Assessing genetic algorithms for variable selection in predictive modeling based on classification: Comparing loss functions and internal models through a simulation study

Catherine M. Bain[1] & Dingjing Shi[1]

[1]Department of Psychology, University of Oklahoma, Norman, OK

Background

- We are seeing an increased importance of variable selection as the size of datasets continue to grow in the behavioral sciences.
- Reasons for variable selection include reducing the burden of data collection, increasing model efficiency, and increasing model generalizability.
- Genetic algorithms are becoming more common methods for variable selection in the behavioral sciences.
- One of the main advantages of genetic algorithms is their adaptability, yet there is little research available comparing the predictive performance of different genetic algorithms in terms of classification models and loss functions.

Methods

Variable Selection Technique

- Genetic Algorithm[4,5]

Classification Models

- Logistic Regression[9]
- Support Vector **Machine (SVM)**[2]
- Random Forest[3]
- Naïve Bayes Classifier[2]

Loss Functions

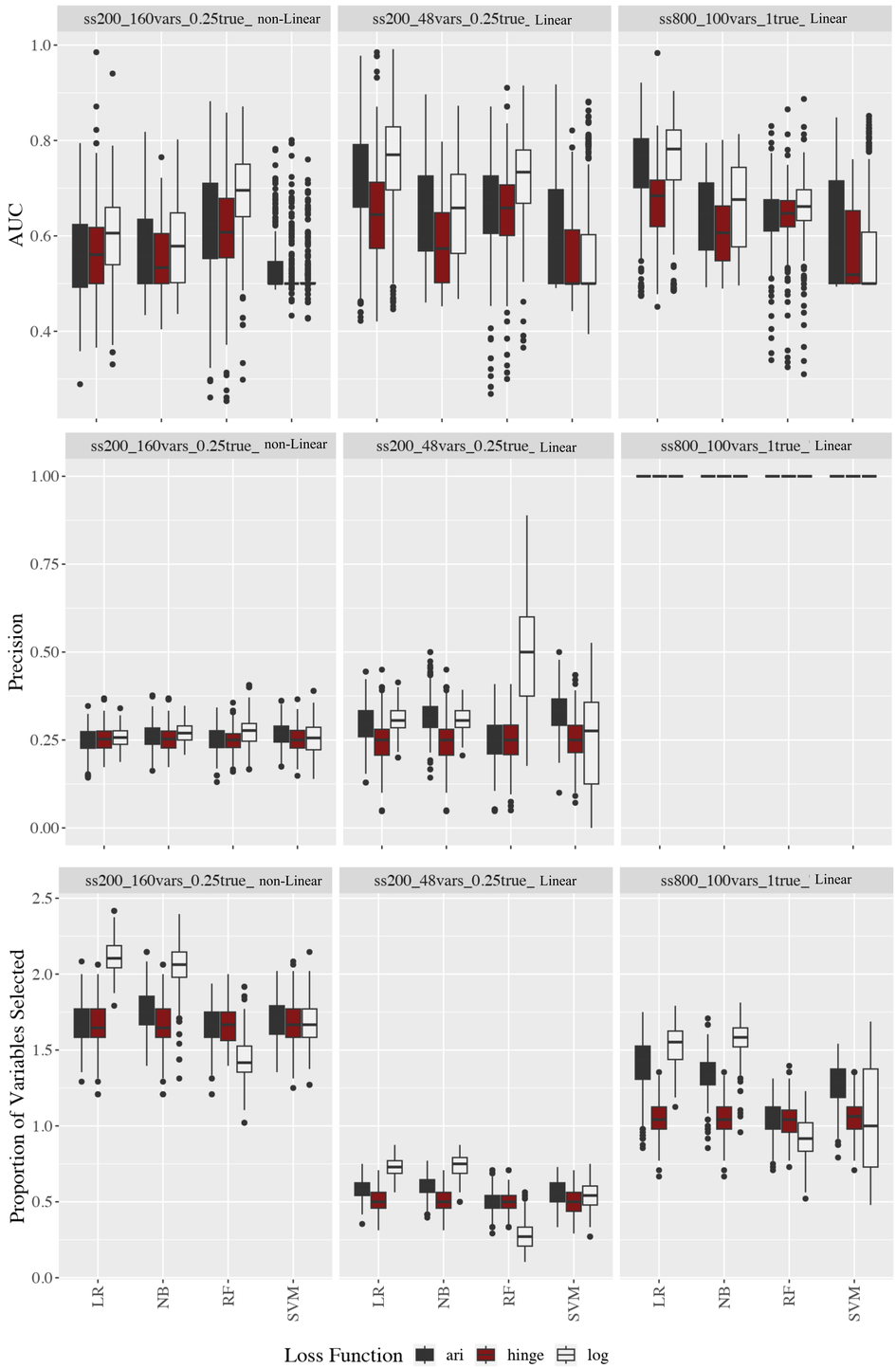
- Adjusted Rand Index[6]
- Hinge Loss[7]
- Log Loss[8]

Simulation Conditions

Data Feature	Levels
Sample Size	200 800
Total Variables	48 100 160
Proportion of True Variables	0.25 0.50 1.00
Linearly Separable Outcome	Yes No

Table 1. This table contains the various data conditions used in our Monte Carlo simulation

Results



Discussion

Classification

- Random forest with log loss consistently provided models with high average AUC and was robust to changes in the data.
- SVM, regardless of loss function, performed poorly, especially in conditions where the number of variables was close to the number of observations
- Naïve Bayes had the best performance with log loss, but not as well with hinge loss.

Variable Selection

- Precision, on average, matched the proportion of true variables in the dataset.
- Random forest with log loss had consistently better precision than chance.

Proportion of Variables Selected

- Random forest with log loss was the most conservative (i.e., always chose the fewest number of variables).
- ARI as the loss function contributed to unstable proportion of variables selection, regardless of internal models.
- Hinge loss, regardless of method, always selected about 50% of items.

Computation Time

- No methods had prohibitive computation times (average of 16 minutes)
- Naïve bayes classifier took the longest (average of 34 minutes)
- Random forest had the second longest computation time, but was rewarded in an increased AUC.

Conclusion

The performance of the genetic algorithm as a variable selection technique is dependent upon the classification algorithm used and the chosen loss function. Our results indicate that if one is using a genetic algorithm for variable selection in a classification problem, we recommend random forest with log loss as it produced models with highest AUCs in most conditions.

References and Contact Information

[2] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2022). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-11, https://CRAN.R-project.org/package=e1071. [3] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22. [4] Luca Scrucca (2013). GA: A Package for Genetic Algorithms in R. Journal of Statistical Software, 53(4), 1-37. https://doi.org/10.18637/jss.v053.i04 [5] Luca Scrucca (2017). On some extensions to GA package: hybrid optimisation, parallelisation and islands evolution. The R Journal, 9/1, 187-206. https://doi.org/10.32614/RJ-2017-008 [6] Vovk, V. (2015). The Fundamental Nature of the Log Loss Function. In L. D. Beklemishev, A. Blass, N. Dershowitz, B. Finkbeiner, & W. Schulte (Eds.), *Fields of Logic and Computation II: Essays Dedicated to Yuri Gurevich on the Occasion of His 75th Birthday* (pp. 307–318). Springer International Publishing. [7] Xu, Y., Akrotirianakis, I., & Chakraborty, A. (2016). Proximal gradient method for huberized support vector machine. *Pattern Analysis and Applications, 19*(4), 989–1005. https://doi.org/10.1007/s10044-015-0485-z [8] Vovk, V. (2015). The Fundamental Nature of the Log Loss Function. In L. D. Beklemishev, A. Blass, N. Dershowitz, B. Finkbeiner, & W. Schulte (Eds.), *Fields of Logic and Computation II: Essays Dedicated to Yuri Gurevich on the Occasion of His 75th Birthday* (pp. 307–318). Springer International Publishing. https://doi.org/10.1007/978-3-319-23534-9_20 [9] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.