Кластеризация точек на плоскости по Евклидовой метрике.

Владимирова Элина

26 декабря 2020 г.

1 Резюме

Была решена задача объединения точек на плоскости в кластеры по Евклидовому расстоянию.

Входные данные - количество кластеров. Посредством генерации определенных наборов пар координат на плоскости и проведения кластеризации задаваемых ими точек двумя методами: hierarchical на обучающей выборке и k-means на всем наборе - оказывается получено изображение осуществленного распределения путем рисования точек каждого кластера уникальным цветом из заданной палитры.

2 Постановка задачи

Заданные точки на плоскости ($\leqslant 10^7$ пар координат) необходимо разбить на известное количество кластеров следующим образом:

1. Осуществить случайную выборку точек (около 0,1% общего количества) из заданного массива;
2. Провести иерархическую кластеризацию на тестовой выборке;
3. Взяв за основу центроиды полученных первичных кластеров, провести кластеризацию методом k-means на остальных точках.

3 Допущения

- Количество точек считается заданным изначально.
- Количество кластеров задается посредством стандартного ввода.
- Генерация координат точек осуществляется функциями, выделенными в отдельный файл *pointGenerator.py*. Сделано это было в основном с целью получения нескольких различных простейших типов расположения точек в тестовых данных: обыкновенный случайный разброс, приближение к прямой, приближение к 3 фрагментам различных прямых и соприкасающиеся и вложенные эллипсы.
- Поскольку кластеризация точек тестовой выборки нужна только для получения первичных центроидов, тестовая выборка генерируется отдельно от общей.

4 Описание решения

Координаты точек основного массива и обучающей выборки генерируются по заданному количеству точек, после чего на тестовом множестве посредством встроенных функций Python производится иерархическая кластеризация. Координаты центроидов полученных кластеров вычисляются как среднее арифметическое координат всех точек кластера и используются как стартовое положение центров для K-Means, проводимого уже на общем массиве точек и написанного вручную.

Все графики строятся с использованием библиотеки *matplotlib.pyplot*. На итоговых изображениях различные кластеры отрисованы различными цветами, точки тестовой выборки выделены размером; также присутствуют отметки первичных (небольшие фиолетовые крестики) и конечных (средних размеров красные крестики) кластеров.

5 Результаты

Наборы точек, получаемые *pointGenerator.py*, успешно распределяются в кластеры способом, часто похожим на разбиение оригинального K-Means (также реализован в главном файле программы, но закомментирован, поскольку использовался только для сравнения с новым кластеризатором). Программа успешно обрабатывает до 10 000 000 точек для 10 кластеров.

Рис. 1: Разбиение случайно разбросанных на плоскости точек в диапазоне [0; 1000] 1000, 10 000 и 1 000 000 точек (+ 1 000 000 точек в приближении)

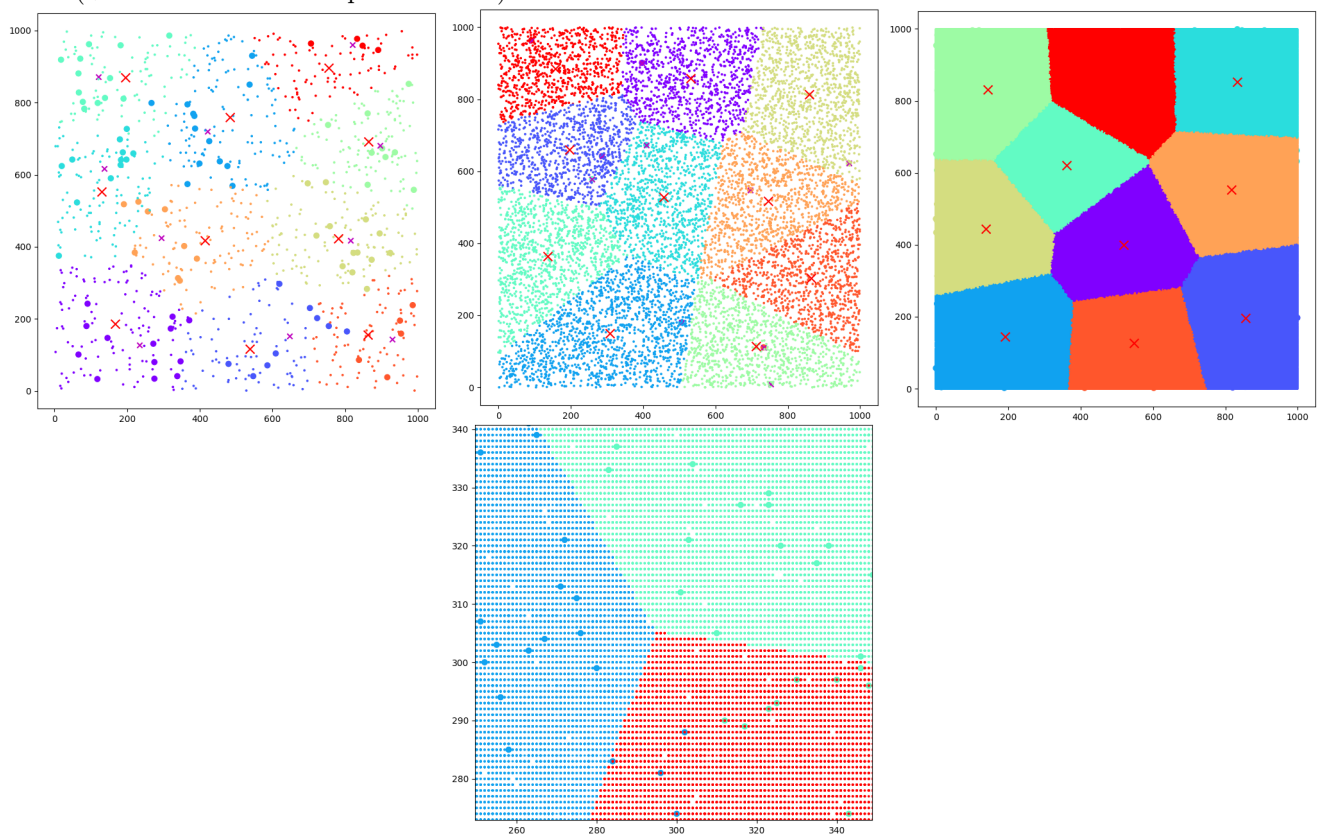
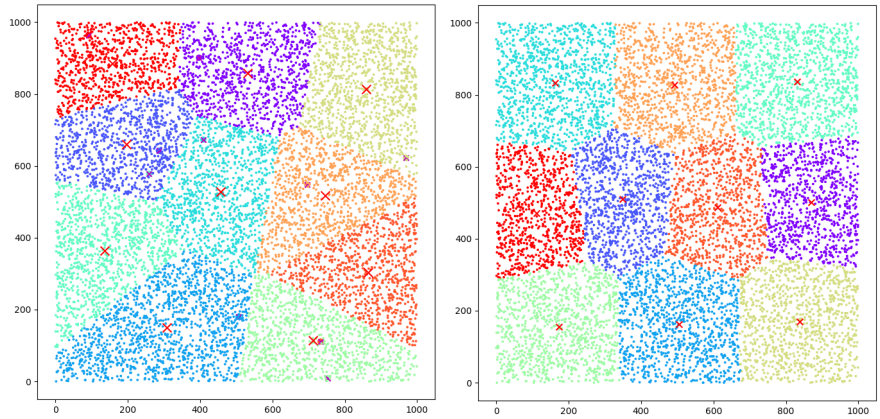
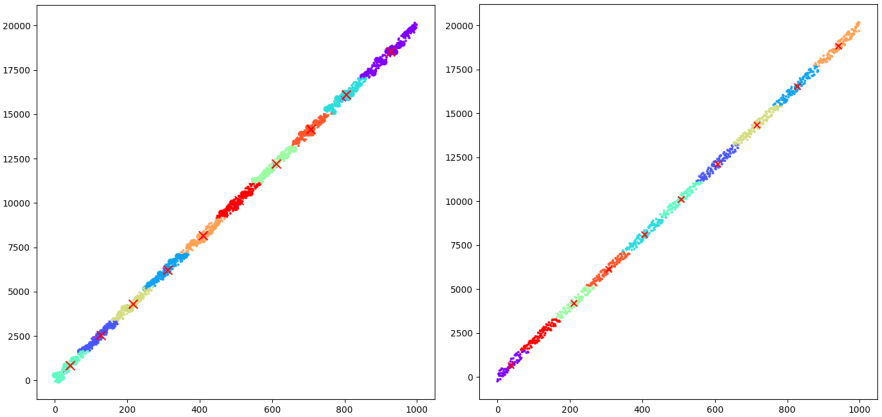


Рис. 2: Разбиение одного и того же массива точек новым кластеризатором и K-Means соответственно



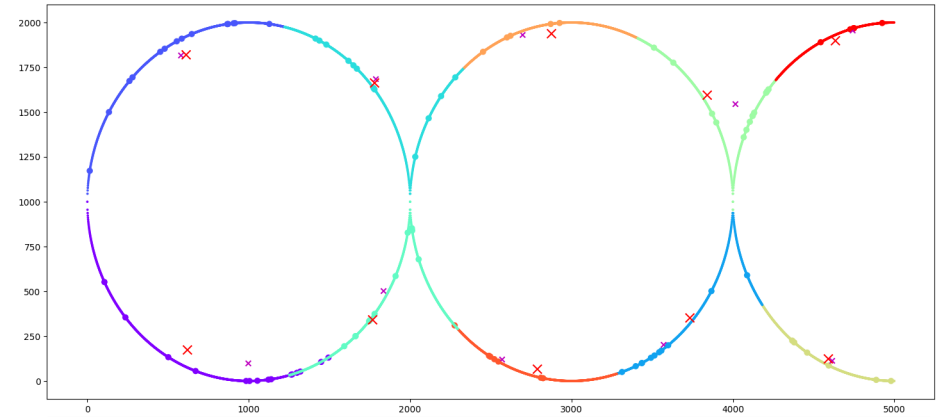
На неограниченной случайной выборке точек новый кластеризатор показывает меньшую систематичность.

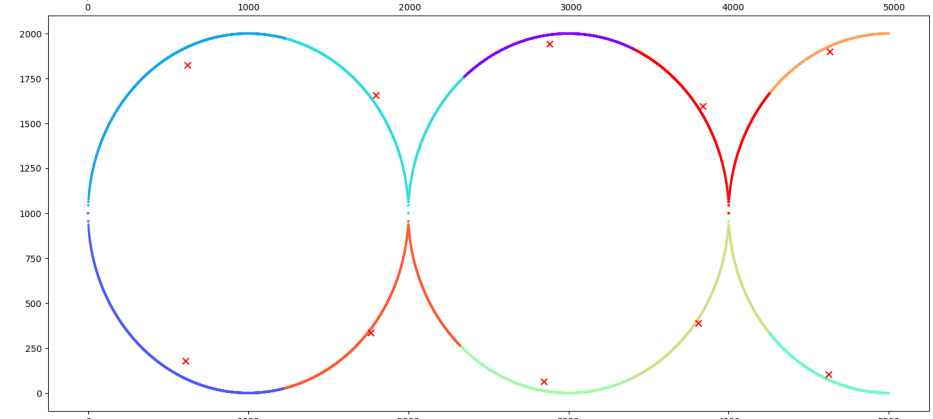
Рис. 3: Разбиение на кластеры точек, собранных в линию, новым кластеризатором и K-Means соответственно



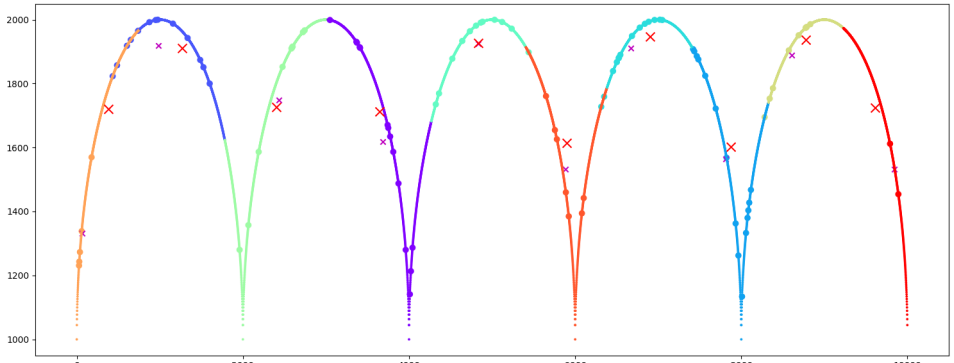
Разбиения прямой новым кластеризатором и K-Means практически идентичны.

Рис. 4: Разбиение более сложных графиков новым кластеризатором и K-Means соответственно





Разбиения практически совпадают.



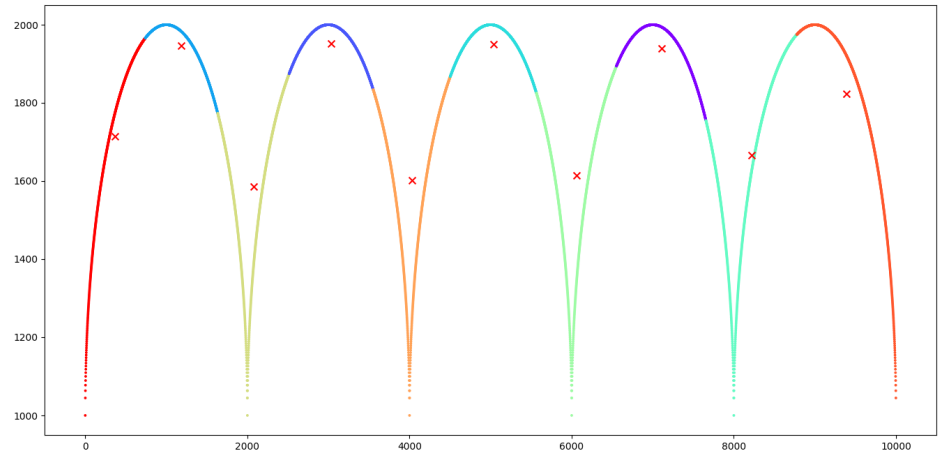
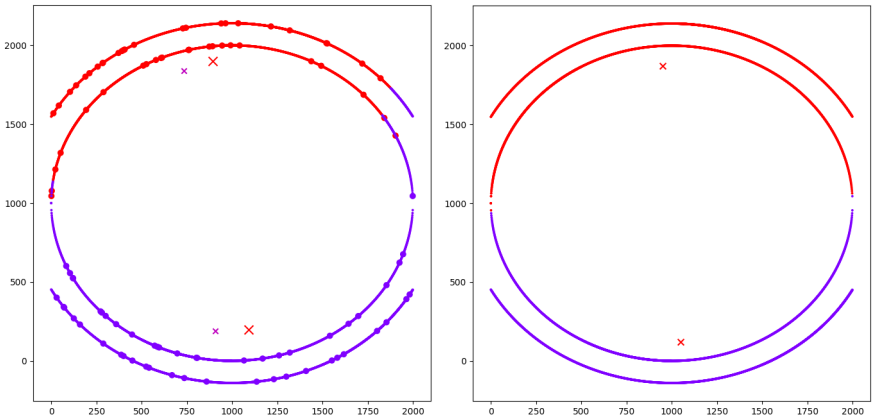
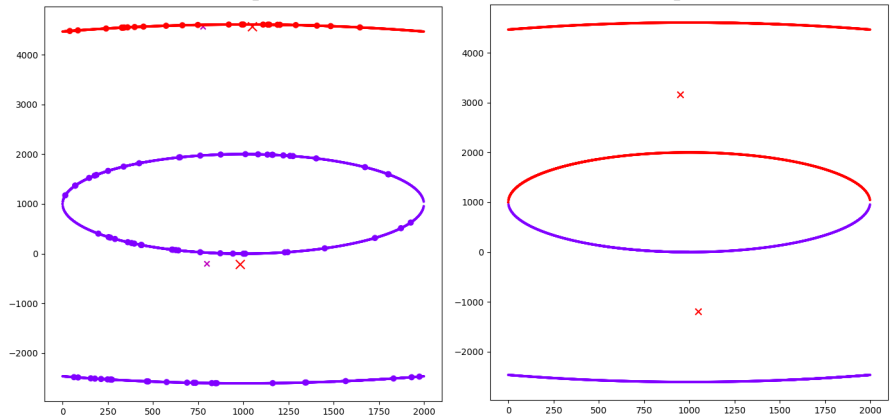


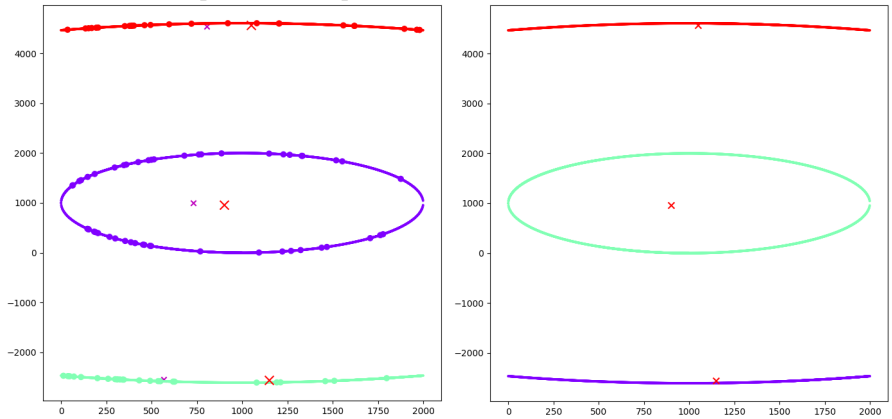
Рис. 5: Разбиение ситуаций со "вложенными"группами на 2 кластера новым кластеризатором и K-Means соответственно



В основном разбиения имеют незначительные различия.



Это наиболее явное различие в разбиениях из всех засвидетельствованных мной.



Однако оно исчезает при попытке разбиения на 3 кластера.

6 Выводы

Как и предполагалось, новый кластеризатор устраняет основные проблемы K-Means - например, более правильно разбирается со "вложенными"группами точек, - однако он также зависит от первоначальной тестовой выборки точек, что имеет свои последствия в основном в виде большей визуальной хаотичности разбиения.