一种用于大规模模式识别问题的神经网络算法*

吴鸣锐, 张 钹

(清华大学 计算机科学与技术系,北京 100084);
(清华大学 智能技术与系统国家重点实验室,北京 100084)
E-mail: wmr@s1000e.cs.tsinghua.edu.cn
http://www.tsinghua.edu.cn

摘要: 许多实际的模式识别问题如对手写体汉字的识别,都属于大规模的模式识别问题.目前,传统的神经网络算法对这类问题尚无有效的解决办法.在球邻域模型的基础上提出一种可用于大规模模式识别问题的神经网络训练算法,试图加强神经网络解决大规模问题的能力,并用手写体汉字识别问题检验其效果.实验结果揭示了所提算法是解决大规模模式识别问题的一个有效且具有良好前景的方法.

关键词: 神经网络;模式识别;字符识别;训练算法;球邻域模型
中图法分类号: TP18 **文献标识码**: A

神经网络发展至今,由于其并行和容错等特点,在很多方面,尤其是在模式识别领域已经表现出一定的优势和潜力.然而,随着研究的深入,人们已经逐渐认识到就目前情况来看,神经网络尚不能有效解决大规模的模式识别问题.大规模的模式识别问题是指问题的特征空间维数高,样本数量大而且类别多.像手写体汉字识别、汉语的音节识别等许多实际问题都是这类问题的典型代表,因此,解决这类问题对于神经网络理论上的完善以及技术上的实用化都具有重要的意义.

张铃等人给出了一种新的M-P神经元的几何意解释——球邻域模型[1],并在此基础上提出了多层前向网络的交叉覆盖设计算法[2].该算法将神经网络的训练转化为几何的覆盖问题,思路独特,而且已经在小规模问题中取得了较好的结果[2].本文以球邻域模型为基础提出一种新的前馈神经网络训练算法,并以手写体汉字为例,进一步研究神经网络处理大规模问题的能力.

用单一的网络同时处理几千个汉字尚有很多困难,在初步研究阶段,较为可行的方法是先将汉字进行粗分类,然后对每个"粗类"中的样本进一步识别.本文尝试识别300个手写汉字,样本空间是256维,每个汉字有130个样本,其中训练样本数目是70,测试样本数目为60,即网络要处理39 000个256维、共分为300类的向量,这样的规模对于现有的神经网络模型来说仍然是个很大的挑战.

本文第1节阐述所采用的构造算法及其基本思想,并从原理上与传统的模式识别方法作简要的比较.第2节介绍实验的数据和结果.第3节给出结论.

1 算法及其基本思想

球邻域模型及其应用的详细内容可参见文献[1],这里只从其中一个角度阐述其基本思想以及

* 收稿日期: 1999-09-09; 修改日期: 2000-03-15
基金项目: 国家自然科学基金资助项目(69823001);国家重点基础研究发展规划973资助项目(G1998030509);高等学校博士学科点专项科研基金(98000335)
作者简介: 吴鸣锐(1973-),男,湖南汉寿人,博士生,主要研究领域为模识识别,组合优化,神经网络;张钹(1935-),男,福建人,教授,博士生导师,中国科学院院士,主要研究领域为人工智能,计算机应用技术.

在大规模模式识别问题中的应用.

模式识别的 Bayes 决策理论中要求知道样本的概率分布,这在实际问题中是很难做到的,所以,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的,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.本文提出的算法就是向这个方向发展的一种初步的尝试,即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.该算法用多个 MP 神经元所对应的多个"球邻域"[1]去覆盖各类的所有训练样本,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.下面,我们具体介绍算法的内容,并以写体汉字识别为例,检验它对于解决大规模模式识别问题的有效性.首先引入若干要用到的记号.

设共有 N 类样本,记第 $C(1\leqslant C\leqslant N)$ 类样本集合为 $S(C)$,其元素个数为 $|S(C)|$,设样本空间的维数为 D.在算法运行之前,首先把所有样本都映射到一个 $D+1$ 维的球面 SP 上[1].记 C 类的第 m 个样本为 $X_{k,m}$,$1\leqslant m\leqslant S(C)$,用 $\langle a,b\rangle$ 表示向量 a,b 的点积.

每个样本都有一个覆盖标记 $Cover(C,i)$,$1\leqslant C\leqslant N$,$1\leqslant i\leqslant|S(C)|$,该标记为 TRUE 时表示样本 $X_{C,i}$ 已被某个球邻域覆盖;当标记为 FALSE 时,则表示该样本尚未被任何球邻域所覆盖.

在训练过程中,每个 MP 单元都有一个类别标号,取值为 1 到 N 之间的整数,表示被它覆盖的样本所属的类别.

设 $1\leqslant C\leqslant N$,$1\leqslant i\leqslant|S(C)$,

令 $\mathrm{Max}(C,i)$ 表示样本向量 $X_{C,i}$ 与其他类中的样本的最大点积;

令 $\mathrm{Min}(C,i)=\min\{\langle X_{C,i},X_{C,j}\rangle\}$,其中 $Cover(C,j)$ 为 FALSE 且 $\langle X_{C,i},X_{C,j}\rangle>\mathrm{Max}(C,i)$.下面给出具体的算法.

1.1 训练算法

首先标记所有样本的覆盖标记为 FALSE,表示初始时所有样本都未被覆盖.a,b 是预先确定的小于 1 的非负常数,且 $a-b=1$.

```
Begin
  For C:=1 to N do        {C 表示类别}
    Begin
      For i:=1 to |S(C)| do
        Begin
          If Cover(C,i)=FALSE
          Then
          Begin
            计算 Min(C,i)和 Max(C,i);
            T←a*Max(C,i)+b*Min(C,i);
            在神经网络的隐层中加入一个 MP 单元,它以 X_{C,i} 为权值向量,以 T 为阈值,且该
```

```
                    神经元的类别标记为C;
                    将新加入的神经元所覆盖的样本的覆盖标记置为TRUE;
                End
            End
        求出将所有类别标号为C的隐层单元的权值向量的中心,记为Center(C),称为该类几何区域的
        "中心";
    End
加入一个输出层单元;
End
```

注:在上面的训练算法中,有几点需要加以说明.由于样本都被映射到球面上,因此点积实际上表示的是距离信息,点积越大,距离越小,可是点积的计算比求距离要方便;每个新加入的MP单元都只覆盖同一类中的若干样本,因此训练结束后,类别标号相同的一组神经元所覆盖的区域就是其所对应类别的样本分布区域的近似,而标号不同的神经元所覆盖的区域是没有交集的;最后还要计算每组标号相同的神经元的权值向量的中心向量,作为该区域的"中心";训练中,a,b为学习参数,它们的作用主要是控制新加入的隐层单元所覆盖区域的大小,可按需要加以调节;输出层单元的作用是在测试时根据隐层单元的输出得到最终的类别,因此,它可以用逻辑方法来替代,详见下面的识别算法.

1.2 识别算法

给定一个$D+1$维向量P(已经映射到球面上),识别其类别.算法如下:

```
Begin
    Result←0;   {Result为输出结果}
    For C:=1 to N do
    Begin
        若类别标号为C的隐层单元中存在输出为1的单元,则Result←C,并终止循环
    End
    If Result=0 Then
    Begin
        对于1≤C≤N,计算P到所有Center(C)的欧式距离,设距离最小所对应的类别为C0,则令Result←
        C0;
    End
    输出Result;
End
```

说明:判断向量P的类别的过程很直接,即首先看它是否被某个隐层单元所覆盖,若存在这样的隐层单元,则该隐层单元的标号就是P的类别;若不存在覆盖P的隐层单元,则计算P距离哪个区域的"中心"最近,距离值最小的"中心"所对应的类别就是P的类别.

下面将上述算法应用于手写体汉字识别,并给出有关的实验和结果.

2 实验及结果

在实验中,每个手写体汉字用方向线素[3]的特征提取方法转化为256的向量.

实验的汉字集合选取如下:原则上尽量选取相似的汉字,但是现在还没有度量多个(300个)汉字相似度的标准方法,这里利用了汉字的区位码,由于每个汉字可用其区位码描述,而且许多字型相似的汉字区位码也相近.例如,汉字"奥"、"懊"和"澳"的区位码分别为1634,1635和1636.实验中选取300个汉字区位码值范围为1601~1918.

在实验中,每个汉字共130个样本,任意选取70个作为训练样本,余下60个作为测试样本,总计39 000个样本.

学习参数的选取:$a=0.1, b=0.9$.

由算法的特性可知,训练之后的网络对于训练样本的识别率必定为100%,因此在测试识别率时只对没有参与训练的样本进行统计.

结果见表1.

Table 1 Experimental results

表1 实验结果

Number of classes①	Number of M-P neurons②	Recognition rate③(%)	Number of training samples④
20	232	96.0	1 400
50	622	95.3	3 500
100	1 741	94.0	7 000
300	6 121	92.4	21 000

①类别数,②M-P单元数目,③识别率,④训练样本总数.

表1中列出的是实验的主要结果.第1列是用于测试的汉字集合的数目,分别为20,50,100,300.第2和第3列分别是隐层单元数目及识别率.最后一列是实验中要处理的训练样本总数,以示问题的规模.

3 结 论

对于本文提出的算法的测试结果可见表1.由于识别率只是对未参与训练的样本进行统计,因此可以看出,该算法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(back propagation)算法相比,有着明显的优势.众所周知,BP算法不仅不能保证对于训练样本的高识别率,而且当训练到一定程度时,经常会遇到"Over-Fitting"的问题.本文提出的算法之所以表现出良好的泛化能力,关键就在于它是从数据本身出发去逼近其分布的几何轮廓.因此,该算法对于解决像手写体汉字的识别这类大规模的问题是有效的.

从表1还可以看出,随着待识别类别的增加,识别率的下降不是很快,这也说明该算法对于解决大规模模式识别的问题有着重要的研究价值和很好的应用前景.

这里只是一个初步的实验,我们还计划做进一步的工作,如对更大规模的问题进行实验;研究如何进一步提高识别率;加入多选机制;改进覆盖算法以减少隐节点数目;改进加入的隐单元的参数(权值和阈值)的确定方法等.另外,还可尝试将该算法用于其他大规模的模式识别问题(如汉语的408个无调音节的识别等),以检验和挖掘该算法的潜力.

致谢 清华大学计算机科学与技术系的马少平教授和金奕江博士为本文提供了样本数据,在此表示感谢.

References:

[1] Zhang, Ling, Zhang, Bo. A geometrical representation of M P neural model and its applications. Journal of Software, 1998,9(5):334~338 (in Chinese).

[2] Zhang, Ling, Zhang, Bo, Yin, Hai-feng. An alternative covering design algorithm of multi-layer neural networks. Journal of Software, 1999,10(7):737~742 (in Chinese).

[3] Ma, Shao-ping, Xia, Ying, Zhu, Xiao-yan. 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).

附中文参考文献：

[1] 张铃，张钹. M-P神经元模型的几何意义及其应用. 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. 软件学报，1999，10(7)：737～742.

[3] 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.

A Neural Network Algorithm for Large Scale Pattern Recognition Problems*

WU Ming-rui, ZHANG Bo

(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*);
(*State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing* 100084, *China*)
E-mail: wmr@s1000e.cs.tsinghua.edu.cn
http://www.tsinghua.edu.cn

Abstract: Many practical pattern recognition problems, such as recognition of handwritten Chinese characters belong to the pattern recognition problems of large scale. Now conventional ANN (artificial neural network) algorithms cannot solve this set of problems efficiently. In this paper, a neural network algorithm based on the sphere neighborhood model is introduced, aiming at enhancing the neural network's ability to solve the pattern recognition problems of large scale. The performance of the algorithm is tested with the handwritten Chinese character recognition problem. Experimental results show that the proposed algorithm is competent and has well prospects to this set of problems.

Key words: neural network; pattern recognition; character recognition; training algorithm; sphere neighborhood model

* Received September 9, 1999; accepted March 15, 2000

Supported by the National Natural Science Foundation of China under Grant No. 69823001; the National Grand Fundamental Research 973 Program of China under Grant No. G1998030509; the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 98000335