DOI: 10.16661/j.cnki.1672-3791.2018.36.018

基于决策树算法的信贷风险评估模型

陆一
(湖北工业大学计算机学院 湖北武汉 430068)

摘 要: 农村商业银行控制运营成本、提升经济效益的重要手段是信贷风险管理,但是银行每天都需要处理大量的信贷业务。本文针对农村商业银行信贷业务中风险较高等问题,设计了一种基于决策树算法的信贷风险评估模型。该模型具有较高的准确率,为银行信贷风险评估提供重要决策依据。

关键词: 决策树 银行信贷 风险评估
中图分类号: F304.4 **文献标识码:** A **文章编号:** 1672-3791(2018)12(c)-0018-02

随着近年来国家对中小金融企业发展的支持,农村商业银行的信贷业务的种类也日益丰富,信贷业务更加复杂,信贷风险管理系统有了更高的要求,因此只有通过先进的管理工具和途径、统一的信息化管理技术,这样才能对信贷业务实行科学、规范化的管理,进一步实现对信贷资产的有效的风险控制和有效监管。

就浙江省来看,日前浙江省农村信用社下的各农村商业银行(或正处于改革下的农村合作银行和农村信用社)都是改革的攻坚阶段,随着经营规模的不断扩大及其信贷业务种类的增加,信贷风险管理的难度也必然会加大,这更要依赖先进的管理工具、统一的信息化技术,这样才能科学规范化地管理信贷业务的过程。为了最大化地实现信息的共享,加强农村商业银行系统内的信息规范化和数据的管理,构建更为安全、有效、规范的信贷风险管理系统成为必然趋势,这样才能满足新形势下应对农村商业银行经营的风险并且不断扩大业务规模,才能可持续地发展农村商业银行。

1 决策树算法介绍

C4.5算法是决策树生成的一种十分经典的算法,这个算法是对ID3算法的优化。针对ID3算法C4.5做的主要改进有以下几点:(1)通过信息增益来决定属性分裂的值。(2)能够处理处理连续性还有离散型数据。(3)把决策树构造完成后能够进行剪枝,简化决策树。(4)能够处理有缺失的样本数据。

在C4.5算法中,只通过属性的信息增长率选择分裂属性的。如公示(1)所示,判断属性的分裂值即split information:

$$SplitInfomation_A(s) = -\sum_{j=1}^{m} \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|} \quad (1)$$

在公式中,训练集S通过训练数学A的划分为m个子集,$|S_j|$表示在第j个子数据集中样本的大小,$|S|$表示总样本大小。

通过计算属性A样本分裂之后的增益信息,来判断该值是否为最佳分裂属性,增益信息计算公式如公式(2)所示,InfoGain值的计算:

$$InfoGain(S, A) = E(S) - E_A(S) \quad (2)$$

在式(2)中,信息增益的计算是计算信息增益率十分重要的一步,E(S)为训练数据集S的熵,$E_A(s)$为属性A分裂后的信息熵。E(S)和$E_A(s)$的计算公式见公式(3):

$$\begin{cases} E(S) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i) \\ E_A(S) = -\sum_{j=1}^{m} \frac{|S_j|}{|S|} E(S_j) \end{cases} \quad (3)$$

2 信贷风险评估模型

2.1 决策属性选择

以桐乡市农村信用合作社个人贷款客户历史数据中,取一小部分数据作为训练集生成决策树,并生成客户信用等级评定模型。指标的合理选取对于模型的有效性有着的影响,为了指标的全面性以及准确性,个人贷款客户信用评估指标体系共分为若干项,经过仔细的调查研究,可以分为年龄、贷款与收入的比值,学历、还贷与收入的比值,是否有违约记录等5个属性值,具体如表1所示。现随机抽取两万名客户的信息进行训练以生成决策树。

年龄属性:是从数据库中的记录的数据获取。

学历属性:也是和年龄属性一样,都是通过数据库的原始数据获取。

贷款金额比:贷款金额比是贷款总金额和年收入的比

表1 决策属性

变量	变量名	计算方法
X1	年龄	根据客户的资料获取
X2	学历	根据客户的资料获取
X3	贷款金额收入比	贷款金额/年收入
X4	还款金额收入比	每月还款金额/(年收入/12)
X5	是否有违约记录	0表示有违约记录,1表示没有违约记录

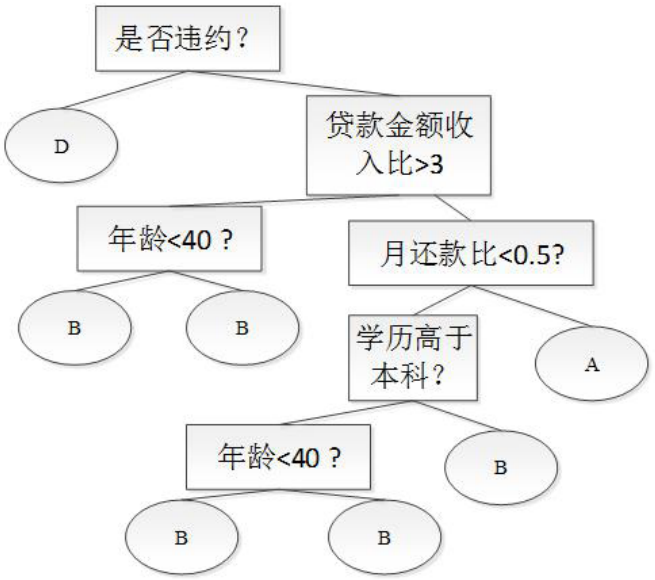

图1 信用模型决策树

值，通过该属性可以判断该客户贷款的金额是否超出能力范围。其计算公式：X3=贷款总金额/年收入。

还款金额比：还款金额比是月还款金额与月收入的比值，这个属性可以判断客户的还款能力。计算公式：X4=月还款金额/（年收入/12）。

是否有违约记录：该属性是判断用户是否有违约记录，具有一票否决权，即表示有任何违约记录，就拒绝贷款。

2.2 决策树剪枝

由于决策树的建立完全是依赖于训练样本，因此该决策树对训练样本能够产生完美的拟合效果。但这样的决策树对于测试样本来说过于庞大而复杂，可能产生较高的分类错误率。这种现象就称为过拟合。因此需要将复杂的决策树进行简化，即去掉一些节点解决过拟合问题，这个过程称为剪枝。

剪枝方法分为预剪枝和后剪枝两大类。预剪枝是在构建决策树的过程中，提前终止决策树的生长，从而避免过多的节点产生。预剪枝方法虽然简单但实用性不强，因为很难精确的判断何时终止树的生长。后剪枝是在决策树构建完成之后，对那些置信度不达标的节点子树用叶子结点代替，该叶子结点的类标号用该节点子树中频率最高的类标记。对于一个叶子节点，这个节点覆盖了N个样本，其中有e个是错误的分类，那么整棵树的错误判断值为，其中0.5为惩罚因子，通过经验判断，惩罚因子一般设定为0.5.那么对于一整颗树的误判率就为：

$$ErrorRation=\frac{\sum_{i=1}^{l}e_i+0.5l}{\sum_{i=1}^{l}n_i} \quad (4)$$

式（4）中，l为所有节点的个数表示该节点中的所有样本，表示节点错误的样本。一旦求得的整个数的误判错误率的均值加上标准差，小于节点的均值误差，那么就可以将子树替换叶子节点，即表示为剪枝。

3 规则描述

根据C4.5决策树模型，对给定的5个属性构建了一颗决策树，其决策树如图1所示，为了更为清晰地了解决策模型，将决策树转变为下列规则。决策树将信用等级分为A、B、C、D这4个等级，其中信用等级为A的表示还款能力强，贷款风险低。D则表示信用等级最低，风险最高。根据决策树模型来看，年龄较小，学历越高则风险越低。年龄较大，学历越低，并且贷款金额和收入比越高风险越大。

4 结语

本文针对农村商业银行信贷业务中风险较高等问题，设计了一种基于决策树算法的信贷风险评估模型。通过决策树模型发现，在农村商业银行信贷业务中，年龄较小，学历越高则风险越低。年龄较大，学历越低，并且贷款金额和收入比越高风险越大。该模型对辅助信贷决策有着重要的作用。

参考文献

[1] 侯斌甲.基于B/S架构的个人信贷系统的研究与实现[D].兰州大学,2016.
[2] 张韶峰.大数据在信贷和保险领域的实践[J].新经济,2016(19):48-50.
[3] 张彦.河南农信社信贷管理系统的设计与实现[D].天津大学,2016.
[4] 曹可雲荻.基于B/S结构的商业银行信贷管理系统的设计与实现[D].吉林大学,2016.
[5] 王宣强.基于REST的银行信贷系统API体系研究[J].金融电子化,2016(3):79-80.
[6] 严晔玮.建设银行宜春分行小企业信贷管理系统的研究与分析[D].云南大学,2016.