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

Gerard G. Dumancas[1*] and Ghalib Bello[2]

[*1]Oklahoma Baptist University, Shawnee, OK, USA 74804, email: gerard.dumancas@okstate.edu
[2]Virginia Commonwealth University, Richmond, VA, USA 23284, email: belloga@vcu.edu

Introduction

The problem of multicollinearity

- Big data analytics and high-performance data mining have become increasingly popular in various fields.
- A typical analytic scenario involves the use of thousands of highly correlated variables.
- Multicollinearity occurs when complex correlation patterns exists among variables, which leads to a number of undesirable consequences.

Previous studies in multicollinearity

- Previous studies have compared machine learning methods in modelling of data with multicollinearity but none have been applied directly to mortality prediction using lipid profile clinical data.
- Garg and Tai compared commonly used statistical methods such as stepwise regression, radial basis function partial least squares (PLS), partial robust M-regression, ridge regression (RR) and principal component regression (PCR) in handling multicollinearity problems.
- Singh and colleagues compared the performance of logistic regression (LR), artificial neural network (ANN), support vector machine (SVM) and decision tree in predicting fault proneness models.
- Similar to the latest study in 2013 by Whalen and Pandey, we compared the performance of several machine learning algorithms to determine their ability to handle multicollinearity.
- The insights gained from this study could be useful in selecting machine-learning methods for automated pre-processing of thousands of correlated variables in biomedical data mining.

Methods

- We compared the predictive performance of the following commonly used machine learning techniques for multicollinear data with a binary dependent variable as outcome: LASSO, Ridge Regression, Elastic Net, Partial Least Squares regression (PLS-DA), random forests, recursive partitioning and regression trees (RPART), gradient boosting, support vector machines, naive bayes, logistic regression, and artificial neural networks using various fine tuning parameters (Table 1).
- We also examined the added benefit of using stacked generalization, an ensemble learning technique, to improve predictive accuracy in the presence of multicollinearity (Figure 1).
- We utilized moderately-correlated lipid profile data and demographic data (age, gender, race/ethnicity) to predict a binary outcome (five-year mortality).
- The data was obtained from the National Health and Nutrition Examination Survey (NHANES). The lipid profile data consisted of: total cholesterol, high density lipoprotein-cholesterol and triglycerides.
- After pre-processing, the analytical dataset consisted of 726 individuals (605 controls/121 cases).
- We further divided this into ~70% training consisting of 483 individuals (403 controls/80 cases) and ~30% testing consisting of 243 individuals (202 controls/41 cases).
- The area under the ROC curve, or simply AUC, was used as a measure of the predictive performance of the various classification algorithms.

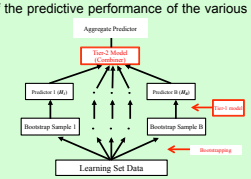

Figure 1. Schematic diagram of the stacking algorithm

Results

- Our results indicate that PLS-DA is the top-performing technique (Figure 2).
- Stacking did not significantly demonstrate superior performance compared to the top-performing individual predictors.
- The performance of PLS-DA as measured by AUC was found to be significantly different compared to LASSO, RPART, random forest, ridge regression-based stacking, and naïve bayes but was not found to be significantly different compared to the rest of the techniques (Table 2).

Table 1. Fine tuning parameters used in several machine learning techniques to achieve the highest AUC values in the test set.

Technique		
ANN	*Hidden layers*	2
	Decay parameter	0.06
Gradient boosting	*Number of trees*	100
	Interaction depth	2
SVM	*Gamma*	0.01
	Epsilon	1
PLS-DA	*Factors*	6
Elastic net	*Alpha*	0.5
Random forest	*Node size*	50
	Number of trees	10
Naïve Bayes	*Laplace*	0.01
Stacking using regularized generalized linear models	*Alpha*	0

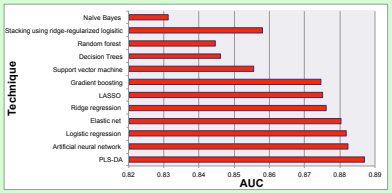

Figure 2. Comparison of AUC values among different machine learning techniques.

- The results identify a number of robust techniques for dealing with multicollinearity and are potentially useful in big data analytics problems in biomedical research wherein complex and non-negligible correlation patterns exist among variables.

Table 2. DeLong's test comparing AUCs to the top performing technique (PLS-DA, *AUC*=0.8870)

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**
RPART	**0.0318**
Random forest	**0.0025**
Stacking using ridge-regularized logistic regression	**0.03382**
Naïve Bayes	**0.0085**

Conclusions

- Our results show that the use of ensemble learning technique may not always improve the performance of a predictive model and that the use of PLS-DA maybe the more preferred choice over the other algorithms.
- Insights gained from this study could be used to design methods for automated pre-processing of thousands of variables used for big data analytics and high performance data mining.
- Our results will also provide essential guidelines for predictive modeling in biomedical data applications with complex correlation patterns.
- We provide a comprehensive assessment by comparing the predictive abilities of 12 machine learning techniques that can be implemented for automated pre-processing of correlated variables in high performance data mining and big data analytics

Bibliography

[1]. Yoo W, Mayberry R, Bae S, Singh K, Peter He Q, Lillard JW. A Study of Effects of MultiCollinearity in the Multivariable Analysis. Int J Appl Sci Technol. 2014 Oct;4(5):9–19.
[2]. Prunier JG, Colyn M, Legendre X, Nimon KF, Flamand MC. Multicollinearity in spatial genetics: separating the wheat from the chaff using commonality analyses. Mol Ecol. 2015 Jan;24(2):263–83.
[3]. Garg A, Tai K. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. Int J Model Identif Control [Internet]. 2013 Jan 1 [cited 2015 Mar 1];18(4):295–312. Available from: http://dx.doi.org/10.1504/IJMIC.2013.053535
[4]. Singh Y, Kaur A, Malhotra R. Comparative analysis of regression and machine learning methods for predicting fault proneness models. Int J Comput Appl Technol [Internet]. 2009 Jan 1 [cited 2015 Mar 6];35(2):183–93. Available from: http://dx.doi.org/10.1504/IJCAT.2009.026529
[5] Whalen S, Pandey G. A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics. ArXiv13095047 Cs Q-Bio Stat [Internet]. 2013 Sep 19 [cited 2015 Jul 5]; Available from: http://arxiv.org/abs/1309.5047
[6]. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Lippincott Williams & Wilkins; 2008. 776 p.
[7]. Buring JE. Epidemiology in Medicine. Lippincott Williams & Wilkins; 1987. 406 p.
[8]. Hennessy S, Bilker WB, Berlin JA, Strom BL. Factors influencing the optimal control-to-case ratio in matched case-control studies. Am J Epidemiol. 1999 Jan 15;149(2):195–7.