

基于神经网络模型的网络借贷高危企业信用风险的识别研究

王茂光, 朱子君

(中央财经大学信息学院, 北京 100081)

摘要: 网络借贷的飞速发展在一定程度上缓解了小微企业融资难的问题, 但也暴露出网络借贷平台信用风险的识别问题。为充分识别高危网贷企业的特征, 以中小型网贷企业为样本, 通过指标筛选, 挑选出与风险识别相关度较高的指标作为指标变量。并利用 BP 神经网络算法模型得出高危网贷企业在不同条件下的信用风险识别率和信用风险分类正确率。实验结果表明, 高危网贷企业的信用风险具有高度识别性, 高召回率、高正确率的特点。

关键词: 高危网贷企业风险识别; 指标筛选; 神经网络; 召回率

中图分类号: F49

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2017.00222

Credit risk identification of high-risk online lending enterprises based on neural network model

WANG Mao-guang, ZHU Zi-jun

(School of Information, Central University of Finance and Economics, Beijing 100081, China)

Abstract: The rapid development of online lending alleviates the difficulty of financing for small and micro enterprises to a certain extent, but it also exposes the credit risk identification problem of online lending platform. In order to fully identify the characteristics of high-risk network lending enterprises, small and medium-sized network lending companies were selected as samples, and indicators that were highly correlated with risk identification were chosen as indicators variables. And by using the BP neural network algorithm model, the credit risk identification rate and credit risk classification accuracy rate of high risk network lending enterprises, under different conditions, were obtained. The results show that the credit risks of high-risk network lending enterprises are highly recognized, and have the characteristics of high recall rate and high accuracy.

Key words: high risk online lending enterprise risk identification, index screening, neural network, recall rate

1 引言

中小微企业融资难的问题一直是世界各国普遍关注的难点问题。随着网络借贷的飞速发展, 虽然融资难的问题得到了极大的缓解, 许多小型公司可以通过网络融资获得资助, 但是同时也

暴露出网络融资高风险的弊端。例如, e 租宝涉嫌非法集资 500 多亿, “校园贷”事件中, 学生在多平台借款, 最后却因无力还款而造成巨大经济压力等。这些网络融资风险产生的主要原因在于目前大多数网络融资的公司规模都较小, 没有较好的管理、经营机制以及完善的信息披露机制。

收稿日期: 2017-11-03; **修回日期:** 2017-12-01. **通信作者:** 朱子君, zzj_2017@163.com

基金项目: 网金中心合作基金资助项目 (No.020676116004); 北京大学合作基金资助项目 (No.020676114004)

Foundation Items: Cooperation Project with Network Finance Center (No.020676116004), Cooperation Project with Peking University (No.020676114004)

公司资质参差不齐,加上相关机构监控力度弱,风险控制模型尚不成熟,无法满足目前网络借贷的需求,从而导致中小微型企业融资难、融资贵的问题依然没有得到彻底解决。因此,如何在网络借贷中运用恰当的风险识别系统是重中之重,即建立适当的风险识别体系,进行风险识别。

互联网技术的迅速发展为网络借贷的信用风险识别提供了有力的技术手段。尤其是大数据的快速发展,打破了传统风险模型模式,使利用大数据进行企业风险划分成为可能。

2 相关工作

在国内研究中,企业的信用风险评估仍处于起步阶段,许多研究已经将关注点定位在企业的风险识别和模型搭建,但是大多数的评价模型仍依赖于定性分析,缺乏量化风控模型。尤其在互联网金融跑路现象盛行的情况下,有效的模型是人们关注的焦点。目前人们研究较多的模型算法有基于贝叶斯、决策树、神经网络等算法。各个学科知识也逐渐被引用到模型搭建中,如统计学、数学、计算机科学技术等。

在风险评估方面,Beaver等^[1]首次将统计方法运用到企业危机预警研究中,指出现金流与负债总额之比、资产负债率是最能体现企业财务状况的指标,并对企业的财务情况进行预测。刘新海^[2]对大数据应用处于国内领先的阿里巴巴企业进行征信体系和征信预测分析。肖曼君等^[3]针对P2P网络借贷采用排序选择模型对平台信用风险进行实证分析。刘宇璟等^[4]针对小微企业创业者素质提出了从创业人格特质、创业能力和创业知识3个维度评价小微企业创业者指标体系。郭建平^[5]构建了基于灰色关联分析的电子商务信用风险评价模型,并对各级最终模型的评价结果提出相应的应对策略。这些方法大多属于传统风险评估方法,多集中于定性分析,缺乏定量分析的有效模型。

神经网络的快速发展,促使许多学者将神经网络方法应用到解决问题的过程中,Atiya^[6]提出了将神经网络应用于对银行破产的信用风险评价,提出运用神经网络的模型将传统金融指标用于对银行风险进行预测。叶斌等^[7]总结了设计BP

网络时确定隐含层层数以及节点数的几种传统方法,并指出了BP网络结构优化的新思路 and 未来发展趋势。熊志斌^[8]运用遗传神经网络模型对我国上市公司财务危机进行预测,并与BP神经网络进行对比。翟万里^[9]针对商业银行企业客户信用风险的主要影响因素,基于改进的BP神经网络建立了商业银行信用风险模型。刘建伟等^[10]论述了深度学习的新进展,对深度学习在网络层数、激活函数选择、模型结构等方面进行了综述。

目前,虽然国内外对于网络借贷的风险关注度都很高,但是风险识别评价还多停留在采用专家评价法或者定性评价法等主观方法,缺少有效的定量模型和算法。在目前网络借贷快速发展的时期,必须对大量的网络数据进行模型的实证分析。

3 模型算法

3.1 数据处理

针对网贷平台风险较高的问题,本文通过收集网贷平台的信用风险数据以及学习出现过“跑路”“经营困难”等问题网贷企业的数据,来训练网贷企业的风险评估模型。本文的数据来自第三方资讯平台“网贷之家”的网贷平台企业数据,数据中主要包含了网贷企业的各项经营和财务指标,其中涵盖了网贷企业的资本情况数据、营业财务数据和市场的打分等各项指标。

数据共包含37个变量,各个具体属性如表1所示。

之后本文对收集到的37个指标进行指标筛选和离散化处理。筛选时采用WOE值和IV值对指标变量进行选取。

定义1 假设 y_i 为在相应属性中高风险企业数量, n_i 为在相应属性中低风险企业数量, py_i 为 y_i 占样本中所有高风险企业的比例, pn_i 是 n_i 占样本中所有低风险企业的比例,则有

$$WOE_i = \ln \left(\frac{py_i}{pn_i} \right) = \ln \left(\frac{\frac{y_i}{\sum y_i}}{\frac{n_i}{\sum n_i}} \right) \quad (1)$$

$$\text{定义2} \quad IV_i = (py_i - pn_i)WOE_i \quad (2)$$

$$IV = \sum IV_i \quad (3)$$

表 1 数据变量属性	
属性	含义
X1	网贷企业名称
X2	网贷企业的法人代表
X3	网贷公司性质、公司类型
X4	网贷企业的开业时间
X5	网贷企业核准开业时间
X6	资讯平台对网贷企业的评价打分
X7	网贷企业的平均收益
X8	网贷企业的注册资本
X9	网贷企业实际缴付的资本
X10	网贷平台是否允许投资自动投标业务
X11	网贷平台是否允许投资者进行债券转让
X12	网贷平台资金是否有托管机构
X13	网贷平台投资的标是否有担保
X14	担保机构名称
X15	网贷平台管理费用的情况
X16	网贷平台充值费用的情况
X17	网贷平台提现费用的情况
X18	网贷平台转让费用的情况
X19	投资期限在 1 个月的标的所占比重
X20	投资期限在 3 个月的标的所占比重
X21	投资期限在 4~6 个月的标的所占比重
X22	投资期限在 6 个月以上的标的所占比重
X23	标的金额在 $[0, 10^5)$ 元所占比重
X24	标的金额在 $[10^5 \sim 10^6)$ 元所占比重
X25	标的金额在 $[10^6, 10^7)$ 元所占比重
X26	标的金额在 1 000 万元及以上所占比重
X27	贵宾客户费用情况
X28	费用支付方式情况
X29	网贷企业的近期交易成交量
X30	网贷平台累计待还金额
X31	平台最近 30 天待还金额
X32	平台净流入金额
X33	平台的投资人数
X34	平台的借款人数
X35	对网贷平台进行评价的人数
X36	用户对于网贷企业的推荐程度 (%)
X37	是否为高危网贷企业

IV 值在 WOE 值的基础上计算得出，更加直观地表示出该属性值对于因变量分类之间的关系程度。

以“平台投资期限”为例计算 WOE 和 IV 值，如表 2 所示。平台收益划分为 1、2、3 三档，平台状态由 0、1 表示。平台收益为 1 且平台状态为 0 的共有 1 051 家平台，以此类推。根据式(1)可计算得出平均收益在“1”下的 WOE_1 值为 0.71，根据式(2)可求得 IV_1 为 0.33，求和后得到的 IV 为 1.4。

表 2 变量“平台投资期限”的 WOE 和 IV 值				
平台收益	平台状态		WOE	IV
	0	1		
1	1 051	715	0.71	0.33
2	80	489	-1.49	0.36
3	12	378	-3.12	0.71
总计	1 143	1 582	—	1.40

通常将 IV 值在 $[0.1, \infty)$ 范围内的自变量作为提取后的指标。同时在数据处理时，本文将同类指标合并（如投资期限占比、标的金额占比）计算其加权平均数；将不便于划分和离散的属性（如公司法人）去除；将缺失值较多的属性变量（如平台的投资人数、平台的借款人人数）去除。经过处理后，再对得到的连续性变量进行离散化处理。最后共整理出 21 个因变量，1 个自变量（企业风险类型），有效数据 2 725 条。具体分类如表 3 所示。

3.2 模型算法

本文的算法运行环境是 Windows 操作系统，使用 Java 编程语言编写，并通过 BeanShell——一种兼容 Java 的脚本语言，进行运行调试。BeanShell 是一个小巧免费的 Java 源码解释器，支持对象式的脚本语言特性，亦可嵌入 Java 源代码中。之所以选择 BeanShell 的原因是 BeanShell 可以快速地检验算法调整的结果，还可以检验单个输入的反馈，所以本文使用 BeanShell 进行实验。

本文算法使用的模型是 BP 神经网络算法，一个典型的神经网络，主要由神经元节点与节点之间的联系（突触）构成：一些节点负责简

表 3		最终变量选取和划分标准	
变量类别	变量名称	数据类型	分类依据
企业基本指标	公司类型	离散变量	私营类=-4, 股份制=-3, 合伙制=-2, 国有制=-1, 集体所有制=1, 联营企业=2, 外商投资类=3, 港澳台投资类=4, 缺失=0
	开业时间	连续变量	(0, 90]天=5, (90, 180]天=4, (180, 360]天=3, (360, 1200]天=2, (1200, ∞)=1, 缺失=0
	核准开业时间	连续变量	(- ∞ , 0]天以下=6, (0, 100]天=5, (100, 180]天=4, (180, 360]天=3, (360, 600]天=2, (600, ∞)=1, 缺失=0
	是否有第三方征信	二元变量	有=1, 无=0
	是否接受过风投	二元变量	有=1, 无=0
	自动投标	二元变量	有=1, 无=0
财务指标	平均收益	连续变量	[0.05, 0.1]=1, [0.1, 0.15]=2, [0.15, 0.2]=3, (0.2, ∞)=4
	平均投资期限	连续变量	(0, 30]天=1, (30, 60]天=2, (60, ∞)天=3
	平均标的金额	连续变量	(0, 5×10^4]元=1, (5×10^4 , 5×10^5]元=2, (5×10^5 , 10^6]元=3, (10^6 , 5×10^6]元=4, (5×10^6 , 10^7]元=5
资本情况	注册资金	连续变量	(0, 5×10^6]元=1, (5×10^6 , 2×10^7]元=2, (2×10^7 , 5×10^7]元=3, (5×10^7 , 10^8]元=4, (10^8 , ∞]元=5, 缺失=0
	实缴资本	连续变量	(0, 10^6]元=1, (10^6 , 5×10^6]元=2, (5×10^6 , 10^7]元=3, (10^7 , 5×10^7]元=4, (5×10^7 , ∞]元=5, 缺失=0
	实缴资本比例	连续变量	(0, 0.01]=1, (0.01, 0.1]2, (0.1, 0.5]=3, (0.5, 1]=4, 缺失=0
平台资本保障	平台管理费	二元变量	有=1, 无=0
	平台充值费	二元变量	有=1, 无=0
	平台提现费	二元变量	有=1, 无=0
	平台 VIP 费	二元变量	有=1, 无=0
	平台转让费	二元变量	有=1, 无=0
	债券转让	二元变量	有=1, 无=0
	保障模式	离散变量	非融资性担保公司=-4, 融资性担保公司=-3, 平台垫付=-2, 小贷公司=-1, 风险准备金=1, 抵押质押=2, 中介平台=3, 其他=4, 缺失=0
评价指标	第三方评价	二元变量	有=1, 无=0
	用户推荐度	连续变量	0%=0, (0, 0.1]=1, (0.1, 0.25]=2, (0.25, 0.5]=3, (0.5, 0.99]=4, 100%=5
结果变量	平台风险状态	二元变量	有=1, 无=0

单地把数据传入或传出网络，另一些节点则负责处理数据。将数据导入网络或从网络导出数据的层级称为输入层和输出层节点，其他节点被称为隐藏层节点。

根据数据可以得出，BP 神经网络的输入层节点有 21 个，输出层节点有 1 个，中间隐藏层节点数可变，共 3 层。BP 神经网络即后向传播算法，通过检查每一个训练实例，并依结果调整自身链接的权值，以使输出值与期望值之间

的误差最小化。本文通过调用 `updateWeight()` 实现权值的更新，使用 `sigmoid` 函数作为激励函数，学习率和迭代次数均可变。

神经网络的基本思路如下。首先将总数据分为训练数据和测试数据两部分，对训练数据进行学习，再用运行测试数据得出结果并验证训练的准确率。一般来讲，训练数据越大，准确性会相应提高。本文使用优化的神经网络算法，提高了神经网络在低比例训练数据下的学

习效率。学习效率（分类正确程度）通过计算分类的正确率和召回率衡量。

定义 3 计算训练结果正确率的比值为正确率和召回率。

各变量假设如表 4 所示，其中， TP 为低风险企业被模型预测为低风险企业的数量， FN 为低风险模型预测为高风险企业的数量，以此类推。则可得

$$\text{正确率} = \frac{TP + TN}{TP + TN + FP + FN}$$

(4)

$$\text{召回率} = \frac{TN}{TN + FP} = \frac{TN}{N}$$

(5)

通过比较分类的正确率和召回率来分析模型分类的合理度和分类的正确程度。

3.3 方法创新点

本文在研究方法上主要采用实证研究法和定量分析法。在实证分析法中首先对目前网络借贷企业的信用数据进行指标量化和离散化，进而利用神经网络评估高风险企业，以确保高风险企业不会成为漏网之鱼，实验结果证明，高风险企业的数据具有一定特征，在神经网络训练过程中，学习成功率很高。

相较于其他常见的模型算法，如决策树模型、贝叶斯模型、SVM 模型等，神经网络具有实现复杂非线性映射的优点，在大量数据训练

下具有高度自学习和自适应能力，同时具有一定的容错能力，在部分或局部神经元受到破坏后不会影响系统的正常工作。因此，本文选择神经网络进行模型训练。

4 实验结果

本文采用实证方法来分析不同学习率（0.1, 0.5, 0.025）和神经网络迭代次数（10, 25, 50, 100, 200, 500）对学习率的影响，并通过调整训练数据和测试数据的比例寻找最佳的测试比例。

1) 在控制变量条件下，测试不同的迭代次数对正确率和召回率的影响，结果如表 5 和图 1 所示。

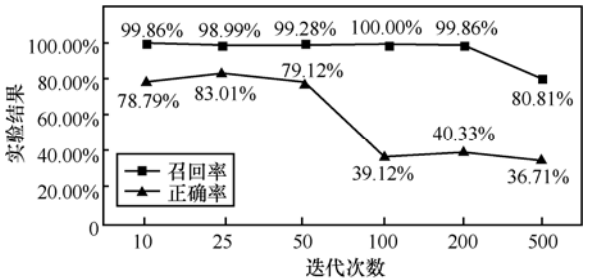


图 1 不同迭代次数下高风险企业识别的表现

可以看出，在迭代次数的增大最初会导致正确率短暂上升，然后下降，其中在迭代次数为 25 达到最佳，整体数据的正确率高达 83%。而召回

表 4 不同情况下的变量假设

实际类别	预测类别		总计
	GOOD(低风险企业)	BAD(高风险企业)	
GOOD(低风险企业)	TP	FN	$P(\text{实际为 GOOD})$
BAD(高风险企业)	FP	TN	$N(\text{实际为 BAD})$
总计	$p(\text{被分为 GOOD})$	$n(\text{被分为 BAD})$	$P+N$

表 5 模型在不同迭代次数下对高风险企业识别的表现

学习率	迭代次数	训练数据	测试数据	高风险企业数	高风险企业分类正确数据	高风险企业识别召回率	神经网络正确率
0.025	10	900	1 825	693	692	99.86%	78.79%
0.025	25	900	1 825	693	686	98.99%	83.01%
0.025	50	900	1 825	693	688	99.28%	79.12%
0.025	100	900	1 825	693	693	100.00%	39.12%
0.025	200	900	1 825	693	692	99.86%	40.33%
0.025	500	900	1 825	693	560	80.81%	36.71%

率在不同的迭代次数下, 比值都较高, 表示迭代次数对高风险识别的影响并不大。

2) 在控制变量情况下, 本文测试不同的学习率对正确率和召回率的影响, 结果如图2所示。

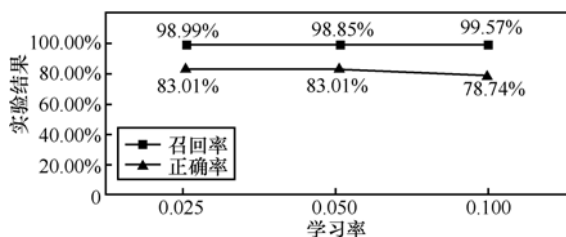


图2 不同学习率下高风险企业识别的表现

通过图2可以发现, 学习率在[0.025,0.1]的范围内, 企业风险识别的正确率和召回率都较高。学习率对风险的敏感程度不大。

3) 在控制变量的情况下, 本文测试训练数据和测试数据比率对正确率和召回率的影响, 结果如图3所示（图例括号中数字表示迭代次数）。

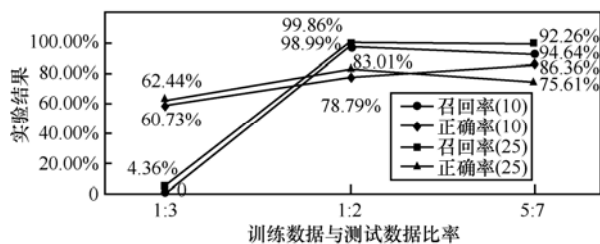


图3 训练数据与测试数据比率对召回率和正确率的影响

通过图3可以发现, 训练数据与测试数据比率对召回率的影响比对正确率的影响大。在比率为1:3的情况下, 召回率几乎为零, 提升到1:2后, 召回率恢复到正常水平。而正确率在不同比率下一直保持在比较平稳的水平。

4) 在控制其他变量情况下, 设置不同的隐藏层节点数（如5, 7, 10）来测试对高风险企业识别的正确率和召回率的影响。其中无关变量的设置为迭代次数等于10, 学习率为0.025。

根据图4可得出, 在隐藏层节点数为5和7的情况下, 高风险企业识别的召回率和正确率均较高; 将节点数提高到10时, 在召回率依然较好的情况下, 正确率明显降低。可见隐藏层节点数也不是越多越好。

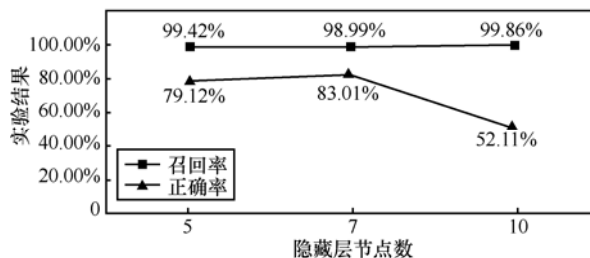


图4 隐藏层节点数对召回率和正确率的影响

5) 为了证明算法的优越性, 本文又对比了在不同的算法下（朴素贝叶斯、决策树、神经网络算法）召回率和正确率之间的差异（训练数据与测试数据比为1:2），结果如图5所示。

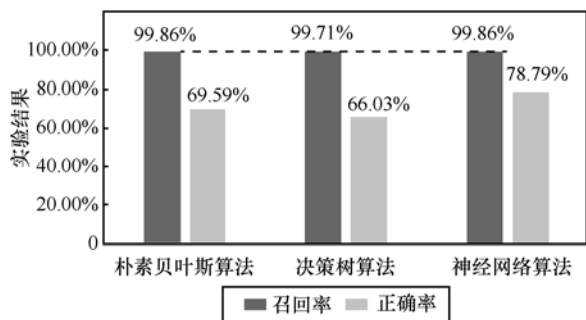


图5 不同算法之间召回率和正确率对比

根据图5可以发现, 不同的算法间召回率都比较高, 可以保证高风险识别率较高。但是贝叶斯和决策树的正确率明显低于神经网络的正确率, 证明了神经网络在少量训练数据下良好的学习能力。

5 结束语

本文通过采用BP神经网络模型, 对网络借贷高危企业进行风险识别, 测试神经网络对企业风险识别的正确率, 以及高危风险企业信用正确分类的召回率。实验结果表明, 高危企业的数据有明显的特征, 高危企业被误识别为低风险企业的可能性很低, 即召回率较高。并且企业风险划分的正确率较高, 误报率也较低, 保证了模型的正确性。进而分析了在选取不同的迭代次数、学习率和训练数据比率的情况下, 对模型正确率和召回率的影响。

下一步工作是进一步识别高风险网贷企业的风险关键性指标, 找出导致网贷企业风险较高的原因, 以及探究并测试不同的神经网络模型对风险识别的影响。

参考文献：

- [1] BARTH M E, BEAVER W H, LANDSMAN W R. Value-relevance of banks' fair value disclosures under SFAS No. 107[J]. Accounting Review, 1996, 71(4): 513-537.
- [2] 刘新海. 阿里巴巴集团的大数据战略与征信实践[J]. 征信, 2014, 32(10): 10-14.
LIU X H. The strategy of big data for Alibaba Group and its practices of credit reference[J]. Credit Reference, 2014, 32(10): 10-14.
- [3] 肖曼君, 欧缘媛, 李颖. 我国 P2P 网络借贷信用风险影响因素研究——基于排序选择模型的实证分析[J]. 财经理论与实践, 2015, 36(1): 2-6.
XIAO M J, OU Y Y, LI Y. Research on the factors affecting credit risk of P2P lending in China: an empirical analysis based on the sort selection model[J]. The Theory and Practice of Finance and Economics, 2015, 36(1): 2-6.
- [4] 刘宇璟, 林裘绪, 窦先琴. 小微企业创业者素质评价指标体系构建[J]. 统计与决策, 2015, (20): 185-188.
LIU Y J, LIN Q X, DOU X Q. The construction of the quality evaluation index system for the entrepreneurs of small and micro enterprises[J]. Statistics and Decision, 2015, (20): 185-188.
- [5] 郭建平. 电子商务信用风险评估模型[J]. 统计与决策, 2016, (11): 69-71.
WU J P. Credit risk assessment model of e-commerce[J]. Statistics and Decision, 2016, (11): 69-71.
- [6] ATIYA A F. Bankruptcy prediction for credit risk using neural networks: a survey and new results[J]. IEEE Transactions on Neural Networks, 2001, 12(4): 929.
- [7] 叶斌, 雷燕. 关于 BP 网中隐含层层数及其节点数选取方法浅析[J]. 商丘职业技术学院学报, 2004(6): 52-53.
YE B, LEI Y. On the method of selecting the number of hidden layers and the number of nodes in BP network[J]. Journal of Shangqiu Vocational Technical College, 2004, (6): 52-53.
- [8] 熊志斌. 运用遗传神经网络模型对我国上市公司财务危机的预测[J]. 统计与决策, 2008, (14): 32-34.
XIONG Z B. Using genetic neural network model to predict financial crisis of Listed Companies in China[J]. Statistics and Decision, 2008, (14): 32-34.
- [9] 翟万里. 基于人工神经网络的商业银行信用风险评估模型研究[D]. 长沙: 长沙理工大学, 2013.
ZHAI W L. Research on credit risk assessment model of commercial banks based on artificial neural network[D]. Changsha: Changsha University of Science and Technology, 2013.
- [10] 刘建伟, 刘媛, 罗雄麟. 深度学习研究进展[J]. 计算机应用研究, 2014, 31(7): 1921-1930.
LIU J W, LIU Y, LUO X L. Progress in research on deep learning[J]. Application Research of Computers, 2014, 31(7): 1921-1930.

作者简介：



王茂光(1974-), 男, 山东招远人, 中央财经大学教授, 主要研究方向为互联网金融风控和征信、软件工程、分布式智能系统。



朱子君(1993-), 女, 吉林长春人, 中央财经大学硕士生, 主要研究方向为互联网金融风控和征信。