

【Abstract】 The self-organizing maps (SOM) is an unsupervised learning algorithm, which is capable of self organization and visualization and has been used in many areas. In this paper, the background of self-organizing maps algorithm and basic algorithm are introduced, the parameter setting-up and disadvantage of SOM are analyzed. The improved algorithms have been reviewed emphatically, and research hotspot and applying fields are also discussed briefly. Finally, some remarks on future research are given.

自组织映射(Self-Organizing Maps, SOM)^[1]算法作为一种聚类和高维可视化的无监督学习算法,是通过模拟人脑对信号处理的特点而发展起来的一种人工神经网络。该模型由芬兰赫尔辛基大学教授Teuvo Kohonen于1981年提出后,现在已成为应用最广泛的自组织神经网络方法,其中的WTA(Winner Takes All)竞争机制反映了自组织学习最根本的特征。

1 SOM 基本原理

SOM^[1] 网络结构如图1所示,它由输入层和竞争层(输出层)组成。输入层神经元数为 n ,竞争层由 m 个神经元组成的一维或者二维平面阵列,网络是全连接的,即每个输入结点都同所有的输出结点相连接。

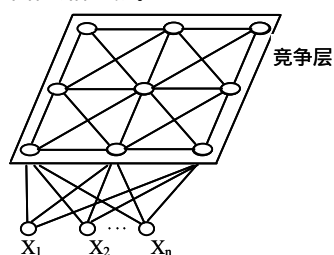


图1 SOM 网络结构

SOM 网络能将任意维输入模式在输出层映射成一维或二维图形,并保持其拓扑结构不变;网络通过对输入模式的反复学习可以使权重向量空间与输入模式的概率分布趋于一致,即概率保持性。网络的竞争层各神经元竞争对输入模式的响应机会,获胜神经元有关的各权重朝着更有利于它竞争的方向调整“即以获胜神经元为圆心,对近邻的神经元表现出兴奋性侧反馈,而对远邻的神经元表现出抑制性侧反馈,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 $50\mu\text{m}\sim 500\mu\text{m}$ 左右的神

经元;远邻是指半径为 $200\mu\text{m}\sim 2\text{mm}$ 左右的神元。比远邻更远的神元则表现弱激励作用,如图2所示。由于这种交互作用的曲线类似于墨西哥人带的帽子,因此也称这种交互方式为“墨西哥帽”。

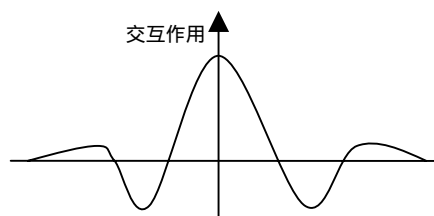


图2 神经元交互模式

SOM 算法的具体过程如下:

- (1)将权值 W_{ij} 赋予小的随机初始值;设置一个较大的初始邻域 N_c ,并设置网络的循环次数 T ;
- (2)给出一个新的输入模式 $X_k: X_k=\{X_{1k}, X_{2k}, \dots, X_{nk}\}$,输入到网络上;
- (3)计算模式 X_k 和所有的输出神经元的距离 d_{jk} ,并选择和 X_k 距离最小的神经元 c ,即

$$\|X_k - W_c\| = \min_j \{d_{jk}\} \quad (1)$$

则 c 即为获胜神经元;

- (4)更新结点 c 及其领域结点的连接权值

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i - w_{ij}(t)) \quad (2)$$

其中 $0 < \eta(t) < 1$ 为增益函数,随着时间逐渐减小;

基金项目:四川省重大基础研究基金资助项目(04JY029-001-4);西南交通大学科技发展基金资助项目(A2004015)

作者简介:杨占华(1981—),男,硕士生,主研方向:数据挖掘,计算智能,信息检索;杨燕,副教授

(5)选取另一个学习模式提供给网络的输入层,返回步骤(3),直到输入模式全部提供给网络;

(6)令 $t=t+1$, 返回步骤(2), 直至 $t=T$ 为止。

在自组织映射模型的学习中,通常取 $500 \leq T \leq 10000$ 。 N_c 随着学习次数的增加逐渐减小。增益函数 $\eta(t)$ 也即是学习率。由于学习率 $\eta(t)$ 随时间的增加而渐渐趋向零,因此,保证了学习过程必然是收敛的。一般要求:

$$\sum_{k=0}^{\infty} \eta(t+k) = \infty, \quad \sum_{k=0}^{\infty} \eta^2(t+k) < \infty$$

其中: $0 < \eta(t+k) < 1$, $k=1, 2, \dots, \infty$ 。

在实际的权系数自组织过程中,一般,对于连续系统取:

$$\eta(t) = \frac{1}{t}$$

对于离散系统,则取:

$$\eta(t+k) = \frac{1}{t+k}$$

无导师学习现在发展的还不成熟,SOM 算法还存在一些局限性,比如:

- (1)网络结构是固定的,不能动态改变;
- (2)网络训练时,有些神经元始终不能获胜,成为“死神经元”;
- (3)SOM 网络在没有经过完整的重新学习之前,不能加入新的类别;
- (4)当输入数据较少时,训练的结果通常依赖于样本的输入顺序;
- (5)网络连接权的初始状态、算法中的参数选择对网络的收敛性能有较大影响。

为此,一些学者提出了不同的改进算法,从不同方面不同程度地克服了这些缺点。

2 SOM 的改进算法

2.1 基于动态确定神经元数目的改进

传统SOM模型存在着许多的不足,特别是需要预先给定网络单元数目及其结构形状的限制。为此,人们提出了多种在训练过程中动态确定网络形状和单元数目的解决方案,比较有代表性的有:Alahakoon提出的GSOM^[2],GSOM在初始时,竞争层由4个神经元构成正方形结构,在训练过程中,对于每一个输入样本 x ,计算其获胜结点 c 的累计误差TE,若TE大于预先指定的生长阈值GT,则在 c 的邻域内找一空闲位置生成一个新结点,若 c 的邻域内无空闲位置,则将 c 的累计误差TE分配给其邻域内的结点。其不足是不能按需要方便地在合适的位置生成新结点;王莉等提出的树型动态增长模型TGSOM^[3],它与GSOM的不同在于它可以按需要方便地在任意合适位置生成新结点,克服了GSOM的缺点;Frltze提出了增长细胞结构(Growing Cell Structure,GCS)^[4]算法,GCS算法从一个由3个神经元构成的三角形结构开始,记录下每个神经元获胜的次数,在下一周期开始前,选出获胜次数最多的神经元,在其最大的一边上增加一个含初始权值的新结点,并重新计算新结点及各邻接结点的获胜次数,同时,可根据结点的获胜次数进行结点的删除操作。Choi等提出了自组织、自创造的神经网络模型(Self-creating and Organizing Neural Networks,SCONN)^[5],SCONN在初始时存在一个激活水平足够高的根结点,找出输入向量 x 的最佳匹配单元 c ,然后比较 $|x - w_c|$ 与 c 的激活水平。若前者大于后者,生成一个 c 的子结点以匹配 x ;否则,修正 c 及邻域结点的权值。

2.2 基于匹配神经元策略的改进

Kohonen竞争学习机制经常会使得竞争层中有些结点始终不能获胜,尽管SOM采用拓扑结构来克服此缺点,但并不是非常有效,为此提出了很多克服此缺点的算法,比较典型

的有:SOFM-CV,SOFM-C,ESOM(Expanding Self-organizing Map),TASOM(Time Adaptive SOM),DSOM^[6-8]。SOFM-CV的思想是:把SOM网络的权值都初始化为 $1/\sqrt{n}$ (n 为输入向量的维数),每个输入向量 x 要经过如下修正后: $\alpha x + (1-\alpha)/\sqrt{n}$ (α 随时间从0逐渐增大),再输入网络。SOFM-C即带“良心”的竞争学习SOFM,它的基本思想是给每个竞争层结点设置一个阈值,每次使竞争获胜的神经元的阈值增加,使经常获胜的神经元获胜的机会减小。ESOM的思想是把更新获胜结点 c 及其领域结点的权值公式(2)修改为下式:

$$w_{ij}(t+1) = c_j(t) \left(w_{ij}(t) + \eta(t)(x_i - w_{ij}(t)) \right) \quad (3)$$

其中 $c_j(t) \geq 1$,由 $w_{ij}(t), x_i, \eta(t)$ 确定。在TASOM中,每个神经元都有自己的学习率和邻域函数,并且能根据学习时间自动地调整学习率和邻域的大小。DSOM的思想是把内源性一氧化氮(NO)的四维动态扩散特性和其在长时间学习过程中的增强作用应用到SOM中,输入向量 x 输入网络后,以某种规则(评价函数)确定竞争层中一组获胜神经元,称为亚兴奋神经元簇。并把每一个亚兴奋神经元作为NO的扩散源。然后计算各亚兴奋神经元所处位置的NO浓度,则NO浓度最高的神经元为最终获胜单元。

2.3 SOM 算法和其它算法的组合

比较有代表性的组合算法有:Xiao等提出了把SOM和微粒群优化(Particle swarm optimization,PSO)算法结合用来对基因数据进行聚类^[9],先用SOM算法对基因数据进行聚类,得到一组权值,然后用此权值初始化PSO算法,用PSO算法对此聚类结果进行优化。Sankar等提出了把粗糙集和SOM结合的RSOM算法,它先用粗糙集理论中的依赖规则获得输入数据的大致聚类情况等知识,然后通过这些知识来确定SOM网络的结构,并对SOM权值进行初始化,用SOM网络对结果进行训练、优化。Hussin等提出了把SOM和自适应共振理论(Adaptive Resonance Theory,ART)模型相结合用来对文档进行聚类^[10],先用SOM算法对文档进行划分,然后用ART对所有的划分进行聚类。孙放等提出了把SOM和多层感知器(Multilayer Perceptron,MLP)结合进行语音识别^[11],首先用SOM算法进行语音特征矢量量化(VQ),用轨迹图训练MLP网络,相当于建立好了参数模板,用此参数模板就可以进行语音识别。

3 SOM 的应用

SOM 算法以其所具有的诸如拓扑结构保持、概率分布保持、无导师学习及可视化等特性吸引了广泛的注意,各种关于SOM 算法应用研究的成果不断涌现,现已被广泛应用于语音识别、图像处理、分类聚类、组合优化(如TSP问题)、数据分析和预测等众多信息处理领域。总之,SOM 算法的应用十分广泛,有着较好的发展前景,值得大家作进一步的研究。

4 总结与展望

无导师学习现在发展得还不成熟,SOM 网络还有很多问题需要进一步研究,比如:SOM 网络从高维映射到低维时会出现畸变,压缩比越大,其畸变程度越大。如何改进SOM 网络的结构,减小SOM 网络的训练时间以及如何确定SOM 算法中的参数问题,也是今后需要深入研究的重点之一。SOM 有良好的自组织、概率保持等特性,其可视化研究将成为今后研究的热点。

(下转第228页)