This article was downloaded by: [159.220.78.19]
On: 04 August 2014, At: 05:28
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

Multinomial Logistic Regression Ensembles

Kyewon Lee [a], Hongshik Ahn [a], Hojin Moon [b], Ralph L. Kodell [c] & James J. Chen [d]

[a] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York, USA

[b] Department of Mathematics and Statistics, California State University, Long Beach, California, USA

[c] Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

[d] Division of Personalized Nutrition and Medicine, Biometry Branch, National Center for Toxicological Research, Jefferson, Arkansas, USA

Published online: 23 Apr 2013.

To cite this article: Kyewon Lee, Hongshik Ahn, Hojin Moon, Ralph L. Kodell & James J. Chen (2013) Multinomial Logistic Regression Ensembles, Journal of Biopharmaceutical Statistics, 23:3, 681-694, DOI: 10.1080/10543406.2012.756500

To link to this article: http://dx.doi.org/10.1080/10543406.2012.756500

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Journal of Biopharmaceutical Statistics, 23: 681–694, 2013
Copyright © Taylor & Francis Group, LLC
ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2012.756500

Taylor & Francis
Taylor & Francis Group

MULTINOMIAL LOGISTIC REGRESSION ENSEMBLES

Kyewon Lee[1], Hongshik Ahn[1], Hojin Moon[2], Ralph L. Kodell[3], and James J. Chen[4]

[1]*Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York, USA*
[2]*Department of Mathematics and Statistics, California State University, Long Beach, California, USA*
[3]*Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA*
[4]*Division of Personalized Nutrition and Medicine, Biometry Branch, National Center for Toxicological Research, Jefferson, Arkansas, USA*

This article proposes a method for multiclass classification problems using ensembles of multinomial logistic regression models. A multinomial logit model is used as a base classifier in ensembles from random partitions of predictors. The multinomial logit model can be applied to each mutually exclusive subset of the feature space without variable selection. By combining multiple models the proposed method can handle a huge database without a constraint needed for analyzing high-dimensional data, and the random partition can improve the prediction accuracy by reducing the correlation among base classifiers. The proposed method is implemented using R, and the performance including overall prediction accuracy, sensitivity, and specificity for each category is evaluated on two real data sets and simulation data sets. To investigate the quality of prediction in terms of sensitivity and specificity, the area under the receiver operating characteristic (ROC) curve (AUC) is also examined. The performance of the proposed model is compared to a single multinomial logit model and it shows a substantial improvement in overall prediction accuracy. The proposed method is also compared with other classification methods such as the random forest, support vector machines, and random multinomial logit model.

Downloaded by [] at 05:28 04 August 2014

Key Words: Class prediction; Ensemble; Logistic regression; Majority voting; Multinomial logit; Random partition.

1. INTRODUCTION

Acute gastrointestinal bleeding (GIB) is an increasing health care problem due to rising nonsteroidal anti-inflammatory drugs (NSAID) use in an aging population (Rockall et al., 1995). In the emergency room (ER), the ER physician can misdiagnose a GIB patient at least 50% of the time (Kollef et al., 1997). While it is best for a gastroenterologist to diagnose GIB patients, it is not feasible due to time and constraints. Classification models can be used to assist the ER physician to

Received May 11, 2011; Accepted October 11, 2011
Address correspondence to Hongshik Ahn, Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA; E-mail: hahn@ams.sunysb.edu

diagnose GIB patients more efficiently and effectively, providing scarce health care resources to those who need it the most. Chu et al. (2008) evaluated eight different classification models on a 121-patient GIB database. Using clinical and laboratory information available within a few hours of patient presentation, the models can be used to predict the source of bleeding, need for intervention, and disposition in patients with acute upper, mid, and lower GIB.

Another application we consider in this paper is determination of stromal signatures in breast carcinoma. West et al. (2005) examined two types of tumors with fibroblastic features, solitary fibrous tumor (SFT) and desmoid-type fibromatosis (DTF), by DNA microarray analysis and found that the two tumor types differ in their patterns of expression in various functional categories of genes. Their findings suggest that gene expression patterns of soft tissue tumors can be used to discover new markers for normal connective-tissue cells. Compared to the GIB data the breast cancer data West et al. (2005) are high-dimensional. The data set contains 4148 variables on 57 tumor patients. A classification method can be applied to classify the data into DTF, SFT, and other types of tumors.

Logistic regression is a model that fits the log odds of the response to a linear combination of the explanatory variables. It is used mainly for binary responses, although there are extensions for multiway responses as well. An advantage for using logistic regression is that a model can be clearly and succinctly represented. Logistic regression is widely used in areas such as medical and social sciences.

LORENS (Logistic Regression Ensembles: Lim et al., 2010) uses the CERP (Classification by Ensembles from Random Partitions: Ahn et al., 2007) algorithm to classify binary responses based on high-dimensional data using the logistic regression model as a base classifier. CERP is similar to random subspace (Ho, 1998), but the difference is that base classifiers in an ensemble are obtained from mutually exclusive sets of predictors in CERP to increase diversity, while they are obtained by a random selection with overlap in random subspace. Although partitioning into mutually exclusive sets of features does not guarantee independence among the exclusive sets of predictors, the correlated variables in different subsets may yield diversification by combining with different sets of other variables.

Although logistic regression is known to be robust as a classification method and is widely used, it requires that there be more observations than predictors. Thus, in general, to use logistic regression for high-dimensional predictor spaces, variable selection is unavoidable. In LORENS, however, each base classifier is constructed from a different set of predictors determined by a random partition of the entire set of predictors, so that there are always more observations than predictors. Hence the logistic regression model can be used without variable selection. Details of finding an optimal partition using cross-validation in the learning phase are given in Lim et al. (2010). LORENS generates multiple ensembles with different random partitions, and conducts a majority voting of individual ensembles to further improve gain in overall accuracy.

LORENS is a useful classification method for high-dimensional data, and is designed for binary responses. Multiclass problems are common; thus, we developed

Downloaded by [] at 05:28 04 August 2014

a method comparable to LORENS for a multiway classification. We expanded LORENS to multiclass problems (mLORENS) and compare the performance with that of a single multiple logistic regression model in this study. The multinomial logistic regression (MLR) model can be easily implemented and it is less computer intensive than the tree-based CERP. Lim et al. (2010) showed that the prediction accuracy of LORENS is as good as that of random forest (RF: Breiman, 2001) or support vector machine (SVM: Vapnik, 1999) using some real data sets and a simulation study.

Prinzie and den Poel (2008) proposed a random multinomial logit (RMNL) model that fits multinomial logit models in different bootstrap samples. They borrowed the structure of RF, and used the idea of bagging. RF generates diversity by randomly selecting variables in each node of a base tree. Thus, various variables are involved in each tree classifier. In RMNL, however, one base classifier is built by only one random selection of variables.

In this paper we investigate the improvement of mLORENS over MLR, which has never been studied before. The GIB data, breast carcinoma data, and simulated data are used. We focus on the improvement of mLORENS over MLR, but to further evaluate mLORENS, it is also compared to RF, SVM, and RMNL. RMNL is included in the comparison because it is also an ensemble classification method using the logistic regression model as a base classifier. We implemented the mLORENS algorithm in R. The software will be available upon request.

Downloaded by [] at 05:28 04 August 2014

2. METHODS

2.1. LORENS (Logistic Regression Ensembles)

Based on the CERP algorithm, Lim et al. (2010) developed LORENS by using logistic regression models as base classifiers. To minimize the correlation among the ensemble of classifiers, the feature space is randomly partitioned into K subspaces with roughly equal sizes. Since the subspaces are randomly chosen from the same distribution, we assume that there is no bias in selection of the predictor variables in each subspace. In each of these subspaces, we fit a full logistic regression model without a variable selection. LORENS combines the results of these multiple logistic regression models to achieve an improved accuracy of classifying a patient's outcome by taking the average of the predicted values within an ensemble. The predicted values from all the base classifiers (logistic regression models) in an ensemble are averaged and the sample is classified as either 0 or 1 using a threshold on this average. Details of the method can be found in Lim et al. (2010).

2.2. mLORENS for Multinomial Logistic Regression Model

Suppose Y is a response variable with $J > 2$ categories. Let $\{\pi_1, \ldots, \pi_J\}$ be the response probabilities satisfying $\sum_{j=1}^{J} \pi_j = 1$. When one takes n independent observations based on these probabilities, the probability distribution for the number of outcomes that occur as each of the J types is multinomial. If a category is fixed as a baseline category, we have $J - 1$ log odds paired with the baseline

category. When the last category is the baseline, the baseline-category logits are $\log(\pi_j/\pi_J)$, $j = 1, \ldots, J-1$. The logit model using baseline-category logits with a predictor x has the form $\log(\pi_j/\pi_J) = \alpha_j + \beta_j x$, $j = 1, \ldots, J-1$. The model consists of $J-1$ logit equations, with separate parameters. By fitting these $J-1$ logit equations simultaneously, estimates of the model parameters can be obtained, and the same parameter estimates occur for a pair of categories regardless of the baseline category (Agresti, 1996). The estimates of the response probabilities can be expressed as $\pi_j = \exp(\alpha_j + \beta_j x)/\sum_h \exp(\alpha_h + \beta_h x)$, $j = 1, \ldots, J-1$.

Some alternative approaches can be considered instead of the preceding multinomial logistic regression (MLR) model. One of the approaches is nested binary models (Jobson, 1992). For each of the classes, the logistic regression model is fit for the class versus the remaining classes combined in this model. The class with the highest estimated response is chosen. We compared the MLR model with this alternative model in section 3, but we do not show it in this paper because the results were not significantly different.

We developed mLORENS, which uses MLR as a base classifier. We anticipate that the proposed method will possess the nice properties of MLR and simultaneously inherit the advantages of CERP handling high-dimensional data sets.

Downloaded by [] at 05:28 04 August 2014

3. APPLICATIONS

The improvement of mLORENS over a single MLR model is investigated and its performance is compared to RF, SMV, and RMNL. The methods are applied to real data sets and to simulation data. Twenty repetitions of 10-fold cross validation (CV) were conducted for each model. For mLORENS, the optimal partition sizes were determined in the training phase. For a single MLR model, variable selection is necessary when the sample size is smaller than the number of predictors. For MLR, a fixed number of predictors was pre-assigned and the variables were selected from the training data. Variable selection was not performed for mLORENS as it is not necessary due to the random partition. ACC (overall accuracy), SENS (sensitivity), SPEC (specificity), PPV (positive predictive value: rate of true positives among positive predictions), and NPV (negative predictive value: rate of true negatives among negative predictions) were compared. The receiver operating characteristic (ROC) curves were created and areas under the curves (AUCs) were compared to assess the performance of the models in terms of sensitivity and specificity. AUC is a scalar measure gauging the performance of the ROC curve. An AUC of 1 represents a perfect prediction. The Mann–Whitney statistic was calculated, which is equivalent to the AUC (DeLong et al., 1988).

Prediction accuracy is calculated by the total number of correct predictions divided by total number of predictions, and sensitivity and specificity are calculated, respectively, for each category as follows. If the number of classes is three, for example, the following table shows the counts of predicted classification and true classification. The sensitivity for class 1 is calculated by $a/(a+b+c)$, for class 2 is $e/(d+e+f)$, and for class 3 is $i/(g+h+i)$. The specificity for class 1 is $(e+f+h+i)/$

$(d+e+f+g+h+i)$, for class 2 is $(a+c+g+i)/(a+b+c+g+h+i)$, and for class 3 is $(a+b+d+e)/(a+b+c+d+e+f)$.

		True class		
		1	2	3
Predicted class	1	*a*	*d*	*g*
	2	*b*	*e*	*h*
	3	*c*	*f*	*i*

The MLR model was implemented into mLORENS in R using the *multinom* function in the *nnet* package. The *multinom* function fits the multinomial logit model via neural networks. Thus, the *multinom* function is computer-intensive. For RF, the *RandomForest* package in R is used. The defaults of 500 trees are generated in the forest, and $m^{1/2}$ variables are randomly selected at each node of the base tree. The *e1071* package in R is used for SVM with the following input variables: tolerance (0.001), epsilon (0.1), degree (3), gamma (1), coef0 (0), and cost (1). The RMNL program is implemented in R by the first author. Average values obtained from 20 repetitions of 10-fold CV were used for all the comparisons for each model.

Downloaded by [] at 05:28 04 August 2014

3.1. GIB Data

Patients with acute GIB were identified from the hospital medical records database of 121 patients. Chu et al. (2008) classified the sample into different bleeding sources, need for urgent blood resuscitation, need for urgent endoscopy, and disposition. In this paper, we focus on the source of bleeding because we are interested in multiway classification. The bleeding source is classified into three classes: upper, mid, and lower intestine. The definitive source of bleeding was the irrefutable identification of a bleeding source at upper endoscopy, colonoscopy, small bowel enteroscopy, or capsule endoscopy. Twenty input variables utilized to predict the source of GIB included prior history of GIB, hematochezia, hematemesis, melena, syncope/presyncope, risk for stress ulceration, cirrhosis, ASA/NSAID use, systolic and diastolic blood pressures, heart rate, orthostasis, nasogastric (NG) lavage, rectal exam, platelet count, creatinine, blood urea nitrogen (BUN), and international normalized ratio (INR).

For mLORENS, random partition into two subsets of predictors was used without a search for an optimal partition size. Since the number of predictor variables is only 20, the fixed partition size was used for this example. Among 121 subjects, 81 fell on upper, 29 on lower, and the remaining 11 fell on mid intestine. The class with the highest probability estimate is chosen to be the right class without considering a threshold for a decision. The prediction accuracy along with sensitivity and specificity was provided in Table 1. Although both MLR and mLORENS worked well on GIB data, mLORENS was better in prediction accuracy. The prediction accuracy for mLORENS was 93%, compared to 89% for MLR. According to McNemar's test, the accuracy of mLORENS was significantly higher than that of MLR (p-value $< .0001$), SVM (linear kernel: p-value $< .0001$, radial kernel: p-value .025), and RMNL (p-value .015). The accuracies of mLORENS and

Table 1 Comparison of mLORENS and MLR for the GIB data (SD in parentheses)

Method	#part.[a]	ACC	Mean AUC[b]		Upper	Lower	Mid
mLORENS	2	.93 (.01)	.90 (.02)	SENS:	.98 (.01)	.89 (.03)	.68 (.09)
				SPEC:	.93 (.02)	.97 (.01)	.98 (.01)
				AUC:	.96 (.01)	.93 (.02)	.83 (.05)
MLR		.89 (.02)	.90 (.02)	SENS:	.92 (.02)	.86 (.04)	.75 (.09)
				SPEC:	.97 (.03)	.95 (.01)	.94 (.01)
				AUC:	.95 (.01)	.90 (.02)	.85 (.05)
RMNL		.92 (.02)	.90 (.03)	SENS:	.96 (.01)	.91 (.02)	.66 (.13)
				SPEC	.96 (.03)	.95 (.02)	.97 (.01)
				AUC	.96 (.01)	.93 (.02)	.81 (.07)
RF		.94 (.01)	.90 (.02)	SENS:	.98 (.01)	.94 (.02)	.61 (.08)
				SPEC:	.93 (.01)	.97 (.01)	.99 (.00)
				AUC:	.96 (.01)	.95 (.01)	.80 (.04)
SVM		.90 (.02)	.88 (.03)	SENS:	.95 (.01)	.86 (.04)	.64 (.13)
Linear				SPEC:	.91 (.04)	.96 (.02)	.96 (.01)
Kernel				AUC:	.93 (.02)	.91 (.02)	.80 (.06)
SVM		.92 (.01)	.88 (.01)	SENS:	.99 (.00)	.90 (.00)	.54 (.07)
Radial				SPEC:	.88 (.01)	.95 (.01)	1.00 (.00)
Kernel				AUC:	.93 (.00)	.93 (.00)	.77 (.04)

Note. Fixed partition sizes are used in mLORENS.
[a]Predetermined number of mutually exclusive subsets in the partition.
[b]Mean of the AUCs from the three classes.

Downloaded by [] at 05:28 04 August 2014

RF were not significantly different (p-value .115). Throughout the paper, the p-values are for two-sided tests. For each of the three categories, AUC was obtained by grouping the data to the given category and the rest. The mean AUC was obtained by taking average of these three AUCs (Hand and Till, 2001). Sensitivity and specificity were not significantly different between MLR and mLORENS, but mLORENS showed higher sensitivities for large classes (upper and lower), and for small class (mid) MLR showed high sensitivity. Specificity for the larger class was higher in MLR than in mLORENS. The mean AUC was high in both mLORENS and MLR. RF and RMNL performances were similar to that of mLORENS, while SVM performed worse than mLORENS.

3.2. Breast Cancer Data

West et al. (2005) studied determination of stromal signatures in breast carcinoma using DNA microarray analysis to examine two fibroblastic tumors: solitary fibrous tumor (SFT) and desmoid-type fibromatosis (DTF). The data contain 57 subjects. Ten cases of DTF and 13 cases of benign SFT were compared to 34 other previously examined soft tissue tumors with expression profiling on 42,000 element cDNA microarrays, corresponding to approximately 36,000 unique gene sequences. The data were obtained from http://smd. stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=436. For the data, 4148 predictors were used in the analysis. Table 2 shows the comparison of mLORENS and other classification methods these data.

Table 2 Comparison of mLORENS and MLR for the breast cancer data (SD in parentheses)

Method	#var.[a]	#part.[b]	ACC	Mean AUC[c]		DTF	SFT	Other
mLORENS	all	207	.92	.92	SENS:	1.00 (.00)	.68 (.08)	.99 (.02)
		(79)	(.02)	(.07)	SPEC:	.99 (.01)	1.00 (.00)	.82 (.04)
					AUC:	1.00 (.01)	.84 (.04)	.91 (.03)
MLR	10		.66	.73	SENS:	.87 (.15)	.43 (.14)	.69 (.08)
			(.07)	(.14)	SPEC:	.93 (.04)	.83 (.05)	.64 (.12)
					AUC:	.90 (.07)	.63 (.07)	.67 (.08)
	20		.67	.76	SENS:	.84 (.13)	.58 (.14)	.66 (.09)
			(.07)	(.11)	SPEC:	.93 (.03)	.79 (.06)	.73 (.10)
					AUC:	.89 (.06)	.69 (.07)	.70 (.07)
	30		.69	.78	SENS:	.88 (.09)	.66 (.12)	.65 (.08)
			(.05)	(.10)	SPEC:	.92 (.04)	.79 (.06)	.79 (.09)
					AUC:	.90 (.04)	.73 (.05)	.72 (.05)
RMNL	25.9		.93	.93	SENS:	1.00 (.02)	.75 (.07)	.98 (.02)
	(2.9)		(.02)	(.02)	SPEC:	1.00 (.01)	.99 (.01)	.86 (.04)
					AUC:	1.00 (.01)	.87 (.03)	.92 (.02)
RF			.86	.85	SENS:	1.00 (.00)	.47 (.07)	.97 (.02)
			(.02)	(.02)	SPEC:	.99 (.01)	.98 (.01)	.70 (.04)
					AUC:	1.00 (.01)	.72 (.03)	.83 (.02)
SVM			.94	.95	SENS:	1.00 (.00)	.83 (.04)	.97 (.02)
Linear			(.01)	(.01)	SPEC:	.98 (.01)	1.00 (.00)	.90 (.02)
Kernel					AUC:	.99 (.01)	.92 (.02)	.94 (.01)
SVM			.86	.84	SENS:	.90 (.00)	.46 (.00)	1.00 (.01)
Radial			(.00)	(.00)	SPEC:	1.00 (.00)	1.00 (.00)	.65 (.00)
Kernel					AUC:	.95 (.00)	.73 (.00)	.83 (.00)

Note. The partition sizes for mLORENS are determined in the learning phase.

[a]Number of selected variables chosen in the training phase.

[b]Mean (SD) number of mutually exclusive subsets of predictors in a partition, chosen in the training phase.

[c]Mean of the AUC's from the three classes.

Downloaded by [] at 05:28 04 August 2014

The partition sizes of mLORENS were obtained in the training phase. The average number of mutually exclusive subsets in a partition was 207 resulting in approximately 20 variables in each subset on average in mLORENS. Since MLR requires variable selection, 10, 20, or 30 variables were selected in the training phase. The variable selection was performed in the learning set using the BW ratio (ratio of between-group to within-group sums of squares) defined as

$$BW(j) = \frac{\sum_i \sum_k I(y_i = k)\left(\overline{x}_{kj} - \overline{x}_{.j}\right)^2}{\sum_i \sum_k I(y_i = k)\left(\overline{x}_{ij} - \overline{x}_{kj}\right)^2}, \quad \text{where } \overline{x}_{.j} = \frac{\sum_i x_{ij}}{N}; \overline{x}_{kj} = \frac{\sum_{I(y_i=k)} x_{ij}}{N_k},$$

for sample i and gene j. This criterion has been shown to be reliable for variable selection from high-dimensional data (Dudoit et al., 2002).

The overall accuracy of mLORENS was higher (.92, SD .02) than that of MLR (.69, SD .05 when 30 variables were selected). mLORENS also outperformed MLR in mean AUC (.92 versus .78). mLORENS was significantly higher in ACC according to McNemar's test (p-value < 0.0001). These results show that without

considering variable selection, mLORENS performs significantly better than MLR which requires variable selection. mLORENS was comparable to RMNL (.93, SD .02), worse than SVM linear kernel (.94, SD .01, p-value .002), but significantly higher (p-value $< .0001$) than RF (.86, SD .02) and SVM radial kernel (.86, SD .00) in accuracy. We observed the same pattern in AUC.

3.3. Simulation

We conducted a simulation study to evaluate mLORENS in comparison with MLR using high-dimensional simulated data and we compared to RF, SVM, and RMNL. We generated two data sets with 120 subjects and 500 independent predictors, one for training and the other for testing. The ratios of the classes were given as 40:40:40. One hundred pairs of these training and test data sets were generated. Fifty of these predictor variables were generated from three different normal distributions, and the remaining 450 predictor variables were generated from one normal distribution. The latter served as noise. The predictors were generated from normal distributions with standard deviation (σ) of 1, 2, or 3. Figure 1 shows the simulation design. Among the 50 significant variables, the first 10 variables were

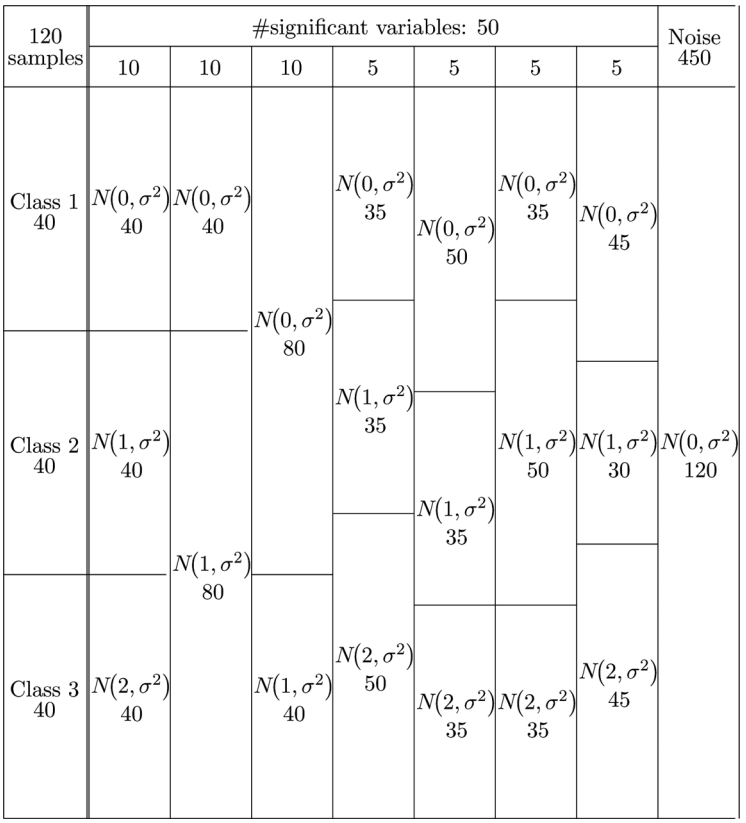

Figure 1 Simulation design for a comparison of mLORENS and MLR.

Downloaded by [] at 05:28 04 August 2014

generated from $N(0, \sigma^2)$ for class 1, $N(1, \sigma^2)$ for class 2, and $N(2, \sigma^2)$ for class 3. The next 10 variables were generated from $N(0, \sigma^2)$ for class 1 and $N(1, \sigma^2)$ for class 2 and 3. For the other significant variables see Figure 1. The 450 noise variables were independently generated from $N(0, \sigma^2)$.

For each simulation data, the model fit in the training set was applied to the test set for evaluating the performance of each classification method. The average of the accuracies from the 100 pairs of simulation data sets from each classification method is provided in Tables 3 through 5, for $\sigma = 1$, $\sigma = 2$, and $\sigma = 3$, respectively. LORENS included all 500 variables without variable selection. An ideal model should be the MLR model consisting of the 50 significant predictors, assuming that we can accurately identify these variables. Since the significant variables are unknown in practice, we compared mLORENS and the ideal MLR model with MLR models with variable selection. The variable selection was performed in the learning set using the BW ratio.

Since the number of significant variables is unknown in practice, the numbers of selected variables were preselected as 10, 30, 50, and 70 for MLR. The given numbers of variables were selected in the training phase. For mLORENS, we tried both fixed numbers of partitions and an optimal number of partitions searched in the training phase. The result was good with the optimal number of partitions from the search method. mLORENS showed substantially higher accuracy compared to the MLR models. As expected, the data with smaller standard deviation gave much higher accuracy than the data with higher standard deviation. For data with standard deviation 1, the model with 10 variables gave a better result than the ideal MLR model consisting of all significant variables (Table 3). It might be because the 50 significant variables have repetition. In fact, 5 to 10 variables were generated from each of the 7 distinct significant features. Thus, MLR with more variables or the correct model may be redundant. According to the result, mLORENS appears to handle the redundancy effectively by random partition. Since the feature space is randomly partitioned to 30 subspaces on average, the MLR model on each subset would contain nearly two significant variables on average. This would minimize the chance of redundancy and the ensemble of these models takes advantage over the single MLR model. When the standard deviation was 3, the MLR with the known 50 significant variables performed significantly better than MLR with variable selection due to the ambiguity of the data (Table 5). The accuracies of mLORENS and the ideal MLR model with all the significant variables were not significantly different when the standard deviation was 3. Otherwise, the accuracy of mLORENS was significantly higher than that of the best MLR model and RMNL with the p-value less than 0.0001 for all values of the standard deviation using McNemar's test. The mean AUC was significantly higher in mLORENS than in any single MLR model. When we compared to RF and SVM, mLORENS was significantly worse when the standard deviation was 1 (p-value $< .0001$), comparable when the standard deviation was 2, and significantly better when the standard deviation was 3 (p-value $< .0001$).

The variable selection in MLR was highly accurate for the data with standard deviation 1, while it was not as accurate for the data with standard deviation 3. Most of the significant variables were selected for the data with standard deviation 1, but the average number of significant variables selected were 7.4, 13.6, 27.4, and

Downloaded by [] at 05:28 04 August 2014

Table 3 Performance of mLORENS, multivariate logit models with variable selection, and multivariate logit model with all significant variables for the simulated data with 500 predictors

Method	#var.[a]	#sig. var.[b]	#part.[c]	ACC	Mean AUC[d]		Class 1	Class 2	Class 3
mLORENS	All		38.0	.92	.94	SENS:	.99 (.01)	.76 (.07)	.99 (.01)
			(7.6)	(.02)	(.05)	SPEC:	.94 (.03)	.99 (.01)	.94 (.03)
						AUC:	.97 (.01)	.88 (.03)	.97 (.01)
MLR	10	10		.83	.88	SENS:	.89 (.07)	.75 (.08)	.86 (.08)
with		(0)		(.04)	(.06)	SPEC:	.94 (.03)	.88 (.05)	.93 (.03)
variable						AUC:	.92 (.04)	.82 (.04)	.89 (.04)
selection	30	29.2		.82	.86	SENS:	.88 (.07)	.72 (.10)	.86 (.08)
		(.8)		(.05)	(.07)	SPEC:	.90 (.04)	.88 (.05)	.95 (.03)
						AUC:	.89 (.04)	.80 (.06)	.90 (.04)
	50	49.2		.74	.81	SENS:	.80 (.08)	.69 (.08)	.74 (.09)
		(.8)		(.04)	(.06)	SPEC:	.88 (.04)	.81 (.05)	.91 (.04)
						AUC:	.84 (.04)	.75 (.04)	.82 (.05)
	70	49.2		.67	.75	SENS:	.72 (.09)	.63 (.09)	.66 (.10)
		(.7)		(.06)	(.07)	SPEC:	.87 (.05)	.76 (.06)	.87 (.04)
						AUC:	.80 (.05)	.70 (.05)	.77 (.06)
MLR	50			.74	.81	SENS:	.80 (.09)	.68 (.10)	.73 (.10)
with significant				(.05)	(.06)	SPEC:	.88 (.05)	.82 (.06)	.91 (.04)
variables						AUC:	.84 (.05)	.75 (.05)	.82 (.06)
RMNL				.86	.89	SENS:	.96 (.05)	.66 (.10)	.95 (.05)
				(.04)	(.03)	SPEC:	.91 (.04)	.96 (.04)	.92 (.04)
						AUC:	.94 (.03)	.81 (.05)	.94 (.03)
RF				.95	.96	SENS:	.99 (.01)	.87 (.06)	.99 (.02)
				(.02)	(.02)	SPEC:	.97 (.02)	.99 (.01)	.96 (.02)
						AUC:	.98 (.01)	.93 (.03)	.98 (.01)
SVM				.94	.95	SENS:	.96 (.03)	.90 (.05)	.95 (.03)
Linear				(.02)	(.01)		.97 (.02)	.96 (.02)	.98 (.02)
Kernel							.97 (.02)	.93 (.02)	.96 (.02)
SVM				.95	.96	SENS:	.97 (.03)	.92 (.05)	.96 (.03)
Radial				(.02)	(.02)	SPEC:	.98 (.02)	.97 (.02)	.98 (.02)
Kernel						AUC:	.97 (.02)	.94 (.03)	.97 (.02)

Note. Predictors were generated from normal distribution with standard deviation 1.

[a]Number of selected variables chosen in the training phase.

[b]Number of significant variables among the selected variables.

[c]Mean (SD) number of mutually exclusive subsets of predictors in a partition, chosen in the training phase.

[d]Mean of the AUC's from the three classes.

28.4, when the model contained 10, 30, 50, and 70 variables, respectively, for the data with standard deviation 3.

We also conducted a simulation with the same design, but with correlated variables. The results were similar except for lower overall accuracy in general. Thus, we do not report the result from this simulation.

The run time of mLORENS was approximately 2 hours to finish 100 repetitions of simulating the high-dimensional data in section 3.3 on a Windows 3.0 GHz machine.

Downloaded by [] at 05:28 04 August 2014

Table 4 Performance of mLORENS, multivariate logit models with variable selection, and multivariate logit model with all significant variables for the simulated data with 500 predictors

Method	#var.[a]	#sig. var.[b]	#part.[c]	ACC	Mean AUC[d]		Class 1	Class 2	Class 3
mLORENS	All		29.6[e]	.71	.78	SENS:	.86 (.06)	.40 (.08)	.86 (.06)
			(8.8)	(.04)	(.11)	SPEC:	.84 (.04)	.87 (.04)	.84 (.04)
						AUC:	.85 (.03)	.64 (.04)	.85 (.03)
MLR	10	9.6		.61	.71	SENS:	.69 (.08)	.46 (.09)	.68 (.09)
with		(.6)		(.05)	(.09)	SPEC:	.83 (.05)	.76 (.06)	.83 (.05)
variable						AUC:	.76 (.05)	.61 (.05)	.75 (.06)
selection	30	20.2		.60	.70	SENS:	.68 (.09)	.46 (.11)	.67 (.09)
		(1.9)		(.06)	(.09)	SPEC:	.83 (.05)	.74 (.07)	.84 (.05)
						AUC:	.75 (.05)	.60 (.06)	.75 (.05)
	50	38.9		.60	.70	SENS:	.70 (.09)	.50 (.09)	.61 (.10)
		(2.0)		(.04)	(.08)	SPEC:	.82 (.05)	.73 (.06)	.86 (.05)
						AUC:	.76 (.05)	.62 (.05)	.73 (.05)
	70	39.5		.55	.66	SENS:	.63 (.09)	.45 (.09)	.56 (.10)
		(2.0)		(.05)	(.08)	SPEC:	.79 (.05)	.70 (.06)	.83 (.05)
						AUC:	.71 (.05)	.58 (.05)	.69 (.05)
MLR	50			.61	.71	SENS:	.70 (.11)	.50 (.10)	.63 (.10)
with significant				(.06)	(.08)	SPEC:	.82 (.06)	.74 (.07)	.86 (.05)
variables						AUC:	.76 (.06)	.62 (.05)	.75 (.05)
RMNL				.64	.73	SENS:	.79 (.08)	.38 (.08)	.74 (.09)
				(.04)	(.03)	SPEC:	.81 (.05)	.80 (.06)	.84 (.05)
						AUC:	.80 (.04)	.59 (.04)	.79 (.04)
RF				.71	.78	SENS:	.83 (.07)	.45 (.09)	.83 (.06)
				(.04)	(.03)	SPEC:	.85 (.04)	.85 (.05)	.86 (.04)
						AUC:	.84 (.03)	.65 (.04)	.84 (.04)
SVM				.68	.76	SENS:	.75 (.08)	.55 (.08)	.73 (.08)
Linear				(.05)	(.03)	SPEC:	.88 (.04)	.76 (.05)	.88 (.04)
Kernel						AUC:	.81 (.04)	.66 (.05)	.81 (.04)
SVM				.70	.78	SENS:	.76 (.09)	.59 (.09)	.75 (.09)
Radial				(.04)	(.03)	SPEC:	.89 (.04)	.77 (.06)	.89 (.04)
Kernel						AUC:	.83 (.04)	.68 (.04)	.82 (.04)

Note. Predictors were generated from normal distribution with standard deviation 2.

[a]Number of selected variables chosen in the training phase.

[b]Number of significant variables among the selected variables.

[c]Mean number of mutually exclusive subsets of predictors in a partition (SD), chosen in the training phase.

[d]Mean of the AUCs from the three classes.

[e]Average (SD) number of subsets in the partition determined in the learning phase.

4. CONCLUSION AND DISCUSSION

A major advantage of LORENS over CERP or other aggregation methods is that the base classifier, logistic regression, is widely used and well understood by a broad base of users in science and medicine. Moreover, LORENS is more computationally efficient than CERP, which is tree-based. Due to the random partition, LORENS is free of the constraint on the dimension of the data. In addition to binary classification problems, multiclass classification problems are not

Downloaded by [] at 05:28 04 August 2014

Table 5 Performance of mLORENS, multivariate logit models with variable selection, and multivariate logit model with all significant variables for the simulated data with 500 predictors

Method	#var.[a]	#sig. var.[b]	#part.[c]	ACC	Mean AUC[d]		Class 1	Class 2	Class 3
mLORENS	All		27.8[e]	.57	.68	SENS:	.68 (.09)	.36 (.08)	.69 (.08)
			(8.5)	(.04)	(.03)	SPEC:	.80 (.05)	.77 (.06)	.80 (.05)
						AUC:	.74 (.05)	.56 (.04)	.74 (.04)
MLR	10	7.4		.49	.62	SENS:	.57 (.10)	.35 (.09)	.55 (.10)
with		(1.5)		(.05)	(.04)	SPEC:	.76 (.06)	.73 (.06)	.76 (.06)
variable						AUC:	.67 (.05)	.54 (.05)	.66 (.06)
selection	30	13.6		.49	.62	SENS:	.55 (.11)	.37 (.09)	.54 (.10)
		(2.3)		(.05)	(.04)	SPEC:	.77 (.07)	.70 (.07)	.76 (.06)
						AUC:	.66 (.06)	.54 (.05)	.65 (.05)
	50	27.4		.50	.62	SENS:	.57 (.10)	.38 (.09)	.54 (.11)
		(3.0)		(.05)	(.04)	SPEC:	.79 (.06)	.69 (.07)	.76 (.06)
						AUC:	.68 (.06)	.54 (.05)	.65 (.06)
	70	28.4		.46	.60	SENS:	.53 (.11)	.37 (.09)	.50 (.10)
		(3.0)		(.05)	(.04)	SPEC:	.76 (.06)	.68 (.07)	.76 (.06)
						AUC:	.64 (.06)	.52 (.05)	.63 (.05)
MLR	50			.57	.68	SENS:	.67 (.10)	.43 (.10)	.60 (.10)
with significant				(.06)	(.04)	SPEC:	.80 (.06)	.72 (.06)	.83 (.05)
variables						AUC:	.74 (.05)	.58 (.06)	.71 (.06)
RMNL				.51	.63	SENS:	.59 (.12)	.36 (.10)	.57 (.09)
				(.05)	(.04)	SPEC:	.76 (.06)	.72 (.08)	.78 (.05)
						AUC:	.68 (.05)	.54 (.05)	.68 (.05)
RF				.53	.65	SENS:	.60 (.09)	.37 (.09)	.63 (.09)
				(.05)	(.03)	SPEC:	.79 (.05)	.72 (.07)	.79 (.05)
						AUC:	.70 (.05)	.55 (.05)	.71 (.05)
SVM				.55	.67	SENS:	.61 (.09)	.45 (.09)	.60 (.09)
Linear				(.05)	(.04)	SPEC:	.82 (.05)	.68 (.07)	.83 (.05)
Kernel						AUC:	.71 (.05)	.57 (.05)	.72 (.05)
SVM				.56	.67	SENS:	.60 (.11)	.46 (.11)	.62 (.10)
Radial				(.05)	(.04)	SPEC:	.84 (.05)	.68 (.08)	.82 (.05)
Kernel						AUC:	.72 (.05)	.57 (.05)	.72 (.05)

Note. Predictors were generated from normal distribution with standard deviation 3.

[a]Number of selected variables chosen in the training phase.

[b]Number of significant variables among the selected variables.

[c]Mean number of mutually exclusive subsets of predictors in a partition (SD), chosen in the training phase.

[d]Mean of the AUCs from the three classes.

[e]Average (SD) number of subsets in the partition determined in the learning phase.

rare in biomedical applications. In this paper, we expanded LORENS to multiclass classification problems.

Our results show that the performance of mLORENS is superior to that of MLR in terms of achieving significantly higher prediction accuracy, although mLORENS appears to favor the majority class when class sizes are very unbalanced. mLORENS shows higher AUC than a single MLR model for high-dimensional data in this paper. Since randomly selected mutually exclusive subsets of predictors

Downloaded by [] at 05:28 04 August 2014

are assigned to each of the random partitions, redundancy of the data is reduced as we discussed in section 3.3. Furthermore, variable selection is not necessary in mLORENS. The correlation between the classifiers is lowered by random partitioning. As discussed in section 3.3, the random partition minimizes the redundancy of the model in each subset of the variables. By integrating these advantages to the ensemble majority voting, the accuracy of the method is greatly improved. Although our focus in this paper was on the improvement of mLORENS over MLR, we also compared it with other popular existing classification methods. The performance of mLORENS was significantly better than that of RF and SVM for hard-to-classify data with larger standard deviation according to our simulation study. Furthermore, mLORENS is less computer-intensive than these methods.

In this paper, the performance of mLORENS was evaluated and compared with a single MLR model. The class with the highest estimate was chosen as the right class without considering a threshold for a decision. Lim et al. (2010) developed a threshold search for two-way classification for LORENS. For a multiway classification, the search for optimal thresholds is multidimensional, which is complicated and computer-intensive. In a future study, we plan to develop a method to determine the optimal thresholds in the learning phase, which is expected to improve the balance between sensitivity and specificity when class sizes are very unbalanced.

Logistic regression is a standard and commonly used model for a binary classification. MLR is an extension of the logistic regression model to a multiway classification. We used the *multimum* function in R for MLR. This function uses the artificial neural network approach, and thus this is computer-intensive. In a future study we plan to develop a more efficient and less computer-intensive algorithm for MLR by applying a standard procedure.

Downloaded by [] at 05:28 04 August 2014

REFERENCES

Agresti, A. (1996). Multicategory logit models. In: *An Introduction to Categorical Data Analysis*. New York, NY: Wiley-Interscience, pp. 286–287.

Ahn, H., Moon, H., Fazzari, M. J., Lim, N., Chen, J. J., Kodell, R. L. (2007). Classification by ensembles from random partitions of high-dimensional data. *Computational Statistics and Data Analysis* 51:6166–6179.

Breiman, L. (2001). Random forest. *Machine Learning* 45:5–32.

Chu, A., Ahn, H., Halwan, B., Kalmin, B., Artifon, E., Barkun, A., Lagoudakis, M., Kumar, A. (2008). A decision support system to facilitate management of patients with acute gastrointestinal bleeding. *Artificial Intelligence in Medicine* 42:247–259.

DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44:837–845.

Dudoit, S., Fridlyand, J., Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 97:77–87.

Hand, D. J., Till, R. J. (2001). A simple generalisation of the area under the curve for multiple class classification problems. *Machine Learning* 45:171–186.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IIII Transactions on Pattern Analysis and Machine Intelligence* 20(8):832–844.

Jobson, J. D. (1992). *Applied Multivariate Data Analysis*. Vol. 1. Regression and Experimental Design, chap. 8.3.6. New York, NY: Springer.

Kollef, M. H., O'Brien, J. D., Zuckerman, G. R., Shannon, W. (1997). A classification tool to predict outcomes in patients with acute upper and lower gastrointestinal hemorrhage. *Critical Care Medicine* 25:1125–1132.

Lim, N., Ahn, H., Moon, H., Chen, J. J. (2010). Classification of high-dimensional data with ensemble of logistic regression models. *Journal of Biopharmaceutical Statistics* 20:160–171.

Prinzie, A., den Poel, D. V. (2008). Random forests for multiclass classification: Random MultiNomial Logit. *Expert Systems with Applications* 34:1721–1732.

Rockall, T. A., Logan, R. F. A., Devlin, H. B., Northfield, T. C. (1995). Incidence of and mortality from acute upper gastrointestinal haemorrhage in the United Kingdom. *British Medical Journal* 311:222–226.

Vapnik, V. (1999). *The Nature of Statistical Learning Theory*. New York, NY: Springer Verlag.

West, R. B., Nuyten, D. S., Subramanian, S., Nielsen, T. O., Corless, C. L., Rubin, B. P., Montgomery, K., Zhu, S., Patel, R., Hernandez-Boussard, T., Goldblum, J. R., Brown, P. O., van de Vijver, M., van de Rijin, M. (2005). Determination of stromal signatures in breast carcinoma. *PLoS Biology* 3(6):e187.

Downloaded by [] at 05:28 04 August 2014