Journal of Data Science **1**(2003), 481-496

Data Mining in Chemometrics: Sub-structures Learning via Peak Combinations Searching in Mass Spectra

Yu Tang[1,2], Yi-Zeng Liang[3] and Kai-Tai Fang[1]
[1]*Hong Kong Baptist University,* [2]*Suzhou University and*
[3]*Central South University*

Abstract: In this paper, a new approach of finding sub-structures in chemical compounds by searching peak combinations in mass spectra is given. Based on these peak combinations, further identification and classification methods are also proposed. As an application of these methods, saturated Alcohol and Ether are classified efficiently by using a variable selection method.

Key words: Mass spectra, peak combination, sub-structure, variable selection.

1. Introduction

Data mining (DM), also named knowledge discovery in databases (KDD) is an extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases (Han and Kamber 2001). It has been widely used in computer science, statistics, business management and biology, etc. Chemistry is an experience-dependent science. Most of the rules and disciplines obtained in this field come from chemical experiments and measurement data. With the growth of chemical measurement and modern information technology, more and more huge databases containing a large amount of chemical compounds information are established, such as the spectral databases, chromatographic

databases, or databases on molecular structures and their properties. How to discover new information from such huge collections is a new challenge.

In this paper, we wish to explore new knowledge from a mass spectral library, especially, we investigate sub-structures in different chemical compounds. By comparing large amount of compounds stored in the mass spectral library, we can define a certain peak combination that represents the character of a certain sub-structure. As long as these connections between peak combinations and sub-structures are established, further investigations can be carried on. As an example, saturated Alcohol and Ether can be classified efficiently by using a variable selection method.

All the data used in this paper are taken from the same mass spectral library as in Liang and Gan (2001). It was established by transferring NIST62 mass spectrum library, which is built in the GCMS-QP5000 of Shimadzu.

The paper is organized as follows. Section 2 introduces the method for finding sub-structures by searching peak combinations in mass spectra. Three case studies are demonstrated in Section 3. The last section gives some discussion.

2. Method for Sub-structure Discovering

2.1 Concept of peak combination

Mass spectrum is one of the most important spectra in Chemometrics. It can identify different chemical compounds by showing different spectral skeleton. But, generally speaking, how to identify efficiently or how to classify different compounds by the analysis of mass spectra is still an open problem as up to now more than 23,400,000 chemical compounds have been found. In the past decades, many authors have used the values of single peaks as variables to identify or classify different chemical compounds, see, for example, Gan *et al.* (2001). However, different peaks occurring in the mass spectrum should not be regarded as independent with each other. In fact, many series of peak combinations in mass spectrum are always relevant, as it will be seen in the later discussion. In Werther *et al.* (1994) and Werther *et al.* (2002), the authors proposed some vari-

ables representing the character of the compounds based on the peak values of the origin mass spectra. Many features (transformations), such as "Modulo-14-features", "Autocorrelation features", "Logarithmic intensity ratios" and "Spectra type features" have appeared in the literature. Especially, the feature "characteristic peak series" presented in Werther *et al.* (2002) shows some prior information of sub-structure contained in the compounds. They used three peak combinations, that is, peak values of 64-65-66-90, 28-55-82-109 and 45-59-73-87-101 to identify the existence of sub-structures "nitrogen-substituted pyrimidine", "$[(HCN)_{1-4}+H]^+$" and "$[C_nH_{2n-1}O_2]^+$", respectively. How to select more of such characteristic peak combinations is the motivation of the present paper.

Mass spectrum is a stick diagram which records the abundances of different ions broken from the vaporized organic molecular after its being bombarded by a stream of electrons. Firstly the organic molecular is to become an energetically unstable molecular ion, and then the molecular ion is subsequently broken into a host of particles. Among them, those fairly larger and unstable particles continue to be broken into smaller ones. During the fragmentation, the charged particles will be accelerated, deflected and detected by the mass spectrometer and finally be recorded in the mass spectrum. Figure 1 is a typical mass spectrum of Heptane.

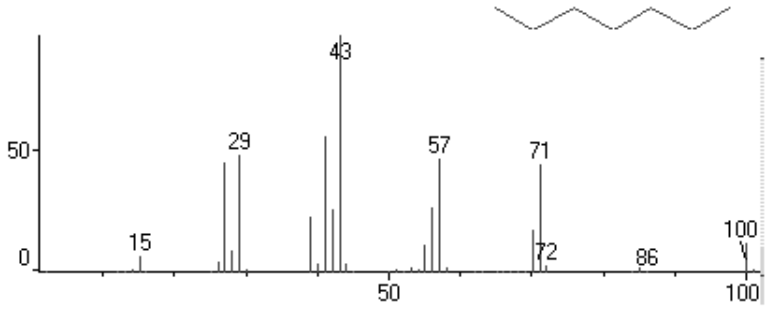

Figure 1: The Mass Spectrum of Heptane

From the formation of the mass spectra, we see that each charged parti-

cle during the fragmentation may leave a line in the mass spectra and these series of particles are produced step by step. Thus, when two compounds contain a common sub-structure, same particles may have chances to appear during the fragmentation and their "trace" will be recorded in the mass spectra. If we check large amount of known compounds in the library, we can select certain position combinations occurring together most frequently in the mass spectra. Tabulating these position combinations in a file, we browse the database again and compare the performance of different compounds on each fixed position combination. We can select certain position combinations on which some of compounds perform the same behavior, i.e. their mass spectra share similar peak values on the corresponding position combinations. We call such a position combination a peak combination. And those compounds are called models. Then we will draw a conclusion that all models on a peak combination belong to the same class in the sense of their existing a common sub-structure. So roughly speaking, a peak combination in mass spectra corresponds to a sub-structure in compounds. In this way, the spectrum-structure information can be established.

Two types of peak combinations are defined as follows. The first type is called "*subsequential type*". It refers to such peak combinations that are formed by recording a series of subsequential ions from big to small. In the mass spectrum, it has the character of the following properties:

a. different positions in such peak combinations scattering far from each other;

b. each position with higher abundance.

Another type is called "*cluster type*" that is typically formed by losing several small particles from a definite ion fragment and has the following properties:

1. the positions of these peak combinations occurring together in an area of mass spectrum, constituting a cluster of a family;

2. at least one of the abundance of different positions should be higher, but some can be fairly low.

As showed in Figure 1, 71-57-43-29 can be regarded as the subsequential type, while 26∼29, 39∼44, and 53∼58 can be regarded as the cluster type.

2.2 Data set construction

As stated in Subsection 2.1, we establish a data file which records the position combinations frequently occurring in the mass spectral library. It will be used for the later comparison. Corresponding to the two types of peak combinations defined above, we adopt two strategies to search them in the mass spectral library.

For the subsequential type, we list all the peak positions with their abundance above a threshold for each selected sample. Then from these positions we select their combinations most frequently occurring in the same mass spectra, say, the frequency ratio of their occurrences together in the total samples is above 10%. For a typical mass spectrum, it always has several scores of peaks, sometimes even more than one hundred. So when the mass spectral library is large enough, it is a hard task to list all these position combinations. In our program, instead of listing all these position combinations, a random selection method is used. That is, we randomly select certain position combinations whose single peak positions occur frequently, then record those position combination whose occurrences in the mass spectra library as a whole are above the threshold value. Though this substitution may leave out many characteristic peak combinations, it might explore some interesting results in our experiment, where the number of selection time is set to be a fairly large number, for example, 100,000 times.

For the cluster type, we record all the successive positions with definite length which most frequently occur in the database. Here we admit of at most one zero value among these positions. For example, if we set fixed length five, and many compounds have non-zero values on 41-42-43-45, then we also regard 41∼45 as a position combination. From now on, the data set is constructed by the above method.

2.3 Sub-structure learning

It should be emphasized that the position combinations in the data set are not necessary to link to certain sub-structures existed in compound. However, if a certain sub-structure does exist in some compounds, it may leave some information related on these position combinations.

Given a compound, whether it is in the library or not, we can compare its

values on all the recorded position combinations with those of compounds listed in the library. If there are some compounds in the library share similar values on a certain position combination with the given compound, we can pick out these compounds. As stated in Subsection 2.1, such position combination is now defined as a peak combination and will link to a certain sub-structure, while these compounds are treated as models. In the program, we use two similarity indices for the behavior of two compounds on each of these peak combinations. The first one is the correlation coefficient. Denote $A = \{a_1, a_2, \cdots, a_n\}$ and $B = \{b_1, b_2, \cdots, b_n\}$ the corresponding value vectors on n-peak combination of two compounds, respectively. The correlation coefficient between A and B is defined as follows:

$$\rho = \frac{\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \overline{b})^2}},$$

where $\overline{a}$ and $\overline{b}$ represent the mean of A and B, respectively, i.e., $\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$ and $\overline{b} = \frac{1}{n}\sum_{i=1}^{n} b_i$. The second one is called coefficient of variation (CV):

$$CV = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(R_i - \overline{R})^2}\Big/\overline{R},$$

where

$$R_i = \frac{a_i}{b_i},\ i = 1, 2, \cdots, n,$$

and $\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$ represents the mean of R_i. It is obvious that if some b_i equals zero while the corresponding a_i is a non-zero value, then compound A and B must not share the common behavior on this peak combination, so we terminate the process. The two criteria can frequently be seen in the

literature, we use both of them for comparison. The threshold values of the two criteria can be fairly arbitrary. Generally speaking, they don't affect the result much. In Case Study II in Section 3, they are set to be 0.99 and 0.15 in the program, respectively. However, when the sample size is not so large, as it will be showed in Case Study I, they are set to be 0.99 and 0.25. And when we want to get more peak combinations for later identification and classification, as in Case Study III, they can be set even looser.

Before ending this section, we summarize the whole procedure for searching peak combinations from mass spectra into the following steps:

Step 1: According to the two types of peak combination, use two corresponding ways of searching methods introduced in Subsection 2.2 to list position combinations.

Step 2: For each position combination found in step 1, select a compound in the library, and calculate two similarity indices between the other samples with the one selected.

Step 3: Record the samples whose correlation coefficient is above the bound and whose coefficient of variation is below the bound. Check these samples whether they share a common sub-structure. If so, then regard the position combination as a peak combination. And these samples with satisfied similarity indices are recorded as models.

Step 4: Iterating step 2 and 3 after all the samples are selected for comparing.

Step 5: Iterating step 2, 3 and 4 after all the position combinations in step 1 are tested.

3. Case Studies

In this section, three case studies are investigated and show that the peak combination approach is very useful in data mining.

Case Study I. We use 34 saturated Alcohol and Ether with molecule weight 102. Using the searching method in Subsection 2.2 and 2.3, we

find several peak combinations in Table 1 which seem to represent the substructure $CH_3CH(OH)-$ existed in compounds (in this paper, the Index means the serial No. in NIST62).

Table 1: Peak combinations in the first experiment

Index (Model)	Value of peaks					Molecular formula
$--$	27	29	41	44	45	
1769	7.2	7.2	13.6	6.8	100.0	$CH_3(CH_2)_3CH(OH)CH_3$
1773	8.8	7.2	18.0	6.8	100.0	$CH_3CH(CH_3)CH_2CH(OH)CH_3$
1777	11.6	14.0	19.2	11.6	100.0	$CH_3CH_2CH(CH_3)CH(OH)CH_3$
1792	8.0	7.6	15.6	6.0	100.0	$CH_3(CH_2)_3CH(OH)CH_3$
1794	8.0	8.4	15.2	6.8	100.0	$CH_3(CH_2)_3CH(OH)CH_3$

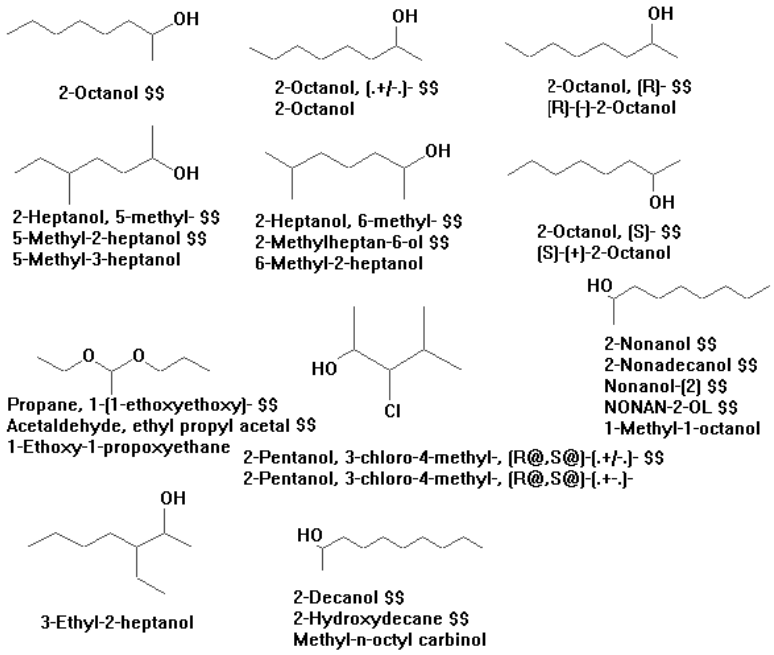

Figure 2: Results selected from 8000 compounds

Here We find that No. 1769, 1792 and 1794 have the same peaks pattern, it is not surprise as they share the same molecular formula $CH_3(CH_2)_3CH(OH)CH_3$.

We also check the above rule in the whole library. For each compound in the library, we compared it with the selected models in the rule, that is, calculate the ρ and CV on the position combination 27-29-41-44-45. If the average of ρ and CV are greater than 0.9 and 0.25, respectively, we claim that the compounds contain the sub-structure $CH_3CH(OH)-$.

Figure 2 lists the compounds it finds from serial No. 5501 to 13501. Most of them contain sub-structure $CH_3CH(OH)-$ in deed. What should be pointed out is that the base peak 45 seems much more important. However, in all the 34 samples, there are five types of Ether with base peak 45 listed in Figure 3. Single peak value may mix up all of these substances. We distinguish them by using peak combinations.

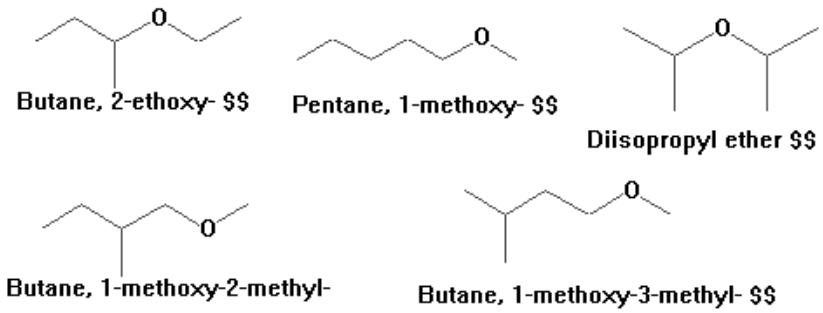

Figure 3: Ether in 34 samples with base peak 45

Case Study II. We use all single alkene and cycloparaffinic hydrocarbons in the library and altogether there are 693 compounds. Similar to Case Study I we only list the result peak combinations in Table 2.

Case Study III. This is the most interesting one. As we know, it is a difficult task to distinguish whether a compound is Alcohol or Ether due to their similar properties. In this experiment we will present an efficient way to distinguish these two classes by using peak combinations.

We select all saturated Alcohol and Ether in the library. Altogether there are 375 samples, with 295 types of Alcohol and 80 types of Ether. We use half of them, i.e. all the even order in the 375 samples sorted by the

Table 2: Peak combinations in the second experiment

Index (Model)	Value of peaks						Possible Sub-structure
−−	54	55	57	67	68	111	
18491	14.80	81.60	60.00	12.00	12.80	15.20	
21972	15.60	80.40	60.40	12.80	13.60	15.60	$CH_2 = CH-$
25488	14.80	84.80	66.00	13.60	14.00	19.20	
28779	16.40	91.60	66.00	15.20	14.40	24.40	
34426	15.20	88.40	71.60	14.00	14.40	22.40	
−−	29	67	82	83	84	111	
28772	29.20	20.40	16.80	66.40	32.40	22.40	
28776	24.00	14.80	12.00	51.60	24.00	15.20	
34414	30.40	19.60	18.80	72.80	31.60	26.80	$-CH = CH-$
34415	23.20	14.40	13.20	54.80	21.20	17.20	
34422	34.80	20.40	16.40	77.20	28.80	22.40	
39519	31.60	21.20	20.00	84.80	30.00	29.20	
−−	43	55	82	97	111		
28770	73.20	100.00	65.60	58.80	26.00		
31652	73.20	90.80	61.60	52.80	21.20		$K_6 - CH <$
34421	66.00	84.80	61.60	50.80	20.00		
34423	67.60	82.80	58.00	51.20	25.20		
37069	62.40	74.40	59.60	50.40	19.60		
−−	53	54	55	56	57	58	
31650	3.60	8.40	97.20	25.20	100.00	3.60	
31651	3.60	7.60	87.60	18.00	93.20	3.60	
31652	3.60	8.00	90.80	18.00	100.00	3.60	
34420	3.20	7.20	85.60	17.20	91.20	4.00	
34421	3.20	8.00	84.80	19.60	100.00	3.60	$K_6 - CH <$
37067	4.00	9.20	87.60	22.40	98.00	3.60	
37071	3.20	7.60	77.60	16.40	100.00	3.60	
50827	3.60	10.00	96.80	22.00	95.20	4.00	
50829	3.60	10.00	98.40	23.20	100.00	4.40	
−−	50	51	52	53	54	55	56
4653	2.00	4.40	2.80	14.40	15.60	100.00	56.00
4664	2.00	4.80	2.80	16.00	16.00	100.00	57.60
4674	1.60	4.00	2.40	13.60	13.60	100.00	61.20
11059	1.60	3.60	2.40	14.40	14.80	100.00	60.00
11074	1.60	3.20	2.00	12.00	12.40	62.00	47.20
11081	2.00	4.00	2.80	15.20	14.80	93.60	62.40

serial No. in NIST62, to establish a classifier, and use all of them (including the training samples) to check the efficiency of the classifier. Firstly, as done in the above two experiments, we also list all the peak combinations the program finds. While here there are some differences. As the second "cluster type" peak combination allows of zero value (as pointed out in Subsection 2.2), it is not used here to define variables. For the first "subsequential type" peak combination, the bound values of the two criteria, i.e. ρ and CV, are set to be a little looser. In fact, in the program they are set

to be 0.7 and 0.3, respectively in order to present more chances to find useful variables which can classify the compounds. As a result, we obtain 1381 peak combinations. In the following, we will define corresponding 1381 variables based on these peak combinations. Also we use a variable selection method to select some of these important peak combination variables to establish a linear model to classify Alcohol and Ether efficiently. If a peak combination forms by positions $P_1, P_2, \cdots, P_l$ with models $a_1, a_2, \cdots, a_n$, then we define a corresponding variable v as follows. For convenience, we use the notation $x(P_j)$ to represent the abundance of compound x on position P_j. Denote

$$m_j = \frac{1}{n}\sum_{i=1}^{n} a_i(P_j), j = 1, 2, \cdots, l.$$

For each compound x, the variable v can be calculated as follows:

$$v_x = \sqrt{\frac{1}{l-1}\sum_{i=1}^{l}(R_j - \overline{R})^2} \Big/ \overline{R}, \tag{1}$$

where

$$R_j = \frac{x(P_j)}{m_j},\ i = 1, 2, \cdots, l,$$

and $\overline{R} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{l} R_j$. However, these 1381 peak combination variables may be highly correlated and may not be efficient enough to distinguish the two classes. We aim to select several of them to establish the classifier. For this reason two steps are added on these peak combination variables. Firstly, we eliminate the highly correlated variables. If the correlation coefficient of two such variables is higher than 90%, one of them will be removed. After this process, 162 variables are left to be further selected from. Then we use a variable selection method proposed by Fan and Li (2001). The main point of this method is to use penalty functions which are symmetric, nonconcave and have singularities at the origin to produce sparse solutions. The authors also provide an algorithm in their paper for optimizing penalized likelihood functions. Here we apply their algorithm to our database. To the end, it recommends six variables. After checking with the origin peak combinations

record, we find these six variables correspond to the six peak combinations listed in Table 3.

Table 3: Peak combinations in the third experiment

Position	Index(Model)	Value of peaks
29 67 82 97 125	26420	36.0 19.6 36.0 56.0 14.4
	29675	29.2 18.8 42.8 57.2 12.0
	40236	30.4 18.8 44.0 72.0 19.2
	46406	39.6 19.2 31.6 76.4 13.2
	55569	27.2 13.2 21.6 55.2 10.8
70 81 84 98 125	26420	34.0 14.0 22.8 13.6 14.4
	40236	39.6 12.4 27.2 16.4 19.2
	46406	34.0 14.8 24.8 20.0 13.2
	55569	23.2 10.4 14.4 12.4 10.8
	57795	28.4 11.6 17.2 11.2 10.4
29 67 83 111 125	26420	36.0 19.6 66.8 28.4 14.4
	29675	29.2 18.8 82.8 27.6 12.0
	40236	30.4 18.8 82.8 36.4 19.2
	46406	39.6 19.2 71.2 33.2 13.2
	55569	27.2 13.2 48.4 24.0 10.8
31 55 82 85 98	26420	16.0 95.2 36.0 18.8 13.6
	29675	20.4 100.0 42.8 14.8 17.6
	32424	22.4 100.0 39.2 15.2 16.0
	37841	18.0 100.0 41.2 22.0 20.0
	40236	20.8 100.0 44.0 22.0 16.4
69 81 84 98 111	26420	68.0 14.0 22.8 13.6 28.4
	32424	82.0 12.4 29.2 16.0 25.2
	40236	77.2 12.4 27.2 16.4 36.4
	46406	61.2 14.8 24.8 20.0 33.2
	55569	46.8 10.4 14.4 12.4 24.0
	57795	62.4 11.6 17.2 11.2 23.6
54 69 81 84 97	26420	13.2 68.0 14.0 22.8 56.0
	29675	12.8 80.0 10.4 32.4 57.2
	32424	14.8 82.0 12.4 29.2 57.2
	40236	12.4 77.2 12.4 27.2 72.0
	46406	13.6 61.2 14.8 24.8 76.4

Define a response variable y. For Alcohol, set $y = -10$; for Ether, set $y = 10$, then the variable selection method proposed in Fan and Li (2001) also gives a linear regression on these six variables, which is listed in the following:

$$Y = 0.0989v_1 + 0.0907v_2 - 0.0632v_3 - 2.6395v_4 + 0.0244v_5 - 0.0368v_6. \quad (2)$$

Here $v_i, i = 1, 2, \cdots, 6$ represents the i-th variable in the sense of (1). Use this regression model, we check all 375 samples, and obtain an extremely satisfied result with only one misclassification. We show it by Figure 4.

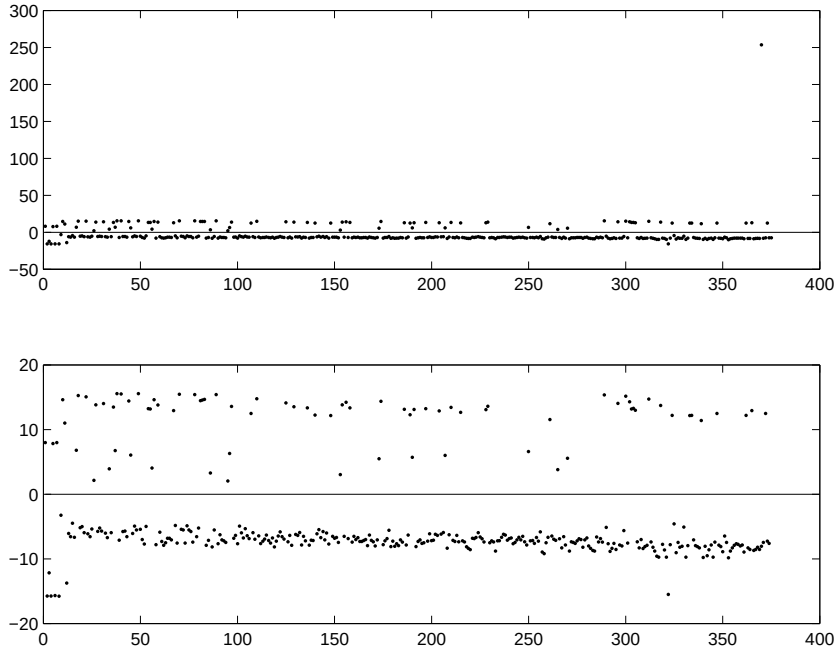

Figure 4: Regression result in the third experiment. Top: for all the samples; below: after removing the outlier.

The misclassification sample is called Nonacosan-10-ol with molecular form $C_{29}H_{60}O$. Its mass spectrum is showed in Figure 5. Deep investigation into this mass spectrum, we can find that it has totally 10 peaks and all of the peaks are separated from each other. It is strange in our knowledge and

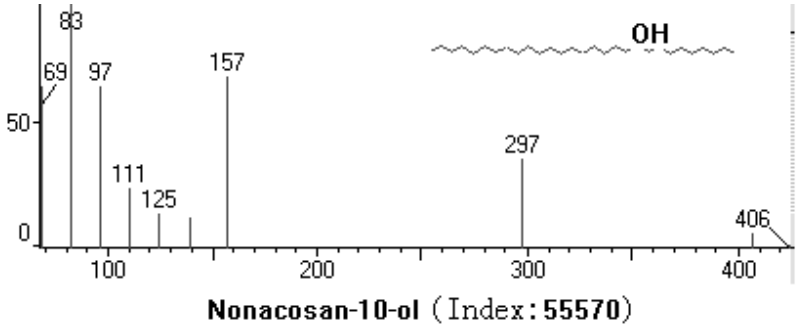

Figure 5: The mass spectrum of Nonacosan-10-ol.

may be considered as an outlier. Thus the program distinguishes it out of the other compounds.

It should be noted that the same process has been added to the Case Study II and the result is even better. Alkene and cycloparaffinic hydrocarbons are classified exactly with five variables. So it has a potential application in classification of chemical compounds.

4. Conclusion and Discussion

In this paper, a new classification approach of dealing with the mass spectra is proposed. By selecting characteristic peak combinations of different sub-structures, different compounds can be identified and classified using prior information obtained from a given library.

What's more, the above approach can be extended directly to determine a mixture with several different compounds, which is mostly useful for the determination and classification in the Chinese herbal medicine. Due to the experiment conditions and other constrains, pure component can't be picked out by using high-resolution chromatography. Thus, the mass spectrum obtained may represent a combination of more than one compound's skeleton. It makes the decomposition of the mixture and one-one match work more complex. However, if we use the information of peak combinations comparison, we may recognize these compounds even when they are not contained in the library samples. We can identify the sub-structures from these peak combinations, and finally reconstruct the compounds the mixture contains by combining these sub-structures.

In short, peak combinations approach can provide a new chance for the analysis of mass spectra. A series of subsequent results will be carried out.

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 thank Prof. Min-Te Chao for valuable comments.

References

Fan J. Q. and Li R. Z. (2001). Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association* **Vol. 96, No. 456**, 1348-1360.

Gan F., Yang J. H. and Liang Y. Z. (2001). Library search of mass spectra with a new matching algorithm based on sub-structure similarity, *Analytical Science* **17**, 635-638.

Han J. W. and Kamber M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Liang Y. Z. and Gan F. (2001). Chemical knowledge discovery from mass spectral database I. Isotope distribution and Beynon table, *Analytica Chimica Acta* **446**, 115-120.

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, *Journal of Chemometrics* **16**, 99-110.

Werther W., Lohninger H., Stancl F. and Varmuza K. (1994). Classification of mass spectra. A comparison of yes/no classification methods for the recognition of simple structural properties, *Chemometrics and Intelligent Laboratory Systems* **22**, 63-76.

Received May 20, 2003; accepted August 2, 2003

Yu Tang
Department of Mathematics
Hong Kong Baptist University
Hong Kong, P. R. China
01400940@hkbu.edu.hk

Yi-Zeng Liang
College of Chemistry and Chemical Engineering
Central South University
Changsha, P. R. China
yzliang@cs.hn.cn

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