基于信息熵的数据集重标识风险评估方法*

陈 磊[1,2],薛见新[1,2],张润滋[1,2],刘文懋[1]

(1.绿盟科技集团股份有限公司,北京 100089;2.清华大学 自动化系,北京 100084)

摘 要:去标识化作为一种隐私保护技术,在数据发布领域得到了广泛应用。然而,在大数据时代下,攻击者可能获得了更多的关联数据,去标识数据集仍然存在重标识攻击的风险。基于 Shannon 信息熵,并结合信息安全风险评估框架,提出了一种综合的重标识风险评估方法。首先,将攻击者可能利用的数据集的各种属性组合归纳为若干个脆弱性,然后逐一对这些脆弱性从可能性和危害性两个维度进行评估。最后,为了综合评估整个数据集的重标识风险,构造了一种基于熵值增量和加权的评估算法。实验结果表明,所提评估方法可全面、直观地反映风险分布与趋势。

关键词:隐私保护;去标识数据集;重标识风险评估;信息熵

中图分类号:TP399 **文献标识码**:A **DOI**:10.19358/j.issn.2096-5133.2020.12.001

引用格式:陈磊,薛见新,张润滋,等. 基于信息熵的数据集重标识风险评估方法[J].信息技术与网络安全,2020,39(12):1-7.

Re-identification risk assessment of de-identified datasets based on information entropy

Chen Lei[1,2], Xue Jianxin[1,2], Zhang Runzi[1,2], Liu Wenmao[1]

(1.Nsfocus Information Technology Co., Ltd., Beijing 100089, China;
2.Department of Automation, Tsinghua University, Beijing 100084, China)

Abstract: As a privacy protection technology, de-identification has been widely used in data publishing scenarios. However, in the era of big data, attackers may obtain more associated data, and there is still a risk of re-identification attacks on de-identified datasets. Based on information entropy and information security risk assessment framework, this paper proposes a comprehensive re-identification risk assessment method. Firstly, the various attribute combinations of a de-identified dataset that attackers may utilize are summarized into several vulnerabilities, and then these vulnerabilities are evaluated one by one from probability and impact dimension. Finally, in order to comprehensively evaluate the re-identification risk of the dataset, this paper constructs a fast evaluation algorithm based on entropy increments and weights. Extensive experimental results demonstrate that the proposed evaluation method can comprehensively and intuitively reflect the risk distribution and trend.

Key words: privacy protection; de-identified datasets; re-identification risk assessment; information entropy

0 引言

在大数据时代下,数据共享、发布和交易等场景需求变得越来越多,一方面促进了数据流通与价值利用,另一方面引发的个人数据与隐私安全事件近年来呈现爆发趋势[1]。

为了应对挑战,在法规层面,全球掀起了数据隐私的立法热潮,如欧盟《通用数据保护条例》(GDPR)、美国《加州消费者隐私法案》(CCPA)等。我国 2017 年实施的《网络安全法》,其中一个章节专门明确个人信息安全;此外,我国《个人信息保护法》在加快立法与制定中。在技术层面,如何平衡数据利用与隐私保护问题,已经成为学术界和工业界的一大研究热点[2]。当前,已经发展出了保留格式加密(Format-Preserving Encryption, FPE)[3]、差分隐私(Differential

* 基金项目:中国博士后科学基金资助项目(2019M660511, 2020M670181)

Privacy, DP)[4]、K-匿名(K-Anonymity)[5]和 L-多样性(L-Diversity)[6]以及去标识化(De-identification)[7]等技术。其中,去标识化技术通过对原始个人信息进行部分屏蔽、泛化和失真等数据变换操作,是一种意图消除"个人身份"的隐私保护技术。由于其处理规则简单灵活且易于并行处理(高效),目前在隐私保护的数据发布和数据挖掘等实际场景中有广泛应用与部署。通常,在工业界习惯称为"数据脱敏"。

然而研究发现,去标识化并不能达到理想的"完全消除个人身份"的状态,例如欧盟 29 条工作组在一份研究报告指出,去标识化是缓解/降低重标识风险(mitigating the risks)的有效手段,但不能完全消除残余风险(Residual Risk)[8]。此外,随着数据资源的公开或个人数据库的泄露(攻击者可通过各种渠道获取更多的攻击资源),一些实际的攻击案例证明经过去标识化处理后得到的数据集仍然存在重标识攻击(Re-identification Attack)的风险与可能性。一个经典的案例:美国马萨诸塞州发布了医疗患者信息数据库,去掉患者的姓名和地址信息,仅保留患者的 ZIP、Birthday、Sex 和 Diagnosis 等信息。研究人员发现,存在一个可访问的公开数据库是州选民的登记表,它包括选民的 ZIP、Birthday、Sex、Name 以及 Address 等详细个人信息。将这两个数据库的同属性段{ZIP, Birthday, Sex}进行关联与链接操作,如图1所示,可恢复出大部分选民的医疗健康信息,从而造成一起严重的医疗隐私泄露事故[5]。因此衡量和评估不同的去标识数据的风险程度与级别是一个重要的问题。

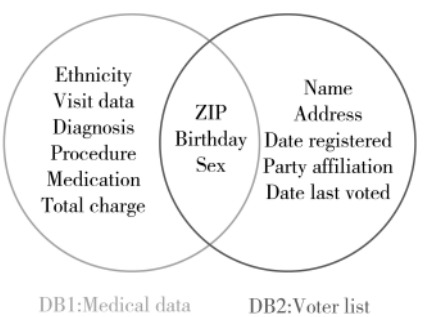

图 1 重标识攻击:两个数据库关联

本文针对该问题,基于 Shannon 信息熵和信息安全风险评估框架,提出了一种综合的重标识风险评估方法。

1 基于信息熵的重标识风险评估方法

本节首先介绍构建的重标识风险评估模型,接着分别对模型的三因素进行评估与刻画,最后阐述重标识风险综合评估方法与步骤。

1.1 评估模型

基于信息安全风险评估框架[9],本文对重标识攻击场景的风险进行建模:重标识风险值由重标识攻击发生的可能性(记为 prob)以及导致的危害性(记为 impact)两部分组成。进一步地,风险的可能性由去标识数据集的脆弱性的严重程度(简称脆弱性的严重程度,记为 v)和利用该脆弱性发起重标识攻击的频率(也称特定脆弱性的攻击频率,记为 f)两个因素有关;基于"属性越敏感,对攻击者的价值越大,其危害也越大"事实,风险的危害性可描述为与脆弱性的严重程度,以及利用该脆弱性实施攻击导致的隐私泄露程度(也称特定脆弱性的隐私泄露程度,记为 l)两个因素相关。评估模型综合以上风险的多个要素,采用乘法原则,那么某个脆弱性导致的风险分数可表述为:

$$\text{Riskscore} = \text{prob} \times \text{impact} \tag{1}$$

其中,$\text{prob} = v \times f$, $\text{impact} = v \times l$。风险值 Riskscore 的取值区间为[0,1],其分值越大,表示风险越大。

由上述模型可知,脱敏数据集的脆弱性分析与评估是关键。下面通过例子进行阐述。如图 2(a)所示,假设有一个公开可访问的去标识数据集,它有 5 条记录(它实际依次对应的身份为"Zhang Shan"、"Li Si"、"Wang Wu"、"Zhao Liu"和"Chen Qi"),列属性有 3 个准标识符(Sex、Age 和 Address)和 1 个敏感属性(Income)。此外,假设存在掌握不同身份数据表的攻击者,如图 2(b)~(d)所示:攻击者 Λ_1 掌握了"Name-Sex"身份数据表;攻击者 Λ_2 掌握了"Name-Age"身份数据表;攻击者 Λ_3 掌握了"Name-Sex-Age-Address"身份数据表。根据"Sex"的关联性,攻击者 Λ_1 可恢复出图 2(a)表的第二行记录的身份——"Li Si",同时获得她的敏感属性 Income 为"20K",以及其他的身份属性信息,即她的"Age"和"Address"。类似地,根据"Age"的关联性,攻击者 Λ_2 可重新标识出图 2(a)表的第一行记录的身份——"Zhang Shan"。攻击者 Λ_3 根据"Sex"、"Age"和"Address"的关联性,可重新标识出图 2(a)表的第一、二、三行记录的身份——"Zhang Shan"、"Li Si"和"Wang Wu"。由此可见,攻击者 Λ_1 根据去标识数据表的"Sex"列,攻击者 Λ_2 根据"Age"列,攻击者 Λ_3 根据"Sex"列、"Age"列和"Address"列分别重新标识出部分记录的

Sex	Age	Address	Income
Male	35	Beijing	70K
Female	20	Shanghai	20K
Male	20	Shanghai	30K
Male	30	Beijing	35K
Male	30	Beijing	40K

(a)去标识化数据集

Name	Sex
Zhang Shan	Male
Li Si	Female
Wang wu	Male
Zhao Liu	Male
Chen Qi	Male

(b)攻击者 Λ_1 掌握的身份数据集

Name	Age
Zhang Shan	35
Li Si	20
Wang wu	20
Zhao Liu	30
Chen Qi	30

(c)攻击者 Λ_2 掌握的身份数据集

Name	Sex	Age	Address
Zhang Shan	Male	35	Beijing
Li Si	Female	20	Shanghai
Wang wu	Male	20	Shanghai
Zhao Liu	Male	30	Beijing
Chen Qi	Male	30	Beijing

(d)攻击者 Λ_3 掌握的身份数据集

图 2 去标识数据集以及攻击者可能掌握的数据集

身份,造成一定的隐私泄露。攻击者利用的这些“单列”以及“列组合”可以看作是去标识数据集上客观存在的各种不同的脆弱性。

定义 1 去标识数据集的脆弱性(简称脆弱性):去标识化数据集 $D(A_1,A_2,\cdots,A_m)$ 准标识符属性(包括 $A_1,A_2,\cdots,A_\omega$)中的一列或多列组合的关联特性可能被掌握不同攻击资源的攻击者利用,它们统称为去标识数据集的脆弱性。它们可分别表示为脆弱性 $\#(A_1)$、$\#(A_2)$、$\#(A_1,A_2)$ 和 $\#(A_1,A_2,\cdots,A_\omega)$。

定义 2 去标识数据集的脆弱性维度(简称脆弱性维度):去标识数据集的脆弱性对应的准标识符个数。例如脆弱性 $\#(A_1)$ 的维度为 1,$\#(A_1,A_2)$ 的维度为 2。

定义 3 特定脆弱性被攻击频率:去标识化数据集 $D(A_1,A_2,\cdots,A_m)$ 的某个特定脆弱性被攻击者攻击的总次数。

定义 4 特定脆弱性的隐私泄露程度:攻击者利用去标识化数据集 $D(A_1,A_2,\cdots,A_m)$ 的某个特定脆弱性,比如 $\#(A_1)$、$\#(A_2)$ 和 $\#(A_1,A_2)$,发起重标识攻击,导致敏感属性和准标识符属性数据泄漏的情况。

1.2 风险三因素的评估与刻画

为了方便后续的风险评估,作以下的攻击场景假设:

假设 1(可获取去标识数据集):攻击者可以访问、查询或下载去标识数据集。比如数据发布场景中,攻击者隐藏在接收者中接收到数据集。

假设 2(攻击者目标):攻击者发起重标识攻击的目标是为了恢复去标识数据集 $D(A_1,A_2,\cdots,A_m)$ 记录对应的身份,以及获取该身份关联的敏感属性值和准标识符属性值。

假设 3(攻击者能力):攻击者掌握身份数据集,且该数据集属性包括去标识数据集的准标识符属性字段列,假设掌握任意该属性列的概率为 p。

1.2.1 脆弱性的严重程度

由定义 2 可知,对于去标识数据集 $D(A_1,A_2,\cdots,A_m)$(前 ω 列是准标识符对应的属性列),维度为 1 的脆弱性为 $\#(A_1),\#(A_2),\cdots,\#(A_\omega)$,即该类型脆弱性的个数为 C_ω^1;维度为 2 的脆弱性为 $\#(A_1,A_2),\#(A_1,A_3),\cdots,\#(A_{\omega-1},A_\omega)$,即该类型脆弱性的个数为 C_ω^2。依次类推,维度分别为 $3,4,\cdots,\omega$ 的脆弱性个数为 $C_\omega^3,C_\omega^4,\cdots,C_\omega^\omega$。因此,脆弱性总个数为:

$$W=\sum_{k=1}^{\omega}C_\omega^k=2^\omega-1 \tag{2}$$

对以上 W 个脆弱性的严重程度进行评估。定性地分析,脆弱性的严重程度与数据集的分布密切相关:若脆弱性对应的列或列组合的数据分布越集中(即每条记录的取值趋向唯一),攻击越有可能发起重标识攻击,即重新标识出大部分记录对应的身份。综合考虑,本文选取信息熵[10]对脆弱性的严重程度进行刻画,有两点原因:(1)信息熵可有效反映数据分布情况。熵值越大,数据分布越集中,脆弱性严重性越大。(2)信息熵可有效反映攻击者获取的信息量。熵值越大,获取平均信息量越大,脆弱性严重性越大。因此,使用信息熵公式可以对每一种脆弱性 $\#(A_1',A_2',\cdots,A_k')$ 的严重程度进行逐一刻画与度量。假设去标识数据集总共有 n 条记录,为了方便统一评估与比较,对信息熵进行归一化,可得:

$$v=\frac{H(\#(A_1',A_2',\cdots,A_k'))}{\log n} \tag{3}$$

其中,由于 $0\leqslant H(\#)\leqslant H_{\max}=\log n$,那么归一化熵值为 $v\in[0,1]$。

1.2.2 特定脆弱性的攻击频率

定性地分析,对于任意一种特定的脆弱性,其

攻击的频率与脆弱性被利用的难易程度成反比关系：脆弱性被利用的难度系数 ρ 越小，其攻击频率 f 越大，所以可刻画为倒数关系：

$$f=\frac{1}{\rho} \tag{4}$$

而脆弱性被利用的难度系数 ρ 与脆弱性的维度 ϕ 密切相关：脆弱性的维度 ϕ 越大，即包含的准标识符个数越多，其被利用的难度系数 ρ 越大。因此，需设计一条衰减函数 $\rho=F(\phi)$，可充分反映两者相互影响的关系趋势。

基于假设 2，攻击者掌握的准标识符个数可看成 n 次(最多掌握 n 个准标识符)独立重复的伯努利试验，其每个属性掌握的概率为 p，即建模为二项分布 (n,p)。从攻击者角度看，大部分攻击者掌握的准标识符个数在 np 附近，掌握 t 个准标识符的概率为 $C_n^t p_0^t (1-p_0)^{n-t}$。从脆弱性角度看，对于一个维度为 ϕ 的脆弱性，掌握 ϕ 个及以上的准标识符的攻击者均可利用。因此，难度系数 ρ 可赋值为所有可能的攻击者利用该脆弱性的概率累加和的倒数，即：

$$\rho=\frac{1}{\sum_{k=1}^{\phi} C_n^k p_0^k (1-p_0)^{n-k}} \tag{5}$$

联合式(4)和式(5)，维度为 ϕ 脆弱性的攻击频率可表述为：

$$f=F(\phi)=\sum_{k=1}^{\phi} C_n^k p_0^k (1-p_0)^{n-k} \tag{6}$$

1.2.3 特定脆弱性的隐私泄露程度

定性地分析，对于任意一种特定的脆弱性，其成功攻击后导致的隐私泄露包括两方面：敏感属性和准标识符属性的泄露数据。对于以上任意一种类型的属性，本文定义其隐私泄露程度为“属性的敏感程度”和“去标识化程度”相乘得到。敏感属性敏感程度 s_{SA} 和准标识符属性敏感程度 s_{QID} 可人工进行预定义(满足 $0<s_{QID}\leqslant s_{SA}<1.0$)，比如敏感属性 $s_{SA}=1.0$，$s_{QID}=0.05$。而对于“去标识程度”，定义以下的评估指标进行刻画：

对于数值型的属性：

$$g=1-\frac{1}{n}\sum_{i=1}^{n}\varphi\left(\left|\frac{x_i'-x_i}{x_{\max}-x_{\min}}\right|\right) \tag{7}$$

其中：

$$\varphi(x)=\begin{cases}x, x\leqslant 1\\ 1, x>1\end{cases} \tag{8}$$

对于类别型的属性：

$$g=1-\frac{1}{n}\sum_{i=1}^{n}\delta(x_i', x_i) \tag{9}$$

其中：

$$\delta(x_i', x_i)=\begin{cases}1, x_i'\neq x_i\\ 0, x_i'=x_i\end{cases} \tag{10}$$

1.3 重标识风险综合评估方法

1.3.1 特定脆弱性的风险与平均风险分数评估

对于去标识数据集 $D(A_1,A_2,\cdots,A_m)$ 的第 i 个脆弱性可表示为 $\#_i$，假设使用上一节的评估方法，分别得到脆弱性的严重程度、攻击频率和隐私泄露频率分别为 v_i、f_i 和 l_i，那么该脆弱性导致的风险分数可表述为：

$$\text{Riskscore}_i=fr(\#_i)=\text{prob}_i\times\text{impact}_i=(v_i\times f_i)\times(v_i\times l_i) \tag{11}$$

其中，prob_i 和 impact_i 分别表示该脆弱性带来风险发生可能性以及导致的危害性。

分别计算去标识数据集 $D(A_1,A_2,\cdots,A_m)$ 所有脆弱性的风险分数，然后综合风险分数求平均，可得到该去标识数据集的平均风险分数：

$$\begin{aligned}\text{Riskscore}_{\text{mean}}&=\frac{1}{W}\sum_{i=1}^{W}fr(\#_i)\\ &=\frac{1}{W}\sum_{i=1}^{W}(v_i\times f_i)\times(v_i\times l_i)\end{aligned} \tag{12}$$

1.3.2 综合风险分数快速评估新方法

上一小节风险分数求平均是一种简单的综合评估方法，然而它存在以下的缺陷：(1)高维数据集的评估消耗时间长。由式(2)知，风险分数评估为 $W=2^{\omega}-1$，评估时间复杂度随准标识符维度 ω 指数级增长。(2)求平均使所有的脆弱性带来风险权重相同，无法客观地凸显风险的主要来源。

针对以上缺陷，本文继续提出了一种新的替代方法——基于增量的快速综合风险评估方法。主要思想是低维的脆弱性风险赋予更多的权重，这符合低维的脆弱性在实际攻击场景中更容易发生。具体步骤描述如下：

(1)识别去标识数据集 $D(A_1,A_2,\cdots,A_m)$ 的脆弱性增量方向及增量大小 $\Phi=(\Delta v_1,\Delta v_2,\cdots,\Delta v_{m_0})$。脆弱性增量发现算法如算法 1 所示。

算法 1. 脆弱性增量发现算法

输入：去标识数据集 $D(A_1,A_2,\cdots,A_m)$ 和准标识符

属性集$(A_1, A_2, \cdots, A_w)$。

输出：脆弱性增量属性 $\Psi=(A_1', A_2', \cdots, A_w')$，熵值增量 $\Phi=(\Delta v_1, \Delta v_2, \cdots, \Delta v_w)$。

1. $C \leftarrow \{1, 2, 3, \cdots, m_0\}$ /* 定义准标识编号集合*/
2. for each i in C do
3. $v_i \leftarrow H(T_n(A_n))/\log_2 n$ /* 计算归一化熵值 */
4. end for
5. $\Psi, \Phi \leftarrow \emptyset$
6. $v_0 \leftarrow 0$
7. while $C \neq \emptyset$ do
8. $v_{\max} \leftarrow \max(v_i), i \in C$ /* 定义准标识编号集合 */
9. $i_{\max} \leftarrow \text{argmax}(v_i), i \in C$ /* 定义准标识编号集合 */
10. $v_{\max} \leftarrow v_{\max} - v_0$ /* 定义准标识编号集合*/
11. $v_0 \leftarrow v_{\max}$ /* 取最大化归一化熵值 */
12. $i_{\max} \leftarrow \text{argmax}(v_i), i \in C$ /* 对应准标识符属性编码 */
13. $\Psi \leftarrow \Psi || A_{i_{\max}}$ /* 连接属性 */
14. $\Phi \leftarrow \Phi || \Delta v_{\max}$ /* 连续属性的熵值增量 */
15. $C \leftarrow C - \{i_{\max}\}$ /* 求补集 */
16. $T_n \leftarrow T_n(A_{i_{\max}})$ /* 已评估的列 */
17. for each i in C do
18. $T_n \leftarrow T_n || T_n(A_{i_{\max}})$ /* 待评估的列之间的连接 */
19. $v_i \leftarrow H(T_n(A_n))/\log_2 n$ /* 计算归一化熵值 */
20. end for
21. end while

(2)脆弱性发生的可能性综合评估。表示熵值增量及利用概率的累加，即：

$$\text{prob}_g = \sum_{i=1}^{m_0} \Delta v_i \times f_i \tag{13}$$

(3)脆弱性导致的隐私泄露危害性综合评估。表示熵值增量及隐私泄露的累加，即：

$$\text{impact}_g = \sum_{i=1}^{m_0} \Delta v_i \times l_i \tag{14}$$

(4)整体数据集的重标识数据风险综合评估。结合式(13)、(14)，得：

$$\text{Riskscore}_g = \text{prob}_g \times \text{impact}_g \tag{15}$$

对本节两种综合评估算法的计算复杂度进行简单的分析与比较：对于平均风险评估，其信息熵(数据集分布)的计算次数为 $M=2^{m_0}$，时间复杂度为 $O(2^{m_0})$，而快速综合风险计算信息熵的次数为 m_0，时间复杂度为 $O(m_0)$。一般来说，$m_0 \geqslant 5$，因此效率上后者显然优于前者。

2 实验结果与分析

2.1 数据集与参数设置

在试验仿真环境中，使用 PC，其参数为 Intel® Core™ i7-6700 CPU@3.40 GHz 处理器，8.00 GB 内存，64 位 Windows 7 操作系统。评估算法通过 Python 代码实现。数据集使用公开的 UCI-Adult 数据集，它是在美国 1994 年人口普查数据库中抽样得到的子数据集，共有 32 561 条记录，已经去除标识符字段，可以看成一个去标识数据集。UCI-Adult 数据集包含 15 个描述属性，本文选择其中的 10 个准标识符属性(Quasi-Identification，QID)和 1 个敏感属性(Sensitive Attribute，SA)——income 属性(取值为'<=50K'和'>50K')作为待评估的原始数据集。实验过程采用的参数 $p=0.3$，num=20。此外，QID 和 SA 的敏感程度分别假设为 $s_{\text{QID}}=0.05$，$s_{\text{SA}}=0.5$。

2.2 重标识风险评估实验

(1)实验 1：敏感属性对重标识风险的影响

该实验有两个数据集，一个为原始的 UCI-Adult 数据集，另一个是不包含'income'属性列的 UCI-Adult 数据集。使用 1.3.2 节提出的快速评估方法，首先将根据算法 1 得到的脆弱性增量属性方向作为脆弱性增量属性 $\Psi=$(age，occupation，hours-per-week，education，relationship，workclass，race，sex，marital-status，native-country)，其对应的熵值增量为 $\Phi=$(0.379 1，0.226 2，0.178 9，0.101 2，0.049 4，0.017 9，0.008 8，0.005 6，0.003 0，0.002 6)。对两个数据集分别进行重标识风险评估，分别得到风险分数为 0.747 5、0.324 0。将其通过式(13)、(14)展开为可能性和危害性两个维度，风险分布可视化结果如图 3 所示。由图可知，两个数据集的可能性值相等，而包含 income 的数据集的危害值相比更大。这说明去除数据集的

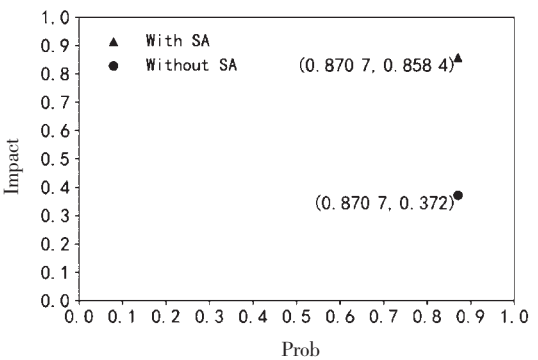

图 3 两个数据集的重标识风险分布

敏感属性，重标识攻击发生可能性不变，但重标识攻击发生后造成的隐私泄露危害显著地降低，这符合预期的评估目标，因为后者没有敏感属性的泄露。

(2)实验 2：不同维度数据集对重标识风险的影响

该实验在 10 个不同维度的去标识数据集(下文简称数据集)上进行重标识风险评估。具体来说，它们的维度分别为 2，3，4，…，11，它们均包含敏感属性‘income’列，其他属性从集合 Ψ=(age，occupation，hours-per-week，education，relationship，workclass，race，sex，marital-status，native-country)中按顺序获取1，2，3，…，10 个准标识符。对这 10 个数据集分别进行评估，可得到可能值、危害值以及风险分数，绘制的趋势曲线如图 4 所示。可以看出，可能值、危害值以及风险分数均随着数据集维度的增大而增大，且维度达到一定数目(大于 6)时三个值增长趋向稳定。这在一定程度上反映了重标识攻击的风险趋势：(1)高维度数据集存在更多的脆弱性，且它包含低维子数据集所有的脆弱性，因此高维的风险更大；(2)高维相比低维数据集，引入了新的脆弱性，但该脆弱性维度也随之增大。根据 1.2.2 节分析，可被攻击者成功利用的概率更低，因此风险随着维度的增加将趋向平缓。进一步地，可得实际应用的启示：应尽量避免高维数据集的发布与共享，遵循“最小化可用原则”，进而可降低隐私泄露的风险。

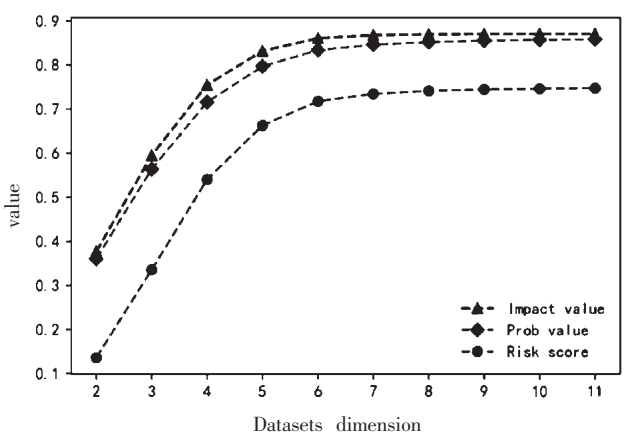

图 4 不同维度的数据集对重标识风险的影响

(3)实验 3：不同去标识程度对重标识风险的影响

该实验在原始 UCI-Adult 数据集以及三个不同去标识化程度的数据集上进行。定义三个不同去标识化处理的数据集：①低去标识程度的数据集。它的 2 个数值型属性均进行量化处理，其量化步长为 5，8 个类别型属性进行随机替换“*”屏蔽处理，屏蔽比例 5%。②中去标识程度数据集。同①的处理，其量化步长为 10，屏蔽比例 20%。③高去标识程度数据集。同①的处理，其量化步长为 10，屏蔽比例为 60%。将这四个数据集进行重标识风险评估，其风险可能性和危害性可展示为二维坐标图，如图 5 所示。由图可以看出，去标识化程度越强，重标识风险攻击发生的可能性以及危害性均呈现出下降趋势。这可以客观反映风险的下降趋势：(1)数据集的去标识化程度越大，那么记录对应的个人属性组合的区分度越差，因此攻击者成功实施重标识攻击的可能性更小；(2)即使攻击者通过重标识攻击可以将数据集的部分记录对应身份复原出来，由于大部分记录的属性已经进行屏蔽和量化处理，因此对隐私泄露的危害相比更低一些。

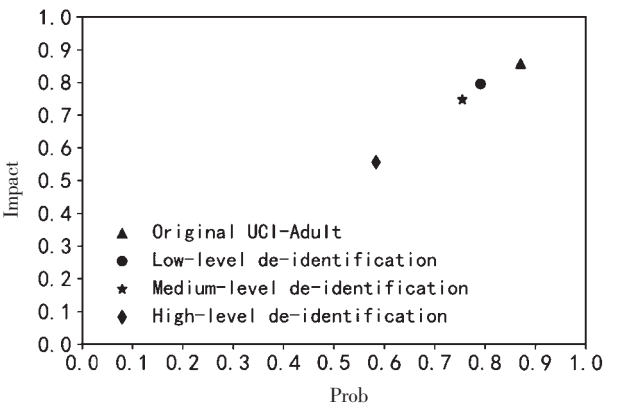

图 5 不同的去标识程度对重标识风险的影响

3 结论

本文基于 Shannon 信息熵和信息安全风险评估框架，提出了一种更为综合的重标识风险评估方法，不仅刻画了风险的可能性维度，同时刻画了其危害性维度。在具体评估中，首先将去标识数据集的若干属性组合建模为不同的脆弱性，然后针对每一种脆弱性进行多个维度的评估与刻画。最后，为了评估整体数据集，构造了一种基于信息熵值增量和加权的评估算法，实验结果验证了提出方案的有效性与可行性。

参考文献

[1] 绿盟科技.2019 年网络安全观察[EB/OL].(2020-01-xx) [2020-06-16].http://blog.nsfocus.net/wp-

content/uploads/2020/01/2019-Cybersecurity-Insights.pdf.

[2] 冯登国,张敏,李昊.大数据安全与隐私保护[J].计算机学报,2014,37(1):250-262.

[3] DWORKIN M.Recommendation for block cipher modes of operation:methods for format-preserving encryption[M].Gaithersburg:NIST Special Publication,2016.

[4] DWORK C.Differential privacy-Encyclopedia of cryptography and security[M].Boston,MA:Springer,2011.

[5] SWEENEY L.K-anonymity:a model for protecting privacy[J].International Journal of Uncertainty Fuzziness & Knowledge Based Systems,2002,10(5):557-570.

[6] MACHANAVAJJHALA A,KIFER D,GEHRKE J,et al.L-diversity:privacy beyond K-anonymity[J].ACM Transactions on Knowledge Discovery from Data,2007,1(1):1-52.

[7] ISO/IEC.Privacy enhancing data de-identification terminology and classification of techniques:20889[S].2018.

[8] Article 29 Data Protection Working Party.Opinion 05/2014 on anonymisation techniques[R/OL].(2014-03-xx)[2020-06-16]./https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/.

[9] 全国信息安全标准化技术委员会.信息安全技术—信息安全风险评估规范:20984[S].2017.

[10] SHANNON C E.A mathematical theory of communication[J].The Bell System Technical Journal,1948,27(3):379-423.

(收稿日期:2020-10-20)

作者简介:

陈磊(1991-),男,博士,绿盟和清华大学联合培养博士后,主要研究方向:数据安全、隐私保护、区块链。

薛见新(1984-),男,博士,绿盟和清华大学联合培养博士后,主要研究方向:图计算、网络攻击溯源、机器学习。

刘文懋(1983-),通信作者,男,博士,高级工程师,主要研究方向:网络安全、云安全、安全机器学习。E-mail:liuwenmao@nsfocus.com。