

ENSEMBLE METHOD HIT RATIO FOR ROBUST TESTS OF SPREAD

ABDUL RAHMAN OTHMAN
School of Distance Education
University Science Malaysia
oarahman@usm.my

TEH SIN YIN
School of Distance Education
University Science Malaysia
syin.teh@gmail.com

An ensemble method was used to combine the outputs of several diverse classifiers to form a potentially stronger solution in a classification system. The SAS system facilitates the building of a program that can perform a composite method that combined logistic regression and discriminant analysis by using PROC LOGISTIC, PROC DISCRIM, DATA step, PROC FREQ and other SAS functions. The aim is to improve the correct classification rate from the same data set. To achieve this, classification from logistic regression and discriminant analysis were integrated to form an ensemble. The averages of posterior probabilities (for the target values) were taken from logistic regression and discriminant analysis, and classified records according to their average posterior probabilities. Then, the classification tables were created by defining predicted values based on the prior probability from the average posterior probabilities. The intended audiences for this paper are those who have working knowledge of Base SAS and have the fundamental grasp of statistics.

Keywords: Ensemble method, posterior probabilities, leave-one-out (L-O-O) classification, hit ratio.

1. Introduction

A database consisting of p -values and attendant information for tests of spread procedures was made available from Keselman, Wilcox, Algina, Othman and Fradette (in press). An ensemble method that combines logistic regression and discriminant analysis was used to determine the importance of the simulation conditions for robust test of spread procedures in generating ideal p -values. In other words, this technique was used to evaluate particular simulation conditions that could give robust Type I error rates. These at rates fell in [0.045, 0.050].

Essentially, the dependent variable have two values, 1 representing p -values falling in [0.045, 0.050] and 0 for p -values falling outside of this interval. And, there were 24 dummy independent variables with 25,257 observations. The preliminary run on logistic regression for the training data set showed that with this particular variables structure, there were zero parameter estimates. This was a sign of the presence of multicollinearity. However, the collinear variables might be still relevant to the model. A possible solution to this problem was to restructure the data by redefining the variables. The 5 new categorical independent variables after restructuring the data were presented in Table 1. They represented the simulation conditions in this study. The *SHAPE* and *TAIL* variables represented the skewness and kurtosis properties of a distribution, respectively. Hence they were actually fixed in the distributions. There were 7 distributions simulated in Keselman, et al. (in press), they were

- 1) The Fleishman (1978) transformation of the standard normal distribution into a skewed platykurtic distribution with skewness, $\gamma_1 = 0.5$ and kurtosis, $\gamma_2 = -0.5$.
- 2) A second Fleishman transformation of the standard normal distribution into a skewed normal-tailed distribution with $\gamma_1 = 0.75$ and $\gamma_2 = 0$.

- 3) The Beta (0.5, 0.5) distribution representing symmetric platykurtic distributions with $\gamma_1 = 0$ and $\gamma_2 = -1.5$.
- 4) A g and h distribution (Hoaglin, 1985) where $g = h = 0$. This is the standard normal distribution with $\gamma_1 = \gamma_2 = 0$.
- 5) A $g = 0$ and $h = 0.225$ long-tailed distribution with $\gamma_1 = 0$ and $\gamma_2 = 154.84$, representing symmetric leptokurtic distributions.
- 6) A $g = 0.76$ and $h = -0.098$ distribution with $\gamma_1 = 2$ and $\gamma_2 = 6$, representing skewed leptokurtic distribution.
- 7) A $g = 0.225$ and $h = 0.225$ distribution. This is also a long-tailed skewed leptokurtic distribution ($\gamma_1 = 4.9, \gamma_2 = 4673.8$), but more severe than (6).

The product of the 7 levels of *DISTR*, 2 levels of *Gsize* and 3 levels of *GSCOND* resulted into 42 combinations of the 5 new independent variables. For each of the 42 combinations, the number of records in group 0 and group 1 were counted for dependent variable (*PVAL05*). Hence, there were 42 combinations multiplied by 2 levels of *PVAL05* equaling 84 records. The total number of counts for the 84 records will be 25,257. The validation data set was similarly restructured. However, being a different data set from the training data set, the count will be different. The total number of counts for the 84 records in validation data set will be 1,329. This meant that at the end of aggregation process, there were two data sets ready for analysis. Logistic regression, discriminant analysis, and a composite method that combined the two methods mentioned earlier, were performed on these data sets.

Table 1: Categorical independent variables available for entry

Variable	Variable Label	Level	Level Label
DISTR	Type of distribution	BETA(0.5,0.5)	Symmetric platykurtic
		FLEISHMAN1	Skewed platykurtic
		FLEISHMAN2	Skewed normal-tailed
		G=.225/H=.225	Skewed leptokurtic (severe)
		G=.76/H=-.098	Skewed leptokurtic
		G=0/H=.225	Symmetric leptokurtic
		N(0,1)	Standard normal
SHAPE	Skewness of distribution	SKEW	Skewed
		SYMM	Symmetric
TAIL	Kurtosis of distribution	LEPT	Leptokurtic
		PLAT	Platykurtic
		NORM	Normal
Gsize	Total group size	120	N=120
		60	N=60
GSCOND	Group size increments	INCR05	Increment of 5
		INCR10	Increment of 10
		EQUAL	Equal sample size

In Section 2, the leave-one-out (L-O-O) classification was defined. Prior and posterior probabilities were used to create the L-O-O correct classification tables were described in Section 3, The algorithm of the ensemble method was given in Section 4, followed by results and discussion in Section 5. And, the conclusion is in the final section.

2. Leave-One-Out (L-O-O) classification

The L-O-O method represents a special case of the cross-validation technique (Gong, 1986). Given n cases available in a data set, a classifier is trained on $(n-1)$ cases and then is tested on the case that was left out (Huberty, 1994; Johnson & Wichern, 2002). This process is repeated n times until every case in the data set have been included once as a cross-validation instance. The results are averaged across the n test cases to estimate the classifier's prediction performance (Lachenbruch, 1975).

Although the L-O-O was initially developed as a validation technique, it was subsequently used commonly as a more believable estimator of the correct classification rate for data that were not normal and homogenous.

3. Prior and Posterior Probabilities

The prior probability is an estimate of the likelihood that a case belongs to a particular group when no information about it is available. In this study, the data set showed group 0 with more observations than group 1. Thus, prior probability was set to be proportional to the size of the groups, i.e. 0.07 (1948/26586). This was arrived at looking at the proportion of the group 1 responses in the preprocessed data set.

The posterior probability is the probability that an observation belongs to a particular group, based on the knowledge of the values of variables collected for it for this purpose. The probability that a case belongs to a particular group is basically proportional to the Mahalanobis distance of each observation from that group centroid. The observation is assigned to the group to which it is the closest as measured by the Mahalanobis distance. In fact, in discriminant analysis, a case is assigned to the most likely group (posterior probability is the largest) based on its discriminant score (Tatsuoka, 1988). These posterior probabilities were then used to create the correct classification table.

4. Algorithm

The algorithm of the ensemble process is as follows:

- 1) Get the posterior probabilities derived from L-O-O of logistic regression for pval05.
- 2) Get the posterior probabilities derived from L-O-O of discriminant analysis for pval05.
- 3) Merge the posterior probabilities from both analyses.
- 4) Take the averages of posterior probabilities for the target values.
- 5) Define predicted values based on the prior probability from the average posterior probabilities.
- 6) Created the classification tables.

Each step is actually performed via separate DATA step or PROC FREQ step. The system was designed in this way so that each step could run separately or as an entire process. SAS codes will be made available upon request.

5. Results and discussions

Table 2 consisted of six 2×2 contingency tables of observed pval05 versus predicted grouping (pred_pval). Rows represent method and columns represent training/validation data set. The correctly-classified observations appeared on main diagonal of each 2×2 table. For the ensemble method on training data set (refer to row 'Ensemble' and column 'Training' in Table 2), 80.75% observations in group 1 was correctly classified into group 1. In the validation data set, 79.55% observations in group 1 correctly classified into group 1. The ensemble method combined the outcomes from logistic regression and discriminant analysis. Each classifier will make a different error, and by combining these classifiers, total error can be reduced. The ensemble classification indicated that the percentage of observations in group 1 was correctly classified into group 1 in ensemble method was higher than logistic regression and discriminant analysis. This meant that the ensemble were more accurate than its component classifiers.

Looking at the correct classification rate group by group, the results revealed that for group 1, logistic regression produced higher rate than discriminant analysis and vice versa for group 0 (Table 2). Due to the exceedingly high correct classification rate of discriminant analysis in group 0, the overall percentage of correct classification for discriminant analysis was also subsequently higher (Table 3). Therefore, an ensemble method was performed.

From Table 3, the ensemble method gave training data set 49.61% of hit ratio, and validation data set 52.15% of hit ratio. Noticed that the hit ratio for ensemble method were less than the hit ratios for logistic regression and discriminant analysis. Unfortunately, the hit ratio for ensemble method was actually too low (lower than 80%) to conclude that model fitted in data well. And, caution was needed as the good prediction was actually for group 0. However, these were the best unbiased estimates of the correct classification rates that could obtained with the present set of data. Hence, based on the percentages of correct classification by group and the hit ratio, the best method was based on logistic regression method.

Table 2: Percentages of Correct Classification by Groups for Training and Validation Data Sets of Simulation Conditions

Method	Training				Validate			
	pval05	pred_pval		Number of Cases	pval05	pred_pval		Number of Cases
		0	1			0	1	
Logistic Regression	0	13821	9576	23397	0	768	473	1241
		59.07	40.93	100.00		61.89	38.11	100.00
	1	570	1290	1860	1	33	55	88
		30.65	69.35	100.00		37.50	62.50	100.00
	Total	14391	10866	25257	Total	801	528	1329
		56.98	43.02	100.00		60.27	39.73	100.00
Discriminant Analysis	0	20914	2483	23397	0	1128	113	1241
		89.39	10.61	100.00		90.89	9.11	100.00
	1	1310	550	1860	1	69	19	88
		70.43	29.57	100.00		78.41	21.59	100.00
	Total	22224	3033	25257	Total	1197	132	1329
		87.99	12.01	100.00		90.07	9.93	100.00
Ensemble Method	0	11028	12369	23397	0	623	618	1241
		47.13	52.87	100.00		50.20	49.80	100.00
	1	358	1502	1860	1	18	70	88
		19.25	80.75	100.00		20.45	79.55	100.00
	Total	11386	13871	25257	Total	641	688	1329
		45.08	54.92	100.00		48.23	51.77	100.00
Priors	0.93	0.07	1.00	Priors	0.93	0.07	1.00	

Table 3: Overall Percentages of Correct Classification for Training and Validation Data Sets of Simulation Conditions

Method	Training	Validation
	Total (Hit Ratio)	Total (Hit Ratio)
Logistic Regression	59.83	61.93
Discriminant Analysis	84.98	86.31
Ensemble Method	49.61	52.15

6. Conclusions

The SAS system facilitates the building of a program to composite logistic regression and discriminant analysis using DATA step and PROC FREQ. This SAS procedure is recommended for those who are not supplied with Enterprise Miner and still want to do ensemble method using statistical procedures. The proposed technique is useful and applicable for combine model predictions or the outputs of several diverse classifiers/experts to form a potentially stronger solution in an ensemble system.

7. References

- [1] Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532.
- [2] Gong, G. (1986). Cross-validation, the jackknife and the bootstrap excess error estimation in forward regression logistic regression. *Journal of the American Statistical Association*, 81(393), 108-113.
- [3] Hoaglin, D. C. (1985). Summarizing shape numerically: The *g*- and *h*- distributions. In D. C. Hoaglin, F. Mosteller & J. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 461-513). New York: Wiley.
- [4] Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.
- [5] Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, New Jersey: Prentice-Hall.
- [6] Keselman, H. J., Wilcox, R. R., Algina, J., Othman, A. R., & Fradette, K. (in press). A comparative study of robust tests for spread: asymmetric trimming strategies. *British Journal of Mathematical and Statistical Psychology*.
- [7] Lachenbruch, P. A. (1975). *Discriminant analysis*. New York: Hafner.
- [8] Tatsuoka, M. M. (1988). *Multivariate analysis* (2nd ed.). New York: Wiley.