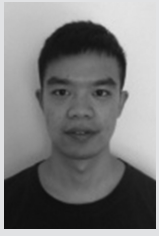
大数据发现银行贷款风险

Uncovering the Risk in Bank Loans by Big Data

曾伟，男，电子科技大学副教授，研究内容包括机器学习、推荐系统、链路预测、国家进出口贸易等。发表学术论文14篇，以第一作者发表学术论文10篇，其中6篇SCI检索论文和4篇国际会议论文。主持两项国际项目（Sino Swiss Science and Technology Cooperation No.EG57-092011和No.TE-70382）和一项国内项目（优秀博士生学术支持计划No.YBXSZC2013029），参与多个国家科学自然基金项目。

孔新川，男，浙江大学经济学硕士，杭州迈宁数据科技有限公司创始人、CEO，拥有多年的企业管理及TMT领域投资经验，对传统金融机构贷后管理有深刻的理解和认识。



陈威，男，电子科技大学互联网科学中心硕士生，主要研究方向为推荐系统、数据挖掘及流行性预测等。

周涛，男，电子科技大学大数据研究中心主任、教授、博士生导师，主要关注统计物理与信息科学、社会学、经济学等领域的交叉科学问题，发表SCI论文200余篇，引用10 000余次，H指数超过50。

最近几年，大数据的商业化应用开始逐步落地[1]，其中，金融方面的大数据应用是被投资人最看好的大数据产业化方向，在个人征信、企业征信、客户画像与精准营销等方面都有成熟的应用。例如花旗银行通过挖掘信用卡数据，实现交叉营销。当客户每次刷卡时，银行根据时间、地点以及过往的购物记录，筛选并推送给客户周边商店、餐厅的折扣优惠，从而获得第二次交易价值。富国银行运用大数据识别欺诈行为，通过研究客户之间发生的历史交易，检测是否存在背离常规操作模式的资金异动，通过综合观察多个数据来源，总结出用户典型的交易习惯，实现实时的可疑交易甄别。在国内，许多商业银行也在大数据领域不断地探索和尝试，例如中国银行的“中银沃金融”利用大数据技术，整合电商平台共享数据、征信数据以及客户经理面谈获取的信息，利用授信审批模型实现自动审批。本文介绍大数据在金融风险管理方面的实际案例。对于以银行为代表的金融机构而言，风险管理贯穿它们的全业务过程，越早发现风险越早采取措施，风险管理的成本越低，给金融机构带来的损失越小。

贷款的风险管理对于传统银行业和新兴的互联网金融行业都起着至关重要的作用。不良资产问题长期困扰着国有银行，成为国有银行面临的主要金融风险，直接威胁国有银行的生存和发展。根据银监会对外公开报告，为改善资产质量，我国政府于1999年和2000年为四大国有商业银行分别成立资产管理公司，剥离不良资产1.3万亿元，使其不良贷款比率平均下降10个百分点。但是，进行资产剥离只能缓解已有不良贷款带来的冲击，剥离后的不良贷款比率仍然远高于中国人民银行的监管水平。对于互联网金融企业，尤其是通过P2P或者分期付款等方式，以高息贷款为实质性业务的企业，风险的控制是成败的关键。无论线上有多大流量，每月有多少流水，风险投资有多大规模，如果贷款违约率控制不了，最终都必然走向失败[2]。因此，建立和完善风险管理体系，提高自身的风险管理水平和管理效率，是商业银行和互联网金融企业持续发展的重要基础。

一方面随着贷款客户数量的增多（来源于个人信用贷款和中小微企业贷款数量的增长），传统的人工管理手段（如业务经理管理自己的客户）已经无法满足目前风险管理在成本和效率上的需求；另一方面，银行系统（数据库）中包含了大量的客户交易转账、存款取款、信用卡消费等多个维度的数据，同时随着互联网的普及，客户在互联网（如微信、QQ等）上会产生大量的外部数据，这为大数据在贷款风险管理方面发挥作用提供了基础。目前，越来越多的银行和互联网金融机构开始探索如何利用大数据的方法进行风险预警的工作，并希望建立一个高度自动化、智能化与银行其他系统密切配合的风险预警系统。

电子科技大学和杭州迈宁数据科技有限公司的联合研究小组，基于银行系统中客户的贷款协议信息、交易流水信息等内部信息以及工商局、法院等外部信息来设计风险预警模型。这里主要针对已放贷款进行贷后风险的管理和预警。对于每笔已放贷款，银行会要求客户在每月或者每个季度（视贷款规定的还款间隔而定）规定的还款日期之前还款，若客户在还款日期前没有还款，则该客户为逾期客户（计算入违约率），否则为正常客户。研究小组希望能够利用客户的当前数据，预测其下个月或者下个季度是否为逾期客户。

客户的贷款协议信息包含了每个客户的基本信息，其中包括贷款笔数、贷款金额、还款卡号余额、本月应还金额等；另外，贷款协议信息还包含客户所在的行业

类别、注册公司的规模等信息。客户的交易流水信息包含每个客户的交易对手、交易金额和交易时间等基本信息。另外，笔者团队也计算了每个客户每月交易金额的平均值、方差和交易时间间隔、收入和支出比例等，并将这些信息作为客户的特征。

进一步地，通过网络爬虫爬取客户的工商数据、法院诉讼和房产抵押等外部数据。工商数据包含了客户实体企业的注册资金、企业规模、法人代表等信息。法院诉讼数据包含了最近客户是否存在诉讼记录，房产抵押数据包含了客户及配偶的房产信息。将这些外部数据也作为客户的特征。

基于以上数据，利用机器学习的方法对客户进行初筛选。采用了线性回归、Logistic回归、SVM、神经网络、决策树等分类器，将每一个单模型都看作一个弱分类器，然后再进行融合[3,4]。通过集成学习，获得更好的分类效果。进一步地，利用复杂网络方法和时间序列分析技术筛选剩下的客户。不断地迭代以上两个步骤，直到算法达到最优，其整体思路如图1所示。

以复杂网络方法为例[5]，如果有N个违约客户，完全随机抽样N个节点所形成的网络几乎全都是孤立节点或者非常小的连通片，客户之间基本没有资金往来关系。但是所有违约客户形成的网络却要比同规模的随机抽样网络连边密集得多。这说明违约是有网络效应的：一方面金融风险本身具有传递性，客户A如果资金出现问题，无法按时还款，则客户A对应的应付客户有可能因为没有收到A的钱，导致资金链出现问题，从这个意义上讲，如果上一个月A向B流入了资金，且上个月A出现了违约或者这个月预测A违约风险很高，都会提高B的预测风险；另一方面，违约还具有社会效应，譬如A违约之后，因为违约额度不高，银行没有及时处理，A就有可能将此消息传播给自己的商业伙伴，从而使得其他人也出现违约的行为。从这个意义上讲，只要A和B有资金关系，不管是流入或者流出，鉴于A的违约行为或者高违约风险，也会提高对B的风险预测。把“因为网络效应而产生的违约风险”做成若干个特征，也放入了客户特征库中进行迭代学习。

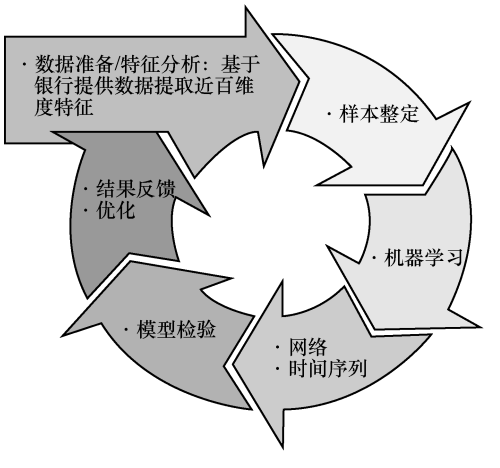


图1 贷后风险预警模型

主要通过两个指标来刻画预测的效果。一是用召回率（recall，可参考参考文献[6]）来度量准确性，即预测出来的高风险客户能够包含银行真实违约客户的比例，这个比例越高越好，最高是100%。二是用查找范围，即预测的高风险客户占整个客户总量的比例，在相同准确性的情况下，查找范围越小越好。如图2所示，与合作银行原有的方法相比（基于Logistic回归和其他单一模型的机器学习方法，未进行特征挖掘和特征学习），研究小组采用的方法使准确性从46.7%上升到88.0%。而银行原来的方法把大约20%的客户判断为高风险客户，研究小组采用的方法则只需要筛查11.2%的客户，相比银行传统的方法有了跨越性的提高。

在中央大力建设信用社会的过程中，中国仍然有很大一段时间是一个信用成本

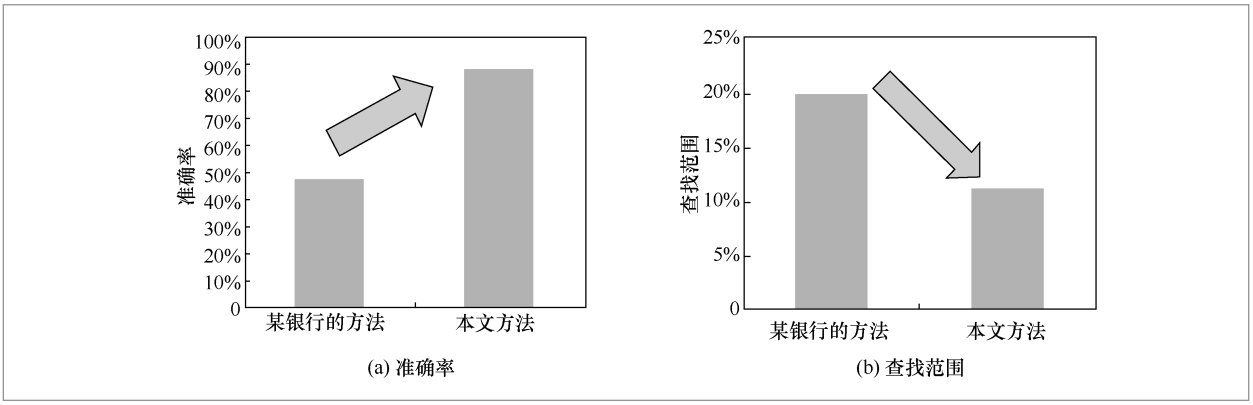


图2 风险模型预测能力对比结果

很低的国家，大家不以违约为耻，反以违约不被追究为荣！在这种情况下，以信用为“担保”的针对个人和中小微企业的贷款风险格外大——而这恰好是很多互联网金融企业的主营业务。大数据的方法通过整合内外数据，并引入深度的特征挖掘和大规模的集成学习，有望在信用机制尚未健全的时候，为金融机构的风险管理提供一架高效的“预警机”。

参考文献

[1] Schoenberg V M, Cukier K. 大数据时代：生活、工作、思维的大变革. 盛杨燕, 周涛译. 杭州: 浙江人民出版社, 2013
Schoenberg V M, Cukier K. Big Data: A Revolution that Will Transform How We Live, Work, and Think. Translated by Sheng Y Y, Zhou T. Hangzhou: Zhejiang People's Publishing House, 2013

[2] 李平, 陈林, 李强等.互联网金融的发展与研究综述. 电子科技大学学报, 2015, 44(2): 245~253
Li P, Chen L, Li Q, *et al*. Review of research and industry development of internet finance. Journal of University of Electronic Science and Technology of China, 2015, 44(2): 245~253

[3] Friedman J. Greedy function approximation: agradient boosting machine.The Annals of Statistics, 2001, 29(5): 1189~1232

[4] Ridgeway G. Generalized Boosted Models: A Guide to The GBMPackage, http://cran. r-project.org/web/packages/gbm/vignettes/gbm.pdf,2007

[5] 汪小帆, 李翔, 陈关荣. 网络科学导论. 北京: 高等教育出版社, 2012
Wang X F, Li X, Chen G R. Network Science: An Introduction. Beijing: Higher Education Press, 2012

[6] Lü L, Zhou T. Link prediction in complex networks: a survey. Physica A Statistical Mechanics & Its Applications, 2011, 390(6): 1150~1170 □