基于时空聚集的网贷反欺诈建模与研究*

俞旭峰[1],王　澎[1],郭　威[2],张子柯[1]

(1.杭州师范大学 阿里巴巴复杂科学研究中心,浙江 杭州 311121;
2.阿里巴巴集团 新零售技术事业群,浙江 杭州 310008)

摘　要:识别突发的团伙欺诈已经成为网贷业务中亟待解决的问题。在特征维度较少的情况下,提出了一种基于时空聚集的网贷反欺诈模型。首先基于用户定位信息和申请贷款的时间,设计了一个适用于网贷场景下的聚集指标:K-N 最近邻指数;然后,将不同时间观察窗口的 K-N 最近邻指数利用基于 LSTM(长短期记忆网络)的 seq2seq(序列到序列)模型提取 embedding(嵌入)特征;最后,利用 LightGBM 模型预测欺诈发生的概率。实验结果表明,所提出的指标能更有效地捕捉坏账,且相比于仅使用基础特征,预测结果的 KS 值和 AUC 都有了较好的提升。

关键词:数据挖掘;金融欺诈识别;时空数据分析;近邻指数;LSTM

中图分类号:TP391　　**文献标识码**:A　　DOI:10.19358/j.issn.2096-5133.2020.02.013

引用格式:俞旭峰,王澎,郭威,等.基于时空聚集的网贷反欺诈建模与研究[J].信息技术与网络安全,2020,39(2):69-74.

Anti-fraud modeling and research of online loans based on time and space aggregation

Yu Xufeng[1], Wang Peng[1], Guo Wei[2], Zhang Zike[1]

(1. Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, China;
2. New Retail Technology Business Group, Alibaba Group, Hangzhou 310008, China)

Abstract: The identification of sudden gang fraud has become an urgent problem in the online loan business. In the case of less feature dimensions, this paper proposes an anti-fraud model of online loans based on spatiotemporal aggregation. Firstly, based on the users' location information and the time of applying for the loan, a clustering indicator suitable for the online loan business, K-N nearest neighbor index is designed; Then, the K-N nearest neighbor index of different time observation windows is used to extract embedding features from seq2seq (sequence to sequence) model based on LSTM (Long Short-Term Memory); Finally, the LightGBM model is used to predict the probability of fraud. The experimental results show that the proposed indicator can capture bad debts more effectively. Compared with only using the basic features, the KS value and AUC of the prediction result are better improved.

Key words: data mining; financial fraud identification; spatio-temporal data analysis; neighbor index; LSTM

0　引言

网贷具有以下 3 个重要的优势:高回报、覆盖面广、需求量大[1],所以最近几年得到持续蓬勃发展。然而,网贷在给借贷者带来便利、及时的金融服务的同时,也给放贷方带来了欺诈者的攻击威胁的风险[2-3]。首先,网贷主要是面向那些没有抵押、在传统信贷体系之外的借贷者;其次,网贷业务中个人数据较敏感,放贷方难以充分获取用户真实数据,所以那些缺少较为全面的反欺诈风控机制的放贷方面临着重大损失的风险[2-4]。

目前,国内外已有不少文献从不同角度来展开网贷反欺诈研究。如文献[5]总结了信用卡风险控制领域常用的统计方法,包括信用卡统计、信用卡债务、信用评分和信用评分率、平均信用卡债务等;文献[6]分析了借贷者社交网络与贷款欺诈的关系;文献[7]提取了贷款者的照片来分析网贷是否成功;文献[8]分析了贷款人的描述性文本对网贷是否成功和欺诈概率的影响;文献[9]使用被提取的贷款者行为欺诈图特征去预测网贷的欺诈概率;文献[10]发现了手机使用情况与网贷欺诈的相关性。

*基金项目:国家自然科学基金(61673151);浙江省自然科学基金(LR18A050001)

本文从特征探索的角度出发,对欺诈行为尤其是团伙欺诈行为的贷前预测进行了探索。首先,利用网贷场景下普遍存在的放贷时用户授权的空间位置与放贷时间,根据团伙欺诈时空聚集的行为特性,提出了一个实用、简洁的聚集指标——K-N 最近邻指数;然后,对 K-N 最近邻指数进行序列学习;最后结合监督学习模型 LightGBM[11],对贷款进行欺诈预测。

1 坏账时空特性分析

1.1 数据描述

本文的网贷交易数据为国内某互联网公司的统计数据。网贷数据的时间长度为 61 天。交易数据仅包含申请成功且具有标签的贷款,记录无误的贷款数为 216 470 笔,其中坏账为 2 654 笔,总体的坏账率约为 1.226%。放贷时的 GPS 定位精度为小数点后两位。具体数据字段如表 1 所示。

表 1 网贷的特征

数据类型	包含信息
基本特征	性别,年龄,贷款金额,是否存在贷款历史
地理特征	身份证上的省份,身份证上的城市,贷款时所在的城市,贷款时的 GPS 定位
时间特征	贷款用户的注册时间、认证时间,贷款的申请时间

1.2 坏账时空分析

本文对前 53 天贷款数据进行分析,后 8 天贷款数据作为测试集。在数据集的特征探索中,特别分析了前 53 天坏账量最大的城市——西安的贷款分布状况。

图 1(a)为西安第 39 天(当日城市坏账率为 1.85%)的贷款空间分布情况。灰色点表示坏账率为 0 的区域(单位大小为 1 km^2),黑色点表示坏账率为 100% 的区域(单位大小为 1 km^2),点的大小表示贷款量(该图中小点为 1 笔,大点为 2 笔)。当日西安仅产生了 1 例坏账,贷款的分布都较为随机。

图 1(b)为西安第 43 天(当日城市坏账率为 18.03%)的贷款空间分布情况,灰色点表示坏账率为 0 的区域(单位大小为 1 km^2),标记为“T”黑色点表示坏账率为 82% 的区域(单位大小为 1 km^2),普通黑色点表示坏账率为 100% 的区域(单位大小为 1 km^2),点的大小表示贷款量(该图中小点为 1 笔,大点为 11 笔),除去非常异常的(108.95°E,34.29°N)区域(图中标记为“T”黑色点)后与图 1(a)相似,贷款的分布较为随机且整体坏账率较低。

(a) 西安第39天的贷款空间分布 (b) 西安第43天的贷款空间分布

图 1 西安不同日期的贷款空间分布

正常贷款行为是较为随机的,但欺诈行为往往时空集中、具有团伙性。如图 2 所示,该图上半部分表示西安地区每日的坏账率,下半部分表示西安地区前 53 天贷款量排名前三的区域(区域大小为 1 km^2)的每日所有贷款量与坏账量,点的大小表示该类贷款的数量。在贷款量排名前三的区域存在一个明显坏账空间集中的区域(108.95°E,34.29°N),该区域存在 55 笔坏账(占西安总坏账的 51.89%)。而且该区域坏账爆发时间也较集中,主要在第 40 ~ 44 天与第 48 ~ 51 天的时间段。

2 反欺诈建模

考虑到坏账之间异常的聚集关系,本文提出了 K-N 最近邻指数,一个能衡量贷款在某一阶段的空间聚集性指标。另外,本文中的观察窗口指的是观

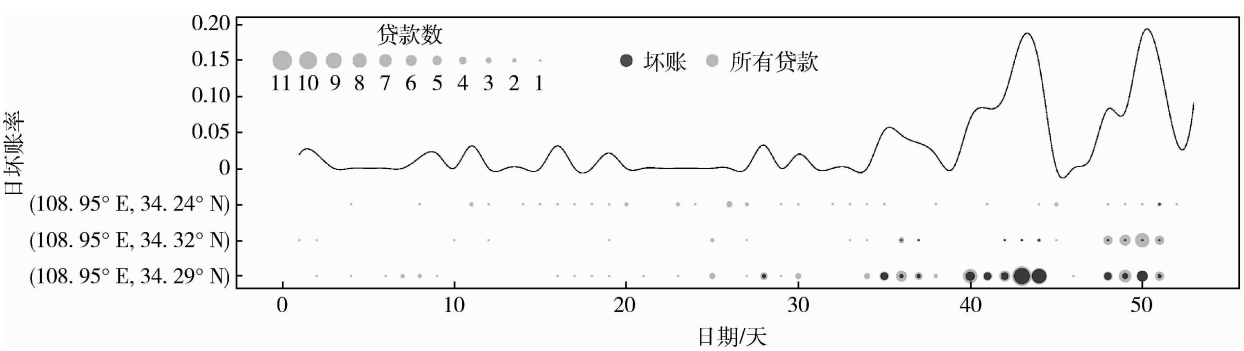


图 2 西安的日坏账率与贷款量排名前三区域的贷款情况

测贷款发生前的时间段。下文中“t 天观察窗口内观测点的邻近点”含义是在观测贷款放贷发生的前 t 天内附近贷款的空间位置。

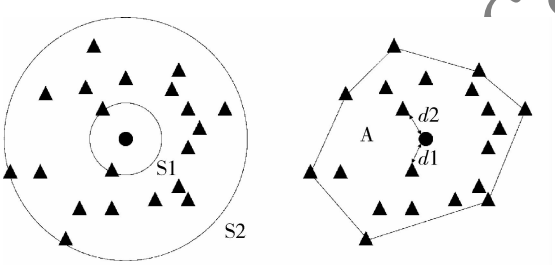
2.1 K-N 最近邻指数

最近邻分析概念[12]最初是由 CLARK P J 和 EVANS F C 提出的,用于比较区域内植物聚落情况。具体地,假定所有的点完全随机分布,则其平均距离为其密度倒数值的一半。该结果与借助图像观测到的实际的点分布格局的比值通常叫做最近邻指数(nearest neighbor index)[13]。

最近邻指数反映了一个区域内点之间的聚集程度,体现的是全局的聚集状况。受最近邻指数的启发,本文改进得到了 K-N 最近邻指数,能反映单个点局部相对聚集情况。

K-N 最近邻指数的设计概念如下:

观测点与邻近点的空间示意图如图 3 所示。图 3(a)与图 3(b)中点的含义相同,圆点为观测点(观测贷款的位置),三角形为一定时间大小的观察窗口内观测点的邻近点(邻近贷款的位置)。图 3(a)根据包含邻近点的个数先后建立两个大小不一的邻近域 S1、S2。S1 包含 k 个邻近点,S2 包含 n 个邻近点($n>k$)。通过图 3(a)的 S2 中 n 个邻近点可以组成图 3(b)的封闭图形。



(a) 观察点大小两个邻近域　　(b) K-N最近邻指数计算示意

图 3　观测点与邻近点的空间示意图

根据图 3(a)中 S2 包含的 n 个邻近点是否能组成封闭图形分成以下两种情况:

(1) n 个邻近点能组成封闭图形,如图 3(b)所示,S1 内 k 个邻近点与观测点距离的平均值 D,S2 内 n 个邻近点随机情况的平均最近邻距离 E(随机情况下平均最近邻距离为其密度倒数值的一半[13]),两者的比值表示观测点的邻近相对聚集情况。

(2) n 个邻近点不能组成封闭图形。可能存在的极端情况,即 n 个邻近点无法形成封闭的图形,呈现的几何状态为绝大数邻近点集中于某一点或者连接成一条线。此类情况本身就是非常聚集的表现,但又很难采用特定的数值进行定值。该极端情况下,本文将观测点的 K-N 最近邻指数得分定为“空值”。因为最终预测模型为 LightGBM 模型,该集成树算法在树节点进行选取最佳特征分裂点时将缺失值样本分别置于左右叶子节点,最终选择分裂增益最大的方向。所以本文将极端情况处理为“空值”,在不失其特性的情况下也是适用于最后的预测。

t 天观察窗口内某观测点的 K-N 最近邻指数具体计算步骤如下:

(1)观测点放贷发生的前 t 天观察窗口内,计算观测点空间最近邻的 k 笔贷款($k<n$)到观测点距离平均值 D:

$$D = \sum^{k} d_i / k \tag{1}$$

其中,k 为 t 天观察窗口区域内 S1 内邻近点数量,d_i 为观测点与观察窗口内 S1 内邻近点的距离。

(2)观测点放贷发生的前 t 天观察窗口内,S2 内 n 个邻近点形成如图 3(b)所示的凸包[14],按最近邻指数的定义[13],计算凸包内(包括边缘)随机情况下整个区域内邻近点最近邻距离的平均距离 E:

$$E = \frac{0.5}{\sqrt{n/A}} \tag{2}$$

其中,n 为 t 天观察窗口区域内 S2 内邻近点数量,A 为观察窗口内 S2 内全部邻近点所围成凸包的面积。如图 3 所示,通过对邻近点的凸包计算[14]得到面积 A。

(3)计算 K-N 最近邻指数 r:

$$r = D/E \tag{3}$$

较低的 r 得分表现为观测点距离观察窗口内的邻近点相对较为接近,观测点与邻近点相对更为聚集。

以第 31 ~ 53 天的贷款作为观测贷款,观察窗口大小范围为 1 ~ 30 天。控制不同的观察窗口 t、k 值与 n 值,K-N 最近邻指数都能很好地区分坏账与正常贷款,具体分析如下:

设置 K-N 最近邻指数中 $k=2$,$n=20$。如图 4 所示,按观察窗口的时间长短,从短到长取了5 天、15 天、25 天的观察窗口,坏账在 K-N 最近邻指数

低分区域的占比都很明显大于正常贷款。不同观察窗口下,坏账更容易得到较低的 K-N 最近邻指数得分。

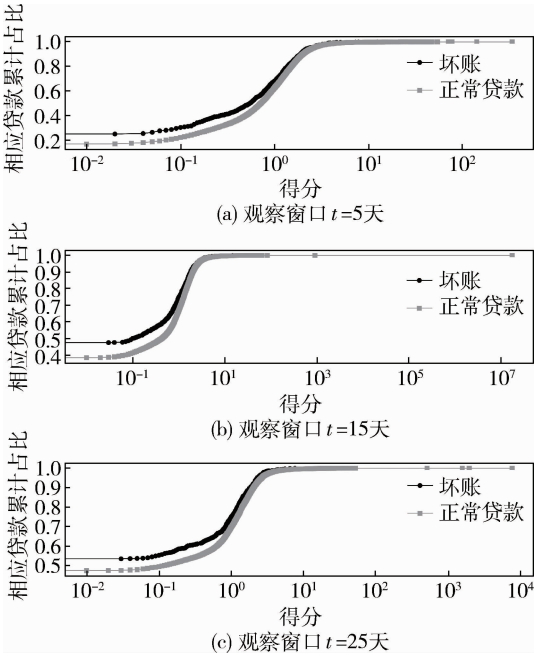


图 4　不同观察窗口下的 K-N 最近邻指数的得分累计分布

图 5、图 6 反映了不同 n 值、k 值对坏账与正常贷款的 K-N 最近邻指数得分中位数的影响。K-N 最近邻指数得分中位数会随着 n 值增大而减小,随着 k 值增大而增大,但坏账的得分中位数都明显低于正常贷款。

图 7 反映了改变 n 值对坏账与正常贷款的 K-N 最近邻指数得分的空值占比的影响。K-N 最近邻指数得分的空值占比随着 n 值增大而减小,但坏账的空值占比都明显高于正常贷款。坏账的邻近点更容易无法形成封闭图形。

所以在较为合适的观察窗口 t、k 值与 n 值情况下,K-N 最近邻指数对坏账与正常贷款有较好的区分能力。

同一个观察点不同观察窗口下存在聚集变化,为了进一步提取不同窗口下 K-N 最近邻指数的序列信息,本文以基于 LSTM 的 seq2seq 模型学习 K-N 最近邻指数序列得到最终向量来表征聚集变化。

2.2　K-N 最近邻指数的序列特征

seq2seq 模型[15]最初使用于自然语言处理领域,核心思想是通过深度神经网络模型将一个作为输入的序列映射为一个作为输出的序列。该模型最初采用的深度神经网络模型为 RNN。而 LSTM 在 RNN 基础上进行了提升,使其能够获取到更长距离的信息,从而学习到长依赖的特征[16]。

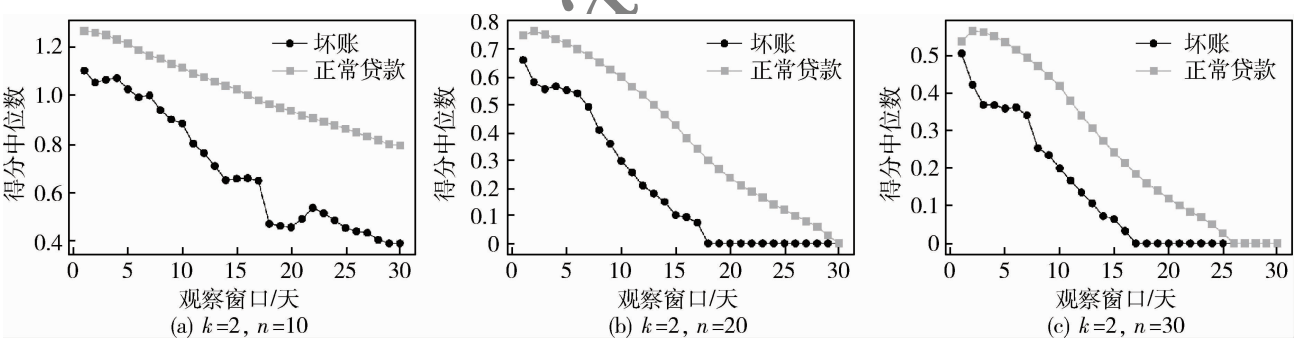


图 5　不同 n 值时的 K-N 最近邻指数中位数

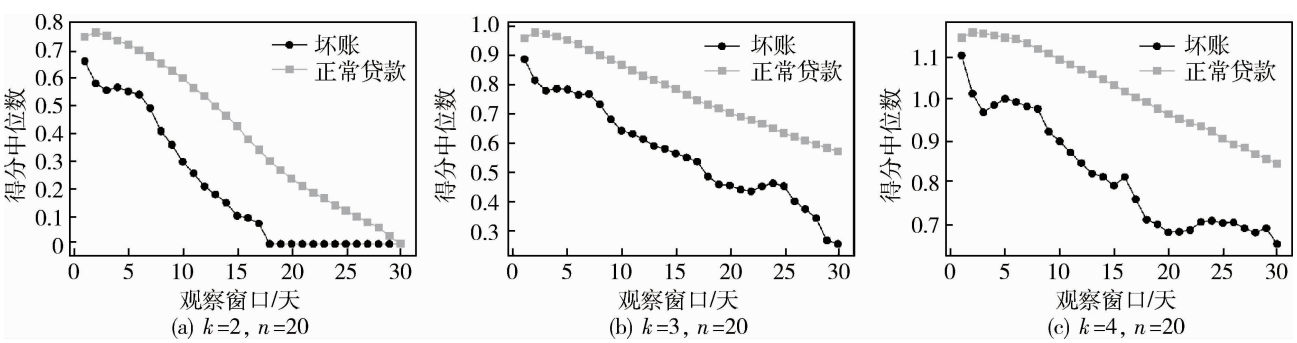


图 6　不同 k 值时的 K-N 最近邻指数中位数

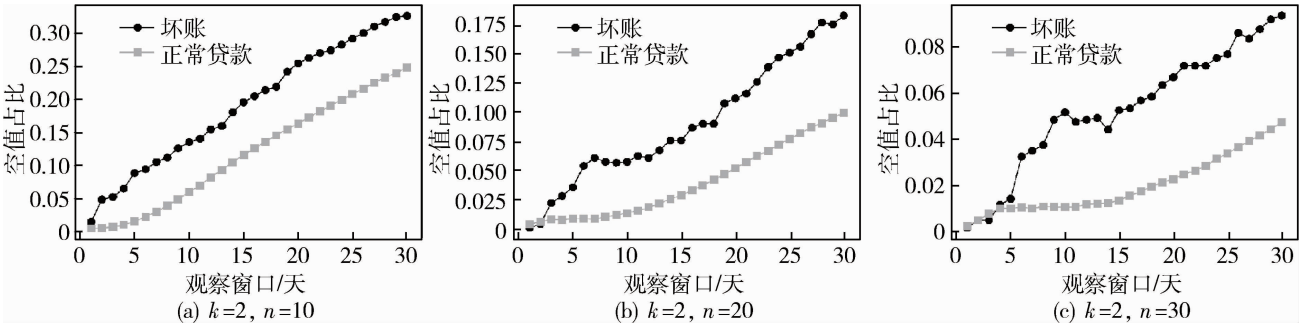


图 7　不同 n 值时的 K-N 最近邻指数空值占比

根据不同时间间隔的贷款对当前贷款作用强弱不同,将不同观察窗口的 K-N 最近邻指数,按观察窗口的长短从大到小组成相应的序列 $L1=\{r_n, r_{n-1}, \cdots, r_2, r_1\}$,并输入到第一个 LSTM 模型组成的编码器(Encoder)。如图 8 所示,编码器隐藏层状态为:

$$h_t = f(h_{t-1}, r_{n-t+2}) \tag{4}$$

其中,c 包含了输入序列 $L1$ 编码后的信息,第二个 LSTM 模型组成的解码器(Decoder)的隐藏层状态为:

$$H_t = f(H_{t-1}, f_{t-1}, c) \tag{5}$$

最后,控制一定的维度数 m,将解码器的隐藏层组成输出序列 $L2=\{f_1, f_2, \cdots, f_{m-1}, f_m\}$。$L2$ 序列作为 $L1$ 序列的 embedding 结果。$L2$ 序列存在原序列 $L1$ 元素间的交互信息,作为新特征对原模型进行提升。

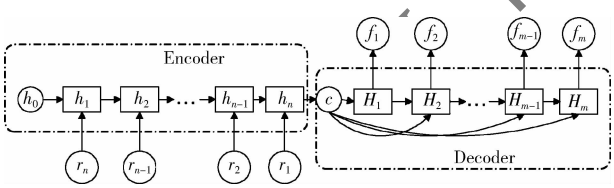


图 8　基于 LSTM 的 seq2seq 模型

3　实验与结果分析

本文利用表 1 的特征进一步建立了如表 2 所示的基础特征。

表 2　模型的基础特征

数据类型	包含信息
基本特征	性别,年龄,贷款金额,是否存在贷款历史
外地人特征	身份证城市与贷款城市是否一致
时间差特征	注册与认证时间差,认证与贷款时间差,注册与贷款时间差

模型评估采用 KS 值与 AUC。AUC 反映的是模型对测试样本整体的预测能力。作为风控建模中最为常见的指标,KS(Kolmogorov-Smirnov)值适用于正负样本极其不平衡的场景,衡量的是好坏样本累计分布之间的差值。好坏样本累计差异越大,KS 值越大,那么模型的风险区分能力越强。

本文将第 31 ~ 53 天作为训练集,第 54 ~ 61 天作为测试集,利用 LightGBM 模型对第 31 ~ 53 天网贷数据进行训练,然后对第 54 ~ 61 天网贷数据进行预测输出欺诈概率。设置 K-N 最近邻指数中 $k=2$, $n=20$,观察窗口 t 为 1 ~ 30 天,得到不同观察窗口下相应的一系列 K-N 最近邻指数 $r_{1\sim30}$。为了能使用 seq2seq 模型提取 embedding 特征又能满足序列的完整性,权衡考虑之下,此时对 K-N 最近邻指数空值进行了"-1"填充,控制输出维度在 $m=5$ 的情况下有了较好的提升。K-N 最近邻指数预测效果如表 3 所示。

表 3　K-N 最近邻指数预测效果对比

特征	KS 值	AUC
基础特征	0.364	0.745
$r_{1\sim30}$	0.210	0.624
基础特征 + $r_{1\sim30}$	0.395	0.762
基础特征 + $r_{1\sim30}$ + embedding 特征	0.407	0.776

相比于仅利用基础特征,K-N 最近邻指数与基础特征组合对于预测有了较好的提升。再对 K-N 最近邻指数序列利用基于 LSTM 的 seq2seq 模型抽取序列信息,K-N 最近邻指数类特征能对仅使用基础特征的 KS 值提高约 11.8%,AUC 提高约 4.2%。

4　结论

本文提出了一个新的适用于网贷时空聚集的指标——K-N 最近邻指标,并结合基于 LSTM 的

seq2seq 模型对不同观察窗口的 K-N 最近邻指标提取序列信息,得到观察点聚集变化的信息。最终采用 LightGBM 模型进行预测,实验结果表明该指标对坏账的预测有了较好的提升,这也说明了该指标的有效性。

参考文献

[1] HUANG R H. Online P2P lending and regulatory responses in China: opportunities and challenges[J]. European Business Organization Law Review, 2018, 19(1): 63-92.

[2] SHEN W. Internet lending in China: status quo, potential risks and regulatory options[J]. Computer Law & Security Review, 2015, 31(6): 793-809.

[3] HOU X H, GAO Z X, WANG Q. Internet finance development and banking market discipline: evidence from China[J]. Journal of Financial Stability, 2016, 22: 88-100.

[4] CAO F Q. Challenges of Internet finance to traditional finance [J]. Financial Forum, 2015 (1): 3-65.

[5] WOOLSEY B, SCHULZ M. Credit card statistics, industry facts, debt statistics [EB/OL]. (2013-04-20)[2019-12-01]. http://www.creditcards.com/credit-card-news/credit-card-industry-facts-personal-debt-statistics-1276.php.

[6] LIN M F, PRABHALA N R, VISWANATHAN S. Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending [J]. Management Science, 2013, 59(1): 17-35.

[7] GONZALEZ L, LOUREIRO Y K. When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans[J]. Journal of Behavioral and Experimental Finance, 2014, 2: 44-58.

[8] DORFLEITNER G, PRIBERNY C, SCHUSTER S, et al. Description-text related soft information in peer-to-peer lending-evidence from two leading European platforms[J]. Journal of Banking and Finance, 2016, 64: 169-187.

[9] MIN W, TANG Z Y, ZHU M, et al. Behavior language processing with graph based feature generation for fraud detection in online lending[C]. Proceedings of WSDM Workshop on Misinformation and Misbehavior Mining on the Web, 2018.

[10] LIU H, MA L, ZHAO X, et al. An effective model between mobile phone usage and P2P default behavior[C]. Computational Science-ICCS 2018. Springer, Cham, 2018: 462-475.

[11] KE G L, MENG Q, TOMAS F, et al. LightGBM: a highly efcient gradient boosting decision tree[C]. Advances in Neural Information Processing Systems (NIPS2017), 2017: 3149-3157.

[12] CLARK P J, EVANS F C. On some aspects of spatial pattern in biological populations[J]. Science, 1955, 121 (3142): 397-398.

[13] 蔡运龙,陈彦光,刘卫东,等. 地理学:科学地位与社会功能[M]. 北京:科学出版社,2012.

[14] 李洪波. Clifford 代数,几何计算和几何推理[J]. 数学进展,2003(4):23-33.

[15] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]. Advances in Neural Information Processing Systems (NIPS2014), 2014: 3104-3112.

[16] ZHENG J, XU C C, ZHANG Z A, et al. Electric load forecasting in smart grids using long-short-term-memory based recurrent neural network[C]. 2017 51st Annual Conference on Information Sciences & Systems. IEEE, 2017.

(收稿日期:2019-12-08)

作者简介:

俞旭峰(1994 -),男,硕士研究生,主要研究方向:异常检测、金融反欺诈建模、数据理论与分析。

王澎(1981 -),男,博士,讲师,主要研究方向:虚假交易行为特征和识别算法设计。

郭威(1982 -),男,硕士,主要研究方向:大数据系统研发。

版权声明