USE OF PLS COMPONENTS TO IMPROVE CLASSIFICATION ON BUSINESS DECISION MAKING

José C. Vega Vilca, Aniel Nieves-González and Roxana Aparicio

Institute of Statistics and Computer Information Systems, University of Puerto Rico, Rio Piedras Campus, Puerto Rico

ABSTRACT

This paper presents a methodology that eliminates multicollinearity of the predictors variables in supervised classification by transforming the predictor variables into orthogonal components obtained from the application of Partial Least Squares (PLS) Logistic Regression. The PLS logistic regression was developed by Bastien, Esposito-Vinzi, and Tenenhaus [1]. We apply the techniques of supervised classification on data, based on the original variables and data based on the PLS components. The error rates are calculated and the results compared. The implementation of the methodology of classification is rests upon the development of computer programs written in the R language to make possible the calculation of PLS components and error rates of classification. The impact of this research will be disseminated, based on evidence that the methodology of Partial Least Squares Logistic Regression, is fundamental when working in a supervised classification with data of many predictors variables.

KEYWORDS

Supervised classification, error rate, multivariate analysis, Logistic Regression

1. INTRODUCTION

In data analysis via supervised classification [13] a classifier is constructed based on the observed data. The data is arranged into an $n \times p$ matrix X, where n is the number of rows (subjects) and p is the number of columns (variables in the study), and a column vector $Y \in R^n$ that contains and indicator of the group to which each of the n subjects belongs to. The goal of constructing the classifier is to place new subjects into one of the G groups established in the given problem. Whenever p (the variables of the predictor matrix X) is large, is generally implied multicollinearity between the variables. Such multicollinearity is defined as a high linear dependence between the predictor variables. In this study it is demonstrated, by case studies, that the multicollinearity should be eliminated in order to construct a better classifier.

The general rules of thumb of data analysis by supervised classification can be summarized as follows:

- Given a new subject characterized by the p variables in the study. Into which of the defined groups (G) does the subject should be classified?
- The new subject should be classified into the group where the probability of belonging to that group is greater than the probability of belonging to the other groups.
- Based on the matrix X and the vector Y one should construct a classifier with a minimum error rate of classification.

The lack of knowledge about the consequences of multicollinearity in the predictor matrix X force the researchers to directly apply the techniques of supervised classification and to construct inefficient classifiers with a high error rate. The classifier error rate is defined as follows.

DOI: 10.5121/ijdkp.2018.8302

Definition 1.1

Let ϵ_C be the error rate of classification for a classifier C, and s be a new subject that does not belong to a group G. Then ϵ_C is the probability

$$\epsilon_C - P(s \in G) | s \notin G \qquad (1)$$

That is, ϵ_C, is the conditional probability that the classifier locates a new subject into a group to which the subject does not belong to.

In this work the multicollinearity problem is solved by transforming the predictor variables into latent variables, also called components. The components are linear combinations of predictor variables that have the property of being orthogonal (not correlated) and are obtained through the application of a method named Logistic Regression by Partial Least Squares (PLS). This method was introduced by Bastien, Esposito-Vinzi, and Tenenhaus [1].

This work states a method to improve the strategies for data analysis in situations where the subjects under consideration (e.g. people, animals, or things), should be classified correctly into groups according to their characteristics to find favorable or unfavorable patterns. For instance, a loan applicant to a bank provides personal information like income, sex, age, number of dependents, expenses, etc. This applicant is evaluated according to the information provided and is classified into potential good or bad borrower with the objective to determine whether the loan should be granted or not granted to the applicant.

The goal of this study is to disseminate the application of Logistic Regression by Partial Minimum Squares, introduced by Bastien, Esposito-Vinzi, and Tenenhaus [1], to eliminate the problem of multicollinearity in the predictor matrix and demonstrate. by means of case study, that the multicollinearity should be eliminated in order to construct a better classifier function, characterized by a minimal error rate of classification

2. MULTICOLLINEARITY

The authors in [11] analyze multicollinearity in multiple regression problems and verify two aspects about multicollinearity: First, it is a problem that makes it difficult to precisely quantify the effect that exerts each predictor variable over the dependent variable. Second, it can be determined by the computation of the Variance Inflation Factor (VIF) and by the condition number (η). The VIF is an indicator of specific multicollinearity of each predictor variable. The VIF is defined as:

$$VIF_j = \frac{1}{1-R_j^2} \; for \; j = 1, \dots, p \qquad (2)$$

where R_j^2 is the coefficient of determination for the linear regression of X_j with respect of the other predictor variables. As a rule of thumb, if $VIF_j \geq 10$, then there is strong multicollinearity. The condition number of the correlation matrix of the predictor variables is an indicator of the global multicollinearity of the predictor variables. The condition number is computed as

$$\eta = \sqrt{\frac{|\lambda_{max}|}{|\lambda_{min}|}} \qquad (3)$$

where λ_{min} and λ_{max} are the minimum and maximal eigenvalue (by moduli) of the correlation matrix of the predictor variables. Generally, if $\eta \geq 25$, then there is strong multicollinearity.
Once the multicollinearity is detected it should be eliminated by means of the method proposed in this work, Logistic Regression by Partial Least Squares (PLS).

2.1. Diagnosis Of Multicollinearity

Fernando Tusell [10] states that there are some indicators and statistical values that help to diagnose multicollinearity in multiple regression. Below, we present three basic rules for multicollinearity diagnosis. The first one is strictly related to multiple regression, and the other two are related to supervised classification.

- A large value for the coefficient of determination and the not significance of most of the parameters. In the presence of multicollinearity the estimated regression coefficients have a sign that is the opposite of what was expected. Moreover, its variance is also high, and because of that one gets the not significance of the parameters. In this case it seems that none of the predictor variables explains the response variable, whereas all of them, as a whole, do explain the response variable. The multicollinearity does not allow to clarify the contribution of each predictor variable.
- An eigenvalue of the correlation matrix with magnitude close to zero (zero in the case of perfect multicollinearity). In this case, because difference between the smallest and the greatest eigenvalue, the condition number of the correlation matrix will be large and therefore the multicollinearity is evident.
- A large value of the VIF for the predictor variables. If for some predictor variable $VIF \geq 10$, then the coefficient of determination for the regression of such variables versus the other variables is greater or equal to 0.9. This indicates dependence between the variables that are supposedly independent. Furthermore, it can be demonstrated that the VIF for each predictor variable is located in the main diagonal of the inverse of the correlation matrix.

3. Logistic Regression PLS

Bastien, Esposito Vinzi y Tenenhaus [1] presented an algorithm that transforms predictor variables (with multicollinearity) into latent variables, also called PLS components (with no multicollinearity). The authors of [1] illustrate their methodology by analyzing a data set named "Bordeaux". This data set corresponds to 34 years of observations of a French wine in terms of quality (Y): good, average, and poor. The predictor variables are: X_1 , the sum of the average daily temperatures (in $°C$); X_2, the duration of sunny weather (in hours); X_3, the number of very hot days; and X_4,the amount of rainfall (in mm). Without any multicollinearity analysis the investigators used the logistic regression as a classifier. They classified the data and found 7 classification errors, therefore the estimated error rate was $\frac{7}{34} = 20.6\%$. Using the method of Logistic Regression PLS the authors transform the four predictor variables into one PLS component and use the logistic regression classifier. They reclassified the data and found 6 errors, ergo the error rate is $\frac{6}{34} = 17.6\%$.

It has been observed that the PLS logistic regression method is efficient albeit the data that is analyzed have low multicollinearity. In no case the variance inflation factor (VIF) was greater than 10. The values of VIF for the predictor variables were: $VIF_1 = 4.7$, $VIF_2 = 4.7$, $VIF_3 = 4.0$ and $VIF_4 = 1.3$. The condition number was $\eta = 4.7$, which is lesser than 25. Thereby, the existence of multicollinearity is minimal or almost none.

Recently, Bertrand, Meyer and Maumy-Bertrand [2] presented a library for R called plsRglm: PLS generalized linear models for R. The library deals with PLS Regression for the case of multiple regression and with PLS logistic regression for the case of supervised classification. They also solve the classification problem for the "Bordeaux" wine data. For that problem the investigators compute all the possible PLS components (four in that case) and select the optimal number of components in the data in order to find the best model for classification. They did that by using the following criteria:

- Akaike Information Criterion (AIC).
- Bayesian Information Criterion (BIC).
- Misclassification error rate.

To select the number of components one must keep in mind that an overly simplistic model (too few components) produces a large approximation error (underfitting) whereas an overly complex model (too many components) produces a large estimation error (overfitting).

3.1. Selection Of The Number Of Components

Three criteria are used to select the number of components PLS: Akaike Information Criterion, Bayesian Information Criterion and the number of bad classifications. The manner in which the AIC and BIC criteria work is explained in [3].

1. The Akaike Information Criterion (AIC) estimates the relative distance between the unknown likelihood function of the data and the adjusted likelihood function of the model. Thus, a smaller AIC values means that the analyzed model is closer to the true model.

2. The Bayesian Information Criterion (BIC) estimates the posterior probability function that a model under a given bayesian configuration is the true model. Hence, a smaller BIC value means that is more probable that the analyzed model is the true model.

3. The Misclassification error rate: After constructing the classifier, the data that was used to construct the classifier is classified. Then the number of misclassifications is counted. Whenever the number of bad classifications is minimum then it is considered that the analyzed model is the best one.

4. Classifiers

We now present seven classifiers that are usually used in supervised classification: logistic regression, linear discriminant analysis, quadratic discriminant analysis, K-nearest neighbors with $K = 3$ and $K = 5$, naive Bayes, recursive partitioning, and regression trees (the latter two are classification trees).

4.1. Logistic Regression:

It is a regression model widely used for data analysis. In this case the response variable is binary and dichotome or in some cases polytome, whereas the predictor variables could be continuous or categorical. The logistic regression is a special case of the Generalized Linear Model (GLM), where the parameter estimation and hence the probability estimation is done using the maximum likelihood method [6].

4.2. Discriminant Analysis:

It is a multivariate analysis technique that constructs a classifier function based on multivariate data that belongs well-defined classes or groups. The goal is to assign new subjects to one of these groups. The classifier function is then constructed as a linear combination of a set of independent or predictor variables. If the covariance matrix of the groups under consideration is homogeneous, then we apply the Linear Discriminant Analysis, otherwise we apply the quadratic Discriminant Analysis [12].

4.3. K-NEAREST NEIGHBORS:

The classifier function K-Nearest Neighbor (KNN) is a simple classifier based on distance. A new subject will be classified into the most frequent class that its K-nearest neighbors belong to. For $K = 3$ and $K = 5$ (the most used values) there is a different classifier function [8].

4.4. NAIVE BAYES

It is a simple but efficient algorithm that predicts the class to which a new subject belongs to. It based on Bayes's theorem and the term naive is used because the algorithm uses bayesian techniques that do not consider possible dependencies between predictor variables [7].

4.5. CLASSIFICATION TREES

It is a classifier that recursively splits up the interval of possible values of the predictor variables. The goal is to construct logical networks and to establish rules that represent the knowledge of the problem through a tree structure. We used Recursive Partitioning and Regression Trees (rpart) as established in [4].

5. CLASSIFIER ERROR RATE

The classifier error rate is defined as the probability that a classifier function classify a new individual into a group that does not belong to (see Eq. (1)). The most commonly used classifier error rates are: the apparent, cross-validation leaving 1 out (cv-n), and cross-validation 10 (cv-10).

5.1. APPARENT ERROR RATE [5].

Although the apparent error rate is used by many investigators, its use is not recommended because is overly optimistic (usually yields low values) and has a high bias. Figure 1 illustrates the computation of the apparent error rate. We followed the following procedure in its computation:

1. A classifier function is constructed using all the data.
2. The classifier function classifies the data that was used to construct the classifier.
3. The number of misclassifications is counted.
4. The proportions of bad classifications are computed. It is the total number of bad classifications divided by the sample size.

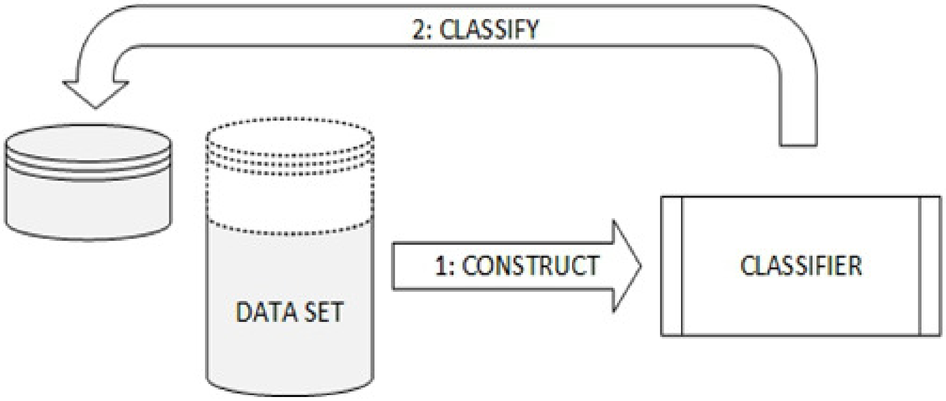


Figure 1: Apparent error rate.

5.2. Error Rate By 10-Fold Cross-Validation [9].

This method yields a more accurate error rate. Figure 1 illustrates the computation of this error rate. The following procedure was used to compute this error rate:

1. The data set is split into 10 subsets.
2. The classifier function is constructed using 9 of the 10 subsets of the sample.
3. The subset not used to construct the classifier is classified using the classifier function.
4. Steps 2 and 3 are repeated until all subsets are classified.
5. The number of bad classifications is counted.
6. The proportion of bad classifications is computed as the number of bad classifications divided by the sample size.

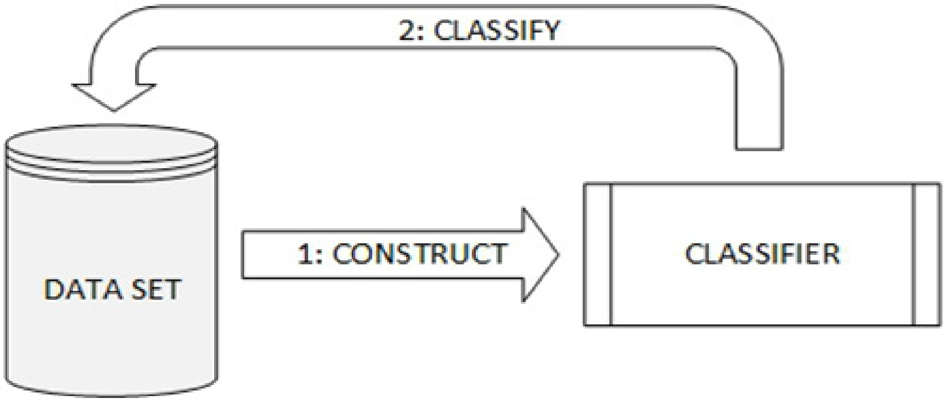


Figure 2: Error rate by cross validation.

5.3. Error Rate By Leave-One-Out Cross-Validation.

Error rate by cross-validation leaving 1 out. This method is also known as error rate by n-fold cross-validation. Akin to cross-validation 10, this method yields a more accurate error rate. Figure 2 shows the computation of the error rate by means of the following steps:

1. The data set is split into n parts, where n is the sample size.
2. The classifier function is constructed using $n-1$ parts of the sample.
3. The individual that was not considered for the classifier construction is then classified.
4. Steps 2 and 3 are repeated until all members of the sample are classified.
5. The number of bad classifications is counted.
6. The proportion of bad classifications is computed as the number of bad classifications divided by the sample size.

6. Datasets

Five different data sets were used in the present work. We describe such data sets below and in Table 1.

6.1 AUSTRALIAN DATA SET

The Australian database contains the characteristics of 690 clients of a financial institution. The dependent variable is "credit card" and there are 14 predictor variables. The dependent variable indicates whether or not the client obtains the credit card approval. The data set is available in https://archive.ics.ici.edu/ml/datasets.

6.2 HOUSEVOTES84 DATA SET

The data set includes the votes of the members of the House of Representatives of the United States of America, with over 16 key votes identified by the Congressional Quarterly Almanac (CQA). The number of predictor variables is 16 and the response variable has two possible values: republican or democrat. The variable number three was eliminated because it has the same values. The data is available in the repository of Machine Learning Databases of University of California at Irvine (UCI), http://www.ics.uci.edu/ mlearn/MLRepository.html

6.3 GERMAN DATA SET

This data set contains 20 variables of financial information of about 1000 loan applicants, and a classifier variable that expresses whether the applicant is a "good" client. The data is available in https://archive.ics.ici.edu/ml/datasets.

6.4 SONAR DATA SET

A database with 208 observations. Each one with over 60 variables and 2 classes. The data is available in the repository of Machine Learning Databases of UCI. https://archive.ics.ici.edu/ml/datasets.

6.5 COLON DATA SET

A data set that consists of microarray experiment results. The data contains 2000 attributes for two types of colon tissue: normal and tumor. The data is available in the Gene Expression Project webpage of Princeton University, http://microarray.princeton.edu/oncology.

Table 1. Data sets description

Name	Subjects	Predictors	Classes	Description
Australian	690	14	2	Clients
House Votes 84	232	15	2	Voters
German	1000	20	2	Clients
Sonar	208	60	2	Sonar signals
Colon	62	2000	2	Microarrays

7. IMPLEMENTATION AND RESULTS

The application of the methodology presented in this study used data from Table 1, each of these data sets were processed in the following manner:

1. Each data set, which are characterized by their original variables, was analyzed. Apparent error rate, leave-one-out cross-validation error rate (cv-n), and 10-fold cross-validation 10 (cv-10) error rate were calculated.
2. Each data set was transformed to PLS components, that were analyzed. First we examined the degree of multicollinearity of the predictor variables by means of the condition number. Second, the predictor variables were transformed to PLS (uncorrelated) components and the number of components used was determined by the

AIC, BIC and the misclassification error rate. These results are shown in Table 2. Finally, apparent error rate, leave-one-out cross-validation error rate (cv-n) and 10-fold cross-validation 10 (cv-10) error rate were calculated.

Table 2. Determination of the number of components PLS, each set of data

Set	**PLS components**	**AIC**	**BIC**	**wrong rated**
Australian η = 3.59	PLS_Comp_0	950.2	954.7	307
	PLS_Comp_1	479.2	488.3	98
	PLS_Comp_2	437.3	**450.9**	**87**
	PLS_Comp_3	**432.8**	451.0	90
	PLS_Comp_4	434.0	456.7	88
	PLS_Comp_5	436.0	463.2	86
House Votes 84 η = 8.58	PLS_Comp_0	322.5	326.0	108
	PLS_Comp_1	106.1	113.0	20
	PLS_Comp_2	47.1	57.4	10
	PLS_Comp_3	33.1	**46.9**	6
	PLS_Comp_4	**32.7**	50.0	**5**
	PLS_Comp_5	34.3	55.0	5
German η = 3.12	PLS_Comp_0	1223.7	1228.6	300
	PLS_Comp_1	985.1	995.0	236
	PLS_Comp_2	967.8	**982.5**	227
	PLS_Comp_3	**965.6**	985.3	**224**
	PLS_Comp_4	966.7	991.2	228
	PLS_Comp_5	968.6	998.0	233
Sonar η = 42.99	PLS_Comp_0	289.4	292.7	97
	PLS_Comp_1	210.8	217.5	55
	PLS_Comp_2	167.4	177.4	38
	PLS_Comp_3	142.6	156.0	27
	PLS_Comp_4	137.0	153.7	**23**
	PLS_Comp_5	123.0	143.1	24
Colon η = inf.	PLS_Comp_0	82.6	84.8	22
	PLS_Comp_1	60.6	64.8	16
	PLS_Comp_2	36.0	42.4	6
	PLS_Comp_3	17.5	26.0	2
	PLS_Comp_4	**10.0**	**20.6**	**0**
	PLS_Comp_5	12.0	24.8	0

Table 2 shows that House Votes 84, Australian and German datasets, have low multicollinearity, since the values of the condition numbers are 6.75, 8.58 and 3.12, respectively, which are all less than 25. Regarding the number of PLS components we observe that the whole Australian data set needs only 2 components from 14 predictor variables, the House Votes 84 dataset needs 3 components PLS from 15 predictive variables, and the German data set needs 2 PLS components from 20 predictor variables. Sonar and Colon datasets have high multicollinearity because their condition numbers are 42.99 and infinity, respectively (both greater than 25). Only 4 PLS components were used for Sonar dataset from 2000 predictor variables. Also, only 4 components were used for Colon dataset from 60 predictor variables.

Tables 3, 4, 5, 6 and 7, show the apparent error rates, leave-one-out cross-validation (cv-n) error rates, and 10-fold cross-validation (cv-10) error rates. These errors were calculated for the original data based on predictor variables and for processed data based on PLS components. In general, we note the following:

1. Apparent error rate is always lower than leave-one-out cross-validation and 10-fold cross-validation error rates, for both, original data and PLS components.
2. For data with low multicollinearity in their predictive variables, such as the Australian, House Votes 84, and German datasets, the calculation of the three types of error rates yielded almost the same value. The difference is that the error rates from transformed data were calculated considering a minimum number of PLS components: 2 components for 14 predictors of Australian, 3 components for 15 predictors of House Votes 84, and 2 components for 20 predictors of German.
3. For data with high multicollinearity in their predictive variables, such as Sonar and Colon datasets, the calculation of the three types of error rates yielded lesser values for processed data using PLS components compared with the error rates for the original data. The difference is that the error rates from transformed data were calculated considering a minimum number of PLS components: 4 components for 60 predictors of Sonar dataset and 4 components for 2000 predictors of Colon dataset.
4. Minimum error rate identifies the best classifier, which is not unique and depends on the data. The error rates that should be used to evaluate a classifier are 10-fold cross validation (cv-10) and leave-one-out cross-validation (cv-n), in that order. For the Australian dataset, the best classifier is logistic regression with original data and LDA with processed data; for House Votes 84 dataset the best classifiers are LDA and Rpart, with original data and logistic regression with processed data; for the German dataset the best classifier is LDA with original data and logistic regression with processed data; for Sonar dataset, the best classifier is knn-3 with original data and knn-3 and logistic regression with processed data. The best classifier for original Colon dataset, is logistic regression and for the processed data, the best classifiers are logistic regression and LDA.

Table 3. Australian dataset error rates

Method	Original data (14 Predictors)			2-component PLS		
	apparent	CV-n	CV-10	apparent	CV-n	CV-10
Reg. Logistic s	12.46	13.91	13.77	12.61	12.60	12.46
LDA	13.91	14.20	14.06	11.59	11.59	12,32
Qda	18.84	20.00	20.29	14.20	14.20	14.78
knn-3	16.38	32.75	32.61	9.86	14.93	13.77
knn-5	22.03	31.16	31.16	10.29	14.49	13.48
Naive Bayes	20.00	20.29	21.16	13.91	13.91	13.91
Rpart	11.74	12.17	14.35	12.03	13.48	12.75

Table 4. House Votes 84 dataset error rates

METHOD	Original data (15 Predictor)			3-component PLS		
	apparent	CV-n	CV-10	apparent	CV-n	CV-10
Reg. Logistics	2.16	6.47	6.90	2.59	2.59	2.59
LDA	3.02	3.02	3.02	3.02	3.02	3.02
Qda	3.45	NA	NA	3.45	3.45	3.45
knn-3	6.03	7.76	7.76	1.72	3.02	3.88
knn-5	7.76	8.62	8.62	2.59	3.45	3.45
Naive Bayes	5.17	5.17	7.33	6.47	6.90	6.90
Rpart	3.02	3.02	3.02	3.45	3.88	6.47

Table 5. German dataset error rates

METHOD	Original data (20 Predictor)			2-component PLS		
	apparent	CV-n	CV-10	apparent	CV-n	CV-10
Reg. Logistics	23.40	25.00	24.90	22.70	22.90	22.80
LDA	2310	24.20	24.50	22.70	22.80	22.90
Qda	22.00	26.90	26.70	2230	22,60	2310
knn-3	19.20	37.40	37.70	15.70	26.80	27.60
knn-5	25.10	35.10	35.40	18.50	25.60	26.00
Naive Bayes	24.50	25.50	26.30	2250	2250	23.00
Rpart	21.80	26.90	26.50	21.20	21.90	25.60

Table 6. Sonar dataset error rates

METHOD	Original data (60 Predictor)			4-component PLS		
	apparent	CV-n	CV-10	Apparent	CV-n	CV-10
Reg. Logistics	0.00	27.40	26.92	11.06	12.50	12.98
LDA	9.62	24.52	25.48	13.46	13.94	14.42
Qda	0.00	24.04	25.48	14.90	15.38	15.38
knn-3	11.06	18.75	20.67	8.17	12.50	12.98
knn-5	13.46	17.31	21.15	9.13	12.02	13.94
Naive Bayes	26.92	32.69	32.69	15.38	17.31	18.75
Rpart	12.50	33.17	29.33	9.62	1635	17.79

Table 7. Colon dataset error rates

METHOD	Original data (2000 Predictor)			4-component PLS		
	Apparent	CV-n	CV-10	apparent	CV-n	CV-10
Reg. Logistic s	0.00	51.61	4.84	0.00	4.84	3.23
LDA	3.23	22.58	19.35	1.61	1.61	3.23
Qda	3.23	NA	6.45	1.61	4.84	6.45
knn-3	8.06	14.52	14.52	6.45	9.68	12.90
knn-5	12.90	16.13	16.13	6.45	9.68	9.68
Naive Bayes	29.03	40.32	64.52	1.61	6.45	8.06
Rpart	8.06	41.94	79.03	9.68	20.97	25.81

8. CONCLUSIONS

1. In each dataset, the choice of the number of PLS components is independent of the degree of multicollinearity of the predictor variables. The Australian dataset has condition number $\eta = 3.59$ and 2 PLS components were selected. The House Votes 84 dataset has condition number $\eta = 8.58$ and 3 PLS components were selected. The German dataset has condition number $\eta = 3.59$ and 2 PLS components were selected. The Sonar dataset has condition number $\eta = 42.99$ and 4 PLS components were selected. The Colon dataset has condition number $\eta = \infty$ and 4 PLS components were selected.
2. BIC was the most frequent selection criterion of the number of PLS components in each dataset. The misclassification error rate criterion was used only for selecting the number of PLS components for the Sonar dataset. The selection criteria of the number of components, i.e. AIC, BIC, and misclassification error rate, for the Colon dataset agreed.
3. The dimensionality of each dataset was drastically reduced by use of transformation components PLS. Australian dropped 14 variables to 2 components PLS, House Votes 84 dropped from 15 variables to 3 components PLS, German dropped from 20 variables to 2 PLS components, Sonar was reduced from 60 variables to 4 PLS components and in Colon dropped from 2000 variables to 4 components PLS.
4. Apparent error rates of each of the classifiers, for all data sets, are on average slightly lower when using PLS components compared to the apparent error rates when using the original predictor variables.
5. 10-fold cross-validation (cv-10) of each one of the classifiers, for all data sets, are on average slightly higher compared with the leave-one-out cross-validation error rates, in both cases, using original variables and using PLS components.
6. 10-fold cross-validation (cv-10) and leave-one-out cross-validation error rates for all data sets using PLS components are generally lower than the equivalent error rates when using all the predictor variables. Here stands the benefit of working with PLS components, to achieve a significant decrease in the rate of error in all the classifiers.
7. There is not an ideal classifier with minimal error rate for any set of data. Analyzing 10-fold cross-validation error rates for each dataset, we found that the best classifier for Australian dataset is discriminant linear, for House Votes 84 and German datasets the best classifier is logistic regression, for Sonar dataset the best classifiers are logistic regression and knn-3, and for colon dataset the best classifiers are logistic regression and linear discriminant.

REFERENCES

[1] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.

[2] Gizem, Aksahya & Ayese, Ozcan (2009) Communications & Networks, Network Books, ABC Publishers.

[3] Philippe Bastien, Vincenzo Esposito-Vinzi, and Michel Tenenhaus. PLS generalised linear regression. Computational Statistics & data analysis, 48(1):17–46, 2005.

[4] F. Bertrand, N. Meyer, and M. Maumy-Bertrand. Package 'plsRglm'. version 1.1.1. R documentation, 2015.

[5] Guillaume Bouchard and Gilles Celeux. Selection of generative models in classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4):544–554, 2006.

[6] L Breiman, JH Friedman, RA Olshen, and CJ Stone. Classification and regression trees. CRC Press, 1984.

[7] Smith C. Some examples of discrimination. Ann. Eugenic, 18:272–282, 1947.

[8] Annette J Dobson and Adrian Barnett. An introduction to generalized linear models. CRC press, 2008.

[9] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schutze. An introduction to information retrieval. Cambridge University Press, 2008.

[10] Brian D Ripley. Pattern recognition and neural networks. Cambridge university press, 1996.

[11] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society. Series B (Methodological), pages 111–147, 1974.

[12] Fernando Tusell. Análisis de regresión. Introducción teórica y práctica basada en R. Adolescence. An age of opportunity, 2011.

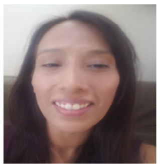
[13] José Carlos Vega-Vilca and Josué Guzmán. Regresión PLS y PCA como solución al problema de multicolinealidad en regresión múltiple. Revista de Matemática Teoría y Aplicaciones, 18(1):09–20, 2011.

[14] William N Venables and Brian D Ripley. Modern applied statistics with S. Springer-Verlag, 2002.

[15] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

AUTHORS

Dr. José C. Vega holds a Ph.D. degree in Computer and Information Sciences and Engineering from the University of Puerto Rico - Mayaguez Campus. He received his MS degree in Statistics from the University of San Marcos, Lima, Peru and his BS in Statistics from the Agraria La Molina University, Lima, Peru. He is a Professor in the Institute of Statistics and Information Systems of the University of Puerto Rico - Río Piedras Campus.



Roxana Aparicio received her Ph.D. degree in Computer and Information Sciences and Engineering from the University of Puerto Rico - Mayaguez Campus in 2012. She received her MS degree in Scientific Computing from the University of Puerto Rico and her BS in Computer Engineering from the University San Antonio Abad, Cusco, Peru. Currently she is professor in the Institute of Statistics and Information Systems of the University of Puerto Rico - Río Piedras Campus.

Aniel Nieves-Gonzalez received his Ph.D. in Applied Mathematics from the State University of New York at Stony Brook in 2010. He has a M.S. in applied mathematics from the University of Puerto Rico and his undergraduate degree is in Computer Science and Physics also from the University of Puerto Rico. He is currently an Assistant Professor at the Institute of Statistics and Computerized Information Systems at the University of Puerto Rico Rio Piedras Campus. He has published papers about mathematical models (differential equations) of complex systems physiological systems like the thick ascending limb (a part of the kidney). He still works in problems related to kidney physiology, but also works in problems related to coral population dynamics and spectral analysis of high frequency financial data. His research interests include dynamical systems, power spectral analysis, wavelet analysis, and parallel computing.