Федеральное государственное автономное образовательное учреждение
высшего образования

«Национальный исследовательский университет

«Высшая школа экономики»

Факультет компьютерных наук
ООП «Прикладная математика и информатика»

Отчёт о прохождении практики

Студент: Кобелев Максим Олегович

Группа: 165

Место прохождения учебной практики: НИУ ВШЭ

Руководитель:

профессор: Факультет экономических наук / Департамент математики, д.ф.-м.н. Лепский Александр Евгеньевич

Москва, 2017

Оглавление

АННОТАЦИЯ **3**

1. ОБЗОР МЕТОДОВ АГЛОМЕРАТИВНОЙ КЛАСТЕРИЗАЦИИ **4**

1.1 Агломеративная кластеризация 4

1.2 Методы агломеративной кластеризации 4

2 АГЛОМЕРАТИВНАЯ НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ **5**

2.1 Особенности и преимущества метода нечеткой кластеризации 5

2.2 Алгоритм агломеративной нечеткой кластеризации 6

3 ДЕТАЛИ И НЕКОТОРЫЕ ОБОСНОВАНИЯ РЕАЛИЗАЦИИ АЛГОРИТМА АГЛОМЕРАТИВНОЙ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ **7**

3.1 Выбор метрики 7

3.2 Вспомогательные радиальные функции 8

3.3 Оценка сложности алгоритма 8

4 СРАВНЕНИЕ АЛГОРИТМОВ НЕЧЁТКОЙ И НЕРАЗМЫТОЙ АГЛОМЕРАТИВНОЙ КЛАСТЕРИЗАЦИИ **9**

4.1 Описание данных 9

4.2 Результаты 9

4.3 Сравнительный анализ. 13

5 ЗАКЛЮЧЕНИЕ **13**

6 ИСПОЛЬЗУЕМЫЕ МАТЕРИАЛЫ **13**

Аннотация

Термин *иерархической кластеризации* объединяет целое множество алгоритмов, выполняющих разбиение данных на группы, визуализация результата которых, в конечном итоге, производится с помощью графов. Суть алгоритмов, выполняющих иерархическую кластеризацию, заключается в последовательном разделении больших кластеров на меньшие, или наоборот, в последовательном объединении меньших кластеров в большие. Таким образом, алгоритмы иерархической кластеризации подразделяются на две группы по типу используемого метода: дивизимные (разделительные) и агломеративные методы (объединительные).

1. Обзор методов агломеративной кластеризации

1.1 Агломеративная кластеризация

Благодаря тому, что в основе структуры работы иерархической кластеризации лежит построение так называемого «таксономического» дерева вложенных кластеров, возможны две различные стратегии для проведения кластеризации объектов – агломеративная (от англ. Agglomerative Nesting) и дивизивная (от англ. Divisive Analysis). Цель этого раздела – поговорить более подробно про агломеративную стратегию и обсудить схемы кластеризации, работающие в её рамках.

Таким образом, используя стратегию объединения, агломеративные алгоритмы иерархичной кластеризации во время работы постепенно объединяют объекты (кластеры) в более крупные кластеры. Общая тенденция агломеративных алгоритмов такова, что в начале каждый объект выполняет роль отдельного кластера. Далее на каждом шагу наиболее похожие кластеры объединяются в один новый кластер, который замещает объединенные. Эта операция будет производиться до тех пор, пока все объекты не буду составлять один кластер.

1.2 Методы агломеративной кластеризации

Агломеративная кластеризация строит дерево слияний, объединяя наиболее похожие кластеры. Но если «похожесть» или другими словами - расстояние между заданными объектами известно, вычислить на первом шагу два наиболее похожих кластера (найти наиближайшие точки) не составит труда, то возникает вопрос каким способом искать «похожие» между собой кластеры, состоящие из уже более чем одной точки. И здесь имеет значение метод, который будет использоваться в алгоритмах для расчёта расстояния между большими кластерами. Следующие методы являются одними из наиболее популярных правил группировки кластеров [2, 3]:

- Метод одиночной связи, (англ. Single linkage). Чаще встречается, как «метод ближайшего соседа».
- Метод полной связи, (англ. Complete linkage). Чаще встречается, как «метод дальнего соседа».
- Центроидный метод, (англ. Centroid linkage).

Подробнее познакомимся с каждым из методов:

1. Метод одиночной связи

Из другого названия этого метода – «метода ближайших соседей», понятно по какому принципу происходит поиск «похожих» или ближайших кластеров для дальнейшего их слияния в один кластер более высокого уровня. Действительно, слияние кластеров произойдет если хотя-бы по одному из элементов каждого кластера находятся достаточно близко друг к другу. Отсюда и формулируется более формальное название этого метода – метод одиночной связи, потому как для слияния двух кластеров достаточна лишь единственная связь между ними. Теперь совсем не сложно формализовать правило, по которому будет действовать метод одиночной связи – расстояние между кластерами будет определяется как минимальное расстояние между всевозможными парами двух точек, одна из которых принадлежит первому из объединяемых кластеров, а вторая – второму.

$$D(X,Y) = \min_{x \in X,\ y \in Y} d(x,y),$$

где $d(x,y)$ – расстояние между $x \in X, y \in Y$; X и Y – различные кластеры.

2. Метод полной связи

В этом методе, в противоположность методу одиночной связи, правило объединения указывает, что сходство между кандидатами на включение в существующий кластер и любым из элементов этого кластера не должно быть меньше некоторого порогового уровня. Это правило более жесткое, чем правило метода одиночной связи. В методе полной связи минимальное расстояние определяется как максимум из множества расстояний между элементом первого кластера и элементом второго кластера.

$$D(X,Y) = \max_{x\in X,\ y\in Y} d(x,y),$$

где $d(x,y)$ – расстояние между $x \in X, y \in Y$; X и Y – различные кластеры.

3. Центроидный метод

Этот метод существенно отличается от двух рассмотренных выше методов одиночной и полной связей. В них «похожие» кластеры объединяются в новый кластер высшего уровня простым объединением входящих в каждый из объединяемых кластеров точек. В центроидном методе, кластеры объединяются в новый, при этом центр нового кластера высчитывается по формуле взвешенного или невзвешенного попарного среднего, поэтому перерасчёт центра производится на основе вычисления расстояния между всеми парами точек принадлежащей каждому из объединяемых кластеров.

$$D(X,Y) = \frac{1}{|X|\cdot|Y|}\sum_{x\in X}\sum_{y\in Y} d(x,y),$$

где $d(x,y)$ – расстояние между $x \in X, y \in Y$; X и Y – различные кластеры.

2 Агломеративная нечеткая кластеризация

2.1 Особенности и преимущества метода нечеткой кластеризации

Чтобы показать очевидные преимущества метода нечеткой кластеризации, вспомним сущность кластерного анализа. Его задача – разбиение множества объектов на относительно однородные группы, которые называют кластерами. Особенностью такого разбиения должно быть то, что объекты, отнесенные к одному кластеру должны быть схожи между собой и наоборот, объекты, отнесенные к разным кластерам, должны очень сильно различаться.

Таким образом, четкая кластеризация представляет собой такое разбиение множества объектов, в которой каждый объект может относиться к одному, и только к одному кластеру.

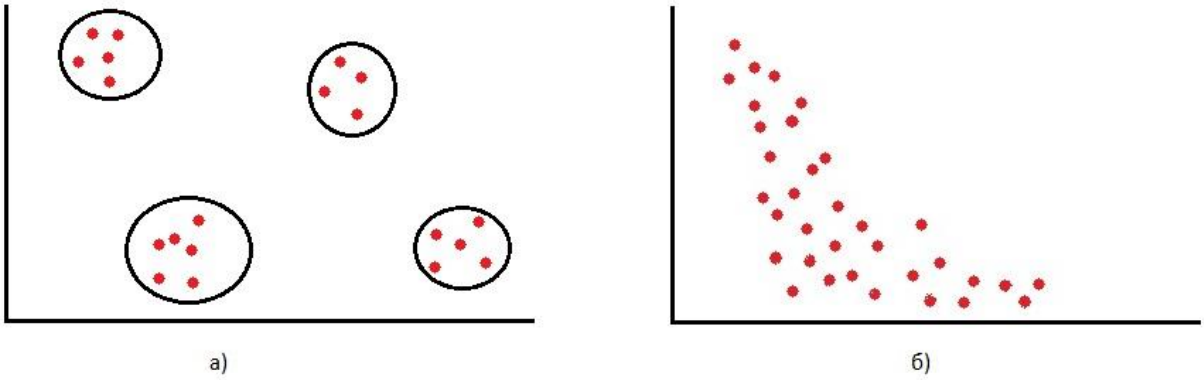


Рис 1.

Рассмотрим идеальную ситуацию, когда объекты расположены как на рисунке 1 а). В этом случае, хорошо применим алгоритм четкой кластеризации, так как видно, что объекты, явно претендующие на объединение в каждый из 4 кластеров, расположены друг относительно друга на большом расстоянии [2].

С другой стороны, на рисунке 1 б) изображена ситуация, которой свойственно встречаться на практике. Здесь очень сложно провести четкое разбиение, так как границы предполагаемых кластеров выражены неявно, а также сложно понять перед запуском алгоритма кластеризации, какое количество кластеров будет оптимально производить разбиение этих данных на группы. Именно поэтому в большинстве случаев лучше применить один из алгоритмов агломеративной кластеризации, так как он не требует задания количества кластеров в качестве входного параметра [2]. Кластеризация производится «на ходу». Использование метода нечеткой кластеризации зачастую предпочтительно, ввиду отсутствия строгого отнесения каждого объекта к какому-то одному кластеру. В нечеткой кластеризации каждый объект может одновременно принадлежать нескольким кластерам, но с разной степенью принадлежности. Здесь, «спорные» точки могут быть включены сразу в несколько кластеров.

2.2 Алгоритм агломеративной нечеткой кластеризации

1. Инициализируем множество кластеров C заданными нам точками:
   $t := 1;\ C_t = \{\{x_1\}, \dots, \{x_n\}\}$.
2. Заполнить матрицу принадлежности U.
3. *for* $t = 2 \dots n$ ($t -$ номер итерации)
4. Найти в C два ближайших кластера:
   $(P, Q) := \arg\min_{U \neq V} R(P, Q)$;
   $R_t = R(P, Q)$
5. изъять кластеры P и Q, и добавить слитый кластер $W = P \cup Q$
   $C_t := C_{t-1} \cup \{W\} \setminus \{P, Q\}$;
6. Пересчитать матрицу степеней принадлежности U.

3 Детали и некоторые обоснования реализации алгоритма агломеративной нечеткой кластеризации

Реализацию алгоритма можно посмотреть в приложенном файле ‘Fuzzy-HAC Final’. Также она доступна по короткой ссылке [http://bit.ly/2we1SwK].

3.1 Выбор метрики

В качестве метрики между объектами (точками многомерного пространства) будем использовать евклидово расстояние.

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x'_i)^2}\,.$$

Методы одиночной и полной связи не приспособлены для иерархического алгоритма нечеткой кластеризации из-за того, что каждая точка одновременно принадлежит всем кластерам, хоть и с разной степенью принадлежности. Отсюда и проблема неоднозначного отнесения объектов к имеющимся кластерам. Именно поэтому эти методы, для корректной работы, требуют некоторых модификаций, позволяющих определять, к какому кластеру принадлежит выбранная точка. Существует несколько вариантов:

- Использование порогового значения. Задаётся некоторое значение. Если функция принадлежности у данной точки на выбранном кластере приняла значение большее, чем пороговое, то считаем, что эта точка принадлежит кластеру.
- Использование максимального значения функции принадлежности. Правило будет определять принадлежность выбранной точки тому кластеру, на котором его функция принадлежности приняла максимальное значение.

Недостатки этих методов в том, что мы начинаем пренебрегать принадлежностью выбранной точки к остальным кластерам. При нескольких итерациях накапливаются ошибки, и в итоге результат произведенной кластеризации может быть очень неправдоподобным.

Именно поэтому был выбран вариант использования центроидного метода для определения близости кластеров. Я использовал метод невзвешенного попарного среднего, по формуле:

$$\delta_{ik} = \sum_{j=1}^{n} f(u_{ij}, u_{kj}),$$

где δ_{ik} - функция поиска ближайших кластеров, основанная на соответствующих строках матрицы принадлежности U для кластеров c_i и c_k, а u_{ij} – степень принадлежности $j-$го элемента $i-$му кластеру.

Здесь f(x, y) – вспомогательная функция, для вариативного поточечного сравнительного анализа кластеров на удалённость, аргументами которой являются значения функции принадлежности из промежутка $[0, 1]$. Она может принимать следующие виды [1]:

$$f_{abs}(x, y) = |x - y|, \qquad f_{sqr}(x, y) = (x - y)^2, \qquad f_{wgt}(x, y) = |x - y|(x + y)\,.$$

Использование первой функции f_{abs} будет представлять собой наиболее естественное поточечное сравнение двух строк матрицы принадлежности (кластерах), в то время как, f_{sqr} будет делать более сильный акцент на более удаленных кластерах, и наконец, f_{wgt} будет фокусироваться на больших разницах больших степеней принадлежности, поэтому приоритет будет только у ближайших соседей.

3.2 Вспомогательные радиальные функции

Агломеративная кластеризация начинает свою работу с объявления каждой точки как отдельного кластера. С этого начинается стартовая проблема принадлежности, так как в этот момент матрица принадлежности принимает вид единичной матрицы. Дело в том, что после объявления каждого объекта в качестве кластера, степень принадлежности этих объектов к своим кластерам равна 1, а по отношению к остальным – 0. Следовательно между любыми такими двумя кластерами похожих найти не получится. Назовём эту ситуацию однозначной принадлежностью. В качестве решения предлагается использовать дополнительную функцию с одним параметром размытия alpha [1].

$$f_{Cauchy}(r,\alpha) = \frac{1}{r^2+\alpha}, \qquad f_{Gauss}(r,\alpha) = e^{-\frac{r^2}{2\alpha^2}},$$

где r – расстояние между выбранной точкой и центром кластера, к которому рассчитывается её принадлежность.

Таким образом, использование этих функций с параметром $\alpha > 0$ для вычисления значения функции принадлежности не допускает появления однозначных принадлежностей, возникающих при использовании простого обратного квадратичного расстояния (в этом случае $\alpha = 0$). Это позволяет нам высчитывать значение функции принадлежности j – го элемента i – му кластеру, даже на начальном этапе по универсальным формулам:

$$u_{ij}^{(\alpha)} = \frac{f_{Cauchy}(d_{ij};\alpha)}{\sum_{k=1}^{c} f_{Cauchy}(d_{kj};\alpha)}, \qquad u_{ij}^{(\alpha)} = \frac{f_{Gauss}(d_{ij};\alpha)}{\sum_{k=1}^{c} f_{Gauss}(d_{kj};\alpha)}.$$

3.3 Оценка сложности алгоритма

Пусть n – количество объектов в представленной выборке. В таком случае имеем $n-1$ итераций объединений внешнего цикла. Внутри него очевидно, самой затратной по времени будет функция поиска ближайших кластеров *(find_similar_clusters).* По ней можно получить верхние оценки на время работы всего алгоритма. Поиск ближайших кластеров происходит за квадратичное время относительно длины массива кластеров. На первых шагах его длина равна $O(n)$, так как в момент инициализации каждый объект является отдельным кластером. Однако само сравнение кластеров реализовано за время, пропорциональное количеству заданных объектов, то есть за время $O(n)$. В таком случае функция *find_similar_clusters* будет в среднем выполняться за время $O(n^3)$. Всего за время работы алгоритма она будет вызвана $n-1$ раз. Поэтому время выполнения всего алгоритма будет составлять $O(n^4)$.

Алгоритм агломеративной нечеткой кластеризации будет работать довольно медленно из-за нестандартно реализованного сравнения кластеров, в отличии от стандартных иерархических алгоритмов неразмытой кластеризации. Это значит, что размер наших входных данных будет ограничиваться несколькими сотнями объектов [1].

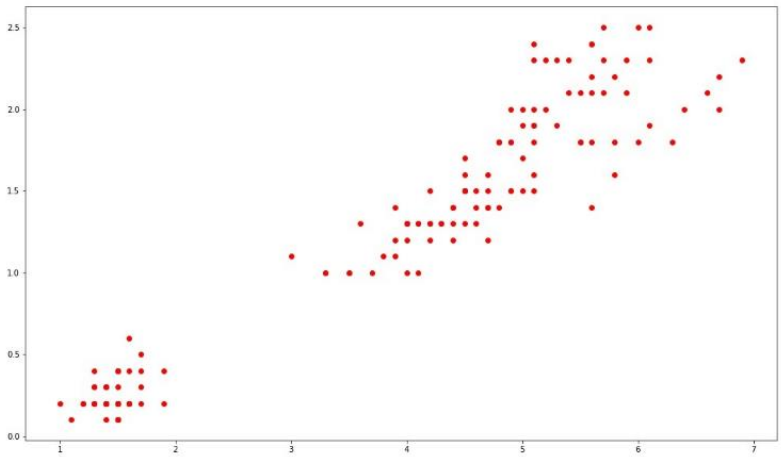
4 Сравнение алгоритмов нечёткой и неразмытой агломеративной кластеризации

Для проведения сравнительного анализа методов чёткой и нечёткой кластеризации были взяты два алгоритма: реализованный алгоритм “Agglomerative Fuzzy Clustering” [1], и стандартный алгоритм иерархической кластеризации реализованный в модуле scipy.cluster.hierarchy [4].

Предлагается анализировать полученные дендрограммы с разными метриками для нахождения расстояния между кластерами ($g_{abs}, g_{sqr}, g_{wgt}$), разными значениями оператора размытия (α), а также разными радиальными функциями (f_{Cauchy}, f_{Gauss} с параметром α).

4.1 Описание данных

В качестве входных данных была использована довольно известная выборка данных “Iris Data”. Данные были взяты с открытого репозитория “UCI machine learning” [5]. Для кластеризации, у объектов были выделены атрибуты “petal_length” и “petal_width”, соответствующие 3 и 4 измерениям. Учитывая структуру распределения объектов на плоскости, взятые атрибуты являются наиболее информативными для проведения кластеризации.



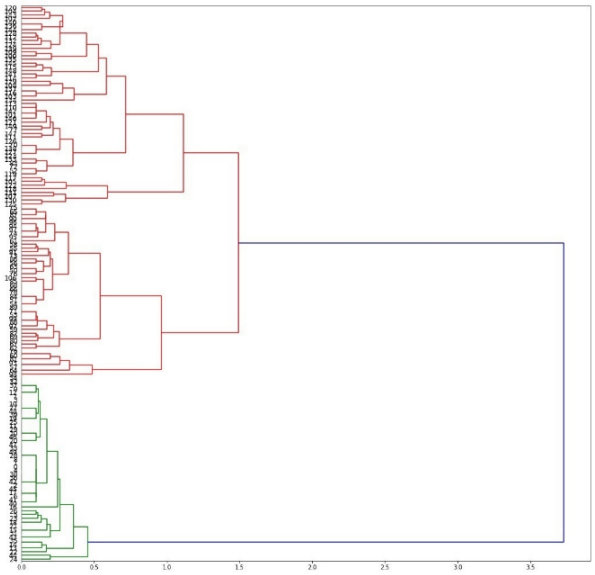
Dataset Iris [3, 4]

4.2 Результаты

Результат работы алгоритма принято демонстрировать с помощью таксономического дерева – *дендрограммы*. Дендрограмма обычно представляет собой дерево (граф без циклов), построенный по матрице мер близости. Такая древовидная диаграмма позволяет отображать результат кластеризации объектов из n-*мерного* пространства на плоскости, в 2-мерном формате.

Важное свойство, позволяющее изображать процесс кластеризации при помощи дендрограммы – свойство *монотонности* [3]. Функция расстояния R обладает свойством монотонности, если при каждом слиянии расстояние между объединяемыми объектами только увеличивается: $R_2 \leq R_3 \leq \cdots \leq R_c$, где R_t – расстояние между ближайшими кластерами, найденными на t –ом шаге (слиянии).

1. **Четкая иерархическая кластеризация стандартным алгоритмом scipy.cluster.hierarchy**

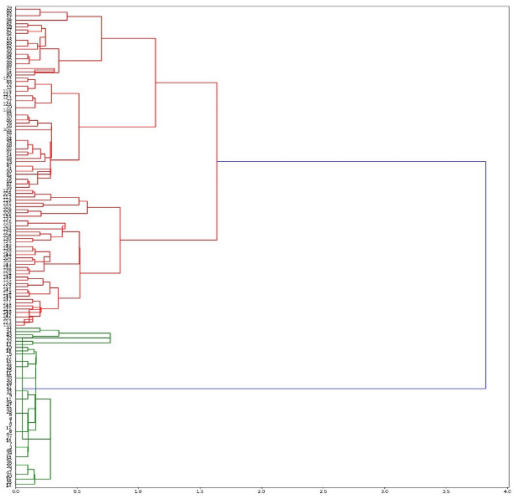


Iris dataset

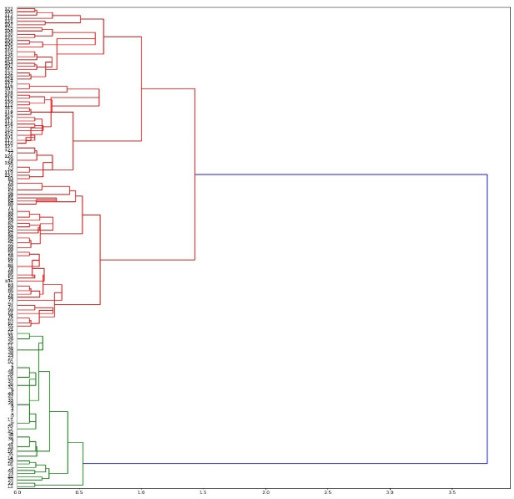
2. **Реализованный алгоритм агломеративной нечеткой кластеризации.**

Стоит заранее заметить, что среди приложенных в этой главе дендрограмм – нет правильных. Дендрограмма лишь помогает выделить детальную кластерную структуру множества объектов. Визуализация таксономических диаграмм позволяет сравнивать результаты проведенных кластеризаций, к сожалению, лишь субъективно. Также она позволяет делать некоторые выводы при сравнении проведенных кластеризаций с разными параметрами.

<u>Сравнение 1:</u>



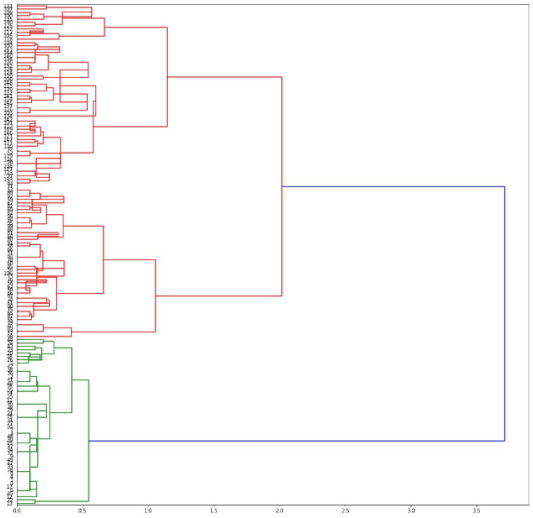
Cauchy, abs, alpha = 0.2



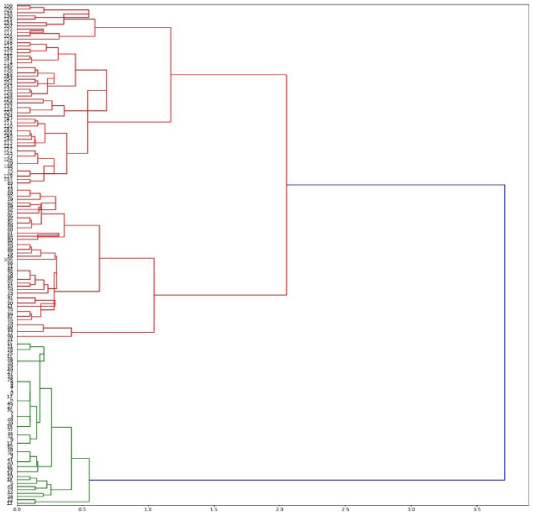
Gauss, abs, alpha = 0.2

Здесь нечёткий алгоритм с параметрами [Cauchy, abs, 0.2] объединяет далёкие объекты зелёного сегмента, что приводит к немонотонной диаграмме на ранних этапах. Красные объекты объединяются монотонно. Его двойник – алгоритм с параметрами [Gauss, abs, 0.2] делает с точностью наоборот: объединяет достаточно далёкие объекты красного сегмента, и хорошо работает на объектах зелёного.

Сравнение 2:



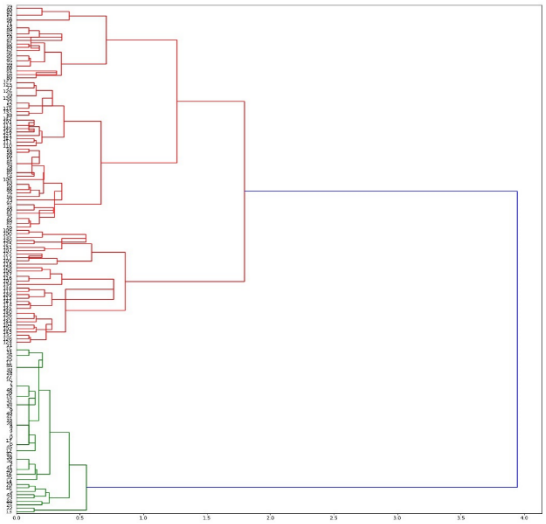
cauchy, abs, alpha = 1



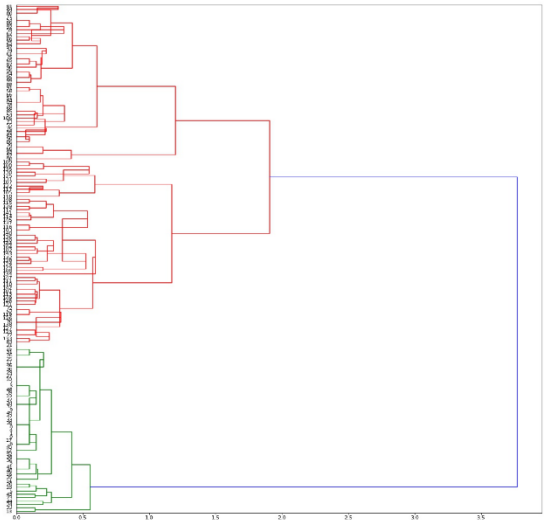
cauchy, sqr, alpha = 1

Показано, что более строгая метрика [sqr] для нахождения близких кластеров, справляется с задачей кластеризации более плавно и монотонно, чем [abs], при тех же остальных параметрах [Cauchy, 1]. Вообще sqr-метрика довольно хорошо проявила себя на всех промежуточных тестах при использовании части данных Iris Dataset.

Сравнение 3:



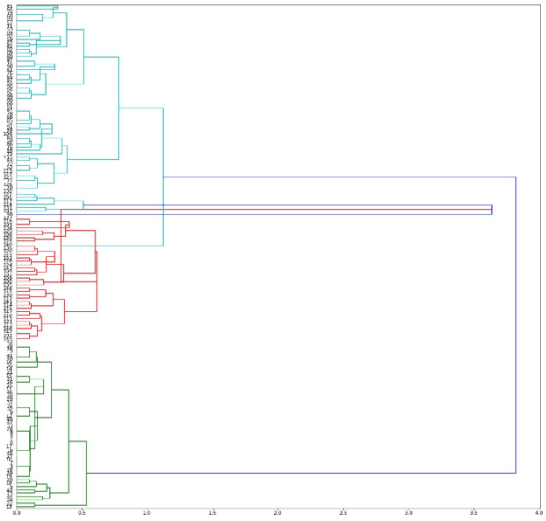
Cauchy, wgt, alpha = 0.2



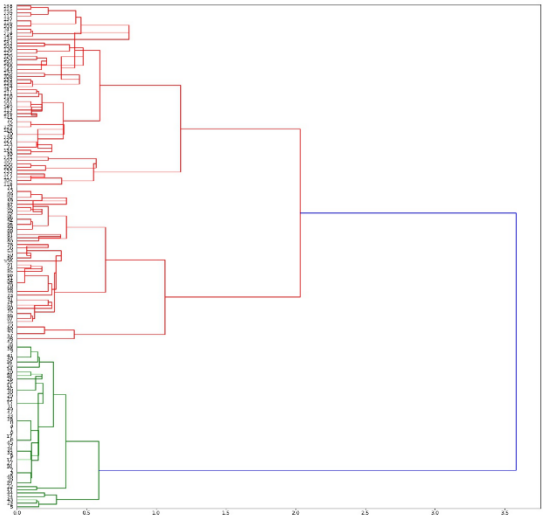
Gauss, wgt, alpha = 0.2

Это сравнение снова демонстрирует нам разницу между методами Cauchy и Gauss для решения проблем принадлежности на начальных этапах. Они одинаково кластеризовали зелёный сегмент при параметрах [wgt, 0.2], что очень удивительно, ведь зеленый сегмент сам по себе представляет собой выборку из около 50 самостоятельных объектов (точек), на плоскости расположенных в левом нижнем углу.

Сравнение 4:



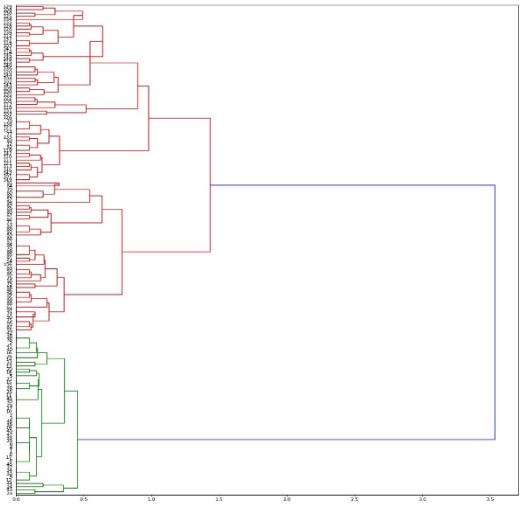
Gauss, sqr, alpha=-0.2



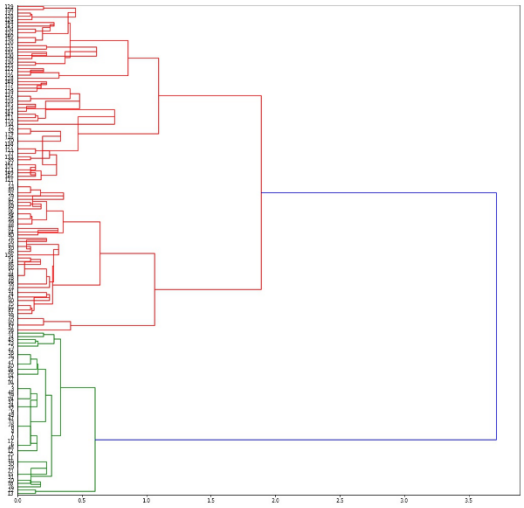
Gauss, sqr, alpha=1

Эти две дендрограммы демонстрируют нам колоссальную разницу в кластеризации на промежуточных этапах, при выборе различных значений оператора размытия. [Gauss, sqr, 0.2] показывает не очень адекватную дендрограмму, объединив очень далёкие объекты в один кластер. Расстояние между ними сравнимо с расстоянием между центрами двух кластеров на последнем этапе кластеризации, когда они сливаются в один единственный.

Сравнение 5:



cauchy, sqr, alpha=0.2



gauss, abs, alpha=1

И наконец, два алгоритма, запущенных с предельно разными параметрами – [Cauchy, sqr, 0.2] и [gauss, abs, 1] показывают практически одинаковые дендрограммы. Можно заметить, что они лишь немного отличаются своей растянутостью по горизонтальной оси, но это является следствием использования разных метрик для поиска расстояния между кластерами, ведь sqr-метрик объединяет на более ранних этапах менее удалённые кластеры, нежели abs.

4.3 Сравнительный анализ.

В итоге, были получены довольно необычные дендрограммы, с заметными клубками и слияниями в обратную сторону. Это следует из того факта, что центроидный метод поиска расстояний между кластерами не обладает свойством монотонности. К счастью, почти на всех тестах, программа сталкивалась с такой проблемой только на начальном этапе кластеризации. В среднем, на последних 10 шагах, когда уже можно явно наблюдать результат кластеризации, график на диаграмме монотонный. Заметим, что немонотонные скачки могут появляться в том случае, когда центры двух очень похожих кластеров находятся на большом расстоянии друг от друга, причем между ними также присутствует ещё один кластер. После слияния двух дальних кластеров, центр нового пересчитывается и в итоге находится уже не так далеко ко всем остальным. Расстояния между центрами образованного кластера и остальных станет гораздо меньше, чем между удалёнными центрами только что слитых кластеров, потому и происходит скачок. Он наиболее заметен в сравнении №4 с параметрами [Gauss, sqr, 0.2].

5 Заключение

Мне был предложен к рассмотрению иерархический алгоритм нечеткой кластеризации, основанный на поиске ближайших кластеров путём добавления каждому объекту параметра - степени принадлежности к каждому существующему кластеру. Сравнительный анализ не выявил явного преимущества нечеткого метода кластеризации перед неразмытым. Нечеткие алгоритмы показывают примерно схожие результаты даже при разных параметрах кластеризации. Было установлено, что для проведения качественной кластеризации объектов важно корректно подобрать сам алгоритм, его тип, метрику, используемые им методы и параметры. За время своей практики, я и старался это сделать. Для этого были исследованы не только сами принципы работы алгоритма агломеративной нечеткой кластеризации. Потребовалось также изучить и все остальные его аспекты, такие как: метрика для объектов, кластеров, методы поиска ближайших кластеров, схемы их объединения, пересчёта центров и прочих проблем, возникающих при кластеризации. Каждая вариация алгоритма была отлажена и запущена на одних и тех же данных с целью поиска оптимального выбора параметров кластеризации.

6 Используемые материалы

[1] Christian Borgelt and Rudolf Kruse (2013) **Agglomerative Fuzzy Clustering.** School of Computer Science, Otto-von-Guericke-University Magdeburg.

[2] Jae-On Kim, Charles W. Mueller (1986) **Factor Analysis:** Statistical Methods and Practical Issues (Eleventh Printing). Clustering analysis: 165-180.

[3] К.В. Воронцов (2010) **Лекции по алгоритмам кластеризации и многомерного шкалирования.** Иерархическая кластеризация: 10-15.

[4] The Scipy community (2017) **Hierarchical clustering** (web resource) [https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html].

[5] Lichman M (2013) **UCI Machine Learning Repository** [http://archive.ics.uci.edu/ml]. University of California, Irvine, CA, USA