Иерархическая кластеризация - это метод анализа данных, который позволяет группировать объекты или наблюдения в иерархической структуре.

1) В начале алгоритма каждый объект состоит в своём кластере

2) Строится матрица близости для всех кластеров

3) Два ближайших кластера объединяются в один.

4) Процесс повторяется до тех пор, пока все объекты не окажутся в одном кластере.

Такие последовательные объединения носят название - агломеративный процесс. В качестве результата её применения выступает дендограмма разбиения. Это такая схема, на которой отображен агломеративный процесс. Помимо этого, при применении такого подхода существуют различные способы измерения расстояния между кластерами.

Визуализация дендограммы занимает достаточно много места и не очень различима. Но в целом мы можем провести визуальный анализ и выделить основные сформированные кластеры.

Такой подход сугубо эвристический, поэтому было бы неплохо иметь некий момент остановки иерархического процесса кластеризации. Он реализован для метода одиночной связи.

В методе рассматривается последовательность минимальных расстояний между кластерами на каждом шаге объединения двух кластеров в один. На некоторой итерации происходит скачок в величине минимального расстояния - именно на этом моменте происходит излишнее объединение уже сформированных кластеров. Аппроксимируя последовательность расстояний сначала линейной, а потом параболической функцией появляется возможность отследить этот момент программно и определить подходящее для задачи число кластеров.

Методы ‘single’ и ‘average’ не подходят для иерархической кластеризации таких данных, потому что выделяют элементы в отдельные кластеры (average) либо последовательно увеличивают один и тот же кластер, прибавляя к нему по одному субъекту.

Методы ‘complete’ и ‘ward’ дают неплохой результат - выделяют по два кластера, содержащие тройку лидеров (Москва, Московская обл., Санкт-Петербург) и тройку с высоким ВРП, но низкой инновационной активностью.

Попробовать рассмотреть все методы связи - целый раздел. В качестве расстояния брать метод ward-а или центроидный.

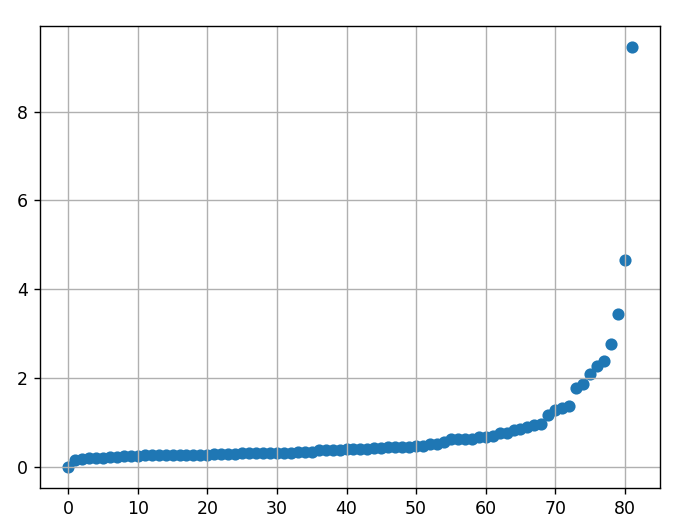
Брать по четыре точки (можно три - шесть: посмотреть разные варианты) и на этих четырёх считать погрешность - переносить их в начало координат и там аппроксимировать.

Эти приближения - это метод локтя. Аналитическое обобщение.

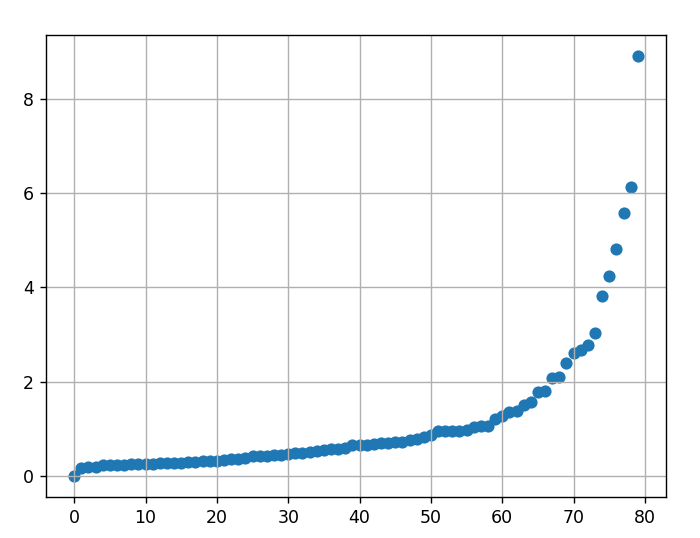
Подготовить несколько картинок дендограмм (4 штуки) сжать их, проанализировать Маргинальное распределение по координатам отдельным для нахождения наиболее подходящего места для центроидов.

Момент e^2 – 2 посмотреть работает ли он?

Single



Complete



1. Описание метода иерархической кластеризации.

Перейдём к реализации метода иерархической кластеризации. Он представляет из себя группировку объектов в иерархическую структуру. В качестве результата выступает дендограмма разбиения - схема, на которой отображена получившаяся структура.

Преимуществами являются простота в понимании результатов, возможность в визуализации иерархии кластеров. Недостатком являются вычислительные затраты в случае большого объема данных. Существует два подхода к реализации данного алгоритма кластеризации: агломеративный и дивизионный.

Первый подход заключается в том, что в начале алгоритма каждый объект состоит в отдельном кластере, затем пары кластеров объединяются по мере продвижения иерархии, пока не останется один общий кластер.

Алгоритм:

(1) В начале алгоритма каждый объект состоит в своём кластере.

(2) Строится матрица близости (proximity matrix) для всех кластеров.

(3) Два ближайших кластера объединяются в один.

(4) Процесс (шаги 2,3) повторяется до тех пор, пока все объекты не окажутся в одном кластере.

Для построения матрицы сходства необходимо использовать определённую метрику для измерения расстояния между кластерами. Существует несколько методов:

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

$$ тут формула $$

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

$$ тут формула $$

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

$$ тут формула $$

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

$$ тут формула $$

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

$$ тут формула $$, где $$ расписать чему равна каждая из составляющих $$

Для кластеризации субъектов будут рассмотрены все методы связи.

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 А.В. Орехова. Его программная реализация была осуществлена на языке Python. В качестве расстояний для метода должны быть использованы метод Варда и метод полной связи.

При втором подходе все объекты объединены в один кластер, который рекурсивно разделяется на более мелкие кластеры. Этот процесс продолжается до тех пор, пока не будет достигнут установленный критерий остановки. Примером такого подхода является метод DIANA, реализованный на языке R.

Алгоритм.

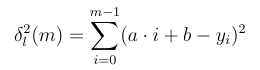
1. Все объекты объединяются в один кластер.
2. Выбирается кластер по некоторому критерию.
3. Выбранный кластер разделяется на два отдельных кластера на основе некоторого критерия.
4. Процесс (шаги 2, 3) повторяются до тех пор, пока не будет достигнут некий критерий остановки.

Марковский момент остановки агломеративного процесса кластеризации.

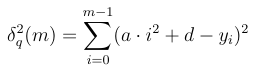
После завершения аггломеративного алгоритма кластеризации мы имеем множество минимальных расстояний {F1, F2, F3, ..., F\_n-1}, которое является линейно упорядоченным относительно числовых значений своих элементов: F1 <= ... <= Fn-1

На некоторой итерации при слиянии кластеров должен произойти резкий скачок числового значения минимального расстояния. В этот момент происходит объединение уже сформированных кластеров, что является излишним. Соответственно, на рассматриваемом шаге (обозначим его буквой k) процесс кластеризации завершается. Метод построен на том, что можно аппроксимировать этот скачок параболической функцией.

Обозначим набор {y1, y2, ..., yk} - множество тренда, полученное преобразованием y\_i = F\_i+q\*i, где q - коэффициент тренда, i - итерация в аггломеративном процессе кластеризации. Рассматривается квадратическая погрешность линейной аппроксимации по m узлам:



И неполная параболическая аппроксимация по m узлам:



Положим, что m = 4. Обозначим рассматриваемые на некоторой итерации узлы: 

Тогда, в качестве критерия используем разность двух погрешностей:



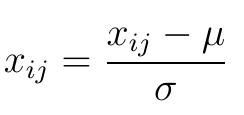
Соответственно, характер возрастания числовой последовательности y\_n изменится с линейного на параболический на k-ой итерации, если для узлов y\_k, y\_k-1, y\_k-2, y\_k-3 справедливо неравенство delta > 0, а для узлов y\_k-1, y\_k-2, y\_k-3, y\_k-4 справедливо неравенство delta <= 0.

Метод используется в случае использования центроидного метода и метода Варда.

Реализация метода иерархической кластеризации на рассматриваемых данных.

Была использована программная реализация метода из библиотеки sklearn на Python. Данные по субъектам взяты за 2020 год с сайта Росстат. Представляют собой 82 объекта (субъекты), каждый из которых описывается шестью признаками (d = 6, n = 82). Три признака описывают экономическое развитие субъекта и три признака - инновационное развитие. 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. [29] Существует несколько методов стандартизации. Наиболее популярный: Z-преобразование.

 \mu - среднее, \sigma^2 - дисперсия. Используем его для стандартизации исходных данных о субъектах.

Метод одиночной связи.

Как результат, получаем дендограмму разбиения, визуальный анализ которой позволяет понять сколько кластеров будет сформировано.