基于 RBF 神经网络的汉字粗分类方法

杨 芳,李红睿,田学东
YANG Fang,LI Hong-rui,TIAN Xue-dong

河北大学 数学与计算机学院,河北 保定 071002
College of Mathematics and Computer Science,Hebei University,Baoding,Hebei 071002,China
E-mail:yangfang@hbu.cn

YANG Fang,LI Hong-rui,TIAN Xue-dong.Coarse classification scheme for Chinese character based on RBF neural network.Computer Engineering and Applications,2009,45(6):170-172.

Abstract: Coarse classification is the key to improve recognition speed.An improved coarse classification scheme based on RBF (Radial Basis Function) neural network for Chinese character is presented in this paper.Four-side code feature and gross meshed feature are respectly applied as coarse feature to compare in this experiment.The GB2312-80 first-level Chinese character samples including printed and handwritten form are objects in this experiment.Experiment results show that proposed method has excellent performance on coarse classification in contrast to Euclidean distance as classifier used in conventional method.

Key words: Radial Basis Function(RBF) neural network;coarse classification;four-side code;gross meshed feature

摘 要:粗分类是提高汉字识别速度的主要手段。将 RBF(Radial Basis Function neural network)神经网络用于汉字粗分类,采用汉字四边码和粗网格作为汉字粗分类的特征以进行比较。分别对 GB2312-80 一级字库印刷体及手写体进行实验,实验结果表明将 RBF 神经网络用于汉字粗分类比通常使用的欧式距离作为分类器有较好的性能。

关键词:RBF 神经网络;粗分类;四边码;粗网格
DOI:10.3778/j.issn.1002-8331.2009.06.048 **文章编号:**1002-8331(2009)06-0170-03 **文献标识码:**A **中图分类号:**391.4

随着 OCR 技术的进步,研究对象集合越来越大,仅汉字字符就达到 20 000 多个。汉字具有字符集大、相似字多、结构变化复杂、字体多样等特点,尤其是手写汉字字形变化丰富,这给汉字字符识别造成了很大的困难。为保证识别率,如果采用将待识字符特征与字典特征逐一比对的方法会使识别速度随字数的增加明显的下降。为了提高整体识别速度,通常都采用多级识别方案。即先以汉字的一些简单特征对汉字进行粗分类,将具有相似特征的汉字分到同一候选类中,经过一级或多级粗分类再进行细识别。通常粗分类的结果,是细识别的输入,所以粗分类的优劣直接关系到系统性能的好坏。

目前系统中通常对汉字提取特征有结构特征[1]、粗网格特征[2]、笔画密度[3]、粗外围特征[4-5]以及四边码[6]特征,通常选用较简单的粗网格特征、笔画密度特征以及四边码作为粗分类特征,并以相对简单的欧式距离作为分类器。这种粗分类方法计算简单,所提取的粗分类特征与细识别特征互补,很好地提高了整体识别速度。但是采用相对简单的欧式距离作为分类器,分类结果与候选类中心距离过于关系密切, 当字形变化较大时,无法达到比较理想的分类效果,从而使得汉字识别率下降。由于 RBF 神经网络的模拟生物特性以及它能较好地拟合数据类,将 RBF 神经网络引入粗分类,并采用粗网格及四边码作为粗分类的特征。同时对一级字库的印刷、手写体汉字分别做实验进行比较。

1 RBF 神经网络

RBF 神经网络它的产生具有很强的生物学背景。在人的大脑皮层区域中,局部调节及交叠的感受野(Receptive Rield)是人脑反应的特点。基于感受野这一特性,Moody 和 DarKen 提出了一种神经网络结构[7]。RBF 神经网络是前馈神经网络中的一类特殊的三层神经网络。网络从输入层到隐含层的变换是非线性的,而从隐含层到输出层的变换则是线性的。其隐含单元的特性函数采用非线性的径向基函数,以对输入层的激励产生局部化响应, 即仅当输入落在输入空间一指定的小范围内时,隐层单元才会做有意义的非零响应[8]。RBF 神经网络拓扑图如图 1 所示。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60772073);河北省教育厅科学技术研究重点项目(the Key Scientific and Technical Research Project of Ministry of Hebei Education of China under Grant No.ZH2007104);河北省科学技术研究与发展计划项目(the Science and Technologies Research Plan of Higher Education of Hebei Province,China under Grant No.06213598);河北大学自然科学基金资助项目(the Natural Science Fundation of Hebei University Under Grant No.2006Q01)。

作者简介:杨芳,女,讲师,主要从事智能图文信息处理研究;李红睿,女,硕士研究生,主要从事智能图文信息处理研究;田学东,男,教授,主要从事智能图文信息处理研究。

收稿日期:2008-01-16 **修回日期:**2008-04-14

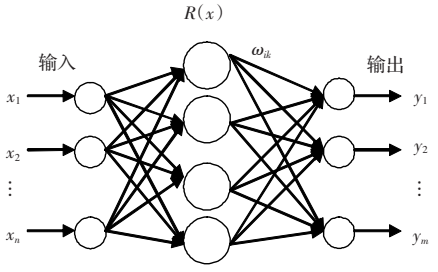


图 1　RBF 神经网络拓扑图

RBF 神经网络的输入为 $X=\{x_1,x_2,\cdots,x_m\}$，输出 $Y=\{y_1,y_2,\cdots,y_n\}$，对应的数学表达式为：

$$y_k(x)=\sum_{i=1}^{l} w_{lk}R(x),\quad k=1,2,\cdots,n \tag{1}$$

其中 l 为隐层单元数。w_{lk} 为输出单元与隐层单元之间的连接权值。R 为非线性径向对称基函数，通常采用高斯函数：

$$R_j(x)=\exp(-\|x-c_j\|^2/2\sigma_j^2)\quad j=1,2,\cdots,l \tag{2}$$

其中 c_j 是第 j 个基函数的中心，与 X 具有相同维数的向量；σ_j 是第 j 个感知的变量(可以自由选择的参数)，它决定了该基函数围绕中心点的宽度。

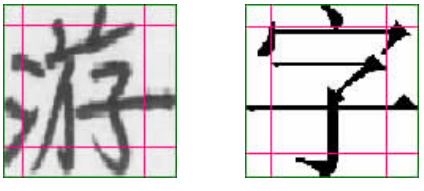
2　汉字粗分类特征提取

特征提取是整个字符分类的核心部分，它直接关系到系统识别性能优劣。由于汉字四边框内含有丰富的结构信息，并且四边框部分的笔画比较稀少，受断笔干扰较小，因此本文选取四边码特征以及对位移和旋转不敏感的粗网格特征[9]作为粗分类特征。

2.1　四边码特征提取

将单个汉字图像大小归一化为 64×64 像素。实验直接对汉字灰度图像进行特征提取以避免二值化对汉字信息的损失。为了使划分后每个边框内都有一定的笔画数目，选取每边图像灰度值大于 200 的像素点数达到 100 处作为边框位置(如图 2 所示)，以此分割出文字上、下、左、右的 4 个部分。这样避免了固定坐标对特征提取的不稳定性，同时提取的特征更能反映文字的信息。

经统计得出汉字笔画宽度一般大于 4 个像素，而位于笔画中间的白点噪声小于 3 个像素。根据图像每一部分内含有的笔画数目进行编码。因此每个汉字的粗分类特征是一个四维向量由上、下、左、右四边笔画数组成，即 $\boldsymbol{X}=(x_1,x_2,x_3,x_4)$，其中 x_1 表示边框上边穿过的笔画数，x_2 表示边框下边穿过的笔画数，同样 x_3,x_4 分别代表边框左右两边穿过的笔画数。例如图 2 中，手写体“游”字的编码是 $\boldsymbol{X}=(2,5,3,1)$，印刷体“字”的编码是 $\boldsymbol{X}=(1,1,2,3)$。



(a)手写体　　(b)印刷体

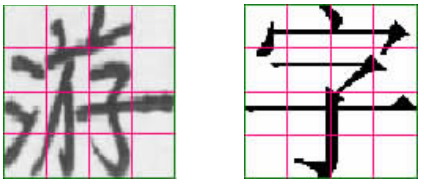
图 2　四边码边框划分示意图

2.2　粗网格特征

粗网格特征是对大小归一化后的汉字分成 $N\times N$ 个网格，统计每个网格中位于汉字笔画上像素数量，计算每个网格像素数量在文字像素总数所占的百分比 $x(k,l)$，以此来作为每个汉字的粗分类特征。其中每个网格各自反映字符的某一部分特征。本文中 $N=4$，这样每个汉字有 16 维特征，即 $\boldsymbol{X}=(x_1,x_2,\cdots,x_{16})$ 其中 $x_i(i=1,\cdots,16)$ 表示划分的 16 个格的像素数目与汉字总像素数百分比 $x(k,l)$。其中 $x(k,l)$ 数学表达式如下：

$$x(k,l)=\frac{\sum_{i=0}^{n/N}\sum_{j=0}^{n/N} f(k\times\frac{n}{N}+i,l\times\frac{n}{N}+j)}{\sum_{i=0}^{n/N}\sum_{j=0}^{n/N} f(i,j)},\quad k,l=0,1,\cdots,n \tag{3}$$

其中 n 为归一化后的汉字的宽度，同样为避免二值化对汉字特征的损失，粗网格特征也是直接对汉字灰度图像提取，$f(i,j)$ 为汉字图像灰度值。



(a)手写体　　(b)印刷体

图 3　粗网格特征提取示意图

3　实验过程及分析

3.1　RBF 神经网络参数选择

实验中分类器采用Ⅱ型 RBF 神经网络(聚类辅助型网络)[7]。采用文献[10]的方法：中心 c_j 采用 k-均值聚类算法中心，即 $c_j=\frac{1}{|\varphi_j|}\sum_{X^\mu\in\varphi_j} X^\mu$，宽度 σ_j 定义为 $\sigma_j=\alpha\frac{1}{\varphi_j}\sum_{X^\mu\in\varphi_j}\|X^\mu-c_j\|$。选定 σ_j 和 c_j 后，对 RBF 神经网络权值 w_{lk} 进行训练。

3.2　实验过程

用四边码作为特征进行实验时，RBF 神经网络的输入层有 4 个节点，输入为汉字的四边码特征 $\boldsymbol{X}=(x_1,x_2,x_3,x_4)$；用粗网格特征进行实验时，RBF 神经网络的输入层有 16 个节点，输入向量为汉字的粗网格特征 $\boldsymbol{X}=(x_1,x_2,\cdots,x_{16})$。将 GB2312-80 一级字库所有汉字的特征通过 K 均值聚类算法进行聚类以确定粗分类类标作为神经网络的输出，并选取不同的隐层节点个数来训练神经网络。使用 K 均值聚类算法将一级字库的所有汉字特征进行聚类，实验表明对于四边码特征和粗网格特征将其分为 3 类、5 类和 10 类数据都分布比较均匀(如图 4 所示为聚 5 类时直方图)，因此将汉字粗分类为 3 类、5 类和 10 类。分别对

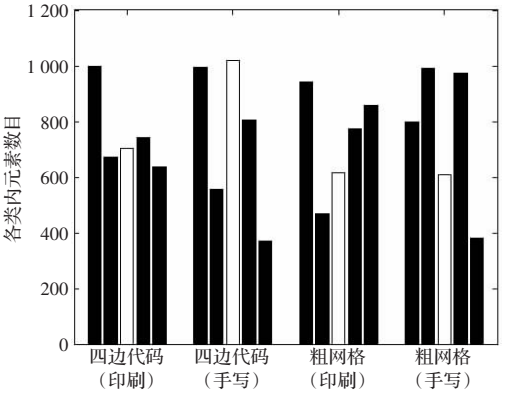


图 4　聚 5 类直方图

一级字库印刷体及手写体汉字做实验,并且选取 7 套印刷体和手写体对神经网络进行训练,3 套用于测试神经网络的分类正确率。

3.3 实验结果分析

在实验过程中分别对采用 RBF 神经网络和欧式距离作为分类器进行测试。粗分 3 类、5 类以及 10 类实验结果如表 1~表 3 所示。在 RBF 神经网络中由于选取的隐层节点数目不同,正确率会有很大的差别,所以表中选取分类效果最好地进行对比。

表 1 粗分 3 类实验结果

特征类型	粗分正确率/(%)	
	RBF 神经网络	欧式距离
四边码(印刷)	95.56	90.61
四边码(手写)	85.11	79.32
粗网格(印刷)	97.22	90.39
粗网格(手写)	96.26	83.55

表 2 粗分 5 类实验结果

特征类型	粗分正确率/(%)	
	RBF 神经网络	欧式距离
四边码(印刷)	97.06	90.85
四边码(手写)	89.75	85.47
粗网格(印刷)	93.42	88.17
粗网格(手写)	92.57	79.48

表 3 粗分 10 类实验结果

特征类型	粗分正确率/(%)	
	RBF 神经网络	欧式距离
四边码(印刷)	91.41	89.63
四边码(手写)	89.60	83.17
粗网格(印刷)	95.07	82.56
粗网格(手写)	90.74	81.69

由 3 个表可以看出,对于印刷体汉字来说,RBF 神经网络与欧式距离的正确分类率相比略高一些,但对于手写体汉字 RBF 神经网络则显示出它的优越性。同时对比表中数据可以发现四边码特征对于印刷体汉字有较好的特性,但对手写体则稍差,这说明了四边码对字形变化的适应性较差。但对于粗网格特征手写体与印刷体正确分类率都比较高,说明粗网格特征对位移及旋转不敏感,可直接用于手写体汉字粗分类。

4 结论

由于 RBF 神经网络能较好地拟合数据类,将它用于汉字粗分类。在实验中分别对汉字图像(包括印刷体和手写体)提取四边码和粗网格特征进行比较,将一级汉字分别粗分为 3 类、5 类和 10 类,并与传统的欧式距离进行对比。实验结果表明将 RBF 神经网络用于汉字粗分类得到了较高的粗分正确率,尤其对于手写体汉字有很好的效果。

将 RBF 神经网络用于汉字粗分类取得较好的实验结果,但实验还需进一步改进。本文对汉字进行归一化时采用较简单的线性归一化方法,没有考虑手写汉字的重心位置和笔划密度,因此如果采用基于重心的归一化方法,再对汉字进行特征提取,粗分类正确率一定会有所提高。同时本文对汉字提取的四边码和粗网格粗分类特征本质上都属于统计特征,但都得到了较高的分类率,如果在识别阶段再结合汉字的结构特征或全局特征,就可较好的用于手写体汉字识别。

参考文献:

[1] 崔怀林,赵树芗.手写体汉字识别粗分类方法的研究[J].电子科技大学学报,1996,25(3):311-315.
[2] 吴成东,刘文涵,傅小菲,等.基于粗网格神经网络的车牌字符识别方法[J].沈阳建筑大学学报:自然科学版,2007,23(4):693-697.
[3] 周鸣芳,汪庆宝.手写体汉字识别的一种粗分类方法[J].北京工业大学学报,1985,11(4):33-41.
[4] 人金昌,赵荣椿,张炜.一种快速有效的印刷体文字识别算法[J].中国图象图形学报,2001,6(10):1011-1015.
[5] 刘传憬.一个实用的多字体多字号印刷汉字 OCR 系统[J].计算机应用研究,1995,4:57-59.
[6] 朱永娇.汉字特征提取的量化研究[J].科学技术与工程,2007,7(10):2402-2405.
[7] 王旭东,邵惠鹤.RBF 神经网络理论及其在控制中的应用[J].信息与控制,1997,26(4):272-284.
[8] 居琰,汪同庆,刘建胜,等.基于集成 RBF 神经网络的小类别手写体汉字识别系统[J].计算机工程与应用,2002,38(23):100-102.
[9] 尹晓峰,阎昌德,苏东庄.汉字基本集特征提取与分布研究[J].中文信息学报,1986,1:63-71.
[10] Schwenker F,Kestler H A,Palm G.Three learning phases for radial-basis-function networks[J].Neural Networks,2001,14:439-458.

(上接 167 页)

5 结束语

描述了一种粒度自动选择的半结构网页信息抽取方法。根据页面的 HTML tag 序列构造后缀树,利用后缀树搜索序列中极大重复和串联模式作为候选模式进行筛选,应用特征参数对候选模式进行过滤,最后选取最佳的模式,并从其实例中抽取网页数据记录的信息。实验结果表明,该方法是有效可行的。目前,致力于进一步改善抽取器的准确性,并应用于 Hidden Web 挖掘系统中。

参考文献:

[1] Bergman M K.The deep Web:surfacing hidden value[EB/OL].(2001).http://www.brightplanet.com/resources/details/deepweb.html.
[2] Chang C H,Kayed M,Girgis M R,et al.A survey of Web information extraction systems[J].IEEE Transactions on Knowledge and Data Engineering,2006.
[3] Chang C H,Lui S C.IEPAD:information extraction based on pattern discovery[C]//WWW2001,Hong Kong,2001.
[4] Ukkonen E.On-line construction of suffix trees[J].Algorithmica,1995,14:249-260.
[5] Stoye J,Gusfield D.Simple and flexible detection of contiguous repeats using a suffix tree[J].Theoretical Computer Science,2002,270(1/2):843-850.