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]
University of California Berkeley[1], Lawrence Berkeley National Laboratory[2]
jonathanwang017@berkeley.edu, {wyoo, asim, kwu}@lbl.gov

1. INTRODUCTION

Scientific applications are increasingly reliant on large distributed workflows to create and analyze vast amounts of data [1, 3]. However, it is becoming more challenging to model application performance due to the large number of variables that affect overall performance. To reduce the time needed to establish a performance model, we explore a set of variable selection techniques to find the best subset of variables for building the performance model.

As a case study, we examined the Palomar Transient Factory (PTF) application, which processes large amounts of astronomy observations through a lengthy processing pipeline [2]. Our prediction task is to use variables about the data objects plus the execution time of the first few steps of the pipeline to forecast the overall execution time of the entire workflow.

2. METHODS

We established a testing baseline using exhaustive variable selection, which finds the optimal subset, and tested several standard variable selection methods, including Recursive Feature Elimination, Univariate F-Test, and Gini Feature Importance.

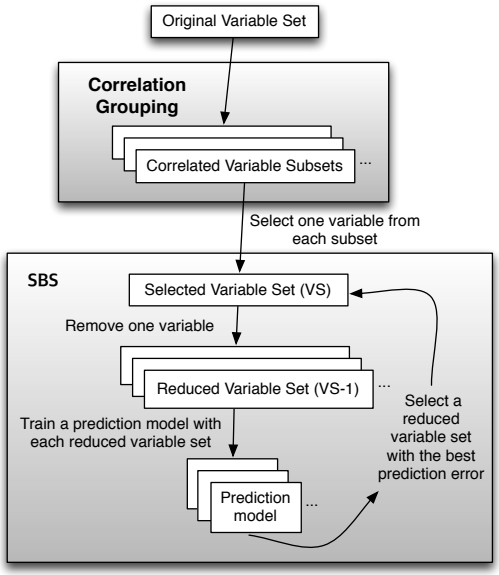

Figure 1: Overview of the variable selection process

As a greedy selection approach, we implemented Sequential Backward Selection (SBS), which starts with the fulls set of variables and removes one variable at each iteration. We selected the subset with the best prediction accuracy to determine the best variable to remove. Apache Spark was used to parallelize subset testing at each iteration.

We also combined SBS with correlation-based grouping to take advantage of multiple correlated variables to further improve the performance of SBS. We eliminate redundant variables quickly by grouping highly correlated variables and selecting the best representative variable from each group. The selected variables are then reduced using the SBS. Correlation grouping was also parallelized across the correlation groups. Figure 1 shows the steps of the variable selection process.

3. RESULTS

Our SBS and SFS implementations performed better and more consistently than existing variable selection implementations. Parallelizing SBS resulted in a significant runtime improvement from about 18 hours (65020 seconds) to less than an hour (2727 seconds). We were unable to compare SBS against exhaustive selection on the full variable set due to runtime limitations. However, by testing SBS on a smaller 10-variable set, we found that SBS achieved similar prediction error to exhaustive search.

Figure 2 shows the results of testing SBS with the full variable set. There is a very visible trend as variables were removed. The decreases in error represent noisy variables being removed, while the flat segments represent redundant variables. After the optimal subset is reached, the error grows rapidly due to key variables being removed, leaving insufficient information to make an accurate prediction.

Figure 3 illustrates the rapid decrease in the training time relative to the loss in the prediction accuracy. There was little loss in the prediction accuracy until the subset reached extremely small sizes.

Our experiments showed that correlation grouping preprocessing returned results comparable to standard SBS. Figure 4 shows the selection process for both methods. The changes in error follow similar patterns, with the plot for correlation grouping being more condensed. At a correlation threshold of 0.8, grouped SBS took only 888 seconds to run as opposed to 2727 seconds for regular SBS.

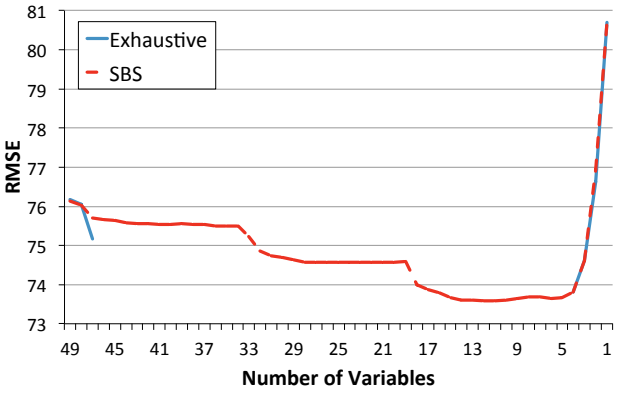

Figure 2: Variable selection process for Sequential Backward Selection

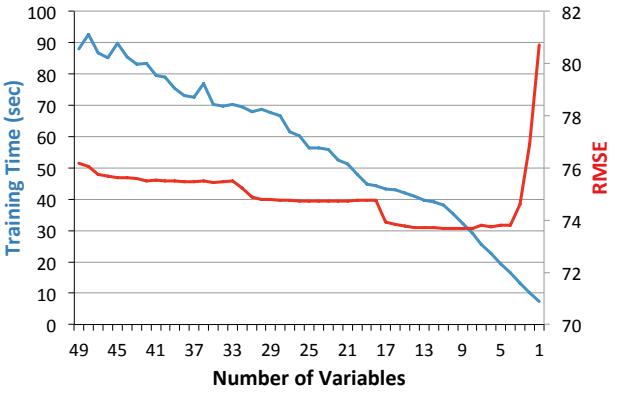

Figure 3: Change in prediction error compared to training time

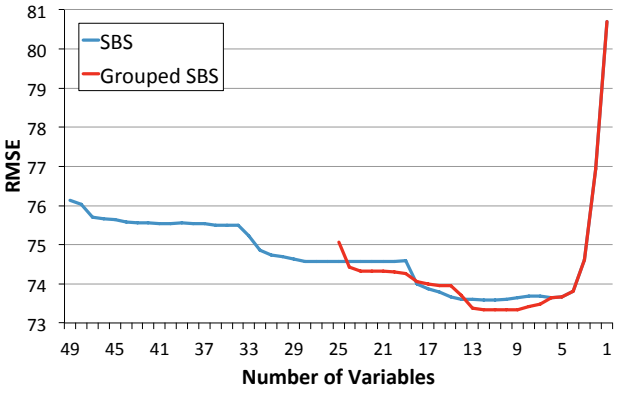

Figure 4: Comparison of SBS with correlation grouping and standard SBS

In order to evaluate overfitting in the variable selection, we used a separate test set to validate our variable subset. The selection trends of of the two datasets are the same, showing that the selected variables were not overfit to the training data.

4. CONCLUSION

Variable selection methods were shown to be effective in reducing model training time and eliminating noisy variables on the PTF analysis pipeline measurement data. Sequential Backward Selection proved to be an effective variable selection method for this measurement dataset as it found a subset comparable to exhaustive selection in significantly shorter time.

We developed a framework to quickly select variables from the PTF data to optimize the prediction accuracy. Due to the high levels of correlation among variables in the dataset, correlation-based grouping in our method was able to further improve the performance of the SBS. In this experiment, it was able to identify the same subset as the SBS in just one-third of the computation time.

By taking advantage of high performance computing resources and variable correlations, we were able to select a variable subset that can result in accurate performance prediction within significantly shorter computation time than existing methods.

5. REFERENCES

[1] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft, Oct. 2009.

[2] N. M. Law, S. R. Kulkarni, R. G. Dekany, E. O. Ofek, R. M. Quimby, P. E. Nugent, J. Surace, C. C. Grillmair, J. S. Bloom, M. M. Kasliwal, L. Bildsten, T. Brown, S. B. Cenko, D. Ciardi, E. Croner, S. G. Djorgovski, J. v. Eyken, A. V. Filippenko, D. B. Fox, A. Gal-Yam, D. Hale, N. Hamam, G. Helou, J. Henning, D. A. Howell, J. Jacobsen, R. Laher, S. Mattingly, D. McKenna, A. Pickles, D. Poznanski, G. Rahmer, A. Rau, W. Rosing, M. Shara, R. Smith, D. Starr, M. Sullivan, V. Velur, R. Walters, and J. Zolkower. The palomar transient factory: System overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 121(886):1395–1408, 2009.

[3] A. Shoshani and D. Rotem, editors. *Scientific Data Management: Challenges, Technology, and Deployment.* Chapman & Hall/CRC Press, 2010.