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)。在其中，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)的转换进行调整来确定的，这些模式在观察的各种基本变换中发生。在数学上每个这些系统表示信号空间的一个子空间，并由几个基向量描述。从子空间的输入向量的距离被定义为它的正交投影在这个子空间的正交补语，参见： Kohonen (2001，p。 6)。因此，如果距离的长度从子空间是零，输入向量是表述作为基向量的线性组合;但一般而言，距离的长度构成错误。在自适应学习中，基向量为旋转以减小误差，即从应有子空间的距离。

对特征集的选择和从主观测到的自动提取必须经常基于启发式规则。在生物学中，各种特征检测器已经发展了很长的进化过程。

对于更复杂的比较，您可能还会寻找其他类型的功能用作向量元素。例如，在文本分析中，完整的文档可以根据其字数统计 (即word 直方图) 区别开来，因此必须非常小心地注意词在不同的文本;cf 沙顿和麦吉尔(1983)。因此，直方图的元素，对应于不同的词，必须是加权的乘法因子派生的相对发生的词。除了使用一个单词的统计熵外，直方图中的单词也可以加权 (因此，可以忽略稀有和非常常见的单词)，方法是基于反向文档频率 (IDF)。"文件频率" 是指文本语料库中有多少文档发生，而以色列国防军则是这个数字的逆。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。

符号字符串构成另一个常见的变量类型。除了文本，字符串变量发生，例如在生物信息学和有机化学: 在基因编码，原子序列的大分子等: cf，例如，Kohonen 和 Somervuo (2002) 和欧雅,Somervuo、Kaski 和 Kohonen (2003)。通常字符串的长度非常不同。某种编辑距离(即，将一个字符串转换为另一条线所需的基本编辑操作数) 是字符串变量之间距离的一个非常有效的定义。这些操作通常必须根据各种错误的统计进行加权。对于非常长的字符串 (如蛋白质序列)，有一些启发式的快捷方式计算，如应用于宽扩展 FASTA 方法 (皮尔逊，1999; 皮尔逊 & 李普曼，1988可能需要)。这些距离度量经常在数据库中预计算。

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。在文本中考虑一个词，在其相邻词的上下文中。如果词汇中的每个单词都由一个随机代码表示，则词的表示之间的相互关联仍然很小。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胞胎，然后发生在三胞胎相同位置的相同随机码的出现。对词的语义值的分析可以基于语境相似性研究，从这些分析中可以得出非常深刻的语言学结论，如 Kohonen 和邢 (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，例如以下内容。考虑一组由有限数量参数描述的信号过滤器;否则，筛选器的结构是相同的。对于相同的输入信号，不同的滤波器通常会产生不同的响应。考虑，例如，用作预测器的滤镜。一个与给定输入有关的初等预测误差，是预测的差值和真实的未来信号值的绝对值。对于输入信号的给定统计信息,两个筛选器之间的距离(即其参数集的间距) 可以定义为其应有的预测错误的根均方根，如 Lampinen 和欧雅 (1989)。请注意，两组筛选器参数可能看起来相当不同，尽管各自筛选器的预测错误的差异可能很小。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定数量的。

一个重要的任务是比较动态现象。如果将模型表示为动态状态，则可能会出现这种情况。锤子、西蒙米凯利、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地图的创建。通过 Miikkulainen、Bednar、崔和 Sirosh (2005)，对视觉皮层中的特征敏感细胞形成了一种非常现代的方法。还可能提到最近版本的 selforganizing 预测 (Kohonen，2005，2006)，其中可以实现全局顺序。有了 plasticitycontrolling 分子的扩散模型。

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力学。利用这些神经网络的抽象，人们能够扩展许多问题，使 SOM 成为一个实用的 dataanalysis 工具。

3。5。 网络体系结构

模型的二维组织通常已经有效地逼近高维数据项的相似关系，就像在一些早期的方法称为多维缩放 (MDS) ( 克鲁斯卡尔 & 愿望，1978; Sammon，1969)。但是，应该注意，在 MDS 方法中,每个数据项都必须以几何方式显示，而 SOM 只使用它在常规网格上显示的相对较小的模型集。这样，大量的计算将被保存。

对于特殊用途，可以使用大于两个维度的网格。一个简单的秩排序也可以进行沿一维网格，但是，将无法显示任何更一般的拓扑关系的数据项。

常规数组。SOM 的大多数应用都基于节点的常规数组。有时，一个使用矩形数组的节点，以简化。然而，六角阵列在视觉上更具说明性和准确性，并建议使用。无论使用什么常规体系结构，最好选择阵列的水平和垂直尺寸长度，以对应两个最大的主分量 (即输入的最高特征值的长度)。相关矩阵)，因为然后 SOM 更好地符合低阶信号统计。长圆形规则阵列的优势在于保证学习中更快更安全的收敛。

循环数组。通常的 SOM 模型，在学习之后，展示了熟悉的边界效果: 相邻模型的间距不像在 SOM 中间的边界那么规则。基于这个原因，一些研究人员建议，网络应该是绕圈循环，无论是环形的还是球形的。后一种选择是由绽放团队 (2005) 进行的。如果数据本身具有简单的循环结构，则循环网络是合适的。如果使用 som (例如，在过程控制中查找 som 中的最佳修正)，循环结构也可能是有利的。在循环映射中，即由于网络边界而没有间断。然而，即使在循环网络中，如果某些模型位于输入分布的强肢附近，也可以在模型的间距上辨别出类似的不规则。因此，违规行为似乎是由于数据分发的四肢，而不是网络边框。

增长的网络。一些研究表明，应使网络的结构更好地符合输入数据。因此，应该使网络结构成为变量，例如，应该允许它在输入数据分布的方向上增长;cf，例如，Fritzke (1994)。如果是这样，可以看到显式出现的数据的层次结构。然而，在定义增长方向的条件方面存在着一些武断性。任意分支标准定义网络中的任意结构。不过，在常规数组中，可视化显示效果更好，并且群集趋势可以由U-矩阵技术唯一地说明，如 Ultsch (1993)。

3。6。 常见问题解答

关于 SOM 结构的最常问的问题之一是数组中需要多少个节点。也许有必要首先说明 SOM 的主要优点是数据空间的可视化，因此群集结构应该变得可见。对于非常小的数据集，使用 SOM 是没有意义的，因为它们存在更好的数据分析方法。

因此，假设我们有足够的数据项，以充分地可视化数据空间的准确性。然后应该认识到 SOM 也是一个量化方法，并且具有有限的分辨率来显示群集结构。有时，数据集可能只包含几个簇，粗略的分辨率就足够了。但是，如果有人怀疑数据中有有趣的精细结构，那么需要一个更大的数组来进行足够的解析。

典型的 SOM 数组的范围从几打到几个百节点，并且数组的水平和垂直维度的关系应该至少服从于输入数据的两个最大主分量的关系。不可能预先猜测或估计数组的确切大小。在看到第一个猜想的质量后，必须由试错方法来确定。

但是，还需要认识到 SOM 通常用作一种直方图，其中一个显示映射到每个节点的输入数据项的数量。这种直方图可以通过灰度或 pseudocolors 的阴影来可视化。这种直方图的统计精度取决于每个节点平均映射多少个输入项。一个非常粗略的经验法则可能是，每个节点平均大约有50个项目是足够的，否则该分辨率受数据稀疏性的限制。因此，必须在分辨率和统计精度之间作出妥协。这些方面应特别在统计研究中加以考虑，因为只有有限数量的样本可用。

另一方面，SOM 可能在工业过程的可视化方面处于最佳状态，在那里可以记录无限量的测量结果。在后一种情况下，你可能会尝试使用一个大的数组，因为你可以计算，并与现代计算机，甚至个人的，有可能处理数以千计的节点在 SOM 数组。

4。 两种主要的 SOM 算法

4。1。 原始的逐步递归 SOM 算法

SOM 算法的原始公式类似于渐变下降过程。但是，必须强调的是，在尝试实现3。1 节中给出的一般学习原则时，启发式引入了此版本的算法 。这个基本形式尚未被证明是衍生从任何能量函数。从随机逼近方法 (罗宾斯 & 蒙罗，1951) 中产生的近似和纯正规但不是非常严格的派生;它在 Kohonen (2001，146–147) 中应用。尽管如此，根据相应的修改后的目标函数，成功地派生了稍微修改的 SOM 算法的收敛性，这是由 Heskes 和弗兰克林·范卡彭 (1993) 提供的,以及Kohonen (1991) 和克里斯托弗鲁特瑞尔(1992)。

原始 SOM 算法与全局有序状态的收敛性已经在一些简单的 lowdimensional 案例中得到了数学证明: 请参见科特雷尔和堡垒(1987)。另一方面，瑞特、Martinetz 和 Schulten (1992) 提供了一个证据，证明如果输入的分布为更一般的度，则可以达到本地顺序。向量是离散值。

下面，仅显示 SOM 的原始形式，因为它是计算最简单和最轻的，实际上，几乎毫无例外，它也会为 highdimensional 输入向量产生有用的结果。对排序结果的正确性和生成图的质量进行了数学分析。科特雷尔、堡垒和 Pagés (1997 ) 已彻底讨论了 SOM 的理论基础。)。

再次考虑图 1。此时，让输入数据项构成序列 {x(t)} 的真实n维欧几里德向量x,其中，t(整数) 表示序列中的一个步骤。让 {mi(t)} 是另一个序列的n维度实向量,表示模型mi的连续计算的近似。这里的i是与mi 关联的网格节点的空间索引。原始 SOM 算法假定以下过程聚合并生成模型的要求排序值:

m i (t + 1) = m i (t) + hci (t)[x(t) − mi(t)]，(3)

where hci (t) 称为邻域函数。此函数类似于通常的平滑过程中应用的内核。下标 c是网格中特定节点 (赢家) 的索引，即与型号为m(t),() 具有最小欧几里德从x(t ) 的距离):

c = argmin{∥x(t) − mi(t) ∥}。 (4)

我

情商(3) 和 (4) 可以被说明为定义一个递归步骤，其中首先输入数据项(t) 定义或选择 bestmatching模型 (优胜者) 在网格中根据 Eq。(4)。然后，根据 Eq。(3)中，此节点上的模型以及其空间邻居在网格中被修改。修改总是在这样的方向进行，修改后的模型将与输入更好地匹配。

不同节点上的修改速率取决于函数的数学形式hci (t) 。邻域函数的一个非常实用的选择hci(t) 是

hci (t) = α (t) exp[−sqdist(c，i)/2σ2(t )]，(5)

where α (t) 是一个单调的 (例如，hyperbolically、指数或分段线性) 递减标量函数的t，sqdist (c，i) 是网格中节点c和i之间几何距离的平方，σ (t) 是t分别的另一个单调递减函数。 σ () 的真正数学形式是不重要的，只要它的值在进程开始时相当大，比方说，在网格直径的一半的顺序上，随后它在大约1000步逐渐减少到它的一小部分。拓扑顺序在这个期间被开发。另一方面，在粗略排序的这一初始阶段之后,最终收敛到模型的几乎最优值，比方说，一个数量级的多步骤的顺序。为了得到足够的统计精度，每个模型必须经常更新。但是，我们必须发出警告: 的最终值 σ 不应转到零，因为否则进程将失去其排序能力。它应该始终保持，比方说，在网格间距的一半以上。 在非常大的 SOM 网格中，σ 的最终值可能是网格较短一侧的五的顺序。

hci 的数学形式也有其他可能的选择 (t) 。其中一个非常简单，在其中，我们有hci = 1，从赢家到一定半径，否则为零。

邻里函数在自组织中具有最核心的作用。然而，尝试用纯神经元件实现它并没有成功。相反，在生物建模中，似乎最好的方法是将某些化学控制剂从细胞活性较高的地方扩散出来;cf，例如，Kohonen (2006)。

建议读者使用一些现成的完整软件包，因为在定义良好的学习过程中有许多参数和很多其他方面需要考虑。并且以下批计算过程在实践中是优选的，特别是，因为它包含较少参量比逐步递归算法，并且收敛得更快。

4。2。 学习过程的稳定状态

假设 SOM 的某些稳定状态的收敛性是真的，我们要求mi 的期望值 (t + 1) 和mi (t) for t →∞必须相等，hci 为非零，其中c = c (x(t)) 是输入的赢家节点的索引x (t) 。换言之，我们必须具有

∀i，Et ( ) ))( {hci (x) m m t =} = 0。 (6)

此处Et 是在t上的数学期望值运算符。在假定的渐近状态中，对于t →∞，m(t) 与t无关，并由m 表示。 ∗i 。如果期望值为

Et (。) 写入t →∞，如 (1/t)  t (。) ，我们可以编写

 h ci x (t )

m ∗ i = t   h 。 ci 。(7)

t

但这仍然是隐式表达式，因为c取决于x (t) 和mi。不过，Eq(7) 应用于用于mi的迭代解决方案的动机，称为SOM ("批处理映射") 的批处理计算。

那么请注意，对于不同的节点i，相同的加数发生大量的时间。因此，可取的是，特别是在非常大的 SOMs，首先计算的平均值xm，所有的x(t ) ，最接近数据空间中的模型mj ，然后按此子集中样本的编号nj 和h吉对其权重。然后，我们获得

 n jh 吉安 x m ，j

m ∗ i =  h，(8)  ,

J

在其中批注xm ,j 用于表示最接近模型mj的输入的平均值，nj 是这些输入的数目。

Eq。(7) 或 Eq(8) 稍后将用于派生迭代批计算算法。

4。3。 模型的初始化

特殊问题涉及为mi选择初始值。 Kohonen (2001) 已经演示了它们可以被选择，即使是随机向量，但如果将初始值选择为常规，则更快的排序和收敛速度将会更短,沿超平面进行的二维向量序列，由 x 的两个最大主要组件(即与两个最高特征值相关联的主组件) 跨越。 cf Kohonen (2001)。此方法称为线性初始化。

初始化模型作为随机向量最初用于证明 SOM 的能力成为有序的，从任意初始状态开始。在实际应用中，人们期望尽快实现最终的排序，所以选择良好的初始状态可以加快算法的收敛速度。

在下一小节中，我们将讨论 SOM 的批计算，这也适用于输入向量的一般距离度量。如果输入向量是欧几里德，建议初始化 principalcomponent 方法。为一般距离措施随机初始化总是可能的;然而，一般度量的初始化优化是非常棘手的;cf 例如，Kohonen 和 Somervuo (2002)，不能在此处详细讨论。

对于基于函数相似性的 SOMs 的初始化，可能需要对模型参数进行随机初始化。

4。4。 SOM 的批计算

在继续的过程中，我们将集中于 SOM 的批计算，因为它比逐步递归算法更快、更安全，也可以推广到 nonvectorial 数据。

考虑图 2，其中显示由圆圈描述的二维六角形节点数组。对于每个节点i，模型mi 都是关联的。另外，包含某些输入向量x(t)的副本的列表与每个节点关联。

就像在逐步递归算法中一样，mi的初始值原则上可以选择为随机向量，最好是从输入向量的域中选取。但是，在欧氏度量的情况下，更好的策略是采取mi 从维超平面中提取的规则二维值序列，由两个最大的主要组件组成。x。

然后考虑输入数据向量集{x(t)}，其中t是向量的整数值索引。将每个x) ((t) 与所有模型进行比较，并将x (t 子)) 副本复制到与节点关联，模型向量与x() 在) 欧氏度量中的最佳匹配。

在这个例子中，我们假设一个邻域函数，它在节点的邻域集 Ni 中具有值 1，由节点i到一定距离的节点组成，否则等于零。

根据情商。(7) 和 (8) 每个模型的平衡值现在必须是x() 的平均值。) ，我们希望通过以下策略来接近这种平衡状态。我们计算x() 的的平均值，即所有 x (( ) 已复制到Ni中所有子列表的联合中。为每个节点i(即在所有节点周围的邻居) 计算了相似的平均值。更新mi 然后意味着mi 的旧值由相应的方法替换,在网格的所有节点上进行一个并行计算操作。这就结束了一个更新周期。

|  |
| --- |
| 图 2。 批处理过程中的一个周期的说明。输入数据项x() 首先分布到与其最佳匹配模型相关联的子列表中，然后将模型的新值确定为表示，或者更笼统地说，作为广义中线(在图中写入) 在子列表的邻域集上。 |

in 图 2但是，出现了一个新术语 "广义中值"，写在 set 行下面。在上面的讨论中，应该用 "平均值" 来替换，但因为图 2 也用于解释一般距离度量的自组织情况，因此在它中使用了 "广义中值" 一词，必须解释下一步。

批处理计算方法，如多次提到的，可以推广为 nonvectorial 数据，如符号字符串，如果一个广义中值在样本x (t) 可以定义。一组数据项中的此类中间值定义为与集中其他项具有最小距离的项。这一概念最初是作者提出的;cf Kohonen (1985)，但也可以从 Kohonen (2001) 中找到。我们以前使用的平均值在子列表的联合在邻里设置 N i 中，现在我们将使用同一角色中的广义中值。 此类 SOMs 是由 Kohonen 和 Somervuo (2002) 和病毒类型的 DNA 在欧雅，Somervuo et (2003)中为蛋白质序列构造的; 欧雅、斯珀伯、布隆博格和 Kaski (2004、2005)。

重申此更新周期，始终首先清除所有列表，然后在这些节点下分发输入向量的新副本，(更新的) 模型与新的输入向量匹配得最好。更新后的模型值迟早会变得稳定，在连续迭代中不再改变，训练可以停止。

如果将方法形成为加权平均值，则在其中权重与hci之间相互关联的情况下，就可以获得更符合逐步递归学习的过程。这里的c是节点的索引，更新的模型和i代表其邻域中节点的指数。

对批处理计算方法的收敛性进行了讨论郑汉钧(1997)。

4。5。 点-产品地图

对于公制向量，SOM 的实际计算基于其点产品。对于欧几里德向量，如果向量中有大量的零元素，则这种方法特别有利，因为它们在相似性的评估中被跳过。但是，对于与输入x进行比较的模型向量mi，必须始终将其规范化为常量长度。而不是 Eq。(1)中，获胜者位置的索引现在由定义

c = argmax。 {点(，mi)。(9)

我

由于 SOM 的计算实际上是由批处理算法进行的，所以将所有输入项映射到各自的赢家节点 (即关联的列表) 中，并以类似的方式进行。唯一的修改是修改 "广义中值" 的定义。可能存在这样做的几种可能性，但在我们的应用程序中，我们已经成功地使用了以下规则，如果在每个节点上映射的项数平均值很高，则至少适用这一点:

在 "点-产品 SOMs" 中，一组项目的 "广义中位数" 被标识为与所有其他项目具有最小点乘积总和的项目。

在每次迭代周期后，mi 的规范化将被定为长度。

5。 SOM 的应用

5。1。 SOM 的主要应用领域

在查看详细信息之前，可能有兴趣了解 SOM 方法的合理性。简言之，到2005年底，我们已经记录了7768份科学出版物: 参见： Kaski、Kangas 等。(1998)，欧雅、Kaski 等(2003) 和Pöllä等(2009)，用于分析、开发或应用 SOM。下面的简短列表提供了主要的应用程序区域:

1。 大统计方法

(a) 探索性数据分析

(b) 文本的统计分析和组织

2。 工业分析、控制和电信:Kohonen、欧雅、仿真、签证和 Kangas(1996)

3。 生物医学分析和应用程序大

4。 财务应用程序:Deboeck 和 Kohonen(1998)

除了这些之外，还可以提到一些具体的应用，例如，对犯罪分子行为的分析、星系的分类 (名、Ratnatunga、& 格里菲斯、1997)、不动产分类等。

对于som 的理论和不同版本或 som 在本文中的应用，是不可能充分考虑的。我们只能参考上面列出的 7768 SOM 出版物 (今天，他们的数量超过 1万)，以及十多本教科书、专著或编辑过的书籍，例如 Allinson、尹、Allinson 和松弛 (2001)，Kohonen (1989，2001)，Miikkulainen(1993)，Obermayer 和 Sejnowski(2001)，欧雅和 Kaski(1999)，瑞德等。(1992)，塞弗特和耆那教(2002)，Tokutaka、田文雄和藤村(1999)，Tokutaka、Ookita 和藤村(2007) 和Van Hulle(2000) 和大量的博士论文。

本杂志的两个特别问题专门讨论了 SOM: 2002 特别问题与副标题 "自组织地图的新的发展",神经网络，卷 1，数字 8–9，10月/2002年11月和2006特别问题 "自组织 Maps—WSOM'05 的进展,神经网络，卷 1，数字 6–7，7月/2006年8月。此外，日志神经计算在卷21中发布了一个特殊的 SOM 问题，数字 1–3，1998年10月。

1997年以来，召开了一系列会议，名为WSOM (SelfOrganizing 地图讲习班) 。它们已在下列地点举办: 芬兰 Otaniemi (1997 和 1999);林肯，英国 (2001)，北九州，日本 (2003);巴黎，法国 (2005);德国比勒费尔德 (2007);圣奥古斯丁，佛罗里达州，美国 (2009)，埃斯波，芬兰 (2011)，智利圣地亚哥 (2012)。

5。2。 建造大型 SOMs 的提示

有了非常大的地图，赢家搜索和更新都是耗时的操作。请注意，对于大型数组，社区的初始半径通常也很大。即使是最终半径也可能是一打网格间距的顺序。然后，如果模型的初始值接近最终状态，则 SOM 的收敛速度至少可以更快。

一个很好的估计，一个大地图的初始值可以通过从一个较小的地图开始，正确地初始化，例如，由主成分的方法 (线性初始化)，让它得到稳定近似，然后添加新的,与 SOM 数组的间隙节点。与间隙节点相关的模型的初始值可以根据已近似收敛的较小阵列进行插补。在这之后，更大的地图必须让在更长的训练过程中仔细地融合。

特别是在作为向量的词直方图中，通常有大量的零元素。如果应用点产品相似性度量，则零元素对其计算没有贡献。可以为每个输入向量 pretabulate 非零元素的索引，此后只考虑那些计算点产品的元素。

优胜者的查寻总可以由数量顺序加速，通过存放老优胜者地点的地址与相应的训练数据一起。对老优胜者的指向可以被利用在下个训练周期，仅限制查寻新的优胜者对老优胜者的邻里通过跟随被存放的指针，在之后对优胜者的指针更新。如果首先计算出较小的映射，则可以通过上面解释的插值方法获得指向更大映射的近似指针。

Eq。(8) 允许非常有效地实现并行计算，在这种情况下，不必为mi的计算新值预留额外内存。在计算新的赢家指针之后，不再需要模型向量的以前的值。它们可以被更新的方法替换为xm,j，并且模型向量的新值可以直接从上面的等式计算出来。

此外，通过将数据划分为共享内存计算机中的不同处理器，可以部分并行搜索结果。

有关更详细的说明，请参见Kohonen 等(2000) 或Kohonen(2001，p 294)。

在矢量的高维度下，利用向量元的表示精度较低，可进一步降低内存需求。注意，截断误差可与白噪声相媲美，在计算高维向量的矢量距离时，统计上平滑。在我们最大的 SOM 中，我们只使用了向量元素的八位精度。

5。3。 SOMs 文件收藏

可以通过 SOM 原理来形成文本文档的相似关系图，当文档的word 集合模型被用作特征时; cfLin、Soergel 和 Marchionini(1991)，Merkl(1995)，Merkl、Tjoa 和卡普佩尔(1994)，Scholtes(1991) 和Zavrel(1995)。最简单的特征包括词的加权直方图，被视为实向量，但通常必须进行一些高维直方图的降维。

一个新的想法，称为 WEBSOM，是构造第一个词类别 SOM基于上下文相似性的词(参见： 部分 3。3)。然后，自由自然文本的单词被聚集到这样一个词类别 SOM 的网格点上，形成一个词类别直方图。此直方图为文档的 SOM 提供了输入功能;cf Kaski、Honkela、Lagus 和 Kohonen

(1998)。

然而，我们放弃了这个想法，原因有两个。首先，一个词类别 SOM 的构造不是唯一的定义。第二，我们发现另一个基于加权词直方图的随机投影的降维原理，加上一些计算技巧 (下面讨论)，结果是数量级更有效。通用名称 "WEBSOM"，然而，后来也用于这个改进版本的文件 SOM。

5。3。1。 按类别分类的文档

|  |
| --- |
| 图 3。 将四路透社文档类映射到 SOM。"命中" 的密度由灰色阴影显示。 |

第一个示例中使用的文本语料库由四不同应用程序领域收集的文档组成，由路透社公司编写。未提供原始文档，但刘易斯、杨、Rose 和 Li (2004) 已经对文本数据进行了预处理，删除了 stop 单词，并将单词减少到它们的茎中; 换句话说，我们的工作是基于在减少的词直方图。我们实验室的 Salojärvi 从这个预处理语料库中选择了一个4000个文档子集，仅限于属于以下类别之一的文章:

1。 公司-工业。

2。 经济和经济指标。

3。 政府和社会。

4。 证券和大宗商品交易和市场。

每个类中有1000个文档。Salojärvi 然后拿起这1960字的直方图，出现至少200次在选定的文本。为了进行有统计学意义的独立实验，一小部分文件被预留用于测试。由曼宁和 Schütze (1999) 解释的熵类型因子对1960维词直方图进行加权。利用4000文档其余部分的加权词直方图作为输入，构建了一个2000节点 SOM。

图 3显示 som 上命中的四分布，当四个类中的每个输入项分别应用于 som 时。显然，地图是有序的，也就是说，四类文件被隔离到一个合理的准确性，和1类，3 和4甚至单独连接，尽管他们都处理密切相关的主题。

5。3。2。 大英百科全书的 SOM

在本例中，我们将强调 SOM 方法在文档语料库的组织和搜索等大量问题中的好处。对于文档，单词和其他术语 (word 组合) 的分布构成了一组可能的特征特征。也许有必要阻止最终的批评，指出虽然文本的语义分析更准确地分类了真实的文本内容，但我们却买不起这样的方法。我们的目标是处理大量的文学文献，因此标准的当代计算机，虽然相当先进的电脑，已经不得不在其能力的极限运作。我们选择的特点，然后展示了 (加权) 字直方图作为特征，结合 SOM，在探索大文本集合的力量。

《大英百科全书》电子版中，有6。8万篇文章被拾起。此外，还收集了4。3万个文本项目，如摘要、更新和其他杂项材料。非常长的文章被分成几个部分，总共产生了11。5万份文件。平均文档包含490个单词。然后由 Kohonen 等来计算此材料的 SOM。 (2000)。

首先对这些文档进行预处理，以删除 HTML 标记、链接和图像。用标准形态分析仪将屈折词形式转换为基形式;否则，每一个屈折形式将被视为一个不同的词。然后用反向文档频率 (IDF)加权这些单词。通过忽略具有非常低的以色列国防军价值的词语 (如 stopwords)，最终接受的词汇量的大小是39058字。这将成为 SOM 的输入向量的维数，但它仍然意味着非常繁重的计算。

在文本分析中，所谓潜在语义索引 (LSI) (Deerwester、Dumais、富尔纳斯、& Landauer、1990) 用于减少单词直方图的度。Kaski(1998) 我们的组发现了某个随机投影，如中介绍的瑞德和 Kohonen(1989)，这个词的直方图基本上是有效的，但计算非常轻。

将39058维特征向量随机投影到1000维空间上，减小了特征空间的维数。为此，每个加权直方图向量乘以相同的 39,058by-1000 矩阵，这些行被永久地设置为一些归一化的随机向量值。研究发现，投影的低维向量之间的相对相互距离与原始 highdimensional 向量的关系大致相同。

此应用程序中 SOM 网格的大小为12096个节点。应用了 som 计算中的几个 speedups: 例如，首先计算一个较小的网格的 som，然后引入间质节点形成一个较大的网格，为它们的

|  |
| --- |
| 图 4。 大英百科全书文章地图的特写视图。当用户单击映射区域中带有 "鲨鱼" 标签的节点时，他或她获得有关鲨鱼的文章列表。该节点 "幼虫" 含有昆虫和幼虫的文章，鸟类区域中部的一个节点上含有关于鸟类的文章。更偏远的区域 (这里没有显示) 包含鲸鱼和海豚的文章，并且在他们之外的区域描述捕鲸、鱼叉和爱斯基摩人，这证明主题在地图上平稳地和连续地改变。通过单击搜索的标题，可以读取完整的文章。 |

模型向量，最后用较大的网格进行较短的微调。

在构造了 SOM 并修复了模型后，每个文档都将映射为一个和唯一的节点，即模型向量与给定文档的特征向量匹配得最好的一个。您可以将原始文档存储在计算机内存系统中的单独缓冲区中，并将每个 SOM 模型与相应文档的地址指针关联。通常会将多个文档映射到每个节点。现在存在着这种系统的几种使用模式。

SOM 的一个简单使用模式是浏览。通常有一个与图形的规定，当显示中的特定节点 "单击" 时，与该节点关联的所有地址指针都将被激活，相应文档的标题以及最终的一些附加如描述性词语等信息被写出。在此之后，所需文档可以全部读出。

图 4显示大型 SOM 部分的特写视图。在映射的某些位置附近写入的关键字由 Lagus 和 Kaski中解释的自动注释过程确定(1999)。

映射通常提供一个表单域，用户可以在其中键入查询。我们将举例说明使用表单域与下一个示例有关的情况。

5。3。3。 近700万项专利摘要的 SOM

接下来，我们描述如何实现从一个非常大的文本语料库中进行内容寻址搜索操作。如前所述，为此，必须向映射界面提供一个表单域，用户可以在其中键入查询，例如，一组关键字或描述性句子。此查询的加权词汇应与某些文档的加权词汇表匹配。

此应用程序中文本语料库的大小 (Kohonen et，2000) 大约是大英百科全书的20倍。它包括6840568个专利摘要用英语写。他们有大约200张光盘，并获得从美国，欧洲和日本专利局作为两个数据库: "第一页" 数据库 (1970–1997) 和 "日本专利文摘" (1976–1997)。摘要的平均长度为132字。在省略了 stopwords、数字和非常罕见的词之后，词汇量的大小最终被接受了43222字，它们被香农熵加权。

SOM 的选定大小为1002240个节点，每个模型的维数 (在形成加权词直方图的随机投影之后) 为500。为了使这个大小的地图适合我们实验室的中央计算机，我们只能为每个向量元素预留一个字节。与500向量元素，准确性仍然是足够的统计比较。这是最大的 SOM，我们的知识有

|  |
| --- |
| 图 5。 背景中描述了近700万个专利摘要的文档结构图，查询的表单域显示在右上角。该查询是 "角膜激光手术"，二十最佳匹配节点已标记为小圆圈。在主地图中，这些圆圈融合在一起。地图的放大部分，其中二十个节点是单独可见的，可以通过数字变焦获得，如左下角的小圆圈所示。通过单击特定节点，可以获得与此节点关联的标题列表，并通过单击所选标题来读取相应的摘要。 |

曾经构建，即使我们使用了许多不同的加速方法，缩短了几十年的计算时间，它花了六周时间来构造地图。另一方面，搜索操作可以在几秒钟内完成，因此地图可以实时使用。图 5 以实例化内容寻址搜索为例。

5。4。 人内源性逆转录

只有大约两的人类 DNA 控制蛋白质的合成。其余的功能在很大程度上是未知的。约八由称为人内源性逆转录 (HERVs) 的部分组成。该名称意味着逆转录病毒序列是内源性的，即在基因组内，在人类。这些 HERVs 是古代感染的残余逆转录，如艾滋病毒。逆转录病毒已经融入了祖先物种的基因组，从而成为它的一部分。提示这些逆转录病毒序列在调节某些基因的活性方面获得了人类基因组的新功能。它们也可能在某些条件下产生蛋白质。

HERVs 茎从几种逆转录。当前的哈维分类是不完整的: 某些序列不能分配给任何类，并且分类对其他类型是不明确的。

HERVs 的相互关系及其与其他相关 DNA 元素的相似性研究了欧雅，Somervuo 等 (2003); 欧雅等 (2004，2005)，使用 SOM 技术。用批计算算法计算了3661哈维序列的集合图，并用 FASTA 方法对整个 DNA 序列之间的距离进行了评价，参见： 皮尔逊 (1999) 和皮尔逊和李普曼(1988)。

DNA 片段的聚类显示在图 6。许多不同类型的簇是可见的。它们在代表 SOM 节点的六边形周围用灰色阴影隔开。U矩阵技术 (Ultsch，1993 ) 已着色了群集边框)。

聚类可视化了与传统生物分类和系统进化分类法已知的相同类型的关系。此外，SOM 检测到一个全新的epsilonretroviral 序列组，并能够揭示三以前存在的哈维家族之间的关系。SOM 检测到一组 ERV9、HERVW 和 HUERSP3 序列，这表明 ERV9 和 HERVW 序列可能有共同的起源。

图 6。 人类内源性逆转录的 SOM。图中的标签是为映射的不同区域手动分配的名称。这些标签描述了每个区域中序列的类 (如 HERVADP、HERVH、HERVRb 等类名称)，从开始时将哈维删除。问号用于标记大多数序列未分类的区域。灰色尺度着色描述地图单位之间的距离;黑色表示大距离和白色小。最黑暗的边界划分三个主要小组和打火机边界不同的小组在主要小组之内，分别。

6。 模型的线性混合对输入数据项的逼近

在本章中介绍了迄今普遍未知的分析; 参见Kohonen(2007) 和Kohonen(2008)。其目的是通过显示，而不是单个赢家模型，来扩展 SOM 的使用，通过一组多个模型共同更准确地定义输入数据项，可以更准确地近似输入数据项。应强调的是，我们不是指根据匹配顺序排序的k优胜者。相反，输入数据项通过优化的线性混合模型 (使用非线性约束) 来逼近。

考虑n维 SOM 模型mi ，i = 1 ，2 ,。。。，p ，其中p是 SOM 中的节点数。他们的一般线性混合物被写作为

k 1 m 1 + k 2 m 2 + ··· + kp m p = M k ，(10) ki 是标量值加权系数时，k是由它们形成的p维度列向量，m是以mi 为其列的矩阵。现在，M k应为某些输入向量x的估计。向量拟合错误是然后

e= M k−x 。 (11)

我们的目标是将e的范数最小化为最小二乘法。但是，在此优化中必须考虑特殊的非线性约束。

最近，人们非常关注将拟合系数限制为非负值的最小二乘法问题。这样的约束是自然的，当项的负值没有意义时，例如，当输入项由只能具有非负值的统计指示器组成，或者是文档的加权词直方图时。至少在这些情况下，约束包含的附加信息将使匹配更有意义。

图 7。 SOM 模型的线性混合，适合新的未知文档。混合中的加权系数ki 通过使用适当模型的灰色相对阴影的着色来显示。

6。1。 lsqnonneg 功能

当前的拟合问题属于更广泛的二次规划或二次优化，近年来已开发了许多方法。muchapplied 的单通算法基于Kuhn–Tucker 定理，如劳森和汉森 (1974) 中所解释的那样，但它涉及的太多，无法在此处全面审阅。请注意，它已在 MATLAB 中实现为名为lsqnonneg的函数。下面，变量k、 M和x必须理解为以 MATLAB 格式定义。然后，我们获取权重向量k作为

k = lsqnonneg (M，x)。 (12)

对于矩阵M的任意秩，可以计算lsqnonneg函数，结果将有意义。然而，必须承认，存在一个罕见的理论案例，其中最优解不是唯一。如果最终最佳混合中 的某些mi 是线性从属，则会发生此情况。实际上，如果对 SOM 的输入数据项是随机的，那么最优解的概率不唯一是微不足道的。无论如何，非零权重的位置是唯一的，即使在这种情况下也是如此!

6。2。 一个文件的描述由 SOM 模型的线性混合物

以下分析适用于大多数 SOM 应用程序。这里是文本数据库的例子。

在文本分析中，一个可能的任务是找出一个文本是否来自不同的来源，因此它的词直方图预期是其他已知直方图的线性混合。

演示如何将模型的线性混合与给定文档进行拟合的示例基于 lsqnonneg 函数。本文的语料来源于路透社的数据，如上文所述。

图 7显示一个典型示例，其中 SOM 模型的线性混合被安装到新的未知文档中。混合中加权系数ki 的值由 SOM 模型的适当位置上的灰度相对深浅点显示。应强调的是，此管接头过程还定义了非零系数的最佳编号。在大量文档收集实验中，这个数字通常很小，不到模型数量的百分比。

当模型在已知的类中下降时，线性混合中模型的权重也表示与该模型关联的类标签的权重。因此，通过对各类类标签的权重求和，然后获得与各种类的输入的类加入。

7。 讨论

自组织地图 (SOM) 原理作为一种分析和可视化工具在探索性数据分析中得到了广泛的应用。从工业过程控制和财务分析到大型文档馆藏的管理，都有大量的实际应用。生物信息学中存在着新的、非常有前途的应用。迄今为止，在文本文件的管理和检索中，最大的应用有两个大尺度的例子。

多个商用软件包以及 SOM 上的大量免费软件都可用。本文作者强烈鼓励使用由 SOM\_PAK 团队 (1990) 和 SOM 工具箱团队 (1999) 开发的两个公共域软件包。这两个包都包含辅助分析程序，特别是使用 MATLAB 功能的 SOM 工具箱，它提供了良好的和通用的图形以及经过彻底验证的统计分析程序的结果。

本文应用了 SOM 的基本版本，大多数应用都是基于此。尽管如此，在 SOM 的不同版本中可能至少存在理论兴趣，其中一些修改已经引入。

在 SOM 网格之上的始终被视为二维、正则和六边形。如果目的是在一个图像中可视化整个数据库的总体结构，则此数组的这种形式是有利的。一个不同版本的网格 (例如，由绽放团队 (2005)) 是循环。如果数据本身具有循环结构，或者目的是为了避免节点的直链数组的边界影响，这样的 "拓扑" 可能会有一些意义。如果 SOM 用于过程控制，则可能是这种情况，以便对所有可能的进程状态进行连续和平滑的描述。

另一种通常建议的 SOM 版本是用结构化图形节点来替换常规网格，其中节点的结构和数目是动态确定的; cf，例如,Fritzke(1994)。

然后，当然，出现了一个问题，是否可以根据相当不同的数学原理定义一个 SOM 样的系统。其中一个有趣的建议是主教、Svensen 和威廉姆斯 (1998) 中引入的生成地形图映射 (GTM)。它是基于网格中节点拓扑关系的直接计算。 Van Hulle (2000) 在构造 SOM 拓扑时使用了 informationtheoretic 措施，这是一种不同的、理论上更深层次的方法。

也许，基本 SOM 算法的主要优点之一是，在合理的时间内，只使用个人电脑就可以计算出真正的大映射。

最后，我们必须提醒，在数据项之间的相似关系表示的传统方法是群集它们根据一些相似性或距离度量。由 Anderberg (1973)，哈迪根 (1975)，和Dubes (1988) 所描述的经典聚类算法()，桫椤和贝利 (1973)，但是，通常是相当繁重的计算，因为每个数据项都必须与所有其他的，也许 reiteratively。对于大量的数据，这显然是耗时。SOM 提供的补救方法是通过一组更小的模型来表示所有数据项集，每一个都代表相似或几乎相似的数据项的子集。在不确定数量的输入数据项的分类中，它们与模型的比较至少可以是数量级或更轻的操作。

尽管如此，我们对 SOM 研究的最初动机是试图解释生物中枢神经系统中发现的抽象特征图。然而，SOM 算法的生物学方面和意义在本文中几乎完全被忽略了。最近的 SOM 研究的主要目标是为实际的数据挖掘应用开发算法和计算程序。

确认

作者感谢他的所有合作者，这些年来都实施了 SOM 程序包和应用程序。Merja 博士欧雅亲切地提供了有关最近哈维研究的图片和相关材料。

引用

Allinson，Allinson，L，& 松弛，J (Eds) (2001)。自组织映射的前进。伦敦，英国: Springer。

阿玛瑞 (1980)。神经场的地形组织。数学生物学通报，42，339–364。

Anderberg，m。 (1973)。应用程序的群集分析。纽约，纽约州: 学术。

主教，沙菲·萨米，Svensen，m，& 威廉斯，c。k。i。 (1998)。GTM: 生成地形图图。神经计算，10，215–234。

程，Y (1997)。Kohonen 的批处理映射的收敛性和排序。神经计算，9，1667–1676。

科特雷尔，m，& 堡，1987。Étude d ' 联合国突汽车组织。年鉴。 庞加莱，23，1–20。

科特雷尔，m，堡垒，& Pagés，G。 (1997)。SOM 算法的理论方面。在WSOM 97 的程序中，自组织映射讲习班。赫尔辛基理工大学。神经网络研究中心。埃斯波，芬兰 (pp 246–267)。

Deboeck，G，& Kohonen，T。 (1998)。具有自组织映射的财务可视化探索。伦敦，英国: Springer，pp，246–267。

Deerwester，Dumais，富尔纳斯，G，& Landauer，K。 (1990)。通过潜在语义分析进行索引。美国信息科学学会学报，41，391–407。

佩里斯 (1850)。超级死 Reduktion positiven quadratischen Formen 麻省理工学院 drei unbestimmten ganzen Zahlen。瑞尼《应用Mathematik，40，209–227。

Forgy，e。w。 (1965)。多变量数据的聚类分析: 效率与分类解读。生物统计学，21，768。抽象。

Fritzke，B。 (1994)。生长细胞结构--一个自我组织的网络，用于无监督和监督学习。神经网络，7，1441–1460。

Gersho (1979)。关于向量量化器的结构。信息理论上的 IEEE 交易记录，IT-25，373–380。

灰色，rm 控股公司 (1984)。矢量量化。IEEE ASSP 杂志，1，4–29。

Grossberg，S。 (1976)。关于视觉皮层特征检测器的开发及其在学习和 reaction–diffusion 系统中的应用。生物控制论，21，145–159。

锤子，B，西蒙米凯利，Sperduti，& Strickert，m (2004)。递归 selforganizing 网络模型。神经网络，17，1061–1085。

哈迪根，J。 (1975)。群集算法。纽约，纽约州: 威利。

Heskes，t。m。d-第三，& 弗兰克林·范卡彭，B。 (1993)。自组织的错误潜能。在ICNN'93 会议上，国际神经网络大会，卷 III (pp 1219–1223)。皮斯卡塔韦，新泽西州: IEEE 服务中心。

Hubel，& 伊利·威塞尔，金斗勋 (1962)。在猫的视觉皮层接受的领域，双目和功能结构。生理学期刊，160，106–154。

Dubes，ak，&，钢筋混凝土 (1988)。数据聚类的算法。恩格尔伍德悬崖，新泽西州: 学徒大厅。

Kaski (1998)。通过随机映射减少维数。在关于神经网络的国际联席会议的 IJCNN'98 会议上(pp 413–418)。皮斯卡塔韦，新泽西州: IEEE 出版社。

Kaski，Honkela，Lagus，K & Kohonen，t。 (1998)。WEBSOM-自组织文档集合的映射。神经计算，21 (1)，101–117。

Kaski，S，Kangas，J，& Kohonen，T。 (1998)。自组织地图 (SOM) 文献目录: 1981–1997。神经计算调查，1，1–176。以电子形式在 [http://www。cis。hut。fi/research/som-bibl/vol1\_4。pdf。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fwww.cis.hut.fi%2Fresearch%2Fsom-bibl%2Fvol1_4.pdf)中提供。

Kohonen (1982a)。自组织形成的拓扑正确的特征图。生物控制论，43，59–69。

Kohonen (1982b)。模式的聚类、分类和拓扑图。在模式识别国际会议论文集(114–128)。华盛顿特区: IEEE 计算机社会学课上出版社。

Kohonen (1985)。中值字符串。模式识别字母，3，309–313。

Kohonen (1989)。自组织和关联内存(第三)。Berlin–Heidelberg–Germany: Springer。

Kohonen (1990)。自组织地图。IEEE、 78、1464–1480的程序。

Kohonen (1991)。自组织映射: 优化方法。在 Kohonen，K。 Mäkisara，仿真，& j Kangas (Eds),人工神经网络，II (pp 981–990)。荷兰阿姆斯特丹: 北荷兰。

Kohonen (1995)。自组织中的不变特征检测器的出现。

在 m。 Palaniswami，Attikiouzel，r。j。 标记 II，d 福格尔，& t。 福田 (Eds),计算智能，动态系统透视(pp。 17–31)。纽约，纽约州: IEEE 出版社。

Kohonen (1996)。adaptivesubspace 自组织 ma 中不变特征探测器的出现。生物控制论，75(4)，281–291。

Kohonen (2001)。自组织映射(第三版)。德国 Berlin–Heidelberg: Springer。

Kohonen (2005)。逐点组织投影。在WSOM05 的会议上，第五关于自组织地图的讲习班先贤祠-索邦大学。巴黎，法国 (pp 1–8)。也可在 [http:/samosuniv-paris1。 frwsomwsom05。 html 。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fsamos.univ-paris1.fr%2Fwsom%2Fwsom05.html)

Kohonen (2006)。自组织神经投射。神经网络，19，723–733。

Kohonen (2007)。通过 SOM 模型的线性混合描述输入模式。在WSOM 2007 cd-rom 程序中，德国比勒费尔德。比勒费尔德大学。也可在 [http://biecoll。ub-bielefeld。de。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fbiecoll.ub-bielefeld.de%2F)

Kohonen (2008)。通过自组织映射进行数据管理。在 Zurada，阿吉日元，& j。 王 (Eds),计算智能: 研究前沿(pp。 309–332)。柏林，海德堡，纽约: Springer。

Kohonen，Hynninen，j，Kangas，j，& Laaksonen，j。 (1996)。自组织映射程序包，报告 A31。赫尔辛基理工大学计算机与信息科学实验室。埃斯波，芬兰。

Kohonen，Kaski，Lagus，K，Salojärvi，j，Honkela，j，Paatero，v，& Saarela，(2000)。大量文档收集的自组织。神经网络上的 IEEE 事务，11，574–585。

Kohonen，Kaski，& Lappalainen，H (1997)。自组织形成的各种不变特征滤波器在适应子空间 SOM。神经计算，9，1321–1344。

Kohonen，欧雅，仿真，签证，& Kangas，J。 (1996)。自组织映射的工程应用。IEEE、 841358–1384的程序。

Kohonen & Somervuo，P。 (2002)。如何为 nonvectorial 数据制作大型自组织地图。神经网络，15，945–952。

Kohonen & 星，H。 (2011)。语境自组织的汉语词汇图。在 j Laaksonen & t Honkela (Eds) 中,自组织映射的进展(pp 16–29)。柏林，海得尔堡: Springer。

克鲁斯卡尔，j。b。，& 愿望，m (1978)。( 07-011) 《社会科学论文》关于定量应用的系列文章，在社会学中,多维度缩放。。伯里克公园，CA: 圣人出版物。

Lagus，K，& Kaski，S。 (1999)。用于描述文本文档映射的关键字选择方法。在ICANN'1999 的会议上，第九届国际人工神经网络大会。 卷 1 (pp 371–376)。伦敦，英国: IEE。

Lampinen，J，& 欧雅，E。 (1989)。空间和时间 AR 模型的自组织映射。在 m。 Pietikäinen，& J。 Röning (Eds),第六 SCIA 的会议，图像分析丑闻大会(pp 120–127)。芬兰赫尔辛基: Suomen Hahmontunnistustutkimuksen Seura。

劳森，铁匠，& 汉森，r。j。 (1974)。解决最小二乘问题。恩格尔伍德悬崖，新泽西州: 学徒大厅。

刘易斯，迪迪，杨，Y，玫瑰，t。g。，& 李，T。 (2004)。RCV1: 一种新的文本分类研究基准集。机器学习研究杂志，5，361–397。

林德，Y，Buzo，& 灰色，rm 控股公司 (1980)。矢量量化的一种算法。通信上的 IEEE 事务、 COM-28、84–95。

林，x，Soergel，D，& Marchionini，G。 (1991)。一种用于信息检索的自组织语义映射。在信息检索方面的第十四届国际 SIGIR & D 会议的议事程序(pp 262–269)。纽约:

劳埃德，某地 (1982)。PCM 中最小二乘法的量化。信息理论上的 IEEE 交易记录，28(2)，129–137。贝尔电话实验室纸。1957，转载于。

克里斯托弗鲁特瑞尔，某地 (1992)。备忘录 4669。英国马尔文: 国防研究局。

Makhoul，J，Roucos，& 地理信息系统，H。 (1985)。语音编码中的矢量量化。IEEE、 PROC-73、1551–1588的程序。

马尔斯堡，性病 (1973)。条纹皮层中定向敏感细胞的自组织。Kybernetik，14，85–100。

曼宁，c。d。，& Schütze，H。 (1999)。统计自然语言处理的基础。剑桥，马萨诸塞州: 麻省理工学院出版社。

Merkl (1995)。基于内容的文档分类，具有高度压缩的输入数据。在 Fogelman-Soulié，加里纳利，P。在关于人工神经网络的国际会议 ICANN'95 的程序中，第二卷。南泰尔，法国，EC2 (pp 239–244)。

Merkl，D，Tjoa，m，& 卡普佩尔，G。 (1994)。一种自组织映射，它学习可重用软件组件的语义相似性。在交流蔡，& t (Eds)，ACNN'94 的会议，第五澳大利亚神经网络大会(pp。 13–16)。澳大利亚: 布里斯班。

梅策尼，m。m。，Kaas，兰伯特，壁，j。t，纳尔逊，r。j。,，m，& Felleman，dj (1983)。成人猴体感皮层区3b 和1的地形重组。神经科学，8，33–55。

Miikkulainen 河 (1993)。Subsymbolic 自然语言处理: 脚本、词典和内存的集成模型。剑桥，马萨诸塞州: 麻省理工学院出版社。

Miikkulainen，Bednar，& Sirosh，J。 (2005)。可视皮层中的计算映射。纽约，纽约州: Springer。

Mountcastle，五 (1957)。猫躯体感觉皮层单神经元的形态和地形特征。神经生理学期刊，20，408–434。

名，A，Ratnatunga，k。u。，& 格里菲斯，鹰爪 (1997)。无分类的星系形态学;自组织映射。天体物理学期刊补充系列，111，357–367。

nas，m。m。，& 库珀，l。n。 (1975)。视觉皮层特征 departecting 细胞发育的理论。生物控制论，19，1–18。

Obermayer，K，& Sejnowski，T。 (2001)。自组织映射形成: 神经计算的基础。剑桥，马萨诸塞州: 麻省理工学院出版社。

欧雅，E，& Kaski，S。 (1999)。Kohonen 映射。荷兰阿姆斯特丹: Elsevier。

1998–2001附录欧雅，m，Kaski，S，& Kohonen，T。 (2003)。selforganizing 地图 (SOM) 文件书目。神经计算调查，3，1–156。以电子形式在 [http://www。cis。hut。fi/research/som-bibl/NCS\_vol3\_1。pdf。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fwww.cis.hut.fi%2Fresearch%2Fsom-bibl%2FNCS_vol3_1.pdf)中提供。

欧雅，m，Somervuo，Kaski，& Kohonen，T。 (2003)。人内源性逆转录病毒序列与中位自组织图谱的聚类研究。在WSOM'03 的程序中，自组织映射讲习班。Hibikino，日本。

欧雅，m，斯珀伯，G，布隆博格，J，& Kaski，S。 (2004)。通过引导中位自组织映射，对人内源逆转录进行分组和可视化。在CIBCB 2004 IEEE 生物信息学和计算生物学计算智能研讨会。7–8 10月。圣地亚哥，美国 (pp 95–101)。

欧雅，m，斯珀伯，G，布隆博格，J，& Kaski，S。 (2005)。基于自组织地图的人类内源性逆转录病毒序列组的发现和可视化。

国际神经系统杂志，15(3)，163–179。

皮尔逊 (1999)。FASTA 程序包。 [ftp://ftp。virginia。edu/pub/fasta。](ftp://ftp.virginia.edu/pub/fasta)

皮尔逊，W，& 李普曼，D。 (1988)。改进的生物序列比较工具。美利坚合众国国家科学院学报，85，2444–2448。

佩雷斯，r，玻璃，r。l。，& Shlaer，r。 (1975)。猫视皮层特异性的发展。数学生物学学报，1，277–288。

Pöllä，m，Honkela，t，& Kohonen，t。 (2009)。SOM 论文的书目。可用于 [http://users。ics。tkk。fi/tho/online-papers/TKK-ICS-R23。pdf。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fusers.ics.tkk.fi%2Ftho%2Fonline-papers%2FTKK-ICS-R23.pdf)

瑞德，H，& Kohonen，T。 (1989)。自组织语义映射。生物控制论，61，241–254。

瑞德，H，Martinetz，& Schulten，K。 (1992)。神经计算和自组织映射: 介绍。读，马: 艾迪生-韦斯利。

罗宾斯，H，& 蒙罗，S。 (1951)。随机逼近法。数理统计的史册，22，400–407。

沙顿，G，& 麦吉尔，mj (1983)。现代信息检索入门。纽约，纽约州: 麦格劳希尔。

Sammon，j。w。 (1969)。数据结构分析的非线性映射。计算机上的 IEEE 事务、 C-18、401–409。

Scholtes (1991)。自然语言处理中的无监督语境学习。在IJCNN'91 的会议上，国际神经网络大会，卷I (pp。 107–112)。皮斯卡塔韦，新泽西州: IEEE 服务中心。

塞弗特，美国，&，妇联 (Eds) (2002)。自组织神经网络: 最近的进展和应用。德国海德堡: 物理学-出版社。

SOM\_PAK 团队 (1990)。 [http://www。cis。hut。fi/research/som\_lvq\_pak。shtml。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fwww.cis.hut.fi%2Fresearch%2Fsom_lvq_pak.shtml)[http://www。cis。hut。fi/research/som\_pak/som\_doc。ps 。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fwww.cis.hut.fi%2Fresearch%2Fsom_pak%2Fsom_doc.ps)

SOM 工具箱团队 (1999)。 [http://www。cis。hut。fi/projects/somtoolbox/documentation/。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fwww.cis.hut.fi%2Fprojects%2Fsomtoolbox%2Fdocumentation%2F)[http://www。cis。hut。fi/projects/somtoolbox/package/papers/techrep。pdf 。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fwww.cis.hut.fi%2Fprojects%2Fsomtoolbox%2Fpackage%2Fpapers%2Ftechrep.pdf)

开花队 (日本) (2005)。日本高级有限公司开花软件包。 [http://www。somj。com/。](https://www.microsofttranslator.com/bv.aspx?from=en&to=zh-CHS&a=http%3A%2F%2Fwww.somj.com%2F)

菅，w。e。 (1979)，& 奥尼尔。神经轴代表了小胡子蝙蝠听觉皮层的目标范围。科学，206，351–353。

Tokutaka，H，田文雄，& 藤村，K。 (1999)。自组织映射的应用-ulti 维信息的二维可视化。日本东京: Kaibundo (日语)。

Tokutaka，H，Ookita，m，& 藤村，K。 (2007)。SOM 和应用程序。日本东京: 日本 (日本)。

桫椤，R，& 贝利，D。 (1973)。群集分析。纽约，纽约州: 麦格劳希尔。

Tunturi，a。r。 (1950)。犬中 ecrosylvian 听觉区传入连接排列的生理测定。美国生理学杂志，162，489–502。

Tunturi，a。r。 (1952)。狗的听觉皮层。美国生理学杂志，168，712–717。

Ultsch (1993)。自组织神经网络用于可视化和分类。在 o Opitz，b Lausen，& r 克拉尔 (Eds),信息和分类(pp 307–313)。德国柏林: Springer。

范 Hulle，m。 (2000)。忠实表示和地形图: 来自 distortionto 基于信息的自组织。纽约，纽约州: 威利。

Vesanto，j，Alhoniemi，E，Himberg，j，Kiviluoto，K，& Parviainen，j。 (1999)。Selforganizing 数据挖掘地图在 MATLAB 中: SOM 工具箱。欧洲模拟新闻。25:54，3月。

Vesanto，j，Himberg，j，Alhoniemi，E，& Parhankangas，j。 (1999)。自组织映射在 Matlab 中: SOM 工具箱。在Matlab DSP 会议程序中。11月 16–17，埃斯波，芬兰 (pp 35–40)。

Voronoi，G。 (1907)。新生力量应用 des paramétres 美国栌 théorie 辞章 quadratiques。瑞尼《应用Mathematik，133，97–178。

Zavrel，J。 (1995)。神经信息检索--用神经网络对文档集进行聚类和浏览的实验研究。硕士论文。荷兰阿姆斯特丹。阿姆斯特丹大学。

福莱尔斯 (1980)。大脑皮层中颜色的表示。自然，284，412–418。