

文章编号: 1000-6788(2003)12-0070-06

## 基于 GA 神经网络的个人信用评估

朱兴德, 冯铁军

(上海财经大学经济信息管理系, 上海 200083)

**摘要:** 提出了基于遗传算法神经网络的个人信用评估模型, 利用标准遗传算法和 Solis&Wets 算法的混合算法同时优化神经网络的结构和权重/阈值系数, 并在探讨个人信用评估指标的基础上, 针对模型实际应用问题提出了解决方案。

**关键词:** 个人信用评估; 神经网络; 遗传算法; Solis&Wets 算法; 个人信用评估指标

**中图分类号:** F830.73

**文献标识码:** A

## Individual Credit Appraisal Based on Neural Networks Optimized by Genetic Algorithm

ZHU Xing-de, FENG Tie-jun

(Dept of Economics Information Management, Shanghai University of Finance and Economics, Shanghai 200083, China)

**Abstract** Individual credit appraisal technology is one of the critical issues of building national individual credit system. An individual credit appraisal model based neural network whose topology and connection weights are optimized by genetic algorithm (GA-NN) is proposed in this paper. And the detail of the model's application is given based the discussion about the individual credit appraisal indexes.

**Key words:** individual credit appraisal; neural networks; genetic algorithm; solis&wets algorithm; individual credit appraisal indexes

随着我国市场经济和金融改革的不断深化, 电子商务的高速发展, 加入 WTO 后带来的冲击的不断临近, 个人信用评估体系的建设和完善已经迫在眉睫。

我国个人信用评估体系存在的问题可以归结为三点: 个人信用信息的不完善、个人信用观念淡薄和个人信用评估技术的不成熟。本文主要就第三个问题进行探讨。

目前我国存在上海和济南两种征信模式。前者是政府主导型, 偏重于个人信用信息的收集; 后者以济南建行为代表, 它基于判别分析法并借鉴了国外商业银行的个人信用计分模型推出了目前国内较为规范的个人信用评分标准, 该标准从个人的“自然情况”、“职业情况”、“家庭情况”、“与建行关系”四大方面, 计 19 个项目、细分 72 档分值进行逐项逐档打分。这种模式的缺点在于主观色彩浓厚, 信息采集范围狭窄, 评估结果缺乏公允; 模型业务针对性强, 推广性差, 信用资源浪费严重。

基于此, 本文提出了一种基于 GA-NN (基于遗传算法的神经网络) 的个人信用评估模型和算法, 在学习并且展现评价专家经验的同时, 减少主观因素影响, 并可根据个人信用体系建设的进程方便的调整模型。本文还针对模型实际应用中的问题提出了解决方案。

### 1 基于遗传算法的神经网络

神经网络是以大量的具有相同结构的简单单元的连接来模拟人类大脑的结构和思维方式的一种可实现的物理系统或可通过计算机进行模拟实现。神经网络处理信息不同于数学方法的精密计算, 它是通过信

收稿日期: 2002-11-20

作者简介: 朱兴德, 上海财经大学经济信息管理系, 副教授; 冯铁军, 浙江省舟山市, 研究生

息样本对神经网络的训练,使其具有类似人脑的记忆、辨识能力,完成各种信息处理功能.人工神经网络具有良好的自学习、自适应、联想记忆、并行处理和非线性转换的能力,避免了复杂数学推导,在样本缺损和参数漂移的情况下,仍能保证稳定的输出.

神经网络(Neural Networks, NN)具有两种基本特性,即模式分类能力和非线性函数的表示能力.它在工程领域的应用已经取得了显著的成效.在国外,神经网络模型已在市场研究、签名核实、借贷评估、公司金融状况分析等方面得到应用,并显示出巨大的生命力<sup>[1]</sup>.但是目前神经网络方法仍然面临三大问题:1)难以确定神经网络的优化结构;2)学习时间较长;3)传统的梯度法或基于损失函数的优化方法,在有随机扰动下不能达到最佳效果,常常收敛于局部最优解<sup>[2]</sup>.

遗传算法(Genetic Algorithm, GA)是由自然界的遗传进化理论发展而来,已经成功地解决了许多复杂的优化问题.它的最大的优点是:即使对多态的和非连续的函数,它也能获得全局最优解.因此许多作者都尝试将遗传算法与神经网络结合在一起,利用遗传算法优化设计神经网络的结构和权重系数.在设计神经网络结构时,遗传算法可以自主的辨识最小的包含最优解的搜索空间<sup>[3]</sup>.

考虑到个人信用评估系统是一个多指标的复杂系统,其输入输出规模已经非常大,因此作者考虑采用基于遗传算法的前向神经网络同时优化个人信用评估神经网络模型的结构和权重/阈值<sup>[4]</sup>(GA-NN).以下详细介绍该算法.

1.1 遗传算法的编码

遗传算法在由码串表示的个体组成的群体上进行遗传算子运算,每个码串代表一种神经网络的结构和权重/阈值矢量.本文采用二值编码和实数值编码的混合编码方法:神经网络的结构采用二值编码,而神经网络的连接权重和阈值系数采用实数值编码.图 1 所示为一个两层前向神经网络及其编码结果,其中‘0’输入神经元输入取值恒为-1,其与其他神经元的连接系数表示被连接神经元的阈值.

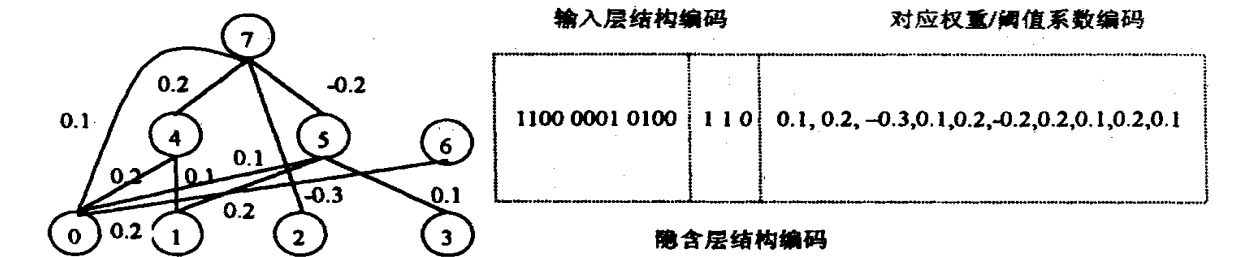


图 1 对前向神经网络的结构和权重/阈值进行混合编码

结构编码中,二值码‘1’表示有相联,二值码‘0’表示无相联.混合编码的二值部分为固定长度,实值部分为变长度,总的码串也为变长度.

在本文提出的上述混合编码方法的基础上,可以对表示神经网络结构的二值码串进行标准的遗传算子运算,从而保留了典型遗传算法方法的优点:计算简单明了,遗传算子对遗传空间的搜索非常有效,易于扩展到大规模的神经网络的优化设计<sup>[5]</sup>.同时对表示神经网络的连接权重系数的实数值编码所进行的 Solis&Wets 运算<sup>[6]</sup>又使新的遗传算法具有 EP 和 ES 的优点.

起始的群体  $P^0 = (a_1^0, \dots, a_n^0)$  中的每个码串  $a_i^0$  是随机产生的,但是实值编码被限定在可能的连接权重系数取值范围之内.起始码串采用均匀分布的随机分布函数产生,并成为下一次遗传搜索过程中个体的父母集合.

1.2 适应度计算和遗传算法的选择算子

将由码串表示的每个个体反编码为相应的神经网络,然后输入所有训练样本,计算神经网络的输出与期望输出之间的平均绝对误差(MAE 误差),将此误差的倒数作为此个体的适应度:

$$Fitness = \frac{1}{\frac{1}{M} \sum_{i=1}^M |Y_i - Y_0|}$$

(1)

其中  $M$  为训练样本总数,  $Y_i$  为第  $i$  个训练样本时神经网络的输出,  $Y_0$  为期望输出.

为了避免未成熟收敛现象,本文对该适应度函数作了线性调整.

$$Fitness = aFitness + b \quad (2)$$

式中系数  $a, b$  的选择需满足以下两个条件:

- 1) 原适应度平均值等于调整后的适应度平均值;
- 2) 调整后适应度函数的最大值要等于原适应度函数平均值的指定倍数.

$$Fitness_{\max} = cFitness_{\text{avg}} \quad (3)$$

其中  $c$  是为得到所期待的最优个体的复制数,  $c$  取值一般为 2.

本文采用轮盘选择算子选取进行下一次遗传运算的父母码串.

### 1.3 遗传算法的交换算子

本文提出的遗传算法只对码串中表示神经网络结构的二值码部分进行标准遗传算法的交换算子的运算, 然后将对应的实值编码进行对等的交换来完成整个码串的交换算子的运算. 本文采用两点杂交, 杂交概率  $p_c$  取为 0.7.

### 1.4 遗传算法的变异算子

本文提出的遗传算法只对码串中表示神经网络结构的二值码部分进行标准遗传算法的变异算子的运算, 类似于遗传算法的交换运算, 同时对表示神经网络权重系数的实值编码进行相应的删除和增加. 鉴于算法中添加了修正机制, 本文选择了较大的变异概率  $p_m = 0.3$ .

这种对混合编码进行的遗传变异算子运算, 对表示神经网络结构的二值码部分产生了较强的变异效果, 而对表示神经网络权重系数的实值码部分变异效果较弱. 因此本文提出了包含 Solis&Wets 算法的混合搜索方法来加强神经网络权重和阈值系数的变异效果, 即对一部分个体采用标准的遗传算子产生后代, 其余的通过 Solis&Wets 算法产生后代. Solis&Wets 算法的具体内容参见文献[6].

### 1.5 修正机制

针对如图 1 所示的 3 层前向神经网络, 结构编码 1100, 0001, 0100, 1, 1, 0 和 1110, 0001, 0110, 1, 1, 0 其实表示的是同一种网络结构. 为了避免种群中同构个体过多而导致算法过早收敛或者搜索失败, 本文在算法中增加了修正机制, 将网络结构的冗余连接全部删除.

另外, 对于类似于 1100, 0100, 0100, 001 这样的结构编码, 其表示的网络中输入和输出不发生任何联系. 修正机制同样对这种无效网络进行了规避处理, 杜绝这种个体的存在.

## 2 GA-NN 模型的应用

本文提出的 GA-NN 模型应用于个人信用评估时需要解决以下问题.

### 2.1 输入设计

个人信用评估系统是一个多指标的复杂系统. 在 GA-NN 实际应用中, 需要解决评估指标和神经网络输入神经元之间的对应关系, 将定性指标转化为定量指标, 并进行正规化处理.

#### 1) 确认输入神经元

综合目前国内外银行和信用机构的个人信用评估内容, 个人信用评估指标可分为自然情况、收入情况、资产/负债情况、道德情况四项, 具体指标如表 1 所示<sup>[7,8]</sup>.

确认输入神经元的步骤如下:

**第一步** 对个人信用评估指标逐层划分, 直至指标层. 按照数据格式, 指标可分为定量指标(如工资)、定性指标(如健康情况)、枚举型指标(如职务).

**第二步** 根据需要合并指标, 即将相关总量指标合并为一个相对指标. 指标合并出于两种考虑: 个人信用评估指标数量庞大, 模型处理的时间将会很长; 指标合并后对评估结果不产生重大影响.

**第三步** 将枚举型指标涵盖的内容根据其对个人信用的影响程度进行概括分类, 每一类对应神经网络的一个输入神经元. 如职务可以划分为首要职位(包括总经理、店主、承包商等)、管理层(包括副总经理、高级经理、经理、工厂经理、办公室经理等)、专家(包括软件工程师、财务人员、医生、教授等)、技术人员(包括飞行员、机械工、电工、管子工等)、非技术员工(包括工人、农业工人、卡车司机、出租车司机等)<sup>[9]</sup>.

**第四步** 将其他非枚举型评估指标一一对应于神经网络的输入神经元.

表 1 个人信用评估指标

要素	指标
自然情况	年龄, 性别, 婚姻, 学历, 家庭构成, 家庭地址, 健康情况, 户口等 .
收入情况	行业, 职业, 职务, 职称, 工作年限, 工作月收入, 其他稳定收入 (如赡养费、抚养费 etc), 其他家庭成员收入等 .
资产/负债情况	资产: 金融流动性资产, 金融投资性资产, 实物资产, 无形资产, 其他资产 (如遗产, 捐赠等); 负债: 短期负债 (信用卡, 分期付款, 账单, 短期借款, 个人借贷), 长期负债 (房屋贷款, 汽车贷款, 助学贷款等) .
道德情况	社会道德: 公检法的个人记录等; 信用道德: 历史还款记录, 公用事业付费记录, 电信付费记录等 .

2) 定性指标的定量化

定性指标一般可以取 0~ 1 之间的数来表征, 如对于健康情况, 1 表示非常好, 0.8 表示很好, 0.3 表示不好等 .

3) 定量指标的正规化

定量指标, 一般不外乎下列几种类型: 成本型 (越小越好型)、效益型 (越大越好型)、适中型 (即不能太大又不能太小为好型)、区间型 (属性值在某一固定区间为好型)<sup>[10]</sup> .

这些指标可以分别用以下方式处理, 其中

$$r_{pi} = u_{di}(x_{pi}) \quad i = 1, 2, \dots, n_1$$

(4)

为评价指标  $u_i$  的属性值  $x_{pi}$  的满意度, 且  $r_{pi} \in [0, 1]$ , 其中  $u_{di}(\bullet)$  是定义在论域  $d_i = [m_i, M_i]$  上的指标  $u_i$  量化的隶属函数,  $m_i$  和  $M_i$  分别表示评价指标  $u_i$  的最小、最大值 .

· 成本型指标量化的隶属函数

$$r_{pi} = u_{di}(x_{pi}) = \begin{cases} 1, & x_{pi} \leq m_i \\ \frac{M_i - x_{pi}}{M_i - m_i}, & x_{pi} \in d_i \\ 0, & x_{pi} \geq M_i \end{cases}$$

(5)

· 效益型指标量化的隶属函数

$$r_{pi} = u_{di}(x_{pi}) = \begin{cases} 1, & x_{pi} \geq M_i \\ \frac{x_{pi} - m_i}{M_i - m_i}, & x_{pi} \in d_i \\ 0, & x_{pi} \leq m_i \end{cases}$$

(6)

· 适中型指标量化的隶属函数

$$r_{pi} = u_{di}(x_{pi}) = \begin{cases} \frac{2(x_{pi} - m_i)}{M_i - m_i}, & x_{pi} \in (m_i, (M_i + m_i)/2) \\ \frac{2(M_i - x_{pi})}{M_i - m_i}, & x_{pi} \in [(M_i + m_i)/2, M_i] \\ 0, & x_{pi} \leq m_i \text{ 或 } x_{pi} \geq M_i \end{cases}$$

(7)

· 区间型指标量化的隶属函数

$$r_{pi} = u_{di}(x_{pi}) = \begin{cases} 1 - \frac{w_{l1} - x_{pi}}{\max(w_{l1} - m_i, M_i - w_{l2})}, & x_{pi} < w_{l1} \\ 1, & x_{pi} \in [w_{l1}, w_{l2}] \\ 1 - \frac{x_{pi} - w_{l2}}{\max(w_{l1} - m_i, M_i - w_{l2})}, & x_{pi} > w_{l2} \end{cases}$$

(8)



其中  $[w_{l1}, w_{l2}]$  为指标  $u_i$  的最佳稳定区间。

## 2.2 输出设计

神经网络的输出表示个人信用评估的结果。目前,国内外的个人信用评估结果表示方式主要有:等级制,百分制,在具体的业务应用中还存在信贷额度、欺诈风险等方式。考虑到神经网络的直接输出是  $[0, 1]$  区间的一个具体数值,因此神经网络的输出形式有两种选择:

1) 多个输出神经元,每个输出神经元对应一种评估结果(如等级),理想的输出神经元取值为 0 或 1, 1 表示评估结果为该神经元对应的评估结果,否则不是。

2) 一个输出神经元,输出取值范围为  $[0, 1]$ , 可理解为个人信用评估分数,且可通过增加乘积项使结果更直观。

比较两种方式,前者的缺点在于:增加了神经网络模型的复杂程度,减缓算法的运行速度;评估结果粗犷;容易出现无法精确判别的情况。后者的缺点在于:对样本数据的数量和正确性要求高。综合考虑,一般取后者。需要注意的是样本的输出值范围要与输出神经元输出值范围一致,可通过对样本输出值进行正规化或者对输出神经元的输出值增加乘积项实现。

## 2.3 样本设计

神经网络模型有效与否与训练样本的设置有重要关系。在组织训练样本集时,通常需要遵循的基本原则是:考虑到各方的情况,如参数之间及参数与结果之间的相互影响关系,尽可能多地为网络提供必要的信息。

如果一个问题中包含有  $m$  个参数,而每个参数各有  $n_i (i = 1, 2, \dots, m)$  个可能值,那么训练样本集中至少应有  $N = n_1 * n_2 * \dots * n_m$  (个)独立样本。这样形成的训练集样本才是完全的。

但是神经网络模型在个人信用评估中的应用起到的是函数逼近的作用,用以通过不完整的样本集构建出较理想的网络模型,针对新的输入输出较正确的结果。但是其样本集同样是越丰富越好,并且样本分布要均匀。

最初的样本输出不可避免的包含了专家的主观意识,这需要随着个人信用评估市场的完善,通过实际的个人信用行为动态的调整个人信用评估结果来修正这些样本。这个过程和模型的调整过程以及国家个人信用体系的完善过程是同步的。

## 2.4 模型调整

由于目前我国个人信用体系的不完善,个人信用数据存在缺漏和虚假的问题,因此神经网络模型在个人信用评估中的应用需要根据个人信用数据的采集情况进行不断调整,从小规模单业务的应用逐步调整到普遍统一应用,从而达到国际水平。从实际应用角度来说,其过程如下:

1) 以银行内部积累的个人信用信息为基础,辅以政府支持的征信机构采集的信用信息,建立最初的个人信用评估样本,训练神经网络模型并进行应用,此时可适当结合专家意见对模型输出的结果进行修正。

2) 银行和其他机构建设并使用个人信用积分系统,根据个人信用行为动态的调整个人信用,在此基础上,周期性的改善样本,并改善模型。

3) 征信机构采集更多的信用信息,将此信息加入样本中,重新构架模型并进行训练。

## 2.5 模型的计算机实现

本文提出的 GA-NN 算法是在 Windows 2000 server 平台上基于 Delphi+ SQL Server 2000 实现的。

该程序基于面向对象编程原理,构建了个体类和群体类。个体类主要实现个体的初始化、适应度计算和修正;群体类主要实现遗传算子运算和 Solis&Wets 运算。同时算法利用动态数组的技术充分减少算法的内存占有量。为了提高通用性和使用方便性,系统还提供了开放式的模型参数设置界面和多种数据读写接口(文件、剪贴板、数据库)。

## 3 仿真算例和结果分析

仿真算例中包含 4 个输入单元, 1 个输出单元, 3 个隐含层, 每个隐含层的最大神经元数量为 5 个, 种

群规模为 30。本文根据建行提供的网上个人信用评估工具, 利用其中的 4 个指标: 工作年限 ( $I_1$ )、个人年经济收入 ( $I_2$ )、是否拖欠银行贷款 ( $I_3$ )、是否学士 ( $I_4$ ) 作为输入, 用专家评分结果 ( $O$  (专家)) 作为实际输出, 设计了训练样本, 并对数据进行了正规化, 如表 2 所示。

表 2 仿真算例使用的样本数据

$I_1$		$I_2$		$I_3$	$I_4$	$O$ (专家)		学习结果
2	0.05	30	1	1	1	22	0.22	0.2186
5	0.125	18	0.9	0	1	32	0.32	0.321803
9	0.225	9	0.45	0	0	31	0.31	0.311248
15	0.375	4	0.2	1	0	15	0.15	0.151444
25	0.625	2	0.1	0	1	25.5	0.255	0.255426
40	1	1	0.05	0	0	10	0.1	0.098452
1	0.025	1	0.05	0	0	10	0.1	0.098452
3	0.075	3	0.15	0	0	11	0.11	0.113785
1	0.025	5	0.25	0	1	21	0.21	0.208126
9	0.225	5	0.25	1	1	23	0.23	0.231574
2	0.05	3	0.15	0	0	11	0.11	0.106041

表 3 最中种群前 1/3 最优个体

结构	连接数	适应度值	MAE 误差
4, 5, 2, 1	28	476.82769	0.002097194
4, 3, 1, 1	16	445.28497	0.002245753
4, 5, 1, 1	23	441.94743	0.002262713
4, 4, 2, 1	24	438.4635	0.002280692
4, 4, 1, 1	23	419.917684	0.002381418
4, 4, 4, 1	28	373.91702	0.00267439
4, 4, 1, 1	19	349.85507	0.002858326
4, 3, 1, 1	14	317.77857	0.003146845
4, 2, 2, 1	14	317.65707	0.003148049
4, 3, 3, 1	24	317.64242	0.003148194

图 2 所示为运行结果, 横坐标分别表示遗传搜索次数; (a) 纵坐标表示每次迭代产生的最优个体的适应度; (b) 纵坐标表示每次迭代产生种群的前 1/3 较优个体的平均 MAE 误差。

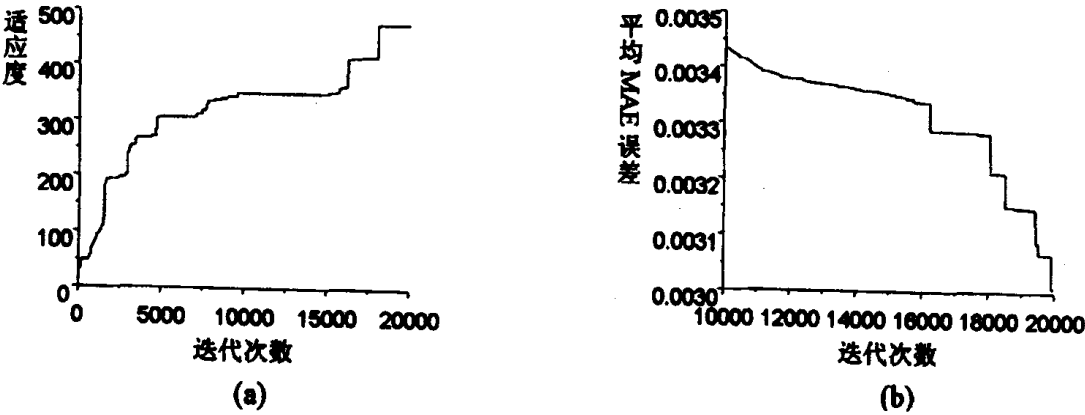


图 2 仿真算例运行结果

从图 2(a) 可以发现, 种群最优个体的适应度上升速度很快, 在算法迭代 10000 次时, 平均绝对误差已经达到 0.003。在迭代过程中存在局部收敛和阶跃现象, 这一方面是由于算法的选择机制始终保留父种群的最优个体; 另一方面是由于在种群适应度较高的时候, 标准遗传算子对混合编码个体的适应度影响较小, 需要等待其他较优个体的出现来促使更优个体的产生。参照图 2(b), 可以发现在最优个体不变的情况下, 种群的其他个体仍在进化。

(下转第 115 页)

下面给出一个紧例:  $\{r_1=1, r_2=0, p_1=4, p_2=p_3=p_4=3, p_5=2, p_6=p_7=1\}$ . 不难验证  $C^H=C_1=C_2=C_3=C_4=10, C^*=9$ , 近似比为  $10/9$

#### 参考文献:

- [1] Yao Enyu, Li Rongheng, He Yong, New progress on optimal partitioning problems[A]. International Conference on Optimization Technique and Application'92[C]. Singapore, World Scientific, 1992, 229- 234
- [2] 姚恩瑜. 集的最优分划问题简介[J]. 运筹学杂志, 1992, 11: 18- 23
- [3] Lee C Y. Parallel machine scheduling with nonsimultaneous machine available time[J]. Disc Appl Math, 1991, 30: 53- 61.
- [5] 徐立新, 张玉忠. 集合核约束分划的贪婪算法分析[J]. 系统工程理论与实践, 1999, 19(4): 129- 132
- [4] Lee C Y. He Yong, Tang Guochun, A note on "Parallel machine scheduling with nonsimultaneous machine available time"[J]. Disc Appl Math, 2000, 100: 133- 135
- [6] Lin Guohui, He Yong, Lu Haiyan, Yao Yujun. Exact bounds of the modified LPT algorithm applying to parallel machines scheduling with nonsimultaneous machine available times[J]. Applied Mathematics-A Journal of Chinese University, 1997, 12B: 109- 116
- [7] He Yong. Parametric LPT-bound on parallel machine scheduling with nonsimultaneous machine available time[J]. Asia-Pacific J of Operational Research, 1998, 15: 29- 36
- [8] He Yong, Kellerer H, Kotov V. Linear compound algorithms for the partitioning problem [J]. Naval Research Logistic, 2000, 47: 593- 601.

(上接第 75 页)

表 3 所示为最终种群的前 1/3 较优个体的情况. 决策者可以综合考虑适应度值和网络结构选择其中的一种进行应用. 最优适应度对应个体的样本学习结果可参见表 2 的末列.

## 4 结论

本文提出利用基于遗传算法的神经网络构建个人信用评估模型, 一方面减少了信用评估中主观因素的影响, 有利于模型的动态调整; 另一方面通过标准遗传算法和 Solis&Wets 算法结合的混合算法同时简化神经网络的结构和优化权重/阈值, 使得神经网络更加精简有效, 并有利于对模型输入参数做进一步的研究.

#### 参考文献:

- [1] Robert R Trippi, Efraim Turban. Neural Networks in Finance and Investing [M]. Probus Publishing Company, 1992
- [2] 李敏强, 徐博艺, 寇纪淞. 遗传算法与神经网络的结合[J]. 系统工程理论与实践, 1999, 19(2): 65- 69
- [3] Koza J R, Rice J P. Genetic generation of both the weights and architecture for a neural network [A]. International Joint Conference on Neural Networks, IJCNN - 91- Seattle [C], 1991: 397- 404
- [4] 黎明, 严超华, 刘高航. 遗传算法优化前向神经网络结构和权重矢量[J]. 中国图象图形学报(A 版), 1999, 4(6): 491 - 496
- [5] Vittorio M. Genetic evolution of the topology and weight distribution of neural networks [J]. IEEE Transactions on Neural Networks, 1994, 5: 39- 53
- [6] Solis F J, Wets J B. Minimization by random search techniques [J]. Mathematics of Operations Research, 1981, 6: 19- 50
- [7] 李平. 加强信用分析, 促进消费信贷的健康发展[J]. 华东经济管理, 2000, 14(5): 53- 54
- [8] John B etc. Managing Credit Risk [M]. 北京: 机械工业出版社, 2001
- [9] Robert H. Cole. Consumer and Commercial Credit Management [M]. Richard D Irwin NC, 1984
- [10] 王宗军. 复杂对象系统多目标综合评价的神经网络方法[J]. 管理工程学报, 1995, 9(1): 26- 33