Guess the Topic

Parallelized Grid Search Optimization for Water Quality Classification

ARTI 503 Parallel AI7-1

Agenda

- Introduction
- Dataset
- Prepossessing
- Machine Learning Models
- parallelized Grid search optimization
- Training run time comparison
- Speed Up and Efficiency (Parallel)
- Accuracy

Introduction

Water is a fundamental and indispensable resource for life on Earth, playing a crucial role in sustaining various ecosystems and supporting human well-being

Abrar Sebiany

Traditional methods used a sequential approach for water quality classification

- Expensive
- Time consuming
- Inefficient

Abrar Sebiany

The solution is

Introduce the parallel approach for water quality classification

- Cost-Effectiveness
- Increased Speed
- Efficient

Abrar Sebiany

Our Contribution

Parallelized Grid Search Optimization for Water Quality Classification

Abrar Sebiany

Grid Search Optimization

A technique used in machine learning to find the optimal hyperparameters for a model

Abrar Sebiany

Machine Learning Algorithms

Three algorithms have been chosen to select the most suitable algorithm for the project

GRADIENT BOOSTING

SUPPORT VECTOR MACHINE (SVM)

RANDOM FOREST

Abrar Sebiany

Dataset

Obtained from

Kaggle

Number of Samples

3276

Classified as

0 Non-Potable

1 Potable

Number of Features

9

Manar Alsayed

Prepossessing

Filling the missing values with the median

Balancing data using SMOTE-Tomek

Reason

To avoid biased

Manar Alsayed

Machine Learning Models

GRADIENT BOOSTING

- Combines several weak learners into strong learners, in which each new model is trained to minimize the loss function.
- **Hyperparameters:**

```
gb_params = {
    'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 1],
    'Max_depth': [3, 5, 7], 'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

SUPPORT VECTOR MACHINE (SVM)

- Find a hyperplane that best separates a dataset into classes.

Hyperparameters:

```
svm_params = {
    'C': [0.1, 1, 10], 'kernel': ['linear', 'rbf'],
    'gamma': ['scale', 'auto'], 'degree': [1,2, 3, 4], 'coef0': [ 0.5, 1.0]
}
```

RANDOM FOREST

- Most commonly used.
- It combines the output of multiple decision trees to reach a single result.

Hyperparameters:

```
rf_params = {
    'n_estimators': [50, 100, 200],'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5], 'min_samples_leaf': [1, 2, 4],
    'max_features': ['auto', 'sqrt', 'log2']
}
```

Joury Alzayat

Parallelized Grid Search Optimization

First:

parallelize the Grid search process using (n_jobs=-1)

Joury Alzayat

Parallelized Grid Search Optimization

First:
parallelize the Grid search process using (n_jobs=-1)

Second:
Parallelize the training models using Parallel() and delayed() from Joblib library:

- Parallel (n_jobs=-1):

which will specify the number of cores that we will use.

- delayed(GridFunction):

create a lazy or deferred function call

Joury Alzayat

Training Run Time Comparison

	Training Run Time (Parallel)		Training Run Time (Non-Parallel)	
	8 cores	16 cores	8 cores	16 cores
Random Forest (RF)	110.403 s	34.202 s	267.387 s	174.415 s
Support Vector Machine (SVM)	5038.879 s	1243.926 s	16180.200 s	5760.418 s
Gradient Boosting (GD)	179.574 s	52.636 s	459.292 s	288.507 s

Noor Aljishi

Speed Up and Efficiency (Parallel)

	Speed Up		Efficiency	
	8 cores	16 cores	8 cores	16 cores
Random Forest (RF)	2.421	5.099	0.302	0.318
Support Vector Machine (SVM)	3.211	4.630	0.401	0.289
Gradient Boosting (GD)	2.557	5.481	0.319	0.342

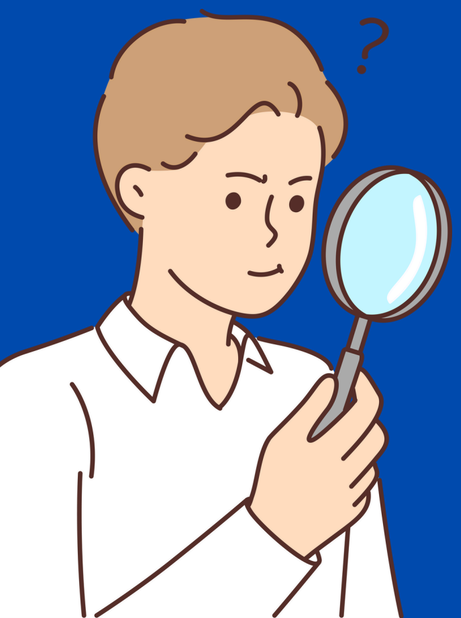
Noor Aljishi

Accuracy

	Accuracy (Parallel)		Accuracy (Non-Parallel)	
	8 cores	16 cores	8 cores	16 cores
Random Forest (RF)	77.1	76.3	76.2	74.6
Support Vector Machine (SVM)	60.9	60.9	60.9	60.9
Gradient Boosting (GD)	76.5	76.2	73.9	74.6

Noor Aljishi

Any Questions?



Thank you for listening

ARTI 503 Parallel AI7-1

Team Members:

- **Abrar Sebiany**
- **Manar Alsayed**
- **Joury Alzayat**
- **Noor Aljishi**