МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное
учреждение высшего образования
Дальневосточный федеральный университет

ШКОЛА ЕСТЕСТВЕННЫХ НАУК

Кафедра информационной безопасности

О Т Ч Е Т
о прохождении учебной практики (учебно-лабораторного практикума)

Выполнил студент
гр. С8117-10.05.01ммзи
_____ Смотрик Е.В.
(подпись)

Отчет защищен с оценкой

_____
_____ С.С. Зотов
(подпись) (И.О. Фамилия)
« 26 » июня 2021 г.

Руководитель практики
Старший преподаватель кафедры
информационной безопасности ШЕН
_____ С.С. Зотов
(подпись) (И.О. Фамилия)

Регистрационный № _____
« 26 » июня 2021 г.

_____ _____
(подпись) (И.О. Фамилия)

Практика пройдена в срок
с « 22 » февраля 2021 г.
по « 26 » июня 2021 г.
на предприятии
Кафедра информационной
безопасности ШЕН ДВФУ

г. Владивосток
2021

Характеристика

Выдана студенту 4 курса, специальности «Компьютерная безопасность», специализации «Математические методы защиты информации», Смотрику Егору Валерьевичу.

Смотрик Егор Валерьевич, в период с 22.02.2021 по 26.06.2021 года, проходил учебную практику (учебно-лабораторный практикум) на кафедре информационной безопасности ШЕН ДВФУ.

За время прохождения практики Егор проявил усердие, тягу к знаниям, огромное желание и трудолюбие, а также неподдельный интерес к изучению материала, требуемого для написания отчета. Приходил на консультацию вовремя с перечнем вопросов, с подробным и исчерпывающим описанием о текущем состоянии практики, со списком отмеченных задач. Внимательно изучал предложенные материалы и литературу на интересующую тематику.

Смотрик Е.В. полностью выполнил предусмотренную программу практики, продемонстрировал умения самостоятельно решать практические вопросы, применяя теоретическую базу, полученную в учебный период, а также при самостоятельном обучении.

При выполнении поставленных задач Смотрик Е.В. характеризуется инициативностью, сообразительностью и ответственностью.

Старший преподаватель кафедры
информационной безопасности _____ Зотов С.С.

ДНЕВНИК СТУДЕНТА

Дата	Рабочее место	Краткое содержание выполняемых работ	Отметки руководителя
22.02.21 – 27.04.21	КИБ	Выбор темы практической работы	
28.04.21 – 30.04.21	КИБ	Поиск материала	
01.05.21 – 05.05.21	КИБ	Анализ найденного материала	
06.05.21 – 14.06.21	КИБ	Реализация алгоритмов кластеризации	
15.06.21 – 20.06.21	КИБ	Написание отчёта по проделанной работе	
21.06.21 – 26.06.21	КИБ	Сдача готового отчета преподавателю	

Студент _____Смотрик Е.В.
подпись Ф.И.О.

Руководитель практики от ДВФУ _____Зотов С.С.
подпись Ф.И.О.

Оглавление

Задание на практику

- Практическое изучение алгоритмов кластеризации данных.

- Написание отчета по практике о проделанной работе.

Введение

Учебная практика (учебно-лабораторный практикум) проходил на кафедре информационной безопасности ШЕН ДВФУ в период с 22 февраля 2021 года по 26 июня 2021 года.

Целью прохождения практики является приобретение практических и теоретических навыков по специальности, а также навыков оформления проведенного исследования в отчетной форме.

Задачи практики:

1. Изучить существующие алгоритмы кластеризации данных.
2. Выбрать датасет и выполнить на нем несколько из алгоритмов кластеризации.
3. Сравнить выбранные алгоритмы между собой.
4. На основе полученных знаний написать отчет по практике о проделанной работе.

1. Кластеризация

Кластерный анализ или кластеризация – это задача группировки набора объектов таким образом, чтобы объекты в одной группе (называемой кластером) были более похожи (в некотором смысле) друг на друга, чем на объекты в других группах (кластерах). Это основная задача исследовательского анализа данных и общий метод статистического анализа данных, используемый во многих областях, включая распознавание образов, анализ изображений, поиск информации, биоинформатику, сжатие данных, компьютерную графику и машинное обучение.

Применение кластерного анализа в общем виде сводится к следующим этапам:

1) Отбор выборки объектов для кластеризации.

2) Определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных.

3) Вычисление значений меры сходства между объектами.

4) Применение метода кластерного анализа для создания групп сходных объектов (кластеров).

5) Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Применение кластеризации несет в себе несколько целей:

- Понимание данных путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа.

- Сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера.

- Обнаружение новизны. Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

Существует также метод группировки набора объектов, называемый классификация. Классификация — один из разделов машинного обучения, посвященный решению следующей задачи. Имеется множество объектов (ситуаций), разделённых, некоторым образом, на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов не известна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

Кластеризация отличается от классификации тем, что изначально не задано множество объектов, для которых известно, к каким классам они относятся, и даже могут быть неизвестны сами классы.

Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин:

- Не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию «по построению». Все они могут давать разные результаты.

- Число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием.

- Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

1.1 K-means кластеризация

Метод k-means — наиболее популярный метод кластеризации. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров: $V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2$, где k – число кластеров, S_i − полученные кластеры, $i = 1, 2, \ldots, k$, а μ_i – центры масс всех векторов x из кластера S_i. Данный алгоритм разбивает множество элементов векторного пространства на заранее известное число кластеров k.

Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение V уменьшается, поэтому зацикливание невозможно.

Проблемы k-means:

- Не гарантируется достижение глобального минимума суммарного квадратичного отклонения V, а только одного из локальных минимумов.

- Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.

- Число кластеров надо знать заранее.

1.2. Агломеративная кластеризация

Иерархическая кластеризация (также графовые алгоритмы кластеризации и иерархический кластерный анализ) — совокупность

алгоритмов упорядочивания данных, направленных на создание иерархии (дерева) вложенных кластеров. Выделяют два класса методов иерархической кластеризации:

- Агломеративные методы: новые кластеры создаются путем объединения более мелких кластеров и, таким образом, дерево создается от листьев к стволу;

- Дивизивные или дивизионные методы: новые кластеры создаются путем деления более крупных кластеров на более мелкие и, таким образом, дерево создается от ствола к листьям. На практике такой подход нигде не применяется из-за больших вычислительных трудностей так как необходимо рассчитать большое количество всевозможных комбинаций деления.

Алгоритмы иерархической кластеризации предполагают, что анализируемое множество объектов характеризуется определённой степенью связности. Как и большинство визуальных способов представления зависимостей графы быстро теряют наглядность при увеличении числа кластеров.

Под дендрограммой обычно понимается дерево, построенное по матрице мер близости. Дендрограмма позволяет изобразить взаимные связи между объектами из заданного множества. Для создания дендрограммы требуется матрица сходства (или различия), которая определяет уровень сходства между парами кластеров.

1.3. Метод кластеризации DBSCAN

Основанная на плотности пространственная кластеризация для приложений с шумами (англ. Density-based spatial clustering of applications with noise, DBSCAN) — это алгоритм кластеризации данных, который предложили Маритин Эстер, Ганс-Петер Кригель, Ёрг Сандер и Сяовэй Су в 1996. Это алгоритм кластеризации, основанной на плотности — если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно

расположены (точки со многими близкими соседями), помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко).

Рассмотрим набор точек в некотором пространстве, требующий кластеризации. Для выполнения кластеризации DBSCAN точки делятся на основные точки, достижимые по плотности точки и выпадающие следующим образом:

- Точка p является основной точкой, если по меньшей мере $min\_samples$ точек находятся на расстоянии, не превосходящем ε, является максимальным радиусом соседства от p), до неё (включая саму точку p). Говорят, что эти точки достижимы прямо из p.

- Точка q прямо достижима из p, если точка q находится на расстоянии, не большем ε, от точки p и p должна быть основной точкой.

- Точка q достижима из p, если имеется путь $p_1, \ldots, p_n$ с $p_1 = p$ и $p_n = q$, где каждая точка p_{i+1} достижима прямо из p_1 (все точки на пути должны быть основными, за исключением q).

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

Теперь, если p является основной точкой, то она формирует кластер вместе со всеми точками (основными или неосновными), достижимые из этой точки. Каждый кластер содержит по меньшей мере одну основную точку. Неосновные точки могут быть частью кластера, но они формируют его «край», поскольку не могут быть использованы для достижения других точек.

Достижимость не является симметричным отношением, поскольку, по определению, никакая точка не может быть достигнута из неосновной точки, независимо от расстояния (так что неосновная точка может быть достижимой, но ничто не может быть достигнуто из неё). Поэтому дальнейшее понятие связности необходимо для формального определения области кластеров,

найденных алгоритмом DBSCAN. Две точки p и q связаны по плотности, если имеется точка o, такая что и p, и q достижимы из o. Связность по плотности является симметричным.

Тогда кластер удовлетворяет двум свойствам:

- Все точки в кластере попарно связны по плотности.

- Если точка достижима по плотности из какой-то точки кластера, она также принадлежит кластеру.

2. Демонстрация работы алгоритмов

2.1. Алгоритм агломеративной кластеризации с помощью дендрограммы

Выполним алгоритм агломеративной кластеризации с помощью дендрограммы и выведем полученное изображение, рис. 1.

```
from scipy.cluster.hierarchy import dendrogram, linkage

Z = linkage(X, "ward")

# строим дендрограмму
dendrogram(Z, leaf_rotation=90.)

#визуализируем наше дерево
fig = plt.figure()
dendrogram(Z,p=200, leaf_rotation=90.)
plt.show()
```

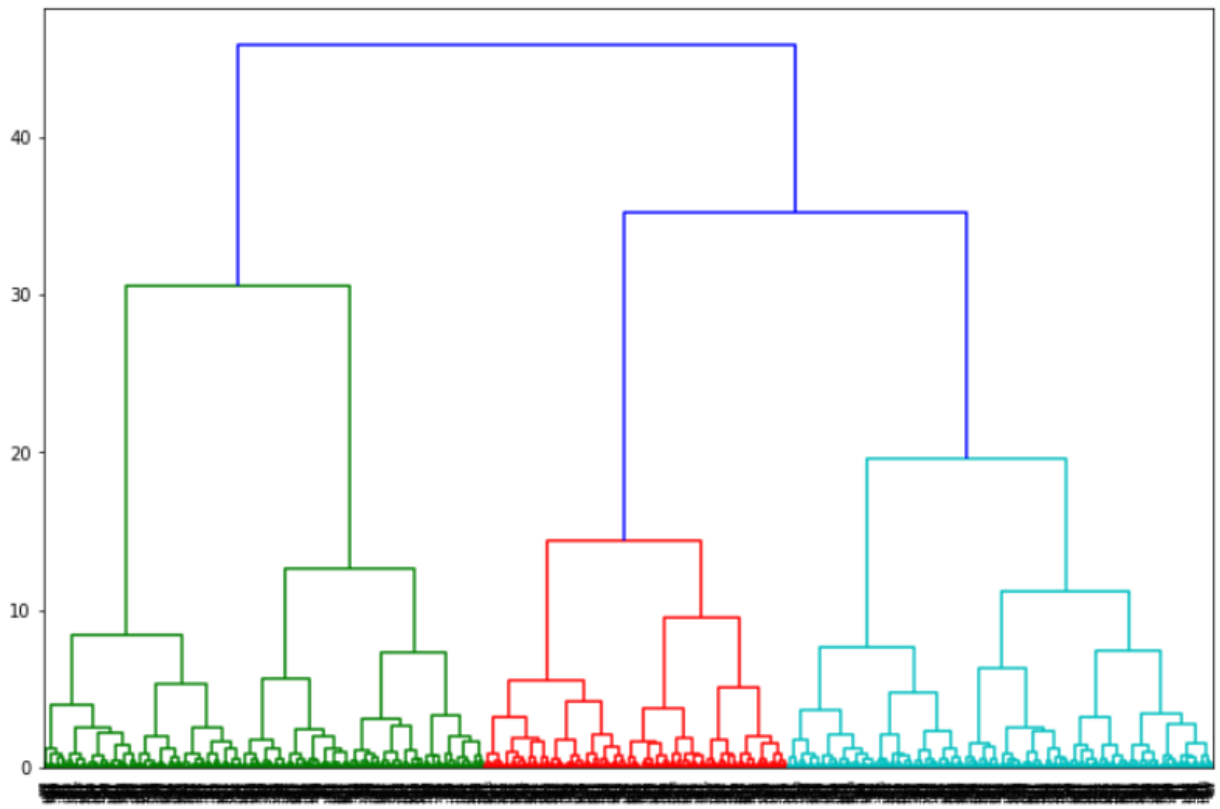

Рис. 1. Дендрограмма.

2.1. Метод кластеризации DBSCAN

Возьмем для кластеризации методом DBSCAN параметры, равные $\varepsilon = 10000$ и $min\_samples = 10$ и проведем кластеризацию нашего датасета. Код данного участка указан на рис. 2.

```
from sklearn.cluster import DBSCAN

dbscan = DBSCAN(eps=10000, min_samples=10)
dbscan.fit(X_data)
```

Рис. 2. Кластеризация датасета методом DBSCAN.

Затем выведем количество получившихся кластеров, как указано на рис. 3. Как мы видим, мы получили 189 кластеров.

```
label_list=dbscan.labels_.tolist()#получили метки кластера
len(set(label_list))#получили количество уникальных меток
```

```
189
```

Рис. 3. Количество кластеров методом DBSCAN.

Теперь выведем таблицу с полученными кластерами, как указано на рис. 4. Значение «-1» в таблице обозначает невозможность определить, к какому кластеру относится данный элемент.

```
data_res['cluster_label_dbscan']=pd.DataFrame(dbscan.labels_)
data_res
```

t_ltm	ct_dst_src_ltm	is_ftp_login	ct_ftp_cmd	ct_flw_http_mthd	ct_src_ltm	ct_srv_dst	is_sm_ips_ports	attack_cat	cluster_label_dbscan
1	1	0	0	0	1	1	0	Normal	-1
1	2	0	0	0	1	6	0	Normal	-1
1	3	0	0	0	2	6	0	Normal	-1
1	3	1	1	0	2	1	0	Normal	-1
1	40	0	0	0	2	39	0	Normal	-1
...	...	...	...	...	...	...	...	...	...
13	24	0	0	0	24	24	0	Generic	161
1	2	0	0	0	1	1	0	Shellcode	-1
3	13	0	0	0	3	12	0	Generic	161
14	30	0	0	0	30	30	0	Generic	161
16	30	0	0	0	30	30	0	Generic	161

Рис. 4. Таблица с кластерами.

Теперь изменим параметры и выведем количество полученных кластеров, как указано на рис. 5.

```
dbscan1= DBSCAN(eps=100000,
                min_samples=10)
dbscan1.fit(X_data)
label_list=dbscan1.labels_.tolist()#получили метки кластера
len(set(label_list))#получили количество уникальных меток
```

```
163
```

```
dbscan1= DBSCAN(eps=100000,
                min_samples=20)
dbscan1.fit(X_data)
label_list=dbscan1.labels_.tolist()#получили метки кластера
len(set(label_list))#получили количество уникальных меток
```

```
120
```

```
dbscan1= DBSCAN(eps=150000,
                min_samples=30)
dbscan1.fit(X_data)
label_list=dbscan1.labels_.tolist()#получили метки кластера
len(set(label_list))#получили количество уникальных меток
```

```
92
```

Рис. 5. Изменение параметров DBSCAN.

Как мы видим, от изменения параметров меняется и количество найденных кластеров.

3. Сравнение методов кластеризации с помощью метрик качества.

Метрики оценки качества кластеризации — инструментарий для количественной оценки результатов кластеризации.

Принято выделять две группы метрик оценки качества кластеризации:

1) Внешние метрики основаны на сравнении результата кластеризации с априори известным разделением на классы.

2) Внутренние метрики отображают качество кластеризации только по информации в данных.

Для оценки качества методов кластеризации сгенерируем проверочные множества, визуализация которых изображена на рис. 6. На рис. 6.а, 6.б и 6.ж предполагается разбиение на 2 кластера, на рис. 6.в, 6.д, 6.е и 6.и предполагается разделение на 3 кластера, а на рис. 6.г и 6.з предполагается отсутствие кластерной структуры. Множество на рис 6.и получено путем объединения множеств на рис. 6.ж и 6.з.

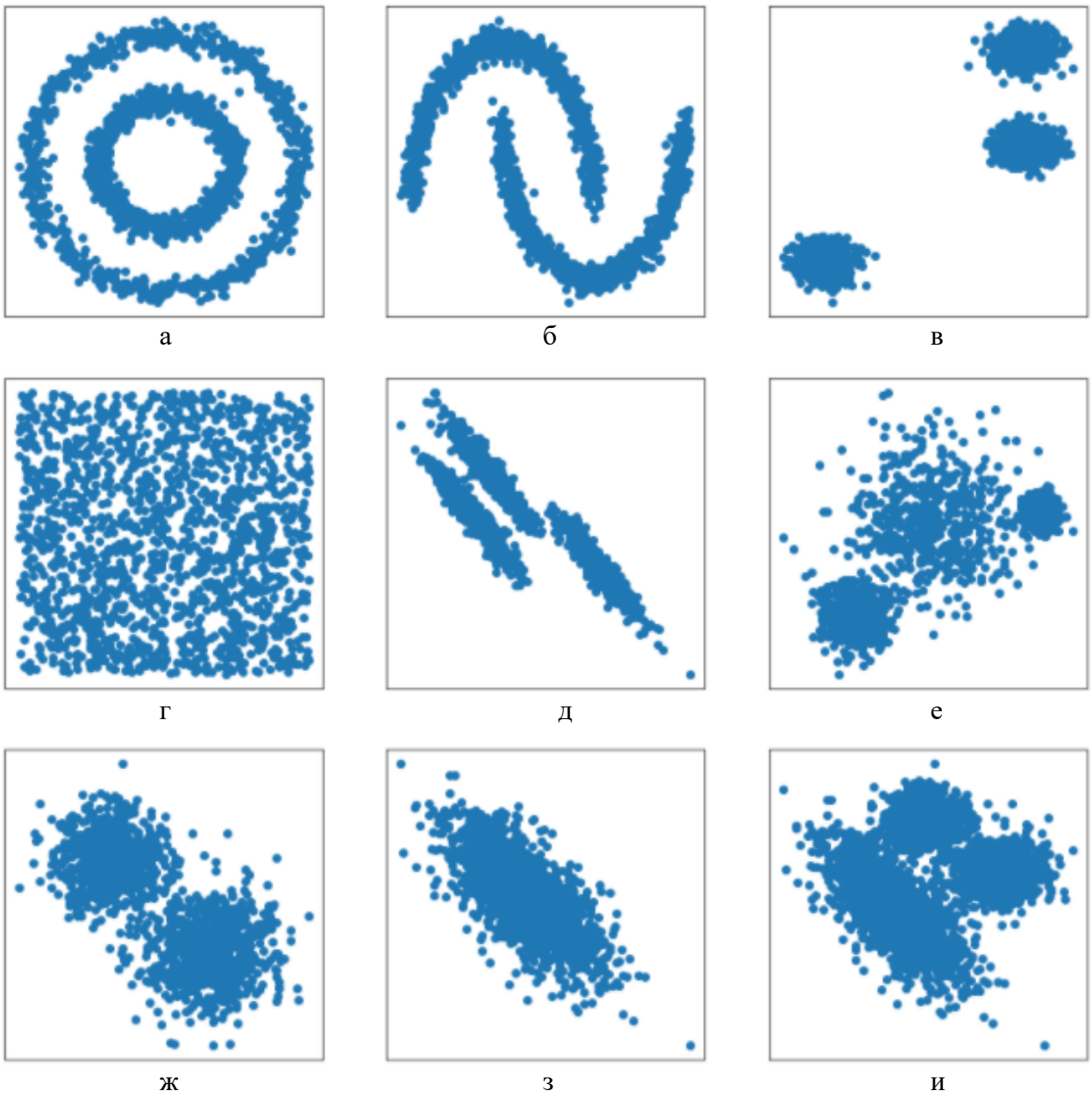

Рис. 6. Визуализация проверочных множеств.

Теперь выполним кластеризацию данных множеств и вычислим метрики. Будем рассматривать следующие метрики:

- Внутренние метрики:
  - Коэффициент силуэта;
  - Коэффициент Девиса-Болдуина;
  - Коэффициент Халинского-Харабаша.
- Внешние метрики:

- Коэффициент однородности;

- Коэффициент полноты:

- V-мера.

3.1. Коэффициент силуэта

Значение силуэта является мерой того, насколько объект похож на его собственный кластер по сравнению с другими кластерами. Силуэт варьируется от -1 до +1, где высокое значение указывает, что объект хорошо соответствует своему собственному кластеру и плохо соответствует соседним кластерам. Если большинство объектов имеют высокое значение, то конфигурация кластеризации подходит. Если многие точки имеют низкое или отрицательное значение, то в конфигурации кластеризации может быть слишком много или слишком мало кластеров. Коэффициент силуэта можно рассчитать с помощью любой метрики расстояния, такой как евклидово расстояние. Результаты вычисления данного коэффициента для различных выборок и различных методов приведен на рис. 7.

```
Коэффициент силуэта для а-го набора данных:
        -для алгоритма KMeans равен                                0.35365974000459466
        -для алгоритма агломеративной кластеризации равен          0.3230343615485682
        -для алгоритма DBSCAN равен                                0.11397816414100963
Коэффициент силуэта для б-го набора данных:
        -для алгоритма KMeans равен                                0.47094206825867163
        -для алгоритма агломеративной кластеризации равен          0.4373686298490737
        -для алгоритма DBSCAN равен                                0.33247143063895745
Коэффициент силуэта для в-го набора данных:
        -для алгоритма KMeans равен                                0.8290743874701529
        -для алгоритма агломеративной кластеризации равен          0.8290743874701529
        -для алгоритма DBSCAN равен                                0.8290743874701529
Коэффициент силуэта для г-го набора данных:
        -для алгоритма KMeans равен                                0.3880871229820807
        -для алгоритма агломеративной кластеризации равен          0.327872572743395
        -для алгоритма DBSCAN равен                                0.14286426496138094
Коэффициент силуэта для д-го набора данных:
        -для алгоритма KMeans равен                                0.5079439603683398
        -для алгоритма агломеративной кластеризации равен          0.48038722424655533
        -для алгоритма DBSCAN равен                                0.478028357286864
Коэффициент силуэта для е-го набора данных:
        -для алгоритма KMeans равен                                0.6427721508140368
        -для алгоритма агломеративной кластеризации равен          0.6257895422762809
        -для алгоритма DBSCAN равен                                0.4879584333076111
Коэффициент силуэта для ж-го набора данных:
        -для алгоритма KMeans равен                                0.629574772353261
        -для алгоритма агломеративной кластеризации равен          0.6278225908121073
        -для алгоритма DBSCAN равен                                0.41990181360123185
Коэффициент силуэта для з-го набора данных:
        -для алгоритма KMeans равен                                0.44037616227790793
        -для алгоритма агломеративной кластеризации равен          0.4020773152350946
        -для алгоритма DBSCAN равен                                0.3355335787077617
Коэффициент силуэта для и-го набора данных:
        -для алгоритма KMeans равен                                0.5130741640308434
        -для алгоритма агломеративной кластеризации равен          0.49163204131731847
        -для алгоритма DBSCAN равен                                0.06794420276574853
```

Рис. 7. Результаты вычисления коэффициента силуэта для различных выборок.

На основании данных значений можно сделать вывод, что лучше всего работают метод K-means и метод агломеративной кластеризации.

3.2. Коэффициент Девиса-Болдуина

Оценка данным методом определяется как средняя мера сходства каждого кластера с его наиболее похожим кластером, где сходство – это отношение расстояний внутри кластера к расстояниям между кластерами. Таким образом, кластеры, которые находятся дальше друг от друга и менее рассредоточены, дают лучший результат. Минимальный коэффициент равен нулю, более низкие значения указывают на лучшую кластеризацию.

Результаты вычисления данного коэффициента для различных выборок и различных методов приведен на рис. 8.

```
Коэффициент Девиса-Болдуина для а-го набора данных:
        -для алгоритма KMeans равен                              1.1845959006331057
        -для алгоритма агломеративной кластеризации равен        1.1964921207413393
        -для алгоритма DBSCAN равен                              989.6898942666963
Коэффициент Девиса-Болдуина для б-го набора данных:
        -для алгоритма KMeans равен                              0.8203830504487548
        -для алгоритма агломеративной кластеризации равен        0.8439722371739317
        -для алгоритма DBSCAN равен                              1.1620081344348578
Коэффициент Девиса-Болдуина для в-го набора данных:
        -для алгоритма KMeans равен                              0.24177016038343804
        -для алгоритма агломеративной кластеризации равен        0.24177016038343804
        -для алгоритма DBSCAN равен                              0.24177016038343804
Коэффициент Девиса-Болдуина для г-го набора данных:
        -для алгоритма KMeans равен                              0.845103853165592
        -для алгоритма агломеративной кластеризации равен        0.9255307915460396
        -для алгоритма DBSCAN равен                              0.6363212494510312
Коэффициент Девиса-Болдуина для д-го набора данных:
        -для алгоритма KMeans равен                              0.7186637808171245
        -для алгоритма агломеративной кластеризации равен        0.7047542752812315
        -для алгоритма DBSCAN равен                              6.07351976430381
Коэффициент Девиса-Болдуина для е-го набора данных:
        -для алгоритма KMeans равен                              0.5719752560698449
        -для алгоритма агломеративной кластеризации равен        0.5799308046351593
        -для алгоритма DBSCAN равен                              1.9023074871893215
Коэффициент Девиса-Болдуина для ж-го набора данных:
        -для алгоритма KMeans равен                              0.5274925707797168
        -для алгоритма агломеративной кластеризации равен        0.5287222619309871
        -для алгоритма DBSCAN равен                              2.4685036750595915
Коэффициент Девиса-Болдуина для з-го набора данных:
        -для алгоритма KMeans равен                              0.805165914951188
        -для алгоритма агломеративной кластеризации равен        0.8606949608256779
        -для алгоритма DBSCAN равен                              17.44379277432413
Коэффициент Девиса-Болдуина для и-го набора данных:
        -для алгоритма KMeans равен                              0.6295844747148145
        -для алгоритма агломеративной кластеризации равен        0.6640761602578619
        -для алгоритма DBSCAN равен                              2.400829755310502
```

Рис. 8. Результаты вычисления коэффициента Девиса-Болдуина для различных выборок.

На основании данных значений можно сделать вывод, что лучше всего работают метод K-means и метод агломеративной кластеризации.

3.3. Коэффициент Халинского-Харабаша

Данная оценка определяется как соотношение между дисперсией внутри кластера и дисперсией между кластерами. Более высокие значения указывают на лучшую кластеризацию. Результаты вычисления данного коэффициента для различных выборок и различных методов приведен на рис. 9.

```
Коэффициент Халинского-Харабаша для а-го набора данных:
        -для алгоритма KMeans равен                              862.2367834243024
        -для алгоритма агломеративной кластеризации равен        735.9377564110198
        -для алгоритма DBSCAN равен                              0.0013722701365677682
Коэффициент Халинского-Харабаша для б-го набора данных:
        -для алгоритма KMeans равен                              1935.081668916595
        -для алгоритма агломеративной кластеризации равен        1612.852206119927
        -для алгоритма DBSCAN равен                              979.4483501812448
Коэффициент Халинского-Харабаша для в-го набора данных:
        -для алгоритма KMeans равен                              37203.36303934008
        -для алгоритма агломеративной кластеризации равен        37203.36303934008
        -для алгоритма DBSCAN равен                              37203.36303934007
Коэффициент Халинского-Харабаша для г-го набора данных:
        -для алгоритма KMeans равен                              1189.7180060365722
        -для алгоритма агломеративной кластеризации равен        924.3445527163302
        -для алгоритма DBSCAN равен                              2.1780563302666933
Коэффициент Халинского-Харабаша для д-го набора данных:
        -для алгоритма KMeans равен                              3781.8526022384194
        -для алгоритма агломеративной кластеризации равен        3211.610566439511
        -для алгоритма DBSCAN равен                              1685.8845553447325
Коэффициент Халинского-Харабаша для е-го набора данных:
        -для алгоритма KMeans равен                              5331.48049113759
        -для алгоритма агломеративной кластеризации равен        4747.224309814294
        -для алгоритма DBSCAN равен                              1412.9017727366913
Коэффициент Халинского-Харабаша для ж-го набора данных:
        -для алгоритма KMeans равен                              4210.120675686922
        -для алгоритма агломеративной кластеризации равен        4174.52154608084
        -для алгоритма DBSCAN равен                              1143.1404451725582
Коэффициент Халинского-Харабаша для з-го набора данных:
        -для алгоритма KMeans равен                              1752.3764681846374
        -для алгоритма агломеративной кластеризации равен        1461.8009730419888
        -для алгоритма DBSCAN равен                              12.258199318789952
Коэффициент Халинского-Харабаша для и-го набора данных:
        -для алгоритма KMeans равен                              3519.7458237062115
        -для алгоритма агломеративной кластеризации равен        2946.0964165084183
        -для алгоритма DBSCAN равен                              365.0999739378853
```

Рис. 9. Результаты вычисления коэффициента Девиса-Болдуина для различных выборок.

На основании данных значений можно сделать вывод, что лучше всего работает метод K-means, а также метод агломеративной кластеризации.

3.4. Коэффициент однородности

Данный коэффициент показывает, насколько кластер состоит из объектов одного класса. Данная метрика равна отношению энтропии класса при условии кластера к энтропии класса. Максимальное значение, равное 1, достигается в том случае, если в кластере объекты одного класса. Результаты вычисления данного коэффициента для различных выборок и различных методов приведен на рис. 10.

```
Коэффициент однородности для а-го набора данных:
        -для алгоритма KMeans равен                              5.1297343607923226e-06
        -для алгоритма агломеративной кластеризации равен        0.0006683304245658186
        -для алгоритма DBSCAN равен                              1.0
Коэффициент однородности для б-го набора данных:
        -для алгоритма KMeans равен                              0.3869800350757354
        -для алгоритма агломеративной кластеризации равен        0.5464776145122131
        -для алгоритма DBSCAN равен                              1.0
Коэффициент однородности для в-го набора данных:
        -для алгоритма KMeans равен                              1.0
        -для алгоритма агломеративной кластеризации равен        1.0
        -для алгоритма DBSCAN равен                              1.0
Коэффициент однородности для г-го набора данных:
        -для алгоритма KMeans равен                              1.0
        -для алгоритма агломеративной кластеризации равен        1.0
        -для алгоритма DBSCAN равен                              1.0
Коэффициент однородности для д-го набора данных:
        -для алгоритма KMeans равен                              0.6181786703273342
        -для алгоритма агломеративной кластеризации равен        0.730886101129279
        -для алгоритма DBSCAN равен                              0.9878677117232207
Коэффициент однородности для е-го набора данных:
        -для алгоритма KMeans равен                              0.7926832731454881
        -для алгоритма агломеративной кластеризации равен        0.9383548456752933
        -для алгоритма DBSCAN равен                              0.5673547809960796
Коэффициент однородности для ж-го набора данных:
        -для алгоритма KMeans равен                              0.9179635522140218
        -для алгоритма агломеративной кластеризации равен        0.9194887311058555
        -для алгоритма DBSCAN равен                              0.9015347839702552
Коэффициент однородности для з-го набора данных:
        -для алгоритма KMeans равен                              1.0
        -для алгоритма агломеративной кластеризации равен        1.0
        -для алгоритма DBSCAN равен                              1.0
Коэффициент однородности для и-го набора данных:
        -для алгоритма KMeans равен                              0.8030811821628727
        -для алгоритма агломеративной кластеризации равен        0.9099781767383747
        -для алгоритма DBSCAN равен                              0.6013612448544557
```

Рис. 10. Результаты вычисления коэффициента однородности для различных выборок.

На основании данных значений можно сделать вывод, что лучше всего работает метод DBSCAN, однако в некоторых случаях данный метод показывает себя хуже других.

3.5. Коэффициент полноты

Данный коэффициент показывает, насколько объекты из класса принадлежат одному кластеру. Данная метрика равна отношению энтропии кластера при условии класса к энтропии кластера. Максимальное значение, равное 1, достигается в том случае, если все объекты класса принадлежат одному кластеру. Результаты вычисления данного коэффициента для различных выборок и различных методов приведен на рис. 11.

```
Коэффициент полноты для а-го набора данных:
        -для алгоритма KMeans равен                                5.129839617030711e-06
        -для алгоритма агломеративной кластеризации равен          0.0007049317174839502
        -для алгоритма DBSCAN равен                                1.0
Коэффициент полноты для б-го набора данных:
        -для алгоритма KMeans равен                                0.38704009418886587
        -для алгоритма агломеративной кластеризации равен          0.5730115092484505
        -для алгоритма DBSCAN равен                                1.0
Коэффициент полноты для в-го набора данных:
        -для алгоритма KMeans равен                                1.0
        -для алгоритма агломеративной кластеризации равен          1.0
        -для алгоритма DBSCAN равен                                1.0
Коэффициент полноты для г-го набора данных:
        -для алгоритма KMeans равен                                5.066300912368923e-16
        -для алгоритма агломеративной кластеризации равен          2.7391942153348627e-16
        -для алгоритма DBSCAN равен                                0.0
Коэффициент полноты для д-го набора данных:
        -для алгоритма KMeans равен                                0.6183158774089489
        -для алгоритма агломеративной кластеризации равен          0.7675815225989412
        -для алгоритма DBSCAN равен                                0.9529489855164172
Коэффициент полноты для е-го набора данных:
        -для алгоритма KMeans равен                                0.8010397321471883
        -для алгоритма агломеративной кластеризации равен          0.9385848414388561
        -для алгоритма DBSCAN равен                                0.8008682127482603
Коэффициент полноты для ж-го набора данных:
        -для алгоритма KMeans равен                                0.9180812900426306
        -для алгоритма агломеративной кластеризации равен          0.9194993435913005
        -для алгоритма DBSCAN равен                                0.721638232186152
Коэффициент полноты для з-го набора данных:
        -для алгоритма KMeans равен                                1.2013234535134215e-15
        -для алгоритма агломеративной кластеризации равен          0.0
        -для алгоритма DBSCAN равен                                6.111592024757734 5e-15
Коэффициент полноты для и-го набора данных:
        -для алгоритма KMeans равен                                0.7815132544891462
        -для алгоритма агломеративной кластеризации равен          0.9058386997451113
        -для алгоритма DBSCAN равен                                0.6926893915395881
```

Рис. 11. Результаты вычисления коэффициента полноты для различных выборок.

На основании данных значений нельзя дать однозначную оценку того, какой метод работает лучше. На различных видах данных получается разный результат.

3.6. V-мера

Данная мера равна среднему гармоническому коэффициента однородности и полноты, то есть равна удвоенному отношению произведения данных коэффициентов к их сумме. Чем ближе к 1, тем качественнее кластеризация. Результаты вычисления данного коэффициента для различных выборок и различных методов приведен на рис. 12.

```
V-мера для а-го набора данных:
        -для алгоритма KMeans равен                             5.1297869883715885e-06
        -для алгоритма агломеративной кластеризации равен       0.0006861433073990412
        -для алгоритма DBSCAN равен                             1.0
V-мера для б-го набора данных:
        -для алгоритма KMeans равен                             0.38701006230219515
        -для алгоритма агломеративной кластеризации равен       0.5594301114962545
        -для алгоритма DBSCAN равен                             1.0
V-мера для в-го набора данных:
        -для алгоритма KMeans равен                             1.0
        -для алгоритма агломеративной кластеризации равен       1.0
        -для алгоритма DBSCAN равен                             1.0
V-мера для г-го набора данных:
        -для алгоритма KMeans равен                             1.0132601824737842e-15
        -для алгоритма агломеративной кластеризации равен       5.478388430669724e-16
        -для алгоритма DBSCAN равен                             0.0
V-мера для д-го набора данных:
        -для алгоритма KMeans равен                             0.6182472662555794
        -для алгоритма агломеративной кластеризации равен       0.748784501536842
        -для алгоритма DBSCAN равен                             0.970094223787309
V-мера для е-го набора данных:
        -для алгоритма KMeans равен                             0.796839594696583
        -для алгоритма агломеративной кластеризации равен       0.9384698294655055
        -для алгоритма DBSCAN равен                             0.6641847294307548
V-мера для ж-го набора данных:
        -для алгоритма KMeans равен                             0.9180224173533105
        -для алгоритма агломеративной кластеризации равен       0.9194940373179565
        -для алгоритма DBSCAN равен                             0.8016175247900129
V-мера для з-го набора данных:
        -для алгоритма KMeans равен                             2.4026469070268403e-15
        -для алгоритма агломеративной кластеризации равен       0.0
        -для алгоритма DBSCAN равен                             1.2223184049515393e-14
V-мера для и-го набора данных:
        -для алгоритма KMeans равен                             0.7921504377071396
        -для алгоритма агломеративной кластеризации равен       0.9079037199053294
        -для алгоритма DBSCAN равен                             0.6438025577646391
```

Рис. 12. Результаты вычисления V-меры для различных выборок.

На основании данных значений нельзя дать однозначную оценку того, какой метод работает лучше. На различных видах данных получается разный результат.

Заключение

Для достижения данной цели, в процессе прохождения учебной практики (учебно-лабораторного практикума) изучил существующие методы кластеризации, а также сравнил их. На основе полученных знаний произвел кластеризацию выбранного датасета.

Также были изучены требования к написанию отчета по практике. В результате прохождения практики был составлен отчет по практике, соответствующий предъявленным требованиям.

В ходе прохождения практики все задачи были выполнены, а цель достигнута.