Белорусский государственный технологический университет

Кафедра «Информационных систем и технологий»

Лабораторная работа №6

**Кластерный анализ**

Выполнил студент

3 курса 2 группы

Процукович К.М.

Проверил

Колесников В. Л.

Минск 2018

# **Цель работы**

Цель лабораторной работы – ознакомиться с методами оптимизации с помощью кластерным анализом, и на их основе освоить возможные способы оптимизации виртуального комплекса.

Для этого нам необходимо сперва классифицировать данные по кластерам (данные, схожие по свойствам). И после анализа мы сможем отбросить лишние данные (те данные, которые сложно отнести к какому-либо кластеру), что позволит нам получить более достоверные данные.

# **Краткое описание объекта исследования**

## **2.1 Введение**

**Кластерный анализ** предназначен для разбиения совокупности объектов на однородные группы, именуемые **кластерами.** Например, в результате кластерного анализа можно получить модель, представленную на рисунке 2.1 ниже:

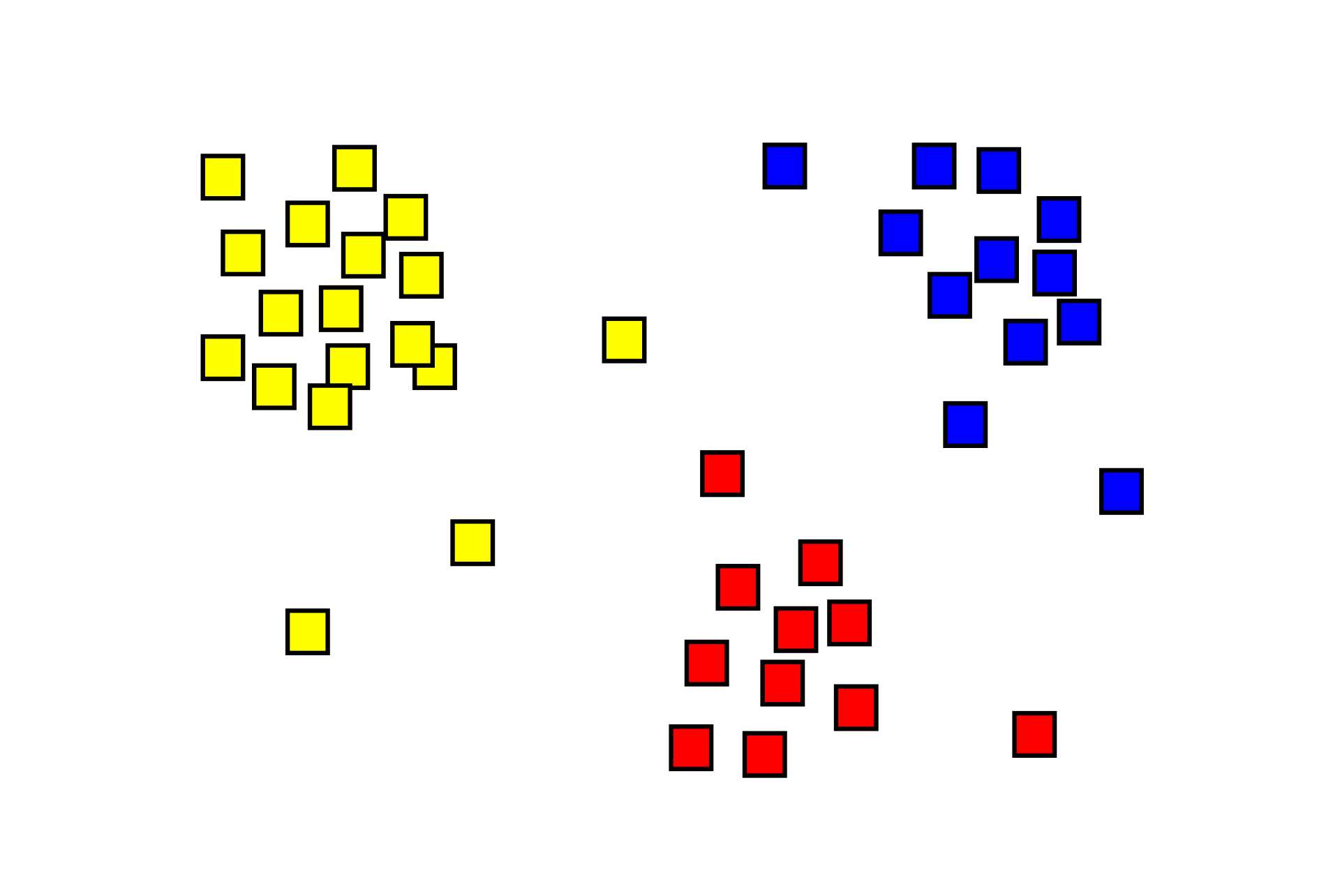


Рис. 2.1 – результат кластеризации

В действительности же термин *кластерный**анализ* включает в себя набор различных [алгоритмов](http://statsoft.ru/home/textbook/glossary/gloss_a.html#Algorithm) классификации из всевозможных областей: археологии, медицине, психологии, химии, биологии, государственном управлении, филологии, антропологии, маркетинге, социологии, геологии и других дисциплинах.

Это позволяет рассудить, что кластерный анализ является универсальным методом для решения проблем во множестве сфер человеческой жизни, но из-за своей многогранности он также приобретает и недостатки, заключающиеся в появлении большого количества несовместимых терминов, методов и подходов, затрудняющих однозначное использование и непротиворечивую интерпретацию кластерного анализа.

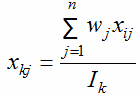
Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры. Например, биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В соответствии с современной системой, принятой в биологии, человек принадлежит к приматам, млекопитающим, позвоночным и животным. Заметьте, что в этой классификации, чем выше уровень, тем меньше сходства между членами в соответствующем классе. Человек имеет больше сходства с другими приматами (с обезьянами), чем с "отдаленными" членами семейства млекопитающих (например, собаками) и т.д.

Однако следует упомянуть о проверке статистической значимости. Фактически, кластерный анализ является не столько обычным статистическим методом, сколько "набором" различных алгоритмов "распределения объектов по кластерам". Существует точка зрения, что в отличие от многих других статистических процедур, методы кластерного анализа используются в большинстве случаев тогда, когда вы не имеете каких-либо априорных гипотез относительно классов, но все еще находитесь в описательной стадии исследования. Следует понимать, что кластерный анализ определяет "наиболее возможно значимое решение". Поэтому проверка статистической значимости в действительности здесь неприменима.

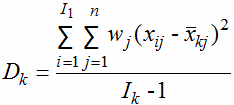
**2.2 Характеристики кластеров**

Основными характеристиками кластеров являются:

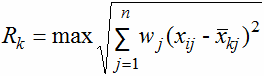
**Центр кластера** – среднее геометрическое место точек в пространстве переменных:



**Дисперсия кластера** – мера рассеяния точек в пространстве относительно центра:



**Радиус кластера** – максимальное расстояние точек от центра:



Наглядно структуру кластера можно рассмотреть ниже (рис. 2.2).

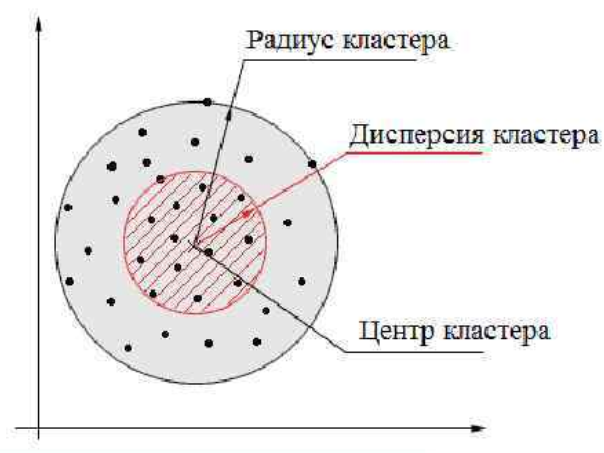


Рис 2.2 ­­– структура кластера

**2.3 Виды кластеров**

|  |  |
| --- | --- |
| Внутрикластерное расстояние, как правило, меньше межкластерного |  |
| Ленточные |  |
| Кластеры с центром |  |
| Кластеры соединенные перемычками |  |
| Кластеры, накладываемые на разряженный фон из редко расположенных объектов |  |
| Перекрывающиеся кластеры |  |
| Кластера, образованные не по сходству, а по иному типу |  |
| Кластеры вовсе могут отсутствовать |  |

## **2.4 Классификация задач кластеризации**

Первый шаг кластеризации, это подготовка данных к кластеризации. К примеру, можно составить таблицы, где строки — это объекты, а столбцы —их признаки.

Вторым шагом будет обработка входных данных. Данные могут анализироваться как путем сравнения объектов исходя из признаков, так и наоборот, путем сравнения признаков исходя их объектов, следует дополнить, что существует и гибридный тип, где совмещены оба выше представленных, но данная методология толком еще не разработана и поэтому почти не применяется.

Также, при кластеризации мы должны понимать наше намерение. Стало быть, мы можем выделить различные цели, такие как:

* **Выявить кластерную структуру**, что позволит упросить дальнейшую обработку данных и принятие решений.
* **Сжать данные**. Разбить данные, выбрав по наиболее типичному представителю из каждого кластера.
* **Выделить нетипичные объекты**, путем нахождения объектов, которые не принадлежат ни к одному из кластеров.

## **2.5 Возможные алгоритмы кластерного анализа**

Третий наш шаг, это выбор алгоритма кластерного анализа.

Существует около 100 разных алгоритмов кластеризации, и общепринятой классификации методов кластерного анализа не существует, однако можно выделить ряд подходов:

* **Вероятностный**
  + K-средних
  + К-медиан
* На основе **искусственного интелекта**
  + Нейронная сеть Кохонена
  + Генетический алгоритм
* **Логический**
* **Графовый**
* **Иерархический**
  + Объединение (древовидная кластеризация)
  + Двувходовое объединение
* Другие

Наиболее часто используемые это:

* кластеризация методом k-средних.
* объединение (древовидная кластеризация);

**2.5.1 Объединение (древовидная кластеризация)**

В данном случае рассматривается задача объединения данных в кластеры или группы больших размеров. Между объектами существует определенные сходства или различия, которые можно выразить как «расстояние между объектами» (рис. 2.3).

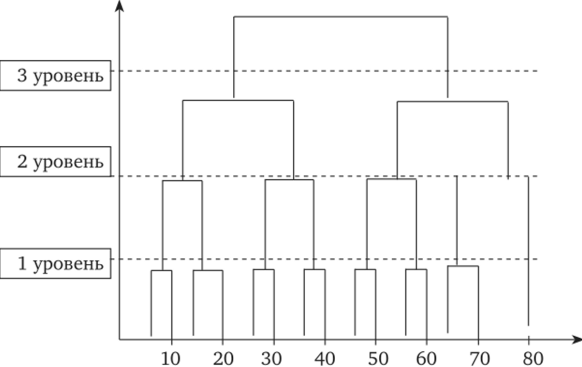


Рис. 2.3 – древовидная кластеризация

Построение дерева начинается с формирования оси абсцисс. На нее наносятся в определенном порядке признаки, характеризующие объект наблюдения, например, величину оплаты покупки. Каждая из покупок уникальна, но их можно объединить в определенные группы, например, все, что до 10 рублей, что до 20 рублей и т.д. Далее повышается порог, относящийся к решению об объединении двух или более объектов в один кластер. В результате связывается воедино все больше и больше объектов и агрегируется большее число кластеров, состоящих из различающихся объектов. На последнем шаге все объекты могут быть объединены вместе. По оси ординат откладываются показатели, характеризующие уровни объединения признаков. В нашем примере это единицы, десятки и сотня.

Исследователь должен определить, на каком уровне детализации следует остановиться, так как в противном случае можно принять за кластер отдельный объект. Высокая степень детализации означает, что исследователь сознательно идет на глубокую детализацию (сегментацию) рынка.

**2.5.2 Метод К-средних**

Метод k-средних – это метод кластерного анализа, цель которого является разделение m наблюдений а k кластеров, при этом каждое наблюдение относится к тому кластеру, к центру которого оно ближе всего (рис. 2.4).

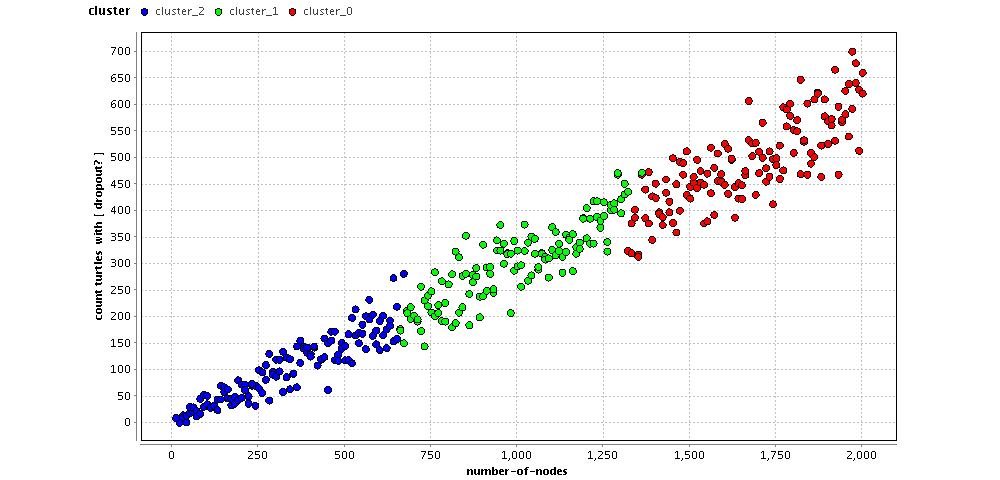
****

Рис. 2.4 – кластеризация методом К-средних

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.

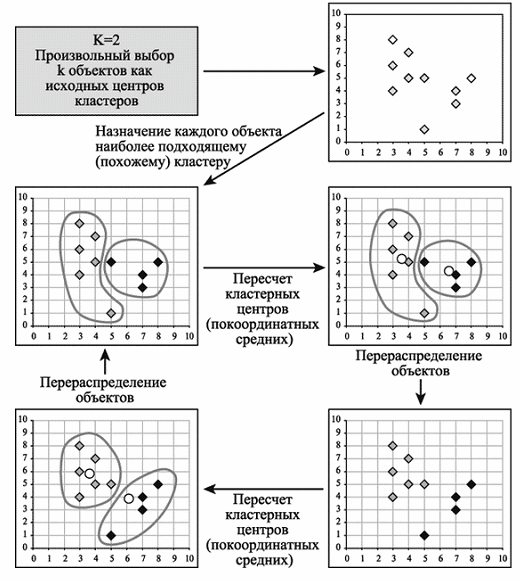


Рис 2.5 – пример работы алгоритма к-средних (к = 2)

Завершающий шаг после получения результатов кластерного анализа, это проверка правильности кластеризации. Для этого следует рассчитать среднее значение для каждого кластера, если кластеризация успешна, то, как минимум, по большей части полученные средние должны отличаться.

# **Организация сбора информации**

Сбор информации будет происходить из результатов кластерного анализа на основе результатов алгоритма кластеризации приложения JMP SAS, которое позволит визуализировать многомерное множество объектов и кластеры, в которые они были распределены в результате действия алгоритма кластерного анализа. Данные будут исследоваться и анализироваться и на их основе будет сделан вывод. Обрабатывая входной параметр, в виде таблицы зависимостей различных параметров производства виртуального предприятия, приложение вернет несколько кластеров, которые будут анализироваться.

У нас уже имеется готовая база данных для анализа (нормализация уже проведена), подготовленная еще на прошлых лабораторных, её мы и будем использовать (рис. 3.1).



Рис. 3.1 – часть базы данных JMP

1. **Эксперимент**

Сперва нам необходимо выбрать входные параметры (рис. 4.1). Мной были выбраны все свойства, кроме тех, которые при моделировании процесса работы виртуального комплекса не были модифицированы.

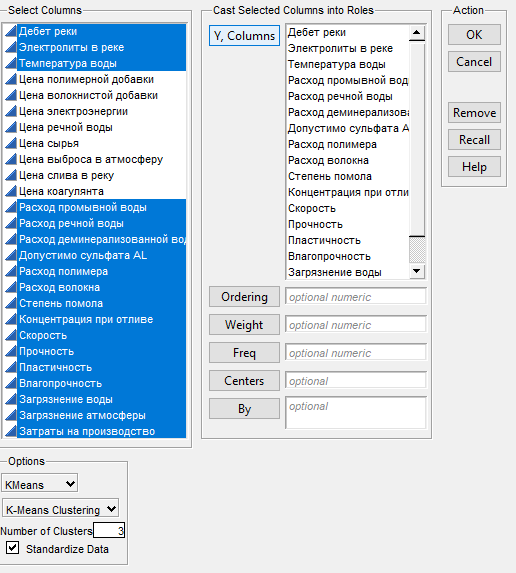


Рис. 4.1 – выбранные параметры для анализа

В результате анализа мы получили данные, которые можем представить, как в 2д, так и 3д формате.

Здесь (рис. 4.2) мы можем видеть, что кластеры, которые содержат больше всего количества параметров несут в себе больше информации, а также скопление таких кластеров образуют пересечения их друг с другом. Это может свидетельствовать о том, что один кластер может оказывать влияние на информацию содержащуюся в другом. Анализ этих кластеров в комплексе даст наиболее развернутый и информативный результат.

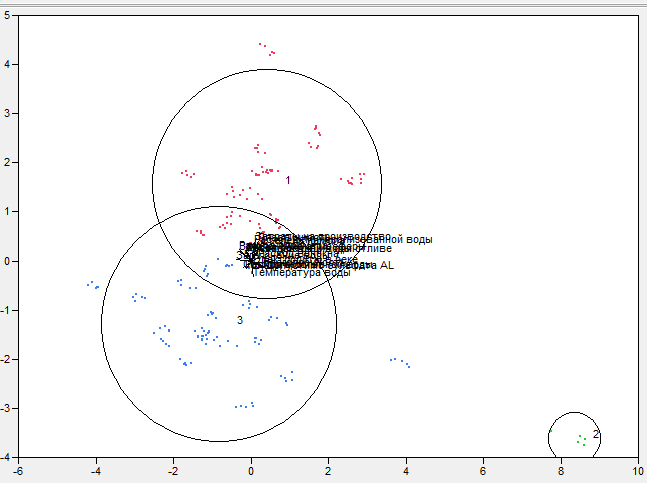


Рис. 4.2 – 2д визуализация разделения данных на 3 кластера

Проведя аналогичные операции, только указав перед началом анализа 4 и 5 кластеров, мы получили и их визуализации (рис. 4.3).

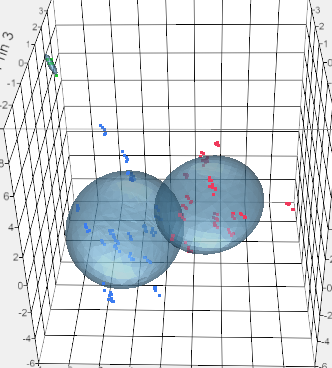
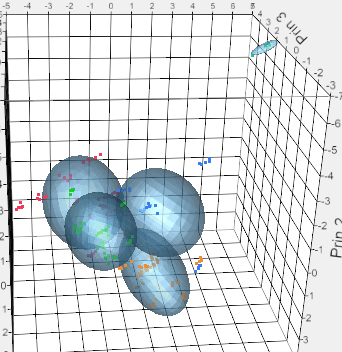
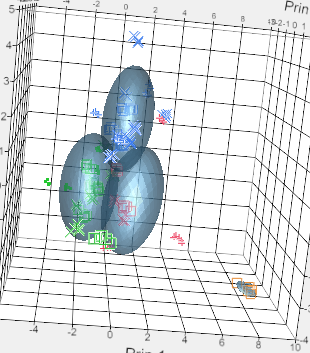
 

Рис. 4.3 – 3д визуализация разделения данных на 3, 4, 5 кластеров соответственно

Однако присутствуют такие объекты, которые не находятся ни в одном из них. Данные объекты – уникальны. Они в БД встречаются очень редко и их значения значительно отличаются от среднего значения по кластерам, поэтому их нельзя причислить ни к одному из кластеров.

1. **Выводы**

Что же можно сказать после проведения анализа?

Проведя анализ три раза, разбив данные на 3, 4, 5 кластеров соответственно, мы имеем неизменное скопление кластеров, которое даже при увеличении их количество не исчезает и один кластер находящийся на удаление от основной массы.

Наше разбиение на кластеры дало нам возможность при поиске информации не пересматривать всю таблицу, а достаточно указать параметр и определить кластер.

Так же встречаются нетипичные объекты, которые не входят ни в один кластер. Однако их количество мало, так что этими значениями можно смело пренебречь.

Вдобавок было обнаружено, что при увеличении количества кластеров интервалы значений и их вероятность в каждом из кластеров уменьшается, поскольку часть значений переносится в созданные дополнительные кластеры.