

基于弱无回溯模式匹配算法的 敏感大数据安全分级检测方法

张海明[†], 王艺霏, 那琼澜, 王森, 温馨, 姚艳丽

(国网冀北电力有限公司信息通信分公司, 北京 100053)

摘 要:随着信息技术的广泛应用,敏感大数据在传输和存储过程中存在隐私泄露、数据滥用、安全风险等问题。为了提高敏感大数据的安全性,提出了一种基于弱无回溯模式匹配算法的敏感大数据安全分级检测方法。建立协方差矩阵构建敏感大数据样本库,设定约束条件同构嵌入敏感大数据子图,采用弱无回溯模式匹配算法回溯匹配敏感大数据查询图索引,降低计算复杂度。利用核函数映射查询图特征向量,提高敏感大数据识别准确度,以剪枝的方式完成敏感大数据特征相似性匹配。通过相似度指数函数计算敏感因子,划分敏感大数据安全等级,完成自动化分级检测。实验结果表明:所提方法的敏感大数据安全分级准确率为100%、特征匹配耗时为2.7 s、收敛速度为0.44。由此证明,所提方法能够有效保护敏感大数据安全,在数据传输和存储中具有较高的实际应用价值。

关键词:弱无回溯模式匹配算法;协方差矩阵;相似度指数函数;敏感大数据

中图分类号:TP399

文献标识码:A

Sensitive Big Data Security Hierarchical Detection Method Based on Weak Backtracking Pattern Matching Algorithm

ZHANG Haiming[†], WANG Yifei, NA Qionglan, WANG Sen, WEN Xin, YAO Yanli

(State Grid Jibei Information and Telecommunication Company, Beijing 100053)

Abstract: With the wide application of information technology, there are some problems in the transmission and storage of sensitive big data, such as privacy leakage, data abuse and security risks. In order to improve the security of sensitive big data, this paper proposes a hierarchical detection method of sensitive big data security based on weak backtracking pattern matching algorithm. Establishing covariance matrix and constructing sensitive big data sample database. Set constraints to isomorphically embed sensitive big data subgraphs, and use weak backtracking pattern matching algorithm to backtrack the index of sensitive big data query graph, which reduces the computational complexity. Kernel function is used to map the feature vector of query graph to improve the recognition accuracy of sensitive big data, and the similarity matching of sensitive big data features is completed by pruning. The sensitivity factor is calculated by similarity index function, and the security level of sensitive big data is divided, thus completing automatic hierarchical detection. The experimental results show that the accuracy of sensitive big data security classification is 100%, the time for feature matching is 2.7 s, and the convergence speed is 0.44. This proves that the proposed method can effectively protect the security of sensitive big data and has high practical application value in data transmission and storage.

Key words: weak backtracking pattern matching algorithm; covariance matrix; similarity index function; sensitive big data

随着信息技术的飞速发展,大数据已成为当今社会中的一个重要组成部分。大数据的产生和存储量均呈指数级增长,其中包含了大量敏感数据,如个人身份信息^[1]、财务数据^[2]、医疗记录^[3]等。敏感大数据的泄露和滥用可能会导致个人隐私的曝光,对个人身份、实名认证信息等构成威胁^[4]。因此,保护敏感大数据的安全至关重要。已有学者探索了多种数据类型的安全检测方法。王伟等^[5]提出了一种基于数据挖掘的信息安全检测方法,结合轨道交通信息系统中异常数据占比较少特点,构建单分类支持向量机模型。采用超平面法将正常数据和入侵数据进行分类,实现网络入侵行为的有效检测。但是,该方法计算算力有待进一步验证。高丽杰等^[6]设计了一种 RSA 公钥加密机制,用于大数据安全性双向检测。根据用户的安全需求度选择不同的数据安全性,引入分类算法聚类评价指标,采用距离函数组合不同权值,获取大数据安全性双向检测结果。但是,该方法在大规模数据集处理中的检测准确率有待进一步提高。

弱无回溯模式匹配算法是一种用在大规模数据集中处理和分析匹配任务的算法,具有计算算力较高的特点,能够在海量数据中快速完成模式匹配。弱无回溯模式匹配算法通过减少或避免回溯操作,提高大规模数据匹配和搜索效率。在敏感大数据的安全分级检测方面,弱无回溯模式匹配算法的应用相对较少。基于此,本文引入弱无回溯模式匹配算法,提出一种新的敏感大数据安全分级检测方法,旨在解决当前大数据环境下的安全问题。建立协方差矩阵捕捉不同敏感大数据之间的相关性,构建敏感大数据样本库,以便更好地表示敏感大数据属性。设定约束条件,将敏感大数据子图同构嵌入到样本库,确保敏感大数据特征得到保留,有助于后续的特征相似性匹配。采用弱无回溯模式匹配算法回溯匹配敏感大数据查询图索引,识别与查询图相似的敏感大数据样本。使用核函数将查询图映射到高维空间,转换查询图特征向量,使敏感大数据更易分类。采用剪枝的方式筛选与查询图相似度高的敏感大数据样本,完成特征相似性匹配。利用相似度指数函数计算敏感因子,根据敏感因子结果大小划分敏感大数据安全等级,以自动化对敏感大数据的分级检测,为后续的安全管理和控制提供数据依据。

1 敏感大数据安全分级检测方法

1.1 构建敏感大数据样本库

样本库是敏感大数据特征匹配的基础,收集业务中产生、使用的敏感大数据分量形成样本库,能够有效提高非线性数据的处理能力,并提高敏感大数据的质量和准确性。设定 c_l 为第 l 个敏感大数据分量,其中 $l = 1, 2, \dots, n$, 建立协方差矩阵^[7], 捕捉不同敏感大数据之间的相关性:

$$V_l = \sum_{l=1}^M c_l \tau \quad (1)$$

式中, M 为敏感大数据总体数量; τ 为敏感大数据累计频率。当 V_l 趋向无穷时,表明敏感大数据分量具有一定的随机性且存在重复数据^[8], 因此需要清洗敏感大数据相关性:

$$Y = \frac{\sum_{c=1}^M z_c V_l}{b} \quad (2)$$

式中, z_c 为敏感大数据第 c 个目标源相似性概率分布的散度; b 为协同过滤参数。标注清洗后的敏感大数据为 c'_l , 构建敏感大数据样本库:

$$W = \frac{c'_l}{M} \times \frac{I + d_c Y}{\omega_c(x, y)} \quad (3)$$

式中, I 为敏感大数据信息熵; d_c 为第 c 个目标源的相似性加权系数; $\omega_c(x, y)$ 为第 c 个目标源的网络参数节点。将敏感大数据相关性划分至 W 中, 以便在后续特征相似性匹配阶段获得更好的结果, 为敏感大数据安全分级检测提供数据支持。

1.2 匹配敏感大数据特征相似性

在构建敏感大数据样本库 W 后, 根据敏感大数据子图同构性设定节点属性约束条件、标签完整约束条件、注入一致约束条件, 协同节点数量、聚类系数、同配性、介数中心度等参数, 保留敏感大数据特征, 提高敏感大数据特征相似性匹配效率。

(1) 节点属性约束

$$\forall (x_i, y_i) \in H_c, (M[x_i], M[y_i]) \in H_c \quad (4)$$

式中, (x_i, y_i) 为敏感大数据子图节点^[9]; H_c 为敏感大数据子图介数中心度。通过节点属性约束条件保持敏感大数据子图中节点属性在 W 中的嵌入过程一致, 确保关键属性信息不混淆。

(2) 标签完整约束

$$\forall (x_i, y_i) \in R_c, l(x_i, y_i) = l(M[x_i], M[y_i]) \quad (5)$$

式中, R_c 为聚类系数^[10]; $l(x_i, y_i)$ 为敏感大数据子图节点的同配性。通过标签完整约束条件确保敏感大数据子图中的所有节点都被正确地嵌入到 W 中, 且没有遗漏或错误节点。

(3) 注入一致约束

$$\forall (x_i, y_i) \in H_c, x_i \neq y_i \Rightarrow M[x_i] \neq M[y_i] \quad (6)$$

通过注入一致约束条件确保敏感大数据子图在 W 中的嵌入过程保持特征相似。

设定约束条件后, 采用弱无回溯模式匹配算法^[11]回溯匹配敏感大数据查询图索引, 使其在匹配过程中避免不必要的回溯, 提高敏感大数据特征相似性。弱无回溯模式匹配算法基于正向匹配和贪婪匹配原则^[12], 根据约束条件从查询图索引中选择最合适的匹配路径, 并在匹配过程中尽可能地避免回溯, 快速识别与查询图相似的敏感大数据样本, 减少计算消耗。使用弱无回溯模式匹配算法进行敏感大数据查询图索引的回溯匹配, 有效提高查询效率和响应速度, 能够适应不同规模和复杂度的查询需求。具体回溯匹配操作如图 1 所示。

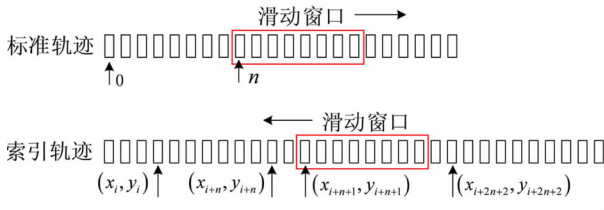


图 1 回溯匹配过程示意图

根据内积匹配准则对约束后的敏感大数据进行映射, 将非线性映射函数 φ 输入到 W 中^[13], 获取新的敏感大数据特征空间:

$$G = \varphi \sum_{i=1}^n l(x_i, y_i) \quad (7)$$

设定 $\eta_l(k_l, k_o)$ 为核函数, 使用核函数将查询图映射到高维空间:

$$\eta_l(k_l, k_o) = \exp \left(- \frac{G \|k_l - k_o\|^2}{\zeta^2} \right) \quad (8)$$

式中, k_l 为子图直径; k_o 为子图的模块距离; ζ 为敏感大数据子图中的边数量。

设定 μ_N 为查询图特征向量, 归一化该特征向量后转换查询图特征向量:

$$B_N = \mu_N \beta_o^n (\varphi(k_o)) \quad (9)$$

式中, β_o 为滞后阶数; φ 为累计和, 使查询图特征向量与查询图索引对应。根据查询图特征向量之间的距离对敏感大数据候选项进行剪枝:

$$f_p = B_N \sum_{i=1}^n (\psi h_i + k_i) \quad (10)$$

式中, ψ 为运行时间; h_i 为存储最短距离; k_i 为子图的核心节点个数。采用余弦相似度^[14]计算子图核心节点特征向量, 匹配敏感大数据候选项特征相似性:

$$\cos(\theta) = \frac{\sum_{j=1}^n (v_j \times f_p)}{\sqrt{\sum_{j=1}^n (v_j)^2} \times \sqrt{\sum_{j=1}^n (f_p)^2}} \quad (11)$$

式中, v_j 为第 j 维的特征向量之间的数值差; τ 为邻居项目。具体弱无回溯模式匹配算法的敏感大数据特征相似性匹配流程如图 2 所示。

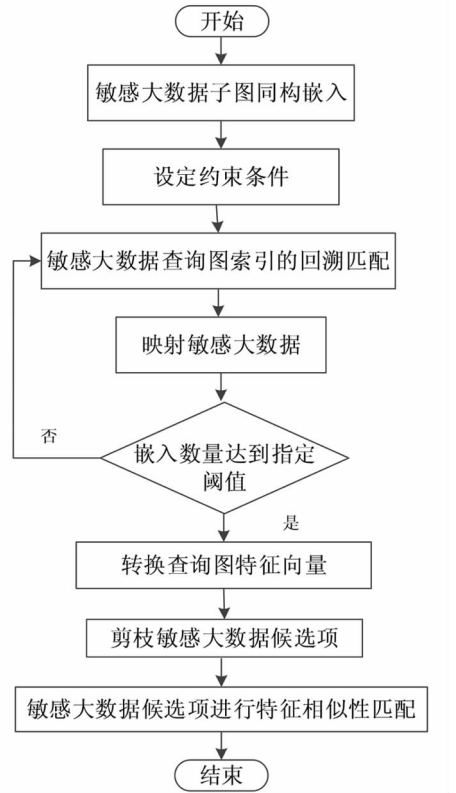


图 2 敏感大数据特征相似性匹配流程图

1.3 敏感大数据自动化分级检测

对特征相似性匹配后的敏感大数据进行特征相似性度量, 为后续敏感大数据自动化分级检测提供数据依据。引入熵参数^[15]量化敏感大数据特征相似性:

$$\epsilon = \cos(\theta) \sum_{j=1}^n c_n^j \lg c_n^j \quad (12)$$

式中, c_n^j 为敏感大数据分量类别 n 在第 j 维下的区间数香浓熵参数, 即量化后的区间数。结合式(12)所示的计算结果, 引入相似度指数函数^[16]计算敏感因子:

$$\kappa = D(t) \sum_{j=1}^n R_j t$$

(13)

式中, $D(t)$ 为相似度函数; R_j 为第 j 维的特征向量导数。

表 1 敏感大数据安全分级内容

κ	安全等级	分级描述
$0 < \kappa \leq 0.05$	非常安全	通过任何技术均不能采集、识别、提取敏感大数据
$0.05 < \kappa \leq 0.4$	一般安全	通过一定的技术识别少数敏感大数据,但无法具体到细化数据
$0.4 < \kappa \leq 0.6$	不安全	通过普通技术识别具有个人信息的敏感大数据,产生一定负面影响
$0.6 < \kappa \leq 1.0$	非常不安全	通过普通技术识别涉密等不予公开的敏感大数据,产生大范围负面影响

κ 值越大则表明敏感大数据越不安全,根据 κ 值大小,及时判断传输和存储过程中隐私泄露、数据滥用、安全风险等情况。至此结合敏感大数据样本库和敏感大数据特征相似性匹配完成对敏感大数据的分级检测,为后续的数据安全保障提供重要依据。

2 实验结果与分析

2.1 实验环境

选取国网冀北电力有限公司信息通信分公司的敏感大数据作为测试对象,结合配电网的实际运行情况,选择电网账户数据包展开测试分析。网络拓扑结构为局域网,带宽为 1 Gbps,通信方式为恒定字节速率,包速率为 2.5 pkt/s,分组负载为 700 bytes。敏感大数据以实时流、批量数据和周期性报告等形式进行传输,接收端与发送端之间距离为 1000 km,发送信息时的比特数为 1 bit,最大传输单元为 1200 bytes。测试过程中的硬件、软件参数如表 2 所示。

表 2 硬件、软件参数详情表

计算机系统	项目	参数
硬件方面	CPU	i3 2120
	主频	3.3 GHz
	物理内存	32 GB
	操作系统	windows 7
	开发语言	Python
软件方面	语料提取工具	NLP
	向量训练工具	Fast Text
	数据库处理工具	SQL server 2022

在电力有限公司信息通信环境场景下,使用

敏感因子 κ 是判断敏感大数据安全分级的指标,通常在 0~1 之间。根据敏感因子结果将敏感大数据划分为四个安全等级,具体如表 1 所示。

Python 中的 NLP 工具从视频弹幕、社交媒体帖子、电子邮件等真实大数据中提取包含“银行账户”“信用卡号码”“财务报表”等词汇的相关文本数据。使用正则表达式数据清洗方法,去除重复数据和无意义信息。使用 Fast Text 词向量训练工具,将清洗后的文本数据用于训练自定义的词向量模型。通过训练模型,将每个词汇表示为一个向量,以便后续的相似性匹配。利用训练好的词向量模型,在预处理的文本数据中进行关键词匹配,寻找与敏感信息相关的词汇、短语以及规则相似性特征。根据匹配结果,使用 Python 内置的字符串搜索方法,标记敏感信息所在的文本数据。使用 SQL Server 2022 数据库处理工具创建表结构,将标记后的敏感信息存入数据库,根据匹配到的信息完成分级。

2.2 实验测试结果

2.2.1 敏感大数据安全分级准确度测试

将电网账户数据包按照视频弹幕、社交媒体帖子、电子邮件划分为 3 个数据集,视频弹幕公开数据条数为 37405 条(敏感数据 20394 条),社交媒体帖子公开数据条数为 29637 条(敏感数据 12963 条)、电子邮件公开数据条数为 28621 条(敏感数据 18309 条)。利用勒索软件向目标服务器发送大量请求,可以识别视频弹幕数据集中少数敏感大数据,能够精准识别社交媒体帖子数据集中所有个人信息敏感大数据,能够获取社交媒体帖子数据集中涉密不予公开的敏感大数据,进而造成电力有限公司信息通信环境下的敏感大数据存在安全风险。采用所提方法对视频弹幕数据集、社交媒体帖子数据集、社交媒体帖子数据集进行相似度指数函数计算,获取 κ 值,划分敏感大数据安全等级,具体结果如表 3 所示。

表 3 敏感大数据安全分级结果

数据集	κ 值	安全等级
视频弹幕	0.37	一般安全
社交媒体帖子	0.55	不安全
电子邮件	0.98	非常不安全

根据表 3 结果可知,视频弹幕数据集 κ 值为 0.37,处于一般安全等级;社交媒体帖子数据集 κ 值为 0.55,处于不安全等级;电子邮件数据集 κ 值为 0.98,处于非常不安全等级。3 个数据集的测试结果与当前电力有限公司信息通信环境下的敏感大数据实际安全等级一致,由此证明所提方法可以准确划分敏感大数据安全等级,确保电力系统稳定运行。

2.2.2 敏感大数据特征匹配耗时测试

为了进一步验证所提方法的敏感大数据安全分级有效性,将文献[5]方法和文献[6]方法作为对比方法,进行匹配耗时测试分析。对不同输入规模的数据集进行敏感大数据特征匹配并记录耗时,观察并分析匹配时间随输入规模增加的趋势,耗时时间越低则表明方法计算量越小,时间复杂度越低。将视频弹幕数据集、社交媒体帖子数据集、电子邮件数据集进行组合,选择整数数据量进行敏感大数据特征匹配。不同方法的匹配耗时测试结果如图 3 所示。

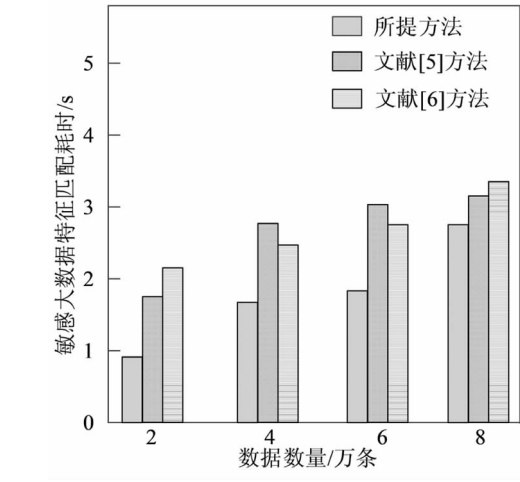


图 3 敏感大数据特征匹配耗时结果

分析图 3 可知,随着电网账户数据数量的增多,不同方法的匹配耗时均呈现上升趋势。当数据数量达到 8 万条时,文献[5]方法的耗时为 3.1 s,文献[6]方法的耗时为 3.2 s,而所提方法的耗时为

2.7 s。这是因为所提方法采用了弱无回溯模式匹配算法回溯匹配敏感大数据查询图索引,设定约束条件从查询图索引中选择最合适的匹配路径,以此提高了敏感大数据特征匹配效率,进而表明所提方法的计算时间复杂度较低。

2.2.3 敏感大数据特征匹配收敛速度测试

由于敏感大数据可能包含大量的数据项和复杂的特征,特征匹配的收敛速度可能会受到多个因素的影响,包括算法的选择、数据集的大小和复杂性,以及计算资源的可用性等。收敛速度越小表明方法能够更快地收敛到准确结果。测试所提方法与文献[5]方法、文献[6]方法的特征匹配收敛速度,测试结果如图 4 所示。

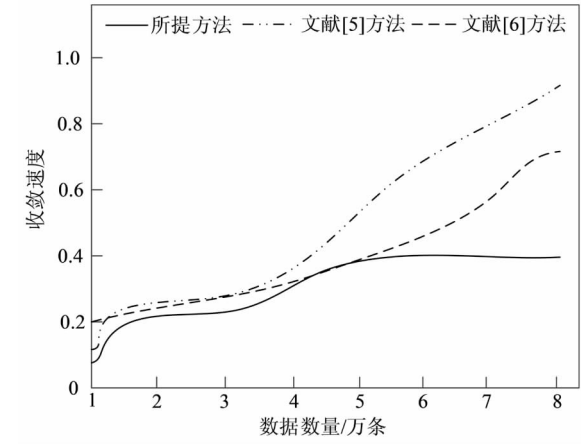


图 4 敏感大数据特征匹配收敛速度结果

分析图 4 可知,电网账户数据数量与收敛速度之间成正比关系,所提方法的收敛速度波动范围较小,当数据达到 5.3 万条时,收敛速度趋于平缓,而文献[5]方法和文献[6]方法的收敛速度波动趋势较大。当数据数量达到 8 万条时,所提方法的收敛速度为 0.44,而文献[5]方法的收敛速度为 0.91,文献[6]方法的收敛速度为 0.74。这是因为所提方法以建立协方差矩阵的方式捕捉不同类型的敏感大数据,获取其特征相关性,并以此为基础构建敏感大数据样本库,进而提高了敏感大数据属性显示效果,由此提升了敏感大数据特征匹配收敛速度。

3 结 论

为了解决敏感大数据传输和存储过程中的安全问题,提出了基于弱无回溯模式匹配算法的敏感

大数据安全分级检测方法。建立敏感大数据样本库,利用弱无回溯模式匹配算法回溯匹配敏感大数据查询图索引,通过余弦相似度计算完成敏感大数据特征相似性匹配。采用相似度指数函数计算敏感因子,完成敏感大数据自动化安全分级检测。测试结果表明,所提方法敏感大数据安全分级准确度较高、特征匹配耗时较短、收敛速度较快。虽然所提方法在敏感大数据的安全分级检测中具有较高的准确度和效率,但在数据传输和存储过程的加密保护有待改进。未来的研究将结合本文成果,引入隐私保护算法、身份认证机制等进一步提高数据传输和存储过程中的敏感大数据安全性,以期为敏感大数据保护和隐私安全领域做出积极贡献。

参考文献

[1] 方朝剑,胡新荣. 基于模糊近似度的隐私敏感数据过滤算法[J]. 吉林大学学报(工学版),2023,53(4):1174—1180.

[2] 崔嵬,杨同军,苗凯. 基于代理重加密的财务数据安全共享技术[J]. 现代电子技术,2023,46(1):74—78.

[3] 张思琪,李晓会,江欣俞,等. 面向医疗数据的 AUR-Tree 差分隐私数据发布算法[J]. 计算机应用研究,2022,39(7):2162—2166+2171.

[4] 黄志清,解鲁阳,张严心,等. 基于区块链的物联网数据服务信誉评估模型[J]. 计算机工程,2022,48(1):33—42.

[5] 王伟,庞婷婷. 基于数据挖掘的城市轨道交通信息安全检测[J]. 微型电脑应用,2022,38(2):113—115

[6] 高丽杰,薛丽香,马照瑞,等. 大数据 RSA 公钥加密安全双向检测仿真[J]. 计算机仿真,2023,40(2):410—414

[7] PERCIVAL W J, FRIEDRICH O, SELLENTIN E, et al. Matching Bayesian and frequentist coverage probabilities when using an approximate data covariance matrix [J]. Monthly Notices of the Royal Astronomical Society, 2022, 510(3): 3207—3221.

[8] IKRAM R M A, GOLIATT L, KISI O, et al. Covariance matrix adaptation evolution strategy for improving machine learning approaches in streamflow prediction[J]. Mathematics, 2022, 10(16): 2971.

[9] 于慧琳,陈炜,王琪,等. 使用子图推理实现知识图谱关系预测[J]. 计算机科学与探索,2022,16(8):1800—1808.

[10] 郁湧,王莹港,罗正国,等. 基于聚类系数和节点中心性的链路预测算法[J]. 清华大学学报(自然科学版),2022,62(1):98—104.

[11] 古险峰,程艳艳,杨立英. 基于回溯模式总结的鲁棒近似子图查询算法[J]. 计算机工程与设计,2022,43(10):2822—2828.

[12] 申滨,孙万平,张楠,等. 基于加权二部图及贪婪策略的蜂窝网络 D2D 通信资源分配[J]. 电子与信息学报,2023,45(3):1055—1064.

[13] 王乐洋,罗鑫磊. 非线性函数协方差传播的 Stein Monte Carlo 方法[J]. 大地测量与地球动力学,2022,42(1):1—4.

[14] WANG F, KONG A W K. Exploiting the relationship between Kendall's rank correlation and cosine similarity for attribution protection[J]. Advances in Neural Information Processing Systems, 2022, 35: 20580—20591.

[15] 严钧,晏婉晨. 基于 Gumbel 分布的熵风险度量的参数估计及渐近行为[J]. 河南师范大学学报(自然科学版),2022,50(1):67—72.

[16] 陶新民,王若彤,常瑞,等. 基于低密度分割密度敏感距离的谱聚类算法[J]. 自动化学报,2020,46(7):1479—1495.