

基于支持向量机与神经网络的中小微企业信贷策略

概述

如何制定针对中小微企业的**信贷策略**从而达到利益最大化对银行来说是至关重要的。本文使用**聚类分析**、**支持向量机**、**对应分析**及**BP神经网络**等方法，通过对所给企业进行行业分类，借助于 MATLAB、R 统计分析软件、Python 及 SPSS 等工具，针对附件中所给数据，比较准确的给出银行对于各企业是否放贷、贷款额度、利率及期限等相关信贷策略。针对问题一，通过对附件一数据进行处理：

1. 删除信誉评级为 D 的企业;
2. 剩余企业按照 A、B、C 三个等级进行分组;
3. 计算企业总盈利额;
4. 是否违约;

按照 A、B、C 三类企业使用聚类分析进行分组，在三类企业中再分别分出优、良两类企业，共计六类企业，相应信贷策略在正文（表 5）及支撑材料（表问题一）中明确给出。

针对问题二，通过使用支持向量机找出能够区分全部企业盈利额数据点的最优超平面，将企业分为优、良两个等级。具体信贷策略及一亿元的细分分法等相关数据见正文（表 8）及支撑材料（表问题二）。

针对问题三，主要考虑新型冠状病毒对企业盈利额的负面作用，同时为消除数量纲对后续计算产生的影响，将企业盈利额从 2019 年 1 月 1 日到 2 月 21 日数据与 2020 年 1 月 1 日到 2 月 21 日数据作为新冠爆发前与新冠爆发后数据，见支撑材料（表问题三），即每个企业对应两个数据，使用对应分析找出受新冠影响较大与较小的企业。对这些企业给出更改过后的信贷策略，详见正文部分（表 9 及表 10）。

另外，针对上述三个问题给出误差分析。为改进问题二、三中所使用的模型，通过 BP 神经网络预测附件二中企业信誉评级与是否违约的情况，将附件一中企业盈利额、客户流失率、信誉评级以及是否违约作为训练集，附件二中企业盈利额、客户流失率作为测试集进行测试，最终测试结果见支撑材料（模型改进），并根据测试结果给出相应的信贷策略，见正文（表 11）。最后给出模型的评价及推广。

关键字：信贷策略聚类分析支持向量机 BP 神经网络对应分析

1 问题的提出

1.1 基本情况

2014 年 9 月 17 日，国务院总理李克强提出进一步扶持小微企业发展推动大众创业万众创新，他在会议中明确指出，小微企业是发展的生力军、就业的主渠道、创新的重要源泉。加大对小微企业、个体工商户特别是在改革中“呱呱坠地”新生者的扶持，让它们在公平竞争中搏击壮大，可形成示范效应，推动大众创业、万众创新，也能增添社会活力和发展内生动力，促进经济稳定增长和民生改善 [1]。各行各业积极配合，特别是国有大型银行及地方控股银行。

目前，银行如何确定是否放贷及贷款额度、利率和期限等信贷策略成为全部中小企业首要关注的问题。

1.2 问题提出

某银行对确定要放贷企业的贷款额度为 10 100 万元；年利率为 4% 15%；贷款期限为 1 年。根据三个附件给出的数据建立数学模型，研究对中小微企业的信贷策略，主要解决下列三个问题：

- * 对附件 1 中给出的 123 家企业的基本信息、进项与销项发票信息进行分析，给出该银行在年度信贷总额固定时对这些企业的信贷策略。
- * 在问题 1 的基础上，对附件 2 中 302 家企业的相关信息进行分析，给出该银行在年度信贷总额为 1 亿元时对这些企业的信贷策略。
- * 由于企业的生产经营和经济效益可能会受到一些突发因素影响，而且突发因素往往对不同行业、不同类别的企业会有不同的影响。因此结合突发因素所带来的影响给出该银行在年度信贷总额为 1 亿元时的信贷调整策略。

2 问题的分析

2.1 背景分析

目前国内中小型企业数量较为庞大，因此，有选择性的资助具有良好前景及较高信誉评级的企业成为国内全部银行的首要目标，从而信贷策略的决定被凸显的尤为重要。在国内大力扶植中小型企业发展的背景下，如何在给定信贷总额的前提下选择值得进行投资的企业成为本文首要的研究目标。

2.2 知识准备

- * 价税合计：销售货物行为中收入与增值税的合计值 [2]，计算公式为
税合计 = 金额 + 税额
- * 销项负数发票：又称负数发票（即红字发票），是企业发生销售货物退回等业务时，冲减销售收入的合法凭证 [3]
- * 贷款利率：是银行等金融机构发放贷款时向借款人收取利息的利率。主要分为三类：中央银行对商业银行的贷款利率；商业银行对客户的贷款利率；同业拆借利率 [4]

2.3 问题的分析

2.3.1 问题一的分析

将全部企业按照信用评级进行分类，并将虚拟变量是否违约进行数值处理，结合销项发票信息与进项发票信息中的价税合计之间的差值进行企业聚类分析（即行业分类），根据不同等级下的企业前景给出相应信贷策略

2.3.2 问题二的分析

在数据处理方面采取与问题一类似的方法，使用支持向量机对数据进行分组，再根据客户流失率对每组数据中的全部企业进行排名，最后将资金分发的数额与排名高低相互联系即可得出最终每个企业所得信贷投资资本

2.3.3 问题三的分析

由于疫情的突发影响，本文将附件二中 2020 年前与 2020 年后的数据进行分离，找出存在 2020 年后数据的全部企业，按照信用评级进行分组，计算出全部企业在疫情前与疫情后的获利情况，将所得数据进行标准化，使用对应分析寻找出受疫情影响较大且具有良好前景的企业，按照类似于问题二的信贷分配方法将资金基于疫情影响程度与企业盈利能力进行合理有效的分配

3 模型的假设

- * 假定计算所得数据之间一定存在某种相似性，从而能够对数据进行分类。

- * 假定支持向量机满足强对偶关系，即 $\min \max(x) = \max \min(x)$ 。
- * 假定所有不同种类数据间存在明显差别，即能够找到一组最优解使得全部数据 4 被有效的分开且不存在任何歧义
- * 假定所有企业之间相互独立，一家企业的决策不会对其他企业产生影响。
- * 假定疫情前与疫情后数据不存在显著的联系，即疫情前数据不会影响疫情后的数据。
- * 由于新型冠状病毒的爆发集中出现在 2020 年后，假定忽略 2019 年新冠的出现对企业盈利额的影响

4 符号表示

v

5 模型的建立

5.1 问题一的分析与求解

5.1.1 对附件一中虚拟变量重新赋值

根据题目所给条件，银行对信誉评级为 D 的企业原则上不予放贷。因此，在对所给数据进行数值化定义前，本文将全部信誉等级为 D 的企业进行删除。重新赋值定义结果如下表所示：

表 1: 虚拟变量的数值化

表中变量	变量取值	重新定义
是否违约	是	1
	否	0
发票状态	有效发票	0
	无效发票	1

5.2 对附件一中数据进行整理

使用 EXCEL 将信誉等级为 A、B、C 的企业数据进行分离，重新创建三个新工作表并保存，针对每一个企业保留三个数据，分别为：1) 是否违

约，值为 0 或 1；2) 每个企业销项发票信息与进项发票信息在价税合计方面的总差值，即可代表企业的盈利水平与前景；3) 在进项与销项信息中作废发票占全部发票的比例，即可代表附件三中的客户流失率，方便在最后针对不同企业给出合理的信贷策略。部分整理所得数据见下表：

表 2: 应用于聚类分析的处理数据

企业盈利额	是否违约	客户流失率	企业盈利额	是否违约	客户流失率	企业盈利额	是否违约	客户流失率
322971720.9	0	0.039297417	10737958.86	0	0.043357076	11332230.96	0	0.047305844
101466764.3	0	0.053434498	152360277.9	0	0.02931283	38266099.99	0	0.076633166
544555633.3	0	0.026875647	13050698.25	0	0.067645517	47297460.34	1	0.015238095
249298673.6	0	0.057100842	-6154738.43	0	0.053996824	36073520.22	0	0.056776557
358113695.7	0	0.023531719	12763898.47	0	0.032625279	31197561.24	0	0.040436616
165328536.7	0	0.065707367	57249825.33	0	0.077235772	30706314.46	0	0.045623145
229373468.6	0	0.060702875	58452951.94	0	0.043114543	5090820.96	0	0.059462395
237731387	0	0.068587106	46119242.76	0	0.029307282	12583897.06	0	0.038942581
45379929.42	0	0.065502183	2794992	0	0.13567282	10407111.66	0	0.039856765
69047648.49	0	0.056296778	371990244.73	0	0.11571618	15268735.06	0	0.036094961
9439313.26	0	0.068786502	2537538	0	0.047949965	19397326.97	0	0.05062241
58345221.86	0	0.090163934	3863593.57	0	0.060814384	5439742.03	0	0.082103321
79479771.23	0	0.048837817	40517628.68	0	0.06732531	16050178.8	0	0.15428167
2483251.49	0	0.067181467	17727501.83	0	0.081510934	506418.78	0	0.050707347
-1210601.15	0	0.04190429	9492988.95	0	0.046735905	6549738.19	0	0.105263158
51359666.57	0	0.03807947	8673451.29	0	0.105633803	4188126.85	0	0.108843337
29320343.4	0	0.088235294	25002771.04	0	0.082183908	2962027.98	0	0.203479452
45860508.41	0	0.035984007	6686641.43	0	0.092293054	6733962.72	0	0.022565321
31966258.11	0	0.072886297	18189205	0	0.01814882	4540320.65	0	0.065710872
10292187.01	0	0.036693847	10611112.38	0	0.041388518	3914272.99	0	0.022433897
9328410.21	0	0.065387349	18149899.37	0	0.042297418	7206381.03	0	0.0625
1789200.36	0	0.053488372	3828671.78	0	0.053477761	1274419.79	0	0.013544018
4808752.14	0	0.097326087	45312.23	0	0.042197432	2566267.58	0	0.025
2314961.47	0	0.058631922	5092019.45	0	0.049883991	2734756.68	1	0.024232561
-540296.53	0	0.037974684	6163820.68	0	0.062733693	4045127.59	0	0.195
1247099.22	0	0.049019608	823881.01	0	0.042829657	2350073.59	0	0.015873016
			3975119.4	0	0.074912892	1192567.49	0	0.09
			3635607.19	0	0.043697479	-856119.96	0	0.064516129
			530993.51	0	0.058333333	290250	0	0.090999991
			-37450137.73	0	0.068965317	954478.89	0	0.0078125
			2685298.78	0	0.090038314	274365.2	0	0.241186047
			1249350.17	0	0.033240997			
			2526484.45	0	0.053435115			

5.3 应用聚类分析将数据分组

5.3.1 数学模型

K-Means 聚类分析：使用 5.1.2 中已经处理好的数据代表每一个企业。将每个企业单独分为一组，随机选定其中的两个企业，使用欧氏距离计算两个企业与其他全部企业的距离，针对两个企业，分别选择距离它们最近的其他企业，形成两组包含两个企业的小组；对两个小组计算中心位置，使用中心位置坐标代表两组数据的值，通过计算中心位置坐标值与其他企业的距离，找出距离两组数据最近的企业点，重复上述做法，直到符合迭代停止的步骤。

数学模型如下：

i 欧氏距离公式： $d_{ij} = \sqrt{(s_i - q_i)^2 + (s_i - q_j)^2 + (s_i - q_j)^2}$ ，其中 $i, j = 1, 2, 3 \cdots 123$ 。

ii 本文将企业按照信誉等级分为三类，在此基础下，针对每一等级将企业分为前景良好值得投资与效益较差两类，使用三次 K - Means 聚类对数据进行分组，在聚类时 k 值取 2。

- iii 聚类分析目标：由于需将数据划分为两组，假设最终分类的两组为 T_1 和 T_2 ，每组数据中心为 t_1 和 t_2 。则有目标函数 $\min \sum_{s \in T_i} |s - t_i|^2$ ，其中 $i = 1, 2$ ，数据中也称为质心，表达式为 $t_i = \frac{1}{|T_i|} \sum_{S \in T_i} S$ ，其中 $i = 1, 2$ 。

iv

$$f_{(x,y)} = \begin{cases} x + y; & x > 10, y < 10 \\ 2x + 4y; & x < 10, y < 10 \\ 4x + 3y; & x < 10, y > 10 \\ x + 22y; & x > 10, y > 10; \end{cases}$$

参考了 [1]

参考文献

- [1] Zheng L, Wang S, Tian L, et al., Query-adaptive late fusion for image search and person re-identification, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1741-1750.
- [2] Arandjelović R, Zisserman A, Three things everyone should know to improve object retrieval, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012: 2911-2918.
- [3] Lowe D G. Distinctive image features from scale-invariant keypoints, International journal of computer vision, 2004, 60(2): 91-110.
- [4] Philbin J, Chum O, Isard M, et al. Lost in quantization: Improving particular object retrieval in large scale image databases, Computer Vision and Pattern Recognition, 2008. CVPR 2008, IEEE Conference on, IEEE, 2008: 1-8.