Finance 金融, 2020, 10(5), 455-464
Published Online September 2020 in Hans. http://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2020.105047

Hans汉斯

基于局部采样样本均衡的P2P借贷违约预警模型

张雪飞

对外经济贸易大学信息学院，北京
Email: zxfuibejyjx@163.com

收稿日期：2020年8月6日；录用日期：2020年8月21日；发布日期：2020年8月28日

摘　要

随着互联网金融的不断发展，P2P网络借贷的借贷人违约风险识别引起金融机构的重点关注，且随着互联网金融整改措施的实施，借贷违约量不断减少，因此在这P2P网络借贷历史违约数据不断减少的环境下，基于不均衡数据的违约预警分析显得尤为重要。本文在BSL不均衡样本抽样算法的基础上，通过Kmeans聚类算法降低抽样时间复杂度，并使用随机森林与其他机器学习分类算法进行对比实验，同时加入借款描述与借款标题的文本分析，最终建立了基于随机森林的P2P网络借贷违约预警模型来实现对于数据不均衡的P2P借贷违约风险识别。在满足高效率、高识别率的同时，满足了增量学习的现实需求，为P2P网络借贷平台提供一定的监管指导意见。

关键词

P2P网络借贷，违约预警，随机森林，样本均衡

P2P Lending Default Warning Model Based on Local Sampling Sample Equilibrium

Xuefei Zhang

E-Commerce in School of Information Technology & Management University of International Business and Economics, Beijing
Email: zxfuibejyjx@163.com

Received: Aug. 6th, 2020; accepted: Aug. 21st, 2020; published: Aug. 28th, 2020

Abstract

With the continuous development of Internet finance, the identification of borrowers' default risk

文章引用: 张雪飞. 基于局部采样样本均衡的 P2P 借贷违约预警模型[J]. 金融, 2020, 10(5): 455-464.
DOI: 10.12677/fin.2020.105047

of peer-to-peer (P2P) lending has attracted the attention of financial institutions, and with the implementation of Internet finance rectification measures, the amount of loan defaults has been decreasing. Therefore, under the environment of decreasing historical default data in P2P lending, default warning analysis based on unbalanced data is particularly important. In this paper, based on BSL unbalanced sample sampling algorithm, K-means clustering algorithm is used to reduce the complexity of sampling time, and random forest is used to compare with other machine learning classification algorithms. At the same time, text analysis of loan description and loan title is added. Finally, a peer-to-peer lending default warning model based on random forest is established to identify P2P loan default risk of unbalanced data. It not only meets the needs of high efficiency and high recognition rate, but also meets the practical needs of incremental learning, and provides certain supervision guidance for peer-to-peer lending platform.

Keywords

P2P Lending, Default Warning, Random Forest, Sample Equilibrium

1. 引言

2007 年 8 月，我国第一家网贷平台“拍拍贷”成立，标志着纯线上无担保网贷平台在中国的正式起步。近年来，P2P 行业在我国发展迅速，将金融服务领域进一步拓宽至互联网金融，增加了社会的金融普惠性。根据网贷之家数据显示，截至 2019 年 12 月底，我国 P2P 平台累计数量已达 6605 家，正常运营 343 家，成交量为 428.89 亿元。然而，由于网贷平台门槛较低、监管主体缺位等内在缺陷，存在许多借款人的违法行为，自 2017 年起，许多平台由于借款人的违约还款行为导致资金链断裂。2019 年 12 月，银监会发布《网络借贷信息中介机构业务活动管理暂行办法(征求意见稿)》，进一步加大对 P2P 的管控约束。随着国家监管制度的不断完善，P2P 行业也将逐渐规范化、成熟化，平台违约率也会越来越低，基于大数据、云计算的违约识别难度也会逐渐增大。基于不均衡数据的违约预测也将成为未来的一大难题。

在互联网金融逐渐发展的时代，学者们纷纷对 P2P 网络借贷违约行为进行违约结构分析，但随着互联网金融的整改措施，基于样本均衡的 P2P 网络借贷将会成为实际应用中的一个难题，传统 SMOTE 算法可以解决部分不均衡样本问题，但无法解决噪声数据导致的边界模糊问题，且鉴于互联网数据的爆发式增长，基于增量学习的机器学习违约行为识别算法也将成为实际应用中的一个降低时间复杂度与业务流程复杂度的重要问题。

2. 文献综述

随着 P2P 网络借贷平台的发展，学者们也纷纷进行了一些有价值的探究。在借款人的学历程度方面，廖理[1]等人研究了借款人学历在违约概率上的影响，研究发现学历较高的借款者违约概率更低。阮素梅[2]发现婚姻状况是否稳定对借款人的违约行为也有重要的影响，并且女性违约率显著小于男性。李广明等人[3]通过提取 P2P 违约借款人的特征，从 13 个变量的统计检验中发现，借贷金额小、借贷期限短、借贷利率低的借款人有较小的可能产生违约行为，并且借款者的学历高低、借款人所在城市、借款人职

业也与是否产生违约行为密切相关。刘博楠[4]表明借贷人个人特征、工作特征、信用特征、借贷特征均会对 P2P 违约风险造成不同程度的影响。

而针对于不平衡数据的处理问题，国内外学者的研究主要集中在两个方面。，首先从数据层面上采用不同的抽样方法进行样本的平衡，如 SMOTE (Synthetic Minority Oversampling Technique)算法[5]，该算法也被广泛应用在不平衡数据的处理中，但原始数据集中的噪声数据可能会使数据边界模糊造成数据分布改变。胡峰[6]等提出 TWDIDO 算法，结合三支决策理论，对边界域和负域中的小类样本进行不同的过采样处理，有效解决不平衡数据的二分类问题，但未进行一定的欠采样处理，对于样本均衡度及其不均衡的数据而言，过采样的迭代次数过高、重复度过大，同样会致使数据边界模糊，从而对原有数据分布造成一定的影响。第二个方面是降维，李杰[7]等人用随机森林的特征选择，对于不均衡样本而言，其标签分类结果无法达到最优。Sherif F. Abdoh 等人[8]用随机森林和 SMOTE 结合递归特征消除并对比 PCA 降维，发现前者的组合分类性能提高都更大。张忠林[9]等人针对边界模糊的问题，提出 BSL 采样算法，借鉴 K 近邻的思想提出边界样本、安全样本和噪声样本的概念，基于边界样本进行插值，但由于 K 近邻思想的全局搜索与过高的时间复杂度，随着样本数据的增长，算法运行时间也将大幅度增长。

而 P2P 借贷人信息中，也存在一些文本信息，例如借款标题和借款描述。李杰[7]等人在众筹违约模型中，选取项目标题、项目发起人简介等文本数据进行文本相似度计算，作为新的特征加入预警模型中。黄承慧[10]等人提出对传统的基于频率的余弦相似度改进，即 TF-IDF 模型基础上分析文本中重要词汇的语义信息，借助外部词典分析词项之间的语言相似度。Mikolov [11]等人提出 word2vec 模型，基于三层神经网络，使得预测词汇相似度的精度得到大幅度提高。Kusner [12]等人提出 WMD (Word Mover's Distance)算法来衡量文本相似度，该算法充分利用了 word2vec 的领域迁移能力，且模型简单，没有任何超参数，并且将问题转为线性规划，有着全局最优解。本文在 P2P 网络借贷领域中，引入文本相似度的计算概念，作为一列新的特征添加到模型中，进一步丰富 P2P 违约风险预警模型。

学者们已经对 P2P 违约风险预警和不均衡样本处理做了大量的研究，但依旧存在一些不足。在 P2P 违约数据越来越少的情况下，传统的欠采样方式容易丢失有用的数据，即 K 近邻距离不同样本标签较大的优异分类型数据，造成信息的缺失。简单的随机过采样方法由于简单的随机复制，容易造成过拟合问题，且模型的普适性不强。SMOTE 算法容易造成边界模糊的问题，对于采样后模型的参数要求较高，不易根据经验选择，仍然具有改变原有数据分布的可能性。借款标题与借款人描述等文本信息，未通过计算文本相似度来全面体现违约模型指标体系。

本文针对 P2P 网络借贷行业违约数据越来越少的情况，改进了张忠林[9]的 BSL 算法，通过 KMEANS 聚类算法降低抽样时间复杂度，在样本均衡的同时既改善了噪声样本造成的数据分布改变和边界模糊的问题，又大大降低抽样的时间复杂度，并使用随机森林与其他机器学习分类算法进行对比实验，同时借鉴李杰[7]的思想加入借款描述与借款标题的文本分析，通过计算文本相似度，进一步完善 P2P 借贷违约风险模型体系。最终选择并建立了基于随机森林的 P2P 网络借贷违约预警模型来实现对于数据不均衡的 P2P 借贷违约风险识别。该模型运行效率较高、违约识别率较为优秀，同时随机森林算法可以较为便捷的进行增量学习，在数据增长的情况可以通过添加一个新树的方式来进行增量学习，满足了实际业务的需求，为 P2P 网络借贷平台提供一定的监管指导意见。

3. 基于样本均衡的 P2P 借贷违约预警模型

本文在对数据进行了一定的剔除、编码和标准化后，改进张忠林[9]等人提出的 BSL 采样方法，首先利用聚类的方式将全局搜索维度改为多个局部搜索，对每一簇利用 K 近邻距离的思想，定义安全样本、

噪声样本和边界样本，使用 BSL 算法对数据重采样，对于不均衡数据进行处理，由于 K 近邻的平均思想，对于猜中或猜错也不再为绝对性选择，而是选择几个进行平均，对于采样而言，改善了 SMOTE 的边界模糊问题。接下来通过不同的分类器对比时间复杂度、违约样本识别率、模型精度，得出基于随机森林的 P2P 违约预警模型。通过测试集验证，结果表明该模型在召回率、F1 值都有一定的提升，同时随机森林也有较好增量学习特性，在数据不断快速增长的背景下，可以较为方便的进行高效率的增量学习。本文的研究思路与技术路线如图 1 所示。

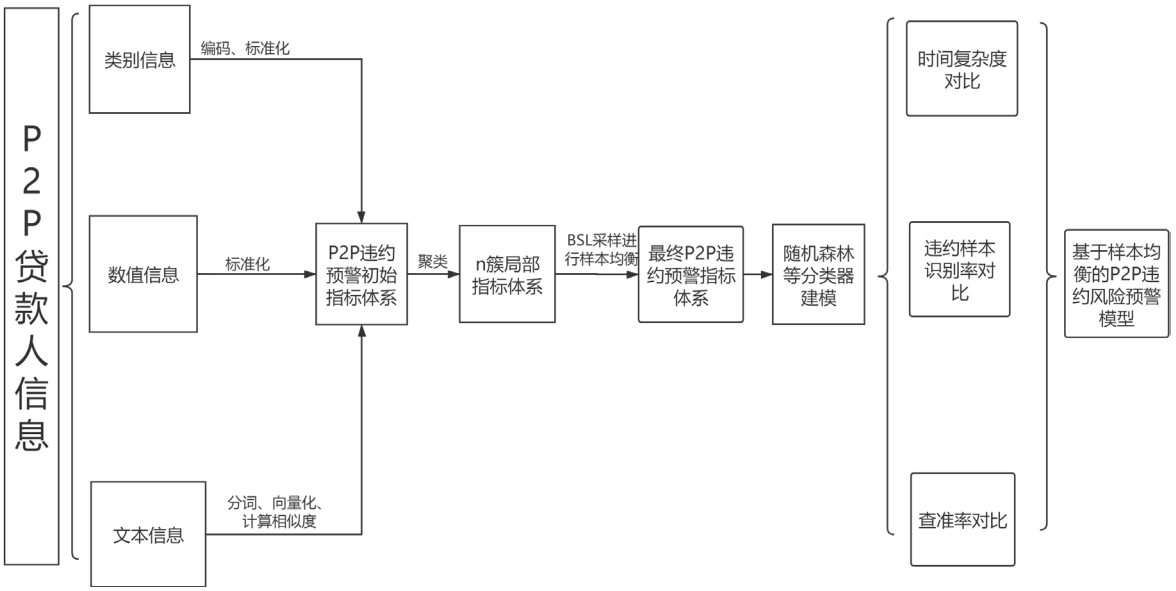

Figure 1. Research ideas and technology line
图 1. 研究思路与技术路线

3.1. 基于局部采样的 BSL 样本均衡算法

SMOTE 算法基本思想是针对少数类样本做处理，并根据样本间的欧氏距离合成新样本，合成的策略是对每个少数类样本 a，从它的最近邻中随机选一个样本 b，然后在 a、b 之间的连线上随机选一点作为新合成的少数类样本。但具有容易产生模糊边界问题，从而影响分类算法性能的缺点。

BSL [9]算法首先根据计算少数样本点的 K 近邻距离，将少数类样本点分为噪声样本、边界样本和安全样本。之后在边界样本选择的基础上进行 SMOTE 插值，对插值的边界进行限制，使得合成的少数样本在一定程度上避免模糊边界的缺陷。但由于采样算法时间复杂度过高，因此本文在采样前进行聚类，在每一个簇数中进行采样，将全局搜索维度改进为局部搜索，可以大大降低时间复杂度，从而更好地进行分类。改进后的 BSL 采样算法具体步骤如下：

算法 1：基于局部采样的 BSL 样本均衡算法。

Input：原始数据集 D，最近邻样本个数 k。

Output：新的样本训练集 T_1'。

Step 1：将原始样本数据集 D 按照聚类的方式，根据轮廓系数的拐点，选择最佳簇数 $n_{cluster}$，将原始数据集 D 划分为 $D_i\left(i=2,3,\cdots,n_{cluster}\right)$。

Step 2：对于每一个局部样本集 D_i 按照 4:1 切分为训练集 T_i^1 和测试集 T_i^2。

Step 3：按照式(1)计算第 i 簇中违约样本样本点 x_{ij} 与所有 T_i^1 中训练样本的欧氏距离，获得该样本点 k 个近邻样本。

$$Dis=\sqrt{\sum_{j=1}^{m}\left(x_{ij}-y_{ij}\right)^{2}} \tag{1}$$

Step 4：对第 i 簇少数类样本进行划分。设在 k 近邻中有 $l\left(0\le l\le k\right)$ 个属于多数样本。

若 $l=k$，x_{ij} 被定义为噪声样本。

若 $2/k\le l\le k$，x_{ij} 被定义为边界样本。

若 $0\le l<2/k$，x_{ij} 被定义为安全样本。

边界样本表示为 $\left\{x_{i1},x_{i2},\cdots,x_{ij},\cdots,x_{in}\right\}$，其中 n 表示为第 i 簇边界样本个数。

Step 5：计算第 i 簇少边界样本点与少数类样本 $x_{ij}\left(j=2,3,\cdots,n\right)$ 的 k 近邻，根据采样倍率 N，根据下式进行线性插值。

$$x_{new}=x'+rand\left(0,1\right)*\left|x'-x_{ij}\right| \tag{2}$$

Step 6：将所有簇中合成的少数类样本与原始训练样本 T_1 合并，构成新的训练样本 T_1'。

3.2. 文本分析

算法 2：WMD(Word Mover’s Distance)。

对于两段文本 D_1 和 D_2，WMD 算法使每段文本中的词都使用 word2vec 算法映射到 embedding 空间中。并且对于文本 D_1 中的每一个词，与 D_2 进行对应，计算每一对词在 embedding 空间中的距离，对所有词的距离进行加总，即该两段文本的文本相似度。WMD 充分利用了 word2vec 的领域迁移能力，具有较为出色的效果，且仅需要词向量作为输入，不需要超参数调节，并且将问题转化为了线性规划，具有全局最优解。

Input：文本 D_1 和文本 D_2。

Output：文本相似度 T_c(Travel Cost)。

Step 1：首先对两段文本 D_1 和 D_2，使用 jieba 分词，并使用哈工大停止词表，去除停止词。

Step 2：使用归一化 BOW (词袋模型)的方法分别表示 D_1,D_2，并使用 word2vec embedding 来表示其中的每一个词。

Step 3：对于每一个 D_1 中的词，设置与 D_2 中的词的权重 w。若词语义较为接近，则赋予较大的权重(即全部移动或移动距离较大)，反之则赋予较小的权重(即不移动或者移动距离较小)。将词向量欧氏距离乘以移动距离，计算出每两个词的转移代价。

Step 4：将 D_1 所有全部转移到 D_2 中，同时 D_2 中所有词也转移到 D_1 中，求全局转移代价和，而全局转移代价和的最小值，即为本文 D_1 和 D_2 的相似度 text similarity。

算法 3：余弦相似度

一个向量空间中两个向量夹角间的余弦值作为衡量两个个体之间差异的大小，余弦值接近 1，夹角趋于 0，表明两个向量越相似，余弦值接近于 0，夹角趋于 90 度，表明两个向量越不相似。欧氏距离衡量的是空间各点的绝对距离，跟各个点所在的位置坐标直接相关；而余弦距离衡量的是空间向量的夹角，更加体现在方向上的差异，而不是位置。欧氏距离能够体现个体数值特征的绝对差异，所以更多的用于需要从维度的数值大小中体现差异的分析，如使用用户行为指标分析用户价值的相似度或差异。余弦距离更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题(因为余弦距离对绝对数值不敏感)。但余弦相似度也有其缺陷，即这类算法没有很好地解决文本数据中存在的自然语言问题，即同义词和多义词。这样对于搜索的精度产生很大的影响。其计算方式如下：

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{3}$$

4. 实验与结果分析

4.1. 实验数据描述

刘礼力[13]选取 2010~2016 人人贷的数据进行违约风险分析，由于近年来互联网金融整改政策严，人人贷 2016 年后也未披露违约信息，因此往年数据参考性不强。本文结合实际情况与平台特征，通过采集器对相关信息进行爬取后，抓取翼龙贷 12,000 条数据，数据量为 12,000*41，违约标签占比 12%，属于不均衡、纬度高的样本。

Table 1. The continuous variables of default warning model
表 1. 违约预警模型连续型变量

统计指标	借款金额	还款期限	借款人年龄	借款利息率
Mean	58,627.9	11.9	43	15
Std	32,614	5.4	9	0.64
Min	97.18	3	24	9.2
Max	950,000	36	66	21
变异系数	0.56	0.48	0.21	0.04

如表 1，对于借款金额、还款年限两个连续性指标而言，其变异系数较大，数据分布较为离散，借款人年龄与借款利率低变异系数较低，数据分布较为紧凑。其中借款人年龄分布区间为 24 岁~66 岁，均值为 43 岁，表明目前 P2P 借款人主要为中年群体。李广明[3]等人指出借款金额、还款期限与借款利息率对于违约风险的重要性，本文加入借款人年龄的特征进行建模分析。

Table 2. The discrete variables of default warning model
表 2. 违约预警模型离散型变量

统计指标	借款人行业	借款类型	信用等级	是否有担保人	借款人性别	借款人所在地	受教育程度	年收入	工作年限
Unique	22	6	7	2	2	27	5	6	4
Top	农业	其他	BB	无	男	河北省	初中及以下	6~12 万元	5 年以上
Freq	6125	11,949	4634	11,001	10,093	1856	9156	5223	8636
异众比率	0.50	0.004	0.61	0.08	0.16	0.85	0.24	0.56	0.28
统计指标	是否有商业保险	借款人职业等级	借款人社保年限	借款人房产	借款人婚姻状况	借款人贷款记录	借款人信用卡额度	借款人征信报告	还款方式
Unique	2	5	5	6	8	4	7	2	3
Top	无	行业人员、学生和职位不确定人员	未缴纳	自由未按揭，价值 0~100 万元	已婚有子女	1-3 年	无信用卡	有	还本付息
Freq	11,475	8237	10,743	8123	9056	5599	8203	8414	11,480
异众比率	0.04	0.31	0.11	0.32	0.25	0.53	0.32	0.30	0.04

如表 2，对于分类型数据，我们主要观察其异众比率。对于借款人所在行业来说，约 50%的借款人从事行业为农业。其次，借贷人的受教育程度偏低，有 24%的借贷人学历为初中及初中以下。同时，借

贷人的担保类信息，担保度较低，例如有92%的借贷人都无担保人，96%的借贷人未购买商业保险，89%的借贷人未缴纳社保，68%的借贷人没有信用卡。根据刘博楠[4]的研究，这些特征结构会对P2P违约风险造成较大的影响。借款人性别中男性占比为84%，冯素玲[14]表明，男性借贷人违约概率显著高于女性概率。最后，借款类型和还款方式的异众比率过低，但通过统计发现，这两个特征对于是否还款违约具有一定的分类影响。

4.2. 数据预处理

由于直接从翼龙贷平台获取到的项目信息存在大量的标签性数据，例如项目ID，所属计划ID，借款人姓名，借款人身份证号等不适用于建模的特征，因此本文首先剔除这一类无用的标志性信息。接下来观察特征信息，其中借款人所在地区、借款人所在省份、借款人所在城市、长期居住地均描述了借款人的地理位置信息，具有较高的相似度，Abolfazl Nadi [15]提出特征过高的分类量会造成加大树模型的过拟合效应，因此本文选取借款人所在省份作为地理位置信息特征。同时，借款人所在行业原始数据标签高达837种，结合翼龙贷平台分析，表明翼龙贷借款人在填写信息时共有四个下拉列表，加上其他的自由填写，导致借款人行业分类量过高，本文通过手动筛选与合并，将其合并为22类。

预测标签y共有三类，为违约，预期和坏账，为了便于二分类的混淆矩阵计算，本文参考廖理[1]的处理方法，将借款状态为“逾期”和“坏账”的所有样本认定为违约样本。

接下来检测缺失值，通过实验发现，只有借款类型一个特征有35个缺失值，笔者通过简单的平均值填充来对缺失值进行处理。

另外原始数据中包含借款标题与借款描述的特征，李杰[7]提出利用自然语言处理，计算借款标题与借款描述的相似度，作为一个新的特征加入到原始数据中，在初步实验中，本文直接将这两列数据删除，在后续模型的优化中，本文将加入这两列数据的相似度。

之后，本文对借款人行业、借款人性别、借款人信用等级等分类标签进行label-enconder编码，将其转化数字的格式，便于分类器计算。对于借款人年龄、借款期限、借款金额等连续型特征和编码后的分类型特征，进行z-score标准化。其标准化表达式如下：

$$Z = \frac{X - E(X)}{\sigma(X)} \tag{4}$$

其中，X表示任意特征，$E(X)$表示其期望，$\sigma(X)$表示其标准差。

4.3. 实验步骤与结果分析

首先采用聚类的方法，根据轮廓系数选择簇数为11簇，对每一簇，分别进行划分训练集T_i^1与测试集T_i^2 ($i = 0,1,2,\cdots,k$表示聚类的第i簇)，其比例为4:1，然后计算少数类样本点(违约样本点)与训练集中所有样本的欧式距离，我们将k值设定为10，即取每一个少数类样本点附近最近的十个点，并判断这十个样本点中违约样本点的数量，如果数量位于$k/2$和k之间，则该样本点为边界样本点，为我们要选择的样本点。接下来继续计算所有边界样本点与少数类样本的k近邻，根据采样倍率N进行插值，最终我们通过BSL采样获取不同簇下新生成的违约样本，并将生成样本与原始样本融合，最终形成新的测试集T_1'。本文根据聚类后不同簇数的样本量，基于经验选择k值为10。

为了验证基于BSL采用的算法针对于P2P违约风险预测的性能，本文首先进行BSL采样前，比较了逻辑回归(LR)，支持向量机(SVM)，随机梯度下降(SGD)，梯度下降树(GBDT)，贝叶斯(Bayes)，决策树(DT)，K近邻(KNN)，随机森林(RF)八个分类器的预警效果，各模型采取十折交叉验证的方法计算其准确度、违约标签的召回率、违约标签的F1值，三个指标的均值以及运行时间，其结果如表3所示：

Table 3. The comparison of different algorithms before BSL sampling
表 3. BSL 采样前不同算法实验对比

模型	准确度	违约标签召回率	违约标签 F1 值	运行时间
LR	0.94	0.51	0.66	0.46s
SVM	0.92	0.29	0.44	8.48s
SGD	0.93	0.48	0.62	0.41s
GBDT	0.95	0.61	0.72	5.45s
Bayes	0.92	0.33	0.47	0.12s
DT	0.94	0.61	0.66	0.20s
KNN	0.935	0.52	0.63	3.70s
RF	0.96	0.64	0.76	0.65s

通过对比发现，随机森林针对于该数据集，在准确度、违约标签的召回率、违约标签的 F1 值以及算法效率上都具有较好的结果。

张忠林[9]所提出的 BSL 采样算法，搜索规模为全局搜索，时间复杂度较高，且随着样本量的不断增加，算法运行时间会大幅度增长，本文借鉴张莉[16]的思想，根据轮廓系数的拐点，采用聚类的方式将样本分为 11 簇，对每一簇分别进行 BSL 采样，将全局搜索改为多个局部搜素，大大降低了时间复杂度。

Table 4. The comparison of bsl sampling and local BAL sampling
表 4. BSL 采样与局部 BSL 采样实验对比

算法名称	采集样本数量	采样时间	RF 的违约标签召回率
全局 BSL 采样	759	2 h	0.71
聚类后的局部 BSL 采样	667	11 min	0.82

表 4 显示，使用聚类后的局部 BSL 采样，在采集违约样本的数量方面，由于全局和局部采样的 k 值设置相同，而局部采样每个簇数中样本量较少，因此使用局部 BSL 采样生成的违约样本数量少于全局 BSL 采样，但可以有效的提升违约标签的召回率，并大幅度降低时间复杂度，因此下文中的 BSL 采样将全部采样聚类后的 BSL 局部采样。

Table 5. The comparison of different algorithms after local BSL sampling
表 5. 局部 BSL 采样后不同算法实验对比

模型	准确度	违约标签召回率	违约标签 F1 值	运行时间
LR	0.90	0.42	0.58	0.80s
SVM	0.87	0.27	0.43	15.9s
SGD	0.89	0.51	0.64	1.02s
GBDT	0.95	0.80	0.85	9.62s
Bayes	0.89	0.43	0.57	0.62s
DT	0.95	0.75	0.76	0.69s
KNN	0.91	0.56	0.69	11.4s
RF	0.96	0.82	0.87	1.3s

表 5 显示加入局部 BSL 采样后，所有模型的运行时间变高，这是因为加入了采样后的数据量变大，针对于 KNN 算法，由于采样和算法的思想较为一致，因此违约数据的召回率得到提高，而针对于基于树思想的算法，例如 GBDT、DT、RF，模型的违约数据召回率提升幅度较大，这是因为基于树思想的算法带有过拟合的特性，而基于 BSL 采样插入的新数据在一定程度上解决了 SMOTE 算法所带来的边界模糊问题，从而在结点生成的过程中使得模型的召回率得到更大程度的提高。

本文根据李杰[7]的加入文本相似度构架违约指标体系的思想，加入项借款标题、借款描述两列文本类型数据。基于两列文本的相似度作为借款人的一项借款特征，且相似度越高，表明填写借款信息越认真，信息越充分，因此将文本相似度作为一个新的变量加入到违约预警模型中。由于借款标题和借款描述，这两列文本具有一定的结构化性质，因此首先选取其句子的余弦相似度可以用来衡量两列文本的相似度，再与 word2vec 进行模型优化。本文首先使用 Python 中文文本分析工具包 jieba 对两列文本进行分词，并使用哈工大停止词表去除停止词。列出每行两列数据的所有词并取并集，构建每个短句的词向量，计算词频，更根据式(3)并返回两个词向量的余弦相似度，计算其文本相似度，得到借款标题-借款描述关联度指标。

接下来分别对使用余弦相似度和 WMD 算法计算出来的加入文本相似度的数据，使用局部 BSL 采样进行样本均衡，并对已经选择好的 RF 模型进行十折交叉验证，将余弦相似度、WMD 算法计算出来的文本相似度与未加入文本相似度的模型对比，其结果如表 6 所示。

Table 6. The comparison of different algorithms of text similarity
表 6. 不同文本相似度算法对比

文本相似度算法名称	采集样本数量	RF 查准率	RF 的违约标签召回率
余弦相似度	721	0.962	0.74
WDM	744	0.954	0.84
无	667	0.96	0.82

通过对比发现，加入了文本相似度特征后，两者采样的样本数量均有所提高，但对于余弦相似度来说，虽然模型的查准率有了小幅度提高，但模型的召回率降低的幅度却较大，因为余弦相似度无法计算同义词的相似性，导致描述中较多的同义词没有得到该有的相似度计算。但 WDM 能在小幅度降低查准率的时候，提高违约标签的召回率，因为基于 word embedding 空间的 WMD 算法充分利用了 word2vec 的领域迁移能力，并将问题转化为求解线性规划，具有全局最优解。。

因此，本文通过在均衡样本的过程中，将 BSL 算法优化为局部搜索，并进行八个分类器的对比，加入文本相似度优化，最终选择基于随机森林的局部采样样本均衡 P2P 违约预警模型，该模型在时间复杂度、违约识别率上具有较好的性能，同时随机森林模型可以较为便捷地应用到增量学习中，在实务中应用可以通过增加一个树的方式进行学习，而不需要从头开始学习，也降低了业务流程的复杂度。

5. 结论

P2P 违约预警是互联网金融风险管理中一个重要的问题，同时随着互联网金融的监管力度不断加大，违约历史数据将越来越少，分类模型的标签不均衡性将会进一步扩大，而基于不均衡样本的违约风险识别也将会成为实务中的一大应用热点。本文首先通过爬取翼龙贷的数据，对数据进行详细的处理与全面的分析，从违约预警结构出发，其次，对于不均衡的样本，改进 BSL 算法，将全局搜索改进为若干个局部搜索，加以文本相似度的实验，最终，科学地构建了基于随机森林的 P2P 网络借贷不均衡数据的违约预警体系，同时随机森林满足了随着数据增加的增量学习需求，可以在实务中通过新添加一个树的方式

来高效地进行学习。

分析结果表明，基于 BSL 采样算法的 P2P 违约预警模型对于不均衡样本的处理性能较好，但基于全局搜索的模型，时间复杂度较高，因此在采样前进行聚类，从每个簇中挑选出边界样本进行插值，能有效降低时间复杂度并对模型进行改进。此模型有效地对具有违约风险的 P2P 借贷项目发出预警，为 P2P 网络借贷平台的资金回收、资金链安全、降低流动性风险提供了可靠的依据，具有较好的现实指导意义；同时，在实证分析中发现，借款期限、借款额度、还款方式、借款人所在地、贷款记录是 P2P 网络平台借贷人违约预警的关键因素。P2P 网络借贷平台可以根据这些指标，加强监管力度。

参考文献

[1] 廖理, 吉霖, 张伟强. 借贷市场能准确识别学历的价值吗?——来自 P2P 平台的经验证据[J]. 金融研究, 2015(3): 146-159.

[2] 阮素梅, 周泽林. 基于 L1 惩罚 Logit 模型的 P2P 网络借贷信用违约识别与预测[J]. 财贸研究, 2018, 29(2): 54-63.

[3] 李广明, 诸唯君, 周欢. P2P 网络融资中贷款者欠款特征提取实证研究[J]. 商业时代, 2011(1): 41-42+58.

[4] 刘博楠. 我国 P2P 网络借贷违约风险的影响因素研究[D]: [硕士学位论文]. 长沙: 湖南大学, 2017.

[5] Chawla, N.V., Bowyer, K.W., Hall, L.O., *et al*. (2002) SMOTE Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 1321-357. https://doi.org/10.1613/jair.953

[6] Hu, F., Wang, L. and Zhou, Y. (2018) Unbalanced Data Oversampling Method Based on Three Decision. *Electronic Journal*, **461**, 135-144. https://doi.org/10.1521/pdps.2018.46.1.135

[7] 李杰, 马士豪, 靳孟宇, Chao-hsien Chu. 基于 SA-SVM 的众筹违约风险预警模型[J]. 统计与信息论坛, 2018, 33(11): 70-77.

[8] Abdoh, S.F., Abo Rizka, M. and Maghraby, F.A. (2018) Cervical Cancer Diagnosis Using Random Forest Classifier with SMOTE and Feature Reduction Techniques. *IEEE Access*, **6**, 59475-59485. https://doi.org/10.1109/ACCESS.2018.2874063

[9] 张忠林, 曹婷婷. 基于重采样与特征选择的不均衡数据分类算法[J]. 小型微型计算机系统, 2020, 41(6): 1327-1333.

[10] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(5): 856-864.

[11] Mikolov, T., *et al*. (2013) Efficient Estimation of Word Representations in Vector Space.

[12] Kusner, M., *et al*. (2015) From Word Embeddings to Document Distances. *International Conference on Machine Learning*, Vol. 37, 957-966.

[13] 刘礼丽. 历史信用记录与当前借款违约风险关系研究——基于 P2P 平台的实证分析[J]. 中国物价, 2020(4): 49-52.

[14] 冯素玲, 赵家玲, 赵书. 女性借款人对降低网贷市场违约风险有积极效应吗?——来自“拍拍贷”的实证研究[J]. 济南大学学报(社会科学版), 2020, 30(2): 91-101+159.

[15] Nadi, A. and Moradi, H. (2019) Increasing the Views and Reducing the Depth in Random Forest. *Expert Systems with Applications*, **138**, Article ID: 112801. https://doi.org/10.1016/j.eswa.2019.07.018

[16] 张莉, 郭军. 基于边界样本的训练样本选择方法[J]. 北京邮电大学学报, 2006, 29(4): 77-80.