

一种基于蚁群算法与粗糙集的混合 BP 神经网络

洪月华

(广西经济管理干部学院 计算机系, 广西 南宁 530007)

摘 要: 为了解决 BP 神经网络对高维冗余样本分类时收敛速度慢、易陷入局部极小值问题, 提出基于蚁群算法与粗糙集的混合 BP 神经网络分类模型. 该混合 BP 神经网络用粗糙集对样本进行约简和降维, 输入层神经元个数得到减少, 降低了训练神经网络的计算复杂度, 用蚁群算法解决了选取神经网络权值和阈值的随机性, 避免了因其而导致的易陷入局部极小值的不足. 对 UCI 数据库中数据集的测试结果说明, 提出的混合 BP 神经网络对高维冗余复杂样本进行分类是可行的, 性能远远比传统 BP 神经网络和蚁群神经网络优越.

关键词: 蚁群算法 ACA; 粗糙集; BP 神经网络

中图分类号: TP183

文献标识码: A

文章编号: 1000-7180(2014)04-0156-04

A Mixed BP Neural Network Based on Ant Colony Algorithm and Rough Set

HONG Yue-hua

(School of Computer, Guangxi Economic Management Cadre College, Nanning 530007, China)

Abstract: In order to solve the problem that when BP neural network classify high dimensional redundant sample the convergence speed is slow and is easy to fall in local minimum problem, a mixed BP neural network model based on ant colony algorithm and rough set is proposed. In the mixed BP neural network, rough set is used to reductive and reduce the dimension of sample, so the input layer neuron number is reduced, which reduce the computational complexity of the training of the neural network, and ant colony algorithm has solved the random of selection neural network weights and threshold, which has avoided the problem of being easy to fall in local minimum problem. The results gotten by testing data sets in the UCI database, the mixed BP neural network is feasible to classify high dimensional redundant sample, and the performance is more better than the traditional BP neural network and ant colony neural network.

Key words: ant colony algorithm; rough set; BP neural network

1 引言

近年来, 研究人员研究得到的数据分类算法有很多, 如: 决策树, 贝叶斯, 支持向量机, 神经网络, 关联规则, K-近邻法等分类方法. 当问题的机理和原因复杂, 不能用规则或公式通过数学模型描述时, 用神经网络分类方法处理往往表现出极大的自适应性、自组织性和灵活性. 当前应用最广泛的神经网络

模型是 BP 神经网络. 但是, 当样本是高维冗余数据时, BP 神经网络极易陷于局部最优解、收敛速度慢、诊断精度低和泛化能力差等问题, 其在分类领域的应用受到了制约. 为了解决这些问题, 许多研究者进行了坚持不懈的研究. 一些学者用蚁群算法优化 BP 神经网络的权值和阈值^[1-5], 获得了不少研究成果. 王宏图^[6-8]等用粗糙集对高维冗余样本进行约简和降维, 减少了 BP 神经网络输入层神经元个数. 依据

收稿日期: 2013-07-07; 修回日期: 2013-09-01

基金项目: 广西自然科学基金青年项目(2012GXNSFBA053178); 广西混杂计算与集成电路设计分析重点实验室开放基金课题(2012HCIC05)

这些研究基础,本文提出一种基于蚁群算法与粗糙集的混合BP神经网络来处理需分类样本是大量高维冗余数据的情况。

2 蚁群算法优化BP网络的原理

蚁群算法ACA是源于自然界现象的仿生算法,是意大利研究者Dorigo M提出,模拟蚂蚁的觅食行为特性,使用其内在的搜索原理,解决组合优化问题。该算法善于全局优化智能搜索,而且有正反馈、鲁棒性强、分布式计算、启发式学习、易和其他算法结合起来使用等特点。

BP神经网络算法实质上是局部搜索算法,学习训练过程采用梯度下降法来修正权值和阈值,使得误差函数E最小。梯度下降法对初始权值和阈值极其敏感,而神经网络初始权值和阈值的选取是随机的,若选择不恰当,且处理数据是庞大的高维冗余数据时,会导致训练过程出现震荡、收敛速度缓慢和陷入局部极值而得不到全局最优解。

神经网络学习训练过程就是一个最优化问题,即寻找一组最优的阈值和权值组合,使得输出层的实际输出和期望输出之间的误差E最小。蚁群算法就是找到最优权值和阈值组合的较好方法,因此本文把其与BP神经网络结合,形成蚁群神经网络,即使用蚁群算法的全局智能寻优能力对神经网络的初始阈值和权值进行优化,提高网络收敛速度,有效地避免了BP神经网络陷入局部极值等缺陷。

3 蚁群神经网络模型的构建及其训练

蚁群算法优化BP神经网络,构建蚁群神经网络模型的基本思路是^[1,3]:假设BP神经网络的 p 个参数有待优化,包括了全部的权值(v_{jk}, ω_{ij})和阈值(α_j, β_k),记为 $q_1, q_2, \dots, q_p$ 。把其中任一参数 $q_i (i = 1, 2, \dots, p)$ 设置为可能取值区间内的 M 个随机非零值,形成集合 I_{q_i} 。假定蚁巢中有 S 只蚂蚁出去寻找食物源,每只蚂蚁随机地从集合 I_{q_i} 中选择一个元素,同时根据规则调整其信息素;当蚂蚁从全部集合中实现选择元素后,就抵达了食物源,然后蚂蚁按原路返回蚁巢。这一过程一次又一次地重复,当所有蚂蚁收敛至同一路径时, p 个网络参数的最优解就找到了,搜索结束。

蚁群算法对BP神经网络进行训练的步骤如下:

Setp1:初始化设置。设最大迭代数是 $\text{count}_{\max}$,集合 $I_{q_i} (i = 1, 2, \dots, p)$ 中元素 j 的初始信息素 $\pi_j(I_{q_i})(t) = c, (1 \leq j \leq M)$,蚁巢中的蚂蚁总数是

S ,训练误差终止条件 E_0 等参数。

Setp2:从蚁巢中释放出所有蚂蚁进行搜索,每只蚂蚁按式(1)的路径选择规则依序在各个集合中挑选出元素一个,直至所有蚂蚁抵达食物源。

路径选择规则:任何一只蚂蚁 $a (a = 1, 2, \dots, S)$ 从集合 I_{q_i} 中随机地选择元素 j 的概率是

$$\text{Prob}(\pi_j^a(I_{q_i})) = \pi_j(I_{q_i}) / \sum_{i=1}^p \pi_i(I_{q_i}) \quad (1)$$

Setp3:计算误差。当全部蚂蚁从每个集合里都挑选出一个元素后,用这些选择出的元素作为BP神经网络的最优参数(权值和阈值)算训练样本的输出误差。若 $E \leq E_{\text{BP}}$ 成立时,说明输出误差已经达到指定的误差要求,退出程序,否则转Setp4。

Setp4:更新信息素规则^[4-5]。上述操作经历的时间是 Δt 个时间单位(迭代次数),对每个集合 I_{q_i} 里的每个的信息素按下式更新。

$$\pi_j(I_{q_i})(t + \Delta t) = \zeta \pi_j(I_{q_i})(t) + \Delta \pi_j(I_{q_i}) \quad (2)$$

式中:

$$\begin{aligned} \Delta \pi_j(I_{q_i}) &= \sum_{a=1}^S \Delta t_j^a(I_{q_i}) \\ \Delta t_j^a(I_{q_i}) &= \begin{cases} B/e^a, & \text{蚂蚁 } a \text{ 在集合 } I_{q_i} \text{ 里选择元素 } j \\ 0, & \text{其他} \end{cases} \\ e^a &= \max_{p=1}^N |z'_p - z_p| \end{aligned}$$

$\zeta (0 < \zeta < 1)$ 是信息素持久性系数, $\Delta t_j^a(I_{q_i})$ 是本次循环中蚂蚁 a 在集合 I_{q_i} 里的元素 j 上留下的信息素。 B 是常量,代表蚂蚁完成一次循环后信息素释放的总数,主要作用是调整信息素的更新速度。 e^a 代表蚂蚁 a 在所有集合里选择出的元素当成BP神经网络的参数时各训练样本的输出误差的最大值。 N 是训练样本总数, z'_p, z_p 分别是神经网络的期望输出和实际输出。据此可见,越小的输出误差,增加的信息素就越多。

Setp5:反复做上述Setp2、Setp3和Setp4,直至迭代次数到达 $\text{count}_{\max}$ 或蚁群中的每一只蚂蚁都收敛到同一条路径,结束寻优,输出最优解。

4 粗糙集与蚁群神经网络集成

为了简化蚁群神经网络的输入,提高数据分类速度和精度,本文用粗糙集的属性约简功能对大量高维冗余数据进行分析处理,找出关键属性,删除噪声数据和无关信息。

4.1 粗糙集理论知识

定义1 设知识表达系统 $S = (U, A, V, F), U$

$= \{x_1, x_2, \dots, x_n\}$ 是论域, 代表对象的非空有限集;
 $A = C \cup D$ 是属性集合且 $C \cap D = \emptyset$, 子集 C 和 D 分别是条件属性集和决策属性集; 属性 $a(a \in A)$ 的值域用 V_a 表示, $\bigcup_{a \in A} V_a = V$; $f: U \times A \rightarrow V$ 表示论域 U 中各个对象的属性值, 是信息函数.

定义 2 在知识表达系统 S 分类过程中, 个体之间相差不大的关系就是不可区分关系. 设 $K \subseteq A$, 论域 U 上 K 的不可区分关系 $IND(K)$ 定义如下:

$$IND(K) = \{(x_i, x_j) \in U \times U \mid \forall a \in K, a(x_i) = a(x_j)\} \quad (3)$$

论域 U 由 $IND(K)$ 形成的划分用 $U/IND(K)$ 表示, 简记作 U/K . 不可区分关系也叫等价关系.

定义 3^[6] 对于知识表达系统 $S = (U, A, V, F)$, 设 $K \subseteq A$, $X \subseteq U$, K 是论域 U 上的不可区分关系, 分别定义 X 的 K 上、下近似集如下:

$$\begin{aligned} K^+ X &= \bigcup \{Y \in U/K \mid Y \cap X \neq \emptyset\} \\ K^- X &= \bigcup \{Y \in U/K \mid Y \subseteq X\} \end{aligned} \quad (4)$$

定义 4^[8] 知识表达系统 S 中, 设 $K \subseteq A$ 且 K 是论域 U 上的不可区分关系, 对于任意 $X \subseteq U$, 则属性 K 关于 X 的正域 $POS_K(X)$ 如下定义为

$$POS_K(X) = \bigcup \{X_i \mid X_i \in U/K, X_i \subseteq X\} \quad (5)$$

定义 5 $S = (U, A, V, F)$ 是一个知识表达系统, C 是条件属性集, D 是决策属性集, $c \in C$; 如果 $POS_C(D) = POS_{C-\{c\}}(D)$, 说明属性 c 在 C 中是不重要的, 否则为重要的. 所有重要的条件属性称为核.

定义 6 知识表达系统 $S = (U, A, V, F)$ 中, C 为条件属性集, 若有 $B \subseteq C$, 对于任意的 $B' \subset B$, 式子 $POS_C(D) = POS_{B'}(D)$ 都为假, 则说明条件属性子集 B 是条件属性集 C 相对于决策属性集 D 的一个约简.

属性约简就是保证分类正确的前提下, 删除冗余属性, 把重要属性保留下来. 大多数情况下约简是不唯一的, 所有约简的交集就构成核.

4.2 基于粗糙集的蚁群神经网络输入数据预处理

使用粗糙集对样本做属性约简处理, 删除冗余样本和冗余属性, 简化蚁群神经网络训练集输入的步骤如下:

Step1: 构成决策表

决策表由这样的表格形成: 表格里方样本数据, 其中列是属性 $a(a \in A)$, 行是论域中的研究对象 x_i , 表格中数据是各研究对象对应的各属性值 $a(x_i)$. 知识系统 S 就是一个决策表.

Step2: 离散化连续属性值

粗糙集理论只能处理和分析离散数据, 因此运用它处理数据之前, 必须把连续属性值进行离散化操作, 本文采用基于统计指标的离散化方法进行处理^[9].

Step3: 属性约简

决策表中的信息由冗余和不相容情况, 并不一定是最简的, 必须要对它进行属性约简, 使得输入蚁群神经网络的样本数据的维数得到降低, 本文选择基于区分矩阵的属性约简方法^[10].

5 仿真实验与结果分析

为了使本文算法(记为 RSACABP)分类结果说服力强, 用蚁群算法优化 BP 网络(记为 ACABP)和传统 BP 神经网络(记为 BP)作为对比算法. 使用 UCI 数据库中的 Statimage(属性 36 个, 样本 6 435 个, 类别 6 个)数据集作为测试数据. 仿真实验环境为: CPU 是 Inter(R) Core(TM)2 Duo 2.93 GHz, 内存 2 GB, 仿真软件用 Matlab7.0. ACA 参数设置: 蚂蚁总数 $S=50$ 只, 权值和阈值的取值范围是 $[-2, 2]$, 最大迭代次数 $\text{count}_{\max}=2\ 000$, 训练误差最小值 $E_0=10^{-4}$.

对 Statimage 数据集中 36 个属性中的连续属性用文献[6]的方法进行离散化和用文献[7]的方法做属性约简, 得到其最小条件属性集(有 17 个属性).

三种算法在 Statimage 测试数据集上分别进行 20 次运算, 获得的平均分类准确率、平均迭代次数和平均运行时间如表 1 所示, 训练次数与训练误差关系如图 1~图 3 所示.

表 1 三种算法分类结果比较表

性能比较	BP	ACABP	RSACABP
分类正确率	64.58%	75.13%	87.64%
平均迭代次数	52 700	1 131	547
平均运行时间/s	5.487 7	2.351 9	1.106 8

由表 1 的统计结果可知, 三个算法在高维冗余数据集 Statimage 中表现出较明显的差异, 本文算法相比其它两个算法, 运行时间最快, 迭代次数最少, 分类正确率最高.

从图 1~图 3 的三条神经网络训练误差曲线可看出, 传统 BP 神经网络算法训练 50 000 多次时, 还没有收敛到目标误差 E_0 , 在训练误差值为 4×10^{-4} 处陷入局部极值而跳不出来, 根本无法获得全局最

优解;ACABP和RSACABP的训练误差都能很好地逼近目标误差 E_0 ,同时后者经过粗糙集处理输入数据后,神经网络训练迭代次数比前者明显少很多。

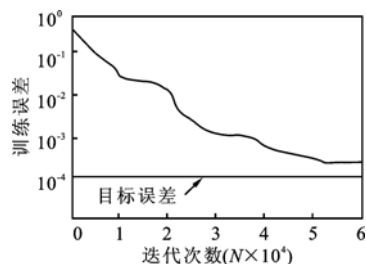


图1 BP神经网络的训练误差曲线

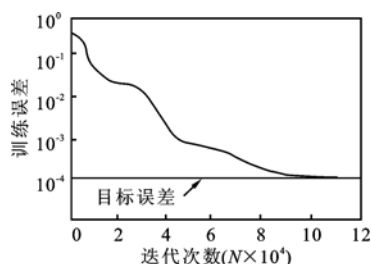


图2 ACABP的训练误差曲线

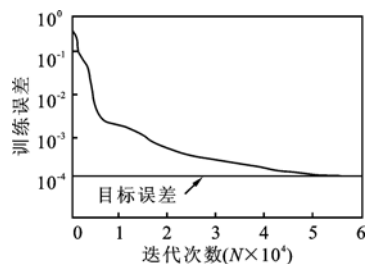


图3 RSACABP的训练误差曲线

6 结束语

本文把蚁群算法、粗糙集和神经网络三者结合起来,提出了一种基于蚁群算法与粗糙集的混合BP神经网络分类模型,先用粗糙集约简样本,BP神经网络的输入得到简化,再用蚁群算法对BP神经网络的权值和阈值进行优化,并以UCI数据库中Stattimage数据集为例,对此模型进行了验证。仿真实

验表明,该分类模型能够加快神经网络收敛速度,网络训练迭代次数明显改善,分类正确率更高,该分类模型可推广到其他高维冗余数据的分类识别中。

参考文献:

- [1] 孙旺,李彦明,杜文辽,等. 基于蚁群神经网络的泵车主泵轴承性能评估[J]. 上海交通大学学报,2012,46(4):596—600.
- [2] 吴文铁,宋曰聪,李敏. 蚁群优化神经网络的网络流量混沌预测[J]. 计算机工程与应用,2012,48(34):97—101.
- [3] 李祚泳,汪嘉杨,郭淳,等. 蚁群算法BP神经网络\基于蚁群算法的BP网络优化算法[J]. 计算机应用,2010,30(6):1513—1518.
- [4] 周美兰,敖雪. 基于蚁群神经网络的电控发动机故障诊断研究[J]. 黑龙江大学自然科学学报,2011,28(3):415—120.
- [5] 曾黄麟,孙勇,陈顺玲. 一种新数控系统温度误差补偿控制方法[J]. 重庆邮电大学学报:自然科学版,2010,22(5):656—659.
- [6] 廖瑛,张绍勇,尹大伟,等. 基于粗糙集遗传神经网络的柴油机故障诊断[J]. 控制工程,2009,16(6):709—712.
- [7] 王宏图,黄振华,范晓刚,等. 粗糙集—神经网络理论在矿井通风系统评价中的应用[J]. 重庆大学学报,2011,34(9):90—94.
- [8] 洪月华,徐霜,梁家荣. 一种基于粗糙集和遗传神经网络的数据分类器模型[J]. 广西科学,2013,20(2):128—131.
- [9] 陈浩. 统计指标离散化方法及应用[J]. 微电子学与计算机,2011,28(11):106—109.
- [10] 白俊清. 基于粗糙集理论的不完备信息系统下的知识获取方法研究[D]. 重庆:重庆大学,2006.

作者简介:

洪月华 女,(1973—),硕士,副教授. 研究方向为人工智能算法在数据挖掘中的应用.