IOWA STATE UNIVERSITY

Department of Industrial and Manufacturing System Engineering

Midwest Big Data Hub

October 29-30, 2019

Fatemeh Amini[1], Mohsen Shahhosseini[2], Guiping Hu[3], and Hieu Pham[4]

Improving Prediction Accuracy of Regression Problems with Optimization-based Ensemble Learning and a Two-layer Feature Selection Method

Introduction

This study proposes two state-of-art optimization-based methodologies to improve prediction accuracy for regression problems.

1. The problem of ultra-high-dimensional datasets, in which the number of predictors exceeds the number of observations, is studied and a hybrid two-layer optimization-based model using Genetic Algorithm (GA) and Elastic Net is proposed. This optimization model considers minimizing prediction RMSE and number of selected predictors using GA with Elastic Net as its fitness function, in the first layer. In the second layer, the best subset of predictors is used to apply simple Elastic Net on, intending to eliminate more predictors.
2. Aggregating multiple learners through an ensemble of models aims to make better predictions by capturing the underlying distribution more accurately. We considered blending as one type of ensemble creating method and designed an optimization-based ensemble learning algorithm that not only intends to reduce variance, but also aims at decreasing the prediction bias. To this end, a bi-level optimization-based algorithm that considers tuning hyperparameters as well as finding the optimal weights to combine ensembles was proposed.

Methodology

Two-layer Feature Selection Method

$$min_{\hat{y},n_p} \quad = w_R * r_{RMSE} + w_p * n_p$$

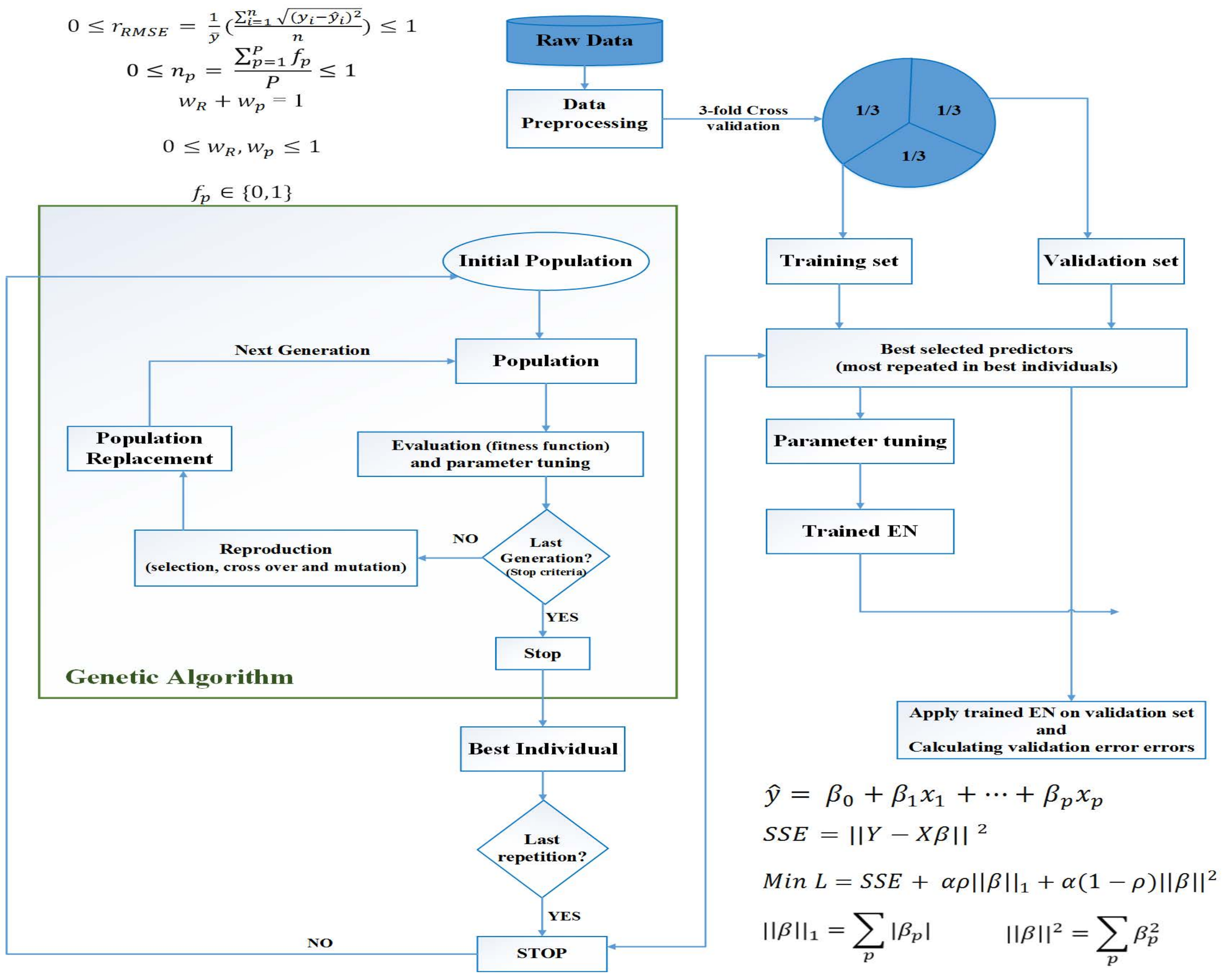
$$0 \le r_{RMSE} = \frac{1}{\bar{y}}\left(\frac{\sum_{i=1}^{n}\sqrt{(y_i-\hat{y}_i)^2}}{n}\right) \le 1$$

$$0 \le n_p = \frac{\sum_{p=1}^{P} f_p}{P} \le 1$$

$$w_R + w_p = 1$$

$$0 \le w_R, w_p \le 1$$

$$f_p \in \{0,1\}$$


Raw Data
Data Preprocessing
3-fold Cross validation
1/3
1/3
1/3
Training set
Validation set
Initial Population
Next Generation
Population
Population Replacement
Evaluation (fitness function) and parameter tuning
Reproduction (selection, cross over and mutation)
NO
Last Generation? (Stop criteria)
YES
Stop
Genetic Algorithm
Best Individual
Last repetition?
YES
NO
STOP
Best selected predictors (most repeated in best individuals)
Parameter tuning
Trained EN
Apply trained EN on validation set and Calculating validation error errors

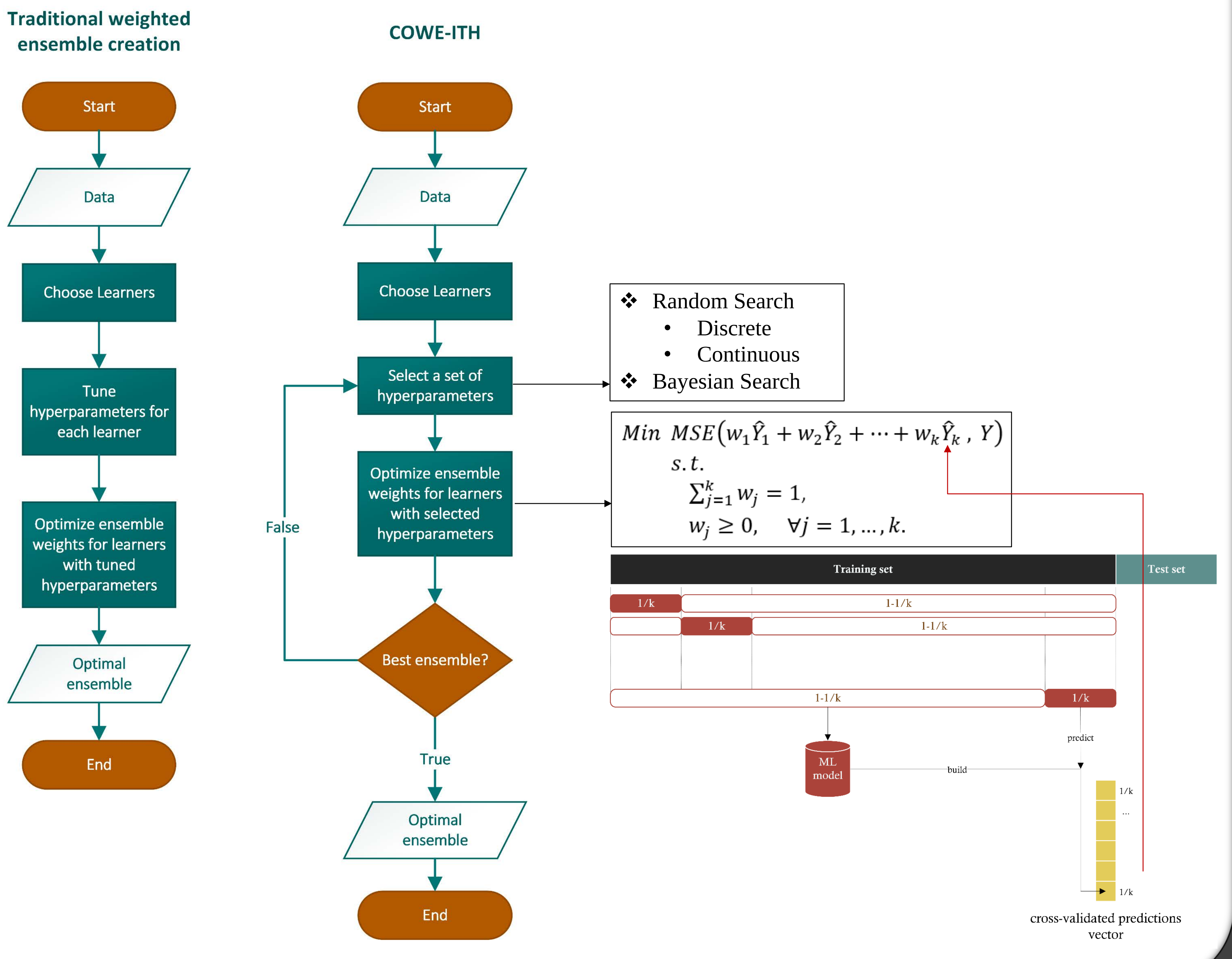
$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$SSE = ||Y - X\beta||^{2}$$

$$Min\ L = SSE + \alpha\rho||\beta||_1 + \alpha(1-\rho)||\beta||^2$$

$$||\beta||_1 = \sum_{p} |\beta_p| \qquad ||\beta||^2 = \sum_{p} \beta_p^2$$

Optimizing Ensemble Weights and Hyperparameters of Machine Learning Models


Traditional weighted ensemble creation
Start
Data
Choose Learners
Tune hyperparameters for each learner
Optimize ensemble weights for learners with tuned hyperparameters
Optimal ensemble
End
COWE-ITH
Start
Data
Choose Learners
Select a set of hyperparameters
Optimize ensemble weights for learners with selected hyperparameters
False
Best ensemble?
True
Optimal ensemble
End
Training set
Test set
1/k
1-1/k
1/k
1-1/k
1-1/k
1/k
predict
ML model
build
1/k
...
1/k
cross-validated predictions vector

- ❖ Random Search
  - Discrete
  - Continuous
- ❖ Bayesian Search

$$Min\ MSE(w_1\hat{Y}_1 + w_2\hat{Y}_2 + \cdots + w_k\hat{Y}_k\ , Y)$$
$$s.t.$$
$$\sum_{j=1}^{k} w_j = 1,$$
$$w_j \ge 0, \quad \forall j = 1, \ldots, k.$$

Experimental Results

Gene ID	Gene Shape Ratio	α_{best}	ρ_{best}	w_R	w_p	FSP
32	1.33	0.017	0.36	0.15	0.85	0.3
37	9.074	0.457	0.36	0.85	0.15	0.3
80	2.89	30.05	0.36	0.15	0.85	0.3
86	3.18	0.435	0.3	0.85	0.15	0.5
89	2.44	3.38	0.23	0.85	0.15	0.3
94	7.33	0.21	0.36	0.15	0.85	0.3
107	3.66	9.74	0.36	1	0	0.3
178	2.63	3.31	0.43	1	0	0.3
181	10.62	3.08	0.1	1	0	0.7
187	2.33	0.56	0.63	1	0	0.5

Model validation

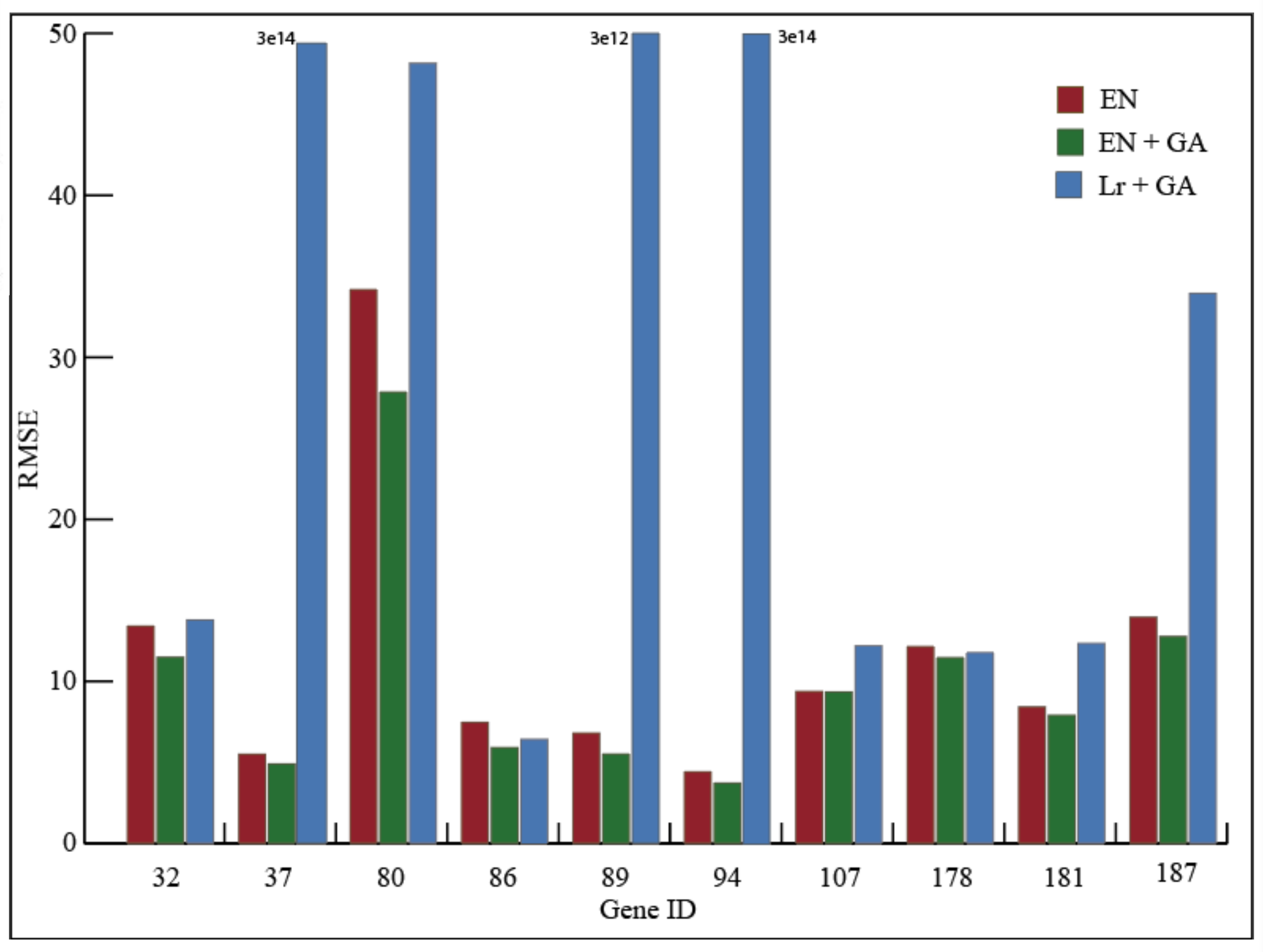
Benchmarks:

- Linear regression + GA (Lr + GA)
- Elastic Net (EN)

Results show that The hybrid Elastic Net-Genetic Algorithm method outperforms Linear Regression-GA and Elastic Net methods in terms of prediction error (RMSE) in predicting the RNA-seq of Maize plant.

Hyper-Parameter Tuning

- All tuned parameters of GA and Elastic Net have been used for final prediction error calculation.
- Gene Shape Ratio = $\frac{\#\ of\ Predictors}{\#\ of\ Observations}$
- FSP (Fraction of Selected Predictors): How many times a specific predictor should be displayed in GA's output to be included in final best subset



	Data sets	Number of Instances	Number of Attributes	Area
1	Airfoil Self-Noise	1503	6	Physical
2	Auto MPG	398	8	Automobiles
3	Boston Housing	506	14	Housing
4	Concrete Compressive Strength	1030	9	Physical
5	Diabetes Data	442	10	Life
6	Energy efficiency	768	8	Computer
7	Forest Fires	517	13	Physical
8	Graduate Admissions	500	9	Education
9	Wine Quality	4898	12	Business
10	Yacht Hydrodynamics	308	7	Physical

Model settings

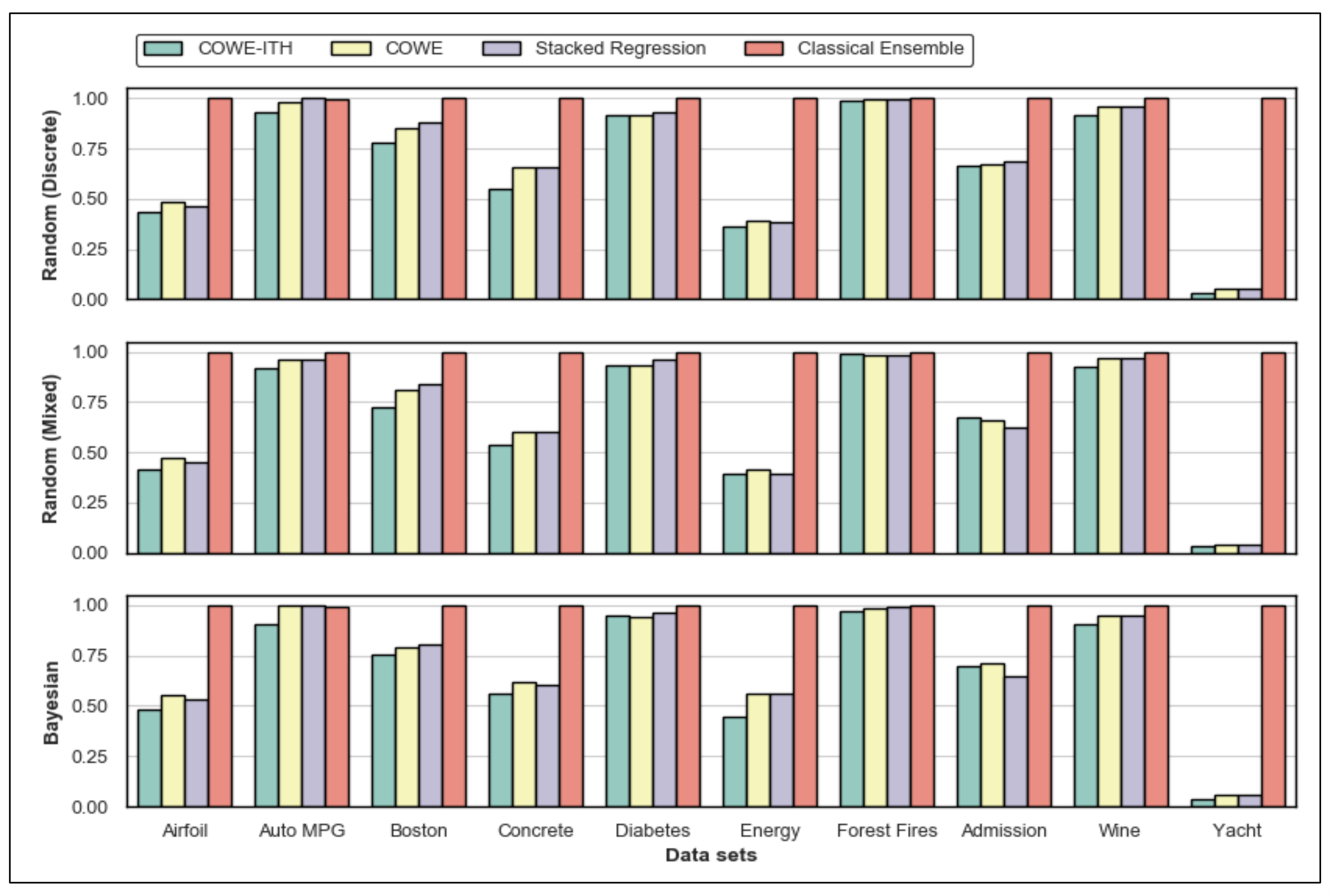
- Four ML algorithms with minimal pre-processing:
  1. LASSO
  2. Random forest
  3. XGBoost
  4. SVM (rbf kernel)
- Similar hyperparameters settings
- 5-fold cross-validation
- Entire process was repeated 5 times
- 20% test set, 80% training set
- **COWE-ITH parameter**: *number of iterations*: closely related to the bias-variance tradeoff. large values vs. small values.

Benchmarks

- Cross-validation Optimal Weighted Ensemble (COWE)
- Classical ensemble
- Stacked regression ensemble

Results

- COWE-ITH is superior than base learners (9/10)
- COWE-ITH outperforms benchmarks (9/10)
- Different hyperparameter values (non-optimal)



Conclusion

- Datasets with high ratio of number of predictors to number of observations are prone to overfitting and single-layer feature selection methods usually are not able to eliminate all irrelevant predictors thus, leading to high prediction error.
- The proposed two-layer feature selection method reduces the number of predictors while maintaining the prediction accuracy.
- The results of applying the two-layer method on multiple Maize genes with various shape ratio show that it outperforms established benchmarks.
- A bi-level nested algorithm that finds the optimal weights to combine base learners as well as the optimal set of hyperparameters for each of them (COWE-ITH) was designed.
- Based on the obtained results, it was shown that COWE-ITH is able to dominate base learners as well as other ensemble creation methods.
- Furthermore, it was demonstrated that the hyperparameters used in creating optimal ensembles are different when they are tuned internally.

1. famini@iastate.edu, 2. mohsen@iastate.edu, 3.gphu@iastate.edu, 4. htpham@iastate.edu