

基于机器学习的信贷风险量化与决策分析

王沼翔¹, 葛琳^{1,2}

(1.郑州航空工业管理学院智能工程学院, 河南 郑州 450046;

2.中国民航大学中国民航信息技术科研基地, 天津 300300)

✉1113696260@qq.com; lingsnow@126.com



摘要: 中小微企业在发展过程中需要信贷融资, 然而部分企业的不良贷款会对金融稳定构成威胁。针对此问题, 可以通过对企业相关财务数据进行分析, 从源头上防范和降低信贷风险。本文采用熵权法、数据统计分析法、K均值聚类和BP神经网络等机器学习算法来给出相应的信贷策略, 规避信贷风险。实验基于2020年全国大学生数学建模竞赛赛题发布的公开数据集, 利用MATLAB R2018a和Python 3.9等工具进行代码编写。测试结果表明, 本文方法可以有效地对中小微企业的信贷风险进行评估并制定相应的信贷策略。

关键词: 风险评级模型; 机器学习; 熵权法; BP神经网络; K均值聚类

中图分类号: TP391 **文献标识码:** A

Credit Risk Quantification and Decision Analysis based on Machine Learning

WANG Zhaoxiang¹, GE Lin^{1,2}

(1.School of Intelligent Engineering, Zhengzhou University of Aeronautics, Zhengzhou 450046, China;

2.Information Technology Research Base of Civil Aviation Administration of China, Civil Aviation University of China, Tianjin 300300, China)

✉1113696260@qq.com; lingsnow@126.com

Abstract: Medium, small and micro enterprises need credit financing in the development process, but non-performing loans of some enterprises pose a threat to financial stability. Aiming at this problem, an analysis of relevant financial data of the enterprise prevents and reduces credit risks from the source. This paper proposes corresponding credit strategies to avoid credit risks by using machine learning algorithms such as entropy weight method, data statistical analysis method, K-means clustering and BP neural network. The experiment is based on the public data set released by the 2020 National College Students Mathematical Modeling Contest, using MATLAB R2018a, Python 3.9 and other tools for code writing. Test results show that the proposed strategies can effectively evaluate the credit risk of medium, small and micro enterprises and formulate corresponding credit strategies.

Keywords: risk rating model; machine learning; entropy weight method; BP neural network; K-means clustering

1 引言(Introduction)

在市场经济运行中, 中小微企业的经营特点决定了其在融资过程中存在一定的风险, 银行在向中小微企业提供贷款

时, 要充分考虑多方面因素并制定合适的信贷策略。面对海量的信贷数据, 如何提高模型与算法对信贷企业风险评估的精确性, 已经成为银行业亟待解决的关键问题。传统信

贷风险管理模式主要依赖人工运用统计方法对企业借贷违约可能性的预测，以被评估者相关指标的历史数据为依据，对被评估者的还贷能力和违约行为进行预测和分析。随着互联网金融的兴起，传统的评估模型与预测算法已不能满足评估信贷风险的实际需要^[1-4]。在大数据技术的作用下，银行可以运用综合分析有效防控信贷风险和不对称问题。借助机器学习 and 深度学习，银行可以分析大量的企业数据，预测信贷风险，提高风险识别能力^[5]。

本文采用熵权法、数据统计分析法和BP神经网络，利用MATLAB R2018a和Python 3.9等工具对有信贷和无信贷记录企业的相关数据集进行分析，给出相应的信贷策略，规避信贷风险^[6-8]。对于有信贷记录企业，本文运用数据统计法分析中小微企业的信贷影响因素，采用熵权法计算每个影响因素的权值，并得出各企业综合信贷风险值，最后通过聚类分析，将各企业划分类别，通过与信誉评级的组合来确定合适的信贷策略方案；对于无信贷记录企业，本文使用有监督学习的神经网络对信贷策略进行预测和分析，得到各个信誉等级数据。实验结果表明，本文所建立的模型可以有效地对中小微企业的信贷风险进行评估并制定有效的信贷策略。

2 数据预处理(Data preprocessing)

本文以2020年全国大学生数学建模竞赛赛题发布的公开数据集为例进行分析，该数据集包含123家有信贷记录企业的21万条进项发票信息、16万条销项发票信息，302家无信贷记录企业的39万条进项发票信息、33万条销项发票信息以及银行贷款年利率与客户流失率关系统计数据。基于Python编程完成对数据集中数据的预处理，分别实现有信贷记录企业和无信贷记录企业的销项、进项发票中的数据计算，通过对每个企业有效发票的金额、税额、价税的累加，得到各企业全部有效发票的总金额、总税额和总价税合计等数据信息。

3 有信贷记录企业信贷分析与决策(Credit analysis and decision-making for companies with credit records)

对于有信贷记录企业，根据企业交易票据中的数据，通过建立数学模型给出合适的信贷策略。首先，构建人脉占有率、市场流通率、市场稳定率、企业盈利率四个影响信贷风险因素的评估因子；其次，基于K均值聚类算法、熵权法建立风险评级模型，以风险级别确定贷款额度；第三，分析客户流失率、贷款利率以及企业评级的关系，以信誉评级确定贷款利率^[2]；最后，按照三类风险评级、三类信誉评级组合成的九种企业定位确定不同额度、不同利率的信贷策略。整体流程图如图1所示，模型分析中所用的数学符号如表1所示。

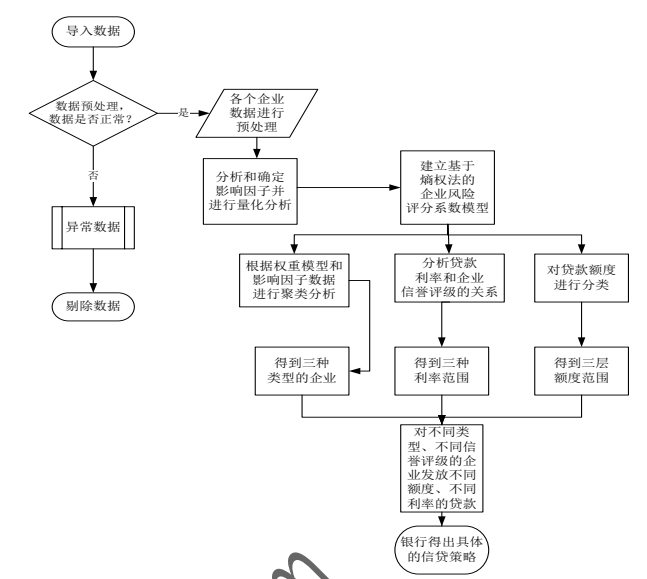


图1 整体流程图

Fig.1 Overall flow chart

表1 符号说明表

Tab.1 Symbol description table

符号	意义
RM	人脉占有率
LT	市场流通率
WD	市场稳定率
YL	企业盈利率
Q_j	熵权法模型权重
LFD	类别综合评分
X'_y	标准化数据
D	划分的额度
L	划分的利率
G	全体企业贷款金额
Σw_i	银行总收益
Cd	基于熵权法风险评分系数模型
$w_{ij}、v_{ji}$	每一个神经元的权值
e_j	中间层神经元误差
d_i	输出层一般误差
s_j	隐含层的神经元输入

3.1 影响因子

对有信贷记录企业的数据根据进项、销项、有效发票和发票总数进行分析，综合考虑企业和银行各方面的因素，在企业实力、供求稳定性等方面设置下列影响因子：

(1)人脉占有率 RM_i

$$RM_i = \frac{J_i + D_i}{\sum (J_i + D_i)}$$

(1)

式中, J_i 、 D_i 分别表示第 i 个企业的进项发票数量和销项发票数量。

(2)市场流通率 LT_i

$$LT_i = \frac{K_{li} + D_{li}}{2 \sum (K_{li} + D_{li})} \quad (2)$$

式中, K_{li} 和 D_{li} 分别表示第 i 个企业的进项发票总金额和销项发票总金额, $\sum (K_{li} + D_{li})$ 为全体企业进货和销售产生的发票总金额。

(3)市场稳定率 WD_i

$$WD_i = \frac{Z_{wi} - (F_{wi} + S_{wi})}{Z_{wi}} \quad (3)$$

式中, S_{wi} 表示负数发票数量, Z_{wi} 、 F_{wi} 分别表示第 i 个企业的发票数总数、作废发票数。

(4)企业盈利率 YL_i

$$YL_i = \frac{X_{yi} - J_{yi}}{X_{yi} + J_{yi}} \quad (4)$$

式中, X_{yi} 和 J_{yi} 分别表示第 i 个企业的销项发票信息中价税合计总和、进项发票信息中价税合计总和。

3.2 基于熵权法的企业风险评级模型

根据人脉占有率 (RM)、市场流通率 (LT)、市场稳定率 (WD)、企业盈利率 (YL) 四个影响因子, 利用熵权法建立评级模型^[3], 计算企业风险评级的度量值 Cd 。

$$Cd = Q_1 RM + Q_2 LT + Q_3 WD + Q_4 YL \quad (5)$$

(1)数据标准化, 其中 x_{ij} 为第 i 个企业第 j 项的评分, 如公式(6)所示。因为各个指标的计量单位不一样, 先对数据进行相对标准化的处理, 减小各项统计数据的相对误差。

$$X'_{ij} = \frac{X_{ij} - \min\{X_{1j}, \dots, X_{mj}\}}{\max\{X_{1j}, \dots, X_{mj}\} - \min\{X_{1j}, \dots, X_{mj}\}} \quad (6)$$

(2)计算第 j 项指标在 i 条件下所占的比重

$$p_{ij} = \frac{X'_{ij}}{\sum_{i=1}^n X'_{ij}}, i=1, \dots, n; j=1, \dots, m \quad (7)$$

(3)计算第 j 项的信息熵值, 其中 $k = \frac{1}{\ln n} > 1$, n 为企业数

$$e_j = -k \sum_{i=1}^n p_{ij} \ln(p_{ij}) \quad (8)$$

(4)计算信息熵冗余度

$$d_j = 1 - e_j \quad (9)$$

(5)计算各个指标的权重

$$Q_j = \frac{d_{ij}}{\sum_{i=1}^n d_{ij}} \quad (10)$$

(6)根据上述计算和分析, 得出企业风险评级模型

$$Cd = 0.41RM + 0.37LT + 0.1WD + 0.12YL \quad (11)$$

3.3 实验与测试

对各个企业代入各个影响因子的权重进行聚类分析, 结果如表2所示。

表2 聚类分析表

Tab.2 Cluster analysis table

影响因子	企业类别			权重
	S ₁	S ₂	S ₃	
RM	0.031031	0.011080	0.007782	0.41
LT	0.015558	0.008434	0.007891	0.37
WD	0.904473	0.931298	0.927491	0.10
YL	0.512995	0.083049	0.882383	0.12
类别综合得分(LFD)	0.003776	0.001348	0.000947	1

将企业聚合为三个类别, 对三个类别进行综合评分 (LFD), 其评分计算如公式(12)所示。将该三类企业重新进行分类并编号, 类别由强到弱编为 S_1 、 S_2 、 S_3 , 根据聚类分析得出三个类别分别对应的部分企业如表3所示。

$$LFD = Q_1 \sum (RM_i) + Q_2 \sum (LT_i) + Q_3 \sum (WD_i) + Q_4 \sum (YL_i) \quad (12)$$

表3 部分企业分类表

Tab.3 Classification table of part of the enterprises

S ₁ 类别企业	S ₂ 类别企业	S ₃ 类别企业
E11	E2	E4
E17	E8	E7
E18	E12	E9
E19	E13	E10
E81	E86	E94
E89	E88	E95

利率和企业的信誉之间存在潜在关系, 在A、B、C三种信誉等级的企业流失率与贷款利率呈正相关且在同等贷款利率水平下, A、B、C三种信誉等级的企业流失率依次降低。本文将年利率和贷款额度分别分为三组, 如表4所示。

表4 额度和利率分组表

Tab.4 Grouping list of quota and interest rate

额度分组	额度/万元	利率分组	利率范围/%
D ₁	10—39	L ₁ (低)	4.25—5.05
D ₂	40—69	L ₂ (中)	5.45—6.25
D ₃	70—100	L ₃ (高)	6.65—7.45

银行根据企业的信贷风险类别和其对应的信誉评级将企业分为九级定位, 对于不同信贷风险级别的企业, 确定不同贷款额度; 对于不同信誉级别的企业, 确定不同贷款利率^[4]。信贷策略各级对应表如表5所示。

表5 信贷策略表

Tab.5 Credit strategy table

企业类别	信誉评级		
	A	B	C
S_1	$D_3 \times L_1$	$D_3 \times L_2$	$D_3 \times L_3$
S_2	$D_2 \times L_1$	$D_2 \times L_2$	$D_2 \times L_3$
S_3	$D_1 \times L_1$	$D_1 \times L_2$	$D_1 \times L_3$

4 无信贷记录企业信贷分析与决策(Credit analysis and decision-making of enterprises without credit records)

本文所分析的数据集涉及302家企业信息中无信誉评级等数据,然而信誉评级在银行信贷策略中具有重要的意义。首先,采用有监督学习的神经网络,对所分析的数据集中123家有信誉评级的企业信息提取的影响因子进行训练;其次,利用训练得到的BP神经网络对无信誉评级的企业进行信誉评级和分析;最后,验证模型是否合理^[6]。

4.1 BP神经网络模型

无信贷记录企业的数据集中无信誉等级数据,故而使用有监督学习的神经网络对信贷策略进行预测和分析,并得到各个信誉等级数据。BP神经网络模型中各符号的含义如下:

网络输入向量 $X_n = (x_1, x_2, \dots, x_n)$; 网络目标向量 $T_n = (y_1, y_2, \dots, y_n)$; 中间层单元输入向量 $S_n = (s_1, s_2, \dots, s_n)$ 、输出向量 $B_n = (b_1, b_2, \dots, b_n)$; 输出层单元输入向量 $L_n = (l_1, l_2, \dots, l_n)$ 、输出向量 $Z_n = (z_1, z_2, \dots, z_n)$; 输出层到中间层的链接权值 $w_{ij}, i=1,2,\dots,p, j=1,2,\dots,p$; 中间层到输出层的链接权值 $v_{jt}, j=1,2,\dots,p, t=1,2,\dots,q$; 中间层每一个神经单元的输出阈值(即偏置向量) $\theta_j, j=1,2,\dots,n$; 输出层每一个神经单元的输出阈值(即偏置向量) $\gamma_t, t=1,2,\dots,n$; 参数 $k=1,2,\dots,m$ 。

BP神经网络模型计算输入输出信息结果是否产生相互影响的权重,并进行综合分析。构建过程如下:

(1)初始化。对每一个神经元的权值 w_{ij} 和 v_{jt} 、阈值 θ_j 和 γ_t 设置区间在 $(-1, 1)$ 内的随机数。

(2)在有信贷记录企业数据集的123家企业中选取一组学习样本 $X_n = (x_1, x_2, \dots, x_n)$ 和目标样本 $T_n = (y_1, y_2, \dots, y_n)$ 。

(3)计算中间层单元的输出 b_j 。

$$s_j = \sum_{i=1}^n w_{ij} \alpha_i - \theta_j, j=1,2,\dots,p \quad (13)$$

$$b_j = f(s_j), j=1,2,\dots,p \quad (14)$$

(4)计算输出层传递神经元 L_t 。

$$L_t = \sum_{j=1}^p v_{jt} b_j - \gamma_t, t=1,2,\dots,q \quad (15)$$

(5)计算网络的目标向量和实际向量的差值。

$$d_t = (y_t - c_t) c_t (1 - c_t), t=1,2,\dots,q \quad (16)$$

(6)计算中间层一般误差 e_j 。

$$e_j = \left[\sum_{t=1}^q d_t v_{jt} \right] b_j (1 - b_j) \quad (17)$$

(7)用输出层各单元一般误差 d_t 与中间层单元输出值 b_j 来修正链接权值 v_{jt} 和阈值 γ_t , 其中 $t=1,2,\dots,q; j=1,2,\dots,p; 0 < \alpha < 1$ 。

$$v_{jt}(N+1) = v_{jt}(N) + \alpha d_t b_j \quad (18)$$

$$\gamma_t(N+1) = \gamma_t(N) + \alpha d_t \quad (19)$$

(8)用中间层神经误差 e_j 来调整链接权值 w_{ij} , 用输入层神经元的输入 x_n 来调整阈值 θ_j , 其中 $i=1,2,\dots,q; j=1,2,\dots,p; 0 < \beta < 1$ 。

$$w_{ij}(N+1) = w_{ij}(N) + \beta e_j \alpha_i \quad (20)$$

$$\theta_j(N+1) = \theta_j(N) + \beta e_j \quad (21)$$

(9)随机选取下个样本代入神经网络进行训练,执行步骤(3),直到网络的训练全局误差达到网络收敛值,学习结束^[7]。本文获得的影响企业实力的因子为四个,输入层神经节点数目为4,而本文神经网络最终输出的只有一个企业实力数据,则输出层节点个数为1。隐含层节点数目的取值如公式(22)所示。式中, n 为输入层节点的数目, l 为输出层节点的数目, m 为隐含层节点的数目。由公式(22)可得隐含层节点数目为2。

$$m = \sqrt{nl} \quad (22)$$

4.2 模型训练

采用独立测试方法,将实验数据分成两部分,其一是训练数据。本文将有信贷记录企业数据集中的123家企业的样本数据按照信誉评价分类,随机抽取100组数据作为实验培训数据,剩下的23组样本数据作为测试数据,用于检测BP神经网络。利用MATLAB将训练样本数据代入程序,并对该神经网络进行训练,得到的训练结果如图2和图3所示。

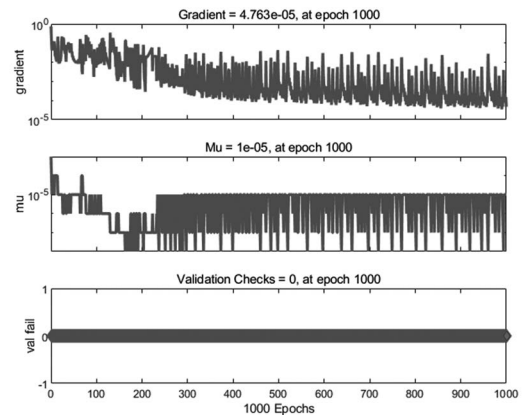


图2 神经网络训练步数

Fig.2 Neural network training steps

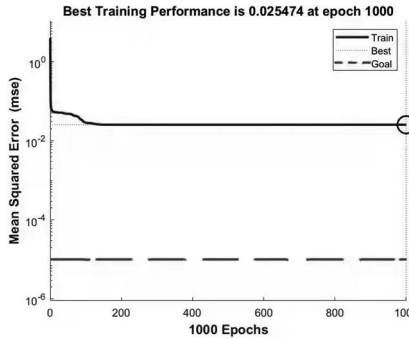


图3 神经网络训练误差

Fig.3 Neural network training error

通过上述训练，可以得到初步的训练结果，为了验证BP神经网络，将测试数据代入已经训练好的网络模型当中进行测试。为得出模型测试准确度，我们构建以下度量模型，测试结果好坏的误差计算如公式(23)所示。式中， μ 为模型误差， y' 为真实值， y'' 为评价值， n 为测试数据集个数。通过上述误差公式计算得出BP神经网络的测试误差为0.112。通过构建并优化BP神经网络模型，计算出无信贷记录企业的四个指标因素值。

$$\mu = \frac{\sum_{i=1}^n |y' - y''|}{n} \quad (23)$$

将无信贷记录企业的四个指标因素值作为输入值代入BP神经网络模型中对企业进行信誉评价，最终得出各个企业的信誉等级，如图4所示。

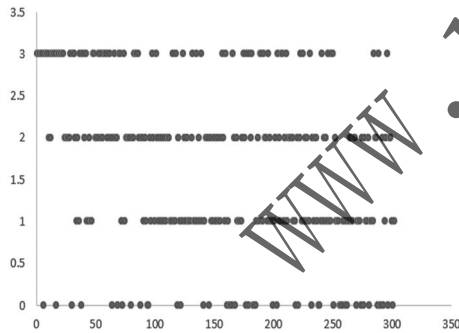


图4 各企业的神经网络信誉评价值

Fig.4 Neural network reputation evaluation value of each company

图4中值为3对应信誉等级为A，数量为76个；值为2对应信誉等级为B，数量为92个；值为1对应信誉等级为C，数量为86个；值为0对应信誉等级为D，数量为48个。因为银行对信誉等级为D的企业原则上不予放贷，故将BP神经网络的信誉等级匹配到无信贷记录企业数据集的各个企业中，将信誉评价为D级的企业剔除。对无信贷记录企业数据集处理后的各项数据利用公式(1)——(4)进行量化分析，计算出各个指标的值，并且利用熵权法求出四个影响因子的权重，则企业风险评级模型如公式(24)所示：

$$Cd=0.44 \times RM+0.42 \times LT+0.02 \times WD+0.11 \times YL \quad (24)$$

4.3 实验与测试

根据无信贷记录企业的数据计算出各个影响因子的权重，分析四个影响因子的数据，进行聚类分析，将企业聚合为三个类别，则得出分级评分如表6所示。

表6 影响因子综合评分表

Tab.6 Score form of comprehensive impact factor

影响因子	企业类别			权重
	S_1	S_5	S_6	
RM	0.003479	0.006186	0.007843	0.44
LT	0.002621	0.002983	0.004858	0.42
WD	0.930065	0.902460	0.922957	0.02
YL	0.895849	0.528507	0.126858	0.11
LFD	0.119777	0.080159	0.037905	1
聚类数量	74	71	103	

由分级数据以及客户流失率可得，当年利率大于0.0745时，各个信誉等级企业的客户流失率为0.5，潜在客户大量流失，其收益也将大幅度下降，所以将客户流失率控制在0.5之内，有利于银行收益的可持续发展^[8]。本文将客户流失率小于0.5的年利率分为三组， S_4 、 S_5 、 S_6 部分对应的企业信贷额度也分为三组。

根据前文数据分析，按照企业的风险评级和实力类别制定九种信贷策略，从而确定不同企业的贷款额度、利率，如表7所示。

表7 信贷策略表

Tab.7 Credit strategy table

企业类别	信誉评级		
	A	B	C
S_4	$D_3 \times L_1$	$D_3 \times L_2$	$D_3 \times L_3$
S_5	$D_2 \times L_1$	$D_2 \times L_2$	$D_2 \times L_3$
S_6	$D_1 \times L_1$	$D_1 \times L_2$	$D_1 \times L_3$

在年利率区间中，均取中间值为相对年利率，分别为4.65%、5.85%、7.05%。设 W_i 为银行在第 i 个企业的收益，收益计算如公式(25)所示：

$$W_i = D_{ai} \times L_{\beta i} \quad (25)$$

按照信贷策略分配金额，得出全体企业的贷款金额总额区间 $G \in [9050, 16316]$ (万元)，银行的收益总额区间 $\sum W_i \in [537.825, 964.2855]$ (万元)。

由于假设中规定是1亿元定额分配贷款金额，9,050万元按照表7的信贷策略对企业进行放贷。剩下950万元有两种选择策略：一是银行利益最大化，将剩余额度优先分配给贷款利率较高的信誉C级企业；二是银行发展持久化，将剩余额度优先分配给信誉A级企业，来降低此类高信誉级别企业的客户
(下转第34页)