CAN NEURAL NETWORKS LEAD TO IMPROVED FORECASTS OF NCAA FOOTBALL?

David Paul, Valiant Pharmaceuticals, Irvine, CA
Samuel L. Seaman, Owen P. Hall, and Andy Sevastopoulos, The Graziadio School of Business & Management,Pepperdine University, Malibu, CA 90263, 310-506-4401

ABSTRACT

Anyone with an interest in college football has probably tried to predict the future for a favorite team. And now, help is just around the corner with a horde of "football prophets" – in the media or popular press, on the internet, or in the statistical literature – ready to assist with that forecast. Recognizing the popularity of the sport and the enthusiasm for prediction of sporting events in general, we have used NCAA football data to compare the classification accuracies of three promising multivariable predictive techniques - neural networks (NN), logistic regression (LR), and linear discriminant analysis (LDA) - when predicting whether or not a team will have a winning season. The current investigation offers an interesting departure from the usual approaches, in that we have applied principal components analysis to reduce dimension before employing the competing predictive models. Whilst only small differences have been observed in predictive accuracies amongst the three techniques, we have noticed an interesting link between a popular estimator of "shrinkage" and the estimated error rate for the logistic regression model?

INTRODUCTION

A 2002 Harris poll suggested that roughly 35% of adults in this country follow college football. Not surprisingly, efforts to forecast various outcomes in the sport have flourished. Football clairvoyants have used everything from intuition to sophisticated dynamic hierarchical Bayesian models to predict winners, point spreads, champions, even the particular play that Notre Dame will use on a given down (http://controls.ame.nd.edu/football/). It is difficult to review and compare predictive accuracies obtained in previous studies of this sort, since the outcome variables used, have been very different across investigations. In the present study, we are not so much interested in finding a "best" forecasting model as we are in comparing three important classification techniques, after dimension reduction, all of them having enormous potential for forecasting the future in a variety of practical applications [8]. For simplicity then, we have chosen to predict a rather innocuous, binary outcome - whether or not a college football team will have a winning season.

Historically, investigators trying to predict a dichotomous outcome in sport, as has been done here, generally have applied either a Logistic Regression model or a Linear Discriminant Analysis [2] [3] [4] [10]. These commonly used procedures, however well known, make certain assumptions about the data (e.g. normality, independence, homogeneity of covariance structures). If the data being analyzed violate those assumptions, the procedures may produce solutions that are not optimal [10]. This has led to an heightened interest in non-parametric approaches to the classification problem, like artificial neural networks, which enjoy tremendous popularity, say advocates, because they require fewer assumptions of the data. The aim of the present investigation, then, is to illuminate for football prophets everywhere, the idiosyncratic effects of football data on the predictive accuracies of discriminant, logistic, and neural network models.

RESEARCH METHODOLOGY

Data

We have used football statistics, available at the NCAA website, for school years 1999 through 2004. Our dependent variable in each year - whether or not a team had a winning season – was created using a winning percentage of 50% or greater as the definition of success. The possible predictor variables that could be used in predictive models like ours are quite numerous. Some of those that have been used, and available at the NCAA website, are: Games Played, Carries, Rushing Yards, Rushing Average, Rushing TD's, Pass Attempts, Pass Yards, Yards/Pass Attempt, Passing TD's, Interceptions %, Yards/Completion, Plays Run, Total Yards, Average yards/Play, Total TD's, Carries, etc... With so many possible independent variables, and given our lack of expertise of any kind with college football, we have been forced to use a dimension reduction algorithm before applying the competing predictive models (we chose not to rely on the intuitive selections made by coaches, sports enthusiasts, or fans). Harrell et al. [5], have warned that p-value based variable selection algorithms can lead to noisy predictive models and, perhaps more troubling, models that contain less than half of all "authentic" predictors. They recommend instead, that an alternative strategy like principal components analysis, be used to reduce dimension. We have followed that advice and have identified, for each season, only the most "important" principal components - based upon estimated eigenvalues. In each season, the first 9 principal components explain roughly 89% of the total variance, with unremarkable contributions being made by any of the remaining principal components.

Topology of the Neural Network

The Neural Network models used in our analyses (one for each season) were of the back-propagation type (initialization parameter values having been determined by trial and error) with fully connected input, hidden, and output layers. Logistic activation functions were employed for all nodes in both the hidden and output layers of each network. A linear function of the "essential" Principal Component scores (obtained in the dimension reduction phase of the analysis) served as the input layer in our models. To minimize the possibility of an upwardly-biased classification rate [1], we restricted the hidden layer to only three neurons for each model. Finally, a single neuron output layer was trained to produce normalized values (0 or1) representing success (winning season) or failure (losing season) for each model.

RESULTS AND CONCLUSIONS

Having carefully reviewed the literature on comparative analyses of classification algorithms, much like the one described here, we were quite convinced that the Artificial Neural Network model would outperform both Discriminant Analysis and Logistic Regression [9]. The estimated leave-one-out error rates [7] for all procedures compared in this study, are remarkably similar. We were at first discouraged by our results, yet upon further reflection, realized that most previous investigations have used a p-value based variable selection algorithm to reduce dimension. Given Harrell's sharp criticism of such procedures [5], we believe it is very possible that our use of principal components may have contributed to improved efficiencies of the LDA and LR models? Since we have not, as yet, performed an exhaustive simulation to test this notion, we are reluctant to draw any such conclusion. And, in deference to Michie et. al. [8], we would be equally uncomfortable about generalizing our results to other types of problems or other kinds of data sets. Rather, we suggest that Discriminant Analysis and

Logistic Regression not be dismissed, regularly, in favor of neural networks simply because of concerns about the assumptions required of the parametric models.

REFERENCES

[1] Baum, E.B., & Haussler, D. (1989). What size net gives valid generalization? From *Advances in Neural Information Processing Systems, Volume 1*. Kauffmann: San Mateo, CA.

[2] Boynton, B. (1997). Are one-run games special? *Baseball research Journal, 26.*

[3] Breiter, D. & Carlin, B. (1997). How to play office pools if you must. *Chance, 10(1),* pp. 5-11.

[4] Goddard, J. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting, 23(1),* pp. 51-66.

[5] Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., and Rosati, R.A. (1984). Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine, Volume 3,* pp. 143-152.

[6] Krane, D.K. (October 4, 2002). Notre Dame's Fighting Irish Remain America's Favorite College Football Team; Fewer people following college football. *Harris Poll #52.*

[7] Lachenbruch, P. (1967). An almost unbiased method for obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics,* 23, pp. 639-645.

[8] Michie, D., Spiegelhalter, D.J., & Taylor, C.C. (1994). *Machine learning, neural and statistical classification.* New York: Ellis Horwood.

[9] Nanda, S., & Pendharkar, P.C. (2001). Development and comparison of analytical techniques for predicting insolvency risk. *International Journal of Intelligent Systems in Accounting, Finance, and Management, 10, pp. 155-168.*

[10] Stern, H. S. (1998). Football strategy: go for it! *Chance, 11(3),* pp. 20-24.

[11] VanHouwelingen, J.C., & LeCessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine, 8,* pp. 1303-1325.