**Задание 1**

*Дано*

Данные хранятся в блоках памяти. Пользователь отправляет команды чтения данных из блоков в некотором порядке. Некоторые из этих команд можно объединить в кластеры, т.е. в группы команд в которых блоки часто читаются последовательно (друг за другом). Пусть уже известны 2 алгоритма кластеризации команд, но неизвестно как сравнить какой из алгоритмов для множества последовательностей команд.

*Пример*

Последовательность команд: 2 5 7 2 5 3 8 7 2 5 7 2 3 5 7 2 5 7

Алгоритм А1 может выделить кластер 2 5 7

Алгоритм А2 может выделить кластер 5 7 2 5

*Найти*

предложить метрику для оценки качества алгоритмов кластеризации в двух случаях:

a) ограничения на память и вычислительные ресурсы отсутствуют,

b) есть ограничения на память и вычислительные ресурсы.

(написать явно формулу и/или алгоритм вычисления метрики)

*Решение*

Среди метрик, используемых в кластерном анализе, можно выделить «вариации на тему» евклидова расстояния, т.е. метрики, которые так или иначе используют разность между координатами многомерных векторов.

В данном случае мы можем применить два подхода:

1) Применить классическую метрику Евклида, т.е. расстояние между двумя *n*-мерными векторами (кластером и фрагментом потока команд) определяется нормой разности этих векторов:

Тогда алгоритм кластеризации пытается минимизировать среднеквадратическую ошибку определения центра кластера **c** (*m* – число элементов в кластере):

Сравнение двух кластеров как раз можно осуществить, сравнив ошибки *Е*, однако такое сравнение необходимо привести к общему масштабу (в условии задачи кластеры определены над векторными пространствами разной размерности). С этой целью введем понятие относительного евклидового расстояния:

Тогда расстояния в пространствах разной размерности можно будет сравнивать. Также от размерности зависит число элементов в кластере. Для начала рассмотрим правило формирования элементов кластера из последовательности команд. Вектор заданной длины должен формироваться из последовательности команд на каждом дискретном такте поступления новой команды – поскольку совпадение с кластером может начаться в любой момент. Т.о. если всего команд *L*, то для заданной размерности векторного пространства *n* число элементов в кластере будет равно . Тогда логичным было бы усреднить ошибку по отношению к числу элементов кластера (за исключением текущего центра, т.е. ):

Для примера из условий задачи:

Элементы кластера А1: [**[2, 5, 7]**, [5, 7, 2], [7, 2, 5], [2, 5, 3], [5, 3, 8], [3, 8, 7], [8, 7, 2], [7, 2, 5], **[2, 5, 7]**, [5, 7, 2], [7, 2, 3], [2, 3, 5], [3, 5, 7], [5, 7, 2], [7, 2, 5], **[2, 5, 7]**]

Элементы кластера А2: [[2, 5, 7, 2], **[5, 7, 2, 5]**, [7, 2, 5, 3], [2, 5, 3, 8], [5, 3, 8, 7], [3, 8, 7, 2], [8, 7, 2, 5], [7, 2, 5, 7], [2, 5, 7, 2], [5, 7, 2, 3], [7, 2, 3, 5], [2, 3, 5, 7], [3, 5, 7, 2], **[5, 7, 2, 5]**, [7, 2, 5, 7]]

Жирным выделены текущие центры кластеров.

Тогда

Видно, что ошибка у кластера во втором случае меньше. Также можем заключить что сложность такого метода сравнения алгоритмов кластеризации возведений в квадрат.

Другим способом сравнения в евклидовой метрике является сравнение текущего центра **c** с центром масс кластера **s**. Каждая координата центра масс кластера вычисляется по формуле:

Затем идет сравнение центра масс и текущего центра, для упрощения вычислений можно сравнивать квадраты относительных расстояний

Для условий задачи

Расстояние от центра масс во втором случае меньше (т.е. второй алгоритм, как и в случае со среднеквадратичной ошибкой, показал лучший результат). Сложность вычислений в этот раз возведений в квадрат плюс затраты на вычисление центра масс, т.е. вычислительная сложность ниже.

2) Второй подход использует другую метрику. Поскольку в данном случае важно полное совпадение координат (речь идет о координатах в отдельности, а не векторе целиком). Если координаты не совпадают, то неважно на сколько – это уже просто другая команда. Тогда можно использовать метрику Хемминга, однако в кластерном анализе используют «Процент несогласия» - меру инвариантную к длине кластера (аналогично относительной евклидовой метрике):

Используя такую метрику можем произвести сравнение лишь на основе среднеквадратичной ошибки.