基于决策树的物流金融信用风险预测

霍云艳　徐鹏赢　李康乐　李　岩

哈尔滨金融学院计算机系 黑龙江 哈尔滨 150030

【摘　要】：随着各种互联网消费贷的兴起，客户的逾期风险预测成为金融行业研究的重要问题。在大数据背景下，对客户数据进行多维度建模，进行信用风险预测，构建信用逾期风险的预测模型，有利于信贷决策。以物流金融平台客户信用风险为研究对象，构建决策树模型对客户信用等级进行划分，以求获得最大化的期望收益。对比常用的分类算法 C4.5、随机森林、随机森林的时序拓展，对客户逾期行为进行分类，并通过实验验证了随机森林的时序拓展方法具有较高的分类准确度。

【关键词】：决策树；物流金融；信用风险

Logistics financial credit risk prediction based on decision tree

Yunyan Huo　Pengying Xu　Kangle Li　Yan Li

Department of Computer Science, Harbin Institute of Finance, Harbin 150030, China

Abstract: With the rise of various Internet consumer loans, customer overdue risk prediction has become an important issue in the study of the financial industry. In the context of big data, multi-dimensional modeling of customer data, credit risk prediction, and construction of a prediction model of credit overdue risk are conducive to credit decision-making. In this paper, taking the credit risk of customers on the logistics financial platform as the research object, the decision tree model is constructed to divide the credit rating of customers in order to maximize the expected returns. Compared with the commonly used classification algorithm C4.5, random forest and random forest time series expansion, the customer overdue behavior is classified, and the time series expansion method of random forest is verified by experiments to have high classification accuracy.

Keywords: Decision tree; Logistics finance; Credit risk

1 引言

物流金融（Logistics Finance）是在物流行业运营过程中，开发和应用的各种金融产品，目的在于有效地组织和协调物流行业中货币资金的流动。这些货币资金的流动具体包括发生在物流过程中的各种存款、贷款、投资、信托、租赁、抵押、贴现、保险、有价证券的发行与交易，以及金融机构所办理的各类涉及物流业的中间业务等。近年来，国内物流行业的飞速发展带动了物流金融服务的爆炸式增长，随着行业的不断发展，有效评价借款人的信用风险已经成为物流金融行业可持续健康发展的关键要素之一，智能金融决策成为研究热点[1]。

目前，不良贷款余额、不良贷款率呈普遍升高的发展态势。信用风险已经成为各个金融行业发展所面临的最重要的金融风险之一。在我国，每年平均有十万亿元以上的新增融资投放和数十万亿元存量贷款到期后的重新发放，后续的信贷危害处理、防控大面积信誉过期都存在着巨大的压力。国内大多数银行因为信贷布局不合理和集中度过高等原因，引发的信用风险会跟着时间推移，而逐渐暴露出来。物流金融作为金融市场中的一个重要组成部分，信用风险是物流金融在其迅速发展过程中面临的主要风险之一。

2 物流金融信用风险

几乎所有金融方式都会面临信用风险，又称为违约风险。物流金融信用风险是借款人(商家或消费者)由于种种原因，没有能力或不愿履行合同规定的条件而构成违约，从而使贷款方遭受损失的可能性。

与发达国家相比，我国的物流金融研究起步较晚，仍处于物流金融的理论探索阶段。1996 年，上海国际金融研究所从物资流动性的视角提出“物资银行”的概念，以物资经营者为对象推动资金良性流动。

随着物流金融业务在国内的具体开展，物流业与金融业的融合愈加深化，物流金融业务取得了突飞猛进的发展，市场规模迅速扩大，2016 年底已达到 8 万亿元的规模。然而，该业务在迅速崛起的同时，也暴露出了一系列的问题，特别是 2012 年以来国内连续爆发多起虚开仓单、重复质押等案件，给银行业、物流业造成重大损失，这些突显出信用体系不健全带来的隐患，给物流金融的发展造成了很大影响，甚

至一度造成物流金融发展的停滞。出于风险管理的考虑，众多银行与物流企业都缩减了物流金融业务规模，使得中小企业融资难的问题更加突出。

近年来,随着深度学习和人工智能的发展,一些智能化方法被应用到信用风险评估的研究中。例如，MALEKIPIRBAZARIM 等[2]使用随机森林算法对国外网络借贷平台 Lending Club 借款人的风险进行预测、马来西亚多媒体大学用随机森林的时序拓展方法对人类活动进行了分类实验。

3 决策树模型

决策树方法是最受欢迎的数据挖掘技术之一，主要应用于分类和预测。决策树方法[3]是以样本为基础的归纳分类和决策方法，它的任务是将样本划分到合适的预定义目标类中。分类算法在企业中有多种多样的应用场景，在物流金融行业中，可以根据客户的信用历史数据将其分类为信用良好客户和高风险客户。在本文中，采用决策树方法来描述物流金融信用风险的问题，通过对比 C4.5 算法、CART 算法，随机森林和随机森林的时序拓展算法，在虚拟数据集上对模型进行训练和测试。

决策树是数据挖掘技术中一种简单有效的分类算法，由一个根节点、多个内部节点和叶节点构成，根节点是整个决策树的开始，内部节点包含特征属性测试条件，用来区分具有不同特征的记录，每个叶节点代表一个类别。决策树根节点到每个叶节点的路径对应一个结果判定，通过对训练样本进行归纳学习，从无规则的实例中推理出分类规则。建立模型的基本流程如下：

步骤 1 创建新节点，对于当前节点的数据集 D，若其样本个数小于阈值，停止递归，返回决策树。

步骤 2 根据划分度量选择最优划分特征属性，把数据集划分子集，生成分支节点。

步骤 3 对新的子节点递归调用步骤 1 和步骤 2，生成决策树。

算法的终止条件一般有三种情况[4]：

情况 1 训练数据集 E 中所有的样本都属同一个类，则将此节点当作一个叶子节点，并以该类标记此节点；

情况 2 无属性可以作为测试属性；

情况 3 训练样本的数量少于用户提供的阈值。

后两种情况中一般以训练样本中占优势的类标记该叶子节点。属性选择度量有信息增益、信息增益率和 Gini 指数等。

3.1 C4.5 算法

C4.5 算法[5]以信息增益率为属性选择标准。属性 A 的信息增益率计算如下:

$$\mathrm{Split(A)} = -\sum_{j=1}^{c} \frac{p_{ij}}{|E_i|} \log_2 \frac{p_{ij}}{E_i} \qquad (1)$$

使用信息增益率作为分支的度量标准，考虑到了分支的数量，使分支的处理更符合实际需求。此外，C4.5 算法改进了过拟合等问题。采用 C4.5 算法产生的分类规则易于理解，准确率较高。但是，在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，导致算法的低效。此外，C4.5 只适合于能够驻留于内存的数据集，当训练集大得无法在内存容纳时程序无法运行。

3.2 随机森林

随机森林是一种基于决策树的集成学习的方法，能够处理高维特征数据，对缺失值和噪声数据都具有很好的容忍度。集成后的分类算法准确率和效率都明显高于原方法。随机森林是一种常用的集成分类算法，其原理类似于集成学习的装袋法，通过组合多个训练集的分类结果来提升分类效果。随机森林算法每次从所有属性中抽取 F 个属性，再从这 F 个属性中选取一个最优的属性作为其分支属性，从而使得模型的随机性增强，提高了模型的泛化能力。

集成模型中构建树时，样本由训练集的有放回抽样得到，此外，在节点分割过程中，选择的分割点是属性的一个随机子集中的最佳分割点，由于这种随机性，随机森林的偏差会增大，但对其取平均值后，方差会有所减小，通常能够补偿偏差的增加，从而产生一个总体更优的模型。

3.3 随机森林的时序拓展

随机森林的时序拓展机制体现在两方面，其一是将传感器数据按照发生的时序重排。众所周知，大多数人类活动都会受到之前的活动的影响，例如，常人在爬山时其运动速度会随时间而变化，攀爬的速度会因疲劳而逐渐减慢。因此，分类器在对活动进行分类时需要考虑之前时间戳的传感器数据[6]。其二是时序随机化，随机森林算法的时序拓展使用 Weka 开发包发布的随机森林版本进行实验，但对其做了调整，Weka 中的随机森林使用随机数作为基学习器，当所有被选择的属性均出自相同窗口时，就会忽略了时序设置。因此随机选取属性个数的计算公式在随机森林的时序拓展中被更新为：

$$m_t = ||\log_2 \sum_{i=1}^{t} M_i + 1|| \quad (4)$$

限制条件为$||m_t - \frac{m_t}{t}|| \notin A_t$，保证至少 50%的分支属性会选自不同的时间窗口。

4 模型建立与分析

4.1 数据来源

用于分类示例的数据集是 UCI 机器学习库的 7 个公开 HAR（Human Activity Recognition，人类活动识别）数据集上，使用了 10 折交叉检验。实验中针对每个数据集设计 3 组观测。人类活动识别是根据运动与环境状况识别人类活动的研究，由于智能家居、移动设备、可穿戴设备、智能织物和辅助机器人的出现，此领域得到了大量关注。主要通过两种方法对 HAR 进行研究，基于视觉（vision-based）的 HAR 和基于传感器（sensor-based）的 HAR。

4.2 实验设计

一组连续数据可以基于设定的时间窗口数量进行合并。一组传感器数据表示为 A={$a_1, a_2, \ldots\ldots, a_m$}。为观察一个大小为 t 的时间窗口内活动的情况，传感器属性的合并集的定义如下：$\widetilde{A_j} = A_1||A_2||\ldots\ldots A_t = j$，其中

$$\{(a_1, \ldots, a_M)^{W=1} \in A_1; (a_1, \ldots, a_M)^{W=2} \in A_2; \ldots\ldots; (a_1, \ldots, a_M)^{W=j} \in A_j;\}$$

其中 t 值越大，表示人类活动的影响时间越长。

在 UCI 机器学习库的 7 个公开 HAR 数据集上，设计 3 组观测（分别为 t 值从 1~3 的情况），t=1 表示无时序影响，t>1 则可观察到时序下延迟的影响，分类准确率很高。

对比 Weka 中随机森林分类结果与随机森林时序拓展分类结果，发现随机森林时序拓展分类性能更加出色。

4.3 结果分析

随机森林模型对空值不敏感，在部分样本的特征属性为空值的情况下，仍可维持分类的准确度，被广泛应用于信用评估。随机森林模型既能够通过随机抽取的方式抽取不同特征变量进行分类，又能够处理大批量、多维度的复杂数据，模型的泛化能力强，不易造成过拟合问题，具有较高的分类准确度。信用评估中样本量较大，原始数据存在较多的空值，预处理后数据仍较为复杂，并且离散型变量占多数，随机森林模型可以很好地处理这样的数据，因而可以选用随机森林算法对贷款违约行为的数据进行拟合预测。

5 结语

信用风险预测的问题被物流金融行业的每家公司所重视。本文选择 C4.5 算法、随机森林、随机森林的时序拓展算法在虚拟的数据集上进行训练和测试，C4.5 算法是经典的决策树分类算法，当数据量不大、数据关系比较简单时，优先采用 C4.5 算法，但在本实验中 C4.5 算法属性的数量急剧增长，实验结果证实在 HAR 中随机森林的时序拓展对人类活动进行分类有更优秀的准确率。

参考文献：

[1] 于晓虹,楼文高.基于随机森林的 P2P 网贷信用风险评价、预警与实证研究[J].金融理论与实践.2016(2):53-58.

[2] MALEKIPIRBAZARI M, AKSAKALLI V.Risk assessment in social lending via random forests[J].Expert Systems with Applications, 2015, 42 (10) :4621-4631.

[3] TAN Xiaomin, FANG Ai, LIANG Bing, YANG Haojie. Locating causes of abnormality of EPG experience based on decision tree. Telecommunications Science.2019.5

[4] 赵蕊.基于 WEKA 平台的决策树算法设计与实现[D]长沙:中南大学,2007.

[5] YANG Xiaojun, QIAN Lufen, BIE Zhi.Comparative Study on Decision Tree Algorithm Based on WEKA Platform. Ship Electronic Engineering. 2018.5

[6] 赵卫东,董亮.机器学习[D].人民邮电出版社.2021:94-95.

[7] Shih Yin Ooi, Shing Chiang Tan, Wooi Ping Cheah. Classify Human Activities with Temporal Extension of Random Forest[A].Neural Information Processing. New York: Springer International Publishing,2016:3-10.

资助项目：2017 年度黑龙江省省属高等学校基本科研业务费科研项目（2017-KYYWF-E0102）

2021 年度黑龙江省省属本科高校基本科研业务费项目（2021-KYYWF-020）