一种基于神经网络覆盖构造法的模糊分类器*

叶少珍[1,2,3+], 张 钹[1,2], 吴鸣锐[1,2], 郑文波[3]

[1](清华大学 计算机科学与技术系,北京 100084)

[2](清华大学 智能技术与系统国家重点实验室,北京 100084)

[3](福州大学 信息科学与技术学院,福建 福州 350002)

A Fuzzy Classifier Based on the Constructive Covering Approach in Neural Networks

YE Shao-Zhen[1,2,3+], ZHANG Bo[1,2], WU Ming-Rui[1,2], ZHENG Wen-Bo[3]

[1](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[2](State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

[3](College of Information Science and Technology, Fuzhou University, Fuzhou 350002, China)

+ Corresponding author: Phn: 86-591-7609281, E-mail: yeshzh@vip.sina.com

http://www.tsinghua.edu.cn

Received 2001-10-08; Accepted 2002-05-13

Ye SZ, Zhang B, Wu MR, Zheng WB. A fuzzy classifier based on the constructive covering approach in neural networks. *Journal of Software*, 2003,14(3):429~434.

Abstract: A geometrical representation of M-P model is firstly introduced, by which the training problem of neural networks may be transformed into the covering problem of a point set. According to this, the geometrical algorithm of neural network training is analyzed. The algorithm may be used for constructing very complicated classifying boundary, but it has higher time complexity. So a fuzzy classifier based on the combination of the covering approach and fuzzy set theory is proposed. The classifier can improve the speed of training and decrease the number of covering sphere-neighborhoods, i.e., decrease the number of hidden nodes of neural networks. The fuzzy set based approach may also provide multi-choices for pattern recognition problems of large scale. Recognition of 700 handwritten Chinese characters is used to test the performance of the approach and the results are promising.

Key words: neural network; pattern recognition; fuzzy classifying; sphere-neighborhood model

摘 要: 首先介绍了一种 M-P 模型几何表示,以及利用这种几何表示可将神经网络的训练问题转化为点集覆盖问题,并在此基础上分析了神经网络训练的一种几何方法.针对该方法可构造十分复杂的分类边界,但其时间复杂度很高.提出一种将神经网络覆盖算法与模糊集合思想相结合的方法,该分类器可改善训练速度、减少覆盖的球领域数目,即减少神经网络的隐结点数目.同时模糊化方法可方便地为大规模模式识别问题提供多选结果.

* Supported by the National Natural Science Foundation of China under Grant No.60135010 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.G1998030509 (国家重点基础研究发展规划(973))

第一作者简介: 叶少珍(1963－),女,福建福州人,博士生,副教授,主要研究领域为模式识别,神经网络,模糊信息系统.

用 700 类手写汉字的识别构造一个大规模模式识别问题测试提出的方法,实验结果表明,该方法对于大规模模式识别问题很有潜力.

关键词: 神经网络;模式识别;模糊分类;球面邻域模型

中图法分类号: TP18 **文献标识码**: A

大规模模式识别问题特指一类较难解决的模式识别问题.该类问题的特点是,特征空间维数高,样本数量大,而且所涉及的类别数目多.许多实际的模式识别问题,如汉字识别、非特定人语音识别都属于这类问题.通常,这类问题所涉及的分类边界十分复杂,因此分类器的构造相应地也十分困难,消耗时间也比较长.

张铃[1]提出了一种 M-P 模型的几何表示,利用这种几何表示,神经网络的训练问题可转化为点集覆盖问题.在此基础上,文献[2]给出一种神经网络训练的几何方法.该方法可构造十分复杂的分类边界,但其时间复杂度很高.因此,对于实际的大规模问题,有必要引入模糊化的方法.其目的是通过减少神经网络的覆盖构件数而降低神经网络结构的规模,因此在不影响原有精确度的情况下,可有效地提高识别的速度.本文提出一种基于神经网络覆盖构造法的模糊分类方法 FCSN(fuzzy-covered with sphere neighborhood),它主要考虑了降低训练和识别时间复杂度的问题.

为了保证一定的完整性,本文首先简单介绍文献[1]中给出的几何表示以及在此基础上文献[2]给出的一种神经网络训练的几何方法.然后介绍 FCSN 算法,包括它的基本思想、实现方法以及其优、缺点.最后给出实验结果和结论.

1 神经网络构造的覆盖方法

不失一般性,可假定输入样本集合为 n 维空间中一有界集合 D,构造如下一一映射:

$T:D\to S^{n+1}, X\in D, T(X)=(X,(d^2-|X|^2)^{1/2})$,其中 $d\geq\max\{|X||X\in D\}$,S^{n+1} 是一个 $n+1$ 维的球面.由于映射 T 是一一映射,所以在后面的讨论中,都假定所有输入样本的模长均相等,即输入样本都处在一个 n 维空间的球面上.

1.1 球面邻域模型

这一节简要介绍M-P 神经元的球面邻域模型.一个 M-P 神经元是一个 n 输入、单输出的处理单元,即

$$y=sgn(W^TX-\Psi),$$

其中 $X=(x_1,x_2,\ldots,x_n)^T$ 表示输入向量;$W=(w_1,w_2,\ldots,w_n)^T$ 表示权向量; Ψ表示阈值.

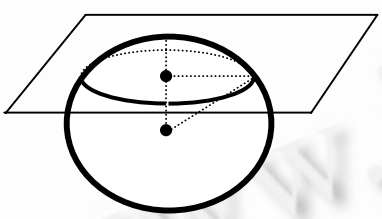

Fig.1 A sphere neighborhood

图 1 球面邻域模型

$$sgn(v)=\begin{cases}1, & v\geq 0,\\ -1, & v<0.\end{cases} \tag{1}$$

$W^TX-\Psi=0$ 所表示的超平面将样本空间分为两个半空间,$H^+:W^TX-\Psi>0$ 以及 $H^-:W^TX-\Psi<0$.又由前面的假定,输入样本均处于一个 n 维的球面 S^n 上,因此,$W^TX-\Psi>0$ 所代表的点集就是球面 S^n 与 H^+的交集,如图 1 所示,这里我们把球面 S^n 与半空间 H^+的交集称为(该 M-P 神经元所对应的)球面邻域.当一个输入向量 X 属于球面邻域时,即 X 被球面邻域覆盖时,神经元的输出为 1 ,否则为–1.

1.2 CSN网络的基本框架

由前面所述,一个 M-P 神经元对应于一个球面邻域,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邻域.在 CSN 网络中,采用同类的训练样本被同一组球面邻域覆盖,而不同类的训练样本被不同组的球面邻域覆盖的具体方法,并考虑用尽量少的球面邻域完成覆盖任务的优化问题.在判断一个输入样本的类别时,只要检测该样本被哪一组球面邻域所覆盖,就可以知道它所对应的类别.特别是对于一个输入样本 X 没有被任何一组球面邻域覆盖,通过定义 X 对各组球面邻域的隶属度函数$\mu_c(X)$,其中 $\boldsymbol{X}$ 是输入样本向量,C 是某一组球面邻域.$\mu_c(X)$的取值如下:若 X 被 C 中某一个球面邻域所覆盖,则$\mu_c(X)=1$,否

则$\mu_c(X)=1/(dist(X,C)*M)$,其中 $dist(X,C)$是 X 与 C 的距离函数,M 是一个正整数,使得 $1/(dist(X,C)*M)$的值小于 1.

综上所述,CSN 网络的基本思路是用由许多简单球面邻域所覆盖的小的几何区域的组合来"勾勒"出输入样本的几何分布.当判断一个输入样本的类别时,CSN 网络的作用实质上是选出隶属度函数$\mu_c(\boldsymbol{X})$取最大值的区域所对应的类别.相应地,我们用 700 类手写体汉字的识别问题构成了一个大规模的模式识别问题,利用 CSN 算法进行测试的结果见表 1.

Table 1 Performance of CSN algorithm

表 1 CSN 算法的性能

Number of classes	Training time (s)	Recognition rate top 1 (%)	Recognition rate top 3 (%)
300	5 481	95.5	99.2
500	16 786	94.5	98.6
700	31 089	93.6	98.3

1.3 CSN网络及算法的分析

由于球覆盖方法的直观性,当人们将样本的分类转变成覆盖问题时,可迅速地、构造性地得到对于训练数据百分之百正确分类的神经网络,而不必像传统的 BP 算法那样反复训练,而且还不一定能得到好的结果.又由于输入向量的类别完全取决于它被哪些球邻域所覆盖,因此这种方法的另一个优点是构造出的网络的行为便于分析.在 CSN 网络中先利用球面邻域模型把神经网络的训练问题转化为几何中的覆盖问题,然后考虑对训练样本具体覆盖方法和用尽量少的球面邻域完成覆盖任务的优化问题,与构造神经网络的几何方法中涉及的两个问题是相同的.从表 1 可以看出,单选和三选的结果都是令人满意的,而且,随着类别数目的增加,识别率下降得不是很快,而且由于三选的识别率很高,这对于进行后处理(如根据上、下文进一步分类)是很有帮助的.因此 CSN 算法具有较强的分类能力.成功的关键是 CSN 覆盖算法构造出的球面领域所覆盖的几何区域能够真实地反映脱机手写汉字各类别样本在空间中的分布情况.

但是从表 1 还可以看出,该算法的训练时间随着类别数的增加而迅速增加.还可看出,尽管在 CSN 网络结构中,考虑了球面领域覆盖数尽量少的优化问题,但对于大模式类的识别问题的实用化,时间复杂性的问题仍然是它的瓶颈.

因此,在本文中将模糊数学的思想[3,4]引入到 CSN 网络中,允许每个球面领域在其边界处覆盖若干个非同类的训练样本,使每个球面领域可覆盖更多的样本点,同时去掉覆盖同类样本极少的球面覆盖领域,这样可减少球面领域的覆盖总数,使网络的结构简单,因而也极大地减少了时间复杂度.这样的处理也比较接近脱机手写汉字的实际分布情况,因为对国标一级的 3 755 个汉字,字符集庞大,同时手写体汉字风格更是因人而异,同一字体的笔画长短、笔画粗细、笔画方向、笔画位置都不相同,尤其对相似字体的交叉是不可避免的.同时,从分析可知,CSN 网络的误识率很大一部分原因来自于未被覆盖的测试点,因此,这里把覆盖得到的每一类所有覆盖中心作为反映脱机手写汉字分布的参考点,利用 RBF 径向基函数为基础定义模糊判决的隶属函数 $U=(\mu_1(X),\mu_2(X),\ldots,\mu_m(X))$,隶属函数各分量的最大者作为识别的结果.

2 FCSN 网络及其实现的算法

2.1 FCSN网络的基本框架

FCSN 网络的基本思路是利用能模糊覆盖同类训练样本的球面覆盖邻域的一组中心作为该类模式在分布空间的参考点,用多类参考点的集合近似地表示输入样本在空间的几何分布.当判断一个输入样本的类别时,FCSN 网络的作用实质上是选出隶属度函数取最大值的参考点所对应的类别.FCSN 网络示意图如图 2 所示.

图中各参数说明如下:设共有 K 个类别,由 FCSN 算法对训练样本集合学习得到覆盖球面领域的中心向量集合为 $M=\{m_i|1\le i\le D$,其中 D 为中心向量总数$\}$,$I(u)$为第 $u(1\le u\le K)$个类别对应的中心点的下标集合;隶属函数向量 $U=(\mu_1(X),\mu_2(X),\ldots,\mu_k(X))$,其中$\mu_c(X)$是样本 X 属于第 $c(1\le c\le K)$个类别的隶属函数,定义如下:

$$\mu_c(\boldsymbol{x}) = \frac{C_c(X)}{\sum_{u=1}^{K} C_u(X)}.$$

这里, $C_c(X) = \sum_{i \in I(c)} \exp(-((d_i)^2 / 2\sigma^2))$, $(d_i)^2 = (\|X - m_i\|)^2$. σ是表示 RBF 函数宽度的参数.

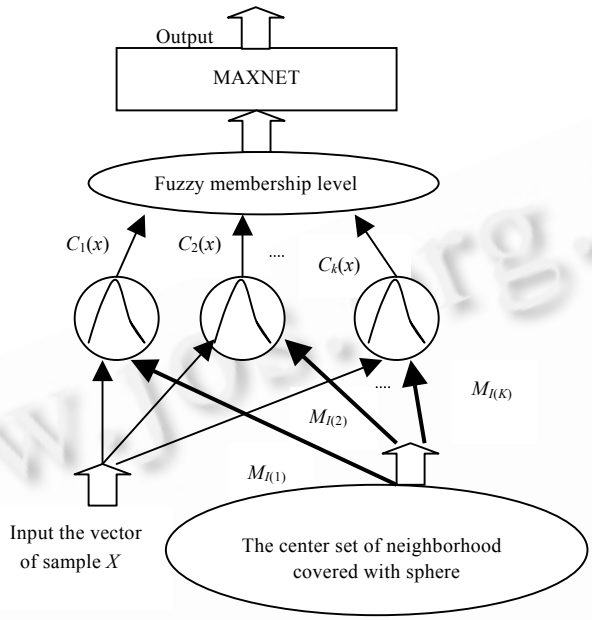

Fig.2 Schematic diagram of a *K*-class FCSN network
图 2 *K*-类 FCSN 网络示意图

2.2 FCSN网络的实现算法

本节给出 FCSN 网络的训练算法,即模糊球面邻域构造算法.

假定有 *K* 类训练样本,那么,如前所述,FCSN 网络的训练要解决的问题是要构造 *K* 组球面邻域,*C*(*i*),*i*=1,2,...,*K*,使得 *C*(*i*)组中的每一个球面领域覆盖大多数第 *i* 类的训练样本,而在球面领域边界处覆盖极小数的非 *i* 类的训练样本.

具体的模糊球面邻域的构造算法思路如下:在构造一个初始球面邻域后,它覆盖了第 *i* 类(1≤*i*≤*K*)的若干样本,然后不断移动其中心并改变其域值,使得在覆盖 *N*1 个非 *i* 类样本的同时有更多的第 *i* 类样本被该模糊球面邻域所覆盖.这样,就可以用尽量少的模糊球面邻域覆盖第 *i* 类的所有训练样本,而且,这样构造出的球面邻域所覆盖的区域是第 *i* 类样本的"高密度"模糊区,即可以对样本在空间的分布给出比较好的模糊近似.因此,FCSN 网络实质上是从几何角度描述样本在空间的分布.

详细的 FCSN 算法描述如下:

对于第 *i* 类训练样本,1≤*i*≤*K*,令 *P*(*i*)表示第 *i* 类训练样本;对于任何一个 *P*(*i*)中的样本点,如果它没有被 *C*(*i*)中任何球面邻域覆盖,则称其为 NCV-point;*N*1 是每个球面领域允许覆盖的最大非 *i* 类样本数.

初始时,*C*(*i*)是空集,*m*=0,*m* 为 *C*(*i*)中球面邻域的数目.因此,此时所有 *P*(*i*)中的点都是 NCV-point.*C*(*i*)的构造算法如下:

(1) 若 *P*(*i*)中的点没有 NCV-point,则转到(9),否则,任选一个 NCV-point $\boldsymbol{X} \in P(i)$,同时 *j*=1;

(2) 构造一个球面邻域 SN,使得其权值 $\boldsymbol{W}=\boldsymbol{X}$(即 SN 以 $\boldsymbol{X}$ 为中心),$\psi=d(\boldsymbol{W})$,其中

$$d(\boldsymbol{W}) = \max_{Y \notin P(i)} \{\langle \boldsymbol{W}, \boldsymbol{Y} \rangle\}. \quad (2)$$

这里,$\langle \boldsymbol{W}, \boldsymbol{Y} \rangle$表示向量的点积,由于 $\boldsymbol{Y}$ 在球面上,所以$\langle \boldsymbol{W}, \boldsymbol{Y} \rangle$越大,$\boldsymbol{W}$ 与 $\boldsymbol{Y}$ 的距离就越小;

$d(\boldsymbol{W})$,$\boldsymbol{W}$ 与 $\boldsymbol{Y}$ 的最小"距离",$\boldsymbol{Y}\notin P(i)$;

构造 SN 后,执行 SNTMP←SN(将 SN 的参数拷贝到一个临时的球面邻域中).注意,虽然此时 $\boldsymbol{X}$ 被球面邻域 SN 覆盖,但是由于 SN$\notin C(i)$,$\boldsymbol{X}$ 仍然是 NCV-point;

(3) 将 SN 的中心(即权值向量 $\boldsymbol{W}$)移动到所有被 SN 覆盖的 $P(i)$中的 NCV-point 的重心,并根据(2)重新计算其阈值;

(4) 如果 SN 覆盖的 $P(i)$中的 NCV-point 的数目多于 SNTMP 覆盖的 $P(i)$中的 NCV-point 数目,则执行 SNTMP←SN,转到(3),否则执行 SN←SNTMP,转到(5);

(5) 设 $\boldsymbol{V}$ 是属于 $P(i)$且没有被 SN 覆盖的所有 NCV-point 中离 $\boldsymbol{W}$ 最近的点,其中 $\boldsymbol{W}$ 是 SN 的权向量,则执行 $W\leftarrow(W+V)/2$;

(6) 如果 SN 覆盖的 $P(i)$中的 NCV-point 的数目多于 SNTMP 覆盖的 $P(i)$中的 NCV-point 数目,则执行 SNTMP←SN,转到(3),否则执行 SN←SNTMP,转到(7);

(7) 将 Y 从整个训练样本集中去掉,$j=j+1$.如果 $j<N1$,转到(2),否则转到(8);

(8) 将 SN 加入 $C(i)$,执行 $m\leftarrow m+1$,转到(1);

(9) 结束.

2.3 FCSN算法小结

FCSN 算法的关键思路是用神经元覆盖区域的组合近似"勾勒"出各类样本分布的几何区域,当判断一个输入样本的类别时,FCSN 网络的作用在实质上是选出隶属度函数$\mu_c(\boldsymbol{X})$取最大值的区域所对应的类别.因而 FCSN 网络用于模式识别问题有许多优点.首先,由于训练过程可以保证每个训练样本都被与它所属类别对应的球面邻域组所模糊覆盖,因此,FCSN 网络的训练不存在不收敛的问题,而且,由于训练算法保证了每个训练样本都被某个主要覆盖同类样本的球面邻域覆盖,因此对于训练样本的识别率基本可达到 100%.另外,对于一个用于实际问题的分类器,经常需要给出多选的结果,这在 FCSN 网络中很容易实现.例如,只要给出隶属度函数值最大的 3 组球面邻域所对应的 3 个类别,即可实现三选,并且可通过隶属函数中高斯函数宽度参数σ的调整,使实验结果更符合单选或多选的情况.可根据这些优点使得 FCSN 算法非常适用于大规模且分类边界复杂的模式识别问题.

虽然 FCSN 算法可以描述样本在空间中任意复杂的分布,而且收敛性也可以得到保证,但是从式(2)可以看出,当构造每一个球面邻域时,都必须计算球面邻域的中心到所有样本的"距离",因此 FCSN 算法的时间复杂度是 $O((K*S)^2)$,其中 K 是类别数目,S 是每类中的训练样本数目.在大规模的模式识别问题中,K 和 S 通常都比较大,这就限制了 FCSN 算法的应用范围.但与 CSN 算法相比,由于模糊化的处理使训练算法复杂度问题有所改善.

3 实验结果

本节介绍用 700 类手写体汉字的识别问题来构成一个大规模的模式识别问题,并用它对文中提出的 FCSN 算法进行测试.

由于字型相似的汉字在区位码上也相近,因此,我们选取区位码在 1 601~2 660 的前 700 个国标一级汉字用于实验.实验中,每个汉字样本采用方向线素[5]的特征提取方法转化为一个 256 维的向量,每个汉字有 130 个样本,任选其中 70 个用来训练,其余 60 个用来测试.因此,共有 49 000 个 256 维、分别属于 700 个不同类别的训练样本,这就构成了一个典型的大规模模式识别问题.实验在 PIII-300 的微机上进行.

3.1 FCSN算法的性能测试

测试结果见表 2.表 1 中的单选识别率和三选识别率都是对测试集样本的识别率.

从表 2 可以看出,单选和三选的结果都是令人满意的,而且,随着类别数目的增加,识别率下降得很慢,而且由于三选的识别率很高,这对于进行后处理(如根据上、下文进一步分类)是很有帮助的,FCSN 算法具有很强的分类能力.从表 2 还可以看出,该算法的训练时间随着类别数的增加而迅速增加,但明显要比 CSN 算法的训练时间小,其原因是经过模糊化处理之后,训练时构造的覆盖球面数有所减少.

Table 2 Performance of FCSN algorithm ($N1$=1)

表 2 FCSN 算法的性能($N1$=1)

Number of classes	Training time (s)	Recognition rate top 1 (%)	Recognition rate top 3 (%)
300	4 050	95.4	99.2
500	12 525	94.3	98.4
700	24 152	93.5	98.0

3.2 参数*N1*的选取

这组实验用于测试 FCSN 算法中参数 $N1$ 的不同取值对算法性能的影响.在这组实验中,类别数目取值为700.实验结果见表 3.

Table 3 The experimental results based on different parameter $N1$

表 3 不同参数 $N1$ 取值的实验结果

$N1$	Training time (s)	Recognition rate top 1 (%)	Recognition rate top 3 (%)
2	23 123	93.4	97.8
5	21 465	92.9	97.6
8	17 436	92.8	97.5
10	12 478	91.4	97.1

从表 3 可以看出,虽然随着 $N1$ 的增大,识别率下降,但三选识别率变化较小.而当 $N1$ 较大时,训练时间也较少,因此实际运用该算法时,须折衷考虑识别率和训练时间,选择 $N1$ 满足实际识别问题的要求.

4 结 论

本文针对大规模模式识别问题,介绍了一种基于神经网络覆盖构造法的模糊分类器设计方法.神经网络覆盖构造法从根本上解决了多层前向神经网络在处理大模式类识别问题时,使用常规BP算法训练不收敛的问题,且对于训练样本的识别率为100%.模糊数学是目前智能系统中普遍使用的软计算的方法之一,其目的是适应现实世界普遍存在的不精确性,通过开拓对部分真实的容忍,达到可处理性、可鲁棒性、低成本求解以及与现实更好的联系.本文提出的 FCSN 分类器是利用以上两种方法的优点,通过集成来实现的,分类器解决脱机手写汉字识别的大模式类识别问题,符合由于种类多、各人的书写变化大而造成的分类边界复杂的情况.实验结果表明,该算法具有直观、分类能力强等优点,并有效地降低了训练过程的时间复杂度,十分适用于大规模的模式识别问题.

References:

[1] Zhang L, Zhang B. A geometrical representation of McCulloch-Pitts neural model and its applications. IEEE Transactions on Neural Networks, 1999,10(4):925~929.

[2] Wu MR. The research on classifier design for pattern recognition problems of large scale [Ph.D. Thesis]. Beijing: Tsinghua University, 2001 (in Chinese with English Abstract).

[3] Yang TN, Wang SD. Fuzzy auto-associative neural networks for principal component extraction of noisy data. IEEE Transactions on Neural Networks, 2000,11(3):808~810.

[4] Zhang D, Pal SK. A fuzzy clustering neural networks (FCNs) system design methodology. IEEE Transactions on Neural Networks, 2000,11(5):1174~1177.

[5] Ma SP, Xia Y, Zhu XY. Handwritten Chinese characters recognizing based on fuzzy directional line element feature. Journal of Tsinghua University (Science and Technology), 1997,37(3):42~45 (in Chinese with English Abstract).

附中文参考文献:

[2] 吴鸣锐.大规模模式识别问题的分类器设计研究[博士学位论文].北京:清华大学,2001.

[5] 马少平,夏莹,朱小燕.基于模糊方向线素特征的手写体汉字识别.清华大学学报(科学与技术),1997,37(3):42~45.