基于 Lasso 和支持向量机的上市公司信用评价

滕树军　天津商业大学理学院
刘丽平　东北财经大学统计学院
刘柏森　东北财经大学统计学院

摘要：随着经济的全球化，作为市场经济交易基础的公司信用研究，已趋于社会化、普遍化。信用关系或者债券关系已经成为一种非常基本的经济关系。而在公司交易规模不断壮大的同时，信用风险也随之而来。本文首先采用 Lasso 方法从可能影响上市公司信用评价的众多财务指标中挑选出现金比率、资产负债率、长期资本负债率、固定资产比率等 17 个重要影响因素，然后再运用支持向量机方法对上市公司信用评价进行预测。实际研究结果表明本文所提出的 Lasso 与 SVM 相结合的新方法的拟合预测效果要优于单纯 SVM 方法的预测效果。

关键词：Lasso；支持向量机；信用评价
中图分类号：F830.91　　文献识别码：A　　文章编号：1001-828X(2018)018-0022-03

一、引言及文献综述

随着经济的全球化，作为市场经济交易基础的公司信用问题，已日渐趋于社会化、普遍化。现代市场经济是建立在信用基础上的经济，从某种意义上说市场经济就是信用经济。在资本市场快速发展的过程中，上市公司在我国经济发展中起着重要作用，已经成为我国国民经济发展的中坚力量。截至 2016 年 12 月，我国沪市和深市上市公司总数量达到 3025 家，总市值达到 508245 亿元，与 2016 年我国 GDP 的比值为 68%。上市公司是我国信贷市场中商业银行的主要授信主体，也是我国资本市场上股票和债券的主要融资主体。商业银行已经把信用风险列为经营管理中所面临风险中的首要风险，同样，作为资本市场上十分重要的融资主体，上市公司如果发生失信事件，将会在资本市场中产生更加剧烈与重大的影响。所以对上市公司进行信用评价，可以使投资主体能够更准确地评价被授信公司的信用状况，有效地减少投资者所面临的投资风险，从而做出准确的判断。

国外很早就对公司信用风险评定展开了研究，并将其研究结果广泛应用于银行、企业及投资机构等。从最开始的借助于专家的经验来评判公司信用情况，到 20 世纪 70–80 年代，发展到以公司财务指标为基础来进行公司信用风险的评定。Beaver[1](1967) 将判别分析方法引入到信用风险分析中，美国学者 Altman[2](1968) 将一元判别模型扩展为多元判别模型。随着不断的研究，Altman、Haldeman 和 Narayanan[3](1977) 将 Z–score 模型进行优化，最终建立了 Zeta 判别分析模型。亚洲金融风暴之后，全世界又兴起了打破旧的信用风险分析方法，随着计算机的快速发展，机器学习理论被广泛应用到企业风险评估当中，主要方法有神经网络、支持向量机 (SVM) 等。

国内对公司信用风险评价的研究要晚一些，应用的方法主要有 Logistic 回归、KMV 与 Logistic 模型的结合、多元自适应回归样条 (MARS) 和支持向量机。胡安冉和孙云 [4](2012) 利用 2010 年股票市场上 6 家 ST 公司以及 4 家已经上市并正常运转的公司财务报表的数据为研究素材，建立了 Logistic 模型，评价了上市公司的信用风险，并验证了其模型的适用性，总体预测准确率为 88%。梁琪 [5](2005) 运用主成分分析法与 logistic 回归分析相结合的方法，对我国沪深两市上市公司的经营失败进行了实证研究，结果表明该方法在模型解释和预测准确率等方面均优于简单的 Logistic 模型分析。孙森和王玲 [6](2014) 利用 KMV 模型计算得到违约距离 (DD)，并将 DD 值与 Z–score 模型中的五个参数作为自变量引入 Logit 模型中，实现 KMV 模型与 Logit 模型的结合，得到了能够评估企业违约可能性的二元选择 Logit 模型，在沪市制造业违约可能性的评估中得到了较为理想的结果。彭颖 [7](2012) 在研究企业信用评估模型研究中，利用上市企业的财务数据，设计了信用分析的指标体系，利用多元自适应回归样条 (MARS) 方法对企业的信用状况建立信用评估模型，依据上市公司 2008 年的财务数据建立 MARS 模型，并与 Logistic 模型进行对比，发现 MARS 模型拟合精度及预测能力均强于 Logistic 模型。

近些年来，SVM 方法已被广泛应用于上市公司财务信用评价预测方法研究中，石秀福 [8](2008) 利用高斯核函数的 SVM 建立上市公司财务风险评价模型，从上市公司 13 个主要财务资料中选出部分指标，建立了 42 种财务风险评价预测模型，并利用这 42 种模型对评估预测精度进行比较研究，说明了基于高斯核函数的 SVM 在上市公司进行财务风险评价预测的优越性。还有其它文献也利用 SVM 方法研究中国上市公司的风险，通过对上市公司的财务比率进行建模和仿真研究，发现 SVM 方法对所选取的样本具有很好的分类效果，在上市公司的风险预测方面具有很强的准确性和可行性。

虽然 SVM 方法比较适合处理具有非线性关系的小样本数据，但当解释变量较多时，SVM 的预测精度不高，因而本文提出 Lasso 方法与 SVM 相结合的方法。首先利用 Lasso 方法对上市公司信用评价的影响因素进行变量选择，剔除对上市公司信用评价不显著的财务指标，从而实现降低数据维度的目的；然后利用支持向量机的非线性运算能力，完成对上市公司信用评价的拟合和预测。实际研究结果表明，这种新的 Lasso–SVM 方法的预测能力要高于直接运用 SVM 方法的预测能力，对于上市公司信用评价问题，有着较好的预测效果。

二、理论准备

1. 基于 Lasso 方法的变量选择

变量选择主要是通过统计方法从繁多的变量中选出对响应变量有很大影响的解释变量，变量选择的结果的好坏严重地影响着所建模型的质量，进而对统计预测精度产生较大的影响。传统的变量选择方法有逐步回归法、AIC 准则、BIC 准则、准则等，其本质上是子集选择法，其特点是无序性和离散性，在选择的过程中，有一些变量被模型剔除，有一些变量被模型选择，当解释变量较多时，子集选择方法的方差通常较高，不能达到降低模型预测误差的目的。

Tibshirani[9](1996) 于 1996 年给出了基于惩罚函数思想的 Lasso

方法，通过给模型参数增加范数的惩罚函数，对系数进行压缩。因为该模型是通过调整参数来选择变量，因此变量的收缩是连续的。该方法的特点是既通过参数估计来进行变量选择，又通过参数连续变化来调整变量连续收缩，自动地选择变量，因而被广泛应用于高维数据的回归分析中。

2. 支持向量机方法

支持向量机是数据挖掘中的一项新技术，是借助于最优化方法来解决机器学习问题的新工具，在解决小样本、非线性及高维度模式识别中表现出许多优势。它的核心是引入该映射的思想与结构风险的概念，通过寻求结构化风险最小来提高学习机的泛化能力，实现经验风险和置信范围的最小化，从而在样本数量较少的情况下，仍能获得良好统计规律的目的，目前该方法已经广泛应用于经济、金融、工程等领域。

三、建模与实证分析

1. 样本数据的选取与处理

本文选择上市公司财务指标来研究企业信用风险，并用被特殊标记 (ST) 的公司作为信用不佳的公司，未被标记 ST 的公司作为信用良好的公司。本文从国泰安数据库中搜集到的数据为 2016 年 1 月份至 12 月份我国沪市和深市中所有上市公司的财务指标数据，其中信用不佳的公司有 130 家，信用良好的公司有 2895 家。对于信用良好的公司，因为其公司数量非常多，而存在缺失值的观测相对较少，因此在对信用良好公司的数据集进行缺失值处理时，本文选择剔除存在缺失值的观测以保证数据的完整性；对于信用不佳的公司，因其数据量有限，本文在处理缺失值时，除删除无任何记录的公司外，其余缺失值选择用信用不佳的公司去除缺失值后的平均值来代替。经处理后的数据集有 113 家信用不佳的公司，有 2664 家信用良好的公司。为了保证数据的平衡性，本文按照 1 : 1 的比例随机抽选信用良好和信用不佳的上市公司，共选择 226 家公司，并从中随机选取了 80 家信用良好的公司与 80 家信用不佳的公司作为试验集，用于建立模型，剩下的 33 对公司作为测试集，用来检验模型效果。

一般而言，企业财务状况与企业信用风险之间存在密切的联系，财务状况的每一个微小的变化都可能对公司产生影响。当公司财务状况良好时，其现金流量控制良好，资本运营通畅，这时公司信用风险相对较小，按时还款的可能性较大。反过来，如果公司财务状况不佳，企业运作、经营都处于不佳状态，很可能出现失信行为。本文研究企业信用风险以及构建模型预测信用风险，选择有代表性的、全面的财务指标作为分析对象。因此，本文选择了涵盖偿债能力、比率结构、盈利能力、经营能力、现金流情况、发展能力以及相对价值这七方面的财务指标作为分析对象 (见表 1)。

表 1　上市公司财务指标体系

指标类别	指标名称	计算公式
偿债能力	X_1：流动比率	流动资产 / 流动负债 ×100%
	X_2：速动比率	速动资产 / 流动负债 ×100%
	X_3：现金比率	现金及现金等价物期末余额 / 流动负债 ×100%
	X_4：利息保障倍数	(利润总额 + 利息费用)/ 利息费用 ×100%
	X_5：资产负债率	负债总额 / 资产总额 ×100%
	X_6：长期资本负债率	非流动负债 / 长期资本 ×100%
比率结构	X_7：流动资产比率	期末流动资产 / 期末所有者权益 ×100%
	X_8：固定资产比率	固定资产 / 资产总额 ×100%
	X_9：流动负债比率	流动负债总额 / 总负债 ×100%
	X_{10}：金融负债比率	(非流动负债合计 + 短期借贷 + 一年内到期的非流动负债 + 交易性金融负债 + 衍生金融负债)/ 负债合计 ×100%
盈利能力	X_{11}：资产报酬率	(利润总额＋财务费用) / 资产总额 ×100%
	X_{12}：投资收益率	投资收益率 =1/ 动态投资回收期 ×100%
	X_{13}：总资产净利润率	净利润 / 总资产余额 ×100%
	X_{14}：投入资本回报率	(净利润 + 财务费用) / 投入资本 ×100%
	X_{15}：长期资本收益率	(利润总额 + 财务费用) / 长期资本额 ×100%
	X_{16}：营业毛利率	(营业收入—营业成本) / 营业收入 ×100%
经营能力	X_{17}：应收账款周转率	营业收入 / 应收账款期末余额 ×100%
	X_{18}：存货周转率	营业成本 / 存货期末余额 ×100%
	X_{19}：流动资产周转率	营业收入 / 流动资产期末余额 ×100%
	X_{20}：固定资产周转率	营业收入 / 固定资产平均净额 ×100%
现金流情况	X_{21}：营业收入现金比率	(经营活动产生的现金流量净额)/(营业总收入)×100%
	X_{22}：现金适合比率	一定时期经营活动产生的现金净流量 /(同期资本支出 + 同期存货净投资额 + 同期现金股利)100%
	X_{23}：营运指数	经营活动现金净流量 / 经营应得现金 ×100%
发展能力	X_{24}：可持续增长率	销售净利率 × 总资产周转率 × 利润留存率 × 权益乘数
	X_{25}：资本积累率	(所有者权益合计本期期末值—所有者权益合计本期期初值)/ 所有者权益合计本期期初值)×100%
	X_{26}：固定资产增长率	本期净增固定资产原值 / 期初固定资产原值 ×100%
	X_{27}：利润总额增长率	(利润总额本年本期单季度金额—利润总额上一个单季度金额)/(利润总额上一个单季度金额)×100%
	X_{28}：总资产增长率	(资产总计本期期末值—资产总计本期期初值)/(资产总计本期期初 ×100%
相对价值指数	X_{29}：市盈率	今收盘价当期值 /(净利润上年年报值 / 实收资本本期期末值)×100%
	X_{30}：市销率	今收盘价当期值 /(营业总收入上年年报值 / 实收资本本期期末值)
	X_{31}：市现率	今收盘价当期值 /(经营活动产生的现金流量净额上年年报值 / 实收资本本期期末值)×100%

* 表中公式来自国泰安数据库

2. 基于 Lasso 回归的变量选择与预测

我们拟使用统计中常用的一类精度，二类精度和总精度三个评价规则来度量各个模型的最终判别效果和预测能力，这三个评价规则定义如下：

一类精度＝信用良好公司被模型正确判为信用良好公司的数量 / 实际信用良好公司数量；

二类精度＝信用不好公司被模型正确判为信用不好公司的数量 / 实际信用不好公司数量；

总精度＝实际信用良好或信用不好公司被模型正确判别的数量 / 被测样本总数量。

3.SVM 支持向量机方法

我们首先运用 SVM 方法对上市公司的财务数据进行分析，此过程可由 R 软件中的 e1071 程序包来实现，参数自动寻优结果为：

best gamma =0.5，cost=4，R^2=66.67%

将训练集数据和测试集数据分别代入模型进行检验，最终得到结果如表 2 所示：从模型解释性与预测精度中，可以看出 SVM 方法在训练集的精度虽然都达到 100%，但在测试集里的一类精度仅为 36.37%，因而总体拟合效果不是很理想。

表 2　上市公司财务数据的分析结果

	SVM 方法		
	一类精度	二类精度	总精度
训练集	100%	100%	100%
测试集	36.37%	96.99%	66.67%
综合精度	81.42%	99.12%	90.27%
	Lasso-SVM 方法		
	一类精度	二类精度	总精度
训练集	98.75%	97.50%	98.13%
测试集	81.82%	87.88%	84.85%
综合精度	93.81%	94.69%	94.25%

4.Lasso-SVM

本文首先把数据进行中心标准化处理，以消除不同量纲的影响，然后利用 R 软件的 Glmnet 程序包，实现通过 Lasso 方法对 Logistic 回归模型进行变量选择。运用广义交叉验证方法，可以得到惩罚参数与变量个数的关系图 (图 1)，该图的横坐标表示惩罚参数值的变化，纵坐标表示模型误差的变化情况，并在图上方给出随着值的变化进入模型的变量个数的变化。当的取值为左侧虚线对应的值时，模型误差最小。

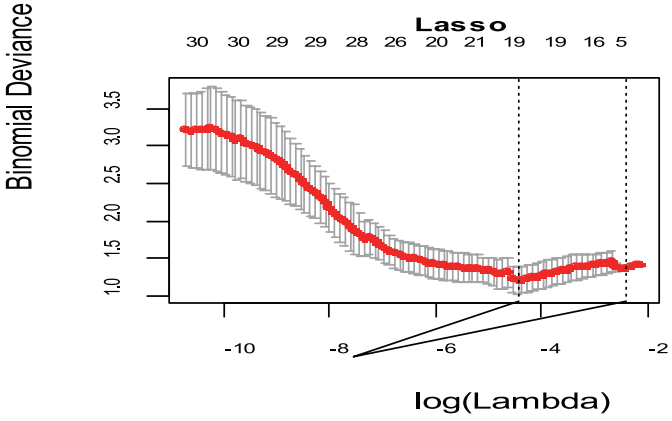

图 1 解释变量个数随的变化走势图

由图 1，我们最终选取了 17 个财务指标：现金比率 (X_3)、资产负债率 (X_5)、长期资本负债率 (X_6)、固定资产比率 (X_8)、流动负债比率 (X_9)、金融负债率 (X_{10})、投资收益率 (X_{12})、长期资本收益率 (X_{15})、营业毛利率 (X_{16})、应收账款周转率 (X_{17})、存货周转率 (X_{18})、流动资产周转率 (X_{19})、固定资产周转率 (X_{20})、营业收入现金比率 (X_{21})、总资产增长率 (X_{28})、市盈率 (X_{29}) 和市现率 (X_{30})。

在运用支持向量机方法时，核函数选取为高斯径向基核函数，参数自动寻优结果为：

best gamma =0.5，cost=4，R^2= 0.8485

将训练集数据和测试集数据分别代入模型进行验证，为便于比较，将最终的分析结果亦列入表 2 中。从模型解释性与预测精度中，可以看出 Lasso-SVM 方法的所有的精度都在 80% 以上，综合精度在 94% 以上，因而 Lasso-SVM 方法的拟合效果要高于直接运用 SVM 方法的拟合效果，能够提高预测精度，拥有更好的预测性能。

四、结语

本文通过对上市公司财务比率数据进行分析，建立信用风险评定模型来预测上市公司的信用风险，分别建立了 SVM 和 Lasso-SVM 模型，通过不同模型选择对上市公司信用风险影响较强的指标，同时根据模型的解释效果和预测效果，选择出更适合评定上市公司信用风险的模型。根据全文研究，可以看出，Lasso-SVM 模型的预测精度都要高于普通的 SVM 模型，这可以说明，在上市公司信用评价问题上，使用 Lasso 方法进行变量选择之后再运用支持向量机方法进行预测有一定的优势，能够提高预测精度，拥有更好的预测性能。

参考文献：

[1]Beaver, W. H. Financial Ratios As Predictors of Failure[J]. Journal of Accounting Research, 1966(4): 71-111.

[2]Altman, E. I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy[J]. The Journal of Finance, 1968(23): 589-609.

[3]Altman, E .I, Haldeman, .R .G, and Narayanan, P. ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations [J].Journal of Banking and France, 1977(1):29-54.

[4] 胡安冉 , 孙云 . 基于 logistic 模型的上市公司信用风险评价的实证研究 [J]. 商 ,2012(24):91.

[5] 梁琪 . 企业经营管理预警 : 主成分分析在 logistic 回归方法中的应用 [J]. 管理工程学报 ,2005(19):100-103.

[6] 孙森 , 王玲 . 基于 KMV-Logit 模型的上市公司违约风险实证研究 [J]. 财会月刊 ,2014(18):64-68.

[7] 彭颖 . 基于多元自适应回归样条的企业信用评估模型研究 [D]. 湖南大学 ,2012.

[8] 石秀福 . 基于支持向量机的上市公司财务信用评价预测方法研究 [D]. 华东师范大学 ,2008.

[9]Tibshirani, R. Regression Shrinkage and Selection via the lasso[J]. Journal of the Royal Statistical Society. Series B(methodological), 1996(58): 267 - 88.

作者简介：滕树军 (1973-)，男，内蒙古赤峰人，天津商业大学理学院讲师，主要从事统计学与运筹学研究。

刘丽平 (1996-)，女，辽宁铁岭人，东北财经大学应用统计在读研究生。

刘柏森 (1973-)，男，吉林抚松人，东北财经大学统计学院讲师，主要从事数学与统计学研究。

基金项目：辽宁省社会科学规划基金重点项目 (L16ATJ001)；天津商业大学教师产学研创新实践活动项目 (2018RC020637)；天津商业大学青年科研培育基金项目 (G12Y100111)。