

基于自增长型多级自组织映射网络的模式识别

傅彦 周俊临

(电子科技大学计算机科学与工程学院 成都610054)

摘要 以自组织映射网络为主要研究对象,描述了自组织映射网络的基本模型。在传统自组织映射网络的基础上,提出了基于自增长型多级自组织映射网络的模式识别方法,能够解决传统自组织映射网的静态结构带来的诸多问题,如在进行训练前必须预先确定网络的模型和神经元数目及其排列方式,若一次分类不准确将严重影响分析结果,等等。而且这种多级结构,还能将输入数据中存在的分级信息直观地表示出来,对于高维数据的分析尤其有利,因此自增长型多级自组织映射网络对大规模模式识别的研究一定会产生极大的促进作用。

关键词 自增长,模式识别,自组织映射网络,高维数据分析

Pattern Recognition Based on Self-Growing Multilevel Self-Organizing Map

FU Yan ZHOU Jun-Lin

(College of Computer Science and Engineering, UEST of China, Chengdu 610054)

Abstract Focusing on the self-organizing map, this paper presents a review on the basic model and puts forward a pattern recognition method that is based on self-Growing multilevel self-organizing map on the foundation of traditional self-organizing map. It can overcome many limitations, which are related to the static architecture of traditional model. For example, traditional model uses a fixed network architecture in terms of number and arrangement of neural processing elements, which has to be defined prior to training, also, if it has error in the classification for the first time, its results may not be corrected, etc. What is more, the new model can intuitively represent the hierarchical relations of the data, in particular, it can benefit the high dimension data analysis greatly. So, self-growing multilevel self-organizing map can promote the research of large scale pattern recognition greatly.

Keywords Self-growing, Pattern recognition, Self-organizing map, High dimensional data analysis

在人工神经网络所提出的各种算法和结构中,自组织映射网络具有能够快速、有效地建立输入信号不同特征的内部表示的特殊属性。除此之外更具新颖性的是,自组织过程还能发现自然语言句子中的语义关联。从这些方面来看,作为结果产生的自组织映射与高级动物大脑皮层中发现的拓扑组织映射非常相像。自组织映射网络经过优化修正模型向量,能像具有智能的生命体一样在各种含有大量噪声信号的模式识别领域中进行工作。这些映射的应用十分广泛,特别是在语音识别领域中已经应用到了实际,并且在机器人、过程控制和电信等领域中的应用也正在进一步向前推进。本文在传统 SOM 网络的基础上提出了自增长型多级自组织映射网,不仅具有传统 SOM 的优点,而且能克服其静态结构带来的许多局限性。

1 自组织映射网络

用来建模神经网络系统的网络结构和信号处理系统大致可分为三类,每一类都基于不同的哲学体系。(1)前馈式神经网络:将输入信号集变换到输出信号集,期望的输出变换经常由外部有监督的修正系统参数来决定。(2)反馈式神经网络:将输入信息定义为反馈系统初始的活动状态,经过状态转换,由渐进的最终状态确定计算结果。(3)自组织映射神经网络:神经网络中相邻的神经元通过相互的侧向作用在它们的活动相互竞争,并且自适应地发展成为不同信号模式的特殊探测器。它的学习称为竞争、无监督学习或自组织学习。

自组织特征映射是一个二维空间排列的人工神经网络,其中每个神经元通过无监督学习过程进行调整,使之对应于特定的不同输入信号模式或模式分类。从根本上而言,在同一时刻只有一个中心神经元或一组局部神经元会对当前输入产生积极的响应。并且,响应的位置区域倾向于越来越固定,就像通过网络建立了某种与不同输入特征对应的坐标系统。因此,神经元在网络中的空间位置或坐标就对应于特定域的输入信号模式,每个神经元或局部神经元组就像是对相同输入模式的一个单独的解码器。所以,与其说神经元在某个位置是否对输入有积极的响应是对输入输出信号进行了精确的变换,不如说它其实是对输入信号提供的各种模式信息进行了不同的解释。

1.1 神经元的侧向交互原理

目前对人脑的研究说明:大脑皮层中神经元是呈二维空间排列的,其输入信号很明显来自两个部分,外部区域的输入和同一区域的反馈输入。神经元之间的信息交互方式有很多种,不过研究表明:相邻近的神经元之间的局部交互的方式是侧向交互。这种侧向交互方式遵从下列有关规则:1)以发出信号的神经元为圆心,对近邻的神经元的交互作用表现为兴奋性侧反馈;2)以发出信号的神经元为圆心,对远邻的神经元的交互作用表现为抑制性侧反馈。

这种规则说明近邻者相互激励,远邻者相互抑制。一般而言,近邻是指从发出信号的神经元为圆心、半径约为50~

傅彦 硕士,副教授,主要从事计算智能方面的研究。

500 μm 左右的神经元;远邻是指半径为200 μm ~2mm左右的神经元。比远邻更远的神经元则表现为弱激励作用。这样,这种局部交互方式如图1所示。由于这种交互作用的曲线类似于墨西哥人带的帽子,所以也称这种交互方式为“墨西哥帽”。

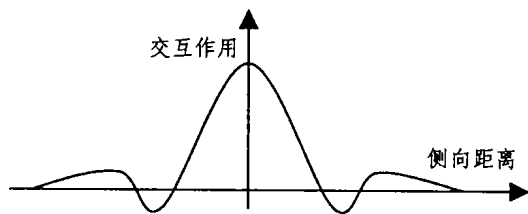


图1 侧向交互模式

神经网络中,邻近的各个神经元之间通过相互的侧向交互作用,从而产生相互竞争的过程,自适应地形成了针对特殊信息的组织结构。很明显,这种结构可以成为检测特殊信息的特殊检测器。这样,神经网络的自组织过程和行为,可以成为一种检测各种不同信息模式的检测识别器,这也是自组织特征映射的意义所在。

1.2 Kohonen 模型

依据大脑对信号处理的特点,在1981年,T. Kohonen 提出了一种神经网络模型,也就是自组织特征映射模型 SOM (Self-Organizing feature Map)。SOM 模型可以用二维阵列表示。其结构如图2所示。

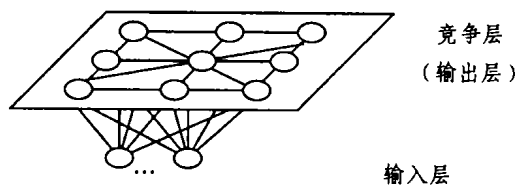


图2 二维阵列 SOM 模型

二维阵列神经网络由输入层和竞争层组成。输入层是一维的神经元。竞争层是二维的神经元。输入层的神经元和二维阵列竞争层的神经元每个都相互连接。二维阵列竞争层也称输出层。

在二维阵列竞争层中,可以清楚看出:每一个输出神经元都和最近相邻的8个神经元相连;当然,最边沿的神经元和3~5个神经元相连,但这只是最边沿的神经元才会这样。而从二维阵列内部一般有:每个输出神经元和8个最相邻的神经元相连。在 SOM 模型中,对于神经元 j ,其外部输入信号可以用 I_j 表示。

自组织映射的目的就是使神经元的权系数的形态表示可以间接模仿输入的信号模式。自组织特征映射 SOM 的学习算法是由两部分组成的,这两部分如下:

1. 最优匹配神经元的选择 设有输入模式 $x, x = (x_1, x_2, \dots, x_n)^T$, 对于自组织特征映射 SOM 网络的输出层神经元 j , 则有权系数向量 $w_j, w_j = (w_{1j}, w_{2j}, \dots, w_{nj})^T, j = 1, 2, \dots, n$ 。神经元的输出 y_j 的初始分布可能是随机的;但随着时间的变化,由于输出层神经元有侧向交互的作用, y_j 的分布就会因对环境的组织而形成“气泡”状。

输入模式 x 和权系数 w_j 的匹配程度是用两者的内积表示的,即用 $x^T w_j$ 表示。内积最大处正是“气泡”中心。当内积 $x^T w_j$ 最大时,则必定有 x 和 w_j 之间的向量差的范数 $\|x - w_j\|$ 最小,这个最小距离就确定了最优匹配的神经元 C ;从而有

“气泡”中心 C , 满足:

$$\|x - w_c\| = \min_{j=1}^n \|x - w_j\| \quad (1)$$

其中: w_c 是神经元 C 的权系数向量。这个式子也就是匹配规则,说明:气泡中心就是神经元 C ,其权系数向量 w_c 同输入模式 x 有最优匹配。

2. 网络权系数的自组织 在 SOM 网络中,每一个输出神经元都接收相同的输入模式 x 。对于输出神经元 j 来说,其最简单的输出是线性的;并且可以用下式表示:

$$I_j = \sum_{i=1}^n w_{ij} x_i \quad (2)$$

$$\frac{dy_j}{dt} = I_j + \sum_{k \in N_c} r_k y_k - g(y_j) \quad (3)$$

其中: $y(t)$ 是随时间 t 变化的递减增益函数; $y_j(t)$ 是输出神经元 j 的输出; $x_i(t)$ 是输入神经元 i 的输入; $w_{ij}(t)$ 是输入神经元 i 和输出神经元 j 在时间 t 时的权系数。对于气泡中心神经元 C ,以其为中心考虑一个邻近的区域 N_c 。 N_c 是以神经元 C 为中心的某一半径范围内全部神经元的集合。在 N_c 区域之内,所有神经元的输出为1;在 N_c 区域之外,所有神经元的输出为0。 r_k 是系数,它和权系数与横向连接结构有关; $g(y_j)$ 是非线性损失,如神经元饱和、分流和泄漏效应等。

在离散系统中,如果以 Δt 为采样周期 T ,则有 $\Delta t = 1T$ 。所以,权系数灵敏度调整方程写成

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t) [x_i(t) - w_{ij}(t)], j \in N_c \quad (4)$$

$$w_{ij}(t+1) = w_{ij}(t), j \notin N_c \quad (5)$$

对于增益函数 $\eta(t)$ 而言,它是一个随时间变化的递减函数。对于离散系统,则取

$$\eta(t+k) = \frac{1}{t+k}, k = 0, 1, 2, \dots, \infty \quad (6)$$

$$\text{或者 } \eta(t+k) = 0.2(1 - \frac{t+k}{10000}), k = 0, 1, 2, \dots, \infty \quad (7)$$

从上面可以看出:离散系统中,随时间的增加或采样周期的推移,增益函数 $\eta(t)$ 的值会越来越小。通常取 $500 \leq t \leq 10000$ 或 $500 \leq t+k \leq 10000$,结束自组织过程。

Kohonen 已经证明:在学习结束时,每个权系数向量 w_j 都近似落入到由神经元 j 所对应的类别的输入模式空间的中心,可以认为权系数向量 w_j 形成了这个输入模式空间的概率结构。所以,权系数向量 w_j 可作为这个输入模式的最优参考向量。

2 自增长型多级自组织映射网络

自组织映射网广泛地应用于将高维数据映射到二维空间来进行分析,并且这种映射尽可能保持了输入数据之间的关联。所以,利用输出空间中模型向量的距离能够表示输入数据间的相似性,通过分析映射我们就能够很好地理解输入数据的结构。

然而,传统的 SOM 仍然存在一些缺点和不足。首先, SOM 网络中神经元的个数和排列方式,在进行训练之前必须是预先定义好的静态结构。显而易见,在含有大量未知输入数据特征的情况下,必须进行大量的尝试来取得合适的 SOM 网络结构,以便得到满意的分析结果。因此,如何在无监督训练过程中找到一个合适的神经网络模型,如何确定神经元的个数和排列方式都是至关重要的。

其次,输入数据之间的分级关系不能直观地在分析结果中体现出来,并且在同一空间中表示的这种关系也很难被确认出来。而在实际中,我们知道分级关系的存在是非常普遍

的。因此,正确地对他们进行辨识也是数据挖掘和模式识别的一项重要任务,而传统的 SOM 无法便利地解决这个问题。

为了以统一的形式解决 SOM 中面临的上述两种局限性,在传统 SOM 的基础上提出一种新颖的人工神经网络体系结构,称为自增长型多级自组织映射网。它是一个多级的 SOM,且网络的规模可以根据实际输入数据的结构自适应地进行扩展,直至整个无监督学习过程结束。

2.1 训练算法

算法思想:实现自增长型多级自组织映射网的思想同自增长型网络,即网络的自适应扩展十分相似。首先我们来看看网络数据是如何进行自适应细分的。在网格数据的分析和计算中,人们往往只对其中的某些特征区域感兴趣,显然如果将所有的网格单元加以细分,不仅增加计算量和计算成本,而且也是对投入资源的一种浪费。自适应的细分方法如下:首先在划分较粗的网格中进行分析,然后将需要重视的特征区域进行进一步的细分,其它区域不变。以此类推迭代下去,直到所有区域达到需要的精度为止。这样既达到了分析的要求效果,又大大节省了计算资源。

自增长型多级自组织映射网如图3,首先将输入数据映射到只含有二个神经元的平面上,进行较粗的聚类,然后根据分析结果对需要细分的类进行进一步划分,即在下一层中相应位置添加二个神经元,扩大网络规模以细化某些分类,继续增加新的更大的层直到分类满意为止。

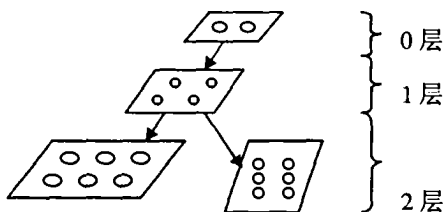


图3 自增长型多级 SOM 结构

2.2 算法初始设置和全局控制

自增长型多级 SOM 的原则是自适应的训练数据。自适应优劣程度的量度,由神经元模型向量与此神经元表示的输入向量间的偏离程度确定。一般来讲,控制整个增长过程有两种不同的策略,一是使用平均量化误差,二是使用绝对值,如量化误差。

神经元 i 的平均量化误差 mqe_i ,是根据它的模型向量 m_i 与 m_i 表示的所有输入向量间的平均欧几里得距离来计算的,如下式,假定 $x=x(t) \in R^n$ 为统计采样的观测向量序列, t 为时间坐标, C_i 为神经元 i 对应输入向量的集合, $\{m_i(t); m_i \in$

$R^n, i=1,2,\dots,k\}$ 为模型向量集合。

$$mqe_i = \frac{1}{n_{C_i}} \sum_{x_j \in C_i} \|m_i - x_j\|, n_{C_i} = |C_i|, C_i \neq \emptyset \quad (8)$$

其中, $\|m_i - x_j\|$ 是欧几里德距离,由下面式子求出:

$$\|m_i - x_j\| = \sqrt{\sum_{l=1}^n (m_{il} - x_{jl})^2} \quad (9)$$

训练的最开始按下式计算第0层的 mqe_0 ,其中 m_0 为所有输入数据的均值, n_I 为输入向量集合 I 中所有输入向量的个数。

$$mqe_0 = \frac{1}{n_I} \cdot \sum_{x_j \in I} \|m_0 - x_j\|, n_I = |I| \quad (10)$$

对 mqe_0 给定一个参数 τ_2 ,如果所有神经元的 mqe_i 都满足 $mqe_i \leq \tau_2 \cdot mqe_0$

则可以结束自扩展,否则对于不满足的神经元需要进行进一步的细分扩展。

作为另一种选择,可以使用神经元 i 的量化误差 qe_i 代替 mqe_i 来进行全局增长控制

$$qe_i = \sum_{x_j \in C_i} \|m_i - x_j\| \quad (12)$$

这时全局停止扩展条件就成了:

$$qe_i \leq \tau_2 \cdot qe_0 \quad (13)$$

虽然,使用数据分布的平均量化误差控制全局自增长比较直观,但是用量化误差却更加接近 SOM 的主要特征,能够为输入空间中向量分布密集的区域提供更多的映射空间,使密集区域进一步放大。因此,我们宁愿使用量化误差进行控制,这样就能对输入空间密集数据簇中的数据进行更好的分析。

2.3 算法训练过程和自增长过程

首先,根据标准 SOM 训练过程新建一个含有较少神经元的 SOM 网络。经过一个固定的常数 λ 次的迭代训练,可以用(11)式求出所有神经元的 qe_i 值。如果一个神经元有较大的 qe_i 说明此神经元对应的输入数据中含有大量的相异部分,即有大量不同类的输入数据映射到同一分类中。因此,我们选择具有最大 qe_i 的神经元当作是出错的神经元,记为 e ,从 e 的邻域中找出一个与它偏离最远的神经元,记为 d ,然后在 d, e 之间插入新的一行或一列神经元,将原有的网络规模进行扩大,新加入神经元的模型向量初始化为它们对应邻域模型向量的均值。

图3展示了插入新的神经元实现 SOM 网络自增长的过程。黑色为新插入的神经元,箭头指向初始化新模型向量的相邻神经元。

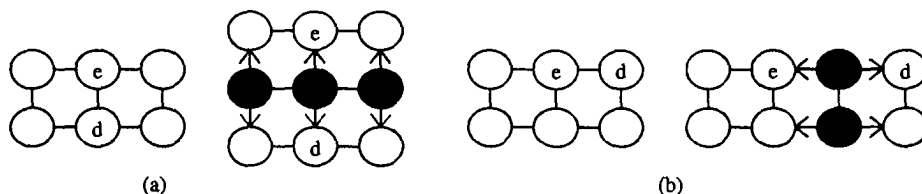


图4 插入新神经元过程

假设 C_i 为输入向量 x_j 中映射到神经元 i 上的向量子集, $C_i \subseteq I, m_i$ 为神经元 i 的模型向量,那么出错神经元为:

$$e = \arg \max_i \left(\sum_{x_j \in C_i} \|m_i - x_j\| \right), n_{C_i} = |C_i|, C_i \neq \emptyset \quad (14)$$

下面在 e 的邻域中开始寻找与 e 偏离最远的 d ,假设 e 的

邻域为 N_e ,那么 d 满足

$$d = \arg \max_i (\|m_i - m_e\|), m_i \in N_e \quad (15)$$

在 d 和 e 之间插入一行或一列神经元,并对它们的模型向量进行初始化。插入后,将学习率和邻域范围重新设置为它

们的初始值,然后进行下一个 λ 次的迭代,得到一个标准的SOM网络。

依次迭代下去,直至最后所有的神经元都满足全局控制的停止条件。

3 仿真实验及分析

表1 输入向量矩阵

属性	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A1	1	2	3	4	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
A2	0	0	0	0	0	1	2	3	4	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
A3	0	0	0	0	0	0	0	0	0	0	1	2	3	4	5	6	7	8	3	3	3	3	3	3	3	3
A4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	4	1	2	3	4

含有4个神经元的中间层和含有70个神经元的最后层分类结果如下:

ABCDE	KLMNOPQR
FGHIJ	STUVWXYZ

4个神经元

BCDE * QR * YZ
A * * * P * * X *
* F * NO * W * * *
* G * M * * * * *
HJKL * TUV * *
* I * * * S * * *

70个神经元

从实验结果可以看出,在中间层将输入分成了4大类,最终将它们按照彼此类似关系,一个个映射到不同的分类上去,越相似神经元离得越近,而且从自增长过程中我们可以看到大类和小类之间的分级结构,并产生了一个合适规模的网络结构和神经元排列来分析输入的向量集。

结束语 本文分析了自组织映射网络的基本原理,提出了不同于传统模型的自增长型自组织映射网络模型,并详细阐述了自增长的训练过程及进行全局控制的数学标准。通过

为了充分展示新模型在模式识别中的应用,我们人工定义了如表1所示的输入向量矩阵,在这些输入向量之间有明显的分级结构,并且将输入向量标记为A到Z。我们采用自增长型自组织映射网络对其进行分类,通过分级的网络能够直观地表示出这种结构。

实验仿真结果,我们看到这种动态的结构模型不仅能够更有效地对高维数据进行模式识别,大大减少因不准确的预先估计带来的不必要计算,而且能将输入数据中的分级继承关系直观地展现出来。

参考文献

- Alahakoon D, et al. Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Trans. Neural Networks*, 2000, 11(1): 601~614
- Kaski S, Kangas J, Kohonen T. Bibliography of self-organizing map (SOM) [J]. *Neural Comput. Surveys*, 1998, 1(3): 1~176
- Suganthan P N. Hierarchical overlapped SOM's for pattern classification [J]. *IEEE Transactions on Neural Networks*, 1999, 10 (1): 193~196
- Roussinov D G, Chen H. Information navigation on the web by clustering and summarizing query results [J]. *Inform. Processing Manage.* 2001, 37(1): 789~816
- Jain A K, et al. Data clustering: A review [J]. *ACM Comput. Surveys*, 1999, 31(3): 264~323
- Floyd S, Jacobson V. Random early detection gateways for congestion avoidance. *IEEE/ACM Transaction on Networking*, 1993, 1(4): 397~413
- Christiansen M, et al. Tuning RED for Web Traffic. In: *Proc. of ACM SIGCOMM Conf.* 2000
- Iannaccone G, May M, Diot C. Aggregate Traffic Performance with active queue management and drop tail. 2001
- Floyd S, Fall K. Promotion the use of end-to-end congestion control in the internet [J]. *IEEE/ACM Transaction on Networking*. Aug. 1999
- Hartling M, Claypool M, Kinichi R. Active Queue Management for Web Traffic: [Technical Report]. May 2002
- Mahajan M, Floyd S, et al. Controlling High-Bandwidth Flows at the Congested Router. In *ICNP*, Nov. 2001

(上接第43页)

结束语 主动队列管理是网络研究中的热点之一,拥塞程度的衡量和丢包策略的选择是实现有效的主动队列管理的关键技术。本文中提出的选择性丢包策略,通过记录连接的一些状态信息,将连接的传输阶段划分为启动阶段和稳定传输阶段,利用假轮转检测发送速率过大的连接,挑选出丢包的候选连接。当产生拥塞时,丢弃候选连接在缓冲区头部的数据包。实验结果表明,采用选择性丢包的 RED 算法,提高了队列的稳定性和网关的吞吐率,显著地降低了丢包率,缩短了 http 的响应时间。

参考文献

- Braden B, et al. Recommendations on queue management and congestion avoidance in the internet. *RFC2309*, 1998