基于无监督聚类和频繁子图挖掘的电力通信网缺陷诊断与自动派单

吴季桦[1]，朱鹏宇[2]，吴子辰[3]，顾彬[3]，洪涛[3]，郭波[3]，王晶[1]，王敬宇[1]
（1. 北京邮电大学网络与交换国家重点实验室，北京 100876;
2. 国网电力科学研究院有限公司，江苏 南京 210012;
3. 国网江苏省电力有限公司信息通信分公司，江苏 南京 210024）

摘　要：缺陷诊断一直是电力通信领域研究的难点之一。基于人工规则的缺陷诊断已经无法应对告警数据的海量增长。基于有监督学习的智能方法需要大量的标注数据和较长的系统构建时间，且大多面向指标性数据，实现部署缺乏可行性。面向告警数据，提出一种基于无监督聚类和频繁子图挖掘实现告警归并和缺陷模式发现的自学习算法，设计了一个自动化完成缺陷诊断及处置的架构。该架构具有良好的可扩展性和迭代更新能力，并部署于实际缺陷自动派单系统中。通过真实场景数据集进行实验验证，结果显示出良好的性能表现，实现了对缺陷的及时发现及精准派单维护。

关键词：电力通信；缺陷诊断；无监督聚类；频繁子图挖掘
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2021253

Fault diagnosis and auto dispatchin of power communication network based on unsupervised clustering and frequent subgraph mining

WU Jihua[1], ZHU Pengyu[2], WU Zichen [3], GU Bin[3], HONG Tao[3], GUO Bo[3], WANG Jing[1], WANG Jingyu[1]
1. State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China
2. State Grid Electric Power Research Institute Co., Ltd, Nanjing 210012, China
3. Information and Communication Branch of State Grid Jiangsu Electric Power Co., Ltd., Nanjing 210024, China

Abstract: Fault diagnosis is one of the most challenging tasks in power communication. The fault diagnosis based on rules can no longer meet the demand of massive alarms processing. The existing approaches based on the supervised learning need large sets of the labeled data and sufficient time to train models for processing continuous data instead of alarms, which are far behind the feasibility of deployment. As for alarm correlation and fault pattern discovery, a self-learning algorithm based on the density-based clustering and frequent subgraph mining was proposed. A novel approach for automatic fault diagnosis and dispatch were also introduced, which provided the scalable and

收稿日期：2021−05−31；修回日期：2021−11−15
通信作者：王晶，wangjing@ebupt.com
基金项目：国家电网公司科技项目（No.5700-202040367A-0-0-00）
Foundation Item: Science and Technology Project of State Grid Corporation (No.5700-202040367A-0-0-00)

self-renewing ability and had been deployed to the automatic fault dispatch system. Experiments in the real-world datasets authorized the effectiveness for timely fault discovery and targeted fault dispatch.

Key words: power communication, fault diagnosis, unsupervised clustering, frequent subgraph mining

1 引言

电力通信网中的海量告警数据显示了网元设备的健康状态以及网元设备间的交互情况。面向告警的缺陷诊断方法先对告警进行告警归并，基于得到的告警归并集合，进一步进行缺陷检测和缺陷定位。

目前国内外主要使用基于规则匹配的方法进行告警归并[1]。随着告警数据的海量增长，基于规则匹配的方法及其相关改进方法难以适应当前的数据环境。Madziarz[2]在移动通信网领域提出了基于 *k*-means 聚类的告警聚类方法，尝试引入无监督聚类以摆脱对规则的依赖。虽然该方法无须大量人力资源的投入，但实际归并效果不理想，且需要业务专家参与预测缺陷的数量，有着极大的局限性。

缺陷诊断分为事件检测和定位。事件检测和定位则基于事件分类。基于人工经验的缺陷诊断方法，主观因素影响较大，并且难以应对指数级增长的海量告警信息。已经有许多研究将人工智能技术运用到电力通信网事件分类和缺陷诊断领域中以摆脱对规则的依赖。人工智能技术应用到缺陷诊断领域时，常针对的是信号等指标性数据，如 Wen 等[3]利用卷积神经网络（convolutional neural networks，CNN）提取信号的特征，Xiao 等[4]利用贝叶斯神经网络以建筑管理系统的测量和人工测试指标作为输入依据，进行变风量通风空调系统的缺陷诊断。这些方法都在各自的数据集上取得了较好的成果。但是电力通信领域缺少大量完整标注的数据，同时，实际的缺陷诊断的主要依据不是指标性的信号数据，而是各个网元上非结构化的告警数据。电力通信网中基于告警完成缺陷诊断的缺陷信息隐藏在告警数据以及其时空关联关系中。

本文在自适应标记筛选及再学习[5]和基于拓扑信息解决时间序列数据异常检测问题[6]的工作基础上，提出了一种基于密度聚类（density-based spatial clustering of applications with noise，DBSCAN）实现告警归并，并且基于频繁子图挖掘（frequent subgraph mining，FSM）完成缺陷模式发现的自学习算法，并设计了一个面向电力通信网告警数据，尽力摆脱对规则的依赖，减轻人力资源投入的自动化缺陷诊断及派单的架构。如图 1 所示，该算法主要应用于告警归并、缺陷诊断以及自动派单模块，模块间松耦合，具有良好的可扩展性。该算法展现出良好的稳健性，具备迭代更新能力，减少缺陷诊断过程对于人工规则的依赖，并在实验中呈现出良好的结果。

本文的贡献总结如下。

（1）提出了面向告警数据进行数据挖掘，基于无监督聚类和频繁子图挖掘算法完成告警归并以及缺陷诊断，智能化完成缺陷模式发现及识别，自动化完成派单检修，具备迭代更新的自学习能力架构，部署在自动派单系统中，以减轻运维压力，实现对缺陷的及时发现和处置。

（2）考虑基于规则的缺陷诊断方法受到人为因素的制约，基于有监督的学习方法受到缺少大量标注数据进行训练的制约，提出一个基于无监督学习以及数据挖掘的模型，在只有少量标注的情况下实现对缺陷模式的及时发现。

（3）考虑告警归并集合内告警存在时空相关，提出了将告警的文本信息向量化映射到向量空间的方法，使得具有相关关系的告警在向量空间之中彼此接近，并使用无监督聚类方法完成告警归并。

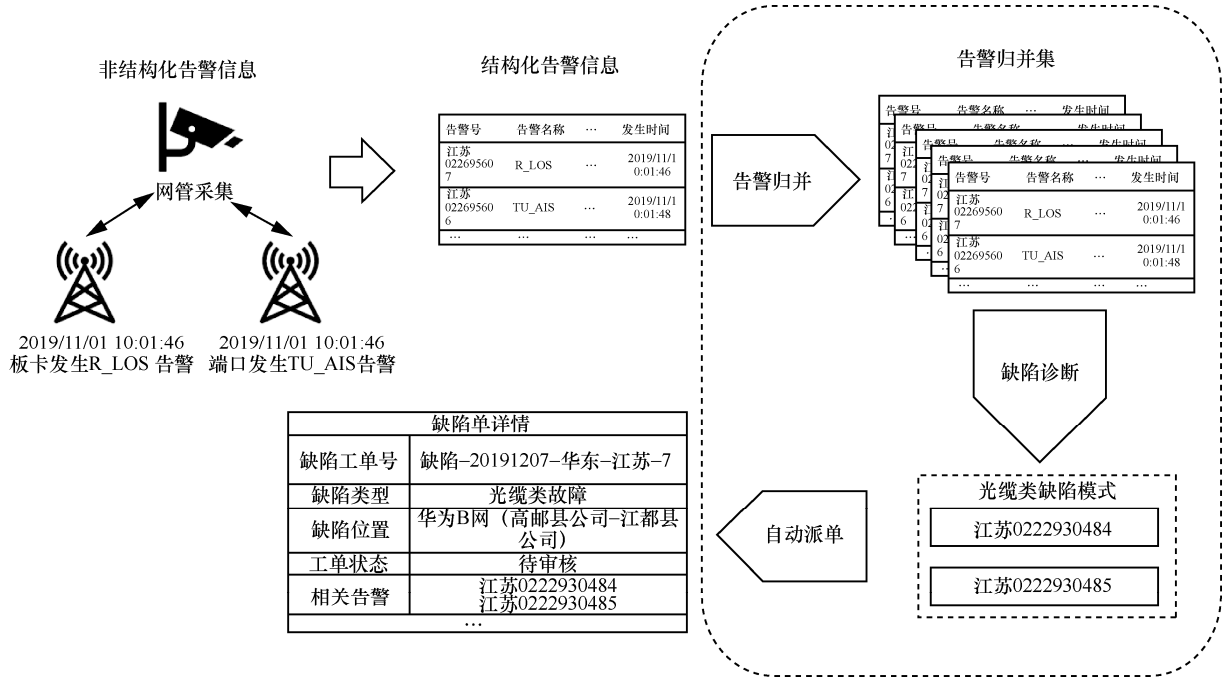


图 1　电力通信网的缺陷诊断和自动派单架构

（4）考虑网络场景中发生告警的节点之间的拓扑关系，提出了对告警及其所处节点的拓扑关系进行模式挖掘，利用频繁子图挖掘方法完成缺陷模式发现。

（5）在真实场景数据集上进行实验，结果表明基于无监督聚类和频繁子图挖掘的缺陷诊断方法取得了良好的性能表现。

2　告警归并模块

2.1　告警归并

告警归并是一个概念性的解释：将某些告警根据某种意义相关联。告警归并是以下几种网络管理任务中的一种通用方法[7]：压缩、计数、抑制。电力通信网中的告警归并是为了后续的缺陷诊断服务的。因此本文场景中的告警归并旨在将可能由同一个缺陷导致或者衍生的告警归并在同一个集合当中。在电力通信的生产场景中，运维人员依靠人工经验的积累梳理出告警衍生关系，并以此为依据完成告警归并。但是这意味着，基于规则完成的告警归并主观成分较大，整理的告警衍生关系也可能不完备。

告警归并任务的目标是将可能由同一个缺陷引起的告警关联在一起。同时，在一段连续时间内，一个缺陷可能会引起一个或多个设备持续输出相似告警。基于这两个前提，本文使用无监督学习来协助完成告警归并，采用基于密度的 DBSCAN 聚类捕获告警簇。

2.2　DBSCAN

DBSCAN 算法作为经典的密度聚类算法，其在无监督密度聚类中的得到了广泛的应用。算法将点分类为核心点和非核心点，定义 1～定义 6 描述了该算法[8]。

定义 1（邻域内点的集合（Eps-neighborhood of a point））$N_{\mathrm{Eps}}(p)$ 表示从 p 点出发，邻域内所有点的集合，即：

$$N_{\mathrm{Eps}}(p)=\{q\epsilon D \mid \mathrm{dist}(p,q)\leqslant \mathrm{Eps}\}$$

其中，D 表示全体点的集合，$\mathrm{dist}(p,q)$ 表示 p 点和 q 点之间的欧氏距离，Eps 表示邻域半径。

定义 2（直接密度可达（directly density-

reachable)) 点 p 被称为从点 q 直接密度可达，当且仅当：

$$p\epsilon N_{\mathrm{Eps}}(q), \left|N_{\mathrm{Eps}}(q)\right| \geqslant \mathrm{MinPts}$$

其中，MinPts 为给定的使 q 成为核心点的邻域内最小点数。

定义 3 （密度可达（density-reachable））如果有一系列的点 $p_1,\cdots,p_n, p_1=q, p_n=p$，$p_{i+1}$ 到 p_i 直接密度可达，那么称点 p 从 q 密度可达。

定义 4 （密度相连（density-connected））如果点 p 和点 q 都从点 o 密度可达，则称点 p 和点 q 密度相连。

定义 5 （簇（cluster））对于集合 D，簇 C 是 D 的一个满足以下条件的子集。

- $\forall p,q:\text{if } p\epsilon C$，并且 q 从 p 密度可达，那么 $q\epsilon C$。
- $\forall p,q\epsilon C$：p 和 q 密度可达。

定义 6 （噪声（noise））噪声为不属于任意一个簇 C_i 的离散点，即：

$$\mathrm{noise}=\{p\epsilon D \mid \forall i: p\notin C_i\}$$

具体而言，在划分簇时，对于给定的边界距离 Eps、最小核心节点数 MinPts 和非空节点集 D，簇 C 构建时首先检测其密度直达性。首先将核心点中具有密度直达关系的点分类给簇 C，之后检测相连性，对剩下的点检测其与簇内任意一点的密度相连性，如果密度相连则归入簇 C。

在分类完成后，对于不属于任何簇的孤立点 p，将其视为噪声[9]。

2.3 告警归并模块设计

DBSCAN 是基于密度的算法，意味着输入的特征应当是对应空间的坐标点，或者是点之间的距离矩阵。在实际背景当中告警是连续的文本信息，因此告警的向量化过程应该体现为特征提取和特征向量之间的权重分配。

本文告警归并的目标对象应当是在时间上相近以及发生设备间有关联关系或者本身其他属性相近的一组告警，也就是 DBSCAN 聚类的目标是将拥有这些特性的属于同一缺陷的告警聚为一个簇。对告警而言，有两方面的信息较为重要：告警本身的相关参数（如告警种类、发生位置、设备类型、设备位置等）以及告警时间。

其中，告警本身的相关参数反映了告警之间的相关程度以及告警在空间上的相近程度，告警时间是当前告警产生的时间，蕴含了缺陷发生的时间信息。对于告警本身的相关参数，使用 One-Hot 方法[10]将其映射为特征向量，对于没有制定权重的 One-Hot 来说，告警之间任意一个特征的差距映射在空间上面距离相同，在 DBSCAN 算法当中作用相同，而通过调整各个特征的权重可以反映不同特征的重要性。

对于时间信息而言，显然不能简单运用 One-Hot 方法，则关键在于统一时间和其他特征在特征向量上的距离关系。告警集合为 D，告警具有 1 个时间特征和 m 个其他经过 One-Hot 转换的相关参数特征。则第 i 条告警对应特征向量表述为 $D_i=(T_i,X_{i,j},\cdots,X_{i,m})$，其中，$T_i$ 为告警 i 的一维时间特征，$X_{i,j}(1\leqslant j\leqslant m)$ 表示告警 i 的第 j 个相关参数对应的向量。则假设告警本身相关参数的每个特征都被赋予同样的权重，如果要使得任意变化一个特征带来的空间坐标之间的距离变化等价于两个相邻告警 a 和 b 的时间差超过了时间窗口的距离变化，设变化的特征为第 t 个特征，特征维度为 n，时间窗口为 WINDOWS 则可以得到：

$$\beta\times\mathrm{WINDOWS}=\sum_{k=1}^{n}\left(X_{a,t,k}-X_{b,t,k}\right)^2 \qquad (1)$$

对于单纯的 One-Hot 向量特征，显然 $\beta=2/\mathrm{WINDOWS}$。

进一步，可以得到：

$$\beta\times\mathrm{WINDOWS}=\sum_{j=1}^{m}\sum_{k=1}^{n_j}\beta_j\left(X_{a,j,k}-X_{b,j,k}\right)^2 \qquad (2)$$

其中，n_j 表示第 j 个特征对应的维度，β_j 表示第

j 个特征的权重。可以通过调整 β 和 $\beta_j(1 \leqslant j \leqslant m)$ 的值来调整时间和其他参数之间的权重。

不同样本的距离综合考虑了告警本身相关参数距离和时间距离。以此对所有告警进行聚类，则最后得到的聚类结果应该是使得时间上较为聚集的相似告警或者时间上极为聚集的较相似告警成为同个簇。

基于无监督聚类告警归并模块的流程如图 2 所示，其中告警的文本化数据的向量化和空间映射过程在以上讨论中已经得到论述。上文证明了在时间上接近以及其他特征接近的告警数据会在向量空间中接近，DBSCAN 算法会将向量空间中接近的告警聚为一个簇，从而完成告警归并的目标。进一步，告警归并的结果将会基于人工审核的缺陷单数据进行有效性评估，以此来调整算法的参数以及评价算法效果。

3　缺陷诊断和自动派单

3.1　缺陷诊断

大型通信网络中的缺陷诊断流程可以被分解为 3 个步骤[10]：故障检测、故障定位、故障诊断。应用于电力通信网的缺陷诊断的技术方法主要有专家系统、神经网络、优化技术、Petri 网络、粗糙集理论、模糊集理论、贝叶斯网络、多 Agent 技术等[11]。

对于由同一种缺陷原因引发的缺陷，应当在设备类型、设备数量、拓扑连接等方面存在相似。类似于将无监督聚类方法应用于告警归并中的前提，数据分布中的相似性给予了人工智能技术发挥其长处的可能。具体地，缺陷诊断任务中数据的相似性体现在拓扑结构上的相似性。对于电力通信网络的缺陷相关告警数据的研究发现，与某一缺陷相关的告警所发生的设备通常具有物理相连关系或者逻辑相连关系。设备及其之上的告警，以及设备间的关联关系可以构成基本图结构。属于同一类缺陷的图结构之间经常存在子图结构的相似甚至相同。因此本文将电力通信的缺陷模式发现问题转化为基于图的模式发现问题进行解决。

3.2　gSpan

作为常见的数据结构，带标记的图可以用来更方便地为表达复杂数据而建模描述。给定一个图数据集 $D=\{G_0,G_1,\cdots,G_n\}$，support(g) 表示 g 作为图 G_i 的子图在数据集 D 中出现的次数。频繁子图挖掘问题就是找到所有子图 g，并且子图 g 可以使得 support(g) ⩾ minSup。其中，minSup 是人为设置的最小支持度阈值[12]。

其中，频繁子图挖掘算法中 gSpan（graph-

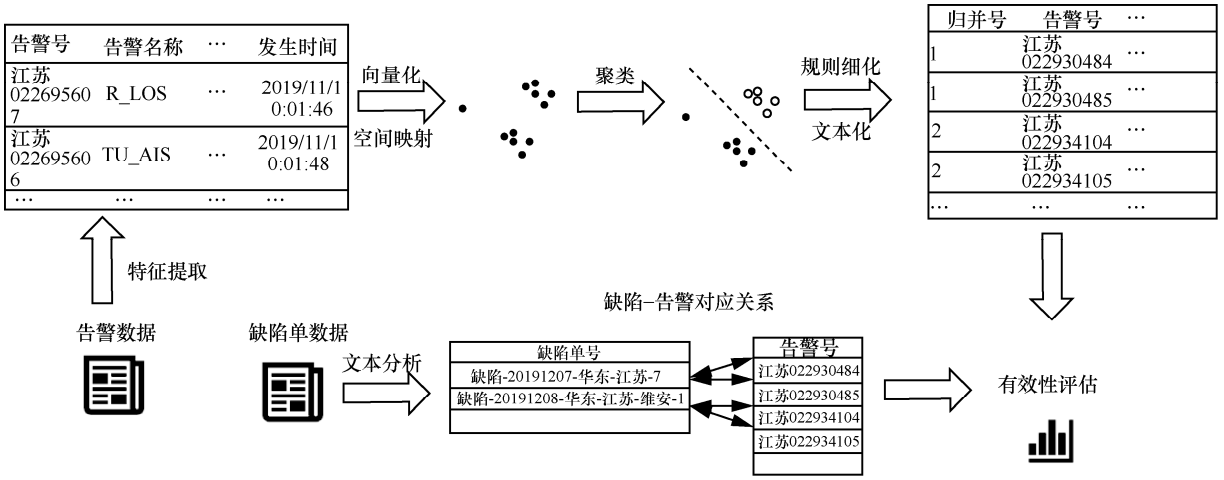


图 2　基于无监督聚类告警归并模块流程

based substructure pattern mining）由于其在时间复杂度以及空间复杂度的优秀表现，在频繁子图挖掘领域中得到了广泛的应用。gSpan 的关键流程包括从规模为 k 的频繁子图集合生成规模为 k+1 的频繁子图候选集，以及检查候选集中的子图是否为同构子图以此修剪冗余部分。

检查子图同构问题是一个 NP 完全问题[13]，因此在 gSpan 中利用最小 DFS 编码和 DFS 字典树解决子图同构的检查。

gSpan 是一个较为复杂的算法，关键的 DFS 编码依据定义 7～定义 10[14]。

定义 7 （DFS 编码）给定由图 G 得到的 DFS 树 T，基于 DFS 编码 $\prec_T$ 构造边序列 e_i。当 $i=0,\cdots,|E|-1$ 时，e_i 小于 e_{i+1} 的编码序。

一条边可以以一个五元组表示，$(i,j,l_i,l_{(i,j),l_j})$。其中，l_i 和 l_j 是顶点 v_i 和 v_j 的标记，$l_{(i,j)}$ 是顶点 v_i 和 v_j 之间的边的标记。

定义 8 （DFS 字典序）给定 $Z=\{\text{code}(G,T)\mid T是G的\text{DFS树}\}$。$Z$ 是一个包含所有带标记的连通图所生成的所有 DFS 编码的集合。给定标记集合 L 的线性序列编码 $\prec_L$，那么 $\prec_T$ 和 $\prec_L$ 的字典结合是集合 $E_T\times L\times L\times L$ 的一个线性序列 $\prec_e$。更多的细节可见文献[15]。DFS 字典序是可比较的线性序列，即：

如果 $\alpha=\text{code}(G_\alpha,T_\alpha)=(a_0,a_1,\cdots,a_m)$ 并且 $\beta=\text{code}(G_\beta,T_b\text{eta})=(b_0,b_1,\cdots,b_n)$，$\alpha,\beta\in Z$，$\alpha\leqslant\beta$ 当以下任意条件满足时成立。

- 对 $k<t,a_t\prec_e b_t,\exists t,0\leqslant t\leqslant\min(m,n),a_k=b_k$。
- 对 $0\leqslant k\leqslant m$ 并且 $n\geqslant m,a_k=b_k$。

定义 9 （最小 DFS 编码）给定图 G，$Z=\{\text{code}(G,T)\mid T是G的\text{DFS树}\}$，基于 DFS 字典序得到的集合中的最小值 $\min(Z(G))$，称为图 G 的最小 DFS 编码。这同样是图 G 的范式标记。

定义 10 （DFS 编码树）在 DFS 编码树中，每一个节点代表了一个 DFS 编码。父节点和子节点之间的关系遵循以下的描述。

给定 DFS 编码 $\alpha=(a_0,a_1,\cdots,a_m)$，任意有效的 DFS 编码 $\beta=(a_0,a_1,\cdots,a_m,b)$ 称为 α 的孩子节点 α，称为 β 的父节点。给定 DFS 字典序时，父子节点之间的关系是固定的。这意味着，先序遍历 DFS 编码树将符合 DFS 字典序。

给定标记集合 L，DFS 编码树应当包含无穷的图。因为本文只考虑有限集中的频繁子图，DFS 编码树的规模也是有限的。DFS 编码树中第 n 层的节点包含(n−1)条边的图的 DFS 编码。通过 DFS 编码树的深度优先遍历，所有的具有最小 DFS 编码的频繁子图都能被发现。特别地，如果节点中包含具有不同的 DFS 编码的重复的图，例如 s 和 s'表示同一个图但是 s 具有更小的 DFS 编码，那么 s'不是最小的 DFS 编码，s'将会被剪枝。

算法 1 和算法 2 描述了 gSpan 算法的伪代码。其中，D 表示图数据库，S 包含了挖掘结果。更多算法细节参考文献[14]。

算法 1 GraphSet_Projection(D,S)

根据支持度对 D 中的标记进行排序
移除不频繁的顶点和边
重新标记剩下的顶点和边
$S^1\leftarrow D$ 中所有频繁的只有一条边的图
根据 DFS 字典序排列 S^1
$S\leftarrow S^1$
for 每一条边 $e\in S^1$ do
    用 e 初始化 s，用包含 e 的图设置 s
    Subgraph_Mining(D,S,s) $D\leftarrow D-e$
    if $|D|<$ minSup
        break

算法 2 Subgraph_Mining(D,S,s)

if $s\neq\min(s)$
return
$S\leftarrow S\cup\{s\}$
枚举每个图中的 s 并且对其子节点计数
for 每个 c，c 属于 s 的子节点 do
    if support(c) ≥ minSup

$s \leftarrow c$

Subgraph_Mining(D_s,S,s)

3.3　缺陷诊断和自动派单模块设计

针对网络拓扑中大规模 KPI 异常检测的场景，文献[6]提出了一种基于图的门控卷积编解码异常检测（graph-based gated convolution codec for anomaly detection，GAD）模型，通过提取节点间的空间特征，以挖掘详细的节点连接状态信息。GAD 运用到大规模网络中获得了良好的表现效果。因此，本文考虑结合时空关系完成电力通信网的缺陷诊断。但是电力通信网场景中某一缺陷所关联的告警往往局限于一个小而准确的范围，需要捕捉到准确的缺陷模式。因此本文采用频繁子图挖掘方法对于缺陷模式进行捕捉。

将频繁子图挖掘方法应用到告警归并集合数据分析领域的首要任务是将归并集合转化成图数据。本文将通信网络中的网元转化成图中的顶点，将网元之间的物理联系（如经过光缆相连）以及网元之间的逻辑联系（如网元与网管之间保持的通信）转化为图中的边，网元上发生的告警转化为图中的标记。由于网元之间的联系是双向的，因此顶点之间的边为无向边。基于以上讨论，本文将告警归并集合转化为顶点带标记的无向连通图，并对此进行频繁子图挖掘。

对于顶点上发生的告警转化为顶点的标记，本文使用告警编码对顶点进行标记。具体来说，给定告警集合 L，标记 l 使用长度为|L|的 One-Hot 编码。例如，对于顶点 v_i 上发生了集合 L 内第 j 位的告警，$l_{i,j}=1$。

特别地，在电力通信网的缺陷诊断的场景下，本文可以对 gSpan 得到的频繁子图模式进行进一步剪枝，对满足以下任一条件的子图，本文不视作可能存在的缺陷模式。

- 只有一个顶点的子图。
- 顶点数大于 2 且度为 1 的节点上没有告警发生的子图。

这是因为实际电力通信场景中的缺陷通常可归类为单网元不衍射到其他网元故障，或者是单网元可衍射到其他网元故障以及网元间介质故障。对于单网元故障，告警应当被网元及其从属的网管采集，至少存在两个顶点；对于单网元可衍射到其他网元故障及网元间介质故障，缺陷模式的最远点应当是故障影响范围的末端，也就是对应最远上报告警的网元。

基于以往面向区间异常检测进行自适应标记筛选和再学习[5]的工作基础，本文利用标记筛选以及基于历史数据实现训练和预测并行的思想，设计了缺陷模式发现以及缺陷诊断及自动派单流程。在模式标注前后，历史缺陷单数据将与对应出现的模式进行关联，使得模式之间的差异能被准确检测到。

基于频繁子图挖掘的缺陷模式发现流程如图 3 所示。经过以上讨论的图数据结构设计，各个顶点上的告警经过告警编码，由告警归并模块得到的归并集合完成拓扑生成，得到带标记的无向连通图。由告警归并集合得到的带标记的无向连通图集合经过子图挖掘并且经过剪枝保留频繁子图模式，再一次进行告警解码后成为待标记的可能缺陷模式。待标记的可能缺陷模式在经过专家标注之后存入知识库，完成缺陷模式的发现。两个发生 R_LOS 告警的端口存在物理关联关系（端口分别从属的站点间存在光缆连接关系）和逻辑关联关系（端口分别从属的网管间存在通道关系）。告警编码时，假设告警集合大小为 9，则告警编码序列长度为 9，若 R_LOS 所处位置为 0，则只发生 R_LOS 的告警编码对应为 100 000 000，没有发生任何告警的告警编码对应为 000 000 000。在子图挖掘并完成剪枝之后得到两个子图，经过告警解码后，子图重新还原为具有高可读性的可能缺陷模式，方便等待人工进行标注。

缺陷诊断和自动派单模块流程如图 4 所示。当新的归并集合到达，得到拓扑生成图 G，使用

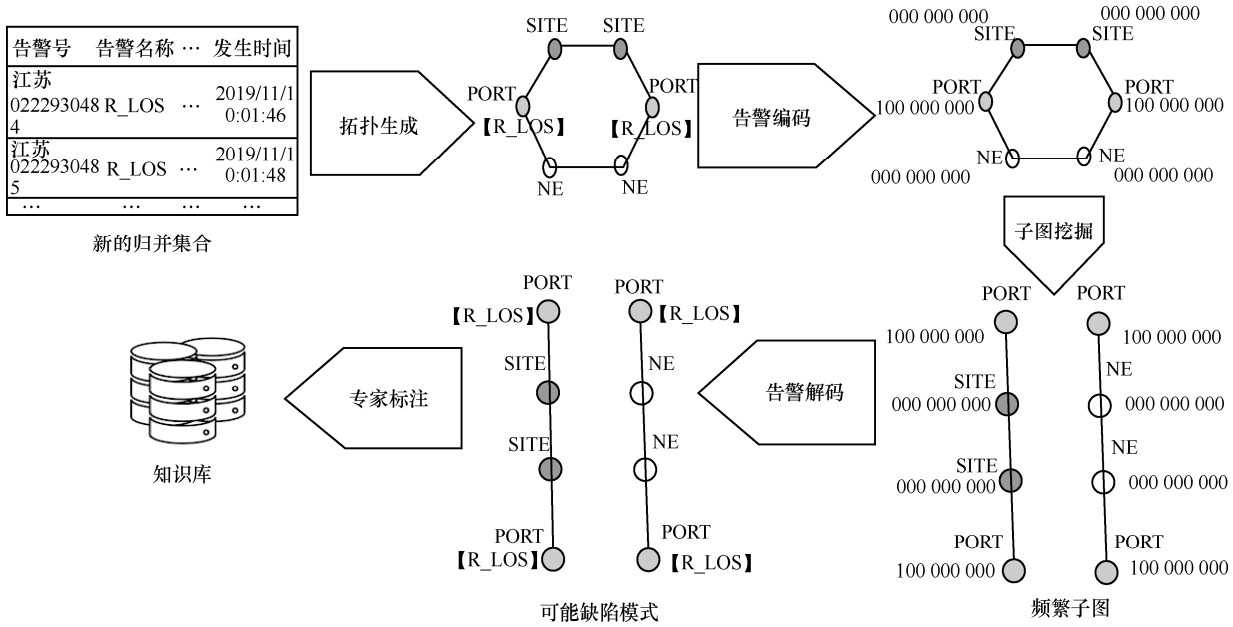


图 3　基于频繁子图挖掘的缺陷模式发现流程

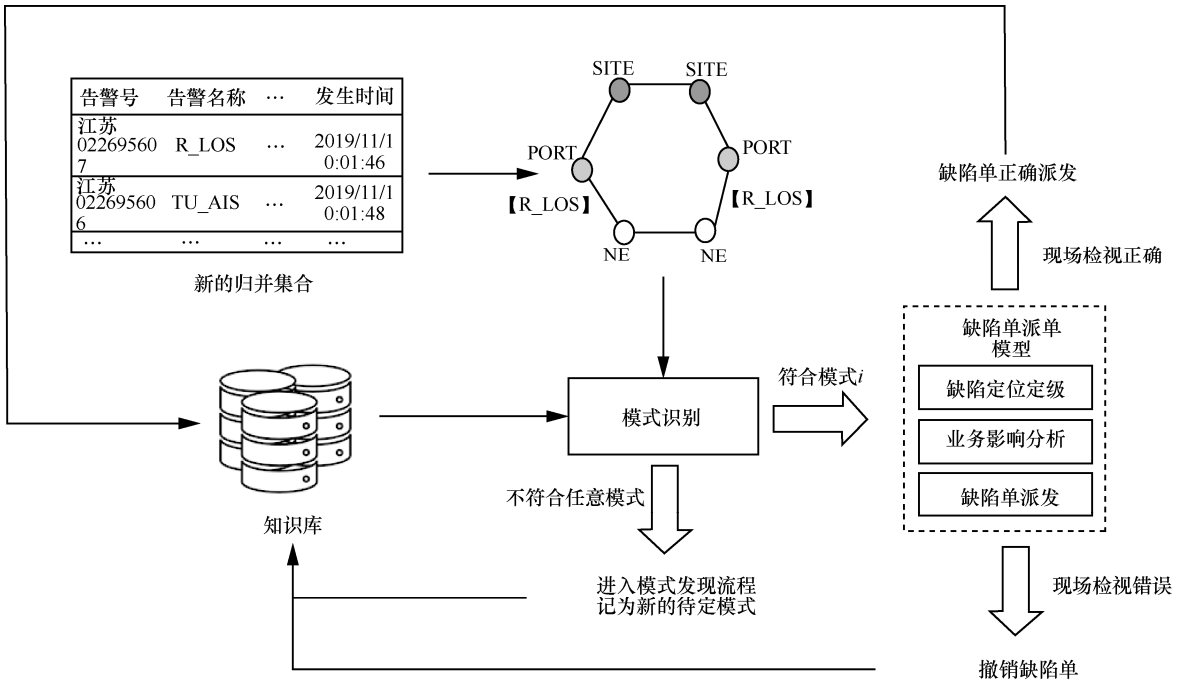


图 4　缺陷诊断和自动派单模块流程 4 实验结果分析

知识库中已标记的缺陷模式进行模式识别。具体而言，若在图 G 中识别到了缺陷模式 i，则根据识别出的模式在图 G 中的映射位置完成缺陷定位，根据知识库中该模式对应的专家标注完成缺陷分类，从而完成缺陷诊断，并基于缺陷定位定级、业务影响分析完成缺陷单派发。若未在图 G 中识别到知识库中的模式，则使图 G 进入模式发现流程，记为新的待定模式。可见本文所提供的缺陷诊断架构具有强大的容错能力并且拥有迭代更新的能力。

4　实验结果分析

4.1　告警归并的有效性验证

归并结果的有效性验证[16]借鉴了聚类方法的评估指标，聚类方法的评价指标[17]分为外部指标和内部指标，内部评价聚类的估计趋势，体现数据的非均匀分布程度。在电力通信系统中，与数据的非均匀程度相比更加关注告警与实际场景的一致性（告警归并结果直接影响后续缺陷处理），因此借助缺陷和告警簇的分布情况通过外部指标来评价归并结果是否准确且完备。

根据以上的讨论，本文中告警归并任务要求将可能由同一个缺陷引起的告警关联在一起。本文使用的数据包括缺陷单数据和告警数据，经过告警流水号进行数据关联。这意味着，同一缺陷单关联的告警应当被归并在一起，且不同缺陷单关联的告警不应被归并在一起。告警归并的评估应该建立在归并集合以及实际缺陷单相关告警数据的一致性评估基础上。本文选择了 V-measure[18]方法进行有效性评估。

$$h=1-\frac{H(C|K)}{H(C)} \tag{3}$$

$$c=1-\frac{H(K|C)}{H(C)} \tag{4}$$

$$v=\frac{2\times h\times c}{h+c} \tag{5}$$

其中，$H(C|K)$ 是给定簇划分条件下类别划分的条件熵，$H(C)$ 是类别划分熵，$H(K|C)$ 是给定类别划分条件下的簇划分的条件熵，即：

$$H(C|K)=-\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n}\log\left(\frac{n_{c,k}}{n_k}\right) \tag{6}$$

$$H(C)=-\sum_{c=1}^{|C|}\frac{n_c}{n}\log\left(\frac{n_c}{n}\right) \tag{7}$$

$$H(K|C)=-\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n}\log\left(\frac{n_{c,k}}{n_c}\right) \tag{8}$$

其中，N 表示实例总数，n_c 表示类别 c 下的实例数，n_k 表示簇 k 下的实例数，$n_{c,k}$ 表示类 c 中被划分到簇 k 的实例数。基于规则匹配、*k*-means[2]、DBSCAN 的告警归并方法的特性和效果对比见表 1。其中 *k*-means 方法需要业务专家的先验知识推测出可能发生的缺陷个数进行预设簇的数目，其他方法不需要进行预设簇；规则匹配方法不具备自学习能力，只能够在规则中学习，而 *k*-means 方法与 DBSCAN 方法可以从数据中进行自学习。

表 1　不同告警归并方法的特性和效果对比

方法	预设簇	自学习	h-score	c-score	v-score
规则匹配	×	×	0.99	0.97	0.98
k-means	√	√	0.96	0.93	0.94
DBSCAN	×	√	**0.99**	**0.99**	**0.99**

h-score、c-score 和 v-score 分别表明了归并结果的同质性、完整性和同质性与完整性的调和平均值，取值为 0 到 1，取值为 1 时为最理想结果。

可以直观地看出，几种方法在信息熵上的表现都能够有效消除不确定性。其中在同质性表现上，规则匹配和 DBSCAN 方法表现最佳，在完整性表现上，DBSCAN 方法表现最佳，综合考虑同质性与完整性的表现，DBSCAN 方法表现最佳且性能表现具有可解释性。*k*-means 方法在缺陷具有突发性的前提中并不适用，因此性能表现都不太理想。规则匹配方法得到的归并结果基于人工经验，因此归并的结果同质性较高，但是对于规则以外的模式无法进行捕获因此完整性欠缺。本文基于无监督聚类的告警归并方法在消除不确定性上表现更强，具有自学习能力，不需要预先人为预测缺陷数目。

基于规则匹配、*k*-means、DBSCAN 的告警归并方法的缺陷一致性对比见表 2。

（1）归并与缺陷一对一

表明归并集合中仅包含一个缺陷且一个缺陷对应的告警被归并到了同一个集合中。归并与缺

表 2　不同告警归并方法的缺陷一致性对比

方法	集合数目	归并与缺陷			
		一对一	一对多	多对一	多对多
规则匹配	163	76.71%	7.53%	13.70%	2.05%
k-means	146	58.71%	30.32%	7.74%	3.23%
DBSCAN	**153**	**84.93%**	**10.96%**	**3.42%**	**0.68%**

陷一对一表明告警被正确归并，显然本文所采用的 DBSCAN 方法显著优于其他方法。

（2）归并与缺陷一对多

表明归并集合中包含多个缺陷但一个缺陷对应的告警被归并到了同一个集合中。归并与缺陷一对多表明部分集合被划分得过大，可以通过细化集合来降低该比例。

（3）归并与缺陷多对一

表明归并集合中仅包含一个缺陷但一个缺陷对应的告警被归并到了多个集合中。可见 DBSCAN 方法比起单纯的规则匹配降低了更多归并与缺陷多对一比例，提高了归并与缺陷一对一比例。

（4）归并与缺陷多对多

表明归并集合中包含多个缺陷且一个缺陷对应的告警被归并到了多个集合中。本 DBSCAN 方法在归并与缺陷多对多上占比最小，表现最优。

集合数目对应着归并告警集合数，也就是对应方法预测的缺陷数目。在集合的数目上，*k*-means 算法需要提前预设集合数目才能运行，预设集合数目设置为缺陷单数目 146，因此生成集合的数目与缺陷总数保持一致，而其他方法生成集合的数目与实际缺陷数目有偏差。除了 *k*-means 方法之外，其他方法不需要设定集合数目，因此集合的数目与实际缺陷数目的一致性部分显示了归并方法的准确性。

评估使用了 146 个缺陷单数据，其中 12 个缺陷单数据存在重复派单的现象，因此归并与缺陷一对多的比例较高。评估数据中的重复派单现象主要来源于：（1）实际环境中缺陷没有得到及时发现和消缺导致一段时间后告警再次产生，由于告警之间时间间隔较长，单一缺陷被归为多个缺陷单；（2）多个站点的共享线路或设备发生缺陷，基于人工或者规则的缺陷诊断将其判断为多个缺陷归档。本系统的自动派单会将算法得到的缺陷单向前归并到已产生但未处理完毕的缺陷单当中，遏制（1）导致的重复派单；相近时间内具有共享线路或设备的多站点缺陷会被归并到同一缺陷中，定位缺陷为该共享线路或设备，遏制（2）导致的重复派单。人工复查证明了算法结果有效核验了原始缺陷单数据，发现了原始缺陷单数据中的重复派单数据。综上，本文提出的基于 DBSCAN 的告警归并方法在归并与缺陷一致性表现上更强，不需要预设集合数目且生成集合与实际缺陷数目较为一致。

4.2　缺陷诊断的有效性验证

频繁子图挖掘得到的待标记模式和缺陷类型的相关程度如图 5 所示，基于 146 个缺陷单及其相关告警数据基于 DBSCAN 完成告警归并后，对于归并集合进行频繁子图挖掘得到的待标记的缺陷模式集合与实际缺陷单数据之间的分布一致性结果。验证实验中制定了 4 种缺陷类型，fiber breaking、power interruption、card abnormal 以及 power abnormal，分别对应的物理意义为光缆类故障、供电设备中断、板卡类故障以及供电设备故障。图 5（a）～（d）给出了子图模式分别与 4 种缺陷类型的相关程度。其中，峰值表示该模式与对应缺陷类型之间存在强相关性。其中图 5（a）和图 5（b）出现了多个峰值，相关程度在 0%～100%，说明 fiber_breaking 和 power_interruption 与多种模式相关，且模式较为复杂，模式间可能存在交叉；图 5（c）出现了多个峰值，相关程度基本只分布在 0%和 100%两个点，说明 card_abnormal 模式简单，但是存在多种模式；图 5（d）只出现了单峰值，说明 power_abnormal 只与单一模式高度相关，其结果与实际环境一致。其中，具有强相关性（图

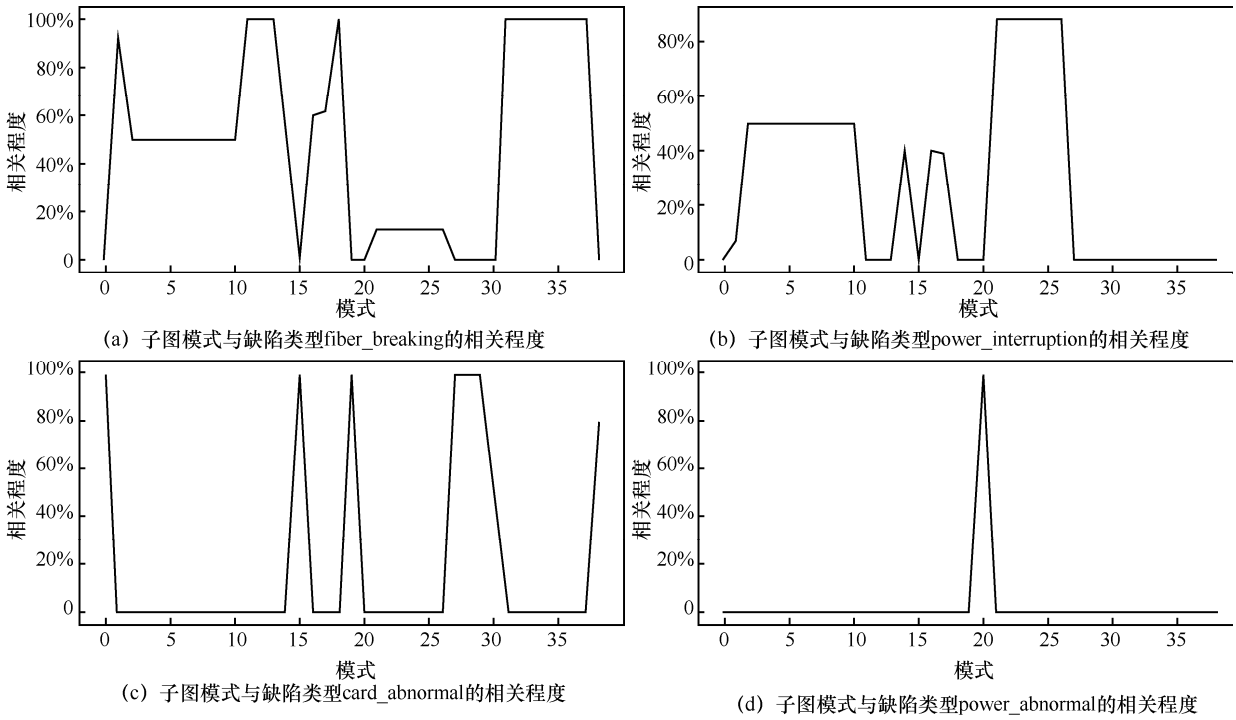


图 5　频繁子图挖掘得到的待标记模式和缺陷类型的相关程度

中相关度为 100%）的待标记的缺陷模式在经过人工审核之后往往是对应缺陷类型下的关键模式。这意味着在没有人工参与的情况下，本文所提供的算法既可以自动化发现可能存在的缺陷模式，同时也可以给予人工标注建议，能够准确捕捉数据之间的相关性，并且可以准确区分不同类型数据。

本文提出的架构已经实际部署在缺陷诊断及自动化派单系统中，基于 gSpan 挖掘得到频繁子图并且经过人工标注选出关键模式之后进行图匹配得到的缺陷诊断混淆矩阵见表 3。混淆矩阵的每一列代表了真实类别，每一列的总数表示预测为该类别数据的数目；每一行代表了数据的预测归属类别，每一行的数据总数表示该类别数据实例的数目。缺陷种类分为 card abnormal、fiber breaking、power abnormal、power interruption。其中，card abnormal 和 power abnormal 预测结果和真实结果完全一致，有两个 fiber breaking 被预测为 power interruption，一个 power interruption 被预测为 fiber breaking。也就是在 card abnormal 以及

表 3　缺陷诊断混淆矩阵

预测	真实			
	card abnormal	fiber breaking	power abnormal	power interruption
card abnormal	28	0	0	0
fiber breaking	0	60	0	1
power abnormal	0	0	12	0
power interruption	0	2	0	13

power abnormal 的缺陷诊断任务上准确率为100%，fiber breaking 类别中 62 个缺陷中有 2 个缺陷诊断错误，准确率为 96.8%，power interruption 类别中 14 个缺陷有 1 个缺陷诊断错误，准确率为 92.9%。实验结果表明本文提出的缺陷诊断方法能够获得较高的准确性。

5 结束语

本文提出的面向电力通信网的缺陷检测和自动派单方法，基于无监督聚类和频繁子图挖掘算法，提供了一个具有自学习和迭代更新能力的架构。该架构为将无监督学习和数据挖掘等人工智能技术引入电力通信领域，减轻了运维压力，降低了人力资源投入，提升了系统安全性和可靠性。架构重点在于告警归并和缺陷诊断及自动派单模块，模块间功能清晰、相互独立，提供向外暴露的接口，具有良好的可扩展性，允许扩展为其他可行算法。本文中两大模块分别基于无监督聚类算法 DBSCAN 和频繁子图挖掘算法 gSpan，摆脱了传统缺陷诊断方法对于人工规则的依赖，并在实验中取得了良好的性能表现。实验结果证明了该架构及其基础算法的可实施性和可部署性，对电力通信网络的智能化进程有一定的理论指导意义。

参考文献：

[1] GARDNER R D, HARLE D A. Methods and systems for alarm correlation[C]//Proceedings of Proceedings of GLOBECOM'96. 1996 IEEE Global Telecommunications Conference. Piscataway: IEEE Press, 1996: 136-140.

[2] MAZDZIARZ A. Alarm correlation in mobile telecommunications networks based on k-means cluster analysis method[J]. Journal of Telecommunications and Information Technology, 2018(2): 95-102.

[3] WEN L, LI X Y, GAO L, et al. A new convolutional neural network-based data-driven fault diagnosis method[J]. IEEE Transactions on Industrial Electronics, 2018, 65(7): 5990-5998.

[4] XIAO F, ZHAO Y, WEN J, et al. Bayesian network based FDD strategy for variable air volume terminals[J]. Automation in Construction, 2014(41): 106-118.

[5] WANG J Y, JING Y H, QI Q, et al. ALSR: an adaptive label screening and relearning approach for interval-oriented anomaly detection[J]. Expert Systems With Applications, 2019(136): 94-104.

[6] QI Q, SHEN R Y, WANG J Y, et al. Spatial-temporal learning-based artificial intelligence for IT operations in the edge network[J]. IEEE Network, 2021, 35(1): 197-203.

[7] JAKOBSON G, WEISSMAN M. Alarm correlation[J]. IEEE Network, 1993, 7(6): 52-59.

[8] SCHUBERT E, SANDER J, ESTER M, et al. DBSCAN revisited, revisited[J]. ACM Transactions on Database Systems, 2017, 42(3): 1-21.

[9] YANG Y C, WANG Y P, WEI Y. Adaptive density peak clustering for determinging cluster center[C]//Proceedings of 2019 15th International Conference on Computational Intelligence and Security (CIS). Piscataway: IEEE Press, 2019: 182-186.

[10] BOULOUTAS A T, CALO S, FINKEL A. Alarm correlation and fault identification in communication networks[J]. IEEE Transactions on Communications, 1994, 42(234): 523-533.

[11] YOUSUF H, ZAINAL A Y, ALSHURIDEH M, et al. Artificial intelligence models in power system analysis[M]//Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications. Cham: Springer International Publishing, 2020: 231-242.

[12] DARRAB S, ERGENC B. Vertical pattern mining algorithm for multiple support thresholds[J]. Procedia Computer Science, 2017(112): 417-426.

[13] HARTMANIS J. Computers and Intractability[EB]. SIAM Review, 1982.

[14] YAN X F, HAN J W. gSpan: graph-based substructure pattern mining[C]//Proceedings of 2002 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2002: 721-724.

[15] YAN X F, HAN J W. gSpan: graph-based substructure pattern mining[C]//Proceedings of 2002 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2002: 721-724.

[16] XIONG H, LI Z M. Clustering validation measures[M]//Data Clustering: Chapman and Hall/CRC, 2018: 571-606.

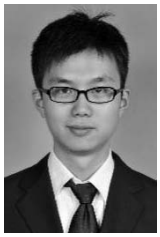
[17] HOU J, LIU W X. Evaluating the density parameter in density peak based clustering[C]//Proceedings of 2016 Seventh International Conference on Intelligent Control and Information Processing (ICICIP). Piscataway: IEEE Press, 2016: 68-72.

[18] NOWOSAD J, STEPINSKI T F. Spatial association between regionalizations using the information-theoretical V-measure[J]. International Journal of Geographical Information Science, 2018, 32(12): 2386-2401.

[作者简介]

吴季桦（1998– ），女，北京邮电大学计算机学院硕士生，主要研究方向为云原生、知识图谱、子图挖掘。

朱鹏宇（1992– ），男，国网电力科学研究院有限公司工程师，主要研究方向为电力通信、人工智能、知识图谱。



吴子辰（1988– ），男，国网江苏省电力有限公司信息通信分公司高级工程师、信通调控中心副主任，主要研究方向为电力通信技术。

顾彬（1983– ），男，博士，国网江苏省电力有限公司信息通信分公司高级工程师，主要研究方向为电力通信技术。

洪涛（1994– ），男，国网江苏省电力有限公司信息通信分公司工程师，主要研究方向为电力光纤通信、计算机网络安全、人工智能技术等。

郭波（1977– ），男，国网江苏省电力有限公司信息通信分公司高级工程师、副总工程师，主要研究方向为电力信息通信技术。

王晶（1974– ），女，北京邮电大学计算机学院副教授，主要研究方向为业务网络、云网络、网络智能等。

王敬宇（1978– ），男，博士，北京邮电大学计算机学院教授、博士生导师，主要研究方向为智能网络、智能运维、边缘计算等。