

基于风险等级的中小微企业信贷模型研究

顾一凡¹, 黄莉媛², 林晨欣², 曹春萍¹

(1.上海理工大学光电信息与计算机工程学院, 上海 200093;

2.上海理工大学管理学院, 上海 200093)

✉guyifan2020@126.com; 948384993@qq.com; 1246506991@qq.com; 2213893844@qq.com



摘要: 为切实解决中小微企业贷款融资和银行对中小微企业贷款策略之间存在的问题, 提出了基于风险等级的中小微企业信贷模型。该模型创新性地将机器学习算法引入传统中小微企业信贷风险及策略的研究当中, 运用PCA降维、K-means聚类确定企业风险等级; 通过Fisher线性判别确定银行信贷利率。应用该模型将123家中小微企业分成五类风险等级, 并给出银行对五类不同风险等级企业的贷款额度及利率, 并通过实验验证模型的有效性和正确性。

关键词: K-means聚类; PCA降维; Fisher线性判别; 信贷模型

中图分类号: TP391 **文献标识码:** A

Research on the Credit Model of Small, Medium and Micro Enterprises based on Risk Level

GU Yifan¹, HUANG Liyuan², LIN Chenxin², CAO Chunping¹

(1.School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;

2.Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

✉guyifan2020@126.com; 948384993@qq.com; 1246506991@qq.com; 2213893844@qq.com

Abstract: In order to effectively solve the problems between loan financing of small, medium and micro enterprises and the bank's loan strategy for them, this paper proposes to build a credit model for small, medium and micro enterprises based on risk level. This model innovatively introduces machine learning algorithms into the research on credit risks and strategies of traditional small, medium and micro enterprises. PCA (Principal Components Analysis) dimensionality reduction and K-means clustering are used to determine enterprise risk level. Bank credit interest rate is determined by Fisher linear discriminant. Based on this model, 123 small, medium and micro enterprises are divided into five risk levels, and the bank's loan lines and interest rates for each level are given. Validity and accuracy of the model are verified through experiments.

Keywords: K-means clustering; PCA dimensionality reduction; Fisher linear discriminant; credit model

1 引言(Introduction)

中小微企业作为我国宏观经济的“毛细血管”, 贡献了我国80%的就业岗位和60%的GDP^[1-2]。但中小微企业自身资金实力弱, 若想发展就需要获得银行提供的贷款支持, 而银行放贷首要考虑的是贷款资金的安全, 因此银行放贷资金安全与中小微企业贷款需求之间的矛盾便成为一个亟待解决的问题。

目前, 银行为了解决这一问题, 采取了如下策略: 对

中小微企业进行风险评估, 对于风险等级在一定标准下的企业, 给予放贷。在郑建华等^[3]提出的研究企业信用评级模型中, 使用了层次分析法进行评级模型的构建。在郝晓露等^[4]提出的商业银行贷款的研究中, 使用了灰色预测模型对贷款利率进行了预测。王薛^[5]使用了AHP——模糊综合评价农村信用社农户贷款风险模型。仔细分析上述模型可以发现, 每个模型都涉及大量计算和推断。计算机在大量数据的计算中表现出了极大的优势, 所以构建基于计算机的信贷模型成为目前

的研究热点。

本文在郑建华等相关工作的基础上,创新性地将机器学习算法引入对中小微企业贷款风险的研究当中。针对这一问题,提出了基于中小微企业风险评估等级的银行贷款模型。该模型首先从企业原始发票信息中提取若干指标;其次通过对指标进行降维,对企业进行聚类,划分出企业风险等级;最后根据企业对应的风险等级,通过构建贷款额度与贷款利率的计算模型,为银行制定相应的贷款策略。

2 企业风险等级确定(Enterprise risk level determination)

从123家企业的进项、销项发票记录中提取信息,定义10个原始风险评价指标^[6-7]。将这10个原始风险评价指标通过PCA降维,得到三个降维之后的评价指标,既保留了原始数据的绝大部分信息,又极大简化了后续的计算量。最后,根据三个降维之后的评价指标,通过K-means聚类,将原123家企业分为五个风险等级,为后续银行信贷策略的确定提供依据。

2.1 原始评价指标定义

从123家企业的进项、销项发票记录中,提取并定义企业实力及企业信誉两大类共10个评价指标。

企业实力评价指标如下定义:

T_{sale} : 企业有效销售次数;

T_{purchase} : 企业有效进货次数;

V_{purchase} : 企业供应稳定性,即有效进货次数与总体进货次数的比值;

V_{sale} : 企业销售稳定性,即有效销售次数与总体销售次数的比值;

C_{purchase} : 企业进货规模,即进项金额总和(营业成本);

C_{sale} : 企业销售规模,即销项金额总和(营业收入);

S_{sale} : 企业对上游企业的影响力,即企业销方销售总额;

S_{purchase} : 企业对下游企业的影响力,即企业购方采购总额。

企业信誉评价指标如下定义:

U_{credit} : 企业信用评级,即将企业A、B、C、D四档原始信用评级折算为4321分;

U_{break} : 企业违约情况,即若企业有违约记录,则该指标为1,否则为0。

2.2 基于PCA降维对原始指标进行简化

主成分分析通过正交变换的方法,将原始线性相关的观测数据转变为若干个线性无关变量表示的数据^[8]。线性无关的变量称为主成分。如此,将原来高维空间中的数据映射到低维空间,降低了计算复杂度。同时,保留了原始数据中的大部分信息,实现了数据降维的功能。

如上10个指标覆盖了123家企业各个方面的信息,可以较为全面地量化其内在的风险。但由于这10个原始指标所构成的高维数据不便于观察和计算,所以采用PCA降维对原始高维指标进行降维。通过计算10个信贷风险指标之间的相关系数矩阵,再求解相关系数矩阵的特征值与特征向量。将特征向量对应主成分的特征值与全体特征向量对应主成分的特征值之和的比值定义为信息贡献率,选取前三个贡献率最大的主成分,其贡献率依次为34.85%、20.22%、13.42%,累积贡献率达68.49%,可以最大程度保留原始10个指标中所包含的企业内在风险的信息,又能解决高维指标数据不便于观察计算的问题。

其中第一主成分为:

$$F_1 = 0.21x_1 + 0.31x_2 - 0.14x_3 + 0.13x_4 + 0.46x_5 + 0.50x_6 + 0.44x_7 + 0.24x_8 + 0.25x_9 + 0.20x_{10} \quad (1)$$

第二主成分为:

$$F_2 = 0.30x_1 + 0.20x_2 - 0.05x_3 + 0.33x_4 - 0.28x_5 - 0.22x_6 - 0.31x_7 - 0.11x_8 + 0.50x_9 + 0.51x_{10} \quad (2)$$

第三主成分为:

$$F_3 = -0.61x_1 - 0.52x_2 - 0.04x_3 + 0.45x_4 + 0.08x_5 + 0.01x_6 + 0.10x_7 + 0.20x_8 + 0.15x_9 + 0.27x_{10} \quad (3)$$

将式(1)第一主成分 F_1 定义为企业交易规模指标,式(2)第二主成分 F_2 定义为企业信誉指标,式(3)第三主成分 F_3 定义为企业盈利能力指标。从三个维度重新审视企业内在的风险等级。

2.3 基于K-means确定123家企业的风险等级

K均值聚类是常用的聚类算法。在未知数据特征标签的情况下,将各项指标接近的样本点聚成一类^[9]。首先,由已确认的分类个数 k 选择 k 个数据对象作为初始聚类中心;然后将其余样本点分配到与之最近的聚类中心所在的类中;进而更新每个类中样本点的均值作为下一次更新的聚类中心。如此往复迭代,直至聚类结果收敛为止。K-means算法流程如图1所示。

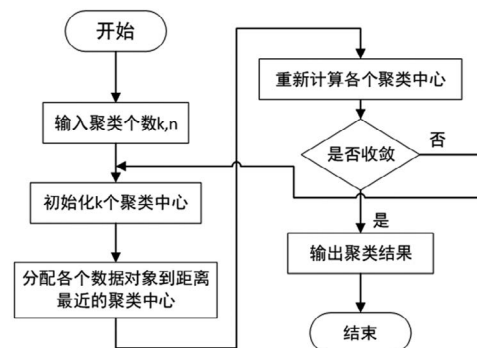


图1 K-means聚类算法流程图

Fig.1 K-means flow chart

将简化后的指标作为训练数据,通过K-means聚类,将

123 家企业聚类成不同风险等级的类别。选取 $k=5$ ，将原始 123 家企业依据降维之后的三个指标分成五个风险等级：高风险企业、较高风险企业、中等风险企业、较低风险企业、低风险企业。以此完成对 123 家企业风险等级的确定。

3 银行信贷模型(Bank credit model)

银行信贷模型分成银行信贷额度确定和银行信贷利率确定两大部分。依据企业有效进货次数 T_{purchase} 等三个指标，通过 Fisher 线性判别预测企业信贷违约概率。根据企业信贷违约概率建立银行信贷额度求解模型。同时，依据银行信贷利率与不同信誉等级下的客户流失率的统计数据数据进行数据拟合，再结合企业信贷违约率等相关数据，建立银行信贷利率求解模型。

3.1 基于 Fisher 线性判别的企业违约率预测^[10]

Fisher 线性判别是一种经典线性判别方法，适用于二分类问题。其核心思想为训练一组样本点，将样本点投影到一条直线上，使得同类样本点的投影点尽可能密集接近，使得异类样本点的投影点尽可能远离。最终训练出一个模型进行线性判别预测，同时给出样本点所属对应两个类别的概率。

由于无法从企业过往违约情况中直接得到企业未来贷款偿还的违约率，故采用 Fisher 线性判别，基于企业有效进货次数 T_{purchase} 、有效销售次数 T_{sale} 、企业信用评级 U_{credit} 作为训练数据训练模型。将企业未来是否违约转化为一个二分类预测问题，即将企业划分为未来会违约的组别和未来不会违约的组别，将预测结果与企业过往违约情况进行比对，得到得分 Z 模型：

$$Z = -0.19 \times T_{\text{purchase}} + 0.14 \times T_{\text{sale}} + 0.042 \times U_{\text{credit}} - 2.168 \quad (4)$$

若得分 $Z > 0$ ，则将该企业预测为不违约；反之，则将该企业预测为违约。由于样本点的分类存在属于对应类别的概率，故将样本点属于违约类别的概率定义为该企业的违约概率 P_y ，将该样本点属于未违约类别的概率定义为该企业的违约概率 P_n 。

3.2 银行信贷额度的确定^[11]

原则上不为信誉评级为 D 的企业发放贷款，故额度模型的建立及之后的计算中，自动剔除信誉评级为 D 的企业。设银行年度贷款总额为固定常数 a 元，计算同一风险等级企业的平均贷款额度 $\bar{y}_k$ 为：

$$\bar{y}_k = \frac{\bar{P}_{n_k}}{\sum_{k=1}^5 \bar{P}_{n_k} \cdot n_k} a \quad (5)$$

其中， $\bar{P}_{n_k}$ 为 k 级企业的平均不违约概率， n_k 为 k 级下企业总数(不包括信誉评级为 D 的企业)。

3.3 银行信贷利率的确定^[12]

本文贷款利率求解步骤如下：

步骤 1：根据银行贷款年利率与不同信誉等级下客户流失率的统计数据，进行对数函数形式的曲线拟合，得到如式(6)所示的银行客户流失率与贷款年利率之间的函数关系。

$$l_{kj} = \begin{cases} 2.239 + 0.669 \cdot \ln(x_k) & (j = A) \\ 2.158 + 0.651 \cdot \ln(x_k) & (j = B) \\ 2.168 + 0.659 \cdot \ln(x_k) & (j = C) \end{cases} \quad (6)$$

其中， l_{kj} 表示第 k 类风险等级且信誉等级为 j 的企业流失率， x_k 表示第 k 类风险的贷款年利率。由于不包含风险等级为 D 的企业，故如上函数中，没有 $j = D$ 的函数关系式。

步骤 2： $1 - l_{kj}$ 可表示为第 k 类风险等级且信誉等级为 j 的企业留存率，将式(6) l_{kj} 代入，则第 k 类风险等级企业的总体流失率 L_k 如式(7)所示。

$$L_k = \frac{\sum_j (1 - l_{kj}) \cdot n_{kj}}{\sum_j n_{kj}} \quad (k = 1, 2, 3, 4, 5) \quad (7)$$

其中， n_{kj} 表示第 k 类风险等级且信誉等级为 j 的企业个数。

步骤 3：以 p_k 为决策变量建立如式(8)所示非线性规划模型。

$$\begin{aligned} \max \text{Income} &= \sum_{k=1}^5 [p_k \cdot y_k \cdot L_k] \\ \text{s.t. } 4\% &< p_k < 15\% \end{aligned} \quad (8)$$

其中， p_k 为第 k 类风险等级的企业贷款利率，目标函数 Income 表示银行的总收入， y_k 表示给第 k 类风险等级企业贷款的总额度，企业贷款利率 p_k 的约束条件为 4% 至 15%。

通过对式(8)非线性规划函数模型在约束条件下求解，可以得到在风险可控的情况下，在贷款利润最大化的前提下，银行贷给不同风险等级企业的各自贷款利率。再结合由式(5)计算得到的不同风险等级下，银行给企业贷款的额度，可以完整给出银行对于不同风险等级下的企业的贷款策略。由于同一风险等级下企业的风险情况近似相等，故按照风险等级组别给出策略，可以极大降低放贷的复杂程度，同时保证风险可控。

4 实验验证(Experiment verification)

4.1 实验数据集

本次研究用数据集包括四个：企业信息数据集、进项发票信息数据集、销项发票信息数据集，以及银行贷款年利率与不同信誉等级下的客户流失率的统计数据。其中，企业信息数据集包括 123 家企业的企业代号、企业名称、企业信用评级以及企业历史违约情况。企业进项发票数据集、销项发票数据集分别包含 123 家企业采购及销售时所开发票的记录，记录中包括发票号码、开票日期、销方单位代号、购方单位代号、金额、税额、价税合计和发票状态。银行贷款年利率

与不同信誉等级下的客户流失率的统计数据集包括贷款年利率对应于不同信誉评级的企业的客户流失率的统计数据。

4.2 计算原始指标

对企业信誉及实力的10个原始评价指标进行计算。同时，为了消除不同指标之间量纲的影响，对每个评价指标进行极大值标准化。部分企业信贷风险指标评分结果如表1所示。

表1 部分企业信贷风险指标评分结果

Tab.1 Scoring results of credit risk indicators of some enterprises

企业 代号	企业实力指标评分								企业信誉 指标评分	
	有效进 货次数	有效销 售次数	供应稳 定性	销售稳 定性	进项金 额总和	销项金 额总和	上游企业 影响力	下游企业 影响力	信誉 评级	违约 情况
E1	9.98	49.93	74.76	91.32	100.00	100.00	100.00	26.63	100.00	100.00
E2	100.00	73.59	91.65	84.68	2.76	15.78	0.27	1.31	100.00	100.00
E3	13.86	100.00	85.98	54.99	0.93	14.18	0.34	10.69	33.33	100.00
E4	1.65	13.58	79.03	90.18	5.50	48.43	2.35	100.00	33.33	100.00
E5	6.59	6.60	85.98	92.92	3.39	5.08	0.94	5.15	66.67	100.00
E6	34.24	5.80	84.43	80.11	5.38	9.87	0.73	3.64	100.00	100.00
E7	37.73	28.44	73.12	43.00	1.05	12.66	0.14	4.88	100.00	100.00
E8	69.47	52.88	89.47	81.42	2.89	9.76	0.52	1.22	100.00	100.00
E9	13.32	37.06	91.16	93.49	0.42	7.93	0.18	3.32	100.00	100.00
E10	12.18	3.42	80.64	89.53	0.10	8.89	0.26	10.95	66.67	100.00

4.3 基于降维后的指标对企业进行不同风险等级的聚类实验及结果分析

基于降维后的三个指标对123家企业进行聚类， k 值为5。不同风险等级企业聚类结果如图2所示。

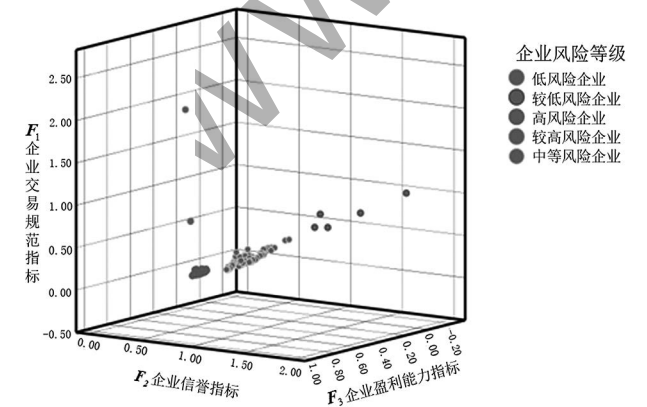


图2 不同风险等级企业聚类散点图

Fig.2 Cluster scatter diagram of enterprises with different risk levels

123家企业根据企业交易规模指标、企业信誉指标、企业

盈利能力指标聚成五类，依次为：低风险企业共1家，较低风险企业共7家，中等风险企业共86家，较高风险企业共2家，高风险企业共27家。通过比对对应企业原始10个指标，其聚类结果与实际情况基本一致。

4.4 企业信贷违约率计算的实验及结果分析

通过Fisher线性判别，根据式(4)，对企业违约情况进行预测，准确率达77.2%。进而计算企业违约概率 P_y 与企业不违约概率 P_n ，部分计算结果如表2所示。

表2 部分企业违约情况

Tab.2 Violations of some enterprises

企业 代号	有效进 货次数	有效销 售次数	信誉评 级情况	违约 情况	预测违 约情况	违约概 率(%)	不违约概 率(%)
E1	9.98	49.93	100	1	1	0.0003	0.9997
E2	100	73.59	100	1	1	0.00852	0.99148
E3	13.86	100	33.33	1	1	0.06223	0.93777
E4	1.65	13.58	33.33	1	1	0.44363	0.55637
E5	6.59	6.6	66.67	1	1	0.03726	0.96274
E6	34.24	5.8	100	1	1	0.00435	0.99565
E7	37.73	28.44	100	1	1	0.0023	0.9977
E8	69.47	52.88	100	1	1	0.00427	0.99573
E9	13.32	37.06	100	1	1	0.00054	0.99946
E10	12.18	3.42	66.67	1	1	0.05326	0.94674

4.5 银行信贷策略的实验及结果分析

通过对贷款额度模型式(5)及贷款利率模型式(8)的求解，得到对五类风险等级企业的贷款策略，完整贷款策略如表3所示。

表3 银行对各类风险等级企业的信贷策略

Tab.3 Bank credit strategy for enterprises of each risk level

信贷风险等级	信贷策略			
	是否可以放贷	贷款利率	企业平均贷款额度(元)	信贷期限
低风险企业	是	0.057728	0.012503a	
较低风险企业	是	0.058643	0.012309a	
中等风险企业	是	0.061097	0.01011a	1年
较高风险企业	是	0.062113	0.00732a	
高风险企业	是	0.062414	0.006958a	

其中，a为银行放贷的总额度。对于风险较低的企业类别，可以获得较大的贷款额度，同时享受较低的贷款利率；而对于风险较高的企业，则在获得较低贷款额度的同时，需

(下转第55页)