控制与决策
Control and Decision

面向复杂网络的异常检测研究进展

苏江军， 董一鸿， 颜铭江， 钱江波， 辛宇

引用本文:
苏江军, 董一鸿, 颜铭江, 等. 面向复杂网络的异常检测研究进展[J]. 控制与决策, 2021, 36(6): 1293−1310.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0055

面向复杂网络的异常检测研究进展

苏江军, 董一鸿†, 颜铭江, 钱江波, 辛 宇

(宁波大学 信息科学与工程学院，浙江 宁波 315211)

摘 要: 异常检测是指识别数据集中显著区别于其他正常模式的数据,广泛应用于欺诈检测、入侵检测、数据分析等领域. 现有的异常检测研究大多是基于非结构化数据点集,而现实中数据间复杂的结构关系构成了复杂网络,在数学形式上表示为图,所以面向复杂网络的异常检测的需求日益增加. 对此,总结了当前复杂网络异常检测的方法与研究进展: 首先提出复杂网络异常检测的必要性与发展历史; 其次,分别从静态图和动态图的视角将复杂网络异常检测分为基于结构、社区、关系学习的静态图异常检测和基于节点、边、子图、全图的动态图异常检测; 然后,分类别地进行概述、分析与比较,并给出复杂网络异常检测的应用场景; 最后,总结未来面向复杂网络异常检测的研究方向.

关键词: 异常检测； 非结构化； 复杂网络； 图； 静态； 动态

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0055 **开放科学(资源服务)标识码(OSID):**

引用格式: 苏江军,董一鸿,颜铭江,等. 面向复杂网络的异常检测研究进展[J]. 控制与决策, 2021, 36(6): 1293-1310.

Research progress of anomaly detection for complex networks

SU Jiang-jun, DONG Yi-hong†, YAN Ming-jiang, QIAN Jiang-bo, XIN Yu

(Faculty Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China)

Abstract: Anomaly detection is to identify data that is significantly different from other normal patterns in the data set, and is widely applied in fraud detection, intrusion detection, and data analysis and other fields. Existing researches on anomaly detection are mostly based on unstructured data point sets, and there are complex structural relationships between data to form a complex network in the real world, and the network is represented as a graph in mathematical form, so the demand of anomaly detection for complex networks is increasing. This paper summarizes the current methods and research advances of anomaly detection for complex networks. First, the necessity and development history of anomaly detection for complex networks are proposed. Then, from the perspective of static and dynamic graphs, the anomaly detection for complex networks is divided into static graph anomaly detection based on structure, community, relationship learning, and dynamic graph anomaly detection based on nodes, edges, subgraphs, and full graphs, and then summarize, analyze and compare by category, and the application scenarios of anomaly detection for complex networks are given. Finally, the future research directions of anomaly detection for complex networks are summarized.

Keywords: anomaly detection; unstructured; complex networks; graph; static; dynamic

0 引 言

随着信息技术的迅速发展和互联网的普及,网络数据呈“指数型”增长. 世界互联网发展报告[1]中统计截止2019年10月,全球互联网用户已达38. 9亿,其中社交网络Facebook与微信月活跃用户数分别高达24.1亿和11.3亿;移动支付网络如支付宝和微信支付用户数分别达到了12亿和9亿. 现实世界中的这种相互影响、彼此关联的关系都可以用复杂网络[2]的形式呈现,并用图模型来表达,其中节点表示用户,用户之间的联系用边来表示. 然而,现实世界中的大量复杂系统往往会受到外界的攻击,赛门铁克发布的2019年互联网安全威胁报告指出,平均每天拦截1.42亿次网络攻击,网络犯罪已成为时下的一个重大社会问题[3].

异常检测[4]是识别网络攻击的有效策略,目的是寻找数据集中显著区别于其他正常模式的数据. 然

收稿日期: 2020-01-12； **修回日期:** 2020-09-07.

基金项目: 浙江省自然科学基金项目(LY20F020009, LZ20F020001)； 国家自然科学基金项目(61602133)； 宁波市自然科学基金项目(202003N4086, 2019A610093)； 宁波大学“海洋生物技术与海洋工程”学科群专项项目(422004582).

责任编委: 薛建儒.

†通讯作者. E-mail: dongyihong@nbu.edu.cn.

而,由于复杂网络数据规模大、结构复杂和异常检测问题的多样性,复杂网络的异常检测面临以下挑战:

1) 拓扑结构高度复杂:复杂网络是复杂系统的抽象,表现在网络节点数量巨大,结构呈现出多种特征;

2) 动态演化特性:复杂网络随时间发生动态变化,表现在节点或边的产生与消失,无论是Web网络、交通网络还是社交网络,都具有动态演化特性;

3) 属性内容多样性:网络中的节点和边具有丰富的属性信息,如社交网络就是一个富属性复杂网络,用户具有个人信息等属性信息;

4) 缺少标签:有监督学习的准确率往往高于无监督学习,但是用于训练模型所用的标记数据十分难以获取,且人工标注费时费力;

5) 异常与领域高度相关:不同领域间的异常定义不同,有些异常的现象在其他领域可能是正常的情况,所以针对具体领域会有不同的异常定义.

传统的异常检测方法包括基于统计和传统的机器学习方法,该类方法的缺点是内容采集慢、效率低,攻击者可以采取相应的方法绕过检测. 由于攻击者难以掌握整个网络拓扑结构,基于图的异常检测应运而生. 随着图计算、机器学习和深度学习等技术的发展,使用基于图的技术[5]来解决复杂网络的异常检测已逐渐成为国内外研究的热点.

本文对复杂网络中基于图的异常检测方法进行系统分类,介绍其中具有代表性的方法,探讨现有方法的局限性和面临的挑战,指明未来的研究方向. 具体贡献如下:

1) 本文专注于使用基于图的技术来检测复杂网络中的异常值,对静态图和动态图的异常检测所面临的挑战和解决方案进行全面地评述、分析和比较;

2) 拓展复杂网络异常检测的方法,整理归纳了近几年涌现的图嵌入、深度自编码器、图卷积网络(GCN)、生成对抗网络(GAN)等解决复杂网络异常检测的新方法;

3) 总结复杂网络异常检测的应用场景,确定现有的复杂网络异常检测存在的问题与挑战,并提出一些新的生成对抗网络研究方向.

1 复杂网络异常检测研究进展

异常检测广泛应用于各个研究领域,通过识别异常值,研究人员可以获得重要的知识,有助于作出更好的数据决策.

早期的异常检测方法是基于统计模型理论,通过假设正常数据对象满足一种特定的分布或概率模型,然后根据对象是否符合该分布或者模型来判定异常值,许多异常检测方法大都源自该思想. Bremer[6]总结了自1984年以来基于统计数据的异常检测,主要包括高斯混合模型、回归模型、内核密度估计(KDE)等,这类方法特别适用于定量数值数据集,具有易于实现和高效性等特点. 然而这类算法要求事先确定数据分布等相关参数,而多数情况下数据分布很难被准确描述或者估计.

20世纪末期,机器学习中的大量算法开始应用于异常检测[7],包括聚类、ID3决策树、分类、贝叶斯网络和最近邻(KNN)等,极大提高了特定领域异常检测[8]的准确性,降低了基于统计模型方法对数据分布的依赖. 但是,现实网络中数据间存在复杂的拓扑结构关系,如在线社交网络、金融交易网络等,传统的机器学习已经不适用于解决复杂网络这种的异常检测.

图模型将实体抽象为顶点,将关系抽象为图中连接顶点的边,为表示实体间的复杂关系提供了强大的手段,基于图的异常检测技术已逐渐成为研究热点[5,9-12]. 1996年,Staniford-Chen等[12]开始尝试使用基于图的方法来检测网络攻击. Noble等[10]在2003年系统地介绍了两种使用Subdue实现的静态图的异常检测方法,包括异常子结构检测和异常子图检测. Pincombe[13]采用ARMA模型来检测动态网络中的异常值,分别使用10个图距离度量来创建网络变化的时间序列,并建模为ARMA过程,通过设置阈值以及顺序比较相邻时段的图形距离度量来识别异常. Eberle等[11]将图中出现的3种变化定义为异常,分别为标签修改、顶点或边缘插入和顶点或边缘删除,通过使用最小描述长度原则发现正常模式,然后使用不同的异常检测方法来发现特定的异常模式. Akoglu等[5]对基于图的异常检测进行了全面的调查,其中包括当时最先进的方法以及一些开放的研究挑战. 基于图的异常值检测方法至关重要,因为它显示了数据间的相互依赖状态和强有力的表示形式,极大提高了复杂网络的异常检测的效率[14].

近年来,图嵌入等深度学习已广泛应用于多个领域并取得了良好效果. 受此启发,结合深度学习与基于图的异常检测方法,如AnomRank[15]、DevNet[16]和LogAnomaly[17]等,都产生了很好的效果.

2 相关概念

2.1 复杂网络中的异常

数据网络可以表示为图$G=(V,E)$. 其中:V表示顶点的集合,E表示边集合. 静态网络中的异常[5]表示给定一个(普通/属性)图数据库,与大多数图对

象(节点/边/子图等)显著不同的图对象被判定为异常. 动态网络中的异常[18]是指给定一个连续的静态图序列(时间快照网络),寻找特定的时间快照对应于图上显著变化或事件的发生,同时找出影响最大的相关的节点、边或子结构.

在实际情况下,异常会被转化为不同的问题,例如异常用户、交易欺诈和异常事件,所以针对不同的具体网络会有不同的异常定义.

2.2　网络的基本概念

定义1 (egonet[19])　egonet又称自我中心网络,网络节点由唯一的一个中心节点(ego)以及这个节点的邻居(alters)组成,边只包括ego与alter之间,以及alter与alter之间的边,详见图1.

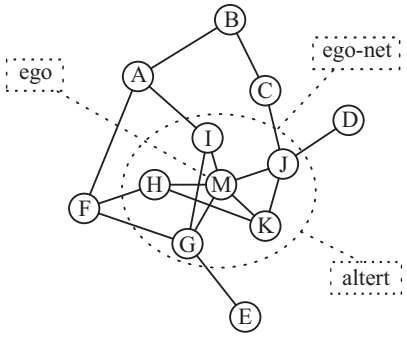


图1　egonet的网络结构

定义2 (属性网络)　属性网络可以表示为$G=(V,E,I)$. 其中:V表示顶点的集合,E表示边数的集合,I表示属性的集合. 在属性图中,顶点、边或两者都可以表征属性信息.

定义3 (时间快照网络)　时间快照网络$G=(G_1,G_2,\ldots,G_n)$表示从时刻1到时刻n的网络状态. 其中:$G_t=\{V^t,E^t,A^t\},1\leqslant t\leqslant n$,$V^t$表示$t$时刻网络中所有节点的集合,$E^t$表示所有边的集合,$A^t$表示节点的属性集合. $e_{ij}^t=1(i,j\in V)$表示在t时刻节点i与节点j之间有边连接,否则$e_{ij}^t=0$. 所以时间快照网络相当于将动态网络拆解成多个静态网络的序列.

3　静态图的异常检测

最初的异常检测是在静态图中展开,利用图的结构或属性信息来查找异常模式以确定异常对象. 根据选择的不同模式分为基于结构、基于社区和基于关系学习的静态图异常检测.

由于静态图不带有时间属性,静态图的异常检测主要面临以下挑战:

1) 网络的规模大:主要表现在复杂网络中的节点和边的数量多,可以用图嵌入方法来对原始网络进行降维特征表示,如Embed[20]、DBMM-1SVM[21].

2) 属性内容复杂:现有的复杂网络的节点或边往往带有富属性信息,可以用属性元路径来游走节点,如CADAHIN[22].

3) 数据噪音的影响:异常与噪音具有一定的相似性,但噪音不是异常值,可以通过噪音数据过滤器预处理原始网络[23].

4) 异常检测准确率低:网络本身的复杂性导致异常检测准确率低. 可以结合多种异常检测方法来检测异常,如GBKD-Forest[24].

5) 异常与领域高度相关:可以定义一些通用的异常,并设计通用的异常检测方法,如ODDBALL[25]和SimRank[26].

3.1　基于结构的异常检测

基于结构的异常检测主要用于识别图结构中罕见的子结构,包括拓扑结构和属性内容方面的子结构. 根据子结构中的关系,基于结构的异常检测可以分为基于特征和基于邻近性的异常检测.

3.1.1　基于特征的异常检测

基于特征的异常检测是指通过特征表示子结构用以在构造的特征空间中进行异常点检测,子结构包括以图为中心和以节点为中心的子结构特征. 这些方法将图数据异常检测问题转化为异常点检测问题,可以应用于多种异常检测场景中.

ODDBALL[25]是在2010年提出的第1个基于egonet检测节点异常的方法,通过egonet总结了几种异常结构以及度量方法. 首先提取每个节点的egonet,然后通过度量方法分别检测3种异常结构. 由于egonet的3种异常结构分别服从幂律分布,通过计算偏离幂律分布的异常得分确定异常egonet中ego. 其中得分函数主要由Out-Line[25]和Out-LOF[27]分别求出,进行归一化后求和. 该检测方法仅适用于加权图中异常检测,其本质是启发式规则,没有明确测试统计量和决策规则.

同属于基于egonet特征的异常检测方法,文献[28]在ODDBALL基础上提出了一种利用egonet度的P值检测异常社区的统计决策规则. 相对于ODDBALL仅能检测异常集团的存在,该方法可以识别组成该集团的节点. 每一个社区egonet度的P值可以通过下式计算:

$$P=\sum_{i=1}^{n}\sum_{j,k:A_{ij}=1,A_{ik}=1}A_{jk}, \tag{1}$$

其中A_{jk}表示第i个egonet的邻接矩阵,且A_{jk}满足

$\{A_{jk}: A_{ij}=1, A_{ik}=1\}$. 异常的社区egonet度的$P$非常低,最后通过两种社区的egonet度的$P$值差异来检测异常社区. 该方法基于统计决策来研究静态网络中的异常,适用于各种静态网络.

相对于基于egonet的方法仅考虑节点的局部特征,GBKD-Forest[24]是一种基于网络全局结构的无监督异常检测方法. 该方法基于图的理论提出了3种类型的结构特征,包括基本图特征、边连接特征和egonet特征. GBKD-Forest通过提取3种类型的结构特征,在Bagging方法内采用随机抽样特征建立KD-Tree以分离异常节点. 与基于距离的邻近度(LOF)算法[27]相比,GBKD-Forest不仅具有最高的精度,AUC(ROC曲线下的面积)也高于其他大多数算法. 同时,算法的分类时间复杂度几乎与节点数呈线性关系,对内存的要求也很低,适用于大规模网络的异常检测.

综上所述,基于特征的异常检测广泛应用于复杂网络异常检测,但因选择的特征不同,导致改进的侧重点与局限性各有差异. 一方面是以图结构为中心的特征[25,28],包括二元组、egonet等;另一方面是以节点为中心的特征[24],包括节点度、中心性度量、边权重等. 结合多种特征可以提高检测准确率,如GBKD-Forest的检测的AUC值高于大多数异常检测算法.

3.1.2 基于邻近性的异常检测

基于邻近性的异常检测又称基于密度的异常检测,利用图结构来量化节点间的亲密度(或接近度),捕获节点之间的简单自相关性关系,其中邻近的节点被认为是同一类(异常或正常). 复杂网络中节点的邻近性(或相似性)度量方法主要包括两大类:基于边结构相似性度量和基于属性内容相似性度量.

SimRank[26]是一种基于图的拓扑结构来衡量静态非属性图中任意两个节点相似度的最经典方法,基于递归的思想,定义节点a与节点b之间的SimRank相似度$s(a,b)$. 现有的基于邻近性异常检测中节点的相似度度量都参照SimRank方法,但是严重依赖图中其他节点的相似度,当复杂网络中的节点数较多时,时间复杂度较高. 此外,SimRank变体P-rank[29]通过将节点的出度和入度关系共同编码为结构相似度计算,从而丰富了SimRank的结构相似度度量.

同属于基于边结构相似性度量的异常检测方法,ASCOS算法[30]是一种非对称结构文本相似性度量. 类似于SimRank,ASCOS递归定义相似性评分,以便可以考虑全局网络结构. ASCOS考虑了两个目标节点之间的所有路径,其相似性得分比SimRank更完整. 为了使ASCOS在计算时间和内存使用方面易于处理,又出现了两种ASCOS变体. 实验结果表明,当目标网络稀疏时,两种变体的异常检测运行时间和计算空间都比直接计算SimRank和ASCOS要小.

SimRank和ASCOS都是基于边结构相似性度量来检测异常的,而属性网络中基于邻近性的异常检测还需要考虑属性内容相似性. CADAHIN[22]是一种发现富属性异质信息网络中节点异常的可约束检测算法框架. 首先,通过将信息丰富的交互数据建模成富属性异质信息网络;其次,提出了带属性元路径、结合边结构和属性内容来计算两个节点u与v的相似度;最后,通过下式计算任意$v\in C$的异常得分:

$$o(v,C,P,W)=\sum_{n=1}^{m} w_n \sum_{u\in C}(1-S_n^p(v,u)). \tag{2}$$

其中:候选集C提供了对异常节点的候选范围的约束,权重向量W提供了对属性元路径重要的约束. CADAHIN综合网络边结构相似性和属性内容相似性来检测富属性异质信息网络中异常节点,得到的结果更加准确且合理. 富属性异质信息网络的复杂性决定了其算法的复杂度较高.

综上所述,基于邻近性的异常检测的关键在于节点间邻近性的度量方法,主要包括属性内容和边结构相似性度量. 而复杂网络中属性相似度的度量主要是通过游走路径来捕获,例如通过属性元路径来捕获节点属性间的相似度. SimRank和ASCOS基于边结构相似性度量来检测异常,但这种度量方法不适用于属性网络. 所以CADAHIN[22]结合了边结构和属性内容两个方面来量化节点间相似性度量,取得更高AUC值的异常检测. 综上,基于邻近性的异常检测关键在于节点间的邻近性度量.

3.1.3 基于特征和基于邻近性的区别与联系

基于结构的异常检测方法包括基于特征和基于邻近性的异常检测方法. 前者利用特征表示的子结构建立一个代表正常行为的模型,与模型不完全匹配的子结构被认为是异常子结构;后者是利用图结构来量化节点间的亲密度(或接近度),捕获这些对象之间的简单自相关性,而那些相差很大的对象被视为异常值. 由于单独使用基于特征或基于邻近度的异常检测方法来检测复杂网络中的异常往往结果欠佳,这激发了人们将基于特征与基于邻近度的方法结合起来解决问题. ODBP[31]是一种集成的异常检测方法,该方法结合了基于特征(COMBN[32])和基于邻近度(LOF[27])的技术,可以有效检测违反特征依赖关系和与对象邻近性的异常. 当两种异常都显著时,ODBP

的异常检测AUC将得到提高;当只能从一个方面捕获异常时,与单一检测方法性能一样好.

3.2 基于社区的异常检测

基于社区的复杂网络异常检测方法依赖于在网络中找到密集连接的节点组或具有跨社区连接的节点、边或者属性,而异常被定义为不直接属于某个特定社区的节点、边或者属性. 基于社区的异常检测主要可以分为基于聚类和基于社区检测的异常检测.

3.2.1 基于社区检测的异常检测

早期基于社区检测的异常检测简单使用社区检测将网络划分为多个社区,在每个社区中根据对象的信息来识别异常值,这种两阶段算法CNA[5]由于仅考虑在社区中的点的信息来检测异常,忽略了节点在整个网络的结构信息,导致检测效率不高.

2010年提出的CODA模型[33]将社区发现和异常检测统一到基于隐马尔可夫随机场的概率公式中. 首先假设所有节点来源于K个社区,异常节点是随机产生的,社区的组成分布包括但不限于高斯分布(连续数据)或多项式分布(文本数据),通过异常对象的分布与均匀分布有很大的差异来检测异常社区. CODA是第1个可以同时分析节点和边连接来识别异常社区的算法,但CODA为每一个节点都分配一个社区,所以用CODA来检测复杂网络中异常会面临维数灾难,带来很高的时间复杂度.

图嵌入是一种将图数据映射为低维稠密向量的过程,能够很好地解决复杂网络带来的维度灾难. Embed模型[20]首先将异常节点定义为将网络不同的社区连接到一起的节点,结合图划分的方法,采用图嵌入进行降维;然后,根据不同节点在特定维度的相似性表示他们在特定集群区域的相似性,从而将图数据异常检测问题转化为所熟悉的异常点检测问题;最后,使用梯度下降法来优化图嵌入. 该方法可以拓展到大规模复杂网络的异常检测,有效解决CODA检测方法带来的维度灾难.

现有的基于社区的异常检测方法通常从局部角度识别网络异常,仅考虑与节点及其一阶邻居相关度量. 为了解决这个问题,CADA方法[34]从全局角度识别异常节点,根据节点连接到不同社区的个数来识别异常节点. 首先使用现有的社区检测算法(Louvain[35]和Infomap[36])将每个节点分配到特定的社区;然后根据节点的邻居节点中属于不同社区的数量为节点分配异常分数;最后设定阈值判断异常节点. 该方法最大的优点就是没有需要学习的参数且具有可拓展性,缺点是会因为复杂网络中的维度灾难而使得异常检测变得困难.

综上所述,基于社区检测的异常检测不是简单将社区检测与异常检测分阶段进行,因为在划分好的社区里检测异常会只考虑节点的局部特征,而忽略了节点的全局特征. 此外,现有复杂网络中的基于社区检测的异常检测都会面临维度灾难,如何解决复杂网络中的高维性是未来的研究方向之一.

3.2.2 基于聚类的异常检测

基于聚类的异常检测首先对复杂网络进行降维处理,得到网络的低维特征向量;然后利用聚类算法将网络的特征向量划分为不同的簇,计算簇内每个对象相对于簇中心的相对距离,并通过可视化来检测出相对距离较大的点. 降维处理与聚类异常检测不是独立进行,而是相互结合使用.

传统的K-means算法需要初始化聚类数K,任意选择初始聚类中心,同时容易受到噪音数据的干扰. Wang等[23]提出了一种改进的K-means算法. 该算法首先使用噪声数据过滤器来预处理数据集以确保在计算初始聚类中心时不包含异常数据点;然后使用基于密度的异常检测方法来识别异常数据点或子集;最后利用聚类时间和精度来评价异常检测效率. 该算法的优点在于能有效处理中小规模的数据集,但无法适用节点数较多的大型复杂网络.

当前大多数基于聚类的异常检测方法都使用评分方案和阈值对异常进行分类,仅适用于具有“已知”集群数量的特定数据集. INCAD[37]是一种基于聚类的异常检测算法,不需要对异常分数或聚类簇数设置阈值,而是基于概率分布进行异常检测和聚类,确保群集的形成不受噪音数据存在的影响,更可靠地定义“正常与异常”行为.

以上基于聚类的异常检测方法过程简单,但一般需要$O(m^2)$的时间,这对于大型数据集而言代价过高. DeepFD算法[38]通过自动编码器将所有的用户节点嵌入到一个潜在空间中,最终使同一欺诈块中可疑用户的表示尽可能接近,而正常用户的表示则均匀分布在剩余的潜在空间中,这样可以准确地检测到欺诈块. 该方法创造性地结合了基于密度和基于聚类的异常检测方法,并且获得了较好的效果.

综上所述,基于聚类的异常检测不是将图聚类与异常检测分开进行,因为异常值本来就会影响图聚类的结果,例如Wang等[23]通过设置噪声数据过滤器来改进聚类结果,然后进行异常检测. 所以基于聚类的异常检测关键在于图聚类和异常检测两者的执行进

程. 此外,将结合基于聚类的异常检测与其他异常检测方法用于检测复杂网络中的异常也是研究的关键,例如DeepFD[38]结合了基于密度和基于聚类两种异常检测方法,具有较高的检测效率.

3.2.3 基于聚类与基于社区检测的区别与联系

社区检测通常是寻找网络中联系稠密的部分,关键在于稠密度的定义,一般通过网络结构特征定义,包括顶点、边、路径和度等特征. 所以,社区检测侧重于网络结构,而忽略了节点的属性.

聚类是指将属于同一类的对象聚在一起,是一种典型的无监督学习方法. 聚类的关键在于计算两个对象间距离,侧重于直接用对象的特征构成的向量来计算距离,没有考虑对象的边和度. 如属性网络聚类侧重找到一堆属性相似的对象,而忽略对象与对象之间的结构联系.

针对复杂网络中的社区检测与聚类的区别是:前者侧重网络结构而忽视属性,主要用于普通网络;后者往往需要先对网络进行降维处理,将得到的特征向量进行聚类以检测异常,主要用于属性网络. 而复杂网络多是带有属性信息的,所以未来将基于聚类和社区检测结合起来用于检测复杂网络中异常是研究的热点之一.

3.3 基于关系学习的异常检测

基于关系学习的异常检测又称基于分类的异常检测,利用对象之间的关系分配类别标签(正常与异常). 常见的基于关系学习的算法[5]包括迭代分类算法(ICA)、吉布斯采样(GS)、循环置信度传播(LBP)、加权表决关系邻分类器(W-VRN). 常见的局部分类器[39]包括朴素贝叶斯、逻辑回归、K-NN、SVM等.

KNN-SVM[40]是一种检测无线传感器网络异常值的方法,利用KNN技术减少训练样本的规模,从而优化SVM训练时间,并通过将数据输入SVM分类核函数来检测异常. KNN-SVM尽管减少了特征空间中的训练时间并实现了更高的检测精度,但是当复杂网络的节点规模较大时,时间复杂度和空间复杂度较高.

虽然支持向量机可以高效地生成决策核函数来检测异常,但无法检测大规模且高维度的复杂网络中的异常值. DBMN-1SVM[21]可检测高维度且大规模复杂网络的异常值,它将深度置信网络(DBN)和SVM组合成混合模型,其中DBN用于将原始的复杂网络表示为低维度的特征数据集,从DBN所学习的特征数据集中训练SVM模型,再将训练出的SVM用于检测异常值. 在混合模型中可以用线性核代替非线性核而不损失准确性,因此具有可伸缩性和计算效率,尤其是检测大规模数据集中的异常值. 但是这种混合模型的特征提取是分开进行的,导致模型准确率不高.

现有基于关系学习的异常检测多是基于混合模型,使用自编码器等深度神经网络模型进行特征表示学习,降低复杂网络维度,然后在低维度的特征数据集中执行传统的分类方法以进行异常检测,但是这种两步混合模型的异常检测准确率不高. OC-NN[41]是一类神经网络模型用来检测复杂网络中异常的集成式方法,结合了深度网络的特征表示以及SVM的分类检测. 通过优化异常检测定制的目标函数学习隐层的特征表示,获得的超平面将所有正常与异常数据点分类. OC-NN的优势在于隐藏层的特征是针对异常检测的特定任务而构建的,与最近提出的使用深度学习特征表示作为异常检测器输入的混合方法大不相同.

综上所述,基于关系学习的异常检测主要是通过分类方法来检测异常,目前常采用深度学习的方法学习网络的低维特征表示,在特征表示的数据集上执行异常检测,如DBN-1SVM[21]. 由于异常和噪声数据会影响特征表示学习,这种分阶段的混合学习方法无法实现较高的检测准确率. OC-NN[41]创造性地使用SVM(如损失函数)来训练神经网络,以优化特征表示学习过程,利用这样的特征数据集进行异常检测能收获更高的准确率.

3.4 静态图的异常检测方法分类总结

基于结构的方法是静态图异常检测中最常用的异常检测方法,它往往从图的结构出发,所以需要图的算法探索采用何种方法检测异常,主要适用于简单图,无法适用于大规模富属性的网络. 基于社区的异常检测方法的时间复杂度介于基于结构和基于关系学习之间,可以定义基于社区的新异常,但它的难点在于社区检测(或聚类)与网络的降维处理. 基于关系学习的异常检测方法可以检测大规模富属性复杂网络的异常,但是需要采用图嵌入的方法来表示网络的低维向量,而针对不同类型的网络选择合适的图嵌入技术本身就是一个难题.

静态图的异常检测方法的比较总结如表1所示,不同的复杂网络的异常值的定义和检测方法不同,应根据复杂网络的具体应用场景以及侧重的特征选取合适的异常检测方法.

表 1 静态图的异常检测方法总结

方法	分类	异常值判断	解决问题	结论
文献[25]	特征	提取egonet的特征	通用的异常检测方法	总结了几种异常结构与度量
文献[28]				明确了检测统计量和决策规则
文献[24]		图结构、边和ego特征异常	提高检测的准确率、降低时间复杂度	检测AUC值高、时间复杂度低
文献[26]	邻近性	边结构的相似度得分	通用的异常检测方法	递归定义网络节点间相似度
文献[30]			降低时间复杂度	优化节点间相似度得分、时间复杂度低
文献[22]		结合属性内容和边结构的相似度得分	属性网络	通过属性元路径捕获属性网络中节点相似度
文献[31]	特征 邻近性	结合特征(COMBN)和密度(LOF)	提高检测的准确率	基于特征和邻近性的集成方法,算法鲁棒性高
文献[33]	社区检测	结合ICM和EM算法获得社区的标签	提高检测的准确率	同时执行社区检测与异常检测的集成方法
文献[20]		将社区连接在一起的节点为异常节点	网络的大规模	定义了新的异常节点且可检测大规模复杂网络的异常值
文献[34]			通用的异常检测方法	无参数且具有拓展性
文献[23]	聚类	计算簇内点到聚类中心的相对距离	噪音影响异常检测	利用噪声数据过滤器预处理数据噪声
文献[37]				同时执行聚类和异常检测的集成方法
文献[38]	聚类 邻近性	聚类后通过LOF检测异常值	网络的大规模	结合了基于聚类和邻近性的集成方法可用于大规模复杂网络
文献[40]	关系学习	SVM分类核函数判断异常	提高检测的准确率	结合KNN与SVM来检测异常
文献[21]			网络的大规模	结合DBN和SVM来检测大规模复杂网络中的异常值
文献[41]			通用的异常检测方法 提高检测准确率	使用一类SVM来训练神经网络以优化特征表示学习过程

4 动态图的异常检测

现实世界中网络是不断变化的,所以动态图的异常检测要求实时检测动态图中的异常行为或者发生异常的时间点,其中异常行为包括异常顶点、边、子图或全图的特征. 除了静态图异常检测带来的挑战外,动态图的异常检测还面临时间属性带来的以下几个挑战:

1) 实时性的要求. 由于动态网络是连续变化的时间序列图,这就要求检测必须快速高效,这也是与静态图异常检测最大的不同. 可以采用自编码、循环神经网络(RNN)和GCN等深度学习模型刻画时间快照网络的低维特征表示,如OCAN[42]、SedanSpot[43]、Spotlight[44].

2) 异常类型的多样性. 动态网络中异常类型具有多样性,包括节点、边、子图、变更和事件,针对具体异常类型采用对应的异常检测方法.

3) 计算复杂度的增加. 随着时间快照的增加,异常检测的时间复杂度也会增加,这是动态异常检测特有的挑战. 可以使用时间复杂度较小的算法来检测异常,如AnomRank[15]和GaHroei[45]等.

4) 数据标签的缺失. 带有标记的异常数据难以获取,人工标记费时费力. 可以利用生成模型生成异常数据,如OCAN[42].

5) 异常的可解释性. 针对具体问题中异常的定义是不同的,可以根据异常的定义设计异常检测方法,如LRGCN-SAPE[46].

根据所选最小单位的差异,将动态图的异常检测分为基于节点、基于边、基于子图和基于全图的动态图异常检测.

4.1 基于节点的动态图异常检测

基于节点的异常检测中的最小单位是单个节点或一组节点[47]. 在每个时间步骤中,提取节点中的特征作为摘要度量,常见的摘要度量包括单个节点的度和egonet网络密度,与其他大多数节点相比表现出不规则行为的节点被检测为异常,从而识别确定顶点的异常时间点. 针对不同的问题领域使用不同的特征

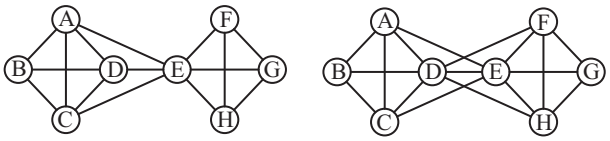
提取方法,但大多数的方法都提供节点异常评分函数来概括每个节点的行为[48].

节点异常评分函数定义为$f:V\to R$. 其中:V表示节点,R为节点异常得分. 根据异常得分函数计算每一个正常节点的异常得分,并且求得均值为$\hat{f}$,则所有异常节点集$V'\subset V$可以被定义为

$$\forall v'\subset V',\ |f(v')-\hat{f}|>c_0, \tag{3}$$

其中c_0是根据正常节点行为计算得出的可接受偏差.

如图2所示,假定节点ABCD组成社区1,节点EFGH组成社区2. t时刻,节点D仅属于社区1;但在$t+1$时刻,节点D既属于社区1,又属于社区2,且无其他顶点有社区演化,所以节点D在$t+1$时刻被判定为异常.



(a) t时刻的网络 (b) t+1时刻的网络

图2 不同时刻网络中节点演变

基于节点的异常检测是最常见的动态图异常检测算法之一. DBMM[49]是一种动态行为混合模型,主要用于检测大规模、动态属性网络中的异常节点. 该方法使用基于特征表示来学习异常节点的特征并推广到未观察到的顶点. DBMM专注于时间异常,在每个时间步长执行节点的异常特征提取,使用正则化计算异常节点的数量,并使用过渡矩阵来计算节点在即将到来的时间步中更改其异常特征的可能性. DBMM在每个时间步中只能检测到异常节点,缺乏更一般的结果,而且该方法的检测精度不高.

为了提高检测精度并避免人工重建特征,M-LSTM方法[50]采用了一种基于多源长短存储网络来检测Wikipedia中的恶意用户. M-LSTM能够捕获用户编辑行为的不同方面,并将每个用户映射到相同的低维嵌入空间. 该方法在包含正常数据(良性用户)和异常数据(恶意用户)的训练数据集上进行训练,不需要启发式规则. 实际上,在收集到的训练数据中恶意用户记录甚少,而且手动为大量恶意用户添加标签很繁琐.

为了解决M-LSTM的训练数据集中缺少带标签的异常数据的问题,采用生成模型的基本思想,在仅给定的正常数据情况下生成异常数据. OCAN[42]是一种用于解决欺诈检测的方法,其训练过程包括两个阶段. 该模型首先采用LSTM-Autoencoder[51]将正常数据根据其在线活动编码到隐藏空间,称为良性用户表示;然后,OCAN训练改进的生成式对抗网络,其中鉴别器被训练成一个分类器,用于区分良性用户和恶意用户,生成器生成潜在的恶意用户;最后经过训练,鉴别器能够检测出恶意用户. 该模型最大的创新就是可以自适应更新用户的表示和动态地预测欺诈用户.

综上所述,基于节点的动态图异常检测旨在找到顶点的子集度量,刻画节点在单个时间快照网络的特征表示. 为了实时检测复杂网络中动态变化的节点,可以用LSTM、RNN、CNN、DNN等深度学习模型来刻画节点的动态演化特性. 如何高效地学习节点的特征表示以及采用哪种深度学习模型来刻画网络的动态演化的特征表示,是基于动态图的节点异常检测的关键,例如OCAN模型[42]采用自编码器来表示节点的特征和LSTM模型来刻画节点的动态演化特性,从而可以实时检测用户欺诈.

4.2 基于边的动态图异常检测

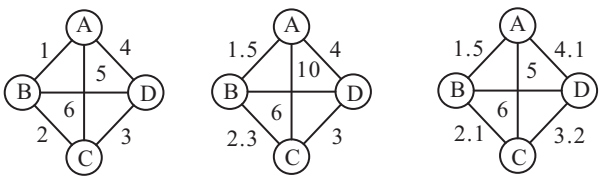
基于边的异常检测中的最小单位为单个边或边的集合[52]. 通常异常边与网络中大多数边相比能够显示出异常演化趋势,可以根据节点之间的边权重演化或边的添加/删除来检测异常的边[53]. 采用评分函数概括图中每条边,从而识别并确定异常边的时间点. 动态图中基于边的异常检测包括两种主要类型:1) 异常边权重随时间波动且出现不连续的峰值;2) 不连接或不属于同一社区中的两个顶点突然出现边的连接.

边异常评分函数定义为$f:E\to R$. 其中:E表示边,R表示边的异常得分. 根据异常得分函数计算每一个正常边的异常得分,求得均值为$\hat{f}$,所有异常边集$E'\subset E$可以被定义为

$$\forall e'\subset E',\ |f(e')-\hat{f}|>c_0, \tag{4}$$

其中c_0是根据正常边行为计算得出的可接受偏差.

如图3所示,可以观察到图中大多数节点的权重变化范围在±0.5波动,但是在$t+1$时刻,连接顶点A与顶点C的边权重波动为±5,所以可以判定边AC在$t+1$时刻为异常边.



(a) t时刻网络 (b) t+1时刻网络 (c) t+2时刻网络

图3 不同时刻网络中边的权重演变

IcLEOD[54]是一种从局部角度检测动态加权图异常的方法. 该方法主要分为两个阶段: 第1阶段为单个对象精心设计本地邻域子图, 该子图包含节点的结构、边的权值及最近邻信息; 第2阶段通过分析和比较不同快照的本地邻域子图的度量来发现异常对象. IcLEOD算法对于演化网络中的异常边检测非常有效, 时间复杂度低, 其最大创新点在于通过基于节点和基于边两个方面来刻画单个时间快照网络的度量, 但是无法实时检测动态网络中的异常.

为了实现实时检测动态网络中的异常边, SedanSpot方法[43]检测图中异常边缘的两个明显迹象: 在活动爆发时发生和连接图中稀疏连接的部分, 并使用亚线性存储器来实时检测异常边缘. 该算法利用这些在亚线性内存中的观察结果进行以下操作: 1) 执行速率调整抽样, 对突发时间内的边缘进行下采样; 2) 使用基于整体随机漫步的边缘异常评分函数, 将传入的边缘与整个(采样的)图进行比较, 检测连接图中稀疏连接部分的边缘.

与SedanSpot检测图中稀疏连接部分的边缘不同, LRGCN-SAPE方法[46]是研究时间演化图中路径分类问题的. 为了捕获时间依赖性和图结构演化, 设计了一个新的神经网络LRGCN, 该神经网络将图快照中的节点相关性视为内部时间关系, 并将相邻图快照之间的时间相关性视为时间关系, 然后联合建模两种关系. 此外, 还提出了一种名为SAPE的新路径表示方法, 可以将任意长度的路径嵌入到固定长度的向量中, 用作分类的标准输入格式. 该模型首先使用两层长期短期记忆LRGCN-SAPE捕获图结构动态性和时间依赖性, 从而获取每个节点的隐藏表示; 然后利用一种自注意机制学习节点的重要性, 并将其编码为统一的路径表示形式; 再将路径表示与完全连接层进行级联; 最后将学习到的路径表示用来预测故障路径, 并且在真实的数据集中得到了较好的效果.

综上所述, 与基于节点的方法类似, 基于边的动态图异常检测旨在找到边的子集度量, 刻画单个时间快照网络的特征表示, 例如边的权重分布与属性值等. 为了能实时检测异常, 可以使用RGCN等深度学习模型. 此外, IcLEOD[54]创新性地结合节点与边来检测动态网络中的异常值, 为更好地度量网络的特征表示以检测动态网络中的异常提供了一个很好的研究方向.

4.3 基于子图的动态图异常检测

基于子图异常检测的最小单位为单个子图或子图的集合[55], 比如社区是一种特殊的子图, 可以采用社区检测来获得子图. 通常异常子图的图结构与其他正常子图的图结构显著不同[56], 当获得了一组时间步长子图时, 就可以根据相邻时间步长子图分配的异常分数确定异常子图或异常时间点. 常见的图结构得分包括平均聚类系数和平均节点度等, 并且这类异常对象的范围包括异常节点和边等特征.

子图异常评分函数定义为$f: G_s \to R$. 其中: G_s表示子图, R表示子图的异常得分. 根据异常得分函数计算每一个正常子图的异常得分, 并且求得均值为$\hat{f}$, 则所有异常子图集$G'_s \subset G_s$可以被定义为

$$\forall g'_s \subset G'_s,\ |f(g') - \hat{f}| > c_0, \tag{5}$$

其中c_0是根据正常子图行为计算得出的可接受偏差.

Chen等[57]讨论了图的社区结构中可能发生的6个基本变化, 包括社区扩张、社区缩减、社区合并、社区分离、社区出现和社区消失.

为了解决动态复杂网络中实时检测异常的关键, SpotLight方法[44]将异常定义为突然出现(或消失)的大型密集有向子图, 它为每个图G提取一个K维SpotLight草图$v(G)$, "异常"子图与草图空间中的"正常"子图相距很远, 利用草图空间中的距离间隙来检测产生异常草图的图为异常图. SpotLight的准确率和召回率都优于其他方法, 但是当原始网络的规模较大时, 算法效率也会随之下降.

大规模动态图的异常检测问题一直是研究的热点问题之一. DPADS方法[58]将静态图的异常检测算法GBAD和并行异常检测算法PLAD拓展到大规模动态图的异常检测中, 通过使用时间滑动窗口对图进行划分, 在初始化阶段选取N个子图, 使用最小描述长度(MDL)原理并行检测正常模式和异常模式, 并行迭代地检测其他子图中的正常结构和异常结构.

综上所述, 查找具有不规则行为的子图需要一种不同于异常顶点或边的方法, 由于无法枚举一个复杂网络中的所有子图, 通常将被跟踪或标识的子图限制为通过社区检测方法找到的子图. Spotlight[44]通过草图来刻画网络的特征度量. 文献[59]从基于节点和社区两个角度来检测动态图中的异常值. 基于向量自回归的以节点为中心的模型分析动态图中的节点行为; 通过跟踪社区的结构变化和动态减少假阳性结果的数量, 开发了两个以社区为中心的模型. 这种结合多种异常检测方法来检测动态图中的异常值是研究方向之一.

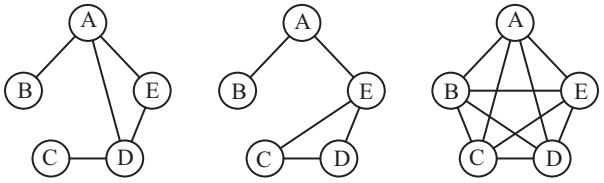
4.4 基于全图的动态图异常检测

基于全图的异常检测中的检测单位为整个图.

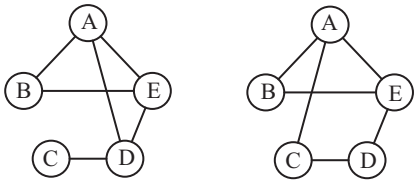
当网络中大多数实体更改其正常的关系模式时,将引起整体图的改变,这种异常类型包括变化和事件[60]. 全图得分通过平均聚类系数和平均节点度定义,测量动态网络中每个时间戳的图的得分来检测异常时间点并判定异常为变更还是事件[48].

4.4.1 基于事件的检测

动态网络中的事件检测任务仅识别异常时间戳[60],不提供异常行为产生原因. 全图的异常评分函数可以通过比较图中顶点和边的数量进行定义. 在图4中,通过$f(G_t)=|E_t|$计算每个时刻图中的边,求得每个时刻图中的边数分别为5、5、10、6、6. 在t时刻,图的结构发生了重大变化,几乎形成了全连接图,且此时间点图结构是孤立的,表示在t时刻检测到了事件.



(a) $t-2$时刻 (b) $t-1$时刻 (c) t时刻



(d) $t+1$时刻 (e) $t+2$时刻

图4 不同时刻网络中事件检测

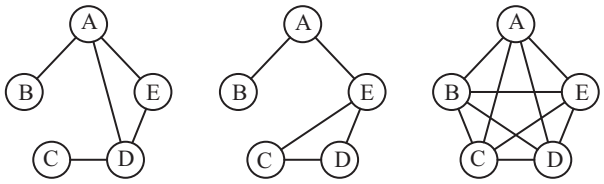
定义4 (事件的检测) 给定图G和图的得分函数$f:G_t\to R$. 其中:G_t表示t时刻网络,R表示t时刻网络的异常得分. 当$|f(G_t)-f(G_{t-1})|>c_0$并且$|f(G_t)-f(G_{t+1})|>c_0$时,在t时刻检测到事件.

异常事件的检测是动态网络异常检测的重要的应用价值体现. AnomRank[15]是一种动态图中快速而准确的异常检测算法. 首先将动态图中的异常分为AnomalyS(结构)和AnomalyW(权重)两种类型,分别根据这两种异常类型的特征设计两种节点异常评分函数,最后将节点得分的一阶和二阶导数作为两种度量异常的标准. 该方法主要用于检测动态图中的异常事件,其性能优于最先进的方法,速度比SedanSpot[43]快49.5倍,精度比DenseAlert[61]高35 %.

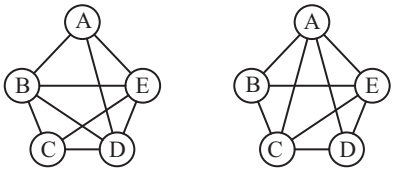
4.4.2 基于变更的检测

动态网络中的另一种异常检测是变更检测[62],它是事件检测的补充. 两种检测的区别是:通过比较相邻时刻边的数量来识别此更改,在t时刻许多新的边添加到图中形成了一个全连接图. 但是在$t+1$时刻,图没有恢复为原始结构,并且图具有新的“正常”行为,新结构的持久性表明在t时刻检测到的是变更,而不是事件.

定义5 (变更的检测) 给定图G和图的得分函数$f:G_t\to R$. 其中:G_t表示t时刻网络,R表示t时刻网络的异常得分. 当$|f(G_t)-f(G_{t-1})|>c_0$并且$|f(G_t)-f(G_{t+1})|\leqslant c_0$时,在$t$时刻检测到变更.



(a) $t-2$时刻 (b) $t-1$时刻 (c) t时刻



(d) $t+1$时刻 (e) $t+2$时刻

图5 不同时刻网络中变更检测

Gahrooei等[45]使用基于概率模型的方法检测网络全局结构变化. 他们使用广义线性模型(GLM)对每个静态属性图建模,假定每个边都是根据指数族(例如伯努利、泊松、正态、伽玛等)中的概率分布生成,有助于对一系列边缘属性进行建模;通过在GLM的参数上构建状态空间模型,使用扩展的卡尔曼滤波器(EKF)[55]来估计和更新随时间变化的参数,捕获图的时间序列之间的依赖性;在$t-1$处估计的参数用于预测在t处的图并且计算皮尔逊残差;使用指数加权移动平均值(EWMA)控制图随时间监控这些残差以检测图流中的全局结构变化.这种非参数概率方法具有从数据本身定义模型结构的灵活性.

综上所述,基于全图动态图异常检测主要包括基于事件和基于变更的检测. 两者最大的区别在于,基于变更的异常时间点后动态图会保持异常的状态为新的正常状态. 基于全图的异常检测最常用的方法是用网络表示学习的方法来学习整个图的特征,将动态图的特征表示构建一个时间序列模型,从而动态检测异常. 例如Pincombe[13]提出了基于ARMA过程的时间序列异常检测技术,该方法使用带权动态图模型来表示通信网络,使用最大公共子图MCS的节点、边和权重等距离特征. 但是该方法只能捕获两个连续时间戳之间的相关性,而不能捕获长期依赖关系. 这也是为什么现在的动态图的异常检测通过LSTM、

GCN和RNN等深度学习模型来刻画动态演化网络的原因.

4.5 动态图的异常检测方法分类总结

基于节点的方法是动态图异常检测中最常用的方法. 以节点为中心的特征刻画网络结构有很多种方法,如节点的度、中心性度量、平均节点度和属性内容等. 这种方法过程简单,当时间快照的长度增加时,该方法的时间复杂度会非常大,无法适用于大规模复杂动态网络.

与基于节点的方法相反,基于边的方法是动态图异常检测中最少见的.因为仅通过边的权重信息来表示网络的特征是不全面的,所以会结合其他异常检测方法来检测动态图中的异常值. 例如IcLEOD[54]结合节点与边来检测动态网络中的异常值,同样,该方法的时间复杂度较高.

相对于前两类异常检测方法,基于子图和基于全图的异常检测方法的时间复杂度较低,因为这两种方法检测的最小单位个数相对更少. 基于子图的异常检测方法通过社区检测限制子图的个数,需要借助于节点和边表示网络的特征,可以基于子图的角度定义一种新的异常,例如Chen[57]定义了6种新的社区结构变化. 基于全图的异常根据发生异常后是否维持现状分为基于变更和基于事件两种角度来检测,该方法也会结合其他异常检测方法来检测异常值,具有较高的检测准确率和较低的时间复杂度.

动态图的异常检测方法分类比较总结见表2. 不同类异常检测方法都有其优缺点,应根据具体应用场景下的复杂网络和待检测的异常类型来选取最合适的异常检测方法,而且需要考虑实时检测网络中的异常值,因为这是动态图异常检测的关键所在.

表 2　动态图的异常检测方法总结

方法	分类	异常值判断	解决问题	结论
文献[49]	节点	用时间快照下节点的特征来判定异常节点	网络的大规模和富属性	检测大规模、动态属性网络中的异常值
文献[50]			实时检测异常	自动学习用户嵌入,且用户嵌入结果可用于各种数据挖掘任务
文献[42]		训练鉴别器对正常用户和异常用户进行分类	缺失数据标签、实时检测异常	自适应更新用户表示和动态检测异常
文献[43]	边	利用异常边缘的两个明显现象来发现异常	实时检测异常	实时检测动态网络中异常边
文献[54]	边、节点	用时间快照下Corenet的度量来发现异常	降低时间复杂度	结合节点与边来检测动态网络中的异常值
文献[46]		利用SAPE学习路径表示实现预测故障路径	异常的可解释性	可以捕获时间依赖性和图结构动力学来联合建模
文献[57]	子图	定义了6种异常社区结构来判断异常社区	异常的可解释性	总结了图中社区的6个变化
文献[44]		用正常与异常的草图空间距离来检测异常子图	实时检测异常	实时检测动态图中的异常
文献[58]		MDL找正常子图模式,滑动窗口检测异常子图	网络的大规模	能检测大规模动态图的异常
文献[59]	子图、节点	向量自回归模型和社区检测确定异常社区	异常的可解释性	结合节点和子图来检测异常
文献[15]	事件	使用节点得分函数的一阶和二阶导数作为度量来检测事件	实时检测异常、降低时间复杂度通用的异常检测方法	实时检测动态异常、且准确率高时间复杂度低和可拓展性
文献[45]	变更	使用全局结构的概率模型来检测变更	属性网络、降低时间复杂度	快速准确检测动态属性网络中异常值
文献[13]	节点、边社区	将网络的特征结合到ARMA模型来检测异常变更	实时检测异常	结合节点、边、社区检测动态图的异常事件

5 分析与比较

复杂网络的异常检测由于其广泛应用于各个领域,引发了新的研究热潮,表3从方法分类、异常类型、主要技术和特点等多方面对近年来复杂网络异常检测方法进行分析比较.

1) 异常类型包括静态图中的结构、社区异常和动态图中的节点、边、子图和全图异常. 属性网络的异常除了结构异常外,还有属性异常以及结构和属性形成的组合异常,丰富的属性信息可以辅助判断其他异常类型. 动态网络最大的特点是自带时间属性,现有的大多数动态网络异常检测都是把网络当作时间快照网络处理,这样针对每一个时间戳网络都可以看

成一个静态网络,因此,可以选择静态网络中的异常类型作为最小单位. 此外,表3中统计了具体网络中的异常类型,包括交易网络、传感器网络、城市时空网络、产品评论网络、时间演化网络、光谱图等. 静态图异常类型最多的是节点与结构异常,动态图异常类型最多的是时间序列异常.

2) 目前,复杂网络的异常检测方法采用的主要技术包括最近邻、矩阵分解、聚类、贝叶斯网络等传统的机器学习,以及图嵌入、LSTM、自编码器、GAN、随机森林、偏差神经网络、GCN和强化学习等深度学习策略. 由此可见,机器学习和深度学习已经广泛应用于复杂网络的异常检测. 特别是近几年来,基于深度学习的网络异常检测有了很大的进展,通过特征表示数据和定义异常分数检测异常. 常见的数据特征表示模型有自编码器、GAN和GCN等,而定义异常分数的方法有基于距离和基于重构误差等两种方法. 由于表示学习和异常检测方法是分开的,可能会产生次优的异常检测. 最近ONE[63]是第一个将特征表示与异常分数学习结合的集成方法,得到了较好的结果. 由于数据集的标签难以获得,现有的复杂网络异常检测大多采用无监督学习,如表3中基于结构、关系学习的静态图异常检测和基于全图的动态图异常检测,但这类方法往往会将数据噪声检测为异常,准确率较低. 而半监督或者有监督复杂网络的异常检测技术由于学习到了异常数据的特征,具有更高的准确率,如表3中基于边的动态图异常检测.

表 3 主要异常检测方法综合比较

方法分类	方法	异常类型	主要技术	特点
基于结构的静态图异常检测	DeepFD[38]	二分图中结构异常	无监督、自编码器、图嵌入、基于密度	深度网络嵌入、可解释性、鲁棒性
	Egonet[28]	静态网络中子图异常	无监督、egonet度、设定阈值	检测异常子图和识别异常子图的节点
	CADAHIN[22]	异质信息网络节点异常	无监督、属性元路径,综合结构与属性评估异常节点	结合网络结构和内容属性来评估节点的异常
基于社区的静态图异常检测	CADA[34]	静态网络中节点异常	无监督、社区检测和余弦相似度	可用于大规模网络,且模型无参数、高效、通用性好
	DevNet[16]	静态属性网络异常检测	半监督、先验知识和神经偏差网络	将异常评分结合到图嵌入、端到端的学习
基于关系学习的静态图异常检测	doOCSVM[64]	无线传感器网络中样本异常	无监督、随机近似函数、OCSVM[65]	可以处理流数据、分布式运行
	DTM[66]	传感器网络污染检测	无监督、最近邻	基于正态分布判定异常
	文献[67]	银行交易网络节点异常	无监督、图嵌入和聚类	结合图嵌入与异常检测
基于节点的动态图异常检测	ONE[63]	属性图中结构与属性异常	无监督、图嵌入	结构和属性图嵌入
	LOF[68]	城市时空异常	无监督、分解、嵌入和时空特征神经网络	图嵌入、神经网络
	OCAN[42]	时间序列异常检测	半监督、自编码器、LSTM、GAN	动态更新模型和动态预测欺诈用户
基于边的动态图异常检测	PIDForest[69]	大规模异构属性边异常	半监督、随机森林	具有可解释性
	LRGCN-SAPE[46]	时间演化图异常检测	半监督、LSTM、GCN和Attention机制	预测异常且具有可解释性
	IRL[70]	时间序列图异常检测	半监督、强化学习和贝叶斯公式	引入强化学习且能实时检测异常
基于子图的动态图异常检测	DeFauder[71]	产品评论图异常检测	无监督、图嵌入	组检测
	MUVAD[72]	多视图异常检测	有监督、最近邻和聚类	识别两种以上类型数据的异常
基于全图的动态图异常检测	文献[45]	动态属性网络中异常变更检测	无监督、GLM、EKF、EWMA	快速且准确检测到异常
	AnomRank[15]	动态网络中异常事件检测	无监督、节点的分数一阶和二阶导数	综合结构和边异常、算法效率极高

3) 复杂网络的异常检测最常见的特点是采用图嵌入技术. 图嵌入[73-74]是一种将图数据映射为低维稠密向量的过程,解决图数据难以高效输入机器学习算法的问题. 针对原始网络的各种挑战,如何生成有效的图嵌入已成为异常检测的关键. 动态图异常检测的关键在于实时检测,难点在于大规模复杂动态属性网络的异常检测.

6 应用场景

复杂网络的异常检测广泛应用于各个领域,包括日志流检测、入侵检测、虚假新闻、垃圾邮件、欺诈交易、传感器网络、时间序列异常、物联网等. 这些应用领域中检测异常值已变得至关重要.

1) 欺诈交易与电信诈骗.

电信网络可以建模为一个图形,其中节点对应交换机,边缘代表光纤链路. 电信欺诈包括3个主要步骤:①通过电信网络进行的欺诈从普通帐户中获取资金;②在欺诈帐户之间进行复杂的交易以隐藏资金来源,即洗钱;③从欺诈性帐户中提取资金. 为了逃避银行欺诈检测系统的检测,欺诈者需要许多欺诈帐户才能使少量资金通过电子银行进行大量交易. 通过分析欺诈帐户和普通帐户的IP地址使用情况发现欺诈帐户,文献[67]提出了一种基于网络嵌入的方法检测由一组通用IP地址访问的欺诈帐户. LRGCN-SAPE[46]是一种新的路径表示方法,在加利福尼亚州的真实信息网络取得了很好的效果.

2) 虚假新闻与垃圾邮件.

社交媒体在给人们带来丰富的新闻信息的同时也会带来一些虚假的新闻,从新闻传播的源头来看,虚假新闻报道可以被视为异常值[75]. 将邮件发送者(或接受者)视为图中节点,邮件往来关系视为图中的边来构建邮件网络,从而复杂网络的异常检测也可以用于检测垃圾邮件. DeFrauder[71]利用潜在的产品评论图并结合行为信号来对候选欺诈组进行检测,这些行为信号可以模拟候选者之间的多方面协作;然后将候选者映射到一个嵌入空间,为每个组分配垃圾邮件评分,以使包含具有高度相似行为特征的垃圾邮件发送者的组获得较高的评分.

3) 物联网与传感器网络.

物联网设备由许多传感器组成,这些传感器根据所需任务融合特定传感器以获取有关区域的信息. 由于传感器中数据可能出现异常,在执行某任务之前,必须识别或检测这些异常值,以免限制任务的整体效率. DTM[66]是一种基于测距距离的NN的方法,已在基于Huber污染的仿真模型中取得很好的效果. 通过检测传感器中的异常可以确保网络路由的质量和提供准确的路由器传输数据,检测网络瓶颈有助于监视计算机网络性能. 文献[76]提出了一种基于可信度反馈的无线传感器网络分布式异常值检测算法,该算法包括评估传感器节点的初始可信度、基于可信度反馈和贝叶斯定理评估最终可信度以及调整异常值3个阶段. 文献[64]使用随机近似核函数将传感器中数据映射到随机大小的低维特征空间,从而识别传感器网络中的异常值.

4) 日志监控与数据流.

异常检测中最重要的应用场景之一就是日志监控,而日志多以数据流形式存在. 例如,Log-Anomaly[17]是由template2vec所支持的将日志流建模为NLP框架,可以同时检测顺序和定量日志异常. IRL[70]是一种基于逆向强化学习的连续数据异常检测框架. 在时间序列数据[77]和数据流[78]中检测异常值至关重要,因为异常值将影响正确结果的快速计算和估计. xStream[79]是一种基于密度的整体异常值检测器,用以检测磁盘驻留静态数据中的异常. LOF[68]是一种基于城市大数据的异常检测的分解方法,在检测人群异常流动和交通事故方面取得了较好的仿真效果.

7 存在的问题与挑战

复杂网络的异常检测已成为网络安全与数据挖掘的一个热门方向,广泛应用于安全、金融、多媒体和物联网等领域. 目前已经对复杂网络的大规模、高维度、动态性和富属性展开了相关研究,但仍存在一些挑战性问题有待于进一步研究.

1) 数据获取的困难度.

数据集的获取是评价一个异常检测算法优劣的重要环节. 但目前只有少数特定领域的公开异常检测数据集,无法作为评估异常检测方法的标准数据集. 现有的评价异常检测方法的数据集多是自获取或合成的数据集,这类数据集会面临以下挑战:①从物联网传感器中获取的数据集要经过一定的特征工程才可用于评价异常检测算法的优劣,因为这类数据具有噪声量大、冗余项多、稀疏度高等特征;②对于数据集中的正常噪音,如果没有很好地定义,则可能会被误认为异常;③无监督的异常检测的准确率往往小于有监督的异常检测,但自获取的数据集往往没有标签,而人工标记面临数据量大的困难. 最近,模型OCAN[42]通过生成对抗机制解决了数据集缺少标签

的问题;Wang等[23]通过噪音数据过滤器来预处理原始网络以减少数据噪音对异常检测的影响,一定程度上缓解了数据获取的困难.

2) 动态性和实时性.

划分时间快照的异常检测方法时间复杂度高,无法精确捕捉到动态网络的演化特征. 而数据流是一种特殊的时间序列数据,具有数据分布动态变化、数据规模大、数据持续到达等特征. 数据流和动态演化图的异常检测中最大的挑战就是数据的动态变化性,要求检测算法具备实时更新和动态适应的能力. 针对数据流的异常检测可以采用滑动窗口限制处理数据点的数量,借助于LSTM、CNN、RNN等深度学习模型来刻画数据流的动态更新,从而做到实时检测异常. 如MAD-GAN[80]使用LSTM-RNN来捕获时间序列分布的传感器数据流异常值. 针对动态演化图的异常检测可以通过图嵌入等深度学习模型来获得演化时刻的图特征表示,再借助LSTM等深度模型将动态演化图转化为时间序列. 此外,还可以借助动态图嵌入实时表示图的特征实现实时检测异常值. 如NetWalk[81]通过自编码器来学习节点的特征表示,在节点特征表示上使用动态聚类技术来动态检测网络异常值. 由于滑动窗口和图嵌入的局限性,以上检测方法只能在一定程度上做到实时性要求,未来可以研究设计自适应滑动窗口来检测数据流中异常值和低时间复杂度的动态图嵌入方法,分别用来实时检测动态数据流和动态演化图中的异常值.

3) 异构网络数据.

异构信息网络由表示对象的多类节点和表示对象间关系的多类边组成,在现实世界复杂网络中是普遍存在的. 例如在物联网中,许多具有不同协议的传感器和设备相互连接,产生了异构的数据流. 当前复杂网络的异常检测大多集中在同构网络,而针对异构网络中节点之间具有丰富的语义信息,如何从语义层面识别异构网络中的异常值是一个具有挑战性的问题. 检测异构网络中的异常值可以从两个角度出发: ① 将异构网络拆分成多个同构网络,对每一种同构网络进行异常检测; ② 通过创建不同查询来确定异常的类型及其范围. 最近,模型QANet[82]将异构网络中每个节点与其节点之间具有的各种元路径视为该节点的属性,然后使用张量分解提取特征,并使用聚类技术检测异常.

4) 多模态信息.

许多实际场景中,属性信息的收集通常有不同的来源,如图像(证件照片、指纹)、文本(银行交易历史记录、用户在线社交媒体帖子)、语音(音频)等,从而产生多模态属性信息. 在多模态属性网络中,属性有时以单独的模态查看时是正常的,但是当共同考虑多个模态时却是异常的. 属性的多模态给属性网络异常检测带来了机遇和挑战: 一方面,不同模态的属性信息可以提高异常检测的准确率; 另一方面,多模态属性信息的复杂分布需要统一的框架来表征,而这通常很难用常规线性模型捕获. 未来可以通过将不同的模态嵌入到一致的语义特征空间中来检测异常. 例如CADAHIN[22]用属性元路径来游走节点以捕获节点的多模态属性信息; CMAD[83]通过在学习的共识潜在特征空间中捕获实例之间的非线性相关性来识别不同模式下的异常实例. 此外,针对无线传感器收集的多模态数据流[84]还具有动态性特点. 由于多模态数据的广泛存在和复杂性,多模态属性信息网络的异常检测是未来研究的重要方向之一.

5) 数据的高维性和海量性.

现实世界的复杂网络中节点规模大、维度高且呈现动态演变的特性,所以复杂网络数据具有高维海量性特征. 如Facebook网络的月活跃用户高达24.1亿人次,而且每一个用户的信息都会动态更新. 检测这种复杂网络中的异常值的关键在于数据的高维性和海量性的处理. 针对网络高维度,可以采用图嵌入等深度学习方法进行网络的低维特征表示. 如Embed方法[20]利用图嵌入对原始社交网络进行低维特征表示,然后采用社区检测的方法检测网络中的异常值. 针对数据的海量性,可以采用分布式算法来降低时间复杂度,如doOCSVM[64]利用分布式算法优化传感器网络中的流数据异常检测时间. 尽管以上方法在一定程度上解决了网络数据的高维海量特性,但无法检测数百万级别节点的社交网络异常值. 所以高维且海量的复杂网络异常检测是未来研究方向之一.

6) 统一的评价标准和通用检测算法.

近年来,复杂网络的异常检测算法大量涌现,但大多数方法仅限于特定领域,导致复杂网络异常检测算法的通用性较差且出现了大量的异常定义,这对于异常定义的统一标准带来了挑战. 表3显示只有极少数算法是通用性算法,如CADA[34]用于检测静态网络中异常节点,ONE[63]用于检测属性网络中异常检测. 现有的异常评价没有统一的标准,通用的复杂网络异常检测算法可以用作异常检测评价标准. 未来的研究方向之一就是建立统一的标准来定义异常检测,建立通用的复杂网络异常检测方法.

7) 异常检测的新方法.

异常检测的关键在于检测的高准确率,虽然复杂网络的异常检测方法多样,但是采用单一异常检测方法的准确率不高. 对此,可以考虑融合多种异常检测方法来提高检测准确率,如GBKD-Forest[24]通过结合3种图类型的结构特征来提高检测准确率. ODBP[31]是一种集成的异常检测方法,该方法结合了基于特征(COMBN[44])和基于邻近度(LOF[27])的技术,可以有效检测违反特征依赖关系和与对象邻近性的异常. 未来可以尝试融合多种异常检测方法的集成学习来提高异常检测准确率,通过组合多个模型以得到更强大的高精度学习模型.

除了融合多个异常检测模型外,最近Chalapathy等[85]对异常值检测的深度学习方法进行了全面研究,回顾了深度学习方法在各种异常值检测应用中的使用方法和有效性. 传统的基于深度学习异常检测方法属于黑盒操作,缺乏对异常值的可解释性. 2019年开始,图神经网络(GNN)由于其较好的性能和可解释性已成为一种广泛应用的图分析方法. 基于GNN的复杂网络异常检测方法是指通过GNN提取整个网络特征的低维向量来预处理原始网络. 文献[86]通过GNN来捕获节点的结构和属性特征以检测社交网络中的节点异常,并且在Twitter数据集上实现了高达98%的AUC值,所以将GNN运用于异常检测具有很大的研究价值.

参考文献(References)

[1] 陈振江. 全球能源互联网大势已成[J]. 国企管理, 2019(14): 100-103.
(Chen Z J. The global energy internet has become a trend[J]. State-owned Enterprise Management, 2019 (14): 100-103.)

[2] 才智杰, 孙茂松, 才让卓玛. 藏文字同现网络的小世界效应和无标度特性[J]. 中文信息学报, 2018, 32(10): 45-52.
(Cai Z J, Sun M S, Cairang Z M. The small-world effect and the scale-free property of the Tibetan characteristics co-occurrence network[J]. Journal of Chinese Information Processing, 2018, 32(10): 45-52.)

[3] O'Gorman B, Wueest C, O'Brien D, et al. Internet security threat report[R]. California: SYMANTEC, 2019: 1-61.

[4] Bhuyan M H, Bhattacharyya D K, Kalita J K. Network anomaly detection: Methods, systems and tools[J]. IEEE Communications Surveys and Tutorials, 2014, 16(1): 303-336.

[5] Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: A survey[J]. Data Mining and Knowledge Discovery, 2015, 29(3): 626-688.

[6] Bremer Ronald. Outliers in statistical data[J]. Technometrics, 1995, 37(1): 117-118.

[7] Faigon A, Narayanaswamy K, Tambuluri J, et al. Machine learning based anomaly detection[P]. US: 10270788. 2019-04-23.

[8] Kurniabudi K, Purnama B, Sharipuddin S, et al. Network anomaly detection research: A survey[J]. Indonesian Journal of Electrical Engineering and Informatics, 2019, 7(1): 36-49.

[9] Brandón Á, Solé M, Huélamo A, et al. Graph-based root cause analysis for service-oriented and microservice architectures[J]. Journal of Systems and Software, 2020, 159: 110432.

[10] Noble C C, Cook D J. Graph-based anomaly detection[C]. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, 2003: 631-636.

[11] Eberle W, Holder L. Anomaly detection in data represented as graphs[J]. Intelligent Data Analysis, 2007, 11(6): 663-689.

[12] Staniford-Chen S, Cheung S, Crawford R, et al. GrIDS — A graph based intrusion detection system for large networks[C]. Proceedings of the 19th National Information Systems Security Conference. Baltimore, 1996: 361-370.

[13] Pincombe B. Anomaly detection in time series of graphs using ARMA processes[J]. Asor Bulletin, 2005, 24(4): 2-12.

[14] Wang H Z, Bah M J, Hammad M. Progress in outlier detection techniques: A survey[J]. IEEE Access, 2019(7): 107964-108000.

[15] Yoon M, Hooi B, Shin K, et al. Fast and accurate anomaly detection in dynamic graphs with a two-pronged approach[C]. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage AK, 2019: 647-657.

[16] Pang G, Shen C, van den Hengel A. Deep anomaly detection with deviation networks[C]. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage AK, 2019: 353-362.

[17] Meng W, Liu Y, Zhu Y, et al. Loganomaly: Unsupervised detection of sequential and quantitative anomalies in unstructured logs[C]. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, 2019: 4739-4745.

[18] Salehi M，Rashidi L. A survey on anomaly detection in evolving data: With application to forest fire risk prediction[J]. ACM SIGKDD Explorations Newsletter, 2018, 20(1): 13-23.

[19] McCarty C. EgoNet: Personal network software[D]. Florida: University of Florida, 2003.

[20] Hu R, Aggarwal C C, Ma S, et al. An embedding approach to anomaly detection[C]. IEEE 32nd International Conference on Data Engineering. Helsinki, 2016: 386-396.

[21] 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.

[22] 张蕊, 张桂发, 郭记明, 等. 富属性异质信息网络的可约束异常检测 [J]. 华中科技大学学报: 自然科学版, 2017, 45(12): 26-31.
(Zhang R, Zhang G F, Guo J M, et al. Constrainable anomaly detection for heterogeneous information networks with rich attributes[J]. Journal of Huazhong University of Science and Technology: Natural Science Edition, 2017, 45(12): 26-31.)

[23] Wang J, Su X. An improved K-means clustering algorithm[C]. IEEE 3rd International Conference on Communication Software and Networks. Xi'an, 2011: 44-46.

[24] Wang K, Chen D. Graph structure based anomaly behavior detection[C]. The 2nd International Conference on Computer Engineering, Information Science & Application Technology. Sanya, 2016: 543-550.

[25] Akoglu L, McGlohon M, Faloutsos C. Oddball: Spotting anomalies in weighted graphs[C]. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Heideberg: Springer, 2010: 410-421.

[26] Jeh G, Widom J. SimRank: A measure of structural-context similarity[C]. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton Alberta, 2002: 538-543.

[27] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying density-based local outliers[C]. Proceeding of the 2000 ACM SIGMOD International Conference on Management of Data. Dallas, 2000: 93-104.

[28] Sengupta S. Anomaly detection in static networks using egonets[J]. 2018, arXiv: 1807.08925.

[29] Brin S, Page L. The anatomy of a large-scale hyper textual web search engine[J]. Computer Networks and ISDN Systems, 1998(30): 107-117.

[30] Che H H, Giles C L. ASCOS: An asymmetric network structure context similarity measure[C]. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara Ontario, 2013: 442-449.

[31] Lu S, Liu L, Li J, et al. Effective outlier detection based on bayesian network and proximity[C]. 2018 IEEE International Conference on Big Data. Seattle, 2018: 134-139.

[32] Babbar S, Chawla S. Mining causal outliers using gaussian bayesian networks[C]. IEEE 24th International Conference on Tools with Artificial Intelligence. Athens, 2012: 97-104.

[33] Gao J, Liang F, Fan W, et al. On community outliers and their efficient detection in information networks[C]. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, 2010: 813-822.

[34] Helling T J, Scholtes J C, Takes F W. A community-aware approach for identifying node anomalies in complex networks[J]. International Conference on Complex Networks and Their Applications. Cham: Springer, 2018: 244-255.

[35] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): 10008-10020.

[36] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences, 2008, 105(4): 1118-1123.

[37] Guggilam S, Zaidi S M A, Chandola V, et al. Integrated clustering and anomaly detection (INCAD) for streaming data[J]. International Conference on Computational Science. Cham: Springer, 2019: 45-49.

[38] Wang H, Zhou C, Wu J, et al. Deep structure learning for fraud detection[C]. 2018 IEEE International Conference on Data Mining. Singapore, 2018: 567-576.

[39] Shibuya H, Maeda S. Anomaly detection method based on fast local subspace classifier[J]. Electronics and Communications in Japan, 2016, 99(1): 32-41.

[40] Xu S, Hu C, Wang L, et al. Support vector machines based on K nearest neighbor algorithm for outlier detection in WSNs[C]. The 8th International Conference on Wireless Communications, Networking and Mobile Computing. Barcelona, 2012: 1-4.

[41] Chalapathy R, Menon A K, Chawla S. Anomaly detection using one-class neural networks[J]. 2018, arXiv: 1802.06360.

[42] Zheng P P, Yuan S H, Wu X T, et al. One-class adversarial nets for fraud detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019(33): 1286-1293.

[43] Eswaran D, Faloutsos C. Sedanspot: Detecting anomalies in edge streams[C]. 2018 IEEE International Conference on Data Mining. Singapore, 2018: 953-958.

[44] Eswaran D, Faloutsos C, Guha S, et al. Spotlight: Detecting anomalies in streaming graphs[C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, 2018: 1378-1386.

[45] Gahrooei M R, Paynabar K. Change detection in a dynamic stream of attributed networks[J]. Journal of Quality Technology, 2018, 50(4): 418-430.

[46] Li J, Han Z C, Cheng H, et al. Predicting path failure in time-evolving graphs[J]. 2019, arXiv: 1905.03994.

[47] Paramasivan B. A study on node based anomaly detection

in wireless sensor networks[D]. Tamil Nadu: Faculty of Information and Communication Engineering, Anna University, 2018.

[48] Ranshous S, Shen S T, Koutra D, et al. Anomaly detection in dynamic networks: A survey[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2015, 7(3): 223-247.

[49] Rossi R A, Gallagher B, Neville J, et al. Modeling dynamic behavior in large evolving graphs[C]. Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome, 2013: 667-676.

[50] Yuan S H, Zheng P P, Wu X, et al. Wikipedia vandal early detection: From user behavior to user embedding[J]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2017: 832-846.

[51] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using lstms[J]. International Conference on Machine Learning, 2015: 843-852.

[52] Ren W, Yardley T, Nahrstedt K. EDMAND: Edge-based multi-level anomaly detection for SCADA Networks[C]. 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids. Aalborg, 2018: 1-7.

[53] Savalle P A, Sartran L, Vasseur J P, et al. Edge-based detection of new and unexpected flows[P]. USA: 10 389 741. 2019-08-20.

[54] Ji T F, Yang D Q, Gao J. Incremental local evolutionary outlier detection for dynamic social networks[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer, 2013: 1-15.

[55] Nguyen L H, Goulet J A. Anomaly detection with the switching Kalman filter for structural health monitoring[J]. Structural Control and Health Monitoring, 2018, 25(4): e2136.

[56] Sudrich S, Borges J, Beigl M, et al. Anomaly detection in evolving heterogeneous graphs[C]. 2017 IEEE International Conference on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data. Exeter, 2017: 1147-1149.

[57] Chen Z Z, Hendrix W, Samatova N F. Community-based anomaly detection in evolutionary networks[J]. Journal of Intelligent Information Systems, 2012, 39(1): 59-85.

[58] 韩涛, 兰雨晴, 肖利民, 等. 一种增量并行式动态图异常检测算法[J]. 北京航空航天大学学报, 2018, 44(1): 117-124.
(Han T, Lan Y Q, Xiao L M, et al. An incremental parallel dynamic graph anomaly detection algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(1): 117-124.)

[59] Wang T, Fang C, Lin D, et al. Localizing temporal anomalies in large evolving graphs[C]. Proceedings of the 2015 SIAM International Conference on Data Mining. Salt Lake, 2015: 927-935.

[60] Yang M, Rashidi L, Rajasegarar S, et al. Graph stream mining based anomalous event analysis[C]. Pacific Rim International Conference on Artificial Intelligence. Charm: Springer, 2018: 891-903.

[61] Shin K, Hooi B, Kim J, et al. Densealert: Incremental dense-subtensor detection in tensor streams[C]. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, 2017: 1057-1066.

[62] Grattarola D, Zambon D, Livi L, et al. Change detection in graph streams by learning graph embeddings on constant-curvature manifolds[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(6): 1856-1869.

[63] Bandyopadhyay S, Lokesh N, Murty M N. Outlier aware network embedding for attributed networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 12-19.

[64] Miao X D, Liu Y, Zhao H Q, et al. Distributed online one-class support vector machine for anomaly detection over networks[J]. IEEE Transactions on Cybernetics, 2019, 49(4): 1475-1488.

[65] Schölkopf B, Platt J C, Shawe-Taylor J, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471.

[66] Gu X Y, Akoglu L, Rinaldo A. Statistical analysis of nearest neighbor methods for anomaly detection[C]. Advances in Neural Information Processing Systems. Vancouver, 2019: 10923-10933.

[67] Liu X, Wang X G. A network embedding based approach for telecommunications fraud detection[J]. International Conference on Cooperative Design, Visualization and Engineering. Cham: Springer, 2018: 229-236.

[68] Zhang M Y, Li T, Shi H Z, et al. A decomposition approach for urban anomaly detection across spatiotemporal data[C]. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, 2019: 6043-6049.

[69] Gopalan P, Levin R, Wieder U. PIDForest: Anomaly detection and certification via partial identification[C]. Advances in Neural Information Processing Systems. Vancouver, 2019: 15809-15819.

[70] Oh M H, Iyengar G. Sequential anomaly detection using inverse reinforcement learning[C]. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, 2019: 1480-1490.

[71] Dhawan S, Gangireddy S C R, Kumar S, et al. Spotting collusive behaviour of online fraud groups in customer reviews[J]. 2019, arXiv: 1905.13649.

[72] Sheng X R, Zhan D C, Lus, et al. Multi-view

anomaly detection: neighborhood in locality matters[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, 2019(33): 4894-4901.

[73] Cai H Y, Zheng V W, Chang K C. A comprehensive survey of graph embedding: Problems, techniques, and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1616-1637.

[74] 涂存超, 杨成, 刘知远, 等. 网络表示学习综述[J]. 中国科学: 信息科学, 2017, 47(8): 980-996.
(Tu C C, Yang C, Liu Z Y, et al. Overview of network representation learning[J]. Science in China: Information Science, 2017, 47(8): 980-996.)

[75] Feng H L, Liang L, Lei H. Distributed outlier detection algorithm based on credibility feedback in wireless sensor networks[J]. IET Communications, 2017,11(8): 1291-1296.

[76] Shu K, Sliva A, Wang S H, et al. Fake news detection on social media: A data mining perspective[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22-36.

[77] Zhang A Q, Song S X, Wang J M, et al. Time series data cleaning: From anomaly detection to anomaly repairing[J]. Proceedings of the VLDB Endowment, 2017, 10(10): 1046-1057.

[78] Zheng Z G, Jeong H Y, Huang T, et al. KDE based outlier detection on distributed data streams in multimedia network[J]. Multimedia Tools and Applications, 2017, 76(17): 18027-18045.

[79] Manzoor E, Lamba H, Akoglu L. xStream: Outlier detection in feature-evolving data streams[C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, 2018: 1963-1972.

[80] Li D, Chen D C, Shi L, et al. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks[J]. International Conference on Artificial Neural Networks. Cham: Springer, 2019: 703-716.

[81] Yu W, Cheng W, Aggarwal C C, et al. Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks[C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, 2018: 2672-2681.

[82] Ranjbar V, Salehi M, Jandaghi P, et al. QANet: Tensor decomposition approach for query-based anomaly detection in heterogeneous information networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(11): 2178-2189.

[83] Li Y, Liu N, Li J, et al. Deep structured cross-modal anomaly detection[C]. 2019 International Joint Conference on Neural Networks. Budapest: IEEE, 2019: 1-8.

[84] 费欢, 肖甫, 李光辉, 等. 基于多模态数据流的无线传感器网络异常检测方法[J]. 计算机学报, 2017, 40(8): 1829-1842.
(Fei H, Xiao F, Li G H, et al. Wireless sensor network anomaly detection method based on multimodal data stream[J]. Chinese Journal of Computers, 2017, 40(8): 1829-1842.)

[85] Chalapathy R, Chawla S. Deep learning for anomaly detection: A survey[J]. 2019, arXiv: 1901.03407.

[86] Chaudhary A, Mittal H, Arora A. Anomaly detection using graph neural networks[C]. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing. Faridabad: IEEE, 2019: 346-350.

作者简介

苏江军 (1994−), 男, 硕士生, 从事数据挖掘、图神经网络与异常检测的研究, E-mail: siyuan2018@foxmail.com;

董一鸿 (1969−), 男, 教授, 博士, 从事大数据处理、数据挖掘与人工智能等研究, E-mail: dongyihong@nbu.edu.cn;

颜铭江 (1995−), 男, 硕士生, 从事数据挖掘、图神经网络与异构网络表示学习的研究, E-mail: yanmj7700@163.com;

钱江波 (1974−), 男, 教授, 博士生导师, 从事机器学习、模式识别与智能系统等研究, E-mail: qianjiangbo@nbu.edu.cn;

辛宇 (1987−), 男, 副教授, 博士, 从事大数据处理、图神经网络与推荐系统等研究, E-mail: xinyu@nbu.edu.cn.

(责任编辑：李君玲)