一种面向网络支付反欺诈的自动化特征工程方法

王成[1)2)3)]，王昌琪[1)2)]

[1)](同济大学 电子与信息工程学院 计算机科学与技术系，上海 201804)

[2)](嵌入式系统与服务计算教育部重点实验室，上海 201804)

[3)](上海智能科学与技术研究院，同济大学，上海 200092)

摘 要 互联网金融欺诈正导致诸多社会经济问题。网络支付是互联网金融中的典型模式之一，此模式中的欺诈交易也是互联网金融欺诈的主要形式之一。通过构建基于机器学习的欺诈检测模型来识别欺诈交易的方法已成为网络支付反欺诈领域的主流思路。在构建欺诈检测模型的过程中，特征工程是最为关键的一步，特征的质量将直接影响模型的性能；通常，这也是最为耗时且对相关领域的专业知识要求最高的步骤。现有网络支付欺诈检测模型在特征工程上主要是领域专家基于业务知识以手动构造的形式来开展。而在网络支付模式下欺诈场景众多，不同场景下的特征构造流程不尽相同。人工特征构建方法已不能满足与日俱增的反欺诈需求。解决此问题的重要方法之一便是自动化特征工程。本文针对网络支付欺诈检测提出了一种轻量化、树结构、高效率、可扩展和可解释的自动化特征工程方法。该方法：（1）对计算条件的要求低且对数据集样本的依赖性小，这一优势是利用树结构模型进行特征构造得以实现；（2）可构造出深度层次的复杂特征和广度层次的各类型特征，这一优势是利用节点处特征构造的新型流程和转换函数权重向量的时效性更新机制得以实现；（3）在网络支付模式不同场景下可实现跨场景复用，这一优势是通过复用和扩展定制化转换函数得以实现；（4）构造出的特征具有可解释性，这一优势得益于基于结合转换函数与树模型的特征构造过程具备可表达性。本文在网络支付典型场景的业务数据集上验证了所设计自动化特征工程方法的有效性。

关键词 网络支付；互联网金融；欺诈检测；自动化特征工程；机器学习

中图法分类号 TP311

An automated feature engineering method for online payment fraud detection

Wang Cheng[1) 2) 3)]，Wang Chang-Qi[1)2)]

[1)](Department of Computer Science and Technology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

[2)](the Key Laboratory of Embedded System and Service Computing, Ministry of Education, China, Shanghai 201804, China)

[3)](Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092, China)

Abstract Internet finance fraud is an increasingly serious social and economic problem. Online payment services (OPSs) are the typical models of Internet finance, and the fraudulent transaction in OPSs is also a typical fraud pattern. The method of identifying fraudulent transactions by constructing a fraud detection model based on machine learning has become a promising idea for online payment anti-fraud. In the process of constructing fraud detection models, the feature engineering is the most critical step. It is also one of the most time-consuming and specialized steps in the relevant area. In the study of feature engineering, the existing online payment fraud detection models are mainly carried out by experts in the form of manual construction based on business knowledge. However, there are many fraud scenarios in OPSs where the process of feature construction is so different. Artificial feature construction methods can no longer meet the increasing demand of anti-fraud. An important way to solve this problem is to automate feature engineering. In the field of Internet financial anti-fraud, the expressibility and interpretability of features play a pivotal role. It is helpful to understand the original source fields and their construction process of important features. This is useful for

本课题得到国家自然科学基金(61972287)资助. 王成，男，1980 年生，博士，教授，博士生导师，主要研究领域为网络服务优化与安全、社交计算与网络、网络空间异常检测和互联网金融反欺诈工程，E-mail: cwang@tongji.edu.cn. 王昌琪，男，1994 年生，硕士生，主要研究领域为互联网金融反欺诈、数据挖掘，E-mail: wcq1051068488@tongji.edu.cn.

mining and analyzing the characteristics of fraud methods and follow-up improvement rules engines. These are of great significance for fraud detection models. Therefore, the interpretability of the model method is particularly important. Usually, the optimization of detection accuracy is carried out under the premise of ensuring interpretability. This paper proposed a lightweight, tree-structure, high efficiency and scalable automatic feature engineering method for fraud detection of online payment. The method is as follows: (1) The method has low requirements on the calculation conditions and little dependence on the dataset samples. To realize this advantage, it used the tree structure model to construct the features. (2) The method can construct complex features in the depth level and various types of features in the breadth level. To realize this advantage, it utilized the unique process of feature construction at the node and the time-dependent updating mechanism of the transforming function weight vector. (3) Cross-scenario multiplexing can be implemented in different online payment scenarios. This advantage is realized by the customized transforming function that can apply to different scenarios. (4) The constructed features are interpretable. The transforming function and the tree model were combined to realize the expressible process of feature construction. Finally, the effectiveness of the method was sufficiently validated on two typical real-world OPS datasets. For the first, we used the method to automatically generate the features required by the fraud detection model. Then we evaluated its effectiveness on two online network payment transaction datasets. It was verified that our method can reduce the time spent on feature engineering steps from days level to hours level compared with the manual methods. In summary, the method proposed in this paper does well in reducing the complexity of feature engineering, reducing labor costs, and improving the overall work efficiency of model development. It provides insights into how to design effective automated feature engineering methods for online payment fraud detection, taking into account both the interpretability and high performance.

Key words Online payment; Internet finance; Fraud detection; Automated feature engineering; Machine learning

1 引言

移动互联网和大数据等信息技术的兴起使得互联网金融应运而生，这给人们的日常生活带来了便利，与此同时，互联网金融欺诈也伴随而生，包括信用卡欺诈、财务报表欺诈、证券和商品欺诈、保险欺诈、抵押欺诈和洗钱等等[1]。互联网金融服务的蓬勃发展导致欺诈事件的发生概率大幅增加[2,3]，不仅给个人和企业带来了不可估量的损失[1,2,3,4]，还会导致诸多的社会经济问题。

网络支付服务是互联网金融的典型模式之一，其已经渗透进越来越多人的生活，与其相关的安全问题也显得尤为重要。在网络支付欺诈手法不断演进、环境场景趋向多变的情况下，反欺诈技术亦亟需向更高效和更普适的方向发展和进步。当前，基于机器学习的欺诈检测方法已成为网络支付反欺诈的主流技术之一。在这类方法中，特征工程通常是极其重要的技术环节。事实上，在应用机器学习技术的大多数领域中，特征工程通常都是一项费时费力的工作，需要数据专家手工进行数据清洗、分析和设计生成特征变量的工作。随着数据技术在各领域应用规模的扩大，人工构造特征的弊端越发显现出来。在网络支付模式下的反欺诈场景中，学习一个检测模型有时会需要输入数百个特征变量；若利用人工构造特征的方法，针对另外一个应用场景则需要重新进行特征工程工作，其中有很大一部分将可能是重复性工作。这会带来大量的人工成本和时间成本浪费，影响模型开发和运营效率；而且这种方法受限于人工经验，可能会遗漏部分有效特征。如何有效利用和复用知识，减少人工和时间成本，自动化生成高效能和全方面的特征，对于互联网金融欺诈检测模型的开发来说意义重大。与此同时，

[1] Tjx data breach: ignore cost lessons and weep, https://www.cio.com/article/2434423/tjx-data-breach--ignore-cost-lessons-and-weep.html.
[2] 2015 global fraud loss survey by the cfca, http://docplayer.net/22825345-2015-global-fraud-loss-survey.html.
[3] Global online payment methods: full year 2016, https://www.researchandmarkets.com/research/398sp7/global_online.
[4] The us sees more money lost to credit card fraud than the rest of the world combined, http://read.bi/18Gin67.

学界和业界关于自动机器学习（AutoML）的研究方兴未艾，其中自动化特征工程也是诸多交叉研究领域的重点。

目前与自动化特征工程相关的工作大致可分为两类：一类隐性自动化特征工程方法，如 2.2.2 节所述，其特征的构造过程是不可解释的。这些方法通过降维或升维方法将原始特征映射到另一个特征空间当中去，实现新特征的生成；或者利用神经网络复杂的中间层计算出输出层新的节点作为新特征。在互联网金融领域中，欺诈检测模型不仅需要检测出异常，还需要了解决策是如何做出来的，即异常背后的原因，以及时更新新的风险策略。对于用户正常的操作而言，如果被拦截也需要得到合理的解释。使用基于机器学习构建的信用风险模型或者反欺诈模型仅仅提高精度是没有意义的，其必须具有可解释性，为决策提供依据。除了模型的可解释性，特征的可表达性和可解释性同样发挥着举足轻重的作用，其有利于了解重要特征的原始来源字段及其构造过程，这对于挖掘分析欺诈手段的特点和跟进改善规则引擎与反欺诈检测模型具有重要的意义。因此，在互金反欺诈领域，模型方法的可解释性尤为重要[4,5,6,7]，对检测精度的优化也通常是在保证可解释性的前提下开展。这类隐性的自动化特征工程方法由于其特征的不可解释性，在网络支付模式下并不适用。另一类是显性自动化特征工程方法，如 2.2.3 节所述，其特征构造的流程都是可见的。其主要可分为两种：一种是通过模型评估来引导搜索的特征空间探索方法，实现新特征的构造；另一种是先生成所有候选特征后，再进行特征选择，得到最终的新特征集合。目前已有的这些显性方法具备特征的可解释性，但在时间和空间上都会产生高额的成本。这对于应用场景更新和模型算法迭代都较快的互联网金融行业来说，并不能快速响应且保持很好的自适应性，所以在网络支付模式下也并不适用。据我们所知，当前尚缺少针对网络支付反欺诈的自动化特征工程方法。

本文针对网络支付反欺诈问题提出一种基于特征构造树的轻量化、高效率、可扩展和可解释的定制化自动化特征工程方法，有效降低对计算条件的要求和对数据样本的依赖性。主要创新性如下：

（1）结合树模型和定制转换函数的自动化特征构造方法，不需要繁琐的预训练过程，在树的构建过程中即可完成新特征的构造，可以快速有效地生成所需特征；并且，特征构造过程具备可表达性，使其构造出的特征具有可解释性。

（2）在定制化特征构造树中增加转换函数的组合功能，使其可在深度层面上构造复杂特征；增加转换函数权重向量的时效性更新机制，使其在广度层面上保证特征构造的广泛性；通过复用和扩展定制化转换函数，实现跨场景复用。

本文在两个在线网络支付交易数据集上，基于已验证有效的人工特征和所设计自动化特征工程方法的生成特征，利用目前主流机器学习模型分别训练并进行欺诈检测，通过性能比较，验证了所提自动化特征工程方法的有效性。

本文其他部分组织如下：第 2 节介绍了互联网金融欺诈检测和自动化特征工程的研究背景和相关工作；第 3 节介绍了定制化自动化特征工程方法的实现细节；第 4 节依次介绍了数据集情况、实验相关的设计细节和相关的评价指标，并对实验结果进行了分析和讨论；最后，第 5 节总结了本文的工作并对未来研究方向进行了展望。

2 研究背景和相关工作

2.1 互联网金融欺诈检测

互联网金融欺诈检测已由传统的黑白名单、规则[8,9]和专家系统向机器学习欺诈检测模型方向发展[10,11]。机器学习在互联网金融欺诈检测系统中的应用已比较广泛。目前业内采用的机器学习欺诈检测方法主要有两大类：基于监督学习的检测[12-16]和基于无监督学习的检测[17-19]。前者基于欺诈和正常交易的样本来训练模型，以判定新来的交易是欺诈交易或合法交易。后者将异常值或异常交易归类为欺诈交易的不同类簇，利用聚类算法[20-21]将交易分组到不同的范围中，并将不属于范围中的交易识别为欺诈。这两种欺诈检测方法都可以预测任何特定交易中欺诈的可能性。本文主要关注于有监督的欺诈检测方法。这方面已有诸多进展：张芸芸等[22]提出了基于 Neo4j 图谱检测信用卡欺诈的方法，将数据集的特征通过图数据库直观展示出来，并且通过 FICO 评分标准建立了 FICO 模型，显著提高了信用卡欺诈的认定率。徐永华等[23]利用采样来的信用卡消费数据训练好一个支持向量机检测系统，然后用支持向量机检测系统对一次信用卡消费行为进行检测，判断是否为欺诈交易行为，检测精度达到 95%以

上。Sahil 等[24]使用十种有监督的机器学习模型来识别欺诈或非欺诈交易并比较它们的准确度等，其中包括随机森林（RF），逻辑回归（LR），支持向量机（SVM），梯度提升树（GBDT），集成分类器（Stacking）等等；其总体结果表明，使用逻辑回归作为元分类器的堆叠分类器最有希望用于预测数据集中的欺诈事务，其次是随机森林和 XGBoost 分类器。Alex 等[25]提出了一种用于真实信用卡欺诈检测问题的定制贝叶斯网络分类器（BNC）算法，其创建是由超启发式进化算法（HHEA）自动执行；该算法将关于 BNC 算法的知识组织到分类中，并针对给定数据集搜索这些组件的最佳组合。Gabriel 等[26]基于交易中涉及的主要实体的历史呈现模型，并且通过检索特征以判定交易是否为欺诈。Alejandro 等[27]扩展了交易聚合策略，并在使用 von Mises 分布分析交易时间周期性行为的基础上创建一组新特征，进而评估不同特征集如何影响信用卡欺诈检测。

基于机器学习方法的欺诈检测系统在其模型开发过程中通常会涉及到建模的多个技术环节，包括数据预处理、特征工程、算法设计和模型选择调参等。上述相关工作主要考虑的是模型算法的设计，而相对较少关注特征工程这一环节；其特征工程通常是根据模型算法手动进行特征构造，其业务逻辑相当繁琐，不同的模型需要的特征构造过程也不尽相同。在与网络支付反欺诈相关的安全领域，与特征工程相关的代表性工作有: Liu 等[28]通过将恶意软件转换为图像进行可视化和分类，来识别特定类型的恶意软件。他们提出了一个新的学习框架，与现有的局部描述符一起使用，通过将它们分组为块并使用新的视觉词袋模型来获得更具区分性和鲁棒性的特征，在三个恶意软件数据库上实现了最新的分类性能。Yang 等[29]提出了两种新颖的方法来识别恶意软件，其中一种是解决恶意软件家族集群问题。他们引入了 t-SNE 算法以可视化特征数据，然后确定恶意软件家族的数量，这种方法准确性更高且具有良好的自适应能力。特征工程在安全领域、互金领域具有重要的作用，其在一定程度上可以决定模型能够达到的性能水平。因此，对于欺诈检测系统的设计，除了模型算法的设计部分，考虑特征工程步骤的潜在提升空间，利用自动化特征工程方法提高这一步骤的效率，以求优化整个模型系统的开发过程。这是本文的研究重点和目标。

2.2 自动化特征工程方法

2.2.1 特征工程概述

特征工程是根据要解决的业务问题，从原始数据中提取出更多信息的过程。特征工程的目的是在不添加新数据的条件下，为模型训练提供更丰富的信息。特征工程通常包括数据清洗、特征设计、特征变换和特征选择等技术环节。特征工程是模型训练的前置条件，模型训练的优劣，通常取决于特征设计和加工的效果。

人工的特征工程方法通常由数据专家依靠其领域内的专业知识，通过迭代试错法和模型评估法来进行，一般具有非常复杂的步骤。业务领域内的专家需要开发大量的代码和脚本，从逻辑上设计、处理特征，其中有很多一部分工作是重复性的，会增加很多时间、人力成本；同时，人工构造特征会存在很多盲区，特征算法专家通常很难在有限的时间内发现并加以构造。在互联网金融风控中，欺诈手段在不断演进，使得模型时效性较强。传统模式下的手动特征构造便显得捉襟见肘。在此背景下，自动化特征工程方法成为解决上述问题以降低风控成本的备受期待的可行方法。

2.2.2 隐性自动化特征工程方法

部分机器学习方法可以自动生成新特征，但是这些特征生成的过程是隐性的，只是在一定程度上间接地体现出了特征工程的思想。比如，有一类降维方法，如主成分分析（PCA）、带核函数的核主成分分析（Kernel PCA）和嵌入方法（Embedding）等，它们将高维的原始数据特征向量映射到低维的特征向量空间中去，将特征转换到另外一个特征空间去。相对而言还有一类升维方法，比如支持向量机（SVM），它是使用核函数隐性地将原始特征向量映射到高维的特征空间当中去；多层神经网络可以自动地通过最小化损失函数来学习出有用的特征，但是其网络结构设计复杂，需要大量的样本来学习网络权重，对于小数据集来说并不适用。

2.2.3 显性自动化特征工程方法

相对的，也存在一些自动化特征工程方法可以显性地生成新的特征，即特征构造的过程是可解释的。按照构造新特征的整体顺序，可以将这类方法分为自顶而下和自底而上的方法。

（1）自顶而下

自顶而下的方法一般是先生成大量的候选特

征，再利用特征选择方法选出特征重要性程度高的特征。这类方法一次性添加大量特征，会增加特征空间的维度，导致计算量过大，并且这类方法在评估大量的候选特征时也可能会产生计算问题，需要合理的筛选策略。目前有一些与自顶而下方法相关的工作：

DSM (Data Science Machine、Deep Feature Synthesis)[30]：应用在关系数据库上，使用一组预定义的转换函数将实体表转换为特征，最后进行特征选择和模型超参数优化。这种方法以主体表为中心一次性生成所有的候选特征再进行选择，随着特征维度的增加会导致计算量过大。

OneBM (One Button Machine)[31]：应用在关系数据库上的，按照关系表之间的关系链接所有的表，再通过预定义的操作运算符执行自动特征构造，最后再执行特征选择操作筛选出最终的特征集。这种方法生成大维度的特征，并且不能够构造一些复杂的组合特征。

ExploreKit[32]：使用一组预定义的结构化操作生成所有可能的候选特征集合，然后利用候选特征的一系列元特征训练一个机器学习模型来评估候选特征的价值。这种方法需要生成所有候选特征集合，并且需要构建出机器学习模型给出候选特征的排名分数进行评估并筛选，计算量较大。

Cognito[33]：以分层的方式探索各种特征的构造，即进行树状的特征探索（广度优先或者深度优先），通过贪婪的勘探策略逐步最大化模型的准确性。这种方法下特征的数量与步骤的数量成指数关系，没有考虑预算的限制，还需要合适的特征选择策略以去除冗余特征。

（2）自底而上

自底而上的方法一般是在构造新特征的同时进行评估，逐步选择有效特征。这类方法就是派生特征的增量评估的过程，取决于添加派生特征的顺序，通常需要合理高效的搜寻探索策略的参与，整个流程才能更加高效。其主要有以下几种：

LFE (Learning Feature Engineering)[34]：使用元特征为每一个转换函数训练一个多层感知机（MLP），给出当前特征下应用当前转换函数带来模型性能提升的可能性。这种方法下多层感知机的训练依赖数据分布的多样性，并且需要大量数据集。

RL (Reinforcement Learning)[35]：根据特征构造的过程提出了转换图的概念，对转换图利用强化学习进行应用转换函数的路径探索学习。状态是当前数据特征集合的向量表征，动作是应用状态转换函数。利用函数近似拟合数据特征集合的状态空间和转换函数动作的 Q 值表。这种方法同样需要大量数据集，并且依赖数据分布的多样性。

FCTree (Feature Construction Tree)[36]：使用决策树在原始特征和构造的特征集合上利用信息增益划分数据节点。每个转换函数根据应用到当前节点上带来的效用进行权重更新。这种就是最普通的特征构造树，没有考虑转换函数的组合，不能进行复杂特征的构造。

本文所提出的方法是属于显性自动化特征工程方法中自底而上的一类特征构造方法。本方法相对于隐性方法具备特征的可解释性；本方法相对于自顶向下的方法计算性能更高；同时在自底而上的方法中利用了树模型的本方法在计算性能和对数据集样本的依赖性上具有明显的优势。通过利用树模型与转换函数相结合的方式，其对生成的新特征根据树模型，以从根节点到叶节点的搜索方式进行了增量评估，其不但可以构造深层次的复杂特征（这类特征的构造过程通常很复杂），也能在广泛性的层面上构造各类型的特征。

2.2.4 在线网络支付模式下应用难点

据我们所知，当前针对网络支付反欺诈问题的自动化特征工程研究相对较为欠缺，其主要有以下难点：

- 自动化特征工程本身存在困难性。已有方法对于时间（计算）和空间（数据集）的依赖非常高，这对于时效性极强的网络支付模式来说显得的成本巨大。互联网金融行业更新发展速度快，模型算法迭代的速度相当快，自动化特征工程需要能够及时响应且保持很好的适应性。

- 网络支付反欺诈应用对于特征的可表达性和可解释性要求较高。了解特征的由来和构造过程对于追根溯源、挖掘分析欺诈手段的特点，跟进改善规则引擎与反欺诈检测模型具有重要作用。需要在显性自动化特征工程的自底而上的方法中利用合理高效的搜寻探索策略来平衡特征构造的成本与其可解释性。

- 网络支付服务涉及面广且场景性强，对规范化和可扩展性需求较高。不同的场景下数据字段大多不同且千变万化，使得自动化特征工程在不同场景下难以实现规范化，已有自动化特征工程相对较少关注规范化和可扩展这方面需求。

3 具体方法

网络支付欺诈检测系统的设计流程如图 1 所示，其包括数据获取、数据预处理、特征工程、模型选择与训练和实时测试与维护模块。本节将对特征工程部分展开，详细介绍所提面向网络支付欺诈检测的自动化特征工程方法的实现细节。

图 1 欺诈检测系统设计流程

表 1 网络支付交易记录可利用原始字段示例

字段名	字段描述
User_ID	用户交易卡号
Transaction_Time	交易时间
Transaction_Amount	交易金额
Pre-trade_Balance	交易发生前的账户余额
Daily_Limit	每天交易总额的金额限额
Single_Limit	每笔交易金额的金额限额
Check	交易的验签方式
Frequent_IP	交易是否使用用户的常用 IP 地址

表 1 是一组网络支付交易记录可利用的原始字段的示意表，特征工程的目的是需要在此基础之上构造出新的特征作为机器学习模型的输入。本文针对网络支付服务，设计采用了定制化特征构造树的自动化特征工程方法来自动进行特征构造。算法有三个主要部分，如图 2 整体框架图所示，包括：第一部分是针对互联网金融网络支付的定制化转换函数设计；第二部分是定制化特征构造树中每个结点处的局部特征构造流程；第三部分是定制化特征构造树中转换函数权重向量的时效性更新机制。

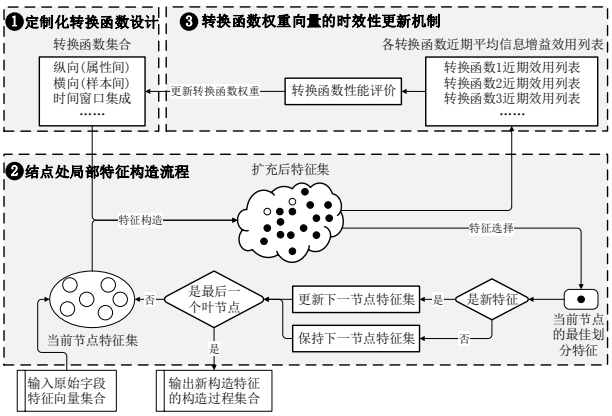


图 2 特征构造树算法整体框架

3.1 定制化转换函数设计

特征构造是原始特征字段进行变换的过程，其会涉及到转换函数的概念，转换函数囊括了代数运算、集成计算等等操作，同时还可以进行特征缩放或者将特征与类别的关系从一个非线性关系转化为线性关系，它可以把特征从原始空间映射到一个全新的特征空间当中去，也可改变原始特征的分布状况，并可以改变原始特征的取值覆盖范围等。这些转换的目的都是生成新的特征。

转换函数的类别可以按照其输入所需的特征数量来划分，其可以被划分为一元转换函数、二元转换函数和多元转换函数。本文使用的转换函数只涉及到了一元转换函数和二元转换函数。按照转换函数的工作域方式，则可以将转换函数主要划分为三类：纵向方式的转换函数、横向方式的转换函数和时间窗口方式的转换函数。

纵向方式的转换函数是作用在单个特征或多个特征属性之间的转换函数，如图 3 所示。转换函数作用于单个特征的，比如可以对虚线框中的交易金额这一列特征求开方值，从而获得一列新特征，相类似的还可以计算平方、sigmod 和 tanh 值等等；转换函数作用在特征之间的，比如可以对两个虚线框中的交易金额和特征 2 这两个字段求差值，获得一列新特征，类似的还可以在特征之间作加法和乘法等。总之，纵向方式的转换函数对每一条交易记录数据的计算环节发生在单个字段列上或者多个字段列之间。

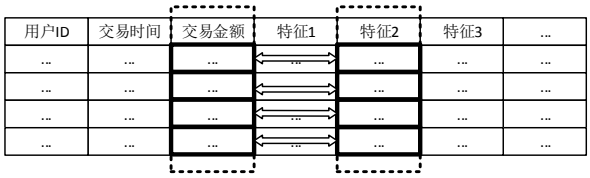


图 3 纵向方式的转换函数的作用域

横向方式的转换函数是作用在一个特征字段下的多个不同样本之间的转换函数，如图 4 所示。比如可以对交易金额这一列特征按用户分组求相邻两笔交易的差，获得用户交易金额差的新一列特征，也可以对交易金额这一列特征按用户分组进行累积求和，获得用户累积交易金额的新一列特征。类似的，还可以计算某个特征的频率、群体累计求和或者累积计数等等。总之，横向方式的转换函数的计算环节发生在一个字段下的多个行之间。

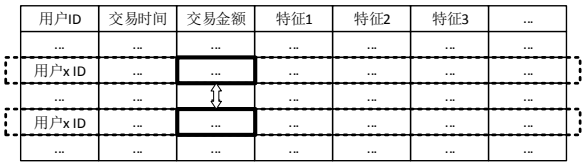


图 4 横向方式的转换函数的作用域

时间窗口方式的转换函数利用了滑动时间窗口的概念，这对于分析在一段时间内的交易行为特征具有重要的作用。其是作用在同一个特征字段上对时间窗口内的多个样本进行操作的转换函数，如图 5 所示。比如可以对交易金额这一列特征按用户分组求一段时间内的累积和，获得用户一段时间内的累积交易金额这一列新特征。类似的，时间窗口方式的转换函数还可以有时间窗口内的求极值、均值、方差、计数、非重计数、最频繁等等。

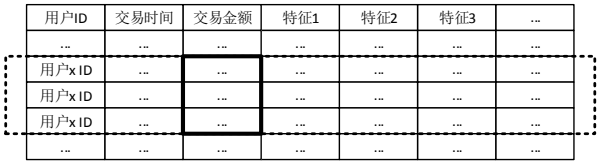


图 5 时间窗口方式的转换函数的作用域

3.2 结点处局部特征构造流程

如图 6 所示的特征构造树，本文定制化的特征构造树在每个节点处不仅仅在交易记录的原始特征集合的基础上构造新特征，还存在转换函数的组合，即在构造出来的新特征的基础上继续构造特征。这里特征构造树保留了父节点上构造出的用来划分数据集的特征，与原始特征组成新的、扩充的特征空间，在此扩充的特征空间上再进行特征构造并选择划分数据集的特征。这种局部特征构造流程增加了转换函数的组合功能，扩充了特征空间的搜寻范围。

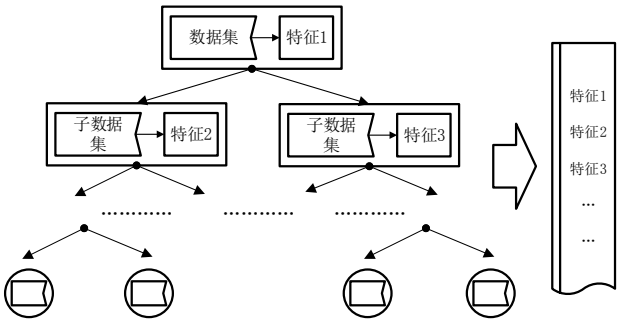


图 6 特征构造树结构

3.2.1 符号定义

在互联网金融网络支付反欺诈中，具体地，假设 $\mathbf{D}$ 是整个网络支付交易数据集，$\mathbf{D}=\{\mathbf{X},\mathbf{Y}\}$。其中 $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$，其中 $\mathbf{x}_i$ 对应的是第 i 条交易记录的各个字段，即一条特征向量，$\mathbf{X}$ 代表所有交易记录的特征向量集合；$\mathbf{Y}=\{y_1,y_2,\ldots y_n\}$，其中 y_i 对应的是第 i 条交易记录是否是欺诈，其取值 $y_i=\{0,1\}$，0 为正常，1 为异常，$\mathbf{Y}$ 代表所有交易记录标签的集合。两者共同组成了整个数据集 $\mathbf{D}$，数据集中交易记录样本的总数为 n。

令 $\mathbf{F}_o$ 表示数据集中的原始字段的特征集合，$\mathbf{F}_a$ 表示当前节点上包含新特征的特征总集合，其既包括原始的特征又包括能够通过转换函数新构造出来的并用来划分数据集的特征，$\mathbf{F}_s$ 表示树中节点处被选择用来划分数据集的新特征及其构造过程的集合。表 2 为各特征集合的示例。

表 2 各特征集合介绍示例

集合	示例
$\mathbf{F}_o$	交易金额、交易时间等原始特征;
$\mathbf{F}_a$	交易金额、交易时间等原始特征; 新特征 1、新特征 2 等新构造的特征;
$\mathbf{F}_s$	新特征 1:mean_window('交易金额'); 新特征 2:var_window('交易金额'); …… 等新特征及其构造过程。

令 $\mathbf{O}$ 表示转换函数的集合，$\mathbf{W}=\{w_1,w_2,\ldots,w_k\}$ 表示转换函数的权重向量，w_i 代表第 i 个转换函数的权重，即被选中的概率，转换函数的总个数为 k。g_f 代表在节点处选择特征 f 作为划分属性得到的信息增益（等同于 ID3 决策树中信息增益的计算方式，信息增益也可以替换为 CART 决策树中的 GINI 指数，本文以下的算法设计和实验部分都是基于信息增益）；g_o 代表在节点处使用转换函数 o 生成的所有特征分别作为划分属性得到的信息增益的均值；$\mathbf{l}_o=\{g_o^{t-m+1},g_o^{t-m+2},\ldots,g_o^t\}$ 代表转换函数 o 的最近 m 次被选中的平均信息增益效用的列表，m 是列表 $\mathbf{l}_o$ 的长度，g_o^t 代表使用 t 时刻选择的转换函数 o 生成的所有新特征，作为划分属性得到的平均信息增益效用值。

3.2.2 构造流程

特征构造树的结构图如图 6 所示。下面具体介绍整个特征构造树进行特征构造的步骤。

步骤 1，初始化转换函数集合$\mathbf{O}$中的转换函数的权值向量$\mathbf{W}$，其中每个$w_i=1/|\mathbf{W}|$；初始化每个转换函数o的最近平均信息增益效用列表$\mathbf{I}_o$，列表的长度设为m，其中的每个值的初始值为 0，初始化$\mathbf{F}_a=\mathbf{F}_o,\mathbf{F}_s=\varnothing$；

步骤 2，在决策树的某个节点上，根据转换函数的权值向量$\mathbf{W}$，依概率选中一个转换函数(权重值越大的转换函数被选中的概率越大)。若其为一元转换函数，则在此节点对应的数据集上，从数据集中所有的s个特征中选择出r个不同的特征，其中$r\le s$，并且$s=|\mathbf{F}_a|$，即特征集合$\mathbf{F}_a$的大小。在这r个特征上应用此转换函数，构造出r个新特征；若其为二元转换函数，则在此节点对应的数据集上，从数据集中所有的s个特征中选择出r组不同的特征对，$r\le \mathrm{C}_s^2$，在这r组特征对上应用此转换函数，构造出r个新特征；

步骤 3，对新构造出来的r个特征和节点中原来的特征$\mathbf{F}_a$，分别计算用每个特征作为划分属性的信息增益g_f，选择信息增益最大的特征$\tilde{f}$作为划分属性，根据特征$\tilde{f}$的具体划分值将数据集划分成左右两部分，并分裂成左右两部分子树，将样本中特征$\tilde{f}$的取值小于具体划分值的样本归并到左子树中，相对地，另外一部分归并到右子树中，它们分别对应左儿子和右儿子节点。若特征$\tilde{f}$为新构造出来的特征，则将特征$\tilde{f}$添加到新构造出来的特征集合$\mathbf{F}_a$当中去，即$\mathbf{F}_a=\mathbf{F}_a\cup\tilde{f}$，且将特征$\tilde{f}$及其构造过程并入集合$\mathbf{F}_s$中去；

步骤 4，根据 3.3 节中转换函数权重向量的时效性更新机制更新转换函数的权重值；

步骤 5，分别进入左儿子和右儿子节点，判断节点中子数据集样本数是否低于设定的最小阈值T，或子数据集样本的纯度是否高于设定的阈值G；若是，则到达叶子节点，结束流程，若不是，则重复步骤 2-4，直至到达叶子节点。当树构造完毕，则进入步骤 6；

步骤 6，整棵树构造完毕后，最终得到特征集合$\mathbf{F}_s$中的特征即为由特征构造树构造出来的新特征及其构造过程。

3.3 转换函数权重向量的时效性更新机制

本文采用信息增益均值来评价各个转换函数构造出的特征的优劣，具体来说，在节点处，通过一个转换函数首先构造出r个新特征，信息增益均值代表的是分别用这些新特征作为数据集的划分属性，得到的信息增益的均值。若一个转换函数的信息增益均值比较高，则其构造出的特征相对来说就是性能较好的，因此应该增加此转换函数的权重，使其具有更高的几率在之后的节点中被选中，相反，信息增益均值较低的转换函数被后续节点选中的几率相应的就应该降低。但是，如果每次选中某个转换函数得到的性能评价都是高的或者都是低的，就会造成某些转换函数的权值变得很高，而某些转换函数的权值变得很低，在后续节点上，转换函数的选择会偏向其中的某个或某几个，造成构造出的特征过于单一的情况。这是一个探索和利用的权衡（Trade-off）问题，既要利用已有的具有更高权重的转换函数，也要考虑到其它转换函数。因此每个转换函数都需要维护一个最近平均信息增益效用列表，根据转换函数最近的性能表现来更新其对应的权重向量，增强时效性，保证权值向量不会收敛到某个或某几个值，使构造的特征更具广泛性。其步骤如下：

步骤 1，根据 3.2.2 节中特征构造的步骤，若当前节点选择出的转换函数为o，则根据其构造出所有r个新特征，分别将其作为数据集的划分属性，按以下公式计算得到平均信息增益g_o，

$$g_o=\frac{1}{r}\sum_{1}^{r}g_{f_i}, \tag{1}$$

其中，g_{f_i}代表使用由转换函数o构造出的第i个新特征f_i作为划分属性得到的信息增益。利用g_o来更新当前转换函数o的最近平均信息增益效用列表$\mathbf{I}_o$，将此平均信息增益g_o添加到列表$\mathbf{I}_o$的末尾，删除列表$\mathbf{I}_o$头的第一个值，即：

t 时刻：$\mathsf{l}_o=\left\{g_o^{t-m+1},g_o^{t-m+2},\ldots,g_o^{t}\right\}$，

$t+1$时刻：$\mathsf{l}_o=\left\{g_o^{t-m+2},g_o^{t-m+3},\ldots,g_o^{t+1}\right\}$，

其中 $g_o^{t+1}=g_o$。

步骤 2，根据当前转换函数 o 的最近平均信息增益效用列表l_o和平均信息增益 g_o，计算出当前转换函数 o 的奖励值 β，其中，l_o^{med} 表示列表l_o中的中值，l_o^{max} 表示列表l_o中的最大值，公式（2）限定了 $\beta\in[0,1]$：

$$\beta=max\left\{0,\frac{g_o-l_o^{med}}{l_o^{max}-l_o^{med}}\right\}. \tag{2}$$

步骤 3，根据当前转换函数 o 的奖励值 β，按照公式（3）更新转换函数的权值向量，再按照公式（4）进行转换函数权值向量的归一化：

$$w_o=w_o*e^{\left(\frac{\beta}{1+\beta}\right)^{\alpha}}, \tag{3}$$

$$w_i=\frac{w_i}{\sum_1^k w_j}. \tag{4}$$

其中，w_o 表示转换函数 o 的权值，公式（3）中 w_o 随奖励值 β 增加而单调增加，也就是说奖励值越高，转换函数权重的增加幅度越大，α 控制着权重更新的速率；公式（4）中 w_i 表示第 i 个转换函数的权值，$\sum_1^k w_j$ 表示所有转换函数的权值总和。

步骤 4，在下一个节点，根据新的转换函数的权值向量W，依概率选择出当前转换函数 o，重复步骤 1-3 直至到达叶子节点。

3.4 定制化特征构造树算法

综合上述自动化特征工程方法的细节，下面给出此定制化特征构造树的算法流程。

算法 1. 定制化特征构造树

输入：训练数据集D；原始字段的特征集合F_o；转换函数集合O；

参数：转换函数 o 的最近平均信息增益效用列表l_o的长度 m；节点样本数阈值 T；节点纯度阈值 P；转换函数权重更新参数 α；

输出：树中节点处被选择用来划分数据集的特征及其构造过程的集合F_s。

1.初始化阶段

转换函数的权值向量W，$w_i=1/|\mathsf{W}|$；

每个转换函数最近平均信息增益效用列表l_o中 m 个值为0；当前节点的特征字段$\mathsf{F}_a=\mathsf{F}_o,\mathsf{F}_s=\varnothing$；

2.特征构造阶段

在当前节点，依据W中概率选择一个转换函数 o，

IF o 是一元转换函数 THEN

在所有的 s 个特征中选择出 r 个不同的特征，（$r\le s$，$s=|\mathsf{F}_a|$），

使用转换函数 o，以此构造出 r 个新特征F_r；

IF o 是二元转换函数 THEN

在所有的 s 个特征中选择出 r 组不同的特征对，（$r\le \mathrm{C}_s^2$，$s=|\mathsf{F}_a|$）；

使用转换函数 o，以此构造出 r 个新特征F_r；

3.特征评估和权重向量更新阶段

FOR each 特征 f in $\mathsf{F}_r\cup\mathsf{F}_a$ DO

计算其作为划分属性得到的信息增益 g_f；

END FOR

选择$\mathsf{F}_r\cup\mathsf{F}_a$中 g_f 最大的特征 f 作为划分属性划分数据集，分裂成左右两部分子树；

IF 特征 f 为新构造出来的特征 THEN

将 f 添加到集合F_a当中去，即$\mathsf{F}_a=\mathsf{F}_a\cup\tilde{f}$，且将特征 $\tilde{f}$ 及其构造过程并入集合F_s中去，即$\mathsf{F}_s=\mathsf{F}_s\cup\tilde{f}$；

FOR each 特征 f in F_r DO

计算其作为划分属性得到的信息增益 g_f；

END FOR

根据公式（1）计算F_r中特征的平均信息增益 g_o；

更新转换函数 o 的最近平均信息增益效用列表l_o；

根据公式（2）计算转换函数 o 的奖励值 β；

根据公式（3）（4）更新并归一化转换函数权值向量W；

4.构造树过程

分别进入左儿子和右儿子节点，

IF 节点中样本数<阈值 T or 样本纯度>阈值 P THEN

此分支到达叶子节点，结束流程；

ELSE

重复步骤2-4，直至到达叶子节点；

FINALLY 所有分支到达叶子结点 RETURN 集合F_s。

4 实验设计与结果分析

4.1 数据集介绍

为了验证所提方法的有效性，本文在两个在线网络支付交易数据集上，利用本方法进行了自动化特征构造的实验，下面先简要介绍下这两个数据集。

表 3 第三方移动支付交易数据集的原始字段特征

特征	数据类型	特征描述
id	String	用户的交易 id
date	Datetime	交易发生的时间，精确到小时
label	Int	交易的欺诈类型标签
f1	Float	交易特征字段 1
…..	Float	其它 296 个交易特征字段

第一个数据集来自某金融科技平台的第三方移动在线网络支付交易的公开数据[1]。数据涉及的时间范围是从 2017 年 09 月 05 日到 2017 年 11 月 05 日，数据集是由两个月的交易记录样本组成的，其中有一部分是带有正负标签（是否是欺诈）的交易记录样本，而另一部分是没有带正负标签的交易记录样本。数据集在经过了严格的脱敏处理后，除去时间（精确到小时）、交易 ID、欺诈标签外，其余共包含 297 个没有具体含义的交易字段特征，并且都是 float 数值类型。此数据集的原始字段特征如表 3 所示。

表 4 银行 B2C 在线网络支付交易数据集的原始字段特征

特征	数据类型	特征描述
User_ID	String	用户的交易卡账号，每个用户有一个唯一的用户 ID
Transaction_Time	String	交易发生的时间，从年开始，精确到秒
Transaction_Amount	Float	交易的交易金额，单位为 RMB
Pre-trade_Balance	Float	交易发生前用户账户的余额，单位为 RMB
Check	String	交易的验签方式
Frequent_IP	Boolean	交易是否使用的是用户的常用 IP 地址
Label	Int	交易的欺诈类型标签

[1] 第三方移动在线网络支付交易数据集地址：https://dc.cloud.alipay.com/index#/topic/data?id=4

第二个数据集来自某商业银行的 B2C 在线网络支付交易记录数据。这份数据包含了从 2017 年 04 月 01 日到 2017 年 06 月 30 日的三个月的交易记录数据，这些交易记录样本中同样有一部分是带有正负标签的，而另一部分是带灰色标签的，即不确定是否是欺诈样本。在此数据集中，我们使用到的数据字段特征共有 7 个，其特征类型和具体含义描述如表 4 所示。

4.2 实验环境

实验的服务器配置是Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz and 128 GB RAM；实验的编程语言环境是Python 3.6。

4.3 数据预处理

首先原始数据存在数据缺失、格式错误等问题，需要对原始数据进行预处理。在第一个数据集中，存在很多缺失值，对于存在很多缺失字段的交易记录，我们将这些交易记录进行删除；对于存在很多缺失值的字段，我们将所有交易记录的这个字段都进行去除，而对于缺失值较少的字段，则使用固定值进行填充。在第二个数据集中，同样一些交易数据记录的时间或金额字段存在缺失值，我们将这些交易记录予以去除。

为了验证我们方法的有效性，还需要对数据集进行划分，选出部分数据集来进行测试。考虑到交易记录数据的有序性，为了防止出现时间穿越现象，对于第三方移动在线网络支付数据集，我们取了后 15 天即从 10 月 20 日到 11 月 05 的交易记录作为测试集以评估我们方法的性能。数据集中没有标签的支付行为样本官方是指极有可能有欺诈的交易记录，实际上在我们的实验中，我们将其作为了欺诈样本来进行训练和测试。整个训练集和测试集的正负样本数据分布如表 5 所示。其中训练集的正负样本比约为 1：58，而测试集的正负样本比约为 1：57。

表 5 第三方移动在线网络支付交易数据集的样本分布

训练集	正样本（欺诈样本）数	12346
	负样本（正常样本）数	721337
测试集	正样本（欺诈样本）数	4501
	负样本（正常样本）数	256547

对于银行 B2C 在线网络支付交易数据集，整个数据集同样也被划分成了训练集和测试集两个

部分，其中训练集包含了 4 月份和 5 月份的共 2459334 条样本记录，测试集包含了 6 月份的共 1042714 条样本记录。数据集中带有灰色样本标签的交易数据记录是不确定是否是欺诈样本的，与第一个数据集不同，这里在实验中将这一部分的数据样本予以去除，所以整个训练集和测试集的正负样本数据分布如表 6 所示。其中训练集的正负样本比约为 1：59，而测试集的正负样本比约为 1：40。

表 6 银行 B2C 在线网络支付交易数据集的样本分布

训练集	正样本（欺诈样本）数	40393
	负样本（正常样本）数	2393817
测试集	正样本（欺诈样本）数	24900
	负样本（正常样本）数	1003539

4.4 分类器

利用本文所述的自动化特征工程方法生成新特征，与交易记录的标签一同输入常用机器学习模型中进行训练就可以得到欺诈检测模型。在互联网金融领域，信用风险模型或者反欺诈模型通常需要具有可解释性，为决策提供依据，一般可以选择决策树、线性回归、逻辑回归、朴素贝叶斯分类器、K-近邻算法和一些决策树集成算法等等。由于本文自动化特征工程算法是以树模型为基本模型进行设计的，所得的新特征应用在树模型上，对于在节点处选择特征进行数据集的划分时会起到更大的作用。所以，本文选择以决策树作为基学习器的 RandomForest、XGBoost 和 LightGBM 三个当前集成效果较好的分类器作为学习模型，以验证本文所设计自动化特征工程方法的性能。

4.5 实验参数设置

这部分我们对实验中的部分参数的设置进行说明，主要有以下三种参数：

（1）3.2.2 节特征构造流程的步骤 2 中，在每个节点处选中转换函数进行特征构造时，每个节点应该构造 r 个新特征。r 值的选择取决于原始特征的个数，其可以任意指定为小于原始特征数的值，在实际应用中，当原始特征数量较小时（一般小于等于 50 个特征），r 值应设置为与原始特征数量相同的数值，保证特征构造的数量；当原始特征数量较大时，r 值可以设置为原始特征数量的 1/3~2/3 之间的数值，增加特征构造的随机性，不同的数据集 r 值应随不同情况进行调试改变。在下面的实验中，我们仅将其设置为不同的数值来做对比实验；

（2）转换函数 o 的最近平均信息增益效用列表 l_o 的长度 m。一般决策树的高度不会超过 10 层，树中的非叶子节点接近 500 个，在数据集一上本文应用了 8 个转换函数，平均每个转换函数被选中的次数接近 60 次；而在数据集二上应用了 14 个转换函数，平均每个转换函数被选中的次数接近 35 次。通常最近几次被选择的列表大小取平均被选择次数的 1/10~1/2 之间的值比较合适，这里是 35/10~35/2 和 60/10~60/2。因此，m 的取值可以设置为 10，在两个数据集上都在区间内，即根据最近 10 次转换函数的信息增益效用值计算奖励值。

（3）公式（3）中权重更新的速率参数 α。图 7 是数据集一上 α 设置为不同数值时各转换函数选取频数分布图。当 α 越大，频数分布越平均，各转换函数被选中的次数越平均，生成的特征类型也会更广泛，但是当 α 过大时，广泛的特征中绝大多数是没有价值的，这使得转换函数的选择效果与随机选择的效果差距很小，因此 α 不应过大，以便在一定程度上保持在当前数据集下转换函数的选择偏好；相反，当 α 较小时，某些转换函数会频繁地被选中，出现收敛的现象，这会导致生成的特征类型单一。由图 7 中可知，当 α 的取值设置为 4 时，兼具特征广泛性和转换函数选择偏好的特点。

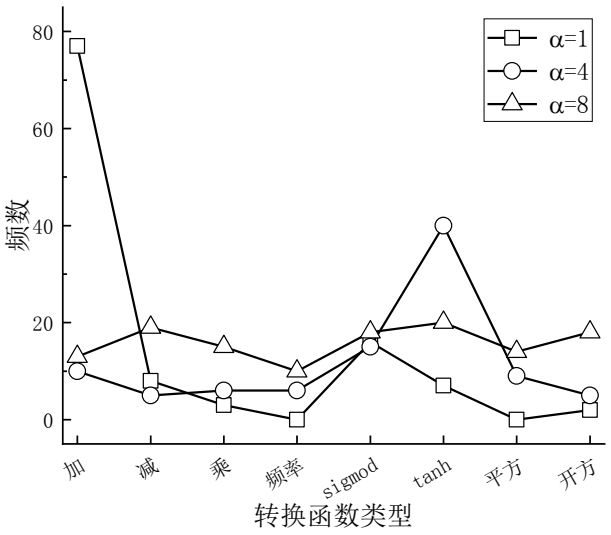


图 7 不同 α 取值下转换函数的频数分布

4.6 评价指标

本文采用了不同假正例率（False Positive Rate、打扰率、误查率）下的真正例率（True Positive Rate、召回率、查全率）、加权真正例率和 ROC 曲线下面积（AUC）来作为欺诈检测性能评价的指标，以此评估本文自动化特征工程方法的性能。

假正例率（FPR）即将负样本预测为正样本占总负样本的比例，在这里指将正常样本预测为欺诈样本占总正常样本的比例；真正例率（TPR）即将正样本预测为正样本占总正样本的比例，在这里指将欺诈样本预测为欺诈样本占总欺诈样本的比例。TPR 与 FPR 成正比关系，本文分别设定 FPR 在 0.1%,0.5%和 1%下时 TPR 的值作为衡量模型性能的指标。

加权真正例率（$TPR_{weighted}$)是我们根据业界应用过的一个评价标准手动设置的一个指标，它的取值由以上三种 FPR 取值下的 TPR 进行加权求得的，它的计算方法表示如下：

$$\begin{aligned} TPR_{weighted} &= 0.4*\left(TPR|FPR=0.001\right) \\ &+ 0.3*\left(TPR|FPR=0.005\right) \\ &+ 0.3*\left(TPR|FPR=0.01\right). \end{aligned} \tag{5}$$

该指标反映了将 FPR 控制在 0.1%,0.5%和 1%时 TPR 的一个综合性能水平。当 FPR 为 0.1%时，其权重为最高的 0.4,这是因为我们的目的就是在更低的 FPR 下取得尽可能高的 TPR，权重的取值也反应了业界对于各 FPR 的重视程度。当然这仅仅是手动设置的一个指标，其权值和各 FPR 的设定不固定，可以根据实际情况作相应调整。

AUC（area under ROC curve）是评价欺诈检测模型性能的常用指标，其对具有类不平衡问题的数据集具有良好的适应性。在 ROC 曲线中，横坐标表示 TPR，纵坐标表示 FPR，对每一个分类阈值，分类器都有对应的 TPR 值和 FPR 值（即对应坐标系上的一个坐标点）。所有坐标点连接成的平滑曲线就是该分类器对应的 ROC 曲线。而 AUC 对应的则是 ROC 曲线下的面积，其取值范围是[0,1]，越接近于 1，则代表其对应的分类器性能越好。

在特征重要性的评估方面，随机森林具有自带的特征重要性计算方法，其对于某个特征 f 的重要性计算步骤如下：首先在随机森林中的每一棵决策树上，使用没有用来进行树的训练的袋外数据计算这棵树的袋外数据误差，将其记为 oob_{err1}；接着在袋外数据上所有数据样本的特征 f 上随机地加入噪声，干扰特征 f 的数据分布，再次计算这棵树的袋外数据误差，将其记为 oob_{err2}；最后假设随机森林中有 N 棵树，那么对于特征 f 的重要性如下公式所述：

$$importance_f = \frac{1}{N}\sum oob_{err2} - oob_{err1}. \tag{6}$$

如果某个特征在加入随机噪声后，袋外数据误差大幅度升高，那么说明这个特征对于模型的预测结果具有重要影响，也就代表这个特征的重要性较高。

4.7 实验结果分析

4.7.1 第三方移动在线网络支付数据集结果

在此数据集上，由于交易记录没有关于用户的个人交易卡账号信息，时间字段只精确到天，所以实验中只使用了纵向方式和部分横向方式的转换函数,没有使用到时间窗口方式的转换函数，具体如表 7 所示，没有使用到横向方式和时间窗口方式的转换函数。

表 7 第三方移动在线网络支付数据集上应用的转换函数

类型	名称	函数	解释
纵向方式	平方	square()	对一列特征的值求平方
	开方	square_root()	对一列特征的值求开方
	sigmod	sigmoid()	对一列特征的值求 sigmod
	tanh	tanh()	对一列特征的值求 tanh
	加法	sum(,)	对俩列特征的值进行相加
	减法	subtraction(,)	对俩列特征的值进行相减
	乘法	multiplication(,)	对俩列特征的值进行相乘
横向方式	差	diff()	对一列特征相邻样本求差
	群体累积	acc()	对一列特征进行累积求和
	群体计数	cum_count()	对一列特征进行累积计数
	频率	frequent()	对一列特征的值求频率

数据集在经过数据预处理后，除去交易 id 号和交易时间字段后的原始特征总数为 285 个。通过改变在每个节点处利用转换函数构造出新特征的个数 r，我们利用本文所述的方法做了两组

实验。其中，实验一中的r设置为 90，即在每个节点处先利用转换函数构造出 90 个候选新特征，由此整个特征构造树构造出来的总的新特征的个数为 73 个；而实验二中的r设置为 190，即在每个节点处先利用转换函数构造出 190 个候选新特征，由整个特征构造树构造出来的总的新特征的个数为 96 个。

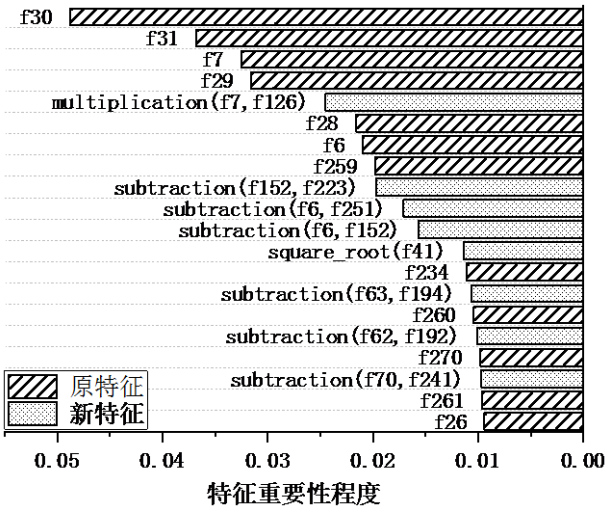


图 8 数据集一上实验设置一的重要性前 20 的特征

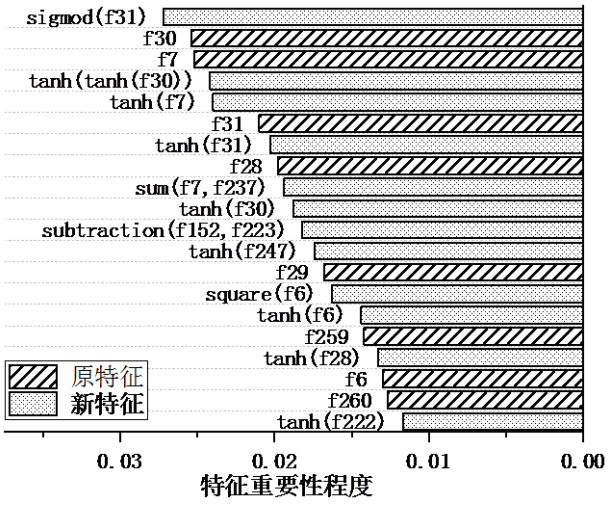


图 9 数据集一上实验设置二的重要性前 20 的特征

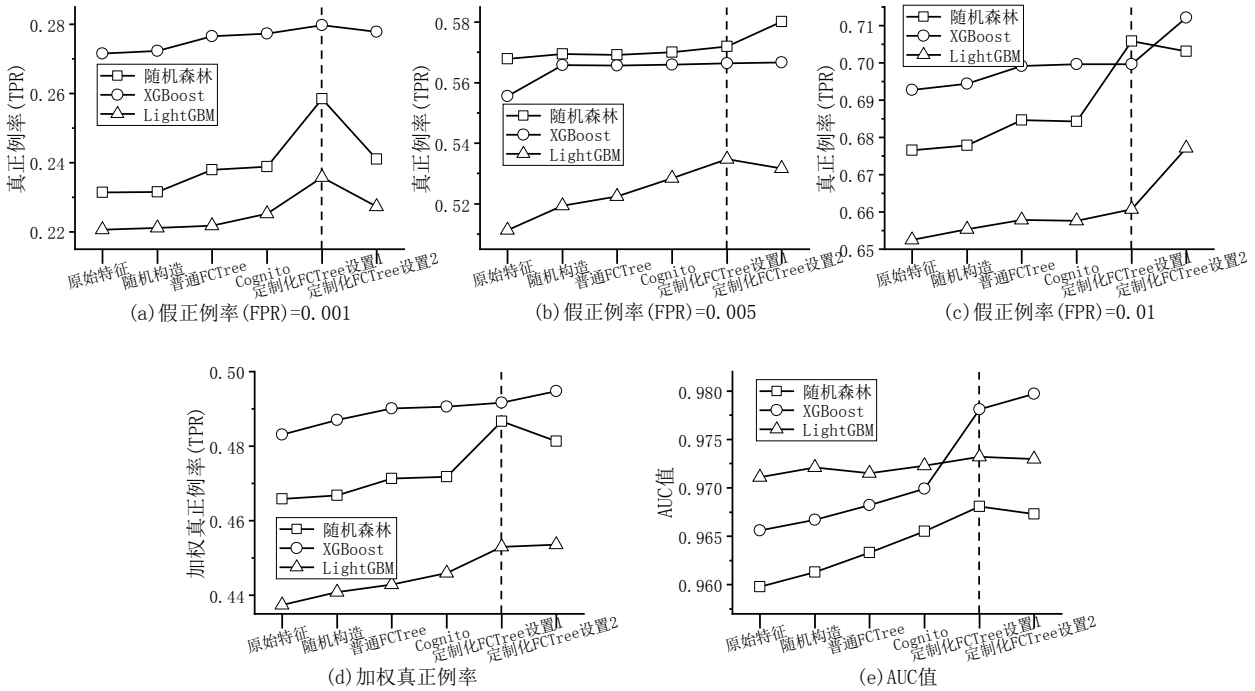
实验选择的机器学习模型为随机森林、XGBoost 和 LightGBM 分类器，其它的设置则保持一致。本文选取以下方法与本工作进行对比，分别是使用原始特征的方法、自底而上方法中的普通特征构造树方法、自顶而下方法中的 Cognito 方法和随机特征构造方法。其中，普通特征构造树方法节点处没有转换函数的组合，不能构造复杂特征，并且不具备转换函数权重向量时效性更新机制；Cognito 方法采用广度优先遍历，树的最大深度设为 6；随机特征构造方法每次随机选择一个转换函数构造特征训练模型，重复 20 次，最后选择对模型性能有提升的一些特征来训练最终的模型。

在两组实验设置下，图 8 和图 9 分别展示了重要性排名前 20 的特征（包括 285 个原始特征和新构造出来的特征）。结果显示实验设置一中特征重要性排名前 20 的有 8 个是由本工作的自动化特征工程方法所构造出来的新特征，其中重要性程度排名最高的一个新特征排在第 5 位；实验设置二中特征重要性排名前 20 的有 12 个是由本工作方法所构造出来的新特征，占据了半数以上，并且重要性程度排名第一位的是一个新特征。由此可见，在此金融科技平台的第三方移动在线网络支付数据集上，基于本工作定制化特征构造树的自动化特征工程方法所构造出来的新特征，在模型的特征评估方面显示出较高重要性，即对模型的影响比较明显，这表明了所构造特征具有价值，也从侧面反应了本工作所提出方法的高效性。同时，图中还展示了每一个新特征的构造过程，比如，通过新特征 multiplication(f7,f126)的形式，可以清晰得知其是由特征 7 和特征 126 相乘得到的，这也直接地体现了特征的可解释性，对于分析风险特征以及解释拦截异常支付交易的缘由具有重要作用。

图 10 中展示了在此数据集上应用不同方法得到的模型性能。由于 ROC 曲线不能明显地展示出各方法的性能差距，子图（a）-（e）展示了前述的其它评价指标： FPR 在 0.1%下的 TPR、FPR 在 0.5%下的 TPR、FPR 在 1%下的 TPR、加权 TPR 和 AUC 值。表 8 中给出了在此数据集上应用上述方法所得出的模型性能，表中粗斜体数值代表比其它对比方法优异。可见，本文方法下的两种设置在不同指标上几乎都优于其它方法。这证明了本文基于定制化特征构造树的自动化特征工程方法所自动构造的特征具有重要参考价值，对于模型欺诈检测性能的提升起到一定作用。

4.7.2 银行 B2C 在线网络支付交易数据集结果

在此数据集上，定制化的特征构造树使用到了全部三种类型的转换函数，具体如表 9 所示。



(a)假正例率(FPR)=0.001 (b)假正例率(FPR)=0.005 (c)假正例率(FPR)=0.01

(d)加权真正例率 (e)AUC值

图 10 第三方移动支付数据集上不同特征构造方法下的模型性能对比

表 8 第三方移动支付数据集上不同模型和特征构造方法下的各评价指标综合对比表

分类器	方法	加权 TPR	TPR|FPR=0.001	TPR|FPR=0.005	TPR|FPR=0.01	AUC
随机森林	原始特征	0.4659	0.2314	0.5678	0.6766	0.9598
	随机构造	0.4668	0.2316	0.5695	0.6778	0.9613
	普通 FCTree	0.4713	0.2379	0.5692	0.6847	0.9633
	Cognito	0.4718	0.2388	0.5701	0.6843	0.9655
	定制化 FCTree 设置 1	***0.4867***	***0.2584***	***0.5719***	***0.7058***	***0.9681***
	定制化 FCTree 设置 2	***0.4814***	***0.241***	***0.5801***	***0.7031***	***0.9673***
XGBoost	原始特征	0.4831	0.2716	0.5555	0.6927	0.9656
	随机构造	0.487	0.2723	0.5658	0.6944	0.9667
	普通 FCTree	0.49007	0.2765	0.5657	0.6992	0.9682
	Cognito	0.4906	0.2773	0.566	***0.6996***	0.9699
	定制化 FCTree 设置 1	***0.4917***	***0.2798***	***0.5664***	***0.6996***	***0.9781***
	定制化 FCTree 设置 2	***0.4948***	***0.2778***	***0.5667***	***0.7122***	***0.9797***
LightGBM	原始特征	0.4374	0.2206	0.5113	0.6525	0.9711
	随机构造	0.4409	0.2212	0.5193	0.6554	0.9721
	普通 FCTree	0.4428	0.2218	0.5225	0.6578	0.9715
	Cognito	0.446	0.2253	0.5285	0.6576	0.9723
	定制化 FCTree 设置 1	***0.4529***	***0.2358***	***0.5347***	***0.6607***	***0.9732***
	定制化 FCTree 设置 2	***0.4536***	***0.2273***	***0.5317***	***0.6771***	***0.973***

表 9 银行 B2C 在线网络支付交易数据集上应用的转换函数

类型	名称	函数	解释
纵向	平方	square()	对一列特征的值求平方
	开方	square_root()	对一列特征的值求开方
	sigmod	sigmoid()	对一列特征的值求 sigmod
	tanh	tanh()	对一列特征的值求 tanh
	加法	sum(,)	对俩列特征的值进行加
	减法	subtraction(,)	对俩列特征的值进行减
	乘法	multiplication(,)	对俩列特征的值进行乘
横向	群体累积	acc()	对一列特征的值进行累积求和
	群体计数	cum_count()	对一列特征的值进行累积计数
	频率	frequent()	对一列特征的值求频率
	个体累积	acc_group()	对一列特征的值，对各用户个体进行累积求和
	个体计数	cum_count _group()	对一列特征的值，对各用户个体进行累积计数
	时间差	diff_time()	对时间特征列的值，对相邻两样本间作时间差
	金额差	diff_amt()	对金额特征列的值，对相邻两样本间作金额差
时间窗口	窗口累积	acc_window()	对一列特征的值按用户个体在窗口内累积求和
	窗口均值	mean_window()	对一列特征的值按用户个体在时间窗内求均值
	窗口方差	var_window()	对一列特征的值按用户个体在窗口内求方差
	窗口计数	cum_count_window()	对一列特征的值按用户个体在窗口内进行计数

原始数据集除去交易卡号和交易时间共有 6 个初始的特征字段，包括交易金额、交易前账户余额、日限额、单笔限额、验签方式和是否是常用 ip。这里在每个节点处首先利用转换函数构造出新特征的个数 r，由于原始特征字段较少，我们将 r 设置固定设为 6。但是这里我们通过抽取原始数据集中负样本（正常交易数据样本）的部分数据，改变训练集中正负样本的比例分布，做了两组对比实验。其中实验一中负样本的抽取比例为 50%，正样本（欺诈交易数据样本）则保留了全部，从而，其正负样本比为 1:30，当前设置下整个特征构造树构造出来了 51 个新特征;而实验二中负样本和正样本都保留了全部，其正负样本比为初始的 1:59，整个特征构造树构造出来的总的新特征的个数为 40。实验选择的机器学习模型为 RandomForest、XGBoost 和 Lightgbm 分类器，同时实验的其它设置保持一致。同样这里利用其它实验组做了参照。

在特征重要性评估上，图 11 和图 12 分别展示了两组实验设置下的模型特征重要性程度排名前 10 的特征(包括原始特征和新构造出来的特征集合)。结果显示，实验设置一中特征重要性程度排名前 10 的特征中，除了是否是常用 ip 和验签方式两个原始特征，另外 8 个都是由本工作的自动化特征工程方法构造出来的新特征，并且新构造出来的特征重要性程度都比较高；实验设置二中特征重要性程度排名前 10 的特征中,除了是否是常用 ip、验签方式和交易金额这三个原始特征，其余 7 个是由本文自动化特征工程方法构造出来的新特征。由此可见，在此数据集上，本工作基于定制化特征构造树的自动化特征工程方法构造出来的新特征都是较为重要的特征，对模型的影响比较明显，这再次从侧面体现本工作提出的自动化特征工程方法的有效性。同样，在图 11 和图 12 中，每一个新特征的原始特征字段由来及其构造过程均可以被清晰表达。以新特征 mean_window(diff_time(time))为例，它的构造过程可以清晰地展现出来：首先计算每个用户的每笔交易与之前一笔交易的时间间隔，再对一个时间窗口内的交易时间间隔计算平均值得到这个新特征。这再次印证了本文显式的自动化特征工程方法具备特征可解释性。

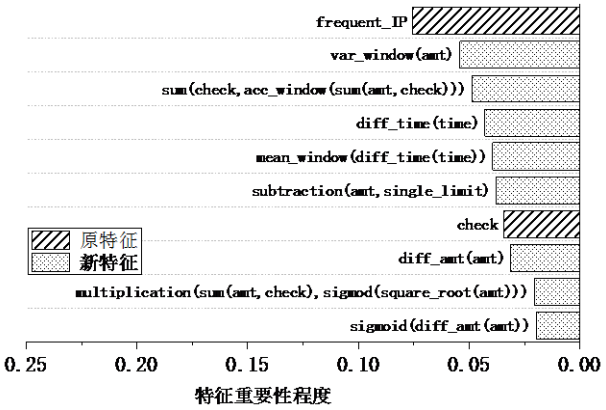


图 11 数据集二上实验设置一的重要性前 10 的特征

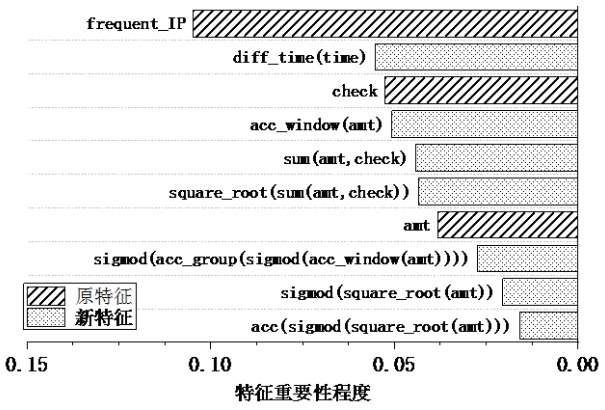
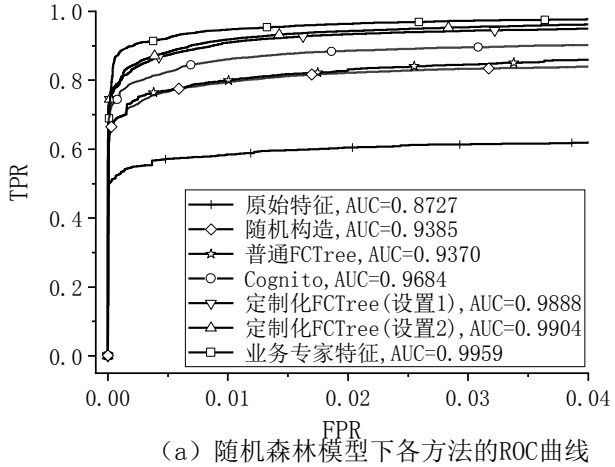
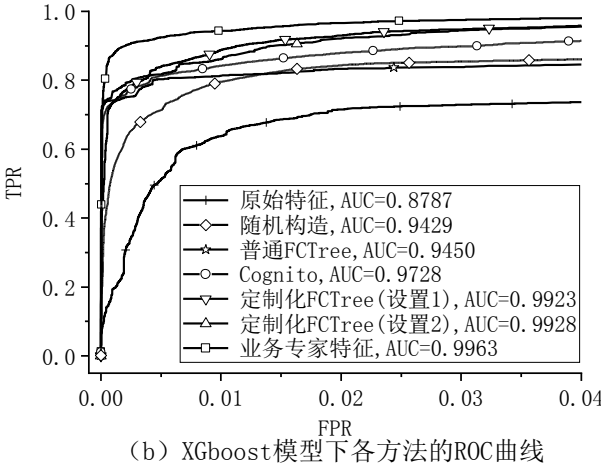


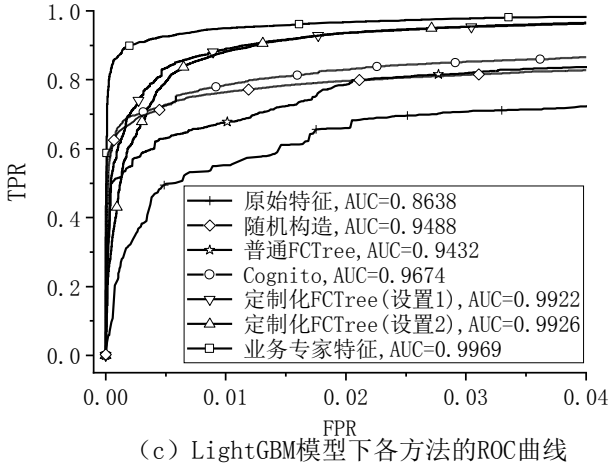
图 12 数据集二上实验设置二的重要性前 10 的特征



（a）随机森林模型下各方法的ROC曲线



（b）XGboost模型下各方法的ROC曲线



（c）LightGBM模型下各方法的ROC曲线

图 13 银行 B2C 网络支付数据集上不同方法构造的特征在各模型的 ROC 表现

图 13 展示了在此数据集上应用不同方法获得的各模型的 ROC 性能对比。由于此数据集原始字段的业务含义以及由业务专家对此根据业务逻辑人工生成的特征均为我们所知，所以在此数据集上我们还与业务专家特征进行了对比。由图中可见，本工作的自动化特征工程方法在两种不同设置下，在模型性能的比较上都优于其他方法，接近于业务专家特征下的模型性能，业务专家涉及到与用户相关的、与用户商户相关的、与商户相关的和与交易数据本身相关的四种类型的特征。在我们的方法中，暂时还未按此分类详细区分，从而，模型性能会受一定的影响。但是，相比业务专家构造的特征，自动化特征构造方法在时间上更具优势。特征构造方法在此数据集（约

243 万的样本和 6 个原始特征）上运行，在不同参数下的平均特征构造时间为 1.2 小时。相对于业务专家特征的数天甚至更长时间而言效率得到大幅提高。

4.7.3 算法特点分析

本文所述基于特征构造树的自动化特征工程算法具有轻量化、高效率、可扩展和特征可解释的特点。下面分别对这几个特点进行总结分析。

本文的方法是基于树模型的特征构造方法，与决策树的构造过程类似，对于决策树，其可以处理异常值、缺失值、支持不同量纲值，也可以处理连续和种类字段，数据的准备阶段往往是简单或者是不必要的，其可以节省大量的预处理步骤。更重要的是树模型的计算量较少，运算速度快，容易转化成分类规则，不需要占用极大的内存维护众多的参数，也不需要复杂的迭代过程，这些都体现其轻量化的特点。

本文方法在两个不同数据集以及不同参数设定下，均可在数小时的时间内构造出模型训练所需的新特征。如表 10 和 11 所示，对于第三方移动在线网络支付数据集，特征构造在约 73 万的样本数和 285 个原始特征的条件下进行，在不同参数下的平均特征构造时间为 4.5 小时；对于银行 B2C 在线网络支付数据集，特征构造在约 243 万的样本数和 6 个原始特征的条件下进行，在不同参数下的平均特征构造时间为 1.2 小时。相对于人工构造特征的数天甚至更长时间来说效率得到大幅提高。在同样是显性的自底而上的自动化特征工程方法中，LFE、RL 方法都需要预训练的步骤，其需要大量的数据集准备工作，工作周期远大于本文方法。图 14 和 15 显示了本工作方法与自底而上的普通特征构造树方法、自顶而下的 Cognito 方法和随机特征构造方法在特征构造时间上的对比，相对于自顶而下的 Cognito 方法，本文方法与普通特征构造树和随机特征构造方法耗时更少，而在后三者中本文方法构造的特征质量更高，模型性能更优。

本文方法在银行 B2C 在线网络支付和第三方移动在线网络支付消费场景上进行了实验分析。由于在转换函数设计环节可以通过扩展转换函数进行跨场景复用，因此，当在网络支付模式下的其它场景上进行特征工程时，仅需针对性的调整已有转换函数、增加一些根据当前场景定制的转换函数即可，对于算法的整体结构和流程是没有其它影响的，这能很好地体现本文方法的可扩展性。

如上实验中所述，每一个新特征的原始特征字段由来和构造过程都可以清晰地展现出来，本工作中显式的自动化特征工程方法具备特征的可解释性，对于后续的分析、模型的改进和风控策略的制定具有重要的意义。

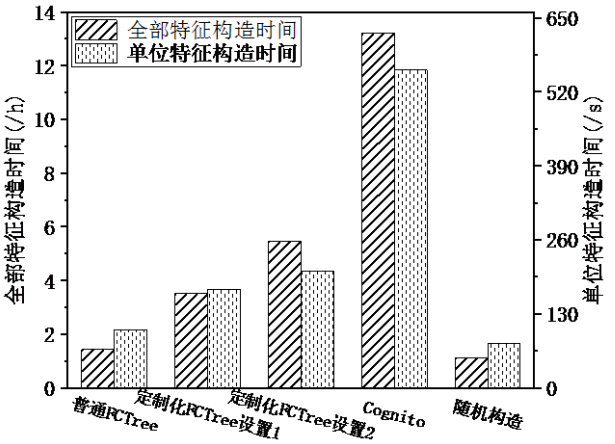


图 14 第三方移动支付数据集特征构造耗时

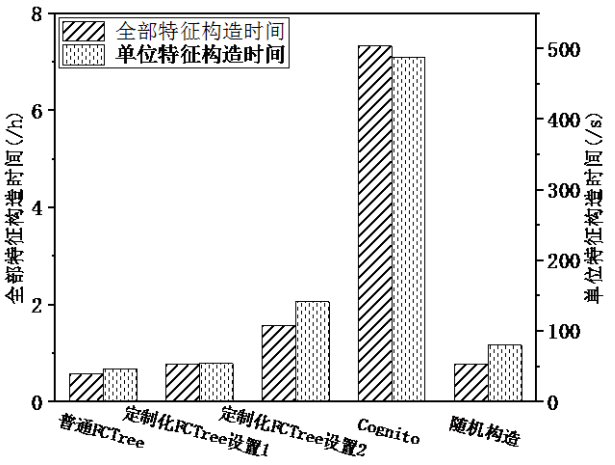


图 15 银行 B2C 线上支付数据集特征构造耗时

表 10 第三方移动在线网络支付数据集特征构造数与耗时

设置	样本量	原始特征数	构造特征数	耗时
每个节点处构造 90 个新特征	733683	285	73	3.51h
每个节点处构造 190 个新特征			96	5.46h

表 11 银行 B2C 在线网络支付数据集特征构造数与耗时

设置	样本量	原始特征数	构造特征数	耗时
50%正常样本	2434210	6	51	0.76h
原样本			40	1.57h

5 结论与展望

针对网络支付反欺诈问题，本文提出了一种定制化的特征构造树的自动化特征工程方法。该方法通过树结构的方法，在数据集进行划分的同时实现
特征的构造；通过定制化的转换函数设计，能够构造出面向网络支付的特征；其在节点处进行局部特征构造时保留构造的新特征，以此作为下个节点的基础特征来构造新特征，从而实现复杂特征的构造；其具有转换函数权重向量的时效性更新机制，避免转换函数权重陷入局部极值，保证特征构造的广泛性。本文利用该方法自动生成欺诈检测模型所需特征，在两个在线网络支付交易数据集上验证了其有效性，并且可以将特征工程步骤的时间花费从人工的数天级别降低到小时级别。本文所提出的方法对于减少特征工程复杂度，降低人工成本，提升模型开发整体工作效率具有参考价值。

该方法存在诸多可扩展和改进之处：（1）方法中转换函数的设计是可以进行扩展而不影响算法的整体结构的，后续可以增加更多的转换函数类型。（2）尝试在更多不同网络支付场景中的具有更大时间跨度的数据集上验证所提方法的有效性和通用性。（3）考虑将此在线网络支付模式下的自动化特征工程方法推广到互联网金融的领域下的其他模式。

参 考 文 献

[1] West J, Bhattacharya M. Intelligent financial fraud detection: a comprehensive review. Computers & Security, 2016, 57: 47-66.

[2] Bhattacharyya S, Jha S, Tharakunnel K K, Westland J C. Data mining for credit card fraud: a comparative study. Decision Support Systems, 2011, 50(3): 602-613.

[3] Chen R C, Chen T S, Lin C C. A new binary support vector system for increasing detection rate of credit card fraud. International Journal of Pattern Recognition, 2006, 20(2): 227–239.

[4] Simon v d Z, Wouter D, Werner v I, Jan V, Mykola P. ICIE 1.0: A Novel Tool for Interactive Contextual Interaction Explanations//Proceedings of the 3rd Mining Data for Financial Applications. Dublin, Ireland, 2018, 11054: 81-94.

[5] Chen Chao-fan, Lin Kang-cheng, Cynthia R, Yaron S, Wang Si-jia, Wang Tong. An Interpretable Model with Globally Consistent Explanations for Credit Risk//Proceedings of 32nd NeurIPS Workshop on Challenges and Opportunities for AI in Financial Services. Montreal, Canada, 2018: 1-10.

[6] Kyle B, Derek D, Ryan K, Brad R. HELOC Applicant Risk Performance Evaluation by Topological Hierarchical Decomposition. CoRR, 2018, abs/1811.10658.

[7] Dennis C, Leo M. V, Jarke J. van W. Instance-Level Explanations for Fraud Detection: A Case Study//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 28-33.

[8] Sun Da-Li. Analysis of association rules and its application in credit card anti-fraud. Chinese Journal of Credit Card, 2007, (22): 38-39 (in Chinese)
(孙大利. 关联规则分析及其在信用卡反欺诈中的应用. 中国信用卡, 2007, (22): 38-39)

[9] Ding Meng-Meng. Internet finance anti-fraud research based on rule engine. Computer Knowledge and Technology, 2018, 14(01): 7-9 (in Chinese)
(丁濛濛. 基于规则引擎的互联网金融反欺诈研究. 电脑知识与技术, 2018, 14(01): 7-9)

[10] Lu Tong-Bo. Application of big data anti-fraud in bank internet financial innovation business. Financial Electronic, 2016, (1): 91-92 (in Chinese)
(鲁统波. 大数据反欺诈在银行互联网金融创新业务中的应用. 金融电子化, 2016, (1): 91-92)

[11] Wang Wei-Wei. The application of artificial intelligence in the field of financial anti-fraud. Chinese Science and Technology Information, 2018, 592(20): 75-77 (in Chinese)
(王威巍. 人工智能在金融反欺诈领域的应用. 中国科技信息, 2018, 592(20): 75-77)

[12] Brause R W, Langsdorf T S, Hepp H M. Neural data mining for credit card fraud detection//Proceedings of the 11th International Conference on Tools with Artificial Intelligence. Chicago, USA, 1999: 103-106.

[13] Chan P K, Fan W, Prodromidis A L. Distributed data mining in credit card fraud detection. Intelligent Systems and their Applications, 1999, 14(6): 67-74.

[14] Zhang Qian, Yan Zhi-Wei, Li Hong-Tao. Research on phishing fraud detection technology. Journal of Network and Information Security, 2017, 03(07): 7-24 (in Chinese)
(张茜, 延志伟, 李洪涛. 网络钓鱼欺诈检测技术研究. 网络与信息安全学报, 2017, 03(07): 7-24)

[15] Wu Wei-Qiang, Hou Qi-Lin. Consumer finance anti-fraud model and method based on machine learning model. Modern Management Science, 2018, (10):51-54 (in Chinese)

(仵伟强, 后其林. 基于机器学习模型的消费金融反欺诈模型与方法. 现代管理科学, 2018, (10):51-54)

[16] Li Yun-Ni. Application of neural network model in bank internet finance anti-fraud. Financial Technology Era, 2018, 276(08): 24-28 (in Chinese)

(李赟妮. 神经网络模型在银行互联网金融反欺诈中的应用探索. 金融科技时代, 2018, 276(08): 24-28)

[17] Hilas C S, Mastorocostas P A. An application of supervised and unsupervised learning approaches to telecommunications fraud detection. Knowledge Based Systems, 2008, 21(7): 721-726.

[18] Bolton R J, Hand D J. Unsupervised profiling methods for fraud detection. London: Department of Mathematics, Imperial College, Technical Report, 2002.

[19] Tasoulis D K, Adams N M, Hand D J. Unsupervised clustering in streaming data//Proceedings of the 6th International Conference on Data Mining. Hong Kong, China, 2006: 638-642.

[20] Quah J T S, Sriganesh M. Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications. 2008, 35(4): 1721-1732.

[21] Weston D J, Hand D J, Adams N M, Whitrow C, Juszczak P. Plastic card fraud detection using peer group analysis. Advances in Data Analysis and Classification, 2008, 2(1): 45-62.

[22] Zhang Yun-Yun, Fang Yong, Huang Cheng. Credit card fraud detection based on Neo4j. Information and Computer (Theoretical Edition) , 2018, 415(21): 28-30 (in Chinese)

(张芸芸, 方勇, 黄诚. 基于 Neo4j 图谱的信用卡欺诈检测. 信息与电脑(理论版), 2018, 415(21): 28-30)

[23] Xu Yong-Hua. Credit card fraud detection based on support vector machine. Journal of Computer Simulation, 2011, 28(8): 376-379 (in Chinese)

(徐永华. 基于支持向量机的信用卡欺诈检测. 计算机仿真, 2011, 28(8): 376-379)

[24] Dhankhad S, Mohammed E A, Far B. Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study//Proceedings of the International Conference on Information Reuse and Integration IEEE Computer Society. Salt Lake City, USA, 2018: 122-125.

[25] S á A G C D, Pereira A C M, Pappa G L. A customized classification algorithm for credit card fraud detection. Engineering Applications of Artificial Intelligence, 2018, 72: 21-29.

[26] Santiago G P, Pereira A C M, Hirata Jr R. A modeling approach for credit card fraud detection in electronic payment services//Proceedings of the 30th Annual ACM Symposium on Applied Computing. Salamanca, Spain, 2015: 2328–2331.

[27] Bahnsen A C, Aouada D, Stojanovic A, Ottersten B E. Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 2016, 51: 134–142.

[28] Y. Liu, Y. Lai, Z. Wang and H. Yan. A New Learning Approach to Malware Classification Using Discriminative Feature Extraction. IEEE Access, 2019, 7: 13015-13023.

[29] H. Yang, S. Li, X. Wu, H. Lu and W. Han. A Novel Solutions for Malicious Code Detection and Family Clustering Based on Machine Learning. IEEE Access, 2019, 7: 148853-148860.

[30] Kanter J M, Veeramachaneni K. Deep feature synthesis: towards automating data science endeavors//Proceedings of the International Conference on Data Science and Advanced Analytics. Paris, France, 2015: 1-10.

[31] Lam H T, Thiebaut J M, Sinn M. One button machine for automating feature engineering in relational databases. CoRR, 2017, abs/1706.00327.

[32] Katz G, Shin E C R, Song D. Explorekit: automatic feature generation and selection//Proceedings of the 16th International Conference on Data Mining. Barcelona, Spain, 2016: 979-984.

[33] Khurana U, Turaga D, Samulowitz H. Cognito: automated feature engineering for supervised learning//Proceedings of the International Conference on Data Mining Workshops. Barcelona, Spain, 2016: 1304-1307.

[34] Nargesian F, Samulowitz H, Khurana U, Khalil E B, Turaga D S. Learning feature engineering for classification//Proceedings of the Twenty-sixth International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 2529-2535.

[35] Khurana U, Samulowitz H, Turaga D. Feature engineering for predictive modeling using reinforcement learning//Proceedings of the Thirty-Second Conference on Artificial Intelligence. New Orleans, USA, 2018: 3407-3414.

[36] Fan W, Zhong E, Peng J. Generalized and heuristic-free feature construction for improved accuracy//Proceedings of the International Conference on Data Mining. Columbus, USA, 2010: 629-640.

Wang Cheng, born in 1980, Ph.D, professor, Ph.D supervisor. His research interests include network service optimization and security, social computing and networks, anomaly detection in Cyberspace, and anti-fraud engineering in Internet Finance.

Wang Chang-Qi, born in 1994, M.S. His research interests include data mining, feature engineering, risk control, and fraud detection, etc.

Background

With the development of internet information technology, online payment services have continued to develop. As an important scenario of Internet finance, the online payment anti-fraud issue has always been a hot topic of both research and application. From the rule system to the machine learning based fraud detection system, anti-fraud methods have been evolving constantly. In the process of using machine learning to detect fraud detection, at present, industry experts need to perform feature engineering manually. Manually constructing features is time-consuming and labor-intensive. In the field of Internet finance, an anti-fraud model usually requires a large number of complex input features. This implies that the method of manually constructing features is impractical. This is not conducive to the evolution of the model over time. Feature engineering plays an important role in the field of security and internet finance. It can determine the level of performance that the model can achieve to a certain extent. Therefore, for the design of the fraud detection system, in addition to the design part of the model algorithm, the potential improvement space of the feature engineering step is considered. And the automated feature engineering method is used to improve the efficiency of this step in order to optimize the development process of the entire model system. Automated feature engineering methods can be divided into implicit methods and explicit methods. Among them, implicit methods lack the interpretability of features, and explicit methods are difficult to balance in time and space. To the best of our knowledge, there is still a lack of automated feature engineering study on the issue of anti-fraud in online payment services. Therefore, this paper introduces the automatic feature engineering method to the anti-fraud issue in online payment scenarios for the first time. More specifically, we design a customized feature construction tree, an automatic feature engineering method, to automate the feature generation process and improve the efficiency of model development. At the same time, the features automatically generated by our proposed method are verified on the real online payment transaction datasets. We have been doing related anti-fraud work on internet finance before. The previous work was to manually construct risk features in the feature engineering stage. The results of this project have realized automated feature engineering. It played a key role in our projects and saved a lot of time.