Mathematics and Statistics 8(2): 100-105, 2020
DOI: 10.13189/ms.2020.080205 http://www.hrpub.org

Comparison Analysis: Large Data Classification Using PLS-DA and Decision Trees

Nurazlina Abdul Rashid*, Norashikin Nasaruddin, Kartini Kassim, Amirah Hazwani Abdul Rahim

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Kedah, Malaysia

Received September 15, 2019; Revised November 10, 2019; Accepted November 17, 2019

Abstract Classification studies are widely applied in many areas of research. In our study, we are using classification analysis to explore approaches for tackling the classification problem for a large number of measures using partial least square discriminant analysis (PLS-DA) and decision trees (DT). The performance for both methods was compared using a sample data of breast tissues from the University of Wisconsin Hospital. A partial least square discriminant analysis (PLS-DA) and decision trees (DT) predict the diagnosis of breast tissues (M = malignant, B = benign). A total of 699 patients diagnose (458 benign and 241 malignant) are used in this study. The performance of PLS-DA and DT has been evaluated based on the misclassification error and accuracy rate. The results show PLS-DA can be considered as a good and reliable technique to be used when dealing with a large dataset for the classification task and have good prediction accuracy.

Keywords Classification, Large Data, PLS-DA, Decision Tree (DT)

1. Introduction

In multivariate classification, the aim of this method is finding the mathematical model who is able to identify the membership or grouping for each sample according to their appropriate class and the basis of a set of measurements. After the process of classification, the model calibrated will be used to find the membership of unknown samples and predicted the defined classes. Classification techniques not only applied to the quantitative variable but also able to handle the qualitative response. In order to classify the category in qualitative response, the mathematical relationship will be used to identify the relationship between a set of descriptive variables. Classification plays an important role in the real world applications in every field for determining the correct result.

The aim of this study is to investigate the performance of two different classification methods using PLS-discriminant analysis and Decision tree analysis for predicting the diagnosis of breast tissues. In medical science, the major problem occurs is getting the correct diagnosis for certain information. Surprisingly, to our knowledge, no research has been carried out on classifying the diagnosis of breast tissues (M = malignant, B = benign) using PLS-DA. Therefore, for the purpose of getting the ultimate diagnosis, this paper attempts to provide a more detailed investigation regarding the effects of the performance of PLS-DA and decision tree to classifying the diagnosis of breast tissues.

1. 1. PLS Discriminant Analysis (PLS-DA)

Partial Least Square Discriminant Analysis is a linear classification method that is based on the PLS regression algorithm [8, 16]. This method is a combination of properties of partial least squares regression and the discrimination power of classification technique. In PLS-DA it is dealing with dependent variable Y and the presence of several dependent Y variables in searching for the latent variable with a maximum covariance with Y variables [1]. One of the advantages of PLS-DA is data variability. The data variability is modeled and called Latent Variables. The latent variable score and loading allowed graphical visualization and show the different data and their relation. For the purpose of identifying the latent variable number, the cross-validation method was applied. However, the problem of PLS-DA has occurred when the variable number was increased. When the number of variables increases, it is difficult to search the proper size of the relevant subspace of variable space [6]. For small datasets, usually, the unstable result appears for covariance but for a large sample, some extensive computation time will be needed. However, in recent studies, the selection of variables in the PLS-based algorithm was to attract more attention among researchers [2, 5, 17].

1.2. Decision tree analysis (DT)

The decision tree [9] is the most important technique in classification problems of breast cancer database and medical field. There are two basic steps in decision trees. First is construct the tree and then applying the tree to the database. The decision tree algorithm creates user-friendly rules that indicate important attributes, requires less calculation and easy to understand contrasted to other algorithms such as Neural Networks [13]. The main advantages of the decision tree are flexible, easy to build, easy to debug and suits for classification and regression. In this research work, the simulation results assure that the priority-based decision tree algorithm is for SEER breast cancer dataset [11]. B. Padmapriya and T. Velmurugan [3], discussed several algorithms such as C4.5, ID3, and CART (Classification and Regression Trees) to classify the data using decision trees. The CART algorithm is chosen to classify the breast cancer data because it provides better precision for medical data sets than ID3. The decision tree gives a powerful technique for classification and prediction in Breast Cancer diagnosis problem [14]. Various supervised learning classifiers are available to classify the data, including Multi-Class classifiers, Decision trees (J48), Naïve Bayes, SMO, KNN, bagging, DNTB, AD Tree & Rep tree are compared, to identify the best classifier using the breast cancer dataset. The experimental results show that the classification result with the decision trees algorithm is more exact than other classifiers by 75.52% was discussed by S. Joshi and A. V. Vidyapeetham [15]. Therefore, this study to investigate the performance of PLS-DA and decision tree to evaluate large dataset for predicting the diagnosis of breast tissues.

Table 1. Wisconsin breast cancer dataset

No.	Variables	Domain
1	Clump Thickness	1-10
2	Uniformity of Cell Size	1-10
3	Uniformity of Cell Shape	1-10
4	Marginal Adhesion	1-10
5	Single Epithelial Cell Size	1-10
6	Bare Nuclei	1-10
7	Bland Chromatin	1-10
8	Normal Nucleoli	1-10
9	Mitoses	1-10
10	Class	1(Benign) or 4(Malignant)

2. Materials and Methods

2.1. Data

The large sample data of breast tissues were obtained from the University of Wisconsin Hospital. The sample will consist of 699 patients diagnose, it contains 458 benign and 241 malignant. The aim is to predict true disease status based on nine different variables. The data is divided into a training sample (used to build the model) which is 70% of patients diagnose and the testing sample is 30% (used to evaluate the performance of the model). The PLS-DA method is then compared with DT to determine the most efficient model. The performance of PLS-DA and DT has been evaluated based on the misclassification error rate and the percentage of testing samples that are correctly classified by the model evaluated by the accuracy rate. The descriptions of the dataset were given in table 1.

2.2. Construction of PLS-DA

PLS-DA used for constructing predictive models when there is a lot of independent variables and high multicollinearity. PLS-DA also allows the series of equations to be analyzed simultaneously while traditional regression may require separate regression equations to analyze.

In a standard variant of PLS-DA, the components are required to be orthogonal to each other. Its components are orthogonal so that PLS-DA is not affected by multicollinearity. This is employed in the package mixOmics, the principal components of PLS-DA can be formulated as the eigenvectors of the non-singular portion of the covariance matrix C, given by:

$$C = \frac{1}{n-1} X^T C_n X \quad (1)$$

where,

C_n is then $n \times n$ centering matrix

The iterative process computes the transformation vectors (also, loading vectors) $a_1 .. . a_d$, which gives the importance of each feature in that component. In iteration h, PLSDA has the following objective:

$$\max \operatorname{cov}(x_n a_n, y_n b_n) \quad (2)$$

where b_n is the loading for the label vector y_n, $X_1 = X$, and X_n and y_n are the residual (error) matrices after transforming with the previous $n - 1$ component.

2.3. Step for built PLS-DA model

Perform the PLS-DA to create a pseudo linear Y value against which to correlate the samples. Specify the number of components, or latent variables (LVs), to use for our data. Then, plot the score between latent variables in order to look up the separation between sample groups. If the samples causing the problems filters that sample. Construct PLS-DA model based on score and weight after the filter. Evaluate the performance of the constructed PLS-DA model based on the minimum misclassification rate.

$$\sum_{i=1}^{n} \frac{error_i}{n_i} \quad \text{where, } i = 1,2,\ldots, n \tag{3}$$

2.4 R-coding for PLS-DA

For constructing PLSDA predictive models, we used package mixOmics for predicting the diagnosis of breast tissues.

```
library(mixOmics)
library(foreign)
mydata<-read.csv("C:\\Users\\USER\\Desktop\\dbc.
csv",header = TRUE)
summary(mydata)
mydata$Class <-as.factor(mydata$Class)
table(mydata$Class)
Y <- as.factor(mydata$Class)
X<-(mydata[,2:10])
```

Next, the training and testing sample is set up, then run a PLS-DA model and look at the prediction for the testing samples:

```
##Data partition
set.seed (1234)
partition<-sample(2, nrow(mydata), replace=TRUE,
prob=c(0.7,0.3))
## For PLS-DA, train and testing the model
train<-mydata[partition==1,]
test<-mydata[partition==2,]
Ytrain<-train$Class
Xtrain<-train[,2:10]
Ytest<-test$Class
Xtest<-test[,2:10]
plsda.train <- plsda(Xtrain, Ytrain, ncomp = 2)
## Then predict
train.predict <- predict(plsda.train, Xtrain, dist =
"max.dist")
Prediction.train <- train.predict$class$max.dist[, 2]
cbind(Y = as.character(Ytrain), Prediction.train)
tab1<-table(Ytrain,Prediction.train)
print(tab1)
test.predict <- predict(plsda.train, Xtest, dist =
"max.dist")
Prediction.test <- test.predict$class$max.dist[, 2]
cbind(Y = as.character(Ytest), Prediction.test)
tab2<-table(Ytest,Prediction.test)
print(tab2)
```

2.5. Construction of Decision Trees

Decision trees are indispensable graphical tools in such settings and displayed in a simple, easy-to-understand format. Each branch of the decision tree represents a possible decision or occurrence. The target variable can be a categorical and continuous variable. Decision tree model will calculate the probability that a given data belongs to each of the target variables or to classify the data by assigning it to the most likely category [4].

Classification trees apply to data where the target variable (outcome) is a classification label, such as the disease status of a patient. Classification trees are decision trees derived using recursive partitioning data algorithms that classify each case into one of the class labels for the outcome. A classification tree consists of three types of nodes, which are root node (the top node of the tree comprising all the data), splitting node (a node that assigns data to a subgroup) and terminal node (final decision or outcome).

2.6. R-coding for DT

We used package foreign for predicting the diagnosis of breast tissues.

```
library(foreign)
file.choose()
DBS3=read.spss("C:\\Users\\User\\Desktop\\dbs
spss.sav",to.data.frame=TRUE)
head(DBS3)
```

The dataset has divided to 70% for training and 30% for testing.

```
##Data partition
set.seed (1234)
partition<-sample(2, nrow(DBS3), replace=TRUE,
prob=c(0.7,0.3))
DBStraining<-DBS3[partition==1,]
DBStesting<-DBS3[partition==2,]
```

Then run a DT model for the prediction and testing samples:

```
##Decision Tree model
library(rpart)
DTM<-rpart(Class~.,DBStraining,method="class")
DTM
plot(DTM)
text(DTM)
rpart.plot(DTM)
rpart.plot(DTM, type=4, extra=101, fallen.leaves=T)
DTMtraining<-predict(DTM,DBStraining,
type="class")
tab<-table(DTMtraining, DBStraining$Class)
print(tab)
sum(diag(tab))/sum(tab)
DTMtesting<-predict(DTM,DBStesting,
type="class")
tab<-table(DBStesting[,10],DTMtesting)
print(tab)
sum(diag(tab))/sum(tab)
```

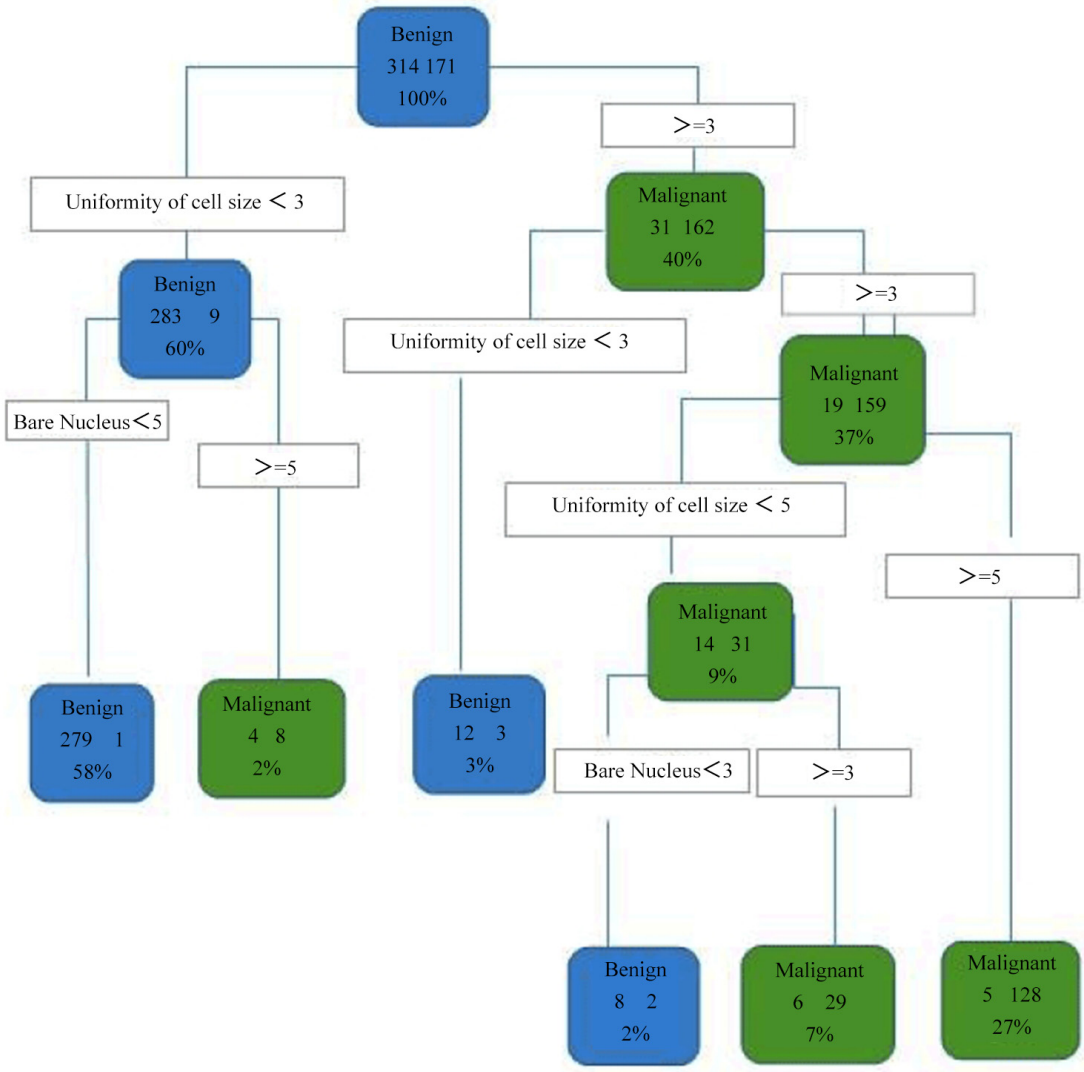

Figure 1. Decision Tree Rules

Figure 1 shows the decision tree rules to classify the diagnosis of breast tissues (M = malignant, B = benign). For benign, the patient's uniformity of cell size is less than 2.5 and the bare nucleus is less than 4.5. The patient's uniformity of cell size is greater than equal to 2.5 and the uniformity of cell shape is less than 2.5. The patient's uniformity of cell size is between 2.5 (included) to 4.5, uniformity of cell shape greater than equal to 2.5 and bare nuclei less than 2.5. Then, malignant shows the patient's uniformity of cell size is less than 2.5 and bare nuclei greater than equal to 4.5. The patient's uniformity of cell size is between 2.5 (included) to 4.5, uniformity of cell shape is greater than equal to 2.5 and bare nuclei greater than equal to 2.5. The patient's uniformity of cell size greater than equal to 4.5 and the uniformity of cell shape greater than equal to 2.5.

3. Result and Analysis

The investigations based on 70% number of training samples and 30% of the testing sample are conducted to compare the performance of the PLS-DA model with DT based on their misclassification and accuracy rate.

Table 2. Classification for training and testing sample

		Actual	Predicted	
Partition	Model	Group	Benign	Malignant
Training	PLS-DA	Benign	306	8
		Malignant	11	160
	DT	Benign	297	5
		Malignant	17	166
Testing	PLSDA	Benign	142	2
		Malignant	6	64
	DT	Benign	136	8
		Malignant	4	66

Table 3. Comparison and performance analysis of the PLS-DA and DT

Partition	Model	Accuracy (%)	Misclassification (%)
Training	PLSDA	(96.08)	(3.97)
	DT	(95.46)	(4.50)
Testing	PLSDA	(96.26)	(3.74)
	DT	(94.39)	(5.60)

Table 2 shows the classification for training and testing samples for a large sample size (n = 699). PLS-DA presents 306 true predict of benign and 160 malignant, while DT predicts 297 of benign class and 166 malignant class for training sample. PLS-DA shows the number of corrected for the testing sample from benign class is 142 and DT is 136. Then, malignant is 64 and 66 respectively. Table 3 summarizes the results of the performance analysis between PLS-DA and DT. Decision tree (DT) has the lowest accuracy rate in training and testing samples compared to PLS-DA but the difference is too small. The difference in accuracy rate in training is 0.62% and testing is 1.87%. PLS-DA presents a lower misclassification rate compared to DT for both classification training and testing. The percentage of corrected PLS-DA is 96.08% for training and 96.26% for testing. Then, training of DT is 95.46% and 94.39% for testing. However, the difference between these two models is small.

4. Conclusion

PLS-DA model was chosen as the best predictive model in predicting the category of breast cancer class since the results show it the best accuracy rate for training and validation samples. In conclusion, the results of this study indicate that PLS-DA can be considered as a good and reliable technique to be used when dealing with a large dataset for the classification task because of the advantage that its components are orthogonal and hence not affected multicollinearity.

Acknowledgments

The research would like to thank the University of Wisconsin Hospital for the breast cancer databases provided the sample data for this study and UiTM Kedah for financial support to publish this paper.

REFERENCES

[1] Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*(16), 3790-3798.

[2] Bu, H. L., Li, G. Z., Zeng, X. Q., Yang, J. Y., & Yang, M. Q. (2007, October). Feature selection and partial least squares based dimension reduction for tumor classification. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering* (pp. 967-973). IEEE.

[3] B. Padmapriya and T. Velmurugan, "Classification Algorithm Based Analysis of Breast Cancer Data," vol. 5, no. 1, 2016.

[4] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984). Classification and regression trees. Belmont, Calif.: Wadsworth.

[5] Helmbold, D. P., & Schapire, R. E. (1997). Predicting nearly as well as the best pruning of a decision tree. *Machine Learning*, *27*(1), 51-68.

[6] Li, Y. Q., Liu, Y. F., Song, D. D., Zhou, Y. P., Wang, L., Xu, S., & Cui, Y. F. (2014). Particle swarm optimization-based protocol for partial least-squares discriminant analysis: application to 1H nuclear magnetic resonance analysis of lung cancer metabonomics. *Chemometrics and Intelligent Laboratory Systems*, *135*, 192-200.

[7] Mansour, Y. (1997, July). Pessimistic decision tree pruning based on tree size. In *Machine Learning-International Workshop Then Conference*- (Pp. 195-201). Morgan Kaufmann Publishers, Inc.

[8] Nkansah, K., Adedipe, O., Dawson-Andoh, B., Atta-Obeng, E., Slahor, J., & Osborn, L. (2015). Determination of concentration of ACQ wood preservative components by UV-Visible spectroscopy

[9] P. A. Punde and M. E. Jadhav, "Advances in Computational Research Mining Of Breast Cancer Database For Classification Using Decision Trees," Vol. 7, No. 1, Pp. 185–186, 2015.

[10] Pérez-Enciso, M., & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Human genetics*, *112*(5-6), 581-592.

[11] P. Hamsagayathri, "Priority Based Decision Tree Classifier for Breast Cancer Detection," pp. 4–9, 2017.

[12] (PDF) Research Challenges and Comparative Study of Various Classification Technique Using Data Mining. Available from: https://www.researchgate.net/publication/267763252_Research_Challenges_and_Comparative_Study_of_Various_Classification_Technique_Using_Data_Mining [accessed Apr 17 2019].

[13] R. Varshney and V. K. Gupta, "Diagnosis of Breast Cancer using Decision Tree Models and SVM," pp. 2845–2848, 2018.

[14] R. Sumbaly, "Diagnosis of Breast Cancer using Decision Tree Data Mining Technique," no. March 2015.

[15] S. Joshi and A. V. Vidyapeetham, "Classification of Breast Cancer Dataset by Different Classification Algorithms," pp. 4–7, 2017.

[16] Wang, T., Wu, H. L., Long, W. J., Hu, Y., Cheng, L., Chen, A. Q., & Yu, R. Q. (2019). Rapid identification and quantification of cheaper vegetable oil adulteration in camellia oil by using excitation-emission matrix fluorescence

spectroscopy combined with chemometrics. *Food Chemistry*.

[17] Yin, S., Wang, G., & Gao, H. (2016). Data-driven process monitoring based on modified orthogonal projections to latent structures. *IEEE Transactions on Control Systems Technology*, *24*(4), 1480-1487