Министерство образования и науки Российской Федерации

ФГБОУ ВО Рыбинский государственный авиационный технический университет имени П.А. Соловьева

Факультет радиоэлектроники и информатики

Кафедра математического и программного обеспечения электронных вычислительных средств

# ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ

по дисциплине

**Математические методы анализа данных**

по теме

Кластерный анализ

Студент группы ИПБ-13 Иванов Р.А.

Преподаватель, доцент Воробьев К. А.

Рыбинск 2017

# Содержание

# 1 Исходные данные

В качестве исходных данных для выполнения класстерного анализа были сгенериррованы на плоскости 50 точек с координатами в диапазоне от (0,0) до (3,8), значение каждой координаты было получено на основе генератора случайных чисел с равномерным распределением(рис. 1).

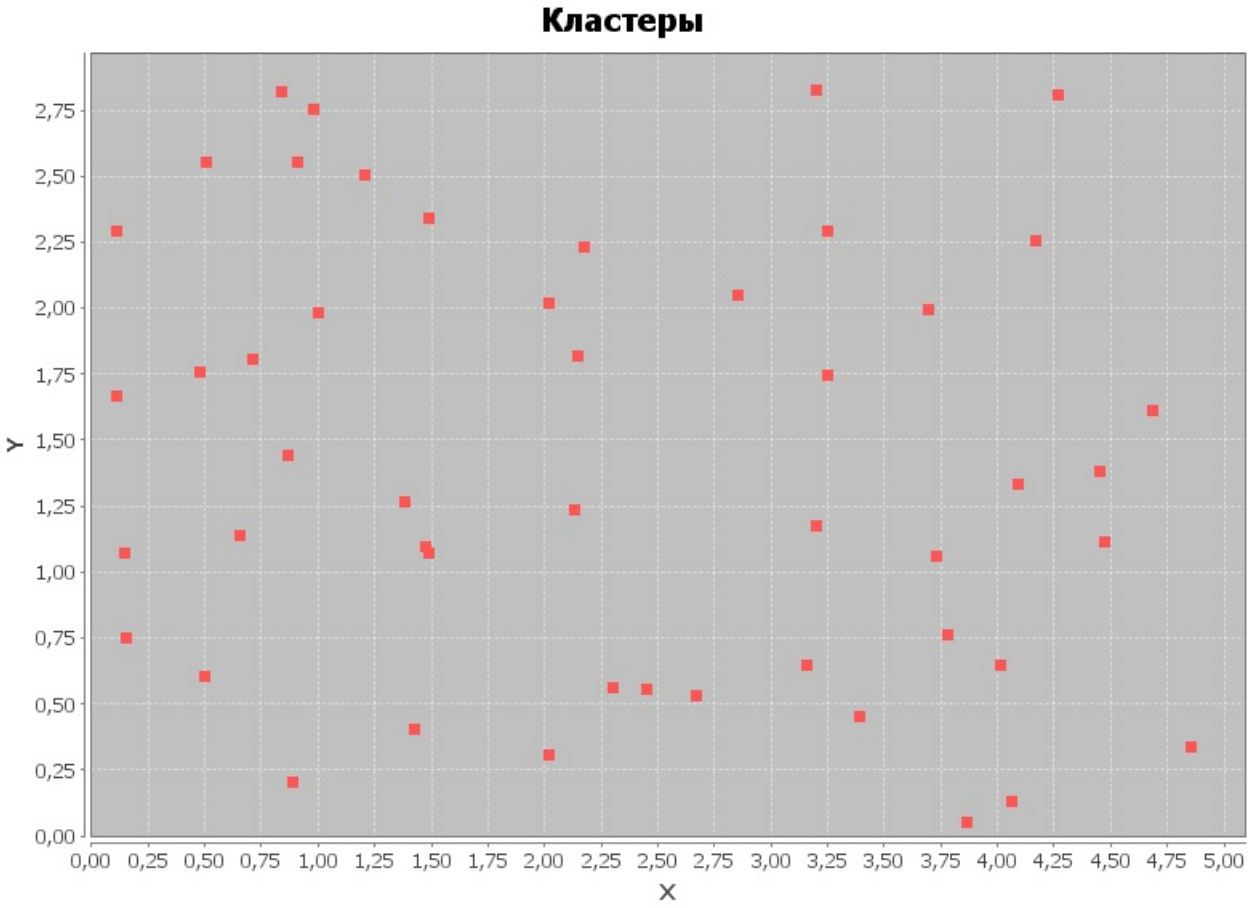


Рис. 1: Исходные данные

# 2 Задача кластерного анализа

Кластерный анализ выполняет следующие основные задачи:

- Разработка типологии или классификации.
- Исследование полезных концептуальных схем группирования объектов.
- Порождение гипотез на основе исследования данных.
- Проверка гипотез или исследования для определения, действительно ли типы, выделенные тем или иным способом, присутствуют в имеющихся данных.

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:

- Отбор выборки для кластеризации. Подразумевается, что имеет смысл кластеризовать только количественные данные.
- Определение множества переменных, по которым будут оцениваться объекты в выборке, то есть признакового пространства.
- Вычисление значений той или иной меры сходства (или различия) между объектами.
- Применение метода кластерного анализа для создания групп сходных объектов.
- Проверка достоверности результатов кластерного решения.

Стоит отметить, что в отличие от задачи классификации, где классы исследуемых объектов уже заранее известны, кластерный анализ не предполагает такого знания.

# 3 Метод ближайшего соседа

Множество методов иерархического кластерного анализа различается не только используемыми мерами сходства и различия, но и алгоритмами классификации. Из них наиболее распространен метод ближайшего соседа. Этот метод известен также под названием метод одиночной связи.

Пусть требуется провести классификацию заданного множества объектов методом ближайшего соседа. Расстояние между двумя классами определяется как расстояние между ближайшими их представителями.

Перед началом работы алгоритма рассчитывается матрица расстояний между объектами. На каждом шаге в матрице расстояний ищется минимальное значение, соответствующее расстоянию между двумя наиболее близкими кластерами. Найденные кластеры объединяются, образуя новый кластер. Эта процедура повторяется до тех пор, пока не будут объединены все кластеры.

В качестве критерия остановки процесса объединения кластеров было выбрано условие, что в рамках одного кластера расстояние между самыми удалёнными точками не должно превышать 3. Было полученно 12 кластеров(рис 2). Между кластерами чётко видны границы, значит можно говорить об удачном анализе.

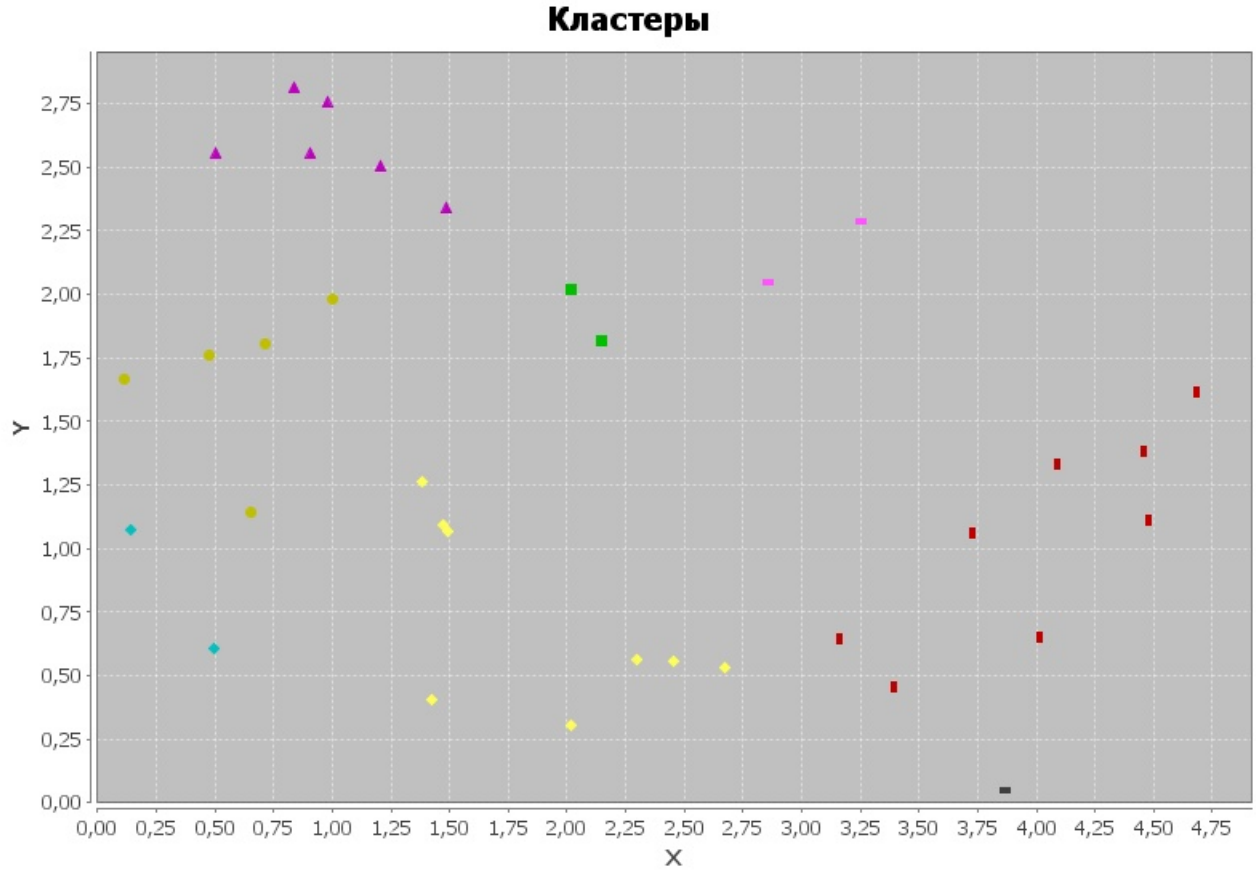


Рис. 2: Результат применения метода ближайшего соседа

# 4 Метод k-средних

Этот метод является наиболее популярным методом кластеризации, но обладает существенным недостатком: необходимо заранее знать количество искомых кластеров. Обычно для этого используют либо другие методы кластерного анализа, либо какие-то экспериментальные или интуитивные значения.

Суть алгоритма в том, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров. Множество элементов разбивается на заранее известное число кластеров случайным образом. На каждой итерации перевычисляется "центр масс"для каждого кластера, полученного на предыдущем шаге, затем элементы заново разбиваются на кластеры 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 не увеличивается, поэтому зацикливание невозможно.

Количество кластеров было взято 12(сколько было полученно предыдущим ме-

тодом). Результат несколько отличается от полученного ранее(рис. 3), это можно объяснить равномерностью распределения точек. Границы между классами видны, но может быть подобраны не лучшим образом, присутствуют довольно большие и довольно маленькие классы. Тут можно сказать, что качество разбиения зависит от начальных центров кластеров и их количества.

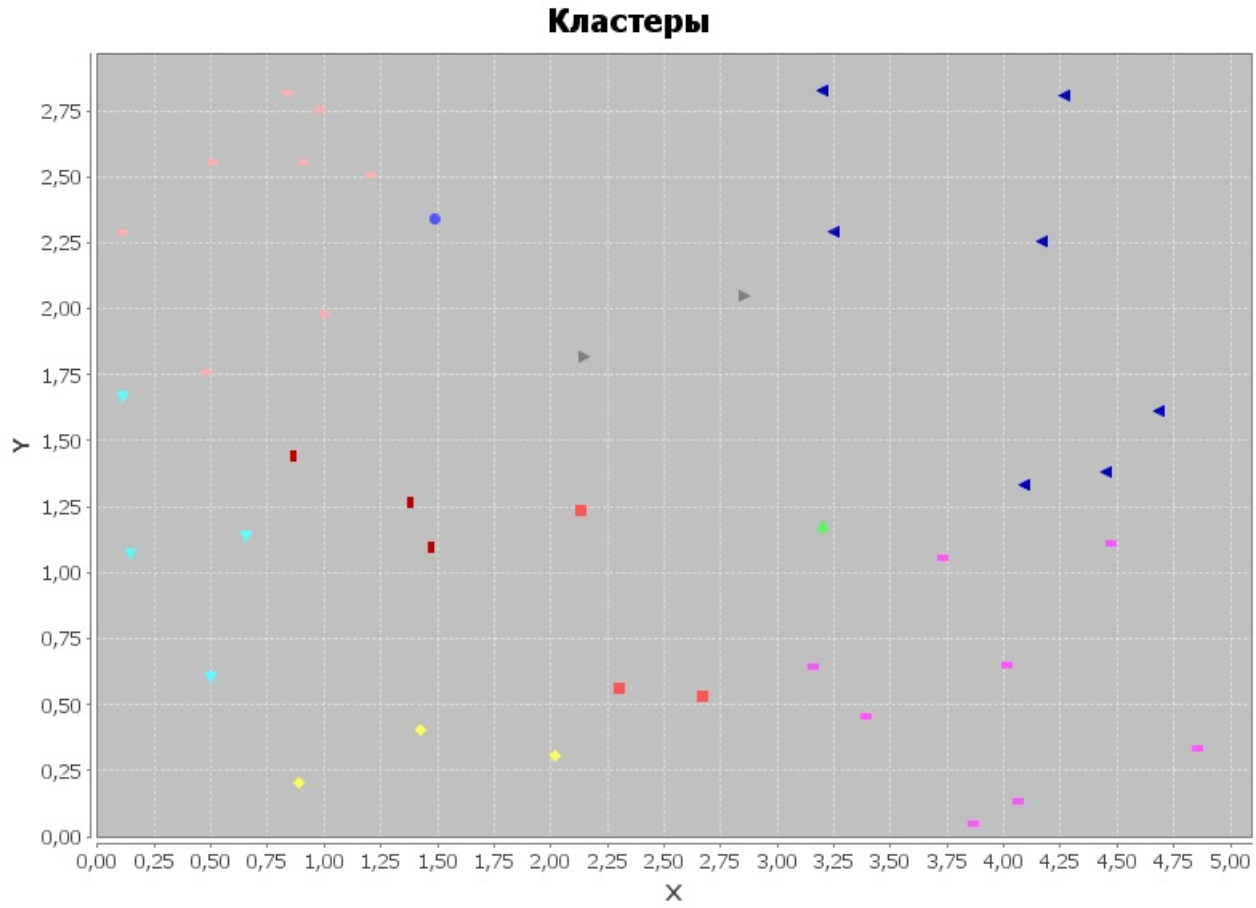


Рис. 3: Результат применения метода k-средних

# 5 Выводы

Результаты, полученные при помощи рассмотренных методов, очень сильно различаются, что вполне очевидн для случая с довольно равномерно распределенными случайными данными. Можно предположить, что при использовании более разнородных данных, к примеру, массива точек с ярко выраженными скоплениями, результаты были бы более похожими друг на друга. Тем не менее, в обоих случаях задача кластерного анализа была решена и получены ответы, адекватно отвечающие задаче разбиения элементов на классы.