第29卷，第3期
2009年3月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 29, No. 3, pp580-584
March, 2009

基于自适应共振神经网络的单粒子激光电离质谱数据分析

林 莺，郭晓勇，顾学军，夏玮玮，郑海洋，张为俊，方 黎

中国科学院安徽光学精密机械研究所环境光谱学研究室，安徽 合肥 230031

摘 要 气溶胶激光飞行时间质谱仪(ALUOFMS)可以在线地对气溶胶单粒子进行物理和化学特性分析，利用双束连续激光对单个粒子的空气动力学粒径进行测量，并通过飞行时间完成单粒子化学成分的检测。该仪器在运行过程中将产生海量的实验数据，对这些数据快速、自动处理并提取有价值的信息是整机系统的关键之一。文章介绍了基于神经网络的自适应共振算法(ART-2a)在随机混和的氯化钠、氯化钙、邻苯二甲酸二正辛酯(DOP)和2,5二羟基苯甲酸(DHB)气溶胶单粒子聚类分析中的成功运用。同以往的质谱分析方法相比，ART-2a可以实现对任意多和任意复杂的输入模式进行自组织，自适应和自稳定的快速识别，更有利于质谱数据的分析。实验结果表明，当警戒值为0.40，学习速率为0.05以及迭代次数为6时，ART-2a可以成功地对这四种物质进行分类，同时得到4类物质的聚类中心，每类的聚类中心都能很好的代表该类物质的特征。

关键词 光谱分析；单粒子测量；激光飞行时间质谱仪；激光解吸附电离；自适应共振神经网络

中图分类号：O434.1 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2009)03-0580-05

引 言

大气气溶胶在全球气候变化、全球和区域污染过程、能见度降低以及人类的身体健康方面扮演着重要角色[1-4]。研究气溶胶单粒子的粒径和化学成分将有助于解析气溶胶的来源和传输过程，以及在大气中的化学反应。气溶胶激光飞行时间质谱仪ALTOFMS(aerosol later time-of-flight mass spectrometer)是近年来发展起来的一个强有力的气溶胶实时在线检测工具，该仪器将激光解吸附电离技术与飞行时间质谱技术相结合，把粒子的质量分辨方法引入到光谱学中，构成了气溶胶激光飞行时间质谱检测技术[5,6]。通常，ALTOFMS每分钟可以采集上百个质谱数据，连续工作一天所获得的数据量将超过十万。对于这些海量实验数据的处理和分析，采用传统的手工方法是不可想象的。因此研究自动、高效的数据分析方法是十分必要的。

目前已用于质谱数据分析的方法有：等级聚类分析(HCA)[7-9]、主成分分析(PCA)[7,10]、模糊C均值(FCM)[8,9]、自组织特征映射神经网络(SOM)[11,12]和自适应共振算法(ART-2a)[13]。其中ART-2a神经网络是一种无监督的矢量分类器，能有效地处理大数据集和高维数据集，并且当某个数据点与当前存在的所有分类都没有达到预设的接近度时，ART-2a将自动为其产生一个新类，而不影响其他已存在的类，十分适合于质谱数据的聚类分析。本文中利用ART-2a成功地对随机混和的多个氯化钠、氯化钙、邻苯二甲酸二正辛酯(dioctlphthalate，DOP)和2,5二羟基苯甲酸(2,5-dihydroxybenzoic acid，DHB)气溶胶单粒子质谱数据进行了聚类分析，研究了不同警戒值、学习速率和迭代次数对聚类结果的影响。

1 实验部分

1.1 实验

我们把实验室环境下产生的NaCl，$CaCl_2$，DOP和DHB气溶胶单粒子的质谱作为样本数据，采用ART-2a算法进行聚类分析。

本实验中样本数据的采集是在我们自行研制的气溶胶激光飞行时间质谱仪上完成的，实验装置见文献[14]。气溶胶粒子产生方法如下：NaCl和$CaCl_2$溶于去离子水，DOP和DHB溶于异丙醇中形成溶液，将上述溶液分别装入德维尔比斯(Devilbiss40#，D40)玻璃喷雾器中，用氮气作为载气，产生气溶胶粒子。形成的气溶胶粒子用容积为10 L的广口瓶收集，通过塑料软管直接与ALTOFMS相连。由于质谱仪内的负压，粒子被自动吸入到ALTOFMS中，当某一单粒子

收稿日期：2007-11-26，**修订日期**：2008-03-06

基金项目：国家自然科学基金项目(20477043)资助

作者简介：林 莺，女，1981年生，中国科学院安徽光学精密机械研究所硕士研究生 e-mail：linying6092@126.com

依次通过粒径测量区内两个相距 70 mm 的波长为 532 nm 激光束时，产生散射光信号。通过计算这两个散射光脉冲的时间差可得到粒子的空气动力学直径。该粒子继续下行到达焦点区时被波长为 266 nm 的 Nd：YAG 激光解吸附电离，从而得到粒子的激光飞行时间质谱。

1.2 数据预处理

任意选取 NaCl，$CaCl_2$，DOP 和 DHB 有效质谱各 100 个。图 1(a)显示的是某个 DOP 气溶胶粒子以飞行时间为横坐标的原始质谱图。每个单粒子质谱包含 4 096 个 8 位数据点。在对质谱数据进行一系列的自动处理(包括数字信号处理、自动基线修正、谱峰识别算法及飞行时间转换质荷比的标定操作等)后，其结果是使每一个粒子质谱列表中包含所有质谱峰的峰强和精确的质荷比值，图 1(b)显示的是该 DOP 气溶胶粒子质谱经预处理后以质荷比为横坐标的质谱图。处理后的质谱图可以表示为 300 维的数据向量，每一维代表一个质量单元，其数值为质谱峰的强度。这样就可以把一个粒子集的质谱转换成包含它们的数据向量的数据矩阵，在矩阵中每一个质谱数据被存储为矩阵中的一行。

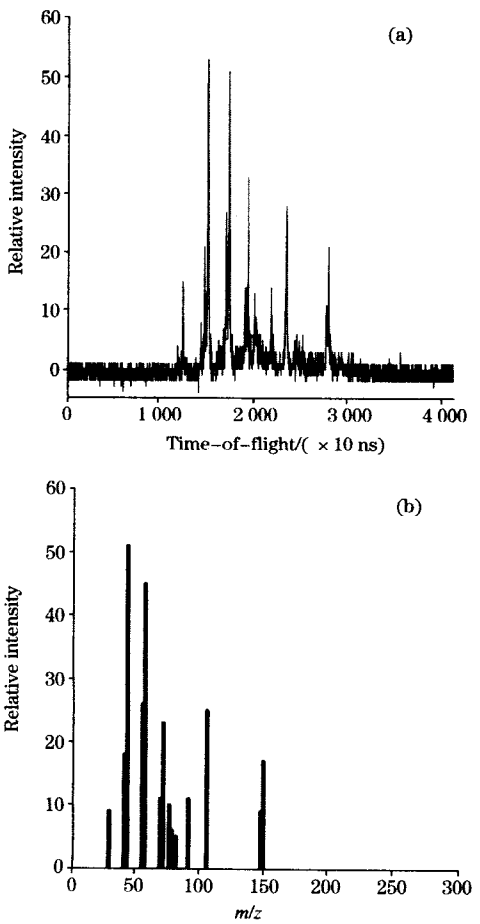

Fig. 1 (a) time-of-flight mass spectrum of DOP and (b) mass spectrum of DOP after pretreatment

1.3 基于神经网络的自适应共振理论

自适应共振理论(adaptive resonance theory，ART)[15, 16]是 Stephen grossberg 和 Carpenter 在 1976 年提出的，其构建了一种竞争型的神经网络，可进行无监督的样本学习。该理论主要思想如下：人脑在处理外界信息时能自动从各种信息中选择对自己最有影响或最感兴趣的信息(即所谓发生共振)，且在接收新的信息时不破坏原有的存储信息。人脑既可以被动地学习(例如应付环境强加给人的、突如其来的事件)，又能主动地有选择地学习(例如集中注意某方面信息，对其他无关信息充耳不闻，视而不见)。既能牢牢地记住旧知识，又能随时接受新信息，达到稳定性与可塑性的完美统一。该理论即以生物学、心理学、认识行为各方面的事实为依据，又进行了精确的数学分析，模拟了人脑的上述功能。根据 ART 理论设计的各种神经网络称为 ART 网，也称 Grossberg 网。

ART-2a 神经网络[13]是 Grossberg 和 Carpenter 等在 1991 年提出的。与其他的 ART 网比起来，该网络在牺牲少量学习精确度的前提下，具有算法简单、计算快速、以及空间复杂度低的优点。ART-2a 算法的核心是权值矩阵 $\boldsymbol{W}(m\times k)$ 的动态形成。其中，m 是输入向量即每个质谱数据的长度，k 是输入向量的个数。权值矩阵的每一列称为权值向量，每个权值向量都对应一个类的聚类中心，用来代表该类。具体算法如下。

(1)初始化网络参数：设置对比增强阈值 θ，$0<\theta<1/m$；设置学习速率 β，$0\leqslant\beta\leqslant1$；设置权向量的初值 α，$0\leqslant\alpha\leqslant1/\sqrt{m}$；设置警戒阈值 ρ，$0\leqslant\rho\leqslant1$。

(2)从数据矩阵 X 中随机选择一个输入向量 x_i。

(3)规格化输入向量

$$p_i=\frac{x_i}{\|x_i\|} \tag{1}$$

(4)对比增强

$$q_{ik}=\begin{cases}p_{ik}, & \text{if } p_{ik}>\theta\\ 0, & \text{otherwuse}\end{cases} \tag{2}$$

(5)重新规格化

$$r_i=\frac{q_i}{\|q_i\|} \tag{3}$$

(6)计算各神经元节点的输入

$$T_j=\begin{cases}\alpha\times\sum_{k=1}^{m}r_{ik} & \text{if } j \text{ is a uncommited node}\\ r_i\times w_j & \text{if } j \text{ is a commited node}\end{cases} \tag{4}$$

(7)竞争选择：选出输入最大的节点 J 作为获胜者

$$T_j=\max j(T_j) \tag{5}$$

(8)共振或重置：如果 j 是新节点(uncommited)，或者是已生成节点(commited)并且满足 $T_j\geqslant\rho$，则保持 J 不变；否则如果 J 是已生成节点并且 $T_j<\rho$，则将重 J 置为一个新的节点。

(9)学习：更新获胜节点 J 的连接矩阵

$$w_j^{\text{new}}=\begin{cases}r_i & \text{if } j \text{ is a uncommited node}\\ S_j & \text{if } j \text{ is a commited node, Eqs. (7)} \sim \text{Eqs. (10)}\end{cases} \tag{6}$$

$$S_J=\frac{t_J}{\|t_J\|} \tag{7}$$

$$t_J=u_J+(1-\beta)\times w_J^{\text{old}} \tag{8}$$

$$u_j = \beta \times \frac{v_i}{\| v_i \|} \tag{9}$$

$$v_{kj} = \begin{cases} r_{ik}, & w_{kj}^{old} > \theta \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

ART-2a 通过设置警戒阈值 ρ，实现类似的样本归类。ρ 可以表示两个样本有多相似才被认为是匹配的，因此，通过改变警戒阈值可以调整模式的类数，ρ 小则模式类别少，反之亦然。如果一个输入向量 x 与某个权值向量 w 形成的角度比 $\arccos(\rho)$ 小，那么就说 x 和 w 发生共振，并且 x 属于 w 所对应的类。否则就认为 x 是一个新类，新类别所对应的权值向量被增加到原来的权值矩阵中。

对于无导师学习网络，输入新数据将会对某个权值向量即聚类中心进行修改，这种修改意味着对新知识的学习和对旧知识的忘却。ART-2a 算法通过学习速率 β 来决定对新知识的学习比例。

若数据矩阵中所有的输入向量 x 都进行了一次归类运算，则记为一次迭代。当这个算法迭代了一定多的次数后，权值矩阵的值改变的非常小，或者对于整个数据集的分类结果固定不变，那么整个分类过程结束。

2 结果与讨论

把 NaCl，$CaCl_2$，DOP 和 DHB 四类物质的共 400 个质谱向量随机打乱顺序，依次输入到 ART-2a 算法中进行运算归类。警戒阈值、学习速率以及迭代次数是运算之前需要确定的经验参数。通过调节警戒阈值，可以控制分类的数目。极端的情况，就是或者分成 400 类，每类只有一个质谱数据，或者把所有的质谱数据分成一类。学习速率的大小可以影响算法的收敛速度，同时也影响到迭代次数的设置。因此需要综合考虑这几个初始参数的大小设置，使分类最优。

若一个质谱在连续两次迭代中被划分到不同的类别，这样的质谱的个数就称为分组改变数。图 2 显示的是在不同的学习速率下(学习速率从 0.05 到 0.95，每隔 0.1 取一个值)，迭代次数与分组改变数的关系。由图 2 可见，对于本次实验数据，在不同的学习速率下，经过 5 次以上的迭代后，分组改变数均为 0，即都达到稳定的分类。所以说对于这些数据，选取不同的学习速率对迭代次数的设置影响较小。根据经验以及参考文献[17]，本实验中学习速率 β 取 0.05，迭代次数取为 6 次。

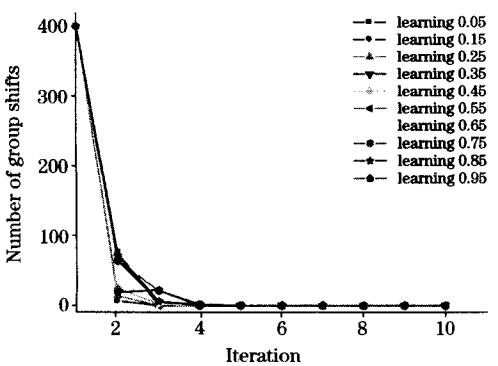

Fig. 2 Number of group shifts as a function of iteration for ten learning rates

图 3 是分类的类别数与警戒阈值(从 0.1 到 0.9，每隔 0.1 取一个值)的关系图。学习速率为 0.05，迭代次数为 6。算法运行 4 次，每次除了质谱向量的输入顺序(每次运行，输入顺序由计算机随机产生)不同外，其他运行条件都相同，得到如图所示 4 条折线。由图 3 可见当警戒阈值为 0.4 时，分类数为 4 或 5 类，符合本实验的样品数。

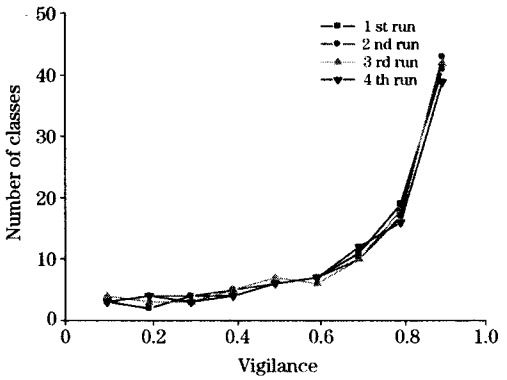

Fig. 3 Number of classes as a function of vigilance for four runs

根据上述分析，警戒阈值 ρ 取 0.4，学习速率 β 取 0.05，迭代次数取 6 次。利用 ART-2a 算法对 400 个气溶胶粒子进行了聚类分析，共得到 4 个分类。每个类别都有一个聚类中心。图 4 显示了这 4 个聚类结果。各聚类的粒子个数、主要质谱峰和物质类型如表 1 所示。

从图 4 的一些特征质谱峰可以判断各聚类对应的物质。第 1 类聚类中心有 $m/z=149(C_6H_4(CO)_2OH^+)$ 的质谱峰，它是 DOP 裂解的特征质谱峰，此外还有一些有机裂解碎片，其质量数分别为 $43(C_3H_7^+)$，$57(C_4H_9^+)$，$71(C_5H_{11}^+)$ 和 $106(C_6H_6CO^+)$，可以断定是 DOP 颗粒。第 2 类是 NaCl 颗粒，这类颗粒的聚类中心在 $m/z=23$ 和 81 有显著强的 Na^+ 和 Na_2Cl^+ 离子信号。第 3 类聚类中心在 $m/z=137$ 和 46 有强的 DHB-OH^+ 和 $HCOOH^+$ 质谱峰，它是 DHB 裂解的特征质谱峰，表明它是 DHB 颗粒。第 4 类是 $CaCl_2$ 颗粒，它的聚类中心在 $m/z=40$ 和 75 有显著强的 Ca^+ 和 $CaCl^+$ 离子信号。表 1 是 4 个聚类信息的概述。从表 1 可以看出，第 1 类和第 3 类的粒子个数为 100，等同于起始的 DOP 和 DHB 粒子的个数。而第 2 类的 NaCl 粒子个数为 101，第 4 类的 $CaCl_2$ 粒子个数为 99。第 2 类除了 100 个 NaCl 粒子，还混杂了一个 $CaCl_2$ 粒子，该粒子对应的质谱图如图 5 所示。从图中可以看出这个粒子不同于其他的 $CaCl_2$ 质谱，它有一个很强的质量数为 23 的峰，相似于 NaCl 类中的 Na^+ 峰，这可能是由样品中的杂质产生的。该粒子的质谱数据在 ART-2a 聚类运算中与 NaCl 泪的聚类中心发生共振，所以被错分到 NaCl 类中。综上所述，当警戒值为 0.40，学习速率为 0.05 以及迭代次数为 6 时，ART-2a 可以成功聚类出这四种物质。

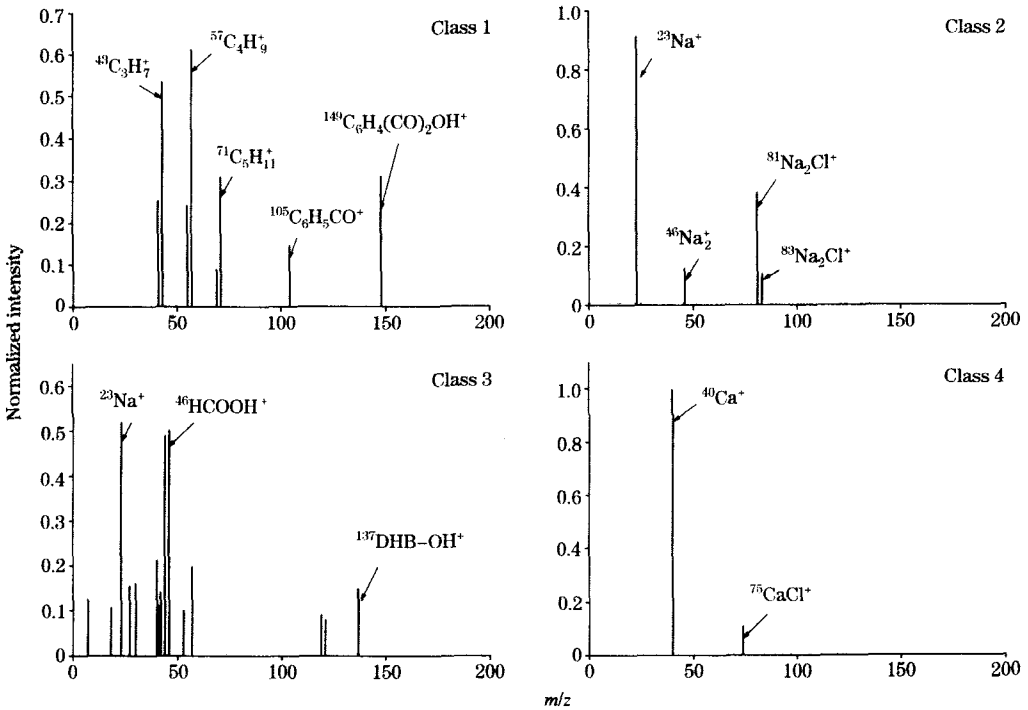

Fig. 4　Normalized weight vectors for 4 classes as determined by ART-2a: Class 1 for DOP, Class 2 for NaCl, Class 3 for DHB and Class 4 for $CaCl_2$

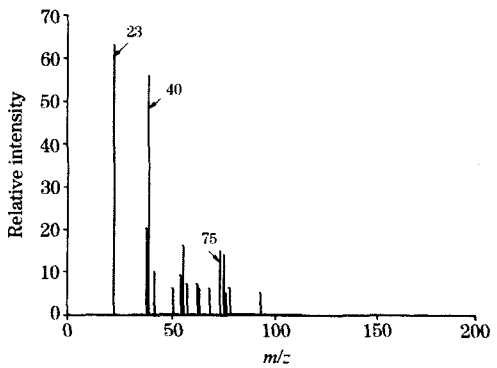

Fig. 5　Mass spectrum of $CaCl_2$ mixed in NaCl Class

Table 1　Summary of classes identified by ART-2a

类别号	每类粒子个数	主要质谱峰 (Da, 按峰强大小排序)	物质类型
1	100	57, 43, 149, 71, 41, 55, 106, 69	DOP
2	101	23, 81, 46, 83	NaCl
3	100	23, 46, 44, 40, 57, 30, 27, 137, 42	DHB
4	99	40, 75	$CaCl_2$

3　结　论

本文使用神经网络 ART-2a 算法实现了对气溶胶激光飞行时间质谱数据的分类。运用该算法对随机混合的 NaCl, $CaCl_2$, DOP 和 DHB 气溶胶单粒子进行了聚类分析，很好地区分了这 4 种不同的物质，并且得到了 4 个气溶胶粒子飞行时间质谱数据的聚类中心，每类的聚类中心都能很好的代表该类物质特征。由于 ART-2a 神经网络具有快速、无监督以及可根据环境变化自适应地产生出新类的优点，因此可将 ART-2a 应用于大量、复杂的大气气溶胶数据在线和离线的聚类分析，将在以后的文章中陆续报道。

参　考　文　献

[1] Christopher A Noble, Kimberley A Prother. Environmental Science and Technology, 1996, 30: 2667.
[2] Moren F, Dolovich M B, Newhouse M T, et al. Aerosols in Medicine: Principles, Diagnosis and Therapy, 2nd ed., Amsterdam: Elsevier, 1993. 321.
[3] Dockery D W, Pope C A. Annu. Rev. Public Health, 1994, 15: 107.
[4] LIAN Yue, LIU Wen-qing, LU Jian-chun, et al(连　悦, 刘文清, 鹿建春, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(2): 198.
[5] Prather K A, Nordmeyer T, Salt K. Analytical Chemistry, 1994, 66: 1403.
[6] Salt K, Noble C A, Prather K A. Analytical Chemistry, 1996, 68: 230.

[7] Xhoffer C, Bernard P, Van Grieken R, et al. Environ. Sci. Technol., 1991, 25: 1470.
[8] Bondarenko I, Treiger B, Van Grieken R, et al. Spectrochim. Acta Part B, 1996, 51: 441.
[9] Treiger B, Bondarenko I, Van Malderen H, et al. Analytical Chim. Acta, 1995, 317: 33.
[10] Hinz K P, Kaufmann R, Spengler B. Aerosol Sci. Technology, 1996, 24: 233.
[11] Kohonen T. Proceedings of the IEEE, 1990, 78(9): 1464.
[12] Kohonen T. Self-Organizing Maps, Berlin, 1995, 2nd. Edition, 1997.
[13] Carpenter G A, Grossberg S, Rosen D B. Neural Networks, 1991, 4: 493.
[14] GUO Xiao-yong, ZHAO Wen-wu, LIN Ying, et al(郭晓勇, 赵文武, 林 莺, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2008, 28(8): 1919.
[15] Grossberg S. Biological Cybernetics, 1976, 23(3): 121.
[16] Grossberg S. Biological Cybernetics, 1976, 23(4): 187.
[17] Denis J P, Kevin P R, Anthony S W, et al. Anal. Chem., 2001, 73: 2338.

Data Analysis of Laser Desorption/Ionization Mass Spectrum of Individual Particle Using Adaptive Resonance Theory Based Neural Network

LIN Ying, GUO Xiao-yong, GU Xue-jun, XIA Wei-wei, ZHENG Hai-yang, ZHANG Wei-jun, FANG Li
Lab of Environmental Spectroscopy, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, China

Abstract On-line measurement of size and chemical composition of single particle using an aerosol laser time-of-flight mass spectrometer (ALTOFMS) was designed in our lab. Each particle's aerodynamic diameter is determined by measuring the delay time between two continuous-wave lasers operating at 650 nm. A Nd : YAG laser desorbs and ionizes molecules from the particle, and the time-of-flight mass spectrometer collects a mass spectrum of the generated ions. Then the composition of single particle is obtained. ALTOFMS generates large amount of data during the process period. How to process these data quickly and extract valuable information is one of the key problems for the ALTOFMS. In the present paper, an adaptive resonance theory-based neural network, ART-2a algorithm, was used to classify mixed mass spectra of aerosol particles of NaCl, CaCl2, dioctylphthalate (DOP), and 2,5-dihydroxybenzoic acid (DHB). Compared with the traditional methods, ART-2a can recognize input patterns self-organically, self-adaptively and self-steadily without considering the complexity and the number of the patterns, so it is more favorable for the analysis of the mass spectra data. Experimental results show that when vigilance parameter is 0.40, learning rate is 0.05 and iteration number is 6, ART-2a algorithm can successfully reveal these four particle categories. The weight vectors for these four particle classes were obtained, which can represent the characters of these four particle classes remarkably.

Keywords Spectral analysis; Individual particles measurement; Laser time-of-flight mass spectrometer; Laser desorption/ionization; Adaptive resonance network

(Received Nov. 26, 2007; accepted Mar. 6, 2008)