基于FNN与GA相融合的数据挖掘方法研究

王 刚,王本年
(铜陵学院 计算机系,安徽 铜陵 244000)

摘 要:模糊神经网络即具有输入信号是模糊量的神经网络,是模糊系统与神经网络相结合的产物,汇聚了二者的优点;遗传算法是一种自适应全局优化概率搜索算法。研究了基于模糊神经网络与遗传算法相融合的一种算法,在应用模糊神经网络进行数据挖掘前,应用遗传算法完成隶属函数的训练,以便更好地进行模糊神经网络学习;经过模糊神经网络学习后,提取相关规则,再次应用遗传算法,进行规则剪枝,提高数据挖掘效率。实验表明,与传统方法相比,该方法能够更快速、更加准确地进行数据挖掘,提取更精确的推理规则。

关键词:数据挖掘;模糊神经网络;遗传算法;隶属函数;规则剪枝
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)02-0119-03

Data Mining Study Based on Fuzzy Neural Networks and Genetic Algorithms

WANG Gang, WANG Ben-nian
(Department of Computer Science, Tongling College, Tongling 244000, China)

Abstract: With fuzzy input signals, fuzzy neural network is a combination of fuzzy system and neural network. Therefore, it asembles the advantages of them. Genetic algorithm can adapt to the overall situation and better probability seeking automatically. Based on the combination of fuzzy neural network and genetic algorithm, studies an algorithm that can train the membership function by applying genetic algorithm before data mining by applying fuzzy neural network. Rules can be obtained after the study of fuzzy neural network while pruned after the application of genetic algorithm; thus the efficiency of data mining is increased. Experiments have shown that, compared with conventional algorithm, this one can mine data more quickly and more accurately, and get the reasoning rules more precisely.

Key words: data mining; fuzzy neural network; genetic algorithm; membership function; rule pruning

0 引 言

数据挖掘(Data Mining)是数据库中知识发现的核心。数据挖掘就是从大量的数据中挖掘出隐含的、未知的、用户可能感兴趣的和对决策有潜在价值的知识和规则[1,2]。

模糊神经网络(FNN)是模糊系统与神经网络相结合的产物,它汇聚了神经网络与模糊系统的优点,集联想、识别、自适应及模糊信息处理于一体;遗传算法(GA)是一种基于自然选择和自然遗传的全局优化算法,采用从自然选择机理中抽象出来的选择、交叉、变异等三种基本的遗传算子对参数编码进行操作,可以实现全局最优搜索。

收稿日期:2007-05-25
基金项目:安徽省自然科学研究项目(KJ2007B382ZC)
作者简介:王 刚(1975-),男,安徽桐城人,讲师,硕士,研究方向为数据挖掘、神经计算等;王本年,教授,博士,硕士生导师,研究方向为人工智能、机器学习、多Agent理论等。

1 模糊神经网络

模糊神经网络[2~4]将模糊逻辑和神经网络结合成一个系统,从而达到以神经网络及模糊逻辑各自的优点弥补对方不足的目的。模糊神经网络在金融时间序列分析、流数据等方面得到了深入研究,展现了其广阔的应用前景。

1.1 模糊逻辑理论概述

模糊集合是模糊概念的一种描述,对于模糊集合而言,一个元素可以既属于该集合又不属于该集合,界限模糊,这种不确定性可以用一个隶属函数来刻画。

模糊集合:设 X 是论域,X 上的一个实值函数,用 $A(\chi)$ 来表示,即 $A(\chi):\chi \to [0,1]$,对于 $\chi \in X$,$A(\chi)$ 称为 χ 对 A 的隶属度,$A(\chi)$ 称为隶属函数,它是 χ 属于 A 的程度的数量指标。通过隶属度函数将相关属性模糊化,作为神经网络的输入,以 $\chi_1,\chi_2,\cdots,\chi_n$ 表示,选择 m 个属性,作为神经网络的输出,仍然使用隶属函数将它们转换为模糊期望输出,以 $y_1,y_2,\cdots,$

y_n 表示，选择合适的样本数据训练神经网络，最终根据输入/输出的结果建立模糊规则，形式如下：

if (χ_1 IS v_1) and (χ_2 IS v_2) and … and (χ_n IS v_n)
then (y_1 IS ω_1) and (y_2 IS ω_2) and … and (y_n IS ω_n)

其中 v_i，ω_i 为模糊属性大小，如描述为大、中、小，高、中、低等。

1.2 模糊神经网络的构造

模糊神经网络一般分为四层结构[2-4]（注：也有学者将其描述为五层结构），如图 1 所示。

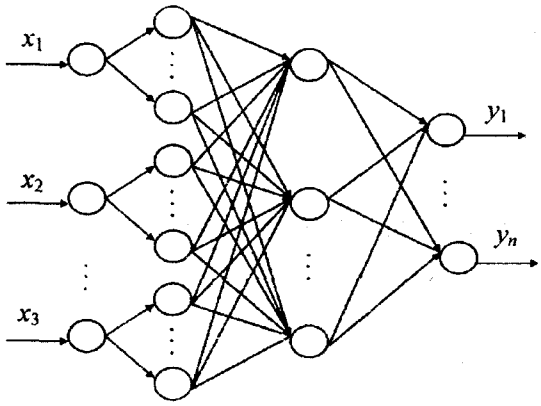

图 1 模糊神经网络系统结构

第 1 层为输入层，设有 n 个属性，所以有 n 个节点；第 2 层为计算隶属函数层，在此层通过模糊隶属度函数计算出每个属性的隶属度值；第 3 层为基于推理规则学习的隐含节点，节点数量根据输入节点和模糊规则样本确定，如果网络的学习记忆能力有限，则可以采用多个隐含层来存储知识；第 4 层为模糊输出层，类似于第 2 层的逆过程，输出语言变量，当然训练网络时误差也由这一层反向传播，不过训练数据必须经过像第 2 层一样的模糊化后才能传回网络，节点数量 = 语言变量数 $\lambda \times m$，m 为要预测的属性数目，注意这里的每 λ 个节点才能表示一个输出属性。模糊神经网络每一层都有明显的物理意义，并且模糊神经网络是一个全局逼近器。在设计模糊神经网络结构时，可以根据问题的复杂程度以及精度要求，并结合先验知识来构造相应的模糊神经网络模型。这样，网络的学习速度就会大大加快，并在一定程度上回避了梯度优化算法带来的局部极值问题[4,5]。

1.3 模糊隶属度函数的确定

在模糊推理中，选择合适的模糊隶属度函数非常重要，但目前还无法用理论证明选择效果的好坏，主要依赖经验选取。常用的有梯形、高斯形以及 S 形等函数。笔者采用高斯函数[6]，它具有连续可微分的非线性特征，且神经网络推理中常将其作为传递函数，其基本表示为：$\mu_{ij}(x_i) = \exp\{-[\frac{x_i - a_i}{b_i}]^2\}$，其中 x_i 表示输入变量，a_i，b_i 分别为隶属函数的中心（通常取 0.2 ~ 0.8）与宽度，在大、中、小三种不同情形下，各隶属函数示意图如图 2 所示。

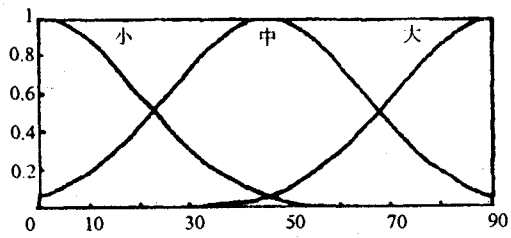

图 2 高斯隶属函数

2 基于 FNN 与 GA 的数据挖掘算法研究

2.1 遗传算法

遗传算法[7]（Genetic Algorithm，GA）是一种借鉴自然界自然选择和进化机制发展起来的高度并行、随机、自适应搜索算法，由 Holland 首先提出。其突出特点在于它包含了与生物遗传及进化很相似的步骤：选择、复制、交叉、重组和变异等。

对于随后进行的网络学习过程可以看作是一个极小化的优化过程，目标函数为神经网络的能量函数 E，优化变量为权值和阈值 W。经典 BP 算法可以使权值、阈值收敛到某个值，但是可能产生的是一个局部最小值。GA 的搜索是同时从每个串开始搜索，大大减少了陷入局部解的可能性，同时 GA 对每一个串进行独立的处理，具有高度的适应能力，利用遗传算法全局搜索能力可以很好地解决上面描述的问题。基本遗传算法主要由染色体编码方法、个体适应度评价、遗传算子及基本遗传算法的运行参数的确定等方面构成。与其它寻优算法相比，遗传算法具有计算简单、功能强大、鲁棒性好的特点。

2.2 模糊神经网络的学习

通过应用遗传算法，修正冗余的隶属函数和网络的节点数，优化模糊推理规则。通过利用遗传算法优化具有全局性的参数、网络结构；利用 BP 算法调节和优化具有局部性的参数。通过粗略学习与精细调节，二者相互融合，提高模糊神经网络推理控制系统的自学习性能和鲁棒性[2,5]。

首先根据实际情况，进行适应度函数的选择，笔者采用系统输出值与系统设定值的绝对值之和的相反数作为适应度函数，即：

$J = -\sum_{i=1}^{m} | y_{out} - y_{set} |$，其中：$y_{out}$ 为系统输出值，y_{set} 为设定值，m 为采样次数；接着进行相关参数的优化，先将各个参数采用二进制编码表示，假定参数的变

量区间为$[\theta_{min},\theta_{max}]$，根据实际基本论域的范围，选择合适的量化因子 k_e，k_c，通过量化因子将它转换到模糊集论域中。为提高遗传算法的收敛性，引入自适应的交叉概率 P_c 和变异概率P_m，P_c、P_m 可以通过下面公式求解：

$$P_c = \frac{k_e}{f_{max} - f} \qquad P_m = \frac{k_c}{f_{max} - f}$$

其中，f_{max}，f 分别为群体中最大合适度与平均合适度，体现了群体的收敛程度，k_e，k_c 为常数；用遗传算法对模糊神经网络进行训练学习。

2.3　基于 FNN 与 GA 的数据挖掘算法思想

综合上述描述，现将基于模糊神经网络与遗传算法的数据挖掘算法概括如下：

STEP1：在第 1 层输入原始属性值，输出值为 x_1，$x_2,\cdots,x_n$；

STEP2：应用训练的隶属函数分别将第 1 层输入的属性值模糊化后输入第 2 层节点，每个节点从左至右分别代表语言变量大、中、小，选择隶属度最大的节点输出为 1，其余为 0；

STEP3：随机产生 n 条二进制字符串，每个字符表示整个网络的一组参数；

STEP4：根据相关公式计算相应的参数的适应度值，对权值和阈初始值进行编码；设置种群规模，构成遗传空间；

STEP5：while(新群体中总数$\leqslant n$)，继续进行下列具体操作：

根据实际情况，按一定概率从群体中选择两个串，设为 S_i，S_j；根据GA算法思想，以概率 P_c 对S_i，S_j 进行交换，得到新串 S_{i1}，S_{j1}；以概率 P_m 对新串的各位产生突变。重复上述过程。直到新的群体产生。

STEP6：继续上述 STEP2，直到群体中的个体性能满足要求为止，群体中适合度最好的字符串所表示的参数就是所要的网络参数。

STEP7：用优化的网络权值训练第 3、第 4 层的权重，笔者采用 S 函数作为激活函数，误差由输出属性的训练集通过隶属函数模糊后反向传回第 4 层，直到所有样本数据学习完毕。

STEP8：当训练神经网络达到期望的准确度时，从被训练的神经网络中提取知识，然后将这些知识用确定或者是模糊的 If - Then 规则表述出来。

STEP9：应用遗传算法进行规则剪枝，这一技术可以参考文献[4,5,8]，计算网络的实际输出与期望输出误差，如果误差在许可范围内，输出规则并发现隐藏的知识，否则，修正相关权值，转到 STEP7。

图 3 是该算法的主要构造过程示意图。

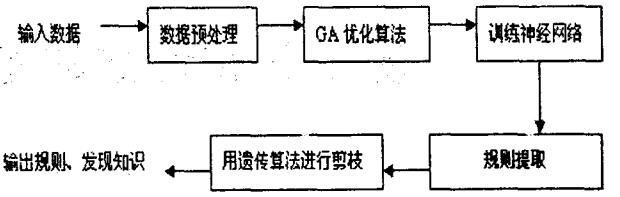

图 3　算法运算过程

3　仿真实验

实验数据来自一园林管理处某稀有植物生长状况调查记录数据库，对原始数据预处理以后，进行了土壤含水量指标、温度指标与该植物生长状况指标之间的关系规则实验，以验证本算法的有效性(见表 1)。

表 1　生长状况相关数据

年份	含水量指标(χ_1)	温度指标(χ_2)	生长状况指标(y_1)
1978	0.68	-5.21	283
1979	-2.13	5.09	647
…	…	…	…
2003	-0.48	6.89	796

在构造模糊神经网络时，以土壤含水量指标、温度指标作为输入，以生长状况指标作为输出，对相关数据进行预处理，确定模糊子集与隶属函数。下面给出在应用遗传算法对隶属函数进行训练前后各隶属函数的曲线图，如图 4、图 5 所示。

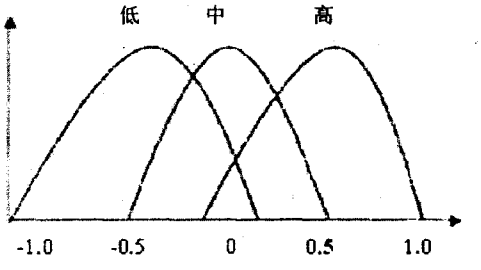

图 4　训练前隶属函数图

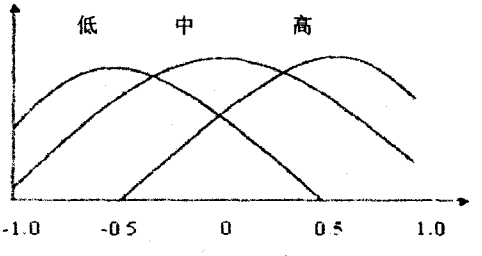

图 5　训练后隶属函数图

经过多次训练，直到网络基本收敛，训练结束。训练结果发现，各相关隶属函数均产生修正，修正后的隶属函数趋于平缓，其分布与给定的样本数据分布趋于一致。最后应用遗传算法进行剪枝，获得更为准确的规则。应用整个算法，获取相关规则结论如表 2 所示。

(下转第 125 页)

表3　IEEE-754浮点数标准

	符号	指数	尾数
单精度	1[31]	8[30-23]	23[22-00]
双精度	1[63]	11[62-52]	52[51-00]

例如,单精度十进制浮点数 -20.59375 采用IEEE-754浮点格式的表示形式为 11000001 10100100 11000000 00000000 ,这种形式可以将其看作一个无符号整数的二进制存储形式,因此所有的负数都大于正数;然后采用基数为 2^8 的CRSORT算法进行递增排序,则将待排序数据调整成为两个部分:前部为负数非递增序列,后部为正数递增序列;最后只需在线性时间内将正数和负数的整体位置互换(负数在调整过程中,其内部数据要倒置),便可以得到正确的有序序列。

(5) 如果对字符串排序,可以利用字符串在计算机存储器的存储特点,将每个字符作为每趟计数排序的操作的对象。由于字符是以ASCII形式存放的,取值范围在闭区间[0,255],便取 2^8 作为基数,对其进行CRSORT排序即可。

3　结束语

文中提出的基于计数的基数排序算法的时间复杂度为 $O(N)$,排序的速度优于快速排序,是一种简单高效的算法。另外,如果待排序的为负整数、字符串以及浮点型等数据,只要对这些数据转换为一种适当的描述形式进行,便可采用这种排序算法,文中对此也作了一定的讨论。

参考文献:

[1] 严蔚敏,吴伟民. 数据结构(C语言)[M]. 北京:清华大学出版社, 1997.

[2] 殷人昆,陶永雷,谢若阳,等. 数据结构(用面向对象方法与C++描述)[M]. 北京:清华大学出版社, 1999.

[3] 唐开山. 二次堆排序算法和提高排序效率的途径[J]. 计算机工程与应用,1998,34(5):45-48.

[4] Cormen T H. Introduction to Algorithms[M]. [s.l.]: The MIT Press, 1990.

[5] Sedgewick R. Algorithms in C++[M]. [s.l.]: Addison-Wesley, 1998.

[6] Baase S. Computer Algorithm: Introduction to Design and Analyse[M]. [s.l.]: Addison-Wesley, 2000.

[7] 王向阳,杨红颖. 分段快速排序法的改进[J]. 小型微型计算机系统,2001,22(11):1382-1385.

[8] 唐向阳. 分段快速排序法[J]. 软件学报,1993,4(2):53-57.

[9] Knuth D E. The Art of Computer Programming 3/Sorting and Searching[M]. 管纪文译. 北京:国防工业出版社,1984.

(上接第121页)

表2　规则提取

含水量指标(χ_1)	高	高	高	中	中	中	低	低	低
温度指标(χ_2)	高	中	低	高	中	低	高	中	低
生长状况指标(y_1)	低	中	低	中	高	中	低	中	低

通过实际数据检查,以上规则是合理的,基本与实际情况一致。实验表明,应用本算法能够获得较为满意的结果。

4　结束语

基于模糊神经网络与遗传算法的数据挖掘方法,充分考虑了数据挖掘过程中数据量大、模糊性强、提取高效准确的推理规则等特点。应用本算法进行数据挖掘,预测精度高,它比单纯的遗传算法或神经网络或模糊系统的效果要好;适合模糊信息的处理,适合对大容量数据的挖掘等。下一步主要就该算法中遗传算法部分的适应值函数、遗传算子、自适应等方面进行深入研究,以提高算法的执行效率。

参考文献:

[1] 潘　笑,万　敏. 基于模糊神经网络的数据挖掘方法研究[J]. 微电子学与计算机,2005,22(12):48-50.

[2] 张　勇,黄金才,张维明,等. 一种遗传模糊神经网络数据挖掘算法[J]. 模糊系统与数学,2006,20(5):131-135.

[3] 钟　珞,饶文碧,邹承明. 人工神经网络及其融合应用技术[M]. 北京:科学出版社,2007.

[4] 周志坚,毛宗源. 一种基于遗传算法的模糊神经网络最优控制[J]. 控制理论与应用,2000,17(5):784-788.

[5] Ishigami H, Fukuda T, Shibata T, et al. Structure optimization of Fuzzy Neural Network by Genetic Algorithm[J]. Fuzzy Sets and Systems, 1995, 71(3):257-264.

[6] Takagi T, Sugeno M. Fuzzy indentification of systems and its applications to modeling and control[J]. IEEE Transactions on Systems, Man and Cybernetics, 1985, 15(1):116-132.

[7] 王小平,曹立明. 遗传算法——理论、应用与软件实现[M]. 西安:西安交通大学出版社,2002.

[8] 李　旸,高正光,李启炎. 基于神经网络与遗传算法的数据挖掘体系结构[J]. 计算机工程,2004,30(6):155-156.