Journal of Data Science **1**(2003), 497-509

Building an Honest Tree for Mass Spectra Classification Based on Prior Logarithm Normal Distribution

Cheng-Jian Xu[1], Ping He[2,3] and Yi-Zeng Liang[1]
[1]*Central South University,* [2]*Hong Kong Baptist University and* [3]*Sichuan University*

Abstract: Structure elucidation is one of big tasks for analytical researcher and it often needs an efficient classifier. The decision tree is especially attractive for easy understanding and intuitive representation. However, small change in the data set due to the experiment error can often result in a very different series of split. In this paper, a prior logarithm normal distribution is adopted to weight the original mass spectra. It helps to building an honest tree for later structure elucidation.

Key words: Classification, data mining, decision tree, mass spectra.

1. Introduction

Mass spectrometry (MS) is a commonly used instrument for the characterization and identification of organic compounds. It is especially useful and efficient when the substance to be measured presents in trace levels. The identification of analyzed compounds in mass spectrometry is usually based on similarity matching techniques against mass spectra library. The most widely used computer programs are Finnigan INCOS dot product (Sokolow *et al.* 1978) and probability based matching (PBM) algorithms (McLafferty and Stauffer 1985). Such techniques only work well when correct reference data are available. When the unknown compound is not included in the library, the matching technique may be less useful. At present, the commonly used mass spectra libraries contain hundred thousands compounds at

most. Compared with about more than twenty millions organic compounds in the world, the volume of mass library is rather limited. At this time, an interpretative system (Luinge 1990) will be more useful. However, the interpretation and deriving structure proposal from MS data is usually not an easy job because of the complicated and widely unknown relationships between MS data and its chemical structures. The initial ambitious goal to construct fully automated systems is far more being reached and is being replaced by the more realistic prospect of developing decision-supported system (Lebedev and Cabrol-Bass 1998). Accordingly, structure elucidation problem may resort to other classification methods, which can reduce the number of candidate structure significantly.

The primary goal of spectra classification is to find correlation between the properties of compounds and their mass spectra. Such relationship is essentially very complicated. Many efficient classifiers have been manipulated to fulfill the above task, such as linear discriminant analysis (LDA) (Varmuza and Werther 1996), classification and regression trees (CART)(Breiman *et al.* 1984), K-nearest neighbor (KNN)(Alsberg *et al.* 1997) and neural network (Eghbaldar *et al.* 1996) and so on. Among these, the decision-tree method, although simple but powerful, has been wildly used in data mining. The pioneer work in Dental projection (Lindsay *et al.* 1980) for mass spectra elucidation is mainly based on decision tree. Compared with neural network, which is often a black box, the decision-tree method has an intuitive representation and the resulting model is easy to understand and assimilated by humans. Moreover, the decision tree provides nonparametric models, no intervention being required from the user, and thus they are very suited for exploratory knowledge discovery. However, one major problem with trees is their high variances (Hastie *et al.* 2001). A small change in the data set can often result in a very different series of splits, which make later interpretation somewhat precarious and difficult.

The signals generated by mass spectra detector will fluctuate at different experiments, which prevents us from using the trees method to discover knowledge in the mass spectra library data directly. In this paper, a probability weighting approach is adopted to weigh the data first, which reduces the sensitivity of data set without losing the characteristic information. The later experiment and comparison show the efficiency of proposed weighting

approaches for building an honest tree.

2. Theory

2.1 Decision tree

The tree-based method is an exploratory technique for discover structure in data, which separate the sample by recursively binary split the data space till the stopping criteria are all satisfied (Friedman 1991). The recursive partitioning fit model takes the form in every region:

$$\text{if } \mathbf{x} \in R_m, \text{then } \hat{f}(\mathbf{x}) = g_m(\mathbf{x}|\{a_j\}_1^p). \tag{1}$$

Here $\{R_m\}$ are disjoint sub-regions representing a partition of the entire data space D. The functions g_m are generally taken to be of quite simple parametric form. The most common one is a constant function

$$g_m(\mathrm{x}|a_m) = a_m. \tag{2}$$

In the process of tree induction, firstly, the entire data space is split into two sub-regions, and in each region the response is fit by equation (1). Here, the split variable and split point are selected by greedy algorithm in term of some criterion. Then one of both of these regions are split into more regions, and this process is continued, until some stopping rule is applied. As far as the criterion of selecting split rules, in classification tree, commonly there are three ones base on different node impurity measures. Suppose a node m, representing a region R_m. The region includes N_m observations $(x_i, y_i), i = 1, \cdots, N_m$ with response taking value from $\{1, 2, \cdots, K\}$. Then $\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$ is the proportion of class k observations in the node m, where $I(\cdot)$ is indicate function. The commonly used different measures of node impurity include:

1. Misclassification $\frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m))$, where $k(m) = \mathrm{argmax}_k \hat{p}_{mk}$.
2. Gini index $\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'}$.
3. Deviance $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.

So at each node, the variable x_j and point s will be selected to split the node if the sum of impurity measures for two sub-regions obtains the minimum after the node are split by the pair (x_j, s).

The tree methods were once successfully used to predict the presence of chlorine and bromine in organic components (Breiman *et al.* 1984). The instability is one of major problems of tree methods, and it was regarded as the price to be paid for estimating a simple tree-based structure for the data (Hastie, Tibshirani and Friedman 2001).

2.2 Data weighting

Due to measuring errors, the mass spectra of a certain organic component are not identical at different experiments. As shown in Figure 1, they often show relative large variances in intensity at some mass to charge ratios. The experimental spectrum of 2-Undecanone is similar but not exactly the same as it is recorded in the standard library. Nevertheless these two spectra indeed refer to one component. Therefore, these differences in mass spectra demonstrate experimental errors instead of diversity of mass spectra.

The tree method is often sensitive to a little change of data set. The major reason for this instability is the hierarchical nature of the process: the effect of an error in the top split is propagated down to all of the splits below it. The high variance is regard as a major problem with tree. Therefore, it is not good to use it directly, because if a very different series of splits are obtained due to experimental errors, it is impossible for us to interpret the classification model. An honest tree for classification of mass spectra is indeed necessary. An honest tree is a tree whose structure should not change greatly with a little change of data set. Some transformation of original data sets is necessary. Since the tree method is often sensitive, it is not good to use it directly.

Grotch (1969) once declared that the intensity of mass spectra closely follows a log normal distribution. Liang and Gan (2001) also reported the same distribution recently. This prior distribution is useful for us to transform the data. Then one can weight the original data accordingly (McLafferty and Stauffer 1985).

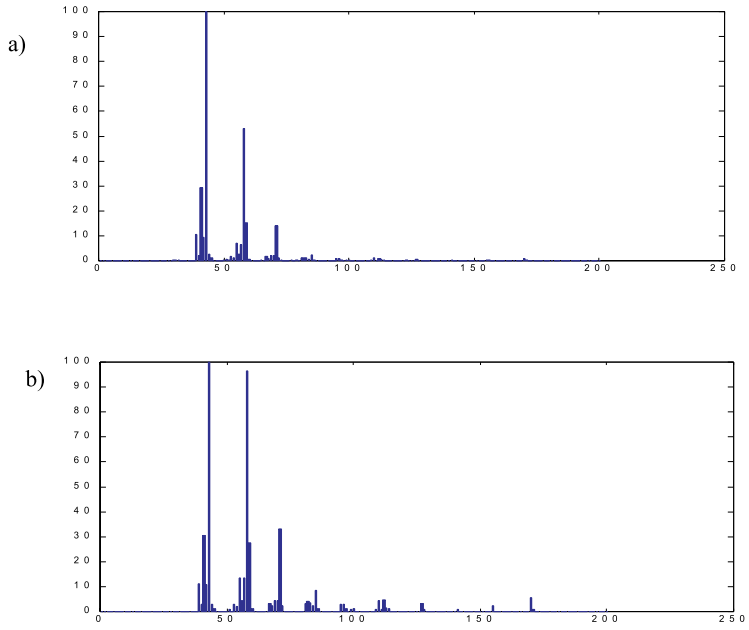

Figure 1. The mass spectra of 2-Undecanone. a) experimental spectra b) standard spectra in NIST library.

A=-1	if intensity I <1
A=0	if $1 < I \leq 3.4$
A=1	if $3.4 < I \leq 9$
A=2	if $9 < I \leq 19$
A=3	if $19 < I \leq 38$
A=4	if $38 < I \leq 73$
A=5	if $73 < I \leq 100$

Here the base peak is defined as 100 in each mass spectrum, then one can transform the data set according to the intensity of each peak. At the first glance, it seems that this transformation may cause the loss of significant amount of information contained in the data set. In fact, most of character of component is maintained. As Pesyna (Pesyna *et al.* 1975) report, in mass range 150-400, 295720 peaks of 6800 spectra have abundance $\geq 1\%$, half of

these peaks are found to be of abundance $\geq 3.4\%$ one-fourth of the 2975720 peaks has abundance $\geq 9.0\%$, one-eighth $(1/2^3) \geq 19\%$; $(1/2^4) \geq 38\%$; and $(1/2^5) \geq 73\%$. The statistical sense about this transformation can be shown as following example, the number of spectra having intensity value bigger than 73 is 2^{-5} of those having values bigger than 1 (Mclafferty and Stauffer 1985).

Such transformation is good for building an honest tree is natural. The high variance is regard as a major problem with tree method. Then one can alleviate it to some degree by trying to use a more stable split criterion or average many trees by Bagging (Breiman 1996), but the inherent instability is often not removed. In this paper, we reduce the variance of data set by above transformation. After such transformation, the intensity has almost same statistical occurrence will be represented by a same number. The quantitative validity of this transform approach well established for document retrieval from libraries by famous PBM (Probability based matching) algorithm (Mclafferty and Stauffer 1985).

3. Experiment

3.1 Data sets

Training data set: Our data set is selected from National Institute of Standards and Technology MS Database containing about the spectra of approximately 62000 component (NIST 62). Total 125 components and the molecular weight of 124 is selected in this study. Only 50 components contain one oxygen element in its formula and their common molecular formula is $C_8H_{12}O$. The other 75 components contain no oxygen element and their common molecular formula is C_9H_{16}.

Testing data set: The test sample is selected from National Institute of Standards and Technology MS Database containing about the spectra of approximately 107000 component (NIST 107). There are a total of 181 components. Only 87 components contain one oxygen element and their common molecular formula is $C_8H_{12}O$. Of the other 94 components, they contain no oxygen element and their common molecular formula is C_9H_{16}.

In the stability test, one can check the stability of tree by data pertur-

bation. We regard a stable tree as an honest tree. One can first deleted one sample at random at each test and observed the change in the structure of decision. Secondly, the heteroscedastic noise is added upon data set to perturb the data set. It is because the noise for mass spectra signal is often proportional to the intensity. The proportionality factor used here is 0.001.

The data analysis process was run in S-PLUS 2000 and executed on a Pentium III 850(Intel) personal computer with 128MB RAM under the Microsoft Windows 98 operating system.

3.2 The stability of tree

The trees built by the full original data set and the weighted data sets are shown in Figure 2. Then we check the stability of tree by delete one sample at random. From Table 1, one can see that the tree structure built by original data sets changed greatly because of perturbation of samples. If the variables actually used in tree construction are completely different, it is difficult or even impossible for us to interpret the model. However, the structure built by weighed data sets changed relatively little with the deleting one sample procedure described above. It demonstrates that weighting procedure is robust to the sample perturbation. This weighing procedure help us to build an stable, we name honest in this tree.

Since the mass spectra are often fluctuated by experiment errors. We also check the stability of tree by adding heteroscedastic noise with data sets.

Inspection of the results in Table 2 reveals that the weighted data set is relatively stable when the noise is added. However, the tree constructed by the original untransformed data show much variation. The above results demonstrate again that the weight process method is more robust in building the classification tree.

3.3 Prediction accuracy

Another 161 components selected from another standard library are used

Figure 2. The tree a) built by original data set. b) built by weighed data set.

to check the prediction accuracy of decision tree. Some component are identical in these two samples, but their mass spectra are not exactly same.

Table 1: Variables actually used in tree construction with one sample deleting

Original data set	Data set with weighting
69,55,106,39	69,67,54,106,106,39,67
69,55,106,39,83	69,67,54,106,106,39,67
69,82,81,106,39,67, 28	69,67,54,106,106,39,67,51,62
69,82,52,106,39,69	69,67,54,106,106,39,67
69,55,106,39	69,67,54,106,106,39,67
69,82,42,106,39,68	69,67,54,106,106,39,67
69,82,51,106,39,26	69,67,54,106,106,39,67
69,82,51,44,106	69,67,54,106, 39, 106,67
69,55,106,44	69,67,54,106,106,39,67
69,55,106,39	69,67,54,106,106,39,67

Table 2: Variables actually used in tree construction with added noise

Original data set	Data set with weighting
69,82,50,106,39,55, 26	69,67,54,106,106,39, 67
69,82,50,106,39,55,26	69,67,54,106,106,39,67
69,82,50,106,39,55,26	69,67,54,106,106,39,67
69,82,42,106,39,28	69,67,54,106,106,39,67
69,82,50,106,39,28	69,67,54,106,106,39,67
69,82,52,106,39,26,55	69,67,54,106,106,39,67
69,82,42,106,39,26	69,67,54,106,106,39,67
69,82,42,106,39,39	69,67,54,106,106,39,67
69,82,42,106,39,26	69,67,54,106,106,39,67
69,82,65,106,39,26	69,67,54,106,106,39,67

Table 3: The training and prediction results of mass spectra data

	Training(125 samples)	Prediction(177 samples)
Original data	88.0%	77.3%
Weighted data	88.7%	82.9 %

The results shown in Table 3 demonstrate the weight process also performs better in training and prediction accuracy. The above weighting process is reasonable since it preserves most of differences from components but not experiment variation. It is because that the intensities having almost same occurrence probabilities were transformed to a same number. Although some information will lost there are at least two advantage for this transformation. Firstly, much variation due to experiment noise is alleviated greatly, Secondly, reducing the redundant information from intensity is good for building an honest tree. The accuracy in intensity is less important than that in mass to charge ratio for library searching and structure elucidation task. Much information coming from intensity is redundant, it is impossible for researcher to get a completely same Mass spectra in different experiments. Therefore, the above transformation is necessary and reasonable.

4. Conclusion

The weighting approach has been proposed for building an honest tree for classification of mass spectra. Its performance was compared with that of the decision tree built by untransformed original data set. Our results show that the proposed method can yield simultaneously solutions with acceptable accuracy and stable partition. The insensitivity to the noise and sample change is of importance for its practical use in mass spectral classification.

Acknowledgment

The first and third authors were financially supported by the National Natural Science Foundation of the People's Republic of China (Grant No 20175036 and 20235020) and the second author was partially supported by Statistics Research and Consultancy Centre, Hong Kong Baptist University. The authors thank Prof. Kai-Tai Fang for his valuable suggestions.

References

Alsberg, B. K., Goodacre, R., Rowland, J. J. and Kell, D. B. (1997). Classification of Pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, K-nearest neighbour, neural and decision-tree methods. *Anal. Chim. Acta.*, **348**, 389-407.

Breiman, L. (1996) Bagging predictors, *Machine learning* **26**, 123-140.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont, California.

Eghbaldar, A., Forrest, T. P., Cambon, D. A. and Guigonis, J. M. (1996). Identification of structural features form mass spectrometry using a neural network approach: application to trimethylsilyl derivatives used for medical diagnosis. *J. Chem. Inf. Comput. Sci.*, **36**, 637-643.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Anals. of Statistics*, **19**, 1-67.

Grotch, S. L. (1969). Peak height distribution of organic mass spectra. In *17th Annual Conference*, Dallas, May 1969, American Society of Mass Spectrometry, pp 459-466.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning, Data Mining, Inference and Predication*, Springer, New York.

Lebedev, K. S. and Cabrol-Bass, D. (1998). New computer aided methods for revealing structural features of unknown compounds using low-resolution mass spectra. *J. Chem. Inf. Comput. Sci.*, **38**, 410-419.

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

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A. and Ledergerg, J. (1980). *Application of Artificial Intelligence for Grganic Chemistry: DENDRAL Project*, McGraw & Hill, New York.

Luinge, H. J. (1990). A knowledge-based system for structure analysis from infrared and mass spectral data. *Trends in analytical chemistry*, **9**, 66-69.

McLafferty, F. W. and Stauffer, D. A. (1985). Retrieval and interpretative computer programs for mass spectrometry. *J. Chem. Inf. Comput. Sci.*, **25**, 245-252.

Pesyna, G. M., Mclafferty F. W., Venkataraghavan R and Dayringer H. E. (1975) Statistical occurrence of mass and abundance values in mass spectra. *Anal. Chem.* **47**, 1161-1164.

Sokolow, S., Karnofsky, J. and Gustafson, P. (1978). The Finnigan library search program: Finnigan application report 2.

Varmuza K. and Werther W. (1996). Mass spectral classifiers for supporting systematic structure elucidation. *J. Chem. Inf. Comput. Sci.,* **36**, 323-333.

Received May 20, 2003; accepted August 2, 2003

Cheng-Jian Xu
College of Chemistry and Chemical Engineering
Institute of Chemometrics and Intelligent Analytical Instruments
Central South University
Changsha, China
xucj2000@263.net

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

Yi-Zeng Liang
College of Chemistry and Chemical Engineering
Institute of Chemometrics and Intelligent Analytical Instruments
Central South University
Changsha, China
yzliang@cs.hn.cn