Analysis of Variable Selection Methods on Scientific Cluster Measurement Data

Jonathan Wang[1], Wucherl Yoo (Advisor)[2], Alex Sim (Advisor)[2], K. John Wu (Advisor)[2]
[1]University of California Berkeley, [2]Lawrence Berkeley National Laboratory

RESEARCH GOAL

To improve the accuracy and training time of performance models using parallelized variable selection methods

APPLICATION

- Palomar Transient Factory (PTF) observes the sky to identify astronomical transients, such as supernovae
- PTF workflow processes image data to identify potential transient objects
- Ongoing effort to understand the resource requirement for the PTF workflows
- Challenge: build a performance model that accurately predicts the total execution time after the first few steps of the workflow

VARIABLE SELECTION

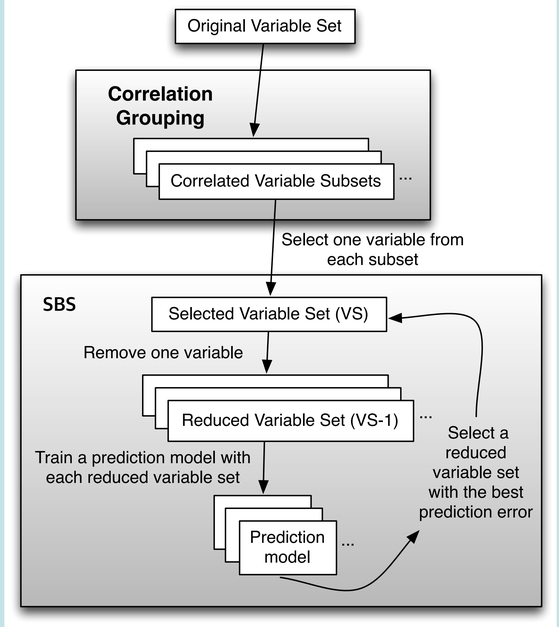

Figure 1: Workflow of 2-phase variable selection process. Stacked sections are parallelized using Apache Spark

- Parallelization
  - Select variables from correlation groups in parallel
  - Test reduced variable subsets in parallel
  - Use 3 nodes with 24 cores each
  - Split into 50 parallel computations (equal to number of variables)

RESULTS

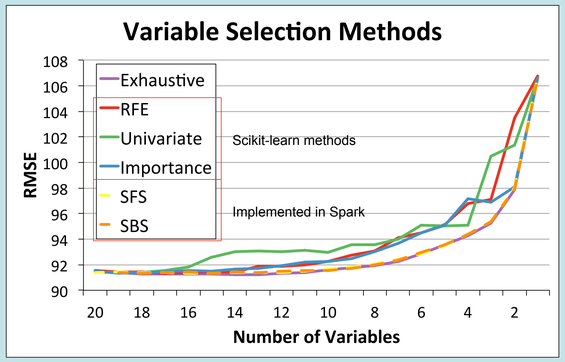

Figure 2: Exhaustive search finds the optimal variable subset. Sequential Backward Selection (SBS) and Sequential Forward Selection (SFS), which were implemented in Spark, achieve lower and more consistent prediction accuracy than existing scikit-learn methods.

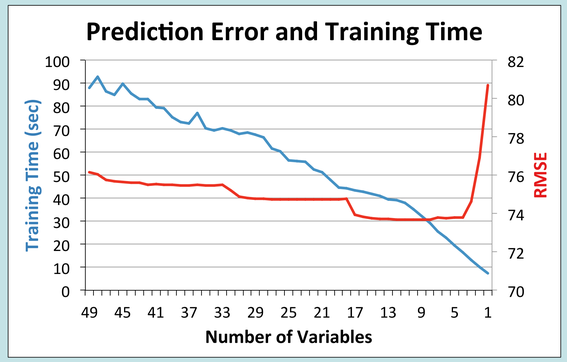

Figure 5: Training time of Gradient Boosting models improves significantly at little cost to the model prediction accuracy until the variable subset becomes too small to make a good prediction.

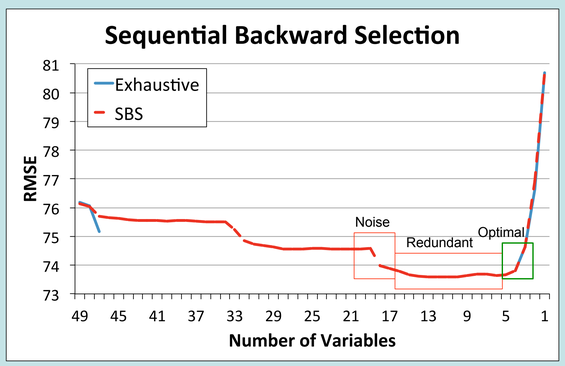

Figure 3: SBS identifies the optimal subset size as well as noisy and redundant variables. Noisy variables negatively impact prediction while redundant variables do not contribute new information. Due to exponential runtime, exhaustive search could not be completed on the larger variable set.

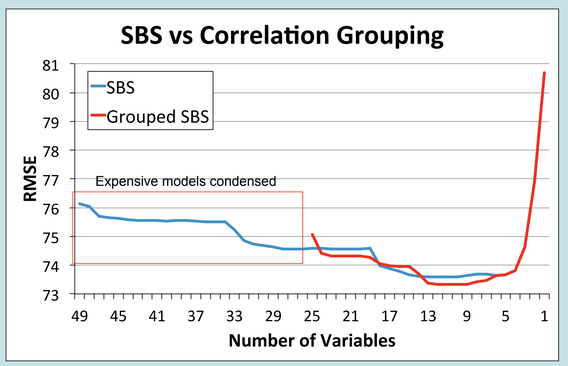

Figure 6: Correlation grouping with correlation threshold 0.8 finds results similar to SBS. Larger, expensive models are eliminated quickly. The improvement is dependent on variable correlations in the data and how the variables are grouped.

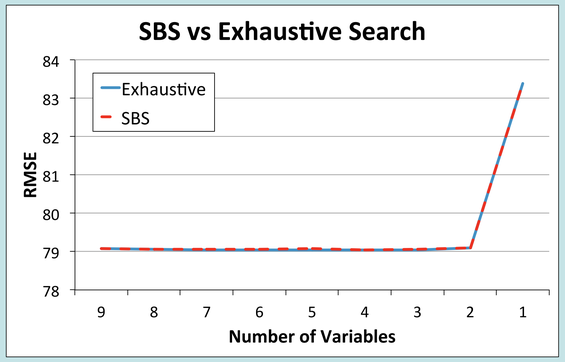

Figure 4: Testing SBS and exhaustive search on a smaller subset of variables shows that SBS is able to achieve similar prediction accuracy to the optimal subset.

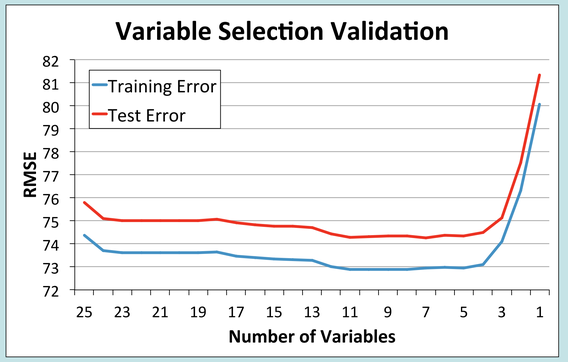

Figure 7: Similar improvement trends show that selected variables are not overfit to data. Variables are selected based on importance to the data rather than model variance.

DISCUSSION

- Gradient boosting selected as machine learning model
  - Non-linear prediction
  - Consistent prediction results
- Variable selection mechanism contains prediction model for training
- Use resulting variable subset to build actual performance model
- RMSE in Figures 2-6 refers to prediction error of training model
- Figure 7 compares error between training model and performance model
  - Optimizing variables for training model optimizes variable subset for performance model

CONCLUSION

- Variable selection identifies optimal subset for Gradient Boosting prediction model
  - Shorter training time
  - Improved prediction accuracy
- SBS approximates exhaustive selection
  - Parallelization improves runtime from **18 hours (65020 sec)** to **45 min (2727 sec)**
- Correlation grouping identifies same variables
  - Reduces runtime of SBS from **2727 sec** to **888 sec** on PTF dataset
  - Correlation threshold determined by balancing runtime improvement and selection accuracy
  - Maximum improvement with many medium sized correlation groups

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. This work was also supported by the U.S. Department of Energy, under Contract No. DE-AC02-05CH11231 and used resources of the National Energy Research Scientific Computing Center.

CONTACT

- Jonathan Wang
- University of California, Berkeley
- Email: jonathanwang017@berkeley.edu