



中文核心期刊
中国科技核心期刊
中国高校优秀科技期刊
RCCSE中国核心学术期刊
首届中国高校特色科技期刊

ISSN 1008-1542

CODEN HKDXFY
DOI:10.7535/hbkjdxsb



河北科技大学学报

JOURNAL OF HEBEI UNIVERSITY OF SCIENCE AND TECHNOLOGY

HEBEI KEJI DAXUE XUEBAO

2022 5

第43卷 第5期 Vol.43 No.5



河 北 科 技 大 学 学 报

第 43 卷 第 5 期 2022 年 10 月

目 次

国家青年/地区	纳米材料/环氧复合涂层的耐蚀性能研究进展
科学基金项目 张 巧, 吴若琳, 张仁辉 (449)
专 栏	单行星排混联式混合动力构型设计及性能验证 初镛坤, 杨 坤, 常依乐, 聂孟稳, 谭树梁 (461)
	基于改进 BAS-BPNN 的 Stewart 平台位姿正解 王淑良, 朱 浩, 赵明伟, 刘丽俊 (473)
	形内自相似层级类蜂窝面外冲击特性研究 李 响, 蔡明杰, 徐兴兴, 周绍国, 焦元辰 (481)
数 学	double-order Hilfer 分数阶共振边值问题解的存在性 孟凡猛, 江卫华, 郭春静 (495)
	基于 ψ - (h, r) -凹算子的非线性分数阶 (p, q) -差分方程的唯一迭代解 王菊芳, 王 斯, 禹长龙 (505)
机械、电子与信息科学	仿生低阻力医用缝合针设计与刺穿过程的数值模拟 潘 盼, 彭培英, 王立新 (516)
	基于模糊聚类的多类簇归属电力实体行为异常检测算法 郭禹伶, 左晓军, 崔景洋, 王 颖, 张光华 (528)
化学与化工	巯基功能化纤维素气凝胶材料的制备及其对 Cr(Ⅲ)的吸附 杨文玲, 张鹏瑞, 李虎林, 邸博洋, 郭清华, 崔洪誉 (538)
交通物流	融合多因素的“时间齿轮”交通流预测模型 ... 兰添贺, 曲大义, 陈 昆, 刘浩敏 (550)
土木建筑工程	GFRP 型材-木组合梁受弯性能试验研究 黄 鹏, 于海丰, 冯佳勋, 郝贵强, 李志强 (560)
编审园地	向本期载文的审稿专家致谢 本刊编辑部 (480)
编读往来	《河北科技大学学报》影响因子上升至 1.772 本刊编辑部 (封二)
	《河北科技大学学报》投稿须知 本刊编辑部 (封三)

JOURNAL OF HEBEI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Vol.43 No.5 Oct. 2022

CONTENTS

• Special Column: National Young Scholar/Local Science Foundation •

Research progress in corrosion resistance of nano-materials/epoxy composite coatings

..... ZHANG Qiao, WU Ruolin, ZHANG Renhui (449)

Configuration design and simulation verification of single planetary gearset power-split hybrid electric

vehicle CHU Yongkun, YANG Kun, CHANG Yile, NIE Mengwen, TAN Shuliang (461)

Forward position and posture solution of Stewart platform based on improved BAS-BPNN

..... WANG Shuliang, ZHU Hao, ZHAO Mingwei, LIU Lijun (473)

Study on out-of-plane impact performance of self-similar hierarchical quasi-honeycomb

..... LI Xiang, CAI Mingjie, XU Xingxing, ZHOU Shaoguo, JIAO Yuanchen (481)

• Mathematics •

Existence of solutions for double-order Hilfer fractional boundary value problems at resonance

..... MENG Fanmeng, JIANG Weihua, GUO Chunjing (495)

Unique iterative solution for nonlinear fractional (p, q) -difference equation based on ψ - (h, r) -concave

operators WANG Jufang, WANG Si, YU Changlong (505)

• Mechanical, Electronics and Information Science •

Design of biomimetic low-drag medical suture needle and numerical simulation on penetrating process

..... PAN Pan, PENG Peiying, WANG Lixin (516)

An abnormal behavior detection algorithm based on fuzzy clustering for multi-categories affiliation
of power entities

..... GUO Yuling, ZUO Xiaojun, CUI Jingyang, WANG Ying, ZHANG Guanghua (528)

• Special Column: Chemistry and Chemical Industry •

Preparation and adsorption of Cr(Ⅲ) on thiol functionalized cellulose aerogels

..... YANG Wenling, ZHANG Pengrui, LI Hulin, DI Boyang, GUO Qinghua, CUI Hongyu (538)

• Transportation Logistics •

"Time gear" traffic flow prediction model with multiple factors integration

..... LAN Tianhe, QU Dayi, CHEN Kun, LIU Haomin (550)

• Civil and Building Engineering •

Experimental study on flexural performance of GFRP profile-wood composite beams

..... HUANG Li, YU Haifeng, FENG Jiaxun, HAO Guiqiang, LI Zhiqiang (560)

文章编号:1008-1542(2022)05-0528-10

基于模糊聚类的多类簇归属电力实体行为 异常检测算法

郭禹伶¹, 左晓军¹, 崔景洋², 王颖¹, 张光华²

(1. 国网河北省电力有限公司电力科学研究院, 河北石家庄 050021; 2. 河北科技大学信息科学与工程学院, 河北石家庄 050018)

摘要: 针对数字化主动电网中电力实体行为复杂化、攻击手段隐蔽化等问题, 提出了一种基于模糊聚类的多类别归属异常检测算法。首先, 对电力实体行为相似性的度量方式进行优化, 并基于优化后的度量方法构建模糊聚类算法, 通过多次迭代得到实体行为对应各类型的隶属度矩阵; 其次, 根据类别软划分隶属度矩阵, 分别计算实体在各个类别内的近邻距离、近邻密度与近邻相对异常因子等参数; 最后, 分析实体在各类簇内的相对异常情况, 判断该电力实体行为是否属于异常行为。结果表明, 与 LOF, K-Means 和 Random Forest 算法相比, 新方法具有更高的异常行为检出数量和更优的异常检测评价指标, 解决了传统异常检测算法样本评价角度单一的问题, 进一步提高了数字化主动电网抵御未知威胁的能力。

关键词: 数据安全与计算机安全; 用户与实体行为分析; 数字化主动电网; 模糊聚类; 异常检测

中图分类号: TP393.0; TM769 **文献标识码:** A **DOI:** 10.7535/hbkd.2022yx05008

An abnormal behavior detection algorithm based on fuzzy clustering for multi-categories affiliation of power entities

GUO Yuling¹, ZUO Xiaojun¹, CUI Jingyang², WANG Ying¹, ZHANG Guanghua²

(1. State Grid Hebei Electric Power Research Institute, Shijiazhuang, Hebei 050021, China; 2. School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China)

Abstract: Aiming at the problems of complex behavior of power entities and concealed attack means in the digital active power grid, a multi-category attribution anomaly detection algorithm based on fuzzy clustering was proposed. Firstly, the similarity measurement method of power entity behavior was optimized, a fuzzy clustering algorithm was constructed based on the measurement value, and the membership matrix of entity behavior corresponding to various classes was obtained through

收稿日期:2022-03-21;修回日期:2022-04-26;责任编辑:王淑霞

基金项目:国家重点研发计划项目(2018YFB0804701);国家自然科学基金(62072239);河北省科技计划项目(20377725D)

第一作者简介:郭禹伶(1987—),女,河北石家庄人,高级工程师,硕士,主要从事网络安全、信息系统检测方面的研究。

通讯作者:张光华教授。E-mail:xian_software@163.com

郭禹伶,左晓军,崔景洋,等.基于模糊聚类的多类簇归属电力实体行为异常检测算法[J].河北科技大学学报,2022,43(5):528-537.

GUO Yuling,ZUO Xiaojun,CUI Jingyang,et al.An abnormal behavior detection algorithm based on fuzzy clustering for multi-categories affiliation of power entities[J].Journal of Hebei University of Science and Technology,2022,43(5):528-537.

several iterations. Secondly, the nearest neighbor distance, nearest neighbor density, and nearest neighbor relative anomaly factor of entities in each category were calculated according to the category softening membership matrix. Finally, the relative abnormal situation of the entity in various clusters was analyzed to judge whether the behavior of the power entity belongs to the abnormal behavior category. The results show that compared with LocalOutlier Factor (LOF), K-means, and RandomForest algorithms, the new method has detected more abnormal behaviors and achieved better anomaly detection evaluation indexes. The problem of a single evaluation angle of samples in traditional anomaly detection algorithms was solved and the ability of the digital active power grid to resist unknown threats was improved.

Keywords: data security and computer security; user and entity behavior analysis; digital active power grid; fuzzy clustering; anomaly detection

随着数字化主动电网的建设与发展,电力设备信息化、数字化和智能化水平进一步提高,物理设备和信息系统耦合加剧,紧密耦合给配电设备和电力网络带来了更大的体系结构风险和更高的网络安全隐患^[1]。主动电网中的电力实体呈现出明显的离散化、开放化、差异化特征,而逐步模糊的网络边界发展趋势使得主动电网面临的网络安全威胁剧增。主动电网中分布的实体种类丰富多样,包括用户设备与电力设施设备,这些实体在提升电网运行效能的同时也带来了新的安全威胁,主动电网的安全形势变得更加复杂^[2]。对各类电力实体进行研究,分析其异常行为趋势^[3]可以发现未知网络安全风险,提高数字化主动电网的网络安全防护能力。

电力实体一般包括电网内的各类信息系统、网络安全防护装置、主机设备等,其异常检测问题属于用户与实体行为分析(user and entity behavior analytics, UEBA)范畴^[4],电力实体行为异常检测的主要目的是检测主动电网的内外部网络安全威胁,进一步提高主动电网的网络安全防护能力。在外部威胁识别方面,PAN 等^[5]设计了一个动态行为残差生成器,通过实体行为分析过滤装置对多种攻击方法进行检测,解决了静态检测模型异常检测种类少的问题。也有学者基于长短期记忆网络(long short-term memory, LSTM)及深度自编码器(deep auto encoder, DAE)建立了专门针对电网传输系统的异常检测模型^[6],该模型能够对传感器实体数据进行分析,从而监控传输保护系统和检测恶意活动。在内部威胁识别方面,JIN 等^[7]基于统计学习方法对智能电表数据进行分析,对电力高阶消耗数据在电网中潜在的电力盗窃行为进行分析,识别其中的盗窃者与受害者。李佳玮等^[8]通过对电力工控系统数据在时间维度上的周期性进行分析,建立了基于高斯混合聚类的时序异常检测模型,通过层次聚类的方式解决异常检测问题。

但上述异常检测方法都对待测样本进行“硬划分”,每个样本只归属于一个类,异常检测结果只与样本所属类簇内的样本有关,缺乏其他类簇内的样本对比,而实际情况中,电力实体行为既有“伪装性”,又有“复杂性”,不应只在一个类簇内进行分析^[9]。模糊聚类^[10]采用模糊数学方法对数据进行分析,最终得到样本类别隶属度矩阵,一个样本可能同时属于多个类别。基于模糊聚类算法处理电力实体行为数据,能够很好地保留样本与各个类簇间的关联信息,通过多角度评估判断电力实体的威胁程度,得到更准确的异常检测结果。ANGELOS 等^[11]使用模糊 C 均值聚类对用户用电数据进行分析,随后,基于样本对应每个类别的隶属度和类别中心的欧式距离设计了一种评价指标,用于区分不同的用电模式,从而得到其中的欺诈者。文献[12]利用模糊聚类对用电数据进行分析。首先,将具有相同用电习惯的电力用户进行聚类;然后,使用孤立森林算法对其中的用电异常情况进行检测,得到了不错的检测效果,但在使用模糊聚类结果时,只选择了其中的一个类别作为样本标签,没能很好地体现出算法的“模糊性”。

针对评价角度过少以及样本类别信息使用不完全的问题,本文提出了一种基于模糊聚类的多类簇归属异常行为检测(abnormal behavior detection based on fuzzy clustering, ABDFC)算法,包括模糊聚类以及多类簇异常检测 2 个过程。首先,基于实体行为频次-逆向实体频次(behavior frequency-inverse entities frequency, BF-IEF)技术建立了实体行为模糊 C 均值聚类算法,得到样本与类别的隶属度矩阵;其次,针对模糊聚类的结果,利用样本多类簇归属的性质设计了异常检测算法。

1 基于模糊聚类的多类簇归属异常检测算法

本文所提出的电力实体异常检测算法包含 2 个主要部分:1)实体行为模糊 C 均值聚类过程;2)多类簇归属异常检测过程。首先,从主动电网中进行电力实体行为的数据采集,并对其进行向量化和标准化操作;

其次,通过模糊聚类过程得到模糊聚类结果,再对模糊聚类结果进行多类簇归属异常检测。整体处理框架如图 1 所示。

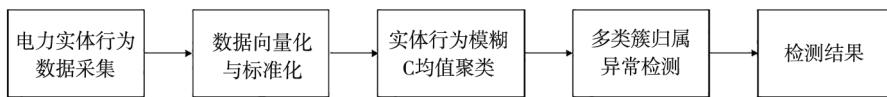


图 1 总体框架

Fig.1 Overall structure

1.1 实体行为模糊 C 均值聚类算法

模糊聚类算法采用模糊界限对待测样本进行软划分,通过隶属度对样本与类别之间的关系进行表述,使得一个样本可以有多个类别标签^[13]。模糊 C 均值(fuzzy c-means, FCM)聚类算法^[14]是模糊聚类算法中最具代表性的算法之一,其通过多次迭代的方式计算样本关于类别的归属度与聚类簇中心点,以达到最大类内相似度与最小类间相似度。模糊 C 均值算法是对普通 K 均值(K-means)算法的一种改进,为了便于与 K 均值算法进行区分,所以改用字母 C 代表聚类簇的个数。传统 FCM 算法主要使用欧式距离对样本与类别中心的偏离程度进行度量,在电力实体行为模糊聚类分析的过程中,发现直接计算行为向量间的距离对具体实体行为的区分度不理想,所以基于 TF-IDF 思想优化了实体行为处理方法,设计了实体行为模糊 C 均值聚类(fuzzy c-means for entities behavior, FCEB)算法。

TF-IDF^[15]是文本分析领域的一种语料库词语加权技术,用来评估一个单词对整篇文档的重要程度。其主要思想是:一个单词在整篇文档中的重要程度与其出现的频率成正比,与其在其他文档中出现的频率成反比。在电力实体行为分析中,参考 TF-IDF 对电力实体的具体行为进行差异化分析,计算实体行为频次-逆向实体频次 BF-IEF。

若有某一行为 B,则 B 的实体行为频次 BF 计算方法为

$$BF = \frac{\text{某实体所有行为中,B 出现的频次}}{\text{该实体所有行为的数量}}。 \quad (1)$$

实体行为频次 BF 为某个行为出现次数与行为数量的比值,一个行为在某个实体的所有动作中出现的次数越多,越能代表该实体,而且越能够区分该实体与其他实体的差别。

行为 B 的逆向实体频次 IEF 计算方法为

$$IEF = \log\left(\frac{\text{数据集中实体总数量}}{\text{行为中包括 B 的实体数量} + 1}\right)。 \quad (2)$$

式中:逆向实体频次 IEF 为电力实体总数量与包含某个行为实体数量比值的对数,对数内的分母进行加 1,是为了防止出现分母为 0 的情况。如果一个行为只在少数实体中出现,则能够更好地区分这些实体与其他实体的不同。出现的范围越广,那么这个行为的区分性就越差。

实体的频次与逆向实体频次的乘积为

$$BF-IEF = BF \times IEF。 \quad (3)$$

将 BF 与 IEF 相乘可以得到实体行为频次-逆向实体频次,该表述方法能够很好地对实体间的行为差异程度进行区别,实体某个行为的表述能力随着它在该实体行为中出现次数的增加而增加,随着它在其他实体内出现次数的增加而减少。使用 BF-IEF 技术对行为进行处理,可以更精确地表述实体行为向量,增强行为表示的准确性。

基于实体行为向量表述方法,再给定第 i 个数据样本 x_i 与第 j 个聚类类别中心 v_j ,即可计算数据样本与聚类类别中心的距离。二者之间的行为向量相似性度量值(behavioral measure, BM)的计算方法如式(4)所示:

$$BM(x_i, v_j) = x_i - v_j = \sqrt{\sum_{i=1}^N \sum_{j=1}^C (x_i - v_j)^2}。 \quad (4)$$

基于实体行为向量表述方法与实体行为向量相似性度量方法,可以构造实体行为模糊 C 均值聚类算法。在模糊聚类中,使用隶属度表示一个样本 x 对于某个类别 c 的隶属程度,一般记为 u_{ij} 。 u_{ij} 表示第 i 个样本对于第 j 个类别的隶属度,其取值范围是 $[0, 1]$ 。当 $u_{ij} = 0$ 时,表示第 i 个样本一定不属于第 j 个类别;当 $u_{ij} = 1$ 时,表示第 i 个样本一定属于第 j 个类别。

基于隶属度可得到模糊聚类的目标函数,假设数据集 $X = \{x_1, x_2, \dots, x_N\}$,聚类类别 $C = \{C_1, C_2, \dots, C_C\}$,第 j 个聚类类别中心 v_j ,则实体行为模糊 C 均值聚类 FCEB 算法的目标函数如式(5)所示:

$$\min_{u_{ij}, v_j} Jm(u_{ij}, v_j) = \sum_{i=1}^N \sum_{j=1}^C u_{ij} m \text{BM}(x_i, v_j)^2. \quad (5)$$

满足

$$\sum_{j=1}^C u_{ij} = 1. \quad (6)$$

式(5)中: m 是模糊系数,取值范围 $[1, +\infty)$,用来调节聚类模糊程度的参数, m 值越大,聚类结果越模糊,一般取 $m=2$ 。式(6)表示某个样本到所有类簇中心的隶属度之和为 1。

利用拉格朗日乘数法,引入 N 个拉格朗日因子将式(5)与式(6)转化为无条件极值问题:

$$\min \mathcal{L}(u_{ij}, v_j) = \sum_{i=1}^N \sum_{j=1}^C u_{ij} m \text{BM}(x_i, v_j)^2 + \sum_{i=1}^N \lambda_i \left(\sum_{j=1}^C u_{ij} - 1 \right). \quad (7)$$

对式(7)中的 u_{ij} 与 v_j 进行求导,可以得到各变量的极值点。

对隶属度参数 u_{ij} 求偏导得:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\text{BM}(x_i, v_j)^2}{\text{BM}(x_k, v_j)^2} \right)^{\frac{1}{m-1}}}.$$
(8)

对聚类中心 v_j 求偏导得:

$$v_j = \frac{\sum_{j=1}^N u_{ij} m x_j}{\sum_{j=1}^N u_{ij} m}.$$
(9)

式(8)是隶属度迭代公式,式(9)是聚类簇中心迭代公式。

所以实体行为模糊 C 均值聚类 FCEB 算法的步骤如下。

1) 初始化 C 个样本作为初始聚类中心 $C = \{C_1, C_2, \dots, C_C\}$,第 j 个聚类类别中心 v_j 。

2) 针对每个样本 x_i ,计算它到第 j 个聚类中心的距离,并更新隶属度矩阵:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left(\frac{\text{BM}(x_i, v_j)^2}{\text{BM}(x_k, v_j)^2} \right)^{\frac{1}{m-1}}}.$$
(10)

3) 针对每个类别,重新计算它的聚类中心

$$v_j = \frac{\sum_{j=1}^N u_{ij} m x_j}{\sum_{j=1}^N u_{ij} m}.$$
(11)

4) 重复第 2、第 3 步,直至达到最大迭代次数或者隶属度矩阵变化小于阈值。

1.2 多类簇归属异常检测

随着网络技术的不断发展,电力实体行为中的异常情况呈现出明显的复杂化、隐蔽化特征^[16]。现有异常检测算法多基于单一维度对行为内容进行分析,采用传统方法对模糊聚类结果进行异常检测,会丧失模糊聚类结果的“模糊”特性。局部异常因子(local outlier factor, LOF)算法^[17]是一种常见的基于密度的单维度异常检测算法,通过比较待分析节点与邻居节点之间局部离群值的大小,从而判断待分析节点是否异常。本文基于 LOF 算法,设计了一种多类簇归属的异常检测(anomaly detection based on multi-categories affiliation, ADMA)算法专门用于解决模糊聚类结果的异常检测问题。

在同一类簇内所有的数据点中,距离待分析点 O 最近的第 k 个点与点 O 之间的距离被称为 K 近邻距离^[18],用 K -nearest neighbor distances 表示,如图 2 所示。点 O 的 K 近邻距离越大,则其周围的点越稀疏,越远离主流数据分布。

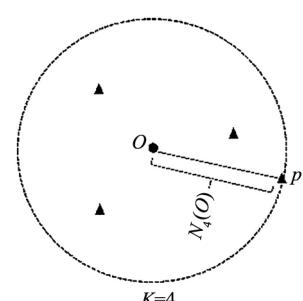


图 2 点 O 的 K 近邻距离

Fig.2 K -nearest neighbor distances of O

值得注意的是如图3所示如果点O在模糊聚类的结果中属于M个类簇,点O的K近邻距离会有M个。从第k点到点O之间的所有点均为点O的K邻域内的点,可以用 $N_K^{(m)}(O)$ 表示。下标K代表近邻个数,由用户输入;上标m代表O在第m簇类内的近邻,m的取值范围是[1,M]。

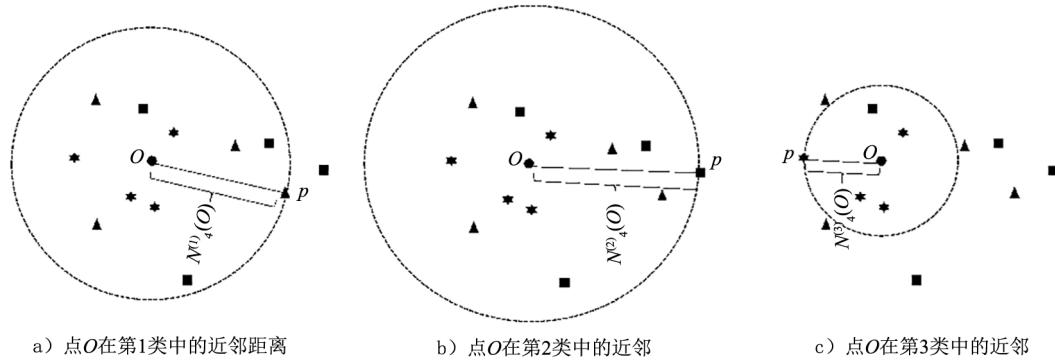


图3 点O在不同类簇内的K近邻距离

Fig.3 K-nearest neighbor distances of O in different categories

点\$p\$到点\$O\$间的可达距离(reach_dist)定义为“点\$O\$的\$K\$近邻距离”和“数据点\$p\$与点\$O\$直接距离”的最大值,即:

$$\text{reach_dist}(p, O) = \max\{\text{KNND}(O), \text{dist}(p, O)\} \quad (12)$$

值得注意的是,式(12)的定义是有方向的,点\$p\$到点\$O\$的可达距离可能不等于点\$O\$到点\$p\$的可达距离。

点\$O\$的近邻密度(nearest neighbor density, NND)用来衡量点\$O\$在所属的聚类簇内与周围其他点相比的疏密程度,定义为每个类簇内\$K\$邻域内的点与点\$O\$平均可达距离的倒数,即:

$$\text{NND}(i)(O) = \frac{1}{\sum_{p \in N_K^{(m)}(O)} \text{reach_dist}(p, O)}, \quad (13)$$

式中:\$m\$代表第\$m\$个簇,对每个点\$O\$的归属聚类簇进行计算,可以得到\$M\$个点\$O\$的近邻密度。平均距离越低,近邻密度越高,近邻密度高意味着该点附近比较稠密。另外,在点\$O\$的邻域内,可能不止包含\$K\$个点,所以要根据实际情况归一化可达距离之和,图4是当\$K=4\$时,点\$O\$的4邻域内包含6个点的情况。

在ADMA算法中,使用点\$O\$近邻相对异常因子(nearest neighbor relative anomaly factor, NNRAF)对点\$O\$的异常程度进行衡量,其异常因子得分为点\$O\$在\$M\$个类簇邻域内样本点的平均近邻密度与点\$O\$近邻密度之比的平均值,即:

$$\text{NNRAF}(O) = \sum_{m=1}^M \frac{u_m \sum_{p \in N_K^{(m)}(O)} \frac{\text{NND}(m)(p)}{\text{NND}(m)(O)}}{|N_K^{(m)}(O)|}, \quad (14)$$

式中:\$m\$代表第\$m\$个簇;\$N_K^{(m)}(O)\$是点\$O\$在第\$m\$个簇的\$K\$近邻内的点;\$\text{NND}(m)(O)\$是点\$O\$在第\$m\$个簇的近邻密度;\$u_m\$是点\$O\$相对于第\$m\$个簇的隶属度。根据NNRAF的定义,如果一个点的近邻相对异常因子小于1,则证明该点所处的位置比较稠密,附近的样本点数量较多,该点属于正常点。如果该点的NNRAF值大于1,则与其在\$M\$簇邻域内的点相比较,该点的局部密度较低,相对近邻节点结构异常,属于异常点。

1.3 ABDFC 异常检测算法过程

基于实体行为模糊C均值聚类(FCEB)与多类簇归属异常检测(ADMA)算法,设计了一种基于模糊聚类的多类簇归属异常检测ABDFC算法。首先,采用BF-IEF技术对电力实体行为进行处理;其次,采用多次迭代的方式得到各实体行为的模糊聚类结果;最后,根据类簇隶属度矩阵对异常行为进行识别。算法的整体过程主要包括模糊聚类和多类簇归属异常检测2个阶段,算法流程如图5所示。

具体步骤如下:

- 1) 标准化现有实体行为信息数据集 \$X = \{x_1, x_2, \dots, x_N\}\$;

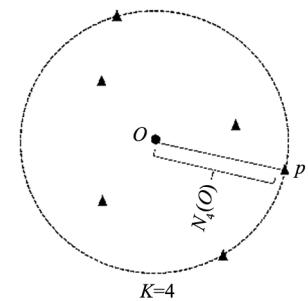


图4 点O的\$K\$近邻内包含多个点的情况

Fig.4 K-nearest neighbor of the point O containing more than one point

- 2) 设定模糊聚类超参数(类别个数 C 与模糊系数 m)；
- 3) 初始化聚类中心, 设第 j 个聚类类别中心 v_j ；
- 4) 针对每个样本 x_i , 利用它到第 j 个聚类中心 v_j 的距离计算 u_{ij} , 并更新隶属度矩阵:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\text{BM}(x_i, v_j)^2}{\text{BM}(x_k, v_j)^2} \right)^{\frac{1}{m-1}}} ; \quad (15)$$

5) 针对每个类别,重新计算它的聚类中心 v_j :

$$v_j = \frac{\sum_{i=1}^N u_{ij} m}{\sum_{i=1}^N u_{ij}^m}; \quad (16)$$

6)重复第4、第5步,直至达到最大迭代次数或者隶属度矩阵变化小于阈值,最终得到隶属度矩阵;

7) 遍历数据集 X 内的样本, 根据每个类别的隶属度计算其 NNRAF 值, 并判断是否为异常点。

$$\text{NNRAF}(O) = \sum_{m=1}^M \frac{um \sum_{p \in N_K^{(m)}(O)} \frac{\text{NND}^{(m)}(p)}{\text{NND}^{(m)}(O)}}{|N_k^{(m)}(O)|}. \quad (17)$$

2 实验仿真

2.1 数据来源

为了验证异常检测算法的有效性,本文通过散布在主动电网重要节点中的网络探针得到了变压器、传感器、智能电表等各类物联网电力实体间的通信数据,部分原始数据样例如下所示:

基于参考文献[4]中的特征工程方法和 NSL-KDD^[19]的数据格式对电力实体间的通信数据进行了标准化预处理,经过处理后的每条连接共有 43 个特征,部分特征如表 1 所示。

表 1 电力实体行为数据特征样例

Tab.1 Example of power entity behavior data characteristics

特征名	数据类型	维度	内容描述
源设备地址	离散型	1	记录连接请求发出者的地址
目的设备地址	离散型	1	记录连接请求接收者的地址
请求时间	离散型	5	请求时间的年、月、日、时、分
连接持续时间	连续型	1	本次连接的持续时间,单位秒
协议类型	离散型	1	协议类型共有 3 种:TCP, UDP, ICMP
返回状态	离散型	1	连接正常或错误的状态
:	:	:	:

除了 1.3 节中所表述的 BF-IEF 加权技术,还使用归一化过程对数据进行特征缩放,经过处理后的数据压缩到 0 到 1 之间,其归一化方法为

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (18)$$

另外根据 NSL-KDD 的攻击判断逻辑对样本数据进行了类别标注，并补充了部分攻击数据，最终得到 5 大类、39 小类共 1 000 条电力实体通信行为数据，数据分布情况如表 2 所示。

2.2 评价指标

在对算法评价的过程中,基于混淆矩阵^[20]对各类指标进行计算,混淆矩阵如表 3 所示。

其中 TP(true positive)代表真正例,是正样本被预测为正样本;FP(false positive)代表假正例,是负样本被预测为正样本;FN(false negative)代表假负例,是正样本被预测为负样本;TN(true negative)代表真负例,是负样本被预测为负样本。基于混淆矩阵可以得到若干评估指标,常用的指标有准确率(accuracy, Acc)、精确率(precision, P)、召回率(recall, R)和 F1 值,具体的计算方法如下所示。

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}, \quad (19)$$

$$P = \frac{TP}{TP + FP}. \quad (20)$$

$$R = \frac{TP}{TP + FN}, \quad (21)$$

$$F1 = \frac{2 \times P \times R}{P + R}. \quad (22)$$

此外还使用了受试者工作特征曲线(receiver operating characteristic curve, ROC)和 ROC 曲线下的面积(area under curve, AUC)对算法进行评价。

3 结果与分析

本次实验在 Intel core i7-9750H@2.6 GHz 处理器,16G 内存,Python 3.7.2 环境下运行。多分类实验的识别结果按照攻击大类进行统计,分别是 Normal,DOS,Probe,U2R,R2L。

首先,进行模糊聚类分析。在参数选择方面,模糊聚类超参数 m 设置为 2,C 值设置为 5。图 6 中,随机选取了 3 个样本维度进行展示,将其隶属度最大的一类作为最终类别。由图 6 可以看出,整体数据的模糊聚类情况有一定的分布规律,Normal 类分布在 xy 轴平面,主要集中在 x 轴、 y 轴及对角线上;Probe 类多分布在 y 轴;其余 3 类多集中在原点附近。

选取了 LOF,K-Means,Random Forest 与本文所提出的 ABDFC 算法进行异常检测效果对比,结果如图 7 所示。在异常行为检出量方面,本文提出算法所检出的异常行为最多,在 1 000 个样本中,共识别得到了非 Normal 类数据 164 条。

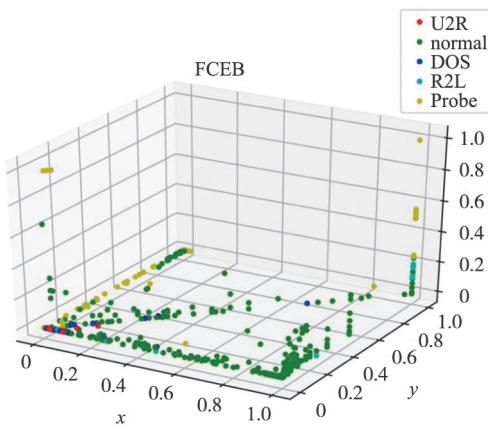


图 6 实体行为模糊 C 均值聚类结果

Fig.6 Fuzzy C-means clustering results of entity behavior

表 2 实验数据的类别分布情况

Tab.2 Category distribution of experimental data

攻击大类	攻击描述	攻击小类	数量
Normal	正常样本	Normal	816
DOS	拒绝服务类攻击	back	6
Probe	端口监视或扫描类攻击	neptune	45
		pod	1
		teardrop	11
		portsweep	36
		satan	17
		ipsweep	31
		nmap	7
U2R	非法的本地超级用户特权访问	perl	17
R2L	来自远程设备的未授权访问	warezclient	12
		guess_passwd	1

表 3 混淆矩阵

Tab.3 Confusion matrix

真实情况	预测结果	
	正样本	负样本
正样本	TP	FN
负样本	FP	TN

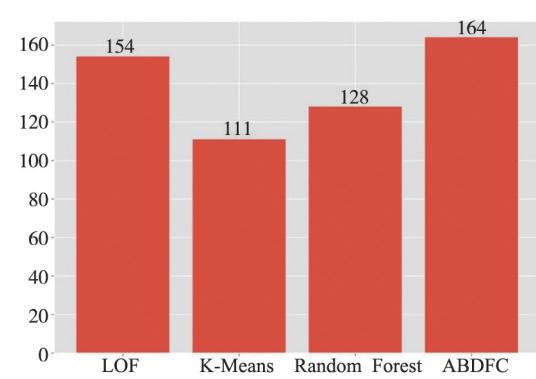


图 7 异常检出量对比

Fig.7 Comparison of abnormal detections

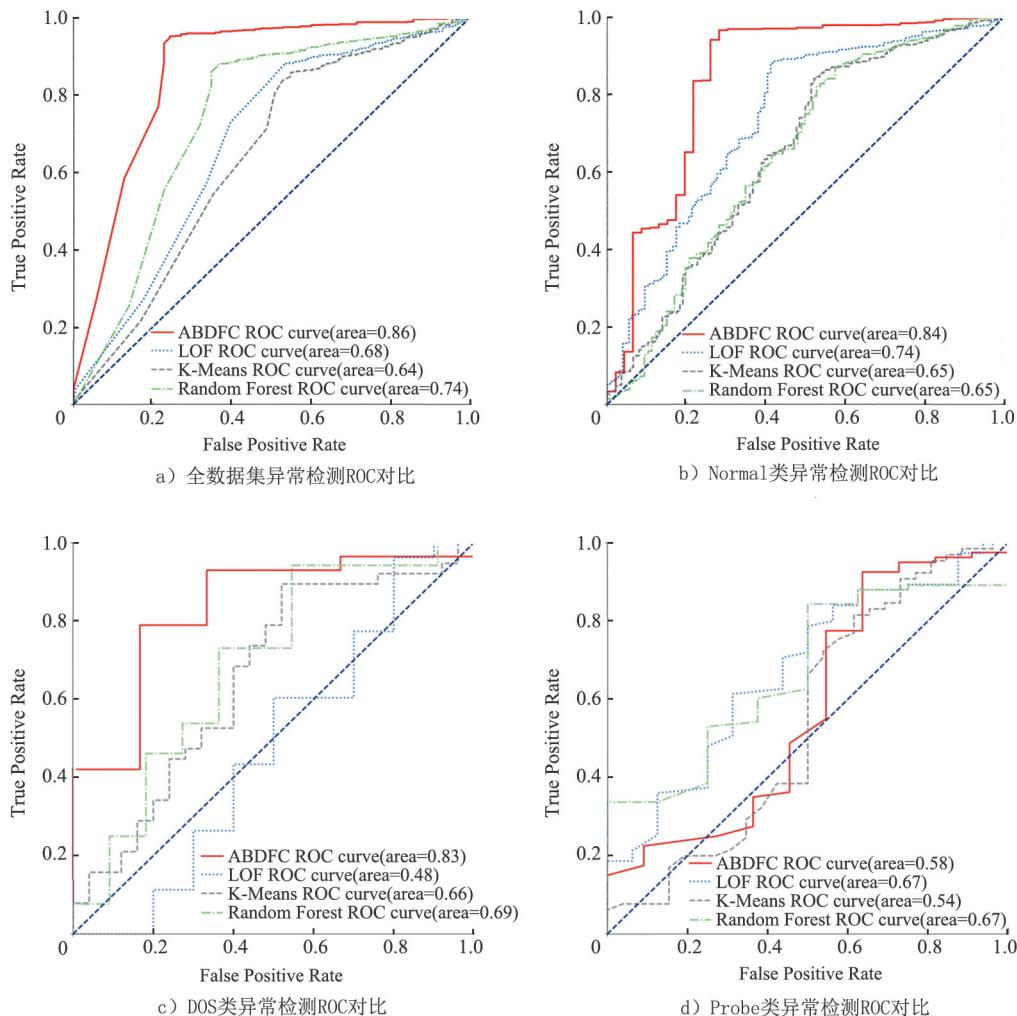
算法的多类别检测将数据集按照类别进行拆分后再进行实验。表 4 是 4 种算法在多类别数据上的评价指标,由表 4 可以看出 ABDFC 算法整体表现最优,随机森林略弱。ABDFC 算法在 R2L 类别与 Probe 类别的检测方面稍有不足,Random Forest 在这 2 个类别检测的各个指标上均表现最好。虽然 ABDFC 算法在多个类别的检测上均取得了不错的效果,但由于 R2L 类别样本数较少,所以模型未能很好地从数据中学习到相应的规律,影响了 ABDFC 算法评价指标的平均值。

表 4 评价指标对比

Tab.4 Comparison of evaluation indexes

类别	算法结果															
	LOF				K-Means				Random Forest				ABDFC			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Normal	0.85	0.89	0.89	0.89	0.83	0.89	0.86	0.87	0.87	0.88	0.88	0.88	0.94	0.97	0.96	0.97
DOS	0.87	0.96	0.89	0.92	0.73	0.72	0.89	0.80	0.84	0.88	0.94	0.91	0.89	0.96	0.91	0.94
Probe	0.79	0.90	0.88	0.89	0.68	0.76	0.85	0.80	0.84	0.94	0.96	0.95	0.84	0.91	0.95	0.93
U2R	0.65	0.64	0.64	0.64	0.59	0.55	0.60	0.57	0.65	0.64	0.64	0.64	0.71	0.79	0.92	0.85
R2L	0.64	0.78	0.78	0.78	0.71	0.89	0.80	0.84	0.82	0.93	0.93	0.93	0.69	0.80	0.89	0.84
平均值	0.76	0.83	0.81	0.82	0.71	0.76	0.80	0.78	0.80	0.85	0.87	0.86	0.81	0.89	0.93	0.90

图 8 分别给出了 4 种算法在进行多类别检测任务时的 ROC。由图 8 可以看出 ABDFC 算法的 ROC 曲线下面积在全量异常识别、Normal 类识别、DOS 类识别实验中性能优势较为明显,ROC 分别达到了 0.86,0.84 和 0.83。由于其他 3 个类别的样本数量较少,所以检测效果略有波动。



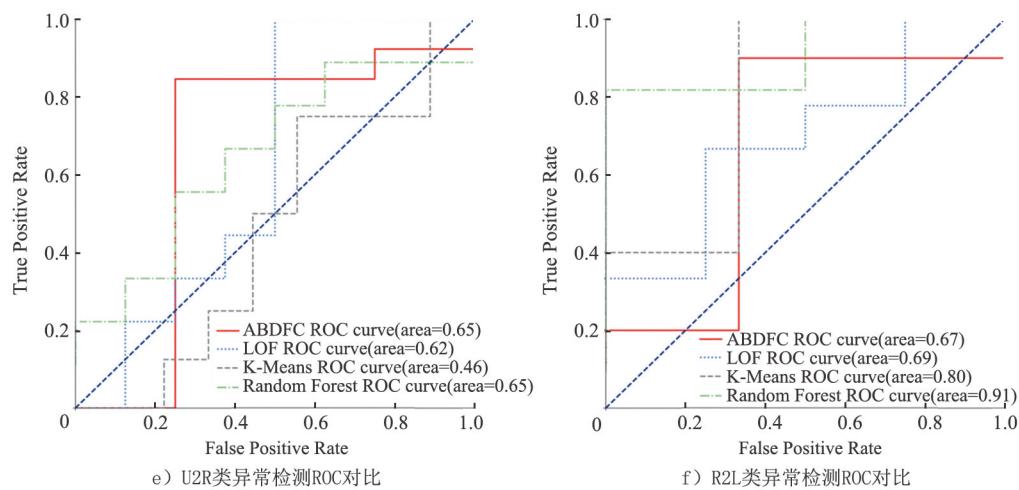


图8 不同算法 ROC 曲线对比

Fig.8 Comparison of ROC curves of different algorithms

从图8各子图的ROC曲线来看,ABDFC算法在全量数据集的异常检测实验中表现最好,其ROC远大于其他算法,这是由于在全量数据集中实体行为种类更多,基于BF-IEF技术所改进的模糊聚类能够对各类电力实体的行为进行更好的区分,从而得到更优的模糊聚类效果。而且在分类别异常检测试验中,实体类别的多样性有所降低,ABDFC算法在多类簇归属信息使用方面的优势有所下降,所以呈现出的检测优势不如全量数据集异常检测试验中大。与传统LOF算法相比,经过电力实体样本多类簇归属性质改进的ABDFC算法进行了更多维度的异常因子计算,可以检测出更多的异常点,对比LOF算法有着更优的表现。Random Forest算法在各个异常检测的实验中对数据的大小变化不敏感,表现比较稳定,但从ROC来看稍逊于ABDFC算法。

此外,还对ABDFC算法的实时识别能力进行了测试,如表5所示。由表5可知,ABDFC算法在进行小批量在线检测方面也优于其他算法,适合作为实体行为增量检测模型进行全天候部署检测。

从实验结果上看,ABDFC算法对于电力实体行为异常检测的整体表现优于LOF,K-Means及Random Forest算法,尤其在Normal样本检测方面的各类指标都比较突出。在实际应用中,电力实体行为分析的主要作用在于对数据进行“正常”与“异常”的二分类验证,所以ABDFC算法比其他3种算法更适用于当前电力实体行为的异常检测。

4 结语

本文提出了一种基于模糊聚类和多类簇归属的异常检测算法,主要应用于电力实体行为分析。首先,基于行为度量方法设计了实体行为模糊聚类算法,并得到电力实体行为与类别间的隶属度矩阵,解决了确定性聚类算法分析角度单一的问题。其次,基于隶属度矩阵设计了多类簇归属异常检测算法,对电力实体行为在各个所属类别内的相对结构进行分析,判断该点的异常情况。结果表明,ABDFC算法解决了传统异常检测算法实体行为硬划分的问题,通过在不同类内分别计算异常程度的方式得到更为准确的检测结果。

本文方法虽然能够利用模糊聚类以及多类簇归属异常检测算法对电力实体行为的异常情况进行分析,但在实验过程中所涉及的超参数较多,因此,对超参数的设置过程和具体影响进行研究,设计自动化模糊聚类异常检测算法是下一阶段的研究方向。

参考文献/References:

- [1] 苏盛,汪干,刘亮,等.电力物联网终端安全防护研究综述[J].高电压技术,2022,48(2):513-525.
SU Sheng,WANG Gan,LIU Liang,et al.Review on security of power internet of things terminals[J].High Voltage Engineering,2022,48(2):513-525.

表5 小批量检测准确率对比

Tab.5 Comparison of evaluation indexes for small amount

参数	算 法			
	LOF	K-Means	Random Forest	ABDFC
准确率/%	90.34	89.50	91.88	93.52

- [2] 吴姗姗,宁昕,郭屾,等.配电物联网在新产业形态中的应用探讨[J].高电压技术,2019,45(6):1723-1728.
WU Shanshan, NING Xin, GUO Shen, et al. Discussion on application of distribution internet of things in new industry form [J]. High Voltage Engineering, 2019, 45(6): 1723-1728.
- [3] DEHGHANI N L, ZAMANIAN S, SHAFIEEZADEH A. Adaptive network reliability analysis: Methodology and applications to power grid[J]. Reliability Engineering & System Safety, 2021, 216: 20-35.
- [4] 崔景洋,陈振国,田立勤,等.基于机器学习的用户与实体行为分析技术综述[J].计算机工程,2022,48(2):10-24.
CUI Jingyang, CHEN Zhenguo, TIAN Liqin, et al. Overview of user and entity behavior analytics technology based on machine learning [J]. Computer Engineering, 2022, 48(2): 10-24.
- [5] PAN Kaikai, PALENSKY P, ESFAHANI P M. From static to dynamic anomaly detection with application to power system cyber security [J]. IEEE Transactions on Power Systems, 2020, 35(2): 1584-1596.
- [6] AHMED A, KRISHNAN V V G, FOROUTAN S A, et al. Cyber physical security analytics for anomalies in transmission protection systems[J]. IEEE Transactions on Industry Applications, 2019, 55(6): 6313-6323.
- [7] TAO Jin, MICHAILIDIS G. A statistical framework for detecting electricity theft activities in smart grid distribution networks[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(1): 205-216.
- [8] 李佳伟,吴克河,张波.基于高斯混合聚类的电力工控系统异常检测研究[J].信息网络安全,2021,21(3):53-63.
LI Jiawei, WU Kehe, ZHANG Bo. Research on anomaly detection of power industrial control system based on gaussian mixture clustering [J]. Netinfo Security, 2021, 21(3): 53-63.
- [9] MCLOUGHLIN F, DUFFY A, CONLON M. A clustering approach to domestic electricity load profile characterisation using smart metering data[J]. Applied Energy, 2015, 141: 190-199.
- [10] MOJARAD M, NEJATIAN S, PARVIN H, et al. A fuzzy clustering ensemble based on clusterclustering and iterative Fusion of base clusters[J]. Applied Intelligence, 2019, 49(7): 2567-2581.
- [11] ANGELOS E W S, SAAVEDRA O R, CORTÉS O A C, et al. Detection and identification of abnormalities in customer consumptions in power distribution systems[J]. IEEE Transactions on Power Delivery, 2011, 26(4): 2436-2442.
- [12] 赵曼,李英娜,李川,等.基于模糊聚类和孤立森林的用电数据异常检测[J].陕西理工大学学报(自然科学版),2020,36(4):38-43.
ZHAO Man, LI Yingna, LI Chuan, et al. Anomaly detection of power consumption data based on fuzzy clustering and isolated forest [J]. Journal of Shaanxi University of Technology (Natural Science Edition), 2020, 36(4): 38-43.
- [13] BIAN Zekang, CHUNG F L, WANG Shitong. Fuzzy density peaks clustering[J]. IEEE Transactions on Fuzzy Systems, 2021, 29(7): 1725-1738.
- [14] 张宁,马盈仓,朱恒东.基于拉普拉斯约束的半监督模糊C均值算法[J].应用数学进展,2021,10(2):433-443.
ZHANG Ning, MA Yingcang, ZHU Hengdong. Semi-supervised fuzzy C-means algorithm based on laplace constraint [J]. Advances in Applied Mathematics, 2021, 10(2): 433-443.
- [15] TANG Zhong, LI Wenqiang, LI Yan, et al. Several alternative term weighting methods for text representation and classification[J]. Knowledge-Based Systems, 2020, 207. DOI: 10.1016/j.knosys.2020.106399.
- [16] SU Tao, SHI Ying, YU Jicheng, et al. Nonlinear compensation algorithm for multidimensional temporal data: A missing value imputation for the power grid applications[J]. Knowledge-Based Systems, 2021, 215. DOI: 10.1016/j.knosys.2021.106743.
- [17] GAO Jianhua, JI Weixing, ZHANG Lulu, et al. Cube-based incremental outlier detection for streaming computing[J]. Information Sciences, 2020, 517: 361-376.
- [18] KONG Xiangzeng, BI Yaxin, GLASS D H. Detecting anomalies in sequential data augmented with new features[J]. Artificial Intelligence Review, 2020, 53(1): 625-652.
- [19] TAVALLAEE M, BAGHERI E, LU W, et al. NSL-KDD Datasets[EB/OL]. <https://www.unb.ca/cic/datasets/nsl.html>, 2022-03-31.
- [20] TÂUTAN A M, IONESCU B, SANTARNECCHI E. Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques[J]. Artificial Intelligence in Medicine, 2021, 117. DOI: 10.1016/j.artmed.2021.102081.

主 编：朱立光
副 主 编：蒋立杰（常务）
张士莹
本期英文编辑：李 辉 冯 民
本期责任编辑：王淑霞
封面设计：王欣欣 李天滢

收载本刊的国内外主要检索系统

- ◆ 中国学术期刊网络出版总库（中国知网）
- ◆ 《中国学术期刊文摘》（CSAC）
- ◆ “中国科技论文统计源期刊”（中国科技核心期刊）
- ◆ 中文科技期刊数据库（维普）
- ◆ 万方数据库
- ◆ 台湾中文电子期刊服务——思博网（CEPS）
- ◆ 美国《化学文摘（网络版）》（CA）
- ◆ 美国《乌利希期刊指南（网络版）》（Ulrichsweb）
- ◆ 美国《艾博思科数据库》（EBSCOhost）
- ◆ 英国《科学文摘》（INSPEC）
- ◆ 英国《高分子图书馆》（PL）
- ◆ 瑞典《期刊公开获取指南》（DOAJ）
- ◆ 瑞士《卫生领域研究网计划》（HINARI）
- ◆ 俄罗斯《文摘杂志》（AJ）
- ◆ 《日本科学技术振兴机构（中国数据库）》（JSTChina）

河北科技大学学报
HEBEI KEJI DAXUE XUEBAO
双月刊，1980年创刊
第43卷 第5期（总第168期） 2022年10月

Journal of Hebei University of
Science and Technology
Bimonthly, Launched in 1980
Vol.43 No.5 (Sum 168) Oct. 2022

主管单位：河北省教育厅
主办单位：河北科技大学
编辑出版：《河北科技大学学报》编辑部
发 行：《河北科技大学学报》编辑部
地 址：河北省石家庄市裕翔街26号
邮 编：050018
电 话：0311-81668291 81668292
网 址：<https://xuebao.hebust.edu.cn>
<http://hbqj.cbpt.cnki.net>
电子信箱：xuebao@hebust.edu.cn
印 刷：石家庄众旺彩印有限公司

Administrated by Hebei Education Department
Sponsored by Hebei University of Science and Technology
Edited and Published by Editorial Department of Journal of
Hebei University of Science and Technology
Distributed by Editorial Department of Journal of
Hebei University of Science and Technology
Address: 26 Yuxiang Street, Shijiazhuang, Hebei
Post Code: 050018
Tel: 86-311-81668291 81668292
URL: <https://xuebao.hebust.edu.cn>
<http://hbqj.cbpt.cnki.net>
E-mail: xuebao@hebust.edu.cn
Printing: Shijiazhuang Zhongwang Color Printing Co., Ltd.

国际标准连续出版物号：ISSN 1008-1542
国内统一连续出版物号：CN 13-1225/TS

发 行 范 围：国内外公开发行
国外发行代号：DK13002
定 价：20.00元



10>