初始权值优化技术在 SOM 网络中的应用

彭雅琴[1]， 陈 俊[1]， 宫宁生[1,2]
(1. 南京工业大学 信息科学与工程学院，江苏 南京 210009；
2. 南京航空航天大学 信息科学与工程学院，江苏 南京 210016)

摘 要：SOM 网络是一种无导师学习方法，被广泛应用于各个领域。网络的性能受很多因素影响，如样本的选择、网络结构、初始权值的选定等。针对网络初始权值选取的不确定性问题，提出了覆盖权值初始化方法来优化 SOM 网络的初始权值：该方法从样本入手，并通过“覆盖”方法得出初始权值，仿真实验结果证明了此方法能有效的提高网络的识别率和稳定性。

关键词：权值优化；SOM 网络；样本分布；归一化方法；权值分布
中图法分类号：TP18 **文献标识码**：A **文章编号**：1000-7024 (2008) 23-6064-02

Implementation of optical weights initialization technology in SOM network

PENG Ya-qin[1], CHEN Jun[1], GONG Ning-sheng[1,2]
(1. College of Information Science and Engineering, Nanjing University of Technology, Nanjing 210009, China;
2. College of Information Science and Engineering, Nanjing University of Aeronautic and Astronautics, Nanjing 210016, China)

Abstract: SOM network is one of the unsupervised learning methods, which is widely applied in various fields. The performance of the SOM network is affected by many factors such as the sample selection, network structure, initial weight and so on. In order to solve the uncertain problem of the initial weight selection, a covering initialization theory is proposed, which is employed to optimize the initial weight of the SOM network. The method starts with training samples and acquires the initial weights by the way of overlay. Experiment results show that the overlay initialization theory can improve the discernment rate and stability of the SOM network effectively.

Key words: weight optimization; SOM network; samples distributing; normalization method; weight distributing

0 引 言

SOM 网络，也称 Kohonen 网络，由芬兰学者 Teuvo Kohonen[1] 提出，其认为：一个神经网络接受外界输入模式时，将会分为不同的对应区域，各区域对输入模式具有不同的响应特征，而且这个过程是自动完成的[2]。SOM 网络正是根据这一看法提出来的，通过不断的发展，现已被广泛应用于语音识别[3]、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。

对于 SOM 网络来说，研究的方向主要有：权值初始化问题，竞争机制、优胜邻域和学习率的设计，输出层网络结构的设计等，每个因素都影响着网络的运行结果，这里针对权值初始化问题做了研究。权值是整个 SOM 网络的记忆因子，网络的学习过程主要是寻找获胜神经元并相应的调整权值，并且初始权值的好坏直接影响网络的收敛情况，所以对初始权值加以适当的优化是非常必要的。

为此本文提出了一种依据样本的覆盖权值初始化方法。此方法在已有的初始化方法基础上，根据竞争算法原理，综合样本分布等因素做了改进，并且通过实验证明了其有效性，能够加速网络的收敛性，提高网络的识别率。

1 SOM 基本原理

1.1 网络结构

SOM 网络由两个层次组成，即输入层和输出层（竞争层）。输入层接受样本输入，节点数和样本维数一样；输出层上，神经元之间相互竞争，通过修改神经元之间的连接权重，使得若干神经元活跃，并最终得出获胜结点。神经元的排列有多种形式，有输出层按一维阵列组织的 SOM 网，其结构最为简单；有输出层按二维平面组织的 SOM 网，是最典型的组织方式，其每个神经元同它周围的其它神经元侧向连接，并标以权重，如图 1 所示。

网络的输入模式为 $x_i^p=(x_1^p,x_2^p,x_3^p,\cdots x_n^p)^T$，$p=1,2,\cdots k$，竞争层神经元向量 $A_{ij}=(A_{1j},A_{2j},A_{3j},\cdots A_{nj})$，$j=1,2,\cdots,m$，神经元 j 与输入神经元之间的连接权向量为 $\omega_{ji}=(\omega_{j1},\omega_{j2},\omega_{j3},\cdots\omega_{jn})^T$，$j=1,2,\cdots,m$。

1.2 学习步骤

(1)初始化：对网络的连接权 ω_{ji} 赋以较小的权值；确定学习

收稿日期：2007-12-05 E-mail：pyqaya@163.com
作者简介：彭雅琴 (1983－)，女，硕士研究生，研究方向为神经网络；陈俊 (1983－)，男，硕士研究生，研究方向为神经网络；宫宁生 (1958－)，男，博士研究生，副教授，研究方向为模式识别、人工智能。

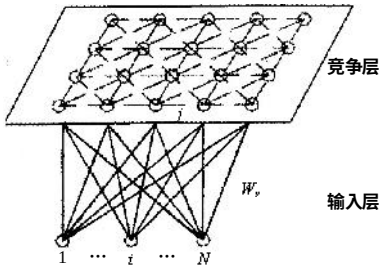

图 1　二维平面线阵

速率的初始值，$\eta(0)(0<\eta(0)<1)$；建立获胜邻域初始值$N_j(0)$；确定学习次数T。

(2)从学习样本中随机选取一个输入模式X_i^p。

(3)寻找获胜节点，计算X_i^p与ω_{ji}之间的欧氏距离，从中找出距离最小的获胜节点。

(4)定义优胜邻域$N_j(t)$，一般初始邻域$N_j(0)$较大，在训练过程中$N_j(t)$随训练时间逐渐收缩；定义学习率$\eta(t)$，一般$\eta(t)$随着时间增大而减小。

(5)调整权值，对优胜邻域$N_j(t)$内的所有节点进行权值修正：$\omega_{ji}(t+1)=\omega_{ji}(t)+\eta(t)(X_i^p-\omega_{ji}(t))i=1,2,\cdots n,j\in N_j(t)$。

(6)选取另一个学习模式提供给网络的输入层，返回步骤(3)，直至k个学习模式全部提供给网络。

(7)$t=t+1$，返回步骤(2)，直至$t=T$为止。

由上述我们可知，SOM 网络的本质就是一个映射，通过竞争算法来完成，其结果就是使一些无规则的输入自动排序，同时使连接权值的分布与输入样本的概率密度分布相似。也就是说 SOM 网络的学习过程就是使权值ω_{ji}越来越接近输入的x_i^p[4]。这是个很重要的结论，从中可以知道权值与样本有着密切的关系，两者越接近，网络的运行效果就好。由此可得如果加强初始权值与样本之间的“密切关系”，则可以提高网络的“天分”[5]从而提高网络的运行能力。而如何把初始权值与样本之间紧密联系起来，正是本文所需解决的问题。

2　权值初始化技术

2.1　现行权值初始化方法

由于初始连接权值对网络的收敛性有重大影响[6]，权值初始化的方法也不断的提出并应用，现行的权值初始化方法大致有几种：

2.1.1　给定权值法

直接给定一组确定的权值：把 SOM 网络的权值初始化为固定的数，向量大小可以相同，或者不等；这是个特殊的权值初始化方法，SOFM-CV 网络中应用了此方法：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：$\alpha x+(1-\alpha)/\sqrt{n}$（$\alpha$随时间从 0 逐渐增大），再输入网络。如果在经典 SOM 网络中应用此方法，很有可能导致不收敛。

2.1.2　随机权值法

SOM 网络的权值一般初始化为较小的随机数，这样做的目的是使权向量充分分散在样本空间当中。但在某些应用中，样本整体上相对集中于高维空间的某个局部区域，权向量的初始位置却随机的分散于样本空间的广阔区域，训练时，距离样本最近的权向量不断被调整，而远离样本的权向量却得不到调整，这样结果可能使样本的聚类结果只为一类。

2.1.3　样本平均数权值法

此类方法为先计算出全体样本的中心向量，在该中心向量基础上迭加小随机数作为权向量初始值。这样就可以将权向量的初始位置确定在样本群中，从而加强网络的学习性能，加快网络的收敛性。此方法在随机数大小的选取上无定性，得出的初始权值很难充分的分布在样本空间中，从而限制了网络的优化性能。

综合以上的权值初始化方法，可以看出初始化权值的两个相关又对立的点：既要保证权值与输入空间的相似性，同时还得保证权值的离散性。本文的初始化方法正是以现行的权值初始化方法为基础，结合 SOM 工作原理，综合提出的。

2.2　覆盖初始化权值法

如何既保证权值与输入空间的相似性，同时又保证权值的随机性呢？为此本文提出了覆盖初始化权值方法，该方法首先通过“归一化”方法优化样本分布，提取样本分布的参数，然后应用多维正态分布覆盖整个样本，最终得到一组更有效的初始权值。

2.2.1　覆盖初始化权值法原理

从前面叙述可知，竞争算法的最后结果就是使学习后的权值ω_{ji}越来越接近输入的x_i^p，这就给了我们重要提示：权值分布是一个独立的分布，但与外界的关系不是独立的，它与样本分布是息息相关的，从而可以考虑由样本分布引导出权值分布。由于在实际应用中，样本分布点比较分散，总体分布不容易把握，由样本分布引导出的权值分布，有好坏差异之分。为此我们可以通过归一化方法来简化样本分布。所谓归一化，就是使向量变成方向不变长度为 1 的单位向量。2 维和 3 维单位向量就可以在单位圆和单位球上直观表示，从高维来看归一化后的样本都被映射到了一个定长为 1 的超球面上了。从其几何意义上我们可以明显看到归一化的优化作用：向量的值虽然改变了，但却能保持尺度的不变性；总体样本特征没有改变，但分布简化了很多；这就意味着原样本分布和归一化的样本分布两者是同性的，且后者的分布更加有序，容易控制。上述处理提取出样本分布后，我们可以得到样本的中心向量、样本的大致范围等。

提取出样本的相关参数后，可以把它引用为权值的分布参数。为了保证权值的离散性，可以根据提取的参数，引用多维正态分布在样本分布范围内随机构造权值向量。产生的随机权值向量以一定概率对称的分布在中心向量的周围，使得权值充分分布在优化位区。

2.2.2　覆盖初始化权值法步骤

(1)使所有样本归一化，应用公式$\hat{X}=\frac{X}{\|X\|}=\left[\frac{x_1}{\sqrt{\sum_{j=1}^{n}x_j^2}}\cdots\frac{x_n}{\sqrt{\sum_{j=1}^{n}x_j^2}}\right]$，得到$\hat{xi}^p=(\hat{x1}^p,\hat{x2}^p,\hat{x3}^p,\cdots\hat{xn}^p)^T$；

(2)从样本分布导出权值分布，计算所有$\hat{xi}^p=(\hat{x1}^p,\hat{x2}^p,\hat{x3}^p,\cdots\hat{xn}^p)^T$的平均值，即中心向量，并依据中心向量，计算每个向量与中心向量的欧氏距离，标记出最大距离$d_{\max}$；

(下转第 6068 页)

整体性能的角度来看，新的多分类器融合系统不失为一种很好的选择。

3 结束语

本文首先介绍了模糊积分用于构造多分类器融合模型的原理，然后利用捕食与被捕食算法来寻找每个分类器的最优模糊测度。通过仿真实验证实该方法能够有效的融合两个分类器，这对多个分类器融合的理论和算法有着重要的研究意义。

参考文献:

[1] 李玉榕.基于模糊积分和遗传算法的分类器组合算法[J].计算机工程与应用,2002,40(12):119-121.

[2] 梁文.基于生态捕食模型的多目标优化问题求解算法[J].中国科学技术大学学报,2005,35(3):360-365.

[3] 李永昆.中立型捕食者-被捕食者系统的周期正解[J].应用数学和力学,1999,20(5):545-550.

[4] 张广全.模糊测度论[M].贵阳:贵州科技出版社,1994:78-85.

[5] Hossein Tahani,James M Keller. Information fusion in computer vision using the fuzzy integral[J]. IEEE Trans on SMC,1990,20(3):733-741.

[6] 哈明虎.模糊测度与模糊积分理论[M].北京:科学出版社,1998:31-45.

[7] 姚明海,李澎林.基于遗传算法和模糊积分的多分类器集成[J].计算机应用与软件,2003,20(8):66-68.

[8] 陈凤德,陈晓星.一类具有功能性反应的中立型捕食者－食饵系统全局正周期解的存在性 [J]. 数学物理学报, 2005,25 (7):981-989.

(上接第 6065 页)

(3) 假设样本各属性之间独立，概率密度为$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-u)^2}{2\sigma^2}}$，以中心向量为中心点构建多维正态分布，并选取$d_{max}$范围内的随机数作为初始权值；

(4)对每个神经单元赋予初值。

覆盖方法通过对现有初始权值技术的改造，既保证了初始权值能够分布在样本的中心位置，又提高了权值的离散性，使得权值能够充分分布在样本中，克服了现有初始权值技术的一些缺点，使得网络加速收敛，同时也有效的提高了网络的识别率。

3 仿真实验结果

Iris(Iris是一种鸢尾属植物)数据集[7]，是一种标准的测试集[8]。数据共有 3 个种类 setosa, versicolor,virginica，每一个种类有 50 个样本，共 150 个样本。每个样本有 4 个属性为：萼片长度，萼片宽度，花瓣长度，和花瓣宽度。对于同一个 SOM 网络，采用相同的实验数据：随机选择 75 个样本作为学习样本，其余 75 个样本作为测试样本。针对不同的权值初始化方法，在样本和其它网络参数不变的情况下，分别进行了 8 次独立的无监督学习的实验，其实验结果如表 1 所示。

从表 1 中的实验数据可以看出，本文提出的权值初始化方法，其运算结果有了明显的提高，网络的平均错分率从 13.7%下降到了 5%，并且网络的稳定性有所提高。

4 结束语

利用 SOM 神经网络进行运算，其性能与很多因素有关：学习算法、网络的结构、学习样本的选择，网络的初始状态，每一个细节都会影响到网络的性能。为此很多学者从各个角度提出了改进方法。在不断的实践过程中，本文发现了初始化权值的重要性，并做了一定的研究。在算法和网络结构等其它因素固定且相同的条件下，分别对二个初始化权值方法做了实验和总结，并证明了本文提出的覆盖初始化方法的有效性。但是权值问题是一个广泛问题，需要根据不同的实际情况做出不同的选择，如何更进一步的加强其在实际应用中的作用还需要以后更深入更长期的研究。

参考文献:

[1] Kohonen T. The self-organizing maps [J]. Proceedings of the IEEE,1990,78(9):1464-1480.

[2] 韩力群.人工神经网络理论、设计及应用[M].北京:化学工业出版社,2004.

[3] 涂晓芝,颜学峰,钱锋.PSO-SOM 分类判别研究及其应用[J].高技术通讯,2006,16(10):1014-1018.

[4] 张立明.人工神经网络的模型及其应用[M].上海:复旦大学出版社,1992.

[5] 肖伟.初始化权值优化技术在机器人学习中的应用[J].电子学报,2005,33(9):1720-1722.

[6] 杨占华,杨燕. SOM 神经网络算法的研究与进展[J].计算机工程,2006,32(16):201-202.

[7] UCI repository of machine learning databases[DB/OL]. http://www.ics.uci.edu/~mlearn/MLRository.html.

[8] Fisher R A. The use of multiple measurements in taxonomic problems[J]. Annual Eugenics ,1936,7,PartII:179-188.

表 1 样本平均数权值法和覆盖权值法的 Iris 实验结果

模型	次数										
	1	2	3	4	5	6	7	8	平均错分数	平均错分率	收敛次数
样本平均数权值法	10	11	12	10	11	9	11	8	10.25	13.7%	1000
覆盖权值法	4	4	3	4	4	3	4	4	3.75	5%	750