Сравнение эффективности работы программ кластеризации из библиотеки Python.

машинный эксперимент, здесь возникает вопрос, какой объем данных необходим, за сколько лет,

Можно провести эксперимент, увеличивая объем информации, так как Вы знаете правильную кластеризацию.

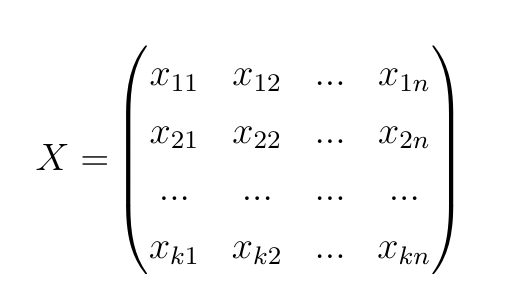
Может быть, стоит провести такой-же анализ для стран Европы, то есть выполнить еще один содержательный пример.

В данном разделе будут рассмотрены основные методы кластеризации и реализованы на языке Python и на языке R, в целях получения эталонного разбиения субъектов РФ.

Основные методы обучения без учителя реализованы на языке Python в модуле “scikit-learn". Визуализация будет осуществляться с использованием языка R. В Python нет возможности осуществить визуализацию из-за большого кол-ва признаков.

1.1 Задача кластеризации.

Кластеризация является особым видом задачи классификации. Она применяется к конечному множеству объектов, которые можно представить в виде точек n-мерного пространства. (k объектов и n признаков соответственно)



Различные подходы к классификации данных можно описать с помощью иерархии, предложенной в публикации об алгоритмах кластерного анализа. [12] Они представлены на схеме 1 (Fig.1)



Figure 1.

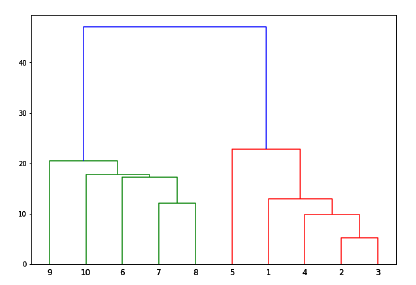
1. Эксклюзивная и неэксклюзивная (чёткая и нечёткая). Под эксклюзивной классификацией понимают такое разбиение, при котором один объект принадлежит только одному кластеру (подмножеству). При не эксклюзивной классификации объект может принадлежать двум кластерам (подмножествам) одновременно. Если рассматриваем распределение людей по возрастным группам имеет место эксклюзивная классификация. В случае классификации по имеющимся у человека аллергиям - не эксклюзивная, так как один человек может иметь сразу несколько аллергических заболеваний. В данном исследовании будет затронута только эксклюзивная классификация.
2. Контролируемое и неконтролируемое обучение. Данные понятия знакомы из машинного обучения (обучение с учителем и без учителя). К первому относится такая задача классификации, при которой известно конечное разбиение. (сам результат будет использован при обучении модели) Так, например, часто рассматривается катастрофа, случившаяся на Титанике: исследуется разбиение на выживших и погибших, которое уже известно. Неконтролируемое же обучение предполагает, что при обучении модели будет использована лишь информация, которая может быть получена при исследовании известных нам данных. Конечный результат в данном случае нам неизвестен (или же он не будет использован в обучении модели). Модели, для которых реализовано обучение без учителя используются в задаче кластеризации и в задаче уменьшения размерности. (PCA - анализ главных компонент и clustering это именно типы алгоритма обучения без учителя). !!!

Так, мы подошли к необходимой нам задаче кластеризации. Рассмотрим типы, на которые она подразделяется.

1. Задача кластеризации подразделяется по подходу к составлению алгоритма на иерархическую, секционную и плотностную. Первая производит серию разбиений, в то время как вторая лишь одно итоговое разбиение. Плотностная кластеризация основана на оценки плотности распределения.

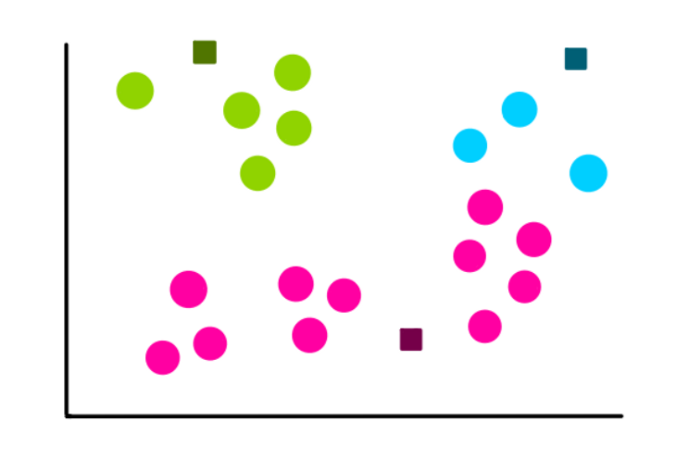
Существует два подхода к иерархической кластеризации: агломеративный и дивизионный.

Первый предполагает, что в начале алгоритма каждый объект состоит в своём кластере и затем происходит объединение, пока все объекты не будут объединены в один кластер. 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. Это методы полной и одиночной связи.

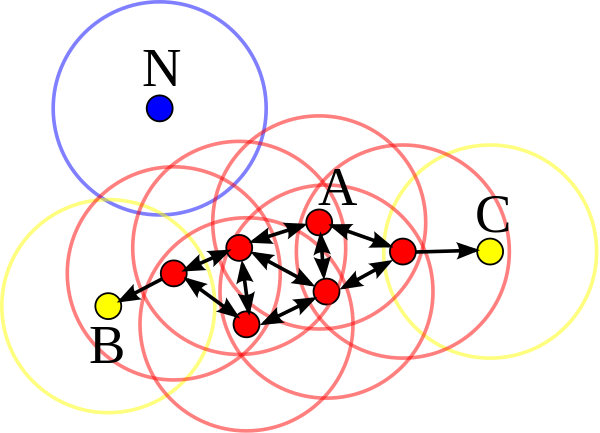


Дивизионный подход реализован посредством объединения всех объектов в один кластер и затем, по некоторому критерию, происходит разбиение до момента, пока каждый из объектов не будет состоять в отдельном кластере.

Секционная кластеризация может быть основана на вычислении квадратичной ошибки, теории графов и т.д. Существует достаточно подходов. Но остановимся на алгоритме k-means и его улучшенной версии k-means++.



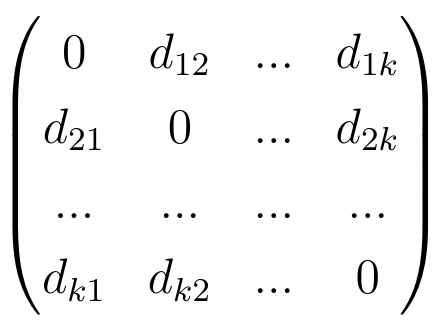
Плотностные методы кластеризации основаны на плотности точек в пространстве данных. Они не предполагают заранее определенное количество кластеров и могут обнаруживать кластеры различных форм и размеров. Наиболее популярный из них метод - DBSCAN.



1.2 Алгоритмы разбиения.

Единственным входным параметром для иерархической кластеризации выступает матрица сходства/близости (proximity matrix), элементами которой являются меры сходства или различий между объектами. Если объекты представлены точками n-мерного координатного пространства, тогда мера сходства/различия оценивается как метрическое расстояние между точками, обозначаемое. 

Соответственно сама матрица расстояний примет вид:



Конечный результат в виде одного кластера не представляет ценности для исследователя, поэтому необходим некий критерий остановки данного процесса. Исследователем может быть проведена эмпирическая оценка дендограммы и выбор числа кластеров, на котором необходимо остановить агломеративный процесс. Такой подход вполне неплох для кластеризации небольшого объема данных, но не имеет никаких математических оснований. Существует статистический критерий завершения алгомеративного процесса кластеризации, предложенный А.В. Ореховым. Именно его я и буду использовать в своей работе при реализации иерархической кластеризации. Он будет применён в разделе II. Помимо этого, при применении такого подхода существуют различные критерии для объединения двух кластеров в один. Это методы полной и одиночной связи.

Для оценки используем внутренние меры оценки качества кластеризации (те, которые основываются исключительно на входных данных):

* Метод Силуэта:

В основе идеи метода лежит вычисление коэффициентов, которые присваиваются каждому объекту в кластере и образуют так называемый **силуэт кластера**. Коэффициенты изменяются 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.

Индекс силуэта – это метрика, используемая для оценки качества кластеризации данных. Он представляет собой показатель того, насколько объект хорошо принадлежит своему кластеру по сравнению с другими кластерами.

Значение индекса силуэта колеблется от -1 до 1. Чем ближе это значение к 1, тем лучше кластеры отделены друг от друга и объекты внутри кластеров компактны. Если значение индекса ближе к -1, это указывает на то, что объекты были присвоены к неправильным кластерам.

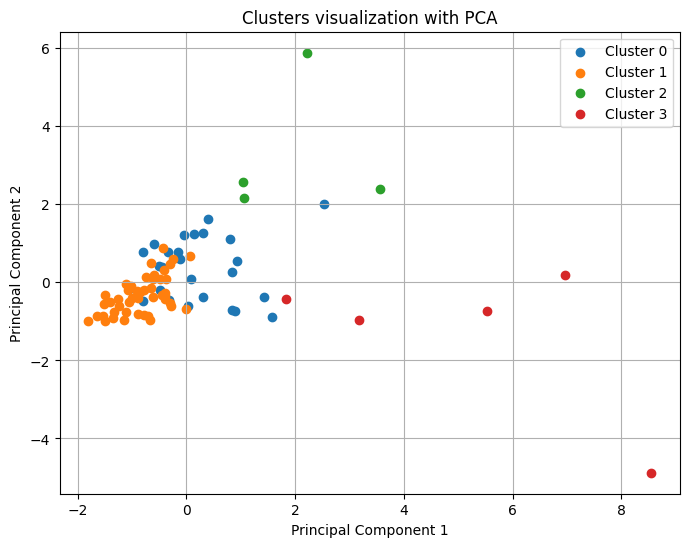
Индекс силуэта рассчитывается для каждого объекта как разность между средним расстоянием до объектов внутри того же кластера и средним расстоянием до объектов ближайшего кластера из других кластеров, деленная на максимальное из двух значений. Для оценки качества кластеризации можно усреднить значения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

Индекс силуэта является полезным инструментом при выборе количества кластеров, а также для сравнения различных алгоритмов кластеризации.

Разбиение, полученное в статье Заварухина:

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)

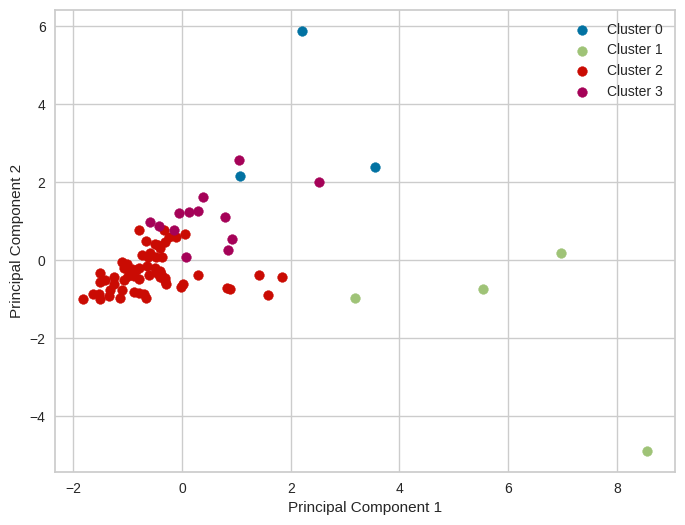
Визуализация этого и последующих разбиений реализуется посредством метода главных компонент (PCA). Он позволяет визуализировать данные в пространстве меньшей размерности и увидеть, как объекты распределены по кластерам.

  
Изучив достаточное количество статей по типологизации регионов, чаще всего в каждой из них присутствуют два почти неизменных кластера - это (2) и (3).

Применение метода k-means.

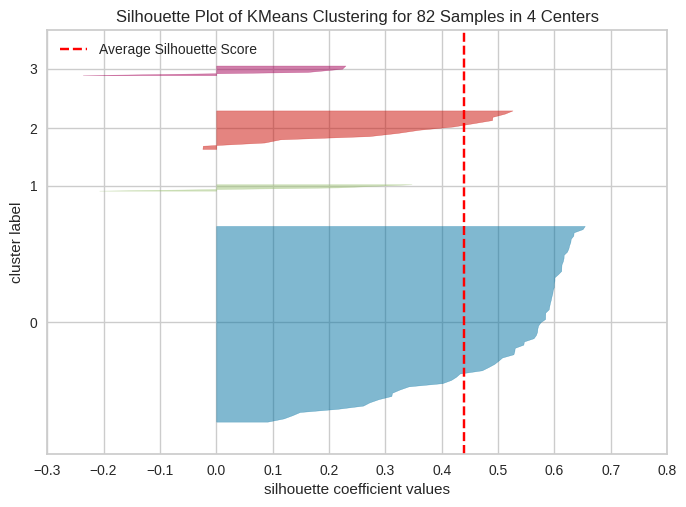
Используем метод KMeans из библиотеки sklearn. У него есть достаточно большое количество параметров. Рассмотрим только те, которые были изменены: число кластеров (n\_clusters) указываем равное 4, параметр init - использовано значение “random” для рандомного выбора центроидов. Параметр n\_init отвечает 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) - мы видим это на визуализации. Следовательно данный параметр должен быть ненулевым, выбрано значение 5. Также задано максимальное количество итераций (max\_iter = 20) - данных не так много и кластеров тоже.

Провожу кластеризацию 100 раз для определения самого часто реализуемого разбиения. Им оказывается разбиение на кластеры размерами (3, 4, 13, 62). Сразу можем заметить, что оно не совпадает с разбиением, представленным в статье. Визуализация самого частого разбиения:

  
  
Как мы видим, кластеры (1) и (0) почти совпадают с исходными. Москва, Московская обл. и Санкт-Петербург образуют группу лидеров инноваций, но не включают в себя Татарстан. Четвёрка: Тюменская область, Магаданская область, Сахалинская область, Чукотский АО образует отдельный кластер всё так же, только перестала включать в себя Еврейская АО. Ну и два оставшихся кластера (2) и (3) - мало чем похожи на эталонные кластеры, но по визуализации можно сразу заметить, что субъекты по своим признакам довольно плотно расположены друг к другу. Имеет смысл исследовать показатели в каждом кластере.

Попробуем посчитать метрики:

Метрика Силуэта(Silhouette Score) даёт неплохой результат 0.44(результат, равный 1 означает идеальную кластеризацию). Безусловно, это разбиение не является лучшим по данной метрике.



Рассмотрим кластеризацию, давшую наивысший результат по мере Силуэта. Большинство из них выделяют Чукотскую АО в один кластер, либо же выдаёт слишком большое количество элементов в кластере (объединяет субъекты из двух эталонных кластеров), а такой результат нас не устраивает.

Индекс Дэвиcа-Болдуина даёт результат: 0.92 (чем ближе к 0, тем лучше). В это же время наименьший результат: 0.79 имеют разбиения, имеющие кластер из одного элемента.

Индекс Calinski–Harabasz: 40.6 (чем больше - тем лучше кластеризация). Самый максимальный результат (47).

Посчитав V-меру, выбираем разбиение с максимальным значением - оно было получено всего лишь в двух случаях из ста: (25, 48, 5, 4).

Приходим к выводу, что получить эталонное разбиение, пользуясь внутренними метриками оценки качества или же опираясь на частоту его появления мы не можем.

Имело так же смысл проверить метрику силуэта для эталонного разбиения: 0.25. DB для него: 1.33. CH: 30. Это говорит о не самой хорошей кластеризации. Проверяя индексы DB и CH приходим к такому же неудовлетворительному результату. Возможно, стоит рассмотреть другие критерии оценки, либо же выбирать иные параметры для модели. Более того, проблема может заключаться в неточности данных или в их недостаточности. Так же не стоит забывать и про то, что за каждый год могут происходить изменения, оказывающие сильное влияние на данные. Возможно, стоит рассмотреть другие года.

Рандомный выбор центроидов в алгоритме k-means не гарантирует достижения уникальной кластеризации. Артур и Васильвицкий[27] предложили улучшенный алгоритм “k-means++”, который использует вероятностный подход для выбора начальных центроидов. Его применение осуществляется с помощью изменения параметра init.

Применение этого алгоритма даёт результат, схожий с результатом, оказавшимся лучшим по рассмотренной метрике (CH). (1, 4, 14, 63). Она не выделяет в один кластер три главных инновационных центра и в целом имеет мало что общего с эталонным разбиением.

Итак, применение метода K-means дало неплохой результат с точки зрения оценки с помощью внутренних метрик, так же это разбиение получилось самым часто реализуемым при рандомном выборе центроидов.