遗传算法在基于二元决策树的模型选择中的应用①

吴 菲② 黄梯云
(哈尔滨工业大学管理学院)

【摘要】模型选择是模型管理的重要部分. 本文提出利用二元决策树实现模型选择,并采用遗传算法构造二元决策树. 最后给出了在趋势预测模型选择中的应用实例,并分析了影响二元决策树错误率的因素.

关键词:二元决策树,模型选择,遗传算法
分类号:C934

0 引 言

模型选择是模型管理的重要部分,模型的自动选择一直是决策支持系统研究领域的重要问题. 迄今,已研究出许多的模型选择方法,这些方法可以被分为两大类:解析法与人工智能法. 解析法是基于目标线性规划模型的方法,依靠模型使用的历史信息来选择模型[1,2];人工智能方法解决许多用传统数学方法难以解决的问题,其中专家系统和人工神经网络是人工智能领域中的两个受到普遍关注的研究课题. 模型选择实际可以认为是一个模式识别过程,与传统的解析方法和专家系统相比,人工神经网络在模式识别和动态系统的预测中更具有优越性[3],为此国内外许多学者在这方面进行了深入的研究[4~7]. 但神经网络也有其局限性,主要表现在以下几点:

1)如何有效地确定神经网络模型中的各种参数到目前为止还未有很好的解决方案,只能根据先前的经验或大量的实验来确定这些参数[8].

2)从数学上看,它是一个非线形优化问题,这就不可避免的会出现局部最小问题[9].

3)算法的收敛速度慢,通常要迭代几千步甚至上万步[9].

70年代,Holland提出了基于生物学进化原理的遗传算法,它模拟达尔文的遗传选择和自然淘汰的计算模型,通过在求解过程中,使群体不断优化,进而找到最优解或准最优解. 作为一种有效的全局并行优化搜索工具,遗传算法具有简单、通用、鲁棒性强和适于并行处理的特点,因而在工农业、经济政治、科学研究等方面取得了广泛的应用.

本文提出了采用遗传算法构造二元决策树对数据进行特征识别,并根据识别的结果选择模型的方法,该方法克服了神经网络在模型选择中的不足,提高了模型选择的准确率. 为此,进行以下三个方面的研究:

1) 基于遗传算法的二元决策树权值矢量求解;

2) 构造二元决策树;

3) 二元决策树在模型选择中的应用.

1 二元决策树

决策树是一种结构简单、搜索效率高的分类器[10]. 目前,出现了采用遗传算法对决策树结构运算,直接构造决策树的方法[11],和通过遗传算法求解树结点的线性权值矢量,进而构造二元决策树的方法[12]. 本文提出一种新的通过遗传算法

① 国家自然科学基金资助项目(79670023).
② 吴 菲:博士生,通讯地址:哈尔滨工业大学管理学院. 邮编:150001. 联系电话:0451-6672238.
本文1998年11月16日收到.

求解树结点的线性权值矢量，并根据树结点的错误率与分割后的错误率减少量构造二元决策树的方法．此方法提高了二元决策树的分类准确率，并且可以人为地控制二元决策树的分类准确率．

二元决策树包含终结点（叶子结点）和非终结点，其中终结点为类属性，非终结点包括决策函数和其左右子树．其决策过程就是从根结点输入值后，根据二元决策树的非终结点的决策函数决定选择哪一个子结点，从而到达终结点，那么终结点对应的类为输入值的类属性[12]．

树中的每个结点可用唯一的由符号{0，L，R}组成的串表示．根结点由字符"0"代表．假设一个结点由串 t 表示，则它的左结点表示为 tL，右结点表示为 tR．

设 $x_t=\{x_1,x_2,\cdots,x_{Nt}\}$ 为到达当前结点 t 的训练样本集，样本大小为 Nt．假设 $x_t=x_t^1\cup x_t^2\cup\cdots\cup x_t^c$，$c$ 代表类的数目，x_t^i 代表 x_t 中属于类 ω_i 的所有样本，

$$x_t^i=\{x|x\in x_t,\text{并且 } x \text{ 来自于 } \omega_i\}$$

x_t 的错误率定义为

$$i(x_t)=\sum_{i=1}^{c}\sum_{j\neq i}p(\omega_i|x_t)p(\omega_j|x_t)$$，这里

$$p(\omega_i|x_t)=\frac{\#x_t^i}{\#x_t}$$

"#"代表样本集样本数目的大小．

当 x_t 仅包含来自同一类的样本时，$i(x_t)$ 为 0．假设线性决策函数 w^Tx 将 x_t 分为 $x_{tL}(w)$ 和 $x_{tR}(w)$ 两部分，则

$$x_t=x_{tL}(w)\cup x_{tR}(w)$$

并且

$$x_{tL}(w)=\{x|x\in x_t,\text{并且 } w^Tx<0\}$$

$$x_{tR}(w)=\{x|x\in x_t,\text{并且 } w^Tx>0\}$$

样本分割后的错误率 $i'(x_t,w)$ 定义为

$$i'(x_t,w)=p(x_{tL}(w)|x_t)i(x_{tL}(w))-p(x_{tR}(w)|x_t)i(x_{tR}(w))$$

这里，

$$p(x_{tL}|x_t)=\frac{\#x_{tL}}{\#x_t},p(x_{tR}|x_t)=\frac{\#x_{tR}}{\#x_t}$$

错误率的减少量为

$$\Delta i(x_t,w)=i(x_t)-i'(x_t,w)$$

为了建立一个二元决策树分类器，对每个非终结点，需要利用遗传算法来寻找一个线性权值矢量 w^*，将到达此结点的样本集最优地分割为两个子集，使得 $\Delta i(x_t,w)$ 为最大．

2 权值矢量的求解

为了求解权值矢量 w^*，首先对每个权值矢量 w 编码．设权值矢量 $w=(w_1,w_2,\cdots,w_n)$．每个系数 $w_i(i=1,2,\cdots,n)$ 对应一个确定的二进制串 A^i，将 A^i 联结在一起则形成对应 w 的二进制串 A，$A=A^1A^2\cdots A^n$．适应度函数定义为

$$f(A)=\Delta i(x,T^{-1}(A))$$

这里 T 代表从权矢量 w 到对应的二进制串 A 的操作．

$$A=T(w),w=T^{-1}(A)$$

对于交叉操作，使用两点交叉方法．设

$$A=a_1a_2\cdots a_l$$

$$B=b_1b_2\cdots b_l$$

其中 l 为串长．假如选择两点位置为 k_1 和 k_2，且 $k_2>k_1$，交叉操作后，

$$A'=a_1\cdots a_{k_1}b_{k_1+1}\cdots b_{k_2}a_{k_2+1}\cdots a_l$$

$$B'=b_1\cdots b_{k_1}a_{k_1+1}\cdots a_{k_2}b_{k_2+1}\cdots b_l$$

两点交叉操作优于一点交叉操作之处在于使得交叉操作有效地作用于每个参数的位串上，充分发挥交叉操作的威力，同时避免了参数位串破坏过大或参数位串改变甚小的不合理情况的出现，并明显地提高了搜索效率．交叉操作按通常接近于 1 的概率 P_C 发生．

变异操作则与前述相同，变异操作发生概率为 P_m，通常较小（$P_m<0.15$）．

终止条件可以规定为产生多少代后停止，或规定为产生一定的后代后解不发生变化而停止．本文采用后一种终止条件．

初始解一般随机产生，对初始解的产生作如下规定：

（1）对样本集进行线性标准化操作；

（2）初始解空间中的任一解分量为随机来自两个不同类的样本差值．假设 x_1 和 x_2 为两个来自不同类的样本，$x_i=(x_{i1},x_{i2},\cdots,x_{im},1)^T,i=1,2$．那么初始解空间权值矢量 $w_j=(x_{21}-x_{11},x_{22}-x_{12},\cdots,x_{2n}-x_{1n},\frac{1}{2}\sum_{i=1}^{n}(x_{2i}^2-x_{1i}^2))^T$，并对 w 进行线性标准化操作．

这样产生初始解空间中每个权值矢量的每个

系数 w_{ij} 的值在 0 与 1 之间,加快迭代收敛速度.

遗传算法寻找最优权值矢量的操作步骤如下:

step1　建立一个初始种群 $p=\{A_1,A_2,\cdots,A_n\}$,A^* 为当前种群中的最优解,$f^*=f(A^*)$,$g^*=g=0$,g 代表进化过程中的繁衍代数,g^* 表示最优解第 1 次产生时的繁衍代数.

step2　对当前种群 p 中的每一个体计算适应度值 $f(A)=\Delta i(p,T^{-1}(A))$

step3　初始化解交配池 M 和后代 O:

$M=\Phi$

$O=\Phi$

$g=g+1$

step4　复制

for i = 1 to n do

用转轮法(roulette wheel method)从 p 中选择父代 A

$M=M\cup\{A\}$

step5　交叉

for i = 1 to n/2 do

从 M 中随机地选择父代 A' 和 A''

$M=M-\{A'-A''\}$

$O=O\cup\{A',A''$ 按照概率 P_c 交叉的后代$\}$

step6　变异

假设 $O=\{A_1^o,A_2^o,\cdots,A_n^o\}$,

for i = 1 to n do

对 A_i^o 的每点,按概率 P_m 变异

step7　替换

假设 A_b^o 和 A_w^o 分别是上一代种群 p 中的最优解和新产生的子代 O 中的最差解

$P=(O-\{A_w^o\})\cup\{A_b^o\}$

step8　假设 A_b 为新一代 p 中的最优解

if $f(A_b)>f^*$

then $A_b^*=A_b$,$f^*=f(A_b)$,$g^*=g$,转 step2;

else if $g-g^*<G$,G 为预先给定的一控制常量,

then 转 step2;

else 返回最优解 A_b^*.

3　构造二元决策树

选择自上而下的方法构造二元决策树.从根结点开始,在每个非终结点,用所分得的子结点集作为标准训练线性权值矢量.对每个新产生的子结点重复此过程直到满足终止条件,则定义此结点为终结点.

对于给定的训练集,参数定义如下:

Q 代表处理过程中的结点集;

N_0 为预定义的到达一个结点的最少样本数;

I_1 为预定义的结点的最小错误率值;

I_2 为预定义的结点分割后的错误率减少量最小值;

当一个结点 x_t 的错误率值 $i(x_t)$ 小于 I_1,到达其的样本数目小于 N_0 或分割后的错误率的减少量 $\Delta i(x_t,w)$ 小于 I_2 时,定义此结点为终结点,其类属性 ω_i,按以下公式计算:

$$i=\arg\max_i(p(\omega_i|x_t))$$

构造二元决策树的算法如下:

step1　打开样本库

step2　对样本进行线性标准化

step3　生成根结点

step4　递归建立二元决策树

1) if 到达结点 x_t 的样本数小于给定最小样本数,即 $\#x_t<N_t$

then 结点 x_t 为叶结点,返回

2) if 结点 x_t 的错误率小于最小错误率,即 $i(x_t)<I_1$, then 结点 x_t 为叶结点,返回

3) 调用遗传算法求解结点的权值矢量 w_t

4) 将到达结点 x_t 的样本分割为两个样本,计算分割后的 $\Delta i(x_t,w_t)$ 值

5) if $\Delta i(x_t,w_t)<I_2$

then 定义结点 x_t 为叶结点,置结点 x_t 的左右子树为空,

else

用 w_t 将 x_t 分为 x_{tL} 和 x_{tR}:

$x_{tL}(w)=\{x|x\in x_t$,并且 $w^Tx<0\}$

$x_{tR}(w)=\{x|x\in x_t$,并且 $w^Tx>0\}$

$Q=Q\cup\{tL,tR\}$

对结点 x_t 的左子树递归调用建立二叉树

对结点 x_i 的右子树递归调用建立二叉树

step5 对每个叶结点，计算它的类属性.

step6 返回决策树 T.

4 二元决策树在模型选择中的应用

模型选择实际可以认为是一个模式识别过程，本文以趋势预测模型为例说明遗传算法基于二元决策树的模型选择. 趋势预测模型选择线性模型、指数模型、生长曲线模型与季节性周期模型 4 种常用的模型，各模型的参数都是介于某一区间的随机值. 本文采用随机生成样本的方法生成 12 期观测数据并根据上述 4 种趋势预测模型结构形成训练样本集.

选择最小错误率 I_1、最小分割错误率减少量 I_2 的值分别为 0.05 和 0.03，遗传算法的交叉概率、变异概率、迭代次数 G、最小分类数的值，依据经验分别选定为 0.9、0.3、100、2. 通过遗传算法构造二元决策树如图 1 所示. 决策树中每一个终结点对应一类数据，代表一个趋势预测模型. 经计算决策树中的权值矢量分别为：

$w1 = [-0.056\,4, -0.022\,7, 0.666\,1, 0.420\,9, -0.164\,7, 0.144\,8, -0.284\,2, -0.426\,9, -0.127\,4, -0.165\,4, 0.039\,1, -0.138\,4]$

$w2 = [-0.668\,8, -0.130\,0, 0.162\,4, 0.138\,3, 0.069\,7, 0.255\,8, 0.093\,8, 0.343\,2, 0.369\,9, -0.133\,9, -0.020\,8, -0.371\,8]$

$w3 = [0.050\,9, -0.013\,9, 0.095\,6, -0.145\,9, 0.606\,2, -0.218\,1, -0.005\,8, -0.528\,7, -0.299\,5, -0.025\,4, 0.084\,2, 0.418\,1]$

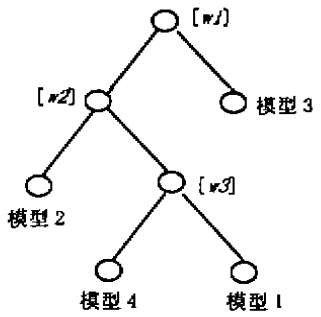

图 1 决策树结构

趋势预测模型结构选择的实验结果如表 1 所示.

表 1 实验结果表明，二元决策树正确完成对趋势预测模型的选择，但二元决策树在模型选择中可能存在错误选择率的问题. 现以股票评估为例说明错误选择率与遗传算法参数之间的关系. 本文选择 159 种股票作为样本，以股票的总资产周转率、销售增长率、净利增长率、总资产增长率、总资产规模、总资产净利率、销售净利率、净资产收益率、股东权益比作为评价标准，评价结果分为绩优、垃圾和普通三种模型.

表 1 实验结果

输入数据												对应模型	正确模型
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	M	M
1.1	2.2	3.2	4.3	5.2	6.6	7.4	8.2	9.1	10	11	12	线性	线性
10	11	12	13	14	15	16	17	18	19	20	21	线性	线性
0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.9	1.7	3.6	9.8	34	指数	指数
0.0	0.0	0.0	0.0	0.0	0.2	0.6	1.3	6.2	17	43	110	指数	指数
1.1	3.4	4.3	3.3	2.0	1.1	2.4	3.3	4.4	3.2	2.2	1.4	生长曲线	生长曲线
0.0	0.8	1.9	1.0	0.0	1.0	2.0	1.0	0.0	0.9	2.1	1.3	生长曲线	生长曲线
0.0	0.0	2.0	4.0	6.3	6.9	7.1	7.3	7.8	7.8	7.9	7.9	季节性周期	季节性周期
5.3	5.3	5.4	15	17	22	44	44	44	44	45	45	季节性周期	季节性周期

参数选择最小错误率 I_1、最小分割错误率减少量 I_2 作为参变量，遗传算法的交叉概率、变异

概率、迭代次数 G、最小分类数作为固定值，分别为 0.9、0.3、100、2，并进行了多次实验，平均结果如表 2 所示. 表 3 为二元决策树与神经网络分别对股票模型分类结果的比较.

表 2　二元决策树分类结果

最小错误率 I_1	最小分割错误率减少量 I_2	错误分类数	错误分类率
0.05	0.03	8.72	5.48%
0.05	0.02	2.84	1.78%
0.05	0.01	0.69	0.43%
0.04	0.03	4.10	2.57%
0.04	0.02	2.14	1.34%
0.04	0.01	0.52	0.33%
0.03	0.02	1.98	1.25%
0.03	0.01	0.45	0.28%
0.02	0.02	1.35	0.85%
0.02	0.01	0.36	0.23%

表 3　二元决策树与神经网络的分类结果比较

	二元决策树	神经网络
错误分类数	0.36	6.84
错误分类率	0.23%	4.3%

从上述结果可以看出，二元决策树比神经网络更具有良好的预测模型选择的能力. 最小错误率 I_1 和最小分割错误率减少量 I_2 直接决定二元决策树的错误分类率，如表 2 中改变最小错误率 I_1 和最小分割错误率减少量 I_2，可以控制二元决策树的错误分类率，使系统成为受控系统，这是神经网络难以实现的.

5　结束语

本文提出了遗传算法基于二元决策树的模型选择方法，并以趋势预测模型为例进行了验证，获得了较好的效果. 遗传算法与神经网络均是具有高度的自组织、自学习能力以及较强鲁棒性的算法. 但神经网络的输出类数等于神经网络的输出层神经元个数，而二元决策树的类由叶结点对应的类属性决定，当二元决策树的结点的错误率大于某一给定值时，可对该结点继续分类，所以二元决策树的叶结点数可能大于分类数，因而分类的准确率较高，并且准确率可以人为控制.

遗传算法在模型选择中的应用是一个新的研究方向，实践证明遗传算法将大大促进模型选择技术的发展和进步，并为模型选择在实践上的应用提供了更为广阔的前景.

参考文献

1　Klein G, Konsynsk B, Bec P O. A liner representation for model management in a DSS. Journal of Management Information System, 1985; 2(2):40～54

2　Klein G. Developing model strings for model management. Journal of Management Information System, 1986;3(2):94～110

3　司　昕. 预测方法中的神经网络模型. 预测,1998; 17(2): 32～34

4　Sohl J E, Venkatkchalam A R. A neural network to forecasting model selection. Informarion & Management, 1995; 29(2): 297～303

5　Tam K Y. Neural network models and the prediction of bank bankruptcy. OMEGA International Journal of Management Science,1991,19(5):429～445

6　Wang J,Malakooti B. A feedforward neural network for multiple criteria decision making. Computer & Operation Research,1992;19(2):151～167

7　杨　璐,黄梯云,洪家荣. 一种基于组合神经网络的时间序列预测方法. 管理工程学报, 1997; 11(1):33～38

8　赵英,赵恒永. 用"基因算法"调节人工神经网络模型. 计算机工程与应用, 1997; 33(6):35～ 38

9　鲍立威. 关于 BP 模型缺陷的讨论. 模式识别与人工智能, 1995; 8(1):34～38

10　陈恩红,王清毅,蔡庆生. 基于决策树学习中的测试生成及连续属性的离散化. 计算机研究与发展,1998; 35(5): 403～407

(下转第 70 页)

4 Toft K B, Prucyk B. Option on leveraged equity: theory and empirical tests. The Journal of Finance. 1997; 52(3): 475～486
5 Leland H E. Corporate debt value, bond covenants, and optimal capital structure. The Journal of Finance. 1994; 49(4): 874～892
6 Leland H E, Toft K B. Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. The Journal of Finance, 1996; 51(3): 576～582
7 Harris M, Raviv A. The theory of capital structure. The Journal of Finance. 1991; 46(1): 297～355
8 Horne J C V, Wachowicz J M. Fundamentals of financial management. New York: Prentice Hall Inc., 1998
9 Sharpe W F, Alexander G J, Bailey J V. Investments. New York: Prentice Hall Inc., 1995
10 余绪缨. 企业理财学. 沈阳: 辽宁人民出版社, 1996
11 汪 平, 刘兴云. 公司理财原理. 上海: 上海财经大学出版社, 1997
12 陈小悦. 资本结构与企业价值. 国际会计与财务研究论丛, 1997. 46～55

Black-Scholes Model Oriented Capital Structure Models

Yang Baochen, Liu Zheng, Zhang Tong
School of Management, Tianjin University

Abstract This paper studies the capital structure in consideration of agency costs, by using Black-Scholes option pricing model, it proposes the corporate valuation model and optimal capital structure model. At last, it gives an numerical example to prove its usefulness.

Keywords: Black-Scholes option pricing model, capital structure, agency costs

(上接第 61 页)

11 肖 勇, 陈意云. 用遗传算法构造决策树. 计算机研究与发展, 1998; 35(1): 49～52
12 Chai Bingbing, Huang Tong, Zhuang Xinhua, Sklansky Jack. Piecewise linear classifiers using binary tree structure and genetic algorithm. Pattern Recognition, 1996; 20(11): 1905～1917

The Application of Genetic Algorithm in Model Selection

Wu Fei, Huang Tiyun
Harbin Institute of Technology

Abstract Model selection plays an important role in model management. Genetic algorithm is employed to construct binary decision tree which is used to select model in this paper. The application sample for trend forecast model selection is given and the factors that influence the inaccurate ration of binary decision tree is discussed.

Keywords: binary decision tree, model selection, genetic algorithm