

Paper 5135-2020

Evolutionary Feature Selection for Machine Learning

Nandini Rakala, Munevver Mine Subasi, and Ersoy Subasi, Florida Institute of Technology

ABSTRACT

We propose an evolutionary feature selection technique for the machine learning predictive modeling task, involving two conflicting goals of minimizing the number of features and maximizing the prediction accuracy of the applied machine learning algorithm, in a multi-objective pareto-based dominance form. The evolutionary feature selection approach involves the steps of population initialization, crossover, mutation, and selection, based on a genetic algorithm mimicking the natural evolutionary process. The machine learning problem is thereby defined as a multi-objective optimization model involving the simultaneous optimization of the given objectives, producing a set of optimal solutions, called the Pareto set, where each solution of this set has a different trade-off between the two objectives. We compare the accuracy and run-times using different feature selection approaches and compare it on real-life datasets to show how the proposed evolutionary multi-objective feature selection approach outperforms the rest, along with theoretical justification based on combinatorics and optimization.

INTRODUCTION

Predictive Modeling is a Machine Learning task of classification or regression consisting of the following steps as shown below in Figure 1.



Figure 1: Machine Learning Predictive Modeling Process Flow.

In this work, we integrate the concept of Multi-Objective Optimization into Feature Selection and the **Machine Learning Classification task**. We first evaluate the machine learning models' performance without feature selection, then apply wrapper-based feature selection approach using the Newton-Raphson with Ridging optimization for forward selection, and finally apply single and multi-objective genetic algorithm on two case studies of real-life datasets using machine learning classification. We apply the Non-dominated Sorting Genetic Algorithm (NSGA-II) algorithm to classify the Chronic Kidney Disease (CKD) Proteomics dataset in a multi-objective approach of evaluating two conflicting objectives of minimizing the number of features, and maximizing the prediction accuracy of the classification algorithm, using RapidMiner Studio. We then provide an application of multi-objective optimization using genetic algorithm on the Iris Dataset available within the SASHELP library, by defining the optimization problem as both single and multi-objective, by maximizing the accuracy and minimizing the model training CPU processing time in parallel, using the SAS® Optimization Autotune Genetic solver within the SAS® Viya Cloud Analytics Server (CAS) environment.

EVOLUTIONARY FEATURE SELECTION

Feature Selection, also referred to as Variable Subset Selection, or Data Reduction, is a data pre-processing step within the Machine Learning framework. If left untouched, using all the original features from the dataset can result in the machine being trained on, and modeled to learning the noisy patterns rather than the actual signal, thereby resulting in a biased model which only performs well on the training dataset and fails significantly when exposed to future unseen data.

There are several feature selection techniques available in literature based on different filter and wrapper approaches. It is often a challenging task to select the right type of feature selection technique for the problem at hand. The filter-based selection approaches involve manually setting a threshold for the correlation coefficient, and hyper-parameter tuning. This often is not an optimal selection and can have various drawbacks such as the solution being stuck in a local optimum and never reaching a global optimum. Therefore, Feature Selection is an inherently multi-objective task. One option when considering multiple objectives by the traditional machine-learning algorithms is to combine different objectives into a single number, represented by a certain trade-off factor known as the scalar cost function. This requires standardizing the values in some way and giving them weights for their importance, and can simplify the optimization problem, but may be sensitive to how individual objectives are weighed. Moreover, choosing the right trade-off factor often depends on right skills and pre-determined knowledge of the dataset in hand. This type of a manual choosing of the given trade-off factor may not be feasible in real-life. The main disadvantage of this approach is that many separate optimizations with different weighting factors need to be performed to examine the trade-offs among the objectives as given by the below equation.

$$R_{reg}(\beta) = R_{emp}(\beta) + F \cdot \lambda(\beta)$$

where,

$R_{reg}(\beta)$: Regularized Risk

$R_{emp}(\beta)$: Empirical Risk

F : Trade-Off Factor

$\lambda(\beta)$: Structural Risk



Figure 2: Evolutionary Cycle of Feature Selection.

Therefore, one must deal with the trade-off between approximation and model complexity in an unbiased and inherently multi-objective fashion, by optimizing both the objectives simultaneously. For example, in feature selection, minimization of the number of features and maximization of feature quality are two common objectives that are likely conflicting with each other, which generates diverse multiple Pareto-optimal models to achieve a desired trade-off among various performance metrics. We, therefore, apply evolutionary feature selection using genetic algorithms to solve the multi-objective optimization problem, where the initial population of solutions is randomly generated using the number of features in the original dataset. The steps followed by the genetic algorithm is described in Figure 2.

MULTI-OBJECTIVE OPTIMIZATION MODEL

Multi-Objective Optimization problems deal with conflicting objectives, i.e. while one objective increases the other decreases. There is no unique global solution but a set of pareto-optimal solutions. In general, we are interested in the following mathematical problem type (Deb 2001):

$$\text{Minimize/Maximize: } Z_m(x); m = 1, 2, \dots, M.$$

$$\text{Subject to: } q_j(x) \geq 0; j = 1, 2, \dots, J.$$

$$r_k(x) = 0; k = 1, 2, \dots, K.$$

$$x_i(L) \leq x_i \leq x_i(U); i = 1, 2, \dots, n.$$

A solution "X" is a vector of "n" decision variables given by:

$$X = (x_1, x_2, \dots, x_n)^T.$$

Feasible Solution

A solution that satisfies all constraints and variable bounds. The set of all feasible solutions is called the feasible region, or "S". The objective space is constituted by the possible values of the "Z" objectives functions for all solutions in "S". (Calle 2017).

Domination

A solution $x^{(1)}$ is said to dominate the other solution $x^{(2)}$ if both conditions (i) and (ii) below are true:

Condition (i): $x^{(1)}$ is no worse than $x^{(2)}$ for all objectives.

Condition (ii): $x^{(1)}$ is strictly better than $x^{(2)}$ in at least one objective.

The mathematical notation for $x^{(1)}$ dominates $x^{(2)}$ is: $x^{(1)} \preceq x^{(2)}$. (Calle 2017).

Non-Dominated Set

Among a set of solutions X, the non-dominated set of solutions X' are those that are not dominated by any member of the set X. (Calle 2017).

Globally Pareto-Optimal set

The non-dominated set of the entire feasible search space "S" is defined as the globally Pareto-Optimal set. (Calle 2017).

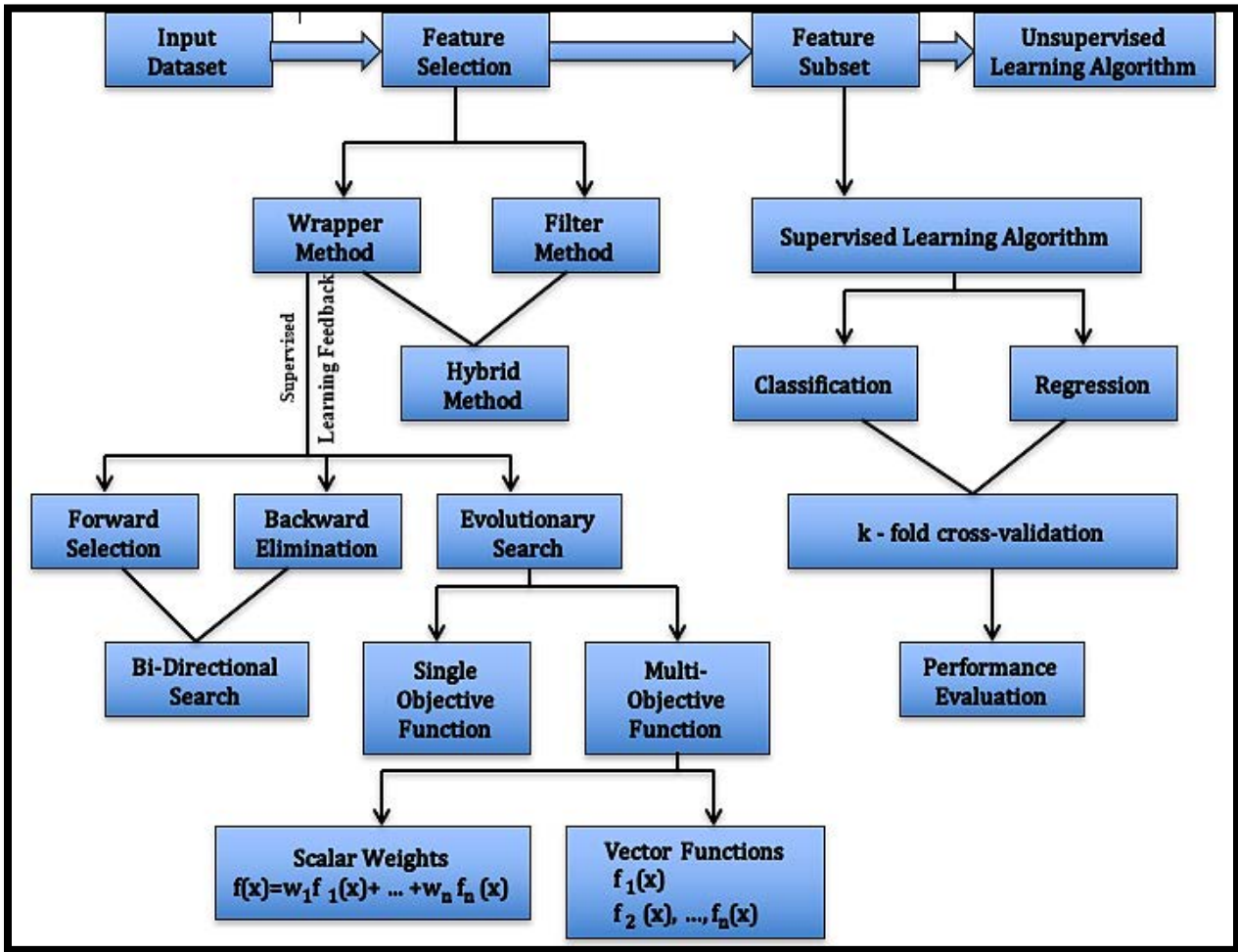


Figure 3: Machine Learning Framework with Filter and Wrapper-based Feature Selection Approaches.

NON-DOMINATED SORTING GENETIC ALGORITHM (NSGA-II)

NSGA-II is a multi-objective evolutionary algorithm. Evolutionary algorithms were developed because the classical direct and gradient-based techniques have the following problems when leading with non-linearities and complex interactions:

- The convergence to an optimal solution depends on the chosen initial solution.
- Most algorithms tend to get stuck to a sub-optimal solution.

NSGA-II has the following three features:

1. It uses an elitist principle, i.e., the elites of a population are given the opportunity to be carried to the next generation.
2. It uses an explicit diversity preserving mechanism (crowding distance).
3. It emphasizes the non-dominated solutions (Deb 2001).

NSGA-II Algorithm

1. Perform a non-dominated sorting in the combination of parent and offspring populations and classify them by fronts, i.e., they are sorted according to an ascending level of non-domination.
2. Fill new population according to front raking.

3. If one front is taking partially like F3, perform crowding-sort that uses crowding distance that is related with the density of solutions around each solution. The less dense are preferred.
4. Create offspring population from this new population using crowded tournament selection (it compares by front ranking, if equal, then by crowding distance), crossover and mutation operators (Calle 2017).


CASE STUDY 1: APPLICATION ON AFRICAN-AMERICAN STUDY OF KIDNEY DISEASE (AASK) PROTEOMICS DATASET

Chronic Kidney Disease (CKD) is a medical condition, defined by reduced Glomerular Filtration Rate (GFR), Proteinuria, or Structural Kidney Disease. Identification and characterization of novel biomarkers and targets of therapy for the CKD patients remains a major focus of the current research in kidney disease and has been the objective of a number of studies, such as the African-American Study of Kidney Disease and Hypertension (AASK) (Subasi et al. 2017). The proposed approach will help develop the knowledge of therapeutic intervention and prognostic study of CKD, by providing deeper insights into medical datasets. We compare the model performance based on KS(Youden) J Statistic, and Misclassification Rates, for several machine learning classification algorithms such as Support Vector Machines (SVM), Random Forest, Logistic Regression, Decision Trees, Gradient Boosting, and Neural Networks, with and without feature selection using the SAS[®] Viya Visual Data Mining and Machine Learning (VDMML) platform (SAS[®] Institute 2019) as shown in Outputs 1 and 2 below.

The AASK Proteomics Dataset characteristics are as described in Table 1 below.

Proteomics Dataset	Values
No. of Instances	116
No. of Features	5751
No. of Classes	2
No. of Positive Samples	68
No. of Negative Samples	48
Data Source	AASK Sponsors

Table 1: AASK Proteomics Dataset Structure.

Model Comparison				
Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
	SVM	SVM	0.4571	0.2500
	Forest	Forest	0.1714	0.5000
	Logistic Regression	Logistic Regression	0.0286	0.5000
	Decision Tree	Decision Tree	0	0.4167
	Gradient Boosting	Gradient Boosting	0	0.5000
	Neural Network	Neural Network	0	0.4167

Output 1: Supervised Classification of AASK Proteomics Dataset without Feature Selection.

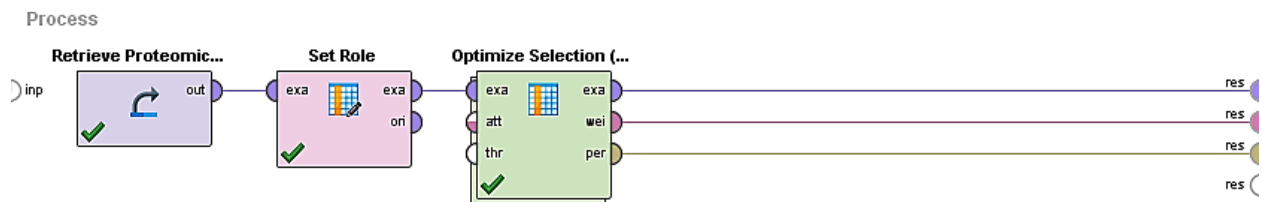
Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
	Forward Logistic Regression	Logistic Regression	0.8571	0.0833
	Forest	Forest	0.1714	0.5000
	Neural Network	Neural Network	0.1714	0.4167
	Gradient Boosting	Gradient Boosting	0	0.5000
	Stepwise Logistic Regression	Logistic Regression	0	0.4167
	Decision Tree	Decision Tree	0	0.4167

Output 2: Supervised Classification of AASK Proteomics Dataset with Feature Selection.

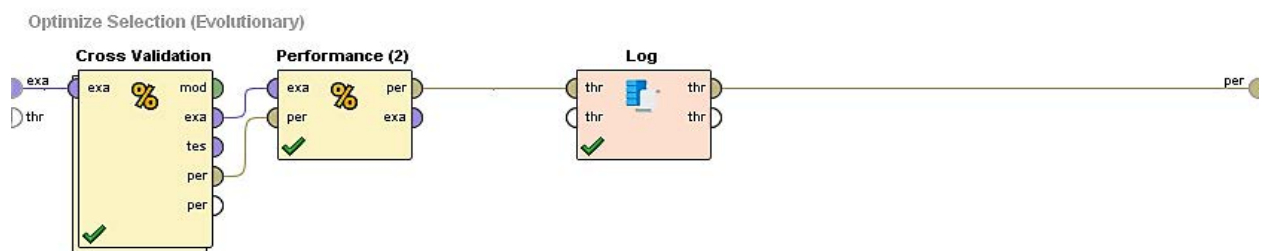
The champion model for the Supervised Classification of AASK Proteomics Dataset with no feature selection resulted in a significantly lower KS(Youden) statistic for the Test Partition of 0.4571 as shown in Output 1. The KS(Youden) or "J" Statistic given by $(Sensitivity + Specificity - 1)$, and was chosen for the classification model evaluation metric for reliability and consistency. The champion model with Feature Selection is Forward Logistic Regression, which uses the gradient based Newton-Raphson Ridging for feature selection. The model was chosen based on the high index of KS(Youden) for the Test partition of 0.86, as shown in Output 2. 91.67% of the Test partition was correctly classified using the Forward Logistic Regression model using the gradient based feature selection. The five most important factors are M2034, M796, M2042, M2040, and M3662.

MACHINE LEARNING CLASSIFICATION USING MULTI-OBJECTIVE NSGA-II EVOLUTIONARY FEATURE SELECTION IN RAPIDMINER STUDIO

Step 1: Retrieve Proteomics Data in the Process Flow



Step 2: Optimize Feature Selection using Evolutionary NSGA-II



Step 3: Run 10-Fold Cross-Validation Experiments

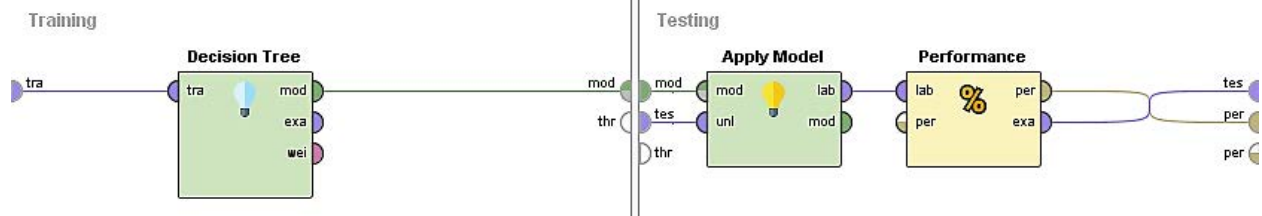
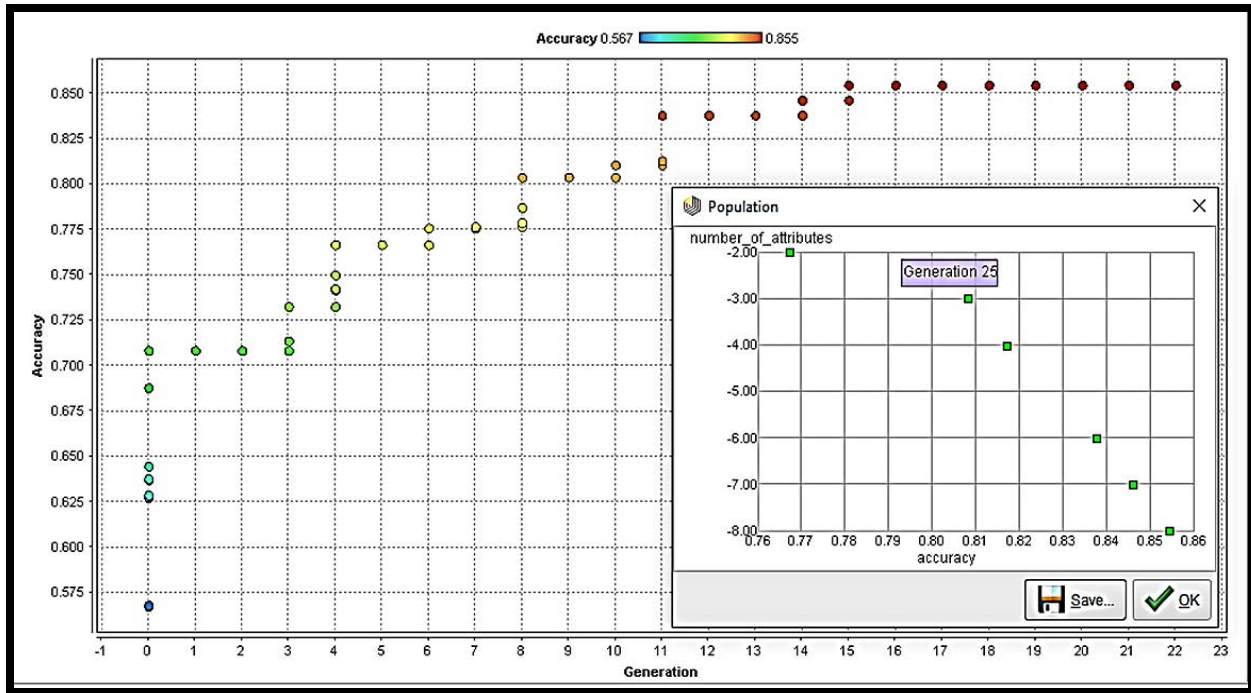


Figure 4: Wrapper-Based Feature Selection for Classification in RapidMiner Studio.

We apply Evolutionary Feature Selection using NSGA-II to minimize the number of features and maximize the classification accuracy in a wrapper-based approach using RapidMiner Studio (Mierswa 2017), as displayed in Figure 4. The Pareto-plot displayed in Output 3 shows that as the number of features increases, the accuracy also increases, and plots the pareto-optimal plot showing that 8 features are sufficient to obtain an overall classification accuracy of about 86% in 25 Generation runs. The number of attributes are displayed on the y-axis, and are negative in nature since in reality we are minimizing the features, which in a multi-objective optimization setting is equivalent to the negative of maximizing the number of attributes. The x-axis shows the accuracy of the wrapper-based classification algorithm. Output 4 provides an overview of the model interpretability showing the combination of features selected from the pareto-optimal plot.



Output 3: Accuracy Results and Pareto-Optimal Plot.

Index	Features	Names	accuracy	number_of_attributes
1	2	M1550, M3156	0.767	2
2	2	M1550, M3156	0.767	2
3	8	M1550, M1554, M2022,...	0.855	8
4	8	M1550, M1554, M2022,...	0.855	8
5	4	M1550, M2994, M3156,...	0.817	4
6	6	M1550, M1554, M3152,...	0.838	6
7	3	M1550, M2994, M3156	0.808	3
8	4	M1550, M2994, M3156,...	0.817	4
9	6	M1550, M1554, M3152,...	0.838	6
10	3	M1550, M2994, M3156	0.808	3
11	7	M1550, M1554, M2022,...	0.846	7
12	7	M1550, M1554, M2022,...	0.846	7
13	2	M1550, M3156	0.767	2
14	2	M1550, M3156	0.767	2

Output 4: Model Interpretability showing the combination of features selected.

CASE STUDY 2: APPLICATION OF SINGLE AND MULTI-OBJECTIVE GENETIC SEARCH BASED OPTIMIZATION FOR CLASSIFICATION OF IRIS DATASET USING SAS VIYA

Lastly, we provide an application of single objective optimization and multi-objective optimization on the Iris Dataset available within the SASHELP library, using the SAS Autotune Genetic Optimization solver on the SAS® Viya CAS platform. The results show how multi-objective genetic search based optimization is robust and efficient in real-life scenarios when dealing with multiple conflicting objectives, and being applicable to complex problem scenarios where the calculus and gradient based methods often get stuck at a local optimum, and in some cases cannot even solve the optimization problem at hand due to the complex non-linear structure of the problem. For the application on single-objective optimization, we minimize the Misclassification Error Percentage as shown in Output 5 for the Iris dataset. For the application on multi-objective optimization, we use the two objectives of minimizing the Misclassification Error Percentage and minimizing CPU Training Processing Time in Seconds as shown in Outputs 6 and 7. The results for the same are displayed as a Pareto-plot in Figure 5, showing that the multi-objective classification approach provides a better solution both in terms of lower error rates and lower processing times.

SINGLE OBJECTIVE GENETIC SEARCH BASED OPTIMIZATION USING GRADIENT BOOSTING ON IRIS DATASET

Prediction Error With Gradient Boosting Tree Analytics for IRIS_AUTOTUNE_23FEB2020:21:28:54							
Tree ID	Number of Trees	Number of Leaves	Misclassification Error	Log Loss	Average Squared Error	Root Average Squared Error	Maximum Absolute Error
0	1	4	0.04867	0.1557	0.01887	0.1374	0.9831
1	2	7	0.02867	0.1286	0.01649	0.1284	0.9828

Tuner Information	
Model Type	Gradient Boosting Tree
Tuner Objective Function	Misclassification Error Percentage
Search Method	GA
Population Size	10
Maximum Iterations	5
Maximum Tuning Time in Seconds	36000
Validation Type	Single Partition
Validation Partition Fraction	0.30
Log Level	2
Seed	12345
Number of Parallel Evaluations	4
Number of Workers per Subsession	0

Output 5: Single-Objective Optimization Using Genetic Search Based Gradient Boosting on Iris Dataset.

MULTI-OBJECTIVE GENETIC SEARCH BASED OPTIMIZATION USING GRADIENT BOOSTING FOR CLASSIFICATION OF IRIS DATASET

SAS Code:

```
/* Define a CAS engine libref for CAS in-memory data tables */
libname mycaslib cas caslib=casuser;
```



```

data mycaslib.iris;
  set sashelp.iris;
  run;

proc cas noqueue;
  autotune.tuneGradientBoostTree result=r /
    trainOptions={
      table={ name='iris', where='Species' },
      inputs={
        "SepalLength",
          "SepalWidth",
          "PetalLength",
          "PetalWidth"
      },
      target='Species',
      nominals={ 'Species' },
      casout={ name='gradboost_iris_evolutionarymodel', replace=true }
    }
    tunerOptions={
      seed=66666,
      secondObjective='TRAININGTIME'
    }
  ;
  print r;
  saveresult r.TunerResults;
  run;
quit;

data ParetoSet;
  set TunerResults (firstObs=2);
  run;

proc sgplot data=ParetoSet;
  title "Pareto Set Produced using Genetic Search Gradient Boosting on Iris
  Dataset";
  scatter x=TrainCpuTime y=MisclassErr;
  run;

```

Prediction Error With Gradient Boosting Tree Analytics for IRIS_AUTOTUNE_23FEB2020:21:28:56							
Tree ID	Number of Trees	Number of Leaves	Misclassification Error	Log Loss	Average Squared Error	Root Average Squared Error	Maximum Absolute Error
0	1	2	0.04000	1.0882	0.2195	0.4685	0.6679

Tuner Information	
Model Type	Gradient Boosting Tree
Tuner Objective Function	Misclassification Error Percentage
Second Objective Function	Training CPU Time in Seconds
Search Method	GA
Population Size	10
Maximum Iterations	5
Maximum Tuning Time in Seconds	36000
Validation Type	Single Partition
Validation Partition Fraction	0.30
Log Level	2
Seed	66666
Number of Parallel Evaluations	4
Number of Workers per Subsession	0

Output 6: Multi-Objective Optimization Using Genetic Search Based Gradient Boosting for Classification of Iris Dataset.

Tuner Results Default and Best Configurations										
Evaluation	Maximum Tree Levels	Number of Bins	Number of Variables to Try	Learning Rate	Sampling Rate	Lasso	Ridge	Misclassification Error Percentage	Training CPU Time in Seconds	Evaluation Time in Seconds
0	5	50	4	0.100000	0.500000	0	1.000000	4.44	0.0480	0.12
104	2	20	3	0.010000	1.000000	5.000000	5.000000	4.44	0.0280	0.05
97	2	20	3	0.505000	1.000000	10.000000	5.000000	6.67	0.0260	0.05
80	2	20	3	0.505000	1.000000	5.000000	5.000000	8.89	0.0230	0.04
8	6	20	3	0.560000	0.100000	4.444444	10.000000	66.67	0.0230	0.03

Tuner Iteration History				
Iteration	Evaluations	Best Objective	Best Second Objective	Elapsed Time in Seconds
0	1	4.44	0.0480	0.12
1	25	4.44	0.0230	0.66
2	46	4.44	0.0230	0.97
3	65	4.44	0.0230	1.21
4	85	4.44	0.0230	1.47
5	104	4.44	0.0230	1.72

Output 7: Iteration History of Multi-Objective Optimization Using Genetic Search Based Gradient Boosting for Classification of Iris Dataset.

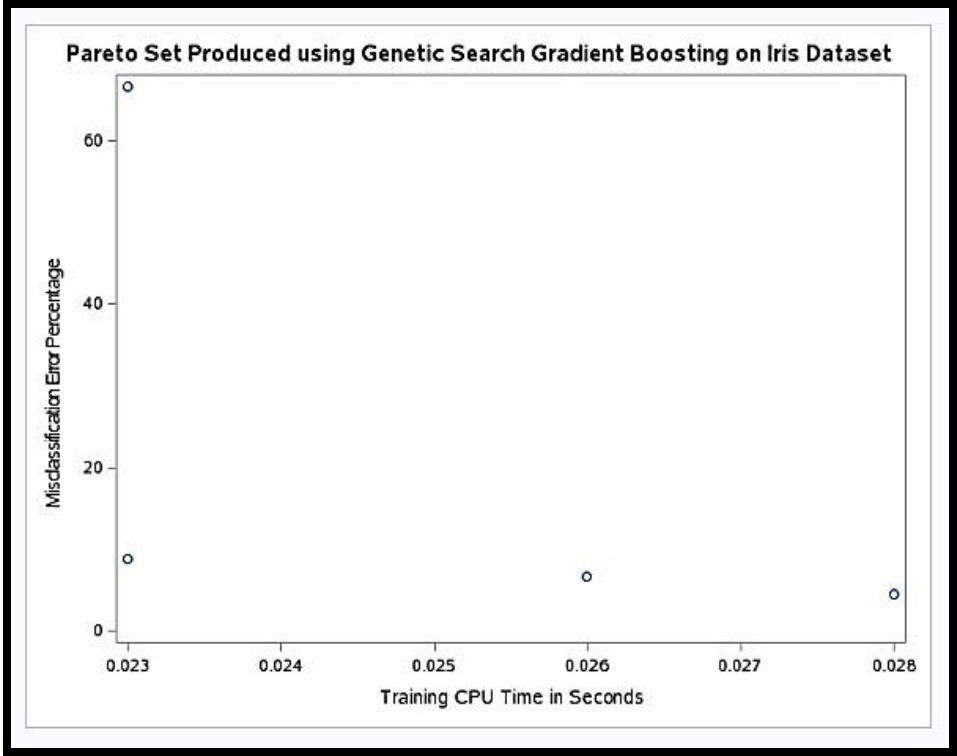


Figure 5: Pareto Plot for the Multi-Objective Genetic Search Based Classification of Iris Dataset.

CONCLUSION

In this paper, we proposed an evolutionary feature selection technique for the machine learning predictive modeling task, involving two conflicting goals of minimizing the number of features and maximizing the prediction accuracy of the applied machine learning algorithm, in a multi-objective Pareto-based dominance form. We provided an evaluation of the machine **learning models' performance** with and without gradient based feature selection, and with evolutionary feature selection using single and multi-objective genetic algorithms. We applied machine learning classification using the proposed approach on two case studies of real-life datasets. We compared the accuracy and run-times using the different classification algorithms and compared it on the AASK Proteomics and the Iris datasets to show how the proposed evolutionary multi-objective feature selection approach outperforms the rest, along with theoretical justification based on combinatorics and optimization. The proposed approach will help develop the knowledge of therapeutic intervention and prognostic study of Chronic Kidney Disease, by providing deeper insights into medical and other real-life scenario datasets. The on-going research also involves the development of a novel multi-objective machine learning classification algorithm based on Logical Analysis of Data.

REFERENCES

African-American Study of Kidney Disease and Hypertension Cohort Study (AASK Cohort). Available at <https://repository.niddk.nih.gov/studies/aask-cohort/>.

Calle, P. "Data Science Techniques." Analytics Lab @OU. October 24th 2017. Available at <https://oklahoamalytics.com/data-science-techniques/nsga-ii-explained/>.

Deb, Kalyanmoy. 2001. "Multi-objective optimization using evolutionary algorithms." John Wiley & Sons, Vol. 16.

Gomez, F., Quesada, A. "Genetic Algorithms for Feature Selection." Neural Designer. 2020. Available at https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection.

Khan, A., Baig, A.R. Feb 2015. "Multi-Objective Feature Subset Selection using Non-dominated Sorting Genetic Algorithm". Journal of Applied Research and Technology, Vol. 13, Issue 1.

Mierswa, I. "Evolutionary Algorithms for Feature Selection". Rapid Miner Studio. December 2017. Available at <https://rapidminer.com/products/studio/>.

SAS® Institute, 2019, "SAS® Visual Data Mining and Machine Learning 8.5: Programming Guide."

Subasi, E., Subasi, M.M., Hammer, P.L., Roboz, J., Anbalagan, V., Lipkowitz, M.S. 19 July 2017. "Classification Model to Predict Rate of Decline of Kidney Function", Frontiers in Medicine. Available at <https://www.frontiersin.org/articles/10.3389/fmed.2017.00097/full>.

UCI Machine Learning Repository. Available at <https://archive.ics.uci.edu/ml/index.php>.

ACKNOWLEDGMENTS

This project was supported in part by an appointment to the Research Participation Program at the U.S. Food and Drug Administration administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. The authors acknowledge the AASK sponsors, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), NIH Office of Research in Minority Health, and King Pharmaceuticals, and thank Michael Lipkowitz for providing the Proteomics Dataset.

RECOMMENDED READING

- *SAS® Visual Data Mining and Machine Learning 8.5: Programming Guide.*
- *An Introduction to SAS® Viya® 3.5 Programming.*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Nandini Rakala
nrakala2015@my.fit.edu
Department of Mathematical Sciences
Florida Institute of Technology
150 W. University Blvd., Melbourne, FL 32901

Munevver Mine Subasi
msubasi@fit.edu
Department of Mathematical Sciences
Florida Institute of Technology
150 W. University Blvd., Melbourne, FL 32901

Ersoy Subasi
esubasi@fit.edu
Department of Computer Engineering and Sciences
Florida Institute of Technology
150 W. University Blvd., Melbourne, FL 32901