Comparison of machine-learning techniques for handling multicollinearity in big data analytics and high-performance data mining

Gerard G. Dumancas, PhD
Oklahoma Baptist University
Shawnee, Oklahoma, USA
+14057308752
Gerard.dumancas@okstate.edu

Ghalib A. Bello, PhD
Virginia Commonwealth University
Richmond, Virginia, USA
+16142180592
belloga@vcu.edu

1. INTRODUCTION

Large datasets involved in big data analytics and high performance data mining often contain thousands of correlated variables. Multicollinearity in big data analytics leads to biased estimation, variance inflation [1] and is a serious impediment to most machine-learning techniques [2]. Very few studies have compared machine-learning methods in modeling of data with multicollinearity and none have been applied directly for mortality prediction using lipid profiles [3-4]. In this study, we provide an extensive comparison of the performance of 12 machine-learning methods for handling multicollinearity using lipid clinical data to predict an individual's 5-year mortality. Results from this study indicate that partial least squares-discriminant analysis (PLS-DA) offers the best solution in handling multicollinearity in this particular scenario. We believe that such a technique can be implemented for automated pre-processing of correlated variables in big data analytics and high performance data mining.

2. DATASET

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health status of US individuals [5]. NHANES lipid profiles (total cholesterol, HDL, triglycerides), demographic (gender, age, ethnicity) and mortality-associated variables (survival time, mortality status) from the National Death Index were used in our analysis. The aim was to use the lipid profile data to predict a binary indicator of 5-year mortality, adjusting for potentially confounding demographic variables. Our final dataset contained 726 individuals. Those who died within the 5 years after participation in NHANES were considered cases (n=121), and the rest considered controls (605). The total sample was divided (by a ~7:3 ratio) into a training set comprising 483 individuals (403 controls/80 cases) and a test set (n=243 [202 controls/41 cases]).

3. DATA ANALYSES

Predictive models built using the training data were then used to predict 5-year mortality in the test set. The models were constructed using the most commonly used machine-learning techniques: PLS-DA, artificial neural networks, boosting, random forests, naïve bayes, support vector machines, recursive partitioning/regression trees, and penalized logistic regression (with ridge, LASSO, elastic net penalties). We also combined the predictions of these techniques to form an aggregate estimator using an ensemble learning method called stacking [6], the idea behind which is that combining predictions of various machine-learning algorithms should lead to better predictive performance than what any individual algorithm is capable of producing. Predictive performance was assessed using AUC. All calculations were performed using packages in *R version 3.1.2* [7].

4. RESULTS/CONCLUSIONS

Stacking did not significantly demonstrate superior performance compared to the top-performing individual predictors, particularly PLS-DA (Table 1). The performance of PLS-DA as measured by AUC was found to be significantly different compared to LASSO, recursive partitioning and regression trees, random forest, ridge regression stacking, and naïve bayes, but was not found to be significantly different compared to the rest of the techniques (Table 2). This study introduces a wide array of methods for handling multicollinearity and provides a benchmark for comparing their performance. The results identify a number of robust techniques, with PLS-DA as the top performing technique, for dealing with multicollinearity. Such a method can be potentially useful in big data analytics wherein complex and non-negligible correlation patterns exist among variables.

Table 1. Comparison of AUC values among different techniques.

Technique	AUC
PLS-DA	0.8870
Artificial neural network	0.8823
Logistic regression	0.8818
Elastic net	0.8803
Ridge regression	0.8761
LASSO	0.8752
Gradient boosting	0.8745
Support vector machine	0.8556
Recursive partitioning and regression trees	0.8460
Random forest	0.8447
Stacking ridge regression	0.8579
Naïve Bayes	0.8312

Table 2. Summary of the p-values of the difference in AUCs between the top performing method (PLS-DA) and other methods.

Prediction method	*P-value*
Artificial neural network	0.4351
Logistic regression	0.2485
Elastic net	0.1702
Ridge regression	0.1090
LASSO	0.0296
Gradient boosting	0.0885
Support vector machine	0.1580
Recursive partitioning and regression trees	0.0318
Random forest	0.0025
Stacking ridge regression	0.03382
Naïve Bayes	0.0085

5. REFERENCES

[1] Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q., and Lillard, J.W. 2014. A study of effects of multicollinearity in the multivariable analysis. Int J Appl Sci Technol. 4, 5 (Oct. 2014), 9–19.

[2] Prunier, J.G., Colyn, M., Legendre, X., Nimon, K.F., and Flamand, M.C. 2015. Multicollinearity in spatial genetics: separating the wheat from the chaff using commonality analyses. Mol Ecol. 24, 2 (Jan. 2015), 263–83.

[3] Garg, A., Tai, K. 2013. Comparison of statistical and machine learning methods in modeling of data with multicollinearity. Int J Model Identif Control. 18, 4 (Jan. 2013), 295-312.

[4] Whalen, S., Pandey, G. 2013. A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics. Cs Q-Bio Stat [cited 2015 Jul 5]. http://arxiv.org/abs/1309.5047

[5] NHANES - National Health and Nutrition Examination Survey Homepage. [cited 2015 Mar 1]. http://www.cdc.gov/nchs/nhanes.htm

[6] Breiman, L. 1996. Stacked regressions. Mach Learn. 24, 1 (July 1996), 49–64.

[7] R for Mac OS X. [cited 2015 Jun 7]. http://cran.r-project.org/bin/macosx/