ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА»

МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ

КАФЕДРА ГАЗОВОЙ И ВОЛНОВОЙ ДИНАМИКИ

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
(ДИПЛОМНАЯ РАБОТА)
специалиста

«Исследование эффективности различных кластерных методов для анализа структуры гамма-адронных семейств.»

Выполнил студент 623 группы
Изофатова Мария Сергеевна

_____

подпись студента

Научный руководитель:
доцент Е.А. Ильюшина

_____

подпись научного руководителя

Москва
2022

Оглавление

Глава 1

Введение

В качестве источника данных в работе используется эксперимент Памир - результат совместного исследования нескольких научных учрежденийй во главе с ФИАНом по изучению адронных взаимодействий (частиц космических лучей) методом больших РЭК (рентгеноэмульсионных камер).

1.1. Физическая постановка задачи.

Основной объект исследования – гамма-семейство - группы генетически связанных частиц с энергией выше 4 ТэВ, которые возникают в результате взаимодействия первичной частицы высокой энергии с ядрами атомов воздуха и последующего ядерно-электромагнитного каскада. Физическая задача состоит в выделении в семействе отдельных групп частиц, родившихся в результате одного последнего взаимодействия.

Для ее решения необходимо подобрать алгоритма кластеризации, который собирает эти частиц в один кластер с наибольшей эффективностью. Алгоритм выбирается на основе анализа модельных данных, и в дальнейшем применяется к экспериментальным.

Перед выбором алгоритма необходимо произвести проверку однородности экспериментальных и модельных данных.

То есть, кластеризация, в данном случае, это разбиение наблюдаемых событий на группы частиц, представляющих собой, преимущественно, продукты отдельных последних (самых нижних по высоте) сильных взаимодействий в атмосфере над установкой.

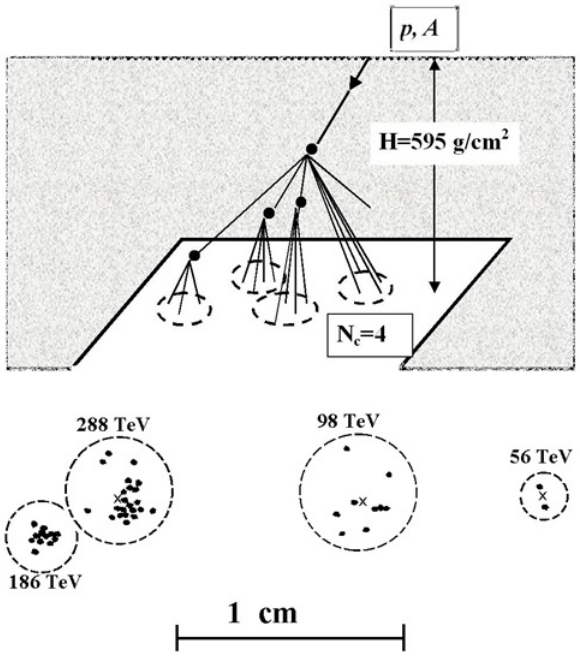


Рис. 1.1. Постановка задачи

1.1.1. Актуальность.

В результате процедуры кластеризации появляется возможность изучать частицы, образованные в более раннем поколении. Кластеризация позволяет повысить чувствительность харак-

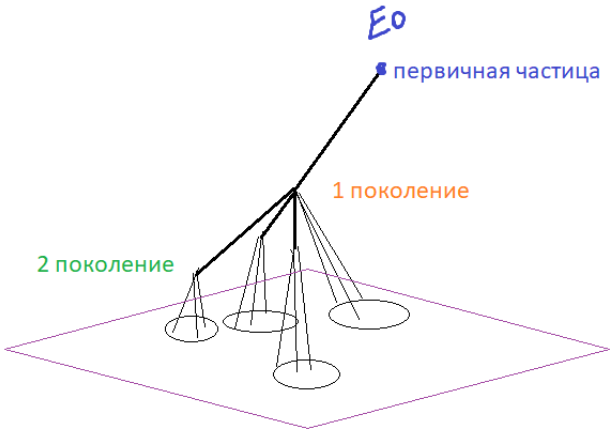


Рис. 1.2. Частица из атмосферы

теристик наблюдаемых семейств (энергия, множественность и расстояние от энергетического центра) к типу первичной частицы, то есть изучать массовый состав первичного космического излучения.
В частности, показать, что число кластеров коррелированно с энергией, а также типом первичной частицы (числом нуклонов).
Кроме того, появляются новые возможности для изучения явления выстроенности наиболее энергичных частиц в семействе (явления компланарности).

1.2. Описание эксперимента

[2] Для изучения деталей внутренней структуры адронов необходимо исследовать их взаимодействия при высоких энергиях.
Существует два источника частиц высоких энергий - ускорители и космические лучи. Верхняя граница энергии частиц космического излучения очень высока. Они дают возможность исследовать взаимодействия адрон-ядро, ядро-ядро при очень высоких энергиях.
Для изучения взаимодействий при энергии $E_0 > 100$ ТэВ применяется метод рентгеноэмульсионных камер (РЭК), устанавливаемых на высотах гор. С помощью детекторов площадями до 1000 м^2 ведется накопление статистики в течение года.
Основной объект исследования - группы генетически связанных частиц (гаммма-квантов и адронов) достаточно большой энергии, которые возникают в результате взаимодействия с ядрами атомов воздуха первичной частицы высокой энергии и последующего ядерно-электромагнитного каскада. Такие группы частиц принято называть семействами.
Особенности экспериментов с космическими лучами - малая интенсивность потоков частиц высокой энергии, круто спадающий энергетический спектр, сложный химический состав первичного космического излучения, наличие каскадных процессов (электромагнитных и ядерных) в атмосфере - вынуждают для анализа экспериментальных данных широко прибегать к модели-

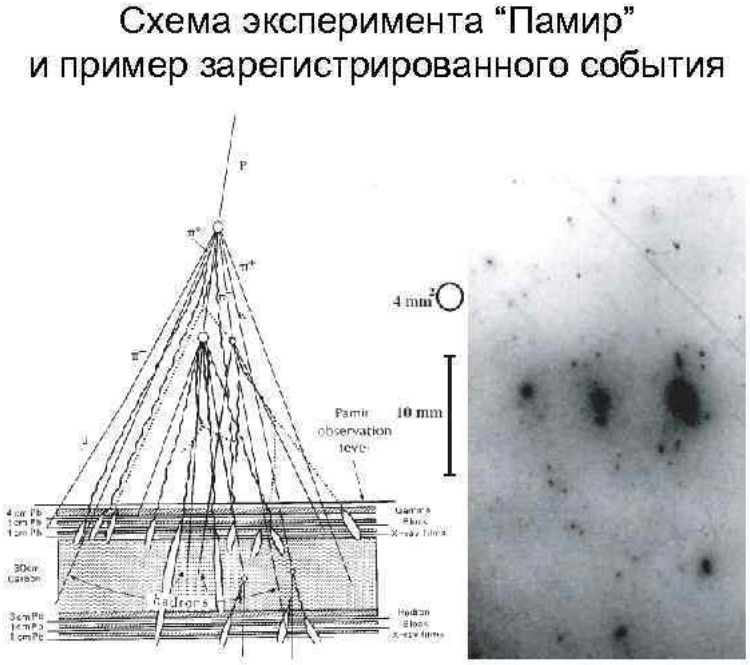


Рис. 1.3. Схема эксперимента "Памир"

рованию регистрируемых явлений и сравнению эксперимента с результатами такого моделирования.

Метод РЭК основан на регистрации электронно-фотонных каскадов (ЭФК), вызываемых гамма-квантами в плотной среде (свинец). Пучок электронов ЭФК приводит к появлению темных пятен на рентгеновской пленке, расположенной на некоторой глубине в поглотителе. Измерения характеристик каскадных частиц, регистрируемых РЭК, дают возможность определить энергию гамма-кванта, вызвавшего ЭФК. Кроме того, при помощи РЭК можно определить угол падения ЭФК на камеру. Как показывают оценки и модельные расчеты, те события, которые регистрируются в эксперименте "Памир"представляют собой смесь вторичных частиц в среднем двух-трех поколений ядерно-магнитного каскада.

1.2.1. Формат данных

В качестве источника данных в работе были использованы файлы экспериментальных (BC.DAT и BCR.DAT) и модельных (BC.MC0 и BCR.MC0) данных в бесформатном виде, а также операторы считывания этих данных в среде программирования FORTRAN (файл READ.FOR).

После считывания данные были конвертированы в два формата xlsx и csv.

В BC.DAT содержится информация о кластерах. В BCR.DAT содержится информация о всех точках, всего около 1000 событий. Экспериментальные данные включают в себя координаты X, Y и значение энергии E для каждой частицы.

Аналогично для файлов, содержащих смоделированные данные в BC.MC0 содержится информация о кластерах, в BCR.MC0 информация о всех точках, всего около 1500 событий.

Модельные данные содержат координаты X, Y, значение энергии E, высоту образования каж-

дой частицы (высоту последнего взаимодействия H), а также число нуклонов первичного ядра A0.
Значения координат X и Y всех пятен почернения отсчитываются от энергетического центра тяжести:

$$X_0 = \frac{\sum\limits_{i=1} E_i x_i}{\sum\limits_{i=1} E_i}, \quad Y_0 = \frac{\sum\limits_{i=1} E_i y_i}{\sum\limits_{i=1} E_i} \tag{1.1}$$

1.2.2. Отбор семейств

Первоначально выбираются только частицы с энергией больше 4 ТэВ, так как изучаемые адронные семейства характеризует высокая энергия.
Из множества смоделированных и экспериментальных семейств выбираются семейства с суммарной энергией по всем частицам превышающей 100 ТэВ.
Для повышения достоверности не рассматриваются семейства, количество частиц в которых меньше трех. Такие малочисленные семейства могут быть связаны с неточностью определения координат.
В соответствии с параметрами изучаемой области пленки почернения рассматриваются только частицы с расстоянием от энергетического центра тяжести меньшим 15 см.

1.2.3. Разбиение данных на группы по энергии.

Определяется суммарная энергия для каждого семейства и семейства разбиваются на интервалы по значениям суммарной энергии (100-200, 200-400, 400-700, >700, >100, 100-700). Впоследствии, в разных интервалах энергии проверяется однородность экспериментальных и модельных данных по физическим величинам.

1.2.4. Статистическая модель

Мы рассматриваем группы генетически связанных частиц - семейства, которые имеют вид пятен почернения на пленке.
Будем считать пятна почернения элементарными событиями и называть частицами. Информация, содержащаяся в файлах BCR.DAT (данные эксперимента) и BCR.MC0 (данные модели) это некоторые случайные выборки. В качестве объектов выступают пятна почернения, которые обладают тремя признаками: координатами X и Y и энергией E. Объекты и признаки формируют матрицу данных.
Далее исследуем выборочные распределения семейств по этим признакам, а также связанным с ними величинам и физически существенным характеристикам семейств, такими как множественность, расстояние и среднее расстояние частицы от энергетического центра тяжести семейства, средняя энергия семейства.

Глава 2

Сравнение распределений экспериментальных и модельных данных.

В силу поставленной задачи выбор оптимального алгоритма выделения групп частиц от последних взаимодействий осуществляется на модельных данных и в дальнейшем применяется к экспериментальным, проводится сопоставление и проверка однородности распределений экспериментальных и модельных данных.

2.1. Построение выборочных распределений для различных наблюдаемых величин.

С целью наиболее информативного сравнения экспериментальных и модельных данных рассмотрим следующие распределения.
Пусть N_γ - число частиц в семействе равно k.
Определим *радиус* R - евклидовое расстояние частицы от энергетического центра тяжести:

$$R = \sqrt{(x_i - X_0)^2 + (y_i - Y_0)^2}$$

1. (sum_energy) Распределение числа семейств по суммарной энергии $\sum E$ всех частиц в семействе:

$$\sum E^k = E_1 + ... + E_k,$$

2. ($mean\_r$) Распределение по средним радиусам частиц:

$$\overline{R} = \frac{R_1 + ... + R_k}{k},$$

3. ($mean\_e\_r$) Распределение по средним произведениям энергии на радиус частицы в данном семействе:

$$\overline{ER} = \frac{E_1R_1 + ... + E_kR_k}{k},$$

4. (lg_mean_r) Распределение по средним десятичным логарифмам радиусов частиц:

$$\overline{lg(R)} = \frac{\lg R_1 + ... + \lg R_k}{k},$$

5. (lg_mean_er) Распределение по средним десятичным логарифмам произведения энергии на радиус частицы:

$$\overline{lg(ER)} = \frac{\lg (E_1R_1) + ... + \lg (E_kR_k)}{k},$$

6. (n_gamma) Распределение по множественности N_γ = k числу частиц в каждом семействе.

Введём понятие относительного порога, равного 0.04. Все частицы в данном i-ом семействе сортируются по убыванию значений их энергий. Далее рассматриваются только те частицы, для

которых значение

$$\frac{E_i}{\sum_{j=1}^{n_i} E_j} > 0.04,$$

где E_i - энергия частиц, n_i - число частиц с более высокой энергией в семействе, чем у i, $\sum_{j=1}^{n_i} E_j$ - сумма энергий всех предшествующих ей более энергичных частиц.

Частицы, для которых удовлетворяется данное неравенство относятся к *надпороговым частицам*. Коэффициент выбран из физических соображений.

Пусть s - число *надпороговых частиц* обозначим s.

7. (n_gamma_thr) Распределение по множественности $N_{\gamma th}$ *надпороговых частиц* - числу частиц s, удовлетворяющих неравенству.

8. (sum_energy_thr) Распределение по сумме энергий всех *надпороговых частиц* в каждом семействе:

$$\sum E_{th}^s = E_1 + ... + E_s,$$

Также были рассмотрены распределения по взаимным расстояниям частиц. Для каждого семейства строился верхний треугольник матрицы расстояний между частицами. Из значений этих расстояний, взятых в каждом семействе составлялось распределение. В качестве расстояний использовались следующие метрики:

$$d_1 = r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

$$d_2 = \left(\frac{E_i E_j}{E_i + E_j}\right) r_{ij}$$

$$d_3 = \sqrt{E_i E_j} r_{ij}$$

d_1 - это обычное евклидовое расстояние. Индексы i и j соответствуют координатам i-ой и j-ой частиц.

9. (d1) Распределение по расстояниям d_1.

10. (d2) Распределение по расстояниям d_2.

11. (d3) Распределение по расстояниям d_3.

Также рассматривались следующие распределения по характеристикам всех частиц: 12. (E) Распределение по энергиям E_i всех частиц всех семейств.

13. (R) Распределение по радиусам частиц R_i.

14. (ER)Распределение по произведению энергии на радиус частицы $E_i R_i$.

15. (lg_r) Распределение по $\lg(R_i)$

16. (lg_e_r) Распределение по $\lg(E_i R_i)$.

2.1.1. Сопоставление распределений в эксперименте и в модели.

Для каждой выборки $X_1, ..., X_k$ из рассмотренных распределений находим следующие величины.

1. ($mean$) Выборочное среднее.

$$\overline{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$$

2. (*var*) Выборочная дисперсия.

$$S_k^2 = \frac{1}{k}\sum_{i=1}^{k}(X_i - \overline{X})^2$$

3. (*skew*) Выборочная ассиметрия.

$$\gamma_1 = \frac{m_3}{m_2^{\frac{3}{2}}} = \frac{m_3}{S_k^3}, \quad m_j = \frac{1}{k}\sum_{i=1}^{k}(X_i - \overline{X})^j,$$

где m_j - выборочный центральный момент j-ого порядка.
4. (*kurtosis*) Выборочный эксцесс.

$$\gamma_2 = \frac{m_4}{m_2^2} - 3 = \frac{m_4}{S_k^4} - 3, \quad m_j = \frac{1}{k}\sum_{i=1}^{k}(X_i - \overline{X})^j,$$

где m_j - выборочный центральный момент j-ого порядка.
5. (*sampling error*) Ошибка выборки.

$$S_{err} = \frac{S_k}{k}$$

где S_k - корень из выборочной дисперсии. 6. (*relative_error*) Относительная ошибка.

$$R_{err}^{mc0} = \frac{|\overline{X}_{mc0} - \overline{X}_{exp}|}{\overline{X}_{exp}}$$

Точность $\overline{X}_{mc0}$ относительно $\overline{X}_{exp}$.

$$R_{err}^{exp} = \frac{|\overline{X}_{exp} - \overline{X}_{mc0}|}{\overline{X}_{mc0}}$$

Точночть $\overline{X}_{exp}$ остностельно $\overline{X}_{mc0}$.
7. (*entropy*) Энтропия.
Пусть случайные события X_i распределены с вероятностями p_i.

$$H = -\sum_{i=1}^{n} p_i log_2 p_i$$

Информация $I = \{log_2 p_1, log_2 p_2, ...\}$.
Как указано в пункте 7, для каждого распределения определяется значение энтропии для экспериментальных и модельных данных. Совокупность энтропий экспериментальных данных будем считать вектором энтропий в экспериментальных данных. Аналогично совокупность энтропий модельных данных будем считать вектором энтропий модельных данных. Дополнительно к проверке одности рассмотренных ранее распределений, проводится проверка однородности векторов энтропий этих распределений.

2.2. Статистические критерии для проверки однорожности.

2.2.1. Двухвыборочный критерий Смирнова.

(*stat D and C_alpha*)

[4] [7] Рассмотрим выборку из распределения экспериментальных данных $X_1, ..., X_{k_1}$ и выборку из распределения модельных данных $Y_1, ..., Y_{k_2}$, dыборочное пространство непрерывно.

Пусть есть некоторое распределение случайных величин:

$X_i \sim \mathcal{F}(x)$, $(i = \overline{1, .., k_1})$ и $Y_j \sim \mathcal{G}(x)$, $(j = \overline{1, .., k_2})$.

Распределение $\mathcal{F}(x)$ и $\mathcal{G}(x)$ неизвестно.

Нулевая гипотеза H_0: $\forall x, \quad \mathcal{F}(x) = \mathcal{G}(x)$.

Альтернативная гипотеза H_1: $\exists x, \quad \mathcal{F}(x) \neq \mathcal{G}(x)$.

Статистика критерия Колмогорова-Смирнова:

$$D_{k_1,k_2} = sup|\overline{F_{k_1}} - \overline{G_{k_2}}|,$$

где $\overline{F_{k_1}}$, $\overline{G_{k_2}}$ - эмпирические функции распределения выборки.

Построим критическую область.

Если гипотеза H_0 справедлива, то $\overline{F_{k_1}}$ близка к $\overline{G_{k_2}}$ и статистика D_{k_1,k_2} принимает малые значения, α - уровень значимости.

$$P(D_{k_1,k_2} > C_\alpha | H_0) \leq \alpha. \tag{2.1}$$

Здесь C_α - критическая точка распределения Смирнова определяется по таблице этого распределения. Если наблюдаемые значения $D_{k_1,k_2} < C_\alpha$ гипотеза H_0 принимается, в противном случае она отклоняется.

Если значения k_1, k_2 велики, то необходимо воспользоваться асимптотическим распределением. Статистика

$$\mathcal{K} = \sqrt{\frac{k_1 k_2}{k_1 + k_2}} D_{k_1,k_2}$$

имеет асимптотическое распределение Колмогорова. Тогда из (2.1) следует:

$$P\big(\sqrt{\frac{k_1 k_2}{k_1 + k_2}} D_{k_1,k_2} > \sqrt{\frac{k_1 k_2}{k_1 + k_2}} C_\alpha \big| H_0\big) \leq \alpha$$

или

$$P\big(\mathcal{K} > \sqrt{\frac{k_1 k_2}{k_1 + k_2}} C_\alpha \big| H_0\big) = 1 - P\big(\mathcal{K} < \sqrt{\frac{k_1 k_2}{k_1 + k_2}} C_\alpha \big| H_0\big) \leq \alpha.$$

Определим критическую точку C_α. $F_\mathcal{K}$ - функция распределения Колмогорова:

$$1 - F_\mathcal{K}\big(\sqrt{\frac{k_1 k_2}{k_1 + k_2}} C_\alpha\big) \leq \alpha.$$

Следовательно:

$$F_\mathcal{K}\big(\sqrt{\frac{k_1 k_2}{k_1 + k_2}} C_\alpha\big) \geq 1 - \alpha.$$

Пусть

$$F_{\mathcal{K}}(y) = 1 - \alpha,$$

тогда $y = y_{1-\alpha}$ квантиль уровня $1 - \alpha$ распределения Колмогорова. Поэтому

$$\sqrt{\frac{k_1 k_2}{k_1 + k_2}} C_\alpha = y_{1-\alpha},$$

$$C_\alpha = \sqrt{\frac{k_1 + k_2}{k_1 k_2}} y_{1-\alpha}.$$

При $\alpha = 0.05$ значение $y_{0.95} = 1.36$ определяется по таблице.
При $\alpha = 0.01$ значение $y_{0.95} = 1.63$ определяется по таблице.

2.2.2. Вывод по применению критерия.

Для векторов энтропий, определенных в пункте 2.2, рассмотренных распределений гипотеза однородности принимается на уровне значимости $\alpha = 0.05$.
Распределения, для которых в результате статистического решения гипотеза однородности принимается в определенных интервалах энергии на уровне значимости $\alpha = 0.01$:
100-200: $\sum E$.
200-400: $\sum E$, $\overline{R}$, $\overline{ER}$, $N_{\gamma th}$, $\sum E_{th}$.
400-700: $\sum E$, $\overline{R}$, $\overline{ER}$, $lg(\overline{R})$, $lg(\overline{ER})$, $N_{\gamma th}$.
>700: $\sum E$, $\overline{R}$, $lg(\overline{R})$, N_γ, $N_{\gamma th}$.
100-700: $\sum E$.
>100: N_γ, $N_{\gamma th}$.
Ниже приводится таблица с числовыми значениями вычисленной статистики (левый столбец) и критического значения (правый столбец) для некоторых распределений в интервалах энергий. Пустые клетки означают, что нужно сравнивать значение статистики с критическим значением из клетки выше.

распределение	**100-**	**200**	**200 -**	**400**	**400-**	**700**	**>700**		**100-**	**700**
$\sum E$	0.084	0.096	0.053	0.136	0.103	0.215	0.218	0.3	0.075	0.074
$\overline{R}$	0.248		0.127		0.17		0.205		0.187	
$\overline{ER}$	0.195		0.099		0.103		0.384		0.131	
N_γ	0.108		0.199		0.426		0.181		0.076	
$\sum E_{th}$	0.194		0.137		0.356		0.357		0.141	
E	0.136	0.027	0.152	0.03	0.205	0.036	0.205	0.034	0.153	0.017
R	0.106		0.055		0.082		0.152		0.034	
ER	0.046		0.096		0.152		0.228		0.051	

Таблица 2.1. Статистика критерия Смирнова и критическое значение

Числовые значения для всех распределений можно найти по ссылке в разделе Выводы (в конце) в таблице UniformityOldBank.xlsx в столбце (*stat D and C_alpha*) первая строка соответствует значению статистики, вторая строка соответсвует критическому значению. Выводы по крите-

риям в файле UniformityOldBankConclusions.xlsx по той же ссылке.

[1] Критерии типа Колмогорова-Смирнова хороши, когда альтернативное распределение таково, что разница между ним и исходным (например, разница в средних) велика. Однако если разница между средними и дисперсиями невелика, но две частотные характеристики заметно отличаются формой, то используются критерии, свободные от распределения.

2.2.3. Непараметрический критерий Манна-Уитни

Критерий Манна-Уитни основан на парном сравнении элементов первой и второй выборок. Статистика Манна-Уитни - это число успехов ($x_i < y_j$) в парных сравнениях.

Нулевая гипотеза H_0: $\forall x, \quad \mathcal{F}(x) = \mathcal{G}(x)$.

Альтернативная гипотеза H_1: $\exists x, \quad \mathcal{F}(x) \neq \mathcal{G}(x)$.

Пусть

$$I(x_i < y_j) = \begin{cases} 1, & \text{если} \quad x_i < y_j \\ 1/2, & \text{если} \quad x_i = y_j \\ 0, & \text{если} \quad x_i > y_j \end{cases} \tag{2.2}$$

Статистика Манна-Уитни вычисляется по формуле

$$U(X_{[n]}, Y_{[m]}) = \sum_{i=1}^{n} \sum_{j=1}^{m} I(x_i < y_j) \tag{2.3}$$

При n, m > 8 можно применить аппроксимацию. То есть в качестве статистики критерия рассмотреть функцию

$$U^*(X_{[n]}, Y_{[m]}) = \frac{U(X_{[n]}, Y_{[m]}) - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \tag{2.4}$$

Если справедлива гипотеза однородности H_0, то статистика $U^*(X_{[n]}, Y_{[m]})$ асимптотически подчиняется стандартному нормальному распределению.

При альтернативной гипотезе H_1: $\exists x, \quad \mathcal{F}(x) \neq \mathcal{G}(x)$ критическая область:
$(-\infty, u_{\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}, +\inf)$, где $u_{\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}$ - квантили стандартного нормального распределения уровней $\frac{\alpha}{2}, 1 - \frac{\alpha}{2}$.

Если численное значение $U^*(X_{[n]}, Y_{[m]})$ попадает в критическую область, то нулевая гипотеза H_0 отвергается, в противном случае нет оснований ее отвергнуть при уровне значимости равном α.

2.2.4. Вывод по применению критерия.

Для векторов энтропий рассмотренных распределений гипотеза однородности принимается на уровне значимости $\alpha = 0.05$.

Распределения, для которых в результате статистического решения гипотеза однородности принимается в определенных интервалах энергии на уровне значимости $\alpha = 0.01$:

100-200: $\sum E$, $N_{\gamma th}$, d_2, d_3, X.

200-400: $\sum E$, $\overline{R}$, $\overline{ER}$, $lg(\overline{ER})$, X.

400-700: $\sum E$, $\overline{R}$, $\overline{ER}$, $lg(\overline{R})$, $lg(\overline{ER})$, $N_{\gamma th}$, X.

>700: $\overline{R}$, $lg(\overline{R})$, N_γ, $N_{\gamma th}$.
100-700: $N_{\gamma th}$, X.
>100: $lg(\overline{ER})$, N_γ, $N_{\gamma th}$, R, $lg(\overline{R})$.
Этот критерий показывает однородность в большем числе распределений, чем критерий Колмогорова-Смирнова.

2.2.5. Двухвыборочный критерий Стьюдента, модификация Уэлча.

$(t - test\ st\ p - value)$
[8] Рассмотрим выборку из распределения экспериментальных данных $X_1, ..., X_{k_1}$ и выборку из распределения модельных данных $Y_1, ..., Y_{k_2}$.
Нулевая гипотеза H_0: $\mu_X = \mu_Y$ (средние выборок равны).
Альтернативная гипотеза H_1: $\mu_X \neq \mu_Y$ (средние выборок не равны).
Критерий проверяет гипотезу для двух независимых неравных выборок с неравными дисперсиями.
Статистика критерия Уэлча:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{k_1} + \frac{S_Y^2}{k_2}}}.$$

Здесь $\overline{X}, \overline{Y}$ - выборочные средние, S_X^2, S_Y^2 - выборочные дисперсии.
При справедливости нулевой гипотезы t приближается к распределению Стьюдента с d степенями свободы:

$$d = \frac{\left(\frac{S_X^2}{k_1} + \frac{S_Y^2}{k_2}\right)^2}{\frac{S_X^4}{k_1^2(k_1-1)} + \frac{S_Y^4}{k_2^2(k_2-1)}}.$$

При превышении значения наблюдаемой статистики t по абсолютной величине критического значения данного распределения (при заданном уровне значимости) нулевая гипотеза отвергается.

2.2.6. Однородность по сдвигу в различных энергетических интервалах.

Гипотеза равенства средних принимается для векторов энтропий рассмотренных распределений характеристик на уровне значимости $\alpha = 0.05$.
Распределения, для которых в результате статистического решения гипотеза равенства средних принимается в определенных интервалах энергии на уровне значимости $\alpha = 0.01$:
100-200: $\sum E$, $N_{\gamma th}$, X.
200-400: $\sum E$, $\overline{R}$, $\overline{ER}$, $lg(\overline{R})$, $lg(\overline{ER})$, X, lg(R).
400-700: $\sum E$, $\overline{R}$, $\overline{ER}$, $lg(\overline{R})$, $N_{\gamma th}$, X, R.
>700: $\sum E$, $\overline{R}$, $\overline{ER}$, $lg(\overline{R})$, $N_{\gamma th}$, X, R.
100-700: $\sum E$, $N_{\gamma th}$.
>100: N_γ, $N_{\gamma th}$, X, R.
В итоге наилучшее совпадение экспериментальных и модельных данных наблюдается в более узких интервалах более высоких энергий в основном в распределениях средних значений.
Вычисленные p-значения находятсяпо ссылке в разделе Выводы (в конце) в таблице

UniformityOldBank.xlsx в столбце (*ttest st pvalue*).
Выводы по критериям в файле UniformityOldBankConclusions.xlsx по той же ссылке.

2.2.7. Двухвыборочный критерий Левене.

(*levene st* $p-value$)

[9] [10] Возьмем выборку из распределения экспериментальных данных $X_1, ..., X_{k_1}$ и выборку из распределения модельных данных $Y_1, ..., Y_{k_2}$.
Нулевая гипотеза H_0: $\sigma_X^2 = \sigma_Y^2$ (дисперсии выборок равны).
Альтернативная гипотеза H_1: $\sigma_X^2 \neq \sigma_Y^2$ (дисперсии выборок не равны).
Случай равенства дисперсий по выборкам является однородностью дисперсии.
Тест Левене менее чувствителен к отклонениям от нормального распределения, чем, например, тест Бартлетта.
Статистика критерия:

$$W = \frac{(k_1+k_2-2)(k_1(Z_{X*}-Z_{**})^2 + k_2(Z_{Y*}-Z_{**}))}{\sum_{i=1}^{k_1}(Z_{Xi}-Z_{X*}) + \sum_{i=1}^{k_2}(Z_{Yi}-Z_{Y*})},$$

где

$$Z_{Xi} = |X_i - \widetilde{X}|, \quad \widetilde{X} - \text{медиана выборки},$$

$$Z_{Yi} = |Y_i - \widetilde{Y}|, \quad \widetilde{Y} - \text{медиана выборки},$$

$$Z_{X*} = \frac{1}{k_1}\sum_{j=1}^{k_1} Z_{Xj}, \quad Z_{Y*} = \frac{1}{k_2}\sum_{j=1}^{k_2} Z_{Yj}, \quad -\text{среднее по выборкам},$$

$$Z_{**} = \frac{1}{k_1+k_2}\Big(\sum_{j=1}^{k_1} Z_{Xj} + \sum_{j=1}^{k_2} Z_{Yj}\Big) - \quad \text{среднее по всему.}$$

Проверяемая гипотеза о равенстве дисперсий отклоняется, если

$$W > F_{\alpha,\ 1,\ k_1+k_2-2,}$$

где $F_{\alpha,\ 1,\ k_1+k_2-2}$ верхнее критическое значение F-распределения
Фишера с $(1,\ k_1+k_2-2)$ степенями свободы и уровнем значимости α.

2.2.8. Однородность по разбросу в различных энергетических интервалах.

Гипотеза равенства дисперсий принимается для векторов энтропий рассмотренных распределений характеристик на уровне значимости $\alpha = 0.05$.
Распределения, для которых в результате статистического решения гипотеза равенства дисперсий принимается в определенных интервалах энергии на уровне значимости $\alpha = 0.01$:
100-200: $\overline{ER}$, E.
200-400: $\sum E$, $\overline{ER}$, $N_{\gamma th}$, $\sum E_{th}$, $lg(\overline{ER})$.
400-700: $\sum E$, $\overline{R}$, $\overline{ER}$, $lg(\overline{R})$, $N_{\gamma th}$, $\sum E_{th}$, Y, R, $lg(R)$.
>**700**: $lg(\overline{ER})$, $N_{\gamma th}$, $\sum E_{th}$, lg(R).
100-700: $\sum E$, $\overline{ER}$, $N_{\gamma th}$, X, R, ER, lg(ER).

>100: $\overline{ER}$, N_γ , $N_{\gamma th}$, X, Y, R.

Вычисленные p-значения можно найти по ссылке в разделе Выводы (в конце) в таблице UniformityOldBank.xlsx в столбце (*levene st pvalue*).

Выводы по критериям в файле UniformityOldBankConclusions.xlsx по той же ссылке.

2.3. Таблица p-значений некоторых распределений. Общий вывод.

распределение	**100-**	**200**	**200 -**	**400**	**400-**	**700**	**>700**		**100-**	**700**
$\sum E$	0.026	0.028	0.65	0.703	0.67	0.659	0.002	0.005	0.014	0.108
$\overline{R}$	0.0	0.012	0.749	0.0	0.247	0.136	0.069	0.004	0.0	0.0
$\overline{ER}$	0	0.197	0.96	0.377	0.656	0.629	0.001	0.02	0.0	0.447
N_γ	0.005	0.01	0.0	0.016	0.027	0.296	0.079	0.978	0.0	0.0
$\sum E_{th}$	0.632	0.002	0.0	0.139	0.0	0.037	0.0	0.001	0.0	0.0
E	0.0	0.17	0.0	0.005	0.0	0.0	0.0	0.0	0.0	0.0
R	0.0	0.0	0.0	0.0	0.1	0.973	0.0	0.0	0.001	0.053
ER	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.001	0.167

Таблица 2.2. p-значения критериев Стьюдента и Левене

Для каждого интервала в левом столбце указано p-значение [1] критерия Стьюдента и в правом столбце p-значение критерия Левене.

В таблице представлены наиболее важные распределения.

По рассмотренным в таблицах значениям можно сделать выводы о выполнении однородности по трем критериям для $\sum E$ во всех энергетических интервалах на уровне значимости $\alpha = 0.01$.

Однорородность $\overline{R}$ и $\overline{ER}$ наблюдается в интервалах энергий **200-400** и **400-700**.

Наблюдаемые отклонения в однородности энергии в интервале **>700**, как выяснилось позднее, связаны с недостаточной *представленностью* частиц высокой энергии. В результате моделирования в семействах не получилось достаточного количества высокоэнергичных частиц.

Отклонения в диапазоне **100 - 200** связаны с тем, что 100% эффективность регистрации частиц в эксперименте начинается лишь со значения E = 6 ТэВ, в то время как в базе экспериментальных и модельных данных рассматриваются частицы с энергией выше E = 4 ТэВ.

В интервале во энергии **>100** наблюдается однородность экспериментальных и модельных данных по трем критериям для множественности N_γ и радиусам R.

Векторы энтропий распределений во всем диапазоне энергий однородны по всем рассмотренным критериям критериям, что соответствует в среднем одинаковой информации в эксперименте и в модели.

[1] Пусть в нашем случае значение случайных величин $X_1, ..., X_k$ равно $x_1, ..., x_k$. Тогда значение статистики, полученное на выборке, будет равно числу $W_{exp} = W(x_1, ..., x_k)$.
Тогда p-значение - это вероятность того, насколько часто статистика критерия при том что гипотеза H_0 верна, будет принимать значение (для других выборок) больше наблюдаемого на данной выборке:
$P(W(X_1, .., X_k) > W_{exp}|H_0) = p$, где $X_1, .., X_k$ - случайные величины.
p-значение равно вероятности ошибки первого рода
Если p-значение $< \alpha$, то гипотеза H_0 отклоняется, α - уровень значимости.

2.4. Гистограммы и эмпирические функции распределений.

Среди всех рассмотренных распределений физических величин есть те, для которых не принимался критерий однородности или критерий равенства средних, или критерий равенства дисперсий. Поэтому был рассмотрен визуальный метод распределения экспериментальных и модельных данных, дополнивший вывод о принятой однородности данных. Визуальный метод сравнения заключался в построении гистограмм и эмпирических функций распределений всех физических величин во всех рассмотренных интервалах суммарных энергий (100-200, 200-400, 400-700, >700, >100). На вертикальных осях гистограмм указаны нормированные значения частот распределений физических величин. Нормирование предполагает деление на число частиц в интервале гистограммы, оно необходимо в связи различным количеством семейств в эксперименте и модели.
Визуально для R и E наблюдается хорошее согласие экспериментальных и модельных данных. Аналогично для остальных распределений, их гистограммы и эмпирические функции можно найти в папке images по ссылке выводах.

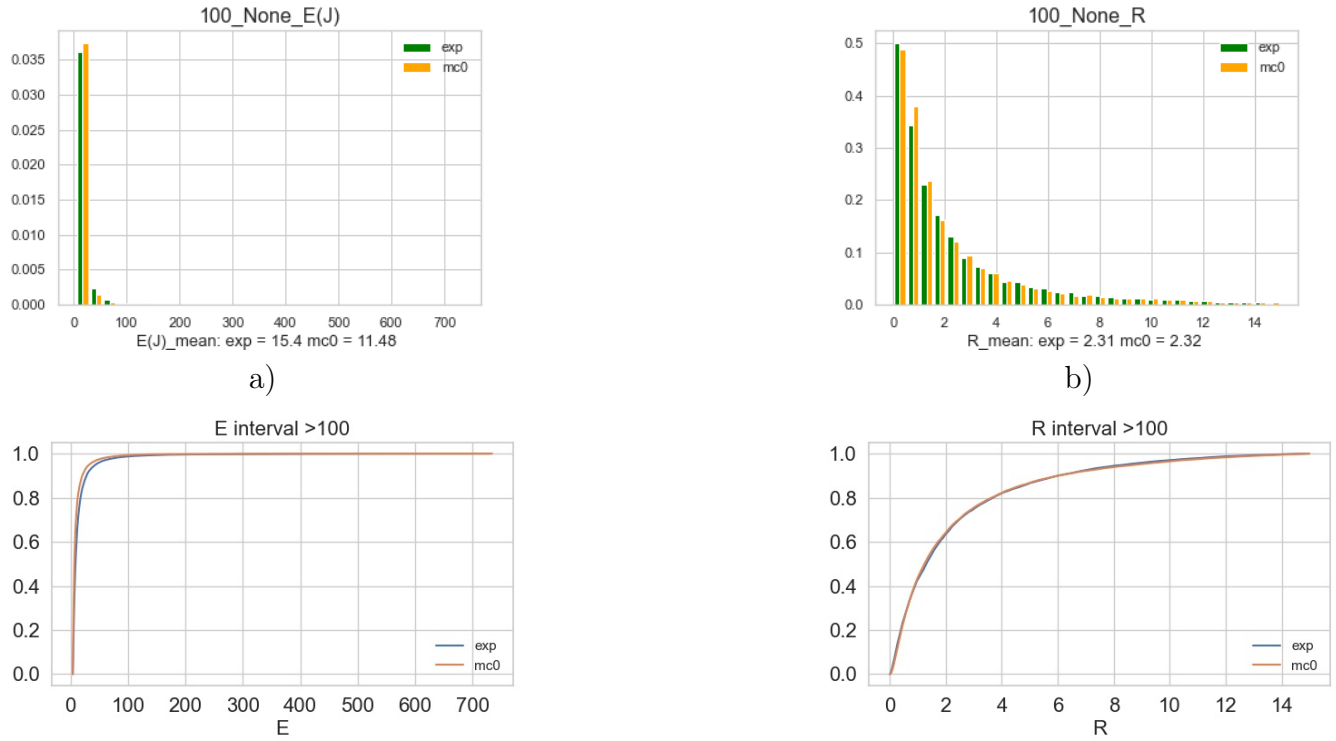


Рис. 2.1. a) гистограмма E в > 100 b) гистограмма R в >100 c) эмпирическая функция E в > 100 d) эмпирическая функция R в >100

Эмпирическая или выборочная функция распределения - это приближение теоретической функции распределения. Пусть $X_1, ..., X_k$ выборка независимых случайных величин объема k, порожденная случайной величиной X, задаваемой функцией распределения F(x). Определим эмпирическую функцию $\overline{F_k}(x)$ распределения случайной величины X:

$$\overline{F_k}(x) = \frac{1}{k}\sum_{i=1}^{k} I(X_i \le x), \quad I - \text{индикатор события } (X_i \le x)$$

Таким образом, значение функции $\overline{F_k}(x)$ в точке x равно относительной частоте элементов выборки, не превосходящих значение x.

Глава 3

Предварительный анализ кластеризуемых данных.

3.0.1. Проверка нормальности для предположения о возможности кластеризации.

Кроме проверки однородности при помощи статистических критериев, в работе была проведена проверка нормальности выборок распределений по X, Y, E при помощи критерия нормальности Шапиро-Уилка [1]. Критерий применялся отдельно ко всему экспериментальному банку данных и отдельно ко всему модельному банку для выборок из распределений по X, Y, E.

Рассмотрим значения, приведенные в таблице, получившиеся в результате статистического решения:

	$\mathbf{X}_{exp}$	$\mathbf{Y}_{exp}$	$\mathbf{R}_{exp}$	$\mathbf{E}_{exp}$	$\mathbf{X}_{mod}$	$\mathbf{Y}_{mod}$	$\mathbf{E}_{mod}$	$\mathbf{R}_{mod}$
значение статистики	0.857	0.846	0.758	0.308	0.831	0.840	0.737	0.327
p-значение	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Таблица 3.1. значения критерия Шапиро-Уилка

Здесь **X**, **Y** - значения координат частицы по оси абсцисс и ординат соответственно, отсчитанные от энергетического центра тяжести, **R** - евклидовое расстояние частицы от энергетического центра тяжести, **E** - энергия частицы.

В таблице (3.1) p-значение меньше заданного уровня значимости $\alpha = 0.05$, поэтому нулевая гипотеза о нормальности отклоняется. Распределения отдельно по X, Y, E не являются нормальными для всего множества экспериментальных и модельных данных.

Нормальное распределение у случайной величины можно предполагать, если на ее отклонение от некоторого заданного значения влияет множество различных факторов, причем влияние каждого из них вносит малый вклад в это отклонение, а их действия независимы или почти независимы. То есть, влияние всех факторов на вид распределения равномерно, никакой из факторов не имеет преимущественного влияния.

В нашем случае из отсутствия нормальности распределения можно предположить наличие фактора, который оказывает сильное воздействие и определяет вид групп частиц (групп пятен почернения), поэтому мы можем предполагать возможность кластеризации данных.

3.0.2. Разделение модельных данных на группы частиц.

Первичные частицы космических лучей, образовавшие семейство, являются ядрами различных химических элементов из таблицы Менделеева. Процессы формирования семейств, образованных различными ядрами, отличаются друг от друга, что приводит к различным распределениям по наблюдаемым физическим величинам.

В случае смоделированных данных известно число нуклонов в ядре первичной частицы. В частности, в банке данных смоделированных семейств представлены три группы: Proton, MgSi, Ferrum.

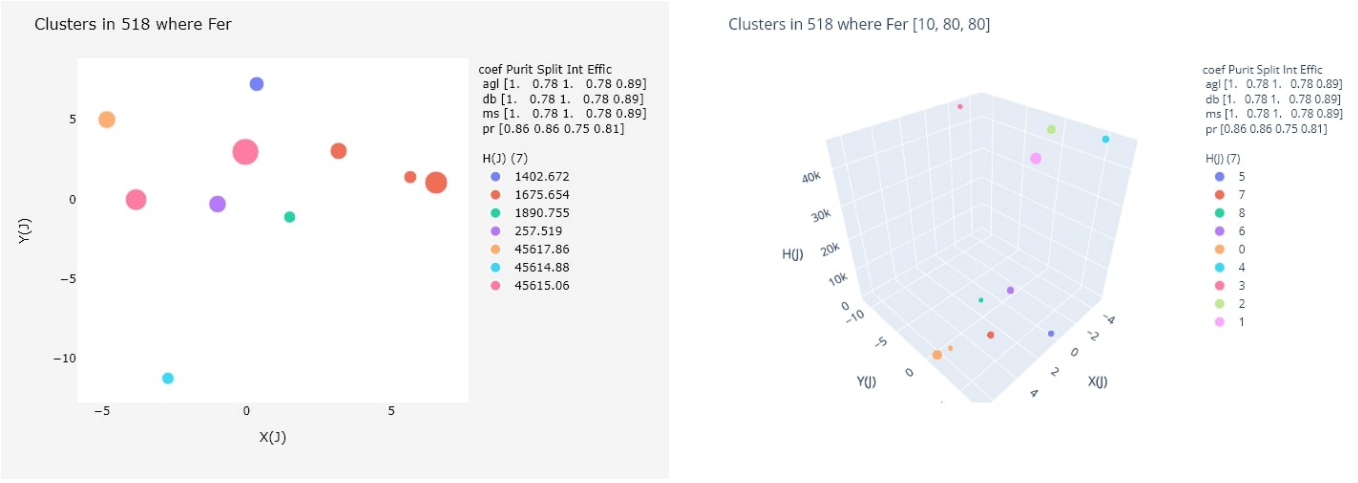
В дальнейшем может быть поставлена задача определения в экспериментальных данных семейств с разными первичными частицами. Это возможно в силу повышения чувствительности физических характеристик к типу первичной частицы для кластеризованных данных, так как происходит переход к более раннему поколению частицы.

3.0.3. Предварительная оценка.

Определенные различия между семействами демонстрирует визуализация, описанная в следующем разделе. Всего было рассмотрено 10 семейств каждой группы частиц Proton, MgSi, Ferrum.
Семейства Ferrum имеют большую множественность в сравнении с остальными семействами.
В семействах Proton встречаются наиболее высоко энергичные частицы, а также более компактное распределение по R.
Однако более точное описание характеристик различных групп частиц требует дополнительного исследования.
Также визуализация помогает выбрать и оценить алгоритм кластеризации.

3.1. Визуализация семейств

Для частиц Proton, MgSi, Ferrum была построена визуализация семейств в двух форматах html и png, она позволяет увидеть различия в семействах.
html визуализация интерактивна, в ней можно рассматривать точки по отдельности, приближать отдельные участки и смотреть информацию о частице при наведении на нее мыши.
Визуализацию можно найти в папке web по ссылке в разделе Выводы (в конце).
Слева расположено 2d изображение семейства. По горизонтальной и вертикальной оси отло-



c) d)

Рис. 3.1. Пример визуализации Ferrum № 518

жены значения X и Y. Размер круга пропорционален величине энергии частицы, цвет соответствует высоте последнего взаимодействия (идеальной кластеризации). В правом верхнем углу расположена таблица значений рассмотренных далее коэффициентов эффективности кластеризации различных алгоритмов. Также при наведении мыши в html формате можно узнать

координаты, энергию, высоту частицы и кластер, определенный алгоритмом agglomerative.
Справа расположено 3d изображение семейства. Две оси соответствуют координатам частицы и одна перпендикулярная им ось соответствует высоте последнего взаимодействия частицы. Частицы, определенные в один кластер алгоритмом agglomerative покрашены в один цвет.
Таким образом, 2d визуализация удобна при предварительной оценке семейств, а 3d визуализация при оценке результата работы конкретного алгоритма.

3.2. Введение расстояния между точками.

В работе рассматривался один вариант расстояния между точками, описанный ранее в качестве одного из распределений:

$$d_3 = \sqrt{E_i E_j} r_{ij}, \quad r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \tag{3.1}$$

Выбор d_3 основан на физических представлениях о рассматриваемых событиях.
Для каждого семейства вычисляется матрица расстояний, которая используется в методах кластеризации. В дальнейшем могут быть рассмотрены другие метрики.

3.2.1. Выбор параметра кластеризации.

В работе рассматриваются алгоритмы кластеризации, которые не требует предварительно известного числа кластеров, но каждом из них необходимо выбирать некоторый параметр, который регулирует кластеризацию.
В связи с видом формулы (3.1) расстояния между частицами для каждого семейства, параметр выбирается по N-ому перцентилю распределения ER произведения энергий на радиусы его частиц (значение ER, ниже которого находится N% элементов выборки из семейства). Значение N варьируется.

3.3. Коэффициенты эффективности кластеризации.

Определим коэффициенты, необходимые для характеристики качества кластеризации. [2]
Доминирующая высота - это высота, встречающаяся у наибольшего числа частиц, или самая популярная для кластера высота.

1. (EF1 - P). Чистота P (Purity) .
Чистота P_i вычисляется для каждого выделенного кластера:

$$P_i = \frac{n_d}{n_c},$$

где
n_d - число частиц в i-ом кластере, имеющих доминирующую высоту,
n_c - число частиц в i-ом кластере.
Чистота P_i может изменяться от $\frac{1}{n_c}$ до 1.

Чистота P получается усреднением P_i по всем кластерам в рассматриваемом семействе:

$$P = \frac{1}{N_c}\sum_{i=1}^{N_c} P_i, \quad N_c \text{ - число кластеров в семействе.}$$

Пример 1.

Пусть есть 2 кластера, $N_c = 2$. В первом кластере 3 частицы $n_c = 3$ с высот h_1, h_1, h_2. Две частицы имеют доминирующую высоту h_1 по определению, $n_d = 2$, тогда $P_1 = 2/3 \approx 0.667$.
Пусть во втором кластере 4 частицы $n_c = 4$ с высот h_2, h_2, h_2, h_3. Три частицы $n_d = 3$ имеют доминирующую высоту h_2, поэтому $P_2 = 3/4 = 0.75$.
Следовательно, $P = \frac{1}{2}(P_1 + P_2) \approx 0.709$.

2. (EF2 - S). Фрагментарность S (Splitting).

$$S = \frac{N_d}{N_c}$$

где
N_d - число **различных** доминирующих высот в семействе по всем кластерам,
N_c - общее число выделенных алгоритмом кластеров.

Пример 2.

Пусть есть 2 кластера, $N_c = 2$. В первом кластере 3 частицы $n_c = 3$ с высот h_1, h_1, h_2. Две частицы имеют доминирующую высоту h_1 по определению, $n_d = 2$.
Пусть во втором кластере 4 частицы $n_c = 4$ с высот h_2, h_2, h_2, h_3. Три частицы $n_d = 3$ имеют доминирующую высоту h_2.
Число различных доминирующих высот в семействе равно $N_d = 2$ - это высоты h_1, h_2.
По определению, фрагментарность равна $S = 2/2 = 1$.

3. (EF3 - I) Целостность I (Integrity).

$$I = \frac{N_d}{N_I}$$

где
N_d - число **различных** доминирующих высот в семействе по всем кластерам
N_I - общее число уникальных высот в семействе.
Частицы, пришедшие с одной высоты относятся к одному взаимодействию. Таким образом N_I - это общее число последних взаимодействий, дающих вклад в рассматриваемое семейство.

Уточнение про различные высоты.

Если алгоритм выделил два кластера, в каждом из которых наибольшее число частиц пришло с одной высоты, то доминирующая высота будет одна. Следовательно, 2 кластера, 1 высота, $I = 1/1 = 1$.

Пример 3.

Пусть есть 2 кластера, $N_c = 2$. В первом кластере 3 частицы $n_c = 3$ с высот h_1, h_1, h_2. Две частицы имеют доминирующую высоту h_1 по определению, $n_d = 2$.

Пусть во втором кластере 4 частицы $n_c = 4$ с высот h_2, h_2, h_2, h_3. Три частицы $n_d = 3$ имеют доминирующую высоту h_2.

Число различных доминирующих высот в семействе равно $N_d = 2$ - это высоты h_1, h_2. Общее число высот $N_I = 3$ это высоты h_1, h_2, h_3

По определению, целостность равна $I = 2/3 \approx 0.667$.

4. (EF4 - E). Эффективность E (Efficiency) .

Эффективность E_i вычисляется для каждого выделенного кластера:

$$E_i = \frac{n_d}{n_I}.$$

где

n_d - число частиц в i-ом кластере, имеющих доминирующую высоту,

n_I - число частиц в семействе с такой высотой.

Эффективность E получается усреднением T_i по всем кластерам в рассматриваемом семействе:

$$E = \frac{1}{N_c}\sum_{i=1}^{N_c} E_i, \quad N_c \text{ - число кластеров в семействе.}$$

Пример 4.

Пусть есть 2 кластера, $N_c = 2$. В первом кластере 3 частицы $n_c = 3$ с высот h_1, h_1, h_2. Две частицы имеют доминирующую высоту h_1 по определению, $n_d = 2$.

Пусть во втором кластере 4 частицы $n_c = 4$ с высот h_2, h_2, h_2, h_3. Три частицы $n_d = 3$ имеют доминирующую высоту h_2.

Всего в семействе $n_h = 3$ частицы с высотой h_1, поэтому $E_1 = 3/3 = 1$ и $n_h = 4$ частицы с высотой h_2, поэтому $E_2 = 3/4 = 0.75$.

Следовательно, $E = \frac{1}{2}(E_1 + E_2) = 0.875$

3.4. Метрика качества кластеризации.

Для сравнения алгоритмов предложена метрика, которая учитывает все введённые коэффициенты качества кластеризации:

$$M = \frac{(E + P)(I + S)}{4}. \tag{3.2}$$

При этом значение $M \in (0, 1]$.

Метрика учитывает то, что необходимо найти оптимальный баланс отдельно между P и E, отдельно между S и I, при этом одновременно иметь высокое значение для этих двух пар коэффициентов (произведение независимых величин).

Глава 4

Методы кластеризации

Кластеризация применяется для выделении в семействе отдельных групп частиц, родившихся в результате одного последнего взаимодействия.
В процессе работы было рассмотрено 7 алгоритмов кластеризации, среди которых наилучшие результаты показали слудующие алгоритмы:
Агломеративная кластеризация,
Основанная на плотности пространственная кластеризация для приложений с шумами (DBSCAN),
Алгоритм Памир.
Так же рассматривались алгоритмы Mean Shift, HDBSCAN, OPTICS, Affinity Propogation. Они описаны в этом разделе.

4.1. Hierarchical clustering. Agglomerative clustering

Метод агломеративной иерархической кластеризации основан на создании вложенных кластеров [6]. Принимает на вход предварительно вычисленную матрицу расстояний, параметр граничного значения для выбора кластеров и метрику определения расстояния между кластерами.

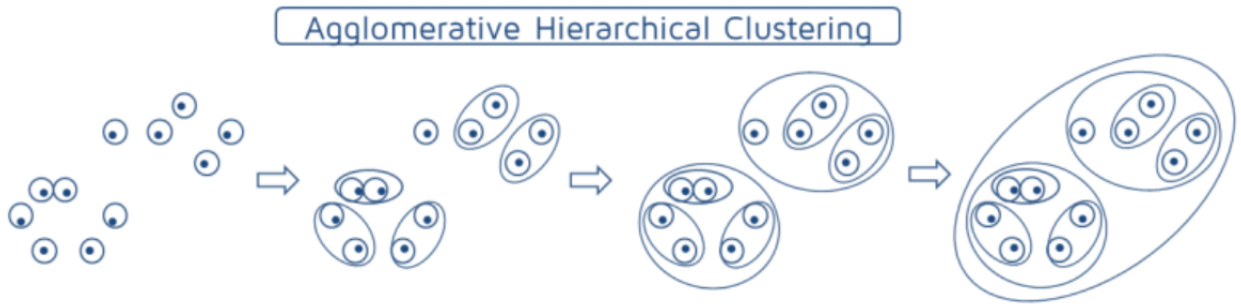


Рис. 4.1. Агломеративная кластеризация

Описание алгоритма
Иерархическая кластеризация - это кластеризация, в которой кластеры получаются вложенными друг в друга. Агломеративный подход состоит в том, что изначально каждый объект считается кластером, а затем происходит их слияние. Слияние продолжается, пока не образуется один большой кластер.
Рассмотрим различные формулы определения расстояний между кластерами.
1. Single linkage - минимум попарных расстояний между точками двух кластеров.

$$d(C_i, C_j) = min_{x_i \in C_i, x_j \in C_j} ||x_i - x_j|| \tag{4.1}$$

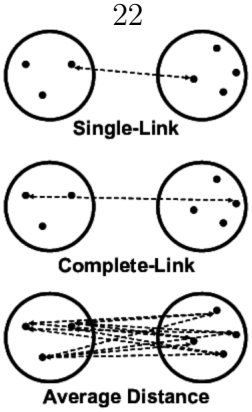


Рис. 4.2. Расстояния для агломеративной кластеризации

2. Complete linkage - максимум попарных расстояний между точками двух кластеров.

$$d(C_i, C_j) = max_{x_i \in C_i, x_j \in C_j} ||x_i - x_j|| \tag{4.2}$$

3. Average linkage - среднее попарных расстояний между точками из двух кластеров.

$$d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x_i \in C_i} \sum_{x_j \in C_j} ||x_i - x_j|| \tag{4.3}$$

Дендрограмма, или дерево склеивания кластеров, показывает итеративный процесс объединения кластеров, причем расстояние между кластерами отображается как высота дуги, которой соединяются соответствующие метки кластеров. С ее помощью можно подобрать параметр граничного значения для выбора оптимального числа кластеров.
Рассмотрим пример. Дендрограмма показывает степень близости отдельных объектов и кластеров. Количество уровней дендограммы соответствует числу шагов слияния или разделения кластеров. Внизу на рисунке расположена примерная шкала расстояний между объектами.

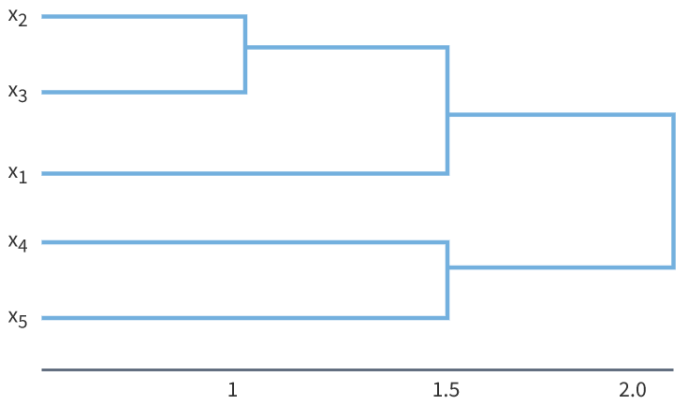


Рис. 4.3. Дендрограмма

Объекты x_2 и x_3 группируются на первом этапе, образуя кластер (x_2, x_3) с минимальным расстоянием между объектами примерно равным 1.

Затем объекты x_4 и x_5 группируются в другой кластер (x_4, x_5) с расстоянием между ними, равным 1,5.
Расстояние между кластерами (x_2, x_3) и (x_1) также оказывается равным 1,5, что позволяет сгруппировать их на том же уровне, что и (x_4, x_5).
Два кластера (x_1, x_2, x_3) и (x_4, x_5)группируются на самом высоком уровне иерархии кластеров с расстоянием 2.

Рассмотрим **пример** применения алгоритма к семейству номер 518. Дендрограмма для него имеет следующий вид:

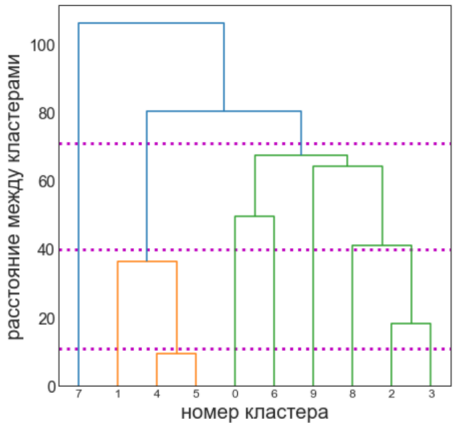


Рис. 4.4. Дендрограмма семейства 518

Здесь на горизонтальной оси числа обозначают номера кластеров. На первой итерации каждой точке присваивается свой собственный номер, который считается номером ее кластера. На вертикальной оси указана величина расстояния между кластерами, вычисленного по формуле (4.3).

$$d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x_i \in C_i} \sum_{x_j \in C_j} ||x_i - x_j||$$

Горизонтальные линии соответствуют двум различным значениям параметра ϵ необходимого для остановки процесса объединения кластеров.
Первая линия ϵ_1 соответствует 10-ому перцентилю распределения ER для семейства 518, вторая линия ϵ_2 соответствует 30-ому перцентилю распределения ER для семейства 518, третья линия ϵ_3 соответствует 80-ому перцентилю распределения ER для семейства 518.
Таким образом, в первом случае объединение остановилось на первой итерации и образовалось 9 кластеров (9 точек пересечения линий), во втором случае объединение остановилось на 3-ей итерации, когда образовалось 7 кластеров, в третьем на 7-ой итерации, когда образовалось 3 кластера.

Полученные значения параметра ϵ и коэффициентов качества кластеризации приведены в таблице:

epsilon	**перцентиль**	**значение**	**P**	**S**	**I**	**E**
1	10	10.527	1	0.778	1	0.778
2	30	39.658	0.929	0.857	0.857	0.857
3	80	71.207	0.778	1	0.429	1

Таблица 4.1. семейство 518

Также в качестве визуализации можно построить график зависимости расстояния между кластерами от номера итерации. В некоторых случаях такой график позваляет определить точку резкого возрастания графика, которая соответствует наилучшему значению для выбора расстояния.

График для семейства номер 518:

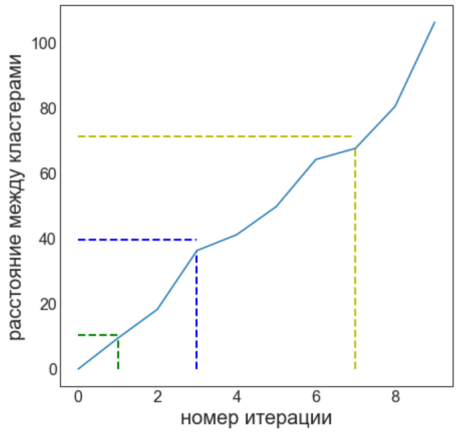


Рис. 4.5. Расстояния

На этом графике зеленая линия соответствует знаяению $\epsilon_1 = 10.527$, первая итерация. Синяя линия значению $\epsilon_2 = 39.658$, третья итерация.

4.2. Алгоритм Памир

Рассмотрим алгоритм Памир, используемый в более ранних исследованиях, посвященных кластеризации семейств. Он был повторно реализован на языке python для сравнения полученных результатов. На этом графике изображены следующие этапы алгоритма:

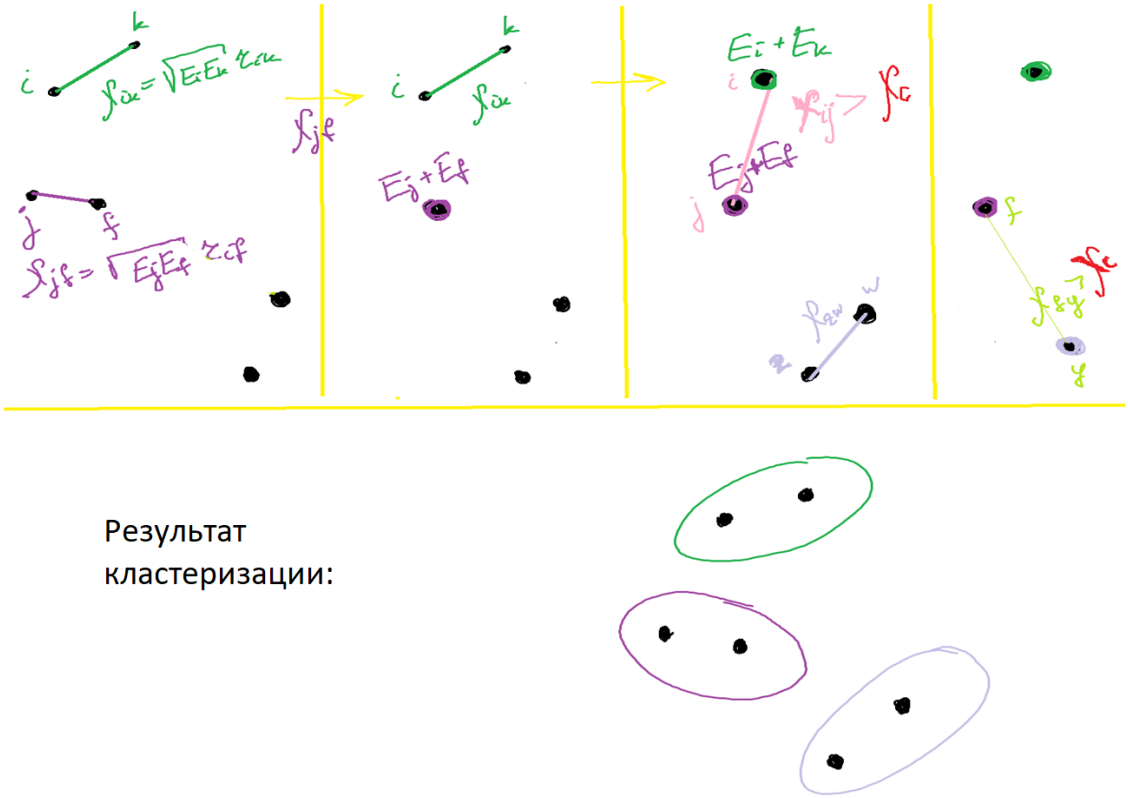


Рис. 4.6. Алгоритм Памир

1) Рассматриваем точки на плоскости, для которых известны взаимные расстояния d3:

$$d_3 = \sqrt{E_i E_j} r_{ij}, \quad \text{где} \quad r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

2) Находим точки с минимальным взаимным расстоянием.
3) Объединяем точки с минимальным расстоянием в одну энергетически взвешенно по координатам x и y, энергия e при этом суммируется.

$$x_{new} = \frac{x_i e_i + x_j e_j}{e_i + e_j} \quad y_{new} = \frac{y_i e_i + y_j e_j}{e_i + e_j} \quad e_{new} = e_i + e_j$$

4) Повторяем пункт 1), пока расстояния между всеми точками меньше заданного граничного расстояния $\chi_c = 48$ или точки не закончатся.
Параметр $\chi_c = 48$ выбран на основе более ранних результатов применения этого алгоритма, также было проверено, что при таком χ_c наблюдается наилучшее значение метрики. Значение χ_c равное 48 соответствует значению характерного поперечного импульса частицы и измеряется в ТэВ*см (размерность метрики).
Реализацию алгоритма можно посмотреть по ссылке в Выводах (в конце).

4.3. DBSCAN

Основанная на плотности пространственная кластеризация для приложений с шумами (Density-based spatial clustering of applications with noise) - это алгоритм, основанный на плотности распределения точек данных. На вход принимает матрицу расстояний и два дополнительных параметра: максимальное расстояние между соседними объекта ϵ и минимальное количество соседних объектов, необходимых для образования кластера $minP$. Эти параметры определяют степень плотности данных. [11] [12]

Описание алгоритма

Выбираем некоторую точку A как необработанную точку.

Отмечаем данную точку A как обработанную.

Находим соседние для точки A объекты, то есть объекты, находящиеся в ее ϵ- окрестности.

Точка A является *основной* точкой, если в ее ϵ-окрестности находится по крайней мере minP точек, считая ее саму. Эти точки называются *прямо достижимыми* из A. Точка C называется *достижимой* из A, если существует такой путь $y_1, .., y_N$, где $y_1 = A$ и $y_n = C$, где каждая точка y_{i+1} *прямо достижима* из y_i (все точки пути, кроме y должны быть основными).

Все точки, не достижимые из основных точек, считаются выбросами.

Если x является основной точкой, то она формирует *кластер* вместе со со всеми точками, достижимыми из этой точки. Каждый кластер содержит по крайней мере одну основную точку. Неосновные точки формируют *край* кластера, поскольку не могут быть использованы для достижения других точек.

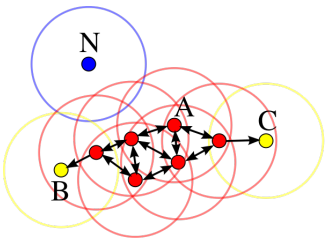


Рис. 4.7. Пример DBSCAN

В качестве примера на рисунке minP = 4. Точка A и другие красные точки являются основными, так как в их окрестностях есть четыре точки. Точки B и C являются достижимыми. Точка N является шумом.

4.4. Mean Shift

Метод сдвига среднего значения (Mean Shift) - это непараметрический анализ пространства признаков для определения местоположения максимума плотности вероятности. Не требует задания числа кластеров, принимает на вход матрицу расстояний и ширину пропускания (bandwidth).

Описание алгоритма

MeanShift это итеративный метод, шагом итерации является сдвиг среднего значения. На каждой итерации скользящее окно смещается в точку с более высокой плотностью (определяются местоположения максимумов (мод) плотности вероятности, задаваемой дискретной выборкой). Когда несколько скользящих окон перекрываются, сохраняется окно, содержащее наибольшее количество точек. Затем данные группируются в соответствии со скользящим окном, в котором они находятся.

Пусть есть текущая оценка x и задана некоторая функция $K(x_i - x)$ - ядерная функция, которая определяет вес ближайших точек для переоценки среднего.

В реализации python используется плоское ядро K:

$$K(x) = \begin{cases} 1 & ||x|| \leq h \\ 0 & ||x|| > h \end{cases}$$

Взвешенное среднее плотности в окне, определенном функцией K равно:

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)} \tag{4.4}$$

где $N(x)$ - окрестность точки x (точки, такие что $K(x_i) \neq 0$), $(m(x) - x)$ - сдвиг среднего значения, h - ширина пропускания или размер окна.

Далее x присваивается значение m(x) и оценка повторяется, пока m(x) не сойдется, т.е. пока $||m(x) - x|| > \epsilon$.

Для описания сгущения точек вводится функция плотности вероятности по заданным точкам x_i:

$$f_K(x) = \sum_i K(x - x_i) = \frac{1}{Nh^d} \sum_i^N K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right), \tag{4.5}$$

где d - число признаков у точек, N - число точек. Плотность вероятности можно переписать:

$$f_K(x) = \frac{c_{k,d}}{Nh^d} \sum_i^N k\left(\frac{||\mathbf{x} - \mathbf{x}_i||^2}{h^2}\right), \tag{4.6}$$

где , k - профиль ядра $K(x) = c_{k,d} k(||x||^2)$.

Точки сгущения соответствуют локальным максимумам функции f(x). Чтобы найти, к какому из центров сгущения относится точка, нужно идти по градиенту f(x) для нахождения ближай-

шего локального максимума.

$$\nabla f_K(x) = \frac{2c_{k,d}}{Nh^{d+2}} \sum_i^N (\mathbf{x} - \mathbf{x}_i) k'\left(\frac{||\mathbf{x} - \mathbf{x}_i||^2}{h^2}\right). \tag{4.7}$$

С учетом предположения существования производной определим $g(x) = -k'(x)$ и $G(x) = c_{g,d} g(||x||^2)$. Тогда получим

$$\begin{aligned} \nabla f_K(x) &= \frac{2c_{k,d}}{Nh^{d+2}} \sum_i^N (\mathbf{x}_i - \mathbf{x}) g\left(\frac{||\mathbf{x} - \mathbf{x}_i||^2}{h^2}\right) \\ &= \frac{2c_{k,d}}{Nh^{d+2}} \left[\sum_i^N g\left(\frac{||\mathbf{x} - \mathbf{x}_i||^2}{h^2}\right)\right] \left[\frac{\sum_{i=1}^N \mathbf{x}_i g\left(\frac{||\mathbf{x}-\mathbf{x}_i||^2}{h^2}\right)}{\sum_{i=1}^N g\left(\frac{||\mathbf{x}-\mathbf{x}_i||^2}{h^2}\right)} - x\right] \\ &= f_G(x) \frac{2c_{k,d}}{h^2 c_{g,d}} \mathbf{m}_{h,G}(x) \end{aligned} \tag{4.8}$$

Первый множитель пропорционален плотности функции вероятности, второй множитель является сдвигом среднего значения. Таким образом, вектор сдвига среднего значения всегда указывает в направлении максимального увеличения плотности. Это свойство впервые было замечено Фукунагом и Хостетлером (Fukunaga and Hostetler).

4.5. Affinity Propagation

Метод распространения близости (Affinity Propagation) принимает на вход матрицу схожести между элементами датасета, то есть матрицу расстояний, и возвращает набор меток кластеров, присвоенных элементам.

Алгоритм является детерминированным, количество кластеров не имеет значения. На матрицу схожести (метрику) не наложено никаких ограничений.

Описание алгоритма

Пусть для каждой точки существует правило $s(i, k)$, которое определяет, насколько точки i и k схожи между собой. Для каждой пары точек эти значения записаны в матрице S. Среди точек выбирается ключевая *лидер*, вокруг которой собираются наиболее похожие на неё. Каждая точка считается не похожей сама на себя $s(k, k) < 0$, таким образом, точка становится ключевой после сравнения с другими точками или в случае, если нет ни одной похожей на неё точки.

Кроме схожести определяются еще два парметра. Ответственность R - матрица с элементами $r(i, k)$ (responsibility) отвечает за то, насколько i может считать k своей ключевой точкой или лидером. Доступность A - матрица с элементами $a(i, k)$ (availability) показывает насколько хорошо k готова стать лидером для i. Ответственность и доступность вычисляются точкой и для самой себя.

Первоначально инициализируются $R = 0$, $A = 0$. Точки присоединяются к лидеру, для которого у них наибольшая сумма $a(i, k) + r(i, k)$. В случае, если $a(i, i) + r(i, i)$ максимален точка сама

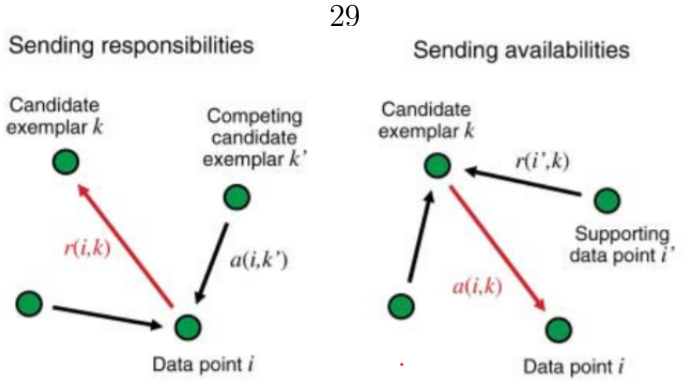


Рис. 4.8. AffinityPropagation

становится лидером.

Иначе говоря, на основе матрицы схожести S каждая точка сначала определяет наиболее похожие по значениям на себя точки - это матрица R, затем подсчитывается, для каждой точки, сколько других точек посчитали ее похожей на себя - это матрица A. Далее выбирается лидер и происходит объединение.

Данная задача является задачей дискретной максимизации с ограничениями. Необходимо найти такой вектор меток $\mathbf{c} = \{c_1, c_2, .., c_N\}, c_i \in 1...N$, который максимизирует функцию

$$S(c) = \sum_{i=1}^{N} s(i, c_i) + \sum_{k=1}^{N} \delta(c_k) \tag{4.9}$$

где $\delta(c_k)$ - член ограничитель, равный $-\infty$, если существует точка i, которая выбрала точку k своим лидером ($c_i = k$), но сама k не считает себя лидером ($c_k \neq k$). Это NP-сложная задача, но для ее решения существует несколько приближенных алгоритмов.

В рассматриваемом методе s_i, c_i, δ_i представляются вершинами двудольного графа, после чего между ними происходит обмен информации, позволяющий с вероятностной точки зрения оценить, какая метка лучше подойдет для каждого элемента.

Формальная запись алгоритма

Вход: матрица S.

Инициализация R =0, A =0;

Начало цикла на T итераций:

Обновить матрицу R;

$$r_{ij} = (1 - \lambda)\rho_{ij} + \lambda r_{ij}, \quad \text{где } \rho_{ij} \text{ - распространяемая ответственность} \tag{4.10}$$

$$\rho_{ij} = \begin{cases} s_{ij} - max_{k \neq j}\{a_{ij} + s_{ij}\} & (i \neq j) \\ s_{ij} - max_{k \neq j}\{s_{ik}\} & (i = j) \end{cases}$$

Обновить матрицу A;

$$a_{ij} = (1 - \lambda)\gamma_{ij} + \lambda a_{ij}, \quad \text{где } \gamma_{ij} \text{ - распространяемая доступность} \tag{4.11}$$

$$\gamma_{ij} = \begin{cases} min\{0, r_{ij} + \sum_{k \neq i,j} max(0, r_{kj})\} & (i \neq j) \\ \sum_{k \neq i,j} max(0, r_{kj}) & (i = j) \end{cases}$$

λ - коэффициент затухания, введеный во избежание численных колебаний.

Конец цикла.

Вычислить:

$$c_i = \underset{k}{argmax}(a_{ik} + r_{ik}). \tag{4.12}$$

Выход: $c_1, c_2, ..., c_N$ - метки отношения к конкретному лидеру для каждой точки.

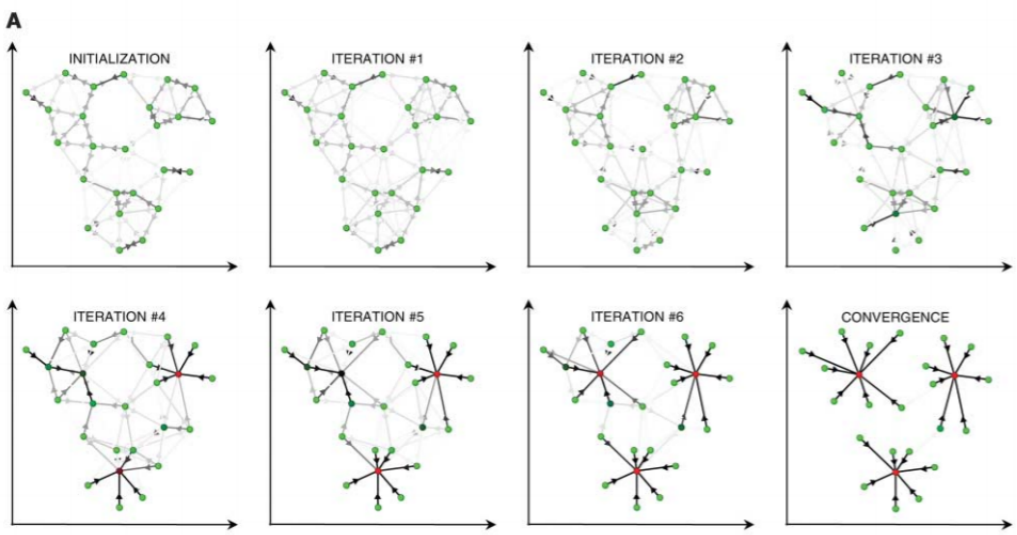


Рис. 4.9. AffinityPropagation

4.6. OPTICS

Упорядочение точек для обнаружения кластерной структуры (Ordering points to identify the clustering structure) - это алгоритм поиска кластеров на основе плотности данных. Учитывает возможность различной плотности данных. Принимает на вход матрицу расстояний и два дополнительных параметра: максимальное расстояние между соседними объекта ϵ и минимальное количество соседних объектов, необходимых для образования кластера $minP$.

Описание алгоритма

Основная идея состоит в том, что точки упорядочиваются специальным образом, который в первую очередь учитывает *соседей* - наиболее близкие точки. Это необходимо для обнаружения кластеров большей плотности. Вводится специальное расстояние для каждой точки, которое характеризует плотность кластера необходимую для помещения точек на этом расстоянии в один кластер.

Иллюстрация проблемы кластеров различной плотности приведена на рисунке. На нем изображено два кластера C_1 и C_2 выделенные при значении окрестности, равном $\epsilon_2 < \epsilon_1$ и один кластер C, выделенный при значении окрестности ϵ_1.

Точка x является *основной* точкой, если в ее ϵ-окрестности находится по крайней мере minP точек, считая ее саму. Вводится *основное* расстояние ρ_c (core-distance) - это минимальное из

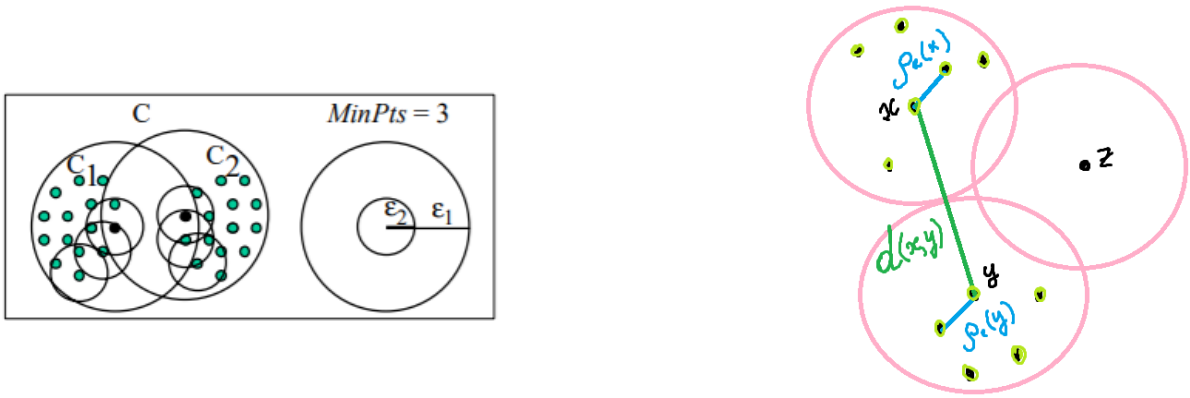


Рис. 4.10. Основные и достижимые расстояния

всех расстояний от точки x до точки в окрестности x:

$$\rho_c(x) = \begin{cases} \text{неопределено} & N_\epsilon(x) < minP \\ \text{минимум из значений расстояний до точек в окрестности} & N_\epsilon(x) \geq minP \end{cases}$$

где $N_\epsilon(x)$ - число точек в ϵ-окрестности точки x.

Достижимое расстояние ρ_r (reachability-distance) между точками x и y - это либо расстояние между этими точками, либо основное расстояние ρ_c для точки x, в зависимости от того, какое из этих расстояний больше. Таким образом:

$$\rho_r(x, y) = \begin{cases} \text{неопределено} & N_\epsilon(y) < minP \\ max(\rho_c(y), dist(x, y)) & N_\epsilon(y) \geq minP \end{cases}$$

Здесь dist(x, y) расстояние между x и y, определенное в матрице расстояний. Расстояни $\rho_r(x, y)$ точки x относительно точки y это наименьшее растояние такое что, x достижима из y. Расстояние $\rho_r(x, y)$ не может быть меньше, чем $\rho_c(y)$ до y, потому что на меньших расстояниях ни один объект не достижим от y.

Достижимое расстояние не определено, если y не является основной точкой в выбранной окрестности ϵ, то есть нет достаточной плотности точек в этой окрестности. Если x и y являются достижимыми после определения расстояний, то они принадлежат к одному кластеру.

4.7. HDBSCAN

Иерархическая основанная на плотности пространственная кластеризация для приложений с шумами (Hierarchical density-based spatial clustering of applications with noise) - алгоритм иерархической кластеризации, основанный на плотности распределения точек данных. На вход принимает матрицу расстояний и дополнительные параметры: число k ближайших соседей, минимальный размер кластера (минимальное число точек для образования кластера).

Описание алгоритма

Алгоритм можно разбить на пять шагов.

Первый шаг. Преобразование пространства в соответствии с плотностью. Необходимо определить области с наибольшей плотностью точек. Для этого определим расстояние до k-ого ближайшего соседа как *основное* расстояние ρ_{ck} (core-distance), минимальное из расстояний до всех точек в окрестности. Далее определим *достижимое* расстояние ρ_{mr} (mutual reachability distance):

$$\rho_{mr}(x, y) = max(\rho_{ck}(y), \rho_{ck}(x), dist(x, y))$$

Здесь dist(x, y) расстояние между x и y согласно матрице расстояний. Важен выбор числа k ближайших соседей.

Рассмотрим выбор расстояний на примере. Нарисуем для данной точки синий круг, пусть у точки пять ближайших соседей, считая ее саму. Тогда основное расстояние - это радиус окружности, проходящей через шестую ближайшую точку. Аналогичное основное расстояние можно построить для зеленой точки, тогда окружность будет иметь больший радиус. Построим основное расстояние и для третьей красной точки. Рассмотрим синюю и зеленую точку, достижимое расстояние между ними равно основному расстоянию для зеленой точки. При этом достижимое расстояние между красной и зеленой точкой равно именно расстоянию между красной и зеленой точкой, так как оно больше основных.

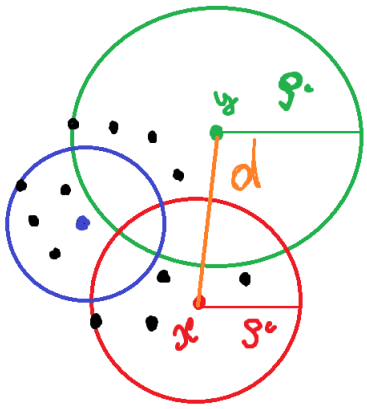


Рис. 4.11. Вложенные кластеры

Утверждается, что введение такого определения расстояния достижимости точки позволяет учесть различие в плотности данных и отнести более разреженные области к одному кластеру.

Второй шаг. Построение минимального остовного дерева. Рассмотрим взвешенный граф с точками данных в качестве вершин и ребрами между точками с весом, равным взаимному расстоянию достижимости между этими точками, определенному ранее. Нам нужно получить подграф с теми же вершинами, имеющий минимальную сумму весов входящих ребер, при этом удаление любого ребра приводит к нарушению связности графа. Для получения такого графа используется алгоритм Прима для построения минимального остовного дерева, в котором последовательно выбирается одно ребро с наименьшим весом.

Третий шаг. Построение кластерной иерархии. Ребра сортируются в порядке возрастания и происходит поиск объединения. Иерархическую кластеризацию можно визуализировать с помощью дендрограммы. Она показывает степень близости отдельных объектов и кластеров.

Возникает вопрос, где нужно провести линию, чтобы определить кластеры, которые она пересечет. Если просто провести линию разреза, то получится фиксированная плотность кластера, так как выбор линии разреза - это выбор растояния взаимной достижимости. Однако нам необходимо рассмотреть кластеры переменной плотности. То есть необходимо иметь возможность разрезать дерево в разных местах для определения кластеров. Именно на этом шаге начинается основное отличие HDBSCAN от алгоритма DBSCAN.

Четвертый шаг. Сжатие дерева склеивания кластеров. Имеющуюся дендрограмму или дерево кластеров сложной структуры нужно сжать в меньшее дерево. Часто бывает что разделение кластера - это одна или две точки, отделяющиеся от кластера. Тогда рассмотрим такое разделение в качестве единого кластера с *выпавшими точками*. Для этого вводится понятие *минимального размера кластера* (minimum cluster size). Если точек в кластере меньше, чем минимальный размер, то отмечаются точки, *выпавшие из кластера* и расстояние, на котором это произошло. Кластеры с достаточным количеством точек остаются без изменений. Таким образом, получится меньшее дерево с небольшим количеством узлов, каждый из которых содержит информацию об уменьшении размера кластера с изменением расстояния. Кластеры на дендрограмме приобретают *ширину* - значение расстояния, на котором точки выпадают из кластера (width varies over the length of the line as points fall out of the cluster).

Пятый шаг. Извлечение кластеров из сжатого дерева. В качестве визуального объяснения выбираются те кластеры, которые имеют наибольшую площадь на дендрограмме. Дополнительное требования состоит в том, что нельзя выбрать в качестве кластера потомка уже выбранного кластера.

Рассмотрим формализацию этого шага. Для начала нужно ввести меру измерения, которая учитывает устойчивость кластеров:

$$\lambda = \frac{1}{d}, \quad \text{где d - расстояние.} \tag{4.13}$$

Определим для данного кластера λ_{birth} - значение, при котором кластер образовался (отделился), и λ_{death} - значение, при котором кластер делится на подкластеры. Также для каждо го кластера для каждой точки p в данном кластере определим значение λ_p - значение, при котором точка выпадает из кластера. Значение λ_p находится между λ_{birth} и λ_{death}, потому что точка либо *выпадает* из кластера, либо покидает его при разделении на меньшие кластеры. Теперь для каждого кластера можно вычислить *устойчивость* (stability):

$$\sum_{p \in cluster} (\lambda_p - \lambda_{birth}) \tag{4.14}$$

Объявим все листовые узлы дерева (узлы, не имеющие дочерних элементов) кластерами. Затем пойдем вверх по дереву. Если сумма устойчивостей дочерних кластеров больше устойчивости кластера, то будем считать устойчивость кластера равной сумме дочерних устойчивостей. Если, наоборот, устойчивость кластера больше суммы устойчивостей его дочерних кластеров, то объявляем кластер выбранным и отменяем выбор всех его потомков. Как только будет достигнут дочерний узел выбор кластеров останавливается. Выбраные в конечном итоге кластеры возвращаются и называются *плоской кластеризацией*.

4.8. Выбор параметров кластеризации.

В методах кластеризации Agglomerative, DBSCAN, MeanShift необходимо выбирать параметр, который будет являться граничным значением для остановки алгоритма. Введение такого значения связано с неизвестным числом кластеров.

В связи с видом формулы (3.1) расстояния между частицами для каждого семейства параметр выбирается по N-ому перцентилю [1] распределения ER энергий и радиусов его частиц. Значение N варьируется в значениях от 5 до 100 с шагом 5.

Для всех семейства (без разбиения на группы частиц) построен график зависимости среднего значения метрики кластеризации M по формуле (3.2) от N. Для значения N, при котором метрика принимает максимальное значение проводится кластеризация.

Графики можно найти по ссылке в разделе Выводы (в конце).

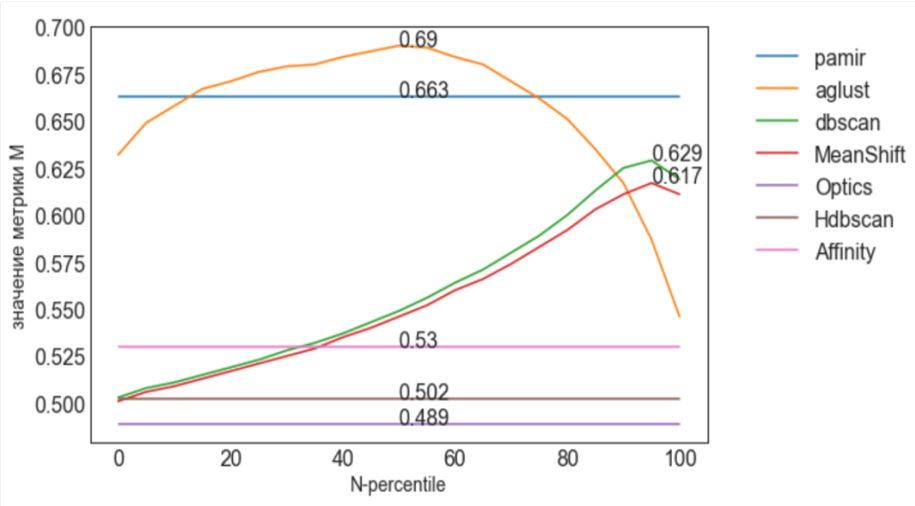


Рис. 4.12. Выбор перцентиля для кластеризации

На вертикальной оси указана величина среднего значения $M = \frac{(P+E)(S+I)}{4}$ для каждого алгоритма, на горизонтальной оси значения N.

Таблица с перцентилями, соответствующими самым высоким средним значениям метрики M:

Группа семейств	**Agglomerative**	**DBSCAN**	**MeanShift**
All	50	95	95

Таблица 4.2. Перцентили кластеризации

Так задаётся параметр граничного расстояния (threshold_distance) для Agglomerative, epsilon-окрестность для DBSCAN и параметр ширины окна (bandwidth) для MeanShift.

В дальнейшем также следует попробовать другие методы выбора наилучших параметров алгоритма.

[1] N-тый перцентиль - это число, ниже которого находится N% значений выборки. В данном случае в качестве выборки выбирается распределение по ER для данного семейства.

Глава 5

Сравнение алгоритмов.

5.1. Сравнение алгоритмпов по метрике кластеризации M.

Все алгоритмы сравнивались по величине среднего значения и распределению метрики $M = \frac{(P+E)(S+I)}{4}$. Рассмотрим гистограммы распределения значений метрики M, в которых происходит сравнение алгоритма Памир (розовый цвет) с остальными методами кластеризации.

Самое высокое среднее значение метрики равное 0.69 наблюдается у алгоритма аггломератив-

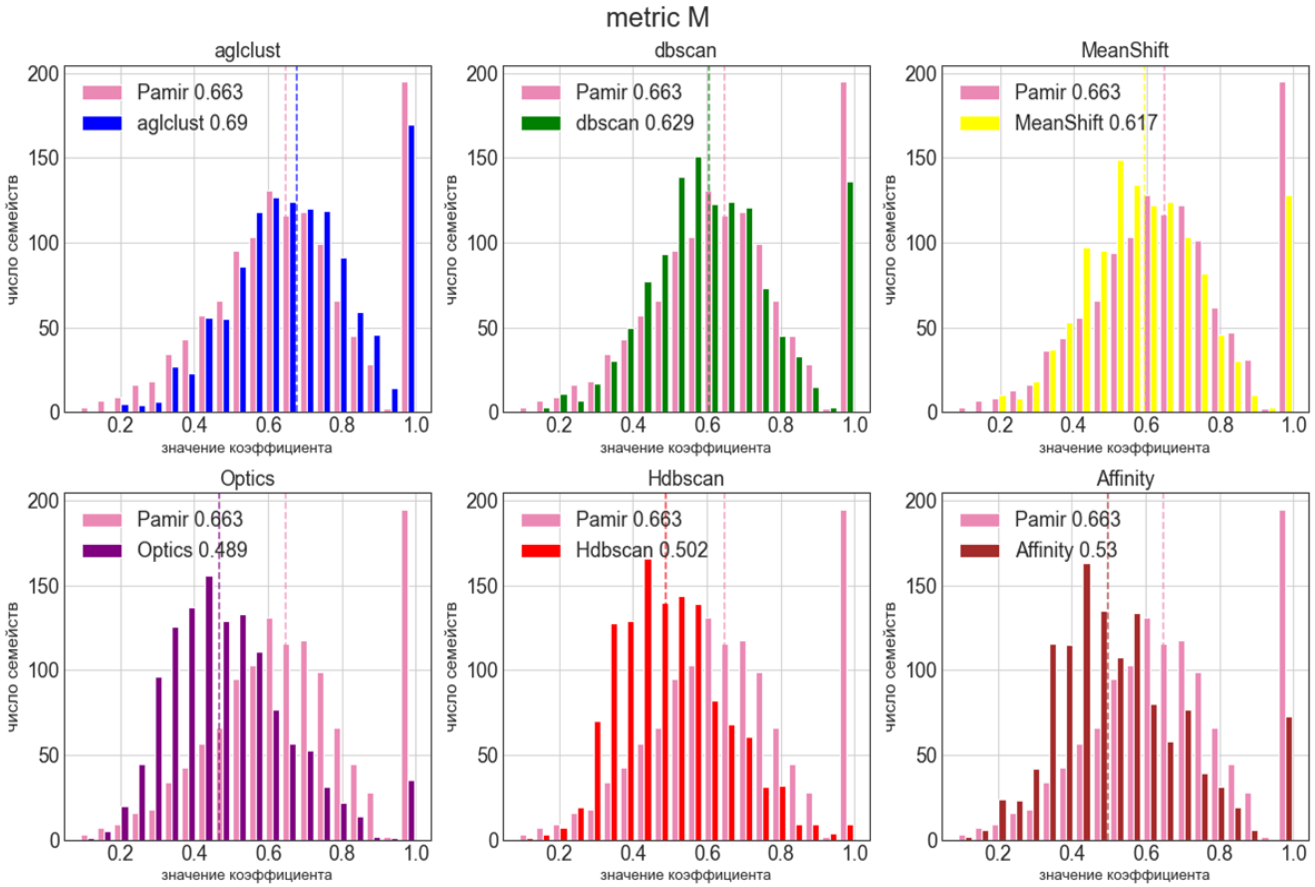


Рис. 5.1. Сравнение распределений значений M

ной кластеризации. Второе по величине значение у алгоритма Памир равняется 0.663. Третье по величине значение метрики наблюдается у алгоритма dbscan и равняетя 0.629. Остальные алгоритмы уступают. Эти средние значения (среднее по всем семействам) метрики M указаны на графике. Вертикальная черта проведена по медиане распределения.

5.2. Сравнение алгоритмпов по индексу Рэнда скорректированному на случайность (adjusted rand score).

Индекс Рэнда рассчитывает меру сходства между двумя кластеризациями, рассматривая все пары выборок и подсчитывая пары, которые приписываются одним и тем же или разным

кластерам в предсказанных и истинных кластеризациях. То есть RI - индекс Рэнда равен отношению числа согласующихся пар к числу пар или по-другому:

$$RI = \frac{TP + TN}{TP + FP + FN + TN},$$

TP - истинно положительные (правильно определены, что принадлежат одной кластеризации), TN - истинно отрицательные (правильно определены, что не принадлежат одной кластеризации), FN - ложно отрицательные (получились в разных кластерах, хотя должны быть в одном), FP - ложно положительные (получились в одном кластере, хотя должны быть в разных).
Индекс Рэнда имеет значение,близкое к 0 для случайной маркировки независимо от количества кластеров и образцов и ровно 1, когда кластеры идентичны (до перестановки).
Затем исходный показатель RI "корректируется на случайность"в показатель ARI по следующей схеме:

$$ARI = (RI - Expected\_RI)/(max(RI) - Expected\_RI)$$

Чтобы получить Expected_RI количество и размер кластеров внутри кластеризации фиксируются, и все случайные кластеризации генерируются путем перетасовки элементов между фиксированными кластерами.
Таким образом,скорректированный на случайность индекс Рэнда $ARI \in [-1, 1]$. Для индекса ARI получены следующие гистограммы распределения значений:

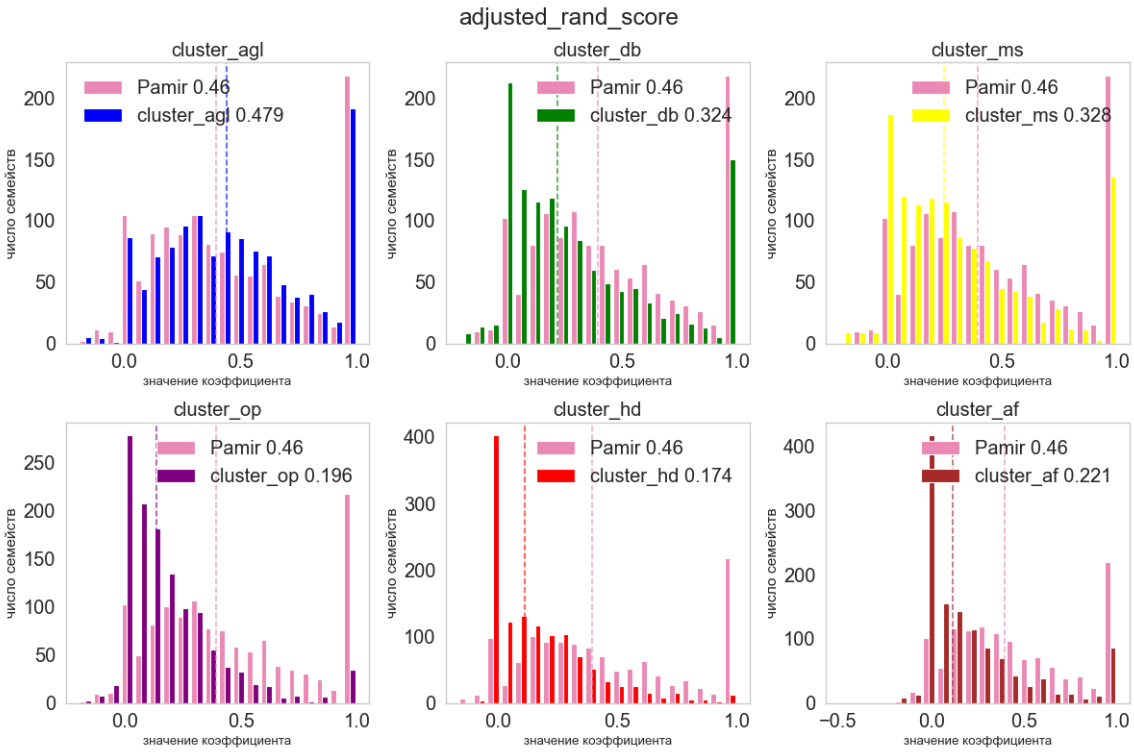


Рис. 5.2. Сравнение распределений значений ARI

Результат сравнения алгоритмов по метрики ARI согласуется с результатом сравнения алгоритмов по метрике M, наибольшее значение наблюдается у тех же алгоритмов.

5.3. Корреляции между числом кластеров и числом последних взаимодействий (высот). Калибровка алгоритмов.

Рассмотрим алгоритмы, для которых наблюдалось наиболее высокое среднее значение метрики M (аггломеративная кластеризация, алгоритм Памир, dbscan) и оценим их *откалиброванность* (другая логика оценки эффективности).

Рассмотрим зависимость числа кластеров по семействам, которое получается после применения алгоритма к модельным данным, от среднего числа высот в семействе с таким числом кластеров.

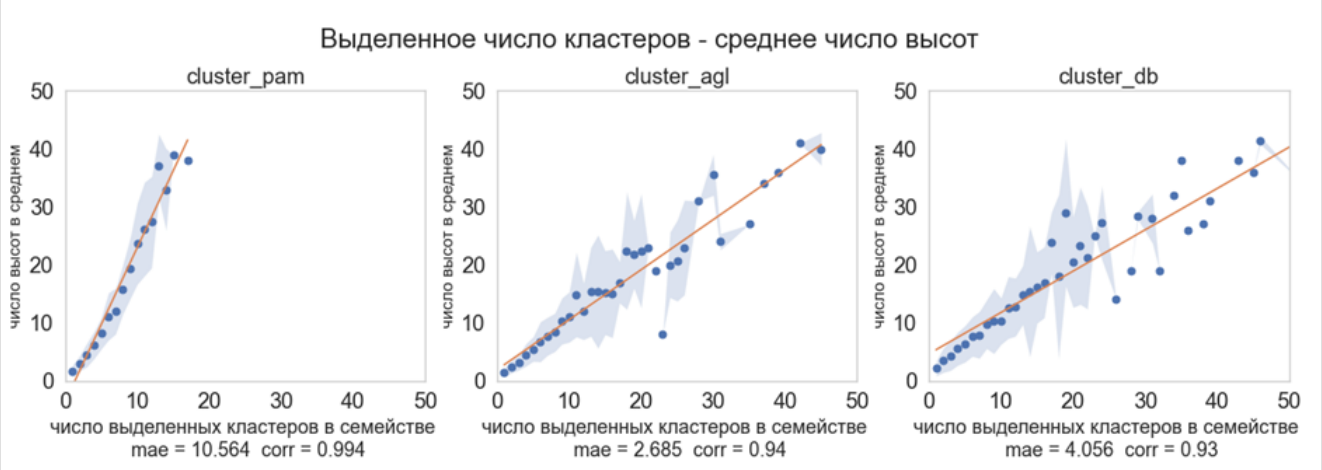


Рис. 5.3. Число кластеров - среднее число высот.

Таким образом, можно увидеть, что для рассматриваемых алгоритмов существует высокая корреляция между числом кластеров и средним числом высот по семействам, что характеризует хорошее качество алгоритма. Под графиками указаны значения, соответствующими коэффициенту корреляции Спирмена corr:

$$corr = 1 - 6\frac{\sum_{i=1}^{n}(rank(v_{1i}) - rank(v_{2i}))^2}{n(n^2-1)}$$

где v_{1i} - число выделенных в семействе кластеров, v_{2i} - число высот в среднем в множестве семейств с числом кластеров, равным v_{1i}.

Также под графиками указана средняя абсолютная ошибка рассматренных выборок:

$$mae = \frac{1}{n}\sum_{i=1}^{n}|v_{1i} - v_{2i}|$$

где n - размер выборки.

Наименьшее значение ошибки наблюдается у алгоритма аггломеративной кластеризации, наибольшее значение у алгоритма Памир. Это означает, что в множестве семейств, в которых алгоритм Памир выделяет 10 кластеров в среднем встречается по 20 уникальных высот последнего взаимодействия. То есть алгоритм аггломеративной кластеризации более точно определяет число кластеров относительно числа высот в модельных данных.

Синяя область на графике изображает коридор ошибки $(h - \sigma(h), h + \sigma(h))$, $\sigma(h)$ - стандартное

отклонение h. Оранжевая линия - прямая, построенная методом наименьших квадратов.
Наибольшее значение корреляции наблюдается у алгоритма Памир, что означает существование функциональной зависимости между числом выделенных кластеров и средним числом уникальных высот для данного множества выделенных кластеров. Для алгоритма аггломеративной кластеризации также наблюдается высокое значение корреляции.
Исходя из этого, алгоритм аггломеративной кластеризации показывает наилучший результат.

5.4. Применение кластеризации к экспериментальным данным. Устойчивость алгоритмов.

Применим алгоритмы аггломеративной кластеризации, Памир и dbscan к экспериментальным данным и посмотрим, как изменится распределения числа кластеров, это позволит оценить устойчивость алгоритмов к изменению типа данных. Для этого рассмаотрим визуальное сравнение гистограмм распределений числа выделенных кластеров экспериментальных и модельных данных и значение кросс энтропии, которая является мерой близости распределений.

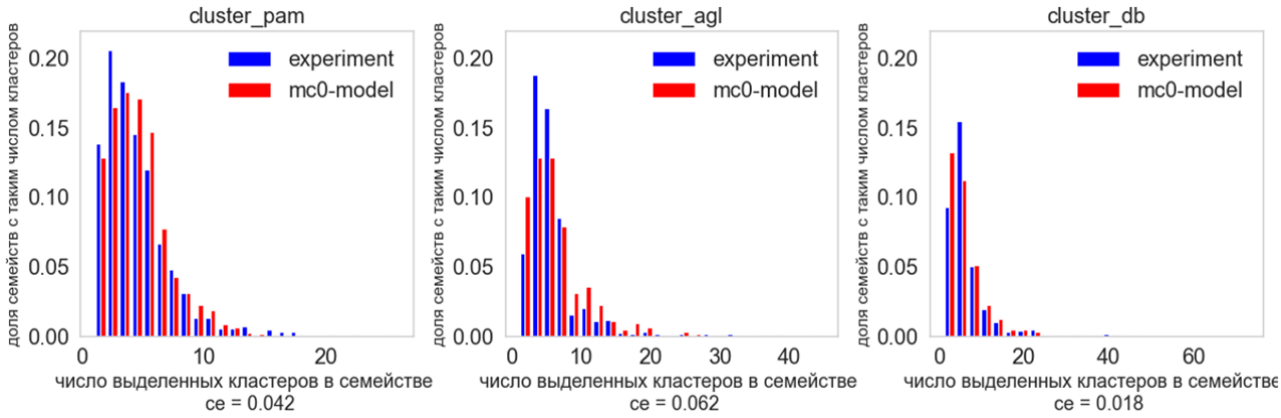


Рис. 5.4. Сравнение распределений экспериментальных и модельных данных.

Под графиками указаны значения, соответствующие значению кросс-энтропии ce:

$$ce = \sum_{i=1}^{n}(a_i - b_i)(ln(a_i) - ln(b_i)) = \sum_{i=1}^{n}(a_i - b_i)ln\frac{a_i}{b_i}, \quad a_i = \frac{n_{1i}}{n}, \quad b_i = \frac{n_{2i}}{n},$$

где n - количество элементов выборки, n_{1i}, n_{2i} - количество элементов выборки 1 и выборки 2 соответственно в интервале разбиения из распределений числа выделенных кластеров экспериментальных и модельных данных.
Для ce < 0.1 распределения принято считать близкими. Таким образом, все рассмотренные алгоритмы можно считать устойчивыми к изменению данных и переносимыми с модельных данных на экспериментальные данные.

5.5. Выводы

Проведено исследование экспериментальных ($\sim$ 1000) и модельных гамма семейств ($\sim$ 1500), произведен последовательный отбор семейств, удовлетворяющих условиям (R $<$ 15 см, E $>$ 4 ТэВ, $\sum E_\gamma > 100$ ТэВ, $N_\gamma > 3$).

Построены распределения физических характеристик экспериментальных и модельных событий. Показана их однородность для различных интервалов суммарной энергии.

Рассмотрено 7 различных алгоритмов кластеризации, не предполагающих известным число кластеров, 3 из которых признаны перспективными.

Предложен метод определения критерия остановки алгоритмов кластеризации (граничное расстояние).

Предложена метрика M эффективности алгоритма кластеризации модельных семейств.

Установлено, что наиболее эффективным с точки зрения этой метрики является алгоритм агломеративной кластеризации.

Проведено сравнение алгоритмов при помощи индеса Рэнда RI и подтверждено согласие результатов с выбранной метрикой M.

Показано наличие значимых корреляций между числом кластеров и числом последних взаимодействий и оценена ошибка алгоритма.

Реализован алгоритм Памир и подтверждены результаты, полученные в более ранних исследованиях.

Построены визуализации семейств и распределений (2d и 3d).

Приложение.

Все таблицы с результатами, визуализации и алгоритмы можно найти по следующей **ссылке**: https://github.com/MaryIzo/coursework.

Таблица UnifotmityOldBank.xlsx содержит p-значения статистических критериев.

Таблица UnifotmityOldBankConclusions.xlsx содержит выводы на основе статистических критериев.

Программная реализация содержится в файлах с расширением .py и .ipynb.

В папке images содержатся все построенные графики.

В папке web содержится html визуализация.

Список литературы

1. Кобзарь А. И. "Прикладная математическая статистика". — — М.: Физматлит, 2006. — 238 с.
2. С.Г. Байбурина, А.С. Борисов, З.М. Гусева, Ф. М. Дунаевский "Исследования ядерных взаимодействий в области энергий 10^{14} - 10^{17} эВ методом рентгеноэмульсионных камер в космических лучах (Эксперимент "Памир")". — Труды ордена Ленина Физического Института им. П.Н. ЛЕбедева том 154 – 218 стр.
3. Тюрин Ю. Н., Макаров А. А. "Анализ данных на компьютере" — Изд. 3-е, перераб. и доп./Под ред. В.Э. Фигурнова - М.: ИНФРА - М, 2002. - 528с., ил.
4. Таблицы математической статистики. Большев Л.Н., Смирнов Н. В. – М.: Наука. Главная редакция физико-математической литературы. 1983. -416 с.
5. Метод Монте-Карло в задачах о взаимодействии частиц с веществом / Н. М. Соболевский ; –Федеральное гос. бюджетное учреждение науки Ин-т ядерных исслед. Российской акад. наук. — Изд. 2-е, испр. и доп. — Москва : Федеральное гос. бюджетное учреждение науки Ин-т ядерных исслед. Российской акад. наук, 2014. — 169 с. : ил.; 30 см;
6. Элбон Крис Машинное обучение с использованием Python. Сборник рецептов: Пер. с англ. — СПб.: БХВ-Петербург, 2019. — 384 с.: ил.
7. Hodges, J.L. Jr., “The Significance Probability of the Smirnov Two-Sample Test,” Arkiv fiur Matematik, 3, No. 43 (1958), 469-86.
8. Welch, B. L. (1947). "The generalization of "Student's"problem when several different population variances are involved".
9. Levene, H. (1960). In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin et al. eds., Stanford University Press, pp. 278-292.
10. Brown, M. B. and Forsythe, A. B. (1974), Journal of the American Statistical Association, 69, 364-367
11. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise” Ester, M., H. P. Kriegel, J. Sander, and X. Xu, In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226–231. 1996
12. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). In ACM Transactions on Database Systems (TODS), 42(3), 19.
13. “Mean shift: A robust approach toward feature space analysis.” D. Comaniciu and P. Meer, IEEE Transactions on Pattern Analysis and Machine Intelligence (2002)