Achieving Accurate Prediction Models: Less is Almost Always More

John D. Morris **Mary G. Lieberman**
Florida Atlantic University

Accurate cross-validated prediction accuracy is posited as the ultimate criterion for prediction model performance. This study investigates and demonstrates, across a wide variety of data sets, the nearly ubiquitous benefit to classification model accuracy of optimal subset selection. Unlike popular "stepwise" methods often used (and abused) in the literature, this study considers only all-possible-subset cross-validated performance as the criterion of accuracy. The superiority of variable subsets is demonstrated for predictive discriminant analysis and logistic regression. Computer programs are also made available.

Among the techniques used for solving classification problems, logistic regression (LR) and predictive discriminant analysis (PDA) are two of the most popular (Yarnold, Hart & Soltysik, 1994). Unlike PDA, LR captures the probabilistic distribution embedded in a categorical outcome variable, avoids violations to the assumption of homogeneity of covariance matrices (in the case of the linear PDA model), and does not require strict multivariate normality. Therefore, when PDA assumptions are violated, we might expect greater cross-validated classification accuracy with LR than PDA.

Although several studies have compared the classification accuracy of LR and PDA, the results have been inconsistent. For example, some studies (Baron, 1991; Bayne, Beauchamp, Kane, & McCabe, 1983; Crawley, 1979) suggest that LR is more accurate than PDA for nonnormal data. However, several researchers (e.g., Cleary & Angel, 1984; Knoke, 1982; Krzanowski, 1975; Lieberman & Morris, 2003; Meshbane & Morris, 1996; Press & Wilson, 1978) found little or no difference in the accuracy of the two techniques with PDA often performing better than LR. Part of the reason these results are in dispute is that one may look at accuracy for all groups or separate-groups. As well, one may consider a cross-validated index of accuracy or the accuracy of reclassifying the calibration sample; these studies are not consistent in respect to the criterion of accuracy used. Specifically, examination of cross-validation accuracy in LR studies is uncommon, and when done is usually of the most basic (also non-unique and unstable) sort (hold-out sample). No computer packages support more appropriate resampling cross-validation methods (variously called PRESS, Lachenbruch *U*, leave-one-out, jackknife and bootstrap).

Whichever method (LR or PDA) is selected, one may consider subsets of all possible variables for purposes or parsimony, and/or to increase cross-validation accuracy of the model (Morris & Meshbane, 1995). The most usual method is to consider accuracy in classification of the sample upon which the model is created (internal) with the objective of parsimony. That is, realizing that some accuracy will be lost in reducing the number of predictor variables in classifying the calibration sample, but compromising that loss with the gain in parsimony afforded by the reduction in size of the prediction model. However, as in multiple regression, an increase in cross-validated prediction accuracy (the most appropriate criterion) is almost always available using a model composed of fewer than all variables available. Thus one may gain both parsimony and some degree of explanatory power for the model. In addition, although traditional methods considering the piecemeal change in performance of models in respect to prediction within the calibration sample have often been used (forward, backward, stepwise, or variants thereof), they are neither optimal, nor unique and are now generally in disfavor.

In the case of PDA an examination of the cross-validation accuracy of all 2^p-1 (where p is the number of predictor variables) subsets of variables has been recommended and utilized (Huberty, 1994; Huberty & Olejnik, 2006; Morris & Meshbane, 1995). In this case the method of cross-validation is the leave-one-out method. In the leave-one-out procedure (Huberty, 1994, *p*. 88; Lachenbruch & Mickey, 1968; Mosteller & Tukey, 1968) a subject is classified by applying the rule derived from all subjects except the one being classified. This process is repeated round-robin for each subject, with a count of the overall classification accuracy used to estimate the cross-validated accuracy. [Clearly the same round-robin procedure can be used to estimate either relative or absolute accuracy in the use of multiple regression and has appeared in that context, with perhaps the earliest reference due to Gollob (1967). In a system intended to select optimal multiple regression predictor variable subsets, Allen (1971) coined the procedure PRESS, and he appears to be the source most often cited in the multiple regression literature.]

Table 1. Data set, number of predictor variables (p), PDA hit-rate for all p variables, number of variables in the best performing subset(s), hit-rate for that subset, and the % change in hit-rate.

#	Data Set Source	p	Hit-rate for p Predictors	# Predictors in Best Subset(s)	Max Hit-rate	% Change
1	Rulon Grps 1 & 2	4	.809	3	**.831**[a]	2.72
2	Rulon Grps 1 & 3	4	.927	3	.934	.75
3	Rulon Gps 2 & 3	4	.830	3	**.836**	.72
4	Block - Grps 1 & 2	4	.679	2	**.743**	9.43
5	Block - Grps 1 & 3	4	.646	4	**.646**	0.00
6	Block - Grps 1 & 4	4	.603	1	**.667**	10.61
7	Block - Grps 2 & 3	4	.553	1	**.632**	14.29
8	Block - Grps 2 & 4	4	.600	2	**.640**	6.67
9	Block - Grps 3 & 4	4	.684	3	**.711**	3.95
10	Demographics	8	.581	3,6	.613	5.51
11	Dropout from 4th	10	.702	2,3,4,5	.787	12.11
12	Dropout from 8th	11	.739	5,6	**.803**	8.66
13	Fitness	10	.588	7	.616	4.76
14	Warncke -Grps 1 & 2	10	.482	2	**.607**	25.93
15	Warncke -Grps 1 & 3	10	.571	3,5,6	.657	15.06
16	Warncke -Grps 2 & 3	10	.402	1,5	**.575**	43.03
17	Bisbey 1& 2	13	.888	5,6,7,8,9,10	.914	2.93
18	Bisbey 2& 3	13	.839	2,4,5,6	**.983**	17.16
19	Talent - Grps 1 & 3	14	.578	7	.698	20.76
20	Talent - Grps 3 & 5	14	.772	8.9	.835	8.16
21	Talent - Grps 1 & 5	14	.746	5,7,8	.797	6.84

[a] Bold when > than LR.

In the case of PDA (and regression) a matrix identity due to Bartlett (1951) allows the task of the requisite *N*-1 matrix inversions to be accomplished with far less computational labor that would otherwise be necessary. However, this mathematical tool is irrelevant to the iterative method of LR optimization, thus *N*-1 LR optimizations must be completed for each of 2^p-1 subsets of predictor variables.

Unlike most LR studies that consider calibration sample statistics as the criterion for model fit (e.g., the Cox & Snell, or Nagelkerke R^2), the criterion for model accuracy is construed in this study, as is typically the case in PDA, as classification accuracy. That is, the proportion of correct leave-one-out cross-validated classifications (hit-rate) for the total sample and each separate group. Thus for a two-group problem, we may order the accuracy of our 2^p-1 candidate LR equations according to three different (total sample and each group) cross-validated classification accuracy criteria.

Method

Analyses from 21 two-group classification problems from Morris and Huberty (1987) were used to illustrate the method and computer program for PDA (Table 1) and LR (Table 2). Although not purported to represent all potential data structures, these data sets have been used in several classification studies as representing a wide variety of number of predictor variables, group separation, and covariance structures. As the number of predictors ranges from 4 to 14, the candidate 2^p-1 cross-validated subsets range from a very modest 15 to 16,383 for the 14 predictor variable problem. However, even in the case of the calculation and sorting of the16K+ cross-validated classification performances, the program executes (on a midrange laptop) in less than 30 seconds.

Result and Conclusions

In the case of both PDA (Table1) and LR (Table 2), one can see that, in all cases, except #5 in PDA, selection of the best performing subset (of the 2^p–1 possibilities) offers a reduction in the number of predictor variables, often by more than half, thus parsimony is well served. One may also note that,

Table 2. Data set, number of predictor variables (p), LR hit-rate for all p variables, number of variables in the best performing subset(s), hit-rate for that subset, and the % change in hit-rate.

#	Data Set Source	p	Hit-rate for p Predictors	# Predictors in Best Subset(s)	Max Hit-rate	% Change
1	Rulon Grps 1 & 2	4	0.803	3	.815	1.49
2	Rulon Grps 1 & 3	4	0.914	3	.934	2.19
	Rulon Gps 2 & 3	4	0.824	3	.830	0.73
4	Block - Grps 1 & 2	4	0.692	1,2	.718	3.76
5	Block - Grps 1 & 3	4	0.620	3,4	.620	0.00
6	Block - Grps 1 & 4	4	0.577	1,2	.628	8.84
7	Block - Grps 2 & 3	4	0.566	1,2	.605	6.89
8	Block - Grps 2 & 4	4	0.587	2	.627	6.81
9	Block - Grps 3 & 4	4	0.684	3	.697	1.90
10	Demographics	8	0.591	4	**.620**[a]	4.91
11	Dropout from 4th	10	0.660	4	.787	19.24
12	Dropout from 8th	11	0.725	3	.782	7.86
13	Fitness	10	0.591	4	**.620**	4.91
14	Warncke -Grps 1 & 2	10	0.446	1	.580	30.04
15	Warncke -Grps 1 & 3	10	0.600	4	**.667**	11.17
16	Warncke -Grps 2 & 3	10	0.425	2	.563	32.47
17	Bisbey 1& 2	13	0.879	6,7,8,9,10	.914	3.98
18	Bisbey 2& 3	13	0.856	5,6,7	.924	7.94
19	Talent - Grps 1 & 3	14	0.621	5	**.733**	18.04
20	Talent - Grps 3 & 5	14	0.787	6,7,8,9	**.858**	9.02
21	Talent - Grps 1 & 5	14	0.740	5	.797	7.70

[a] Bold when > PDA.

particularly with larger models, multiple sets of predictors and size models often achieve maximum accuracy. In addition, one can see that due to the reduction in the number of predictor variables, cross-validation accuracy increased from less than 1% all the way up to more than 40%. Only in data set #5 (PDA & LR) did the reduced model perform the same as the full model. In the case of LR, still offering the same accuracy, but with increased parsimony, and in the case of PDA, offering no advantage. The mean increase in cross-validated hit-rate due to the reduction in the number of predictor variables over all 21 data sets was about 5% for LR and 10% for PDA. Thus one can have parsimony and increased accuracy. Through this procedure and computer programs, researchers will be able to make better decisions about optimally accurate classification model construction.

Although not the focus of this study, it is difficult to ignore potential comparisons between PDA an LR performance. As was stated, greater parsimony and accuracy is afforded in almost every case by selecting an optimally performing subset. As has been previously documented, cross-validation performance was often very close between PDA and LR. However, if one considers only the optimally performing subsets the advantage seems to go to PDA herein. PDA is best in 12 data sets, LR in 5, and performance is the same in 4.

Further consideration of the advantage of the availability of multiple optimally performing subsets should also be noted. Missing data is almost always a difficulty in dealing with real data. First, a model depending on a smaller number of variables has not only the philosophical advantage of parsimony, but may also afford the opportunity to accommodate missing data; there is more opportunity for the model to be applicable as the number of variables decreases. Moreover, if several equally performing (or nearly so) superior subsets are available, the opportunity to accommodate the missing data of an individual score vector is increased; one can use alternative models for alternate missing data configurations, unless, of course, that variable that is missing is including in all of the best subsets. Table 3 illustrates the top 20 (of 256 possibilities) subset accuracies for PDA prediction of dropout from high school from 8 predictors. One can see that the optimal subset contains 4 variables, but many subsets are close, such that performance is maintained. Such information can aid in handling missing data. That is, one might argue that if the four variables in the best performing model are available for a subject (SCHOOLS8, MATH8,

Table 3. Ranked 20 best (of 255) performing subsets, and total model.

	Variables Included in the Model:							
Hit-Rate	SCHOOLS8	REPEATS8	READING8	MATH8	LANG8	SCIENCE8	SOCST8	DSFS8
0.753	√			√		√		√
0.747	√			√				√
0.747	√							√
0.747	√				√	√	√	√
0.747	√			√	√	√	√	√
0.741	√		√		√		√	√
0.741	√			√		√	√	√
0.735	√		√		√	√		√
0.735	√			√			√	√
0.735	√	√	√				√	
0.735	√			√	√		√	√
0.735	√					√	√	√
0.735	√		√	√	√			√
0.735	√	√						√
0.735	√	√	√					
0.728	√	√				√	√	
0.728	√					√		√
0.728	√	√				√		√
0.728	√	√	√			√		√
0.728	√	√				√		
Total Model								
0.679	√	√	√	√	√	√	√	√

Note: SCHOOLS8: Accumulated # of schools attended by grade 8.
REPEATS8: Accumulated # of Grades repeated by grade 8.
READING8: 8th Grade Reading grade.
MATH8: 8th Grade Math grade.
LANG8: 8th Grade Language grade.
SCIENCE8: 8th Grade Science grade.
SOCST8: 8th Grade Social Studies grade.
DSFS8: Accumulated # of D and F grades over all subjects by grade 8.

SCIENCE8, DSFS8) it should be used. However, if for some reason MATH8 and SCIENCE8 (as well as REPEATS8, READING8, LANG8, and SOCST8) are missing from a case, then a model that demonstrates essentially the same performance is available using only SCHOOLS8 and DSFS8. In this case, the SCHOOLS8 is the number of schools the child had attended by the 8th grade and DSFS8 is the number of "D" and "F" grades the child had accumulated, whereas MATH8 and SCIENCE8 are grades in those specific subjects in the 8th grade. So, as an example, for a teacher, or school not reporting subject grades, but the more "global" accumulated variables of SCHOOLS8 and DSFS8 are retained in the county database, the alternate model could be used with the expectation of attaining essentially the same accuracy.

Note, however, in this case, that if SCHOOLS8 is not available, optimal accuracy appears improbable. It is a "don't leave home without it" variable. The programs used herein (one for PDA and one for LR) for the examination and ordering of the 2^p-1 possible subsets of predictor variables by their leave-one-out accuracy are available from the senior author at: jdmorris@fau.edu. They are available as Intel based EXE files (compiled from FORTRAN).

References

Allen, D. A. (1971). *The prediction sum of squares as a criterion for selecting predictor variables* (Tech. Rep. No. 23). Lexington: University of Kentucky, Department of Statistics.

Baron, A. E. (1991). Misclassification among methods used for multiple group discrimination – The effects of distributional properties. *Statistics in Medicine*, *10*, 757-766.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *Annals of Mathematical Statistics*, *22*, 107-111.

Bayne, C. K., Beauchamp, J. J., Kane, V. E., and McCabe, G. P. (1983). Assessment of Fisher and logistic linear and quadratic discrimination models. *Computational Statistics and Data Analysis*, *1*, 257-273.

Cleary, P. D. & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variables. *Journal of Health and Social Behavior*, *25*, 334-348.

Crawley, D. R. (1979). Logistic discriminant analysis as an alternative to Fisher's linear discriminant function. *New Zealand Statistics*, *14*(*2*), 21-25.

Gollub, H. F. (1967, September). *Cross-validation using samples of size one*. Paper presented at the annual meeting of the American Psychological Association, Washington, D. C.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. New York: Wiley.

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.

Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics*, *38*, 191-200.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, *70*, 782-790.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, *10*, 1-11.

Lieberman, M. L, & Morris, J. D. (2003, April). *Comparing classification accuracies between predictive discriminant analysis and logistic regression in specific data sets*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Meshbane, A., & Morris, J. D. (1996, April). *Predictive discriminant analysis versus logistic regression in two-group classification problems*. Paper presented at he meeting of the American Educational Research Association, New York.

Morris, J. D., & Huberty, C. J (1987). Selecting a two-group classification weighting algorithm. *Multivariate Behavioral Research*, *22*, 211-232.

Morris, J. D., & Meshbane, A. (1995). Selecting predictor variables in two-group classification problems. *Educational and Psychological Measurement*, *55*, 438-441.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*, 699-705.

Mosteller, F. & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.). *Handbook of social psychology* (Vol. 2, pp. 80-203). Reading, MA: Addison-Wesley.

Yarnold, P. R., Hart, L. A. & Soltysik, R. C. (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analysis. *Educational and Psychological Measurement*, *54*, 73-78.

Send correspondence to: John D. Morris
Florida Atlantic University
Email: jdmorris@fau.edu