多层前向神经网络在手写体数字识别应用中的研究

胡 健 汪庆宝 涂承宇

(北京工业大学电子工程学系, 100022)

摘 要 针对手写体数字的特点并从实用性的角度出发, 采用了一种融字符特征抽取和识别于一体的五层结构的前向传播网络. 为了减少连接权值的个数, 在网络中采用了权值共享和部分映射. 在系统的训练过程中, 利用误差函数的二阶导数加速网络的收敛, 取得了较好的结果, 比传统BP算法快4～5倍. 在此基础上, 利用二阶导数的信息对网络进行了神经损伤即鲁棒性实验, 使网络权值数大大减少.

关键词 神经网络, 模式识别, 字符识别, 鲁棒性

分类号 TP 391.4, O 235

在我们日常生活中, 使用阿拉伯数字字符的场合很多, 如邮政编码、支票金额、票据数字和各种数字数据等, 其中邮政编码是最典型的例子之一. 国外在这方面的研究异常活跃, 早在1955年西方国家就提出了印刷体数字的OCR. 70年代初, 美国为邮政部门研制了一种对信件进行光学阅读和地址解释的高性能样机. 八十年代初日本在OCR方面居于领先地位. 现在世界上许多国家都拥有邮件自动分检设备, 但在实际应用中这些设备只能对手写印刷体数字有高的识别率, 对非限制性手写体数字的识别还不尽人意.

我国对数字识别的研究工作开始于60年代, 尽管随后出现了各种各样的识别方法, 但与国外相比, 仍有一定的差距, 因而继续深入地研究手写体数字的识别具有深刻的理论意义和实用价值.

1 识别方案的确定

在传统模式识别方法中, 用统计模式识别和句法模式识别的方法进行的识别研究均有报导, 但效果都不十分理想. 从80年代初至今 , 应用这两种方法的手写体自动识别技术有了很大改进, 如多级识别系统、引入人工智能的混合识别系统等. 但是, 如果要求高速、准确地识别手写体数字, 传统的方法很难适应.

人工神经网络以其抗噪声、容错(变形)、自适应、自学习能力强, 融若干预处理和识别处理于一体, 识别速度极快等特点备受人们普遍重视. 于是利用人工神经网络进行手

收稿日期: 1995—12—12

写体数字识别的研究便应运而生.

用神经网络进行字符识别有两种方法. 一种方法是充分利用神经网络的特点，直接把整幅图像送入网络，由网络自动实现特征提取直至识别. 这种系统用硬件(神经网络芯片)实现后，分类速度极快，能更好地满足实际应用的需要. 但这种网络互连较多,待处理信息量大. 另一种方法是先进行字符特征抽取， 然后用所获得的特征来训练神经网络分类器. 这种方法要在字符特征抽取上花费一定的时间，而且系统的识别效果与字符特征的抽取有关，分类速度较慢. 作者从实用性出发，采用第一种方法，利用一个五层前向网络来进行手写体数字识别的研究. 网络的结构如图 1 所示. 网络输入是标准 16 × 16 点阵数字图像，输入经过三个隐含层向前传播至 10 个输出单元，每个单元代表 0 ～ 9 中的一个数字. 前两个隐含层包含有可训练的特征检测器.

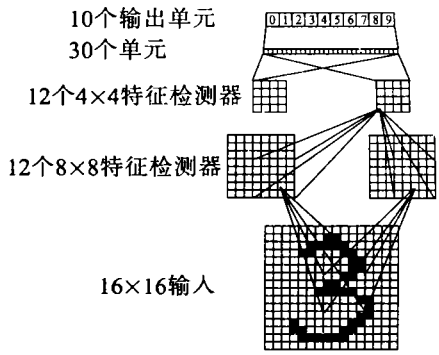

图1 识别网络的结构

为减少权值个数，克服网络训练过程中占用内存空间过大，运算时间过长的缺点，及增强网络抗变形输入模式的能力，该网络在连接中采取了“权值共享”和“ 部分映射”两个措施，其有效性将在第 4 节中阐明. 具体做法如下:

将输入的 16 × 16 点阵分割成 64 个 5 × 5 的子块，每个子块分别映射到第一隐含层的 12 个检测器中(每个检测器为 8 × 8 点阵),每个检测器的所有 64 个单元均含有 25个相同的权值,这样它们在网络的各个不同地方可检测到同一特征. 权值共享及 5 × 5 的接收子块将第一隐含层的权值由全连接方式的 16 × 16 × 8 × 8 × 12 ≈ 200 000 个减少到 25 × 12 = 300 个.

在第一隐含层向第二隐含层的映射过程中，用部分映射代替全映射，即第二隐含层中的每一个检测器仅由第一隐含层中抽取的 8 个检测器映射而成， 这样的组合共有 12 组. 在映射过程中，仍采用权值共享的 5 × 5 子块映射方法，从而使权值数由原来的 64 × 16 × 12 × 12 = 130 000 个减少到 25 × 8 × 12 = 2 400 个.

第三隐含层包含 30 个单元，与第二隐含层构成完全映射，权值数为 16 × 12 × 30=5 760 个.

最后一层是输出层,含有 10 个输出单元，输出层和第三隐含层之间的权值数为30 × 10 = 300 个.

在对输入的 16 × 16 点阵进行分割时不是任意的. 一方面要突出有用的信息，另一方面要对输入信息进行一定的处理，强调各个数字自身的特征. 虽然不同人所写的阿拉伯数字的大小、形状、相对于表格的位置、旋转角度不相同，但一般而言，整个数字的重心相对比较稳定. 作者基于这种考虑，仔细分析了 10 个阿拉伯数字，发现各个数字的中心部位比其它部位更能反映出其自身的特征. 或者说，数字符号的信息主要蕴涵在其图像的中心部位. 因此，在本文所采用的识别系统里，由第一层向第二层的映射中，充分突出了数字图像的中心部位的作用.

2　网络的训练

传统的 BP 学习算法是利用梯度下降法在权值空间搜索，使代价函数最小. 这种方法存在收敛速度慢和容易陷入某个局部极小值点两个问题. 在最优化的数学方法中解决这两个问题的一种有效方法是采用牛顿法. 从理论上说，牛顿法具有良好的收敛性——平方收敛，但计算工作量庞大，计算复杂度为 $O(n^3)$. 因此在实际应用中，一般采用准牛顿法而不直接采用牛顿法. 准牛顿法的权值调整公式为:

$$\Delta W = -\overline{\lambda H}^{-1}\nabla E(W) \tag{1}$$

H 是二阶导数 Hessian 矩阵:

$$H_{ij} = \frac{\partial^2 E}{\partial W_i \partial W_j}$$

$(\overline{H^k})^{-1}$是近似 Hessian 逆矩阵，有多种计算方法，其中最有效的方法是 BFGS 法. 这时的 $(\overline{H^k})^{-1}$计算公式为:

$$(\Delta\overline{H^k})^{-1} = \frac{[\Delta W^k - (\overline{H^k})^{-1}\Delta g^k](\Delta W^k)^{\mathrm{T}} + \Delta W^k[\Delta W^k - (\overline{H^k})^{-1}\Delta g^k]^{\mathrm{T}}}{(\Delta g^k)^{\mathrm{T}}\Delta W^k} -$$

$$\frac{[\Delta W^k - (\overline{H^k})^{-1}\Delta g^k]^{\mathrm{T}}\Delta g^k \Delta W^k (\Delta W^k)^{\mathrm{T}}}{[(\Delta g^k)^{\mathrm{T}}\Delta W^k][(\Delta g^k)^{\mathrm{T}}\Delta W^k]} \tag{2}$$

但这种方法的复杂度仍达到 $O(n^2)$，对图 1 所示具有 8 760 个权值的网络已超出一般微机的内存限制.

为了既可利用二阶导数的附加信息，使算法快速收敛到极小值点，又不增加太多的计算复杂度和存储空间，作者在参考大量文献的基础上，采用著名的 Levenberg and Marquardt 近似

$$H = EE^{\mathrm{T}} + \lambda\Omega \tag{3}$$

式中 λ 是正常数，控制 H 矩阵的计算、Ω 是一适当选择的矩阵. 令矩阵 Ω 具有下面的对角线形式:

$$\Omega = \begin{vmatrix} e_1^1(e_1^1)^{\mathrm{T}} & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & e_{M_1}^1(e_{M_1}^1)^{\mathrm{T}} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & e_1^L(e_1^L)^{\mathrm{T}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & e_{M_L}^L(e_{M_L}^L)^{\mathrm{T}} \end{vmatrix} \tag{4}$$

且 λ 足够大，则 H 矩阵就变成了对角线矩阵

$$H = \lambda\Omega \tag{5}$$

采用这种近似后，计算 H 矩阵时每次的运算量将减少到 $O(n)$的量级.

采用对角线近似后，Hessian 矩阵由于只有对角线元素，因此可以简化求逆矩阵运算.

直接利用 Hessian 项，把权值调整公式改写成：

$$\Delta W_{ij} = -a\frac{\frac{\partial E}{\partial W_{ij}}}{\left|\frac{\partial^2 E}{\partial W_{ij}^2}\right|} \tag{6}$$

在实际计算中，Hessian 项是一个非常小的量. 为了保证计算的稳定性，能够有效地调整权值，可以采用常用于改善 Hessian 矩阵状况的数字优化方法，即给Hessian 项加上一个小常数 μ. 这样权值调整公式可写成：

$$\Delta W_{ij} = -a\frac{\frac{\partial E}{\partial W_{ij}}}{\left|\frac{\partial^2 E}{\partial W_{ij}^2}\right| + \mu} \tag{7}$$

在实际训练中，作者采用此式修正权值. 式中的 Hessian 项，即二阶导数可由单个神经元的输入输出关系及前向网络的结构推出：

$$\frac{\partial^2 E}{\partial W_{ij}^2} = \frac{\partial^2 E}{\partial net_i^2}\left(\frac{\partial net_i}{\partial W_{ij}}\right)^2 = \frac{\partial^2 E}{\partial net_i^2} O_i^2 \tag{8}$$

$$\begin{aligned}\frac{\partial^2 E}{\partial net_i^2} &= \frac{\partial}{\partial net_i}[(O_i - t_i) f'(net_i)] \\ &= \frac{\partial (O_i - t_i)}{\partial net_i} f'(net_i) + (O_i - t_i)\frac{\partial f'(net_i)}{\partial net_i} \\ &= f(net_i)^2 - (t_i - O_i) f''(net_i)\end{aligned} \tag{9}$$

对于隐含层节点

$$\begin{aligned}\frac{\partial^2 E}{\partial net_i^2} &= f'(net_i)^2 \sum_i W_{ii}^2 \frac{\partial E}{\partial net_i} - \\ &\quad f''(net_i) \sum_i \frac{\partial E}{\partial net_i} W_{ii}\end{aligned} \tag{10}$$

式中

$$\frac{\partial E}{\partial net_i} = -(t_i - O_i) f'(net_i)$$

$$\frac{\partial^2 E}{\partial net_i^2} = f'(net_i)^2 - (t_i - O_i) f''(net_i)$$

3 试验结果及分析

作者采用了 40 个人的自由手写体数字共 650 个样本进行实验，其中 200 个作为训练集，450 个作为测试集. 训练结果如下：

3.1　网络的收敛情况

采用传统的 BP 学习算法, 各层的学习率均取 0.25, 网络的收敛情况如图 2 所示. 而采用伪牛顿法, 即以 (7)式调整权值, 各层的学习率仍取 0.25, μ 取 1. 网络的收敛情况如图 3 所示. 可以看出, 采用伪牛顿法要比传统 BP 算法快 4 ～ 5 倍.

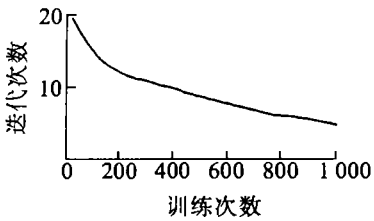

图2　BP算法网络收敛情况

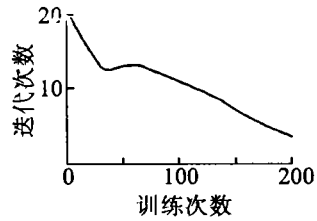

图3　伪牛顿法网络收敛情况

3.2　识别情况

对利用伪牛顿法训练好的网络进行识别测试的结果为: 训练集 200 个样本, 正确识别率 94%, 误识率 0%, 拒识率 6%; 测试集 450 个样本, 正确识别率 86.7%, 误识率0.67%, 拒识率 12.6%.

4　识别系统的优化

人脑具有鲁棒性, 一个成熟的神经网络也应具有鲁棒性. 通过对已训练好的网络进行鲁棒性测试可以看出, 某些神经元受损并不影响网络的识别. 由此设想可否最佳程度地对神经进行损伤, 而不影响网络的记忆.

神经损伤的普遍作法是消去那些对训练误差影响最小的参数, 即小特征法. 以下简单推导分析哪些参数具有小特征值:

对代价函数微分得出代价函数随权值参数矢量 W 波动而改变的式子为

$$\Delta E = g_i \Delta W_i + \frac{1}{2}\sum_i h_{ii}\Delta W_i^2 + \frac{1}{2}\sum_{i\neq j} h_{ij}\Delta W_i \Delta W_j + O(\|\Delta W\|^2) \qquad (11)$$

式中 ΔW_i是 ΔW的分量, g_i 是代价函数 E 梯度 G 的分量, h_{ii}是代价函数 Hessian 矩阵的元素.

由此式可以找到一组参数, 消去它们之后使误差 E 的增加最小. 但这个问题在一般情况下是不可解的, 因为矩阵 H 非常大, 很难计算. 因此, 本文中采取以下一些简单的近似来简化计算.

4.1　正交近似

假设由消去的若干参数产生的误差变化是分别消去每个参数所产生的误差变化之和.

4.2　极值近似

假设参数消去是在训练收敛后进行的, 因此参数矢量在误差 E 的最小 (或局部最小)点,

方程式右端的第一项可以忽略不计.

4.3 二次项近似

假设代价函数接近二次式, 因此方程式的最后一项可以忽略不计. 这样方程式 (11)可简化成

$$\Delta E = \frac{1}{2}\sum_i h_{ii}\Delta W_i^2 \tag{12}$$

因而可以将 h_{ii}作为第 i 个样本的特征值, 若该特征值很小, 说明消去第 i 个参数后不会使误差过多地增加.

用这种方法消去参数的步骤如下:

1) 训练网络直到得到满意的结果;

2) 对每一参数计算二阶导数;

3) 计算每一参数的特征值;

4) 根据特征值挑选参数, 消去那些特征值小的参数;

5) 重复第一步.

实验结果如图 4 所示.

对消去的参数 (设置为 0)加以分析, 发现有的特征检测器整块的权值为零, 也就是说它们所检测的特征对识别不起大的作用. 因此可以通过加强网络的部分映射来减少权值, 由此也证明了本文所选择网络结构的正确性.

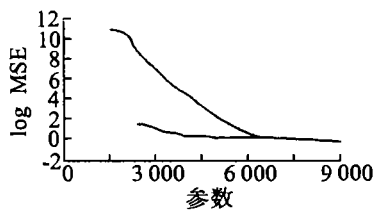

图4 网络参数与均方误差的关系

5 结论

1. 伪牛顿法能有效地对网络进行学习和训练, 学习速度比BP算法快 4 ～ 5 倍, 并取得了较好的识别结果.

2. 利用最佳神经损伤实验可以合理地设计网络结构, 使网络以尽可能少的连接, 获得尽可能高的识别率.

参 考 文 献

1 Hertz John. Introduction to the Thorey of Neural Computation. Addision-Wesley Publishing Commpany, 1991. 115 ～141

2 Stefanos Kollias and Diinitris Anastassiou. Adaptive Training of Multilayer Neural Networks Using A Least Squares Estimation Technique. In: International Conference on Neural Networks. 1988, (Ⅰ) ～383

3 Eduard Sackinger, Member, IEEE, Bernhard E. Bosser, Member, IEEE, Jane Bromley, etc. Application of the ANNA Neural Network Chip to High-Speed Character Recognition. IEEE Transactions on Neural Networks, 1992, 3(3)

4 Himmelblau D M. Introducing efficient Second Order Effects into Back propagation Learning. In: IEEE. International Joint Conference on Neural Network, 1990, (Ⅰ): 631

5 Sue Becker, Yann le Cu. Impproving The Concergence of Back-Propagation Learning With Second Order Methods. In: Proceedings of the 1988 Conectionist Models Summer School. Pittsburg, 1988, 229 ～ 237

6 Yann Le Cun, John S. Denker and Sara A. Solla. Optimal Barain Damage. In: Neural Information Processing systems. Vol. 2, 598 ～604

The Application of the Feedforword Neural Networks with Multi-layer in the Recognition of Handwritten Numerals

Hu Jian　Wang Qingbao　Tu Chengyu

(Department of Electronic Engineering, Beijing Polytechnic University, 100022)

Abstract A feedforword network of 5-layer is presented, which accomplishes numeral features extraction and recognition. This network employed weight share and partial reflection to reduce the total weight numbers. In trainning process, an extension BP algorithm with the second derivative of cost function is used. The algorithm has nice convergence properties, which performs four or five times faster than the conventional BP algorithm. In addition, the network weight numbers are greatly reduced by using the second derivative information for neural damage or robustness test.

Keywords neural networks, pattern recognition, character recognition, robustness