研究与开发

基于关联规则特征提取的网络行为被害性识别集成优化模型

周胜利[1]，阮琳琦[1]，徐睿[2]，张熙康[2]，赵泉喆[2]，连远博[3]
（1. 浙江警察学院，浙江 杭州 310053;
2. 杭州电子科技大学网络空间安全学院，浙江 杭州 310018;
3. 台州市税务局，浙江 台州 318001）

摘　要：网络行为被害风险识别对电信网络诈骗反制预警具有重要意义。针对被害人网络行为特征规则挖掘不足、行为序列间关系难以确定等问题，提出一种基于关联规则特征提取的网络行为被害性识别集成优化模型。模型首先抓取用户访问网站时产生的交互式流量数据包，提取网络流量中的隐性和显性行为特征，再利用频繁模式增长算法挖掘特征间关联规则并重构特征序列，最后结合粒子群优化的随机森林算法，建立基于网络流量分析的电信网络诈骗被害性分析模型。实验表明，相比于普通二分类模型，所提模型具有更好的精确率和召回率，能够有效提升被害性的识别准确率。

关键词：网络流量分析；关联规则；网络行为；电信网络诈骗
中图分类号：TP311
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2023180

An integrated optimization model of network behavior victimization identification based on association rule feature extraction

ZHOU Shengli[1], RUAN Linqi[1], XU Rui[2], ZHANG Xikang[2], ZHAO Quanzhe[2], LIAN Yuanbo[3]
1. Zhejiang Police College, Hangzhou 310053, China
2. School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou 310018, China
3. Taizhou Taxation Bureau, Taizhou 318001, China

Abstract: The identification of the risk of network behavior victimization was of great significance for the prevention and warning of telecom network fraud. Insufficient mining of network behavior features and difficulty in determining relationships, an integrated optimization model for network behavior victimization identification based on association rule feature extraction was proposed. The interactive traffic data packets generated when users accessed websites were captured by the model, and the implicit and explicit behavior features in network traffic were extracted. Then, the association rules between features were mined, and the feature sequences were reconstructed using the FP-Growth algorithm. Finally, an analysis model of telecom network fraud victimization based on network traffic analysis was established, combined with the stochastic forest algorithm of particle swarm optimization. The experiments show that

收稿日期：2023−05−20；**修回日期：**2023−09−06
基金项目：国家社会科学基金资助项目（No.23BGL272）；浙江省自然科学基金公益研究计划项目（No.LGF20G030001）；公安部科技计划项目（No.2022LL16）
Foundation Items: The National Social Science Foundation of China (No.23BGL272), Zhejiang Natural Science Foundation and Public Welfare Research Program (No.LGF20G030001), Ministry of Public Security Science and Technology Plan Project (No.2022LL16)

compared with general binary classification models, the proposed model has better precision and recall rates and can effectively improve the accuracy of network fraud victimization identification.

Key words: network traffic analysis, association rule, network behavior, telecom network fraud

0 引言

随着互联网技术的快速发展，新型犯罪模式层出不穷，其中以电信网络诈骗犯罪为代表的网络犯罪占比正逐年增加。公安部数据显示，2021 年全国法院一审审结电信网络诈骗犯罪案件 2.5 万余件，其中涉及的电信网络诈骗犯罪类型超过 50 余种。除了关注犯罪主体、犯罪行为和侦查办案过程外，被害人群体同样属于电信网络诈骗犯罪反制预警的重点考量范畴。因此，被害人行为特征规律挖掘正逐步成为电信网络诈骗犯罪治理的新研究方向。但目前针对被害人网络行为特征的研究存在特征选取表面化、特征间内在规则挖掘不足、行为序列间关系难以确定等问题。针对上述不足，本文在充分研究被害人网络行为特征的基础上，提出了一种基于关联规则特征提取的网络行为被害性识别集成优化模型（an integrated optimization model of network behavior victimization identification based on association rule feature extraction，VIC-RF&A），通过分析用户行为显隐性特征，挖掘潜在关联规则，识别用户被害风险，进而助力犯罪打击防控，守护人民群众的财产安全。本文主要贡献如下。

（1）采用频繁模式增长（frequent pattern growth，FP-Growth）关联规则算法，以较优的时间效率解决网络行为稀疏、特征间关系不明确等问题。

（2）结合粒子群算法建立优化的随机森林模型，进行流量数据分类，实现网络行为分析与被害性判断。

1 国内外相关研究

面对电信网络诈骗犯罪的严峻形势，开展网络行为被害性研究对于电信网络诈骗反制预警具有重要理论与实践意义，为此，国内外学者从网络流量分析和网络行为特征等多个角度开展网络行为被害性识别研究。

在网络流量分析方面，早期研究主要采用支持向量机、神经网络、自回归滑动平均等模型[1]，实现对异常流量的检测，但普遍因网络流量的非线性、时变性及混沌性特征而效果不佳。针对单一模型的缺陷，吴蕾等[2]提出小波分解结合支持向量机的方法，实现多维度时间序列的流量预测。向昌盛等[3]通过小波分解并分析混沌特性的方法来重构网络流量向量，解决了单一模型难以应对流量变化的难题。席荣康等[4]建立基于自注意力机制和时空特征的流量分析模型，从网络流量角度对隐匿在网络中的犯罪行为进行预测，并有效解决了此前研究中存在的尺寸单一、数据信息丢失等问题。总体而言，目前通过网络流量对电信网络诈骗进行分析的研究，在实验数据集上能够取得较好的实验效果，但仅从网络流量角度对网络犯罪进行分析过于片面，在真实环境中的效果还有待验证提升，因此还需进一步从其他角度进行更深入的分析。

在用户网络行为特征挖掘方面，相关学者[5-7]主要通过分析用户网络行为特征来构建预测模型，以进一步预测用户后续网络行为。在网络行为检测方面，早期研究主要使用 k 均值、支持向量机、隐含狄利克雷分布模型等[8-11]算法，对用户产生的异常网络行为进行检测。但受限于使用的算法，对网络行为的检测效果还有待提升，且相关方法存在着特征选取单一、行为序列间关系难以确定等问题。针对特征选取单一问题，胡增富等[12]通过聚类算法处理初始数据，从中提取多个特征并建立用户行为检测模型。但模型的准确性严重受制于预设参数，可靠性不足。王晓东等[13]

提出一种基于图神经网络和循环神经网络的深度学习模型来检测用户行为，有效解决了常规单一模型难以处理网络行为时间信息、检测效率低下等问题，实现对用户异常行为的准确检测。徐啸炀等[14]对用户与网站交互产生的网络流量开展研究，提出基于网络行为流量分析的电信网络诈骗犯罪被害性识别模型，能有效分类被害人网络行为和识别被害性，但是模型在时间开销和运行效率上存在不足，对大规模网络流量的检测能力有待提升，难以应对海量数据的诈骗被害性识别分析需求。

综上所述，目前针对网络行为被害性识别的研究，主要集中在网络攻击所产生的异常流量识别问题上，在网络行为被害特征提取方面，以显性的特征为主，缺乏对深层隐性特征的挖掘，并存在一定程度的特征冗余；在网络行为分析方面，缺少对特征间关联规则的分析，存在行为稀疏等问题。针对上述问题，本文从被害人视角出发，在充分研究网络流量特征与网络行为特征的基础上，提出了 VIC-RF&A 模型。

2　问题定义

定义 1　隐性行为特征序列。从网络流量中提取的特征，定义为 $H=\{T,N,L\}$。其中，$T=\{t_i \mid i=1,\cdots,k\}$ 为时间类特征集合，$N=\{n_i \mid i=1,\cdots,k\}$ 为数据流特征集合，$L=\{l_i \mid i=1,\cdots,k\}$ 为交互情况特征集合。

定义 2　显性行为特征序列。通过抓取超文本传输协议（hypertext transfer protocol，HTTP）数据包，提取各字段内容后，基于行为分类字典筛选获得的，能够准确反映用户在有关网站上所进行的若干可能导致诈骗被害的行为。本文显性行为特征序列定义为 $D=\{\text{act}_i \mid i=1,\cdots,k\}$，$\text{act}_i$ 为用户的行为，可以是单击、评论、注册、登录、交易等行为。

定义 3　行为向量化。通过提取网络流量中的显性、隐性特征，实现用户网络行为向量化编码。其中显性行为特征集合 D 通过预设行为字典方式，进行字典匹配并二值化编码。合并显性行为特征集合 D 与隐性行为特征集合 H 成为行为特征序列 F。

定义 4　特征间关联规则。当行为特征之间出现某种固定的指向性关系时，该行为特征序列的被害性会显著增大。此时的固定指向性关系即特征间关联规则。

定义 5　特征重构序列。采用 FP-Growth 算法挖掘已标记的被害行为序列获得关联性表，并根据关联性表中的规则，结合原有的行为特征序列，重新完成行为向量化。定义重构后的序列为特征重构序列 Re_F。

定义 6　根据特征重构序列 Re_F，建立分类模型识别并判断目标用户所进行的网络行为是否存在诈骗被害性。

3　VIC-RF&A 模型设计

根据以上定义，可将 VIC-RF&A 模型分为三大模块：网络行为特征挖掘、关联规则挖掘以及用户行为诈骗被害性识别。VIC-RF&A 模型框架如图 1 所示。模型首先抓取用户访问网站时产生的网络流量数据包，并根据定义 1 与定义 2，从网络流量中提取显隐性行为特征，然后根据定义 3 形成行为特征序列。之后，根据定义 4，利用 FP-Growth 算法进一步挖掘行为特征序列内部的关联规则，并根据定义 5 重构行为特征序列，提升所提取特征的表征能力。最后，根据定义 6 并结合粒子群算法优化模型训练参数，建立随机森林模型，实现对用户网络行为被害性的精准识别。

VIC-RF&A 模型的伪代码如下。其中，需要输入网络流量数据包集合 D；第（1）～（2）行代码即模型的步骤 1，用于生成行为特征序列 F；第（3）～（4）行代码即模型的步骤 2，用于挖掘关联规则并重构行为特征序列 F 并生成行为特征

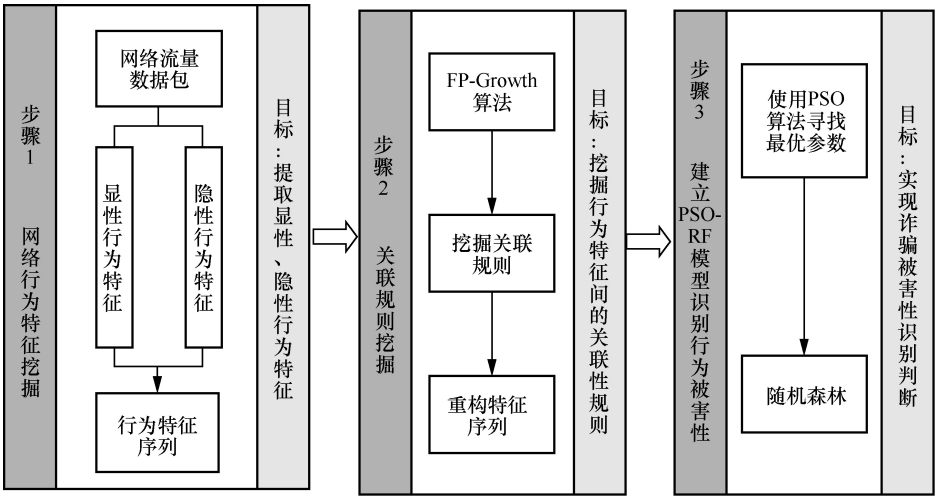


图 1　VIC-RF&A 模型框架

序列 Re_F；第（5）～（11）行代码即模型中步骤 3，其中第（5）～（8）行代码是模型的训练阶段，第（9）～（11）行是模型的测试阶段。

模型 1　VIC-RF&A 模型

输入　网络流量数据包集合 DATA

输出　网络行为被害性评估结果集合 RES

begin

（1）RES=[]//定义被害性评估结果集

（2）*F*(*H*,*D*)=Feature_Extraction(DATA)//挖掘显隐性特征

（3）Association_Rules=Get_Association_Rules(*F*)//生成关联规则列表

（4）Re_F=Get_ByRules(*F*, Association_Rules)//根据挖掘完毕的关联规则，重构并生成行为特征序列 Re_F

（5）Train_DATA=Get_Train_Data(Re_F)//划分模型训练数据

（6）Test_DATA=Get_Test_Data(Re_F)//划分模型测试数据

（7）max_features, n_estimators =PSO(Train_DATA)//粒子群算法寻找最优参数

（8）RF=TrainRandomForest(Train_DATA, max_features, n_estimators)//使用最优参数训练随机森林模型

（9）for *i* in Test_DATA://对每一条测试序列进行测试

（10）Answer=RF.predict(*i*)//实现被害性分析

（11）RES.append(Answer)//收集风险评估结果

end

3.1　网络行为特征序列提取

用户在浏览网站时会产生大量的网络流量数据。通过对流量数据进行分析，挖掘网络流量数据中的显性、隐性行为特征，有利于提高网络行为特征序列构建的全面性和合理性，进而提高模型的整体可靠性，具体如下。

（1）挖掘提取显性行为特征：主要分析数据包的字段内容，通过对比预设字典得到。显性行为特征主要用来反映网络用户在浏览网站中所采取的宏观操作。

（2）挖掘提取隐性行为特征：主要分析流量包的物理结构，从交互时间、数据包个数及长度（数据包长度即数据包字节大小）、通信数据流中数据包的报文段中负载数据长度等多个角度来提取，并进行离散化处理。

隐性行为特征参数见表 1，显性行为特征参数见表 2。行为特征序列 *F* 即同一次网站访问操作所产生的数据包中挖掘出的显性和隐性行为特征的线性组合。

表 1　隐性行为特征参数

特征参数	特征含义
SOURCE_IP	数据包的源 IP 地址
DESTINATION_IP	数据包的目标 IP 地址
SUM_BYTE	数据包总长度
SUM_PACKAGE	数据流总包数
DESTINATION_PORT	目标端口
INTERACTION_TIME	持续时长
SEND_PACKAGE	数据流发送总数据包数
RECEIVE_PACKAGE	数据流接收总数据包数
SEND_LENGTH	发送数据包长度
REVEIVE_LENGTH	接收数据包长度
SEND_TCP_LENGTH	发送 TCP 头字段长度
REVEIVE_TCP_LENGTH	接收 TCP 头字段长度
LOAD_LENGTH	数据流中数据包总负载数据长度
SEND_LOAD_LENGTH	数据流中发送数据包负载数据长度
REVEIVE_LOAD_LENGTH	数据流中接收数据包负载数据长度

表 2　显性行为特征参数

特征参数	特征含义
风险关键词	是否访问具有风险性关键词的网站
银行卡操作	是否存在操作银行卡行为
诈骗游戏行为	是否参与网站诈骗性游戏行为
转账支付行为	是否参与网站转账支付性行为
透露个人信息行为	是否存在浏览网站透露个人信息行为
其他未知行为	是否存在以上定义之外的行为

3.2　关联性规则挖掘

通过分析网络流量数据可知，用户访问正常网站与访问诈骗网站的行为存在某种相似性。仅将用户的网络行为孤立起来研究是片面的，会对模型的精确性和科学性产生影响。因此，采用关联规则算法挖掘用户显性、隐性行为特征之间的内在关系，能提高行为特征与用户被害性间的关联性，提升风险预测、被害性判断的可靠性。本文采用 FP-Grwoth 算法挖掘特征间关联性，并根据结果重构特征序列。

（1）挖掘频繁项，获取关联规则

初步观察数据可知，标签为诈骗被害的行为特征序列与标签为非诈骗被害的行为特征序列，在特征之间的关联性上存在一定的差异。同时，在单条行为特征序列中，孤立的某个行为特征（如某交互情况特征 l_i）的出现不能有效证明该条行为特征序列会导致高诈骗被害风险。但如果行为特征序列 F 中的某项行为特征与其他行为特征间的关联性，可以在大量标签为诈骗被害的关联规则中得到验证时，则在一定程度上说明了该行为特征序列具有高被害可能。因此，一旦找出频繁项构成的频繁集，就可以直截了当地获取其中的关联规则。首先利用式（1）求出置信度：

$$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)} \tag{1}$$

其中，$\text{support\_count}(A \cup B)$ 是包含项集 $A \cup B$ 的行为特征数，$\text{support\_count}(A)$ 是包含项集 A 的行为特征数，根据式（1），并通过以下步骤得到关联规则：

- 对于每个挖掘出的行为特征频繁集 L，产生 L 的所有非空子集；
- 对于 L 的每个非空子集 s，如果 $\frac{\text{support\_count}(L)}{\text{support\_count}(s)} \geqslant \min\_\text{conf}$，则输出规则 $S \Rightarrow (L - S)$。其中，$\min\_\text{conf}$ 是最小置信度阈值。

（2）重构特征序列

将行为特征序列 F 与挖掘得到的规则进行线性组合，重新构造行为特征序列 Re_F。原始的行为特征序列 F 注重表达用户网络行为的宏观操作特点以及网络流量的微观物理特点。重构后的特征序列 Re_F 进一步反映该特征序列的内在关联，表达了特征之间的组合与联系，对判断网络诈骗被害性更有利。

3.3　电信网络诈骗被害性识别

粒子群优化（particle swarm optimization，

PSO）算法是一种寻找最优解的算法。算法思路是通过初始化一些随机的粒子，定义其初始位置、速度等有关参数，更新粒子位置以寻找全局最优解。本模型拟将随机森林的有关参数作为随机粒子的位置坐标向量。通过迭代寻找粒子的最优位置坐标，来完成随机森林最优参数组成的寻找。每个粒子的位置和速度计算式如式（2）～式（3）所示：

$$S_{i,d}^{k+1} = S_{i,d}^{k} + v_{i,d}^{k+1} \tag{2}$$

$$\begin{aligned} v_{i,d}^{k+1} &= w^{k+1} v_{i,d}^{k} + \\ & c_1^{k+1} r_1 \left(P_{i,d} - s_{i,d}^{k} \right) + c_2^{k+1} r_2 (G_d - s_{i,d}^{k}) \end{aligned} \tag{3}$$

其中，$S_{i,d}^{k}$ 表示第 i 个粒子 k 次迭代的 d 维坐标分量，$v_{i,d}^{k}$ 第 i 个粒子 k 次迭代的 d 维速度分量，w^k 表示第 k 次迭代时的惯性权重，c_1^k 和 c_2^k 表示第 k 次迭代时的两个加速因子，r_1 和 r_2 表示[0,1]范围内的两个随机数，$P_{i,d}$ 表示第 i 个粒子 d 维分量的局部最优值，G_d 表示 d 维分量的整体最优值。

VIC-RF&A 模型使用关联规则挖掘模块提取的行为特征序列 Re_F 为训练数据，并使用 Boostrap 自助算法进行数据抽样，构建训练数据集和测试数据集，最后通过建立粒子群优化的随机森林模型来实现诈骗被害性判断。VIC-RF&A 模型中使用的粒子群优化算法优化的随机森林（particle swarm optimization of random forest，PSO-RF）结构如图 2 所示。相关待优化参数有最大特征数（max_features）、决策树子树数（n_estimators），因此通过粒子群优化算法待优化的参数维度 d 等于 2。由于随机森林的预测能力极大地依赖于模型的训练参数，而通过使用粒子群优化算法能够自动获得基于训练数据的最优随机森林训练参数，所以该方法不但能免去模型训练过程中设置不同训练参数的对比实验的环节，还能提高模型在不同训练数据下的泛化能力。

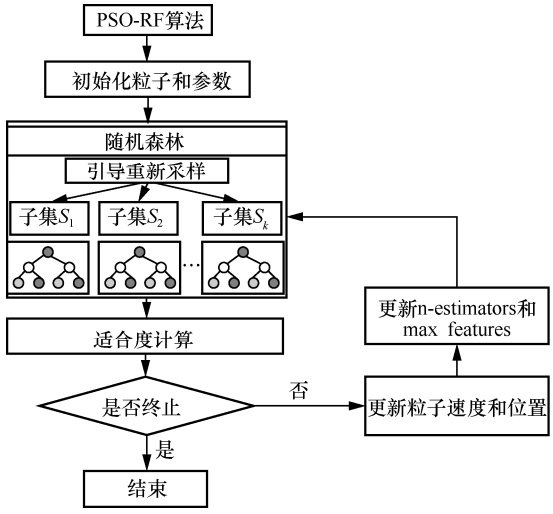


图 2　粒子群优化算法优化的随机森林结构

适应度函数是影响 PSO 算法的最主要因素。该函数决定了对随机粒子位置最优解的评价标准。VIC-RF&A 模型选择预测值与期望值之间误差矩阵的范数作为适应度函数，如式（4）所示：

$$F = k\left(\sum_{i=1}^{n} \text{abs}(y_i - a_i) \right) \tag{4}$$

其中，y_i 是预期值，a_i 是预测值，abs 是绝对值函数，k 为系数。

4　实验与分析

4.1　实验数据与环境

本实验通过自行抓包构建实验数据集，数据集生成流程为：通过使用 wireshark 软件，对模拟用户访问诈骗和非诈骗网站进行操作时产生的 HTTP 数据包进行动态采集（其中诈骗网站地址数据来源于公安数据库，非诈骗网站数据通过互联网搜集），并将捕获结果保存为 pcapng 格式文件，再使用 python 脚本提取流量包中的显性和隐性行为特征，形成总体行为特征序列。数据采集具体步骤如下。

步骤 1　以管理员权限打开 Wireshark，选择 WLAN 模式并开始抓包。

步骤 2　将目标网站（包括诈骗网站和非诈骗网站）地址复制到浏览器地址栏打开，并进行

注册、登录及访问等操作。

步骤 3　操作结束后终止抓包，并将抓包数据保存为 pcapng 格式文件。

步骤 4　重复步骤 1～步骤 3 完成对所有目标网站的抓包工作。

步骤 5　通过编写的 python 脚本，对保存的 pcapng 格式文件数据进行基于定义 1、定义 2 以及表 1、表 2 的处理，得到行为特征序列，并将结果存储至数据库中。

本实验采集的数据集共包含 3 194 条行为特征序列，其中标签为非网络诈骗被害（由非诈骗网站产生）的有 1 143 条，标签为网络诈骗被害（由诈骗网站产生）的有 2 051 条。实验数据情况见表 3。

表 3　实验数据情况

标签类型	访问网站类型	数量/个
网络诈骗被害类	非法博彩游戏	513
	虚假投资	709
	虚假论坛	343
	虚假借贷	486
非网络诈骗被害类	网页游戏	389
	金融投资	259
	生活交友	302
	体育彩票	193

4.2　实验评价指标

（1）ROC 曲线

受试者操作特征（receiver operator characteristic curve，ROC）曲线指在不同的实验判断标准下，以虚报概率(y/N)为横轴，以击中概率(y/SN)为纵轴的图线。ROC 曲线的对角线 $y=x$ 是基于随机猜测策略的分类器对样本进行分类的结果。当某个模型的曲线越接近对角线，表明该模型的判断分类能力越差；曲线越靠近左上角，表明该模型的分类效果越好。

ROC 曲线下面积（area under the ROC curve，AUC）是评价实验价值高低与模型分类效果的重要参考。AUC 值表示在数据集中分别随机选择一个正例和负例样本时，模型预测为正例的概率大于负例的概率的比例。AUC 取值范围为 0～1，AUC 值越接近 1，表示模型的性能越强。

（2）精确率、召回率和 F1 值

使用精确率 P（precision）、召回率 R（recall）和 F1 值综合评价模型识别被害性的效果。定义行为存在被害性为正类，定义行为存在非被害性为负类。TP 指将正类预测为正类的情况，FN 指将正类预测为负类的情况，FP 指将负类预测为正类的情况。精确率衡量了模型在预测结果为正例的样本中实际为正样本的比例，高精确率表示模型较少将负类样本预测为正类。精确率计算如式（5）所示。

$$P = \mathrm{TP} / (\mathrm{FP} + \mathrm{TP}) \tag{5}$$

召回率衡量了模型正确检测到的正例样本比例，高召回率表示模型能够识别出更多的正类样本。召回率计算如式（6）所示。

$$R = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \tag{6}$$

F1 值综合考虑了精确率和召回率的性能，高 F1 值表示模型的整体效果较好。F1 值计算式如式（7）所示。

$$\mathrm{F1} = 2 \times (P \times R) / (P + R) \tag{7}$$

4.3　关联规则挖掘及特征序列重构

通过 FP-Growth 算法挖掘行为特征间的规则。为提高关联规则有效性、避免输出结果过于庞大，设置支持度参数为 0.9，共得到 49 条关联规则，部分关联规则结果见表 4（其中诈骗游戏行为可从非法博彩游戏类诈骗网站中捕获）。

接下来，重构特征序列，步骤如下。

步骤 1　读取关联规则结果，并根据支持度从高到低排序，选取支持度较高的前 N 条作为重构判断依据 ass_rules。

步骤 2　从数据集 data 中取出一行数据 each_data。

步骤 3　从提取出的关联规则集 ass_rules

表 4 部分关联规则结果

规则	前项	后项	支持度	置信度
1	接收数据包负载数据长度	接收数据包长度	0.944	0.945
2	接收数据包长度	接收负载数据长度	0.908	0.998
3	接收数据包长度	接收 TCP 头字段长度	0.901	0.992
4	接收字节数	接收 TCP 头字段长度	0.901	0.990
6	诈骗游戏行为	透露个人信息的行为	0.953	0.996
7	转账支付行为	诈骗游戏行为	0.931	0.989
8	银行卡操作	转账支付行为	0.927	0.981
9	银行卡操作，转账支付行为	诈骗游戏行为，透露个人信息的行为	0.926	0.998
10	诈骗游戏行为	透露个人信息的行为，转账支付行为	0.930	0.971

中依次取出规则 rule，将 rule 的规则前后项 single_rule 与取出数据 each_data 中的对应项进行对比，当数据 each_data 中同时存在规则前后项 single_rule 时，判定这条规则成立，并将 flag=1 拼接到数据后；否则判定这条规则不成立，并将 flag=0 拼接到数据后。

步骤 4 重复步骤 2 与步骤 3，直到数据集 data 中每一项都完成判断。

本次实验中 N 设置为 2。在经过上述步骤后，可得重构后的特征序列 Re_F。本文将重构后的特征序列 Re_F 用于分类实验。部分重构序列结果见表 5。

4.4 实验结果分析

本次实验分为 3 个子实验，包括多个经典机器学习模型间的性能比较实验、长短期记忆（long short-term memory，LSTM）网络模型与 VIC-RF&A

表 5 部分重构序列结果

数据包总长度/byte	数据包总个数/个	平均负载数据长度/byte	…	是否为诈骗网站	Rule1	Rule2
446 065	715	529.867		true	1	1
167 457	120	1 301.47		true	1	1
166 41	37	355.757		true	1	1
195 40	27	629.704		true	1	1
9 414	14	578.429		true	1	1
7 491	14	441.071		true	1	1
6 777	12	470.75		true	1	1
1 747	4	342.75		true	1	1
1 824	3	514		true	1	1
593	2	202.5		true	1	1
897	2	354.5		true	1	1
106 970	180	500.278		true	1	1
2 318	4	485.5		true	1	1
69 960	115	514.348		true	0	0
2 451	2	1 131.5		true	0	0

模型对比实验、显性行为特征对 VIC-RF&A 模型识别能力测试实验。

实验 1：多个经典机器学习模型间的性能比较。实验 1 将数据集的 60%作为训练集，20%作为测试集 1，20%作为测试集 2 进行实验。对比朴素贝叶斯模型、KStar 模型以及序列最小优化（sequential minimal optimization，SMO）模型，从精确率、召回率、F1 值对 VIC-RF&A 模型进行评价。多模型在测试集 1 上的对比结果见表 6，多模型在测试集 2 上的对比结果见表 7，4 模型混淆矩阵见表 8。从表 6 与表 7 可以看出，相较于朴素贝叶斯模型、KStar 模型，本文模型在精确率、召回率和 F1 值的指标上均具有明显优势。从表 8 可以看出，对比四者的混淆矩阵情况，在正例（即诈骗网站）和反例（即正常网站）样本的预测上，VIC-RF&A 模型正确预测的占比更高，预测情况更加准确。

VIC-RF&A 模型、朴素贝叶斯模型、Kstar 模型以及 SMO 模型的 ROC 曲线对比如图 3 所示。对于测试集 1，由图 3（a）可知，VIC-RF&A 模型的 ROC 曲线较其他模型更加远离图像对角线，AUC 最大；对于测试集 2，由图 3（b）可知，VIC-RF&A 模型的 ROC 曲线近似且略高于 SMO 模型，AUC 最大。综合来看，相较于传统机器学习模型，VIC-RF&A 模型具有较好的可靠性和分类识别精确度，模型预测结果比较理想。

实验 2：LSTM 模型与 VIC-RF&A 模型对

表 6　多模型在测试集 1 上的对比结果

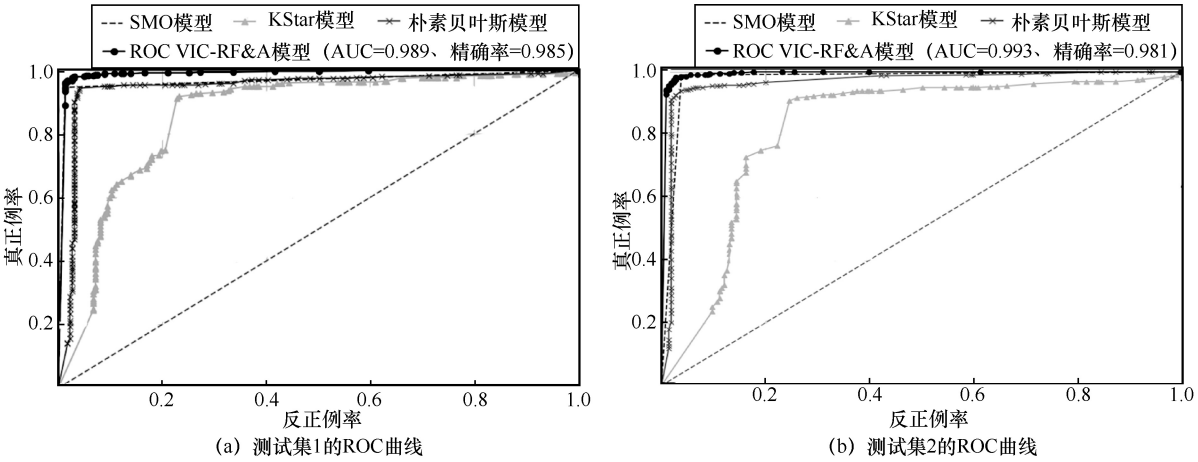
模型	VIC-RF&A	朴素贝叶斯	SMO	KStar
精确率	0.988	0.847	0.963	0.867
召回率	0.971	0.770	0.961	0.869
F1 值	0.979	0.774	0.961	0.868

表 7　多模型在测试集 2 上的对比结果

模型	VIC-RF&A	朴素贝叶斯	SMO	KStar
精确率	0.983	0.807	0.952	0.853
召回率	0.979	0.618	0.950	0.854
F1 值	0.981	0.612	0.950	0.853

表 8　4 模型混淆矩阵

混淆矩阵		测试集 1							
		VIC-RF&A		朴素贝叶斯		SMO		KStar	
		正例	反例	正例	反例	正例	反例	正例	反例
实际类别	正例	404	12	276	140	394	22	380	36
	反例	5	218	7	216	3	220	51	172
混淆矩阵		测试集 2							
		VIC-RF&A		朴素贝叶斯		SMO		KStar	
		正例	反例	正例	反例	正例	反例	正例	反例
实际类别	正例	415	9	184	240	401	23	384	40
	反例	7	208	4	211	9	206	53	162



（a）测试集1的ROC曲线　　（b）测试集2的ROC曲线

图 3　ROC 曲线对比

比实验。由前文分析可知，网络行为特征序列间存在一定的依赖关系，而 LSTM 模型擅长处理存在依赖关系的序列，故选择使用 LSTM 模型与 VIC-RF&A 模型进行对比实验 2。实验 2 将数据集的 60%作为训练集，40%作为测试集，进行对比实验。VIC-RF&A 与 LSTM 对比结果见表 9，VIC-RF&A 与 LSTM 对比实验混淆矩阵见表 10。

由表 9 结果可知，VIC-RF&A 模型在精确率、召回率、F1 值与 AUC 4 项指标上均具有一定优势。由表 10 结果可知，在正例和反例样本的预测上，VIC-RF&A 模型正确预测的占比更高，预测情况更加准确。综合来看，VIC-RF&A 模型与 LSTM 模型都具有较好的可靠性和分类识别精确度。但 LSTM 模型的预测效果受模型整体结构和参数的影响较大，而 VIC-RF&A 模型中使用粒子群优化算法来确定随机森林模型的训练参数，泛化能力更强。此外，LSTM 模型在训练时需要大量的参数，相较于 VIC-RF&A 模型的开销更大。

表 9　VIC-RF&A 与 LSTM 对比结果

模型	精确率	召回率	F1 值	AUC
VIC-RF&A	0.997	0.970	0.983	0.991
LSTM	0.994	0.929	0.961	0.960

表 10　VIC-RF&A 与 LSTM 对比实验混淆矩阵

混淆矩阵		VIC-RF&A		LSTM	
		正例	反例	正例	反例
实际类别	正例	783	24	750	57
	反例	2	469	4	467

实验 3：显性行为特征对 VIC-RF&A 模型识别能力测试。此外，针对表 2 中列举的部分显性网络行为（如银行卡操作、转账支付行为等）在用户使用加密流量情况下采集困难的问题，进行对比实验 3。将特征序列 F 中相关显性行为特征（即操作银行卡行为、参与网站转账支付性行为和浏览网站透露个人信息行为）剔除后得到特征序列 F_eli，并分别将特征序列 Re_F 与特征序列 F_eli 对应数据集重新进行划分，其中 60%作为训练集，40%作为测试集，测试特征序列中的显性行为特征对 VIC-RF&A 模型识别网络诈骗被害性能力的影响。特征序列 Re_F 与 F_eli 对比结果见表 11，特征序列 Re_F 与 F_eli 对比混淆矩阵见表 12。

表 11　特征序列 Re_F 与 F_eli 对比结果

特征序列	精确率	召回率	F1 值	AUC
特征序列 Re_F	0.982	0.977	0.979	0.991
特征序列 F_eli	0.907	0.908	0.907	0.949

表 12　特征序列 Re_F 与 F_eli 对比混淆矩阵

混淆矩阵		特征序列 Re_F		特征序列 F_eli	
		正例	反例	正例	反例
实际类别	正例	824	16	792	48
	反例	13	425	70	368

由表 11 结果可知，在剔除部分显性行为特征后，VIC-RF&A 模型的识别能力会有一定程度的下降。从表 12 可以看出，对于正例和反例样本的预测，特别是对反例的误报率有一定程度的上升。

综合以上分析可知，VIC-RF&A 模型能够有效挖掘特征间关联性规则，并结合粒子群算法优化的随机森林模型实现网络诈骗被害性识别分析，其综合效果优于通过单一分类器或单一隐性行为特征来判断网络行为被害性的模型，具有更高的精确率和可靠性。

5　结束语

本文针对被害人网络行为特征规则挖掘不足、行为序列间关系难确定等问题，提出了一种基于关联规则特征提取的网络诈骗被害性识别集成优化模型。模型利用关联规则挖掘显隐性网络行为的内在相关性，重构网络行为特征序列，再结合粒子群优化的随机森林模型识别网络行为被害性。实验证明，本模型能够可靠、高效地实现网络行为被害性分类识别。未来研究可从深层次网络行为特征挖掘和网络行为时序排列等方面改进完善模型，以进一步提升模型识别效果。

参考文献：

[1] 陈易平, 俞龙, 谌颃. 大数据环境下基于小波神经网络和 ARMA 模型的流量异常检测[J]. 重庆理工大学学报(自然科学版), 2019, 33(10): 149-154.
CHEN Y P, YU L, CHEN H. Traffic anomaly detection based on wavelet neural network and ARMA model in big data environment[J]. Journal of Chongqing Institute of Technology, 2019, 33(10): 149-154.

[2] 吴蕾, 曾慧平, 王海威. 网络非平稳流量多尺度时间序列预测数学建模[J]. 计算机仿真, 2021, 38(8): 356-359, 434.
WU L, ZENG H P, WANG H W. Mathematical modeling of multi-scale time series prediction for network non-stationary traffic[J]. Computer Simulation, 2021, 38(8): 356-359, 434.

[3] 向昌盛, 陈志刚. 面向海量数据的网络流量混沌预测模型[J]. 计算机科学, 2021, 48(5): 289-293.
XIANG C S, CHEN Z G. Chaotic prediction model of network traffic for massive data[J]. Computer Science, 2021, 48(5): 289-293.

[4] 席荣康, 蔡满春, 芦天亮, 等. 基于自注意力机制和时空特征的 Tor 网站流量分析模型[J]. 计算机应用, 2022, 42(10): 3084-3090.
XI R K, CAIM C, LU T L, et al. Tor website traffic analysis model based on self-attention mechanism and spatiotemporal features[J]. Journal of Computer Applications, 2022, 42(10): 3084-3090.

[5] 郭旭, 朱敬华. 基于用户向量化表示和注意力机制的深度神经网络推荐模型[J]. 计算机科学, 2019, 46(8): 111-115.
GUO X, ZHU J H. Deep neural network recommendation model based on user vectorization representation and attention mechanism[J]. Computer Science, 2019, 46(8): 111-115.

[6] 都奕冰, 孙静宇. 融合项目嵌入表征与注意力机制的推荐算法[J]. 计算机工程与设计, 2020, 41(3): 682-688.
DU Y B, SUN J Y. Recommendation algorithm based on item embedding characterization and attention mechanism[J]. Computer Engineering and Design, 2020, 41(3): 682-688.

[7] 向卓元, 刘志聪, 吴玉. 基于用户行为自适应推荐模型研究[J]. 数据分析与知识发现, 2021, 5(4): 103-114.
XIANG Z Y, LIU Z C, WU Y. Adaptive recommendation model based on user behaviors[J]. Data Analysis and Knowledge Discovery, 2021, 5(4): 103-114.

[8] 叶春明, 李志, 郑科栋, 等. 一种基于用户行为状态特征的流量识别方法[J]. 计算机应用研究, 2015, 32(2): 560-564, 578.
YE C M, LI Z, ZHENG K D, et al. Traffic classification based on analysis to users' activities status[J]. Application Research of Computers, 2015, 32(2): 560-564, 578.

[9] 钱叶魁, 陈鸣, 郝强, 等. ODC: 在线检测和分类全网络流量异常的方法[J]. 通信学报, 2011, 32(1): 111-120.
QIAN Y K, CHENM, HAO Q, et al. ODC: a method for online

detecting & classifying network-wide traffic anomalies[J]. Journal on Communications, 2011, 32(1): 111-120.

[10] ERFANI S M, RAJASEGARAR S, KARUNASEKERA S, et al. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning[J]. Pattern Recognition, 2016(58): 121-134.

[11] 郭志民, 彭豪辉, 牛霜霞, 等. 基于用户与网络行为分析的主机异常检测方法[J]. 北京交通大学学报, 2018, 42(5): 40-46.
GUO Z M, PENG H H, NIU S X, et al. Analyzing user and network behaviors for host-based anomaly detection[J]. Journal of Beijing Jiaotong University, 2018, 42(5): 40-46.

[12] 胡富增, 王勇军. 基于数据挖掘的计算机用户行为分析与识别[J]. 自动化技术与应用, 2020, 39(6): 42-47.
HU F Z, WANG Y J. Analysis and recognition of computer user behavior based on data mining[J]. Techniques of Automation and Applications, 2020, 39(6): 42-47.

[13] 王晓东, 赵一宁, 肖海力, 等. 使用GNN与RNN实现用户行为分析[J]. 计算机科学与探索, 2021, 15(5): 838-847.
WANG X D, ZHAO Y N, XIAO H L, et al. User behavior analysis with RNN and graph neural networks[J]. Journal of Frontiers of Computer Science & Technology, 2021, 15(5): 838-847.

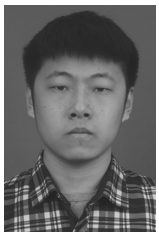
[14] 周胜利, 徐啸炀. 基于网络流量的用户网络行为被害性分析模型[J]. 电信科学, 2021, 37(2): 125-134.
ZHOU S L, XU X Y. Victimization analysis model of user network behavior based on network traffic[J]. Telecommunications Science, 2021, 37(2): 125-134.

[作者简介]

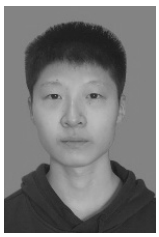
周胜利（1982− ），男，博士，浙江警察学院副教授，主要研究方向为网络安全、网络监察管理。

阮琳琦（1996− ），女，浙江警察学院讲师，主要研究方向为网络安全。

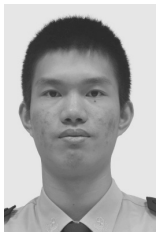
徐睿（2000− ），男，杭州电子科技大学网络空间安全学院硕士生，主要研究方向为网络安全、机器学习。



张熙康（2000− ），男，杭州电子科技大学网络空间安全学院硕士生，主要研究方向为网络安全、自然语言处理。



赵泉喆（2000− ），男，杭州电子科技大学网络空间安全学院硕士生，主要研究方向为网络安全、数字取证。



连远博（2001− ），男，台州市税务局工程师，主要研究方向为网络安全。