Тестовое задание от команды CoreML 1 этап

Hierarchical agglomerative single link clustering

- Реализовать
- Эффективно по времени и памяти
- Найти данные, на которых алгоритм даёт некорректный результат
- Продемонстрировать работоспособность

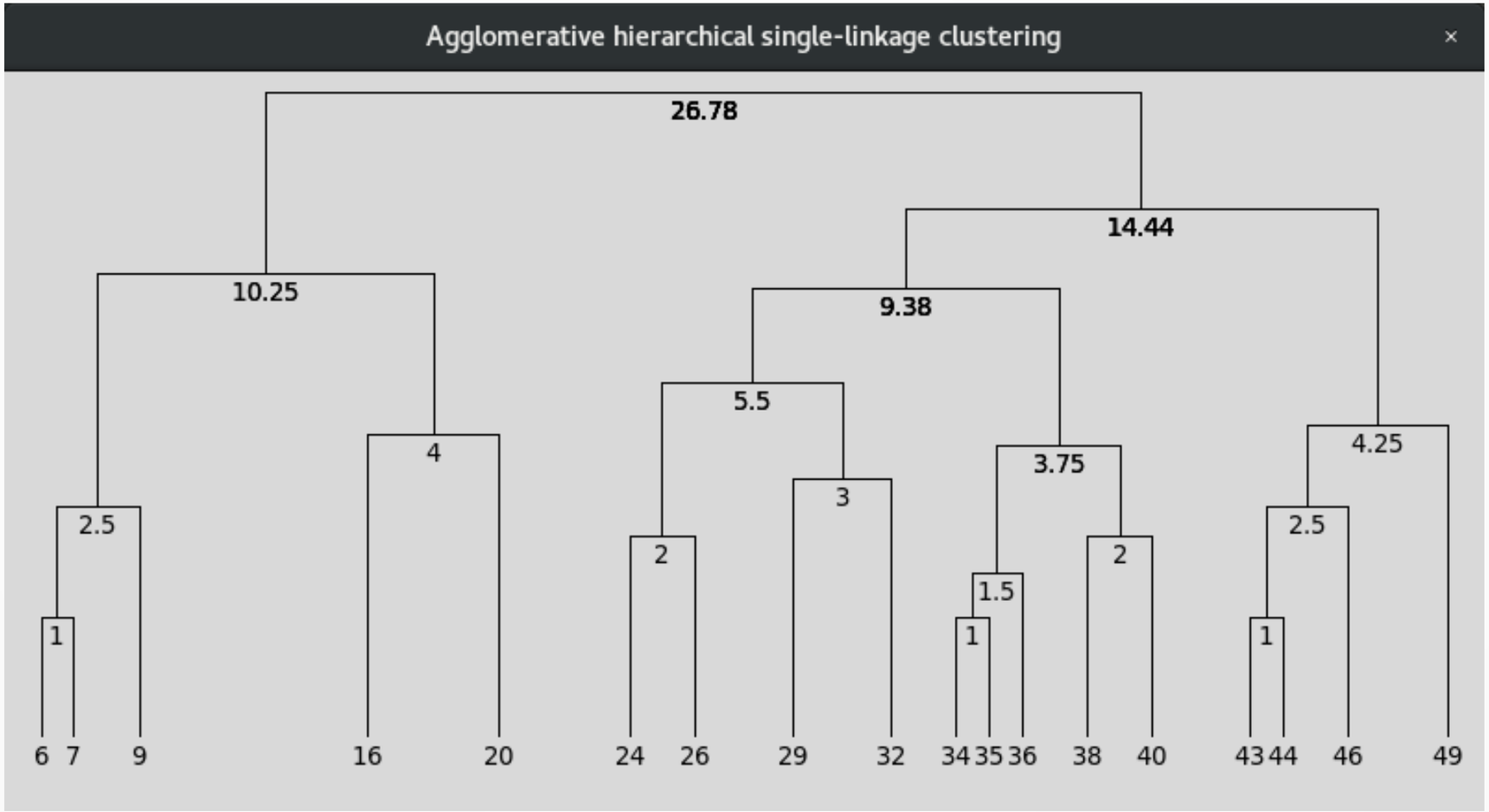
Решение предоставил
Шамов Илья

VK: vk.com/arqtty

Git: github.com/ARQtty

Визуализация работы алгоритма

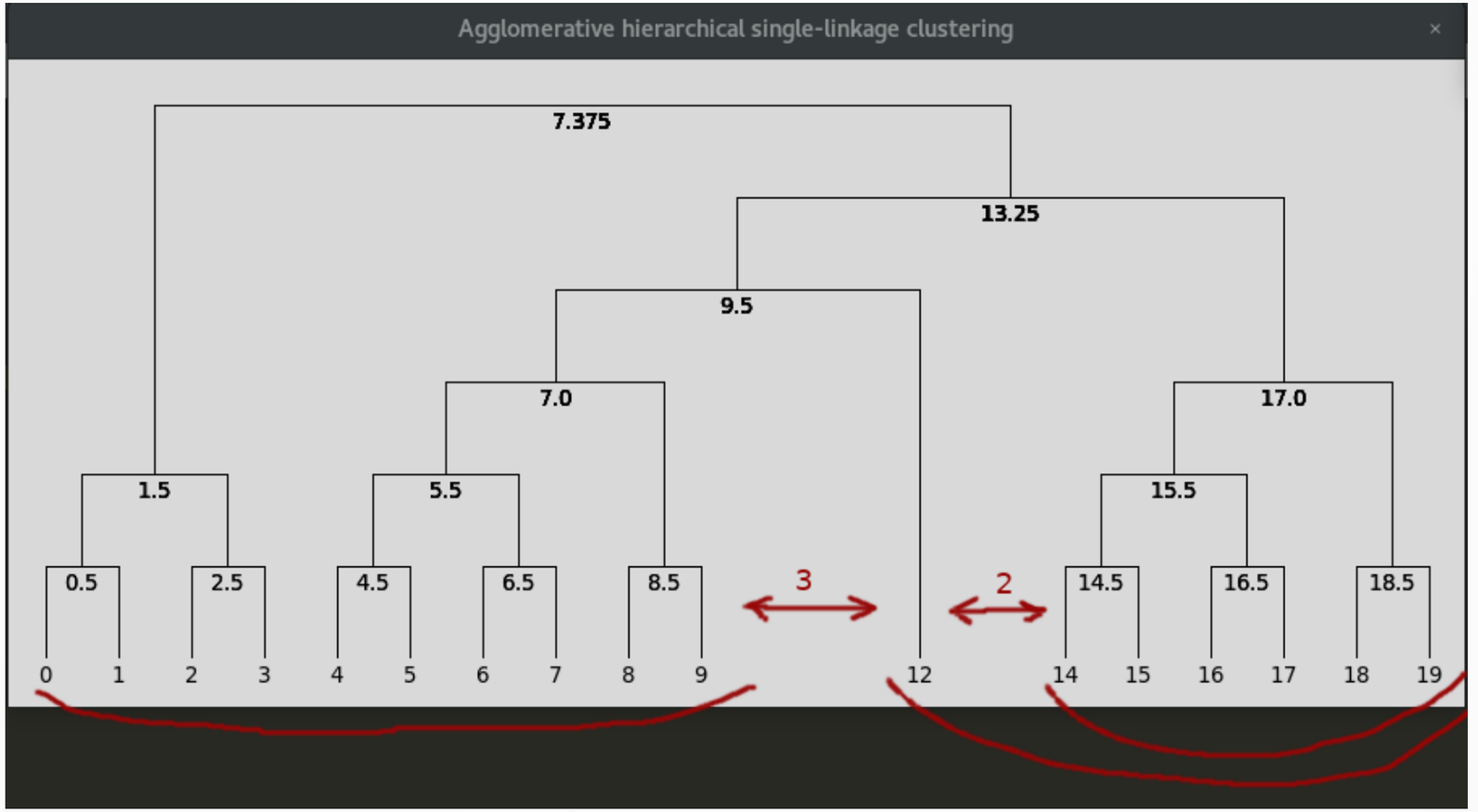
Результатом работы алгоритма является дендрограмма, представляющая иерархию кластеров исходных данных. На нижнем уровне представлены исходные данные (координаты точек на прямой). В родительских узлах дерева указывается расстояние между листьями



Недостатки метода кластеризации

Chaining и влияние шумов

При равномерном распределении данных внутри кластеров кластер может “заваливаться” на верхних уровнях в сторону, начиная присоединять к себе данные, которые скорее относятся к другому кластеру. Так уровень за уровнем chain-effect накапливается и приводит к неправильным (с точки зрения разбиения объектов на классы) результатам



В родительских узлах обозначена координата центра кластера

Описание решения

Входные данные: массив координат точек на прямой
Асимптотика времени: O(NlogN)
Асимптотика памяти: O(N)

Алгоритм:

- Для каждой точки вычисляется расстояние до следующей после неё
- Для каждой точки создаётся объект-узел, хранящий её индекс в исходном массиве данных, расстояние до следующей точки, центр кластера, которому она принадлежит (изначально – координата точки) и индексы левого и правого соседей
- Строится приоритетная очередь узлов с приоритетом по расстоянию до следущей точки
- Пока есть более чем один кластер
  - Берётся точка с min расстоянием до следующей и её пара (O(1))
  - Две точки объединяются в одну
  - Данные актуализируются (O(logN))
    - Обновляется расстояние от созданного кластера до следующей точки
    - Обновляется расстояние от предыдущей точки до нового кластера
    - Обновляются ссылки нового кластера на его соседей

Исходный код решения

github.com/ARQtty/single-linkage-clustering