基于决策树法的我国商业银行信用风险评估模型研究

徐晓霞, 李金林
(北京理工大学 管理与经济学院, 北京 100081)

摘 要:结合我国商业银行的实际, 基于数据挖掘中决策树 C4.5 算法的分析框架建立了商业银行的信用风险评估模型,通过此模型可以根据贷款企业的财务指标,得出企业是否违约的分类。此分类将对商业银行信用风险控制工作具有很好的指导意义。还通过案例进行了实际的风险评估分析。

关键字:信用风险; 决策树; C4.5 算法
中图分类号 F224.9 文献标识码:A 文章编号:1009-3370(2006)03-0071-04

一、引 言

国有商业银行是中国金融体系的主要支柱,是国民经济投融资体系中最重要的组成部分,因此,国有商业银行的稳健、安全运行历来是管理部门关注的焦点,也是理论和实务界研究的热点。世界银行对全球银行业危机的研究表明, 导致银行破产的最常见原因就是信用风险。目前,贷款业务仍是我国商业银行最重要的资产业务,因此,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。

对信用风险评估方法的探索可以追溯到 20 世纪 30 年代,大致经历了比例分析、统计分析和人工智能 3 个发展阶段。在传统的比例分析方法出现以后,基于统计判别分析的模型都是在 Fisher 于 1936 年做出的启发性研究之后提出的[1],其中较为经典的方法包括 Altman 的 Z-score 模型以及在此基础上改进的 ZETA 模型[2]。统计分析法的引入克服了传统比例分析法综合分析能力差、定量分析不足等缺点,但也存在一些问题: 由于统计方法研究的是样本趋于无穷大时的渐进理论, 故要求样本数据有一定的规模; 方法的可用性与建立分类模型时所需的多个假设紧密相关, 如多元判别分析法就要求数据服从多元正态分布, 而现实中大量数据是不符合这些假定的, 因此统计方法在现实应用中很难达到满意的效果。20 世纪 80 年代以来,人工智能的技术如专家系统、神经网络等被引入信用风险评估中,克服了统计方法对假设要求强的缺点,尤其是神经网络法,不仅具有非线性映射能力和泛化能力, 而且还具有较强的鲁棒性和较高的预测精度。但是神经网络法有其自身的缺陷:网络结构难以确定;训练时容易陷入局部极值,训练效率不高。

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。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 2 类样本数量比较接近, 在信用风险评估的实证分析中即要求“履约企业”和‘违约企业’数量比较接近。这和我国贷款企业总的违约情况是有一定差距的, 银监会 2005 年 5 月 16 日公布的最新统计数字显示, 一季度末,全部商业银行不良贷款率为 12.4%,这样在运用 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,本文引入了数据挖掘中的一种基于决策树的 C4.5 算法,并将其用于商业银行信用风险评估中,经过实证分析,取得了较好的效果。

二、决策树学习算法

1.决策树分类器

决策树方法 20 世纪 60 年代起源于对概念学习建模;20 世纪 70 年代后期 Quinlan 发明用信息增益作为启发策略的 ID3 算法[3],从样本中学习构造专家系统;1993 年 Quinlan 在 ID3 算法基础上研究出了改进的决策树归纳包(C4.5),是目前被普遍采用的数据分类方法。本文即引用了 C4.5 算法对我国商业银行贷款企业的财务数据进行实证分析建立企业是否违约的分类模型, 并应用于对企业信用风险的评估和预测。分类通常是被认为是把一组事物分成若干子集合, 而子集合内的成员相互之间比其它成员之

收稿日期:2005-09-02
作者简介:徐晓霞(1979—),女,硕士研究生,研究方向为系统工程,E-mail: xxxia8008@bit.edu.cn

间具有更大的"相似性",而这一任务的实现是通过分类模型来完成的,在对已有数据学习的基础上构造出一个分类函数或一个分类模型,即分类器。既可以用此模型分析已有的数据,也可以用它进行预测。该函数或模型能够把数据库中的数据记录映射到给定类别中的某一个,从而可以应用于数据预测。构造分类器需要根据给定样本数据集作为输入。数据集由一组数据库记录构成,每个记录是一个由有关字段值组成的特征向量,我们把这些字段称作属性,把用于分类的属性叫做类标签,也就是样本集的类别标记。如$(X_1, X_2, K, X_n; C)$,其中 X_i 表示字段值;C 表示类别。图1是通用的 Top-Down 决策树构建算法:

```
输入:节点 n,数据集 D,分割算法(Classification Algorithm,CL)
输出:以节点 n 为根节点的基于数据集 D、分割方法 CL 的决策树
procedure BuildTree(n,D,CL)
    初始化树的根节点;
    在 D 中计算 CL 来求解节点 n 的分割标准;
    if(节点 n 满足分割条件)
      选择最好的效果将 D 分为 D1、D2;
      创建节点 n 的子节点 n1、n2;
      Tree(n1,D1,CL);
      Tree(n2,D2,CL);
    endif
  end
```

图1 Top-Down 决策树构建算法

由算法可知,分割算法 CL 是决策树算法的关键。根据分割方法的不同,目前决策树算法可分为2类:基于信息论的方法和最小 GINI 指标方法。对应前者的算法有 ID3、C4.5,对应后者的有 CART、SLIQ、SPRINT。

2.基于信息论的 C4.5 算法

研究表明,一般情况下,树越小则树的预测能力越强,要构造尽可能小的决策树,关键在于选择恰当属性。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、正交法等多种方法。C4.5 算法使用信息增益比作为启发规则来构造决策树(同样一组样本数据,可以有很多决策树能与之符合),以信息增益比最大为原则来选择好的属性。

(1)信息熵 信息熵(单位为比特)采用公式为

$$\mathrm{Info}(S)=-\sum_{i=1}^{c} p_i \log_2(p_i)$$

S 为样本集;c 是样本分类的个数;p_i 是样本中第 i 类的概率。

(2)信息增益 属性的信息增益是指该属性分割后熵的消减期望值

$$\mathrm{Gain}(S,A)=\mathrm{Info}(S)-\sum_{V\in \mathrm{Values}(A)} \frac{|S_v|}{|S|}\mathrm{Info}(S_v)$$

$|\cdot|$为集合"·"中元素的个数;S 为样本集;A 为样本中的某一属性;S_v 为属性 A=v 时样本对应的样本集;$v\in \mathrm{Values}(A)$为 v 属于 A 的某一个值。

(3)信息增益比

$$\mathrm{GainRatio}(S,A)=\mathrm{Gain}(S,A)/\mathrm{Splitinfo}(S,A)$$

其中,$\mathrm{SplitInfo}(S,A)=-\sum_{v\in \mathrm{Values}(A)}\left(\frac{|S_v|}{|S|}\log_2\frac{|S_v|}{|S|}\right)$

3.K 次迭代交叉验证

一般需要对通过学习数据集来构造的决策树进行修剪,使其具有更好的泛化能力,而对子树是否需要修剪,可使用交叉检验、统计测试或最小描述长度等方法进行判别。

C4.5 算法运用了交叉验证法对学习的结果进行验证。K 次迭代交叉验证是将数据集分成 K 个没有交集的子集,K 个子集中一个用作测试集,而其余的 K-1 个数据集作为训练集,最后对 K 个数据子集的错误数计算平均值,K 次迭代验证是对监督学习算法的结果进行评估的方法。经过大量的统计试验,K 一般取10效果比较好。这种验证方法在小数据集的情况下尤其适用。

三、基于决策树 C4.5 算法的我国商业银行信用风险评估模型实证分析

1.实证分析的样本来源

本文从某银行的信息系统中随机抽取了某行业(2004年)100个贷款企业的资料作为建模样本,其中有81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81个企业中有64个企业贷款履约,17个企业贷款违约,不良贷款率为20.1%,与该行业的整体贷款不良率非常接近,因此,随机抽取的81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。

2.指标体系的建立

适当地选择财务指标建立反映企业信用风险的指标体系,是信用风险评估的基础。

依据全面性、有效性和可操作性的原则,选择了5个方面的9项指标构建了信用风险评估指标体系。

(1)负债水平 资产负债率 X_1。适度的资产负债率表明企业投资人、债权人的投资风险较小,企业经营安全稳健,具有较强的筹资能力。

(2)流动能力和偿债能力 流动比率 X_2、总债

务/ebitdaX_3。流动比率越高,表明企业流动资产周转越快,偿还流动负债能力越强。但需要说明的是,该指标过高,表明企业的资金利用效率比较低下,对企业经营发展不利。总债务/ebitda是指总债务相对于当年的息、税、折摊前收益的大小。总债务与ebitda的比率反映以企业所创造的税前利润和留在企业内部的固定资产折旧费用、摊销费用在支付利息前对总债务的保障能力。该指标越小,企业还债能力越强,反之,企业还债的能力就比较弱。

(3)赢利能力　净资产收益率X_4、销售(营业)利润率X_5。企业获利能力是企业信用的基础,企业只有盈利,才有可能按时偿还债务。

(4)经营效益和资金利用效率　总资产周转率X_6、流动资产周转率X_7。企业的资产管理状况对于企业的信用风险水平有直接的影响。

(5)发展能力　销售(营业)增长率X_8、资本积累率X_9。销售(营业)增长率越大,表明销售(营业)收入增长速度越快,企业发展形势较好,企业的信用风险较小。资本积累率展示了企业的发展潜力。该指标值越高,说明企业的资本积累越多,投资者投入企业资本的保全性和增长性越强,企业应付风险的能力越强,企业的信用风险相对较小。

上述的9项指标即模型中的9项属性。

3.模型的构造和检验

在此模型中类标签为C:0代表违约,1代表履约

运用数据挖掘软件WEKA3.4(主要支持的数据挖掘任务是分类和总结),根据C4.5算法对得到的容量为81的样本数据建立模型,即以信息增益比最大为原则选取节点生成决策树,如图2所示。

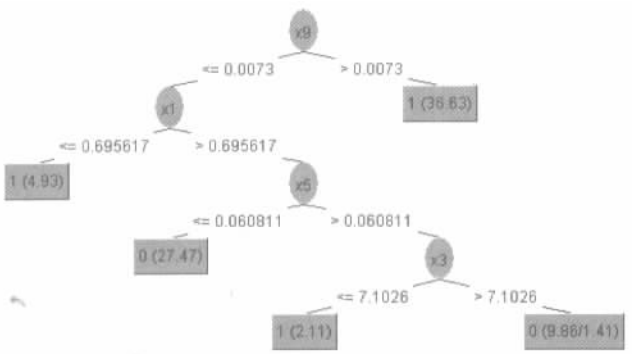

图2　基于C4.5算法的决策树模型

最终选出了4个好的属性作为决策树节点,x_9的信息增益比最大,因此为根节点,也就是说在这9个财务指标中资本积累率对于违约和履约分类的贡献程度是最大的。括号中的数代表平均有几个样本数据根据条件被分到了相应的类中,如:如果x_9>0.0073,平均有36.63个样本数据被正确地分到了1-履约类中;自上向下逐层通过测试条件到达x_3节点,如果x_3>7.1026,平均有9.86个样本数据被分到了0-违约类中,但其中有平均1.41个样本数据分类错误。

用10次迭代交叉验证法来验证模型的误差率,验证结果如下:

正确分类的样本数	72	88.8889 %
错误分类的样本数	9	11.1111 %

=== 混淆矩阵 (Confusion Matrix)===

a	b	<-- 分类
14	3	a = 0
6	58	b = 1

总体来看,分类正确的数据有72个,错误的9个,错误率是11.11 %。对于17个违约数据3个分类错误,即3个企业被误分到了履约企业中;64个履约数据6个分类错误,即有6个履约企业被误分到了违约企业的类中。这里截取了交叉验证结果的一部分:

=== 10次迭代交叉验证 ===

样本	真实值	预测值	错误	概率分布
1	2:1	2:1		0.098 *0.902
2	2:1	2:1		0.098 *0.902
3	2:1	2:1		0.098 *0.902
4	2:1	2:1		0 *1
5	2:1	2:1		0.098 *0.902
6	2:1	2:1		0.098 *0.902
7	1:0	2:1	+	0.098 *0.902

……

可以看到,前6个数据分类正确。第7个数据分类错误,实际属于第一类,属性值为0-违约,但是模型预测分到了第二类,属性值为1-履约。在决策树模型的验证结果中还给出了概率的分布情况,对于第7个数据,由决策树模型的判断该数据属于第一类的概率为9.8%,属于第二类的概率为90.2%,因此将此数据分到了第二类。在模型结果中,给出了分类的概率分布是有很大的实际意义的,这对于已知企业财务状况,预测企业是否违约以及违约的概率是至关重要的。

4.结果分析

由于数据较少,运用C4.5算法采用了交叉验证的方法来检测模型,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,Altman指出第一类错误的损失为第二类错误损失的20~60倍,从直观上也可以清楚的看到,银行误贷款给了一个将会违约的企业由

此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2个方面来评价模型,一是整体的错误率,二是第一类错误率。运用决策树对这81个数据建模并评估,整体的错误率是11.11%,第一类的错误率相对较高,这和数据量少也是有一定的关系。为了降低第一类的错误率,在C4.5算法的基础上又加入了附加算法,引入了惩罚函数,给第一类错误加以了适当的惩罚系数,得到17个违约企业只有2个分类错误,有效的降低了第一类错误率,而整体错误率没有太大变化,也就是说第二类错误率有所上升,由于第一类错误带来的损失是巨大的,因此在整体错误率变化不大或是可接受的情况下,我们尽可能地追求第一类错误率最小。

对于最新引入信用风险评估的SVM算法,虽然效果不错,但是SVM算法要求数据是类别分布均匀的数据,而对于信用风险评估,这一点是不太容易满足的,对于一组新的企业财务数据,根据已有的模型来分析将来是否会违约,我们是无法确定所选出的这些企业违约和履约是分布均匀的。而采用本文引入的C4.5算法,不但克服了这方面的缺陷,而且执行效果还不错。

四、结束语

决策树C4.5算法是基于信息熵理论的有效的分类算法。它运用了交叉确认的模型验证方法,对于我国商业银行信用评估样本数据少是非常适用的。决策树C4.5算法对数据分布无任何要求,应用于商业银行信用风险评估的效果也比较好,因此具有良好的发展前景,值得我们深入研究,以后的工作将从以下两个方面展开:①在本文的研究中,C4.5算法只是分析了"违约"和"履约"的两类分类问题,可以将此方法推广到更为复杂的多类信用评级问题上,以便更好的反映借款人的信用状况,为我国商业银行的贷款决策提供更有力的工具。②C4.5算法目前只适用于根据企业财务数据进行静态的企业信用等级的分类。对于我国商业银行信用风险的控制工作,由于企业贷款周期长,企业的财务报表中的各指标是动态变化的,这就要求有一个动态模型与之对应。开发出动态的C4.5算法是研究的新方向。

参考文献:

[1] Altman EI. Corporate financial distress: a complete guide to predicting, Avoiding, and dealing with bankruptcy[M]. New York: John Wiley & Sons, 1983.

[2] Hellmann T, Stiglitz J. Credit and equity rationing in markets with adverse selection [J]. European Economic Review, 2000, 44: 281-304.

[3] Qualian J R. Programs for machine learning [J]. SanMateo,CA:MorganKaufmann Publishers,1993.

[4] 科罗赫, 加莱, 马克. 风险管理[M]. 北京: 中国财政经济出版社, 2005.

[5] Mehmed Kantardzic. 数据挖掘——概念、模型、方法和算法[M]. 北京: 清华大学出版社, 2003.

[6] Witten I H. Data mining:practical learning tools and techniques with java implementations[M]. China Machine Press, 2005.

[7] 肖北溟. 国有商业银行信贷风险管理研究:博士论文[D]. 北京: 北京理工大学, 2005.

[8] 刘云焘,吴冲,王敏,等. 基于支持向量机的商业银行信用风险评估模型研究[J]. 预测, 2005(1):52-55.

[9] 张剑飞. 数据挖掘中决策树分类方法研究[J]. 长春师范学院学报, 2005(1):96-98.

A Model Based on Decision Tree for Credit Risk Assessment in Commercial Banks

XU Xiao-xia, LI Jin-lin

(School of Economics and Management Beijing Institute of Technology, Beijing 100081)

Abstract: Based on the reality of commercial banks in China at present, a credit risk assessment model is built based on the analysis frame of decision tree C4.5 algorithm. Conclusions of classification can be drawn as to whether or not enterprises break a contract according to the financial data of loan enterprises through this model. This classification will be of great significance to commercial banks' management of controlling credit risks, and examples are provided to illustrate the theory of this mechanism.

Key words: credit risk; decision tree; C4.5 algorithm

[责任编辑: 孟青]