Journal of Statistical Science and Application, April 2016, Vol. 4, No. 03-04, 65-84
doi: 10.17265/2328-224X/2015.0304.001

Comparison of Different Classification Methods on Glass Identification for Forensic Research

Suchismita Goswami* and Edward J. Wegman
Computational and Data Science, George Mason University, Fairfax VA 22030

Classification methods play an important role in investigating crime in forensic research. Here we assess the relative performance of several classification methods, such as Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Mixture Discriminant Analysis (MDA) and Classification Tree (CT) on glass identification data. We present a different approach to investigate the relative performance of the classifiers by invoking the tests of statistical significance and the receiver operating characteristic (ROC) curves in addition to estimating the probabilities of correct classification (PCC). The area under the receiver operating characteristic curve (AUC), the error rate and its 95% confidence interval are used to measure predictive power of these algorithms. Dimensionality reduction of data has been conducted using principal component analysis (PCA) and Fisher's linear discriminant analysis (FDA) and two major components were identified. Among all the classification methods mentioned above, the LDA and the QDA are observed to be statistically significant. The Box's M test ($P<0.0001$), which is used to test the homogeneity of covariance matrices, showed that the homogeneity of covariance could not be assumed for LDA. This suggests that for glass types, window and non-window, the QDA is superior to all methods. The CT, however, has been found to outperform FDA when all six categories of glass are considered.

Keywords: Forensic Research, PCA, LDA, QDA, CT

Introduction

Glass is available in many different forms and chemical compositions. It can be commonly found in windows or doors of a house, in kitchen utensils, in cars or vehicles. The property of the glass, particularly the refractive index, depends on composition and treatments of the glass. The typical glass contains oxides of Si, Mg, Ba, Na and other oxides [Terry et al., 1983]. In forensic investigation, glass fragments are investigated in order to determine whether the glass fragments obtained from an individual belong to window or non-window glass [Bottrell, 2009]. Several procedures, based on compositions and refractive index (RI), are available to identify the glass types. In many cases it is difficult to get a well-defined boundary separating the window from non-window glass just looking at the composition and RI, as the composition and RI of window and non-window based glass overlap to a certain extent [Terry et al., 1983]. The motivation of this study is to use and analyze different machine learning techniques in order to facilitate crime investigations.

A wide range of classification or machine learning techniques applied to glass identification dataset have included linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularized discriminant

Corresponding Author: Suchismita Goswami, e-mail: sgoswam3@masonlive.gmu.edu

analysis (RDA), and k-nearest neighbor (kNN) method [Aeberhard et al., 1994]. In order to investigate the relative performance of the classification methods, they estimated the probabilities of correct classification (PCC) of classifiers and evaluated each of classifiers by the leave-one-out method with the real data. However, the leave-one-out method is very computationally expensive and it yields a non-stratified sample as it has only one instance in the test set. They reported that the PCC for LDA, QDA, RDA and 1NN is 71.2%, 62.6%, 74.2% and 81.0%, respectively, suggesting that kNN could be the efficient method to classify the glass data. Although kNN is the most popular and powerful classification technique, it is not a robust technique for noise as well as outliers in the training data set, and for it to be effective the training data set needs to be relatively large. Neural network ensemble has been proposed to edit the training data sets in order to improve the performance of kNN classifier [Jiang and Zhou, 2004, Ferri et al., 1999]. Besides, it is not very clear which model is the most effective model for providing an optimally separating hyperplane. Also the statistical procedures to measure the predictive power of these algorithms are not taken into account.

Here we present a different approach to assess the relative performance of the classifiers by invoking the tests of statistical significance and the receiver operating characteristic (ROC) curves in addition to estimating PCC. The objective of this research, therefore, is to analyze decision boundaries for various classifiers on a glass identification dataset and predict a suitable classification method. In order to decrease the computational cost and increase the memory usage of high dimensional data for many classification algorithms, dimensionality reduction techniques can be used [Janecek et al., 2008]. These techniques also improve the clarity in data visualization. Here we perform two dimensionality reduction techniques, principal component analysis (PCA) and Fisher discriminant analysis (FDA) on the glass dataset. We then employ five different algorithms or methods, such as logistic regression (LR), LDA, QDA, mixture discriminant analysis (MDA) and classification tree (CT) to investigate the location of decision boundaries on the glass identification dataset with reduced dimensions. We show that for glass types, window and non-window, the QDA is superior to all methods. However, the CT is found to outperform FDA when all six categories of glasses are considered.

We organize the remainder of this paper as follows. At first, we present the descriptive statistics to describe the distribution of data using parallel coordinates. The dimensionality reduction of data is presented in section Dimension Reduction using PCA and FDA. Several classification methods applied to the reduced dimensional data are discussed in section Classification Methods. Results of the different classification methods and the related discussion are presented in section Results and Discussion. The relative performance of classification methods is shown in section Classifier Performance for Two Classes. We then present results for multiple categories in section Multiple Categories using classification trees and Fisher's LDA before the concluding remarks in section Conclusions.

Descriptive Statistics and Exploratory Data Analysis

Glass Identification Data is obtained from [Lichman, 2013]. This dataset has 214 instances and 11 variables. Out of 214 total instances, 163 are window glasses and 51 are non-window glasses with no missing values. The quantitative variables are RI and elemental compositions, consisting of oxides of Na, Mg, Al, Si, K, Ca, Ba and Fe. The type of glass is class attribute, such as building windows float processed, building windows non-float processed, vehicle windows float processed, containers, tableware and headlamps. Among the window glasses, 87 are float processed and 76 are non-float processed. The float processed group consists of 70 building windows and 17 vehicle windows. The non-float processed group contains 76 building windows and 0 vehicle

windows. There are 13 containers, 9 tableware and 29 headlamps among 51 non-window glass.

Fig. 1 is a parallel coordinate plot, which is an efficient way to represent multidimensional data, illustrating some interesting data structures, such as the one-dimensional features (marginal densities) and two-dimensional features (correlation and nonlinear structures) [Wegman, 1990]. It is observed from the parallel plot that the Si, Al, Fe, Na, and RI are approximately normally distributed. However, the distribution of the Ba appears to be right skewed. Clusters are detected on the Mg. The other interesting feature is the crossing between Na and RI, Na and Mg, Al and Mg, suggesting a negative correlation. The high level of Al would tend to have low level of Mg, and the high level of Mg would tend to have low level of Na. We can see an approximate parallelism and relatively fewer crossings between K and Si and Si and Al, suggesting a positive correlation. Also one can observe the negative slope connecting the low Ca to moderate to high K which suggests the presence of an outlier. The left side of the relationship between Fe and Ba shows an approximate hyperbolic boundary, and the right side displays the crossover effect illustrating that for low Fe or low Ba, there seems to be a very little correlation.

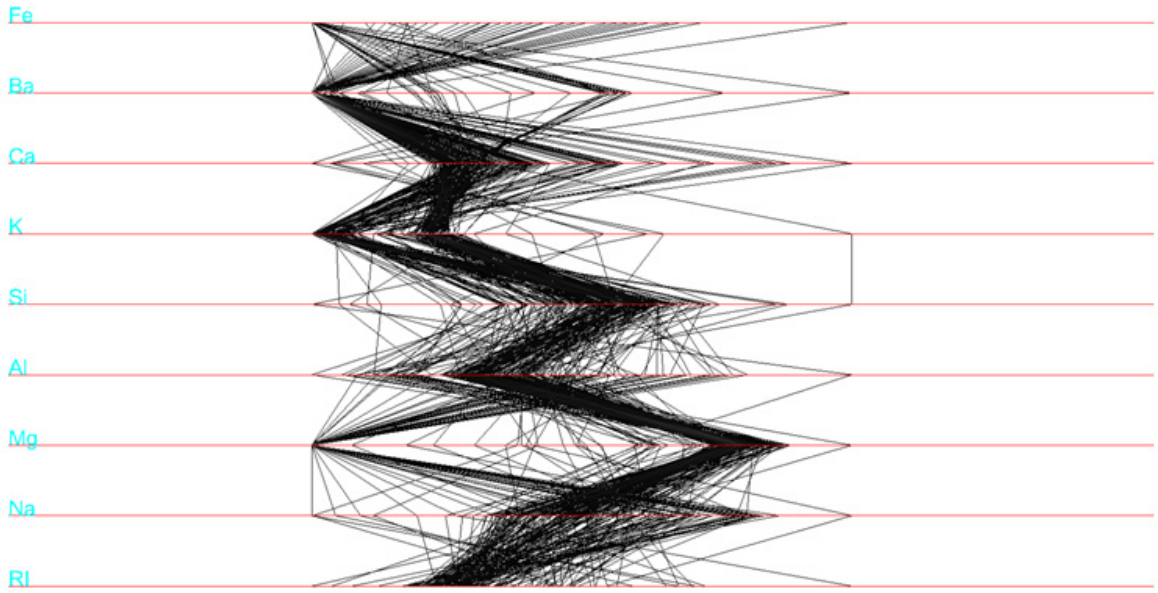

Figure 1. Parallel coordinate plot describing the marginal densities and correlations of variables.

Effect of composition and RI on glass type

In order to investigate the effect of composition and RI on the types of glass, the scatter plots are color coded [Wegman and Dorfman, 2003]. Here we present three scatter plots, Al-Mg, Al-RI and Ca-RI. The building windows float processed glass is brushed with red, building windows non-float processed with green, vehicle window float processed with blue, containers with yellow, tableware with magenta and headlamps with cyan. In figure (see right panel of Fig.2 (a)), the glass with a high level of Mg and a low level of Al are likely to be classified as building windows float processed, and a high level of Mg and a medium level of Al as building windows non-float processed. Glass with a medium RI and a low Al are likely to be classified as building windows float processed and a high level of Al and a low level of RI, however as headlamps (see right panel of Fig. 2 (b)). Glass with a high level of RI and Ca appears to belong to windows non-float processed and a low level of RI and Ca to headlamps (see right panel of Fig. 2(c)).

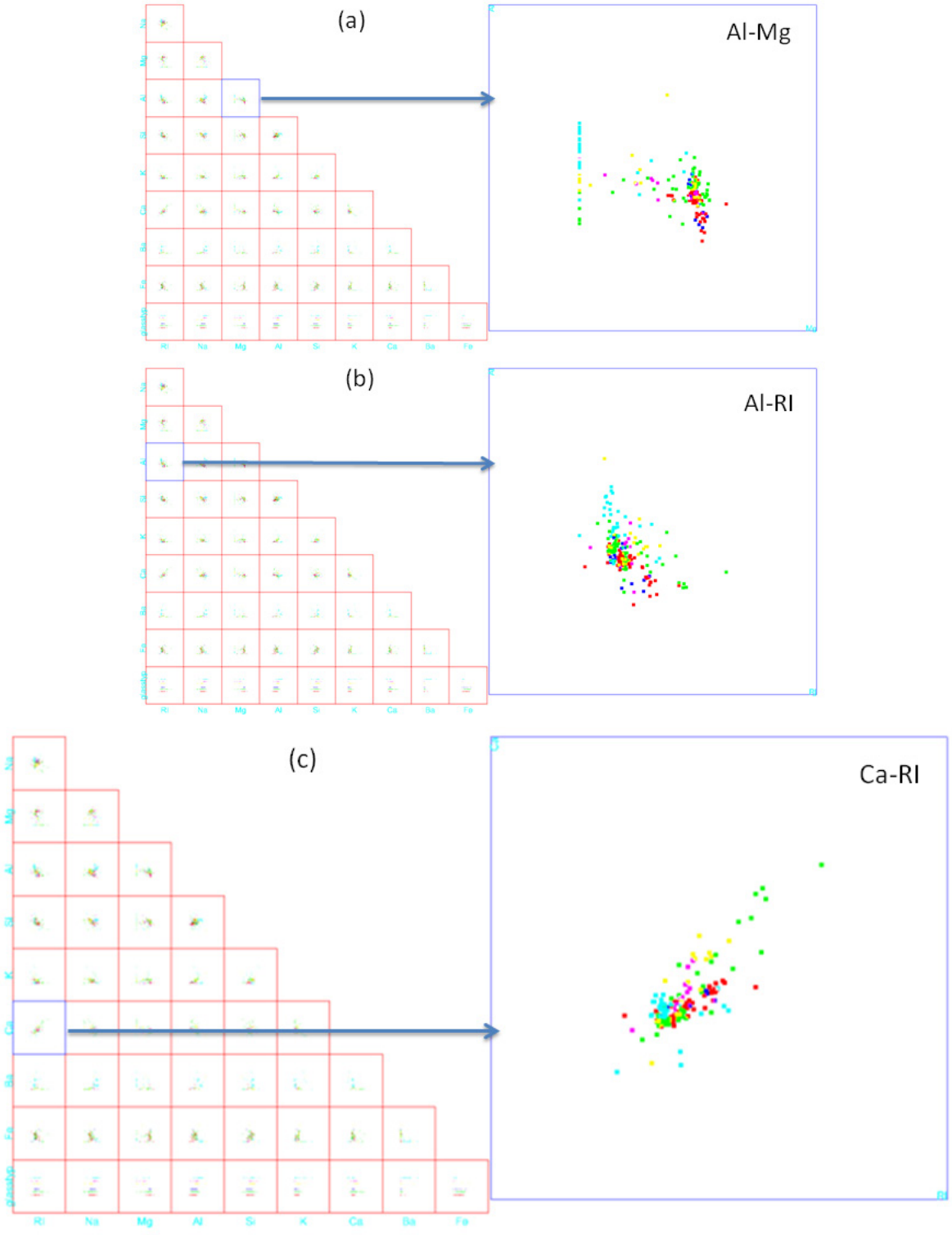

Figure 2. (a) The right panel showing the color coded scatter plot of Al and Mg. (b) The right panel showing the color coded scatter plot of Al and RI. (c) The right panel showing the color coded scatter plot of Ca and RI.

Dimension Reduction using PCA and FDA

PCA

It is a well-known dimensionality reduction method. Here the data are $X_i \in R^9, i = 1,\ldots, 210$. This method uses eigenvalue decomposition of a 9×9 covariance matrix of the data, X. The first principal component is the linear combination of variables that explains the maximum variance in the data set.

The first and second principal components can be written as [Hastie et al., 2011, James et al., 2013]:

$$Y_1 = b_{11}\ x_1 + b_{12}\ x_2 + b_{13}\ x_3 + b_{14}\ x_4 + b_{15}x_5 + b_{16}x_6 + b_{17}x_7 + b_{18}x_8 + b_{19}x_9 \quad (1)$$

$$Y_2 = b_{21}\ x_1 + b_{22}\ x_2 + b_{23}\ x_3 + b_{24}\ x_4 + b_{25}x_5 + b_{26}x_6 + b_{27}x_7 + b_{28}x_8 + b_{29}x_9 \quad (2)$$

Where Y_1 is the first principal component, Y_2 is the second principal component, x_i are the original variables, b_{ij} the weight or loadings associated with variables. In matrix notation, the above equations for glass dataset can be written as, $Y = b\,X$, where Y is a matrix (9×210) of principal components , X is a matrix (9×210) of original variables and its variance-covariance matrix is Σ, b is a matrix (9×9) of loadings, $var(Y) = b\Sigma b^T$ and $\Sigma = b^T D b$; where D is a diagonal matrix and diagonal entries are eigen values, $D = diag\ (d_1, d_2, ..., d_9)$. The data, X, is standardized, and it is assumed that b is an orthonormal matrix [Shlens, 2005].

All variables except Fe that are correlated sufficiently (r>0.3) have been included in PCA. The variance vs. component plot, Fig. 3 (a), shows a reasonable drop after three components. However, we identify only two components based on the values of the loadings. It was observed that Al, Ba, and Na contribute only to the principal component 1 (PC1) (see Fig. 3(b) and table 1) and Ca, RI, Mg and k contribute to the principal component 2 (PC2). With these variables in mind we have named the principal component one as composition, and the principal component two as composition dependent refractive index. The first component accounts for 29.63%, the second component 26.16% and third component 15.15% and fourth component 11.31% of the total variance. The first four components account for 82.25% of total variance and the first two components account for around 56% of total variance in original variables.

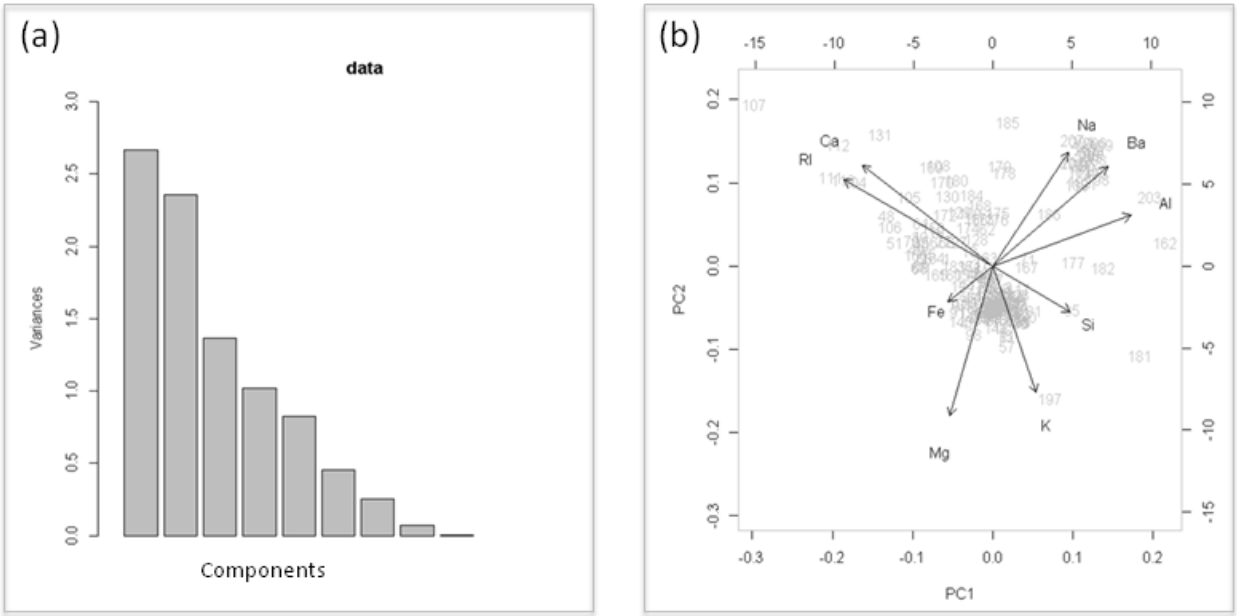

Figure 3. (a) Barplot of variances explained by the principal components with observations for Fe removed. (b) Biplot (scaled) of the first two principal components with observations for Fe removed.

Table 1

Component loadings for rotated components

	Component		
	1	2	3
Ba	0.838	- 0.029	0.013
Na	0.783	0.006	- 0.468
Al	0.725	- 0.196	0.356
Ca	- 0.238	0.951	0.058
RI	- 0.289	0.790	- 0.416
Mg	- 0.617	- 0.619	- 0.436
K	- 0.272	- 0.604	0.251
Si	0.022	- 0.171	0.872

Extraction Method: Principal Component Analysis

Rotation Method: Varimax with Kaiser Normalization.

Rotation converged in 5 iterations

FDA

In this method, the aim is to obtain the uncorrelated linear combination of original variables, termed as discriminant functions. This procedure maximizes the between-to-within class variance. In doing so, maximum discrimination among classes is obtained. The first two discriminant functions are:

$$Y_1 = c_{11}\ X_1 + c_{12}X_2 + c_{13}\ X_3 + \ldots + c_{1p}\ Xp \quad (3)$$

$$Y_2 = c_{21}\ X_1 + c_{22}X_2 + c_{23}\ X_3 + \ldots + c_{2p}\ Xp \quad (4)$$

Where Y_1 is the first discriminant function, Y_2 is the second discriminant function, X_i, i = 1, 2, 3,…, p are the input variables and c_{1i}, i = 1, 2, . . ., p are the weights associated with input variables for the discriminant functions. In matrix notation, the equations can be written as $Y = c^T X$. The covariance matrix of X, Σ is defined as Σ = B + W where B is the between-class variance of X and W is the within-class variance of X. The between-class variance of Y is $c^T Bc$, and the within-class variance of Y is $c^T Wc$. Fisher-LDA maximizes the objective:

$$F(\mathbf{c}) = \frac{\mathbf{c^T B\ c}}{\mathbf{c^T W\ c}}$$

Here **c** is obtained by the eigenvector of $W^{-1}B$ corresponding to the largest eigenvalue [Hastie et al., 2011]. At most min (p, K–1) positive eigen values exist where p is the number of input variables and K is the number of classes. We observe that the first linear discriminant explains about 87% of the between-class variance and the second linear discriminant explains 8.05% of the between-class variance. Therefore, the first two discriminants explain 94.93% between-class variance in the glass identification data, suggesting that the two discriminant functions are sufficient to adequately fit the data.

Visualizing the difference between PCA and FDA

Both PCA and Fisher's LDA can be used for dimensionality reduction. Fisher's LDA is a supervised technique and uses class information. On the other hand, PCA is an unsupervised learning technique. These two approaches are very different. In fact, PCA preserves most of the variability in the data while Fisher's LDA captures most of the between-class variance in the data. The upper and lower panels of Fig. 4 show LDA and PCA, respectively. We can see that container is distinctly separated from tableware in FDA, however some

overlap exists in PCA. Here we consider PCA over FDA as PCA models total variability of data as opposed to the difference between class, and PCA is also superior to FDA under conditions of colinearity.

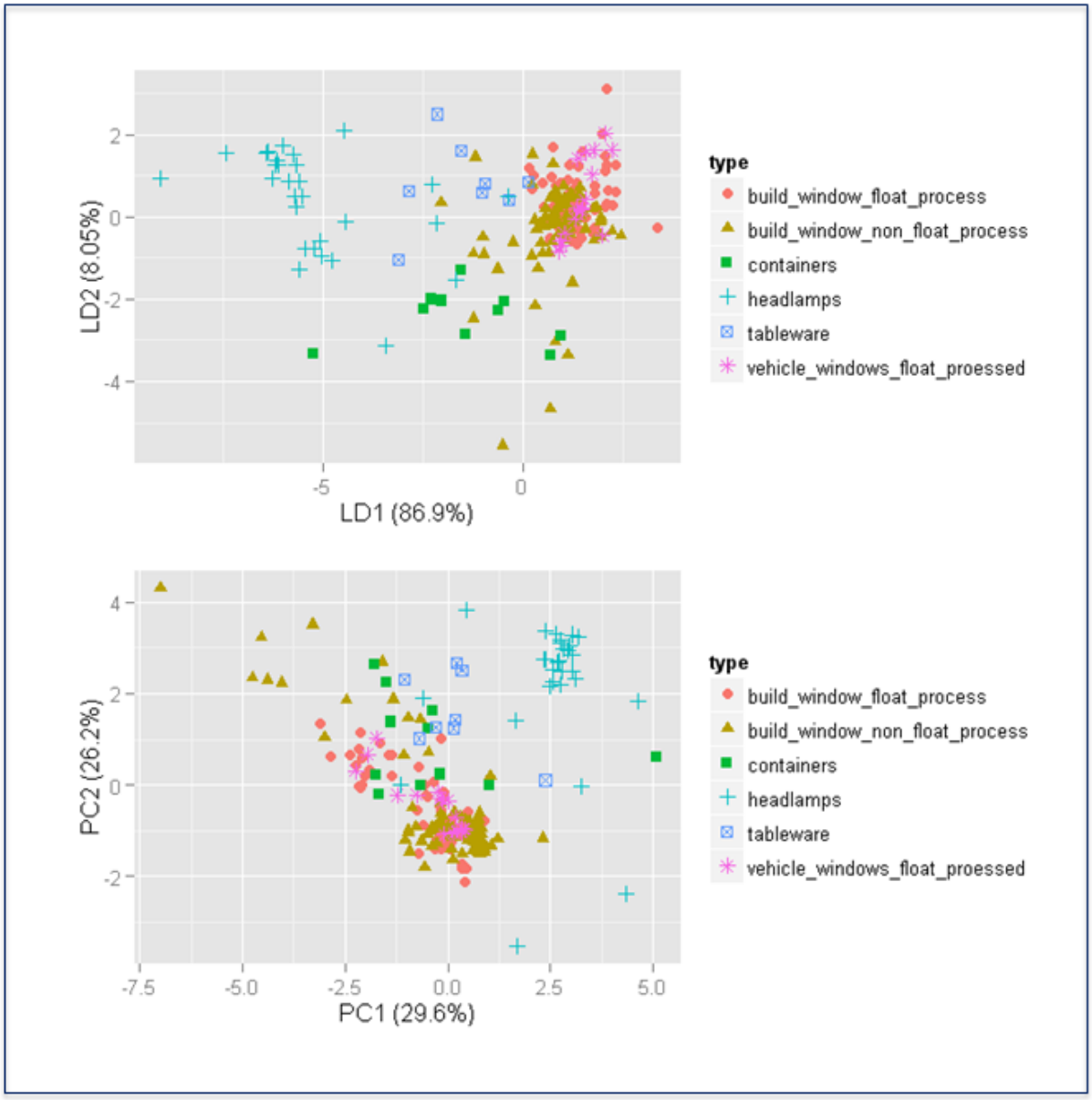

Figure 4. The top panel showing the 2D projection of glass identification data for Fisher's LDA.The bottom panel showing the 2D projection of glass identification data for PCA.

Classification Methods

We investigate the binary classification problem (window vs. non- window) by randomly splitting glass identification dataset into training set (70%) and test set (30%). We then build a classifier on training dataset and evaluate each of classifiers using an independent test set. Besides we compare the performance of several models that are built on training glass dataset with reduced dimensions. The models are LR, LDA, QDA, MDA and CT. As PCA and a number of classification methods, for example, LDA, QDA, MDA are sensitive to outliers, we estimate Mahalanobis distance. Cases (#173, #172, #107, #185, #150) with Mahalanobis distance

greater than $\chi^2(9) = 27.88$ are identified as outliers and are eliminated prior to analysis.

Logistic Regression Model

We define the training dataset as $\chi = \{(x_i, y_i) : i = 1, \ldots, n)\}$, where $x_i \in R^p$ are input variables and $y_i \in \{1:$ window, 0: non-window$\}$ denote the class of the ith observation. The logistic regression model has the following form [Hastie et al., 2011]:

$$\log\left\{\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right\} = \alpha + \beta^T x \tag{5}$$

where $\alpha \in \Re$ and $\beta \in \Re^p$ are unknown parameters. The classification boundary is given by $\{x: \alpha + \beta^T x = 0\}$. The log likelihood is [Hastie et al., 2011]:

$$l(\alpha,\beta) = \sum_{i=1}^{n} \log\left\{[P(Y=1|X=x_i)]^{y_i} \quad [1 - P(Y=1|X=x_i)]^{1-y_i}\right\}$$

$$l(\alpha,\beta) = \sum_{i=1}^{n} \log\left\{ y_i(\alpha + \beta^T x_i) - \log(1 + \exp(\alpha + \beta^T x_i)) \right\}$$

The parameters α and β are estimated by maximum likelihood.

Linear Discriminant Analysis

The LDA model assumes that the input variables given each class are normally distributed and the classes have equal covariance as $X|Y \sim N(\mu_j,\Sigma_j)$, and $\Sigma_j = \Sigma$, $j = 1$ (window), $j = 0$ (non-window) and $X = [x_1, x_2, \ldots, x_n]^T$. The conditional density of X given class $Y = j$ can be written as [James et al., 2013]:

$$P(X = x|Y = j) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_j)^T \Sigma^{-1}(x-\mu_j)\right\} \tag{6}$$

Based on Bayes' rule, the posterior probability of membership in the predicted class is estimated by reversing the conditional probabilities as:

$$P(Y = j \mid X = x) = \frac{P(X=x|Y=j)\,P(Y=j)}{P(X=x)}$$

where P(X=x) is the marginal density of X. In comparing two classes, we plug the normal density into the log posterior odds [Hastie et al., 2011] and we have,

$$\log\left\{\frac{P(Y=1\mid X=x)}{P(Y=0\mid X=x)}\right\} = \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + (\mu_1^T\Sigma^{-1} - \mu_0^T\Sigma^{-1})x = \gamma_0 + \gamma^T x \tag{7}$$

$$\text{where} \quad \gamma_0 = \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 \quad \text{and} \quad \gamma = \Sigma^{-1}(\mu_1 - \mu_0)$$

π_1 and π_0 are prior probabilities of being in window class and in non-window class, respectively. Decision boundary is $\gamma_0+\gamma^T x = 0$. If $\gamma_0+\gamma^T x > 0$, LDA rule classifies input set to window and to non-window otherwise. Parameters are estimated by the maximum-likelihood estimation (MLE). The likelihood function is as follows [Hastie et al., 2011]:

$$L(\mu_1, \mu_0, \Sigma, \pi_1, \pi_0) = \prod_{i=1}^{n} p(x_i, y_i) = \prod_{i=1}^{n} \left\{ \left[\frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1) \right\} \pi_1 \right]^{y_i} \left\{ \left[\frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x_i - \mu_0)^T \Sigma^{-1} (x_i - \mu_0) \right\} \pi_0 \right]^{1-y_i} \right\}$$

The parameters π_1, π_0, μ_1, μ_0 and Σ are estimated as follows [Pohar et al., 2004]:

$$\hat{\pi}_1 = \frac{n_1}{n}, \quad \hat{\pi}_0 = \frac{n_0}{n}, \quad \hat{\mu}_1 = \frac{1}{n_1} \sum_{y_i = 1} x_i, \quad \hat{\mu}_0 = \frac{1}{n_0} \sum_{y_i = 0} x_i,$$

$$\hat{\Sigma} = \left[\sum_{y_i = 1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^T + \sum_{y_i = 0} (x_i - \hat{\mu}_0)(x_i - \hat{\mu}_0)^T \right] / n$$

where n_1 and n_0 are the number of observations in the window class and non-window class, respectively.

Quadratic Discriminant Analysis

The QDA model assumes that the input variables are normally distributed within each class and each class has different covariance matrix as $X|C_j \sim N(\mu_j, \Sigma_j)$, $j = 1$ (window), $j = 0$ (non-window). The class-conditional density of X in class $Y = j$ is [Hastie et al., 2011]:

$$P(X = x \mid Y = j) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma_j|^{1/2}} \exp\left[-\frac{1}{2} (x - \mu_j)^T \Sigma^{-1} (x - \mu_j) \right] \qquad (8)$$

where Σ_j is the covariance matrix for class *j*.

Substituting the normal density into the log posterior odds of window glass versus non-window glass we have [Hastie et al., 2011],

$$\log\left\{ \frac{P(Y=1|X=x)}{P(Y=0|X=x)} \right\} = \log\left(\frac{\pi_1}{\pi_0} \right) - \frac{1}{2} \log|\Sigma_1| + \frac{1}{2} \log|\Sigma_0| - \frac{1}{2} \mu_1^T \Sigma_1^{-1} \mu_1 + \frac{1}{2} \mu_0^T \Sigma_0^{-1} \mu_0 + (\mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1}) X - \frac{1}{2} X^T (\Sigma_1^{-1} - \Sigma_0^{-1}) X \qquad (9)$$

Note that the log odds of window versus non-window is quadratic function of x which is set to 0 in order to obtain the decision boundary.

Mixture Discriminant Analysis

A single Gaussian to model a class, as in LDA and QDA, may not be sufficient to represent the data. With this in mind, it is extended to a mixture of Gaussians. For class k, the within-class density [Hastie et al., 2011] is

$$P(X|Y = k) = \sum_{r=1}^{R_k} \pi_{kr} \Phi(X|\mu_{kr}, \Sigma) \qquad (10)$$

and where the rth mixture density has prior probability of π_{kr} and Σ is equal across all classes and subclasses. The joint density is:

$$P(X=x, Y=k) = d_k \sum_{r=1}^{R_k} \pi_{kr} \Phi(X|\mu_{kr}, \Sigma)$$

where d_k is the prior probability of class k. The MLE of d_k is the proportion of training samples in class k. π_{kr}, μ_{kr}, and Σ are estimated using the EM algorithm.

Classification Trees

It is a nonparametric method which uses recursive binary partitioning to create a binary tree. The CART algorithm is as follows [Breiman et al., 1984].

1) Initialize the tree containing the training data
2) Obtain a set of binary splits based on one variable
3) Select the best split at a node by estimating impurity functions, the Gini index or entropy
4) Obtain the right-sized tree using Independent test set, or 10-fold cross-validation, or 1-SE rule
5) Assign every terminal node to a class

Given node t, the Gini index is defined as [Breiman et al., 1984]

$$i(t) = \Sigma_{i \neq j}\, p(i|t)\, p(j|t) = 1 - \Sigma_j p^2(j|t)$$

where p(j|t) is the probability that cases belongs to the jth class given that node is t.

Results and Discussion

Logistic Regression Model

We now fit a logistic regression model using the training data set with two input variables, X_1 and X_2 obtained from the principal components analysis. Table 2 shows the coefficient estimates for a logistic regression model. There are statistically significant effects of composition and composition dependent refractive index on the log odds of window glass. A 1- unit increase in composition is associated with a decrease in the log odds of window glass by 2.9726 units (P = 4.47e-07) holding other variable fixed. Similarly, a 1- unit increase in composition dependent refractive index will lead to a decrease of 0.9729 in the log odds of window glass (P = 0.000213) holding composition at a fixed value. So the posterior probabilities are:

$$P(Y = 1|X = x) = \frac{\exp(1.7267 - 2.9726\ X1 - 0.9729\ X2}{1+\exp(1.7267 - 2.9726\ X1 - 0.9729\ X2)}$$

$$P(Y = 2|X = x) = \frac{1}{1+\exp(1.7267 - 2.9726\ X1 - 0.9729\ X2)}$$

The decision boundary is given as $1.7267 - 2.9726\ X_1 - 0.9729X_2 = 0$

Classification rule:

$$Y = \begin{cases} 1 & \text{If } 1.7267 - 2.9726X_1 - 0.9729X_2 > 0 \\ 0 & \text{If } 1.7267 - 2.9726X_1 - 0.9729X_2 \leq 0 \end{cases}$$

Table 2

The estimated coefficients for the LR model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.7267	0.3418	5.051	4.39e-07 ***
X1	-2.9726	0.5889	-5.048	4.47e-07 ***
X2	-0.9729	0.2627	-3.703	0.000213 ***

The decision boundary represented by a solid line obtained by logistic model fitted to the training data is shown in Fig. 5(a). It appears that the window and non window glasses are not distinctly separated. In this model, the training error rate is 7.89 and the estimated PCC is 92.11%. The sensitivity, percentage of window glass that are correctly classified, is 97.5% and the specificity, percentage of non-window glass that are

correctly classified, is 71.88%. Training error rate, sensitivity, and specificity, PCC are estimated from the confusion matrix for training dataset. The confusion matrices for training and test dataset are given in Table 3 and Table 4, respectively.

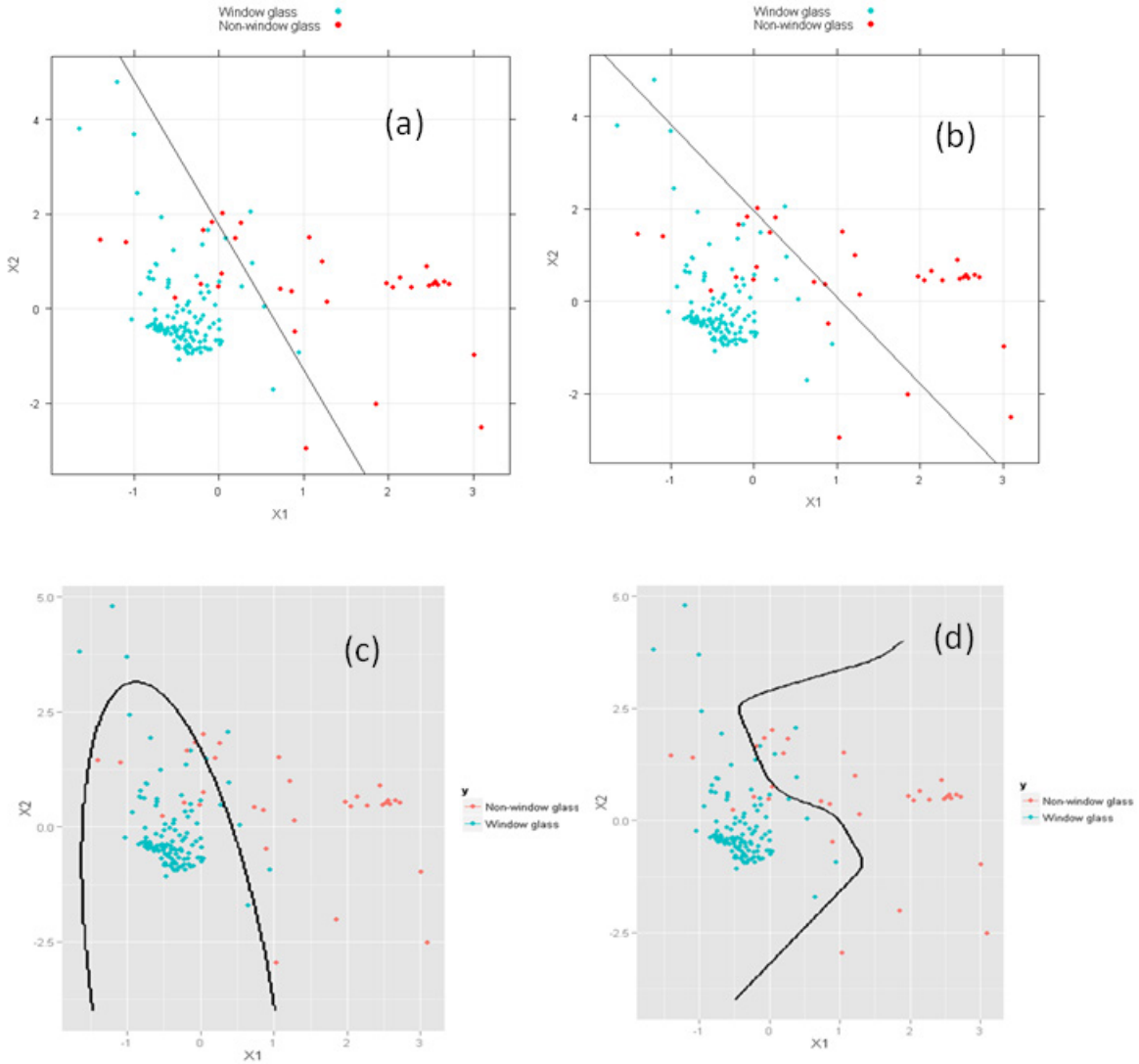

Figure 5. The decision boundaries for the training data obtained by the following classification methods: (a) Logistic Regression. (b) Linear Discriminant Analysis. (c) Quadratic Discriminant Analysis. (d) Mixture Discriminant Analysis. The window and non-window glass are displayed as green and red, respectively.

Table 3

The confusion matrix for training dataset

	Window glass	Non-window glass
Window glass	117	9
Non-window glass	3	23

Table 4
The confusion matrix for test dataset

	Window glass	Non-window glass
Window glass	40	2
Non-window glass	1	14

Linear Discriminant Analysis

We fit LDA to the training data with two input variables, X_1 and X_2. The estimated prior probabilities are $\pi_1 = 0.7895$, $\pi_0 = 0.2105$. The class-specific means and covariances are estimated as:

$$\text{For window glass:} \quad \hat{\mu}_1 = \begin{pmatrix} -0.4076 \\ -0.0999 \end{pmatrix} \quad \text{and} \quad \hat{\Sigma}_1 = \begin{pmatrix} 0.1280 & -0.1157 \\ -0.1157 & 0.9338 \end{pmatrix}$$

$$\text{For non-window glass:} \quad \hat{\mu}_0 = \begin{pmatrix} 1.2042 \\ 0.4447 \end{pmatrix} \quad \text{and} \quad \hat{\Sigma}_0 = \begin{pmatrix} 1.6380 & -0.6456 \\ -0.6456 & 1.3542 \end{pmatrix}$$

The estimated common covariance is:

$$\hat{\Sigma} = \begin{pmatrix} 0.8830 & -0.3807 \\ -0.3807 & 1.1440 \end{pmatrix}$$

and the estimated parameters of the model are:

$$\hat{\gamma}_0 = 2.4841 \quad \text{and} \quad \hat{\gamma} = \begin{pmatrix} -2.3707 \\ -1.2648 \end{pmatrix}$$

The decision boundary can be written as $2.4841 - 2.3707X_1 - 1.2648X_2 = 0$. The LDA classifier predicts window glass if $2.4746 - 2.3707X_1 - 1.2648X_2 > 0$ and non-window glass otherwise. The decision boundary is shown in the scatter plot (Fig. 5 (b)), suggesting that there is some overlap between the window and non-window glasses. In this case, the training error rate is 8.55% with PCC of 91.45%, sensitivity of 99.17% and specificity of 62.5%.

Quadratic Discriminant Analysis

We fit the QDA model to the training data. Decision boundary which is a quadratic function of input variables is $4.024X_1^2 + 0.7317X_1X_2 + 0.1483X_2^2 + 4.7604X_1 + 1.4008X_2 - 2.7832 = 0$. The QDA classifier predicts the window glass if $4.024X_1^2 + 0.7317X_1X_2 + 0.1483X_2^2 + 4.7604X_1 + 1.4008X_2 - 2.7832 > 0$ and non-window glass otherwise. The classification boundary (see Fig. 5 (c)) is shown as a solid line fitted to the training data. It is observed that the decision boundary is quadratic. Fig. 5 (c) shows window and non window glasses are better separated as compared to those in LR and LDA. The training error rate is 10.53% with PCC of 89.47%, sensitivity of 93.33% and specificity of 75% (see Table 5).

Mixture Discriminant Analysis

The classification boundary using MDA model to the training data is shown as a solid line in Fig. 5(d). Here the decision boundary is non linear. It shows also a more distinct separation between window and non window glasses. The training error rate is 7.24% with PCC of 92.76%, sensitivity of 96.67% and specificity of

78.13% (see Table 5).

Classification Trees

We fit CART model to the training data with a two-class (window glass and non-window glass) outcome variable and two input variables, X1 and X2. The tree diagram is shown in Fig. 6 (a). The tree has 9 terminal nodes. In order to verify that this is the sub tree based on the lowest cross-validation error rate, we plot the error rate as a function of size. Fig. 6(b) shows that the tree with 9 terminal nodes has the lowest cross-validation error rate. An equivalent representation of this tree is shown in Fig. 7. In this procedure the input set is recursively partitioned into rectangles [Hastie et al., 2011, James et al., 2013]. In this case, the decision boundary is non-linear, and the training error rate is 5.92% with PCC of 94.08%, sensitivity of 95.83% and specificity of 87.5% (see Table 5). CT has the highest PCC compared to other classifiers.

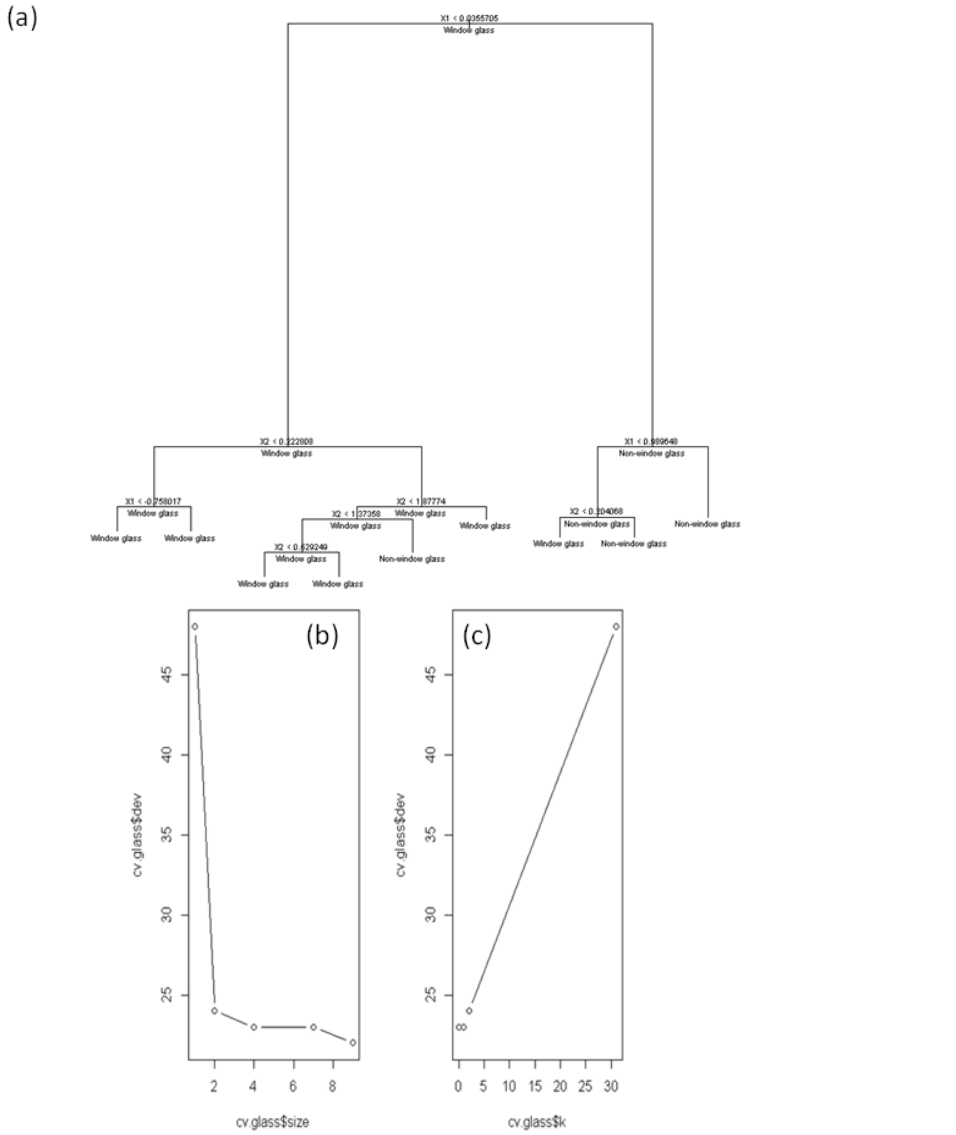

Figure 6. (a) The tree diagram showing nine terminal nodes. (b) The Cross-validation as a function of the number of terminal nodes. (c) Cross-validation as a function of k (fold).

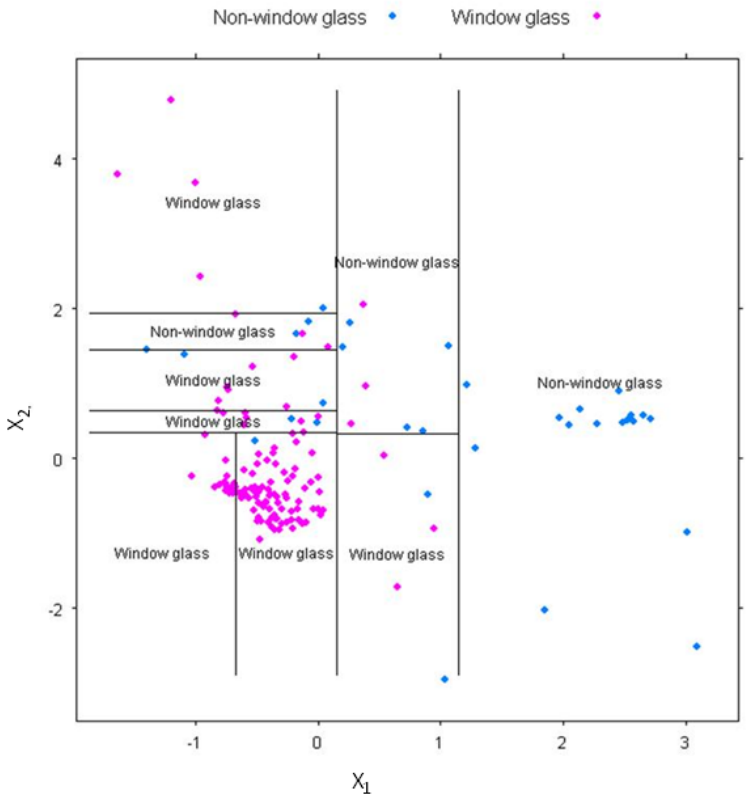

Figure 7. Decision tree showing the non-linear decision boundary. Here the tree procedure recursively partitions the input set into rectangles.

Table 5

PCC, Sensitivity and specificity on training dataset

Training			
Technique	PCC	Sensitivity	Specificity
LR	92.11%	97.50%	71.88%
LDA	91.45%	99.17%	62.50%
QDA	89.47%	93.33%	75.00%
MDA	92.76%	96.67%	78.13%
CT	94.08%	95.83%	87.50%

Classifier Performance For Two Classes

We estimate the receiver operating characteristic (*ROC*) curve on the test dataset to measure model performance. In addition, we consider confidence interval approach and hypothesis testing paradigm. ROC curves (see Fig. 8) are plots of the true positive rate against the false positive rate [Fawcett, 2006]. The true positive (TP) and false positive (FP) rates of a classifier are defined as follows.

TP ≈ Proportion of window glass which are correctly classified

FP ≈ Proportion of non-window glass which are incorrectly classified

Here we calculate the area under the ROC curve (AUC) for classifiers LR, LDA, QDA, MDA, and CT. The AUC is a portion of the area of the unit square. Fig. 8 shows the areas under five ROC curves, LR, LDA, QDA, MDA, and CT. Classifier QDA has greater area compared to other classifiers suggesting that classifier QDA appears to be superior to all.

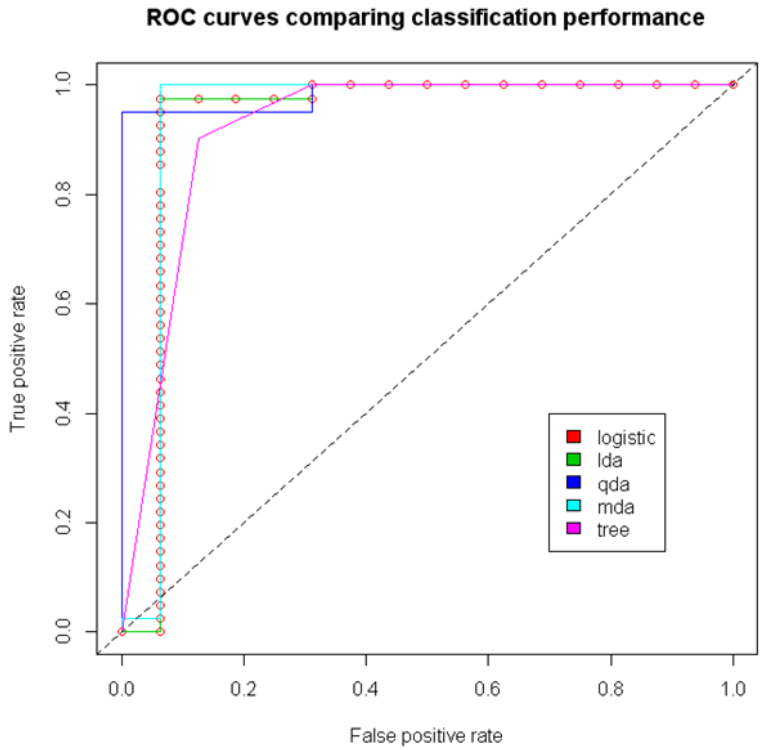

Figure 8. Roc curves for the LR, LDA, QDA, MDA and CT classifiers on the test data.

Note that the training and test samples have 152 and 57 observations, respectively. Both samples are sufficiently large (n>30) and the error rate corresponds to the sample mean. By the central limit theorem, the sampling distribution of the training and test error rates are approximately normal. The estimated classification error rate for training and test data, and their 95% confidence interval are summarized in table 6.

Table 6

Confidence intervals of error rates for glass identification data

Technique	Estimated Error Rates	
	Training	Test
LR	0.079 [95% CI: (0.036, 0.1219)]	0.0526[95% CI : (-0.0054, 0.1106]
LDA	0.0855[95% CI: (0.041, 0.13)]	0.0877[95% CI : (0.0143, 0.1611)]
QDA	0.1053[95% CI : (0.0565, 0.1541)]	0.0702[95% CI: (0.00387, 0.1365)]
MDA	0.0724[95% CI : (0.0312, 0.1136)]	0.035[95% CI: (-0.0127, 0.0827)]
CT	0.0592[95% CI : (0.0217, 0.0967)]	0.0175[95% CI: (-0.0165, 0.0515)]

The 95% confidence limit for the test set error rate for the method LDA is (0.0143, 0.1611) yielding a lower-bound error of 1.43% and an upper-bound error of 16.11 %. Similarly, the 95% confidence limit for the test set error rate for the method QDA is (0.00387, 0.1365) resulting in a lower-bound error of 0.387% and an

upper-bound error of 13.65%. Although the CT has the lowest estimated within test data classification error rate, analyses reveal that among all the machine learning models, LDA and QDA have only statistically significant test set error rates.

We also evaluate the difference between the test error rates of the two classification models, LDA and QDA using the following statistic [Roiger & Geatz, 2002]:

$$\frac{|e1 - e2|}{\sqrt{\{e(1-e)\left(\frac{2}{n}\right)\}}}$$

where e_1 is the error rate for the model LDA, e_2 is the error rate for the model QDA and $e = (e_1 + e_2)/2$ and n is the size of the test set. The estimated statistic is 0.3465, suggesting that the test set error rates between LDA and QDA built with the same training data (p-value = 0.3645) do not have statistically significant differences. We now check the normality assumption and the homogeneity of covariance. A bivariate scatterplot (Fig. 9) with two input variables, X_1 and X_2, is generated to evaluate the normality assumptions. The plot appears to be approximately elliptical indicating that data are from the multivariate normal. The Box's M test (Table 7), however, indicates a significant difference in the covariance matrices among classes (p-value < 0.0001), suggesting that the homogeneity of covariance cannot be assumed for LDA. Therefore, the analyses reveal that QDA is superior to all other models, which is consistent with the results obtained from the ROC curves.

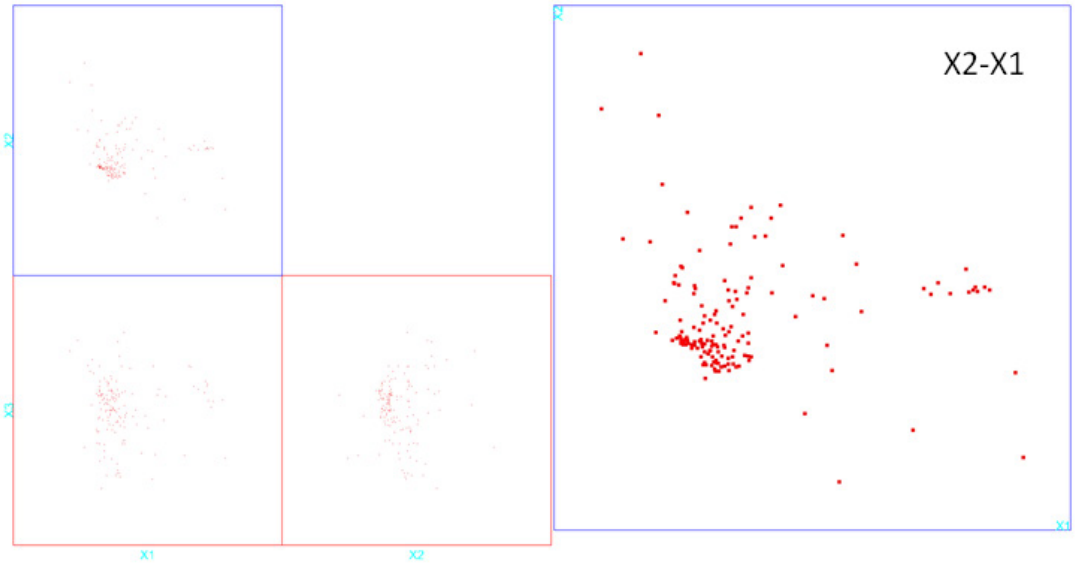

Figure 9. Scatter plot of X2 vs. X1 indicating data might be from multivariate normal.

Table 7

Tests on Null Hypothesis

Box's M	874.006
F Approx.	18.007
df1	45
df2	2.592E4
Sig.	0.000

Tests null hypothesis of equal population covariance matrices

Multiple Categories

Classification Trees

For multiple categories, two classification methods, CT and FDA, have been used to fit the data. In CT method, nodes are classified as bwfp, bwnfp, vwfp, vwnfp, cont, tabw, and headl. "bwfp" means building windows float processed. Building windows nonfloat processed and vehicle windows float processed have been abbreviated as "bwnfp" and "vwfp", respectively. Also, "vwnfp", "cont", "tabw" and "headl" have been used to describe vehicle windows nonfloat processed, containers, tableware and headlamps, respectively. We fit the training data with Gini index as shown in Fig. 10(a). We then examine the cost-complexity penalty (CP) table (see Table 8) for evaluating the overall model fit. The cost-complexity penalty measure is defined as $R_\alpha(T) = R(T) + \alpha|T|$ where R(T) is the resubstitution estimate of the misclassification rate of a tree, T, |T| is the number of terminal nodes of the tree and α (≥ 0) is the complexity parameter. For a given α, $R_\alpha(T)$ is minimized in order to find a subtree $T(\alpha)$ [Breiman et al., 1984]. Resubstitution error is the error rate obtained from training data. The cross-validation error is plotted as a function of cp parameter (Fig. 10(b)). The cross-validation error (Xerror) in the CP table is estimated using 10-fold cross-validation. To pick the right sized tree, the 1-SE rule is used. In Table 8, the observed minimum cross-validated error is 0.53608 with its estimated standard error, 0.060299. The maximum error is 0.596379. The CP (complexity parameter) value for the tree with maximum error is 0.041237. This occurs at tree size with terminal nodes 4 and splits 3. Based on one standard error rule of cross-validation, the optimal tree has 4 terminal nodes (3 splits). Fig. 10(b) also reveals that a tree with four nodes is the best model. We prune the tree using the optimal value of CP. The pruned tree is shown in Fig. 10(c). Terminal nodes in the pruned tree for Gini index are labeled bwfp, bwnfp, and headl. The estimated error rates for the models with Gini index is given in Table 9. The within training error rate is 27.63% and its 95% CI: (0.2052, 0.3474).

Fisher's LDA

This technique is used for visualizing high-dimensional data with multiple classes. In Fisher's LDA, no assumptions are made about distribution of data. Rao (1948) modified FDA (1936) by introducing assumptions of normality and common covariance matrix [Cook and Swayne, 2007]. FDA results in optimal separation between two classes under this assumption. Here data are projected onto a low-dimensional subspace to maximize the ratio of between-class to within-class variance. Based on Bayes' theorem, the posterior probability of membership in the predicted class given the discriminant function score can be estimated by reversing the conditional probabilities [Hastie et al., 2011]. Fig. 11 shows the plots of six class centroids in the two-dimensional subspace spanned by the first two discriminant functions for the forensic glass training data. Here the decision boundaries are higher dimensional affine planes as opposed to line. Also, note that black dots are the centroids of classes. Three classes (red, light cyan and brown) are close together and far from other three classes on the first discriminant function (X axis). Five classes are far from violet class on the second discriminant function (Y axis). Table 10 summarizes the within training error rates and the within test error rates estimates.

Classifier Performance for Multiple Classes

From the Table 10, it is observed that CT has the lowest within test error rate. We also find that there appears an insufficient evidence to indicate a significant difference in test set error rates of CT and FDA built with the same training data (p-value=0.1508). Both models appear to perform well. The Box's M test indicates

that there is a significant difference in the covariance matrices among classes (p-value < 0.0001) and therefore, the homogeneity of covariance cannot be assumed for FDA. In conclusion, the classification tree algorithm is superior to the Fisher's LDA.

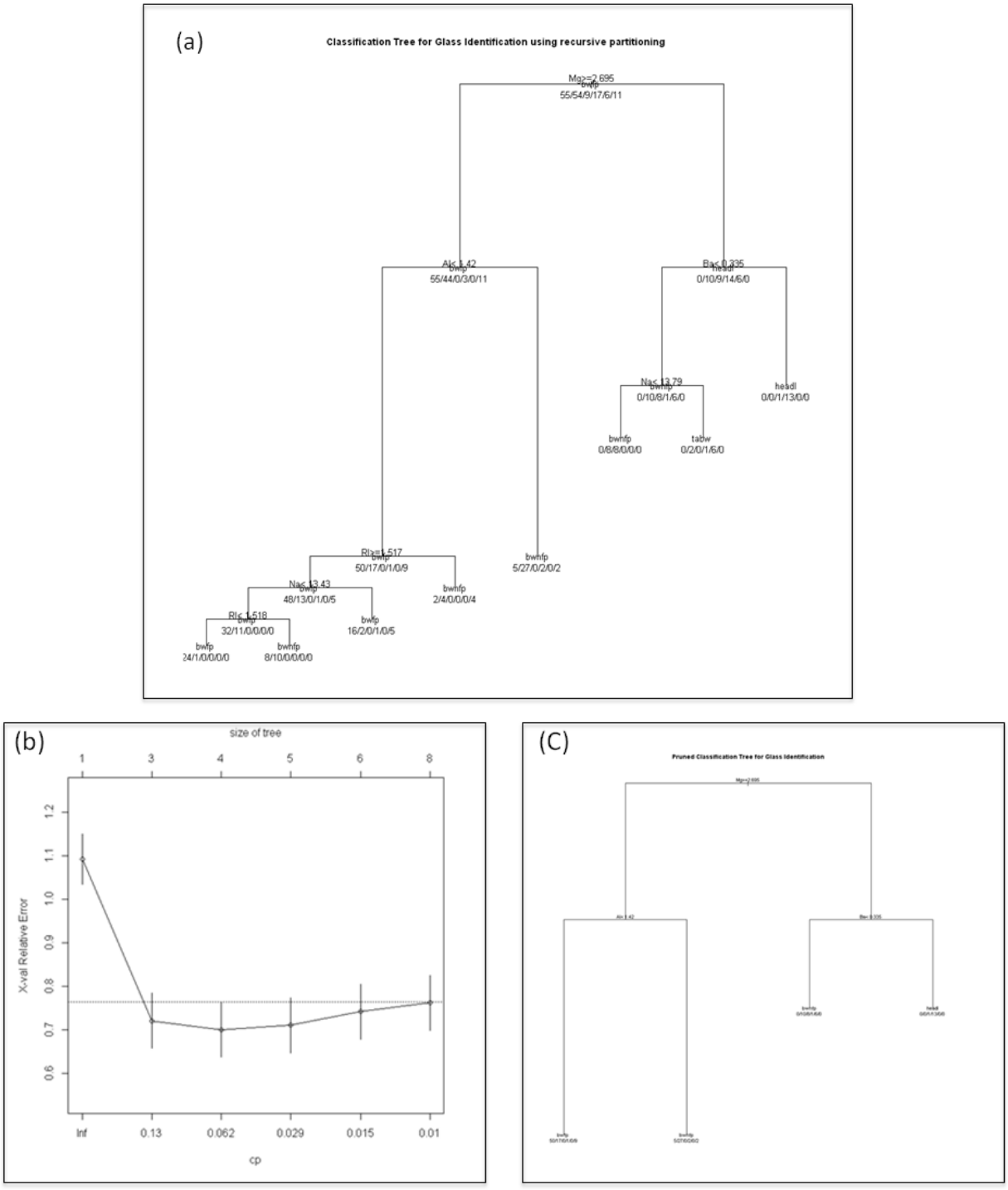

Figure 10. a) The unpruned tree obtained using GINI index. b) Cross-validation error as a function of terminal nodes. c) The pruned tree corresponding to the 1-SE rule.

Table 8

CP table with Gini Index

	CP	nsplit	Rel error	xerror	xstd
1	0.206186	0	1	1.15464	0.055969
2	0.082474	2	0.58763	0.59794	0.061743
3	0.041237	3	0.50515	0.53608	0.060299
4	0.020619	4	0.46392	0.54639	0.060571
5	0.010309	5	0.4433	0.5567	0.06083
6	0.01	6	0.43299	0.54639	0.060571

Table 9

Estimated error rates using Gini index

Technique	Estimated Error Rates	
	Training	Test
Gini Index	0.2456[95% CI: (0.1339, 0.3574)]	0.2763 [95% CI: (0.2052, 0.3474)]

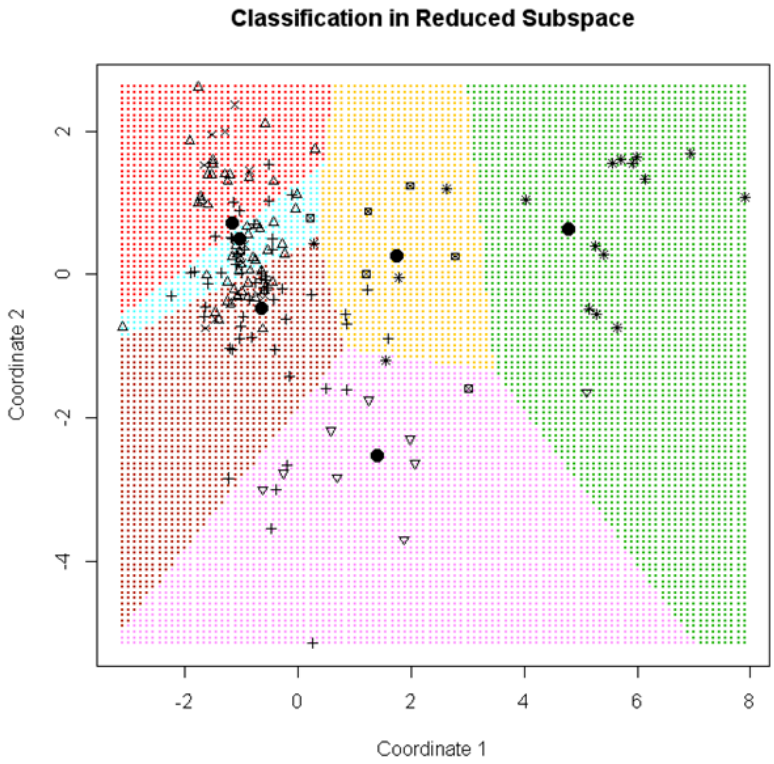

Figure 11. Fisher's LDA decision boundaries for training data in the two-dimensional subspace spanned by the first two FDA coordinates.

Table 10

The error rate estimate

Technique	Estimated Error Rates	
	Training	Test
Classification tree	0.2763 [(0.2052, 0.3474)]	0.2456 [(0.1339, 0.3574)]
Fisher's LDA	0.3684 [(0.3622, 0.3746)]	0.3333 [(0.3171, 0.3495)]

Conclusions

The relative performance of classification methods, such as LR, LDA, QDA, MDA and CT is assessed by invoking the tests of statistical significance and the receiver operating characteristic (ROC) curves. Glass identification data set is used and split into two subsets: training (70%) and test (30%). The classifiers are built on the training dataset, and the relative performance of the classifiers is obtained on the test data. The performance measures for different classification methods are the area under the receiver operating characteristic curve (AUC), the error rates and its 95% confidence interval. Among all the classifiers, the LDA and QDA perform significantly very well on the reduced dimension. However, the Box's M test ($P<0.0001$) indicates that homogeneity of covariance cannot be assumed. Despite the fact that CT has the highest estimated probability of correct classification (PCC), our analyses suggest that for glass types, window and non-window, the QDA outperforms all methods. For multiple classes, the classification tree, however, is found to perform the best when all six categories of glasses are considered.

References

Aeberhard, S., Coomans, D. and Vel, O. De (1994) The Performance of Statistical Pattern Recognition Methods in High Dimensional Settings, IEEE Signal Processing Workshop on Higher Order Statistics.

Bottrell, M. C. (2009) Forensic glass comparison: background information used in data interpretation, Forensic Science Communications 11(2) [http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/april2009/review/2009_04_review01.htm/] (accessed 15 October 2014).

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984) Classification And Regression Trees (The Wadsworth statistics/probability series) published by Chapman & Hall.

Cook, D. and Swayne, D.F. (2007) Interactive and Dynamic Graphics for Data Analysis: with R and GGobi (Use R!)

Fawcett, T. (2006) An introduction to ROC analysis, Pattern Recognition Letters 27, 861-874.

Ferri, F. J., Albert, J. V., Vidal, E. (1999) Considerations about sample-size sensitivity of a family of edited nearest-neighbor rules. IEEE Transaction on Systems, Man and Cybernetics – Part B, 29, 667-672.

Hastie, T., Tibshirani, R. and Friedman, J. (2011) The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics).

James, G., Witten , D., Hastie , T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with Applications in R (Springer Texts in Statistics).

Janecek, A. G. K ., Gansterer, W. N., Demel, M. A., and Ecker, G. F. (2008) On the relationship between feature selection and classification accuracy, JMLR: Workshop and Conference Proceedings, 4, 90-105

Jiang, Y. and Zhou, Z. (2004) Editing training data for kNN classifiers with neural network ensemble, Advances in Neural Networks-ISNN, 3173, 356-361

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Pohar M, Blas M, Turk S. (2004) Comparison of logistic regression and linear discriminant analysis: A simulation study, Metodoloski Zvezki, 1, 143-161

Roiger, R. J. and Geatz, M. W (2002) Data Mining A Tutorial-Based Primer (Addison-Wesley).

Shlens J. (2005) A Tutorial on principal component analysis [www.cs.cmu.edu/~elaw/papers/pca.pdf] (accessed 15 October 2014)

Terry, K. W., Riessen, A. van and Lynch, B. F.(1983) Identification of small glass fragments for forensic purposes, Government Chemical Laboratories, Western Australia [http://crg.aic.gov.au/reports/9.80.pdf](accessed 15 October 2014)

Wegman, E. J. (1990) Hyperdimensional data analysis using parallel coordinates, Journal of the American Statistical Association, 85, 664-675

Wegman, E. J. and Dorfman, A. (2003) Visualizing cereal world, Computational Statistics and Data Analysis, 43, 633-649