计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700003

结合图卷积神经网络和集成方法的推荐系统恶意攻击检测

刘 慧[1,2] 纪 科[1,2] 陈贞翔[1,2] 孙润元[1,2] 马 坤[1,2] 邬 俊[3]
1 济南大学信息科学与工程学院 济南 250022
2 山东省网络环境智能计算技术重点实验室(济南大学) 济南 250022
3 北京交通大学计算机与信息技术学院 北京 100044
(liuhui370285@qq.com)

摘 要 推荐系统已被广泛应用于电子商务、社交媒体、信息分享等大多数互联网平台中,有效解决了信息过载问题。然而,这些平台面向所有互联网用户开放,导致不法用户利用系统设计缺陷通过恶意干扰、蓄意攻击等行为非法操纵评分数据,进而影响推荐结果,严重危害推荐服务的安全性。现有的检测方法大多都是基于从评级数据中提取的人工构建特征进行的托攻击检测,难以适应更复杂的共同访问注入攻击,并且人工构建特征费时且区分能力不足,同时攻击行为规模远远小于正常行为,给传统检测方法带来了不平衡数据问题。因此,文中提出堆叠多层图卷积神经网络端到端学习用户和项目之间的多阶交互行为信息得到用户嵌入和项目嵌入,将其作为攻击检测特征,以卷积神经网络作为基分类器实现深度行为特征提取,结合集成方法检测攻击。在真实数据集上的实验结果表明,与流行的推荐系统恶意攻击检测方法相比,所提方法对共同访问注入攻击行为有较好的检测效果并在一定程度上克服了不平衡数据的难题。

关键词:攻击检测;共同访问注入攻击;推荐系统;图卷积神经网络;卷积神经网络;集成方法
中图分类号 TP391

Malicious Attack Detection in Recommendation Systems Combining Graph Convolutional Neural Networks and Ensemble Methods

LIU Hui[1,2],JI Ke[1,2],CHEN Zhenxiang[1,2],SUN Runyuan[1,2],MA Kun[1,2] and WU Jun[3]
1 School of Information Science and Engineering,University of Jinan,Jinan 250022,China
2 Shandong Provincial Key Laboratory of Network Based Intelligent Computing(University of Jinan),Jinan 250022,China
3 School of Computer and Information Technology,Beijing Jiaotong University,Beijing 100044,China

Abstract Recommendation systems have been widely used in most Internet platforms,such as e-commerce,social media,and information sharing,which effectively solve the problem of information overload. However,these platforms are open to all Internet users,leading to illegal manipulation of rating data through malicious interference and deliberate attacks by unscrupulous users using system design flaws,affecting the recommendation results and seriously jeopardizing the security of recommendation services. Most existing detection methods are based on manually constructed features extracted from rating data for shilling attack detection,which is challenging to adapt to more complex co-visitation injection attacks,and manually constructed features are time-consuming and need more differentiation capability. In contrast,the scale of attack behavior is much smaller than normal behavior,bringing imbalanced data problems to traditional detection methods. Therefore,the paper proposes stacked multilayer graph convolutional neural networks end-to-end to learn multi-order interaction behavior information between users and items to obtain user embeddings and item embeddings,which are used as attack detection features,and convolutional neural networks are used as base classifiers to achieve deep behavior feature extraction,combined with ensemble methods to detect attacks. Experimental results on real datasets show that the method better detects co-visitation injection attacks and overcomes the imbalanced data problem to a certain extent compared with popular malicious attack detection methods for recommendation systems.

Keywords Attack detection,Co-visitation injection attack,Recommendation systems,Graph convolutional neural networks,Convolutional neural networks,Ensemble methods

基金项目:国家自然科学基金(61702216,61772231,61671048,61672262);山东省重大科技创新工程(2018CXGC0706)
This work was supported by the National Natural Science Foundation of China(61702216,61772231,61671048,61672262) and Key Research and Development Program of Shandong Province(2018CXGC0706).
通信作者:纪科(ise_jik@ujn.edu.cn)

1 引言

随着互联网的发展，个性化推荐系统已经成为各种在线应用程序的重要组成部分，如电商平台、视频平台等[1-2]。这些平台会基于用户的评级评论、点击行为、兴趣爱好等信息预测用户对项目的偏好，向用户推荐符合用户偏好的物品[3-5]。广泛应用的推荐方法包括协同过滤方法和基于图的推荐方法等[6-7]。这些方法的基本假设是，如果两个用户在过去表现出相似的偏好，未来也将有相似的偏好。然而由于推荐系统的开放性和系统漏洞[8-10]，尽管来自用户的这些输入丰富了推荐系统的数据库，但也使系统容易受到多种类型的恶意攻击。

恶意用户为了获取更多的平台流量曝光，将自己的项目展现在更多的消费者面前，要么通过实施托攻击向评分系统注入足够数量的精心设计的虚假用户概貌（例如评级、评论），并根据经验对目标项目进行更高（推广攻击）或更低的评分（贬低攻击）[11]；要么通过实施共同访问注入攻击向系统中注入虚假的共同访问，以欺骗推荐系统[8]。因此，推荐系统的恶意攻击行为不仅损害了消费者的利益，扰乱平台的公平性，而且动摇了虚拟市场中客户和用户的信心。

为了解决上述问题，研究人员提出了许多方法，如聚类技术、统计技术、分类技术等[12]。尽管这些方法对检测恶意攻击是有效的，但它们的局限性在于大部分都是基于评级数据和人工构建的精心设计的特征进行的托攻击检测，其为所有用户计算评级特征，如 Top-N 近邻用户相似度（Degree of Similarity With Top Neighbors，DegSim）、用户评级数量和数据库中的平均评级数量偏差（Length Variance，LengthVar）等。检测性能很大程度上取决于人工构建的特征的质量且大多数特征仅对某些类型的托攻击有效[11,13]。面对大规模的真实数据时，由于计算成本高昂，检测性能仍然是有限的[14]。因此，人工方式构建的特征具有非线性小、通常难以提取、区分能力不高、需要较高的知识成本的缺陷，不足以处理复杂攻击的影响，如共同访问注入攻击[7]。另外，现实世界里推荐系统中恶意攻击数据的数量远远少于正常数据的数量，因此推荐系统恶意攻击检测实际上可以被描述为不平衡分类问题。传统的检测方法对少数类不敏感，不能有效地检测出相关攻击。深度学习算法的引入可以弥补人工构建特征的不足。

综上，文中提出结合图卷积神经网络和集成方法的推荐系统恶意攻击检测方法。我们认为用户项目交互蕴含着丰富的潜在行为特征，因此通过分析数据集用户点击行为数据构建用户-项目二部图，并利用图卷积神经网络（Graph Convolutional Neural Networks，GCN）显示编码多阶交互信息构建用户特征和项目特征，自动提取特征对点击行为进行分类，处理共同访问注入攻击。本文的主要贡献如下：

（1）基于 GCN 的嵌入表示学习模块在二部图上通过图学习传播嵌入显示编码多阶用户项目交互信息，形成嵌入向量表示，捕获用户项目交互的行为特征和攻击行为蕴含的隐性交互。其通过图学习从点击行为中学习特征而不是手工设计特征，弥补了人工构建特征难以提取、需要知识成本高及区分能力弱的不足。

（2）将基于卷积神经网络（Convolutional Neural Networks，CNN）的攻击行为分类器与集成方法装袋算法（Bootstrap Aggregating，Bagging）相结合，实现了用户和项目之间细粒度交互行为特征的自动提取，很好地解决了不平衡分类问题。

（3）在真实数据集上的实验结果表明所提算法性能优于目前比较流行的攻击检测算法，能够有效地检测出共同访问注入攻击。

2 相关工作

2.1 推荐系统恶意攻击

恶意用户对推荐系统发起不同类型的攻击，其中一种攻击是托攻击[11]。协同过滤技术能够利用用户的历史评分等信息寻找与其相似的近邻，根据多个最近邻的信息为目标用户产生推荐结果。而托攻击正是利用这一点向评分系统中注入足够数量且精心设计的虚假用户概貌来产生对恶意用户有利的推荐结果。托攻击的评分信息形式[15]如图 1 所示。

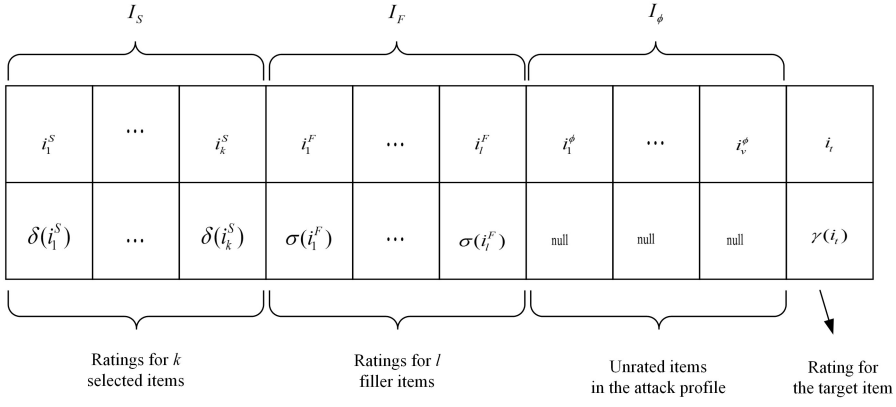


图 1 托攻击的评分信息形式

Fig. 1 Form of shilling attack profile

不法商家首先在服务中注册大量的虚假账户，然后，每个虚假账户给予一个精心挑选的项目子集即选择评分项 I_S 和填充评分项 I_F 特定的评分，使得 I_S 和 I_F 的评分尽可能模仿正常用户的评分从而伪装成正常用户，再通过给予目标项 i_t 偏离真实分数的评分，从而企图成为多个用户的近邻，使目标项被推荐给更多的用户，达到恶意攻击的目的，影响推荐结

果。因此，托攻击检测的目的是检测恶意用户。一些研究表明[16-17]，利用用户项目评级数据的个性化推荐系统很容易受到托攻击的影响。例如：一些不法商家会委托提供专门服务的组织对其产品评最高分或对其竞争对手的产品评最低分，以提升或降低目标产品被系统推荐的频率。

另一种攻击是共同访问注入攻击[7]。在推荐系统中，如果用户访问了两个项目，那么这两个项目就是该用户共同访问的。推荐系统的关键思想是过去经常共同访问的两个项目很可能在未来被共同访问[18-19]，从直观上看，如果两个项目被频繁地共同访问，它们就更相似。具体来说，两个流行的推荐任务是 I2I 推荐和 U2I 推荐，在 I2I 推荐中，当用户访问项目 i 时，系统会显示与项目 i 相似的前 N 个推荐项目。在 U2I 推荐中，系统通过考虑用户的访问历史，向用户推荐排名前 N 的项目。恶意用户利用这一点通过向系统中注入大量的虚假的共同访问以欺骗推荐系统。共同访问注入推广攻击如图 2 所示。

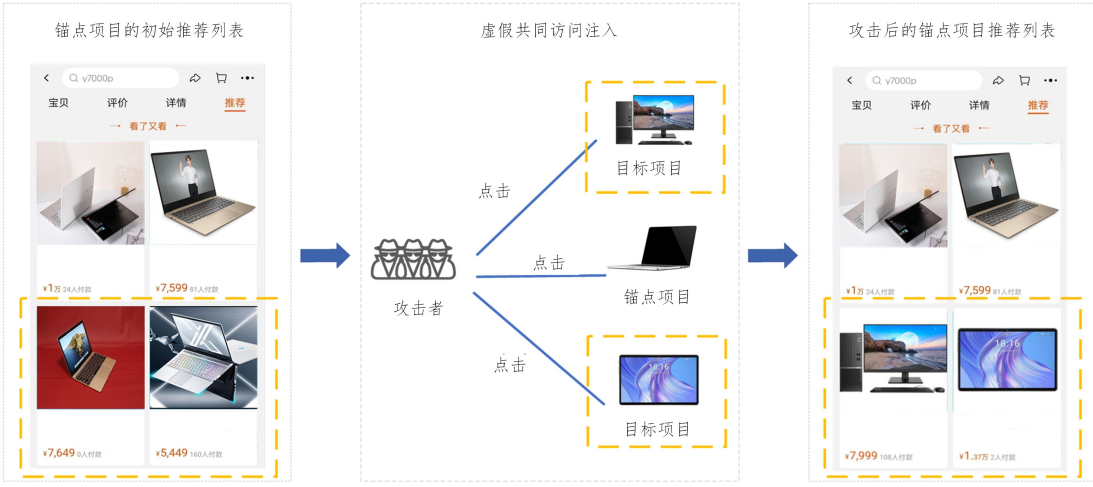


图 2　共同访问注入推广攻击

Fig. 2　Co-visitation injection promotion attack

图 2 展示了共同访问注入攻击的推广攻击，它的目的是将目标项尽可能地推荐给更多的用户，如图所示，经过推广攻击后两个目标商品出现在锚点商品的推荐列表中。不法商家首先从系统中选择一组想要改变其推荐列表的锚点项目（锚点项目通常是爆款商品），然后通过雇佣恶意用户或使用脚本文件共同点击目标项目和锚点项目向系统中注入足够数量的虚假共同访问，使得目标项与锚点项目之间的相似度大于锚点项目与其推荐列表中最后一个项目之间的相似度，从而使目标项目出现在锚点项目的 I2I 推荐列表中，推广目标项目进而影响推荐结果[7]。而降级攻击的目的是使目标项从锚点项目的 Top-N 推荐列表中删除，其通过推广不在锚点项推荐列表中的项目直到目标项不在锚点项目的推荐列表中达到向尽可能少的用户推荐目标商品的目的。共同访问注入攻击检测的目的是检测恶意攻击行为。现实世界中，不法商家通过雇佣一批恶意用户协同点击目标商品和爆款商品，从而建立目标商品与爆款商品之间的关联关系，提升目标商品与爆款商品之间的 I2I 关联分。不法商家通过这种方式诱导用户以爆款的心理预期购买名不符实的商品。

推荐系统恶意攻击不仅损害了消费者的利益，降低用户的购物体验，还严重扰乱了平台的公平性，影响平台和其他商家的信誉。因此，恶意攻击的检测对于保证推荐系统的鲁棒性非常重要。

2.2　推荐系统恶意攻击检测

近年来，出现了一些关于推荐系统恶意攻击检测的研究（尤其是托攻击）。许多先前的检测方法都是基于从评级矩阵中提取的评级行为的表示来设计的，其可以简单地分为基于监督学习的、基于无监督学习的和基于半监督学习的[12]。

基于监督学习的攻击检测被视为分类问题。最初，Burke 等[13]开发了几种评级属性，利用训练好的 K 最近邻分类器检测托攻击。近年来已有学者将深度学习技术应用到推荐系统恶意攻击检测领域。Li 等提出 DegreeSAD[20]方法，从物品受欢迎程度属性提取特征，通过朴素贝叶斯算法检测攻击者。Dou 等[21]提出了一种先令攻击检测模型 CoDetector，联合分解用户-物品交互矩阵和用户-用户共现矩阵，将学习到的包含网络嵌入信息的用户潜在因子作为特征来检测攻击者。Yang 等[22]提出挖掘关联图来识别异常的统一检测框架，以尽可能地处理共同访问注入攻击。Wang 等[23]提出了一种基于图卷积神经网络和目标项识别的群先令攻击检测方法。Zhang 等[24]提出了一个基于 GCN 的用户表示学习框架 GraphRfi，以统一的方式执行鲁棒推荐和攻击检测。GCN 精准地捕获了用户偏好和节点信息，随机森林利用用户表示和评级预测误差很好地进行了攻击检测。基于无监督学习的方法无需训练过程，Metha 等[15]提出了 PCAVarSel 算法，通过低秩矩阵分解预测用户的评分，并对低秩矩阵进行主成分分析，选择出 t 个主成分最小的一批用户判定为恶意用户。Zhang 等[25]研究了一种基于隐马尔可夫模型和分层聚类的方法来揭示攻击。基于半监督学习的方法能够充分利用部分标记数据，实现攻击检测。Wu 等[3]基于常用的统计特征，提出 HySAD 方法，该方法利用极大似然估计参数值，使用类似最大期望算法迭代求解，能够有效地检测出托攻击。Cao 等[26]提出 SemiSAD 方法，首先在少量已标记用户上训练朴素贝叶斯分类器，然后将未标记的用户与 EM-λ 融合，改进初

始的朴素贝叶斯分类器来检测攻击者。

文献[22]表明，现有的检测方法大多都是基于从评级数据中提取的人工构建的特征进行的托攻击检测，评级数据具有丰富的信息来表征用户的基本评级行为。然而，检测性能很大程度上取决于提取的特征的表示。此外，人工构建的特征缺乏普遍性，从而限制了这些方法在真实场景中的应用。

3 问题定义

假设用 P 表示推荐系统中点击行为对应的属性信息集，p_x 为 P 中的一个点击行为样本。为 p_x 设置一个标签，$y_x \in \{0,1\}$，其中 1 代表该条点击行为是推荐系统恶意攻击产生的异常点击行为，0 代表正常点击行为，(p_x, y_x) 是一个训练样本。根据上述定义，n 个训练样本组成了训练数据集，如式(1)所示：

$$D_{\text{train}} = ((p_1, y_1), (p_2, y_2), \cdots, (p_n, y_n)) \tag{1}$$

本文利用数据集 D_{train} 构建模型 f，定义损失函数对模型进行优化，判断当前点击行为 $p_{x'}$ 是否是攻击行为的标签 $y_{x'}$。

4 结合图卷积神经网络和集成方法的推荐系统恶意攻击检测

本章将介绍本文提出的结合图卷积神经网络和集成方法的推荐系统恶意攻击检测方法，其框架图如图 3 所示。其采用 GCN，基于 CNN 的攻击行为分类器和集成方法 Bagging 作为构建模块，因为 GCN 能够充分利用用户-项目交互图中的节点信息和局部结构信息[27]，既能学习节点特征又能学习节点与节点之间的关联关系，捕获深度交互行为信息和用户偏好以及攻击行为蕴含的隐性交互，CNN 能够自动进行特征提取，在分类任务中取得了出色的表现[28]。而 Bagging 算法常与统计分类和回归算法相结合来提高算法的准确率和稳定性[29]，这里我们将其与深度学习模型相结合。

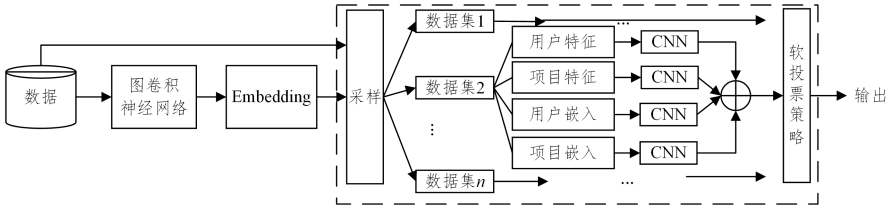
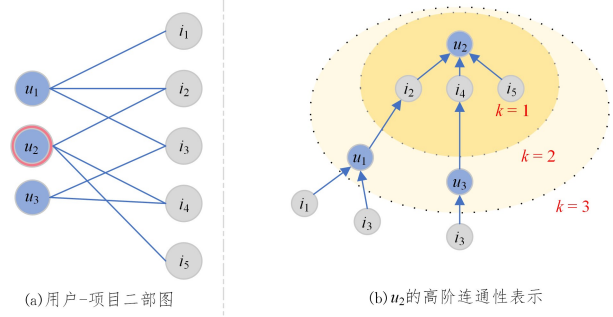


图 3 算法整体框架

Fig. 3 Overall architecture of the proposed algorithm

4.1 基于 GCN 的嵌入表示学习

我们将给定的数据集建模为图，将用户、项目分别视为节点，将用户和项目之间交互关系视为边，由此得到了用户-项目二部图(交互图)，如图 4(a)所示。图 4(b)给出了从 u_2 扩展而来的高阶连通性表示，其包含了交互信息的丰富语义，如行为特征。例如：从二阶交互来看，u_2 和 u_3 都与 i_4 进行过交互，因此 u_2 和 u_3 具有行为相似性；从三阶交互来看，相较于 i_1，u_2 可能对 i_3 更感兴趣，因为有两条路径到 i_3，仅有一条路径到 i_1。由此可见，二部图中包含了丰富的行为信息及用户偏好信息。



(a)用户-项目二部图　(b)u_2的高阶连通性表示

图 4 用户-项目二部图和 u_2 的高阶连通性表示

Fig. 4 Illustration of user-item bipartite graph and high-order connectivity of u_2

为了捕获上述高阶交互信息，受图卷积神经网络[23]和基于图的协同过滤算法[30-32]的启发，我们在二部图上利用交互信息生成用户、项目嵌入表示，并通过多层 GCN 的堆叠，实现高阶交互信息汇聚，产生最终的用户、项目嵌入表示，其架构图如图 5 所示。

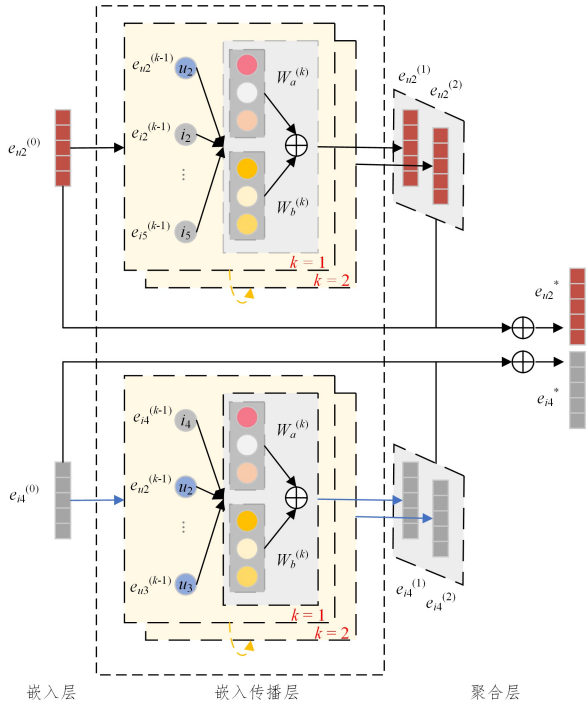


图 5 基于 GCN 的嵌入表示学习框架图

Fig. 5 GCN-based framework for embedding representations learning

4.1.1 嵌入层

这一部分主要初始化用户、项目的嵌入向量表示。我们分别用嵌入向量 $\boldsymbol{e}_u \in \boldsymbol{R}^d$ 和 $\boldsymbol{e}_i \in \boldsymbol{R}^d$ 描述用户 u 和项目 i，其中 d 表示嵌入尺寸。嵌入查找表如式(2)所示，N 和 M 分别表示用户数量和项目数量，每个用户和项目都可以通过该嵌入查找表映射成一个初始化的嵌入表示。

$$\boldsymbol{E}=[\boldsymbol{e}_{u1}^{(0)},\cdots,\boldsymbol{e}_{uN}^{(0)},\boldsymbol{e}_{i1}^{(0)},\cdots,\boldsymbol{e}_{iM}^{(0)}] \tag{2}$$

4.1.2　嵌入传播层

我们堆叠多层 GCN 沿着交互图结构捕获用户-项目多阶交互行为信息，细化用户、项目嵌入。

消息构造：堆叠 k 个嵌入传播层，用户或项目能够接收从其 k 跳邻居传播的信息，在第 k 跳，从 i 到 u 传递的消息定义如式(3)所示：

$$\boldsymbol{m}_{u\leftarrow i}^{(k)}=pui(\boldsymbol{W}_a^{(k)}\boldsymbol{e}_i^{(k-1)}+\boldsymbol{W}_b^{(k)}(\boldsymbol{e}_i^{(k-1)}\odot\boldsymbol{e}_u^{(k-1)})) \tag{3}$$

考虑到节点自身的特征，我们添加节点的自连接，u 的自连接消息如式(4)所示：

$$\boldsymbol{m}_{u\leftarrow u}^{(k)}=\boldsymbol{W}_a^{(k)}\boldsymbol{e}_u^{(k-1)} \tag{4}$$

其中，$\boldsymbol{W}_a^{(k)}$，$\boldsymbol{W}_b^{(k)}$ 是第 k 层可训练权重矩阵，$\boldsymbol{e}_i^{(k-1)}$ 和 $\boldsymbol{e}_u^{(k-1)}$ 是从前面消息传播步骤中生成的项目和用户嵌入表示，储存其 $(k-1)$ 跳邻居的信息。通过元素乘积⊙编码 i 和 u 之间的相互作用，pui 设置为拉普拉斯范数，如式(5)所示，其中 N_u 和 N_i 表示用户 u 和项目 i 的第 k 阶邻居，引入 pui 能够将邻接矩阵进行归一化，防止多层卷积聚合后信息值过大产生的偏差。

$$pui=\frac{1}{\sqrt{|N_u||N_i|}} \tag{5}$$

消息聚合：聚合从 u 的第 k 阶邻居传播的消息以细化 u 的表示，聚合函数定义如式(6)所示：

$$\boldsymbol{e}_u^{(k)}=\mathrm{LeakyReLU}(\boldsymbol{m}_{u\leftarrow u}^{(k)}+\sum_{i\in N_u}\boldsymbol{m}_{u\leftarrow i}^{(k)}) \tag{6}$$

我们采用 LeakyReLU 作为激活函数，增加模型的非线性。通过 k 层 GCN 的堆叠，实现了高阶交互行为信息的嵌入传播，得到了第 k 阶的用户表示 $\boldsymbol{e}_u^{(k)}$。类似地，我们可以得到第 k 阶的用户表示 $\boldsymbol{e}_i^{(k)}$，如式(7)所示：

$$\boldsymbol{e}_i^{(k)}=\mathrm{LeakyReLU}(\boldsymbol{m}_{i\leftarrow i}^{(k)}+\sum_{u\in N_i}\boldsymbol{m}_{i\leftarrow u}^{(k)}) \tag{7}$$

在 k 层传播之后，我们可以得到每一层的用户表示和项目表示即$\{\boldsymbol{e}_u^{(0)},\cdots,\boldsymbol{e}_u^{(k)}\}$和$\{\boldsymbol{e}_i^{(0)},\cdots,\boldsymbol{e}_i^{(k)}\}$。

4.1.3　聚合层

由于不同层的表示强调不同的交互信息，我们将所有层的表示拼接起来，构成融合了多阶细粒度交互信息的最终用户嵌入表示 $\boldsymbol{e}_u^*$ 和项目嵌入表示 $\boldsymbol{e}_{i^*}$，如式(8)、式(9)所示。

$$\boldsymbol{e}_u^*=\boldsymbol{e}_u^{(0)}\|\cdots\|\boldsymbol{e}_u^{(k)} \tag{8}$$

$$\boldsymbol{e}_i^*=\boldsymbol{e}_i^{(0)}\|\cdots\|\boldsymbol{e}_i^{(k)} \tag{9}$$

4.2　基于 CNN 的攻击行为分类器

本节设计了一个基于 CNN 的网络来检测攻击行为，其框架图如图 6 所示。该分类器能够对融合了高阶交互行为信息的用户嵌入和项目嵌入以及用户特征和项目特征进行自动特征提取，捕获局部上下文信息，进一步优化向量表示，提取更深层的行为特征。

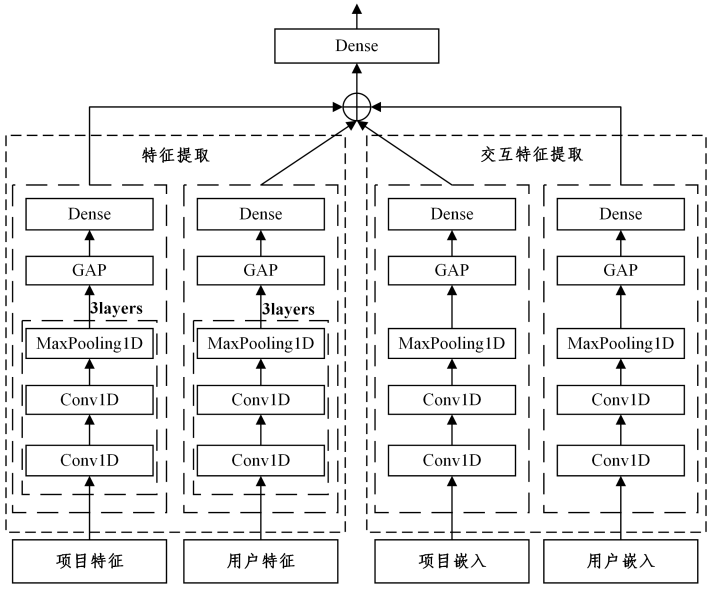


图 6　基于 CNN 的攻击行为分类器框架图

Fig. 6　Framework of CNN-based classifier for malicious attacks

该网络是一个并行网络，分别通过卷积和池化操作捕获细粒度交互行为特征，并将池化结果进行全局平均池化，防止过拟合，通过全连接层得到用户、项目高级特征，然后将高级特征向量进行拼接，通过全连接层采用 Sigmoid 作为激活函数输出对结果的预测值，该激活函数如式(10)所示：

$$S(x)=\frac{1}{1+\mathrm{e}^{-x}} \tag{10}$$

4.3　基于集成方法 Bagging 的模型融合

由于现实世界中攻击行为仅占很小一部分，因此推荐系统恶意攻击行为检测问题是不平衡分类问题。集成方法在面对不平衡分类问题上已经被证明是有效的[33]，集成算法 Bagging 常与其他分类、回归算法相结合，来提高其准确率和稳定性，降低结果的方差。这里我们将其与深度学习模型相结合，以解决不平衡分类问题。

具体方法流程如图 5 所示。这里我们对原训练集正常样本采取不放回抽样，将正常样本分成 N 个子集，且每一个子集样本数量与攻击样本的数量相同。通过重复组合少数类攻击样本和同样数量的多数类正常样本得到若干个新训练集，如图 7 中的数据集 1、数据集 2 等，一个新训练集对应一个基于 CNN 的攻击行为分类器，因为训练集的样本分布不同，训练得到的分类器具有多样化，如图 7 中的分类器 C_1，分类器 C_2 等，最后采取结合策略分类器权重软投票算法将各个分类器的预测值进行整合。

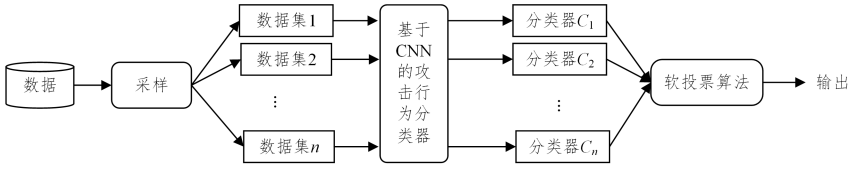


图 7 集成体系架构

Fig. 7 Ensemble architecture

分类器权重软投票算法：为每个分类器 C_i 赋予特定的权重 w_i，T 是分类器的个数，C_i^j 代表第 i 个分类器预测得到的属于类别 j 的概率，$H^j(x)$ 表示 x 属于类别 j 的概率。最终输出如式(11)所示，式(12)是点击行为的预测标签 $\hat{y}$。

$$H^j(x)=\frac{1}{T}\sum_{i=1}^{T}w_iC_i^j(x) \tag{11}$$

$$\hat{y}=\arg\max[H^0(x),H^1(x)] \tag{12}$$

5 实验

5.1 数据集

本文的数据集来源于阿里云天池实验室的真实数据集[1)]。该数据集中恶意用户通过协同点击目标项目和爆款项目实施共同访问注入攻击。一条数据代表一次点击行为所对应的属性信息。原数据集包括每条数据对应的用户 id、项目 id，用户特征、项目特征以及标签等属性信息，表 1 列出了数据集的基本统计信息。

表 1 数据集基本统计

Table 1 Basic statistics of datasets

数据集	正样本数	负样本数	用户数	项目数
训练集	450000	50000	357133	210369
测试集	45000	5000	45899	37964

我们从原始训练集 45 万条数据中随机选取 80%构成训练集，并将剩余的作为验证集，测试集是数据集本身提供的 5 万条数据。

5.2 基线方法

本节将本文提出的算法和以下 5 个基准的托攻击检测算法进行性能比较。

PCAVarSel[15]：无监督学习方法，利用主成分分析计算概貌的主成分系数得分检测攻击。

SemiSAD[26]：半监督学习方法，通过改进朴素贝叶斯分类器来检测攻击者。

DegreeSAD[20]：监督学习方法，利用物品受欢迎程度属性通过朴素贝叶斯算法检测攻击者。

CoDetector[21]：监督学习方法，利用包含网络嵌入信息的用户潜在因子作为特征来检测攻击者。

GraphRfi[24]：监督学习方法，用基于 GCN 的用户表示学习框架以统一的方式执行鲁棒推荐和攻击检测。

5.3 评价指标

为了评估本文方法的性能，我们采用了准确率 ACC(Accuracy)、查准率 P(Precision)、查全率 R(Recall)和 F1(F-Measure) 4 个常用指标，公式如下：

$$ACC=\frac{TP+TN}{TP+FP+FN+TN} \tag{13}$$

$$P=\frac{TP}{TP+FP} \tag{14}$$

$$R=\frac{TP}{TP+FN} \tag{15}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{16}$$

其中，TP 表示判定为恶意攻击行为的集合中真实的恶意攻击行为的比例，FP 表示判定为恶意攻击行为的集合中真实的正常行为的比例，TN 表示判定为正常行为的集合中真实的正常行为的比例，FN 表示判定为正常行为的集合中真实的恶意行为的比例。

5.4 参数设置

在基于 GCN 的嵌入表示学习部分，所有的嵌入大小都固定为 64，我们使用 Xavier 初始化器来初始化用户、商品嵌入表示模块模型参数，并将嵌入传播深度 k 设置为 2，批处理大小设置为 1024，epoch 设置为 400。在基于 CNN 的攻击行为分类器部分，将软投票策略中每个分类器的权重设置为该分类器的准确率 ACC，在实验过程中采用 Adam 优化器，批处理大小设置为 256，epoch 设置为 60，学习率设置为 0.001。

5.5 实验结果

为了验证本文方法的有效性，我们在真实数据集上进行实验，实验结果如表 2 所列。

表 2 多种算法的比较

Table 2 Comparison of multiple methods

方法	攻击行为			正常行为		
	Precision	*Recall*	*F-Measure*	*Precision*	*Recall*	*F-Measure*
PCAVarSel	0.0930	0.0994	0.0961	0.9115	0.9054	0.9085
SemiSAD	0.0712	**0.9873**	0.1328	0.4167	0.0007	0.0014
DegreeSAD	0.5904	0.3410	0.4323	0.9505	0.9816	0.9658
CoDetector	0.5423	0.4898	0.5147	0.9614	0.9685	0.9650
GraphRfi	**0.9966**	0.4008	0.5717	0.6771	**0.9989**	0.8071
Ours	0.7941	0.7158	**0.7529**	**0.9688**	0.9794	**0.9740**

对比后发现：PCAVarSel 算法效果最差，只有当恶意用户之间相同的未评分项比正常用户多时，它才能检测出攻击者，共同访问注入攻击不依赖于评分，因此该条件不成立。SemiSAD 依赖于用户概貌和人工构建的特征，由于数据不平

1) https://tianchi.aliyun.com/dataset/123862

衡问题，恶意用户概貌较少，且其人工构建的特征依赖于评分，效果不佳。DegreeSAD 算法明显优于 PCA，但该算法的检测性能很大程度上取决于人工构建的统计特征的表示质量。相比之下，CoDetector 和 GraphRfi 算法效果要优于以上算法。CoDetector 融合了评级和结构特征信息可以显著提升性能，验证了结构特征的有效性，然而它依然依赖于评级数据。GraphRfi 的攻击检测模块结合 GCN 得到的用户表示和预测误差利用随机森林进行，其效果验证了用户表示学习的有效性，但由于攻击数据仅占数据集的很小一部分，存在严重的不平衡分类问题，导致攻击行为的 Precision 较大，Recall 和 F-Measure 较小，整体性能不佳。

总体上看，本文方法优于基于人工构建特征的 SemiSAD 和 DegreeSAD 方法，验证了通过图学习从用户行为中自动提取融合交互行为和结构信息的多层特征可以弥补人工构建特征区分能力弱、难以提取、需要较高知识成本的不足。另外，本文算法在检测共同访问注入攻击方面优于其他基准方法，这表明流行的托攻击检测方法并不适用于共同访问注入攻击的检测，验证了本文方法在共同访问注入攻击场景中的优越性能。

5.6　超参数分析

本文方法的关键超参数为图卷积嵌入传播深度 k。为了了解这一超参数对实验结果的影响，本文进行了调参实验。保持其他参数固定，设置嵌入传播深度 k 分别为 1，2，3，实验结果如图 8 所示。

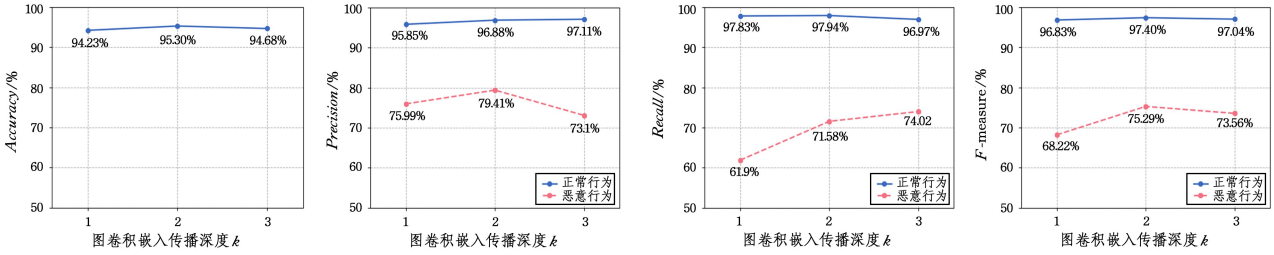


图 8　嵌入传播深度 k 的影响

Fig. 8　Effect of embedding propagation depth k

图 8 显示了不同嵌入传播深度 k 对不同分类指标的影响。显然增加嵌入传播深度可大大提升模型性能，嵌入传播深度 k 为 2 比 k 为 1 在所有方面都都有所改进，嵌入传播深度 k 为 3 在 F-Measure 和 Accuracy 上都比 k 为 1 时有所改进，这说明用户和商品的交互行为信息是由二阶或三阶嵌入传播承载的。但我们发现嵌入传播深度为 3 相比 k 为 2 的评价指标出现了一定程度的下降，我们认为这是太深的架构可能会在表示学习中引入噪声导致过拟合造成的。总的来说，两个嵌入传播层足以捕获用户项目之间的交互行为信息和攻击行为蕴含的隐形交互信息。

5.7　消去分析

为了验证本文方法不同模块对检测性能的影响，我们设计了 4 种模型的变体，对该算法进行消去分析。

(1)去掉基于 GCN 的嵌入表示学习模块(M1)。去除基于 GCN 的嵌入表示学习模块得到的融合了高阶交互行为信息的用户嵌入和项目嵌入，直接将用户特征和商品特征输入基于 CNN 的攻击行为分类器。

(2)去掉用户、项目特征向量(M2)。直接将用户嵌入和项目嵌入输入基于 CNN 的攻击行为分类器。

(3)去掉基于 CNN 的攻击行为分类器(M3)，不进行进一步的特征提取，直接对用户项目特征及其嵌入表示进行二分类。

(4)去掉基于集成方法 Bagging 的模型融合模块(M4)。不对数据集进行采样和模型融合，直接采用基于 CNN 的攻击行为分类器进行二分类。

表 3　消去分析

Table 3　Ablation Study

方法	*Accuracy*	攻击行为			正常行为		
		Precision	*Recall*	*F-Measure*	*Precision*	*Recall*	*F-Measure*
M1	0.9251	0.6178	0.6584	0.6374	0.9618	0.9547	0.9582
M2	0.8847	0.4316	0.4824	0.4556	0.9417	0.9294	0.9355
M3	**0.9929**	0.9823	0.2336	0.3774	0.9215	0.9995	0.9589
M4	0.9240	**0.9878**	0.2426	0.3895	0.9224	**0.9997**	0.9595
Ours	0.9530	0.7941	**0.7158**	**0.7529**	**0.9688**	0.9794	**0.9740**

表 3 列出了消去分析的实验结果，我们可以得到 3 个结论：

(1)移除任何部分，算法的分类评价指标 F-Measure 都会出现一定程度的下降，这说明了算法各元素的有效性。

(2)去掉嵌入表示学习模块，分类指标下降明显，反映了用户项目多阶交互行为信息对攻击行为检测的重要性，体现了通过 GCN 嵌入表示自动学习提取行为特征的优越性能。

(3)去掉模型融合模块，由于存在严重的不平衡分类问题，导致攻击行为的 Precision 较大但 Recall 和 F-Measure 较小，整体性能较差，反映了模型融合能够很好地解决不平衡分类问题，证明了其有效性。

结束语　随着推荐系统的广泛应用，准确识别推荐系统恶意攻击行为是关系到推荐系统可信度的重要课题。考虑到目前大多数的检测方法是基于人工构建的特征和统计方法进行的，因此本文针对人工构建特征的不足提出结合图卷积神经网络和集成方法的推荐系统恶意攻击检测方法。该方法利用 GCN 学习用户项目多阶交互信息，自动提取行为特征和用户偏好集成到用户嵌入和项目嵌入中，结合用户项目本身特征，通过基于 CNN 的攻击行为分类器和 Bagging 算法来检测推荐系统恶意攻击。在真实数据集上进行的实验表明，本

文方法优于目前流行的推荐系统恶意攻击检测算法，对共同访问注入攻击行为有较好的检测效果。

未来的工作中，我们计划利用点击行为被识别为正常行为的概率作为权重，确定点击行为在推荐系统中的贡献，构建一个鲁棒的推荐系统，即使在有推荐系统恶意攻击的情况下也能产生稳定的推荐，这具有重要的现实意义。

参 考 文 献

[1] LUO X, ZHOU M C, XIA Y, et al. Generating highly accurate predictions for missing QoS data via aggregating nonnegative latent factor models[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 27(3): 524-537.

[2] LUO X, ZHOU M C, LI S, et al. An inherently nonnegative latent factor model for high-dimensional and sparse matrices from industrial applications[J]. IEEE Transactions on Industrial Informatics, 2017, 14(5): 2011-2022.

[3] WU Z, WU J, CAO J, et al. HySAD: A semi-supervised hybrid shilling attack detector for trustworthy product recommendation [C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012: 985-993.

[4] GÜNNEMANN S, GÜNNEMANN N, FALOUTSOS C. Detecting anomalies in dynamic rating data: A robust probabilistic model for rating evolution[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 841-850.

[5] GÜNNEMANN N, GÜNNEMANN S, FALOUTSOS C. Robust multivariate autoregression for anomaly detection in dynamic product ratings[C]// Proceedings of the 23rd International Conference on World Wide Web. 2014: 361-372.

[6] LU J, WU D, MAO M, et al. Recommender system application developments: a survey[J]. Decision Support Systems, 2015, 74: 12-32.

[7] YANG G, GONG N Z, CAI Y. Fake Co-visitation Injection Attacks to Recommender Systems[C]// NDSS. 2017.

[8] XING X, MENG W, DOOZAN D, et al. Take This Personally: Pollution Attacks on Personalized Services[C]// USENIX Security Symposium. 2013: 671-686.

[9] CALANDRINO J A, KILZER A, NARAYANAN A, et al. You might also like: Privacy risks of collaborative filtering [C]// 2011 IEEE Symposium on Security and Privacy. IEEE, 2011: 231-246.

[10] FANG M, YANG G, GONG N Z, et al. Poisoning attacks to graph-based recommender systems[C]// Proceedings of the 34th Annual Computer Security Applications Conference. 2018: 381-392.

[11] GUNES I, KALELI C, BILGE A, et al. Shilling attacks against recommender systems: A comprehensive survey [J]. Artificial Intelligence Review, 2014, 42(4): 767-799.

[12] SI M, LI Q. Shilling attacks against collaborative recommender systems: a review[J]. Artificial Intelligence Review, 2020, 53: 291-319.

[13] BURKE R, MOBASHER B, WILLIAMS C, et al. Classification features for attack detection in collaborative recommender systems[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006: 542-547.

[14] YANG Z, CAI Z, GUAN X. Estimating user behavior toward detecting anomalous ratings in rating systems[J]. Knowledge-Based Systems, 2016, 111: 144-158.

[15] MEHTA B, NEJDL W. Unsupervised strategies for shilling detection and robust collaborative filtering[J]. User Modeling and User-Adapted Interaction, 2009, 19: 65-97.

[16] O'MAHONY M, HURLEY N, KUSHMERICK N, et al. Collaborative recommendation: A robustness analysis [J]. ACM Transactions on Internet Technology (TOIT), 2004, 4(4): 344-377.

[17] MOBASHER B, BURKE R, BHAUMIK R, et al. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness[J]. ACM Transactions on Internet Technology (TOIT), 2007, 7(4): 23.

[18] LIU B, KONG D, CEN L, et al. Personalized mobile app recommendation: Reconciling app functionality and user privacy preference[C]// Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015: 315-324.

[19] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems [J]. Computer, 2009, 42(8): 30-37.

[20] LI W, GAO M, LI H, et al. Shilling attack detection in recommender systems via selecting patterns analysis[J]. IEICE Transactions on Information and Systems, 2016, 99(10): 2600-2611.

[21] DOU T, YU J, XIONG Q, et al. Collaborative shilling detection bridging factorization and user embedding [C] // Collaborative Computing: Networking, Applications and Worksharing. Springer International Publishing, 2018: 459-469.

[22] YANG Z, SUN Q, ZHANG Y, et al. Inference of suspicious co-visitation and co-rating behaviors and abnormality forensics for recommender systems [J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 2766-2781.

[23] WANG S, ZHANG P, WANG H, et al. Detecting shilling groups in online recommender systems based on graph convolutional network [J]. Information Processing & Management, 2022, 59(5): 103031.

[24] ZHANG S, YIN H, CHEN T, et al. Gcn-based user representation learning for unifying robust recommendation and fraudster detection[C]// Proceedings of the 43rd international ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 689-698.

[25] ZHANG F, ZHANG Z, ZHANG P, et al. UD-HMM: An unsupervised method for shilling attack detection based on hidden Markov model and hierarchical clustering[J]. Knowledge-Based Systems, 2018, 148: 146-166.

[26] CAO J, WU Z, MAO B, et al. Shilling attack detection utilizing semi-supervised learning method for collaborative recommender system[J]. World Wide Web, 2013, 16: 729-748.

[27] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral

filtering[J/OL]. Advances in Neural Information Processing Systems，2016，29. https://proceedings.neurips.cc/paper_files/paper/2016.

[28] TAN M，LE Q. Efficientnet：Rethinking model scaling for convolutional neural networks[C]// International Conference on Machine Learning. PMLR，2019：6105-6114.

[29] BREIMAN L. Bagging predictors[J]. Machine Learning，1996，24：123-140.

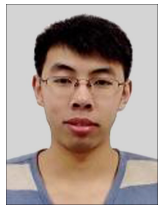
[30] WANG X，HE X，WANG M，et al. Neural graph collaborative filtering[C]// Proceedings of the 42nd international ACM SIGIR Conference on Research and Development in Information Retrieval. 2019：165-174.

[31] HE X，DENG K，WANG X，et al. Lightgcn：Simplifying and powering graph convolution network for recommendation[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020：639-648.

[32] PENG S，SUGIYAMA K，MINE T. Less is More：Reweighting Important Spectral Graph Features for Recommendation[C]// Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022：1273-1282.

[33] LIU X Y，WU J，ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems，Man，and Cybernetics，Part B(Cybernetics)，2008，39(2)：539-550.

LIU Hui，born in 1999，postgraduate，is a member of CCF(No. 19053G). Her main research interests include recommendation system.



JI Ke，born in 1989，Ph. D，associate professor，is a member of CCF(No. 78936M). His research interests include machine learning，recommendation system，etc.