自组织映像在质谱分析中的应用

贺平
（香港浸会大学数学系　香港）

摘　　要

随着质谱仪的问世和发展， 人们得到了大量的有价值的质谱数据，如何从这些质谱数据中挖掘关于化学结构或化学性质的信息？如何通过质谱分析，自动地鉴别化合物和识别化学结构？各种数据挖掘的方法已被用于解决这些问题，主要包括检索和分类两类方法。但它们在用于质谱分析时都有各自的缺陷。针对这些缺点我们提出先利用自组织映像的方法对质谱进行聚类和可视化，获得一些关于质谱所蕴含的化学结构的初步信息，在进一步的作研究。例如将聚类得到的类所反映的子结构作为分类时的响应变量有的时候比人为定的子结构作响应变量更合理。自组织映像法的目的是一个将高维数据非线性的投到一个预先定义好的二维拓扑中。它通过竞争学习的方法达到了降维，聚类，可视化的目的。

关键字 质谱分析， 自组织映像法，聚类

1：引言

通过对质谱数据的分析，自动地鉴别化合物以及识别化合物的结构属性一直是化学计量学中一个重要的任务。质谱分析是先将物质离子化，按离子的质荷比(m/z)分离，然后测量各种离子谱峰的强度而实现分析目的的一种分析方法。被测的物质通过质谱仪会得到相应的质谱图，它是质荷比(m/z)对谱峰强条形图。例如图 1 是乙醛(C_2H_4O)的质谱图。不同的物质有不同的质谱，利用这一性质，可以进行定性的分析。化学家已经从质谱中发现了关于分子量的大量信息以及一些关于分子结构或子结构的信息。

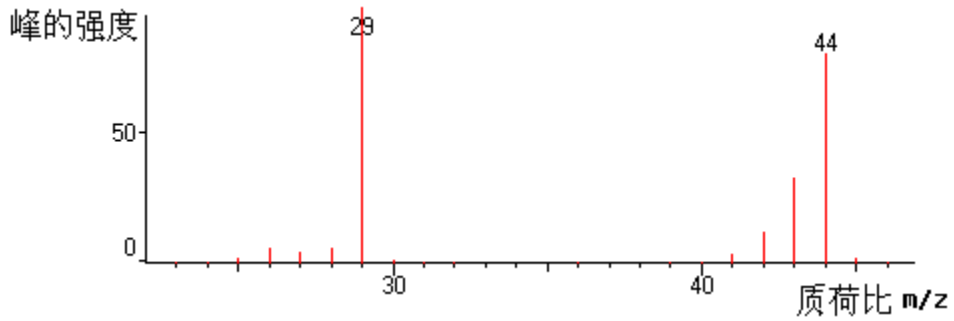


图 1：C_2H_4O 质谱图

质谱仪的问世带来了大量有价值的数据。目前，NIST 98 质谱数据库收集了大约 100,000 化合物的化学结构以及它们相应的质谱；Wiley/NBS 库含有超过

130,000 化合物的化学结构和质谱。(NIST 98 MS Database and Wiley/NBS collection)。如何发掘这些质谱中包含的化学结构的信息？如何建立一个通过质谱就可以自动地鉴别化合物及其化学结构的系统？这些仍然是化学计量学中重要的研究领域并吸引了越来越多的计算机以及统计行业的研究工作者。

各种各样的数据挖掘的方法已被用于质谱分析。目前最普遍的有两类：检索和分类的方法。许多基于统计中相似性比较的检索方法已成功的用于某些化合物的鉴定（Mclafferty et al. 1994 和 Stein et al. 1994）。然而这类方法的成功受到很大的限制：只有当被检测的化合物在检索数据库中，它才可能得到比较好的结果。目前可得到的数据库虽然包含了大量数据，大约 100,000 或 130,000 化合物，但比起自然界中超过 20,000,000 化合物还是很有限的。被检测的化合物的质谱很可能与参考数据库中的所有质谱都差别很大，使得 基于相似性比较的检索方法失去效用。分类方法则可以弥补这一缺陷。它通过对已知数据库的学习，建立一系列的能够通过质谱自动识别化学子结构的分类器。将未知化合物的质谱用于这一系列的分类器，就大致可判断出这些化合物具有哪些子结构以及不具有哪些子结构。而且分类方法比检索的方法大大减少了计算量，分类的方法每次只要将要识别的质谱用于已建立好的分类器就可以了，而不需要像检索的方法一样每次都要将质谱与所有数据库中的质谱比较，当质谱库中的数据量比较大时就需要很大的计算量。但质谱分析中用分类方法也有缺陷：它太依赖于子结构选择的好坏。在训练分类器时某种子结构的有无一般作为响应变量，这种子结构是人为从化学结构中选出来的，虽然这些子结构代表了一定的化学含义但是有可能用来训练分类器的质谱并不反映这种子结构的信息而反映了另一种子结构或另一类化学性质。这种现象的产生是因为人们也不确实地知道质谱中到底包含了化学结构中的哪些信息，而是利用原有的经验和化学意义抽取了一些子结构作为响应变量。目前还有很多的子结构不能被很好的识别有可能就是质谱并不包含这些子结构的信息而实质上反映了一些没有用在这里的子结构和化学特性的信息。用分类方法的另一缺点是它忽略了子结构之间的相互作用会对质谱产生影响。一个分类器被用来识别某一种子结构的有无，另一种分类器被用来识别另一种子结构的有无，子结构之间的关系被完全分离。基于以上两点我们利用自组织映像的方法对质谱数据进行聚类和可视化得到对质谱的初步认识然后再作近一步的分析。通过聚类可以大致看到训练样本的质谱中大致含哪些子结构的信息，并且因为自组织的映像通过临近神经元的互相刺激的作用体现了类（子结构）之间的相互作用。本文第二部分介绍自组织映像的基本想法和算法，第三部分将自组织映像用于质谱数据并讨论其结果。

2： 自组织映像的基本算法

自组织映像是由 kohonen 提出的一种神经网络的学习方法（kohonen 1984；kohonen,1995b;kohonen,1995c;kohonenn,et al 1996b）。这种方法已经被广泛的用于高维数据的聚类和可视化。它既属于聚类的方法也属于非线性投影的方法。它通过将样本投影到一个事先定义好的二维拓扑中达到聚类和可视化的目的。它是 kohoneng 依据大脑对信号处理的特点提出的一种竞争学习型的神经网络。它分为竞争和学习两部分：在竞争阶段，网络中的所有神经元都接受同样的样本，并参与

竞争。再根据某个准则找出最有活性的“胜利者”。学习阶段主要是基于胜利的神经元，获胜的神经元以及它邻近的神经元都要以某种方式向这个输入样本学习。下面就介绍一下自组织映像基本算法

自组织映像法要预先设定一个二维的神经元排列，每一个神经元 $\mathbf{m_i}$ 是一个向量被称为权重矢量

$$\mathbf{m_i} = (m_{i1}, m_{i2} \ldots m_{in})$$

它是与输入样本同维的矢量（n 维）。并且这些神经元通过矩形的或菱形的拓扑互相连接，神经元之间的距离也可以通过拓扑的关系来定义。如图 2，拓扑关系用神经元之间的线表示。

在基本的自组织映像法中拓扑的关系和神经元的个数是在算法一开始就选定的。神经元的个数决定了模型的大小而模型的大小又决定了模型预测的能力所以在定神经元个数的时候必须考虑到模型训练的准确性和预测能力两方面的因素。

首先要对预先设定好的神经元付初值，初值可以是随机选择的一组向量，它适用于对输入数据一无所知的情况；初值也可以是输入数据中原始的样本，这样做的好处在于神经元所代表的点自动的与数据落在同一输入空间；初值也可以是原始变量的一些现性组合例如用主成分分析得到的主要成分，这有助于将自组织映像图向含信息量最大的方向拉。

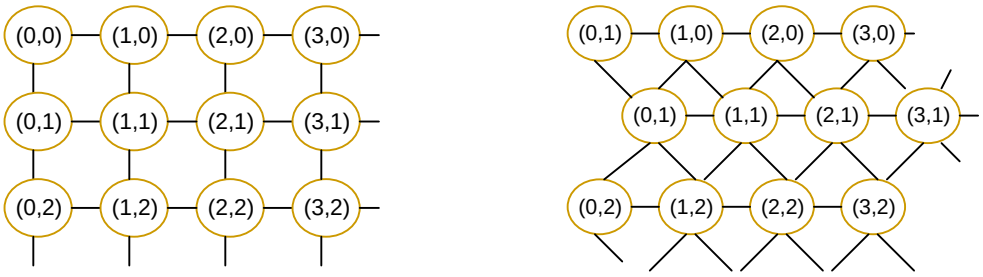


图 2　矩形拓扑　　菱形拓扑

给定二维排列的神经元和各神经元的初值，就进入到自组织映像的训练（竞争学习）阶段。训练是一个随着时间的迭代过程。当一个输入样本向量进入时，最能表现这个样本向量的神经元将在竞争中获胜并向这个样本向量更好的学习，而且这个获胜神经元的邻近的神经元也被允许学习，这就使得邻近的神经元会逐渐的特定的表现相似的输入量。

在训练阶段，一样本向量 $\mathbf{x}$ 从输入空间中随机抽出，并基于某个相似性尺度计算这个样本向量与网络中所有单元的相似程度（称每个带权重矢量 $\mathbf{m_i}$ 的神经元为单元）。其中与输入样本向量最相近的单元为竞争中的获胜者，记为 C:

$$c = c(\mathbf{x}) = \arg\min_i \{\| \mathbf{x} - \mathbf{m_i} \|^2\}$$

通常这个相似尺度是基于距离的，例如欧几里的距离

$$\| x - m_i \| = \sqrt{\sum_{j=1}^{n} (x_j - m_{ij})^2}$$

找到获胜元后，获胜元和它邻近的单元的权重向量都会自动更新使得更新后的权重矢量更接近输入变量。各个单元的学习量是由一个邻域核函数 h 控制的。这个核函数是关于各神经元与获胜元距离的递减函数。单元 $\mathbf{m_i}$ 的更新准则如下：

$$\mathbf{m_i}(t+1) = \mathbf{m_i}(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m_i}(t)]$$

这里 t 表示时间，就向上面提到的训练过程是一个随时间的迭代过程。x(t)是在 t 时刻从输入样本中抽出的一输入向量。h_{ci} 是围绕获胜元 m_c 的一非增邻域函数它包含一个时间的递减函数学习率α(t)。各种邻域函数已被用于自组织映像，例如围绕获胜元 m_c 的高斯邻域函数为

$$h_{ci}(t) = \alpha(t) \cdot \exp(-\frac{\| r_i - r_c \|^2}{2\sigma(t)^2})$$

学习率是一个时间的递减函数，两种形式的函数通常被用到：一是时间的线性函数，

$$\alpha(t) = \alpha(0)(1 - t/rlen)$$

另一个是时间的反比例函数。

$$\alpha(t) = C\alpha(0)/(C + t)$$

这里 rlen 是训练中的总的运行时间，C 是预先定义好的常数。时间的线性函数是从 1 线性的减到 0 而反比例函数则减少的很快，表明模型在前期就建立的比较好，后期不用学习或改动太多。

重复这个竞争学习的过程，神经网络中的神经元就会逐渐的对样本中不同的类或样本中一些特定区域变得越来越敏感从而达到聚类和可视化的效果

3：自组织映像在质谱数据中的应用：

在这一节里我们提出用自组织映像法对质谱数据进行聚类。因为有可能质谱并不包含某些现有的人们认为质谱能够体现的子结构的信息，如果一定要根据这类子结构用质谱分类是不合理的。其次质谱数据都是高维数据，通常样本物质的分子量有多大，样本就有多少维的变量。而且化学家认为这些子结构相互影响应该在质谱中有所体现。针对这种高维的数据，并且类与类存在彼此影响的关系，我们选择自组织映像法。因为自组织映像法是一个将高维降到两维的非线性投影的方法，同时实现在二维拓扑图上的聚类和可视化。并且它的竞争学习的过程中相邻的神经元也会跟着学习正好体现了类之间彼此存在着一定的关系。

实验数据：醇和醚是同分异构体，从质谱库 NIST62.LIB(V1.0P/N 225-01860-93)中抽取所有的分子式从 $C_5H_{12}O$ 到 $C_9H_{20}O$ 的这两类化和物，其中醇有 148 个醚有 53 个。我们将质荷比从 1 到 144 所对应的峰强度作为输入变量（注：$C_9H_{20}O$ 的分子量为 144）。

实验过程：选定的二维拓扑结构为菱形拓扑。其中神经元的按 10 行 7 列排列共 70 个神经元。神经元的权重矢量的初始值为原始输入变量的线性组合。经过自组织映像的训练，我们可得到 70 个神经元的输出权重矢量。

实验结果：为了可视化结果，我们选用 U 矩阵表现神经元之间的距离：相邻的神经元之间的距离被计算，并用不同的颜色表示这些距离。颜色越深表是神经元所表示矢量权重之间的距离越远，浅的颜色则表示了两个神经元之间的距离近。所以颜色浅的区域可以看成是类而深的区域则是孤立点。同时我们还计算了这些神经元所代表的是醇还是醚。如下图：

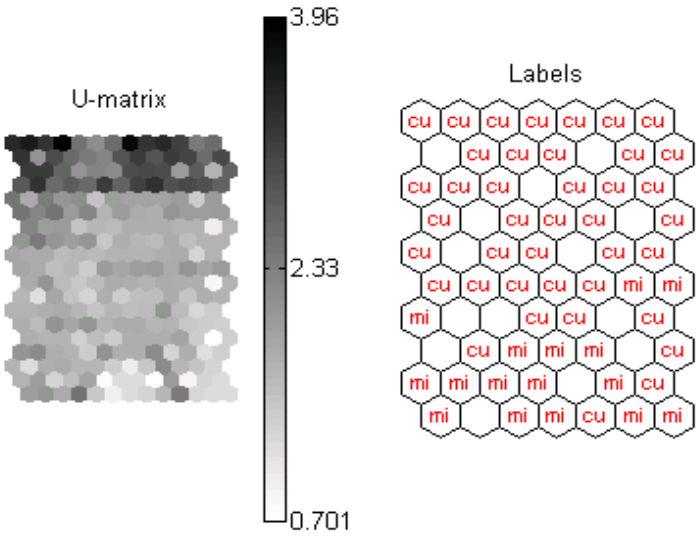


图 3 自组织映像的 U 矩阵图和标志图

从 U 矩阵图可以看出偏下的神经元距离较近可以分为一类，偏上的神经元之间距离较远。从右边的图可以看出大致可以分为醇醚两类但在神经元的右下角部分还是有混杂。结合 U 矩阵图可知各种醇之间的距离要大些，各种醚之间的距离要小些。为了更清楚地表示神经元和醇与醚分类之间的关系，我们给出了“ 碰撞 ”(hit)图。它是一个计数图：找出样本的胜利神经元，并给该神经元加一分。对所有样本做这个计数过程，得‘ 碰撞 ’图。图中附颜色的菱形的大小表示对各神经元计数的大小。绿色表示醇，红色表示醚。可以看到右下角可以分为醇醚两类，但这两类的距离比较近。从分子结构看对应于这块的醇和醚结构比较简单。这里的醇全部为带一个或者两个支链的醇，醚为直链或者带一个支链的结构。在图的左下角和右中方醇和醚有混杂的现象，我们发现都是结构比较简单的醚与带一个十字结构的支链的醇混在一起。而结构更复杂的醇和醚又会分开。所以我们在做醇醚分类时，预先对醇醚做更小的划分可能会比较合理即进行局部分类。

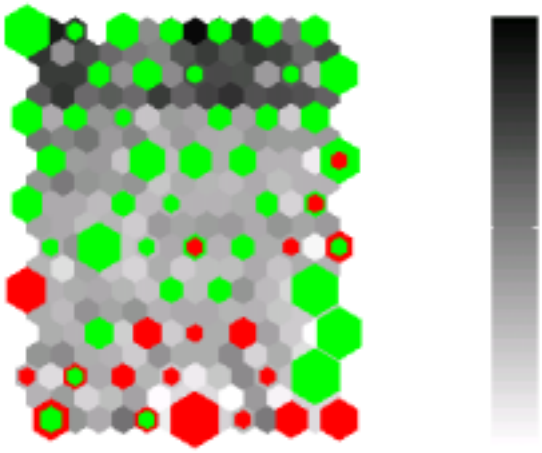
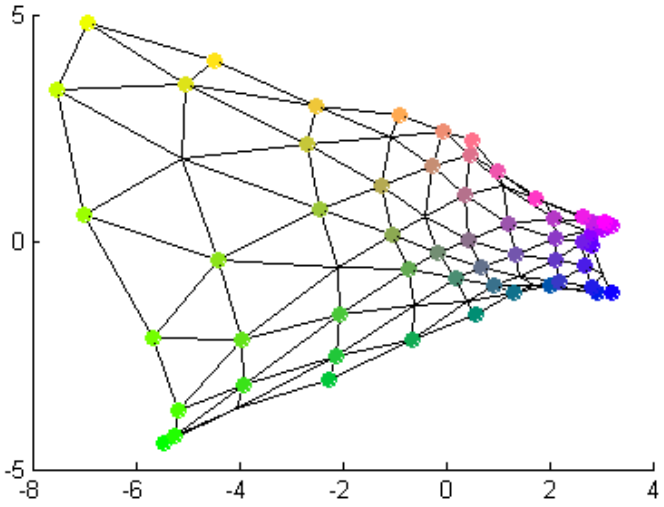


图 4 自组织映像的碰撞图

我们还将所得到的神经元权重矢量投到由原始数据的得到两个主要成分上（用主成分分析法得
到）。得到下图：



点表示神经元在二维主成分图上的位置，线表示他们的距离，点的大小代表了神经元的碰撞计数，颜色代表大致的类。可以看出它反映的信息与图 4 大致相同。

4：总结

这篇文章主要说明了我们为什么将自组织映像的方法应用于质谱分析并介绍了自组织映像的基本算法。这项工作的目的在于想尝试性的找出质谱数据中真正含有关于哪些子结构或化学性质的信息。在以后的工作中我们将把自组织映像法用于更多的实际质谱数据中，并且利用自组织映像法中可视化的一些方法对所得到的类的信息进行近一步的说明，看是否可以给出其化学含义的解释。

参考文献

Mclafferty, F.W., Hertel, R.H.(1994) Probability based matching of mass spectra. *Org. Mass Spectrum*. **8,** 690-702

NIST (1992). *NIST Mass Spectral Database.* National Institute of Standards and Technology, Gaithersburg MD 20899, USA.

NIST (1998). *NIST'98 Mass Spectral Database*. National Institute of Standards and Technology, Gaithersburg MD 20899, USA.

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

Kohonen T. (1984) Self-Organization and Associative Memory ,Sprinnger

Kohonen T. (1995b) Emergence of invariant-feature detectors in self-organization. In Palaniswami, M., Attikiouzel, Y. Marks II, R. J. Fogel, D., and Fukuda, T., editors, *Computational intelligence. A dynamic system perspective,* 17-31. IEEE Press, New York,NY

Kohonen, T. (1995c) *Self-Organizing Maps*. Springer, Berlin.

Kohonen, T., Oja, E., Simula, O., Visa, A., and Kangas, J. (1996b) Engineering application of the self-organizing map. Proceeding of the IEEE, 84:1358-1384