|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| logoМИНИСТЕРСТВО 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 | | | | | | | | | |  |  | Руководитель практики | | | | | | |  | |
|  | | | | | | | | | |  |  | Старший преподаватель кафедры информационной безопасности ШЕН | | | | | | | | |
|  | | | |  | С.С. Зотов | | | | |  |  |  | | | | |  | С.С. Зотов | | |
| (подпись) | | | |  | (И.О. Фамилия) | | | | |  |  | (подпись) | | | | |  | (И.О. Фамилия) | | |
| « | 26 | » | июня | | | | | | 2021 г. |  |  |  | | | | | | | | |
|  | | | | | | | | | | | | | | | | | | | | |
| Регистрационный № | | | | | | | |  | |  |  | Практика пройдена в срок | | | | | | | | |
| « | 26 | » | июня | | | | | | 2021 г. |  |  | с | « | 22 | » | февраля | | | | 2021 г. |
|  | | | | | | | | | |  |  | по | « | 26 | » | июня | | | | 2021 г. |
|  | | | | | |  |  | | |  |  | на предприятии | | | | | | | | |
| (подпись) | | | | | |  | (И.О. Фамилия) | | |  |  |
|  | | | | | | | | | |  |  | 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**

[Характеристика 2](#_Toc75193983)

[ДНЕВНИК СТУДЕНТА 3](#_Toc75193984)

[Задание на практику 5](#_Toc75193985)

[Введение 6](#_Toc75193986)

[1. Кластеризация 7](#_Toc75193987)

[1.1 K-means кластеризация 9](#_Toc75193988)

[1.2. Агломеративная кластеризация 9](#_Toc75193989)

[1.3. Метод кластеризации DBSCAN 10](#_Toc75193990)

[2. Демонстрация работы алгоритмов 13](#_Toc75193991)

[2.1. Алгоритм агломеративной кластеризации с помощью дендрограммы 13](#_Toc75193992)

[2.1. Метод кластеризации DBSCAN 13](#_Toc75193993)

[3. Сравнение методов кластеризации с помощью метрик качества. 16](#_Toc75193994)

[3.1. Коэффициент силуэта 18](#_Toc75193995)

[3.2. Коэффициент Девиса-Болдуина 19](#_Toc75193996)

[3.3. Коэффициент Халинского-Харабаша 20](#_Toc75193997)

[3.4. Коэффициент однородности 21](#_Toc75193998)

[3.5. Коэффициент полноты 22](#_Toc75193999)

[3.6. V-мера 23](#_Toc75194000)

[Заключение 25](#_Toc75194001)

## 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: , где – число кластеров, − полученные кластеры,, а – центры масс всех векторов из кластера . Данный алгоритм разбивает множество элементов векторного пространства на заранее известное число кластеров .

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

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

- Точка достижима из , если имеется путь с и , где каждая точка достижима прямо из (все точки на пути должны быть основными, за исключением ).

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

Теперь, если 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 и связаны по плотности, если имеется точка , такая что и , и достижимы из . 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.

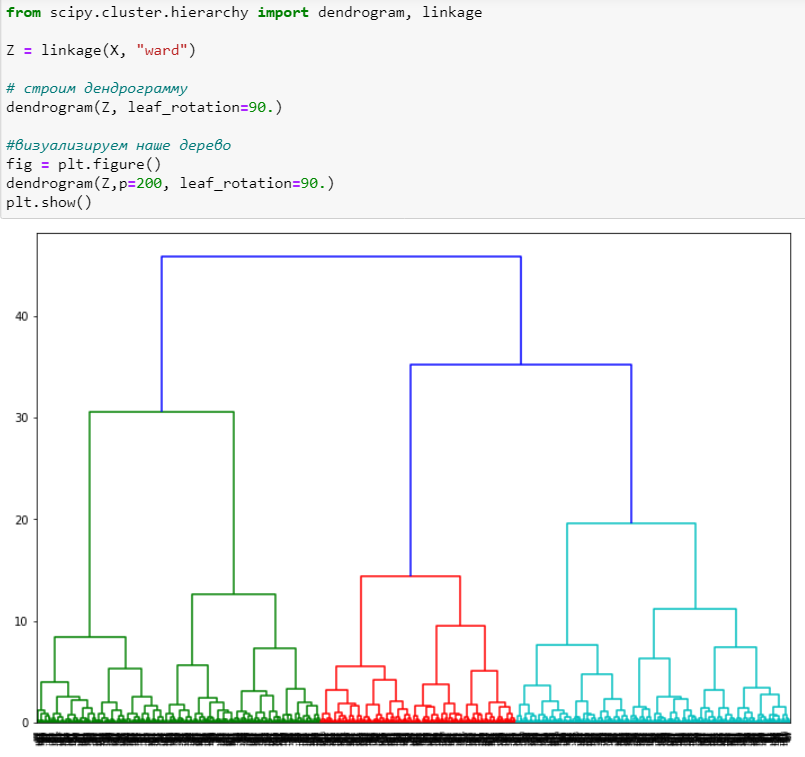


Рис. 1. Дендрограмма.

### 2.1. Метод кластеризации DBSCAN

Возьмем для кластеризации методом DBSCAN параметры, равные и и проведем кластеризацию нашего датасета. Код данного участка указан на рис. 2.

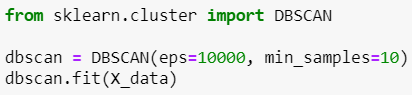


Рис. 2. Кластеризация датасета методом DBSCAN.

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

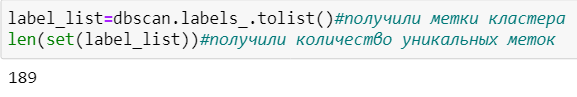


Рис. 3. Количество кластеров методом DBSCAN.

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

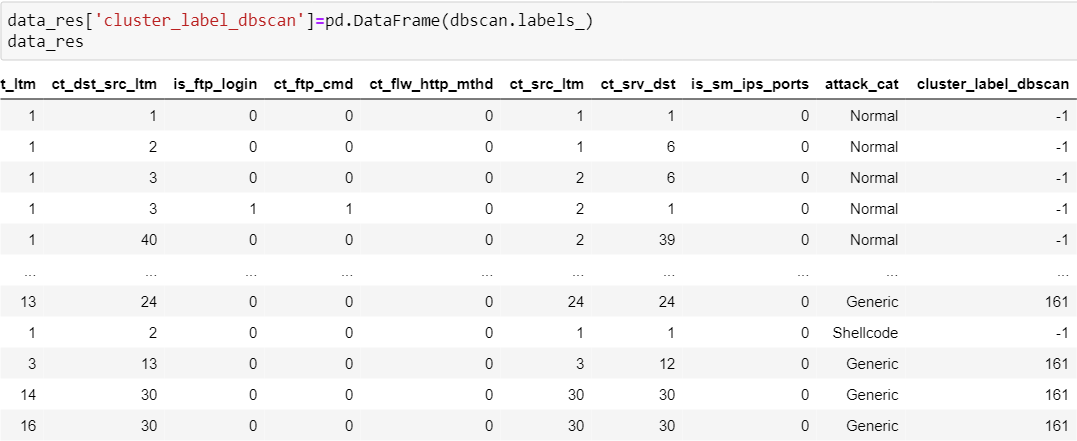


Рис. 4. Таблица с кластерами.

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



Рис. 5. Изменение параметров DBSCAN.

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

## 3. Сравнение методов кластеризации с помощью метрик качества.

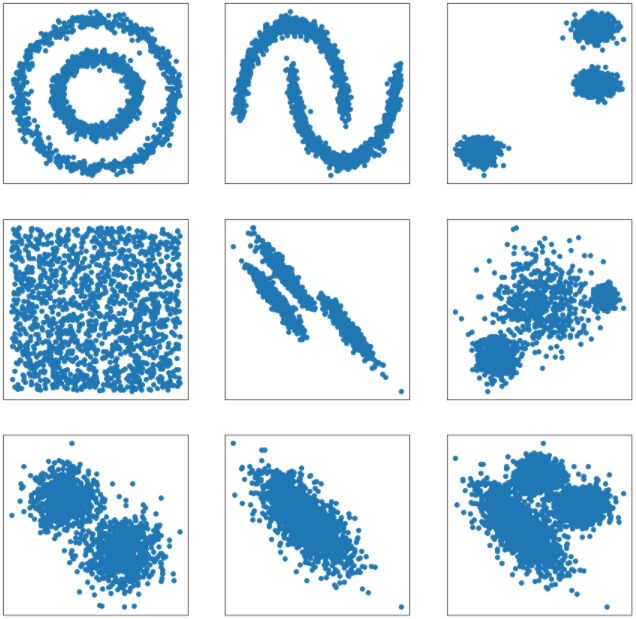
Метрики оценки качества кластеризации — инструментарий для количественной оценки результатов кластеризации.

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.



и

з

ж

е

д

г

в

б

а

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.

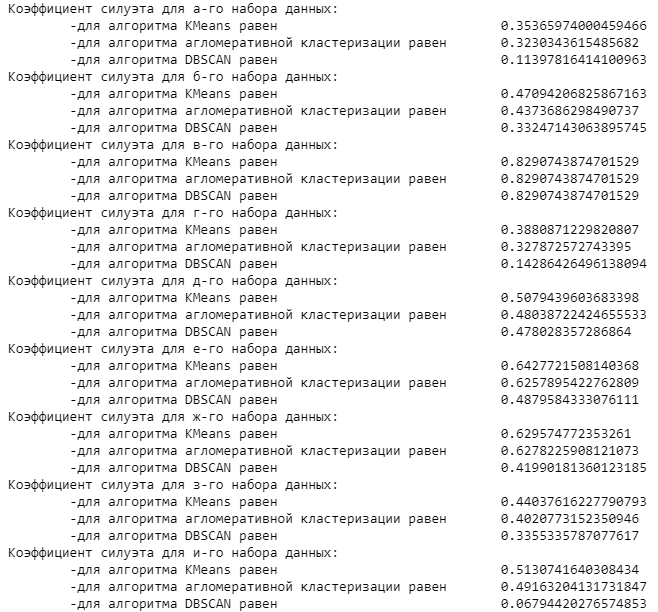


Рис. 7. Результаты вычисления коэффициента силуэта для различных выборок.

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.

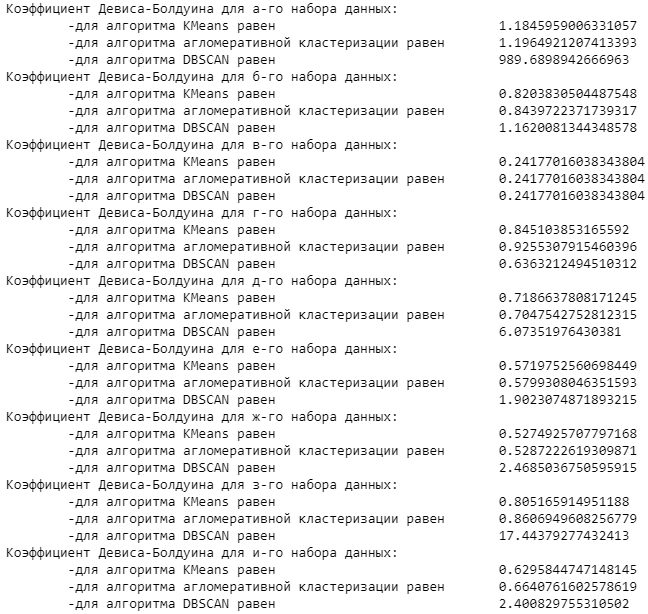


Рис. 8. Результаты вычисления коэффициента Девиса-Болдуина для различных выборок.

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.

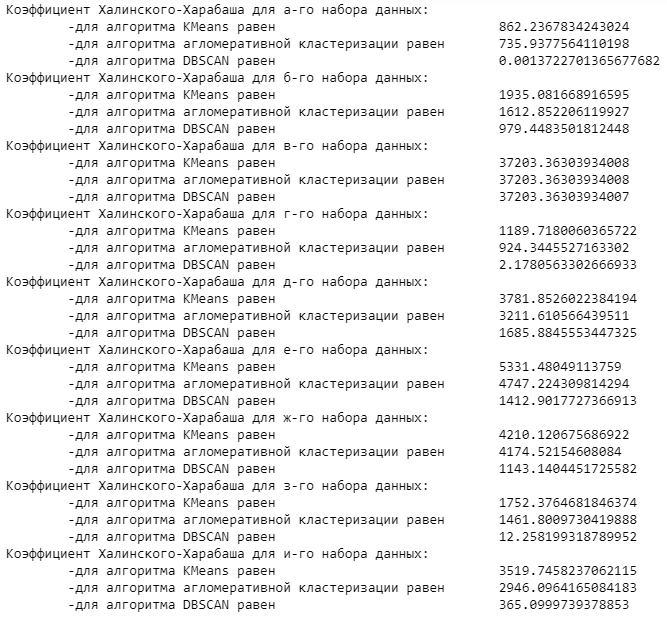


Рис. 9. Результаты вычисления коэффициента Девиса-Болдуина для различных выборок.

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.

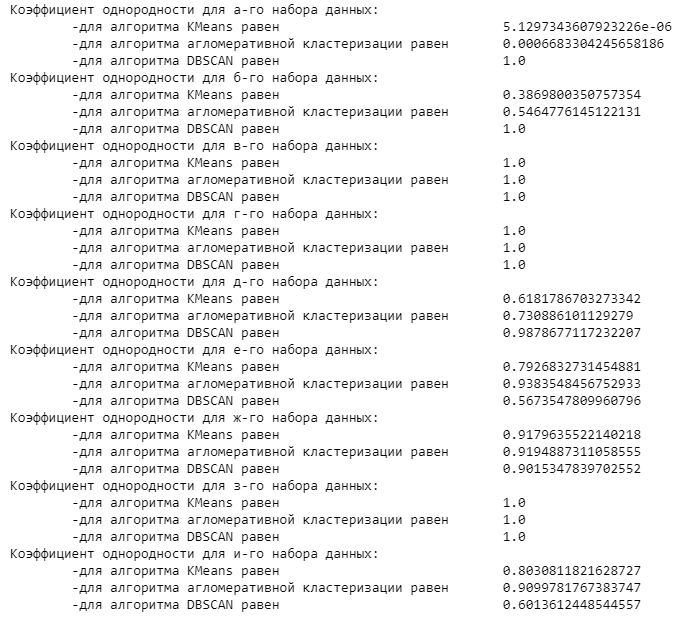


Рис. 10. Результаты вычисления коэффициента однородности для различных выборок.

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.

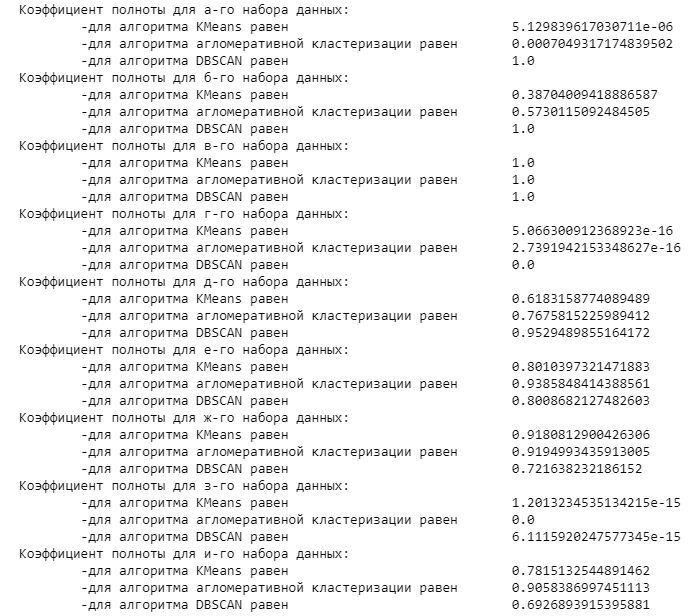


Рис. 11. Результаты вычисления коэффициента полноты для различных выборок.

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.

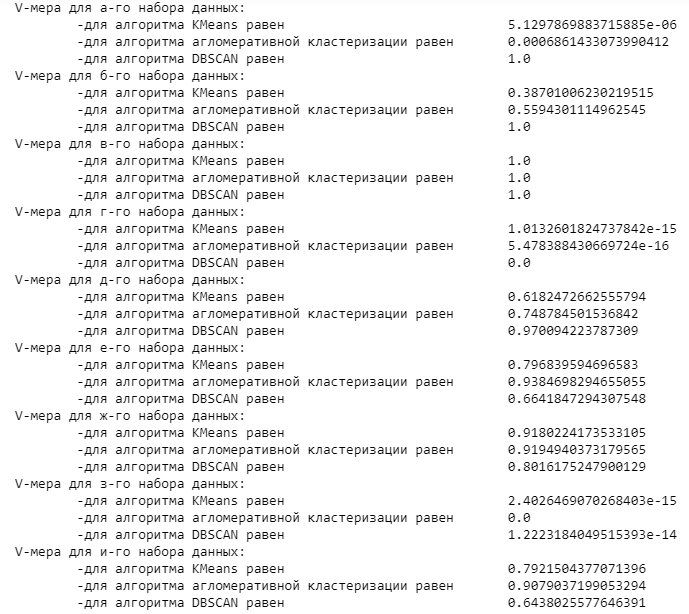


Рис. 12. Результаты вычисления V-меры для различных выборок.

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.