

自组织映射拓扑保持的增强

周向东

(福州大学 数学与计算机科学学院, 福州 350108)

(zhou_xiangdong@hotmail.com)

摘 要:在自组织映射(SOM)中,网格各单元的权值向量仅仅是根据各单元和最佳匹配单元(BMU)之间的距离进行更新的,因而输入数据间的拓扑关系不能得到很好的保持。为此提出了两种改进方案。在第一种改进方案中,各单元的权值向量根据各单元和 BMU 之间对应各坐标的差进行更新。实验结果表明,这种改进方案可以很好地保持拓扑关系,但输入数据的分布密度却不能得到较好的体现。在第二种改进方案中,各单元的权值向量同时根据各单元和 BMU 之间对应各坐标的差与距离进行更新。实验结果表明,这种改进方案不仅能使拓扑关系得到比 SOM 更好的保持,而且较好地体现了输入数据的分布密度,并加快了训练的收敛速度。

关键词:自组织映射;拓扑保持;最佳匹配单元;权值向量;分布密度

中图分类号: TP183 **文献标志码:** A

Enhancement of topology preservation of self-organizing map

ZHOU Xiang-dong

(School of Mathematics and Computer Science, Fuzhou University, Fuzhou Fujian 350108, China)

Abstract: In the Self-Organization Map (SOM), the weight vectors of the units in the grid are updated only according to the distance between the units and the Best Matching Unit (BMU), so the topological relationship between input data can not be preserved very well. Therefore, two improved schemes were proposed. In the first improved scheme, the weight vectors of the units were updated according to the differences of the corresponding coordinates between the units and the BMU. Experimental results show that this improved scheme can preserve topological relationship very well, but the distribution density of the input data can not be reflected quite well. In the second improved scheme, the weight vectors of the units were updated both according to the differences of the corresponding coordinates and the distance between the units and the BMU. Experimental results show that this improved scheme can not only preserve topological relationship better than SOM, but also reflect the distribution density of the input data quite well and accelerate the convergence speed of the training.

Key words: Self-Organization Map (SOM); topology preservation; Best Matching Unit (BMU); weight vector; distribution density

0 引言

自组织映射(Self-Organizing Map, SOM)^[1-2]算法可用于构建一个从输入数据空间(通常是高维的)到单元阵列(通常是低维的)的有序映射。这种映射在一定程度上保留了输入数据间的拓扑关系。这意味着 SOM 算法能够从大量复杂的数据中抽取必要的信息。从应用信息处理的角度看, SOM 算法可以作为一个通用的非线性主成分分析工具,并证明了其在各种复杂数据的可视化、压缩和挖掘等领域的价值。

传统的 SOM 算法存在着许多不足的地方。首先是需要预先给定网格单元数目及其结构形状。这就需要通过实验对不同的网格大小进行比较,以获得最优的结果,而这又是非常耗时的。为此,人们提出了多种在训练过程中动态确定单元数目和网络形状的解决方案。如文献[3]提出的生长自组织映射算法(Growing Self-Organizing Map, GSOM),文献[4]提出的增长细胞结构算法(Growing Cell Structure, GCS)等。其次是使用 SOM 时,几乎都有一个在所有节点中搜索最佳匹配单元(Best Matching Unit, BMU)的步骤。当分析大型数据集时,这会使得 SOM 的运行变得很慢。为此,人们通过构建有效的搜索结构来加快对 BMU 的搜索,如文献[5]的树结构自组织映射(Tree-Structured Self-Organizing Map, TS-SOM)和文献[6-7]提出的进化树(Evolving Tree, ETree)等。另外, SOM 对拓

扑关系的保持也不能令人满意,虽然提出了很多定性和定量的衡量方法^[8-13],但很少有这方面的改进措施。

SOM 保持拓扑关系的关键之处在于领域关系的确定。而领域关系的远近在 SOM 中是根据 BMU 和其他单元之间的距离远近来确定的。在对各单元的权值向量进行更新时,领域关系越近,所做的更新也越大。然而,仅凭距离关系并不能完全刻画各单元之间的拓扑关系。因此, SOM 对各单元的权值向量所做的更新就有很大的局限性。反映在图形上,就是当把各单元以其权值向量的各个分量为坐标放到输入数据空间时,所形成的图形和它们在网格上所形成的图形相比出现了较大的变形,因而拓扑关系也就没有得到很好的保持。

为此,本文提出了两个方案对 SOM 作了改进。在改进方案一中,不是根据各单元和 BMU 之间的距离关系,而是根据它们之间对应各个坐标的关系,来决定对各单元权值向量所作的更新。在改进方案二中,不仅根据它们之间对应各个坐标的关系,同时也根据它们之间的距离关系,来决定对各单元权值向量所作的更新。

改进方案一使输入数据的拓扑关系得到了很好的保持,但输入数据的分布密度却没有得到较好的体现。改进方案二不仅使输入数据的拓扑关系得到了较好的保持(虽然没有方案一那么充分),而且也使输入数据的分布密度得到了较好的体现。

1 SOM 及其不足

SOM 可把高维的数据非线性地投影到低维网格上。这种投影保持了输入空间的拓扑关系,因而通过训练,相似的数据样本将被投影到网格上相同的或相邻近的神经元或单元上。

网格通常是二维的,由一个神经元阵列实现。在文献中,网格通常采用矩形和六边形这两种拓扑领域。以下讨论主要针对矩形拓扑领域。每个神经元和一个权值向量相联系,权值向量和输入空间具有相同的维数。具体地说,输入样本是欧几里得空间中的高维的实数向量, $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_n] \in \mathbf{R}^n$ 。用于一个 n 维训练数据集的 SOM 上的单元的权值向量也是 n 维的实数向量, $\mathbf{m}_i = [m_{i1} \ m_{i2} \ \cdots \ m_{in}] \in \mathbf{R}^n$ 。训练时,对每一个输入样本,从映射神经元中确定它的 BMU。确定一个输入样本 \mathbf{x} 的 BMU 的方法是识别和 \mathbf{x} 最相似的单元,通常采用一个距离函数来度量这种相似性,距离越小,它们越相似。计算距离的一个典型方法就是使用欧几里得距离函数:

$$d(\mathbf{x}, \mathbf{m}_i) = \|\mathbf{x} - \mathbf{m}_i\| = \left(\sum_{k=1, n} (x_k - m_{ik})^2 \right)^{1/2}$$

一旦识别了 BMU,就要对 BMU 及其邻居的权值向量进行更新以缩小它们和输入样本之间的差距。更新以 BMU 为中心,而调整的量随着各单元和 BMU 距离的增加而减少。对邻居单元 \mathbf{m}_i 的更新规则用公式表示就是

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

其中: $0 < \alpha(t) < 1$ 是学习速率函数; $h_{ci}(t)$ 是邻域函数,通常取作高斯函数 $h_{ci}(t) = \exp\left(\frac{-d(r_c, r_i)^2}{2\sigma^2(t)}\right)$ (c 是 BMU 的编号)。

$\alpha(t)$ 和 $h_{ci}(t)$ 的宽度随着训练步数的增加而逐渐减小。

从以上对 SOM 的描述可以看出,在确定 BMU 后,对各单元权值向量的更新是根据它们在网格中和 BMU 之间的距离来进行的。但是,距离并不能完全刻画各单元和 BMU 之间的位置关系。就二维情况而言,除了距离关系外,还有方位角关系。当然,也可以通过在网格中各单元和 BMU 之间 x 坐标的差和 y 坐标的差来完全刻画它们之间的位置关系。

由于距离没有完全刻画各单元之间的位置关系,因此,采用 SOM 后输入数据间的拓扑关系并没有得到很好的保持或者训练中收敛的速度比较慢。这种情况当输入数据空间和网格的维数相同时就更加明显。

为此,本文提出了两种方案对 SOM 作了改进。

2 两个改进方案

2.1 改进方案一

在改进方案一中,针对网格是二维的情况,笔者曾考虑根据各单元和 BMU 之间的距离和方位角来对各单元的权值向量进行更新,但训练的结果出现了发散。于是,改用各单元和 BMU 之间 x 坐标的差和 y 坐标的差进行更新。当输入数据空间也是二维时,用公式表示就是:

$$m_{i1}(t+1) = m_{i1}(t) + \alpha(t) h_{ci1}(t) [x_1(t) - m_{i1}(t)]$$

$$m_{i2}(t+1) = m_{i2}(t) + \alpha(t) h_{ci2}(t) [x_2(t) - m_{i2}(t)]$$

$$h_{ci1}(t) = \exp\left(\frac{-(hor_i - hor_c)^2}{2\sigma^2(t)}\right)$$

$$h_{ci2}(t) = \exp\left(\frac{-(ver_i - ver_c)^2}{2\sigma^2(t)}\right)$$

其中 hor 和 ver 分别表示单元在网格中的横坐标和纵坐标。

对在 $0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1$ 范围内随机均匀分布的二维数据集用 SOM 和上述方案一分别进行训练。数据集有 200 个训练样本,网格采用 $a \times b = 6 \times 6$ 的正方形阵列,取:

$$\alpha(t) = 0.5 \times \exp(-2 \times (t/t_m))$$

$$r = 0.5 \times \sqrt{(a-1)^2 + (b-1)^2}$$

$$\sigma(t) = 2 \times r \times \exp(-4 \times (t/t_m))$$

其中 t_m 为总的训练次数(这里取为 200)。训练结果如图 1 所示。

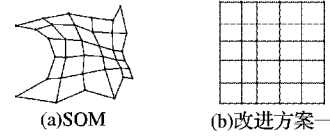


图 1 采用随机均匀分布数据集的训练结果

第 2 个数据集采用 x 方向和 y 方向的均值和方差为 $(\mu_1, \sigma_1, \mu_2, \sigma_2) = (0.5, 0.25, 0.5, 0.25)$ 的二维正态分布,共有 10000 个训练样本,网格采用 $a \times b = 10 \times 10$ 的正方形阵列,训练结果如图 2 所示。

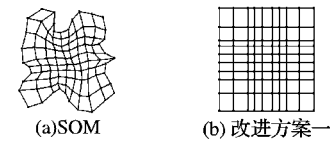


图 2 采用一个正态分布数据集的训练结果

第 3 个数据集采用 x 方向和 y 方向的均值和方差 $(\mu_1, \sigma_1, \mu_2, \sigma_2)$ 分别为 $(0.25, 0.1, 0.25, 0.1)$ 和 $(0.75, 0.1, 0.25, 0.1)$ 的两个二维正态分布,各有 5000 个训练样本,网格采用 $a \times b = 10 \times 10$ 的正方形阵列,训练结果如图 3 所示。

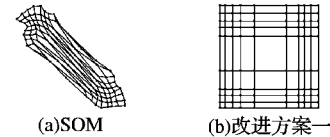


图 3 采用两个正态分布数据集的训练结果

从以上实验可以看出,经过训练,改进方案一比 SOM 具有更加规则的几何图形,每个格子都是标准的矩形。

对于拓扑关系的保持,提出了很多定性和定量的衡量方法^[8-13],采用较多的是由 Villmann 等提出的方法。矩形网格的拓扑保持定义如下^[13]:

设 A 是一个 d_A 维的矩形网格, $M \subseteq \mathbf{R}^d$ 是一个数据流形。 M 的一个映射 $M_A = (\Psi_{A \rightarrow M}, \Psi_{M \rightarrow A})$ 是拓扑保持的,如果从 M 到 A 的映射 $\Psi_{M \rightarrow A}$ 和从 A 到 M 的逆映射 $\Psi_{A \rightarrow M}$ 是邻域保持的。

1) 映射 $\Psi_{M \rightarrow A}$ 是邻域保持的当且仅当在 M 中相邻的位置 $\mathbf{m}_i, \mathbf{m}_j$ 根据 A 上的最大范数 $\|\cdot\|_{\max}$ 属于 A 中相邻的顶点 i, j 。

2) 映射 $\Psi_{A \rightarrow M}$ 是邻域保持的当且仅当根据 A 上的欧几里得范数 $\|\cdot\|_E$ 或范数 $\|\cdot\|_{\Sigma}$ 在 A 中相邻的顶点 i, j 分配到了 M 中相邻的位置 $\mathbf{m}_i, \mathbf{m}_j$ 。

$\mathbf{m}_i, \mathbf{m}_j$ 在 M 中相邻是指它们的 Voronoi 多面体的交集非空。2) 中和一个顶点相邻的有上、下、左、右 4 个顶点,而 1) 中还增加了对角线上的 4 个顶点。

根据以上定义,训练后网格越接近于矩形,拓扑关系就保持得越好。因此以上实验说明了采用改进方案一能使拓扑关系得到很好的保持,大大优于 SOM。

但从图形中也能看出,输入数据的分布密度在图 3 中没有得到体现。

2.2 改进方案二

为解决改进方案一的不足提出了第二种改进方案,即不仅根据各单元和 BMU 之间对应各个坐标的关系,同时也根据它们之间的距离关系,来决定对各单元权值向量所作的更新。用公式表示为:

$$m_{i1}(t+1) = m_{i1}(t) + \alpha(t) h_{ci}(t) h_{ci1}(t) [x_1(t) - m_{i1}(t)]$$

$$m_{i2}(t+1) = m_{i2}(t) + \alpha(t) h_{ci}(t) h_{ci2}(t) [x_2(t) - m_{i2}(t)]$$

其中: $h_{ci}(t)$ 中的 $\sigma(t)$ 和第 2.1 节中的相同, 而 $h_{ci1}(t)$ 和 $h_{ci2}(t)$ 中的 $\sigma(t)$ 取为:

$$\sigma(t) = r \times \exp(-4 \times (t/t_m))$$

采用改进方案二对以上三个数据集进行训练的结果如图 4 所示。

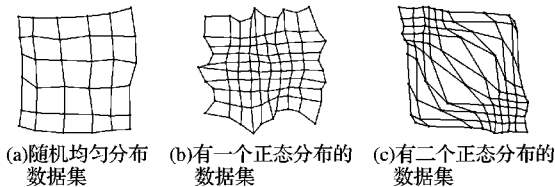


图 4 改进方案二的训练结果

从以上结果可以看出,采用改进方案二能比 SOM 得到更接近于矩形的网格,因而能使拓扑关系得到更好的保持。而且改进方案二也能体现出输入数据的分布密度,只是拓扑关系的保持不如方案一。

当输入数据空间的维数大于二维时,可以采用主成分分析法找出二个主成分,对这二个主成分采用方案一或方案二进行更新,而对其他成分仍采用 SOM 进行更新。

3 实验结果

以下实验用 VC++ 实现。

在第一个实验中,数据集采用 x 方向和 y 方向的均值和方差 ($\mu_1, \sigma_1, \mu_2, \sigma_2$) 分别为 (0.25, 0.1, 0.25, 0.1)、(0.75, 0.1, 0.25, 0.1) 和 (0.5, 0.1, 0.75, 0.1) 的 3 个二维正态分布,各有 5000 个训练样本,网格采用 $a \times b = 10 \times 10$ 的正方形阵列,训练结果如图 5 所示。

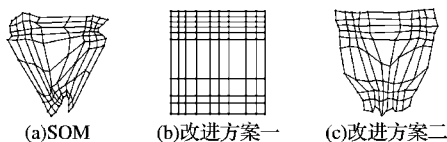


图 5 采用三个正态分布数据集的训练结果

可见,采用改进方案一虽然使拓扑关系得到了很好的保持,输入数据的分布密度却没有得到较好的体现,而改进方案二不仅对输入数据拓扑关系的保持优于 SOM,而且也使输入数据的分布密度得到了较好的体现。

在第二个实验中,把改进方案二用于扭曲问题^[14]。所谓扭曲问题是指网格各单元权值向量的初始值的分布出现了扭曲。对于这种情况采用 SOM,收敛的速度会变得很慢。实验中,仍采用文献[14]中的各参数值,即 $\alpha = 0.05$, $\sigma = 5$, 网格采用 $a \times b = 30 \times 30$ 的正方形阵列。而 $h_{ci1}(t)$ 和 $h_{ci2}(t)$ 中的 σ 取为 2.5, 训练样本 1000 个, 在 $0 \leq x_1 \leq 1$, $0 \leq x_2 \leq 1$ 范围内随机均匀分布。训练结果如图 6 所示。

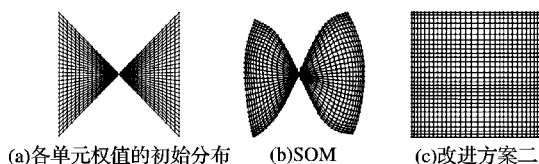


图 6 扭曲问题的训练结果

由此可见,改进方案二只用了 1000 个训练样本就使扭曲状态得到了很好的纠正。而 SOM 即使用 20000 个训练样本,其结果也和图 6(b)差不多。而在文献[9]中,即使采用改进的非对称领域函数,也需要大约 3000 个训练样本才能使扭曲状态得到纠正。这个实验充分说明了改进方案二大大加快了训练的收敛速度。

4 结语

SOM 实现了对输入数据间拓扑关系的保持,但这种保持是很不充分的,因为它仅仅根据网格各单元和 BMU 之间的距离关系对各单元的权值向量进行更新。本文提出了两种改进方案对 SOM 作了改进。在第一种改进方案中,根据各单元和 BMU 之间对应各个坐标的关系,对各单元的权值向量进行更新。实验表明该方案使拓扑关系得到了很好的保持,但没能较好地体现出输入数据的分布密度。在第二种改进方案中,对各单元权值向量的更新同时考虑了各单元和 BMU 之间对应各个坐标的关系与距离关系。实验表明该改进方案不仅使拓扑关系得到了优于 SOM 的保持,而且使输入数据的分布密度也得到了较好的体现,还加快了训练的收敛速度。

参考文献:

- [1] KOHONEN T. Self-organizing maps[M]. Berlin: Springer-Verlag, 1995.
- [2] KOHONEN T. The self-organizing map[J]. Proceedings of the IEEE, 1990, 78(9): 1464-1480.
- [3] ALAHAKOON D, HALGAMUGE S K. Dynamic self-organizing maps with controlled growth for knowledge discovery[J]. IEEE Transactions on Neural Networks, 2000, 11(3): 601-614.
- [4] FRITZKE B. Growing cell structures - A self-organizing network for unsupervised and supervised learning[J]. Neural Networks, 1994, 7(9): 1441-1460.
- [5] KOIKKALAINEN P, OJA E. Self-organizing hierarchical feature maps[C]// 1990 IJCNN International Joint Conference on Neural Networks. Washington, DC: IEEE, 1990: 279-284.
- [6] PAKKANEN J, IIVARINEN J, OJA E. The evolving tree - A novel self-organizing network for data analysis[J]. Neural Processing Letters, 2004, 20(3): 199-211.
- [7] PAKKANEN J, IIVARINEN J, OJA E. The evolving tree-analysis and applications[J]. IEEE Transactions on Neural Networks, 2006, 17(3): 591-603.
- [8] BAUER H-U, PAWELZIK K. Quantifying the neighborhood preservation of self-organizing feature maps[J]. IEEE Transactions on Neural Networks, 1992, 3(4): 570-579.
- [9] RITTER H, MARTINETZ T M, SCHULTEN K. Neural computation and self-organizing maps[M]. Upper Saddle River: Addison Wesley, 1992.
- [10] DER R, VILLMANN T. Dynamics of self-organized feature mapping[C]// Proceedings of the International Workshop on Artificial Neural Networks: New Trends in Neural Computation, LNCS 686. Berlin: Springer-Verlag, 1993: 312-315.
- [11] DEMARTINES P, BLAYO F. Kohonen self-organizing maps: Is the normalization necessary? [J] Complex Systems, 1992, 6(2): 105-123.
- [12] ZREHEN S. Analyzing Kohonen maps with geometry[C]// ICANN '93: International Conference on Artificial Neural Networks. Berlin: Springer-Verlag, 1993: 609-612.
- [13] VILLMANN T, DER R, HERRMANN M, et al. Topology preservation in self-organizing feature maps: Exact definition and measurement[J]. IEEE Transactions on Neural Networks, 1997, 8(2): 256-266.
- [14] TAKAKAKI A. Self-organizing feature maps with asymmetric neighborhood function[J]. Neural Computation, 2007, 19(9): 2515-2535.