一种新的基于神经网络覆盖分类算法

黄国宏 邵惠鹤
(上海交通大学自动化研究所,上海 200030)

摘 要 为了克服传统神经网络算法在处理分类问题时训练时间长、泛化能力弱的不足,提出了一种新的基于构造型神经网络覆盖分类算法,该算法通过在超球面上对样本数据进行聚类分析,找出同类样本中未被覆盖样本的最大密度点,然后在特征空间里做超平面与球面相交,得到球面领域覆盖,从而将神经网络训练问题转化为点集覆盖问题,同时也考虑了神经网络规模的优化问题。实验结果证明了该算法的有效性。

关键词 模式识别 神经网络 最大密度覆盖 MP 神经元 构造型神经网络
中图法分类号:TP18 **文献标识码**:A **文章编号**:1006-8961(2004)10-1165-04

A New Classification Method Based on Neural Network Covering Algorithm

HUANG Guo-hong, SHAO Hui-he
(*The Institute of Automation in Shanghai JiaoTong University, Shanghai* 200030)

Abstract In order to overcome the shortcoming of the longtime training and the frail generalization power of classical neural networks, this paper proposes a new covering classification algorithm based on constructive neural networks. The algorithm starts with the sample data directly and clustering analysis is executed on a hypersphere to find a sample with the max density, and then the intersection between the positive half-space of the hyperplane and sphere, called "sphere neighborhood", is obtained, by which the training problem of neural networks may be transformed into the covering problem of point sets. Thus the new algorithm can reduce the traditional learning complexity. At the same time, the optimization of the neural network is also considered and computer simulation results show that the proposed neural network is quite efficient.

Keywords pattern recognition, neural networks, max density covering, MP neuron, constructive neural networks

1 引 言

近年来,随着计算机的发展,人工智能的兴起,模式识别已成为一个越来越热的研究方向。它所研究的理论和方法在很多科学和技术领域中得到了广泛的重视和应用,推动了人工智能系统的发展。几十年来,模式识别取得了大量的成果,在很多地方得到了成功的应用。但是,由于模式识别涉及到很多复杂问题,现有的理论和方法对于解决这些问题还有很多不足之处。

基金项目:国家"十五"863 项目(2001AA413130)
收稿日期:2004-02-16;**改回日期**:2004-06-15

贝叶斯决策作为一种传统的模式识别方法,它要求知道样本的概率分布,这在实际问题中是很难做到的,而且在模式识别的统计方法中有许多理论(如似然法)都是有关样本概率分布的估计,其中,许多估计方法都是用形式和数目预先确定的分布函数的线性组合(如高斯函数)去逼近样本的分布。现有的前馈神经网络在求解模式识别问题时,其中心思想是建立样本和其类别的映射,然而其训练方法是基于预先给定的评价函数的极小化,其本质也是用形式和数目预先确定的多个函数(即隐层单元的输出函数)的组合去逼近这一映射。可以看出,这两种方法都是从已知形

式和数目的函数的组合入手来分析，而没有直接从样本数据本身入手来推测有关性质。

以球领域模型[1,2]为基础，提出一种新的基于最大密度球领域覆盖的神经网络训练算法，该算法直接从样本数据本身来逼近它在空间中分布的状况，并以此为依据构造神经网络，利用多个 MP 神经元所对应的多个“球领域”去覆盖所有训练样本。由于神经元的有关参数和个数都是直接由训练样本通过样本空间的分布直接决定的，所以该算法的出发点不是要得到样本分布的解析表达式，而是用这些神经元的覆盖区域的组合近似“勾勒”出各类样本分布的几何区域。而当判断一个新的样本应该属于哪一类时，只需判断它被哪个几何区域所覆盖，该区域所对应的类别就是答案。因此，该算法不要求预先固定隐层单元的个数，这就为神经网络的构造提供了很大的灵活性。

2 基于覆盖思想的构造型神经网络

对于一个给定的训练样本集 $\boldsymbol{A}=[\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N]$，首先通过变换[3~5]$T$ 把所有的样本点向上投影到一个 n 维空间的球面 S^n 上。

$T:\boldsymbol{X}\in\boldsymbol{A},T(\boldsymbol{X})=(\boldsymbol{X},\sqrt{f^2-\|\boldsymbol{X}\|^2})$，其中 f 为一实数，且满足大于等于样本空间中所有矢量的模。

一个 MP 神经元是一个 n 输入、单输出的处理单元，即

$$y=\mathrm{sgn}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}-\theta) \tag{1}$$

其中，$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 表示输入向量；$\boldsymbol{W}=(w_1,w_2,\cdots,w_n)^{\mathrm{T}}$ 表示权向量；θ 表示阈值。

$$\mathrm{sgn}(v)=\begin{cases}1 & v\geqslant 0\\ -1 & v<0\end{cases} \tag{2}$$

当所有的点都处于同一球面上时，$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}-\theta>0$ 对应于一个超平面与球面的交集，即所谓的“球领域”，如图 1 所示。$\boldsymbol{W}$ 的方向即超平面的法方向，当它为单位向量时，θ 为超平面到球心的距离。因此对球面上的点分类，只要构造若干组球领域，将所有样本都覆盖住，然后在判断某一样本的类别时，只要看它被哪组球领域覆盖即可。在构造神经网络时，一个球领域对应一个神经元，引入一层神经元对输入点“覆盖”，再用另外一层神经元构造逻辑“AND”和“OR”结构来分析类别，这样就可以构造出满足识别要求的神经网络模型。

球领域覆盖方法的优点是它的直观性，把神经网络训练问题转换为点集覆盖问题，使得人们可以迅速地，构造型地得到对于训练数据百分之百正确分类的人工神经网络，从而克服了传统神经网络反复循环迭代、“过学习”等缺点；另外由于输入向量的类别完全取决于它被哪些球领域所覆盖，因此这种方法的另一优点是构造出的网络的行为便于分析。此外该网络还具有一定的泛化能力，当输入向量有轻微的畸变或噪声干扰时，只要输入向量的覆盖特性不变，没有移动到别的类别对应的球领域中，分类结果也不会变。

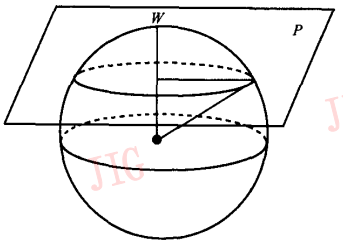

图 1 球领域

3 神经网络最大密度覆盖学习算法

3.1 算法思想

如前面所述，一个 MP 神经元对应于一个球面领域，因此前馈神经网络的训练问题可以转化为点集覆盖问题，即训练的过程实质上是构造许多组球面领域。因此，提出一种基于空间最大密度覆盖的多层前向神经网络算法，该算法就是使同类训练样本被同一组球面领域覆盖，而不同类的训练样本被不同组的球面领域覆盖，并考虑用尽量少的球面领域完成对样本数据的覆盖任务。

首先在特征空间中找到同类样本密度最大的样本点，然后以该样本点为法向量做一超平面

$$P:\ \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}-\theta=0$$

P 与球面相交，那么这时，$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}-\theta>0$ 就表示落在 P 正半空间的部分，这个部分恰好是球面上的某个“球形领域”。这个“球形领域”的中心就是 $\boldsymbol{W}$，其半径为 $r(\theta)$ 是 θ 的单调下降函数。因此对球面上的点分类，只要构造若干球形领域，将它们都覆盖住，然后在判断某一样本的类别时，只要看它被哪个球领域覆盖住即可。

3.2 算法实现

由于多类识别问题总可以转化为多个两类覆盖

问题的组合，因此下面只讨论两类样本的分类问题。

考虑 n 维样本空间中有两类样本 C_1 和 C_2，样本个数分别为 n_1 和 n_2。首先把所有样本投影到一个 n 维的球面上，记 C_i 类样本的第 m 个样本为 $\boldsymbol{X}_{C_i,m}$，其覆盖标记为 $Cover(C_i,m)$，$1\leqslant m\leqslant n_i$，$i=1,2$，当该标记为 1 时，表示样本 $\boldsymbol{X}_{C_i,m}$ 已被某个球形领域覆盖；当标记为 0 时，表示该样本尚未被任何球形领域所覆盖，且初始值全部设置为 0。在训练过程中，每个 MP 神经元都有一个类别标号，取值为整数 1 或 2，表示被它覆盖的样本所属的类别，并用 $\langle \boldsymbol{a},\boldsymbol{b}\rangle$ 表示 $\boldsymbol{a},\boldsymbol{b}$ 向量的内积。

下面给出样本数据点 $\boldsymbol{X}_{C_i,m}$ 处的密度指标定义

$$D_{C_i,m}=\frac{1}{2\pi n_i r_{C_i,m}^2}\times \sum_{n=1}^{n_i}\exp\left[-\frac{(\boldsymbol{X}_{C_i,m}-\boldsymbol{X}_{C_i,n})^{\mathrm{T}}(\boldsymbol{X}_{C_i,m}-\boldsymbol{X}_{C_i,n})}{(r_{C_i,m}/2)^2}\right] \tag{3}$$

$$r_{C_i,m}=\min(\|\boldsymbol{X}_{C_i,m}-\boldsymbol{X}_{C_j,n}\|) \tag{4}$$

其中，正数 $r_{C_i,m}$ 为样本点 $X_{C_i,m}$ 距离异类样本的最短距离，定义了该点的一个邻域，半径以外的数据点对该点的密度指标贡献甚微。显然，如果一个样本数据点具有高密度值，则该数据点一定有多个邻近的样本数据点。

定义

$$d_{C_i,m}^{(1)}=\max_{i\neq j}\{\langle \boldsymbol{X}_{C_i,m},\boldsymbol{X}_{C_j,n}\rangle\} \tag{5}$$

$$d_{C_i,m}^{(2)}=\min\{\langle \boldsymbol{X}_{C_i,m},\boldsymbol{X}_{C_i,n}\rangle > d_{C_i,m}^{(1)}\} \tag{6}$$

$$1\leqslant m\leqslant n_i,\ 1\leqslant n\leqslant n_j,\ i,j=1,2$$

其中，$d_{C_i,m}^{(1)}$ 表示样本 $\boldsymbol{X}_{C_i,m}$ 与异类样本的最小距离；$d_{C_i,m}^{(2)}$ 表示样本 $\boldsymbol{X}_{C_i,m}$ 与同类样本的最大距离。

训练算法如下：

(1) 求训练样本集 $\boldsymbol{A}$ 中样本的最大模 f，将 $\boldsymbol{A}$ 的点向上投影到中心为原点，半径为 f 的球面上；

(2) 计算所有未被覆盖的样本数据点的密度 $D_{C_i,m}$，求其最大值，得到具有最大密度的样本数据点 $\boldsymbol{X}'$；

(3) 计算该样本数据点 $\boldsymbol{X}'$ 处的 $d^{(1)}$ 和 $d^{(2)}$，求得 $d=ad^{(1)}+bd^{(2)}$，a 和 b 为参数且满足 $a+b=1$；

(4) 以样本点 $\boldsymbol{X}'$ 作为法向量 $\boldsymbol{W}$，d 作为阈值 θ 作超平面：$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}-\theta=0$ 与球面相交；

(5) 求出满足 $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}-\theta>0$ 的样本点，并且把相应的覆盖标记（$Cover(C_i,m)$）设置为 1；返回步骤 2 处继续重复执行，直到所有的覆盖标记都被置为 1。

4　实验及结果

考虑平面上双螺旋线的识别问题，如图 2 所示。双螺旋线数据是由极坐标系统下方程 $r=\theta$ 和 $r=-\theta$，$\pi/2\leqslant\theta\leqslant 6\pi$，两条曲线互相缠绕而成，每条曲线上任意取若干样本点作为训练样本和检测样本。

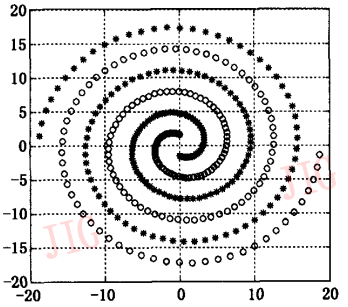

图 2　曲线为双螺旋线
（。和 * 表示两类训练样本点）

在文献[6]中，作者试图利用 BP 算法求解双螺旋问题，但是没有成功；文献[7]提出“生成-收缩”法，来求解双螺旋问题，但正确率只有 89.6%。由此可知此问题的难度。表 1 给出了基于最大密度覆盖法所得到的 matlab 仿真实验结果。

表 1　采用最大密度覆盖算法的实验结果

训练样本数	覆盖数	测试样本数	训练时间(s)	识别率(%)
100	27	20 000	0.938	93.384 0
200	31	20 000	1.122	99.593 4
400	32	20 000	1.722	99.707 5
864	33	20 000	1.879	100
1 700	33	20 000	1.981	100
2 160	33	20 000	2.149	100
4 300	33	20 000	3.250	100

由表 1 可以充分说明本文算法的有效性，它最大的优点就是训练速度快，因为算法本质上是非迭代的，因此，可以在极短的时间内给出识别结果，这比传统的其他的神经网络算法快很多，具有一定的实际应用潜力。当训练样本点增加到一定数量时，识别率可以达到 100%，因此可以通过增加训练样本数来提高样本识别的正确率。

5　结　论

本文在 MP 神经元模型的基础上，提出了最大密度覆盖学习算法，虽然该算法实现的神经网络规

模比文献[2]的规模略大，但是本算法几何意义更加明确，实现起来更加容易，不但对于训练样本的识别率为 100%，而且具有良好的泛化能力，这与传统的 BP 算法相比，具有明显的优势，克服了 BP 算法训练时间长，和“Over-Fitting”等问题。而本文提出的算法之所以表现出良好的泛化能力，关键在于它是从数据本身出发去逼近其分布的几何轮廓。同时，本文算法很好地解决了双螺旋线的分类问题，通过实验可知该算法的有效性。

参 考 文 献

1 张铃，张钹. M-P 神经元模型的几何意义及其应用[J]. 软件学报，1998，**9**(5)：334～338.

2 张铃，张钹，殷海风. 多层前向网络的交叉覆盖设计算法[J]. 软件学报，1999，**10**(7)：737～742.

3 吴鸣锐，张钹. 一种用于大规模模式识别问题的神经网络算法[J]. 软件学报，2001，**12**(6)：851～855.

4 陶品，张钹，叶榛. 构造型神经网络双交叉覆盖增量学习算法[J]. 软件学报，2003，**14**(2)：194～201.

5 Zhang L，Zhang B. A geometrical representation of McCulloch-Pitts neural model and its application[J]. IEEE Transactions on Neural Networks，1999，**10**：925～929.

6 Baum E B，Lang K J. Constructing hidden units using examples and queries[A]. In：Lippman R P *et al* eds. Neural Information Processing [M]，San Mateo，CA：Morgan Kaufmann Publishers，Inc，1991：904～910.

7 Chen Q C. Generating-shrinking algorithm for learning arbitrary classification[J]. Neural Networks，1994，**5**(7)：1477～1489.

黄国宏 1975 年生。2001 年获燕山大学控制理论与控制工程专业硕士学位，现为上海交通大学控制理论与控制工程专业博士研究生。主要研究方向为机器学习、统计学习理论、图像处理、模式识别等。
E-mail：h_guohong@163.com

邵惠鹤 1936 年生。博士生导师。华东理工大学毕业。主要研究方向为工业过程控制、多变量约束控制、最优控制以及现场总线和工业以太网。

更 正

本刊 2004 年第 8 期第 959 页图 1 由于印刷问题，图不够清晰，现重新刊登如下：

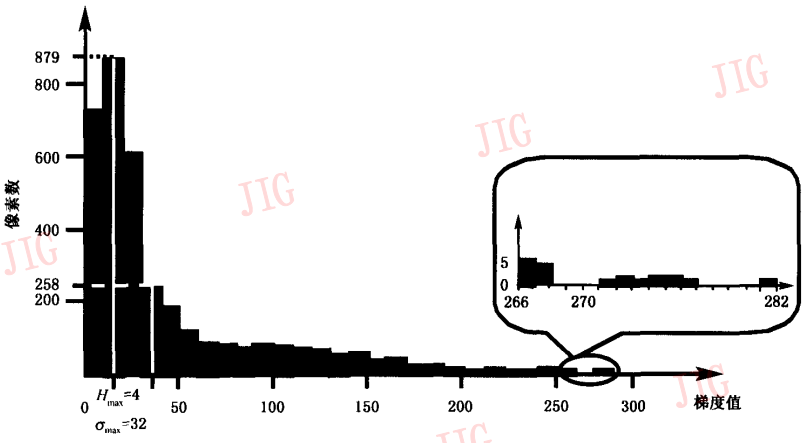

图 1 Lena 图像梯度直方图