

The Classification Tree Combined with SIR and Its Applications to Classification of Mass Spectra

Ping He^{1,2}, Kai-Tai Fang¹, and Cheng-Jian Xu³

¹*Hong Kong Baptist University*, ²*Sichuan University*
and ³*Central South University*

Abstract: A new approach combining classification tree (CT) with sliced inverse regression (SIR) is proposed and applied to the classification of mass spectra in this paper. The classification tree has been widely used to generate classifiers from the mass spectral data because of its powerful ability in automatic variable selection and automatic interaction detection. However, it is often weak on presenting the linear and global relationships among variables. When the variables enter a model with the form of linear combination, the classification tree can not detect the form and leads to a low accuracy. SIR is an effective method to find useful linear combinations of predictor variables to regress the response variable. So merging CT and SIR harmoniously can inherit both advantages of them. Experiments in the paper show that the proposed approach can improve classification accuracy of decision tree and get better result than other classical classification methods.

Key words: Classification of mass spectra, classification tree, data mining, sliced inverse regression.

1. Introduction

Mass spectrometry, an instrumental technique which is used to character and identify chemical compounds, produces large amounts of valuable data for the chemical structure elucidation. In a mass spectrometer, molecules

of the investigated substance are ionized and the produced ions are separated according to their mass-to-charge ratio (m/z , mostly $z=1$) and the abundances of these separated ions are measured. Then a mass spectrum, bar plot with m/z versus abundance of ions (peak height) is provided. At present, the NIST 98 MS Database contains about 100,000 mass spectra with their corresponding chemical structures and the Wiley/NBS collection includes even more mass spectra to about 130,000 (NIST 98 MS Database and Wiley/NBS collection). How to extract the chemical structure information contained in these mass spectra? How to identify compounds and recognize structural properties automatically from mass spectral data? These are still open problems in chemometrics and have been attracted by many authors.

One group of strategy which has been widely applied to above problems is the classification methods based on multivariate statistics and data mining. Consider peak heights at integer mass-to-charge ratio as input variables and chemical structure properties, for example, the presence or absence of a certain chemical substructure, as response variables, classification methods can be used to find relationship between the structure properties and mass spectra or to train classifiers for automatic recognition of structural properties. There have been many kinds of classification techniques applied in this field, such as K-nearest neighbor classification, linear discriminant analysis, principal component analysis, neural networks, and classification tree. K-nearest neighbor (KNN) is a local averaging method. Linear discriminant analysis (LDA) and principal component analysis (PCA) are both globally linear method. Neural network (NN) is the learning method based on the artificial intelligence. Among these classification techniques, classification tree (CT) is the tree-based method which is effective in capturing the local character.

However, not all methods are effective because of the complicated characters of mass spectral data. First, the dimension of a typical piece of the data is very high. In the development of classifiers, it is evident that the generation of suitable spectral features is a crucial step (Werther *et al.* 1994). Therefore, the variables considered consist not only the original peak heights, but also a set of additional appropriate spectral features. In order not to miss all important information in the classification procedure,

it is common, just to stay on the safe side, to consider all kinds of spectral features in the development of classifiers. This makes the number of variables in certain such studies up to several hundreds (Werther *et al.* 2002). As the result the common phenomenon “curse of dimensionality” (Bellman 1961) in data mining also occurs for the mass spectral data. Secondly, it is known that complicated and widely unknown relationship between mass data and chemical structure do exist, various interactions may also present in the data. These types of interactions are even more difficult to model. For these reasons, many of the often used traditional classification methods are inefficient.

In view of these characters, a classification algorithm designed to develop the classifiers of mass spectral data should have two features: first, it can effectively select the variables which are important for classification; second, it takes various interactions into account during the modeling process. Although KNN, LDA and PCA are very useful classification methods in general, their applications to mass spectral data leave something to be desired. However, in many case studies, it has been shown that the classification tree is a powerful classification technique satisfying the above two demands. In particular, the approach proposed by Breiman *et al.* (1984) has been used to generate classifiers in analysis of mass spectral data. Moreover, its comprehensibility makes it more charm than neural networks, which usually provides result difficult to explain with subject matter knowledge.

It is known that the classification tree is weak at capturing classifiers of linear functions of the variables. In a typical mass spectral classification exercise, combinations of original peak heights are very likely to play the important role. For example, one of the numerical transformations successfully used for mass spectra classification is the summation of intensities at masses differing by a multiple of 14, which in most cases corresponds to a CH_2 group (Crawford and Morrison 1968). The transformation called Modulo-14 summation feature is simply a linear combination of the variables. Unfortunately the classification tree can not detect this feature, because a classification tree is typically based on a “divide and conquer” approach and the most useful variable is selected to split into a few (mostly 2) regions at every step. A natural extension is to add new variables, e.g., using certain transformations. An important class of such transformations

consists the linear combinations of the original ones. However there are also technical considerations when we add these linear combinations as the candidates of predictors. For example, we need to address such issues as how to find useful combinations and how to add new variables into the tree model. In this paper, we propose an approach which combines the classification tree with a suitable non-linear regression technique called the sliced inverse regression (SIR). This new method recursively adds some new variables produced by SIR into a set of predictors during the process of tree induction. SIR, Li (1991), is an effective method in nonlinear regression which uses the principal components from high dimensional data. SIR has been applied to many fields successfully, but it is seldom used in chemometrics. Experiments in the paper show that our new approach can effectively improve the prediction accuracy of classification tree in mass spectral data.

The paper is organized as the follows. In Section 2 we briefly introduce the decision tree and SIR and propose our approach. Mass spectrum and mass spectral data are described in Section 3. Two spectral data sets are studied in Section 4 to illustrate our procedure. The last section gives conclusion that our approach can improvement performance of the original classification tree.

2. Methodology

2.1 Classification tree

Classification tree has been popularized in the statistical community since the book of Breiman *et al.* (1984) was published. The method separates the sample by recursively splitting the data space till the stopping criteria are all satisfied (Friedman 1991). Suppose a data set consists of p inputs x_1, \dots, x_p and one response \mathbf{y} , for each of N observations, that is, $(x_{(i)}, y_{(i)})$ for $i = 1, \dots, N$, with $x_{(i)} = (x_{i1}, \dots, x_{ip})$. In the binary partitions process, the algorithm of classification tree firstly creates a node and finds a splitting rule to the given data set. The splitting rule including a split variable x_j and split value " λ " assigns observations to either left or right branch of the node. The observations with $\{x_j < \lambda\}$ are assigned to the left branch and those with $\{x_j > \lambda\}$ to the right respectively. Then

the data space is separated into two sub-regions. Secondly, repeat the same step in the each subspace till all the stopping criteria are satisfied. The extensive details of the algorithm can be seen in reference (Breiman *et al.* 1984).

There are many criteria based on different measures of node impurity, such as misclassification error, Gini index and deviance, to determine the choice of splitting rules in classification tree (Hastie *et al.* 2001). The criterion used in this paper is minimizing deviance. Suppose output variables \mathbf{y} take value $1, \dots, K$ and a node m , representing a region R_m , consists N_m observations $(x_{(i)}, y_{(i)})$, impurity measure deviance D_{R_m} is defined as

$$D_{R_m} = - \sum_{k=1}^K p_k \log p_k. \quad (1)$$

where p_k is the proportion of class k observations in this node m , i.e.

$$p_k = \frac{1}{N_m} \sum_{x_{(i)} \in R_m} I(y_{(i)} = k), \quad (2)$$

where $I(\cdot)$ is the indicator function. Then at each node, representing R , the classification tree performs a greedy algorithm to seek the splitting variable x_j and split value λ by solving

$$(x_j, \lambda) = \min_{x_j, \lambda} \left\{ \frac{N_1}{N} D_{R_1(x_j, \lambda)} + \frac{N_2}{N} D_{R_2(x_j, \lambda)} \right\}, \quad (3)$$

where N is the number of the observations contained in the region R , i.e. $N = |R|$, and $R_1(x_j, \lambda)$, $R_2(x_j, \lambda)$ are the two sub-regions obtained by using split pair (x_j, λ) :

$$R_1(x_j, \lambda) = \{\mathbf{x} | x_j \leq \lambda, \mathbf{x} \in R\}, N_1 = |R_1(x_j, \lambda)|$$

$$R_2(x_j, \lambda) = \{\mathbf{x} | x_j > \lambda, \mathbf{x} \in R\}, N_2 = |R_2(x_j, \lambda)|.$$

The decision tree is conceptually simple yet powerful. However, it has difficulty in capturing linear structure among the variables. If there exist

some linear combinations of the predictors in separating classes from the start, the classification tree will lead a weak classifier because the effect of an error caused by a single variable in the top split can be propagated down to all of the splits below it. So adding some linear combinations of the variables should improve the predictive power of the tree.

2.2 Sliced inverse regression

In this subsection, we discuss the basic idea of sliced inverse regression (SIR) and present its computational algorithm, as introduced by Li (1991). SIR is an effective method in non-linear regression since it combines the good parts of principal component analysis, inverse regression and the information including the dependent and independent variables.

Suppose \mathbf{X} is a $n \times p$ matrix with n observations and p variables and \mathbf{y} is a $n \times 1$ univariate output variable. Express \mathbf{X} and \mathbf{y} in terms of their rows, columns and elements by

$$\mathbf{X} = \begin{pmatrix} x_{(1)}' \\ \vdots \\ x_{(n)}' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = (x_1, \cdots, x_p), \quad \mathbf{y} = \begin{pmatrix} y_{(1)} \\ \vdots \\ y_{(n)} \end{pmatrix}.$$

Li (1991) introduced the following model

$$\mathbf{y} = g(\mathbf{X}\beta_1, \cdots, \mathbf{X}\beta_K, \epsilon),$$

where g is an unknown function, and β_1, \cdots, β_K are K unknown projection directions and the random error ϵ is independent of \mathbf{X} , but its probability distribution is unknown. Obviously, \mathbf{y} is a function of some linear combinations of x_1, \cdots, x_p . The K directions are used to generate the first K principal components, called effective dimension reduction (e.d.r.) direction which is our primary interest.

Based on the inverse regression, SIR inverses the role of \mathbf{y} and \mathbf{X} and treats the \mathbf{y} as if it were the independent variables and treats \mathbf{X} as if it were the dependent variable. It has been proved that if \mathbf{X} has been standardized to \mathbf{Z} , the standardized inverse regression curve $E(\mathbf{Z}|\mathbf{y} = y)$ is contained in

the linear space spanned by the standardized e.d.r direction β_1, \dots, β_K . For details of the theory, reader are referred to Li (1991). The covariance matrix $\Sigma_\eta = \text{Cov}(E(\mathbf{Z}|\mathbf{y}))$ is degenerate in any direction orthogonal to the β_k 's. Therefore, the normalized eigenvectors, β_k ($k = 1, \dots, K$), associated with the first largest K eigenvalues of Σ_η are the standardized e.d.r directions. In the approach, Σ_η are estimated by the slice method. The details are shown in the following algorithm of sliced inverse regression:

1. Standardize \mathbf{X} into $\tilde{\mathbf{X}}$, where $\tilde{x}_{(i)} = \hat{\Sigma}_{xx}^{-1/2}(x_{(i)} - \bar{x})$ ($i = 1, \dots, n$), where \bar{x} and $\hat{\Sigma}_{xx}$ are the sample mean and sample covariance matrix of \mathbf{X} respectively. $\hat{\Sigma}_{xx}^{-1/2}$ is the inverse of the positive definite square root of Σ_{xx} .
2. Sort the data $\mathbf{y} = (y_{(1)}, \dots, y_{(n)})'$ from the smallest to the largest.
3. Divide data set into H slices according to the sequence of \mathbf{y} : n_h is the sample size in the slice $I_h, h = 1, 2, \dots, H$.
4. Within each slice, compute the sample means i.e., $\hat{x}_h = n_h^{-1} \sum_{y_{(i)} \in I_h} \tilde{x}_{(i)}$.
5. Compute the sample covariance matrix for the slice means of \hat{x}_h , weighted by the slice sizes

$$\hat{\Sigma}_\eta = n^{-1} \sum_{h=1}^H n_h (\hat{x}_h)(\hat{x}_h)'$$

6. Find the eigenvalues of $\hat{\Sigma}_\eta$, $\lambda_1 \geq \dots \geq \lambda_p$ and associated eigenvectors $\hat{\beta}_1, \dots, \hat{\beta}_p$. The i th eigenvector $\hat{\beta}_i$ is called the i th SIR direction.
7. Project $\tilde{\mathbf{X}}$ along the first K SIR directions to form principal components.

Similar to principle component analysis, SIR chooses the first K ($K < p$) directions, where K is often determined by the accumulative contribute rate of eigenvalues $\sum_{i=1}^K \lambda_i / \sum_{i=1}^p \lambda_i$. In practice, K is commonly chosen so that the

cumulative contribute rate up to 80% \sim 90% or more. Note that there are some differences between principal component analysis (PCA) and SIR. SIR constructs the new explanatory variables considering the information of both \mathbf{X} and \mathbf{y} whereas PCA's new variables are determined by only the values of \mathbf{X} . The directions found by SIR are the ones that make \mathbf{y} change quickly.

In classification, H , the number of slices, equals to the number of groups and each group forms a slice. Then we can find linear combinations of original variables by the above algorithm. Obviously, the new variables can be directly used in construction of a decision tree. Based on the above discussion we proposed the following method combining classification tree and SIR.

2.3 Approach combining classification tree and SIR

In our approach the vector of predictors are expended by the principal components found by SIR. However, only adding new variables into the decision space prior to the tree induction, the same defects of subtree discussed earlier still exists in subspace after splitting the decision space. So we add new attributes recursively, that is, after the original region is split into two sub-regions, we apply SIR to current data set in each sub-region to find the new principal components and add them into the vector of predictors to join the next selection of the splitting rule.

Given a data set same with that in subsection 2.1, algorithm of the Classification Tree combined with SIR (CTS) includes six steps, as following:

1. Assume current node m , representing the region R_m , contain N_m samples. Apply SIR to the N_m samples and find the principal components to form the new variables.
2. Add the new variables to the set of predictors. (Let \tilde{p} be the number of predictors at this step)
3. Generate candidate splitting value set $\{\lambda_{ij}\}$ for each predictor x_j , $j = 1, \dots, \tilde{p}$. Here, $\lambda_{ij} = (\tilde{x}_{ij} + \tilde{x}_{i+1,j})/2$, $i = 1, \dots, N_m - 1$ and $\tilde{x}_{1j}, \dots, \tilde{x}_{N_m j}$ is a sequence of ordered value obtained by sorting samples according to the value of the predictor x_j .

4. Select the splitting rule (x_j, λ) using equation(3) from candidate splitting pair set $\{(x_j, \lambda_{ij})\}, i = 1, \dots, N_m - 1, j = 1, \dots, \tilde{p}$.
5. Generate two child branches and nodes to represent R_{m1} and R_{m2} which are obtained by splitting the region R_m , $R_{m1} = \{x_{(i)} | x_{ij} < \lambda \ \& \ x_{(i)} \in R_m\}$, $R_{m2} = \{x_{(i)} | x_{ij} > \lambda \ \& \ x_{(i)} \in R_m\}$.
6. Calculate the deviance of the two sub-regions and memorize the number of samples in the two sub-regions. If $D_{R_{m1}} \leq \alpha$ or $D_{R_{m2}} \leq \alpha$, the corresponding branch is terminated as a leaf. If $N_{m1} \leq \beta$ or $N_{m2} \leq \beta$, the corresponding branch is also terminated as a leaf. Then the leaf is assigned as the class k with the maximum proportion P_k (defined in equation (2)). Otherwise, repeated the Step1 - Step5 for each child node until the deviance of each child node is less than α or the number of samples contained in each child node is less than β .

Usually $\alpha=0.05 \sim 0.1$, $\beta = 5$. Smaller α or β will generate a bigger tree with higher accuracy in training but easily lead to over-fitting.

In this approach, SIR directions at every step should be reserved to make sure that the samples need to be predicted can be projected along the same directions to produce new variables during the process of prediction. Then the samples can be inferred according to the training tree. We have succeeded to apply this approach to classification of mass spectra as shown in subsection 4.

3. Mass Spectral Data

3.1 Mass spectrum

Mass spectrum is one of the most important spectra in chemometrics. It is the graphical representation of mass-to-charge ratio (m/z , mostly $z=1$) of separated ions versus abundance (peak height). For example, Figure 2 shows a mass spectrum of C_2H_4O , called acetaldehyde. A mass spectrum is produced by instrumental technique, Mass spectrometry, which is commonly use to character and identify chemical organic. The distribution of peaks in a mass spectrum is very characteristic for a compound, although not unique. Main information obtained from a mass spectrum is molecular mass,

and hints about substance class and parts (substructures) of the molecular structure.

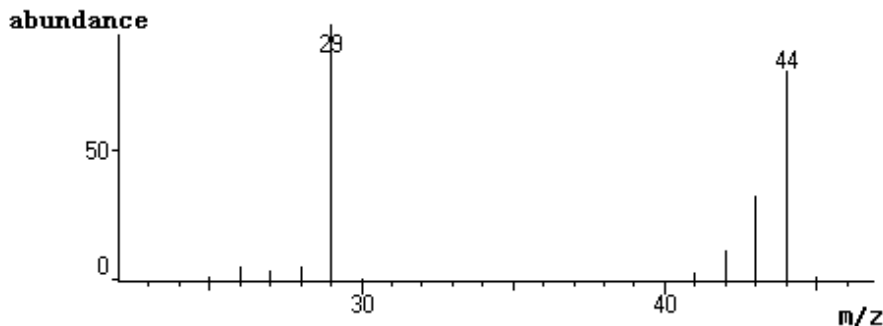


Figure 1. Mass spectrum of C_2H_4O

In our work, mass spectra and chemical structures are taken from the mass spectral library same with (Liang and Gan 2001). This library was established by transferring NIST62 mass spectrum library, which is built in the GCMS-QP5000 of shimadzu. The aim of spectra evaluation is either finding the relationship between mass spectra and chemical structures or identifying structure properties of an unknown compound from its spectrum. So in classification of mass spectral data, commonly, the height of peak at every m/z or spectral features (defined in the next subsection) are always considered as input variables, x_1, \dots, x_p , and the chemical structure properties, for example, the presence or absence of a certain chemical substructure, as response variables.

3.2 Spectral feature

In the following experiments, the vector of predictors includes not only the original peak heights of mass spectra but also a set of features obtained by transforming the original peaks. These spectral features are included because the following considerations:

1. As shown in earlier work, appropriate spectral features are simpler in molecular structures than the original peak heights of mass spectra

(Lohninger and Varmuza 1987 and Varmuza 1998).

2. Although tree structure can detect interactions of variables and SIR can find the useful linear relationships, some simpler relationships are not presented yet.
3. Decision tree has powerful ability on variable selection. If a spectral feature is irrelevant, it cannot influence the model too much. On the contrary, if a useful feature is added in the process of tree induction, the accuracy will be improved and the size of tree will be reduced.

Let I_m be the intensity of a peak at mass-to-charge ratio m/z . In the following, mass-to-charge ratio m/z is simplified as mass m . They are listed below:

A. *Intensities at a single mass normalized to local ion current*

Features of this group emphasize isolated peaks even if possessing only low intensities (Erni and Clerc 1972):

$$x_j = 100I_m / \sum I_k, \quad k = m - \Delta m, \dots, m + \Delta m.$$

B. *Logarithmic intensity ratios*

The logarithmic intensity ratio (Werther *et al.* 1994) is defined by

$$x_j = 100(L_m + \ln 100) / (2 * \ln 100).$$

$$L_m = \ln(I_k / I_{k+\Delta m}), \quad I_k = \max(I_m, 1), \quad I_{k+\Delta m} = \max(I_{m+\Delta m}, 1).$$

C. *Modulo-14 summation*

One of the first numerical transformations successfully used for mass spectra is the summation of intensities at masses differing by multiple of 14, which in most cases corresponds to a CH_2 group (Crawford and Morrison 1968). They are obtained by

$$x_j = 100s_j / s_{max},$$

$$s_{max} = \max(s_1, \dots, s_{14}) \quad s_j = \sum_k I_{l+14k} \quad l = 1 \dots 14, \quad k = 0, 1, 2, \dots$$

D. Autocorrelation

This feature can characterize mass differences between peaks as well as periodicities in a spectrum (Wold and Christie 1984):

$$x_j = 100 \sum_m I_m I_{m+\Delta m} / s_0, \quad s_0 = \sum_m I_m I_m.$$

4. Experiment

In this section we apply the approach proposed in subsection 2.3 to mass spectral data. Two case studies are discussed.

4.1 Data preprocessing

In practice, at every step of finding SIR directions, the variables for which the range of all samples is less than 1 are removed in order to avoid the arithmetic problem caused by calculating the inverse of matrix. Mass spectral data are inaccurate data. They are influenced by noise. Although we do not know the exact value of noise, it is sensible to think the noise is not less than 1. So if the range of the variable is less than 1, it can be regard as useless for classification. Even if the variable contributes to the classification, it can be detected by decision tree because we do not discard it in process of tree induction.

4.2 Example 1

We use a simple sample to illustrate the approach proposed in this paper and to compare results with other classification methods. Alcohol and aether are isometric compounds. From library NIST62, we select all 201 compounds that belong to the two groups and molecular formulas range from $C_5H_{12}O$ to $C_9H_{20}O$. The number of alcohol is 148 and that of aether is 53. From these 201 compounds, we select 141 of them at random as training sample and the remaining ones as testing sample. We assign heights of peaks at masses from 1 to 144 be the original variables because the molecular weight of $C_9H_{20}O$ is 144. Denote the alcohol by "0" and the aether by "1". We introduce the process of the new method as follows.

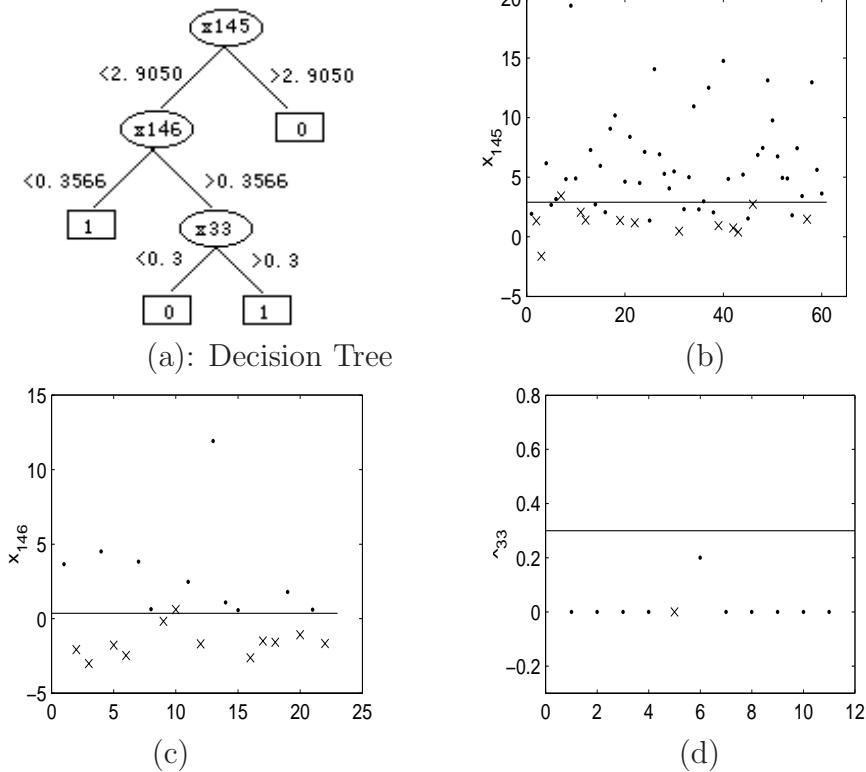


Fig 2: Result obtained by approach combining DT and SIR
 Note: “.” denotes alcohol, “x” denotes aether, “—” denotes splitting rule.

First, as mentioned in section 4.1, 79 variables are obtained in the training sample by removing the variables with ranges less than 1. These variables are used to find SIR direction. In this example, we only select the first SIR direction β to produce the new variable because the first eigenvalue's contribute rate is greater than 90%. Let $\mathbf{X}\beta$, the first β selected, be the 145-th variable, here, to fix ideas, the design matrix \mathbf{X} has been standardized. The splitting rules based on all the 145 variables uses x_{145} at the first node, and the value 2.905 is chosen to be the split value. The whole space is divided into two subspaces according to this splitting rule. At “ $x_{145} > 2.905$ ”, the algorithm has reached the stopping criterion and in this subspace, since the proportion of class “0” is great than that of class “1”, i.e. $P_0 > P_1$ (defined in equation (2)), the terminal node is set to be

“0”. For the case “ $x_{145} < 2.9050$ ”, the algorithm goes one step further to find SIR direction for the current data set. It turns out that a new variable, x_{146} , is added into decision the space, and the splitting pair $(x_{146}, 0.3566)$ is found. Here, the branch “ $x_{146} < 0.3566$ ” reaches the stopping boundary and is terminated as leaf “1”. Finally, for the case “ $x_{145} < 2.9050$ & $x_{146} > 0.3566$ ”, the same process is done, and all branches reach stopping boundary. The full graph (tree) is shown in Figure 2(a). The directions found by SIR at every step are shown in Appendix 1.

In prediction, first we use the same transformations on test samples with the SIR directions to generate additional variables to make the predictors. Then the first splitting rule is used to divide all the 60 testing sample. The result is shown in Figure 2(b). 38 of them fall into the subspace “ $x_{145} > 2.9050$ ” are predicted as “0”. Here, one testing sample predicted incorrectly. The remaining 22 samples are separated by the splitting pair $(x_{146}, 0.3566)$ as shown in Figure 2(c). There are 11 fall into the branch “ $x_{146} < 0.3566$ ” and therefore are predicted as “1”. Here all 11 cases are predicted correctly. Finally, the third splitting rule is applied to the remain 11 samples. From Figure 2(d), we can see the third split rule does not work in this situation because values of the 33-th variable of all 11 samples are less than 0.3. There is one sample misclassified.

We use leave-one-out cross validation to compare the predictive ability of our approach CTS with that of CT and PCA. The result is shown in the column of Table 1, where the result of classification tree is calculated by the program in S-plus 2000. It is observed that the predicting ability is improved by our approach.

Table 1: The Result of Leave-one-out Cross Validation

| Data | CTS(%) | CT(%) | PCA(%) |
|----------------------------------|--------|-------|--------|
| Alcohol and aether | 92.54 | 89.05 | 82.59 |
| Nitrogenous (No features) | 84.86 | 80.00 | 82.43 |
| Nitrogenous (Including features) | 90.29 | 87.43 | 84.57 |

In order to find the relationship between the substructure and mass spectra, the classification tree obtained, using all 201 training samples with CTS, is shown in Figure 3(a), where we use 4 decision nodes to classify the

two groups. The direction found by SIR at every step is shown in Appendix 2. From the first SIR direction used to form x_{145} , we observe that variables $x_{27}, x_{39}, x_{41}, x_{44}, x_{51}, x_{53}, x_{55}, x_{67}, x_{69}$ contributed most and x_{19}, x_{33}, x_{47} take second place. Moreover, $(x_{27}, x_{41}, x_{55}, x_{69}), (x_{39}, x_{53}, x_{67})$ and (x_{19}, x_{33}, x_{47}) are all CH_2 groups because the masses differ by 14 or multiples of 14 in each group. The values of the SIR direction corresponding with the first two groups are about 0.2 and those corresponding with x_{19}, x_{33}, x_{47} are about ± 0.15 . It can be inferred that x_{145} mainly reflects the combination of some CH_2 groups. Also, x_{146} formed at second step also reflects the combination of some CH_2 groups. In this direction $x_{33}, x_{47}, x_{44}, x_{53}, x_{55}, x_{67}$ are most important. In terms of the SIR direction used to get x_{147} , we can observe $x_{59}, x_{60}, x_{29}, x_{30}, x_{33}, x_{39}, x_{41}, x_{42}$ play the important role. The values of this SIR direction corresponding with x_{59}, x_{60} are smaller than -0.2 and those with $x_{29}, x_{30}, x_{33}, x_{39}, x_{41}, x_{42}$ are greater than 0.2. When we consider $(x_{29}, x_{30}, x_{33}), (x_{39}, x_{41}, x_{42}), (x_{59}, x_{60})$ as three local groups, x_{147} may reflect the importance of the combination groups formed by some adjacent peaks.

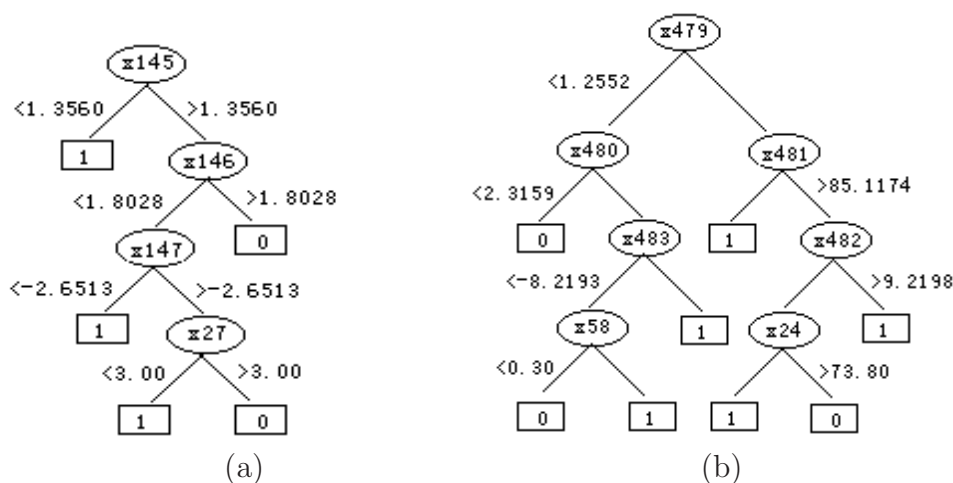


Figure 3. Decision tree

Note: (a) the first sample (b) the second sample

4.3 Example 2

In this experiment, 350 samples from nitrogenous compounds and other 350 samples from no-nitrogenous compounds are selected at random. The range of their molecular weights is from 100 to 200. To illustrate the power of our approach and to supply a comparison base against other methods, two kinds of experiments are considered. In one first case, the predictors are selected as the original peak heights. In another case, the predictors include the original peak heights as well as the spectral features. The heights of original peak at masses from 1 to 200 are the first 200 variables and the next 85 predictors are feature A (defined in subsection 3.2) for masses 15-100 with $\Delta m=3$; Feature B (defined in subsection 3.2) are obtained in the mass interval [15 100] with $\Delta m =1$ and 2, resulting 170 features; Modulo-14 summation are performed in the mass interval 15-100, resulting in 14 features; The mass spectral feature autocorrelation, defined also in subsection 3.2, are calculated for masses in 15-100 with $\Delta m=1, 2$ and 14-20, resulting in 9 features. In this way 478 variables are obtained. We compare the predictive ability of our approach with that of CT and PCA using leave-one-out cross validation. The result is shown in the last two columns of Table 1. The CTS obtained, using all 700 training samples, is shown in Figure 3(b). In this tree only 7 decision nodes were used compared with 31 decision nodes produced by only using CT. We do not display these SIR directions because of the limit of space, but it is available from one of the authors. From the result, we can observe that better classification accuracy can be obtained by CTS which incorporates the appropriate features to predictors. In fact, the improvement in accuracy can be significant. So using variables known to be important from experience and other studies is necessary.

5. Conclusion

In classification of mass spectra, the Classification tree is a very useful tool because mass spectral data are high dimensional and full of complex interactions. However, the classification tree is very weak in representing the linear and simpler relationship among variables and global factors. Our approach, which combines the classification tree and SIR, can make up the defect of classification tree and show improvement in predicting power.

Experimental results show that this approach can improve the classification accuracy of classification tree and reduce the size of tree. They also show that introducing the appropriate features into attributes can improve the classification ability of tree structure.

Acknowledgement

The work was partially supported by the Hong Kong UGC grant RGC/HKBU 2044/02P and Statistics Research and Consultancy Centre, Hong Kong Baptist University. The authors are thankful to valuable comments from Prof. Yi-Zhen Liang and Prof. Min-Te Chao.

Appendix 1

1. SIR direction found at the first step

| | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|-------|
| x2 | x14 | x15 | x17 | x18 | x19 | x26 | x27 | x28 | x29 |
| -.0070 | .0286 | .0673 | .1167 | .1418 | .1670 | .0981 | .2256 | .1240 | .1258 |
| x30 | x31 | x32 | x33 | x36 | x38 | x39 | x40 | x41 | x42 |
| .0889 | .0237 | .0355 | -.1617 | -.0211 | .1065 | .2358 | .1671 | .1983 | .0831 |
| x43 | x44 | x45 | x46 | x47 | x50 | x51 | x52 | x53 | x54 |
| .0602 | .2367 | -.0052 | -.0002 | -.1392 | .0824 | .1959 | .1303 | .2597 | .1093 |
| x55 | x56 | x57 | x58 | x59 | x60 | x61 | x65 | x67 | x68 |
| .2354 | .0797 | -.0385 | .0974 | -.0310 | -.0233 | -.1292 | .0996 | .2231 | .0875 |
| x69 | x70 | x71 | x72 | x73 | x74 | x75 | x77 | x79 | x81 |
| .1992 | .1092 | .0584 | .1051 | -.0269 | .0196 | -.0777 | .0784 | -.0043 | .1307 |
| x82 | x83 | x84 | x85 | x86 | x87 | x88 | x89 | x95 | x96 |
| .0800 | .1476 | .0975 | .0491 | .0832 | .0581 | -.0377 | -.1338 | .0506 | .0458 |
| x97 | x98 | x99 | x100 | x101 | x102 | x110 | x111 | x112 | x113 |
| .1122 | .0893 | .0513 | .0644 | .0557 | -.0162 | .0189 | .0728 | .1313 | .0611 |
| x115 | x116 | x117 | x126 | x128 | x129 | x130 | x142 | x144 | |
| -.0570 | -.1346 | -.0235 | .0983 | .0398 | -.0148 | -.0640 | .0277 | -.1360 | |

2. SIR direction found at second step

| | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| x14 | x15 | x18 | x19 | x26 | x27 | x28 | x29 | x30 | x31 |
| -.0223 | -.0313 | .2231 | .2213 | -.1675 | .0690 | -.0048 | -.1688 | -.1788 | -.1132 |
| x32 | x33 | x38 | x39 | x40 | x41 | x42 | x43 | x44 | x45 |
| .0403 | -.1898 | -.1376 | -.0155 | -.1119 | -.0928 | -.1461 | -.0629 | .2261 | .0367 |
| x46 | x47 | x53 | x54 | x55 | x56 | x57 | x58 | x59 | x60 |
| .0105 | -.1230 | .2167 | -.0800 | .1961 | -.1422 | -.1584 | -.0955 | .0236 | .0384 |
| x61 | x65 | x67 | x68 | x69 | x70 | x71 | x72 | x73 | x74 |
| -.1134 | -.0283 | .3362 | -.0169 | .2552 | .0454 | -.0748 | -.0589 | -.0955 | -.0352 |
| x75 | x79 | x82 | x83 | x84 | x85 | x86 | x87 | x88 | x89 |
| -.0785 | -.0700 | .1680 | .1240 | .0571 | .1186 | .0208 | .0953 | -.0342 | -.1049 |
| x97 | x98 | x99 | x101 | x102 | x112 | x115 | x116 | x117 | x129 |
| .2501 | .0391 | .0831 | -.0388 | -.0836 | .2374 | -.0368 | -.0857 | .0841 | .0159 |
| x130 | x144 | | | | | | | | |
| -.0284 | -.1207 | | | | | | | | |

Appendix 2

1. SIR direction found at the first step

| | | | | | | | | | |
|--------|--------|-------|--------|-------|--------|--------|--------|--------|--------|
| x2 | x13 | x14 | x15 | x16 | x17 | x18 | x19 | x26 | x27 |
| -.0085 | -.0113 | .0645 | .0999 | .0772 | .1128 | .1183 | .1556 | .1033 | .2264 |
| x28 | x29 | x30 | x31 | x32 | x33 | x36 | x37 | x38 | x39 |
| .1368 | .1188 | .0998 | .0124 | .0434 | -.1518 | -.0146 | .0318 | .1194 | .2284 |
| x40 | x41 | x42 | x43 | x44 | x45 | x46 | x47 | x50 | x51 |
| .1248 | .1932 | .0936 | .0923 | .2126 | -.0100 | .0183 | -.1564 | .1013 | .2059 |
| x52 | x53 | x54 | x55 | x56 | x57 | x58 | x59 | x60 | x61 |
| .1291 | .2553 | .1127 | .2312 | .0889 | -.0275 | .0959 | -.0371 | -.0357 | -.1303 |
| x65 | x66 | x67 | x68 | x69 | x70 | x71 | x72 | x73 | x74 |
| .1157 | .0661 | .2226 | .0870 | .1961 | .0997 | .0626 | .0999 | -.0240 | .0063 |
| x75 | x77 | x79 | x81 | x82 | x83 | x84 | x85 | x86 | x87 |
| -.0864 | .0736 | .0139 | .0913 | .0812 | .1466 | .1014 | .0466 | .0772 | .0237 |
| x88 | x89 | x95 | x96 | x97 | x98 | x99 | x100 | x101 | x102 |
| -.0436 | -.1309 | .0606 | .0482 | .1168 | .0871 | .0543 | .0629 | .0388 | -.0655 |
| x110 | x111 | x112 | x113 | x114 | x115 | x116 | x117 | x126 | x128 |
| .0231 | .0783 | .1114 | .0277 | .0238 | -.0294 | -.0863 | -.0233 | .1017 | .0337 |
| x129 | x130 | x142 | x144 | | | | | | |
| -.0015 | -.0914 | .0273 | -.1216 | | | | | | |

2. SIR direction found at second step

| | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| x2 | x13 | x14 | x15 | x16 | x17 | x18 | x19 | x26 | x27 |
| -.0788 | -.1760 | -.0949 | -.1007 | -.0058 | .1058 | .1518 | .1411 | -.0043 | .1507 |
| x28 | x29 | x30 | x31 | x32 | x33 | x36 | x37 | x38 | x39 |
| .0235 | .0388 | .0084 | .0021 | -.0711 | -.2935 | -.0399 | -.0410 | .0393 | .1492 |
| x40 | x41 | x42 | x43 | x44 | x45 | x46 | x47 | x50 | x51 |
| .1043 | .1838 | .0725 | .0352 | .2052 | .0166 | .0108 | -.2998 | .0603 | .1756 |
| x52 | x53 | x54 | x55 | x56 | x57 | x58 | x59 | x60 | x65 |
| .1109 | .2531 | .0900 | .2376 | .0812 | .0066 | .1009 | -.0488 | -.0241 | .0722 |
| x66 | x67 | x68 | x69 | x70 | x71 | x72 | x73 | x74 | x75 |
| .0679 | .2273 | .0998 | .1899 | .0972 | .1033 | .1091 | -.0975 | -.0192 | -.0930 |
| x77 | x79 | x81 | x82 | x83 | x84 | x85 | x86 | x87 | x88 |
| .0753 | -.0880 | .0897 | .0863 | .1526 | .0977 | .0860 | .0901 | .0039 | -.0421 |
| x95 | x96 | x97 | x98 | x99 | x100 | x101 | x102 | x110 | x111 |
| .0630 | .0500 | .1187 | .0987 | .0480 | .0667 | -.0095 | -.1414 | .0313 | .0808 |
| x112 | x113 | x114 | x115 | x116 | x126 | x128 | x129 | x130 | x142 |
| .1123 | .0281 | .0282 | .0420 | -.1199 | .1064 | .0349 | .0543 | .0363 | .0282 |

3. SIR direction found at third step

| | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| x14 | x15 | x19 | x26 | x27 | x28 | x29 | x30 | x31 | x32 |
| -.0354 | -.0364 | .0859 | -.1463 | -.1354 | -.0930 | -.2702 | -.2279 | -.1033 | -.0328 |
| x33 | x38 | x39 | x40 | x41 | x42 | x43 | x44 | x45 | x46 |
| -.2377 | -.1129 | -.2235 | -.1340 | -.2873 | -.2396 | -.1569 | -.0507 | -.0891 | -.1117 |
| x47 | x53 | x54 | x55 | x56 | x57 | x58 | x59 | x60 | x65 |
| -.1371 | .1383 | -.1263 | .0572 | -.1560 | -.2202 | -.0474 | .2375 | .2596 | -.0527 |
| x67 | x69 | x70 | x71 | x72 | x73 | x74 | x79 | x83 | x84 |
| .1800 | .0541 | -.1105 | .0840 | -.0709 | -.0871 | -.0357 | -.1005 | .1525 | -.0541 |
| x85 | x87 | x88 | x97 | x98 | x99 | x101 | x102 | x115 | x116 |
| .1400 | -.0607 | -.1116 | .0827 | .0578 | -.0786 | -.0381 | -.1317 | .0738 | -.1182 |
| x129 | | | | | | | | | |
| -.0926 | | | | | | | | | |

Note: Because the variables that may produce the arithmetic problem has been removed at every step, the i of x_i denotes the position at original variables.

References

- Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press.
- Breiman, L., Friedman, J. H., Olsen, R. A. and Stone, C. J. (1984). *Classification and Regression Tree*. Chapman & Hall, New York.
- Crawford, L. R. and Morrison, J. D. (1968). Computer methods in analytical mass spectrometry. Empirical identification of molecular class. *Anal. Chem.*, **40**, 1469-1474.
- Erni, F. and Clerc, J. T. (1972). Strukturäufklärung organischer verbindungen durch computerunterstützten vergleich sektraler daten. *Helv. Chim. Acta.*, **55**, 489-50.
- Friedman, J.H. (1991) Multivariate adaptive regression splines. *Anal. of Statistics*, **19**, 1-67.
- Hastie, T., Tibshirani, T. R. and Friedman, J. H. (2001). *The Elements of Statistical Learning-Data Mining, Inference, and Prediction*. Springer, New York.
- Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316-327.
- Liang, Y. Z. and Gan, F. (2001). Chemical knowledge discovery from mass spectral database I. Isotope distribution and Beynon table. *Analytica Chimica Acta.*, **446**, 115-12.
- Lohninger, H. and Varmuza, K. (1987). Selective detection of classes of chemical compounds by gas chromatography/mass spectrometry/pattern recognition: polycyclic aromatic hydrocarbons and alkanes. *Analytical Chemistry*, **59**, 236-244.
- NIST, NIST'98 Mass Spectral Database. (1998). *Nation Institute of Standards and Technology*, Gaithersburg, MD 20899.
- Varmuza, K. (1998). Chemometric detectors for selective detection in gas chromatographic analysis. *Trends In Analytical Chemistry*, **7**, 50-53.
- Wiley/NBS Mass Spetral Database 4th edition: Electronic Publishing Division.* John Wiley & Sons, Inc. New York.

- Werther, W., Lohninger, H., Stancl, F. and Varmuza, K. (1994). Classification of mass spectra. A comparison of yes/no classification methods for recognition of simple structural properties. *Chemometrics and Intelligent Laboratory systems*, **22**, 63-76.
- Werther, W., Demuth, W., Krueger, F. R., Kissel, J., Schmid, E. R. and Varmuza, K. (2002). Evaluation of mass spectra from organic compounds assumed to be present in cometary grains. Exploratory data analysis. *J. Chemometrics*, **16**, 99-11.
- Wold, S. and Christie, O. H. J. (1984). Extraction of mass spectral information by a combination autocorrelation and principal components models. *Anal. Chim. Acta.*, **165**, 51-59.

Received May 20, 2003; accepted August 2, 2003

Ping He
Department of Mathematics
Hong Kong Baptist University
Hong Kong, P. R. China
01400894@hkbu.edu.hk

Kai-Tai Fang
Department of Mathematics
Hong Kong Baptist University
Hong Kong, P. R. China
ktfang@hkbu.edu.hk

Cheng-Jian Xu
College of Chemistry and Chemical Engineering
Central South University
Changsha, P. R. China
xucj2000@263.net