Advances in Applied Mathematics 应用数学进展, 2018, 7(1), 10-19
Published Online January 2018 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2018.71003

Hans汉斯

Rating of Small and Micro Businesses Based on SVM

Chao Li, Haihui Wang

Beihang University, Beijing
Email: lichao1509118@buaa.edu.cn, whh@buaa.edu.cn

Received: Dec. 15th, 2017; accepted: Jan. 11th, 2018; published: Jan. 18th, 2018

Abstract

With the transformation of the economy, small and micro businesses have injected new vitality into the economic development of our country. In order to enhance the risk management level of small and micro businesses and to improve their survival rate, effective risk assessment must be carried out. We can use the method of support vector machine (SVM) for the risk rating of small and micro businesses. Based on the established risk assessment index system, we can evaluate the risk level of some small and micro businesses through SVM classification training on selected sample data. Therefore, it is convenient for the enterprise managers to take effective risk control measures to promote the benign development of small and micro businesses according to the risk level.

Keywords

Small and Micro Businesses, Risk Rating, Support Vector Machine, Classification Training

基于SVM的小微企业评级

李　超，王海辉

北京航空航天大学，北京
Email: lichao1509118@buaa.edu.cn, whh@buaa.edu.cn

收稿日期：2017年12月15日；录用日期：2018年1月11日；发布日期：2018年1月18日

摘　要

随着经济不断转型，小微企业为我国经济发展注入了新的活力。为了增强小微企业的风险管理水平，从而提高其成活率，就必须进行有效的风险评价。对我国小微企业进行风险评级，可以运用支持向量机

文章引用: 李超, 王海辉. 基于 SVM 的小微企业评级[J]. 应用数学进展, 2018, 7(1): 10-19.
DOI: 10.12677/aam.2018.71003

(SVM)方法，并依据建立的风险评价指标体系，通过对选取的样本数据进行SVM分类训练，评估我国部分小微企业风险等级水平。从而方便企业管理者根据风险水平，采取切实有效风险控制措施，促进小微企业良性发展。

关键词

小微企业，风险评级，支持向量机，分类训练

1. 引言

近年来，我国股权众筹行业高速发展，项目数量和平台数量都大幅增长。股权众筹在增加投资者理财方式的同时，更完善了小微企业的融资渠道，从而提高了金融市场的效率，促进了我国创新创业，推动了经济结构转型。股权众筹虽然有优势有活力，可以降低获取信息的成本，但也不可避免地存在着虚假信息，各利益相关者信用风险问题很大，所以股权众筹行业乱象丛生[1]。尤其是这两年互联网众筹平台跑路、倒闭、作假事件屡屡发生，互联网金融行业的信用风险不容乐观。如何对股权众筹信用风险进行分析，并采取量化的方法对其进行风险评级，进而提高小微企业抵御风险的能力，已成为学术界、工业界研究的重点[2]。

从过去的研究来看，对小微企业风险的研究大部分集中在定性分析上，即使涉及到定量方法，也多是层次分析法、模糊评价法等传统方法。这些方法不适合股权众筹平台这种影响因素繁多复杂的系统，而且评价结果的可靠性和精确性无法保证。随着互联网技术的不断创新发展，风险评价也向着标准化、系统化和精确化方向推进，出现了新的将计算机模拟与风险评价相结合的方法。支持向量机法(SVM)作为一种备受推崇的数据挖掘技术，已经在文本识别、人脸识别、银行信用风险评估等诸多实践领域得到广泛应用[3]。本文将 SVM 方法引入我国小微企业风险评级中，以提高风险评价精度和效率。

2. 构建小微企业风险评级指标体系

小微企业所面临的风险因素多而复杂，本文从科学、全面、量化和可操作等角度出发，建立一套科学合理的风险指标体系。根据小微企业的特点，将风险评价指标体系分为 5 个一级指标和 18 个二级指标，指标体系具体构建为行业市场环境、产品竞争力、企业基本情况、管理团队素质和财务数据表现五个一级指标，其相应的二级指标以及与一级指标的关系如图 1。

3. SVM 的基本原理

支持向量机(SVM)核心思想是建立一个最优分类线或最优超平面作为决策曲面，将多类样本正确地进行分类。即先通过一定量的样本进行训练，通过不断检验与优化得到较高的训练精度，确定一个最优的决策函数再对分类问题进行处理。我们以二分类问题为例分析一下具体过程[4]。

3.1. 线性可分问题

图 2 中分别是两类样本，支持向量机方法就是寻找最优分类线将两类样本分开，分类线表示为

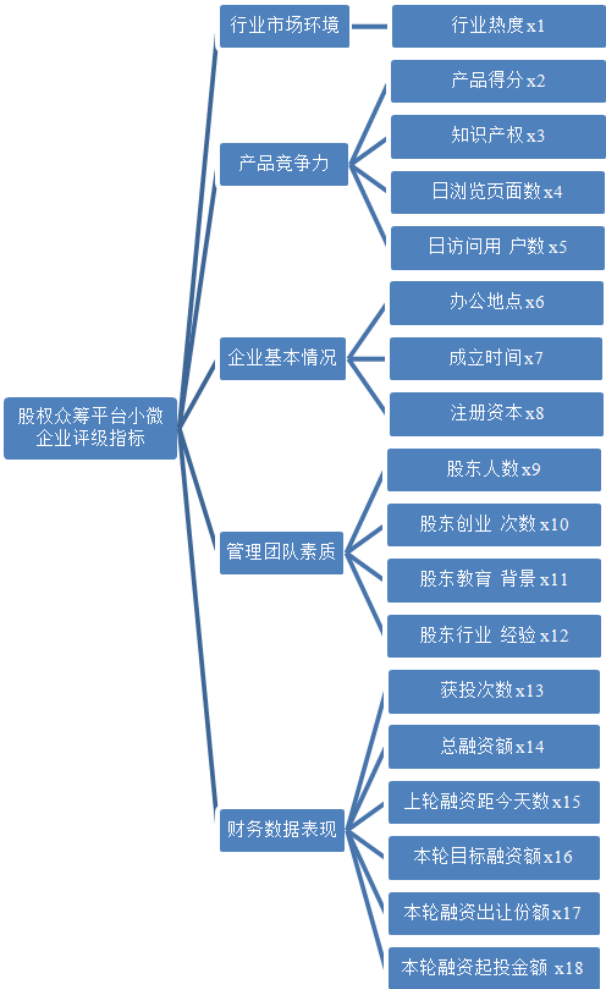

Figure 1. Risk rating system for small and micro businesses in China

图 1. 我国小微企业风险评级指标体系

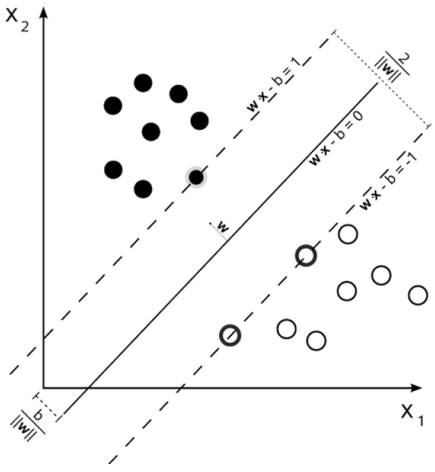

Figure 2. Finding the optimal classification line for two types of samples

图 2. 寻找两类样本的最优分类线

$(w\cdot x)+b=0$，最优分类线代表着最大分类间隔，可用如下最优化问题表示：

$$\min_{w,b}\quad \frac{1}{2}\|w\|^2 \tag{1}$$

$$\text{s.t.}\quad y_i\left(\left(w\cdot x_i\right)+b\right)+1\geq 1, i=1,\cdots,l \tag{2}$$

其中的约束要求各数据点(x_i,y_i)到分类面的距离大于等于1。其中，y_i为数据的分类。

引入Lagrange函数可将上述问题转化为对偶问题：

$$\min_{\alpha}\quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_i y_j\alpha_i\alpha_j\left(x_i\cdot x_j\right)-\sum_{j=1}^{l}\alpha_j \tag{3}$$

$$\text{s.t.}\quad \sum_{i=1}^{l}y_i\alpha_i=0 \tag{4}$$

$$\alpha_i\geq 0 \tag{5}$$

求解上述问题得到最优解：

$$w^*=\sum_{i=1}^{l}y_i a_i^* x_i,\ b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i\left(x_i\cdot x_j\right) \tag{6}$$

将参数代入原式，即可得到最优分类平面。

3.2. 线性不可分问题

对于线性不可分问题，原来对间隔的要求不能达到。引入松弛变量ξ_i，使约束条件弱化为：$y_i\left(\left(w\cdot x_i\right)+b\right)+1\geq 1-\xi_i$。但是，我们仍然希望该松弛变量$\xi_i$最小化。于是，在优化目标函数中使用惩罚参数$C$来引入对$\xi_i$最小化的目标。此时模型为：

$$\min_{w,b}\quad \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i \tag{7}$$

$$\text{s.t.}\quad y_i\left(\left(w\cdot x_i\right)+b\right)+1\geq 1-\xi_i, i=1,\cdots,l \tag{8}$$

以此为原问题，其对偶问题为：

$$\min_{\alpha}\quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_i y_j\alpha_i\alpha_j\left(x_i\cdot x_j\right)-\sum_{j=1}^{l}\alpha_j \tag{9}$$

$$\text{s.t.}\quad \sum_{i=1}^{l}y_i\alpha_i=0 \tag{10}$$

$$0\leq\alpha_i\leq C \tag{11}$$

求解得到最优解为：

$$w^*=\sum_{i=1}^{l}y_i a_i^* x_i,\quad b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i\left(x_i\cdot x_j\right) \tag{12}$$

3.3. 非线性问题

对于非线性问题，可以将低维空间中的曲线(曲面)映射为高维空间中的直线或平面。数据经这种映射后，在高维空间中是线性可分的。设映射为$x'=\phi(x)$，则高维空间中的线性支持向量机模型为：

$$\min_{\alpha}\quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_i y_j\alpha_i\alpha_j\left(\phi\left(x_i\right)\cdot\phi\left(x_j\right)\right)-\sum_{j=1}^{l}\alpha_j \tag{13}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \tag{14}$$

$$0 \le \alpha_i \le C \tag{15}$$

由于数据被映射到高维空间，$\phi(x_i)\cdot\phi(x_j)$的计算量比$x_i \cdot x_j$大得多。此时引入了“核函数”：

$$K(x_i, x_j) = \phi(x_i)\cdot\phi(x_j) \tag{16}$$

由上式可见，核函数的作用是，在将x映射到高维空间的同时，也计算了两个数据的在高维空间的内积，使计算量回归到$x_i \cdot x_j$的量级。

4. 基于 SVM 的风险评级

4.1. 收集并处理样本数据

根据图 1 风险评价指标体系搜集相关的样本数据，本文选取 130 家小微企业作为研究对象，并搜集他们 2017 年的原始数据资料，部分企业数据见表 1。

观察数据集中每一个评级的数量分布，结果见表 2。

可见，各个评级的样本分布相对均衡。

Table 1. Original data of the first 10 enterprises
表 1. 前 10 家企业原始数据

	X1	X2	X3	X4	X5	X6	X7	X8	X9	
coolook	0.35	241	8	2200	100	1	36	13.38	6	
e 代泊	0.75	805	10	320	90	2	46	292.83	10	
看孩子	0.46	487	6	30	20	1	22	125.08	13	
骑乐无穷	0.47	379	19	3800	580	2	51	114	2	
潮牌 pr	0.47	85	7	140	50	8	23	100	4	
悟空音乐	0.57	185	27	300	60	3	32	200	5	
脉圈	0.34	461	9	450	88	1	27	119	3	
慧金天下	0.56	484	2	450	90	2	30	5649.72	7	
羽乐圈	0.57	553	27	8000	66	4	37	123.21	8	
聘宝	0.35	513	4	2000	400	7	44	488.81	6	
	X10	X11	X12	X13	X14	X15	X16	X17	X18	评级
coolook	3	6	12	0	0	0	830	0.020	4	1
e 代泊	6	8	17	4	7800	258	1340	0.069	4	1
看孩子	1	8	12	1	300	626	350	0.090	4	0
骑乐无穷	4	6	10	0	0	0	500	0.100	3	2
潮牌 pr	2	8	8	0	0	0	300	0.070	2	2
悟空音乐	3	6	6	2	1243	604	350	0.100	2.5	2
脉圈	5	8	13	1	500	686	200	0.035	3	1
慧金天下	3	9	10	1	500	755	400	0.074	3.5	1
羽乐圈	4	6	6	0	0	0	300	0.130	2	1
聘宝	1	9	12	1	500	504	600	0.170	3	1

我们进一步观察每一个指标的取值分布，见表 3。

观察发现数据取值范围差异很大，因为 SVM 要用到距离，所以必须做标准化处理，这里采用 Min-Max 标准化方法处理，结果见表 4。

可以看到所有数据都在 0~1 之间，可以进行后续的距离计算。

4.2. 组建样本训练集，训练模型

把前 91 家公司数据作为训练集，后 39 家公司数据作为测试集，用 SVM 来进行模型构建。SVM 有两个主要的参数可以设置：核函数参数 kernel 和约束惩罚参数 C。其中约束惩罚函数 C 为对超过约束条件的样本的惩罚项。C 值越大，惩罚越大，支持向量机的决策边界越窄。我们选用最简单的线性核函数，C 采用 200，训练得到最初的模型。

4.3. 模型性能评估

首先，要得到我们训练的模型在测试集上的预测结果，然后对模型的性能进行评估。模型准确率、

Table 2. Rating distribution
表 2. 各评级分布

0	58
1	46
2	26

Table 3. Statistical data of various indicators
表 3. 各项指标统计数据

	mean	std	min	25%	50%	75%	max
X1	0.524461538	0.114513203	0.34	0.46	0.56	0.58	0.75
X2	458.6769231	161.6385198	85	379	484	548	805
X3	12.09230769	16.05776692	0	3	8	15	113
X4	1907.830769	7455.064861	11	267	450	1000	60,000
X5	548.2153846	2505.925452	11	80	110	200	20,000
X6	2.461538462	1.905332566	1	1	2	4	8
X7	30.47692308	16.60494714	3	21	28	37	102
X8	654.3972308	1213.803404	1	114	200	583.33	5649.72
X9	5.015384615	3.026970792	1	3	4	7	13
X10	2.123076923	1.335359224	0	1	2	3	6
X11	7.169230769	1.359083489	5	6	7	8	10
X12	9.661538462	3.37007774	3	7	10	12	22
X13	1.292307692	1.123713072	0	0	1	2	4
X14	3602.753846	7914.165782	0	0	500	3300	37,800
X15	318.7230769	280.2317152	0	0	270	525	1087
X16	1134.692308	1788.638536	100	300	600	1500	13,000
X17	0.330229231	1.873030315	0.0033	0.0667	0.1	0.12	15.25
X18	3.769230769	2.105873887	1	2	3	4.5	10

Table 4. Statistical data of various indicators after standardization
表 4. 标准化后各项指标统计数据

	mean	std	min	25%	50%	75%	max
X1	0.449906191	0.279300496	0	0.292682927	0.536585366	0.585365854	1
X2	0.518995726	0.224497944	0	0.408333333	0.554166667	0.643055556	1
X3	0.107011572	0.142104132	0	0.026548673	0.07079646	0.132743363	1
X4	0.031619643	0.124273865	0	0.004267449	0.007318008	0.016486356	1
X5	0.026875551	0.125365223	0	0.003451899	0.004952724	0.0094552	1
X6	0.208791209	0.272190367	0	0	0.142857143	0.428571429	1
X7	0.277544678	0.167726739	0	0.181818182	0.252525253	0.343434343	1
X8	0.115671733	0.214881142	0	0.020004532	0.03522922	0.103090612	1
X9	0.334615385	0.252247566	0	0.166666667	0.25	0.5	1
X10	0.353846154	0.222559871	0	0.166666667	0.333333333	0.5	1
X11	0.433846154	0.271816698	0	0.2	0.4	0.6	1
X12	0.350607287	0.177372513	0	0.210526316	0.368421053	0.473684211	1
X13	0.323076923	0.280928268	0	0	0.25	0.5	1
X14	0.095310948	0.209369465	0	0	0.013227513	0.087301587	1
X15	0.293213502	0.257802866	0	0	0.248390064	0.482980681	1
X16	0.080208706	0.13865415	0	0.015503876	0.03875969	0.108527132	1
X17	0.021442622	0.122848244	0	0.004158277	0.006342356	0.007654115	1
X18	0.307692308	0.233985987	0	0.111111111	0.222222222	0.388888889	1

召回率和 f1-score 见表 5。

具体的评级预测结果见表 6。

该混淆矩阵中对角线的元素表示模型正确预测数，对角元素之和表示模型整体预测正确的样本数。而非对角线元素上的值则可以反映模型在哪些类的预测上容易犯错，例如评级 2 均有三次被预测为 0 和 1。

最终分类正确率：0.866666666667。

4.4. 模型性能提升

各项指标可以进一步提高，我们进行参数调优，可以改善模型性能，进一步提升模型效果。

首先是核函数，通过循环依次取不同的核函数，准确率见表 7。

可以看到，最好的是“rbf”这个核函数，可将正确率提升至：0.883333333333。

核函数确定后，我们再来看 C 的取值。第一次给定这样一个列表：c_list = [0.01,0.1,1,10,100,500]，循环依次取值，准确率见表 8。

可以看到，在 C = 500 左右模型效果提升很大，可以进一步探究 C = 500 左右的正确率变化情况。所以给定这样一个列表 c_list2 = [100,200,300,400,500,600]，循环依次取值，准确率见表 9。

可以发现 C = 500 时模型效果确实不错，可将正确率提升至：0.933333333333，评价指标及混淆矩阵见表 10 和表 11。

为了进一步提升模型的效果，我们重点关注惩罚因子 C 是否会有更优的取值。因为求 SVM 最优参

Table 5. Model evaluation index
表 5. 模型评价指标

	precision	recall	f1-score	support
0	0.85	0.96	0.9	24
1	0.88	0.96	0.92	23
2	0.88	0.54	0.67	13
avg/total	0.87	0.87	0.86	60

Table 6. Model rating results
表 6. 模型评级结果

	0	1	2
0	23	0	1
1	1	22	0
2	3	3	7

Table 7. The corresponding accuracy of each kernel function
表 7. 各核函数对应准确率

kernel	Accuracy
linear	0.866667
rbf	0.883333
poly	0.633333
sigmoid	0.4

Table 8. The corresponding accuracy of each C value
表 8. 各 C 值对应准确率

C	Accuracy
0.01	0.4
0.1	0.4
1	0.433333
10	0.65
100	0.816667
500	0.933333

Table 9. The corresponding accuracy of each C value
表 9. 各 C 值对应准确率

C	Accuracy
100	0.816667
200	0.883333
300	0.9
400	0.916667
500	0.933333
600	0.933333

Table 10. Model evaluation index
表 10. 模型评价指标

	precision	recall	f1-score	support
0	0.92	0.96	0.94	24
1	0.96	0.96	0.96	23
2	0.92	0.85	0.88	13
avg/total	0.93	0.93	0.93	60

Table 11. Model rating results
表 11. 模型评级结果

	0	1	2
0	23	0	1
1	1	22	0
2	1	1	11

数本质上是一个二次凸规划问题，即求解函数最优值，而遗传算法最主要的用处就是函数最优化。因为遗传算法是一种高效的随机搜索算法，而且克服了诸如网格搜索法等容易陷入局部最优的缺点，可以通过多次尝试，找到全局最优解。同时遗传算法也会同时考虑 kernel 参数，避免模型出现过拟合现象[5]。我们基于遗传算法的 SVM 算法如下：

步骤 1：初始化种群，随机生成初始种群个体；

步骤 2：将种群中各个体基因串解码为相应核函数编号、核函数参数和错误惩罚因子，并将参数代入 SVM，以训练数据和测试数据对其进行训练和测试；

步骤 3：按照适应度计算法则，计算每个个体的适应度值；

步骤 4：判断是否满足终止条件，如果满足终止条件，退出循环，遗传优化结束，得到优化参数组合，否则转到步骤 5；

步骤 5：执行选择算子，按照最优保存、最差取代的原则进行；

步骤 6：执行交叉算子和变异算子，交叉概率取 0.7，变异概率取 0.1，形成新一代个体后，返回步骤 2 继续执行。

我们用这个模型来进行参数优化，最终选取核函数“rbf”以及惩罚因子 C = 700，评价指标及混淆矩阵见表 12 和表 13。

正确率可以达到：0.966666666667。

4.5. 模型对照

为了更好的说明模型效果，我们以逻辑回归模型作为对比，且选择调优后的参数，这里 C = 1000，penalty = “l1”，solver = “liblinear”，评价指标及混淆矩阵见表 14 和表 15。

正确率可以达到：0.85，效果不如 SVM 模型。

5. 结论

通过对我国小微企业风险评级的研究，我们发现：

1) 与逻辑回归方法相比，SVM 方法在准确率等各项评价指标都明显更优，而且可以通过参数调优不断提升指标，从而更好的对企业风险进行评级。该方法对小微企业风险评价与预测这一复杂的领域有

Table 12. Model evaluation index
表 12. 模型评价指标

	precision	recall	f1-score	support
0	1	0.96	0.98	24
1	0.96	1	0.98	23
2	0.92	0.92	0.92	13
avg/total	0.97	0.97	0.97	60

Table 13. Model rating results
表 13. 模型评级结果

	0	1	2
0	23	0	1
1	0	23	0
2	0	1	12

Table 14. Model evaluation index
表 14. 模型评价指标

	precision	recall	f1-score	support
0	0.92	1	0.96	24
1	0.78	0.91	0.84	23
2	0.86	0.46	0.6	13
avg/total	0.85	0.85	0.84	60

Table 15. Model rating results
表 15. 模型评级结果

	0	1	2
0	24	0	0
1	1	21	1
2	1	6	6

很重要的指导作用。

2) 对 SVM 参数直接调优可以得到更好的效果，本文引入了遗传算法进行参数调优，可以得到最优化的结果。

参考文献 (References)

[1] 侯合银. 高新技术创业企业风险的系统分析: 辨识与规避[J]. 科技管理研究, 2008, 28(10): 132-135.

[2] 袁莉, 李宏男. 基于 SVM 的建筑企业信用评价研究[J]. 价值工程, 2009, 28(3): 141-144.

[3] 姚奕, 叶中行. 基于支持向量机的银行客户信用评估系统研究[J]. 系统仿真学报, 2004, 16(4): 783-786.

[4] 章兢, 张小刚. 数据挖掘算法及其工程应用[M]. 北京: 机械工业出版社, 2006.

[5] Li, L.M., Wen, G.R. and Wang, S.C. (2008) Parameters Selection of Support Vector Regression Based on Genetic Algorithm. *Computer Engineering and Applications*, **44**, 23-26.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2324-7991，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：aam@hanspub.org