26th International Conference on Computer and Information Technology (ICCIT 2023)
Cox's Bazar, Bangladesh

Ameliorating Performance of Random Forest using Data Clustering

Paper ID: 463

Author and Presenter

Ummay Maria Muna

Supervisor

Prof. Dr. Dewan Md. Farid

Co-authors

Shanta Biswas

Syed Abu Ammar Muhammad Zarif

Introduction

Random Forest and K means Clustering Both are excellent method, **but has its own drawbacks**

To increase the accuracy of a Random Forest model, it **requires more trees**, which results **slow model**

Also, the higher the number of the trees, the more **difficult the description of information** is

To solve this issue, we combined the K-means clustering with Random Forest

Motivation

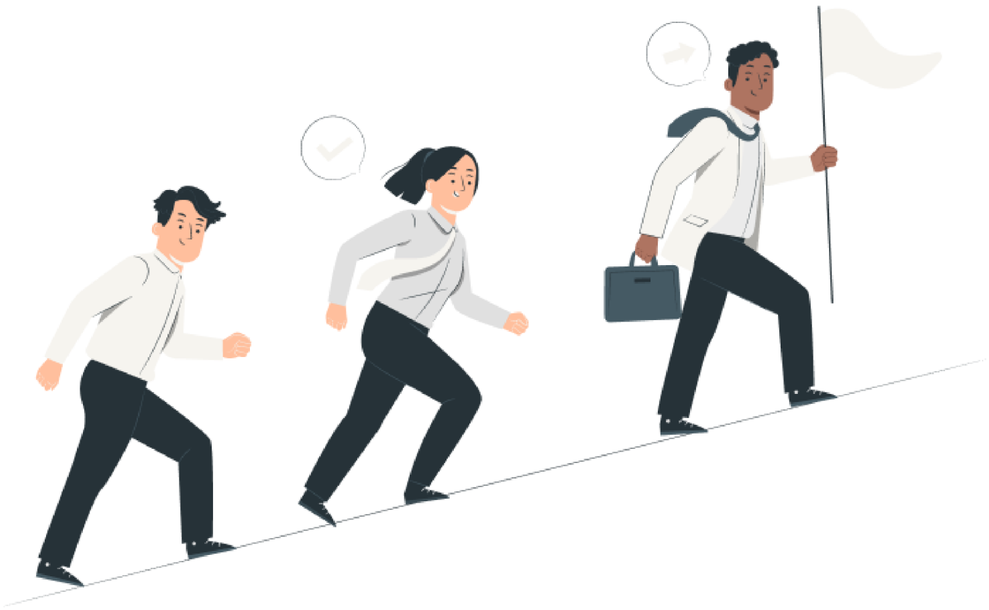
Enhancing Random Forest **Performance**

Noise and **Outlier separation** through Clustering

Improved Data **Homogeneity**

Presents more **structured** and **refined** data subset for modeling

Detecting **Localized Pattern**



Related Works

Author(s)	Works
Arafat et al.	Combined **under-based sampling** and **Random Forest** to balance multi-class imbalanced data and improve classification accuracy.
Jie et al.	Utilized **K-Means clustering** and **Random Forest** to enhance gas content prediction accuracy in CBM reservoirs with weak correlations between parameters.
Hojin et al.	Developed an algorithm using **high-dimensional data** and **ensembles of classifiers** generated from an **optimal number** of **random partitions** within the feature space.
Farid et al.	Clustered high-dimensional biological big data using an **ensemble clustering strategy** incorporating **feature selection** and **grouping mechanisms**.
Evgeny et al.	Induced **decision trees**, providing a **global search across the solution space** since the traditional Decision Tree algorithms often struggle to achieve optimal results
Farid et al.	Applied **optimal adaptive ensemble learning** on High-Dimensional imbalance genomic data.

Datasets

Diabetes Dataset

- Data about **Diabetic Patients**
- **Numeric Data**
- Over **1000+** Samples
- **9 columns** including pregnancies, glucose, blood pressure, insulin etc.

Cardiovascular Disease

- Data about **Cardiovascular related diseases**
- **Numeric Data**
- Over **2000** samples
- **9 columns** including pregnancies, glucose, blood pressure, insulin etc.

Heart Dataset

- Categorized Data about **Patients of Heart Disease**
- **Numeric Data**
- Over **300+** Samples
- **14 columns** including age, sex, cholesterol, blood pressure etc.

Heart StratLog

- Data about **Reading of ECG from different patients**
- **Numeric Data**
- About **1000** Samples
- **14 columns** including age, sex, diastole rate, systole rate etc.

Terminologies

A Clustering

Unsupervised Learning where **similar data** points are **grouped together**

B Ensembling

Instead of relying on a single model, ensembling **leverages the predictions of several models** to improve overall accuracy and robustness.

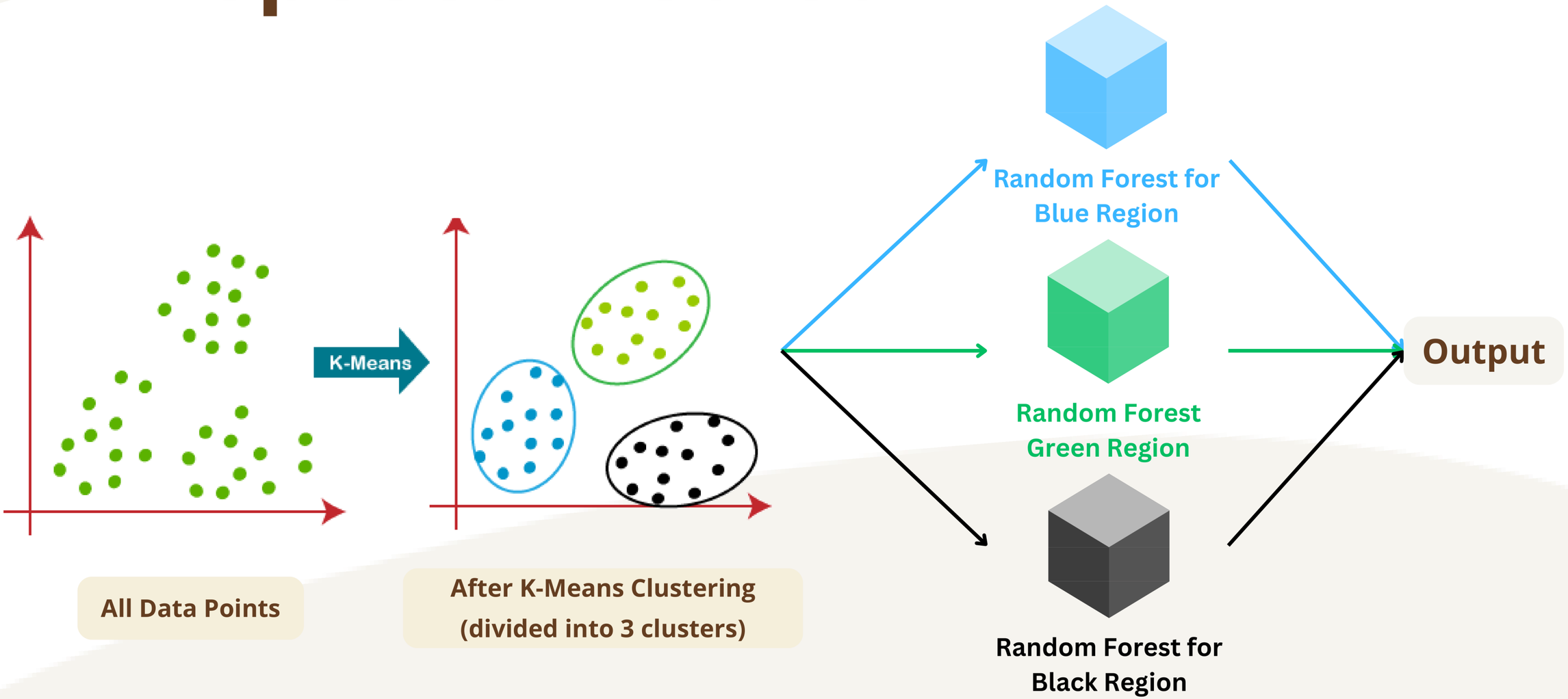
B1 Random Forest

Collection of decision trees, Each tree makes predictions based on the data it receives. Two key ingredients: Bagging and Random Subspace Method

B2 Random Subspace Method

By considering only **a random subset** of features at **each split**, trees become more **diverse** and **less prone** to overfitting.

Proposed Method



Strategy

Define Cluster number

Choosing an optimized number of Cluster suitable for the dataset using elbow method

Cluster the entire Dataset

Running K-means Clustering to cluster the entire dataset

Random Forest on each Cluster

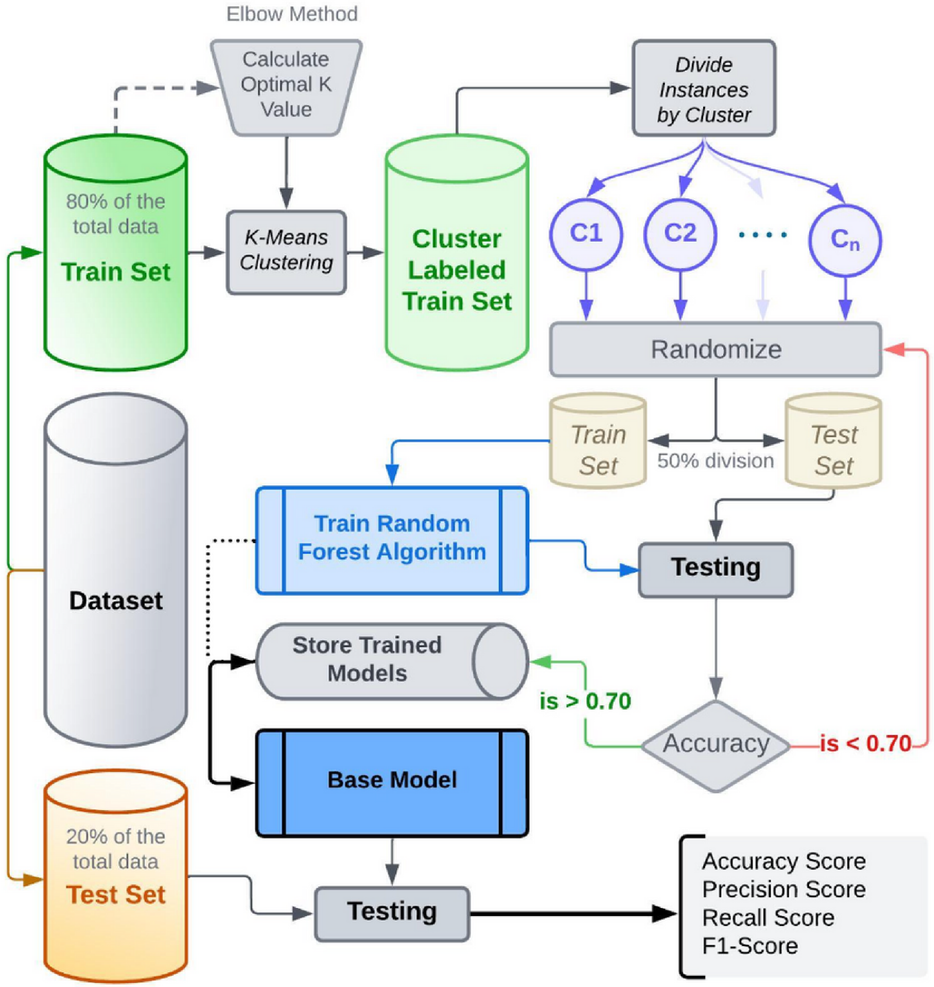
Running Random Forest using Random Subspace method and obtaining the results by majority voting

Choosing RF with >70% accuracy

Not let a RF pass until it gets above 70% accuracy. It is restrained in another iteration with a different train set.

Workflow

- Find optimal **K value** followed by **clustering the 80%** of total dataset (train set) and **store the fit model (C) and each clustered data** separately.
- Fit **random forest** model for **each cluster** with **randomized data** points each time and pick the models that perform outstanding result. **Store those models (R)**
- With saved model(s) **C** and **R**, we can now test the model with the **leftover 20%** (test set) of the total dataset and find metrics.
- With **C** and **R**, we have our proposed model



Results

PERFORMANCE METRICS FOR CLASSICAL RANDOM FOREST

Datasets	Accuracy	Precision	Recall	F1-Score
Diabetes	0.72	0.61	0.62	0.61
Breast Cancer	0.96	0.94	0.94	0.94
Heart Disease	0.85	0.87	0.89	0.88
Cardio	0.77	0.87	0.7	0.78
Heart Statlog	0.8	0.78	0.67	0.72

TABLE II

PERFORMANCE METRICS FOR PROPOSED MODEL

Datasets	Accuracy	Precision	Recall	F1-Score
Diabetes	0.76875	0.69231	0.63158	0.66055
Breast Cancer	0.96522	0.97561	0.93023	0.95238
Heart Disease	0.88587	0.90196	0.89320	0.89756
Cardio	0.74449	0.83333	0.67460	0.74561
Heart Statlog	0.77533	0.79121	0.69231	0.73846

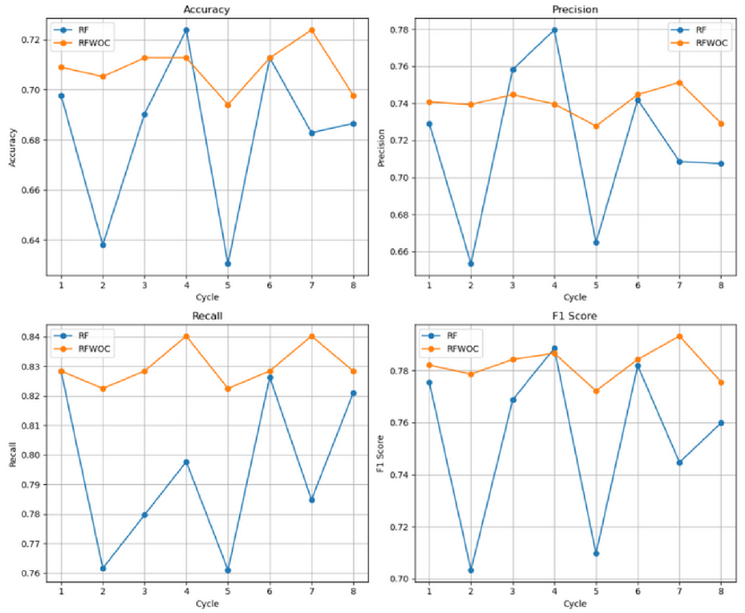


Fig. 8. Performance comparison on nba-logreg Dataset on 8 random runs.

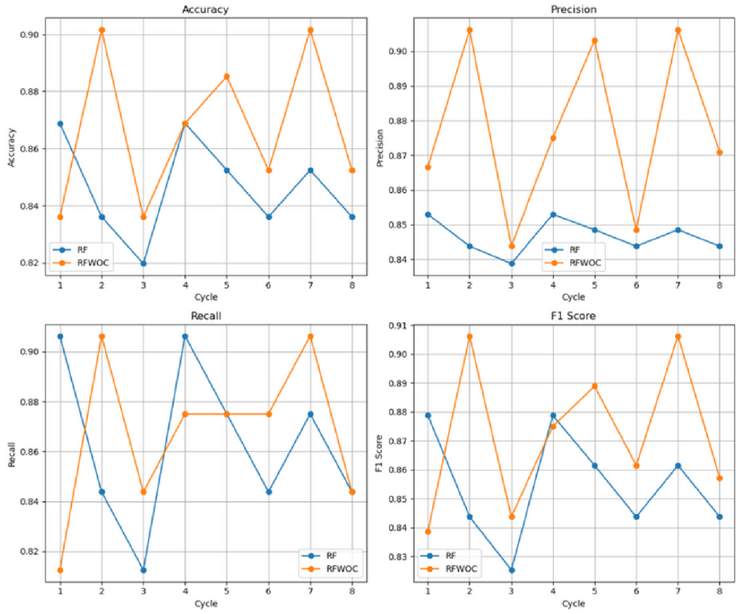


Fig. 7. Performance comparison on Heart Disease Dataset on 8 random runs.

RF (Traditional)

RFWOC

Analysis

- Performance of the model on all the datasets were promising except one which is cariovascular dataset.
- Datasets should be normalized and well defined for the model
- Different run gives different result, however RFWOC give the higher accuracy.
- Since the datasets are suffled two times, means the model is not biased.

Findings

Dataset with less instances does not perform well in this approach

Take longer time to run when a RF struggles to achieve more than 70% accuracy after multiple iterations

For some particular dataset performance differences are negligible

End of Presentation

Thank you

Feel free to ask questions

IEEE
Advancing Technology for Humanity

26th International Conference on Computer and Information Technology (ICCIT 2023)
Cox’s Bazar, Bangladesh