CSYE7105 - HIGH PARALLEL MACHINE LEARNING & AI

PROJECT PROPOSAL

TEAM 21

Asteroid Data Classification and Prediction Through Parallel Computing

MEMBERS:
Akshit Hasmukh Kumar Jain
Uday Kiran Dasari

Introduction

Background

The Asteroid Dataset Classification and Prediction project aims to analyze a comprehensive dataset containing various attributes of known asteroids. By utilizing advanced machine learning algorithms, classification of different types of asteroids based on their properties and predict their potential trajectories and behavior is extremely helpful. Through the application of data-driven models, we can enhance our understanding of the unique characteristics of these space objects and their likelihood of impacting Earth. The insights derived from such research can assist space agencies, researchers, and policymakers in devising effective strategies for planetary defense and risk mitigation.

Motivation

Asteroids, being remnants of the early solar system, carry vital information about its formation and evolution. Moreover, their potential threat to Earth necessitates a comprehensive understanding of their characteristics and behavior. By leveraging advanced data analysis techniques, we can enhance our ability to predict and classify the trajectory and properties of asteroids, thus enabling better preparedness and mitigation strategies. This project aims to contribute to the ongoing efforts to safeguard our planet from potential asteroid impacts and to deepen our knowledge of the cosmos.

Goal

The primary goal of this project is to develop a model for asteroids classification using parallel computing methods for pre-processing and model training for classifying asteroids based on their properties. By employing Dask and other parallel methods, we aim to improve the efficiency of the model enabling faster classification of asteroids. The objective is to compare the speed-up performance achieved through parallel processing on various hardware configurations, aiming to enhance the predictive accuracy and efficiency of asteroid classification models.

The project involves data preprocessing, exploratory data analysis, feature engineering, model development, and performance evaluation. By harnessing the power of machine learning, this project endeavors to contribute to the broader field of planetary science and space exploration, reinforcing our ability to protect our planet from potential celestial threats.

The successful execution of this project can pave the way for the development of more robust and accurate predictive models, thereby bolstering our capabilities in understanding and mitigating the potential risks associated with asteroids. Additionally, the project's findings may provide valuable insights for future space missions, enabling better planning and decision-making for exploration and resource utilization beyond Earth's orbit.

Methodology

1. Data preprocessing and cleaning:

- Extract relevant data focusing on the 45 quantitative parameters for each asteroid.
- Utilize parallel processing with dask to distribute these tasks across multiple cores or nsodes for tasks like data extraction, cleaning, and normalization.

2. EDA analysis:

- Perform exploratory data analysis (EDA) using histograms to visualize variable distributions and identify outliers.

3. Data preprocessing and cleaning:

- Evaluate ML models on the asteroid dataset. Utilizing different machine learning approaches such as Decision Trees, Random Forests, Support Vector Machines, XGBoost and Dense models for performing classification and prediction.
- Employ Dask-ML parallel processing for model training and evaluation to expedite the process. Utilize frameworks like TensorFlow or PyTorch that support parallel computation on GPUs for faster model training.

4. Scope for Parallelization:

- Identify computationally intensive tasks suitable for parallel processing.
- Implement parallelization using Dask and multiprocessing methods to distribute tasks across multiple cores or nodes efficiently.

5. Performance Evaluation:

- Measure the execution time of critical tasks in both serial and parallel implementations.
- Compare the speed-up performance achieved and efficiency by different parallel methods on different hardware configurations by using multiple CPUs/GPUs.

6. Analysis and Visualization:

- Analyze the speed-up results obtained from parallel processing compared to the non-parallel approach.
- Visualize the performance gains using graphs or charts to present the efficiency improvements clearly.

7. Optimization and Iteration:

- Fine-tune the parallel implementation by optimizing task distribution and resource allocation.
- Iterate on the methodology based on performance insights to further enhance efficiency.

Description of the Dataset

This dataset comprises of 958524 records, which are observations related to asteroids. Each record contains a unique asteroid with its associated data such as orbital parameters, physical properties etc. It features a collection of 45 quantitative parameters for each asteroid. These parameters include eccentricity e, semi-major axis a, perihelion distance q, inclination i, absolute magnitude H, diameter among many other characteristics. The dataset's diversity in variables allows a detailed and nuanced modelling approach. This dataset sourced from Kaggle, serves an excellent resource in honing our skills in using ML models and analysing the performance of these models when considering parallelism, offering a real-world context for Parallel Processing. It is a classification problem since we need to categorize asteroids based on the features.

Data Sources: https://www.kaggle.com/datasets/sakhawat18/asteroid-dataset/data

References:

Hossain, Mir Sakhawat & Zabed, Md. (2023). Machine Learning Approaches for Classification and Diameter Prediction of Asteroids dx.doi.org/10.1007/978-981-19-7528-8_4 Accessed 12 March 2024.