自组织映射的基本要素

摘要：自组织映射 (SOM) 是一种自动数据分析方法。广泛应用于工业、金融、自然科学、语言学等领域的聚类问题和数据挖掘。其最广泛的应用，正如本文中所举例的，可以在大文本数据库管理和生物信息学中找到。SOM与在数字信号处理和传输中有着广泛应用的经典矢量量化 (VQ) 相关。跟VQ一样，SOM 使用一个模型的有限集合来表示各输入数据项的分布情况。不过，在 SOM 中，这些模型是自动与常规 (通常是二维) 网络中的节点有序地关联起来，从而使得相似性较大的模型会自动与网格中相邻的节点关联起来，而相似性小的模型被放置在网络中彼此相隔较远的位置。该组织是模型的一种相似关系图，有了这个关系图就可以观察得到数据的位置关系，特别是高维数据项。如果数据项属于某些预先确定的类别，则可以根据这些类别对模型 (和节点) 进行校验。然后，未知输入项就可以根据对应节点进行分类了，节点中的模型在用于构建SOM的一些度量上是最相似的。本文介绍的一个新发现是，一个输入项可以更准确地由几个最佳匹配模型的线性混合来表示。这通过最小二乘法拟合过程是有可能的，但其中线性混合模型中的系数要求为非负值。

关键字：自组织映射 SOM 数据分析 大脑映射 相似度 矢量量化

1. 大脑映射

在过去的一百多年里，人们一直认为，大脑的各个皮层区域是专门对应不同的认知功能模块。然而，直到后来，Mountcastle (1957) 以及 Hubel 和伊利·威塞尔 (1962) 有了新的发现，大脑中的某些单个神经细胞是有选择性地对一些特定的感官刺激做出反应。这些细胞通常形成局部集合，它们在集合中的位置顺序会对应于特定刺激物的某些特征值。这种细胞系统被称为大脑映射。

刚开始，人们认为大脑映射是由基因决定的，就像身体的其他形态和组织一样。一直到许多这些映射，至少他们的内部结构和特征级别被发现是取决于感官经验和其他事件之后，才有了新的认识。经过研究经验的报导，特别是由Merzenich(1983)等人的报导，大脑映射的研究有了很大的转变。

在二十世纪七十年代的一些理论生物学家中，例如Grossberg(1976)，Nass(1976) ，Cooper(1975) ，Perez、Glass和 Shlaer(1975)，提出了这样的问题：通过学习(即适应模拟感官刺激)，特征敏感细胞在人工系统中是否也能自动形成。然而，Malsburg (1973)和后来的Amari (1980)已经表明，它们的拓扑排序也会从输入数据中得来。

在输入驱动自组织的第一个成功理论证明中，上述建模方法值得一提。其中的特征敏感细胞的出现就是由所谓的竞争性学习神经网络实现的。在细胞的子集中，对传入的输入信号有最强激活的细胞的适应性，使其能适应特定的输入特性或它们的组合。

然而，早期的仿生大脑映射模型并不适合于实际的数据分析。它们的一个固有缺陷就是得到的映射是被分割的。它们是由一些顺序是不连续地跳跃的、随机的小组件组成。因此在整个映射队列上的全局排序没能实现。虽然这种局部排序在生物学上是司空见惯的，但许多代表这种抽象功能的大脑映射，如Suga 和 O’Neill (1979)，Tunturi (1950, 1952)，和 Zeki (1980)分别报导的音质映射、色彩映射和声纳回波映射，都是要有全局组织的。但这些模型也没能扩大规模，它们不能用于大型网络和高信号维度，尽管计算能力大大提高。

回顾以往，从早期的自组织神经模型来看，有一个重要的因素缺失了。它是一个控制因子或函数，其数量取决于局部信号的活动，但它本身并不对信号造成影响。其唯一目的是控制网络中神经连接选定的子集的可塑性 (通过信号可修改性)。因此，在神经模型中，仅通过一些经连接 (即神经连接) 的其他节点的活动来控制节点的活动是不够的。人们需要额外的各种控制因素来调解信息而不是调解活动。众所周知，这样的一个信息是在神经领域中通过例如化学上的信使分子传递的。

另一方面，如果上述神经和化学作用被考虑到，至少以抽象形式被考虑到，那就有可能将自组织系统提升到能突破现代计算机的性能极限的水平。

2. 经典矢量量化 (VQ)

通过竞争性学习实现最佳调谐的特征敏感滤波器实际上在很早之前在信号处理中已经以抽象的形式有所展示。我的意思是经典矢量量化 (VQ)，它的基本思想由 Lloyd (1957) (以标量形式) 和 Forgy (1965) (以向量形式)已经介绍过。事实上，矢量空间的最佳量化可追溯到1850年，称为二维和三维空间中的狄利克雷细分以及任意维空间中的沃罗诺伊细分；参见狄利克雷(1850) 和 沃罗诺伊 (1907)。从那以后，VQ 成为现代数字信号处理的标准技术。

在矢量量化中，向量值输入数据的空间，例如特征向量，被分成有限数量的连续区域，每个区域通过单个模型向量进行最佳表示，最初在 VQ 中称为码书向量。(后者来自数字信号传输，其中 VQ 用于传输信息的编码和解码)

在最佳分区中，码书向量是构造出来的，这样在最佳匹配的码书向量(称之为胜利者)中的输入数据项的平均距离(以某种标准度量)就被最小化了，即平均量化误差最小化。

为了简单起见，VQ 仅用欧氏距离来说明。将输入数据项构成n维欧几里德向量，用x表示。码书向量用 表示，索引用下标i表示。特定的码书向量用下标c表示索引，称为胜利者，即与x的欧几里德距离最小:



(1)

若 p (x) 是x的概率密度,平均量化错误 E被定义为：

。 (2)

其中 dV是数据空间V的体积差。作为能量函数的目标函数 E可以通过梯度下降过程最小化。然而，问题是高度非线性的；尽管如此，例如，本书作者已经表明，它收敛到一个局部最低限度；参见 Kohonen (1991)。

如果输入数据项集是有限集，那么批处理计算方法也是可行的。它被称为Linde–Buzo–Gray (LBG)算法，参见Linde、Buzo 和Gray (1980)，但它是由 Forgy (1965) 设计的。关于上述VQ已有丰富的相关文献，也称为 "k-均值分类 "。经典的一些参考，可参见 Gersho (1979)，Gray (1984) 和 Makhoul、Roucos 和 Gis (1985)这些例子。

图1. 自组织映射说明。一个输入数据项X会被传播到模型集合中，模型集合中与X最匹配。网络中位于的邻域(较大圆圈)中的所有模型相对其他模型与X的匹配效果更好。

3. 自组织网络: 常规

3.1. SOM的机理

大概1981 -1982年左右，本书作者引入了一种新的非线性投影映射，称之为自组织映射（SOM），它与VQ类似，但其中的模型（对应于VQ中的码书矢量）变成了空间的，是全局有序的（Kohonen，1982a，1982b，1990,2001）。

SOM 模型与常规的，通常是二维网络的节点(图1) 相关联。SOM 算法构建了以下模型:

较相似的模型将与网格中较接近的节点相关联，而不相似的模型会逐渐被放置在网络中位置相隔较远的地方。

如果在刚开始，中心思想就以以下简单的说明方式表述出来，那么理解 SOM的学习规则和数学含义就更容易了:

每个输入数据项都应选择与输入项最匹配的模型，该模型及其网格中的空间邻居子集应进行相应的修改以更好地匹配。

和在VQ 中类似，所要做的修改主要集中在包含胜利者模型所选定节点上。另一方面，因为在网格中胜者周围的整个空间邻域会被修改一次，且由于平滑作用，在该邻域中的模型的局部排序的程度将提升。连续的不同的输入会引起模型的不同子集的修正。局部排序操作将逐渐扩展到整个网格。然而，真正的数学过程比那更复杂一些。

生成SOM模型的有序集的实际计算可以通过以下任一个主要类型的算法来实现: 1.初始SOM算法中的模型由递归的，逐步逼近过程来计算。为了达到合理稳定状态所需的步骤数，其中输入数据项可以按周期性或随机序列，一次一个地应用于该算法中。 2.另一方面，在批处理方式过程中，所有输入数据项都作为一个批处理应用于算法，并且所有模型都在单个并行操作中进行更新。这个批处理过程通常需要多重复几次，通常这样之后模型会完全稳定。即使仅是到达近似稳定的状态，时间上也比逐步计算方式的时间短一个数量级。

应该强调的是，只有批量学习版的SOM才可用于实际应用，因为它不涉及任何学习速率参数，它的收敛速度会快出一个数量级，并且更安全。另一方面，逐步学习规则最初建立仅仅是为了理论论证理由，便于与其他自组织模型进行比较。而且，我们不可能用常规的度量标准来使用逐步学习，但我们会发现批处理学习也能解决这个问题。

下面将对SOM算法作更详细的描述。

很多商业软件包以及关于SOM的大量免费软件都已经有可用的了。本书作者极力推荐使用合理的公共领域软件包。例如，由我们开发的两个免费软件包：*SOM\_PAK* (Kohonen、Hynninen、Kangas、& Laaksonen、1996；SOM\_PAK 团队，1990) 和*SOM Toolbox*(SOM工具箱团队，1999；Vesanto、Alhoniemi、Himberg、Kiviluoto & Parviainen 1999;；Vesanto、Himberg、Alhoniemi & Parhankangas，1999)，都可从互联网上下载。两个包都包含辅助分析程序，特别是使用MATLAB函数的 SOM 工具箱，提供了多种图形工具。

与大多数仿生的映射模型不同，SOM中的拓扑排序始终可以在整个映射上全局具体化。

从展示出来的空间顺序中，可方便快速地对输入数据的相似关系进行可视化检查，并便于对数据样本进行检验和确认。而且，通过对模型的适当的校验，数据的聚类和分类变就会得清晰明了。

本文的余下部分集中讨论SOM原理和应用。SOM 作为一种可视化工具在探索性数据分析中得到了广泛的应用。从工业处理控制和金融分析到大型文档库的管理，都存在大量的实际应用。生物信息学中也存在着新的、富有前景的应用。迄今为止，最大的应用是在文本文件的管理和检索中，本文就包含有两个这样的例子。

许多版本的SOM算法已被推广多年。它们数量太多，这里就不能一一罗列；参考文献中会提到广泛的书目。参见：(Kaski、Kangas、& Kohonen、1998； Oja、Kaski & Kohonen，2003； Pöllä、Honkela & Kohonen，2009)。另请参见，第7节讨论。

3.2. SOM的校正

如果输入项落入有限的类别中，则可以对应这些类别建立不同的模型，并提供相应的符号标签。这种类型的校正可以通过两种方式进行: 1.如果输入项的数量足够多，则可以先研究所有输入数据项与各种模型所做匹配的分布。大多数输入样本与该模型匹配时，会根据类别标出一个特定模型。在匹配程度相同的情况下，有一个做法是，可以对该模型的较大的邻域进行多数选举(选举匹配的样本多的模型)。2. 如果只有少量的输入数据项可用，那么上述的多数选举就没有意义 (例如，会有很多的平局，或者某些模型没有命中)，则可以用所谓的k-nearest-neighbors (kNN)方法。对于每个模型，将搜索与之最接近(在SOM构造中应用的度量标准上)的k个输入数据项，并且对其进行多数选举以确定节点的最可能分类。在平局的情况下，k的值会增加，直到该平局打破为止。通常，k的选择可以顺序的从6到100，具体取决于输入数据项的数量和SOM组的大小。

当将一个新的未知输入项与所有模型进行比较时，将用最佳匹配模型来标识它。然后将输入项的分类理解为模型最佳匹配。

3.3. 关于"相似性匹配"

SOM存在很多版本，使用不同的"相似性"定义。这个属性值得简短的讨论。"相似性"和"距离"通常是相反概念。

对相似性的意义认知是非常模糊的。例如，人们可以谈论两个人或两个历史时代的相似之处，尽管这种比较通常是基于主观的观点。

如果同样的比较要自动地实现，那么它只能基于一些非常有限的分析，称为统计属性。如果我们处理科学或技术中的具体对象，情况就会清晰很多，因为那时候我们可以，基于数学概念上的属性向量之间的距离测量的非相似定义。统计数字通常也表示为实际向量，由数值结果或其他统计指标组成。各种谱系和其他变换也可以看作是它们的分量的多维向量。

尝试比较此类向量的第一个问题通常是它们的元素的不同的缩放。对于度量标准比较，一个简单的补救方法是规范化刻度，以便不同维度中的变量的方差，或者它们的最大值和最小值分别变为相同。在这之后，一些标准距离测度，如欧几里德，或更一般，Minkowski距离等可以尝试，根据数据的性质选择。事实证明，欧几里德距离 (规范化) 已经适用于大多数实际研究，因为 SOM能够在展示更复杂的相互依存的变量。

向量之间的相似性的自然度量一般是内积。在SOM研究中，通常使用点积。这一措施也使用比欧几里德距离更吻合生物的神经模型。但是，对于与输入x进行比较的模型向量，必须始终将其规范化为定长。如果矢量维数很高，并且输入向量被规范化为定长，则SOMs基于欧氏距离和基于点积的之间的差别是微不足道的。(对于构建欧几里德和点积SOMs，分别参见4.1节和4.5节)。另一方面，如果向量中有大量的零元素，则点积的计算速度相应地更快。此属性可以有效地使用，特别是在本文末尾所讨论的文档映射的快速计算中。

在进入更深一步之前，有必要强调一个基本事实。一张图片，通常是一组像素或其他结构元素，像这样输入向量一般不适用。图像中的自然变化，如平移、旋转、尺寸变化等，以及由于不同光照条件下而造成的变化通常是很广泛的，以至于基于它们的外观进行对象的直接比较结果是不敏感的。相反，自然项的分类应基于其特征的进行提取和分类，这些特性必须是尽可能的不变。这种类型的特征可以包括颜色图谱、傅立叶变换中图像的扩展、小波、主分量或某些图像算子的特征向量等。如果可以通过一组受限制的不变特征来描述输入对象，则输入表示的维数和计算负载将减少很多。

有一种特殊的非相似性或者距离度量，应用在被称为自适应子空间的SOM (ASSOM)中。参见：Kohonen (1995，1996，2001)和Kohonen、Kaski 和 Lappalainen(1997)。其中，某些基本系统与节点相关联，并且这些系统发展为对某些类别的局部特征（例如平移不变，旋转不变或尺寸不变）进行恒定响应的特定滤波器。它们的参数是通过对相同输入模式(不管它们是什么)的转变而调整适应来确定的，这些模式在被关注的各种基本变换中产生。在数学上这些系统每一个都表示信号空间的一个子空间，并由几个基向量来表示。一个输入向量到子空间的距离被定义为在正交补子空间上的正交投影。参见： Kohonen (2001，p.6)。因此，如果到子空间的距离长度是零，输入向量可表述为基向量的线性组合；但一般而言，那样距离的长度会产误差。在自适应学习中，可旋转基向量来减小误差，即到应有子空间的距离。

对特征集的选择和从原始观测中自动提取特征通常必须基于启发式规则。在生物学中，各种特征检测器都是在很长的进化过程中发展起来的。

对于更复杂的比较，您可能还会寻找其他种类的特征用作向量元素。例如在文本分析中的，完整的文档可以根据其单词的统计 (即单词的直方图) 区别开来，因此必须非常细致地关注词在不同的文本中的关系性。参见：Salton和McGill(1983)。因此，直方图的元素，对应于不同的词，必须是从由单词显露出的相关性而产生的乘法因子加权。除了使用一个单词的统计熵外，直方图中的单词也可以通过基于反向文档频率 (IDF)的方法加权 (因此，可以忽略稀有和非常常见的单词)。"文件频率" 是指特定的词汇在文本语料库中文档出现多少次，而以IDF则是这个数字的逆。在适当的权重下，构成特征向量的词直方图可以被限制为一些数以百计的维度。

符号字符串构成了另一种常见的变量类型。除了在文本中会产生字符串变量外，例如在生物信息学和有机化学中: 在基因编码，大分子的原子序列等都有字符串。参见例子，Kohonen 和 Somervuo (2002) 和Oja,Somervuo、Kaski 和 Kohonen (2003)。通常字符串的长度差异很大。某种编辑距离，即将一个字符串转换为另一字符串所需的基本编辑操作数，是字符串变量之间的距离的一个非常有效的定义。这些操作通常必须根据各种误差的统计进行加权。对于非常长的字符串，如蛋白质序列，一些启发式的快捷计算方式，如应用于广泛流传的FASTA 方法(Pearson，1999；Pearson & Lipman，1988)，可能是需要的。这些距离度量在数据库中经常是预先计算的。

还有其他更抽象的相似性度量方法。其中之一是单词的上下文相似性。考虑在一份文本中的一个词，在其相邻词的上下文中考虑。如果词汇中的每个单词都由一个随机的代码表示，则词的表示之间的相互关联度仍然很小。但是，两个局部上下文的相似性度量，例如，文本中连续三个单词的三次重复，然后连续产生三个同样位置上的一模一样的随机代码。从这里的分析中就可以得出非常深刻的语言学结论：对词的语义值的分析可以基于语境相似性研究，如 Kohonen 和Xing (2011)。

一种非常特殊的相似性度量方法是功能相似性，这可能意味着例如以下所讲的内容。考虑到信号过滤器受一组有限参数约束；否则，过滤器的结构是相同的。对于相同的输入信号，不同的过滤器通常会产生不同的响应。应该是因为滤镜用作了测试器。一个与给定输入有关的初始预测误差，是预测值和未来信号真实值的差的绝对值。对于给定输入信号的统计信息，两个过滤器之间的距离(即其参数集的间距) 可以定义为其应有的预测错误的均方根，如 Lampinen 和Oja (1989)建议的那样。注意，尽管各筛选器的预测误差的差别可能很小，但两组筛选器参数可能看起来差异很大。在实际应用中，人们可能会发现不同种类的功能相似之处是不限数量的。

一个重要的任务是比较动态现象。如果将模型用来表示动态状态，就有可能会出现这种情况。Hammer、Micheli、Sperduti 和 Strickert (2004 ) 对动态 SOMs 进行了非常有趣的讨论。

3.4. 建模中的抽象层次

生物建模。最早的 SOM 模型，倾向于复制详细的神经网络结构，是为了描述和解释大脑映射的创建。一种非常现代的很接近特征敏感的，对视觉皮层上的细胞形成的方法由 Miikkulainen、Bednar、Choe和 Sirosh (2005)制作出来。可能会有人提到最新版本的自组织预测(Kohonen，2005，2006)，这种自组织预测可以实现全局排序。有了它，塑性控制分子的扩散模型会被涉及到。

数学抽象。在抽象的数学模型中，忽略了网络连接、突触可塑性和化学控制等细节，通过矩阵向量函数和计算来表示活动的动态。利用这些神经网络的抽象，人们能够扩展到许多问题，使SOM成为一个实用的数据分析工具。

参考文献：

Allinson, N., Yin, H., Allinson, L., & Slack, J. (Eds.) (2001). Advances in self-organizing maps. London, UK: Springer.

Amari, S. (1980). Topographic organization of nerve fields. Bulletin of Mathematical Biology, 42, 339–364.

Anderberg, M. (1973). Cluster analysis for applications. New York, NY: Academic.

Bishop, C. M., Svensen, M., & Williams, C. K. I. (1998). GTM: the generative topographic mapping. Neural Computation, 10, 215–234.

Cheng, Y. (1997). Convergence and ordering of Kohonen’s Batch map. Neural Computation, 9, 1667–1676.

Cottrell, M., & Fort, J. C. (1987). Étude d’un processus d’auto-organization. Annales de l’Institut Henri Poincaré, 23, 1–20.

Cottrell, M., Fort, J.C., & Pagés, G. (1997). Theoretical aspects of the SOM algorithm. In Proceedings of the WSOM 97, workshop on self-organizing maps. Helsinki University of Technology. Neural Networks Research Centre. Espoo, Finland (pp. 246–267).

Deboeck, G., & Kohonen, T. (1998). Visual explorations in finance with self-organizing maps. London, UK: Springer, pp. 246–267.

Deerwester, S., Dumais, S., Furnas, G., & Landauer, K. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 391–407.

Dirichlet, G. L. (1850). Über die Reduktion der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. Journal für die Reine und Angewandte Mathematik, 40, 209–227.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. Biometrics, 21, 768. abstract.

Fritzke, B. (1994). Growing cell structures—a self-organizing network for unsupervised and supervised learning. Neural Networks, 7, 1441–1460.

Gersho, A. (1979). On the structure of vector quantizers. IEEE Transactions on Information Theory, IT-25, 373–380.

Gray, R. M. (1984). Vector quantization. IEEE ASSP Magazine, 1, 4–29.

Grossberg, S. (1976). On the development of feature detectors in the visual cortex with applications to learning and reaction–diffusion systems. Biological Cybernetics, 21, 145–159.

Hammer, B., Micheli, A., Sperduti, A., & Strickert, M. (2004). Recursive selforganizing network models. Neural Networks, 17, 1061–1085.

Hartigan, J. (1975). Clustering algorithms. New York, NY: Wiley.

Heskes, T. M., & Kappen, B. (1993). Error potential for self-organization. In Proceedings of ICNN’93, international conference on neural networks, vol. III (pp. 1219–1223). Piscataway, NJ: IEEE Service Center.

Hubel, D. H., & Wiesel, T. H. (1962). Receptive fields, binocular and functional architecture in the cat’s visual cortex. Journal of Physiology, 160, 106–154.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering of data. Englewood Cliffs, NJ: Prentice-Hall.

Kaski, S. (1998). Dimensionality reduction by random mapping. In Proceedings of IJCNN’98, international joint conference on neural networks (pp. 413–418). Piscataway, NJ: IEEE Press.

Kaski, S., Honkela, T., Lagus, K., & Kohonen, T. (1998). WEBSOM—self-organizing maps of document collections. Neurocomputing, 21(1), 101–117.

Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981–1997. Neural Computing Surveys, 1, 1–176. Available in electronic form at http://www.cis.hut.fi/research/som-bibl/vol1\_4.pdf.

Kohonen, T. (1982a). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59–69.

Kohonen, T. (1982b). Clustering, taxonomy, and topological maps of patterns. In Proceedings of the sixth international conference on pattern recognition (pp. 114–128). Washington, DC: IEEE Computer Soc. Press.

Kohonen, T. (1985). Median strings. Pattern Recognition Letters, 3, 309–313.

Kohonen, T. (1989). Self-organization and associative memory (3rd ed.). Berlin–Heidelberg–Germany: Springer.

Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78, 1464–1480.

Kohonen, T. (1991). Self-organizing maps: optimization approaches. In T. Kohonen, K. Mäkisara, O. Simula, & J. Kangas (Eds.), Artificial neural networks, II (pp. 981–990). Amsterdam, Netherlands: North-Holland.

Kohonen, T. (1995). Emergence of invariant-feature detectors in self organization.

In M. Palaniswami, Y. Attikiouzel, R. J. Marks II, D. Fogel, & T. Fukuda (Eds.), Computational intelligence, a dynamic system perspective (pp. 17–31). New York, NY: IEEE Press.

Kohonen, T. (1996). Emergence of invariant-feature detectors in the adaptivesubspace self organizing ma. Biological Cybernetics, 75(4), 281–291.

Kohonen, T. (2001). Self-organizing maps (3rd ed.). Berlin–Heidelberg, Germany: Springer.

Kohonen, T. (2005). Pointwise organizing projections. In Proceedings of the WSOM05, 5th workshop on self-organizing maps Panthéon-Sorbonne University. Paris, France (pp. 1–8). Also available at http://samos.univ-paris1.fr/wsom/wsom05. html.

Kohonen, T. (2006). Self-organizing neural projections. Neural Networks, 19, 723–733.