheet_name='2020', usecols=list(range(0,7)))
data_2020.columns = ['subject', 'VRP', 'INVEST', 'FUNDS', 'COEFF', 'RESEARCH', 'PRODUCT']

'''Applying the method'''
clustering = Hierarchical_clustering(data_2020, 'ward')
clustering.find_labels(0)
clustering.preprocess()
clustering.Markov_moment(q = 1.2)
clustering.print_clustering()
clustering.draw_sequence_of_distances()
```

Листинг 1: Критерий остановки агломеративного процесса кластеризации