Research on Integrated Mining Method for Edge Clustering Data in Complex Sensor Networks*

GENG Dezhi*, LI Junli
(Department of Information Technology and Engineering, Jinzhong College, Jinzhong Shanxi 030600, China)

Abstract: The existing edge clustering data mining methods of complex sensor networks do not consider the heterogeneity of data, which is easy to cause edge information interference, resulting in low mining accuracy. An integrated mining method for edge clustering data in complex sensor networks is proposed based on *K*-Means. Firstly, the weights of data edge clustering are determined through graph theory knowledge to reduce the interference of irrelevant edge information on data mining parameters caused by the heterogeneity of complex sensor network data. Then, data mining parameters for complex sensor networks are obtained based on the above operation, providing reliable input for integrated mining of edge clustering data in complex sensor networks. Finally, the similarity of the side vectors is calculated, the region radius is determined, the cluster center is obtained, and integrated mining by improving is achieved the *K*-Means clustering algorithm. The results show that the proposed method is more accurate in mining the minimum intra cluster distance of complex network data, and the mining accuracy of the proposed method for complex networks is higher than 98%.

Key words: complex sensor network; edge clustering data; data integration mining; weight; similarity degree; *K*-Means clustering
EEACC: 6140; 6150P **doi**: 10.3969/j.issn.1004-1699.2025.11.024

复杂传感网络边聚类数据集成化挖掘方法研究*

耿德志*, 李俊丽
(晋中学院信息技术与工程系, 山西 晋中 030600)

摘 要: 现有复杂传感网络边聚类数据挖掘方法未考虑数据异构性, 容易造成边缘信息干扰, 导致挖掘精度较低。提出一种基于 *K*-Means 的复杂传感网络边聚类数据集成化挖掘方法。首先, 通过图论知识确定数据边聚类的权值, 以减少因复杂传感网络数据异构性所造成的不相关边缘信息对数据挖掘参数的干扰。然后, 以上一操作为基础, 获取复杂传感网络数据挖掘参数, 为复杂传感网络边聚类数据集成化挖掘提供可靠的输入。最后, 计算边向量的相似度, 确定区域半径, 获取聚类中心, 通过改进 *K*-Means 聚类算法实现集成化挖掘。结果表明, 所提方法挖掘复杂网络数据的最小化簇内距离较为准确, 将该方法应用于复杂网络中, 挖掘精度高于 98%。

关键词: 复杂传感网络; 边聚类数据; 数据集成化挖掘; 权值; 相似度; *K*-Means 聚类
中图分类号: TP393 **文献标识码**: A **文章编号**: 1004-1699(2025)11-2100-07

复杂传感网络是由大量智能传感器和通信节点组成的网络, 能够收集和传输各种环境信息。随着传感器和通信技术的进步, 复杂传感网络在实际应用中表现出巨大的潜力[1-2]。在复杂传感网络中, 传感节点之间的连接关系可以表示为图的边[3], 边聚类数据集成化挖掘的任务旨在将节点之间相似或相关的边分组或聚类在一起, 由此能够整合和分析来自不同传感器节点的数据, 实现复杂传感网络数据的高效利用, 并提取其中潜在的有用信息, 进而提高数据的价值和利用效率[4-6]。而相对于传统的数据集成化挖掘任务, 复杂传感网络涉及多个传感器和通信节点, 其数据类型和格式可能各不相同, 包括数值、文本、图像等, 具有数据异构性的特征, 给数据集成化挖掘带来了一系列的挑战。因此, 为了更好地利用和应用复杂传感网络中的数据, 结合聚类方法研究数据挖掘问题, 提高对大型复杂传感网络数据分类能力, 以发现更有价值的数据, 具有一定的研究意义。

聚类方法分析在相关领域是非常重要的[7]。对于复杂传感网络, 不论其规模大小, 簇内部的节点间都存在着一定程度的联系。通过分析和研究复杂传感网络中的聚类特征, 就可以得到网络中节点与节点之间存在的某种关系, 有助于更好地理解和把握整个网络的结构, 为发现有价值数据提供重要参考和指导。当前对于复杂传感网络聚类算法可以分为两种基本类别: 首先是基于聚类优化的方法[8]; 再者就是启发式感知方法[9]。其中, 聚类优化的方法通常利用某种聚类准则, 通过迭代的方式来调整簇的分配; 后者则是通过一系列启发式规则和技巧

项目来源: 山西省基础研究计划青年科技研究基金项目(20210302123356)
收稿日期: 2023-11-27 **修改日期**: 2023-12-18

来寻找满足特定标准的聚类结果。两种方法均可有效解决复杂传感网络聚类问题。针对数据挖掘问题,相关学者也给出了一些较好的成果。

文献[10]提出基于人工智能神经网络的数据挖掘方法。该方法预处理了数据,并且采用关联法计算输入输出神经元的关联,从而提取数据特征,并根据相关度排序 RBF 神经网络节点,基于此展开数据挖掘。将该方法应用到复杂传感网络中,其可有效提高计算效率。但是容易受到边缘信息干扰,导致该方法的搜索精度较低,数据挖掘效果不佳。文献[11]通过获取不同属性数据的度量值,并且融合相关性的数据,结合自适应滤波方法抑制上行链路干扰,实现数据挖掘优化。该方法应用到复杂传感网络中,具有较快的挖掘速度,但却忽略了边缘信息干扰,导致数据挖掘精度较差。文献[12]提出自编码网络数据挖掘方法。该方法通过无向图与标准化实现谱聚类,并利用自编码网络,获取各维度隐藏特征,实现数据挖掘。但将该方法应用到复杂传感网络中,也容易受到边缘信息干扰,导致挖掘精度较低,且挖掘过程中簇内距离较长,挖掘效率较低。文献[13]提出混合入侵数据挖掘方法。该方法通过设计混合入侵数据检测流程,提取混合入侵数据特征,并计算数据特征邻域距离、可达距离和可达密度,基于计算结果,实现数据挖掘。该方法在复杂传感网络的应用中,由于未考虑干扰问题,导致数据挖掘结果存在较大偏差。

针对以上问题,本文提出一种复杂传感网络边聚类数据集成化挖掘方法。该方法主要通过图论知识确定数据边聚类的权值,有效解决了因复杂传感网络数据异构性所造成的不相关边缘信息对数据挖掘参数的干扰;并获取了数据挖掘参数,计算了边向量相似性,考虑到 *K*-Means 聚类算法对孤立点的敏感性,确保挖掘效果,基于密度思想改进 *K*-Means 聚类算法;将挖掘参数作为输入,依据边向量相似性,通过改进 *K*-Means 聚类算法实现集成化挖掘,有效提升复杂传感网络数据集成化挖掘效果。

1　复杂网络数据挖掘参数的获取

1.1　基于图论的复杂传感网络数据权值确定

复杂传感网络中的数据量通常很大,并且其数据类型和格式存在差异性,具有数据异构性的特征。复杂传感网络是一种复杂的网络结构,通常由大量的节点和边组成,用于在不同的传感器之间收集、传输和处理数据。在复杂传感网络中,数据的来源可以是不同类型和格式的传感器,这些传感器可能采集不同种类的数据,如温度、湿度、压力等。由于数据类型和格式的差异,数据之间可能存在异构性,因此在进行数据挖掘或分析时需要考虑如何有效地处理和比较这些异构的数据。基于图论的数据权值确定方法可以通过赋予不同的权值来充分考虑数据的重要性和相似性程度,从而进行有效的边聚类。这样可以减少因数据异构性导致的不相关边缘信息对数据挖掘结果的干扰,提高聚类结果的准确性和适应性,增强数据集成化挖掘的效果。

图论知识可将任何复杂传感网络抽象成一个图 $G=(V,E)$,在该图中,网络中的个体被看作为节点 V,而个体间的关系看作为边 E,通过分析节点集合中任意两个节点之间的连接,得到不同类型的网络属性,这为进一步研究网络中的信息提供一种有效的实现方法。在图论知识下,如果两个节点之间的连接没有方向性,即连接是双向的,那么称该网络为无向网络;相反,如果连接有方向性,则称该网络是有向网络。此外,如果网络中的每条边都带有一个值,这个值可以被称为边的权重,从而使网络成为有权网络;相反,如果所有边的权重都相同,或者都被视为 1,那么网络为无权网络。

在图论中,通过邻接矩阵表示拓扑结构信息,这种表示方法可以完成整个图的识别[14]。邻接矩阵是一个方阵,其中矩阵元素代表图中的一个节点,矩阵的值表示对应节点之间的连接关系。如果两个节点之间存在边,则相应的矩阵元素为 1,带有权重值 a_{ij},其计算公式为:

$$a_{ij}=\frac{x_i x_j}{\sum_{i,j=1}^{N} x_i x_j} \tag{1}$$

式中:x_i 为复杂传感网络抽象成一个图 $G=(V,E)$ 的节点的数据。

如果两个节点之间不存在边,则相应的矩阵元素为 0,无权。由此通过邻接矩阵,以获取关于节点之间连接情况,以确定网络权值。一般情况下,复杂传感网络图 $G=(V,E)$ 中的邻接矩阵,可用 $\mathbf{graph}=(a_{ij})_{N\times N}$ 描述,则 i 和 j 两节点相连时,$a_{ij}\geqslant 1$,则其表示为节点度参数;$1>a_{ij}>0$,其表示为聚集系数;反之,$a_{ij}=0$,其表示为网络效率。邻接矩阵 $\mathbf{graph}=(a_{ij})_{N\times N}$ 为:

$$\mathbf{graph}=(a_{ij})_{N\times N}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \cdots & \cdots & a_{ij} & \cdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix} \tag{2}$$

通过式(2)的邻接矩阵的元素确定复杂传感网络数据权值,公式为:

$$A_{ij}=\frac{\mathrm{cov}(a_{ij},a_{i+1,j+1})}{\mathrm{std}(a_{ij})\times\mathrm{std}(a_{i+1,j+1})} \quad (3)$$

$$a_{ij},a_{i+1,j+1}\in \mathbf{graph}$$

式中:std(·)为标准差函数。

基于复杂传感网络数据权值,确定复杂传感网络数据挖掘参数,并获得复杂传感网络中的边向量的相似度,为复杂传感网络边聚类数据集成化挖掘提供可靠的输入,实现有效精准的数据集成化挖掘。

1.2 复杂传感网络数据挖掘参数计算

基于确定后的复杂传感网络数据权值,排除干扰信息,有针对性地计算复杂传感网络数据挖掘参数,为复杂传感网络边聚类数据集成化挖掘提供可靠的输入,以获得精准有效的复杂传感网络对象半径,排除孤立边和边界边,为完成边聚类数据集成化挖掘奠定基础。复杂传感网络的数据挖掘参数计算具体描述如下:

①节点度

在复杂传感网络中,节点的度表示与该节点直接相连的节点的数量。基于此,利用邻接矩阵,可以计算每个节点的度,其等于其所在行(或列)所有非零值的和。在有权网络上,节点度分为入度和出度,入度边的方向指向该节点,而出度边的方向指向 1 步邻居节点。节点的入度表达式为:

$$d_{i,\mathrm{in}}=\sum_{i\in\boldsymbol{\Gamma}_i}A_{ij} \quad (4)$$

式中:$\boldsymbol{\Gamma}_i$ 为节点 i 的直接邻居集合,$a_{ij}=1$ 为节点 i 和 j 存在一条边相连,且边的方向指向节点 i。节点的出度 $d_{i,\mathrm{out}}$ 为:

$$d_{i,\mathrm{out}}=\sum_{j\in\boldsymbol{\Gamma}_i}A_{ij} \quad (5)$$

式中:连接 i 与 j 两节点边的方向指向节点 j。基于节点的入度和出度获得节点的平均度,其表达式如下:

$$\langle k\rangle=\frac{1}{N}\sum_{i=1}^{N}d_{i,\mathrm{in}}+d_{i,\mathrm{out}} \quad (6)$$

式中:N 为点总数。

②网络效率

在不完全连通的复杂传感网络中存在一些孤立节点,孤立节点与网络中的其他节点之间不存在直接或间接的连接关系。孤立节点的存在导致在计算节点间的最短距离时出现较大偏差,因为孤立节点到网络中其他节点的距离非常大,导致节点间最短距离无法计算。因此,利用网络效率 E 作为平均路径长度,通过 E 表示最短距离,表达式如下。

$$E=\frac{1}{N(N-1)}\sum_{i,j\in N,i\neq j}\frac{1}{L_{ij}} \quad (7)$$

式中:L_{ij} 表示节点 i 与节点 j 间的距离。

③聚集系数

聚集系数分为节点和网络两种聚集系数,其中,节点聚集系数计算公式为:

$$C(i)=\frac{\chi(\chi-1)}{3E_i} \quad (8)$$

式中:χ 为实际边的数量;E_i 为最大边数量。

当节点的聚集系数与最大边数量相等时,说明该节点的邻居节点之间构成了一个完全团。

通过计算每个节点的聚集系数,并求取所有节点的聚集系数均值,得到准确的复杂传感网络聚集系数。网络聚集系数可以评估整个网络中节点之间的紧密程度、连接模式以及网络的结构特点。聚集系数计算公式为:

$$\langle C\rangle=\frac{1}{N}E_i\sum_{i=1}^{N}C(i) \quad (9)$$

式中:N 节点总数。

复杂传感网络聚集系数$\langle C\rangle$取值在 0 到 1 范围内,当$\langle C\rangle$的值接近 1 时,说明网络是一个全耦合网络,且 $C(i)=1$,节点的任意两个邻居间都相连。反之接近 0 时,说明网络中所有节点都没有直接的连接,每个节点都是孤立的且 $C(i)=0$。若一个复杂传感网络具有较高的聚类系数,说明这个网络有明显的簇结构。

2 复杂传感网络的边聚类数据集成化挖掘方法设计

根据上述内容所获得的挖掘参数,将其作为边聚类数据集成化挖掘方法的输入,计算复杂传感网络中对象半径,排除孤立边和边界边,确定聚类中心。然后基于复杂传感网络数据权值,实现边向量相似度的计算,完成边分簇,以此为依据,实现复杂传感网络边聚类数据集成化挖掘。

在复杂传感网络中,不同节点之间可能存在大量的连接关系。为获得更全面和准确的边关联信息,通过权重计算节点的相似性,从而推导出边的相似性,以此分簇,具体设计如下:

假设一对边分享同一节点,即复杂传感网络中的边 $\boldsymbol{e}_{ik}$ 与 $\boldsymbol{e}_{jk}$ 有共享节点 k,称 k 为主节点,i、j 节点为进入节点。则包含邻居的一个节点 g 定义为:$n_+(i)=\{g\mid d(i,g)\leqslant 1\}$,$d(i,g)$ 表示 i 与 g 间的最短路径长度。由此,通过确定后的复杂传感网络数据权值,复杂传感网络中边向量的相似度可以定义为 $S(\boldsymbol{e}_{ik},\boldsymbol{e}_{jk})$,其计算公式如下:

$$S(\boldsymbol{e}_{ik},\boldsymbol{e}_{jk})=\frac{a_{ij}\mid n_+(i)\cap n_+(j)\mid}{\mid n_+(i)\cup n_+(j)\mid} \quad (10)$$

依据边向量的相似度完成分簇,为进一步提高边聚类的数据集成化挖掘效果,依据簇间的重叠系数合并相邻的簇为 $D=(T_i,T_j)$,消除簇间的边界,提高数据的集成化程度,并且起到过滤数据的目的。

假定复杂网络中的簇 T_i 与 T_j 的重叠系数为 $S=\frac{|T_i\cap T_j|}{|T_i\cup T_j|}$,其中,$T_i\cap T_j$ 为两个簇公共节点的数目,$T_i\cup T_j$ 为两个簇所有节点的数目和。

为减小孤立节点对 *K*-Means 聚类算法的影响,以提高其稳健性和准确性,改进 *K*-Means 聚类算法。利用式(7)获取的复杂传感网络中边的相似性计算结果,依据改进的 *K*-Means 聚类算法实现复杂传感网络中数据的集成化挖掘。基于密度思想改进后的 *K*-Means 聚类算法流程如下:

输入:将 $\boldsymbol{D}'$ 作为复杂传感网络数据集合,即 $\boldsymbol{D}'=\{\langle k\rangle,E_i,C(i),\langle C\rangle\}$,最小邻域点数 MinPts。

输出:边聚类结果。

①根据复杂传感网络数据集合 $\boldsymbol{D}'$ 输入的最小邻域点数 MinPts 计算出复杂传感网络中对象半径 ε,计算公式表示为:

$$\varepsilon=\sqrt{\frac{E_iV\times\text{MinPts}\times\Gamma\{\pi^n\times[C(i)\langle C\rangle n+1]\}}{\langle k\rangle m\times\sqrt{\pi^n}}} \tag{11}$$

式中:m 为对象的数量,n 为维度数量,Γ 为伽马函数,V 为数据集 $\boldsymbol{D}'$ 的容积。

②计算出数据集中每条边间的距离,将其存储在相异度矩阵,再选择一个起始边,计算起始边与其他所有边的距离 $D(\boldsymbol{e}_{ik},\boldsymbol{e}_{jk})$,依据 ε 值,查找满足小于 ε 范围内的边,若在 ε 内含有超过 MinPts 数目的点,则该对象为核心对象;反之小于 MinPts,且至少有一个核心对象在其邻域之内的为边界对象;不满足上述条件的则均为孤立对象。由此完成核心对象的搜寻,将孤立边和边界边从数据集中删除。假设边界边个数为 m_b,孤立边个数为 m_o,则删除后数据集 $\boldsymbol{D}'$ 中边聚类数据对象核心边的个数为 $R=(m-m_b-m_o)$。

③取每个簇中核心边的中心为聚类中心 R^n。

④执行 *K*-Means 聚类算法,输出复杂传感网络边聚类结果。

依据确定的复杂传感网络数据权值,获得边的相似性,通过聚类计算复杂传感网络中的边向量,将具有相似特征和属性的边向量划分到同一个集合,形成一个簇。利用改进的 *K*-Means 聚类算法实现复杂传感网络中的数据的集成化挖掘,则实现过程如下:

步骤 1　从复杂传感网络中任意选取 z 个聚类质心点 $\eta_1,\eta_2,\cdots,\eta_z\in\boldsymbol{R}^n$,作为复杂传感网络集成化聚类的初始中心;

步骤 2　对于每一个边向量 $\boldsymbol{D}_{i,j}$,计算其应该属于的类 $\boldsymbol{C}^n$:

$$\boldsymbol{C}^n=\arg\min_m\left\|\frac{\boldsymbol{D}_{i,j}-\eta_z}{S(\boldsymbol{e}_{ik},\boldsymbol{e}_{jk})}\right\| \tag{12}$$

步骤 3　对于复杂传感网络中的每一个类 $\boldsymbol{C}^n$,计算该类的质心:

$$\boldsymbol{\eta}_m=\sum_{\boldsymbol{D}_m\subset\boldsymbol{C}^n}\boldsymbol{D}_m/|\boldsymbol{C}^n| \tag{13}$$

步骤 4　设置复杂传感网络数据的集成化挖掘的目标函数,公式为:

$$\boldsymbol{J}(c,\boldsymbol{\eta})=\sum_{o=1}^{O}\sum_{\boldsymbol{D}_{i,j}\subset\boldsymbol{C}^n}\left\|\frac{\boldsymbol{D}_{i,j}^o-\boldsymbol{\eta}_m^o}{S}\right\| \tag{14}$$

式中:o 表示迭代次数。

如果 $\boldsymbol{J}\geqslant\varepsilon$,返回步骤 2,继续迭代;

步骤 5　若 $\boldsymbol{J}<\varepsilon$,输出此时划分的 $\boldsymbol{C}^n$;

步骤 6　输出复杂传感网络数据的集成化挖掘结果。

至此通过上述步骤实现复杂传感网络边聚类数据集成化挖掘方法设计。该方法首先为减少因复杂传感网络数据异构性造成不相关边缘信息的干扰,通过图论知识确定边聚类数据权值。其次,基于确定后的权值,有针对地计算挖掘参数及边向量相似性。然后,将挖掘参数作为输入,确定区域半径,获取聚类中心。最后,利用改进 *K*-Means 聚类算法,依据获得边向量相似性,划分复杂传感网络数据簇,进行集成化挖掘,输出挖掘结果,实现复杂传感网络数据的集成化挖掘。

3　仿真分析

3.1　方案设计

为了验证本文挖掘方法对于复杂传感网络数据的集成化挖掘性能,从网络中选取 8 个复杂传感网络作为仿真数据集,其中 2 号复杂传感网络节点分布如图 1 所示。

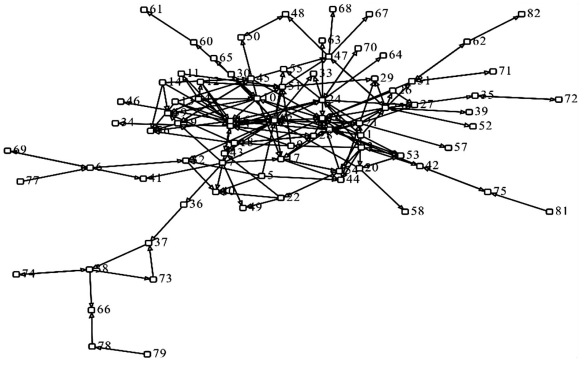

图 1　复杂传感网络节点分布图

各复杂传感网络基本信息如表 1 所示。

表 1 复杂传感网络基础数据

复杂传感网络序列号	节点数/个	边数/条	簇数目/团
1	8	16	2
2	11	18	3
3	12	23	2
4	13	24	2
5	19	58	3
6	18	26	4
7	11	27	3
8	16	32	2

基于上述信息,为验证所提方法的有效性,采用所提方法挖掘复杂传感网络数据集中的簇,分析簇挖掘结果与实际簇数量是否相同。为进一步验证所提方法的优越性,选取文献[10]的人工智能神经网络挖掘方法、文献[11]的上行链路干扰挖掘方法、文献[12]的自编码网络挖掘方法作为对比方法,将最小化簇内距离、模块度、挖掘精度作为评价指标,通过对比不同方法在同一数据集上的同一指标所得结果,验证本文方法的挖掘性能。为确保仿真所得结果的可靠性,对相关初始参数进行设置,具体如表 2 所示。

表 2 参数设置

参数	数值	参数	数值
密度阈值	0.5	迭代次数	100
邻域半径	0.1	聚类质心点数量	4

3.2 结果分析

3.2.1 复杂网络数据挖掘结果

采用本文方法对复杂传感网络数据集进行集成化挖掘,挖掘结果如表 3 所示。

表 3 列出了复杂传感网络数据集的真实挖掘划分结果,每一个中括号内都是当前复杂网络的一个簇,中括号的数量表示当前网络的簇数目。通过表 2 结果可以看出,采用本文方法可以实现复杂传感网络中数据的有效挖掘,证明本文方法对于复杂传感网络具有较高的挖掘性能。本文方法可以有效挖掘复杂传感网络数据集中的簇,簇挖掘结果与实际簇数量及内容均相同,表明本文方法的挖掘精度较高。这是因为所提方法在进行进行挖掘前,通过图论知识确定数据边聚类的权值,衡量了边的重要性和相关性,有效减少不相关边缘信息对数据挖掘结果的干扰,提高挖掘结果的准确性。

表 3 复杂传感网络数据综合挖掘结果

复杂传感网络序列号	簇数目/团	实际结果	挖掘结果
1	2	[1,2,3,][4,5,6,7,8]	[1,2,3,][4,5,6,7,8]
2	3	[1,2,3,4][5,6,7,8,9,][10,11]	[1,2,3,4][5,6,7,8,9,][10,11]
3	2	[1,2,3,4,5,6,7][8,9,10,11,12]	[1,2,3,4,5,6,7][8,9,10,11,12]
4	2	[1,2,3,4,5,6,7,8][9,10,11,12,13]	[1,2,3,4,5,6,7,8][9,10,11,12,13]
5	3	[1,2,3,4,5][6,7,8,9,10,11,12,13][14,15,16,17,18,19]	[1,2,3,4,5][6,7,8,9,10,11,12,13][14,15,16,17,18,19]
6	4	[1,2,3,4,5][6,7,8,9][10,11,12,13,14,15][16,17,18]	[1,2,3,4,5][6,7,8,9][10,11,12,13,14,15][16,17,18]
7	3	[1,2,3,4][5,6,7,8][9,10,11]	[1,2,3,4][5,6,7,8][9,10,11]
8	2	[1,2,3,4,5,6,7,8][9,10,11,12,13,14,15,16]	[1,2,3,4,5,6,7,8][9,10,11,12,13,14,15,16]

3.2.2 最小化簇内距离

最小化簇内距离是衡量不同方法挖掘性能的重要指标,通过最小化簇内距离,可以衡量划分至复杂传感网络中的相同簇内的数据间的相似度,统计 4 种方法的最小化簇内距离,统计结果如图 2 所示。

通过图 2 结果可以看出,采用本文方法挖掘复杂传感网络中的数据,所获取的最小化簇内距离明显低于另三种方法。采用本文方法挖掘复杂传感网络数据的最小化簇内距离低于 0.15,采用其他三种方法挖掘复杂传感网络数据的最小化簇内距离均高于 0.2。对比四种方法所得结果可知,采用本文算法挖掘复杂传感网络数据,挖掘结果中的簇内数据间距较小,可以令聚类获取的相同簇内的数据具有较高的相似性,具有优越的挖掘性能。这是因为所提

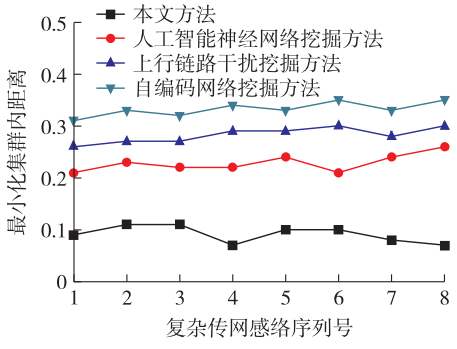

图 2 最小化集群内距离比较

方法在进行挖掘之前,为获得更全面和准确的边关联信息,通过节点的相似性推得边的相似性,确定了复杂传感网络边向量的相似性计算结果,为挖掘提供了可靠的支撑,使相似的边分配到同一个聚类中,有效降低了所得挖掘结果中的簇内数据的间距。

3.2.3　复杂传感网络挖掘结果的模块度对比

模块度是一种用于衡量簇结构质量的指标,其利用划分完成的网络与具有同原网络相同度序列但度不相关的配置模型进行对比,以度量发现的簇质量。其计算公式为:

$$Q=\sum_{c}^{n_c}\left[\frac{l_c}{M}-\left(\frac{d_c}{2M}\right)^2\right] \tag{15}$$

式中:n_c 为簇的数量,M 为复杂传感网络的总边数,l_c 为簇内总边数,d_c 为节点的总度数。模块度可用于评估簇结构的紧密程度,较高的模块度值表明簇内部边比随机预期更多,表示簇结构质量更好。则基于该指标,采用 4 种方法挖掘复杂传感网络数据的模块度,统计结果如表 4 所示。

表 4　不同方法的模块度

复杂传感网络序列号	模块度			
	本文方法	人工智能神经网络挖掘方法	上行链路干扰挖掘方法	自编码网络挖掘方法
1	0.825	0.680	0.401	0.524
2	0.811	0.713	0.432	0.593
3	0.922	0.372	0.446	0.589
4	0.918	0.724	0.431	0.611
5	0.833	0.402	0.438	0.609
6	0.832	0.710	0.620	0.403
7	0.807	0.401	0.667	0.609
8	0.846	0.346	0.438	0.618

将表 4 中模块度数据绘制成图,直观地观察数据的变化情况,如图 3 所示。

通过图 3 结果可以看出,采用本文方法挖掘复杂传感网络数据的模块度明显高于其他三种方法。本文方法挖掘复杂传感网络数据,挖掘结果模块度较高,表明本文方法挖掘结果中的簇结构更加明显,表明本文方法具有较高的挖掘性能。这是因为所提方法基于密度思想改进了 *K*-Means 聚类算法,通过确定区域半径、最小邻域点数来优化聚类结果,使得相同簇内的数据更好地归为一个聚类中心,提高了模块度,使簇结构更加明显。

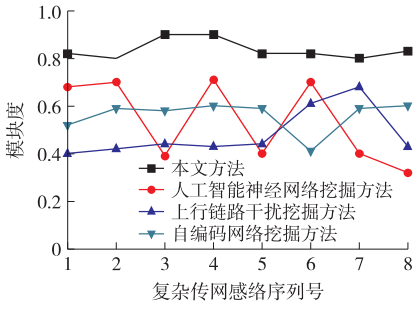

图 3　模块度对比结果

3.2.4　工业物联网应用分析

为了进一步验证本文方法的应用性,将该方法应用于某企业的工业物联网中。该网络连接方式采用 Wi-Fi 连接,通信协议为 Modbus TCP/IP。网络中数据构成的节点之间有一条边相连,如果两节点之间没有联系,则认为两节点之间没有边相连。由此,可将整个网络划分成不同的小圈子,也被称为簇。然而,在工业物联网中,某个节点存在出现多重身份的问题,导致簇之间出现重叠现象。因此,采用上述方法对其进行挖掘,以验证此方法是否可有效应对该现象,以实现精准的挖掘。应用的工业物联网结构如图 4 所示。

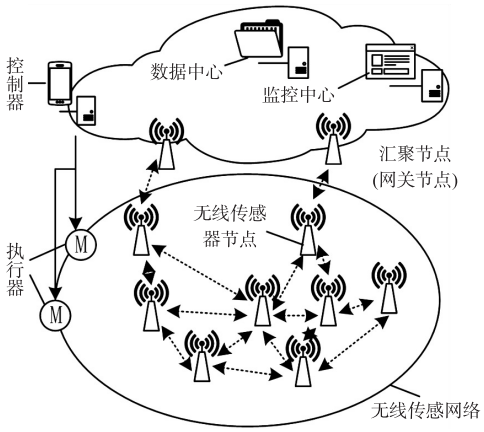

图 4　工业物联网结构

根据图 4 可知,该企业的工业物联网有 8 个无线传感器节点,这 8 个节点分别对应不同类型的工业传感数据,如压力、温度、电流等,将其依次编号为 1、2、3、4、5、6、7、8。基于该工业物联网,采用 4 种方法对工业物联网进行挖掘,分析不同方法的挖掘精度,如图 5 所示。

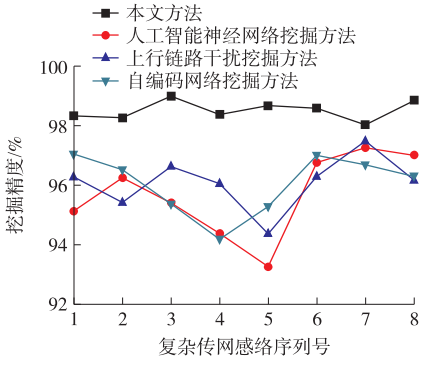

图 5　挖掘精度对比结果

通过图 5 结果可以看出,将本文方法应用于工业物联网的复杂网络中,挖掘精度高于 98%,采用其他三种方法挖掘工业物联网中的数据,挖掘精度均低于 98%。综上结果验证本文方法具有较高的挖掘精度,利用本文方法具有的挖掘精度,可以实现本文方法的良好应用。本文方法可有效应用于工业物联网中,其具有良好的挖掘性能。

4 结论

为提高边聚类数据集成化挖掘的精度,提出一种基于 *K*-Means 的复杂传感网络边聚类数据集成化挖掘方法。该方法通过图论知识确定边的复杂网络数据的权值,减少不相关边缘信息对数据挖掘结果的干扰,并基于此获取复杂传感网络数据挖掘参数。然后计算复杂传感网络相似性,确定区域半径以及最小邻域点数,通过改进 *K*-Means 聚类算法实现复杂网络中数据的集成化挖掘。仿真结果表明:①本文方法挖掘复杂传感网络数据的最小化簇内距离低于 0.15,表明该方法具有优越的挖掘性能。②本文方法挖掘复杂传感网络数据,挖掘结果模块度较高,表明该方法挖掘结果中的簇结构更加明显,具有较高的挖掘性能。③本文方法挖掘工业物联网中时,挖掘精度均高于 98%,表明该方法具有较高的挖掘精度。

参考文献:

[1] 邓鹏,李刚,马会茹,等. 离子强度响应光子晶体传感器的制备与性能[J]. 中国粉体技术,2020,26(5):28-32.

[2] 孙佳鑫,万振凯. CNT 纱线传感器配置算法在健康监测中的应用[J]. 兵器材料科学与工程,2022,45(2):129-133.

[3] 刘新娥,杜鹏. 透过式自组传感网络的设计及其在网格社区领域的应用[J]. 电子器件,2023,46(2):557-560.

[4] 党晓婧,张林,吕启深,等. 基于边聚类的电力通信网 Sybil 攻击检测算法[J]. 沈阳工业大学学报,2022,44(5):502-506.

[5] 王进,余薇,孙开伟,等. 基于聚类提升树的多标签学习[J]. 江苏大学学报(自然科学版),2021,42(4):428-437.

[6] 冯麟涵,杨俊杰,焦立启. 基于 RBF 神经网络的船舶冲击谱速度数据挖掘与预报[J]. 振动与冲击,2022,41(13):189-194,210.

[7] Wang F F,Hu H F. Coverage Hole Detection Method of Wireless Sensor Network Based on Clustering Algorithm [J]. Measurement,2021,179(5):109449.

[8] Thirumoorthy K,Muneeswaran K. A Hybrid Approach for Text Document Clustering Using Jaya Optimization Algorithm[J]. Expert Systems with Applications,2021,178(15):115040.

[9] Peng Z,Barzegar B,Yarahmadi M,et al. Energy-Aware Scheduling of Workflow Using a Heuristic Method on Green Cloud. Scientific Programming,2020(1):8898059.

[10] Zhang S,Duan C,Elhoseny M. Clustering Optimization Algorithm for Data Mining Based on Artificial Intelligence Neural Network[J]. Wireless Communications and Mobile Computing,2022(1):1304951.

[11] 冀江瑀,陈秀波. 上行链路干扰下双层网络中心数据挖掘方法[J]. 计算机仿真,2022,39(6):386-390.

[12] 王舒梵,严涛,姜新盈. 谱聚类欠取样下自编码网络不平衡数据挖掘[J]. 计算机系统应用,2021,30(10):331-335.

[13] 范玲楠. 多信道舰船无线移动通信网络混合入侵数据挖掘方法[J]. 舰船科学技术,2021,43(14):166-168.

[14] Xie J,Miao Q G,Liu R Y,et al. Attention Adjacency Matrix Based Graph Convolutional Networks for Skeleton-Based Action Recognition[J]. Neurocomputing,2021,440(14):230-239.

耿德志(1982—),男,汉族,山西交城人,硕士,晋中学院信息技术与工程系副教授,研究方向:智能信息处理、数据挖掘技术,2004 年参加工作。工作以来长期讲授《计算机基础》《计算机高级应用》等计算机公共课程以及《操作系统》《VB 程序设计》《数据挖掘导论》《数据库原理与应用》等专业课程,在计算机专业教学方面有着丰富的经验,参与过《大学计算机基础》教材的编写,主持省级研究生教改课题 1 项,参与校级教改课题 3 项。近年来在《山西大学学报自然科学版》《计算机仿真》《激光杂志》等高水平期刊发表学术及教改论文 10 余篇,出版学术专著 1 部,gengdezhi2000@ 163.com;

李俊丽(1982—),女,汉族,山西高平人,博士,晋中学院信息技术与工程系教授。长期从事数据挖掘与并行计算、机械产品生产大数据分析等领域的研究工作,目前主要利用 Spark 集群系统的强大数据处理能力,研究大数据挖掘方法和性能优化技术。近年来发表学术论文 20 余篇,其中 SCI 收录 3 篇,EI 收录 1 篇,参与国家自然科学基金项目 2 项。