基于神经网络和决策树相结合的信用风险评估模型研究

赵静娴[1,2]，杜子平[2]
(1.天津大学 管理学院，天津 300072；2.天津科技大学 经管学院，天津 300222)

摘 要：文章提出了一种将神经网络和决策树相结合的信用风险评估模型NN-DT。该方法依据属性重要性将贷款企业的财务指标进行排序，然后通过RBF神经网络进行属性裁减生成决策树，从而得出企业是否违约的分类。最后以判别分析以及C4.5算法为参照方法进行了实证研究，结果表明，NN-DT模型显著地提高了预测精度。

关键词：决策树；RBF神经网络；信用风险
中图分类号：F830.5；F224 **文献标识码**：A **文章编号**：1009-3370(2009)01-0076-04

一、前 言

20世纪80年代以来，随着金融全球化趋势及金融市场的波动性加剧，各国银行和投资者受到了前所未有的信用风险的挑战。世界银行对全球银行业危机的研究表明，导致银行破产的主要原因就是信用风险，因此国际金融界对信用风险的关注日益加强。目前我国商业银行信用风险管理中，信用风险的测量与评估是最为重要和关键的环节。目的在于通过对企业目前的财务、管理、发展等方面的分析来判断企业违约可能，为贷款提供决策支持。在西方国家和新型经济国家，有关这个领域的研究受到诸多学者的重视。目前我国商业银行仍然是企业融资的主要金融中介，因此面对经济调整和企业经营风险的现状，商业银行如何有效地防范以债务方违约为主要特征的信用风险，如何评估信用风险就显得尤为重要。

当前被广泛应用于信用风险研究的模型主要有统计模型和人工智能模型两大类。传统的统计模型包括多元判别分析模型和对数回归模型。统计模型最大的优点在于其具有明显的解释性，存在的缺陷在于过于严格的前提条件，如判别分析模型要求数据分布服从多元正态分布、同协方差。但就信用风险分析数据而言，无论直观判断还是对已有数据分析，结果均表明企业样本不满足统计假设。

随着信息技术的发展，人工智能和机器学习的一些分类和预测的算法也被引入到金融信用风险评估领域中来[1,2]，主要包括人工神经网络和决策树等方法。

人工神经网络是一种高效分类器，不需要先验知识，具有良好的容错性、自适应性和很强的泛化功能。但是人工神经网络建模方法复杂，内部规则可理解性差，不易从中提取规则。

决策树是基于统计理论的非参数识别技术，通过计算机实现决策树算法将统计分析和计算机运算相结合不仅保持了多元参数、非参数统计的优点，而且可以自动进行变量选择，降低维数，分类结果表达形式简单易懂，并可有效地用于对数据的处理。但决策树方法对具有某些特征的条件属性特别敏感，这些属性的存在严重影响了决策树的分类精度及所产生的评估规则的简明性，剪枝将大大增加运算量，同时也不能从根本上解决根结点选择不当的问题。现有的将决策树和神经网络相结合的方法大多着眼于用决策树模拟出神经网络的内部规则，方法较为复杂[3,4]。

近几年，随着机器学习理论的不断发展，基于支持向量机(Support Vector Machine，SVM)的一种专门针对小样本学习的算法被引入到了信用风险评估中。姚奕和叶中行(2004)利用SVM研究银行客户信用评估系统[5]。沈翠华和高万林(2004)利用SVM对企业信用等级进行分析[6]。SVM具有分类面简单、泛化能力强、拟合精度高等特点。但是由于SVM的主要思想是以建立一个超平面作为决策曲面，使得正例和反例之间的隔离边缘最大化。这就要求在机器学习过程中两类样本数量比较接近[7]。在信用风险评估的实证分析中要求“违约企业”和“非违约企业”数

收稿日期：2007-12-25
基金项目：国家自然科学基金(70671074)；天津科技大学科研基金(20080303)
作者简介：赵静娴(1981—)，女，天津大学研究生，天津科技大学教师。E-mail：nzjx2005@163.com

量比较接近。这和我国贷款企业总的违约情况是有相当大的差距的,因此在运用 SVM 算法建模抽取数据的过程就会有一定的主观性和局限性。

针对上述情况，本文提出了一种将神经网络和决策树相结合的信用风险评估模型 NN-DT,经过实证分析,取得了较好的效果。

二、NN-DT 模型的建立

1.决策树[8,9]

决策树是应用最广的归纳推理的算法之一,是一个类似于流程图的树结构，其中每个内部结点表示在一个属性熵的测试，每个分支代表一个测试输出,而每个树叶节点代表类或类分布。决策树通过把实例从根节点排列到某个叶子节点来分类实例,叶子节点即为实例所属的分类，树上每个节点说明了对实例的某个属性的测试，节点的每个后继分支对应于该属性的一个可能值。

决策树算法包括 ID3、C4.5、SLIQ、SPRINT 等。决策树算法的关键是属性的选择标准。属性选择依赖于各种对例子子集的不纯度度量方法。不纯度度量方法包括信息增益、信息增益比、正规增益等多种方法。假设 S 为样例集，目标属性具有 n 个不同的值,则对 S 分类所需的期望信息为

$$Entropy(s)=\sum_{i=1}^{n}-p_i\log_2 p_i \qquad (1)$$

其中,p_i 是 S 中属于类别 i 的比例。

一个属性 A 相对于样例集合 S 的信息增益 Gain(S,A)定义为

$$Gain(S,A)=Eneropy(S)-\sum_{v\in Value(A)}\frac{|S_v|}{|S|}Entropy(S_v) \qquad (2)$$

其中,Value(A)为属性 A 的取值集合；

v 为 Value(A)中的某个值；

|S|为样例总数；

$|S_v|$为属性 A 取值为 v 的样例数。

信息增益标准有一种偏好较细划分的倾向。因此,无论测试属性在实际意义上是否对数据分类最有意义,只要其划分数据的类别最多,在信息增益标准下它将是所选属性。通过对在信息增益的基础上加入了一个被称作分裂信息(Split Information,SI)的项来惩罚分类过细的属性,可以克服信息增益标准的缺点。

信息增益率 GainRatio(S,A)被定义为

$$GainRatio(S,A)=\frac{Gain(S,A)}{SI} \qquad (3)$$

$$SI=-\sum_{i=1}^{n}\frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|} \qquad (4)$$

其中,S_1 到 S_n 是 n 个值的属性分割了 S 而形成的 n 个样例子集。

当$|S_i|\approx|S|$时,信息增益率会非常大,因此信息增益标准有倾向不均匀分布的偏好。

由于训练例子集中的噪音、错误项,或干扰属性的影响，以此训练集生成的决策树常常包含了这些错误的信息,它能够正确分类训练集中的数据,但在分类测试例子集时精度不高。而且据此生成的决策树规模较大。这种现象被称为过度拟合(Overfitting)。通常要在生成决策树之后运用剪枝技术来处理这类过度拟合问题。

2.RBF 神经网络[10]

径向基神经网络(RBF Neural Networks)是一种三层前向网络。第一层输入层由信号源结点组成。第二层为隐含层，单元数视所描述问题的需要而定。第三层为输出层,它对输入模式的作用做出响应。

从隐含层输出至输出空间的变换是线性的。隐含单元的变换函数是 RBF 函数,它是一种局部分布的中心对称衰减的非负非线性函数。这样网络的权就可以由线性方程组直接解出或用递推最小二乘方法递推计算,从而大大加快学习速度,避免局部极小点问题。

3.NN-DT 模型

为避免噪音和干扰属性对数据分类的影响,本文提出在建立决策树之前先对属性进行重要性排序,再利用神经网络不需先验知识的“黑箱”分类特点,及其分类效能高的优势,对属性进行裁减,选择出对数据分类最有效的若干属性建立决策树，并抽取出规则。

(1)以正规增益 NG 为评价标准对属性进行重要性排序。

其中

$$GR(A,S)=\frac{IG(A,S)}{\sum_{i=1}^{n}\frac{|S_i|}{|S|}\log\frac{|S|}{|S_i|}}$$

(2)用 RBF 神经网络对其中最重要的若干属性进行训练并检验其预测精度，然后按属性重要性次序向两端分别加减一个邻近的属性再进行训练及检验并和原检验结果比较，如此反复直到找到分类效果最佳的 m 个属性为止。

(3)以 NG 值最大的属性作为根节点的测试属性，对属性的每个值创建分枝，并且据此划分样本。

(4)在各内节点计算剩余属性的 NG,选择 NG 最大的属性作为此分枝的下一个测试属性。

(5)重复第四步直到结点属性各分支下的训练样例同属一类或所有属性均已用过为止生成决策树。当某分支下的训练样例分属不同分类,而所有属性均已用过时,可将包含该结点训练样例最多的分类作为此分支的分类。

三、实证分析

1.样本及变量选择

本文从某银行的信息系统中随机抽取了某行业(2005 年)118 个贷款企业的资料作为建模样本,其中有 102 个企业的财务数据资料完整,可以作为分析研究的对象。在这 102 个企业中有 81 个企业贷款履约,21 个企业贷款违约,不良贷款率约为 20%,与该行业的整体贷款不良率非常接近,因此,随机抽取的 102 个企业财务资料具有代表性,可以用此样本对总体进行统计推断。

参照工商银行信用等级评定指标及可获得企业数据指标的条件,我们选取了 X_1~X_{10}10 个变量为基本变量,这些变量反映了企业偿债、盈利、管理、成长能力等方面的情况,分别为:

X_1 速动比率;

X_2 利息保障倍数;

X_3 资产负债;

X_4 销售收入利润率;

X_5 资本收益率;

X_6 每股收益;

X_7 总资产周转率;

X_8 存货周转率;

X_9 资本积累率;

X_{10} 营业增长率。

2.交叉有效性估测

由于本文的数据样本数量较小,为充分利用样本所提供的信息能够得到较好结果,采用交叉有效性数据估测方法:

将整个样本集合随机分为 N 个含有大致相同数量样本的子集。取第 i 个集合用于检验,其余的 N-1 个集合作为第 i 个训练样本集,因此将有 N 个数据集合可用于构建模型。本文取 N=4,并利用错判率最小原则选择最佳模型。

3.结果分析

运用 NN-DT 模型,在交叉有效性估计下本样本空间的最优树如图 1 所示。

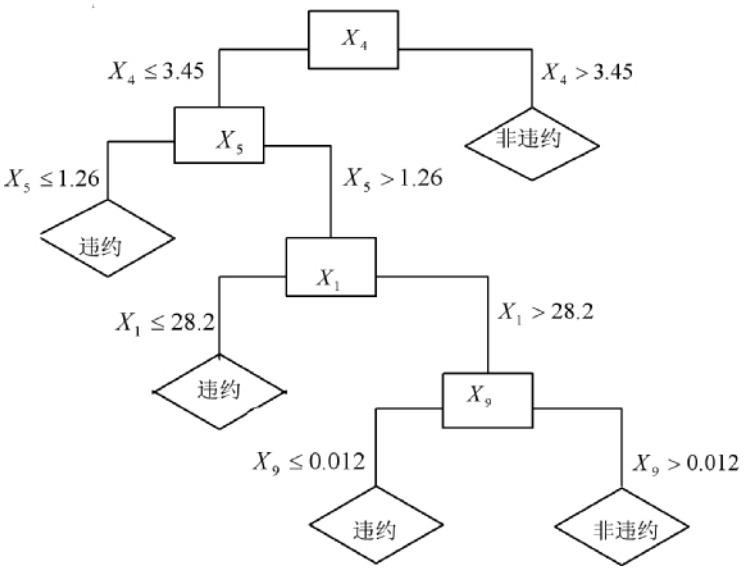

图 1 决策树模型

表 1 错判比率对照

	第一类错误比率	第二类错误比率	总错判比率
C4.5	10.7%	11.3%	11.18%
LDA	16.2%	12.7%	13.2%
NN-DT	5.3%	6.8%	6.5%

同时本文运用线性判别模型 LDA 以及决策树中 C4.5 算法对样例的计算精度与本文提出的 NN-DT 计算精度进行了对比,如表 1 所示。

其中第一类错误是指将“违约”企业评判为“履约”企业。第二类错误是指将“履约”企业评判为“违约”企业。从直观上可以清楚地看到,银行误贷款给一个将会违约的企业所带来的损失是远远大于拒绝贷给一个将会履约的企业的损失。因此我们可以从两个方面来评价模型,一是整体的错误率;二是第一类错误率。通过表 1 可以看出,无论针对总体,还是第一类错误,NN-DT 模型都较另外两种方法有更高的预测精度。

四、总 结

本文提出的 NN-DT 算法对数据分布无要求,十分适用于银行信用风险评估。同时它结合神经网络和决策树算法的优势,可以有效选择出对信用评估最重要的若干属性,避免了有害冗余属性的干扰和繁琐的决策树剪枝计算,生成的信用评价规则相对于直接运用决策树更简明。而且通过仿真试验表明,NN-DT 算法较 LDA 和 C4.5 算法在预测精度上也有明显的提高。

参考文献：

[1] Altmen E, Fraydman H, Kao E D. Introducing recursive partitioning for financial classification: the case of financial distress[J]. Banking and Finance,1985,11(2):269-291.

[2] H L Jensen. Using neural networks for credit scoring[J]. Managerial Finance,1992,18(6):15-26.

[3] Schmitz G P J, Aldrich C, Gouws F S. A NN-DT: an algorithm for extraction of decision trees from artificial neural networks[J]. IEEE Trans on Neural Networks,1999, 10(6):1392-1401.

[4] Qiangfu Zhao. Evolutionary design of neural network tree-integration of decision tree, neural network and GA [J]. Proceedings of the 2001 Congress on Evolutionary Computation, 2001,1:240-244.

[5] 姚奕,叶中行.基于支持向量机的银行客户信用评估系统研究[J].系统仿真学报,2004,16(4):783-786.

[6] 沈翠华,高万林.基于支持向量机的企业信用评估模型[J].管理信息化,2004: 73-74.

[7] 徐晓霞,李金林. 基于决策树法的我国商业银行信用风险评估模型研究[J].北京理工大学学报(社会科学版), 2006(6):71-74.

[8] Haixun Wang, Yu P S. SSDT: a scalable subspace-splitting classifier for biased data [J]. ICDM 2001 Proceedings, IEEE International Conference Proceedings, 2001,29 Nov.-Dec.:542-549.

[9] Tom M Mitchell Mitchell, Tom M (Tom Michael).Machine learning [M].McGraw-Hill Education (Asia) , 2003:42-47.

[10] M D Powell. Radial basis is functions for multivarialte interpolation: a review in algorithms for the approximation of functions and data[J]. Oxford: Clarendon Press, 1987.

A Model Based on Combination of Neural Network and Decision Tree for Credit Risk Assessment

ZHAO Jing-xian[1,2], DU Zi-ping[2]

(1.School of Management, Tianjin University, Tianjin 300072;

2.School of Economics and Management, Tianjin University of Science & Technology, Tianjin 300222)

Abstract: The paper builds a credit risk assessment model based on combination of neural network and decision tree. The method ranks financial data of loan enterprise based on the importance of the attributes, prunes the attributes using RBF neural network and builds a decision tree. Therefore conclusions of classification can be drawn as to whether or not enterprises break a contract. Finally comparing NN-DT with methods of discriminate analysis and C4.5 algorithm, studies are empirically carried out, and NN-DT largely improves the prediction precision.

Key words: decision tree; RBF neural network; credit risk

[责任编辑:孟青]