Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ (ТУСУР)

Кафедра автоматизации обработки информации (АОИ)

Вариант 3

Отчёт к Лабораторной работе №3

по дисциплине «Параллельные вычисления и системы»

Выполнил:

Студент группы 422–M1

\_\_\_\_Белоус Г.В.

Принял:

К.т.н., Доцент кафедры АОИ

\_\_\_\_Аксёнов С.В.

Оглавление

[1 Введение 3](#_Toc116226139)

[1.1 Цель лабораторной работы 3](#_Toc116226140)

[1.2 Задание на лабораторную работу 3](#_Toc116226141)

[2 Ход выполнения работы 4](#_Toc116226142)

[3 Вывод 7](#_Toc116226143)

# Введение

## 1.1 Цель лабораторной работы

Изучить инструменты MPI и написать программу выполняющую кластеризацию данных из предоставленной выборки алгоритмом K-Means, в каждом потоке должен запускаться алгоритм кластеризации для разного количества кластеров и в итоге оцениваться по необходимой метрике.

## **1.2 Задание на лабораторную работу**

Обучающая выборка (DS\_2019\_public.xlsx) загружается одним из процессов. Значения атрибутов привести к диапазону от 0 до 1. Далее этот процесс отправляет загруженные данные всем остальным процессам. Все процессы начинают запускают несколько раз алгоритм кластеризации данных алгоритмом K-Means (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html). Каждый раз при этом вычисляется значение индекса качества. Среди решений выбирается лучшее (это решение имеет либо самый большой, либо самый маленький индекс – смысл выводится из уравнения индекса). Далее все лучшие индексы процессов сравниваются и процесс с лучшим результатом выводит свой результат.

Содержание отчета:

1. Титульный лист
2. Задача по полученному варианту.
3. Вариант 3 (последняя цифра в зачетке 2 или 9) RS индекс в задании использовать функции MPI\_Isend, MPI\_Gather
4. Листинг программы
5. Примеры входных и выходных файлов
6. Описание результатов тестирования
7. Выводы

# Ход выполнения работы

Программа для обработки данных была написана на Python с использованием платформа Google Colaboratory для блокнотов Jupyter. Кроме среды запуска блокнотов Python и R Colab позволяет совместно использовать свободный доступ к ограниченному количеству GPU и TPU.

Для работы с данными были подключены необходимые фрейморки mpi4py, предоставляющий основные инструменты MPI на языке Python, и sklearn, позволяющий использовать алгоритм K-Means. Так же, в соответствии с вариантом, данные передавались с помощью функций MPI\_Isend, MPI\_Gather.

MPI\_Send блокирующая, в отличии от неблокирующей функции MPI\_Isend, после отправки данных полностью останавливают выполнение программы до тех пор, пока отправленное сообщение не будет получено. Соответственно, вторая не останавливают исполнение программы.

При выполнении операции сборки данных MPI\_GATHER каждый процесс, включая корневой, посылает содержимое своего буфера в корневой процесс. Корневой процесс получает сообщения, располагая их в порядке возрастания номеров процессов.

Листинг программы можно увидеть в прикрепленном файле “PC\_Lab3.ipynb”.

Исходя из задания, для классификации с помощью алкоритма K-Means, была написана программа использующая необходимые функции MPI, входные данные из колонок KWHCOL и BTUELCOL были нормализированы. Оценить качество кластеризации необходимо было с помощью RS(R Squared), высчитывающийся по формуле 2.1:

(2.1)

где – сумма квадратов расстояний внутри кластера;

– сумма квадратов расстояний по всему множеству;

– сумма квадратов расстояний по всему множеству, .

Работа программы со значениями центра кластеров для различного количества и массив со значениями RS индекса показаны на рисунке 2.1:

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 2.1 – Результат работы программы

Точность предсказаний занесена в таблицу 2.1:

Таблица 2.1 – RS индекс для результатов расчета по алгоритму K-Means с разным количество кластеров.

|  |  |
| --- | --- |
| Количество кластеров | RS индекс |
| 2 | 0.001 |
| 3 | 0.011 |
| 4 | 0.057 |
| 5 | 0.112 |
| 6 | 0.772 |
| 7 | 0.851 |

# Вывод

В ходе выполнения лабораторной работы изучены инструменты MPI и написана программа выполняющую кластеризацию данных из предоставленной выборки алгоритмом K-Means, в каждом потоке запускается алгоритм кластеризации для разного количества кластеров и в итоге оцениваться по необходимой метрике.

Принцип работы программы заключается в кластеризации данных алгоритмом K-Means для различного количества кластеров, расчёт метрики по формуле 2.1 в каждом потоке и отправки значения неблокирующей функции MPI\_Isend и последующей сборкой данных с помощью функции MPI\_Gather/

На основе полученных результатов метрик из таблицы 2.1 можно сделать вывод о том, что в случае с выбором количества кластеров от двух до семи, максимальным индексом будет в случае с наибольшем количеством кластеров, так как RS индекс считает расстояния. Однако, можно заметить, что разница индекса для пяти и шести кластеров различается примерно в семь раз.