计算机科学 2005Vol. 32No. 11

基于粗糙集和神经网络的分类器及其在LPR中的应用*)

张年琴[1,2] 苗夺谦[1,2] 李道国[1,2]

(同济大学计算机科学与工程系 上海 200092)[1] (国家高性能计算机工程中心同济分中心 上海 200092)[2]

摘 要 粗糙集和神经网络在模式识别中都可用于分类,但是都有局限性。虽然两者没有太多的共同点,将它们结合起来却能相互补充,起到比单个理论更好的分类效果。本文从理论上给出了用粗糙集约简算法减少BP网络中的一个神经元或连接时网络输出能产生的最大误差。接着将粗糙集和BP网络结合起来设计分类器,并通过车牌数字识别验证了该分类器的有效性。实验说明该分类器比单独用粗糙集和神经网络设计的分类器识别率高、识别时间短。

关键词 粗糙集,BP网络,分类器,车牌,字符识别

Classifier Based on Rough Set and Neural Network and its' Application in LPR

ZHANG Nian-Qin[1,2] MIAO Duo-Qian[1,2] LI Dao-Guo[1,2]

(Department of Computer Science and Engineering, Tongji University, Shanghai 200092)[1]

(Tongji Branch, National Engineering & Technology Center of High Performance Computer, Shanghai 200092)[2]

Abstract In pattern regconition, both rough set theory and neural network can be used to classify patterns. The two theory don't have much common point, but they can be complementary when we combine them in designing a classifier and the recognition effect can be better. This article gives the max output error when the rough set is used to reduce a nerve cell or a connection in Back-Propogation(BP) neural network. Then, rough set theory and BP network are integrated into one classifier. we validate this classifier's validity by using it in the character recognition of licence plate. The experiment shows that this classifier is better than classifiers using rough set or BP neural network in recognition rate and recognition speed.

Keywords Rough set, BP neural network, Classifier, Licence plate, Character recognition

1 引言

粗糙集理论是一种新的处理含糊和不确定性问题的数学工具。它作为一种软计算方法,与模糊方法、遗传算法、神经网络等一样,是有发展潜力的智能信息处理方法。它在数据库中的应用迅速发展起来。

神经网络以其优越的并行处理能力、容错能力和泛化能力,在模式分类和识别中得到广泛的应用。人工神经网络的基本方式是尽量建立一个系统,它和生物神经系统的工作方式类似。神经网络中连接的性质和连接之间数据的交换取决于具体的应用。

粗糙集理论和神经网络没有太多的共同点,但是将两者结合起来却能相互补充,起到比单个理论更好的分类效果。神经网络特有的学习能力可以解决模式识别中的分类问题,但是它不能确定哪些知识是冗余的,哪些是有用的,所有在输入的特征相量很大的情况下,网络结构复杂,训练和识别的过程都很长。粗糙集理论对特征可以进行离散化和约简,因此可以增强神经网络的泛化能力,提高网络的执行效率。使粗糙集和神经网络区别于其他机器学习方法的是它们在噪声条件下仍然能得到好的结果。

汽车牌照图像的自动识别系统是交通管理系统中的关键技术。牌照上的字符由于受天气变化、光照等的影响噪声很多,如何提高识别系统的准确性和实时性是关键。粗糙集神经网络分类系统结合了这两个理论的优点,经实验论证对车牌字符的识别效果不错。

本文第2部分介绍基于粗糙集的神经网络分类系统;第3部分是基于粗糙集的神经网络分类系统在车牌字符识别中的应用;第4部分是实验结果;最后是结论。

2 基于粗糙集和神经网络的分类器

2.1 粗糙集在预处理中的应用

粗糙集理论是Pawlak[1]提出的一种处理模糊和不确定性的数学工具,已经成功用于数据挖掘、模式识别、过程控制等领域[2]。根据Pawlak的定义,若属性集 $P\in C$ 是给定决策表的条件属性集 C 的约简,则 P 应满足两个条件:① $POS_P(D)=POS_C(D)$;② 对任意的 $a\in P$, $POS_{P-\{a\}}(D)=POS_C(D)$。现有的属性约简算法有:一般属性约简算法、基于差别矩阵的约简算法、基于信息熵的约简算法[3]等等。属性的输入数据可能是连续的或者泛化能力不强,这时需要先对其进行离散化处理。相应的离散化和约简算法我们实验室都已经实现。

2.2 基于粗糙集的神经网络理论

神经元是人工神经网络的基本处理单元,它是一个多输入多输出的非线性元件。如果神经元 i 连接到神经元 j,则神经元 j 的输入计算公式如下:

$$input_j=\sum w_{ij}\times output_i \tag{1}$$

其中 w_{ij} 是神经元 i 和神经元 j 之间的连接的权值。

训练过程中权值的调整公式如下:

$$w_{ij}^{new}=w_{ij}^{old}+\alpha(t)\cdot g(input_j) \tag{2}$$

g 是任意的作用函数,包括阈值型函数、线性型函数、S型函数(Sigmoid)、辐射基函数等。$\alpha(t)$ 是一个学习因子,在训练过程中开始值很大,随着时间函数逐渐减小。现在最流行的学习算法是反向传播算法(BP算法)。据大略估计,90%的神经网络的应用都是基于BP算法的。在这个算法中,根据下

*)国家自然科学基金项目(60175016,60475019)。张年琴 硕士研究生,研究方向为模式识别、数据挖掘、粗糙集理论。苗夺谦 教授,博士生导师,研究方向为人工智能、模式识别、数据挖掘、粗糙集理论、主曲线。

面的公式对权值进行调整。

$$w_{ij}^{new}=w_{ij}^{old}+\alpha\cdot err_j\cdot f'(input_j) \tag{3}$$

其中 f' 是由 s 型函数的导数，α 是学习系数，它在整个学习过程中是个常数，err_j 是神经元 j 的一个误差。因为 s 型函数的特性，$f'(x)=f(x)\cdot(1-f(x))$ 的计算是非常容易的。

从神经网络中去掉那些冗余的神经元和它们之间的连接可以降低网络的复杂度。粗糙集可以用来确定一个网络结构对解决一个问题是否是完备的，网络中是否存在一些冗余的神经元。如果一个新的网络结构，在一定程度上改善了原网络结构的性能，那么做结构上的调整就是成功的。定义神经网络的向量空间 N，在这个向量空间上的一个操作是将一个神经网络 $S_i\in N$ 映射到另一个神经网络 $S_j\in N$。在这里，通过减少一个或多个神经元，粗糙集就可以被看成是调整神经网络的一个操作。

对于三层 BP 网络，即输入层、隐含层和输出层，连接存在于输入层神经元和隐含层神经元、隐含层神经元和输出层神经元中。我们定义输入层和隐含层神经元连接的权值为，w_i^h，$i=1,2,\cdots,n$；$h=1,2,\cdots,H$，而隐含层和输出层神经元之间连接的权值为 w_h，$h=1,2,\cdots,H$，$o=1,2,\cdots,O$，其中 H 是隐含层神经元的个数，n 是输入层神经元的个数，O 是输出层神经元的个数。定义网络的输出为[4]：

$$Y_o=f(\sum_{h=1}^{H}w_h f(\sum_{i=1}^{n}x_i w_i^h)) \tag{4}$$

其中 f 是一个 S 型函数。因为我们的目标是尽可能多地去除不需要的网络连接，所以得到除去一个连接时对网络输出的影响是非常重要的。

将 Y_o 看成是单个变量（i^{th} 输入神经元到 h^{th} 隐含神经元的连接）的函数，则 Y_o 对网络权值的导数为：

$$\frac{\partial Y_o}{\partial w_i^h}=Y_o\times(1-Y_o)\times w_h\times x_i\times(1-f(\sum_{i=1}^{n}x_i w_i^h))f(\sum_{i=1}^{n}x_i w_i^h) \tag{5}$$

同理：

$$\frac{\partial Y_o}{\partial w_h}=Y_o\times(1-Y_o)\times f(\sum_{i=1}^{n}x_i w_i^h) \tag{6}$$

由式(5)和拉格朗日中值定理得：

$$Y_o(w)=Y_o(w_i^h)+\frac{\partial Y_o(w_i^h+\delta(w-w_i^h))}{\partial w_i^h}\times(w-w_i^h) \tag{7}$$

其中 $0<\delta<1$。将 w 取 0，我们得到：

$$|Y_o(0)-Y_o(w_i^h)|\leqslant|w_i^h\times\frac{\partial Y_o(w_i^h)}{\partial w_i^h}| \tag{8}$$

假设 S 型函数属于区间[−0.5,0.5]，则：

$$\left|\frac{\partial Y_i(w)}{\partial w_i^h}\right|\leqslant|w_h x_i|/16 \tag{9}$$

从式(8)和式(9)我们可以得到：

$$|Y_o(0)-Y_o(w_i^h)|\leqslant|w_i^h w_h x_i|/16 \tag{10}$$

这个等式给出了当省略权值 w_i^h 时网络输出变化的上限值。如果 $x_i\in[0,1]$ 则等式(10)变成：

$$|Y_o(0)-Y_o(w_i^h)|\leqslant|w_i^h w_h|/16 \tag{11}$$

同样地，如将 Y_o 看成是单个变量 $v(h^{th}$ 隐含神经元到 o^{th} 输出神经元的连接）的函数，我们可以得到：

$$\left|\frac{\partial Y_i(v)}{\partial w_h}\right|\leqslant 1/8 \tag{12}$$

$$|Y_i(0)-Y_i(w_h)|\leqslant|w_h|1/8 \tag{13}$$

等式(12)和等式(13)给出了当用约简算法去掉网络中的神经元和连接时输出能产生的最大误差。

以上我们论证了将粗糙集理论用于神经网络的构造可以使网络结构简单，并给出了用粗糙集理论中的约简算法减少一个神经元和连接时对结果造成的影响。如果 w_h 很小的话，那么误差也是很小的。当然网络的最简单结构就是用约简算法得出的约简中元素的个数作为输入层神经元的个数。下面我们就用这种最简单的结构设计分类器。

2.3 基于粗糙集和神经网络的分类器的构造算法

现在我们开始基于粗糙集的神经网络的构造，它的算法如下[5]：

(1)提取训练集和测试集的特征向量，输入决策表。

(2)选择决策表中用于学习的那部分记录形成训练表 table1，用粗糙集约简算法计算其所有可能的约简。

(3)选择几个区别比较大的约简结果，分别对 table1 进行更新，形成约简表 SimTable1，SimTable2……等等。

(4)构造三层 BP 神经网络，输入层神经元的个数待定。

(5)将输入层神经元的个数定为 SimTable1 的字段数，并在此约简表的基础上训练这个网络。

(6)对选择的约简表 SimTable2…等重复 4 操作，得到多个训练好的神经网络。

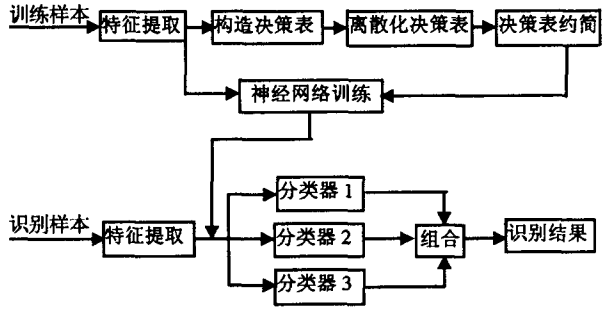
(7)选择决策表中用于测试的那部分记录，输入神经网络，测试各个网络的性能。

(8)提取代识样本的特征向量，将相应的部分输入各个网络进行识别，利用投票法或其他的分类器组合方法给出识别结果。

3 基于粗糙集和神经网络的分类器在车牌字符识别中的应用

3.1 车牌字符识别流程

从原始汽车牌照图片中切割出牌照后，要先对牌照进行校正，使牌照称为长方形的规则牌照。对校正后的牌照进行归一化、灰度化和二值化，从二值化后的牌照中切割出字符。对切割出字符的识别流程如下：



3.2 字符特征决策表的建立和属性约简

粗糙集使用决策表作为知识表达系统。决策表 $S=(U,R,V,f)$，$R=C\cup D$ 是属性集合，子集 C 和 D 分别称为条件属性和决策属性，U 是论域，$V=\bigcup_{r\in R}V_r$ 是属性值的集合，V_r 表示属性 $r\in R$ 的值域，信息函数

$$f:U\times R\to V。$$

由于样本有限，我们仅对车牌数字字符设计了粗糙集神经网络分类器。

作为例子，我们选择了数字字符图像的粗网格特征。切割出的字符经过归一化和二值化后用于提取。粗网格特征属于统计特征，又称局部灰度特征。粗网格特征的提取步骤如下(以黑底白字)为例：把字符图像分成 $N\times N$ 个网格，统计每个网格中白色像素的个数作为这个小格的特征值，将所有网格特征值组合在一起形成一个 $N\times N$ 维的特征向量。

根据粗网格构成的决策表有 $N\times N$ 个条件属性和一个决策属性，即字符。若原字幅图像归一化成 64×64 的标准，

将字符图像分成 4×4 个网格，则每个网格白色像素最多为 256 个。则构成的决策表条件属性 $C=\{C1,C2,\cdots,C16\}$，且属性值值域为 $V_c=\{0,1,\cdots,256\}$，决策属性 $D=\{d\}$ 且属性值域 $V_d=\{0,1,2\cdots,9\}$。

建立决策表后，需先对其进行离散化处理。在这里我们采用了简单的等距离离散化方法，经过实验，距离间隔为 20 的时候效果最佳。则离散化后属性值的值域为 $V_d=\{0,1\cdots,16\}$。

我们使用基于差别矩阵的约简算法，可以得出全部的约简。从约简结果选择区别比较大的几个约简，分别用来构造分类器。

3.3 BP 神经网络的构造

在进行 BP 网络的设计时，一般应从网络的层数、每层中神经元的个数、激活函数、初始权值以及学习效率几个方面进行考虑[6]。经过实验，选择如下：

(1)网络的层数：我们选择了三层网络，即输入层、隐含层和输出层。

(2)各层神经元个数：输入层为约简结果中属性的个数，即约简特征向量元素格数。隐含层神经元个数为 30 时，网络收敛效果比较好。输出层神经元个数为 10，即对应于 10 个数字。

(3)初始权值的选取：取在(−1,1)之间的随机数。

(4)学习速率：学习速率的范围是 0.01～0.8，我们选择的是 0.1。

3.4 字符识别算法

使用训练后的网络识别待识样本：

(1)取一个待识样本，由某个约简结果属性组成向量。$X=[x_1,x_2,\cdots,x_n]$

(2)将向量输入相应的 BP 网络，得到某个输出结果，如 $D1=[0.1,0.8,0.1,\cdots,0.0]$。

(3)同样的可以在其他的 BP 网络上，得到另两个输出结果，如 $D2=[0,0.7,0,0.1,0.2,\cdots,0]$，$D3=[0,0.9,\cdots,0.1]$。

(4)由神经网络组合的投票法的到识别结果，比如是字符“1”。

4 实验结果分析

实验的编程环境为 Visual C++ 6.0。实验的对象为车牌数字字符集(因为汉字和字母样本数量太少)。样本是拍摄的车牌图像经字符自动分割和归一化算法得到的[7]。训练样本 200 个(每个字符 20 个)，测试样本为 200 个(每个字符 20 个)，对测试样本进行 10 次试验，取平均值作为最终结果。

我们采用了三种方法对样本进行测试。

(1)基于粗糙集规则匹配：由粗糙集属性约简和值约简得到匹配规则，输入样本特征进行匹配。

(2)基于 BP 网络：属性没有经过粗糙集离散化和属性约简，直接送入 BP 网络进行训练。

(3)基于粗糙集和 BP 网络：即本文中论述的方法。

三种方法实验结果比较如下表所示。

方法	匹配时间(s)	平均正确识别样本个数	平均拒识样本个数	平均误识样本个数	识别率(%)
1	5.2	145	30	25	72.5
2	4.3	162	15	23	81
3	2.8	183	8	9	91.5

结论 用粗糙集理论对属性约简后，神经网络的神经元数减少，连接也随之减少。本文从理论上得出从网络上移除一个连接对网络输出影响的极限值。实验论证，经过粗糙集离散化和约简处理后，神经网络的结构简单，学习效率高。基于多个约简结果的分类器的组合比单个分类器效果好。总之，基于粗糙集和神经网络结合的分类系统识别率高，识别需要的时间短。

参 考 文 献

1 王国胤. Rough 集理论于知识获取[M]. 西安：西安交通大学出版社，2001

2 边肇琪，张学工，等. 模式识别(第二版)[M]. 北京：清华大学出版社，1999

3 苗夺谦，胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展，1999，6

4 Setiono R. Extracting rules from Pruned neural networkd for breast cancer diagnosis. Artificial Intelligence in Medicine，1996，8(1)：37～51

5 Swiniarski R W，Skowron A. Rough set methods in feature selection and recognition[J]. Pattern recognition letters[J]，2003，24：833～849

6 Pandya A S，Macy R B，著，徐勇，等译. 神经网络模式识别及其实现[M]. 电子工业出版社，1999

7 刘智勇，刘迎建. 车牌识别(LPR)中的图像提取及分割[J]. 中文信息学报，2003，14(4)

(上接第 171 页)

表 2 不完备信息系统 2

U	a	b	c	d	e
x_1	1	1	*	0	1
x_2	1	0	1	0	0
x_3	1	1	0	1	0
x_4	*	0	*	1	0
x_5	1	0	*	1	2
x_6	0	1	0	1	2

由于 $M_c(4,5)=\lambda^2$ 是 $M_a(4,5)$、$M_b(4,5)$、$M_c(4,5)$、$M_d(4,5)$中唯一最小的，因此属性 c 是核属性。因为 $M_c \cap M_d = M_c$，所以属性 d 是可约的。由于 $M_c \cap M_a < M_c$，可知属性 a 不可约，同样，由于 $(M_c \cap M_a) \cap M_b < (M_c \cap M_a)$ 可知属性 b 不可约。因此，最后得到的约简结果为 $\{a,b,c\}$。

结论 本文提出了一种针对不完备信息系统的等价类矩阵表示方法，同时给出了相关的理论研究结果以及属性约简方法。由于采用矩阵表示，使得等价类的描述简洁明确，给属性约简带来方便。更重要的是为进一步研究属性约简开辟新的思路。

参 考 文 献

1 Pawlak Z. Rough set approach to multi-attribute decision analysis. European Journal of Operational Research，1994，72：443～459

2 Grzymala-Busse J W，Hu M. A comparison of several approaches to missing attribute values in data mining. In：Proc. of the 2nd Int'l Conf. on Rough Sets and Current Trends in Computering. Berlin：Springer Verlag，2000. 378～385

3 Guan J W，Bell D Z，Guan Z. Matrix computation for information systems. Information Sciences，2001，131：129～156

4 曾黄麟. 粗集理论及其应用(修订版). 重庆：重庆大学出版社，1998