计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200047

基于粒关联的数据聚合信息级别推演方法

李金辉[1]　曹利峰[1]　汪小芹[2]　白金龙[1]　陈　阳[1]
1 河南省信息安全重点实验室　郑州 450000
2 中国电子科技集团公司第七研究所　广州 510277
(jineh1214@163.com)

摘　要　为解决大数据聚合而引起敏感信息泄露的问题,对数据之间的关联性进行了深入的分析,提出了基于粒关联的数据聚合信息级别推演方法。根据数据属性的依赖关系,挖掘出高关联度的数据对象,进而依据数据对象关联属性的敏感级别模糊集可能性测度推演用户访问多信息系统时由数据聚合推导高敏感级别信息的可能性。这种方法有助于为用户制定数据访问策略,控制对关联数据的分析,降低信息泄露的风险。

关键词:数据分析;粒关联;关联规则;聚合推演;信息泄露
中图分类号　TP309

Information Level Inference Method for Data Aggregation Based on Granular Association

LI Jinhui[1],CAO Lifeng[2],WANG Xiaoqin[2],BAI Jinlong[2] and CHEN Yang[2]
1 Henan Provincial Key Laboratory of Information Security,Zhengzhou 450000,China
2 China Electronics Technology Group Corporation Seventh Research Institute,Guangzhou 510277,China

Abstract　To address the issue of sensitive information leakage through the existence of big data aggregation,this study analyzes the correlation between data deeply and proposes an information level inference method for data aggregation based on granular association. The method mines highly associated data objects based on the dependencies of data attributes,and then deduces the possibility of inferring highly sensitive information from data aggregation when users access the multi-information system based on the fuzzy set possibility measurement of the sensitivity level of the associated attributes of the data objects. This approach aids in establishing access policies for users,controlling the control the analysis of associated data,and reducing the risk of information leakage.

Keywords　Data analysis,Granular association,Association rules,Aggregation inference,Information leakage

1　引言

大数据、云计算、区块链等新技术的不断涌现,掀起了数字化经济的潮流,各行各业智能化、数字化程度日益提升,加速了互联网中数据收集、聚合和共享的需求[1]。企业之间建立了彼此独立的信任域,然而,良好的数据交互共享可以打破各信任域之间的信息壁垒,挖掘出数据更深层次的价值[2]。但多个信息系统之间可能存有用户权限不对等、数据之间存在隐式的关联关系等可能造成信息泄露的问题。新冠疫情期间,各地的健康宝数据展现了跨域数据共享的能力。但在一过程中,敏感数据的机密性受到了极大挑战。具体来说,不同组织在安全要求和数据管理级别上存在差异,导致了数据采集和处理权限的不均衡。一些权属部门拥有广泛的数据采集权限和较高的数据处理权限,而其他部门则可能仅需有限的数据访问权限,由此形成了高安全域和低安全域的差别。这种差异可能导致高安全域向低安全域开放其无权访问的数据,即敏感数据暴露。此外,在数据发布或共享时,机构部门可能会对高敏感信息进行匿名化处理,然而攻击者可能通过来自不同信息来源中数据对象之间的关联关系,推导出高敏感信息。例如,某攻击者获取了某学生的选课信息及其出勤记录,那么攻击者可从其他信息源所发布的包含选课记录及出勤情况与课程成绩的匿名化数据中,推断出该同学的成绩,但数据中并未公开其信息。除了教育领域中学生成绩的泄露外,其他领域也存在由数据对象之间的关联关系所导致的高敏感信息被推导获取的问题。因此,如何评估数据聚合过程中隐私泄露的风险成为以上场景下的关键科学问题,分析全局数据,识别隐式关联关系,从关联数据的聚合中推断高敏感级别信息造成信息泄露的可能性,评估用户跨信息系统时的访问策略具有重要的研究价值。这些操作对于维护多信息系统的正常运行至关重要。

由数据聚合导致的安全问题主要指在用户进行数据访问时,所访问的权限以内的数据聚合后能够推导出敏感级别超出用户权限的数据信息,造成用户越权访问行为,从而引发信息泄露的风险。如图1所示,每个信任域中都存储有敏感级别不同的数据,其中P代表公开数据,S代表秘密数据,C代表机密数据,它们的敏感等级逐步提升。而域中用户的权限不同,假设信任域A与信任域B之间的用户可以相互访问对方域中的信息。域A中敏感级别为P和S的数据 $Data_{2\_D_A}$,$Data_{3\_D_A}$ 和 $Data_{4\_D_A}$ 融合在一起后能够推导出敏感级别为C的数据 $Data_{9\_D_A}$,信任域B中敏感等级为P的数据 $Data_{1\_D_B}$ 和

通信作者:曹利峰(caolf302@sina.com)

$Data_{2\_D_B}$ 聚合时，可以推导出信息敏感等级为 S 的数据 $Data_{6\_D_A}$。说明这些数据信息具有强关联性，若数据聚合在一起，则能够推导出敏感等级超过用户访问权限的数据信息，它们应是不被允许共同访问的。当用户对域空间下数据进行访问时，就可能推导出高于该用户权限的数据信息，从而造成信息泄露，引发安全问题。因此应当推演全域数据进行聚合推导更高敏感等级信息的可能性，以调整用户对权限内数据的访问，防止用户在对数据进行访问时造成越权访问行为，降低敏感信息泄露的风险。

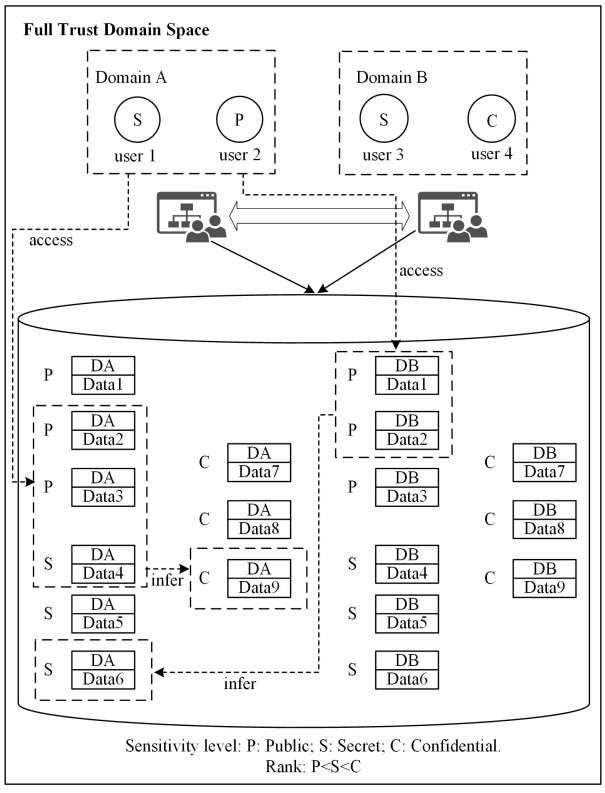

图 1 数据聚合问题示意图

Fig. 1 Diagram of the data aggregation problem

2 相关研究

由数据聚合所导致的可推导出高敏感级别数据信息属于数据的逻辑推理问题，这是一个复杂的弱形式化问题，当用户在没有直接访问的情况下泄露了某些受保护的数据信息时，就会发生逻辑推理，其中直接访问是指数据允许某个用户访问或查询[3]。用户所获得的聚合信息使得信息系统在没有特殊保护措施的情况下可以推断出特定的隐私和机密信息。为保护在用户访问信息系统时数据信息的安全，共有两种防护策略，分别是响应式的防护方式和预防性的防护方式[4]。响应式的防护方式主要是当信息系统监测到异常的数据访问模式或数据库中某些敏感数据被非法获取时做出反应，对其访问连接进行阻断或向管理员发送警报。而预防性的防护方式侧重于在信息泄露事件发生之前采取措施来降低风险。从计算的消耗来看，响应式的防护方式成本更高，对系统的存储和计算能力都有较高的要求，而预防性的防护方式更便于用户使用，通过提前设置规则和防护机制，能够在安全事件发生之前进行复杂的实时分析和处理，不会出现资源过度消耗的问题。

在信息技术的演变历程中，此类问题最早出现于数据库体系中，不法分子为获取某数据库中的隐私敏感信息，利用相似聚类逻辑推理来分析数据的相似特性推断敏感信息，利用关联推理来拼凑数据之间的隐蔽联系，从而获得受保护的数据信息。Cao 等[5]通过对客体之间的关联性进行深入研究，提出了一种基于属性关联的客体聚合信息级别推演方法，来指导网络边界访问控制策略的实施，降低信息系统泄密的风险。Cao 等[6]针对该问题，对同一安全域中的客体资源进行相似性分析，提出了基于聚类分析的客体资源聚合信息级别推演方法，来指导等级化网络区域边界访问控制策略的制定，控制主体对相似客体的受限访问。

由于数据量的爆炸式增长，进入大数据时代，不同来源的数据可能包含了同一主体相关的信息，这意味着有更多的数据可供挖掘利用，使得数据泄露风险进一步增大。不法分子可通过整合多源数据，分析用户多场景下的行为数据，挖掘其关联性来推测隐蔽数据。此外，针对数据的相似性，运用分类预测模型与聚类分析技术发现数据中的潜在特征和规律，进而推断受保护的信息。企业或组织之间的数据共享和第三方合作导致数据共享过程中缺乏严格的隐私保护措施，使得可直接访问的数据被恶意利用于推断超出共享范围的受保护信息，从而造成用户的个人隐私信息暴露。Liu 等[7]根据信息物理融合系统 CPS 的客观规律及业务流转间复杂的关联关系，通过分析现有的数据泄露威胁建立了数据窃取和数据推断两种数据泄露的威胁范式，并提出面向分类分级的数据推断关联性进行评估来对跨安全等级数据共享时进行安全防护。Hao 等[8]提出了一种基于改进的 K-means 聚类算法的并行关联规则挖掘方法，建立数据对象准则函数，利用改进的 K-means 聚类算法对大数据进行聚类，并构建可拓的物元关系数据库模型实现对大数据关联关系的并行挖掘。

人工智能的出现使得关联挖掘达到了更高层次的数据处理水平，同时对数据安全带来了更大的威胁。恶意用户可利用深度学习模型的神经网络架构提取数据特征，并与其相关数据聚合后推断出受保护信息。此外，并可通过整合多源数据构建知识图谱并进行推理，综合分析医疗、社交、消费等数据，挖掘出更深层次受保护的信息。Liu 等[9]提出了一种特征选择与突出分类器框架，通过提取数据与属性特征的关联规则，利用 Boruta 从中选择出高度相关的最优特征，并采用随机森林为不同特征分配权重，最后基于样本数据和最优特征，利用贝叶斯优化的 KNN 模型来对数据进行预测。Lu 等[10]在对多模态数据进行建模时引入了超图模型[11]，将数据对象之间的一对一关系扩展为多对多关系，更准确地描述了对象之间的关联关系。Wang 等[12]对实体之间的关联信息进行提取和利用，提出了知识图谱补全的 AiTransE 模型，用来计算首尾实体与直接关系的关联程度，并将线性加权的关联度引入目标函数来提高模型性能。

利用人工智能技术来挖掘关联数据以发现访问数据聚合时推导高敏感级别信息的问题，在挖掘不同类型的数据特征、处理多维度的数据特征上起到了重要作用，但仍存在诸多缺陷，在实际应用中用户难以确定模型所挖掘的特征是否合理，以及难以对模型的决策依据进行有效评估和审核。此外，人工智能模型的训练和更新需要大量的计算资源和时间，随着数据的动态更新，需要模型重新进行训练，造成较高的计算成本且限制了模型对数据动态变化的适应能力。将粒计算融入

至数据安全领域，能够有效防止逻辑推理所导致的数据敏感信息泄露，实现更高层次的动态数据分析与推理，能够从微观的粒度层面深入剖析数据的内在特征及关联关系。Lin 和 Zhang[14]基于数据多粒度来将个人数据以多层次粒度进行分层分级划分，以明确隐私保护。他们根据不同层次粒度之间的序关系来指定用户的管理权与使用权，使数据得到更具针对性的隐私保护。Cao 等[15]在文献[6]的基础上针对由相似数据推导出敏感信息导致信息泄露的问题，提出了基于粒分析的大数据聚合信息敏感性推演方法。该方法依据数据属性形成数据粒集，利用聚类建立相似粒子云，依据属性模糊集可能性测度及粒对敏感粒子云的贡献度对信息泄露问题发生的可能性进行推演，通过对相似数据的分析来降低信息的泄密风险，但并未对关联数据进行分析。故本文针对关联数据的分析，在文献[5]的基础上提出了基于粒关联的数据聚合信息级别推演方法，通过对数据进行粒化，挖掘不同粒度下数据属性之间的关联关系，进而发现高关联度的数据对象，最终计算由数据聚合而导致推导出更高敏感等级数据的可能性。

3 基于粒关联的数据分析方法

多信息系统下存储着大量的、形态各异的数据信息，数据之间的逻辑推导关系主要表现为：

1) $data_i \rightarrow data_j$；

2) $data_i \rightarrow data_j, data_j \rightarrow data_k \Rightarrow data_i \rightarrow data_k$；

3) $data_i \xrightarrow{p_1} data_k, data_j \xrightarrow{p_2} data_k \Rightarrow (data_i, data_j) \xrightarrow{p(\geqslant\{p_1,p_2\})} data_k$；

4) $data_i \xrightarrow{p_1} data_k, data_j \xrightarrow{p_2} data_k \Rightarrow data_i \xleftrightarrow{p(\leqslant\{p_1,p_2\})} data_j$ 等。

其中，$data_i$ 表示数据，$\rightarrow$ 表示推导关系，p_i 表示推导概率。针对以上数据之间的逻辑推导关系，提出将数据进行粒化，进而对粒化数据的属性之间所存在的隐式关联关系进行分析。

3.1 粒与粒特征

令 U 为某领域对象的全域空间，$U=\{x_1, x_2, \cdots, x_{|U|}\}$，$AT$ 为属性空间，$AT=\{a_1, a_2, \cdots, a_{|AT|}\}$，$a_i$ 为属性集的第 i 个属性，属性集 AT 由条件属性集 C 和决策属性集 D 两部分组成，并且 $C \cup D=AT$，$C \cap D=\varnothing$。

定义 1（粒） 令 $a \in C$，vc 为属性 a 在 C 上某个数据对象的取值；$m()$ 为 C 上的意义集函数，即 a 为属性类型、vc 为属性值；(a, vc) 为 C 上的一个原子公式，记作 a_{vc}；a_{vc} 的组合为 C 上的合式公式，记作 $\varphi=\prod a_{vc}$，则 $(\varphi, m(\varphi))$ 称为 U 上的一个属性粒 cg(condition granules)。其中，$m(\varphi)=\{x, x \in U, |x| \approx \varphi\}$，即 U 上所有满足 φ 的数据对象集合。按照属性值进行数据划分，称为数据粒化，最终得到各类属性粒。条件属性集 C 对论域 U 的划分 $U/IND(C)=\{E_1, E_2, \cdots, E_t\}$，$E_j(1 \leqslant j \leqslant n)$ 为条件属性 C 对 U 的一个属性粒 cg。决策属性集 D 对论域 U 的划分为 $U/IND(D)=\{D_1, D_2, \cdots, D_n\}$，$D_j(1 \leqslant j \leqslant n)$ 为决策属性 D 对 U 的一个决策粒 dg(decision granules)，决策粒是一种特殊的属性粒，即 $dg \subseteq cg$。由属性粒 $cg_i(cg_i \in U/C_1)$ 与属性粒 $cg_j(cg_j \in U/C_2)$ 所形成的关联关系为 $cg_i \rightarrow cg_j$，由属性粒 $cg(cg \in U/C)$ 与决策粒 $dg(dg \in U/D)$ 所形成的关联关系为 $cg \rightarrow dg$。

定义 2（粒结构） 令 $cg=(\varphi, m(\varphi))$，$\alpha_i \subseteq \varphi$，$cg_i=(\alpha_i, m(\alpha_i))$，则称 $cgs=\{cg_i, L, \angle\}$ 为属性粒结构。其中，L 为粒结构的层次，$\angle$ 为粒层之间的偏序关系。

属性粒结构是一个层次化结构。粒层之间的偏序关系 $\angle$ 包括两类，一类为属性包含关系，另一类为属性推导关系。同层属性粒间也包括两类关系，一类为相似关系，另一类为相互独立关系，相互独立的属性粒之间不存在相似关系。

图 2 为包含关系的粒结构，通常为单属性粒、多属性粒组成的层次化结构。若同层粒之间为相似粒，即 cg_α 与 cg_β 相似，则形成的上层属性粒 $cg_{\alpha \| \beta}$ 为下层属性粒的合取，即 $cg_{\alpha \| \beta}=cg_\alpha \wedge cg_\beta=(\alpha \wedge \beta, m(\alpha) \wedge m(\beta))$，表示相似粒共性的属性特征与意义。若同层粒之间为相互独立粒，即 $\alpha \cap \beta=\varnothing$，则形成的上层属性粒为下层数据粒的析取，即 $cg_{\alpha \| \beta}=cg_\alpha \vee cg_\beta=(\alpha \vee \beta, m(\alpha) \vee m(\beta))$，表示独立属性粒的融合特征与意义。

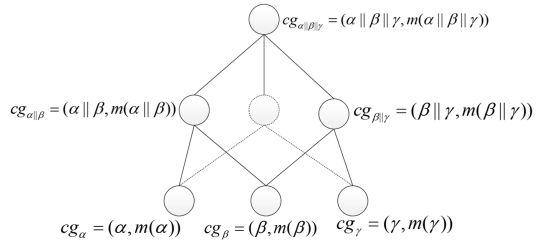

图 2 包含关系属性粒结构

Fig. 2 Inclusive relational condition granules

图 3 为推导关系的粒结构，是以粒推导关系为基础构建的层次化结构。图 3(a)中，若 $cg_\alpha \xrightarrow{\lambda_1} cg_\gamma, cg_\beta \xrightarrow{\lambda_2} cg_\gamma$，即 cg_α 以依赖度 λ_1 推导出 cg_γ，cg_β 以依赖度 λ_2 推导出 cg_γ，则 $cg_\gamma=((\alpha, \beta \xrightarrow{\exists \lambda \lambda \geqslant \{\lambda_1, \lambda_2\}} \gamma), ((m(\alpha), m(\beta) \xrightarrow{\exists \lambda \lambda \geqslant \{\lambda_1, \lambda_2\}} m(\gamma)))$，表示多属性粒聚合推导单一属性粒的特征与意义。图 3(b)中，如果 $cg_\alpha \xrightarrow{\lambda_1} cg_\gamma, cg_\alpha \xrightarrow{\lambda_2} cg_\rho, cg_\beta \xrightarrow{\lambda_3} cg_\gamma, cg_\beta \xrightarrow{\lambda_4} cg_\rho$，即 cg_α 分别以依赖度 λ_1 和 λ_2 推导出 cg_γ 和 cg_ρ，cg_β 分别以依赖度 λ_3 和 λ_4 推导出 cg_γ 和 cg_ρ，那么 $cg_\alpha \xrightarrow{\exists \lambda' \lambda' \leqslant \{\lambda_1, \lambda_2\}} cg_{\gamma \| \rho}$，即 $cg_{\gamma \| \rho}=\{\alpha \xrightarrow{\exists \lambda' \lambda' \leqslant \{\lambda_1, \lambda_2\}} \gamma \| \rho, m(\alpha) \xrightarrow{\exists \lambda' \lambda' \leqslant \{\lambda_1, \lambda_2\}} m(\gamma \| \rho)\}$。同样，$cg_\beta \xrightarrow{\exists \lambda'' \lambda'' \leqslant \{\lambda_3, \lambda_4\}} cg_{\gamma \| \rho}$，即 $cg_{\gamma \| \rho}=\{\beta \xrightarrow{\exists \lambda'' \lambda'' \leqslant \{\lambda_3, \lambda_4\}} \gamma \| \rho, m(\beta) \xrightarrow{\exists \lambda'' \lambda'' \leqslant \{\lambda_3, \lambda_4\}} m(\gamma \| \rho)\}$，那么也必然存在，$cg_\alpha, cg_\beta \xrightarrow{\exists \lambda \lambda \geqslant \{\lambda', \lambda''\}} cg_{\gamma \| \rho}$。图 3(b)表示多属性粒聚合推导分散属性粒的特征与意义。

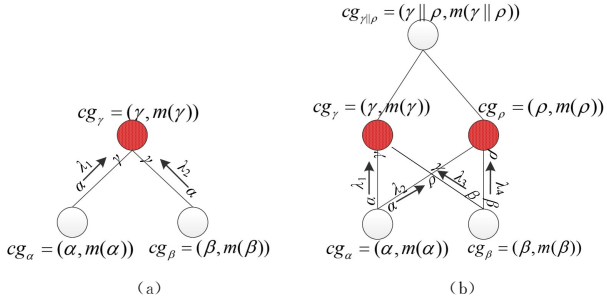

图 3 推导关系属性粒结构

Fig. 3 Inference relational condition granules

定义 3（支持度与置信度） 令属性粒规则集的项为 $\langle cg_i \rightarrow cg_j, sup, conf \rangle$，其中 sup 和 $conf$ 分别为属性粒关联规则 $cg_i \rightarrow cg_j$ 的支持度和置信度(又称依赖度)：

$$sup(cg_i \rightarrow cg_j) = \frac{|cg_i \wedge cg_j|}{|U|}$$

$$conf(cg_i \rightarrow cg_j) = \frac{|cg_i \wedge cg_j|}{|cg_i|}$$

其中，$|\cdot|$表示集合的基。

3.2 基于粒关联的数据属性关联关系挖掘算法

为了提高发现属性粒之间关联关系的效率，提出了基于粒关联的数据关联关系挖掘算法，利用属性约简的方法挖掘数据属性之间的关联推导关系，即属性粒与决策粒之间的关联关系。

定义 4（概念对象） 令 bD 为同一信任域下的数据集合，$bD=\{bd_1, bd_2, \cdots, bd_l\}$ $(bd_i \subseteq U)$，其中，$bd_i=\{x_1, x_2, \cdots, x_n\}$，$S=\{s_1, s_2 \cdots, s_k\}$ 为敏感集，设置 bd_i 为一个数据列表，每个列表中存储一组数据信息并包含多个数据对象。令 x_i 的属性集为 $\omega_{ic}=\langle a_{i1}, a_{i2}, \cdots, a_{ik}, d_i\rangle$，$d_i$ 为 x_i 的决策属性，ω 中属性对应的属性值为 $v_i=\langle v_{i1}, v_{i2}, \cdots, v_{ik}, v_{id}\rangle$，定义数据概念对象 xg_i 为 $\langle x_i, \langle \omega_i \cdot v_i, s_{x_i}\rangle$。多个数据对象可能存在于同一属性粒中。

定义 5（粒特征矩阵与特征值矩阵） $U/C=\{E_1, E_2, \cdots, E_l\}$，$U/D=\{D_1, D_2, \cdots, D_k\}$。$E_i$ 的特征向量为 $\vec{\boldsymbol{E}}_i=\langle index_i, obj_i, reg_i, \delta_i\rangle$，$\vec{\boldsymbol{E}}_{ic}=\langle e_{i1}, e_{i2}, \cdots, e_{ij}, \cdots, e_{im}\rangle$ 为 E_i 的特征值向量，其中 $e_{kj}=f(x_k, a_j)$ $(\forall x_k \in E_i, \forall a_j \in C)$，$f(x_k, a_j)$ 表示 x_k 在 a_j 属性上的属性值，$index_i \in x_k$ $(\exists x_k \in E_i)$ 是特征索引，obj_i 为数据对象 xg 的集合，$obj_i=\{xg_k \mid x_k \in E_i\}$。若 $E_i \subseteq POS_C(U/D)$，即等价关系划分的正域，则 $reg_i=P$。若 $E_i \subseteq BND_C(U/D)$，即等价关系划分的边界域，则 $reg_i=B$。$\delta_i=\delta_C(E_i)=\{D_j \mid E_i \cap D_j \neq \varnothing\}$。粒特征矩阵和特征值矩阵定义如下：

$$\boldsymbol{D}_E=\begin{bmatrix}\vec{E}_1\\ \vec{E}_2\\ \vdots\\ \vec{E}_l\end{bmatrix}=\begin{bmatrix}index_1 & obj_1 & obj_2 & reg_1\\ index_2 & obj_2 & reg_2 & \delta_2\\ \vdots & \vdots & \vdots & \vdots\\ index_l & obj_l & reg_l & \delta_l\end{bmatrix}$$

$$\boldsymbol{M}_{EC}=\begin{bmatrix}\vec{E}_{1c}\\ \vec{E}_{2c}\\ \vdots\\ \vec{E}_{lc}\end{bmatrix}=\begin{bmatrix}e_{11} & \cdots & e_{1m}\\ e_{21} & \cdots & e_{2m}\\ \vdots & \ddots & \vdots\\ e_{l1} & \cdots & \boldsymbol{e}_{lm}\end{bmatrix}$$

定义 6（属性分配约简与重要度矩阵） 定义 $Att_{\min}$ 为最小辨识属性集，$Att_{\min}=\{Att_0, Att_1, \cdots, Att_r\}$，其中，对于 $\forall Att_i \in Att_{\min}$，$\exists D(E_i, E_j)$ 满足 $Att_i \in D(E_i, E_j)$，且对于 $\exists Att_i \in Att_{\min}$，$\forall D(E_i, E_j)$ 满足 $Att_i \in D(E_i, E_j)$。当 $(E_i, E_j) \in D_l^*$ 时，$D_l(E_i, E_j)=\{a_k \in AT \mid f(E_i, a_k) \neq f(E_j, a_k)\}$。最小辨识属性集中的元素彼此不相同，并且它们之间不存在包含关系。令 $Important_i$ 表示属性 a_i 在分配约简中的属性重要度，则属性重要度矩阵定义为：

$$\boldsymbol{M}_{\text{Important}}=[\boldsymbol{AT}\ \boldsymbol{IM}]=\begin{bmatrix}a_1 & Important_1\\ a_2 & Important_2\\ \vdots & \vdots\\ a_m & Important_m\end{bmatrix}$$

其中，$\boldsymbol{AT}$ 表示属性重要度的属性向量，$\boldsymbol{IM}$ 表示重要度向量。

根据以上定义，提出了基于粒关联的数据属性关联关系挖掘算法，如算法 1 所示。

算法 1 基于粒关联的数据属性关联关系挖掘算法

输入：全域对象 U

输出：关联规则集 R

```
1. e=1
2. for i=1 to |U| do
3.     for j=1 to e do
4.         if xg_i.<ω_i · v_i>=E_j.<ω · v> then
5.             xg_i→E_j
6.         else e++, xg_i→E_e
7.     end for
8. end for
9. for i=1 to e do
10.    E_i→𝔼
11. end for
12. k=1
13. for i=1 to |U| do
14.    for j=1 to k do
15.        if xg_i.⟨d_i⟩=D_j.⟨d⟩ then xg_i→D_j
16.        else k++, xg_i→D_k
17.    end for
18. end for
19. for i=1 to d then
20.    D_i→𝔻
21. end for
22. for i=1 to |U| do
23.    for j=1 to e do
24.        if E_j⊆D_i then C_(D_i)←E_j
25.        if E_j∩D_i≠∅ then C¯(D_i)←E_j
26.    end for
27. end for
27. for i=1 to e do
29.    if E_i⊆C_(D_i) then reg_i=POS
30.    else reg_i=BND
31.    E⃗_i⟨E_i.⟨x⟩[1], E_i.⟨x⟩, reg_i, δ(E_i)⟩
32.    E⃗_ic←⟨E_i, ⟨ω_i · v_i⟩⟩
33.    D_E←E⃗_i
34.    M_EC←E⃗_ic
35. end for
36. for i=1 to e do
37.    for j=i+1 to e do
38.        if δ_i≠δ_j(δ_i, δ_j∈D_E) then {
39.        compute D(E_i, E_j) by M_EC
40.        if Att_min=∅ then Att_min←D(E_i, E_j)
41.        else UpdataAtt(D(E_i, E_j), Att_min)
42.        }
43.    end for
44. end for
45. compute M_D by Att_min
46. compute M_Important, M≥_Important by M_D
47. compute Red by M≥_Important, Att_min
48. for E⃗_i, E⃗_ic to D_E, M_EC do
49.    for i=1 to |Red| do
50.        cg←⟨Red_i · v_i⟩
51.        if i=|Red| then {
```

```
52.        for j=1 to |Ei · ⟨δ(Ei)⟩| do
53.          dg←⟨δ(Ei)[i]⟩
54.          rule←⟨cg,dg⟩
55.        end
56.      }
57.    end for
58. end for
59. for i=1 to |rule| do
60.    compute sup,conf by rulei
61.    if sup≥min_sup,conf≥min_conf then
62.      R←rulei
63. end for
64. return R
```

该算法主要分为 3 部分:第一部分(1－21 行)对全域空间中的数据对象进行预处理,根据数据的条件属性和决策属性,进行初次粒划分,得到粒集合$\mathbb{E}$和$\mathbb{D}$,其中包括了多个属性粒。第二部分(22－35 行)根据上、下近似集对各个粒进行进一步划分,得到粒的正域 POS 和边界域 BND,进而构造粒特征矩阵 $\boldsymbol{D}_E$ 和特征值矩阵 $\boldsymbol{M}_{EC}$,第三部分(36－64 行)依据最小辨识属性集的分配约简思想得到差别矩阵 $\boldsymbol{M}_D$ 和条件属性集的约简 Red,得到属性粒的关联关系,并将大于等于最小支持度 min_sup 和最小置信度 min_conf 的关联规则记录至关联规则集 R 中。

算法中所涉及到的符号解释如表 1 所列。

表 1　符号解释

Table 1　Symbol interpretation

Symbol	Interpretation
U	全域对象
xg_i	数据概念对象
ω	属性
v	属性值
d	决策属性
E_i	一个属性粒 cg
$\mathbb{E}$	属性粒集合
D_i	一个决策粒 dg
$\mathbb{D}$	决策粒集合
$\underline{C}(D_i)$	上近似集
$\overline{C}(D_i)$	下近似集
reg_i	等价关系划分
POS	正域
BND	边界域
$\delta(E_i)$	该属性粒数据所对应的决策属性值
$\vec{E}_i$	粒特征向量
$\boldsymbol{D}_E$	粒特征矩阵
$\vec{E}_{ic}$	粒特征值向量,划分该粒的属性值
$\boldsymbol{M}_{EC}$	特征值矩阵
$Att_{\min}$	最小辨识属性集
Red	属性约简
R	关联规则集

3.3　基于关联推导关系的高关联度数据对象发现算法

本文研究的重点是关联数据聚合时导致高敏感级别信息泄露的问题,故针对由算法 1 所挖掘得出的属性粒之间关联关系中前置条件的属性粒,对其所包含的数据对象之间的关联度进行挖掘。

定义 7(数据对象的关联度)　令 k 为 R 中规则的个数,R_i 为第 i 个规则,数据对象 xg_i 和 xg_j 的关联度为 $assoc(xg_i,xg_j)$,定义为:

$$assoc(xg_i,xg_j)=\sum_{i=1}^{k}w_i conf(R_i)$$

其中,w_i 为 R_i 的权重,定义为:

$$w_i=Info(cg_i)S^{\max}/\sum_{i=1}^{k}Info(cg_i)S^{\max}$$

其中,$Info(cg_i)$为属性粒的信息量,$S^{\max}$为属性粒中属性最高敏感级别的量化值。令 $minassoc$ 为数据对象最小关联度阈值,若满足 $assoc(xg_i,xg_j)\geqslant minassoc$,则数据对象 xg_i 和 xg_j 被称为强关联数据对象,简称关联对象。

根据以上定义,本文提出了一种基于关联推导关系的高关联度数据对象发现算法,旨在识别具有高度关联的数据对象。如算法 2 所示。

算法 2　基于关联推导关系的高关联度数据对象发现算法

输入:全域对象 U,粒关联规则 R

输出:高关联度数据对象集合 Q

```
1. for i=1 to |U|−1 do
2.   for j=i+1 to |U| do
3.     for l=1 to |R| do
4.       if[xgi,xgj]∈Rl.⟨obj⟩ then {
5.         wl=Info(cgl)S^max/Σ_{i=1}^{k} Info(cgi)S^max
6.         assoc(xgi,xgj)=wl * conf(Rl)
7.         if assoc(xgi,xgj)≥min_assoc then
8.           Q←⟨(xgi,xgj),assoc(xgi,xgj)⟩
9.       }
10.    end for
11.  end for
12. end for
13. return Q
```

3.4　关联数据属性的动态更新算法

随着数据对象等动态增长,数据属性也随之增加,属性粒将会随之改变,那么数据对象之间的关联会随属性粒的变化而变化,其中包括了数据对象属性之间的关联程度以及可能出现的新的属性关联关系。因此,对关联数据的动态更新能够更好地维护全局数据的关联性。本文在挖掘数据属性关联关系的基础上提出了关联属性的动态更新算法,如算法 3 所示。

算法 3　关联属性的动态更新算法

输入:新增数据 ΔU

输出:动态更新的关联规则集 R^{t+1}

```
1. e=|𝔼|
2. for i=1 to |ΔU| do
3.    if Red^t.⟨ω⟩⊆Δxgi.⟨ωi⟩ then
4.      if Red^t.⟨v⟩=Δxgi.⟨vi⟩ then
5.        Δxgi→E_Red
6.    for j=1 to e do
7.      if Δxgi.⟨ωi · vi⟩=Ej.⟨ω · v⟩ then
8.        Δxgi→Ej
9.      else e++,Δxgi→Ee
10.   end for
11. end for
12. l=|𝔻|
13. for i=1 to |ΔU| do
14.   for j=1 to l do
15.   if Δxgi.⟨di⟩=Dj.⟨d⟩ then
16.     Δxgi→Dj
17.     else l++,Δxgi→Dl
18.   end for
```

19. end for
20. k=1
21. for i=1 to e do
22. for j=1 to e do
23. if $xg_i.\langle \Delta\omega_i \cdot \Delta v_i\rangle = E_j.\langle \Delta\omega \cdot \Delta v\rangle$ then
24. $E_k^{t+1} \leftarrow xg_i$
25. else k++, $E_k^{t+1} \leftarrow xg_i$
26. end for
27. end for
28. for $item_1$, $item_2$ in E^{t+1}, D^{t+1} do
29. $item_1 \rightarrow \mathbb{E}$
30. end for
31. compute D_E^{t+1}, M_{EC}^{t+1}
32. compute Red^{t+1} by D_E^{t+1}, M_{EC}^{t+1}
33. compute $rule^{t+1}$ by Red^{t+1}
34. compute R^{t+1} by R^t, Δrule, min_sup, min_conf
35. return R^{t+1}

该算法首先根据 *Red* 所包含的属性集来对新增数据进行筛选，将与 *Red* 中属性值相同的 xg 直接化为已有的属性粒，将不满足的重新进行粒化分，更新 D_E^{t+1}，M_{EC}^{t+1}，最终得到更新的关联规则集 R^{t+1}。

4 关联数据聚合信息敏感级别的推演

在研究粒关联推导关系以及高关联度数据对象发现的基础上，通过关联属性敏感级别模糊集可能性测度推演关联对象推导出更高敏感级别信息的可能性。

定义 8(关联属性敏感模糊集) 令 U 为全域空间，AT 为 U 上的属性空间，V 为 AT 的值域；令 $F_s=\{f_{s_1}, f_{s_2}, \cdots, f_{s_k}\}$ 为 AT 上的关联属性敏感模糊集，f_{s_i} 为敏感程度为 s_i 的关联粒，关联粒表示包含高关联度数据对象所属的属性粒，关联属性为关联粒中具有属性关联关系的属性。令 X 为 AT 上取值的变量，与 X 有关的可能性分布为 $\prod x$，$\prod x$ 的可能性分布函数记为 π_x，并在数值上定义为 F_s 的隶属度，即 $\forall c \in AT$，$\pi_x(c)=u_s(c)$。

定义 9 设 F 为敏感程度 $S(f_{s_i})$ 的 AT 上的模糊集，$\prod_X$ 是与变量 X 有关联的可能性分布，而 X 在 AT 中取值，则 F 的可能性测度定义为：

$$Poss(X \text{ is } F) \cong \bigvee_{c\in AT}(F(c) \wedge \pi_X(c))$$

其中，$F(c)$ 为 F 的隶属函数，$\pi_X(c)$ 为与 X 有关的可能性分布函数。

定义 10 关联对象聚合推导出更高敏感程度信息的可能性 P，取决于关联属性集在高于最高敏感级别模糊集上的可能性测度与其权重的乘积之和。

$$P=\sum_{i=1}^{k} w_i(Poss(X \text{ is } F | F\in F_s, S(F) > \max(S(a_1), S(a_2), \cdots))$$

其中，k 为强关联聚合对象集 Q_i 的个数；$Poss()$ 为高于所有关联属性中最高敏感程度的模糊集可能性测度。w_i 为关联粒的权重，定义为：

$$w_i = I_i(Q_i^1; Q_i^2) / \sum_{j=1}^{k} I_j(Q_i^1; Q_i^2)$$

其中，$I(Q_i^1; Q_i^2)$ 为 Q_i^1，Q_i^2 的互信息。互信息定义为：

$$I(Q_i^1; Q_i^2) = H(Q_i^1) - H(Q_i^1 | Q_i^2)$$

其中，$H(P)$ 为定义 P 的信息熵，P 为全域 U 的一个等价划分，信息熵的定义为：

$$H(P) = -\sum_{i=1}^{n} p(cg_i)\log_2 p(cg_i)$$

定义 11 若 $P \geqslant \tau$，则认为关联对象聚合时可能会导致推导出更高敏感级别的信息。其中，τ 指关联对象推导出敏感级别为 S^{max}(关联数据对象的最高敏感级别)的信息可能性阈值，即对满足该条件的数据对象同时进行访问时可能存在越权访问、信息泄露等安全问题，需要限制这些数据的共同访问，并对用户的访问策略进行改进。

依据上述定义，本文提出关联数据聚合信息敏感级别的推演算法，如算法 4 所示。

算法 4 关联对象聚合推导高敏感性信息可能性推演算法

输入：Q_s

输出：$\langle Q_s, S_i, P\rangle$

1. $S^{max} = \max(S(Q))$
2. $j = \min(\{i | f_{s_i} \in F_s, S_i > S^{max}\})$
3. for i=j to $|F_s|$ do
4. P=0
5. for l=1 to $|Q_s|$ do
6. $w_l = I_l(Q_l^1; Q_l^2) / \sum_{j=1}^{k} I_j(Q_l^1; Q_l^2)$
7. compute $Poss_l(X \text{ is } F)$ by $\prod x$
8. $P += w_l * Poss_l(X \text{ is } F)$
9. end for
10. if $P \geqslant \tau_i$ then return $\langle Q_s, S_i, P\rangle$
11. end for

5 实验仿真分析

为了验证本文所提出的基于粒关联的大数据聚合信息级别推演方法的执行效果，对本文提出的方法进行了仿真实验。仿真实验环境为：Python 3.7.16，Intel(R) Core(TM) i7-10875H CPU @2.30 GHz，32.0 GB 内存，系统环境为 Windows 10。本实验的主要目的是探究发现低敏感级别的关联数据之间可能导致用户推导高敏感级别数据造成信息泄露的问题。为满足实验对数据间多种属性关系的要求，采用实验室和互联网中属性特征易划分的科研性数据以及人工合成数据。依据实验数据的特征提取数据属性，并量化数据属性和数据属性值，构造数据关联关系的先验知识库、属性及属性集合的可能性分布以及各个敏感等级上的模糊集。实验评估指标包括算法的执行效率以及算法推演的准确性等。

5.1 关联关系图谱的生成

根据算法 1 对全域数据中存在的属性之间的关联关系进行挖掘，并获取具有属性关联推导关系的属性粒，然后根据算法 2 对属性粒中所包含的数据对象之间的关联度进行挖掘。在此基础上，构建属性粒之间的关联关系图谱以及具有较高关联度的数据对象之间的关联关系图谱。设定最小支持度 $minsup=0.04$、置信度 $minconf=0.5$、关联度 $minassoc=0.12$ 对数据规模为 1 000 的数据进行实验仿真并生成关联关系图谱。属性粒之间的关联关系图谱如图 4 所示。其中的节点为一个属性粒，包含了一个或多个数据对象，图中使用属性集标识属性粒节点，用节点的大小放映其重要度，即支持度。节点之间的有向边表示节点之间的关联关系，用有向边颜色的深浅反映节点之间的联系性，即置信度。具有较高关联度的数据对象之间的关联关系图谱如图 5 所示，因数据量庞大，

故随机抽取了一部分数据进行展示，其中每个节点都代表了一个数据对象，节点之间的连线表示节点之间的关联关系，其颜色的深浅反映数据对象之间的关联度。

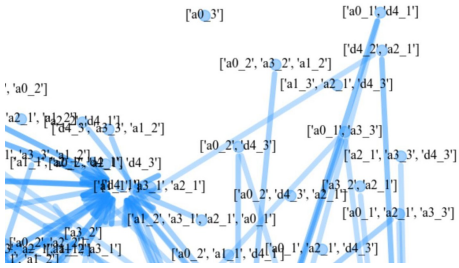

图 4　属性粒之间的关联关系图谱

Fig. 4　Association relationship graph between attribute granular

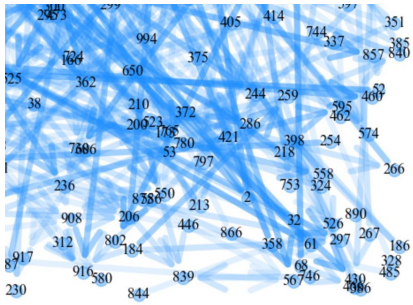

图 5　数据对象之间的关联关系图谱

Fig. 5　Association relationship graph between data objects

5.2　算法性能分析

利用选取的数据信息，从以下两个方面对数据关联关系挖掘算法及高敏感等级信息可能性推演算法共同执行的性能进行评估。

(1)数据规模变化情况下的算法执行效率

针对所选取数据数量的变化，在 $minsup=0.04$、$minconf=0.5$、$minassoc=0.12$ 的情况下对该算法的执行效率进行评估，比较结果如图 6 所示。

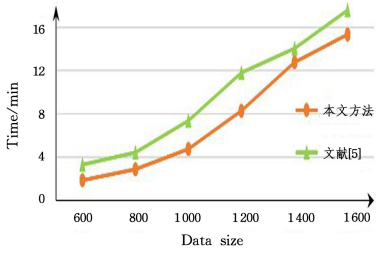

图 6　数据规模变化情况下的算法执行效率

Fig. 6　Execution efficiency of the algorithm with changing data size

从图中可以看出，在数据规模相同的情况下，本文相较于文献[5]所提出的算法效率更高。随着数据规模的递增，数据对象之间的关联复杂度增加，使得算法的执行时间增加。该算法的执行效率呈递增关系，与算法的时间复杂度有关。进行实验测试的数据数量为 600 时，其属性空间大小为 3 000，算法的执行时间为 110.95 s，此时算法的执行速度较为迅速，所挖掘出的规则数量为 227，具有高关联度的关联对象规模为 15 133。数据数量为 1 500 时，其属性空间大小为 7 500，算法的执行时间为 919.09 s，所挖掘的规则数量为 159，关联对象规模为 35 316，此时算法的执行速度较数据数量为 600 时有大幅提升，说明了随着数据规模的增加，得到的关联对象的数量更多，其复杂度越大。该实验下的数据数量的增多，数据属性变得更复杂，导致关联规则的支持度与置信度发生了变化。由于阈值统一设定，因此数据规模的增大反而使得挖掘到的关联规则的数量减少。

(2)阈值变化情况下的算法执行效率

针对最小支持度 $minsup$、置信度 $minconf$ 等参数变化，设定实验数据个数 n 为 1400，关联度 $minassoc=0.12$，以此为条件对算法进行评估，该算法的执行效率如图 7 所示。

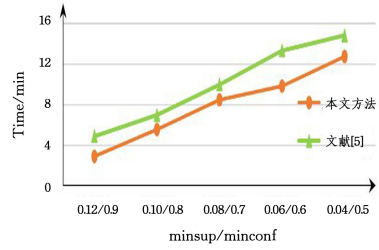

图 7　阈值变化情况下的算法执行效率

Fig. 7　Execution efficiency of the algorithm with changing data size

从实验结果可知，随着 $minsup$ 与 $minconf$ 的递减，算法的执行时间增加，当设定 $minsup=0.12$、$minconf=0.9$ 进行实验测试时，所挖掘出的关联规则数量为 16，具有高关联关系的数据对象关系的规模为 74 311。当设定 $minsup=0.04$，$minconf=0.5$ 进行实验测试时，所挖掘出的关联规则数量为 163，具有高关联关系的数据对象关系的规模为 28 944。阈值设定的降低使得关联规则的数量增多，即属性之间的关联关系增多，将导致数据对象之间的关联关系增多，但挖掘出的关联规则的支持度与置信度的值偏低，造成数据对象之间的关联度随之下降，即具有高关联关系的数据对象的规模变小，但算法的运行复杂度加大，其执行时间也相应增加。

5.3　算法推演准确性分析

为了说明算法的执行效果，对其进行准确性验证，采用的衡量指标为推演算法的准确率(CR)，错误率(WR)及误差率(ER)。设 T 为由本文算法推导出来的存在安全推演问题的实验数据集合，N 为标准情况下存在安全推演问题的数据集合，指标计算公式如下：

$$CR=|T\cap N|/|N|$$

即 T 中包含在 N 中的数据对象占实验数据总数的比例。

$$WR=|T-(T\cap N)|/|N|$$

即推导错误的数据所占的比例，也即 T 中未包含在 N 中的数据对象占实验数据总数的比例。

$$ER=|N-(T\cap N)|/|N|$$

即未推导出的存在安全推演问题的实验数据所占的比例。

在设定敏感级别数据推演的可能性阈值 τ 为 0.8，0.7，0.6，0.5 的情况下分别对规模为 1 000，1 200 和 1 400 的实验数据进行聚合推演更高级别敏感信息可能性的实验仿真，统计分析算法的准确率和错误率，其中设定 $minsup=0.04$，$minconf=0.5$，$minassoc=0.12$，实验结果如图 8、图 9 所示。

由仿真实验结果可知，可能性阈值 τ 越高说明推演过程中忽略了较多的关联对象的数据，因此算法得到的准确率较低，遗漏了较多的敏感数据；τ 越低说明能够推导出更多的会造成安全问题的关联对象，此时算法得到的准确率较高，但容易出现对正常数据的误判，导致错误率提高。由以上分析可

知，合理的设定可能性阈值 τ 是保证推演准确性的关键。该实验仿真得到的数据显示，当 $\tau=0.6$ 时，算法推演的准确率大致维持在 90%以上，算法的错误率不超过 2%，从实际效果来看，能够较好地发现安全问题的存在。

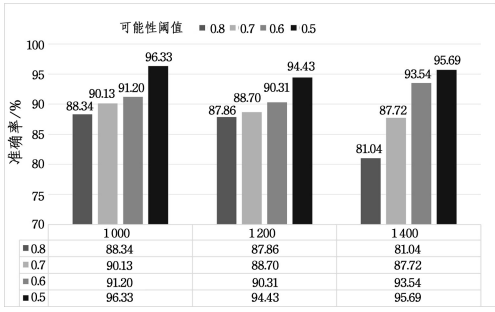

图 8 算法准确率示意图

Fig. 8 Diagram of lgorithm accuracy

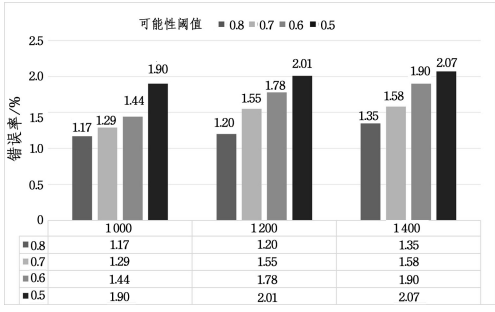

图 9 算法错误率示意图

Fig. 9 Diagram of algorithm error rate

结束语 本文利用基于粒关联的数据聚合信息级别推演方法分析了大数据之间由于逻辑推理而存在的关联性，基于关联属性敏感模糊集、关联粒的可能性测度以及其权重分析关联对象聚合推演出更高敏感等级数据信息的可能性。本研究对于控制多信息系统用户对关联数据的限制访问，、定跨域用户的访问策略，防止信息泄露等安全问题的发现具有重要意义。

参 考 文 献

[1] WENBO Z, XIAOTONG H, ZHENSHAN B. A secure and efficient multi-domain data sharing model on consortium chain[J]. The Journal of Supercomputing, 2022, 79(8): 8538-8582.

[2] SALEHI A S, RUDOLPH C, GROBLERM. A Dynamic Cross-Domain Access Control Model for Collaborative Healthcare Application[C]// 2019 IFIP/IEEE Symposium on Integrated Network and Service Management. 2019: 643-638.

[3] POLTAVTSEV A A, KHABAROV R A, SELYANKINO A. Inference Attacks and Information Security in Databases[J]. Automatic Control and Computer Sciences, 2021, 54(8): 829-833.

[4] POLTAVTSEV A A, KHABAROV R A, SELYANKINO A. Comparative Analysis of Methods for Protection against Logical Inference[J]. Automatic Control and Computer Sciences, 2022, 55(8): 984-990.

[5] CAO L F, CHEN X Y, DU X H, et al. A Level Inference Method for Aggregated Information of Objects Based on Associated Attributes[J]. ACTA Electronica Sinca, 2013, 41(7): 1442-1447.

[6] CAO L F, CHEN X Y, DU X H, et al. A Level Inference Method for Aggregated Information of Objects Based on Clustering Analysis[J]. Journal of Electronics & Information Technology, 2012, 34(6): 1432-1437.

[7] LIU T, WANG Z J, LIU Y, et al. Data inference: data leakage paradigms and defense methods in cyber-physical systems[J]. Scientia SincaInformationis, 2023, 53(11): 2152-2179.

[8] HAO L, WANG T, GUO C. Research on parallel association rule mining of big data based on an improved K-means clustering algorithm [J]. International Journal of Autonomous and Adaptive Communications Systems, 2023, 16(3): 233-247.

[9] LIU X, ZHANG Z, ZHANG G. Using improved feature extraction combined with RF-KNN classifier to predict coal and gas outburst[J]. Journal of Intelligent & Fuzzy Systems, 2023, 44(1): 237-250.

[10] LU B, FAN Q, ZHOU X L, et al. A multimodal multi-label classification method based on hypergraph[J]. Computer Engineering& Science, 2024, 46(9): 1667-1674.

[11] GAO Y, ZHANG Z, LIN H, et al. Hypergraph learning: Methods and practices[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(5): 2548-2566.

[12] WANG Y, ZHAO E, WANG W. A Knowledge Graph Completion Method Based on Fusing Association Information[J]. IEEE Access, 2022, 10: 50500-50507.

[13] HU X C, LI Y, CHEN Z J, et al. Review of research of granular fuzzy rule-based modeling[J]. CAAI Transactions on Intelligent Systems, 2024, 19(1): 22-35.

[14] LIN A J, ZHANG M T. Differential privacy protection based on multi-granularity data[J]. Journal of Soochow University (Philosophy and Social Sciences), 2024, 45(2): 182-192.

[15] CAO L, LU X, GAO Z, et al. An Information Sensitivity Inference Method for Big Data Aggregation Based on Granular Analysis[C]// 2019 Big Data. 2019.

LI Jinhui, born in 2000, postgraduate. His main research interests include information security and access control.

CAO Lifeng, born in 1981, Ph.D, professor, Ph.D supervisor. His main research interests include information security and blockchain.