文章编号：2095-4980(2025)05-0482-07

数据显隐性关系驱动的敏感数据泄露风险预测

梁　花[a]，靳　敏[b]，严　华[a]，韩世海[a]，李　玮[b]

(国网重庆市电力公司 a.电力科学研究院，重庆 401123；b.数字化部，重庆 400014)

摘　要：随着物联网(IoT)、大数据以及人工智能(AI)技术的快速发展，海量数据正在以前所未有的规模被生成和利用。这些数据中包含大量敏感信息，如何安全存储敏感数据成为亟待解决的现实问题。现有的数据存储方案通常侧重于敏感数据的直接保护，忽视了敏感数据与非敏感数据之间显性和隐性关联所带来的泄露风险。为此，本文从信息熵的角度深入分析了数据间的显性和隐性关系，提出一种快速评估显隐性关系并预测敏感数据泄露风险的方法。通过引入信息提升比(LR)和信息掌握概率(PIC)，能够有效识别非敏感数据对敏感数据泄露风险的影响。仿真实验中，统计特性数据集(SPD)中的单属性LR最大为0.308，联合属性LR可提升至0.891；敏感数据泄露风险的检测概率显著提高，最高达到23.2%。仿真结果表明，该方法能够有效识别并应对因显隐性关系带来的安全风险，显著提升敏感数据存储的整体安全水平。

关键词：信息熵；显隐性关系；风险评估；数据关联性；安全策略

中图分类号：TP309.2　　**文献标志码：**A　　**DOI：**10.11805/TKYDA2024383

Sensitive data leakage risk prediction driven by data explicit and implicit relationships

LIANG Hua[a]，JIN Min[b]，YAN Hua[a]，HAN Shihai[a]，LI Wei[b]

(a.Electric Power Research Institute, Chongqing 401123, China；b.Digitization Department，Chongqing 400014, China, State Grid Chongqing Electric Power Company)

Abstract：With the rapid development of Internet of Things(IoT), big data, and Artificial Intelligence (AI) technologies, massive amounts of data are being generated and utilized on an unprecedented scale. These data contain a large amount of sensitive information, and how to securely store sensitive data has become a realistic problem that needs to be solved. The existing data storage schemes usually focus on the direct protection of sensitive data, while ignoring the leakage risks associated with explicit and implicit associations between sensitive and non-sensitive data. The explicit and implicit relationships among data are deeply analyzed from the perspective of information entropy, and a method is proposed to quickly assess the explicit and implicit relationships and predict the leakage risk of sensitive data. By introducing the information Lift Ratio(LR) and the Probability of Information Control(PIC), the method can effectively identify the influence of non-sensitive data on the risk of sensitive data leakage. In the simulation experiments, the maximum single-attribute LR in the Statistical Property Dataset(SPD) is 0.308, and the joint-attribute LR can be up to 0.891, and the predicted value of the sensitive data leakage risk is significantly improved, up to 23.2%. The simulation results show that the method can effectively identify and cope with the security risks caused by explicit and implicit relationships, thus significantly improving the overall security level of sensitive data storage.

收稿日期：2024-08-16；**修回日期：**2024-10-17

基金项目：国网重庆市电力公司重点研发项目(2023渝电科技59号)

引用格式：梁花,靳敏,严华,等. 数据显隐性关系驱动的敏感数据泄露风险预测[J]. 太赫兹科学与电子信息学报, 2025,23(5):482-488. DOI:10.11805/TKYDA2024383.

Citation format： LIANG Hua,JIN Min,YAN Hua,et al. Sensitive data leakage risk prediction driven by data explicit and implicit relationships[J]. Journal of Terahertz Science and Electronic Information Technology, 2025,23(5):482-488. DOI:10.11805/TKYDA2024383.

Keywords: information entropy; explicit and implicit relationship; risk assessment; data correlation; security policy

随着信息技术的飞速发展，尤其是物联网(IoT)和人工智能(AI)的广泛应用，数据不仅成为推动社会进步和经济发展的核心驱动力，其安全性的重要性也愈发凸显[1-3]。物联网设备的大规模部署产生了海量实时数据，这些数据跨越了物理世界与数字世界的界限，极大丰富了数据的维度和复杂度。而AI技术的引入，使得从这些数据中挖掘出有价值的信息成为可能，但同时也对数据的安全存储和隐私保护提出了前所未有的挑战。在物联网环境中，设备间的互联互通使敏感数据的暴露面大大增加。传感器、智能家居、工业控制系统等物联网设备直接收集并传输用户的个人信息、环境数据、设备状态等敏感信息[4]，一旦这些设备或传输通道被黑客攻击或不当利用，敏感数据就可能会轻易泄露[5]。AI技术的运用，能够提升数据处理的效率和准确性，但也加剧了数据泄露的风险。AI算法需要大量数据进行训练和优化，这增加了数据被集中处理和存储的需求，但同时也为潜在的攻击者提供了更多的攻击路径。此外，AI的自动化决策能力也可能在不加注意的情况下，基于敏感数据做出不当的决策，进而引发数据泄露或滥用。

传统的敏感数据存储方案主要关注于敏感数据本身，依赖数据加密、访问控制、防火墙等技术手段，这些措施在一定程度上提高了敏感数据的安全性。但随着数据量的爆炸性增长和数据关系的复杂化，敏感数据与非敏感数据之间的显性和隐性关联日益成为安全防护的薄弱环节[6-8]。显性关系，如数据库中的表结构、字段间的直接依赖等，虽然易于识别和管理，但往往被忽视其潜在的安全风险。而隐性关系，如多字段间的依赖关系等，则更加隐蔽和难以察觉，更可能成为敏感数据泄露的重要渠道。因此，如何有效评估敏感数据与非敏感数据之间的显性和隐性关系，预测并防范由此带来的数据泄露风险，成为当前敏感数据存储安全领域亟待解决的重要问题。信息熵作为信息论中的核心概念，为量化信息的不确定性提供了有力的工具。通过信息熵的视角，深入分析数据间的信息关联，揭示那些不易被察觉的隐性关系，从而为敏感数据的安全存储提供新的思路和方法[9-10]。本文从信息熵的角度深入分析了数据间的显性和隐性关系，提出一种快速评估显隐性关系并预测敏感数据泄露风险的方法。

1　相关理论及基础概念

1.1 数据显隐性关系

数据的显隐性关系，在数据科学、计算机科学及相关领域中，并没有一个直接对应的历史由来概念，因为它更多地是作为一个分析框架或工具来理解和解释数据之间的关系。本文对其做出以下定义：

1) 显性关系

显性关系是指那些可以直接观察、测量或通过常识判断得出的关系，这些关系往往不需要复杂的模型或推理过程就能被明确识别。而在更高级的数据分析中，显性关系可能涉及数据之间的直接相关性或因果关系，这些关系可通过统计方法或数据挖掘技术直接揭示，如，邮政编码与住居区域存在直接对应的显性关系，同一区域传感器测得的日照强度与地面温度间具有高度相关的显性关系。

2) 隐性关系

隐性关系则是指那些不直接显露在数据表面，需通过复杂的模型、推理或数据挖掘技术才能揭示的关系。这些关系可能隐藏在数据的深层次结构中，或需要通过多源数据的整合和关联分析才能发现。该属性往往揭示了数据之间的内在联系和潜在规律，对于理解数据的本质和预测未来的趋势具有重要意义，如，用户的收入水平与用户居住区域、就诊医院等信息间存在一定的隐性联系；用户的日常行为模式(如作息规律、用餐时间等)及环境数据(如室内温湿度、光照强度等)可以揭示出用户的社会层次及生活质量的要求等。

1.2 信息熵与互信息量

信息熵是信息论中的一个核心概念，用于度量信息的不确定性或随机性。互信息量是衡量2个随机变量之间共享信息量的一个度量，用来表示当一个随机变量的值已知时，另一个随机变量不确定性的减少量。在通信系统中，信息熵可用来衡量传输信息的效率，而互信息量则可用来评估接收端从发送端接收到的信息中实际获取了多少有用的信息。将信息熵和互信息量关联起来，有助于更全面地理解信息论中的基本概念和它们在实际应用中的作用[11]。

1.2.1 信息熵

信息熵用于描述信源输出符号的不确定度，即信源所含信息量的多少。信息熵越大，说明信源发出的信息

越难以预测，即信息的随机性越大。

假设一个离散随机变量其值域为$\{x_1,x_2,\cdots,x_n\}$，且每个取值对应率为$P(X=x_i)=p_i$，其中$i=1,2,\cdots,n$，且满足$\sum_{i=1}^{n}p_i=1$。则X的信息熵$H(X)$定义为：

$$H(X)=-\sum_{i=1}^{n}p_i\log_2(p_i) \tag{1}$$

若X是数据集$\mathbb{Z}$中的某一则记录某一属性的取值，其值域仍为$\{x_1,x_2,\cdots,x_n\}$。当数据集$\mathbb{Z}$中记录数量N足够大时，可将属性值x_i出现的频率f_i近似地视为随机变量X取值为x_i的概率p_i。使用$N(x_i)$表示属性值x_i在数据集中出现的次数，则有：

$$p_i\approx f_i=\frac{N(x_i)}{N} \tag{2}$$

1.2.2 互信息量

互信息量用于衡量2个随机变量之间的信息共享或相互依赖程度。互信息量可看作是一个随机变量已知，另一个随机变量减少的不肯定性，或是从一个随机变量中可获得的关于另一个随机变量的平均信息量。其表达式为：

$$I(X;Y)=\sum_i\sum_j P(x_i,y_j)\log_2\frac{P(x_i,y_j)}{P(x_i)P(y_j)} \tag{3}$$

式中：X和Y为2个随机变量；$P(x_i,y_j)=P(X=x_i\wedge Y=y_i)$。

若X和Y是数据集$\mathbb{Z}$中的某一则记录中某两属性的取值，当数据集$\mathbb{Z}$中记录数量N足够大时，将属性值组合(x_i,y_j)出现的频率$f_{i,j}$近似地视为$P(x_i,y_j)$，并使用$N(x_i,y_i)$表示属性值组合(x_i,y_j)在数据集中出现的次数，则有：

$$P(x_i,y_j)\approx f_{i,j}=\frac{N(x_i,y_j)}{N} \tag{4}$$

2 敏感数据泄露风险预测

随着社会对隐私保护意识的不断增强，确保敏感数据本身的安全性已成为各方关注的焦点。但基于数据的显隐性关系，即通过非敏感数据推导出敏感数据的风险仍存在。这种风险在大数据分析和机器学习模型中尤为突出，因为这些技术可通过分析大量非敏感数据，推断出与敏感信息相关的模式或关系。

2.1 显隐性关系分析

为量化非敏感属性对敏感属性信息贡献的程度，提出信息提升比(LR)指标，该指标体现了在给定非敏感属性信息后，敏感属性不确定性的降低比值。此度量标准基于条件熵和信息增益的基本原理进行构建，能客观地评估非敏感属性对于揭示敏感属性信息的作用。

假设一个数据集，其包含敏感属性S和非敏感属性集$A=\{A_1,A_2,\cdots,A_d\}$。定义非敏感属性对敏感属性的LR如下：

1) 单个非敏感属性的LR：

$$LR(A_i\rightarrow S)=\frac{I(A_i;S)}{H(S)} \tag{5}$$

式中：$H(S)$为敏感属性S的信息熵；$I(A_i;S)$为非敏感属性A_i与敏感属性S的互信息量。

2) 非敏感属性组合的LR：

当考虑多个非敏感属性的组合时，LR可扩展为：

$$LR(A\rightarrow S)=\frac{I(A_1\&A_2\&\cdots\&A_n;S)}{H(S)} \tag{6}$$

式中$I(A_1\&A_2\&\cdots\&A_n;S)$为联合所有非敏感属性$A_1,A_2,\cdots,A_d$提升对敏感属性$S$的互信息量。

多个属性的组合通常会增加LR，因为不同的属性可能携带互补的信息。当单独考虑某个非敏感属性时，它提供的信息量可能有限，但当多个非敏感属性联合在一起时，它们可能会揭示更多关于敏感属性的细节。属性之间的组合能够捕捉到单个属性无法体现的复杂关系，因此组合属性对敏感属性的互信息量通常更大。通过将多个属性共同考虑，系统可更有效地推断或预测敏感属性，从而提高LR，增加对敏感信息的洞察力。

LR 越大，泄露的可能性就越高。这是因为较高的 LR 意味着非敏感属性(或属性组合)与敏感属性之间有更强的关联性。当攻击者获取到这些非敏感属性时，可以通过分析这些属性来推测或预测敏感属性的值，从而增加敏感数据泄露的风险。

2.2 信息掌握概率

从概率论视角，攻击者获得特定属性(如敏感数据等)的概率，即信息掌握概率(PIC)。这是一个复杂但至关重要的过程，直接关联到风险评估领域。通过人工评分进行数值化衡量，是这一过程中一个既直观又实用的方法，它允许专家基于经验、专业知识以及对系统环境的深入理解，为每个样本的特定属性分配一个分数或等级，对潜在威胁进行量化分析。具体的专家评分如式(7)所示：

$$Poi=\sum_{i=1}^{m}(w_i\times S_i)+f(other) \tag{7}$$

式中：m 为参与评分的专家人数；w_i 为第 i 个专家的评分权重，该权重反映了专家在整体评分中的权威性和重要性层次；S_i 为第 i 个专家对评分对象给出的打分结果；$f(other)$ 用于综合考虑和量化一些难以衡量的因素对总评分的影响，如评分属性的整体安全性印象、与已知威胁模式的匹配度等，该函数的具体实现会根据不同的评分场景和专家团队的特点而有所调整。

在获取了单个属性被窃取风险的总评分后，进一步利用式(8)将总评分转换为相应的概率值，以便直观地表示该属性被窃取的可能性。

$$P(x)=\frac{Poi_x}{Poi_{\max}} \tag{8}$$

式中：Poi_x 为属性 x 的人工评分；$Poi_{\max}$ 为评分体系中设定的最高评分。

在通过人工评分确定每个属性的 PIC 后，需要深入研究这些属性间的联合 PIC。需要明确的是，多个属性的联合 PIC 受单个属性 PIC 的约束。具体而言，任何一组属性的联合 PIC 均不会超出它们各自 PIC 的最小值，且在实际情况中，由于属性间可能存在的相互关联，这一理论的上界往往难以实现。另一方面，联合 PIC 的下界则直接关联于各属性 PIC 的乘积。因此多个属性的联合 PIC 将在由乘积定义的下界和受属性间关系影响的上界之间取值。

$$P(x)\times P(y)\leqslant P(x\&y)\leqslant \min\left(P(x),P(y)\right) \tag{9}$$

为了近似估算多个属性的联合 PIC，一种有效的方法是在其理论上的上界和下界之间，通过随机采样的方式生成大量联合 PIC 样本，然后对这些样本取平均值。对于涉及 3 个或更多属性的复杂场景，上述方法同样适用。关键在于深入理解属性间的相互依赖关系，以及在多维概率空间中实施有效的随机采样策略。

2.3 风险预测

在数据风险值计算过程中，通过构建模型和算法，科学地计算出数据泄露风险的量化值。该风险值不仅反映当前信息系统的安全状况，还可以为后续的安全策略制定、资源调配及应急响应提供重要的参考依据。通过设定风险阈值，系统能够自动判断当前风险水平是否超出可接受范围，这有助于及时发现并应对潜在的数据泄露风险，保护敏感信息的安全与完整。

假设要计算 A_1、A_2、A_3 3 个属性以及它们与隐私属性(称之为 S)之间的关系。这些属性可能是用户数据中的不同部分，如工作单位(A_1)、家庭地址(A_2)、个人职业(A_3)等，而隐私属性 S 可能是用户的银行账户信息或健康记录等高度敏感的信息。对于每个属性，需要评估攻击者通过非法手段(如黑客攻击、数据泄露、社会工程学等)获得它们的概率。这些概率基于人工评分方法获得，并表示为 $P(A_1)$、$P(A_2)$、$P(A_3)$。

现实中属性之间往往存在依赖关系，如知道工作单位(A_1)可能会增加获取隐私属性 S 的概率，因此，需要分析这些依赖关系，并使用具体的风险值表示。本文将获取属性 A_1 时获得隐私属性 S 的风险值定义为 $Risk(S|A_1)$：

$$Risk(S|A_1)=LR(A_1\rightarrow S)\times P(A_1) \tag{10}$$

当攻击者获得的属性增多时，需对式(10)进行扩展。攻击者可能获得属性组合($A_1\&A_2\&A_3$)时，获得隐私属性 S 的风险值为：

$$Risk(S|A_1,A_2,A_3)=LR(A_1\&A_2\&A_3\rightarrow S)\times P(A_1\&A_2\&A_3) \tag{11}$$

式(11)存在一个明显的矛盾：当攻击者获取的属性增多时，其 LR 是增加的，但 PIC 却是下降的。这样有可能

导致获取的属性越多，信息泄露的风险越小，这显然是不符合常理的。对式(11)改进：

$$Risk(S|C)=\max_{C_i\subseteq C}\left[LR(C_i\rightarrow S)\times P(C_i)\right] \tag{12}$$

式中：C为攻击者可能获取的属性组合；C_i为C的子集。

通过分析A_1、A_2、A_3等属性对S的LR，可反映各属性在揭示S信息方面的贡献。随后，综合考虑这些属性的联合效应，评估它们共同作用下获取S概率的增量影响。这一过程不仅考虑了单个属性的独立作用，还深入探究了属性间的相互增强或削弱效果，如某些属性的存在可能降低其他属性对S预测的准确性，反之亦然。

为了判断是否存在隐私泄露风险，还需设定一个门限值(θ)。这个门限值可根据组织的安全政策、行业规范、法律法规等因素确定。如果计算出的$Risk(S|C)$大于这个门限值，则认为存在显著的敏感泄露风险；反之，则认为风险较低或可忽略。

3 实验仿真

通过Matlab这一数据处理与仿真工具，探索并分析属性LR和属性风险泄露的概率。通过计算并比较不同属性对敏感属性的LR，旨在识别出对风险预测最有影响力的特征，从而优化预测模型的性能。实验采用2个开源数据集：SPD数据集[12](可在https://crises-deim.urv.cat/opendata/获取)和Adult数据集[13](可在https://archive.ics.uci.edu/ml/datasets/adult获取)。SPD数据集详细记录了患者的年龄、性别、民族、种族、邮政编码、所在城市、住院时长、入院季节和总费用，采用用数字1~10进行表示。Adult数据集包含15个属性，分别为年龄、工作类别、受教育程度、受教育时间、婚姻状况、职业、社会角色、种族、性别、资本收益、资本支出、每周工作时间、国籍以及收入，也采用数字1~15进行表示。在SPD数据集中，总费用(10)作为敏感属性，其余作为非敏感属性；Adult数据集中使用收入(15)作为敏感属性，其余作为非敏感属性。

SPD数据集单属性LR和风险的获得如表1所示，从表中可以发现，医院(1)、年龄(2)和邮政编码(6)对获取敏感属性(10)具有较大的LR，即这些属性对于敏感属性的获得具有较大的提升。LR与PIC的乘积即可得到敏感属性风险泄露的概率，表1中的属性风险泄露概率可以衡量非敏感属性对敏感属性的泄露风险。

表1 单属性LR和泄露风险(SPD数据集)

Table1 Information LR and leakage risk of single attribute (SPD dataset)

item	1	2	3	4	5	6	7	8	9
LR	0.243	0.112	0.005	0.007	0.013	0.308	0.07	0.073	0.012
PIC	0.89	0.79	0.98	0.63	0.62	0.76	0.85	0.56	0.70
risk	0.216	0.088	0.005	0.005	0.008	0.232	0.059	0.041	0.008

Adult数据集的结果如表2所示，同样展示了非敏感属性对敏感属性的LR、单个属性可能获取的概率和该非敏感属性对敏感属性的泄露风险概率。

表2 单属性LR和泄露风险(Adult数据集)

Table1 Information LR and leakage risk of single attribute (Adult dataset)

item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
LR	0.123	0.028	0.503	0.116	0.116	0.198	0.116	0.208	0.010	0.046	0.141	0.058	0.075	0.010
PIC	0.63	0.41	0.54	0.47	0.63	0.47	0.67	0.74	0.41	0.78	0.28	0.34	0.58	0.63
risk	0.078	0.012	0.272	0.055	0.073	0.093	0.078	0.154	0.004	0.036	0.040	0.020	0.044	0.006

为表现出实验的多样性和代表性，通过设计单个非敏感属性与敏感属性组合和多个非敏感属性与敏感属性组合的方式进行实验。

通过SPD数据集进行多属性联合的方式计算LR和风险泄露概率。属性组合为：医院(1)+年龄(2)、医院(1)+住院时间(8)、医院(1)+年龄(2)+住院时间(8)、医院(1)+年龄(2)+邮政编码(6)+住院时间(8)，不同组合对应的LR和风险泄露概率如图1所示。图中结果表明，随着非敏感属性数量的增多，非敏感属性组合对敏感属性的LR越大，敏感泄露概率增大或保持不变(与非敏感属性获取概率有关)。当攻击者获取到全部4个非敏感属性时，LR高达0.89，意味着该情况下攻击者获得敏感属性的几率非常高。

通过Adult数据集进行多属性联合的方式计算LR和风险泄露概率。属性组合包括年龄(1)+受教育程度(4)、年龄(1)+关系(8)、婚姻状况(6)+关系(8)、年龄(1)+婚姻状况(6)+关系(8)+资本收益(11)，不同组合对应的LR和风险泄露概率如图2所示。图中结果呈现规律与图1基本相似，但攻击者获取到全部4个非敏感属性时，LR仅为0.39，说明Adult数据集对敏感属性的保护程度高于SPD数据集。

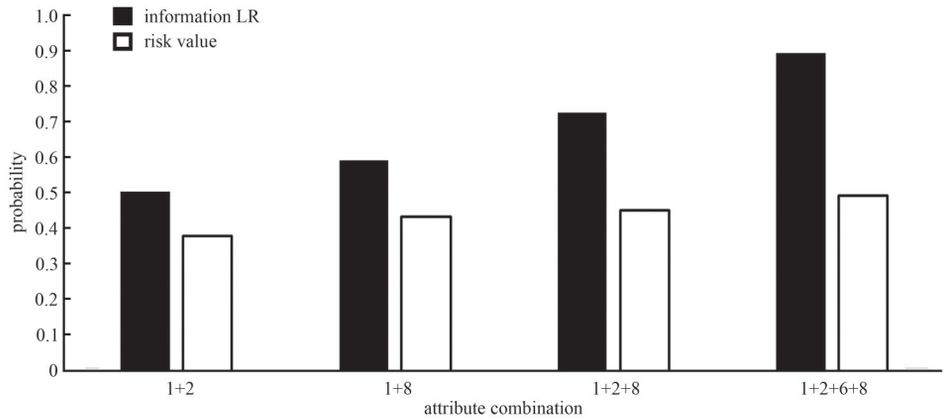

Fig.1 Multi-attribute joint calculation of information LR and risk leakage probability(SPD dataset)
图1 多属性联合计算信息提升比和风险泄露概率(SPD数据集)

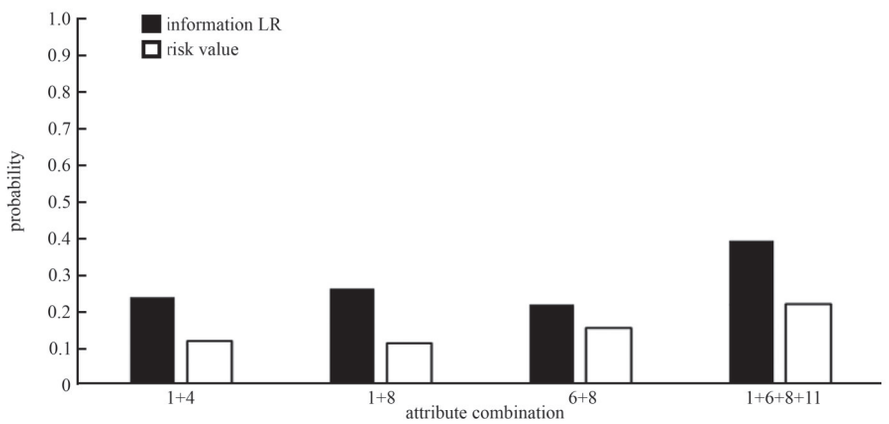

Fig.2 Multi-attribute joint calculation of information LR and risk leakage probability(Adult dataset)
图2 多属性联合计算信息提升比和风险泄露概率(Adult数据集)

本文对属性进行抑制和泛化，进行单个属性和多属性风险泄露分析。抑制是指将敏感信息的某些部分进行隐藏或遮蔽，如，通过删除或模糊化数据的最后几位，降低数据的可识别性。泛化则是将具体值替换为一个范围或类别，如，将年龄以10岁或20岁为间隔归类为“20~30岁”或“30~40岁”，以减少信息的详细程度。两者结合使用，可有效降低敏感属性的泄露风险，提升数据的安全性。以SPD数据集为例，对其中的属性医院(1)、年龄(2)、邮政编码(6)和住院时间(8)进行抑制或泛化处理。方案一：对医院(1)最后2位编号进行抑制，年龄(2)以10岁间隔进行泛化，对邮政编码(6)最后2位编码进行抑制；方案二：对医院(1)最后3位编号进行抑制，年龄(2)以20岁间隔进行泛化，对邮政编码(6)最后3位编码进行抑制。将抑制和泛化后的方案与之前的方案进行对比试验，结果如表3所示。实验表明，通过抑制和泛化技术可有效降低对敏感属性的LR，进而对敏感属性进行保护。

表3 处理前后多属性联合LR(SPD数据集)

Table3 Information LR of multi-attribute before and after processing(SPD dataset)

item	1+2	1+8	1+2+8	1+2+6+8
original	0.500	0.588	0.722	0.891
plan1	0.268	0.104	0.116	0.119
plan2	0.163	0.074	0.084	0.081

4　结论

本文聚焦于敏感数据存储的安全问题，指出当前方案常忽略敏感数据与非敏感数据间的显性和隐性信息关联，这对数据安全构成新挑战。为应对此挑战，本文从信息熵的角度出发，深入分析了数据间的显性和隐性关系，并提出了一种创新的评估方法。该方法能快速评估数据间的显隐性关系，并有效预测敏感数据的泄露风险。

仿真实验的结果显示，该评估方法不仅能有效识别由显隐性关系带来的安全风险，还能为提升敏感数据存储的安全水平提供支撑，为解决当前信息社会中的敏感数据安全问题提供了新的思路和有效途径。

参考文献：

[1] 尚浩,朱恒华,李双,等. 一种基于图神经网络的地质钻孔数据保护方案[J]. 地球科学, 2023,48(8):3151-3161. (SHANG Hao, ZHU Henghua,LI Shuang,et al. A geological borehole data protection based on graph neural networks[J]. Earth Science, 2023,48(8):3151-3161.) DOI:10.3799/dqkx.2021.232.

[2] 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. (FANG Chaojian, HU Xinrong. Privacy-sensitive data filtering algorithm based on fuzzy approximation[J]. Journal of Jilin University (Engineering and Technology Edition), 2023,53(4):1174-1180.) DOI:10.13229/j.cnki.jdxbgxb.20220081.

[3] 臧国全,张盼盼,柴文科,等. 个人通信数据的敏感性识别与隐私计量研究[J]. 图书情报知识, 2024,41(2):110-120. (ZANG Guoquan, ZHANG Panpan, CHAI Wenke, et al. Sensitivity identification and privacy measurement of personal communication data[J]. Document, Informaiton & Knowledge, 2024,41(2):110-120.) DOI:10.13366/j.dik.2024.02.110.

[4] 王辉,陈宇,申自浩,等. 结合对比监督和排序树的轨迹数据差分隐私保护方案[J]. 计算机工程与科学, 2023,45(10):1797. (WANG Hui, CHEN Yu, SHEN Zihao, et al. Differential privacy protection scheme for trajectory data combining contrastive supervision and sorting tree[J]. Computer Engineering and Science, 2023,45(10):1797.) DOI:10.3969/j.issn.1007-130X.2023.10.009.

[5] 张志强,王海宝,周文涛,等. 基于身份认证的智能电网安全防护技术[J]. 太赫兹科学与电子信息学报, 2021,19(2):330-333. (ZHANG Zhiqiang,WANG Haibao,ZHOU Wentao,et al. Safety protection technology of smart grid based on identity[J]. Journal of Terahertz Science and Electronic Information Technology, 2021,19(2):330-333.) DOI:10.11805/TKYDA2019496.

[6] 魏方,张莹,辛昊儒,等. "显性-隐性"众包地理信息在多尺度景观感知研究中的应用[J]. 西部人居环境学刊, 2023,38(1):124-132. (WEI Fang, ZHANG Ying, XIN Haoru, et al. The application of crowdsourcing technology in landscape perception research[J]. Journal of Human Settlements in West China, 2023, 38(1): 124-132.) DOI: 10.13791/j. cnki. hsfwest.20230118.

[7] SATHISH KUMAR G, PREMALATHA K, UMA MAHESHWARI G, et al. No more privacy concern: a privacy-chain based homomorphic encryption scheme and statistical method for privacy preservation of user's private and sensitive data[J]. Expert Systems with Applications, 2023(234):121071. DOI:10.1016/j.eswa.2023.121071.

[8] REN Yongjun,HUANG Ding,WANG Wenhai,et al. BSMD:a blockchain-based secure storage mechanism for big spatio-temporal data[J]. Future Generation Computer Systems, 2023(138):328-338. DOI:10.1016/j.future.2022.09.008.

[9] WANG Mengzhu, LIU Junze, LUO Ge, et al. Smooth-guided implicit data augmentation for domain generalization[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 36(3): 4984-4995. DOI: 10.1109/TNNLS.2024.3377439.

[10] YANG Zhigang, CAO Xia, WANG Honggang, et al. VRIL: a tuple frequency-based identity privacy protection framework for metaverse[J]. IEEE Journal on Selected Areas in Communications, 2024,42(4):933-947. DOI:10.1109/JSAC.2023.3345425.

[11] PAN Meihong, XIN Hongyi, SHEN Hongbin. Semantic-based implicit feature transform for few-shot classification[J]. International Journal of Computer Vision, 2024,132(11):5014-5029. DOI:10.1007/s11263-024-02113-8.

[12] SÁNCHEZ D,MARTÍNEZ S,DOMINGO-FERRER J. Comment on "unique in the shopping mall:on the reidentifiability of credit card metadata"[J]. Science, 2016,351(6279):1274. DOI:10.1126/science.aad9295.

[13] BECKER B,KOHAVI R. UCI machine learning repository[Z]. 1996.

作者简介：

梁　花(1990-)，女，硕士，工程师，主要研究方向为数据安全防护体系.email:stategridcqlianghua@outlook.com.

靳　敏(1989-)，女，硕士，工程师，主要研究方向为网络安全攻防技术与体系建设.

严　华(1986-)，男，学士，工程师，主要研究方向为数据安全、存储介质安全.

韩世海(1975-)，男，学士，高级工程师，主要研究方向为网络安全攻击技术.

李　玮(1989-)，男，学士，工程师，主要研究方向为数据安全救援与销毁技术.