Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное

учреждение высшего образования

«Самарский государственный аэрокосмический университет

имени академика С.П. Королева

(национальный исследовательский университет)»

Факультет информатики  
Кафедра технической кибернетики

**ОТЧЁТ**  
по НИР бакалавра

на тему

Кластеризация больших объёмов данных.

Студент гр. 6409 Костанян О.Г.

Руководитель работы Куприянов А.В.

САМАРА 2015

**Реферат**

Отчет по НИР бакалавра: 24 стр., 4 источника, 4рисунка, 2 таблицы.

КЛАСТЕРИЗАЦИЯ, АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ, КАЧЕСТВО КЛАСТЕРИЗАЦИИ, ТАКСОНОМИЯ, K-MEANS, BIG DATA.

Целью данной работы является изучение и проведение экспериментов кластеризации больших объёмов данных.

Были изучены основные алгоритмы кластеризации больших объемов данных. Изучены основные функции кластерного анализа математического пакета Matlab R2014a, проведена оценка качества работы алгоритмов для заданной выборки.

Оглавление

[Введение 4](#_Toc439895071)

[1 Основные понятия 5](#_Toc439895072)

[**1.1** **Метрика** 5](#_Toc439895073)

[**1.2** **Дендрограмма** 9](#_Toc439895074)

[**1.3** **Свойства кластеризации** 10](#_Toc439895075)

[2 Тип кластерной структуры 12](#_Toc439895076)

[3 Алгоритмы кластеризации 13](#_Toc439895077)

[**3.1** **Графовые методы кластеризации** 13](#_Toc439895078)

[**3.2** **Иерархическая кластеризация (таксономия)** 15](#_Toc439895079)

[**3.3** **Статистические методы кластеризации** 16](#_Toc439895080)

[4 Функционалы качества кластеризации 18](#_Toc439895081)

[5 Тестирование алгоритмов 19](#_Toc439895082)

[6 Заключение 23](#_Toc439895083)

[Список использованных источников 24](#_Toc439895084)

Введение

На сегодняшний день в мире накоплено огромное количество разной неструктурированной информации, из которой не сразу можно получить сколь-нибудь значимые знания. С целью упрощения восприятия огромных объемов данных разработаны много алгоритмов кластеризации.

Кластеризация позволяет упростить дальнейшую обработку данных, сократить объем хранимых данных, выделять нетипичные объекты, строить иерархию множества объектов.

В данной работе будут приведены основные понятия, алгоритмы, примеры работы программ. Будет проведена оценка качества работы алгоритмов.

1. Основные понятия

**Кластерный анализ** — задача разбиения заданной [выборки](http://www.machinelearning.ru/wiki/index.php?title=%D0%92%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0) объектов (ситуаций) на непересекающиеся подмножества, называемые [кластерами](http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80&action=edit), так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

**Кластер** — объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица, обладающая определёнными свойствами.

В данной работе рассмотрим три метода кластеризации больших объемов данных:

* Графовые методы кластеризации
* Иерархическая кластеризация (таксономия)
* Статистические методы кластеризации

Решение задачи кластеризации данных, вообще говоря, является неоднозначной, так как не существует точной постановки задачи кластеризации, существует много критериев качества кластеризации, существует много эвристических методов кластеризации, число кластеров, как правило, неизвестно заранее, результат кластеризации существенно зависит от метрики, которую эксперт задает субъективно.

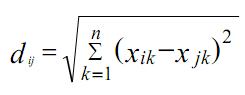
* 1. **Метрика**

Пусть объекты i и j принадлежат множеству M и каждый объект описывается k признаками, тогда будем говорить, что на множестве M задана метрика, если для любой пары объектов, принадлежащих множеству M, определено неотрицательное число dij , удовлетворяющее следующим условиям (аксиомам метрики):

1. Аксиома тождества: dij = 0 ⇔ i≡ j.
2. Аксиома симметричности: dij = dji ∀ i, j.
3. Неравенство треугольника: ∀ i, j, z∈M, выполняется неравенство diz ≤ dij + dzj .

Пространство, на котором введена метрика, называется метрическим. Наиболее используемыми являются следующие метрики:

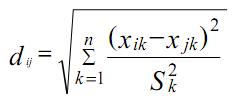
1. Метрика Евклида:



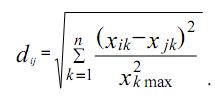
2. Метрика нормированного Евклида.

Нормализованные евклидовы расстояния более подходят для переменных, измеряемых в различных единицах или значительно различающихся по величине.

Если дисперсии по характеристикам отличаются друг от друга, то:

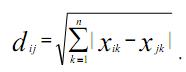


Если масштаб данных различен, например, одна переменная измерена в стэнах, а другая в баллах, то для обеспечения одинакового влияния всех характеристик на близость объектов используется следующая формула подсчета расстояния:



3. Метрика city-block:

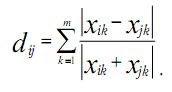
Метрика city-block (манхэттенская метрика, получившая свое название в честь района Манхэттен, который образуют улицы, расположенные в виде пересечения параллельных прямых под прямым углом; как правило, применяется для номинальных или качественных переменных):



4. Метрика на основе корреляции: dij=1- |rij|.

Расстояния, вычисляемые на основе коэффициента корреляции, отражают согласованность колебаний оценок, в отличие от метрики Евклида, которая определяет схожесть в среднем.

5. Метрика Брея-Картиса, которая также используется для номинативных и ранговых шкал, обычно данные предварительно стандартизируются:

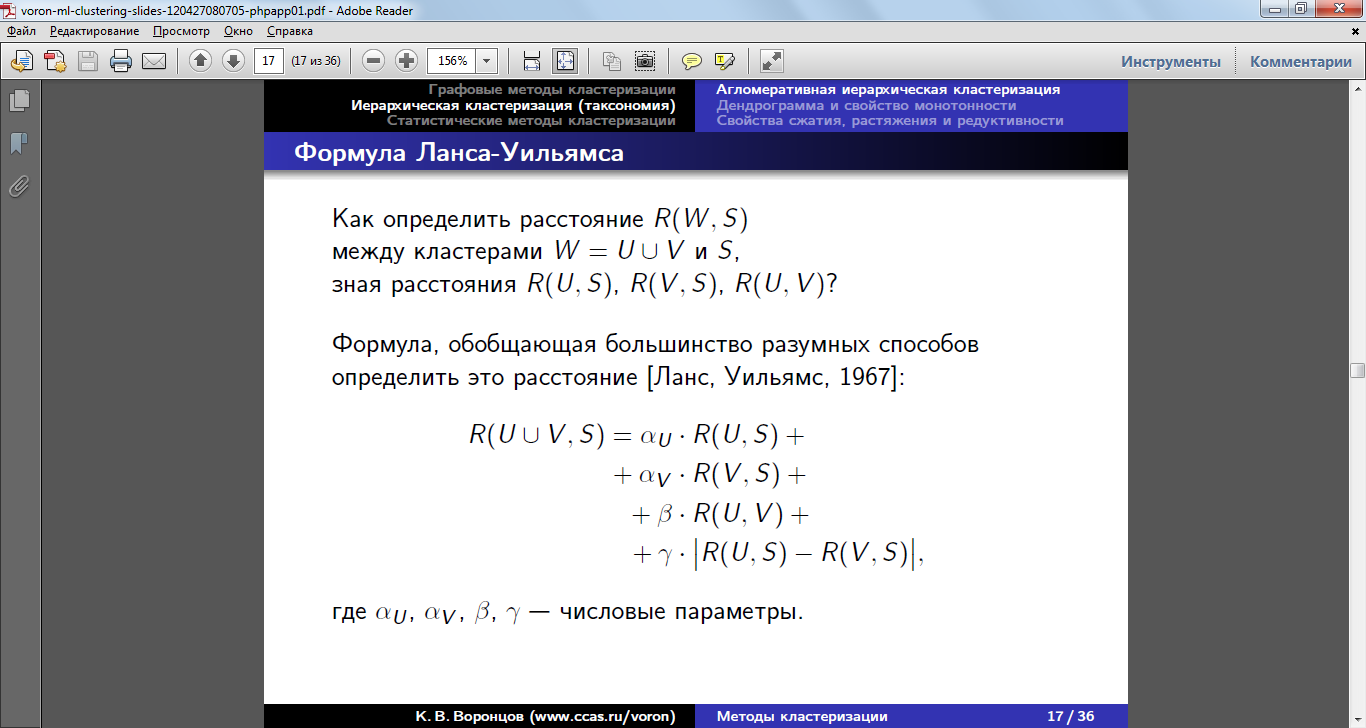


Выбор метрики определяется задачей исследования и типом данных. Помимо приведенных выше методов, разработаны метрики для ранговых и дихотомических переменных и т.д.

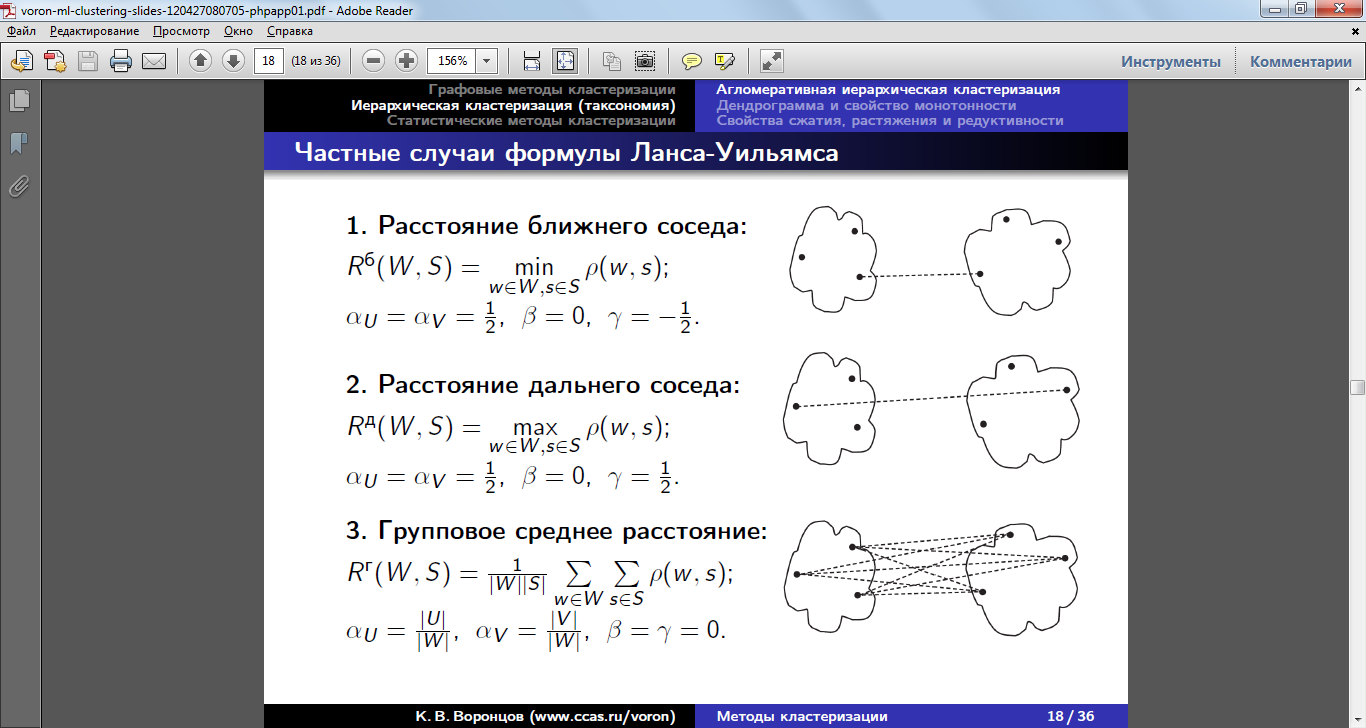
Для определения «естественного» числа кластеров применяется критерий разбиения на классы в виде отношения средних внутрикластерных расстояний к межкластерным расстояниям.

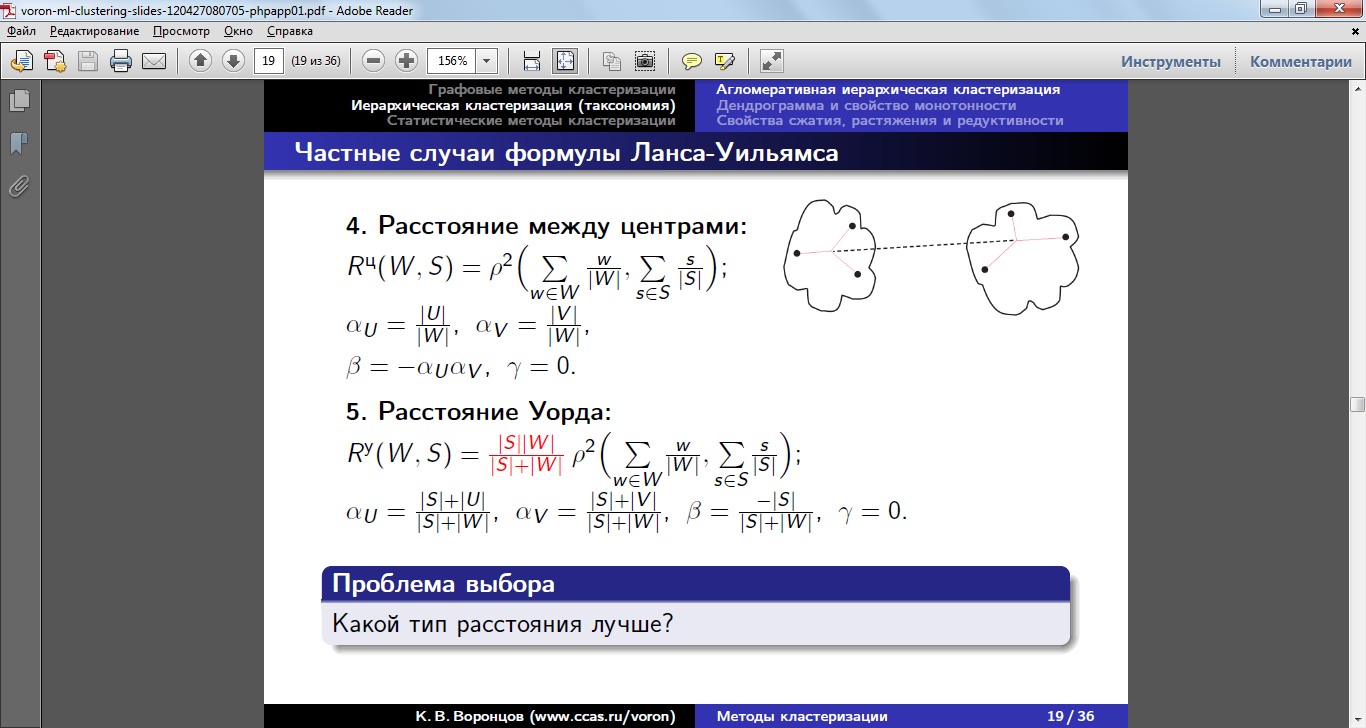
При решении задачи кластеризации встает вопрос: как определить расстояние R(W, S) между кластерами W = U ∪ V и S, зная расстояния R(U, S), R(V, S), R(U, V).

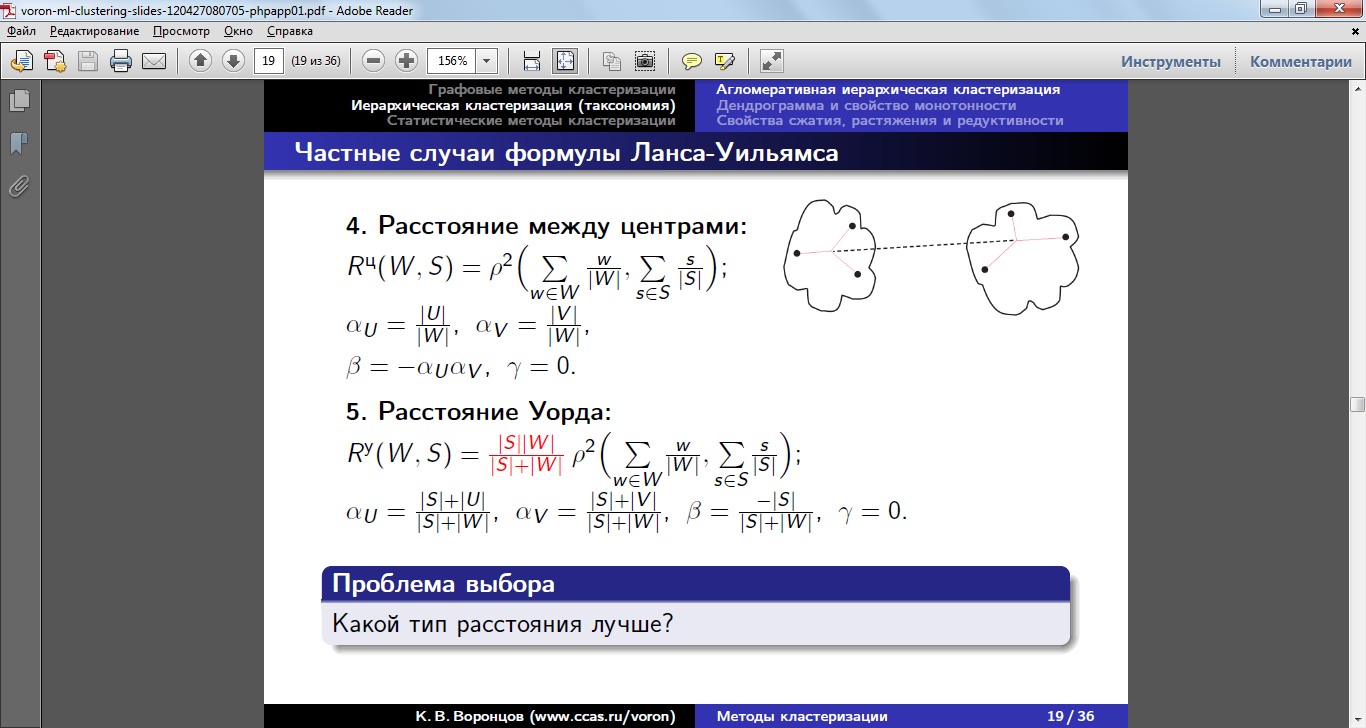
Формула, обобщающая большинство разумных способов определить это расстояние [Ланс, Уильямс, 1967]:



Наиболее часто используемыми стратегиями выбора расстояния между кластерами являются следующие:



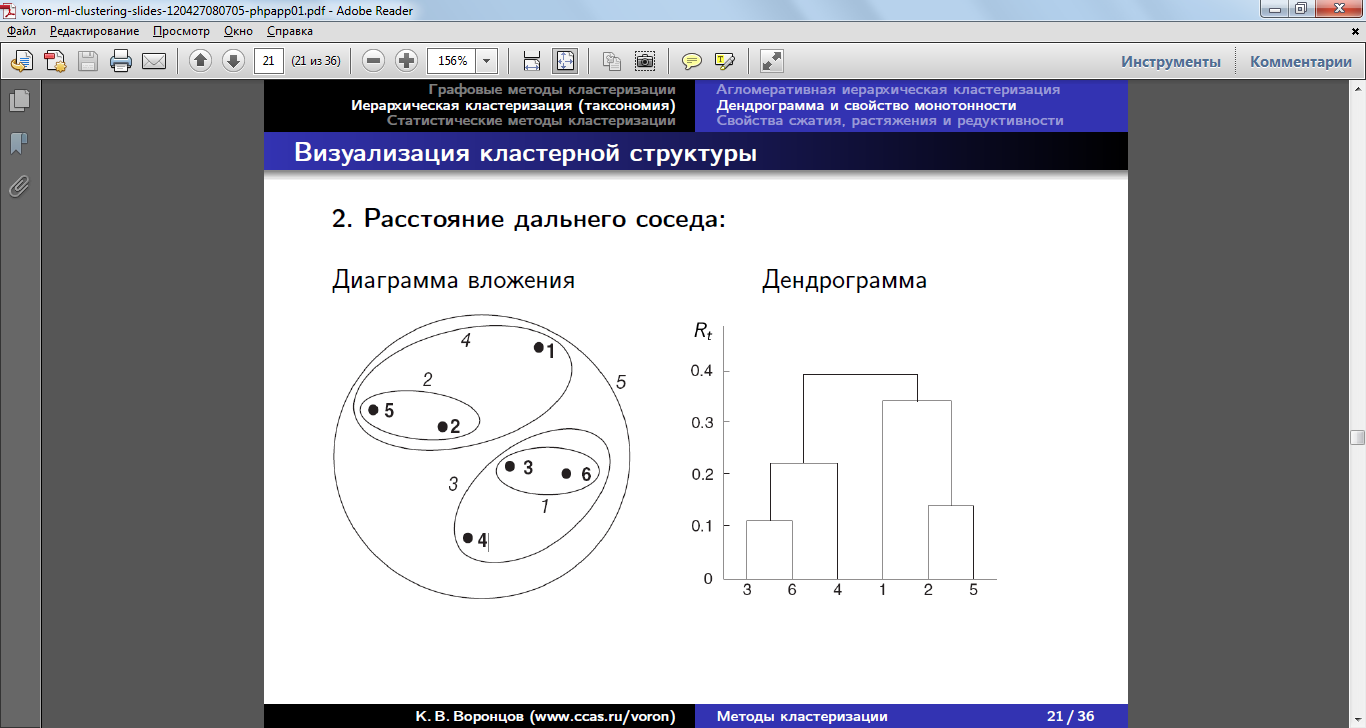
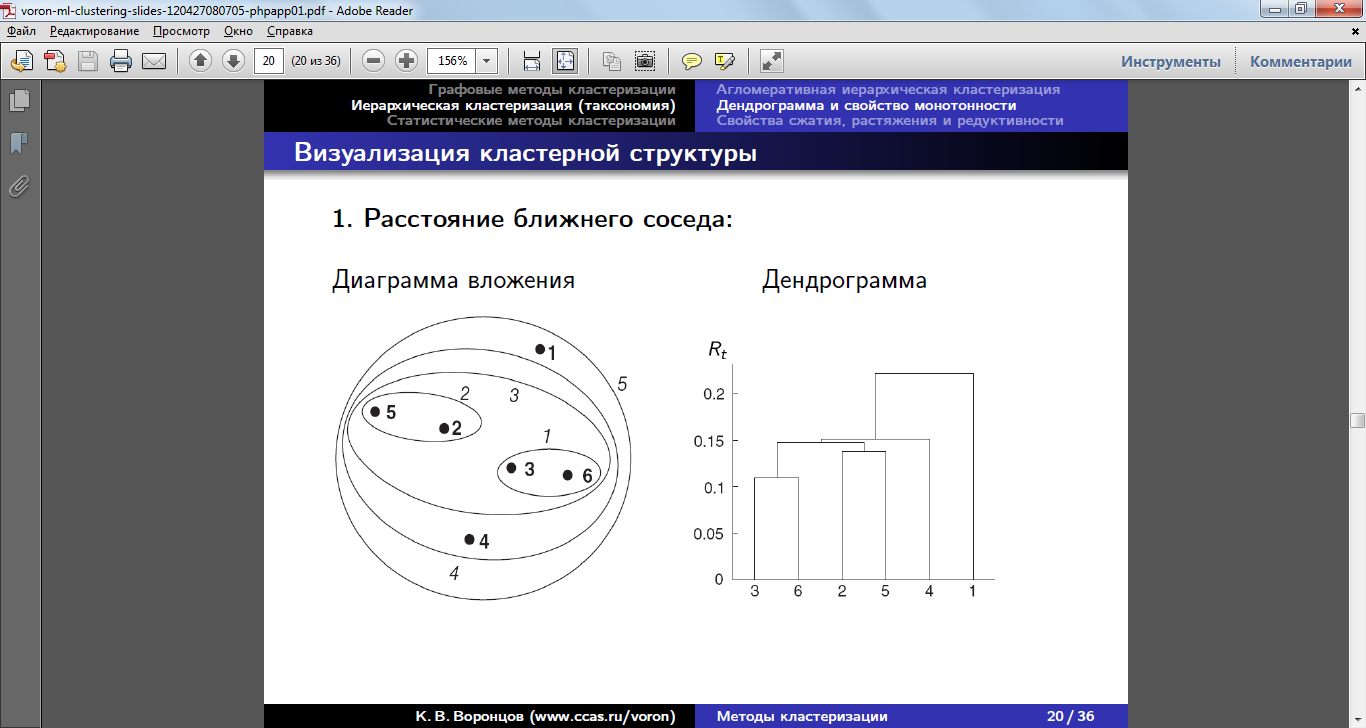


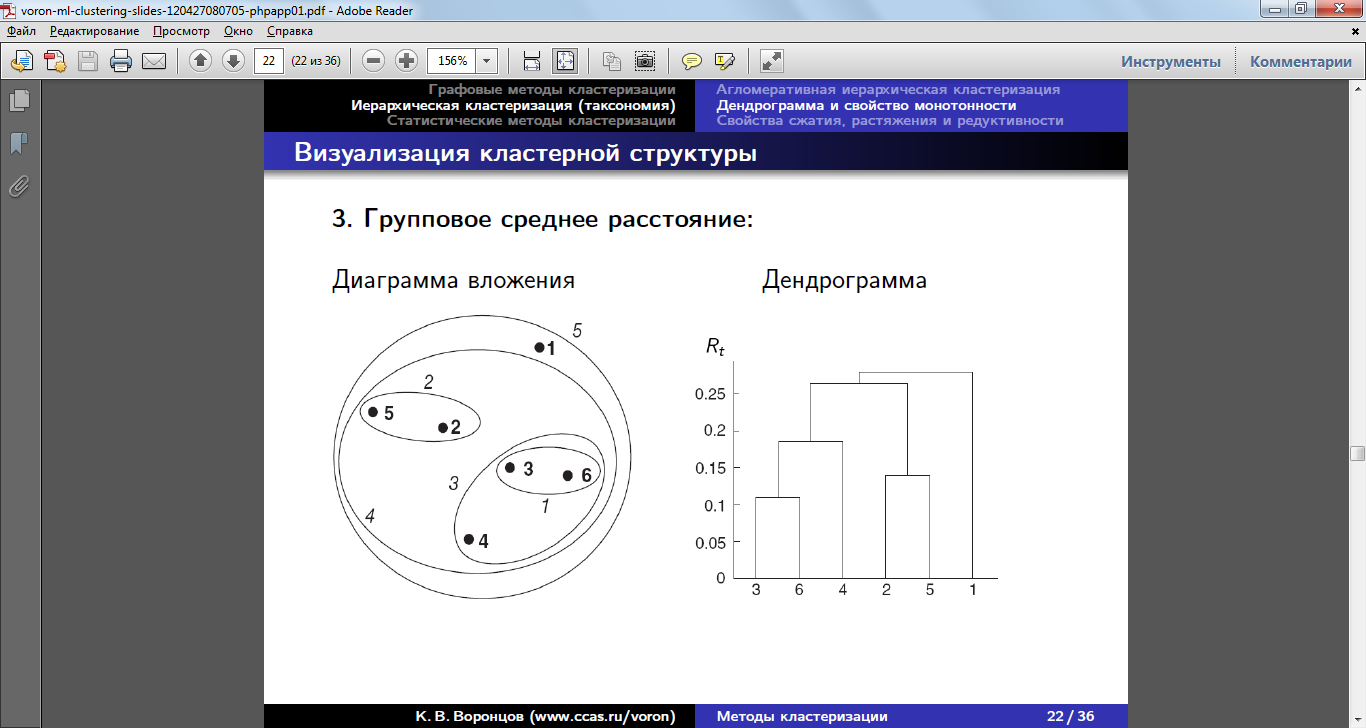


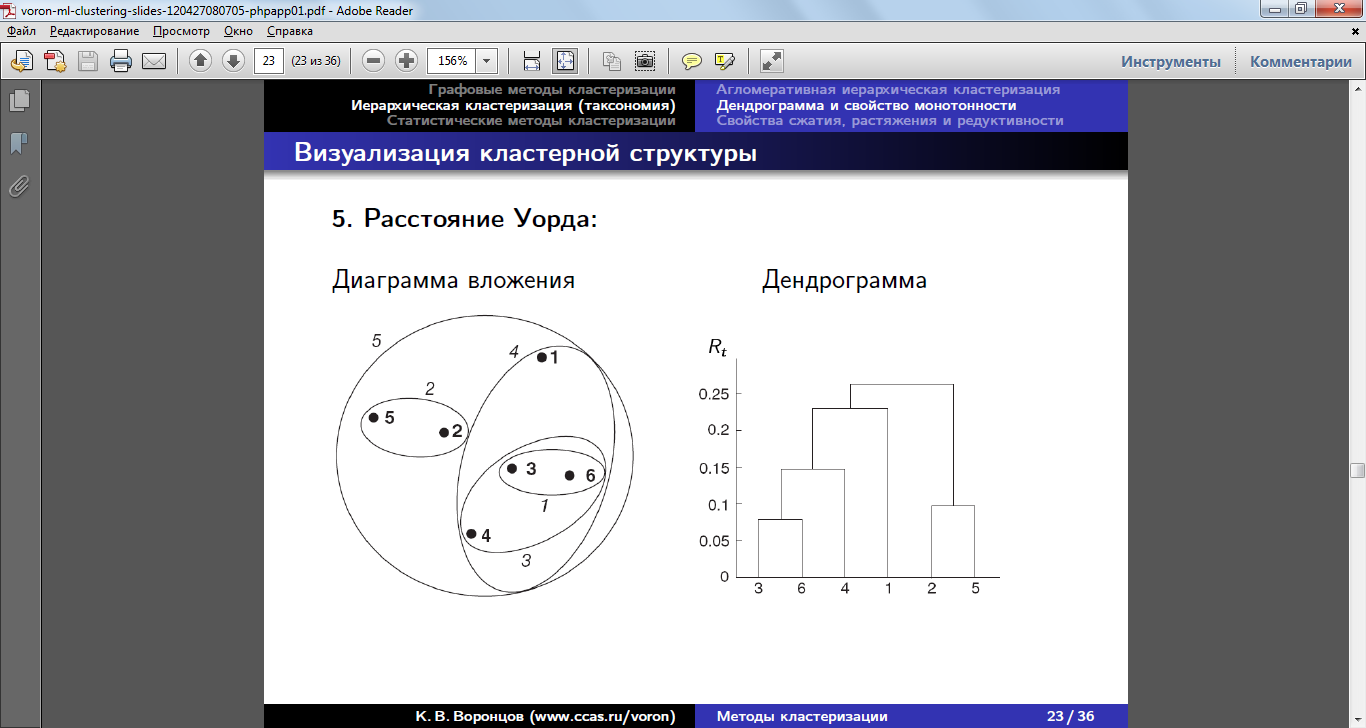
Сравним визуально с помощью диаграммы вложений и дендограмм приведенные метрики.

* 1. **Дендрограмма**

**Дендрограмма** — графическое изображение в двухмерной проекции степени сходства объектов.







* 1. **Свойства кластеризации**

Ниже приведены свойства, которые помогут выбрать нужное кластерное расстояние.

Кластеризация монотонна, если при каждом слиянии

расстояние между объединяемыми кластерами только

увеличивается: R2 < R3 < . . . < Rn.

Если кластеризация монотонна, то дендрограмма не имеет

самопересечений.

Rц не монотонно; Rб, Rд, Rг, Rу - монотонны.

Кластеризация сжимающая, если Rt < ρ(μU, μV ), ∀t.

Кластеризация растягивающая, если Rt > ρ(μU, μV ), ∀t.

Иначе кластеризация сохраняет метрику пространства.

Свойство растяжения наиболее желательно, так как

оно способствует более чёткому отделению кластеров.

Rб - сильно сжимающее;

Rд, Rу - растягивающие;

Rг, Rц - сохраняют метрику пространства.

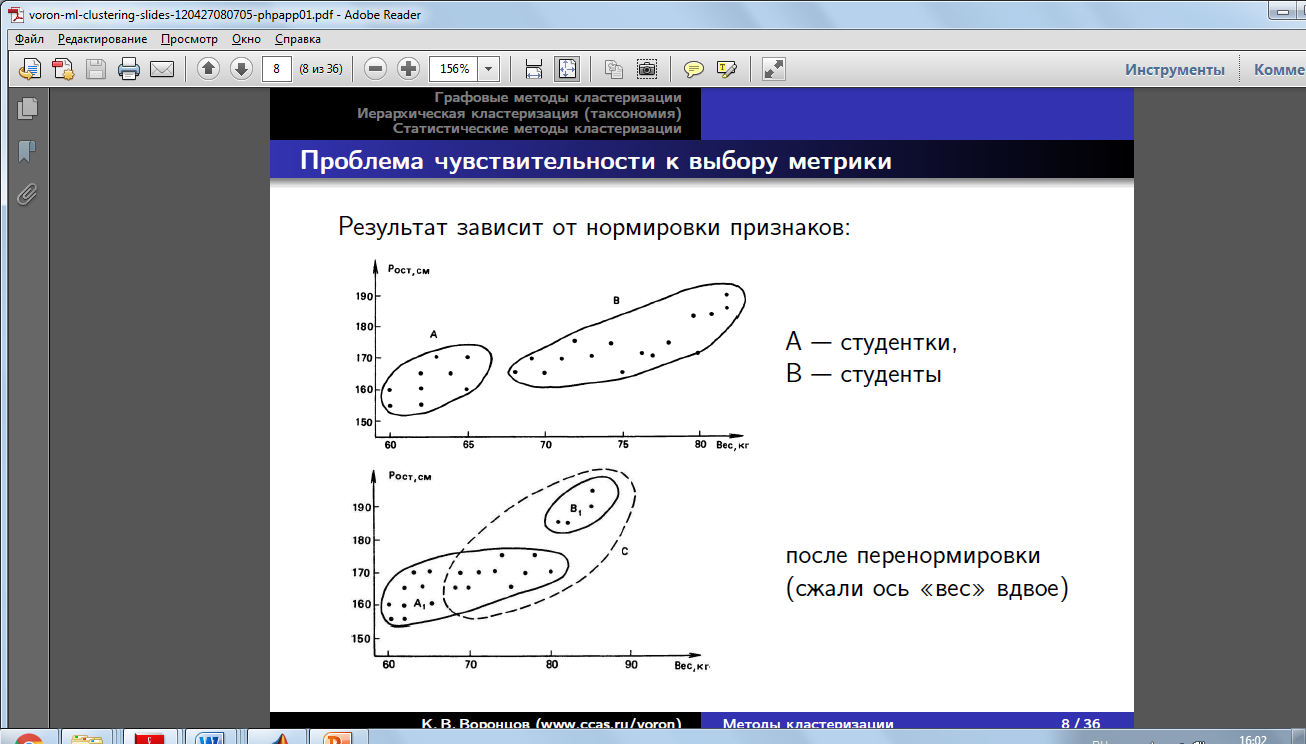
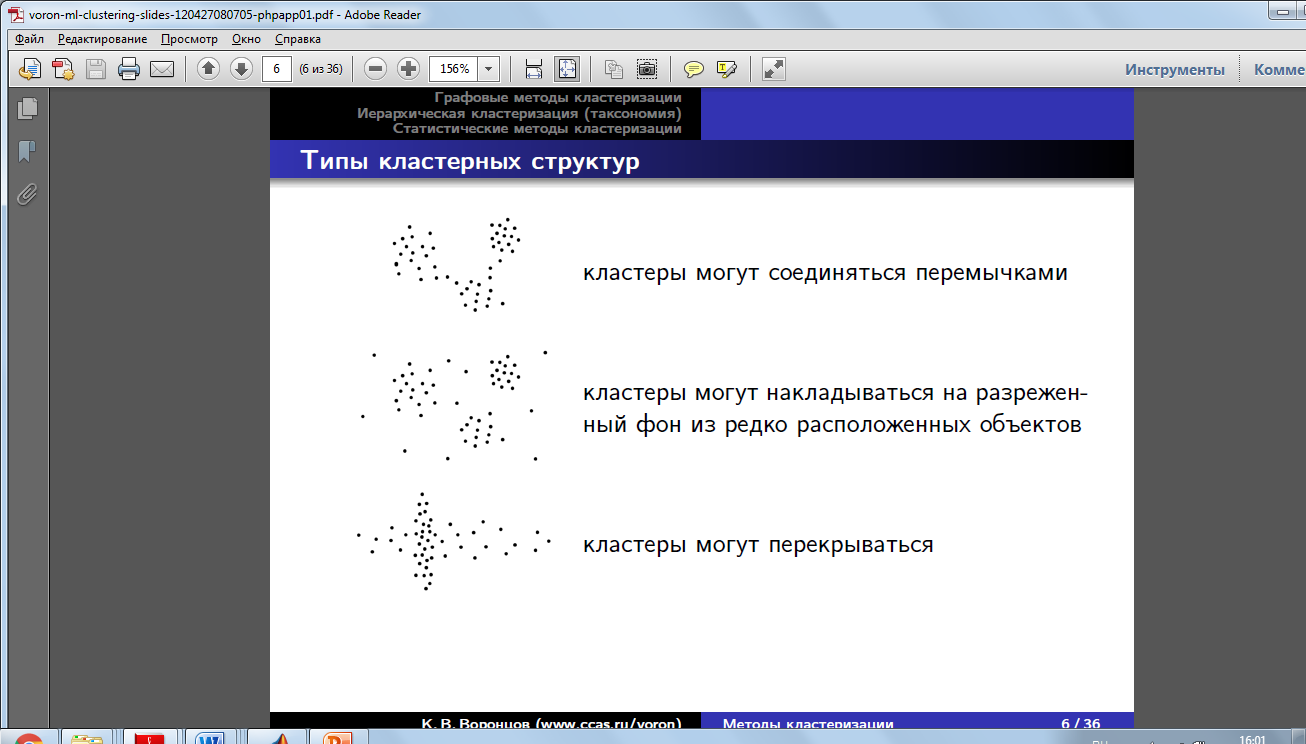
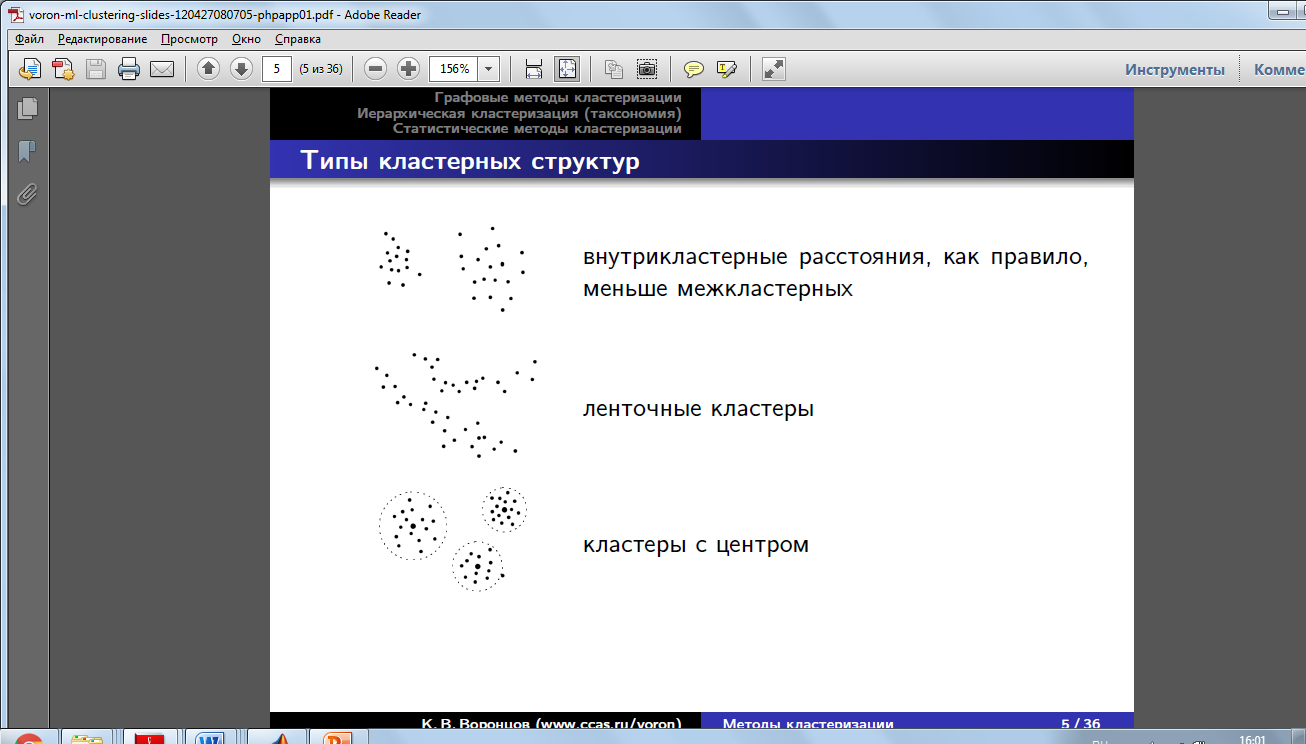


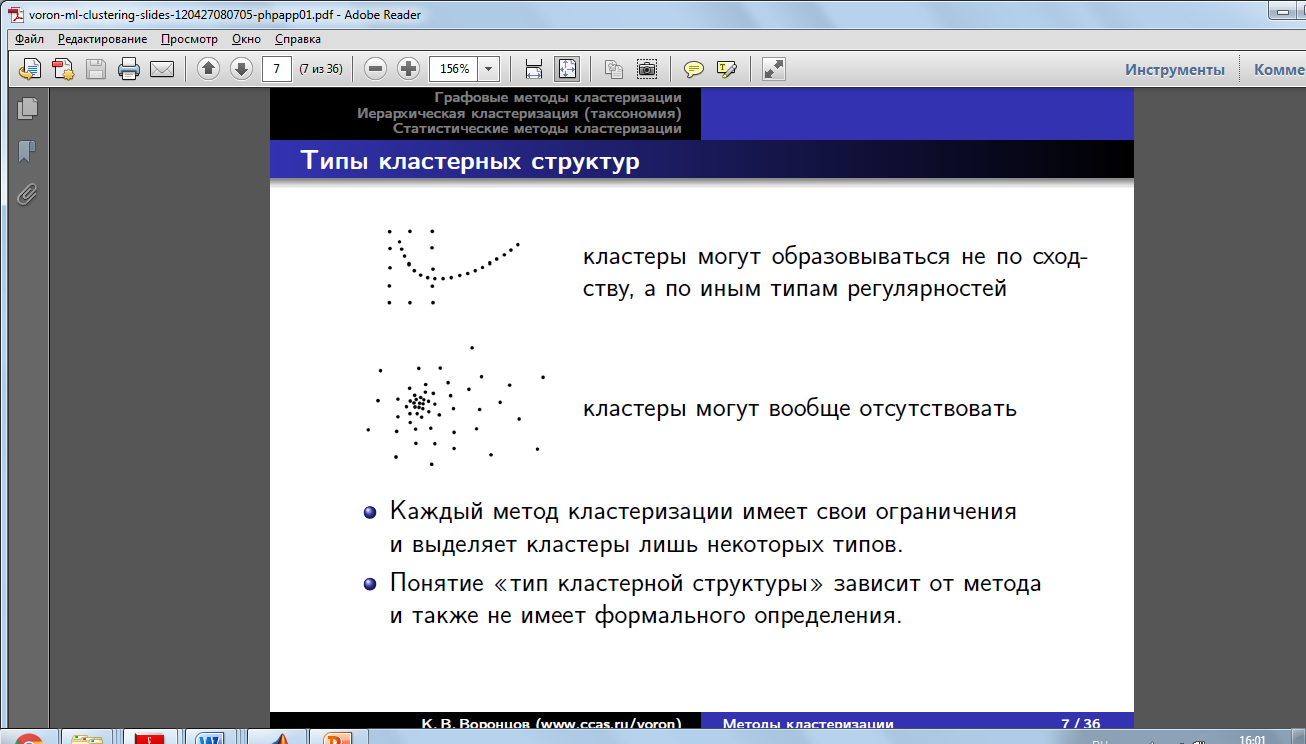
Рисунок 1 – чувствительность к выбору метрики

На Рисунке 1 видно, что результат кластеризации сильно зависит от выбранной метрики.

1. Тип кластерной структуры

Приведем ниже некоторые «типы кластерных структур»





Каждый метод кластеризации имеет свои ограничения

и выделяет кластеры лишь некоторых типов.

Понятие «тип кластерной структуры» зависит от метода

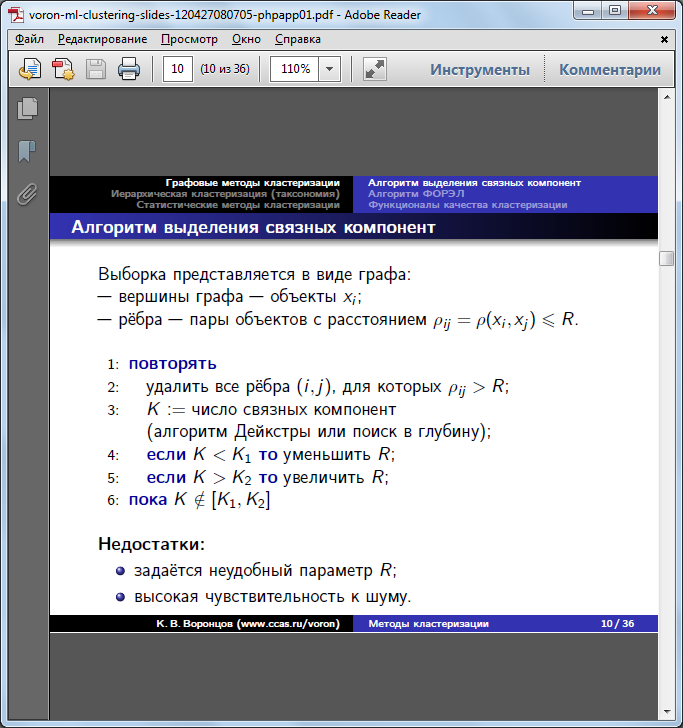
и также не имеет формального определения.

1. Алгоритмы кластеризации

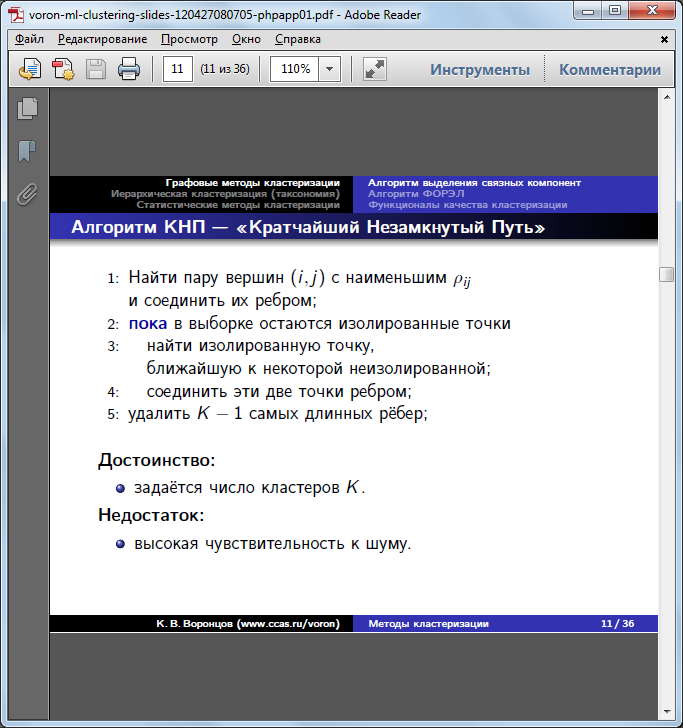
Как говорилось выше, в данной работе мы рассмотрим следующие виды кластеризаций:

* Графовые методы кластеризации
* Иерархическая кластеризация (таксономия)
* Статистические методы кластеризации
  1. **Графовые методы кластеризации**

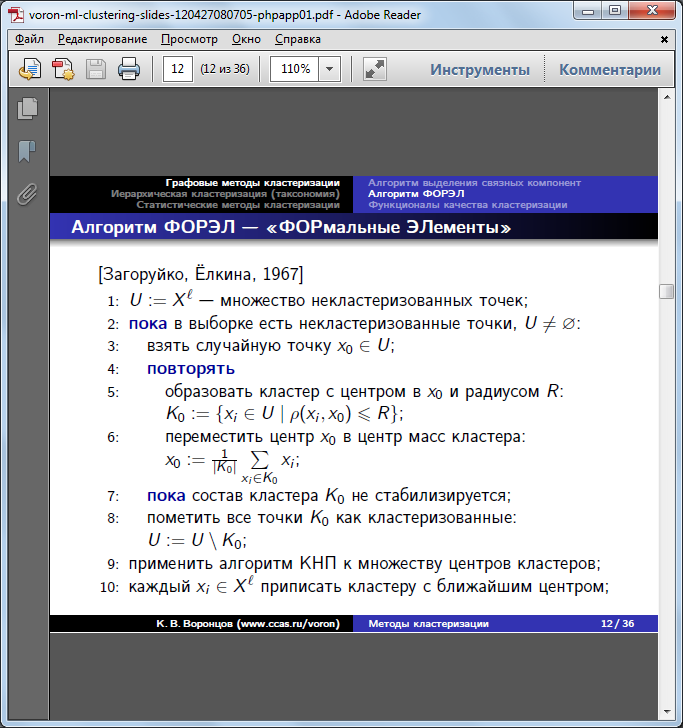
Алгоритм выделения связных компонент.

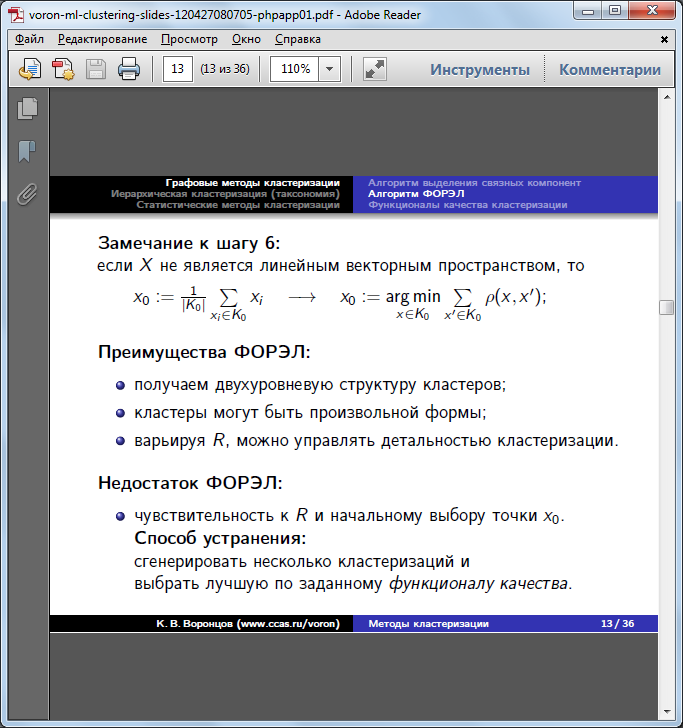


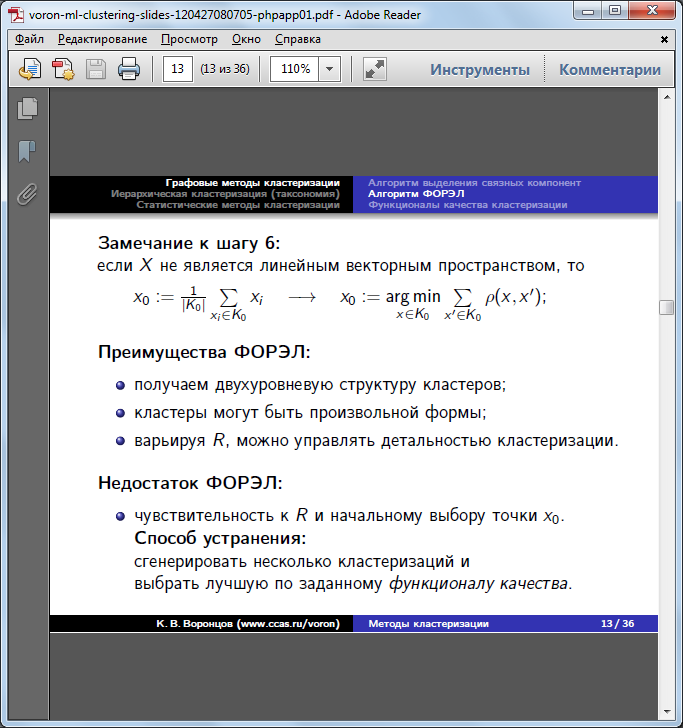
Алгоритм КНП – «Кратчайший Незамкнутый Путь»



Алгоритм ФОРЭЛ – «ФОРмальные ЭЛементы»

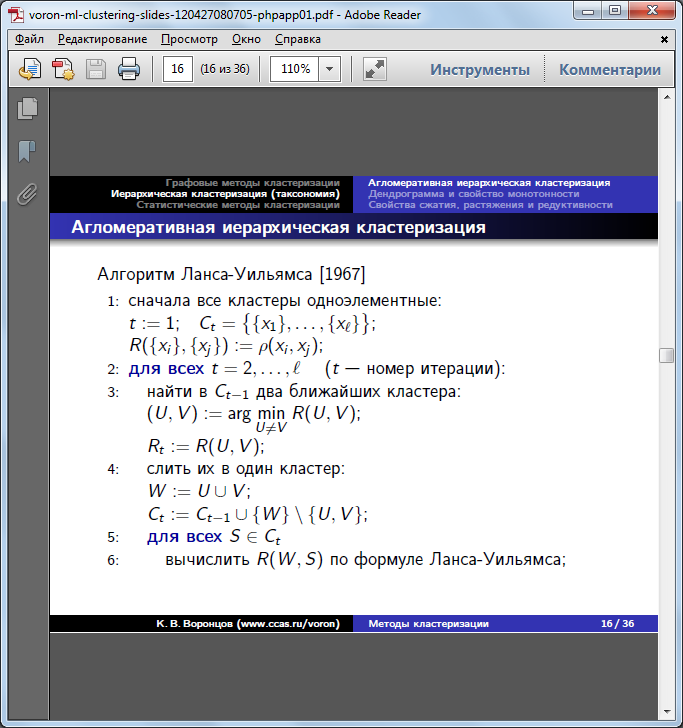




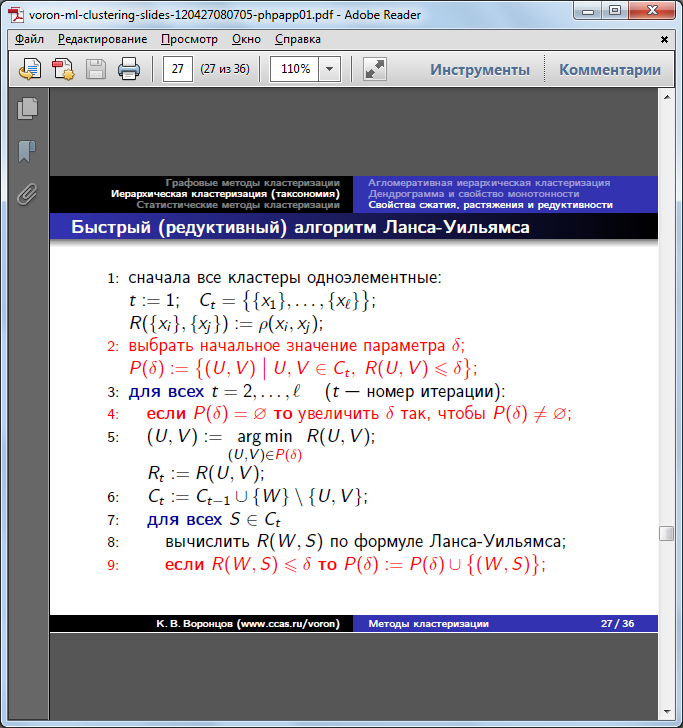


* 1. **Иерархическая кластеризация (таксономия)**

Алгоритм Ланса-Уильямса



Быстрый (редуктивный) алгоритм Ланса-Уильямса:

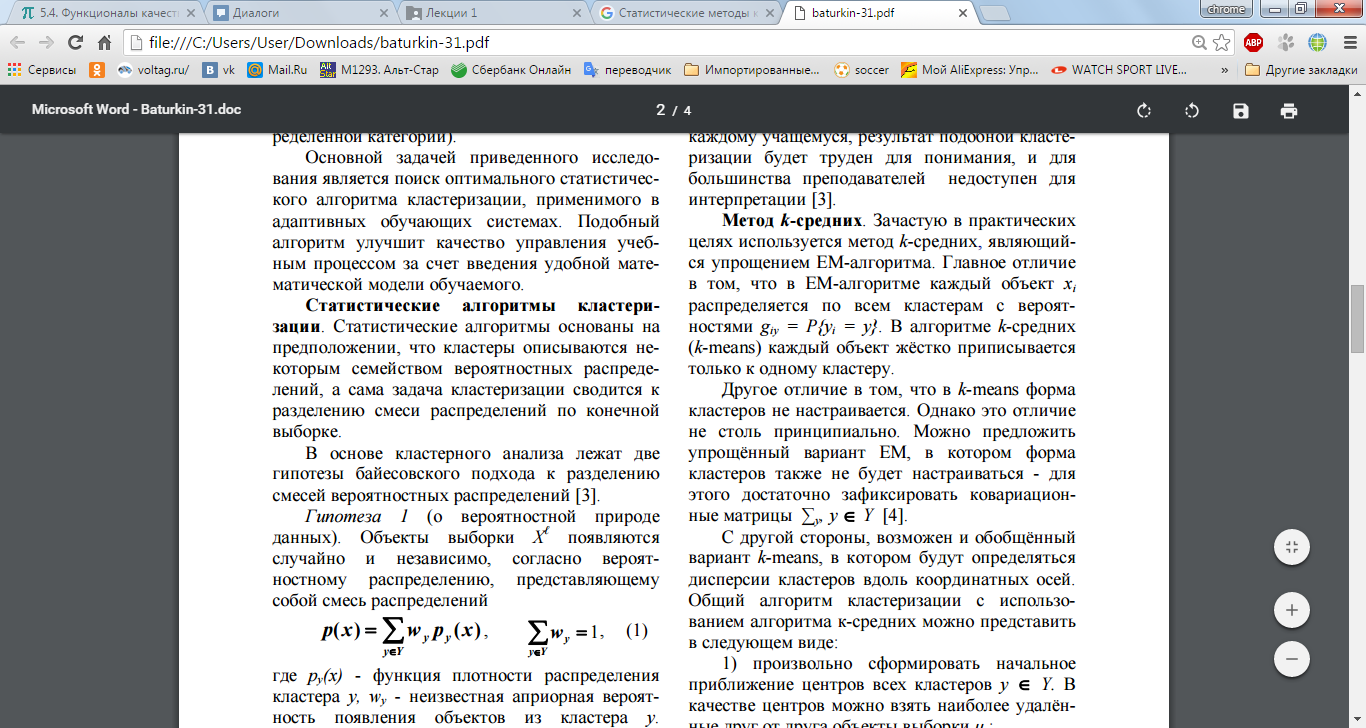


* 1. **Статистические методы кластеризации**

Статистические алгоритмы основаны на предположении, что кластеры описываются некоторым семейством вероятностных распределений, а сама задача кластеризации сводится к разделению смеси распределений по конечной выборке.

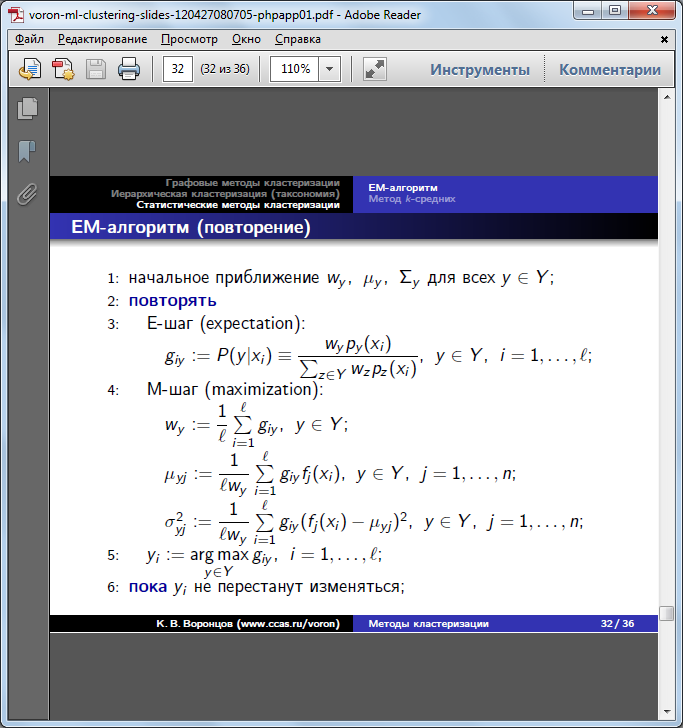
В основе кластерного анализа лежат две гипотезы байесовского подхода к разделению смесей вероятностных распределений.

Гипотеза 1 (о вероятностной природе данных): объекты выборки появляются случайно и независимо, согласно вероятностному распределению, представляющему собой смесь распределений.



Гипотеза 2 (о форме кластеров): объекты описываются n числовыми признаками f1(x),…,fn(x), X=R n . Каждый кластер y∈Y описывается n-мерной гауссовской плотностью py(x)=N(x;μy,∑y) с центром μy=(μy1,..., μyn) и диагональной ковариационной матрицей ∑y = diag(σ2y1, . . . , σ2yn). При этих предположениях задача кластеризации совпадает с задачей разделения смеси вероятностных распределений и для её решения можно применить EM-алгоритм.

EM-алгоритм:

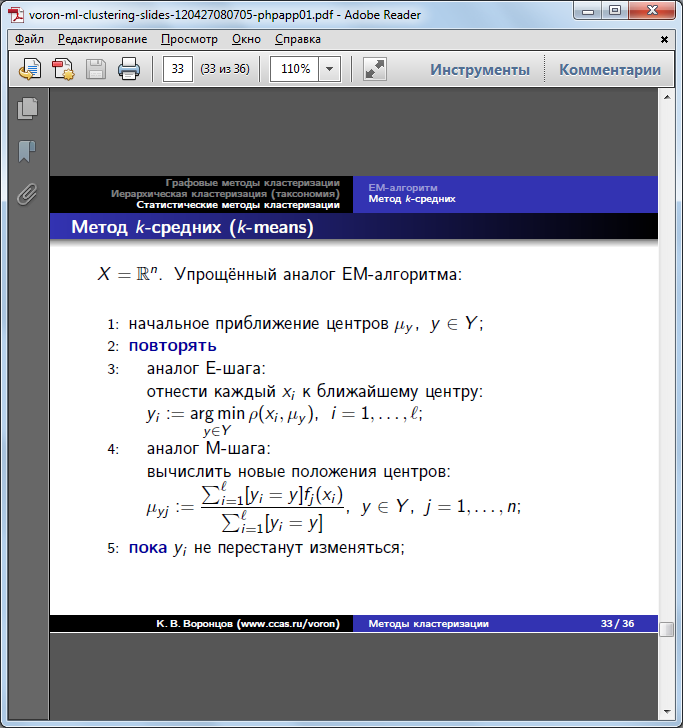


Метод k-средних (k-means)

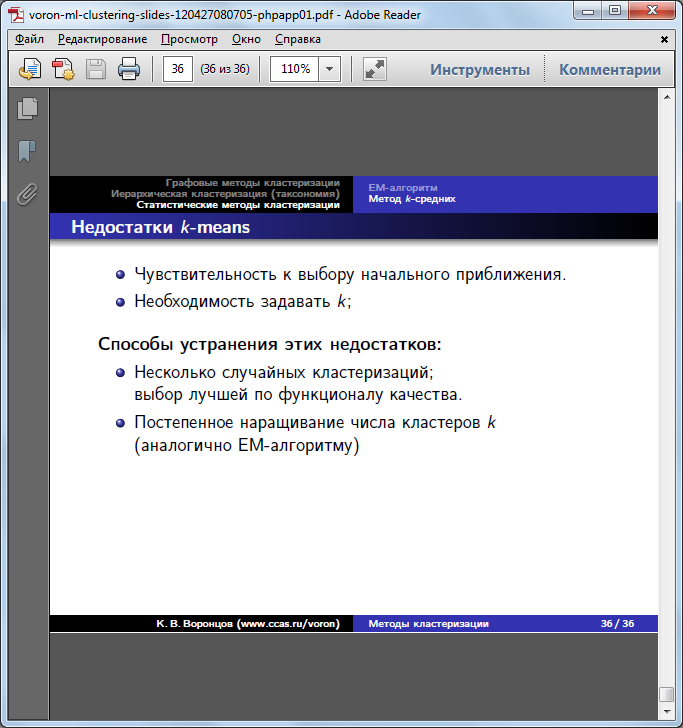
Алгоритм представляет собой версию [EM-алгоритма](https://ru.wikipedia.org/wiki/EM-%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC), применяемого также для разделения смеси [гауссиан](https://ru.wikipedia.org/wiki/%D0%93%D0%B0%D1%83%D1%81%D1%81%D0%B8%D0%B0%D0%BD%D0%B0" \o "Гауссиана). Он разбивает[множество](https://ru.wikipedia.org/wiki/%D0%9C%D0%BD%D0%BE%D0%B6%D0%B5%D1%81%D1%82%D0%B2%D0%BE) элементов [векторного пространства](https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D1%81%D1%82%D0%B2%D0%BE) на заранее известное число кластеров *k*.

Основная идея заключается в том, что на каждой итерации перевычисляется [центр масс](https://ru.wikipedia.org/wiki/%D0%A6%D0%B5%D0%BD%D1%82%D1%80_%D0%BC%D0%B0%D1%81%D1%81) 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 центра масс кластеров. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение *V* не увеличивается, поэтому зацикливание невозможно.



Недостатки k-means:

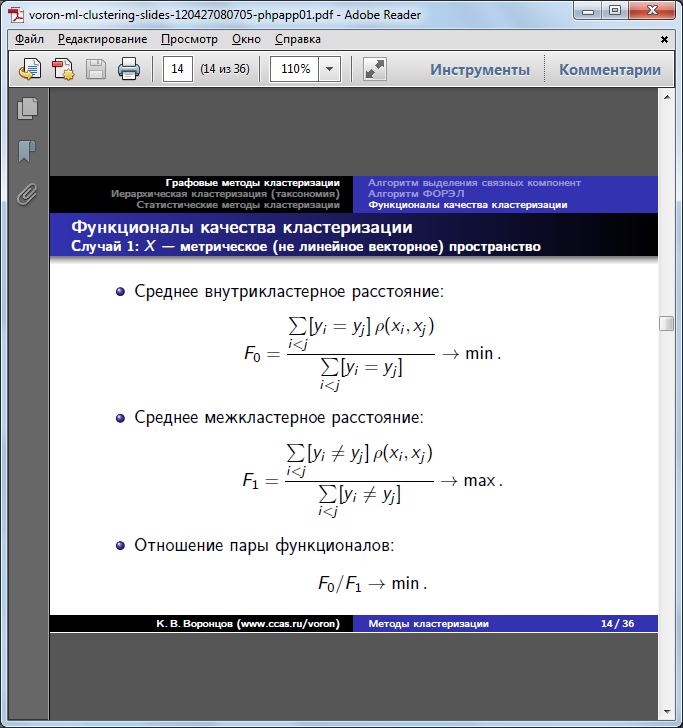


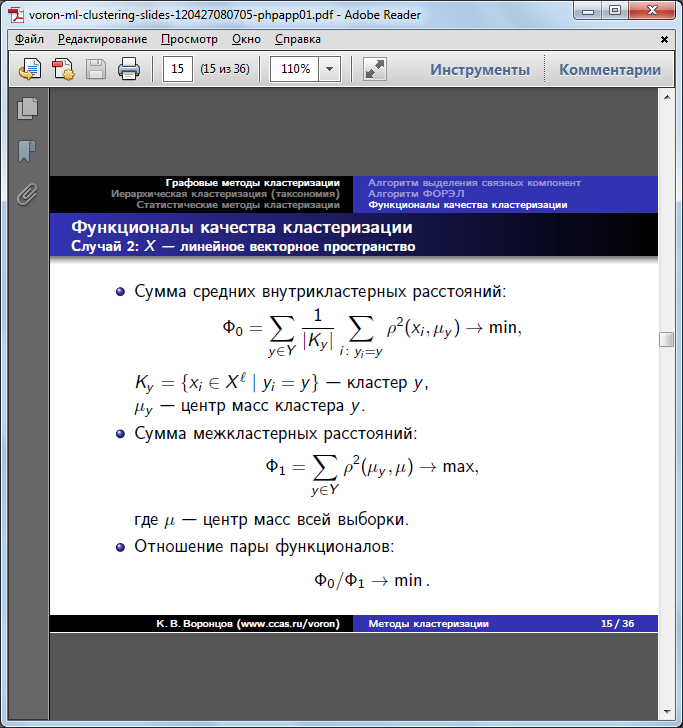
1. Функционалы качества кластеризации

С целью качественного сравнения одного кластерного разбиения с другим вводится понятие функционала качества кластеризации.

Функционалом он называется потому, что чаще всего разбиение задается, вообще говоря, набором дискриминантных функций . Тогда под наилучшим разбиением понимается то разбиение, на котором достигается экстремум выбранного функционала качества.

Выбор того или иного функционала качества, как правило, осуществляется весьма произвольно и опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо строгую формализованную систему.





1. Тестирование алгоритмов

Для тестирования выше приведенных алгоритмов воспользуемся математическим пакетом Matlab R2014a.

Смоделируем случайную выборку распределенную по нормальному гауссовскому закону распределения и проведем кластеризацию алгоритмом k-means.

Критерием кластеризации является минимум внутрикластерной суммы расстояний точек кластера до его центроида. Сумма расстояний находится по всем кластерам. В качестве расстояний точек кластера до его центроида используется квадрат евклидового расстояния.

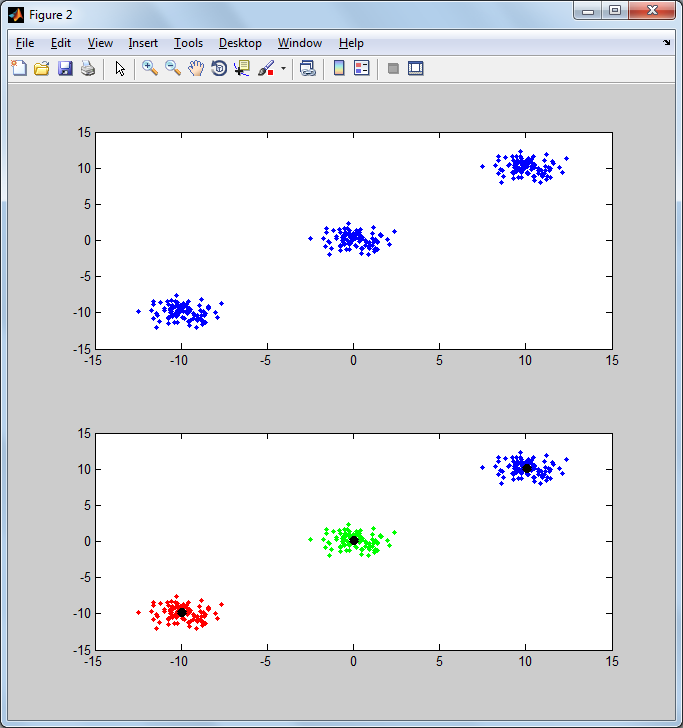


Рисунок 2 – результат работы алгоритма k-means

На рисунке 2 видно, что алгоритм k-means сработал корректно и выделил три кластера разными цветами.

Немного усложним задачу, и сгенерируем более сложную кластерную структуру.

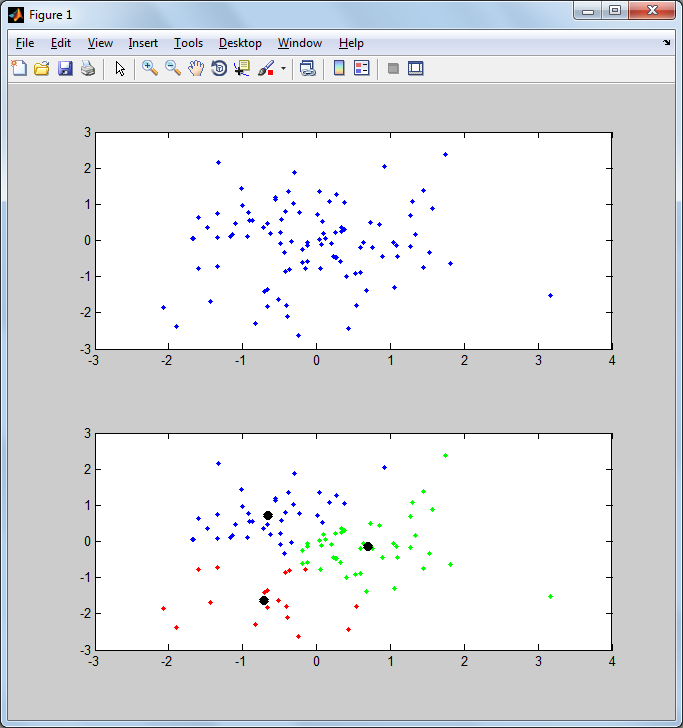


Рисунок 3 – результат работы алгоритма k-means

|  |  |  |  |
| --- | --- | --- | --- |
| Итерации | Фаза оптимизации | Кол-во задействованых объектов | Общая сумма расстояний |
| 1 | 1 | 100 | 107.899 |
| 2 | 1 | 10 | 97.9344 |
| 3 | 1 | 8 | 87.1423 |
| 4 | 1 | 4 | 85.5288 |
| 5 | 1 | 1 | 85.4465 |
| 6 | 2 | 0 | 85.378 |

Таблица 1 - информация по каждой итерации k-means

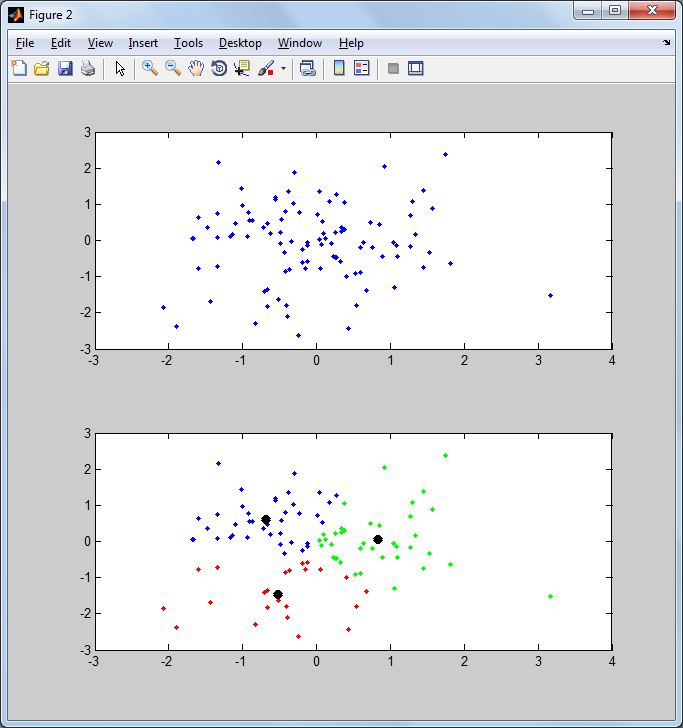


Рисунок 4 – результат работы алгоритма k-means

|  |  |  |  |
| --- | --- | --- | --- |
| Итерации | Фаза оптимизации | Кол-во задействованых объектов | Общая сумма расстояний |
| 1 | 1 | 100 | 96.0016 |
| 2 | 1 | 11 | 88.1638 |
| 3 | 1 | 2 | 87.8551 |
| 4 | 1 | 4 | 86.8368 |
| 5 | 1 | 2 | 86.5195 |
| 6 | 1 | 3 | 86.1405 |
| 7 | 1 | 1 | 85.8788 |
| 8 | 2 | 0 | 85.6209 |

Таблица 2 - информация по каждой итерации k-means

Видим, что результат работы алгоритма при повторном запуске немного отличаются друг от друга, это объясняется тем, что алгоритм k-means зависит от начальных условий. Причем результат работы алгоритма в первом случае лучше, чем во втором, т.к общая сумма расстояний объектов до своих центроидов в первом случае меньше.

1. Заключение

В процессе выполнения научно-исследовательской работы бакалавра, я ознакомился с основными понятиями кластерного анализа. В ходе работы я изучил основные алгоритмы кластеризации и провел тестирование данных алгоритмов на смоделированной случайной выборке.

В результате, мною были построены графики и сделаны выводы о качестве работы алгоритма над заданной выборкой. В дальнейшем я планирую более подробно изучить данную область т.к. она высоко оценивается исследовательскими компаниями и имеет перспективы развития, а также кажется мне довольно интересной.

Список использованных источников

1. MATLAB.Exponenta [Электронный ресурс]. URL: <http://matlab.exponenta.ru/statist/book2/14/kmeans.php> (дата обращения: 15.12.2015).
2. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. URL: [www.machinelearning.ru](http://www.machinelearning.ru/) (дата обращения: 17.12.2015).
3. Воронцов К.В. [Алгоритмы кластеризации и многомерного шкалирования](http://www.ccas.ru/voron/download/Clustering.pdf). Курс лекций. МГУ, 2007.
4. С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. Прикладная статистика: классификация и снижение размерности. М. Финансы и статистика, 1989.