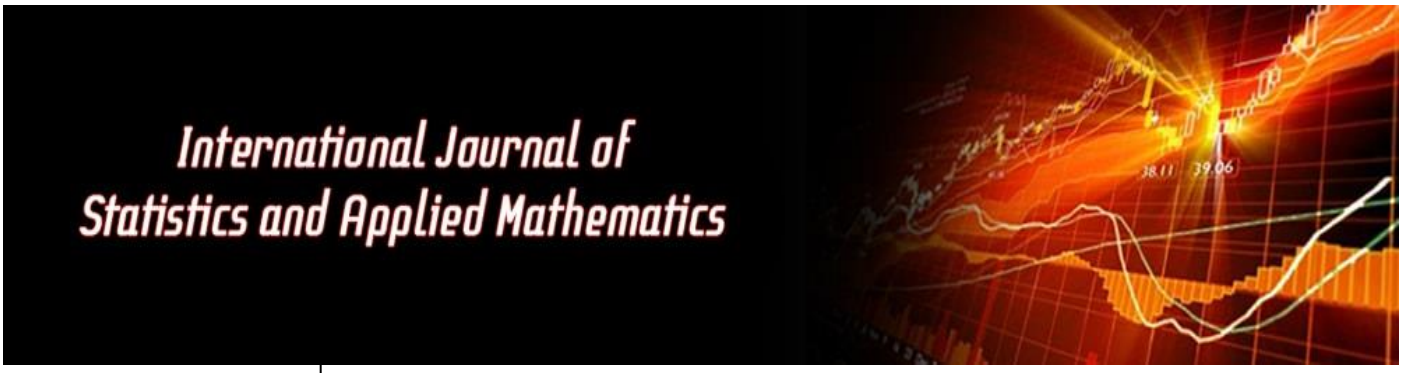
International Journal of Statistics and Applied Mathematics 2018; 3(2): 321-325



ISSN: 2456-1452
Maths 2018; 3(2): 321-325
© 2018 Stats & Maths
www.mathsjournal.com
Received: 20-01-2018
Accepted: 22-02-2018

S Padmanaban
NIRRH Field Unit, Indian Council of Medical Research, KMC Hospital, Chennai, Tamil Nadu, India

Martin L William
Department of Statistics, Loyola College, Chennai, Tamil Nadu, India

Correspondence
S Padmanaban
NIRRH Field Unit, Indian Council of Medical Research, KMC Hospital, Chennai, Tamil Nadu, India

Variable selection for identification of fatty liver cases: Nonparametric discriminant analysis

S Padmanaban and Martin L William

Abstract
A forward-model-building algorithm for discriminating two populations in a nonparametric setting has been recently introduced by Padmanaban and William (2016) wherein the restrictions existing in the traditional discriminant analysis have been removed. This latest approach is applicable without assumptions on the two underlying populations. As an application of this approach, we consider the discrimination and classification of fatty liver cases from others using some observable variables that are believed to be related to the health condition of the liver. The discriminant model is developed with a sample of 160 cases drawn from a case-control study. We shall also compare the discriminant model performance with that of logistic regression.

Keywords: classification, discriminant, Kolmogorov-Smirnov statistic

1. Introduction
Discrimination of objects of two populations and the corresponding problem of effectively classifying objects to the two has engaged the attention of scholars for the past many decades. Generally, variables are included in the discriminant based on a comparison of means in the two populations. Further, classification of an object to one of the two populations is based on the distances of the object's discriminant value from the means of the two populations. Applying the technique in a non-parametric setting requires equality of variance-covariance matrices of the two populations, though this condition is not needed for multivariate normal populations. In real-time applications where several variables are observed, joint normality or equality of variance-covariance matrices is not guaranteed. In the case of joint normality, one can test the equality of the variance-covariance matrix and if equality is accepted, can apply the classical discriminant function; if equality is rejected, quadratic discriminant function is applicable. If the data are from non-normal populations, the distribution-free linear discriminant analysis is applicable, but there is no easy procedure available for testing the equality of variance-covariance matrices. In most situations, practitioners assume equality and proceed. Thus, the above gap between theory and practice has remained for long.

The classical theory of discriminant analysis has seen much developments over the past many decades. The construction of discriminant models basically involves identifying the important variables that are material in discriminating the two populations. An early work in this regard is that of Habbema and Hermans (1977) [7] who considered variable selection in discriminant analysis by error rates. Chang (1983) addressed the question of separation of a mixture of two multivariate normal populations using principal components. Another interesting work is a paper by Bensmail and Celeux (1996) [3] wherein Gaussian discriminant analysis was addressed via eigen-value decomposition. A stepwise algorithm using 'Bayesian Information Criterion' was suggested by Murphy *et al.* (2010) [11] following Raftery and Dean (2006) [14] who proposed a similar approach for model-based clustering. These papers focus mainly on parametric settings and their scope for applications is restricted.

Discriminant analysis to non-parametric settings has been another direction of research pursued by various authors. Stepwise variable selection and evaluation with different criteria was considered by Pfeiffer (1985) [13]. Hastie *et al.* (1994) [8] considered nonparametric discriminant analysis with nonlinear classifiers to handle situations with a large number of

input variables. Baudat and Anouar (2000) [1] gave a nonlinear linear discriminant analysis via kernel approach which is theoretically close to support vector machines. Nonparametric discriminant analysis with adaptation to nearest-neighbour classification was developed by Bressan and Vitria (2003) [4]. Chiang and Pell (2004) [6] combined genetic algorithms with discriminant analysis for identifying key variables. The focus in all these works was identification of the variables that are effective in discriminating the populations.

Recently, Padmanaban and William (2016) [12] proposed a discriminant analysis procedure for a distribution-free context which does not require the condition of equality of the variance-covariance matrices. Along with the basic theoretical framework, the above-referred paper provides a model-building algorithm (called 'forward-model-building process', a term usually used in building predictive models) for selecting the important variables that make up the discriminant function. In the above-mentioned paper, a different approach has been taken for two-population discriminant analysis, while adhering to the spirit and objective of classical discriminant analysis. For a discussion on the 'model performance' measure to evaluate the classification ability of the discriminant model and the decision rule for identifying the optimal cut-off point for classification, we refer to the above paper of Padmanaban and William (2016) [12].

This paper consists of four sections apart from the current introductory section and is arranged as follows: In Section 2, we review the basic theoretical framework, optimal discriminant function, model performance measure and the variable-selection algorithm to build an efficient discriminant model, which have been introduced by the authors recently. Section 3 discusses the phenomenon of fatty liver, a problem that is widely prevalent in current times, and the possible associated factors. As a real-life application of the variable selection algorithm, the classification and discrimination of fatty liver cases from others using measurements on eight observable variables is considered in Section 4 using a sample of 160 cases drawn from a case-control study. We shall also compare the discriminant model performance with that of logistic regression.

2. A Review of the Recently Introduced Procedure

Let π_1 and $\pi_2$be two populations whose relative sizes are given by the proportions p_1 and p_2. We consider the problem of discriminating the members of the two classes using observed data on a random vector, say, $X = (X_1, X_2,..., X_p)^T$.Denoting the mean-vectors of X in the two populations as $\mu_1 = E_1(X)$ and $\mu_2 = E_2(X)$ and the variance-covariance matrices of X in the two populations as Σ_1 and Σ_2, we have the following results from Padmanaban and William (2016) [12]:

1. For a random vector X and another random object W, the relationship between the unconditional and conditional mean vectors and variance-covariance matrices is given by

$$E(X) = E_W[E_{X|W}(X)] \text{ and } V(X) = E_W\{V_{X|W}(X)\} + V_W\{E_{X|W}(X)\} \quad (2.1)$$

2. The overall variance-covariance matrix of the combined population is given by

$$\Sigma = p_1\Sigma_1 + p_2\Sigma_2 + p_1(1-p_1)\,\mu_1\,\mu_1^T + p_2(1-p_2)\,\mu_2\,\mu_2^T - p_1 p_2(\mu_1\,\mu_2^T + \mu_2\,\mu_1^T) \quad (2.2)$$

In Discriminant Analysis, the multivariate observations (X) are transformed to univariate observations (Y) by considering linear combinations of the X_i's. The 'X-based optimal discriminant' is given by

$$Y = (\mu_1 - \mu_2)^T \Sigma^{-1} X \quad (2.3)$$

Suppose $X_{(s)}$ be a subset of the variables used to build the optimal discriminant. Denote the mean vectors of $X_{(s)}$ in the two populations as $\mu_{1(s)}$ and $\mu_{2(s)}$ and the 'overall' variance-covariance matrix of $X_{(s)}$ as $\Sigma_{(s)}$. The $X_{(s)}$-based optimal discriminant is

$$Y_{(s)} = (\mu_{1(s)} - \mu_{2(s)})^T \Sigma_{(s)}^{-1} X_{(s)} \quad (2.4)$$

Typically, the parameters are replaced by the sample estimates in practice. The performance of the $X_{(s)}$-based optimal discriminant is measured by the two sample Kolmogorov-Smirnov Statistic based on the $Y_{(s)}$ measurements given by

$$KS_{(s)} = \max_y \left(F_{1(s)}(y) - F_{2(s)}(y) \right) \quad (2.5)$$

where $F_{1(s)}(\cdot)$ and $F_{2(s)}(\cdot)$ are the (empirical) survival functions of $Y_{(s)}$ for the two populations.

Given two subvectors $X_{(s1)}$ and $X_{(s2)}$, the optimal $X_{(s1)}$-based discriminant is said to be 'more efficient' than the optimal $X_{(s2)}$-based discriminant if $KS_{(s1)} > KS_{(s2)}$. If there exists a random subvector $X_{(s^*)}$ for which $KS_{(s^*)} > KS_{(s)}$ for every other random subvector $X_{(s)}$, then the corresponding optimal discriminant $Y_{(s^*)}$ is the 'most efficient' discriminant.

The practical issue is finding the 'most efficient' discriminant as it is computationally prohibitive when there are a large number of predictor variables. This is true of every model-building situation involving a large number of predictor variables and therefore, different algorithms are suggested to 'build' improved models sequentially instead of finding the 'best' from 'all possible' models.

Padmanaban and William (2016) [12] developed a 'forward model-building' algorithm to build a 'sequence' of models, starting with a single variable and 'select' variables one-by-one based on their ability to add to the discriminatory ability of the model. The sequence of steps involved in the algorithm is briefly presented below:

Let $X_1, X_2, ., X_p$ be the candidate input variables.

Step1: The scores on 'p' discriminants $Y_{(1)}, Y_{(2)},. . .,Y_{(p)}$, where $Y_{(i)}$ is the discriminant based on single input variable X_i, are obtained for each individual record in the combined data. Let the Kolmogorov-Smirnov Statistic for $Y_{(i)}$ be denoted as $KS_{(i)}$. If

$$KS_{(i)} > KS_{(j)} \text{ for every } j \neq i$$

Then X_i is the top discriminator between the two populations. The significance of this $KS_{(i)}$ statistic is evaluated and if found significant at a desired level, X_i first 'enters' the model and model building continues.

Step 2: With X_i having been already selected, take one additional variable at a time and obtain (p–1) discriminants having input-pairs $(X_1,X_i),...., (X_{i-1},X_i), (X_{i+1},X_i),...., (X_p,X_i)$. Denote the discriminants as $Y_{(1,i)}, Y_{(2.i)},..., Y_{(i-1,i)}, Y_{(i+1,i)}, ... Y_{(p,i)}$ and the corresponding Kolmogorov-Smirnov statistics as $KS_{(1,i)}, KS_{(2,i)},...,KS_{(i-1,i)}, KS_{(i+1,i)},..., KS_{(p,i)}$. If for some 'm',

$$KS_{(m,i)} > KS_{(j,i)} \text{ for every } j \neq m, \text{ and } KS_{(m,i)} > KS_{(i)}$$

Then X_m enters the model as the second variable. It is to be noted that the significance of $KS_{(m,i)}$ is guaranteed because of the significance of $KS_{(i)}$ in the first step. In contrast, if

$$KS_{(m,i)} > KS_{(j,i)} \text{ for every } j \neq m, \text{ but } KS_{(m,i)} \leq KS_{(i)},$$

Then X_m does not enter the model, nor any of the remaining X_j's enter, as its entry leads to reduced discriminatory ability and the model building stops with only one input variable. Clearly no other variable can enter.

At every subsequent step that is considered, one more additional variable enters provided the maximum KS value at that step exceeds the maximum KS value of the previous step. If it is equal to or less than the previous maximum, the process stops. When the process stops at the $(k+1)^{th}$ step, the optimal discriminant functionis the one obtained in the k^{th} step with the maximum KS value, leading to significant and maximum discrimination between the two populations. We denote the final subset of variables reached in this process as $X_{(S*)}$ and the 'final' efficient discriminant as $Y_{(s*)}$.

Classification or Prediction Rule

The classification or prediction rule to allocate an object to one of the two populations is based on the optimal cut point at which the KS statistic value is attained. Let y_0 be the point such that

$$KS_{(s*)} = \max_{y} \left(F_{1(s*)}(y) - F_{2(s*)}(y) \right) = F_{1(s*)}(y_0) - F_{2(s*)}(y_0)$$

This point $y_0$gives maximum differentiation between the distributions of the $Y_{(s*)}$ scores in the two populations and is the 'efficient cut-point'.

Now, let the means of the final efficient discriminant $Y_{(s*)}$ in the two populations π_1 and π_2 be denoted as $\mu_{1Y(s*)}$ and $\mu_{2Y(s*)}$ and, let $\mu_{1Y(s*)} > \mu_{2Y(s*)}$. For membership-prediction, we proceed as follows:

If $y_{(s*)}$ is the value of the final efficient discriminant $Y_{(s*)}$ for an object, then the following classification rule is to be applied:

≥

$$\text{Classify object to: } \begin{cases} \pi_1 \text{ if } \mathcal{Y}_{(s*)} \geq \mathcal{Y}_0 \\ \pi_2 \text{ if } \mathcal{Y}_{(s*)} < \mathcal{Y}_0 \end{cases}$$

3. The Phenomenon of Non-Alcoholic Fatty-Liver Disease

Overview

Non-alcoholic fatty liver disease (NAFLD) is a very common disorder and refers to a condition where there is accumulation of excess fat in the liver of people who drink little or no alcohol. The most common form of NAFLD is a non-serious condition called fatty liver. In fatty liver, fat accumulates in the liver cells. Although having fat in the liver is not normal, by itself it probably does not damage the liver.

A small group of people with NAFLD may have a more serious condition named non-alcoholic steatohepatitis (NASH). A study on the epidemiology of NAFLD and NASH in the United States and the rest of the world is reported in Sayiner *et al.* (2016). In NASH, fat accumulation is associated with liver cell inflammation and different degrees of scarring. NASH is a potentially serious condition that may lead to severe liver scarring and cirrhosis. Cirrhosis occurs when the liver sustains substantial damage, and the liver cells are gradually replaced by scar tissue, which results in the inability of the liver to work properly. Some patients who develop cirrhosis may eventually require a liver transplant. For an expanded review of NAFLD, we refer to Benedict and Zhang (2017).

Incidence

Non-alcoholic fatty liver disease (NAFLD) is a distinct hepatic condition and one of the most common causes of chronic liver disease globally. Prevalence of the disease is estimated to be around 9-32% in the general Indian population, with a higher incidence rate amongst obese and diabetic patients. More details in the Indian context may be found in Kalra *et al.* (2013) [10]. A similar study on people of different races and ethnicities has been reported in Kalia and Gaglio (2016) [9].

Methodology

The patients attending gastroenterology out-patient department and confirmed by us as having fatty liver are selected for study group. Those patients diagnosed as not having fatty liver are selected as control group. A sample of 80 from each group is included for this study.

Potential Factors Associated with Fatty liver

1. AGE – A general perception is that older a person, higher is the likelihood of diseases that affect the condition of inner organs including fatty liver condition.
2. TRIGLYCERIDE (mg/dl) – Elevated TGL is found in obese people and diabetics and those who consume alcohol and is associated with heart and blood vessel disease.
3. GGT (U/L) – The gamma-glutamyl transferase test may be used to determine the level of alkaline phosphatase (ALP) which is likely to indicate liver disease.
4. HEIGHT (cm) – Although Height has no apparent relationship to condition of nay inner organ, it being a variable used to measure BMI makes it a candidate variable.
5. WEIGHT (kg) – Increase in weight is generally associated with life-style diseases and could be associated with liver diseases.
6. BMI – Body mass index is an indicator of the health of a person and could be associated with fatty liver.
7. WAIST CIRCUMFERENCE (cms) – The waistline is generally considered to be another indicator of the health of a person, including the condition of one's liver.
8. FLI – The Fatty Liver Index is based on Waist Circumference, Body Mass Index, Triglyceride and GGT and was initially developed to detect fatty liver.

Objective: This study aims to relate the above factors to fatty liver through nonparametric discriminant analysis developed by Padmanaban and William (2016) [12]. We wish to identify the significant factors that are associated with fatty liver and also give a decision rule to classify people as 'fatty liver cases' or notbased on the discriminant scores using the observable variables.

Study Design: Case control study

Sample Size: Study group (with Fatty liver): 80
Control group (without Fatty liver): 80

4. Prediction of Fatty liver through Efficient Discriminant Analysis

A sample of the data on the eight variables listed under

'Potential factors' along with the outcome (Fatty liver = 1, No Fatty liver = 2) is given below:

Table 1

Fatty liver	AGE	TGL(mg/dl)	GGT(U/L)	HEIGHT(cms)	WEIGHT(KG)	BMI	WAIST (cms)	FLI
1	63	76	29	153	90.3	38.6	123	93.6094277
2	44	100	22	154	56	23.6	90	25.2277107
1	53	175	46	149	68.3	30.8	101	82.6382847
1	42	63	21	146	77.2	36.2	102	69.5757162
2	43	201	32	144	51	24.5	93	53.2939213

We apply the forward-model-building algorithm developed in this paper and get the following results.

Step 1: The KS statistics for models with single variables are found to be
$KS_{(X1)} = 0.35$, $KS_{(X2)} = 0.2125$, $KS_{(X3)} = 0.15$, $KS_{(X4)} = 0.2625$, $KS_{(X5)} = 0.425$, $KS_{(X6)} = 0.4125$, $KS_{(X7)} = 0.4125$, $KS_{(X8)} = 0.4125$
X_5 enters the model in the first step. The KS value of 0.425 is found to be statistically significant.

Step 2: The KS statistics for models with one additional variable with X_5 are found as
$KS_{(X1,X5)} = 0.5625$, $KS_{(X2,X5)} = 0.4125$, $KS_{(X3,X5)} = 0.45$, $KS_{(X4,X5)} = 0.425$, $KS_{(X6,X5)} = 0.4375$, $KS_{(X7,X5)} = 0.4625$, $KS_{(X8,X5)} = 0.5$
X_1 enters the model in the second step.

Step 3: In this step we get
$KS_{(X1,X2,X5)} = 0.55$, $KS_{(X1,X3,X5)} = 0.6$, $KS_{(X1,X4,X5)} = 0.5625$, $KS_{(X1,X5,X6)} = 0.575$, $KS_{(X1,X5,X7)} = 0.5375$, $KS_{(X1,X5,X8)} = 0.5375$
X_3 enters the model in the third step.

Step 4: In this step we get
$KS_{(X1,X2,X3,X5)} = 0.6125$, $KS_{(X1,X3,X4,X5)} = 0.6125$, $KS_{(X1,X3,X5,X6)} = 0.6125$, $KS_{(X1,X3,X5,X7)} = 0.6125$, $KS_{(X1,X3,X5,X8)} = 0.5875$
X_2 enters the model in the fourth step. [Note that any one of the variables X_2, X_4, X_6, X_7 qualify the 'entry' criterion, but suppose thatwe select the first of these (in the order of arrangement of the variables) namely X_2]

Step 5: In this step we get
$KS_{(X1,X2,X3,X4,X5)} = 0.6125$, $KS_{(X1,X2,X3,X5,X6)} = 0.6125$, $KS_{(X1,X2,X3,X5,X7)} = 0.6$, $KS_{(X1,X2,X3,X5,X8)} = 0.575$
No variable satisfies the 'entry' criterion and therefore, the model building stops with Step 4 with four variables selected in the order of X_5, X_1, X_3, X_2.
The 'Efficient Discriminant' obtained at the end of Step 4 of our algorithm is:

$$Y = 0.0428*AGE + 0.00168*TGL + 0.0117*GGT + 0.0783*WEIGHT \quad \ldots (4.1)$$

Membership-Prediction Rule: If 'y' denotes the measured value of the 'Efficient Discriminant' Y of (4.1) for an individual, then the prediction rule is as follows:

$$\text{Classify individual to: } \begin{cases} \textit{Fatty liver Group} & \textit{if } y > 8.096 \\ \textit{NoFattyliver Group} & \textit{if } y \le 8.096 \end{cases}$$

We observe form (4.1) that, higher AGE, higher TGL, higher GGT and higher Weight indicate the likelihood of fatty liver for an individual.

We now provide an alternative selection due to the 'equal' KS values occurring in Step 4.

Step 4a: Suppose, we select X_7 to enter the model in the fourth step. Then, we have the following Step 5a.

Step 5a: In this step we get
$KS_{(X1,X2,X3,X5,X7)} = 0.6$, $KS_{(X1,X3,X4,X5,X7)} = 0.6125$, $KS_{(X1,X3,X5,X6,X7)} = 0.6125$, $KS_{(X1,X3,X5,X7,X8)} = 0.6$
No variable satisfies the 'entry' criterion and therefore, the model building stops with Step 4a with four variables selected in the order of X_5, X_1, X_3, X_7.
The 'Efficient Discriminant' obtained at the end of Step 4a is:

$$Z = 0.0347*AGE + 0.0132*GGT + 0.0530*WEIGHT + 0.0412*WC \quad \ldots (4.2)$$

Membership-Prediction Rule: If 'z' denotes the measured value of the 'Efficient Discriminant' Z of (4.2) for an individual, then the prediction rule is as follows:

$$\text{Classify individual to: } \begin{cases} \textit{Fatty liver Group} & \textit{if } z \ge 9.5913 \\ \textit{NoFattyliver Group} & \textit{if } z < 9.5913 \end{cases}$$

We observe form (4.2) that, higher AGE, higher GGT, heavier Weight and larger Waist Circumference indicate the likelihood of fatty liver for an individual.

5. Comparison with Logistic Regression Model

Denoting 'fatty liver' as the outcome of interest, we build a logistic regression model using the forward stepwise method of model building.
Step 1: Waist Circumference entered with high significance and with a positive coefficient.

Step 2: Weight entered with high significance and with a positive coefficient.

Step 3: Age entered with high significance and with a positive coefficient.

Step 4: GGT entered with high significance and with a positive coefficient.
The model building stops with step 4 and we have the following logit equation from the model:

$$\log\left(\frac{p}{1-p}\right) = -17.44 + 0.057*\text{Age} + 0.023*\text{GGT} + 0.095*\text{Weight} + 0.079*\text{WC} \quad (4.3)$$

where 'p' is the probability of fatty liver occurrence. The KS for this model is found to be 0.4625, lower than the KS obtained for the 'Efficient Discriminant Models' obtained in (4.1) and (4.2). Thus, we find that the non-parametric

discriminant method performs better than binary logistic regression method in predicting 'fatty liver presence'.

It is interesting to note that the discriminant model (4.2) is based exactly on the same variables as the binary logistic regression model (4.3), but has higher discriminatory ability as measured by the Kolmogorov-Smirnov Statistic. The model (4.1) is, however, based on a different subset of variables, which overlaps with the variables in the logistic regression model. Still, performance-wise, both discriminant models (4.1) and (4.2) are equally good. Our investigation points out that even when we allow X_4 or X_6 to enter the model (instead of X_2 or X_7), the model building stops with the fourth step and the resulting models are all equal in performance.

The better performance of discriminant model(s) in our study points out the promising nature of our nonparametric discriminant model as an alternative to logistic regression. We also infer that it could identify some important discriminating variables which logistic regression fails to bring out.

6. References

1. Baudat G, Anouar F. Generalized Discriminant Analysis using a Kernel Approach. Neural Computation. 2000; 12(10):2385-2404.
2. Benedict M, Zhang X. Non-alcoholic fatty liver disease: An expanded review. World J. Hepatol. 2017; 9:715-732.
3. Bensmail H, Celeux G. Regularized Gaussian discriminant analysis through eigen value decomposition. J. Amer. Statist. Assoc. 1996; 91:1743-1748.
4. Bressan M, Vitria J. Nonparametric Discriminant Analysis and Nearest Neighbor Classification. Pattern Recognition Letters. 2003; 24:2743-2749.
5. Chang WC. On using Principal Components before Separating a Mixture of two Multivariate Normal Distributions. J. Roy. Statist. Soc. Ser C. 1983; 32:267-275
6. Chiang LH, Pell RJ. Genetic algorithms combined with discriminant analysis for key variable identification. J. Process Control. 2004; 14:143-155.
7. Habbema JDF, Hermans J. Selection of variables in discriminant analysis by F-Statisticand error rate. Technometrics. 1977; 19:487-493.
8. Hastie T, Tibshirani R, Buja A. Flexible Discriminant Analysis by Optimal Scoring. Amer. Statist. Assoc. 1994; 89:1255-1270.
9. Kalia HS, Gaglio PJ. The Prevalence and Pathobiology of Nonalcoholic Fatty Liver Disease in Patients of Different races or Ethnicities. Clin. Liver Dis. 2016; 20:215-224.
10. Kalra S, Vithalani M, Gulati G, Kulkarni CM, Kadam Y, Pallivathukkal J, Das B, Sahay R, Modi, K.D. (2013). Study of prevalence of nonalcoholic fatty liver disease (NAFLD) in type 2 diabetes patients in India (SPRINT). J. Assoc. Physicians India. 2016; 61:448-453.
11. Murphy TB, Dean N, Raftery AE. Variable Selection and updating in Model-Based Discriminant Analysisfor High Dimensional Data with Food Authenticity Applications. The Annals of Applied Statistics. 2010; 4(1):396-421.
12. Padmanaban S, Martin William L. A nonparametric discriminant variable-selection algorithm for classification to two populations International Journal of Applied Mathematics and Statistical Sciences. 2016; 5(2):87-98.
13. Pfeiffer KP. Stepwise Variable Selection and Maximum Likelihood Estimation of Smoothing Factors of Kernel Functions for Nonparametric Discriminant Functions evaluated by Different Criteria. J. Biomed. Informatics. 1985; 18:46-61.
14. Raftery AE, Dean N. Variable Selection for Model-Based Clustering. J. Amer. Statist. Assoc. 2006; 101:168-178.
15. Sayiner M, Koenig A, Henry L, Younossi ZM. Epidemiology of Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis in the United States and the Rest of the World. Clin. Liver Dis. 2016; 20:205-214.