文章编号:1001 - 9081(2007)02 - 0330 - 03

优先排序神经网络 K 网覆盖分类研究

朱世交[1,2],王 真[1],廖明军[3]
(1. 同济大学 计算机科学与工程系,上海 20092; 2. 同济大学 半导体与信息技术研究所,上海 20092;
3. 同济大学 交通运输工程学院,上海 20092)
(mediate@163.com)

摘 要:从高维空间特征点覆盖的角度,讨论了优先度排序神经网络(PONN)算法,提出了非各向同性的 K 网覆盖算法(KPA)算法,最后给出标准测试集和应用测试集的比较结果,并对其与各向同性覆盖中心适配选择算法(CASA)进行了分析与比较,实验结果表明 KPA 算法在样本连续性构造方面优于 CASA 算法。

关键词:优先排序神经网络; 模式识别; 拓扑空间
中图分类号: TP181 **文献标识码**:A

Research on K-classification covering for PONN

ZHU Shi-jiao[1,2], WANG Zheng[1], LIAO Ming-jun[3]
(1. *Department of Computer Science and Engineering, Tongji University, Shanghai* 20092, *China*;
2. *Institute of Semiconductors and Information Technology, Tongji University, Shanghai* 20092, *China*;
3. *College of Transportation Engineering, Tongji University, Shanghai* 20092, *China*)

Abstract: From the aspect of coverage in the High-Dimensional Space (HDS), we discussed algorithms of Priority Ordered Neural Network (PONN), and promoted K-Partitioning Algorithm (KPA) based on anisotropy in HDS. Benchmark testing and application were made on KPA as well as the comparison with Center Adaptive Selection Algorithm (CASA). The results of experiment prove that the constructive method of KPA is superior to CASA especially in continuous samples.

Key words: Priority Ordered Neural Network (PONN); pattern recognition; topology space

0 引言

随着信息时代的发展,当前的机器处理的数据特征较原来有了巨大的变化,主要归纳为:高维、高数据量、非结构化等,虽然当前人们接触到的数据量很大,但从数据中寻找到需要的知识存在困难[1]。人工神经网络(ANN)作为一种并行计算的方法在分类、聚类、以及优化计算等方面都有着广泛的应用[2]。在众多类型的神经网络中,前馈网络最受人们青睐,其训练方法大致可分为基于搜索的方法,如 BP 算法,以及使用优化步骤的方法 SVM[3],但这些方法在构造网络及新样本加入时往往需要大量训练时间,虽然 SVM 在当今得到了广泛应用,但其最优划分在大量样本情况下不能在多项式时间范围内得到,属于 NP 难问题[1];对于海量数据,理论上可行,但实践中存在困难。传统神经网络的这些缺陷限制了其应用范围。

传统模式识别方法中只注意到事物类别的划分,而没有意识到事物的联系,同类样本往往存在“同源”或“同类”,而且同类类别个体的变化是连续的,在此基础上文献[4]提出了仿生模式识别,把分类识别认为是高维空间的同类样本覆盖问题,提出了与人类认知学习概念模式类似的优先度排序神经网络(Priority Ordered Neural Network, PONN)的概念[5],这种方式更接近于人类的认识。

本文针对 PONN 对样本特征空间的覆盖算法进行研究,提出了高维空间中非各向同性的 K 网覆盖方法,该方法以数据自身在空间中的分布为基础,先规划出同层、同类超球,接下来对超球内近邻点进行领域覆盖,生成以大的超球为界限同层同领域 K 网覆盖。该方法无需预先知道隐含层神经元个数,构造灵活,而且网络构造具有层次性,符合人对事物认识的特征。

1 PONN 对知识的分类表示方法

1.1 PONN 基本模型定义

PONN 的基本思想是空间类别划分具有层次特征,类别分类由小类分层判别组成,网络结构如图 1 所示。

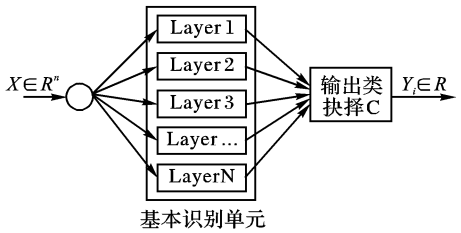

图 1 PONN 网络层次构造单元

网络基本结构单元由各个不同层次中间层神经网络单元构成,通过不同层次数据分类,形成一个塔式结构网络,数据类别划分依赖数据本身;输出类别依赖于前端不同层次神经元并行计算的激活状态。为阐述方便,基本符号定义如下:

收稿日期:2006 - 09 - 01

作者简介:朱世交(1980 -),男,安徽五河人,博士研究生,主要研究方向:优先排序神经网络的理论及其在仿生识别中的应用; 王真(1980 -),女,山东泰安人,博士研究生,主要研究方向:数据压缩、模式识别; 廖明军(1974 -),男,湖南邵东人,博士研究生,主要研究方向:交通信息工程与控制.

R^n:表示 n 维欧氏空间, 样本点 $x_i \in R^n$;

S_i:表示 n 维空间中第 i 个特征类子空间,其中 $S_i \subset R^n$, $i \in \{1,2,\cdots,m\}$;$\{\psi_i^k\}$,k 层第 i 类别的神经元集合,$k \in I, i \in \{1\cdots m\}$;单个神经元 $\psi_{i,j}^k$,表示 i 类 k 层第 j 个神经元;

L_k:表示构造得到的 k 层神经元集合,定义 $L_k := \{\{\psi_i^k\} \mid k \in I, i \in \{1\cdots m\}\}$;

O_k:神经元 k 层输出序列,定义如下:$O_k := \{0,\cdots,I_q^p,0,0 \mid q \in \{1\cdots m\}, p = 0 \text{ 或 } 1\}$

Y_i:类别 i 的判别输出。

神经网络对类别划分问题,本质上就是寻找满足条件的未知函数 $f: R^n \to R$。传统网络没有表示知识层次划分的能力,因此往往不符合人类认知能力,人类认识事物往往是根据事物流形曲线特征来完成的[6],PONN 分层覆盖体现了这种流形变化,对空间内数据样本,存在不同层次覆盖,在类别覆盖过程中表现为空间区域的外覆盖,形式化描述如下:

覆盖层次的定义,对于 m 类样本的空间区域 $\{S_i \mid x_i \in S_i, i = 1\cdots m\}$,对每个覆盖层次,样本空间区存在一定序列的吸引点分别定义为:$\{x_{i,j} \mid x_{i,j} \in S_j, i = 1\cdots p, j = 1\cdots m\}$,以及由这些吸引点构造的外覆盖区域 $\{D(x_{i,j}) \mid x_{i,j} \in S_j, i = 1\cdots p, j \in \{1\cdots m\}\}$,覆盖区域的定义:$D = \{f \mid f: R^n \to R\}$,当 D 的域值在 $[\theta_1\theta_2]$ 区间内时,表示在覆盖区域内,$\theta_1\theta_2$ 为空间区域映射的上界和下界。

1.2　网络构造基本原理

对神经元 $\psi_{i,j}^k$,规定单个神经元 $\psi_{i,j}^k$ 在高维空间中是有限空间区域外覆盖,近邻空间测度采用欧式度量空间,通过空间特征点 x_l 与空间支撑点 x_{lo} 的距离比较进行测度,公式描述如下:

$$\rho(\psi_{i,j}^k) = \sqrt{\sum_{l=1}^{n}(x_l - x_{lo})} \tag{1}$$

当 $\rho(\psi_{i,j}^k)$ 测度小于阈值 Γ_0,$\Gamma_0 \in R$(阈值由支撑点与异类样本最近距离确定),则认为在区域覆盖范围内;反之,则表示在区域之外,即不被此神经元 $\psi_{i,j}^k$ 覆盖。形式化描述如下:

$$F(\rho) = \begin{cases} 1, \rho < \Gamma_0 \\ 0, \text{其他} \end{cases} \tag{2}$$

对于构造得到的神经元 $\psi_{i,j}^k$,同层神经元空间覆盖定义为 $C(\psi_i^k)$,特别注意的是,同层覆盖情况下的不同神经元的覆盖区域不重叠,即 $C(\psi_i^{k1}) \cap C(\psi_i^{k2}) = \varnothing$,$k1 \neq k2$,由此得到的同类样本外覆盖在高维空间维系上表现为各向同性,即超球体覆盖同类子样本点;各向同性有利于高维空间覆盖体的构造,但是随着维数的增加,超球的体积也就越大;同时,考虑到同类样本变化存在同源渐变,所以可以利用连续性,在超球内使用各向异性神经元覆盖特征点。

网络训练依据训练样本本身的特征属性,起始阶段网络内部的层次为空,即 $k = 0$(不同于传统网络需预先确定神经元数目);随着训练样本进入网络,网络层数 k 逐渐增大,神经元数不断增多,构成层内神经元 $\{\psi_i^k\}$ 集合,要求同层非同类神经元在高维空间覆盖区域不可重叠。

网络识别基本原理。网络接受待识别数据输入 $x \in R^n$,各层响应输出序列 O_k,定义判别:

$Y_i := \{i \mid i = \{0 \text{ 或 } q := \{\min(k) \& \{O_k > 1\}\}\}\}$

通过 Y_i 对样本 $x \in R^n$ 进行识别:

$$Y_i = \begin{cases} \text{类别 } i, i > 0 \\ \text{拒绝}, i = 0 \end{cases}$$

网络从不同层神经元给出样本识别结果。

2　PONN 的 K 网覆盖算法

针对同类样本具有同源特点,在构造 PONN 网络时,先通过欧式度量规则,把本类与异类样本区分开;其次在同类样本中获得覆盖本类样本最多的最大超球体,最后在超球体内部进行样本点的连接,最后形成超球内点的连接网络,故称 K 网覆盖。

设当前层要处理的类为 S_i 类,样本中心向量点为 Xo_i,以 $\rho(Xo_i, S_j), j \neq i$ 为当前中心点到其他类别的空间欧式距离,取最小 ρ 作为此类在空间中超球的半径,即 $r = \min\{\rho(Xo_i, S_j), j \neq i\}$;通过超球半径 r 作为样本空间阈值,得到类 i 在高维空间中以 Xo_i 为球心,r 为半径的超球范围内 S_i 的子类 B_i,$B_i \subseteq S_i$,接下来使用以下规则在子类 B_i 中使用 Xo_i 为球心进行类别覆盖。

对 $X_i \in B_i, X_i \neq Xo_i$,用欧式测度找点 $X_j = \{X_k \mid \min(\rho(X_k, X_j)), X_k \in B_i\}, X_k \neq Xo_i$,确定中心点到当前点的最短距离 $r_x = r - \min(\rho(X_i, Xo_i), \rho(X_j, Xo_i))$,对于高维空间两样本点之间的连接,使用类似于超香肠神经元[5] 式3 对 X_i,X_j 进行覆盖。

$$d^2(x,x_i,x_j) = \begin{cases} \| x - x_i \|^2, q(x,x_i,x_j) < 0 \\ \| x - x_j \|^2, q(x,x_i,x_j) > \| x_i - x_j \| \\ \| x - x_i \|^2 - q(x,x_i,x_j)^2, \text{其他} \end{cases}$$

其中:

$$q(x,x_i,x_j) = \left\langle x - x_i, \frac{x_j - x_i}{\| x_j - x_i \|} \right\rangle \tag{3}$$

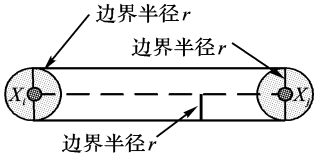

图 2　覆盖相邻点的神经元(2 维)

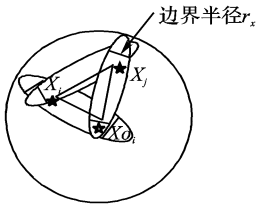

图 3　包含中心点的相邻覆盖点(2 维)

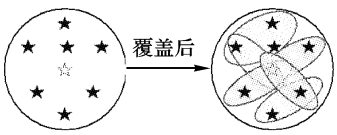

图 4　B_i 中相邻覆盖点覆盖示例(2 维)

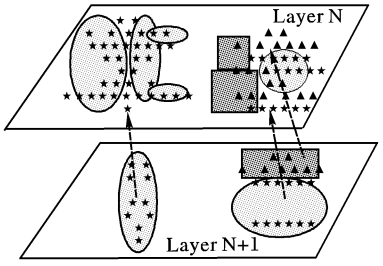

图 5　神经元覆盖的 N、$N+1$ 层情况

当前覆盖区域 S 定义为:

$S(x;x_i,x_j,r) = \{x \mid d^2(x,x_i,x_j) < r^2\}$,其覆盖区域如图 2 所示。

以上是相邻点连接,除中心点之外的其他点连接之外,还要把点与中心点连接,含邻近三个点连接形式如图 3 所示。

对整个同类样本的覆盖而言,使用邻近神经元覆盖同类样本,在构造近邻覆盖神经元之前,需把已有覆盖神经元覆盖类 B_i 中的点去除掉,整个神经元的覆盖情况如图 4 所示。

本层覆盖不到的神经元通过下一层的神经元覆盖来完成,对于所有的空间点覆盖神经元分层结构表示,如图 5 所示。

2.1　K 覆盖算法

K 覆盖算法(K-Partitioning Algorithm, KPA)为分别选取不同测试样本类中的中心向量点作为支撑点,获得同类样本

的超球覆盖,然后在超球内进行 K 网络近邻点连接。算法描述如下:

1) 初试化:$i = 1, i \in \{1,2,\cdots,m\}, k = 1$;

2) 从当前层次的样本队列中,选取测试样本空间中心点 $x_0 = \frac{1}{card(S_i)} \sum_{l}^{card(S_i)} x_l, x_l \in S_i, j = 0$;

3) 以 x_0 为支撑点,使用 $\psi_{i,j}^k$ 神经元覆盖 S_i,覆盖得到 S_i 的子集合为 $W, W \subseteq S_t$;

4) 若 $W = \phi$,则选取 $x_0 = \{x \mid \max(\rho(x_0, x)), x \in S_i\}$,继续步骤 3(此时同类样本至少有一个点 x_0);

5) 选取此类与异类样本的最小距离 ρ,作为超球的半径 r;

6) 在超球范围内,以 x_0 为中心以 r 为半径进行内部的邻域覆盖,找到超球内的近邻样本点 x_i, x_j,利用式(3)描述的方法构造当前覆盖层神经元族 $\{S(x; x_i, x_j, r)\}$。

7) 选取 $S_t = S_t \backslash W$,如果 $S_i = \phi$,从样本队列中去除 S_i 类别,否则,S_i 继续进入下一层样本队列;

8) 判别当前层次样本队列是否为空,如果为空,则继续步骤 9;否则 $i = i + 1$,继续 2);

9) 判别下一层次的样本队列是否为空,是,则结束;否则 $k = k + 1$,继续 2)。

同类样本连续性、紧致性体现了事物的本身特征,*KPA* 算法正是利用事物高维空间特征分布,通过超球内不同方向神经元覆盖,从而实现了对原有样本的 K 覆盖,用较少体积的 K 覆盖代替大体积的各向同性超球覆盖。

2.2 中心适配选择算法

中心适配选择算法(Center Adaptive Selection Algorithm, CASA)的基本思路是获取当前样本点的中心 x_0,以及获得与异类样本点最近的半径 r,以 x_0 为中心,r 为半径的超球覆盖同类样本点。算法描述如下:

1) 定义起始 $i = 1, i \in \{1,2,\cdots,m\}, k = 1$;

2) 从当前层次的样本队列中,选取测试样本空间中心点 $x_0 = \frac{1}{card(S_i)} \sum_{l}^{card(S_i)} x_l, x_l \in S_i, j = 0$;

3) 以 x_0 为支撑点,用 $\psi_{i,j}^k$ 神经元覆盖 S_i,覆盖得到 S_i 子集为 $W, W \subseteq S_t$;

4) 若 $W = \phi$,则从新选择样本中心点 $x_0 = \{x \mid \max(\rho(x_0, x)), x \in S_i\}$,继续 3)(此时同类样本至少有一个点 x_0);

5) 选取 $S_t = S_t \backslash W$,如果 $S_i = \phi$,从样本队列中去除 S_i 类别,否则,S_i 继续进入下一层样本队列;

6) 判别当前层次样本队列是否为空,如果为空,则继续 7);否则 $i = i + 1$,继续 2);

7) 判别下一层次的样本队列是否为空,是,则结束;否则 $k = k + 1$,继续 2)。

3 实验结果与数据分析

实验中选择了 Iris 三分类问题进行类别分类比较,三类植物长度特征,其中包括萼片长度、宽度、花瓣长度和宽度。共计 150 个样本点,随机获取一定比例的数据集合用于训练,剩下作为测试集。

图 6 描述了 KPA 、CASA 算法下 PONN 神经网络构造和识别情况,显示网络在不同训练样本点情况下的变化情况,其中 CASA 算法使用的神经元数比 KPA 算法的神经元数量要少,但是对于小样本数量训练集 KPA 算法具有比 CASA 更好的识别效果,同时在实际应用中能够使用的训练样本点相对于整个样本空间而言是有限的,所以 KPA 算法比 CASA 在实际应用中更有应用价值。

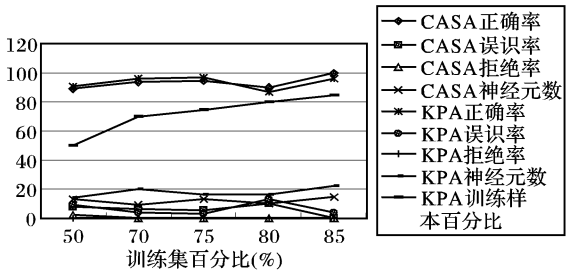

图 6 KPA 与 CASA 对不同训练集样本情况

表 1 3/4 训练集葡萄酒分类比较

	正确率(%)	误识率(%)	拒绝率(%)	神经元数	时间(s)
KPA	83.721	16.279	0	57	0.06
CASA	81.395	16.279	2.326	43	0.055

表 1 中描述了实际应用中同地区的不同 3 类葡萄酒分类测试情况,共计样本 178 个,每类个数分布为 59、71、48,样本特征包括:酒精度、苹果酸、花色素苷等 13 个主要化学特征。值得说明的是测试结果是通过多次不同条件测试得到的平均结果,主要是为了对比两种不同算法对网络构造和识别的效果。从表 1 中的数据来看,KPA 算法识别率比 CASA 算法识别率高,但 KPA 比 CASA 算法使用的神经元数要多,构造时间开销大,对于样本的空间连续性 KPA 比 CASA 具有更好的表现,能够减少误识率、拒识率,可以看出 KPA 比 CASA 算法在构造 PONN 网络时更具潜力。

通过实验可以看出:

1) KPA 比 CASA 算法具有更好的表现事物空间特征的连续特征;

2) 结合待识别事物特征的高维空间特征分布,有利用于提高网络构造性能。

PONN 网络不同于传统网络的类划分方法,而由高维空间覆盖角度来分析和构造网络,也就克服了传统网络例如 BP 网络,难于训练、中间层神经元难于确定、知识表示缺陷等问题。当然,如何更好地构造 PONN 网络,还需利用待识别事物本身特征的空间分布特征作为先验知识,这有待下一步研究。

参考文献:

[1] BLUM A, RIVEST RL. Training a 3-node neural network is NP-complete[J]. Neural Networks, 1992, 5: 117 - 127.

[2] 俞宗泉. 人工神经网络发展五十五年[J]. 自动化与仪表, 1998, 5(13): 1 - 4.

[3] BURGES CJC. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discvery, 1998, 2(2): 121 - 167.

[4] 王守觉. 仿生模式识别(拓扑模式识别)——一种模式识别新模型的理论与应用[J]. 电子学报, 2002, 30(10): 1417 - 1420.

[5] WANG S. Priority Ordered Neural Networks with Better Similarity to HumanKnowledge Representation[J]. Chinese Journal of Electronics, 1999, 8(1); 1 - 4.

[6] SEUNG HS, LEE DD. The Manifold Ways of Perception[J]. Science, 2000, 290: 2268 - 2269.