计算机科学
COMPUTER SCIENCE

**一种面向电商网络的异常用户检测方法**

杜航原，李铎，王文剑

引用本文

杜航原，李铎，王文剑. 一种面向电商网络的异常用户检测方法[J]. 计算机科学, 2022, 49(7): 170-178.

DU Hang-yuan, LI Duo, WANG Wen-jian. Method for Abnormal Users Detection Oriented to E-commerce Network [J]. Computer Science, 2022, 49(7): 170-178.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600092

# 一种面向电商网络的异常用户检测方法

杜航原[1]　李　铎[1]　王文剑[1,2]
1 山西大学计算机与信息技术学院　太原 030006
2 计算智能与中文信息处理教育部重点实验室(山西大学)　太原 030006
(duhangyuan@sxu.edu.cn)

**摘　要**　在电商网络中,异常用户往往表现出与正常用户截然不同的行为特征,检测异常用户并分析其行为模式对维护电商平台秩序具有十分重要的现实意义。通过分析异常用户的行为模式,将电商网络抽象为异质信息网络并转化为用户-设备二分图,然后在此基础上提出了一种面向电商网络的异常用户检测方法——自监督异常检测模型(Self-Supervised Anomaly Detection Model,S-SADM)。该方法具有自监督学习机制,采用自编码器编码获取用户节点表示,通过优化联合目标函数来完成反向传播,同时采用支持向量数据描述对用户节点表示进行异常检测。经过网络的自动迭代优化,不仅使用户节点表示具有监督信息,还获得了较稳定的检测结果。最后,在真实网络数据集和半合成网络数据集中对 S-SADM 进行实验,结果表明了该方法的有效性和优越性。

**关键词**:异常检测;电商网络;异质信息网络;自监督学习;自编码器;支持向量数据描述
**中图法分类号**　TP183

## Method for Abnormal Users Detection Oriented to E-commerce Network

DU Hang-yuan[1],LI Duo[1] and WANG Wen-jian[1,2]
1 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2 Key Laboratory of Computational Intelligence and Chinese Information Processing(Shanxi University),Ministry of Education,Taiyuan 030006,China

**Abstract**　In the e-commerce network,abnormal users often show different behavioral characteristics from normal users. Detecting abnormal users and analyzing their behavior patterns is of great practical significance to maintaining the order of e-commerce platforms. By analyzing the behavior patterns of abnormal users,we abstract the e-commerce network into the heterogeneous information network,and convert it into a user-device bipartite graph. On this basis,we propose a method for detecting abnormal users oriented to e-commerce network——self-supervised anomaly detection model(S-SADM). The model has a self-supervised learning mechanism. It uses an autoencoder to encode the user-device bipartite graph to obtain user node representations. By optimizing the joint objective function,the model completes backpropagation,and uses support vector data descriptions to perform anomaly detection on user node representations. After the automatic iterative optimization of the network,the user node representation has supervised information,and we obtain relatively stable detection results. Finally,S-SADM is validated on 3 real network datasets and a semi-synthetic network dataset,and the experimental results demonstrate the effectiveness and superiority of the method.

**Keywords**　Anomaly detection,E-commerce network,Heterogeneous information network,Self-supervised learning,Autoencoder,Support vector data description

## 1　引言

随着互联网的不断普及和发展,许多不良商家通过操纵大量用户在各大电商网络平台上进行虚假评论、恶意刷单等欺诈活动,来诱导顾客购买有缺陷的产品,严重损害了消费者的利益。为了消除这些异常用户所带来的负面影响,学术界提出了许多关于异常检测的方法,并且这些方法成功获得了应用。Hu 等[1]提出的基于密度的局部离群点检测算法通过引入信息熵来挖掘同质信息网络中的局部离群点。Ren 等[2]结合 KNN 离群点检测算法和多层次随机森林模型来检测

到稿日期:2021-06-10　返修日期:2021-10-25
基金项目:国家自然科学基金(61902227,62076154,U1805263);中央引导地方科技创新项目(YDZX20201400001224);山西省自然科学基金(201901D211192);山西省高校科技创新项目(2019L0039)
This work was supported by the National Natural Science Foundation of China(61902227,62076154,U1805263),Special Foundation from the Central Finance to Support the Development of Local University(YDZX20201400001224),Natural Science Foundation of Shanxi Province,China(201901D211192) and Science Foundation of the Higher Education Institutions of Shanxi Province,China(2019L0039).
通信作者:王文剑(wjwang@sxu.edu.cn)

同质信息网络中的异常行为。文献[3]提出了一种基于后缀树的异常检测算法，该算法根据时间序列中异常点的周期性来检测异常点。文献[4]采用自编码器和概率神经网络集成实现异常点检测目标。

目前的研究工作大多面向同质信息网络，许多方法依旧存在局限性，不适用于异质信息网络[5]，因此面向异质信息网络的异常检测应运而生。Liu 等[6]提出了一种面向异质网的基于张量表示的动态异常点检测方法，该方法构建张量索引树分类样本并聚类，根据聚簇是否变换进行动态异常点判断，被有效应用于异质信息网络中。文献[7]提出了一种面向异质信息网络的异常检测方法，该方法基于异常用户的两大行为模式，自适应地从异质网络中学习区分嵌入，并提出了一种关注机制来学习不同类型节点的重要性，根据聚合模式的不同来区分异常用户。文献[8]提出了一种面向异质信息网络社区分布的离群点检测方法，该方法基于所有对象类型的流行社区分布模式并联合非负矩阵分解来检测离群点。文献[9]提出了一种面向电商网络的欺诈检测模型 DeepFD，该模型采用自编码器提取异质图中欺诈账户的行为模式，并通过 DBSCAN 方法来检测欺诈块。Zheng 等[10]提出了一种联合嵌入方法，用于捕获异质图结构，并通过聚类方法检测社交网络中的欺诈账户群。Dong 等[11]提出了一种结合自编码器和神经随机森林的端到端可训练联合模型，用于检测电商网络中的虚假评论。

异质信息网络是现实世界结构的一种抽象，它着重表现多种实体和实体之间的关联关系。在电商平台中，每天都会产生成千上万的交易记录，这些交易记录实际上是由用户、设备、产品和商家等实体之间的各种关联关系组成，包括注册、登录、购买、评价等行为。通过分析电商网络中异常用户的行为模式，重点关注用户和设备两类实体之间的关联关系，即用户在不同设备上的登录活动，将电商网络抽象为异质信息网络并转化为用户-设备二分图。在此基础上，本文提出了一种面向电商网络的异常用户检测方法——自监督异常检测模型(Self-Supervised Anomaly Detection Model，S-SADM)。它采用自编码器编码获取用户节点表示，并完成正向传播。根据异常用户的两大行为模式，我们定义了用户节点之间的行为相似度以及用户节点表示之间的行为相似度，将重构误差、用户节点和用户节点表示的行为相似度差异和支持向量数据描述(Support Vector Data Description，SVDD)的目标函数作为联合目标函数进行优化，并完成反向传播，接下来采用 SVDD 对用户节点表示进行异常检测。网络的自动迭代优化不仅使得用户节点表示同时保留了较完整的二分图结构和用户行为特征，还自动为异常检测提供了监督信息，并获得了较稳定的检测结果。

本文第 2 节对电商网络异常用户的行为特性描述以及自编码器等相关工作进行了介绍；第 3 节对 S-SADM 的构建进行了详细的介绍；第 4 节对 S-SADM 的算法流程进行了总结；第 5 节对 S-SADM 进行了实验验证和对比分析，并对其时间复杂度进行了分析；最后总结全文并展望未来。

本文的主要贡献如下：

(1)针对电商平台中存在的异常用户检测问题，将电商网络抽象为异质信息网络并转化为用户-设备二分图，在此基础上提出了自监督异常检测模型 S-SADM。通过网络的自动迭代优化，所提方法能够较好地保留二分图结构和用户行为特征，并提升异常用户检测的检测性能。

(2)在 3 个真实网络数据集和 1 个半合成网络数据集上对 S-SADM 和几种现有基线方法进行了对比实验与分析，证明了本文方法的有效性和优越性。

## 2 相关工作

### 2.1 异常用户的行为模式

在电商平台中，许多不法商家注册了大量新用户并在某些特定的设备上登录，进而实施集体性欺诈活动，异常用户便在这种环境下应运而生。Liu 等[7]通过对电商平台用户数据进行分析，提出了异常用户的两大行为模式：设备聚集性和活动聚集性。

(1)设备聚集性。图 1 分别给出了不同类型的用户 在不同设备上的登录情况。图 1(a)给出了正常用户的登录情况，其中点的分布较为均匀，说明正常用户的行为是独立的。图 1(b)给出了异常用户的登录情况，其中点的分布较为密集且规律，说明某些特定设备连接了大量的异常用户。在电商网络中，异常用户控制了大量的计算设备以进行欺诈活动，但该行为代价十分高昂，出于成本控制，大多数异常用户通常会在几组特定的设备上进行登录。

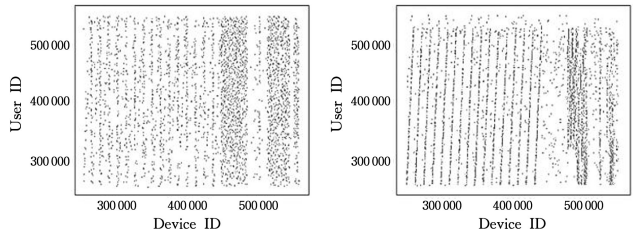


图 1 不同用户在不同设备上的登录情况

Fig. 1 Logins of different users on different devices

(2)活动聚集性。图 2 分别给出了新注册用户在不同时间段的登录情况。图 2(a)给出了正常用户的登录情况，其中点的分布较为均匀，说明正常用户的登录是随机的。图 2(b)给出了异常用户的登录情况，其中点在某个时间段的分布较为密集且规律，说明大批量异常用户通常在某个时间段内集中登录。在电商网络中，异常用户需要在短时间内完成任务，因此会在某个时间段内爆发集体性活动。

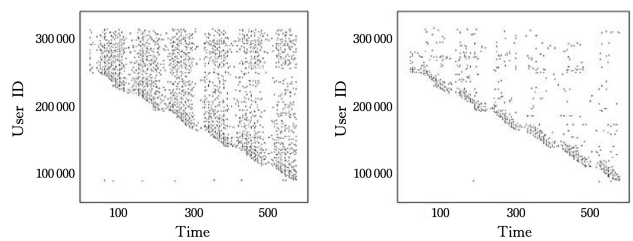


图 2 用户在不同时间段登录设备的情况

Fig. 2 Situation of users logging in to the device at different time periods

### 2.2 自编码器

自编码器的概念最早由 Rumelhart 等[12]提出，随后，文献[13]对其进行了详细的阐述。自编码器主要包括两部分，即编码器(Encoder，E)和解码器(Decoder，D)，且其结构对称，即编码器的隐层数量和解码器的隐层数量相同。自编码器的目的就是在输出层重建输入信号，使输出信号 $\boldsymbol{y}$ 和输入信号 $\boldsymbol{x}$ 尽可能相同。编码器的编码过程为：

$$\boldsymbol{z}=\sigma_e(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1) \tag{1}$$

解码器的解码过程可以描述为：

$$\boldsymbol{y}=\sigma_d(\boldsymbol{W}_2\boldsymbol{z}+\boldsymbol{b}_2) \tag{2}$$

其中，$\boldsymbol{W}_1$，$\boldsymbol{b}_1$ 为编码权重和偏置，$\boldsymbol{W}_2$，$\boldsymbol{b}_2$ 为解码权重和偏置；$\boldsymbol{z}$ 为隐空间中的特征表示；$\sigma_e$ 为激活函数，目前比较常用的有 ReLu，Sigmoid 和 Tanh 等，$\sigma_d$ 可以是与 $\sigma_e$ 相同的激活函数[14]。因此，自编码器的优化目标是最小化 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 之间的误差，即最小化重构损失：

$$J(\boldsymbol{W},\boldsymbol{b})=\sum(L(\boldsymbol{x},\boldsymbol{y}))=\sum\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \tag{3}$$

编码过程是通过一种确定性的映射将输入信号转换为隐空间中的特征表示，解码过程则是尽量将隐空间中的特征表示重新映射为输入信号[14]。自编码器中的参数，即权重和偏置，通过最小化目标函数来学习获得。

自编码器具有重建过程简单、可堆叠多层和以神经科学为支撑的优点[14]。如今，基于自编码器的方法(如图像分类[15-16]、异常检测[17-18]和模式识别[19]等)被广泛应用于各种研究领域，且取得了成功。

由于所提方法面向电商网络数据，而现有的原始数据往往存在信息缺失的情况，许多有监督方法需要通过手工方法标记大量数据来获取足够的训练样本，非常浪费人力和物力。因此，基于无监督的异常检测受到了广泛的关注，而基于自编码器的异常检测方法由于其良好的性能被广泛应用于电商网络欺诈检测[9]和社交网络异常用户检测等领域[10]。

## 3 面向电商网络的异常用户检测方法

本节提出了一种面向电商网络的异常用户检测方法——自监督异常检测模型，并从基本定义、模型结构和优化阶段 3 个方面对其进行了详细介绍。

### 3.1 基本定义

在电商平台中，每个用户在登录、浏览、购买和评价时都会留下记录。如图 3 所示，这些记录涵盖了用户、设备、产品和商家等不同实体之间的关联关系，其中也包括了异常用户的行为轨迹。为了更直观地分析电商网络中不同用户的行为特征，将电商网络抽象为异质信息网络。

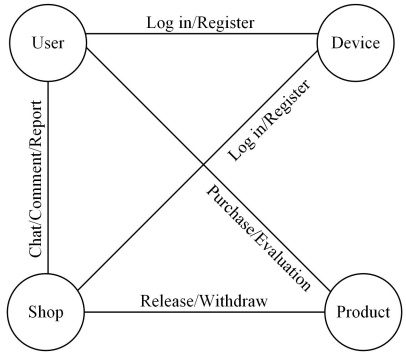


图 3 电商网络中不同实体之间的关联关系

Fig. 3 Relationship between different entities in e-commerce network

**定义 1**(异质信息网络[6]) 给定一个有向图 $G=(V,E;\tau,\varphi;A,R)$，$V$ 代表节点集，$E$ 代表边集，$\tau$ 表示对象类型映射函数，$\varphi$ 表示关系类型映射函数，$\tau(v)\in A$ 表示每个对象 $v\in V$ 都属于一个特定的对象类，$\varphi(e)\in R$ 表示每个关系 $e\in E$ 都属于一种特定的关系类。当节点类型数量 $|A|>1$ 或边的类型数量 $|R|>1$ 时，这样的信息网络被称为异质信息网络，反之为同质信息网络。

2.1 节中，Liu 等[7]分析了电商平台用户在不同设备上的登录情况。因此，我们将重点关注用户和设备两类实体之间的关联关系，将异质信息网络转化为用户-设备二分图，并将其用于接下来的研究工作中。

**定义 2**(二分图结构) 给定一个有向图 $G=(X,Y,E)$，其中，$X=\{x_1,x_2,\cdots,x_m\}$ 表示 $m$ 个用户节点的集合，$Y=\{y_1,y_2,\cdots,y_n\}$ 表示 $n$ 个设备节点的集合，$E=\{e_{ij}\}_{i=1,2,\cdots,m}^{j=1,2,\cdots,n}$ 表示从 $X$ 到 $Y$ 的有向边的集合。如果从 $x_i$ 到 $y_j$ 存在一条边，则 $e_{ij}=1$，否则 $e_{ij}=0$。因此，二分图结构可以表示为矩阵 $\boldsymbol{S}=[\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_m]^{\mathrm{T}}$，其中 $\boldsymbol{s}_i=[e_{i1},e_{i2},\cdots,e_{in}]$。

### 3.2 S-SADM 的模型结构

本节将详细介绍 S-SADM 的模型结构，图 4 给出了 S-SADM 的完整框架。

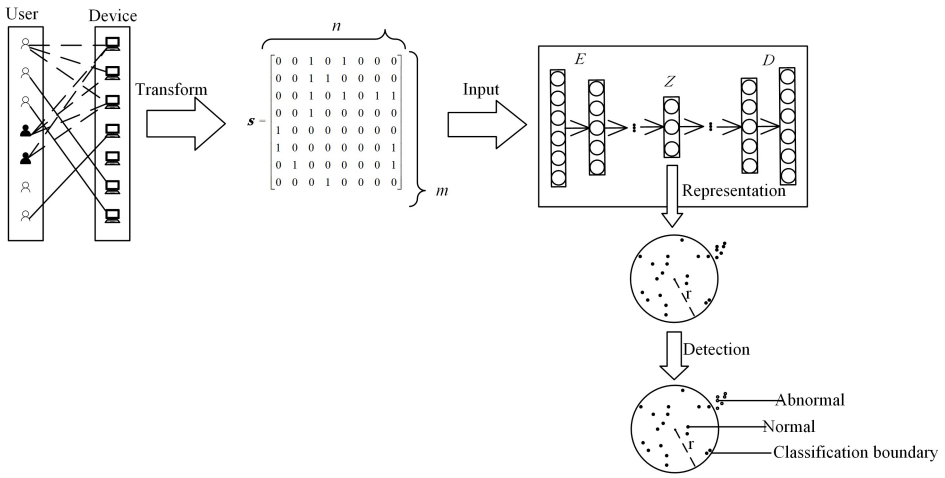


图 4 面向电商网络的异常用户检测方法 S-SADM 的完整框架

Fig. 4 Complete framework of abnormal user detection method S-SADM oriented to e-commerce network

大多数用户在登录电商平台时通常仅在个人设备上进行登录,而异常用户通常在几组特定的设备上进行登录。这种行为在二分图中表现为用户节点倾向于与一小部分设备节点相互关联,导致二分图结构的 $\boldsymbol{S}$ 矩阵极其稀疏,因此采用自编码器编码获取低维稠密的用户节点表示,并完成前向传播。

将二分图结构矩阵 $\boldsymbol{S}$ 作为自编码器的输入,则编码过程的形式化表示为:

$$\boldsymbol{Z}=\sigma(\boldsymbol{WS}+\boldsymbol{b}) \tag{4}$$

然后,将用户节点表示 $\boldsymbol{Z}$ 解码重构为二分图结构矩阵 $\hat{\boldsymbol{S}}$,解码过程的形式化表示为:

$$\hat{\boldsymbol{S}}=\sigma(\boldsymbol{WZ}+\boldsymbol{b}) \tag{5}$$

其中,$\boldsymbol{Z}=[\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_m]^{\mathrm{T}}$,$\boldsymbol{z}_m$ 为第 $m$ 个用户节点在隐空间中对应的用户节点表示,$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 为编码权重和偏置,$\sigma$ 为激活函数,编码器和解码器部分均使用 Relu 激活函数。

为了更加准确地检测异常用户,需要进一步捕获不同用户节点的行为特征。2.1 节中,Liu 等[7]提出了异常用户的两大行为模式,即设备聚集性和活动聚集性。接下来根据这两大模式定义原始空间中用户节点之间的行为相似度。

由设备聚集性可知,异常用户在很大程度上会共享设备。在二分图中表现为异常的用户节点具有许多共同连接的设备节点,这些设备节点使它们之间的相似度较高。而正常用户节点的行为是独立的,总体上相似度较低。用户节点之间的设备相似度定义为:

$$sim\_d_{ij}=\frac{|N_i\cap N_j|}{|N_i\cup N_j|} \tag{6}$$

其中,$N_i:=\{y_j\in Y:e_{ij}=1\}$ 表示与用户节点 $x_i$ 有关联关系的设备节点的集合。

根据活动聚集性可知,异常用户群体会在一天中的某个时间段内爆发集体性活动,因此将一天划为 24 个时间段,统计用户在每个时间段内登录不同设备的次数 $T$。将每个用户的登录行为描述为 $t_i=[T_0,T_1,\cdots,T_{23}]$,则用户节点之间的活动相似度定义为:

$$sim\_t_{ij}=\frac{t_i\cdot t_j}{|t_i|\times|t_j|} \tag{7}$$

结合两类相似度,定义在原始空间中用户节点之间的行为相似度为:

$$sim_{ij}=sim\_d_{ij}\times sim\_t_{ij} \tag{8}$$

由于用户节点 $x_i$ 和用户节点 $x_j$ 在隐空间中的表示分别为 $z_i$ 和 $z_j$,因此用户节点表示之间的行为差异可以通过欧氏距离来定义:

$$dis_{ij}=\|\boldsymbol{z}_i-\boldsymbol{z}_j\|_2^2 \tag{9}$$

进一步将欧氏距离转化为用户节点表示之间的行为相似度,定义如下的映射函数:

$$\widehat{sim}_{ij}=\exp(-dis_{ij}) \tag{10}$$

其中,$\widehat{sim}_{ij}$ 范围为(0,1)。对于用户节点 $x_i$ 和 $x_j$ 而言,当两者距离接近 0 时,则 $\widehat{sim}_{ij}$ 近似于 1,代表 $x_i$ 和 $x_j$ 的行为差异较小。当两者距离足够大时,$\widehat{sim}_{ij}$ 近似于 0,代表 $x_i$ 和 $x_j$ 的行为差异较大。通过缩小 $\widehat{sim}_{ij}$ 和 $sim_{ij}$ 之间的差异,可以获取具有用户行为特征的节点表示。

然后对用户节点在隐空间中的表示进行约束。Ruff 等[20]提出了深度支持向量数据描述(Deep Support Vector Data Description,Deep-SVDD)模型,该模型训练了一个神经网络,通过迭代优化 SVDD 的目标函数,试图将大部分节点表示到核心为 $\boldsymbol{c}$、半径为 $r$ 的最小体积超球隐空间中,而异常节点表示在超球外,根据节点的几何分布有效地检测异常点。

为了使 S-SADM 具有自监督机制,即用户节点表示能够自动为异常检测工作提供监督信息,通过优化联合目标函数 $L$(见式(19))完成反向传播,并对自编码器网络中的权重和偏置实现更新。

接下来,采用支持向量数据描述对用户节点表示 $\boldsymbol{Z}$ 进行异常检测。

首先,计算超球隐空间的核心 $\boldsymbol{c}$:

$$\boldsymbol{c}=\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{z}_i \tag{11}$$

然后,计算各个用户节点表示与核心之间的欧氏距离:

$$d_i=\|\boldsymbol{z}_i-\boldsymbol{c}\|_2^2 \tag{12}$$

并形成距离集合 $D=\{d_1,d_2,\cdots,d_m\}$。

为了寻找合适的超球半径 $r$,通过 $3\sigma$ 准则对集合 $D$ 的正态分布情况进行讨论。下文给出 $3\sigma$ 准则的定义。

若 $x\sim N(\mu,\sigma^2)$,则有:

$$P\{|x-\mu|<\sigma\}=0.6826 \tag{13}$$

$$P\{|x-\mu|<2\sigma\}=0.9545 \tag{14}$$

$$P\{|x-\mu|<3\sigma\}=0.9973 \tag{15}$$

由式(15)可知,正态变量 $x$ 的取值在区间$(\mu-3\sigma,\mu+3\sigma)$之外的概率小于 0.003,一般认为这一事件的概率非常低。

根据该准则,在集合 $D$ 中将该区间以外的 $d_i$ 剔除,并在余下的集合中选择最大值作为半径 $r$,保证了绝大多数节点能够表示在超球隐空间内。最后,将各个用户节点表示与核心间的距离 $d_i$ 与半径 $r$ 进行对比,若 $d_i$ 大于半径 $r$,则该用户节点为异常用户,否则为正常用户。

S-SADM 通过网络的自动迭代优化,使得用户节点表示自动学习到监督信息,并获得较稳定的检测结果。

### 3.3 优化阶段

S-SADM 的联合目标函数共由 3 部分组成。第一部分目标函数是自编码器的重构误差,即原始输入 $\boldsymbol{S}$ 和重构输出 $\hat{\boldsymbol{S}}$ 之间的差异。通过最小化重构误差,可以较完整地保留二分图结构。目标函数 $L_{\mathrm{rec}}$ 的定义如下:

$$L_{\mathrm{rec}}=\sum_{i=1}^{m}\|\hat{\boldsymbol{s}}_i-\boldsymbol{s}_i\|_2^2 \tag{16}$$

第二部分目标函数是 $sim_{ij}$ 和 $\widehat{sim}_{ij}$ 之间的差异。通过最小化该目标函数,获取具有用户行为特征的用户节点表示。目标函数 $L_{\mathrm{sim}}$ 定义如下:

$$L_{\mathrm{sim}}=\sum_{i=1,j=1}^{m}\|\widehat{sim}_{ij}-sim_{ij}\|_2^2 \tag{17}$$

在文献[20]中,Deep-SVDD 通过优化 SVDD 的目标函数,将大部分用户节点围绕核心表示在最小超球内,异常用户节点则表示在最小超球外。为了约束用户节点表示,将第三部分目标函数[20]定义为:

$$L_{\mathrm{svdd}}=\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{z}_i-\boldsymbol{c}\|_2^2 \tag{18}$$

最后，为了同时在用户节点表示中保留二分图结构和用户行为特征，并对其进行约束，根据式(16)－式(18)定义联合目标函数。目标函数 $L$ 的定义为：

$$L=L_{rec}+\alpha(L_{sim}+L_{svdd}) \tag{19}$$

其中，$\alpha$ 为超参数。通过最小化 $L$，可以获得具有监督信息的用户节点表示。

## 4 S-SADM 算法的流程

本节给出了 S-SADM 的算法流程，为接下来的实验验证及对比分析提供了理论指导。

**算法 1** 面向电商网络的异常检测方法 S-SADM

输入：异质信息网络数据

输出：异常检测评价指标结果

1. 通过定义 2，将异质网络数据转换为用户-设备二分图并构建二分图结构矩阵 **S**
2. for i，j=0 to m
3. 通过式(8)计算用户节点之间的行为相似度
4. end for
5. 将矩阵 **S** 作为自编码器的输入，优化器选择随机梯度下降，设置 epoch，batch size，learning rate 的大小
6. for i=0 to epoch
7. 通过式(4)计算用户节点表示，完成前向传播
8. 通过式(10)计算用户节点表示之间的行为相似度
9. 通过式(19)优化联合目标函数 L，完成反向传播
10. 通过式(11)计算超球隐空间核心
11. for i=0 to m
12. 通过式(12)计算各个用户节点表示与核心之间的距离，形成集合 D
13. end for
14. 通过式(13)－式(15)的原理，剔除 D 中在区间外的 $d_i$，选择余下的集合中最大的值作为半径 r
15. for i=0 to m
16. if $d_i > r$
17. 该节点为异常用户点
18. else
19. 该节点为正常用户点
20. end if
21. end for
22. return 异常检测评价指标结果
23. end for

## 5 实验结果及分析

为了验证 S-SADM 的有效性和优越性，选取几种现有的基线方法与其进行对比实验。实验环境为 Intel Core i7-7700 CPU 3.60 GHz，内存为 8 GB，操作系统为 Win10 64 bit。

### 5.1 基线方法

首先对实验中的基线方法进行介绍，包括孤立森林等 3 种经典异常检测方法、基于自编码器的经典异常检测方法以及近几年提出的异常检测方法 DeepFD[9] 和 FraudNE[10]。具体介绍如表 1 所列。

表 1 基线方法

Table 1 Baseline methods

| 方法名称 | 方法介绍 |
|---|---|
| IF | 孤立森林[21](Isolation Forest，IF)是一种无监督学习异常检测方法，该方法认为当随机树森林为某些特定的点集体产生较短的路径长度时，它们很可能是异常的 |
| KNN | $K$ 最近邻算法[2]($K$-Nearest Neighbor，KNN)通过距离的比较来区分异常数据，如果某数据点的 $k$ 个最近邻点都属于异常，那么该数据点属于异常点 |
| LOF | 局部异常因子算法[1](Local Outlier Factor，LOF)通过将样本的局部密度与其邻居的局部密度进行比较，可以识别出密度远低于其邻居的样本，该样本则被认为是异常样本 |
| AE+IF | 训练自编码器并获取数据的网络表示，并在此基础上采用 IF 进行异常检测 |
| AE+KNN | 训练自编码器并获取数据的网络表示，并在此基础上采用 KNN 进行异常检测 |
| AE+LOF | 训练自编码器并获取数据的网络表示，并在此基础上采用 LOF 进行异常检测 |
| DeepFD | 文献[9]通过自编码器对用户-项目二分图编码以获取具有行为特征的低维用户节点表示，并采用 DBSCAN 对用户节点表示进行欺诈块检测 |
| FraudNE | 文献[10]通过深度联合网络嵌入来检测异常用户，它充分利用了用户-项目二分图的拓扑信息，将所有用户节点和项目节点同时嵌入同一个低维空间中，并采用聚类方法检测欺诈块 |

### 5.2 数据集

本文实验在 Kaggle 上的 3 个真实电商网络数据集和 1 个半合成电商网络数据集中进行，数据集如表 2 所列。下文对 4 个数据集进行描述。

(1)真实数据集 1 记录了某电商平台上的用户购买记录，其中包括用户 ID、用户年龄、用户性别、登录设备 ID、采购时间等。本实验在原始数据集中随机采样，从而构成该数据集。

(2)真实数据集 2 记录了某电商平台上的用户购买记录，其中包括用户 IP 地址、用户邮箱地址、用户电话和登录设备 ID 等。本实验在原始数据集中随机采样，从而构成该数据集。

(3)真实数据集 3 记录了某大型在线商店 2019 年 10 月—2020 年 4 月的用户购买记录，其中包括用户 ID、商品 ID、登录设备 ID 和采购时间等。本实验在原始数据集中随机采样，从而构成该数据集。

(4)半合成数据集记录了某电商平台上的用户购买记录，其中包括用户 ID、用户年龄、用户性别、登录设备 ID 和采购时间等。它在原始数据中随机采样，并加入了部分异常用户的行为记录，其中异常用户数量占所有用户数量的 5%。

表 2 电商网络数据集

Table 2 E-commerce network data sets

| 数据集 | 用户节点数 | 设备节点数 | 边数 |
|---|---|---|---|
| 真实数据集 1 | 189 | 282 | 5000 |
| 真实数据集 2 | 43 | 105 | 5000 |
| 真实数据集 3 | 236 | 275 | 3000 |
| 半合成数据集 | 60 | 88 | 4000 |

### 5.3 评价指标

由于一般异常检测的数据集都是不平衡数据集，因此评价指标更为复杂。通过5个评价指标对各种方法进行性能衡量，分别是精确率(Precision，$P$)、召回率(Recall，$R$)、$F1$-*measure*、$AUC$ 和 $G$-*mean*。

异常检测问题可以作为分类问题。因此，使用混淆矩阵进行度量[22]，如表3所列。其中，$TP$(True Positives)表示正类样本被正确分类的数量，$FN$(False Negatives)表示正类样本被错误分类的数量，$FP$(False Positives)表示负类样本被错误分类的数量，$TN$(True Negatives)表示负类样本被正确分类的数量[23]。

表3 混淆矩阵

Table 3 Confusion matrix

| 真实情况 | 检测结果 | |
|---|---|---|
| | 正例 | 反例 |
| 正例 | 真正例($TP$) | 假负例($FN$) |
| 反例 | 假正例($FP$) | 真负例($TN$) |

(1) 精确率为：

$$P=\frac{TP}{TP+FP} \tag{20}$$

(2)召回率为：

$$R=\frac{TP}{TP+FN} \tag{21}$$

(3)一般来说，精确度和召回率之间是矛盾的，这里引入F1-measure作为综合指标，以平衡准确率和召回率的影响并较为全面地评价一个分类器。F1-measure为：

$$F1=\frac{2\times P\times R}{P+R} \tag{22}$$

(4)AUC是ROC曲线下的面积，是分类方法的一个重要评价指标。ROC曲线的纵轴是“真正例率”(True Positive Rate，TPR)，代表实际的正例中被正确检测的比例。ROC曲线的横轴是“假正例率”(False Positive Rate，FPR)，代表实际的负例中被错误检测的比例。即：

$$TPR=\frac{TP}{TP+FN} \tag{23}$$

$$FPR=\frac{FP}{FP+TN} \tag{24}$$

(5)G-mean：用于不平衡数据集中，即：

$$G\text{-}mean=\sqrt{R\times TNR} \tag{25}$$

其中，$TNR$(True Negative Rate)代表检测出的负例占所有负例的比例，即：

$$TNR=\frac{TN}{TN+FP} \tag{26}$$

### 5.4 实验验证及对比分析

本实验在tensorflow框架上进行，隐空间维度设为10。实验结果如表4－表7所列，表4中加粗数据为该指标的最优结果。接下来对实验结果进行对比分析。

首先对4项实验中的经典异常检测方法和基于自编码器的经典异常检测方法进行对比分析。根据实验结果可得，后者在不同指标上的表现结果较前者有一定提升，说明自编码器能够较好地提取网络数据特征并将其成功应用于异常检测中。

表4 不同方法在真实数据集1中的评价指标结果

Table 4 Evaluation index results of different methods in real data set 1

| 方法名称 | $P$ | $R$ | $F1$-*measure* | $AUC$ | $G$-*mean* |
|---|---|---|---|---|---|
| IF | 0.9912 | 0.8958 | 0.9410 | 0.4497 | 0.8976 |
| KNN | 0.9912 | 0.9099 | 0.9488 | 0.4569 | 0.9119 |
| LOF | 0.9911 | 0.9046 | 0.9459 | 0.4542 | 0.9065 |
| AE+IF | **0.9915** | 0.8958 | 0.9411 | 0.4498 | 0.9011 |
| AE+KNN | 0.9912 | 0.914 | 0.9499 | 0.4535 | 0.9108 |
| AE+LOF | 0.9912 | 0.9099 | 0.9494 | 0.4699 | 0.9066 |
| DeepFD | 0.9529 | 0.9748 | 0.9638 | 0.5258 | 0.9866 |
| FraudNE | 0.9633 | 0.9714 | 0.9681 | 0.5523 | 0.9827 |
| S-SADM | 0.9792 | **0.9795** | **0.9788** | **0.7310** | **0.9869** |

表5 不同方法在真实数据集2中的评价指标结果

Table 5 Evaluation index results of different methods in real data set 2

| 方法名称 | $P$ | $R$ | $F1$-*measure* | $AUC$ | $G$-*mean* |
|---|---|---|---|---|---|
| IF | 0.5844 | 0.6171 | 0.5295 | 0.5326 | 0.7537 |
| KNN | 0.598 | 0.6126 | 0.5293 | 0.5316 | 0.7572 |
| LOF | 0.5932 | 0.6107 | 0.5274 | 0.5298 | 0.7554 |
| AE+IF | 0.5847 | 0.6399 | 0.5422 | 0.4964 | 0.7701 |
| AE+KNN | 0.6189 | 0.6049 | 0.5347 | 0.5329 | 0.7688 |
| AE+LOF | 0.6163 | 0.6044 | 0.5369 | 0.5435 | 0.7631 |
| DeepFD | 0.8106 | 0.8879 | 0.8410 | 0.5220 | 0.9421 |
| FraudNE | 0.8306 | 0.8991 | 0.8622 | 0.5421 | 0.9465 |
| S-SADM | **0.9174** | **0.9103** | **0.9144** | **0.6015** | **0.9523** |

表6 不同方法在真实数据集3中的评价指标结果

Table 6 Evaluation index results of different methods in real data set 3

| 方法名称 | $P$ | $R$ | $F1$-*measure* | $AUC$ | $G$-*mean* |
|---|---|---|---|---|---|
| IF | 0.9901 | 0.8953 | 0.9403 | 0.4497 | 0.8974 |
| KNN | **0.9901** | 0.9153 | 0.9513 | 0.4598 | 0.9175 |
| LOF | 0.9902 | 0.9006 | 0.9433 | 0.4524 | 0.9027 |
| AE+IF | 0.9901 | 0.8954 | 0.9404 | 0.4497 | 0.9134 |
| AE+KNN | 0.9905 | 0.9005 | 0.9586 | 0.4543 | 0.9273 |
| AE+LOF | 0.9902 | 0.9137 | 0.9462 | 0.4558 | 0.9319 |
| DeepFD | 0.9580 | **0.9788** | 0.9683 | 0.5000 | 0.9830 |
| FraudNE | 0.9487 | 0.9635 | 0.9580 | 0.5008 | 0.9811 |
| S-SADM | 0.9748 | 0.9766 | **0.9741** | **0.6486** | **0.9839** |

表7 不同方法在半合成数据集中的评价指标结果

Table 7 Evaluation index results of different methods in semi-synthetic data set

| 方法名称 | $P$ | $R$ | $F1$-*measure* | $AUC$ | $G$-*mean* |
|---|---|---|---|---|---|
| IF | 0.8932 | 0.8481 | 0.8700 | 0.4472 | 0.8709 |
| KNN | 0.8940 | 0.8610 | 0.8771 | 0.4541 | 0.8842 |
| LOF | 0.8986 | 0.8629 | 0.8802 | 0.4824 | 0.8848 |
| AE+IF | 0.8936 | 0.849 | 0.8789 | 0.4625 | 0.8717 |
| AE+KNN | 0.8939 | 0.8605 | 0.8791 | 0.4551 | 0.8907 |
| AE+LOF | 0.9036 | 0.8691 | 0.8791 | 0.5165 | 0.893 |
| DeepFD | **0.9116** | 0.9413 | 0.9245 | 0.497 | 0.9701 |
| FraudNE | 0.9100 | 0.9490 | 0.9294 | 0.4991 | 0.9711 |
| S-SADM | 0.9109 | **0.9512** | **0.9312** | **0.5217** | **0.9725** |

对4项实验中的S-SADM和基线方法进行对比分析。表4列出了不同方法在真实数据集1中的评价指标结果，S-SADM在召回率、F1-measure、AUC和G-mean 4项指标上的表现更为出众。表5列出了不同方法在真实数据集2中的评价指标结果，在该实验中，S-SADM在不同指标上均表现出了优于基线方法的能力。表6列出了不同方法在真实数据集3中的评价指标结果，其中KNN在精确率这项指标上达到最优，DeepFD在召回率这项指标上表现最好，S-SADM在F1-measure、AUC和G-mean3项综合指标上的表现远优于基线

方法，在召回率指标上的表现仅次于DeepFD。表7列出了不同方法在半合成数据集中的评价指标结果，其中DeepFD在精确率这项指标上表现最好，而S-SADM在该项指标上仅次于DeepFD，并在其余评价指标上有着优于基线方法的表现。

分析实验结果可知，S-SADM在不同数据集中的不同评价指标上都展现了其有效性和优越性，说明具有自监督学习机制的异常检测方法有着较强的检测性能和泛化能力。

为了对所有方法的泛化能力进行全面的比较，对不同方法在不同数据集下的F1-measure评价指标结果进行了Friedman检验。

Friedman检验的原假设和对立假设分别为H0和H1。

H0：不同方法在不同数据集中的F1-measure评价指标结果无差异，即不同方法的性能相同。

H1：不同方法在不同数据集中的F1-measure评价指标结果存在差异，即不同方法的性能存在差异。

首先对不同方法在不同数据集上的F1-measure评价指标结果进行排序，并对每一行序值取平均值，得到平均序值，如表8所列。

表8 序值表

Table 8 Ordinal number

| 方法名称 | 真实数据集1 | 真实数据集2 | 真实数据集3 | 半合成数据集 | 平均序值 |
|---|---|---|---|---|---|
| IF | 9 | 7 | 9 | 9 | 8.50 |
| KNN | 6 | 8 | 5 | 7 | 6.50 |
| LOF | 7 | 9 | 7 | 4 | 6.75 |
| AE+IF | 8 | 4 | 8 | 8 | 7.00 |
| AE+KNN | 4 | 6 | 3 | 5 | 4.50 |
| AE+LOF | 5 | 5 | 6 | 6 | 5.50 |
| DeepFD | 3 | 3 | 2 | 3 | 2.75 |
| FraudNE | 2 | 2 | 4 | 2 | 2.50 |
| S-SADM | 1 | 1 | 1 | 1 | 1.00 |

假定在$N$个数据集上比较$k$种方法，令$r_i$表示第$i$种法的平均序值。$r_i$服从正态分布，其均值为$(k+1)/2$，方差为$(k^2-1)/12$，则有$\chi^2$分布：

$$\tau_{\chi^2}=\frac{12N}{k(k+1)}\left(\sum_{i=1}^{k}r_i^2-\frac{k(k+1)^2}{4}\right) \tag{27}$$

根据式(27)可推导F分布：

$$\tau_F=\frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}} \tag{28}$$

其中，$\tau_F$为服从自由度为$k-1$和$(k-1)(N-1)$的F分布[24]。在显著性水平$\alpha=0.05$、自由度为8时，通过计算，$\tau_F=13.94$，大于临界值2.355[24]，证明H0的假设内容不成立，即不同方法的性能存在差异。

为了更好地对各种方法进行区分检验，通过Friedman检验图对不同方法进行分析，如图5所示。

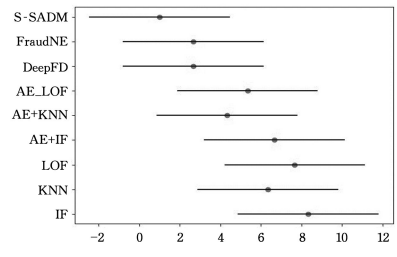


图5 Friedman检验图

Fig. 5 Friedman test chart

由图5可知，不同方法的横线段均有交叠区域，说明不同方法的性能不存在显著性差异，同时说明了S-SADM的检测性能与基线方法相比有一定的提升。

接下来，对超参数$\alpha$进行敏感度分析。当$\alpha$取(0,1)中的不同值时，观察S-SADM在不同数据集中的评价指标结果是否发生变化，实验结果如图6—图9所示。

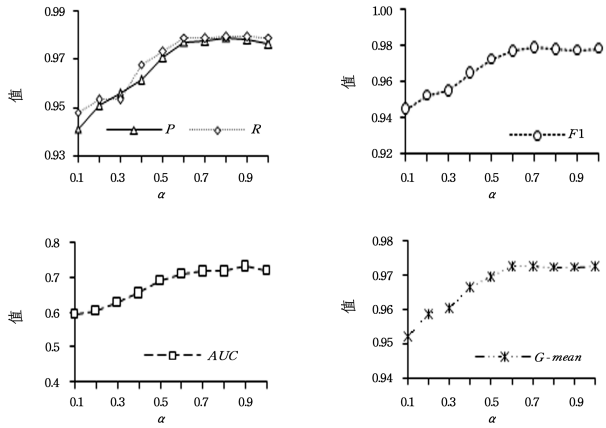


图6 S-SADM在数据集1中$\alpha$变化时的不同评价指标结果

Fig. 6 Different evaluation index results of S-SADM when $\alpha$ changes in data set 1

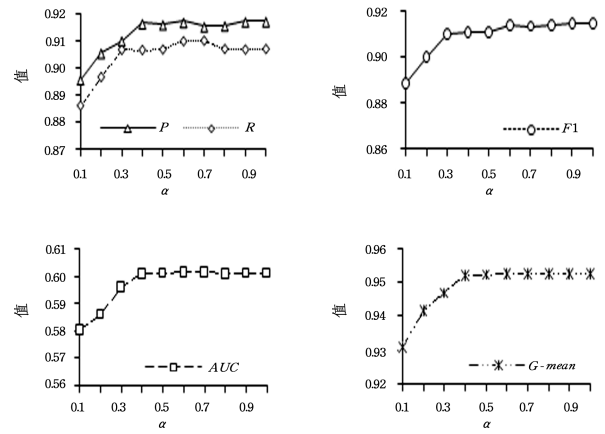


图7 S-SADM在数据集2中$\alpha$变化时的不同评价指标结果

Fig. 7 Different evaluation index results of S-SADM when $\alpha$ changes in data set 2

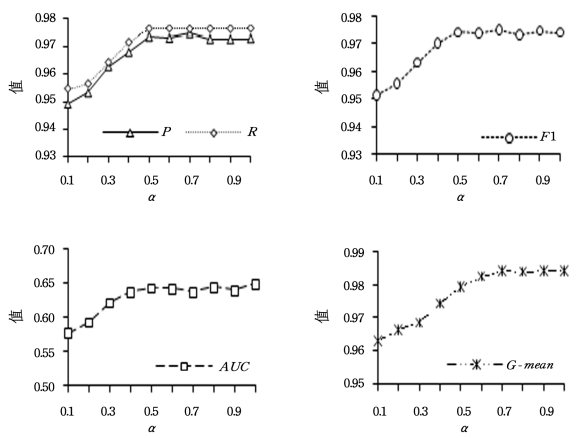


图8 S-SADM在数据集3中$\alpha$变化时的不同评价指标结果

Fig. 8 Different evaluation index results of S-SADM when $\alpha$ changes in data set 3

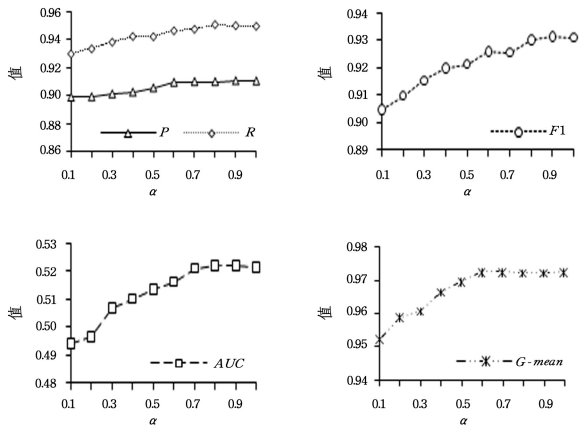


图 9 S-SADM 在半合成数据集中 $\alpha$ 变化时的不同评价指标结果

Fig. 9 Different evaluation index results of S-SADM when $\alpha$ changes in semi-synthetic data set

根据实验结果分析可知，随着 $\alpha$ 值逐渐增大，不同评价指标的值不断增大，当 $\alpha$ 增大到一定值之后，不同评价指标值的变化趋于稳定，说明超参数 $\alpha$ 在一定范围内对训练 S-SADM 存在影响，同时证明了设计联合损失函数 $L$ 的合理性。

此外，对隐空间维度进行敏感度分析。设 $\alpha=1$，隐空间维度 $d$ 的范围为(2,10)。当 $d$ 取不同值时，观察 S-SADM 在不同数据集中的 F1-measure 评价指标结果是否发生变化，实验结果如图 10 所示。由图 10 可知，随着 $d$ 不断增大，不同数据集中的 F1-measure 缓慢升高，当 $d$ 增大到一定值之后，变化趋于稳定。根据实验结果分析可得，隐空间维度 $d$ 在不同实验中表现出较低的敏感度，当 $d$ 在一定大小范围内时，在低维空间中获得了较完整的数据表征，使得 S-SADM 的评价指标结果趋于稳定。

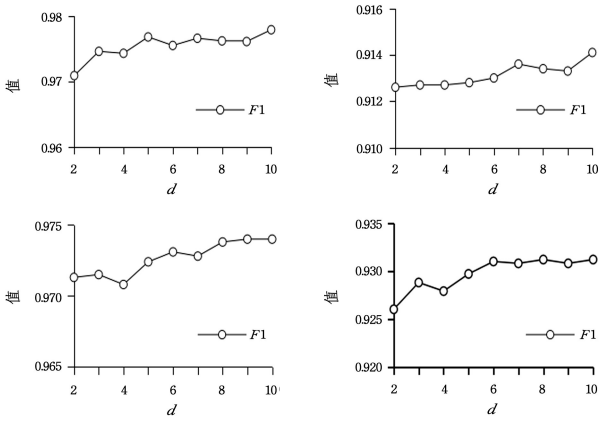


图 10 S-SADM 在不同数据集中 $d$ 变化时的 F1-measure 评价指标结果

Fig. 10 F1-measure evaluation index results of S-SADM when $d$ changes in different data sets

下文对激活函数在 S-SADM 中的效果进行讨论，并通过 S-SADM 在不同数据集下的 F1-measure 评价指标结果进行对比分析。由表 9 可知，ReLu 函数在精确率、F1-measure 和 G-mean 3 项评价指标上达到最优；由表 10 和表 12 可知，ReLu 函数在不同评价指标上均有突出的表现；由表 11 可知，ReLu 函数在精确率、F1-measure、AUC 和 G-mean 4 项评价指标上效果最好。综上所述，当 ReLu 函数作为 S-SADM 中的激活函数时，S-SADM 的检测性能最佳。

表 9 S-SADM 在真实数据集 1 中激活函数变化时的 F1-measure 评价指标结果

Table 9 F1-measure evaluation index results when activation function of S-SADM changes in real data set 1

| 激活函数 | *P* | *R* | *F1* | *AUC* | *G-mean* |
|---|---|---|---|---|---|
| ReLu | **0.9915** | 0.9795 | **0.9788** | 0.7310 | **0.9869** |
| Sigmoid | 0.9795 | **0.9808** | 0.9778 | **0.7536** | 0.9866 |
| Tanh | 0.9776 | 0.9784 | 0.9780 | 0.7071 | 0.9866 |

表 10 S-SADM 在真实数据集 2 中激活函数变化时的 F1-measure 评价指标结果

Table 10 F1-measure evaluation index results when activation function of S-SADM changes in real data set 2

| 激活函数 | *P* | *R* | *F1* | *AUC* | *G-mean* |
|---|---|---|---|---|---|
| ReLu | **0.9174** | **0.9103** | **0.9144** | **0.6015** | **0.9523** |
| Sigmoid | 0.9158 | 0.9069 | 0.8782 | 0.6000 | 0.9523 |
| Tanh | 0.9158 | 0.9069 | 0.8782 | 0.6000 | 0.9523 |

表 11 S-SADM 在真实数据集 3 中激活函数变化时的 F1-measure 评价指标结果

Table 11 F1-measure evaluation index results when activation function of S-SADM changes in real data set 3

| 激活函数 | *P* | *R* | *F1* | *AUC* | *G-mean* |
|---|---|---|---|---|---|
| ReLu | **0.9754** | 0.9744 | **0.9759** | **0.6933** | **0.9839** |
| Sigmoid | 0.9734 | 0.9757 | 0.9745 | 0.6733 | 0.9831 |
| Tanh | 0.9748 | **0.9766** | 0.9741 | 0.6486 | 0.9830 |

表 12 S-SADM 在半合成数据集中激活函数变化时的 F1-measure 评价指标结果

Table 12 F1-measure evaluation index results when activation function of S-SADM changes in semi-synthetic data set

| 激活函数 | *P* | *R* | *F1* | *AUC* | *G-mean* |
|---|---|---|---|---|---|
| ReLu | **0.9116** | **0.9512** | **0.9312** | **0.5217** | **0.9725** |
| Sigmoid | 0.9105 | 0.9400 | 0.9247 | 0.4951 | 0.9627 |
| Tanh | 0.9106 | 0.9339 | 0.9218 | 0.4916 | 0.9555 |

根据第 4 节中的算法流程可知，S-SADM 的时间复杂度为 $O(m)+O(epoch*2m)$，即 $O(n^2)$，较经典异常检测方法的时间复杂度更高，与其余基线方法相比具有相同的时间复杂度。在时间复杂度相同的情况下，该方法在不同数据集中具有更加突出的表现。

**结束语** 电商网络包括了不同实体间的关联关系，其本身具有异质信息网络的结构。通过分析异常用户行为模式，提出了一种面向电商网络的异常检测方法 S-SADM，该方法具有自监督机制。通过实验验证及对比分析可知，S-SADM 在不同评价指标上均优于基线方法，证明了该方法的可行性和优越性。

本文提出的方法 S-SADM 虽然表现出了良好的异常检测性能，但忽略了对异质信息网络中节点属性信息和边属性信息的思考。此外，本文重点关注了用户与设备之间的关联关系，在电商网络中不同实体之间依旧存在着其他的关联关系，我们希望今后通过分析实体间的行为模式，来进一步扩展异常用户检测的研究工作。

## 参 考 文 献


[1] HU C P, QIN X L. A density-based local outlier detection algorithm[J]. Journal of Computer Research and Development, 2010, 47(12): 2110-2116.

[2] REN J D, LIU X Q, WANG Q, et al. Multi-layer intrusion detection method based on KNN outlier detection and random forest [J]. Journal of Computer Research and Development, 2019, 56(3): 116-125.

[3] RASHEED F, ALHAJJ R. A framework for periodic outlier pattern detection in time-series sequences[J]. IEEE Transactions on Cybernetics, 2013, 44(5): 569-582.

[4] CHAKRABORTY D, NARAYANAN V, GHOSH A. Integration of deep feature extraction and ensemble learning for outlier detection[J]. Pattern Recognition, 2019, 89: 161-171.

[5] WU S, WANG S R. Information-theoretic outlier detection for large-scale categorical data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(3): 589-602.

[6] LIU L, ZUO W L, PENG T. Dynamic outlier detection method based on tensor representation in heterogeneous network[J]. Journal of Computer Research and Development, 2016, 53(8): 1729-1739.

[7] LIU Z Q, CHEN C C, YANG X X, et al. Heterogeneous graph neural networks for malicious account detection[C]// ACM International Conference. New York: ACM, 2018: 2077-2085.

[8] GUPTA M, GAO J, HAN J W. Community distribution outlier detection in heterogeneous information networks[C]// Joint European Conference on Machine Learning & Knowledge Discovery in Databases. Berlin: Springer, 2013: 11-29.

[9] WANG H B, ZHOU C, WU J, et al. Deep structure learning for fraud detection[C]// IEEE International Conference on Data Mining. NJ: IEEE, 2018: 567-576.

[10] ZHENG M Y, ZHOU C, WU J, et al. FraudNE: a Joint Embedding Approach for Fraud Detection International[C]// Joint Conference on Neural Networks. NJ: IEEE, 2018: 4739-4746.

[11] DONG M Q, YAO L N, WANG X Z, et al. Opinion fraud detection via neural autoencoder decision forest[J]. Pattern Recognition Letters, 2020, 132: 21-29.

[12] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[13] BOURLARD H, KAMP Y. Auto-assoation by multilayer perceptrons and singular value decomposition[J]. Biological Cybernetics, 1988, 59(4): 291-294.

[14] YUAN F N, ZHANG L, SHI J T, et al. Overview of auto-encoding neural network theory and application[J]. Chinese Journal of Computers, 2019, 42(1): 203-230.

[15] LI E Z, DU P J, SAMAT A, et al. Mid-level feature representation via sparse autoencoder for remotely sensed scene classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(3): 1068-1081.

[16] GENG J, WANG H Y, FAN J C, et al. Deep supervised and contractive neural network for sar image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 4: 1-18.

[17] HASAN M, CHOI J. NEUMANN J, et al. Learning temporal regularity in video sequences[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. NJ: IEEE. 2016: 733-742.

[18] RIBEIRO M, LAZZARETTI A E, LOPES H S. A study of deep convolutional auto-encoders for anomaly detection in videos[J]. Pattern Recognition Letters, 2018, 105: 13-22.

[19] GAO S H, ZHANG Y T, JIA K, et al. Single sample face recognition via learning deep supervised autoencoders [J]. IEEE Transactions on Information Forensics and Security, 2015, 10(10): 2108-2118.

[20] RUFF L, VANDERMEULEN R, GOERNITZ N, et al. Deep one-class classification[C]// International conference on machine learning. Cambridge MA: JMLR, 2018: 4390-4399.

[21] LIU T, TING K M, ZHOU Z H. Isolation-based anomaly detection[J]. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1): 1-39.

[22] NAGANJANEYULU S, KUPPA M R. A novel framework for class imbalance learning using intelligent under-sampling[J]. Progress in Artificial Intelligence, 2013, 2(1): 73-84.

[23] JIANG K, LU J, XIA K L. A novel algorithm for imbalance data classification based on genetic algorithm improved smote[J]. Arabian Journal for Science & Engineering, 2016, 41(8): 3255-3266.

[24] ZHOU Z H. Machine Learning [M]. Beijing: Tsinghua University Press, 2016: 42-43.


**DU Hang-yuan**, born in 1985, Ph.D, associate professor, master supervisor. His main research interests include cluster analysis and complex network theory.

**WANG Wen-jian**, born in 1968, Ph.D, professor, Ph.D supervisor, is a senior member of China Computer Federation. Her main research interests include machine learning, data mining and computational intelligence.

（责任编辑：杨雪敏）