

基于粗糙集理论的信贷风险评估模型研究

郑路远

湖南大学数学与计量经济学院, 湖南 长沙 410012

摘要: 信贷风险是现代商业银行所需面对的首要风险, 特别是我国商业银行由于缺乏基础数据, 无法采用国外银行的先进信贷风险评估模型, 长期以来一直使用传统方法进行信贷风险评估, 因此急需探索一个适用于我国国情的信贷风险评估模型。为此, 首先建立一套包含财务指标与非财务指标的信贷风险评估指标体系, 然后根据粗糙集理论能够处理不可区分关系的特点, 结合我国具体国情, 提出了基于粗糙集理论的信贷风险评估模型, 并给出数据预处理、属性简化、决策规则集的生成、对象分类及规则预测精度验证的实现方法。最后以多家公司的信贷情况为测试实例, 采用基于粗糙集理论的信贷风险评估模型进行测试, 测试结果表明, 信贷正常公司的预测准确率达到83.33%, 非正常公司的预测准确率达到100%, 能够为银行的信贷决策提供有效的参考。

关键词: 信贷风险评估; 粗糙集理论; 模型

中图分类号: TN202

文献标识码: A

文章编号: 1000-2324(2016)02-0316-05

Research on Credit Risk Assessment Model Based on Rough Set Theory

ZHENG Lu-yuan

College of Mathematics and Econometrics/Hunan University, Changsha 410012, China

Abstract: Credit risk is the primary risk for which modern commercial banks are facing, especially in our country due to lack of basic data, so commercial banks can't use the foreign advanced credit risk assessment model to have to use the traditional one so as to explore a suitable for China's national conditions of the credit risk assessment model. Aiming at this problem, first of all, a set of financial indicators and non-financial indicators of credit risk assessment index system should be established and then according to the characteristic, which theory of rough set is able to handle indistinguishable relationship, combining with China's specific national conditions, the credit risk assessment model based on rough set theory is put forward and presents a simplified data preprocessing, attribute and decision rule set the generation rules, object classification and prediction accuracy of the implementation of the method. Finally the multiple companies credit conditions are tested for some cases with the credit risk assessment model based on rough set theory. The results show that the prediction accuracy in credit normal companies reaches 83.33%, it is 100% in abnormal companies. Which could provide an effective reference for bank credit decisions.

Keywords: Credit risk assessment; rough set theory; model

近年来的美国股市崩盘、拉美债务危机以及美国“次贷危机”, 引起各国对金融风险管控的高度重视^[1]。目前, 欧美发达国家于2007年开始执行《新巴塞尔资本协议》, 该协议反映了当前银行领域在金融风险管控方面的最新技术和方法, 能够有效的对信贷风险实现管控。而我国与欧美发达国家银行业的信贷风险管控水平相差较大, 因此我国银行业急需进行信贷风险管控理论的研究, 同时借鉴国际银行业的优秀信贷风险管控经验, 全面提高我国银行业的信贷风险管控能力。

1 信贷风险评估指标体系的建立

为了更好的进行信贷风险管控, 建立科学合理的信贷风险评估指标体系, 在参考了国外学者对信贷风险评估指标体系研究成果的基础上^[2], 结合我国具体国情, 选取的指标体系分为财务指标和非财务指标, 财务指标如表1所示, 其值为连续型。从表1中可以看出, 财务指标主要选取了目前企业通用的财务指标, 各财务指标的计算值也按照通用公式进行计算。

表1 财务指标表
Table 1 Financial indicators

偿债能力 Ability to pay debt	营运能力 Ability to operate	盈利能力 Ability to profit	发展能力 Ability to develop	现金流量 Circulated cash
流动比率 C ₁	总资产周转率 C ₅	资产净利率 C ₈	主营业务增长率 C ₁₂	现金流动负债比率 C ₁₅
速动比率 C ₂	流动资产周转率 C ₆	净资产收益率 C ₉	总资产增长率 C ₁₃	现金债务总额比 C ₁₆
资产负债率 C ₃	应收账款周转率 C ₇	销售净利率 C ₁₀	净资产增长率 C ₁₄	主营业务收入现金含量 C ₁₇
利息倍数 C ₄		每股收益 C ₁₁		总资产现金回收率 C ₁₈

收稿日期: 2015-01-06

修回日期: 2015-03-10

作者简介: 郑路远(1994-),男,浙江长兴人,本科,主要研究方向:应用数学. E-mail:zheng@163.com

数字优先出版:2016-04-08 http://www.cnki.net

为了弥补财务指标对企业信贷风险评估的不足,采用了行业发展和企业情况作为非财务指标,其中行业景气指数以 100 为分界,大于 100 说明经济上行,小于 100 说明经济下行;企业情况各指标的取值如表 2 所示。

表 2 企业的非金融指标

Table 2 Non-financial indicators in companies

项目	董事长学历	资产重组类型	有无担保	审计意见类型
Item	Chairman's education	Asset restructuring type	Any guarantee	Audit opinion type
0	高中	无重组	无	
1	大专	资产出售与转让	有	无保留意见
2	本科	资产置换		保留意见
3	硕士	债务重组		拒绝表示意见
4	博士及以上	其它		否定意见

2 基于粗糙集理论的信贷风险评估模型的建立

在不满足统计假设的条件下,采用粗糙集理论产生的决策比较简单,为不准确数据的研究分析及挖掘数据内在联系方面提供了较为有效的方法^[3-5],因此与传统评估方法相比,在信贷风险评估模型中采用粗糙集理论,能够较为准确的进行信贷风险评估。在实际应用中,主要分为数据预处理、属性简化、决策规则集的生成、对象分类及规则预测精度验证五个步骤。

2.1 数据预处理

数据预处理就是对商业银行掌握的信贷主体数据进行数据的正确性及完整性检查,对数据中的噪声进行处理并对连续属性进行离散化,使经过处理的数据满足粗糙集理论的要求,主要分为缺省值的处理和连续属性离散化两步。

缺省值处理:由于商业银行掌握的数据表一般缺项较少,为了不影响数据表中包含的信息,采用 Conditioned Mean Completer 算法,缺项值由与该缺项数据的决策属性值相同的数据项中取均值获得。

连续属性离散化:粗糙集理论要求属性值必须是离散型数据,由于本文选择的财务指标属性值分布较均匀,所以使用等频率算法进行属性值离散。具体为将某一具体属性值由大到小进行排序,然后依据给定的离散数 k ,将 m 个属性值均分为 k 段,各段都包含有 m/k 个属性值,然后得到断点集,就完成了连续属性的离散化。

2.2 属性简化

目前粗糙集属性简化中常用的基于区分函数的简化算法和基于属性重要性的简化算法,在数据较多时,计算量过大,所以本文采用遗传算法来完成属性简化^[6-8]。算法中区分矩阵的一项由候选约简的表示位串来代表,也就是对象的分辨属性集,某位为 1 时代表该属性存在,为 0 时代表该属性不存在,算法的适应函数如式 1 所示:

$$F(v) = \frac{n - L_v}{n} + \frac{C_v}{(m^2 - m)/2} \quad (1)$$

式中, v 表示某分辨属性集的位串; n 表示条件属性的数量,也就是属性集的长度; L_v 表示位串 v 中值为 1 的数量; C_v 表示位串 v 可以区分的对象数量; m 表示训练样本的数量。适应函数包含两部分,第一部分表示希望 L_v 的取值尽量小,后一部分表示希望可以区别的对象尽量多。在进行初始种群的设计时,可将专家或核等必要的属性增加进种群中,以提高算法收敛的速度。算法的具体执行步骤如下:

- ① 计算条件属性 C 对于决策属性 D 的依赖度;
- ② 设 $core(C) = \Phi$, 然后逐个去除属性 $c \in C$, 若 $\gamma_{C-c} \neq \gamma_C$, 就能够确定 $core(C)$ 为其中的一个核; 若 $\gamma_{core(C)(D)} = \gamma_C(D)$, 则说明 $core(C)$ 为最小属性约简, 如不是则继续执行步骤 3;
- ③ 由随机生成的长度为 n , 数量为 m 的二进制位串表示的个体组成初始种群, 然后计算出初始

群体中所有个体的适应度;

④依据赌轮盘的方法进行个体选择,同时根据变异概率 P_m 和交叉概率 P_c 生成下一代群体,在变异过程中应保证与核属性相对应的基因位不发生变化;

⑤计算下一代群体中全部个体的适应度;

⑥在获得下一代个体后,如果新一代个体中的某个个体的适应度小于上一代某个个体的适应度,则使用上一代最好的个体替换新一代最差的个体,以保证算法能够收敛;

⑦如果连续 t 代群体中最高适应度个体的适应度不再提高,则停止计算并将最优个体输出,也就是说此时的属性已经是最简化的。

2.3 决策规则集的生成

根据属性简化表,决策规则采用“if..., then...”的表达形式,即当属性满足一定的条件要求时,就可以得出相应的决策规则。但为了去掉表达决策规则时的多余属性值,需要进行属性值约简,本文采用计算决策规则的覆盖度和可信度进行值约简,覆盖度和可信度的计算如式2、式3所示:

$$\beta_R(D) = \frac{|[x_k]_R \cap D_k|}{|D_k|} \tag{2}$$

$$\alpha_R(D) = \frac{|[x_k]_R \cap D_k|}{|[x_k]_R|} \tag{3}$$

式中, $\beta_R(D)$ 表示覆盖度, $\alpha_R(D)$ 表示可信度。 D_k 表示第 k 个决策规则的决策属性类, $[x_k]_R$ 表示对规则的全部条件属性分类。通过选取覆盖度和可信度比较高的决策规则来完成值约简。

2.4 对象分类

完成决策表的属性约简及值约简后,得到了最终的全部决策规则。银行可以根据决策规则对任意一个贷款对象进行分类,但依据决策规则得到的某一贷款对象与其信息数据的匹配程度可能会有以下几种情况:1)新贷款对象仅匹配某一条规则;2)新贷款对象能够匹配多条规则,且匹配结果一致;3)新贷款对象能够匹配多条规则,但匹配结果不相同;4)新贷款对象无法匹配任何一个规则。

对情况 1 和情况 2,根据规则集对贷款对象的判定结果仅有一个,所以能够确定贷款对象的分类;但对于情况 3 和情况 4,无法根据规则集对贷款对象进行分类,本文分别采用投票法和最近相邻法来解决情况 3 和情况 4,具体如下:

投票法:决策规则集用 R 表示,让 R 为对象的所有可能决策类分配一个代表其可信度的量值。通常,对象都被划分到改值最大的类中。假设待进行分类的对象为 x ,投票的具体过程如下:

①扫描整个决策规则集 R ,激活规则集 $R(x)$ 并找出与对象 x 匹配的全部规则;

②各个规则 $\alpha \rightarrow \beta \in R(x)$ 为其后件的决策类指定一个量值 $votes(\alpha \rightarrow \beta)$ 作为 $\alpha \rightarrow \beta$ 的票数: $votes(\alpha \rightarrow \beta) = support(\alpha \wedge \beta)$,这就是在决策规则集中同时满足规则 α 和规则 β 的对象数目;

③计算对象 x 相对于各个决策类的可信度 $certainty(x, \beta)$ 如式 4 所示,投票数 $votes(\beta)$ 和 $R_\beta(x)$ 分别如式 5、式 6 所示。

$$certainty(x, \beta) = \frac{votes(\beta)}{\sum_{\beta} votes(\beta)} \tag{4}$$

$$votes(\beta) = \sum_{\alpha \rightarrow \beta \in R_\beta(x)} votes(\alpha \rightarrow \beta) \tag{5}$$

$$R_\beta(x) = \{\alpha \rightarrow \beta : \alpha \rightarrow \beta \in R(x)\} \tag{6}$$

$$Sim_i = \sum_{j=1}^n w_j (|\phi_j^* - \phi_{i,j}| / \phi_j^*) \quad (7)$$

最近相邻法: 各个规则和测试仪样本的相似度如式 7 所示, 其中, ϕ_j^* 表示测试样本的第 j 个评价指标的具体值, $\phi_{i,j}$ 表示规则库中第 i 个规则的第 j 个评价指标具体值, w_j 表示第 j 个评价指标的权重大小, 本文取所有权重相同。 Sim_i 表示规则库中的第 i 个规则和测试样本的相似度, 数值越小表示两者的相似程度越高。

2.5 决策规则预测精度检验

决策规则建立后, 应依据测试样本进行规则检验, 以验证所建立的规则是否科学。在规则检验中, 建立的规则配比准确率越高、测试样本的数量越大, 则说明建立的信贷风险评估模型的可行性越高。

3 结果与分析

为验证本文提出的基于粗糙集理论的信贷风险评估模型是否准确, 选取 2015 年的 60 家 ST 公司作为信贷违约样本, 60 家信贷正常公司作为信贷正常样本, 然后从中随机抽取 96 家公司 (48 家信贷违约, 48 家信贷正常) 作为评估模型的训练样本, 剩下的 24 家公司作为测试样本, 用于检测评估模型的准确性。本文选取的样本中, 缺少部分数据项, 具体如表 3 所示, 表中缺陷数据采用 Rosetta 软件中的 Conditioned Mean Completer 算法补全。

表 3 部分原始数据

Table 3 Partial original data

公司代码	流动比率	速动比率	资产负债率	利息保障倍数	总资产周转率
Company code	Current ratio	Quick ratio	Asset-liability ratio	Multiple of interest safeguard	Total asset turnover
C600715	0.1972	0.5381(缺项)	0.9146	-241.5600	0.0048
C600335	1.0362	0.5365	0.7384	-34.978 (缺项)	0.4321
C000595	0.1415	0.6437	0.7124	-4.7859	0.4628
C000002	1.5854	0.5658	0.7357	11.5078	0.2872

对于信贷风险评估体系中的连续性指标, 运用 Rosetta 中的 Equal Frequency (等频率算法) 进行指标离散, 将每个指标分为 4 段, 分别以 1、2、3、4 表示各段的离散值, 部分指标离散后的数值如表 4 所示。

表 4 离散化后的部分属性数据

Table 4 Partial data of the properties after discretization

公司代码	流动比率	速动比率	资产负债率	利息保障倍数	总资产周转率
Company code	Current ratio	Quick ratio	Asset-liability ratio	Multiple of interest safeguard	Total asset turnover
C600715	3	3	2	4	1
C600335	3	2	1	3	1
C000595	4	2	1	1	3
C000002	3	2	1	3	4

在选用的信贷风险评估指标体系中, 有很多指标是多余的, 采用 Rosetta 软件的遗传算法对评估指标体系中的属性进行简化, 最终选取了 $\{C_6, C_8, C_{15}, C_{17}, C_{19}\}$ 五个属性作为简化后的条件属性。在简化指标的基础上, 设置规则的覆盖度大于 0.05, 可信度大于 0.75, 最后一共得到了 30 条决策规则, 部分决策规则下所示:

①流动资产周转率 C_6 (1) AND 资产净利率 C_8 (2) AND 现金流动负债比率 C_{15} (1) AND 主营业务收入现金含量 C_{17} (4) AND 行业景气指数 C_{19} (3) $\Rightarrow D(0)$;

②流动资产周转率 C_6 (2) AND 资产净利率 C_8 (3) AND 现金流动负债比率 C_{15} (3) AND 主营业务收入现金含量 C_{17} (3) AND 行业景气指数 C_{19} (2) $\Rightarrow D(0)$;

③流动资产周转率 C_6 (2) AND 资产净利率 C_8 (1) AND 现金流动负债比率 C_{15} (2) AND 主营业务

收入现金含量 C_{17} (3) AND 行业景气指数 C_{19} (4) => D(1);

④流动资产周转率 C_6 (1) AND 资产净利率 C_8 (3) AND 现金流动负债比率 C_{15} (4) AND 主营业务收入现金含量 C_{17} (2) AND 行业景气指数 C_{19} (2) => D(1);

获得决策规则集后, 使用未参与训练的余下 24 个样本公司进行测试, 即将这 24 个样本按照决策规则进行分类, 然后与该公司的实际信贷情况进行对比, 具体如图 1 所示:

		Predicted		
		0	1	
Actual	0	10	2	0.833333
	1	0	12	1.0
		1.0	0.857143	0.916667
ROC	Class	Undefined		
	Area	Undefined		
	Std. error	Undefined		
	Thr. (0, 1)	Undefined		
	Thr. acc.	Undefined		

图 1 测试结果
Fig.1 Test results

从图 1 中的测试结果可以看出: 信贷情况正常公司的 12 个样本中, 有 10 个预测正确, 2 个样本错误的预测成了信贷违约, 正确率达到 83.33%; 信贷情况非正常的 12 个公司, 信贷情况的预测全部正确, 正确率达到了 100%; 信贷风险判别的综合正确率达高达 91.67%。说明基于粗糙集理论的信贷风险评估模型具有良好的预测精度。

4 讨论

本文建立的基于粗糙集理论的信贷风险评估模型具有良好的预测精度, 但同时也存在一些问题: 一是在信贷风险评估指标体系中没有能够反映宏观经济情况的指标, 这是因为该指标需要大量样本数据, 目前建立的模型还难以实现; 二是本文仅将信贷风险分为两级, 但实际银行是将信贷评估划分为五级, 如何将粗糙集理论应用在五级信贷评估中。这些都还需在后续的工作中继续研究。

5 结论

针对我国商业银行采用的传统信贷风险评估方法的不足, 提出了基于粗糙集理论的信贷风险评估模型。首先建立了一套信贷风险评估指标体系, 然后给出了粗糙集理论应用在信贷风险评估模型中的各步骤实现方法, 最后采用 24 家公司的信贷情况作为测试样本, 测试结果表明信贷风险判别的综合正确率达高达 91.67%, 可为银行的信贷决策提供有效的参考。

参考文献

[1] Douglas RE, John DF. Corporate Financial Management[M]. Beijing: Chinese People's University Press, 2015

[2] Li RenPu, Zheng Ouwang. Mining Classification Rules Using Rough Sets and Neural Networks[J]. European Journal of Operational Research, 2014,3(4):443-448

[3] 肖厚国, 桑琳, 宫悦, 等. 基于遗传算法的粗糙集属性约简及其应用[J]. 计算机工程与应用, 2015, 44(15): 228-230

[4] Slowinski R, Zopounidis C, Dimitras AI. Rough Set Predictor of Business Failure [J]. Soft Computing in Financial Engineering, 2014, 5(8): 402-424

[5] 朱小刚. 基于计算机粗糙集的数据挖掘设计与应用分析[J]. 山东农业大学学报: 自然科学版, 2015, 46(5): 765-768

[6] 吴山产, 毛锋, 王文渊. 基于粗糙集两种离散化算法的研究[J]. 计算机工程与应用, 2014(26): 67-69

[7] 薛锋, 柯孔林. 粗糙集与神经网络系统在商业银行贷款五级分类中的应用[J]. 系统工程理论与实践, 2015, 3(5): 40-45

[8] 于达仁, 胡清华, 鲍文. 融合粗糙集和模糊聚类的连续数据知识发现[J]. 中国电机工程学报, 2014, 24(6): 205-210