Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет «Высшая школа экономики»

Факультет компьютерных наук
Образовательная программа «Науки о Данных»

О Т Ч Е Т
по проектной работе

Методы инициализации центроидов в методе к-средних

Выполнил студент гр. мНОД21_ИССА

Комарова Анна Сергеевна
(ФИО)

_____
(подпись)

Руководитель проекта:

Профессор Миркин Борис Григорьевич
(должность, ФИО руководителя проекта)

_____
(оценка)

_____ _____
(Дата) *(подпись)*

Москва 2023

Содержание

Общее описание проекта

Данный исследовательский проект представляет собой вычислительный эксперимент по сравнительному анализу способов инициализации центроидов в методе кластеризации k-средних, а также применение нового подхода по генерации Гауссовских кластеров.

Выполнение проекта

Ход выполнения проекта

Для проведения эксперимента мною были сгенерированы синтетические данные на основе метода, предложенного в статье [1]. Данный подход по генерации Гауссовских кластеров отличается от других подходов наличием всего одного параметра *a*, который позволяет управлять распределением как внутри кластеров, так и между ними. Центры кластеров генерируются из *M*-мерного равномерного распределения на гиперкубе $a[d_1, d_2]^M$, где *M* – количество признаков, d_1, d_2 – границы области, инициализируются $d_1 = -1, d_2 = 1$, а *a* – параметр пересечения кластеров, чем меньше *a,* тем выше взаимное перемешивание кластеров и тем более трудно разделимыми они будут.

В сравнительной части данного эксперимента внимание уделялось следующим подходам к инициализации центров кластеров:

1. Метод случайных точек. В этом подходе k случайных точек данных выбираются из набора данных и используются в качестве начальных центроидов. Этот подход очень изменчив и предусматривает сценарий, при котором выбранные центроиды не будут хорошо расположены во всем пространстве данных.
2. k-means++. В данном подходе первый центроид назначается случайно выбранной точкой данных, а последующие центроиды выбираются из оставшихся точек данных на основе вероятности, пропорциональной квадрату расстояния от ближайшего существующего центроида до данной точки. Таким образом, реализуется попытка отодвинуть центроиды как можно дальше друг от друга, покрывая как можно большую часть занятого пространства данных с момента инициализации.
3. Новый подход, предлагаемый в этой работе – метод панома. Алгоритм инициализации кластеров в предлагаемом методе заключается в следующем: генерация *2k* точек, где *k* -количество кластеров, из равномерного распределения на расстоянии от общего центра не менее, чем половина максимального расстояния (расстояние от общего центра до самой удаленной точки выборки). Далее из этого множества отбрасываются *k* точек, в условные кластеры которых попало меньше всего объектов выборки, а оставшиеся k становятся центроидами, которые далее будут использоваться в качестве инициализированных центров кластеров в алгоритме kmeans.

Результаты проекта

Сравнение подходов производилось по следующим параметрам: количество итераций до сходимости алгоритма и качество кластеризации на основе индекса ARI. Индекс ARI рассчитывает меру сходства между двумя кластеризациями путем рассмотрения всех пар образцов и подсчета пар, отнесенных к одинаковым или разным кластерам в предсказанной и истинной кластеризациях. Скорректированный индекс ARI будет иметь значение, близкое к 0.0 для случайной маркировки и ровно 1.0, когда кластеризации идентичны.

Ниже представлены результаты сравнения трех перечисленных подходов по инициализации центроидов для метода к-средних для разных уровней пересечения классов.

	a=1		*a*=0.75		*a*=0.5	
	n_inter	ARI	n_inter	ARI	n_inter	ARI
Метод случайных точек	12.9	0.712	14.6	0.637	22.8	0.472
k-means++	9	0.744	13.8	0.649	14.8	0.476
Паном	20.6	0.676	19.5	0.634	28.1	0.528

Таблица 1. Сравнение подходов инициализации центроидов в методе к-средних

Предложенный метод инициализации центроидов Паном уступает по скорости сходимости методам, реализованным в стандартных библиотеках, - метод случайных точек и k-means++, но показывает качество, близкое к эти методам при значениях показателя пересечения кластеров *a*=0.75 и даже превосходит упомянутые алгоритмы при значениях *a*=0.5.

Использованные методы и технологии

В данном проекте использовались следующие библиотеки: scikit-learn [2]. Код написан на Python3 и размещен в публичном репозитории [3].

Заключение

В результате данной работы я ознакомилась с новым методом генерации синтетических данных для задачи кластеризации, реализовала предложенный подход по инициализации центроидов.

Список литературы

1. Kovaleva, E. V., & Mirkin, B. G. (2015). Bisecting K-means and 1D projection divisive clustering: A unified framework and experimental comparison. *Journal of Classification*, *32*, 414-442.
2. https://scikit-learn.org/stable/
3. https://github.com/annakomarovabs/Practice