一种基于决策树和遗传算法-BP 神经网络的组合预测模型

梁 栋,张凤琴,陈大武,李小青,王梦非
(空军工程大学信息与导航学院,西安 710077)

摘 要:设计了基于决策树和 GA-BPNN(遗传算法-BP 神经网络)组合预测模型,通过决策树分类贡献优先特征选择方法解决了 BP 神经网络的输入参数难选取的问题;利用改进遗传算法的全局择优能力,解决了 BP 神经网络由于随机选取初始权值导致易陷入局部极小值的缺陷。实验证明,组合预测模型能自学习专家经验,准确地对职业能力进行智能预测。

关键词:能力预测;BP 神经网络;决策树;遗传算法
中图分类号:TP391.4 **文献标志码**:A **文章编号**:2095-2783(2015)02-0169-06

A composite prediction model based on decision tree and GA-BPNN

Liang Dong, Zhang Fengqin, Chen Dawu, Li Xiaoqing, Wang Mengfei

(*Information and Navigation College*, *Air Force engineering University*, *Xi'an* 710077, *China*)

Abstract: A forecast model combining decision tree and GA-BPNN (genetic algorithm-BP neural network) model is designed. Decision tree is provided with feature selection method to solve the input parameter selection issue of BP neural network. The modified genetic algorithm with the global searching ability is applied to prevent BP neural network falling into local minimum caused by random selection of initial weight value. It is proven in simulation experiments that the composite pattern can learning expert specialist proficient by itself and predict vocational ability intelligently and accurately.

Key words: vocational ability prediction; BP neural network; decision tree; genetic algorithm

人工神经网络是人脑的抽象计算模型,是人工智能的一种,由于其具有高度非线性映射的学习和归纳能力,被广泛应用于各类预测模型中[1]。BP 神经网络是误差反向传播的人工神经网络,是神经网络模型诞生以来运用最为广泛的模型之一,大约 70%~80%的神经网络模型是 BP 神经网络或其变种[2]。BP 神经网络具有一些自身无法克服的缺陷[3],对其输入参数的选择没有有效的方法,在面对特征较多的输入数据集时,会导致学习不稳定;同时,BP 神经网络对初始权重非常敏感,若初始权重选择不当,会导致神经网络陷入到局部极小值中,无法得到全局最优解。职业能力的预测对预测结果的精确度有很高的要求,需要构建更为精确的预测模型。本文利用决策树分类模型在特征选择方面拥有的优势,利用信息增益算法有效选出并验证最优特征组合,弥补了 BP 神经网络无法确定输入参数的缺陷。利用改进遗传算法良好的全局择优性能,对 BP 神经网络的权值进行初始化,通过将三者结合起来,组成组合神经网络预测模型,实验证实,组合预测模型具有优良的性能,能很好完成能力预测的任务。

1 GA-BPNN 组合模型框架

组合神经网络模型主要由基于决策树的分类贡献优先特征选择方法、基于遗传算法的神经网络权值初始化方法、面向能力预测的 BP 神经网络 3 部分组成,核心是具有实现任何复杂非线性映射能力的 BP 神经网络,其具体的流程框架如图 1 所示。

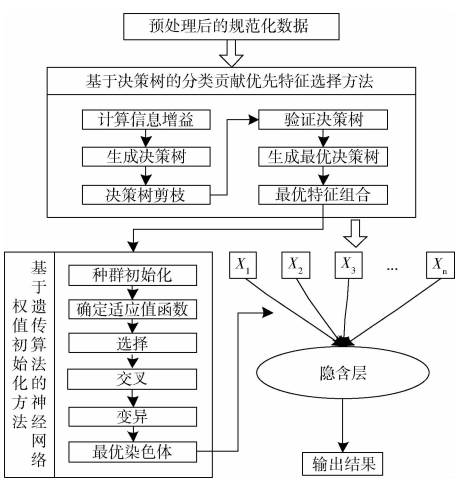

图 1 组合模型流程框架图

收稿日期: 2014-09-23
基金项目: 国家自然科学基金资助项目(61174162);陕西省自然科学基金资助项目(2014JM8345)
作者简介: 梁栋(1990—),男,硕士研究生,主要研究方向为神经网络、机器学习
通信联系人:张凤琴,副教授,主要研究领域为分布式数据库、信息安全、数据挖掘,332184321@qq.com

从图1中可以看出:基础数据经过预处理变为组合模型可直接使用的规范化数据;经过决策树分类贡献优先特征选择方法,可提取出最优特征组合作为BP神经网络的输入参数;通过改进的遗传算法可为最优特征组合选择最优的初始权值。最后经由BP神经网络得到输出结果。

2 基于决策树的分类贡献优先特征选择方法

训练数据中通常包含很多变量,其中有些变量与目标能力无关或影响很小。变量过多时,神经网络很难正常工作,而且会增加其过拟合的可能性[4],因此在将数据输入BP神经网络进行训练前,需要根据目标能力对变量进行精简,选择合适的特征变量,确定BP神经网络的输入参数。

分类贡献优先的特征选择方法主要是依据特征的信息增益,其基本思想是对于训练数据集进行计算并比较其中每个特征的信息增益,选出信息增益大的特征。

设训练数据集为 D,$|D|$ 表示数据集的样本容量,即数据集包含的所有样本的个数。设有 K 个类 C_k,$k=1,2,\cdots,K$,$|C_k|$ 为类 C_k 中包含的样本个数,$\sum_{k=1}^{K}|C_k|=|D|$。设特征 $\boldsymbol{A}$ 有 n 个不同取值 $\{a_1,a_2,\cdots,a_n\}$,由特征 $\boldsymbol{A}$ 不同的取值可将数据集 D 划分为 n 个子集 $D_1,D_2,\cdots,D_n$,$|D_i|$ 为 D_i 的样本个数,其中 $\sum_{i=1}^{n}|D_i|=|D|$。若子集 D_i 中属于类 C_k 的样本集合为 D_{ik},即 $D_{ik}=D_i\cap C_k$,$|D_{ik}|$ 为 D_{ik} 的样本个数,信息增益的计算方法如下。

输入:训练数据集 D 和所有特征集合 $\{A_1,A_2,\cdots,A_n\}$;

输出:每个特征 A_j 对训练数据集 D 的信息增益 $g(D,A_j)$。

步骤1 计算数据集 D 的经验熵 $H(D)$:

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|}。$$

步骤2 遍历特征集合,依次计算每个特征 $\boldsymbol{A}_j$ 对数据集 D 的条件熵 $H(D|\boldsymbol{A}_j)$:

$$H(D|\boldsymbol{A}_j)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)=$$

$$-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}。$$

步骤3 计算信息增益 $g(D,\boldsymbol{A}_j)$:

$$g(D,\boldsymbol{A}_j)=H(D)-H(D|\boldsymbol{A}_j)。$$

步骤4 重复步骤2和3,直至计算完特征集合中所有特征的信息增益。

将计算出的数据集的所有特征的信息增益从大到小排序,根据具体的应用要求按比例选出排序靠前的特征作为最优特征组合。

3 基于遗传算法的神经网络权值初始化方法

标准的遗传算法与BP神经网络形成优势互补,在能力分类预测方面得到了成功的应用[5],但是算法仍然存在一些缺陷,主要表现为收敛速度慢、搜索效率低、搜索前期容易陷入封闭竞争。为使遗传算法更好地满足神经网络权值初始化的要求,在组合模型中发挥更好的作用,需对遗传算法的部分关键环节进行改进。

3.1 改进的选择策略

在标准的遗传算法中,选择策略主要是根据个体的适应值高低决定其被选入下一代群体的概率,因此适应值低的个体入选的几率微乎其微,很容易导致出现超级个体,使整个算法停滞不前。为使算法尽快收敛,并保证算法以1的概率收敛到全局最优解,对选择方法的判别准则进行改进。

在选择下一代群体时,将当前群体中适应值最高的个体无条件地进入下一代,即可以保留当前最优解,防止算法反向波动。在父代和子代中随机选取个体 i 和 j,入选为下一代的概率分别为:

$$p(i)=\begin{cases}1, & f(i)\geqslant f(j),\\ \exp(\dfrac{f(i)-f(j)}{T}), & \text{其他};\end{cases}$$

$$p(j)=\begin{cases}0, & f(i)\geqslant f(j),\\ 1-\exp(\dfrac{f(i)-f(j)}{T}), & \text{其他}。\end{cases}$$

式中:$f(i)$ 与 $f(j)$ 分别为个体 i 和 j 的适应值;T 为控制参数。在每一代选择结束后,T 值按比例 α 下降,随着迭代次数的增加,T 值不断减小,当 $T\to 0$ 时,有

$$\exp(\frac{f(i)-f(j)}{T})\to 0;$$

此时个体 i 和 j 被选中的概率变为:

$$p(i)=\begin{cases}1, & f(i)\geqslant f(j);\\ 0, & \text{其他};\end{cases}$$

$$p(j)=\begin{cases}0, & f(i)\geqslant f(j),\\ 1, & \text{其他}。\end{cases}$$

改进的选择方法如下。

```
ReSelect()
1. 适应值最高个体直接入选下一代
2. 在父代和子代群体中随机选择 i 和 j
3. if f(i)>=f(j)
4. 个体 i 入选下一代
5. T=T× α
6. else
7. if rand(0,1)<=min(1,exp(f(i)-f(j))/T)
8. 个体 i 入选下一代
9. else
10. 个体 j 入选下一代
11. T=T× α
```

12. end

13. end。

3.2　改进学习模式

在标准的遗传算法中,优秀父代的优良基因得到继承,可使子代也具有较高的适应值。考虑父代本身可以通过自调整提高自身适应值,经过遗传会使子代的性能得到相应的提高,学习模式需引入学习算子,先给出如下定义[6]。

定义 1　设 x 是群体 $P(K)$ 的位串, $f(x)$ 是位串 x 的适应值函数, $\overline{f}$ 是群体适应值的平均,若 $f(x)>\overline{f}$,则称 x 是优良位串。

定义 2　设 x_{im} 是位串 x_i 的第 m 位,若 $x_{im}=x_{jm}\ \forall i,j\in\{1,2,\cdots,k\}$ 且 $i\neq j$,则称位 m 是 K 个优良位串 $x_1,x_2,\cdots,x_k$ 的优良位。

定义 3　设 $\hat{H}$ 是群体中任意一个模式, $\overline{f}(H)$ 是模式 H 的平均适应值,若 $\overline{f}(H)>\overline{f}(\hat{H})$,则称 H 是群体的优良模式。

设 x 是群体中适应值相对较低的位串, p_1 为学习率,学习算子 S 为:

$$S:(x,H)\rightarrow y。$$

$$y_i=\begin{cases}x_i, & \text{rand}(0,1)>p_1;\\ H_i, & \text{rand}(0,1)\leqslant p_1。\end{cases}$$

式中:rand(0,1)表示取(0,1)之间的一个随机数,位串 x 学习模式 H 变为 y ,位串 y 以概率 p_1 进入优良模式 H ,因此提高整个算法的性能,算法具体步骤如下。

改进的学习模式算法如下。

SeStudy()

1. 初始化参数 P_c、P_m、P_1、N;

2. 种群初始化 $P(K)$, $K=0$

3. 计算 $P(K)$中位串适应值

4. 令 solution= $P(K)$中最大适应值

5. while (不满足终止条件)

6. 搜索 $P(K)$中优良模式 H

7. $f(H$ 存在)

8. 用学习算子修正 $P(K)$中适应值低的位串

9. 对 $P(K)$进行选择、交叉、变异操作

10. $K=K+1$

11. 计算 $P(K)$中位串适应度

12. if(solution< $P(K)$中最大适应值)

13. solution= $P(K)$中最大适应值

14. return solution。

4　面向能力预测的 BP 神经网络构造方法

4.1　BP 神经网络结构设计

1)隐含层数的确定。

BP 神经网络可以包含一个或多个隐含层,不过理论上已经证明,对于大部分的预测应用,一个隐含层即可满足功能需求,因为可以通过增加隐含层包含的节点数实现任意非线性映射。

2)隐含层节点个数的确定。

隐含层所包含的节点个数对神经网络的性能有很大的影响,若隐含层含有较多节点则神经网络的性能会较好,但是训练时间会相应增加,收敛速度较慢;若隐含层节点数过少则会使神经网络的预测精确度下降。通常可如下确定隐含层的节点个数。

① $\sum_{i=0}^{n}C_M^i>k$,其中 k 为训练样本数, n 为输入层节点数, M 为隐含层节点数,当 $i>M$ 时,规定 $C_M^i=0$。

② $M=\sqrt{n+m}+a$, m 和 n 分别为输入层节点数和输出层节点数,其中 $a\in[0,10]$ 。

③ $M=\log_2 n$,其中 n 为输入层节点个数。

3)输出层神经元个数。

输出层神经元的个数由具体问题来确定。在职业能力预测模型中,对于目标属性为标称型数据时的分类预测,一个输出节点即可满足要求;对于目标属性为连续性数值的数据拟合问题,根据需要确定能力种类和个数,设置相应的输出节点即可。

4.2　改进的学习算法

权值的调整是神经网络学习算法的核心,调整权重的技术主要涉及动量和学习率[7]两个重要的参数。所谓动量(momentum)是指各个单元的权重向某个方向改变的趋势。每个权重记住自己是已经变大还是变小,而动量设法使它继续朝相同的方向发展。若动量值较大,则神经网络会对那些扭转权值方向的训练样本反应较慢;若动量值较小,则神经网络允许权值更自由的来回震荡。学习率(learning rate)控制着权重的变化速度。通常,提高学习率的最佳方法是开始时比较大,随着神经网络训练过程的进行而慢慢变小,这样当神经网络越来越接近最优解时,就可以通过微调来达到最优权值。标准 BP 神经网络所采用的最速下降法收敛速度较慢,可以通过设置动量和更改学习率提高收敛速度。

学习率的变化可以通过误差 e 增减来判断[8]。当误差以较小的方式趋于目标时,说明修正方向是正确的,这是学习率增加;当误差增加超过一定范围时,说明前一步修正进行的不正确,此时减小步长,并撤销前一步修正过程。学习率的增减公式为:

$$\eta(n+1)=\begin{cases}k_{\text{inc}}\eta(n), & e(n+1)<e(n);\\ k_{\text{dec}}\eta(n), & e(n+1)>e(n)。\end{cases}$$

在误差反向传播的基础上引入动量项来控制权值的改变量,可以比原有算法具有更好的性能,在说明算法之前先规定一些定义。

设训练样本数据集为 $D=\{(\boldsymbol{x}^{(p)},\boldsymbol{t}^{(p)})\}_{p=1}^{N}$,其中 $\boldsymbol{x}^{(p)}$ 表示输入向量, $\boldsymbol{t}^{(p)}$ 表示输出向量, N 为训练样本的个数,取误差平方和作为网络的误差函数:

$$E=\frac{1}{2}\sum_{p=1}^{N}\sum_{k=1}^{m}[\boldsymbol{t}_k^{(p)}-\boldsymbol{y}_k^{(p)}]^2。$$

式中: $\boldsymbol{y}_k^{(p)}=\boldsymbol{y}_k(\boldsymbol{x}^{(p)})$ 。

由链式求导法则得：

$$\frac{\partial E}{\partial v_{kj}} = \sum_{p=1}^{N}[\boldsymbol{y}_k^{(p)} - \boldsymbol{t}_k^{(p)}]\psi(\sum_{i=1}^{d}\boldsymbol{\omega}_{ji}x_i + \boldsymbol{\omega}_{j0});$$

$$\frac{\partial E}{\partial v_{k0}} = \sum_{p=1}^{N}[\boldsymbol{y}_k^{(p)} - \boldsymbol{t}_k^{(p)}];$$

$$\frac{\partial E}{\partial \boldsymbol{\omega}_{ji}} = \sum_{p=1}^{N}\sum_{k=1}^{m}[\boldsymbol{y}_k^{(p)} - \boldsymbol{t}_k^{(p)}]\psi(\sum_{i=1}^{d}\boldsymbol{\omega}_{ji}x_i + \boldsymbol{\omega}_{j0})[1-\psi(\sum_{i=1}^{d}\boldsymbol{\omega}_{ji}x_i + \boldsymbol{\omega}_{j0})]x_i;$$

$$\frac{\partial E}{\partial \boldsymbol{\omega}_{j0}} = \sum_{p=1}^{N}\sum_{k=1}^{m}[\boldsymbol{y}_k^{(p)} - \boldsymbol{t}_k^{(p)}]\psi(\sum_{i=1}^{d}\boldsymbol{\omega}_{ji}x_i + \boldsymbol{\omega}_{j0})[1-\psi(\sum_{i=1}^{d}\boldsymbol{\omega}_{ji}x_i + \boldsymbol{\omega}_{j0})]。$$

设 η 为学习率，α 为动量常数，则权值向量的变化为：

$$\Delta \boldsymbol{v}_{kj} = -\eta\frac{\partial E}{\partial v_{kj}} + \alpha\Delta \boldsymbol{v}_{kj}^{\text{old}};$$

$$\Delta \boldsymbol{v}_{k0} = -\eta\frac{\partial E}{\partial v_{k0}} + \alpha\Delta \boldsymbol{v}_{k0}^{\text{old}};$$

$$\Delta \boldsymbol{\omega}_{ji} = -\eta\frac{\partial E}{\partial \boldsymbol{\omega}_{ji}} + \alpha\Delta \boldsymbol{\omega}_{ji}^{\text{old}};$$

$$\Delta \boldsymbol{\omega}_{j0} = -\eta\frac{\partial E}{\partial \boldsymbol{\omega}_{j0}} + \alpha\Delta \boldsymbol{\omega}_{j0}^{\text{old}}。$$

增加动量和学习率的改进算法步骤如下。

步骤 1　对权重和偏差 v_{kj} 、ω_{ji} 、v_{k0} 和 ω_{j0} 进行初始化，其中 $k=1,2,\cdots,m$，$j=1,2,\cdots,n$，且 $i=1,2,\cdots,d$，设置学习率为 η，动量为 α。

步骤 2　对于输入向量 $\boldsymbol{x}^{(p)}$，计算神经网络输出 $\boldsymbol{y}_k^{(p)}$，根据预期的目标输出计算其误差，若误差大于设定的阈值，则进入步骤 3，否则转入步骤 4。

步骤 3　若迭代次数超过上限，进入步骤 4，否则更新权重和偏差 $\boldsymbol{v}_{kj}^{\text{new}} = \boldsymbol{v}_{kj}^{\text{old}} + \Delta\boldsymbol{v}_{kj}$ 、$\boldsymbol{v}_{k0}^{\text{new}} = \boldsymbol{v}_{k0}^{\text{old}} + \Delta\boldsymbol{v}_{k0}$、$\boldsymbol{\omega}_{ji}^{\text{new}} = \boldsymbol{\omega}_{ji}^{\text{old}} + \Delta\boldsymbol{\omega}_{ji}$ 、$\boldsymbol{\omega}_{j0}^{\text{new}} = \boldsymbol{\omega}_{j0}^{\text{old}} + \Delta\boldsymbol{\omega}_{j0}$，返回步骤 2。

步骤 4　训练停止。

5　实验分析

对企事业单位来说，提高工作生产效率，不仅要“唯才是用”，而且要“人尽其能”，随着信息化程度的不断提高，在各类业务系统数据库中积累了大量人员基础数据，如何将基础数据转化为能力数据，是人力资源管理领域关切的问题。传统能力数据的获取方法主要依赖于专家组进行评估[9]，在基础数据量大、能力数据实时性要求高的情况下，无法及时提供实时、准确、可靠的能力信息，给用人单位和企业带来影响。因此，需要构建能力预测模型，通过自学习专家经验，将基础数据自动转化为能力数据，为企事业单位提供辅助决策支持。

实验数据采用某企业员工的基本信息和业务表现数据。能力预测的目标是判断该员工能否完成公司指定的任务，鉴于用人单位的数据和信息保密性要求，本文无法展示职业数据中心中真实的基础数据，所采用的数据集在一定程度上进行了脱密处理，未使用真实的特征字段名，数据项为经过数据预处理后满足 BP 神经网络输入要求的规范化数据。数据集包含 10 000 个样本，每个样本包含 24 个属性，包括年龄、身高、体重等基本元素和与训练相关的能力参数，每个属性值的精确度保留到小数点后 10 位。所有数据可分为两类，目标属性为 1 表示完成了任务的正样本，目标属性为 0 表示未完成任务的负样本，正样本与负样本各 5 000个，各类样本中 80%的数据作为训练数据，另 20%作为验证数据，用于测试模型的分类精确度。实验数据存放在 Excel 表格中，如图 2 所示。

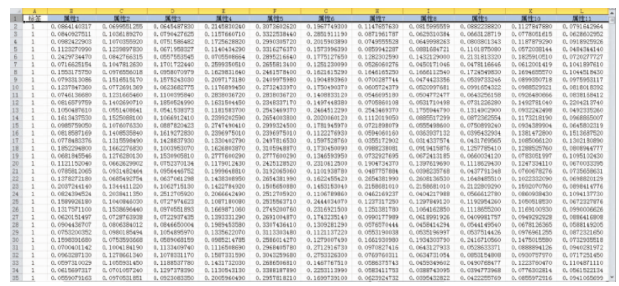

图 2　实验数据截图

计算所有字段的信息增益，并依次选取信息增益值最大的属性作为分类节点，使用 C4.5 算法构建决策树，对训练样本和测试样本的分类结果分别如表 1 和表 2 所示。

表 1　训练样本分类预测表

训练样本分类	预测类别	
	0	1
0	2 786	1 214
1	162	3 838

表 2　测试样本分类预测表

训练样本分类	预测类别	
	0	1
0	269	731
1	173	827

从表 1 和表 2 中可以计算得出未经过剪枝的决策树模型对训练样本的分类正确率为 82.8%，对测试样本的分类正确率为 54.8%。从预测结果来看，模型对训练样本的分类达到了比较理想的水平，但是对测试样本的分类错误率较高，这是由于过度拟合使模型泛化能力变弱，过度的记忆训练样本，丧失了较好的分类能力。决策树剪枝是避免噪声、提高分类精度的关键技术，但是决策树剪枝的程度同样会对预测精度造成影响，不同修剪程度下决策树模型的预测情况如表 3 所示。

表 3　测试样本分类预测表

修剪程度	训练样本正确率/%	测试样本正确率/%
70	87.6	53.7
75	83.9	54.6
80	82.7	63.4
85	81.3	71.6
90	75.4	71.3
95	66.3	69.4

由表 3 可知，训练样本的预测正确率会随着修剪程度的增大而不断降低，对于测试样本来说，随着修剪程度从 70 不断增加，预测正确率不断提升，在修剪程度为 85 左右时达到最高，然后开始下滑，因此可以初步判定最优的修剪程度在 85～90 之间，具体的修剪程度需要进一步细化，具体情况如表 4 所示。

表 4　测试样本分类预测表

修剪程度	训练样本正确率/%	测试样本正确率/%
82	82.1	69.7
83	82.1	70.6
84	81.7	72.4
85	81.3	71.6
86	79.4	71.3
87	78.3	71.4
88	76.8	71.1

根据结果可知，当修剪程度为 84 时，测试样本的分类正确率最高，此时训练样本的正确率为 81.7%，只降低了 1 个百分点，因此选取 84 作为决策树修剪程度值，由此生成的决策树为最优决策树，根据信息增益比，按分类贡献的百分比对各属性进行排序，结果如图 3 所示。

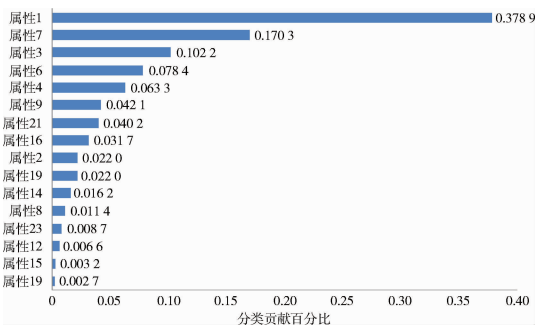

图 3　特征分类贡献度排序图

选择分类贡献度大于 1%的 12 个特征作为最优特征组合，属性编号分别为 1、7、3、6、4、9、21、16、2、19、14、8，由之后的组合模型可知，这些选出的特征将作为神经网络的输入变量。

为验证 GA-BP 神经网络组合模型在“作战”能力预测方面的可用性以及性能优势，采用经过处理的基础训练数据集对模型进行验证，并对预测结果进行分析。主要进行两个方面的实验：一是将无特征选择方法的 BP 神经网络模型预测性能与最优特征组合作为神经网络输入参数的模型进行对比，二是将单一 BP 神经网络模型和单一 GA 算法与 GA-BP 组合神经网络模型性能进行对比分析。

实验采用图 2 中的数据集，使用 Matlab 2011a 进行编程，程序运行主机 CPU 为 Intel Pentium Dual E2220，主频为 2.4 GHz，内存为 1 GB。

根据 BP 神经网络算法，输入变量采用最优特征组合，设置隐含层神经节点数为 9，学习率为 0.002 3，动量因子为 0.000 03，最大迭代次数为 2 000，误差容限为 0.000 01，训练样本和测试样本比例为 8∶2，正负样本的比例为 1∶1，因此一共有 8 000 个训练样本，2 000 个测试样本，单一 BP 神经网络训练预测误差曲线如图 4 所示。

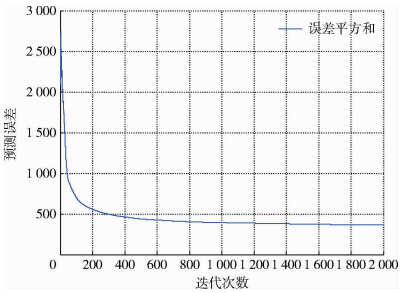

图 4　BP 神经网络训练误差曲线图

从图 4 中可以看出，迭代次数大于 400 后，误差曲线的变化率变小，整个神经网络接近收敛，此时预测总正确率为 85.80%。

1）特征选择方法对比试验。

利用图 3 所示的实验结果，设置 BP 神经网络的输入参数为包含 12 个特征的最优特征组合，将其与单一 BP 神经网络模型各运行 10 次，以百分比为单位绘制两种模型的预测错误率变化曲线图，如图 5 所示。

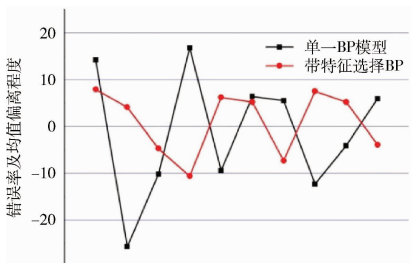

图 5　错误率及均值偏差对比图

增加特征选择方法后，模型的预测精度得到明显提高，通过曲线的振幅可以看出，特征选择方法能有效排除噪声的干扰，使预测的稳定性得到明显提升。带特征选择方法的模型比单一 BP 神经网络在性能上得到了大幅度提升，在能力预测方面稳健的表现能使系统运行更加稳定。

2）单一模型与组合模型性能对比实验。

单一的遗传算法虽然可以得到最优解，但是由于其采用穷举式的搜索方法，数据量庞大时，算法运行时间相当长，其迭代次数曲线如图 6 所示。

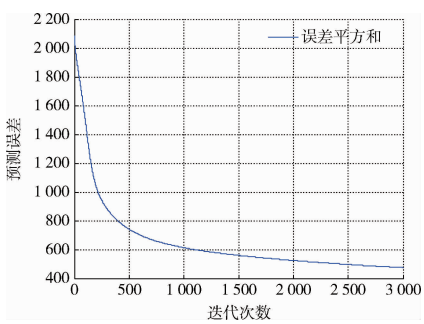

图 6　遗传算法误差曲线图

与单一的BP神经网络模型相比，在达到相同精确度的情况下，遗传算法需要迭代1 000次以上，收敛速度过慢，程序运行时间太长。GA-BP组合神经网络模型使两者取长补短，在收敛速度和预测精度方面都可达到很好的效果。组合模型在隐含层节点数不同时的迭代次数和预测误差见表5。

表5 训练样本分类预测表

隐含层节点数	收敛迭代次数	预测正确率/%
3	60	75.3
4	80	86.5
5	68	88.15
6	166	94.05
7	86	92.5
8	94	91.95
9	107	92.05
10	129	91.75

从表5中可以看出：当隐含层节点数为6时，组合模型的预测正确率最高，此时的收敛迭代次数为166次，比单一BP神经网络和单一遗传算法收敛速度快很多；组合模型预测正确率为94.05%，是所有单一模型中最高的，其训练表现和误差变化如图7所示。

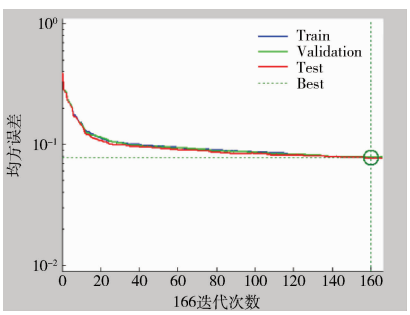

图7 组合模型误差曲线图

从图7中可以看出，误差最小值出现在第160次迭代时，训练数据和测试数据误差的变化比较统一，神经网络在收敛过程中较为稳定，未出现较大波动，说明经过权值初始化后，神经网络不再会收敛于局部极小值，而是越来越接近全局最优解。误差变化梯度与直方图如图8所示。

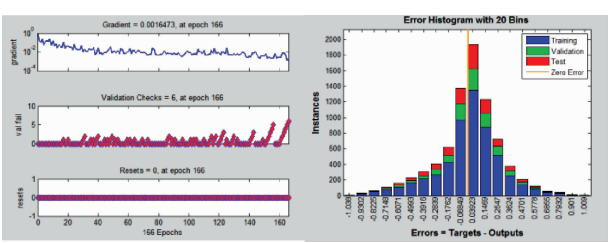

图8 组合模型误差梯度与直方图

综上所述，组合模型在职业能力预测时，收敛速度快，预测正确率高，模型具有稳定性，可以很好地达到系统的需求。

6 结 论

本文根据企事业单位能力数据获取的需要，为实现对专家经验的自学习，以及智能预测职业能力，提出了基于决策树和GA-BPNN的组合预测模型。模型很好地利用决策树在选择最优特征的优势和遗传算法的全局择优性能，弥补了BP神经网络的不足，提高了学习收敛速度和预测准确度。实验证明，组合模型能很好地完成能力预测的功能，且具有很好的稳定性。

[参考文献] (References)

[1] 周玉，钱旭，张俊彩. 可拓神经网络研究综述[J]. 计算机应用研究，2010，27(1)：1-5.
Zhou Yu，Qian Xu，Zhang Juncai. Survey and research of extension neural network [J]. Application Research of Computers，2010，27(1)：1-5. (in Chinese)

[2] 孙志军，薛磊，许阳明，等. 深度学习研究综述[J]. 计算机应用研究，2012，29(8)：2806-2810.
Sun Zhijun，Xue Lei，Xu Yangming，et al. Overview of deep learning [J]. Application Research of Computers，2012，29(8)：2806-2810. (in Chinese)

[3] 吕琼帅. BP神经网络的优化与研究[D]. 郑州：郑州大学，2011.
Lü Qiongshuai. BP Neural Network Optimization and Research [D]. Zhengzhou：Zhengzhou University，2011. (in Chinese)

[4] 徐丽萍，姜志旺. 基于粗糙集及信息增益的数据挖掘预测算法[J]. 中国科技论文，2012，7(7)：552-559.
Xu Liping，Jiang Zhiwang. A new data mining and prediction algorithm based on rough set and information gain [J]. China Sciencepaper，2012，7(7)：552-559. (in Chinese)

[5] 楼旭伟，楼辉波，朱剑锋. 基于遗传算法径向基神经网络的交通流预测[J]. 中国科技论文，2013，8(11)：1141-1144.
Lou Xuwei，Lou Huibo，Zhu Jianfeng. Prediction of traffic flow of optimized radial basis function neural network based on genetic algorithm [J]. China Sciencepaper，2013，8(11)：1141-1144. (in Chinese)

[6] 陈明. 神经网络原理与实例精解[M]. 北京：清华大学出版社，2013：111-112.
Chen Ming. Neural Network Theory and Instance Analysis [M]. Beijing：Tsinghua University Press，2013：111-112. (in Chinese)

[7] Linoff G S，Berry M J A. Data Mining Techniques：for Marketing，Sales，and Customer Relationship Management [M]. 2nd ed. New York：John Wiley & Sons，2011.

[8] 田少杰，洪跃，李阳. 基于改进BP神经网络的预测系统开发[J]. 工业控制计算机，2012，25(11)：77-81.
Tian Shaojie，Hong Yue，Li Yang. Forecast system based on improved BP neural network [J]. Industrial Control Computer，2012，25(11)：77-81. (in Chinese)

[9] 王鹏. 网络化作战条件下系统评估方法研究[J]. 指挥控制与仿真，2011，33(3)：24-27.
Wang Peng. Research on evaluation method of network operation system [J]. Command Control and Simulation，2011，33(3)：24-27. (in Chinese)