面向敏感性攻击的多敏感属性数据逆聚类隐私保护方法

张 冰,杨 静,张健沛,谢 静
(哈尔滨工程大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 针对传统 l-多样性模型仅考虑等价类中敏感值形式上的差异,而忽略敏感值的敏感度差异,且难以抵御一种新的攻击方式——敏感性攻击的问题,提出了一种使用逆文档频率 IDF 度量敏感值的敏感度的方法,并使用属性分解的方式构造敏感组,以避免多敏感属性数据表的 QI 属性泛化造成的高信息损失.同时,还提出了一种面向敏感性攻击的多敏感属性($l_1,l_2,\cdots,l_d$)-多样性隐私保护算法 MICD,该算法通过敏感度的逆聚类实现敏感组中敏感值的敏感度差异,以提高多敏感属性数据表抵御敏感性攻击的能力.实验结果表明,MICD 算法能够较好的抵御敏感性攻击,且具有较小的信息损失量.

关键词: 隐私保护; 敏感性攻击; 逆聚类; 多敏感属性; (l_1, l_2, …, l_d)-多样性; 敏感度差异
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2014)05-0896-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.05.010

A Multi Sensitive Attribute Data Inverse Clustering Privacy Preserving Algorithm for Sensitivity Attack

ZHANG Bing, YANG Jing, ZHANG Jian-pei, XIE Jing
(*College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*)

Abstract: In allusion to l-diversity model not considering sensitivity differences between the sensitive attributes, a new attack pattern which named sensitivity attack was proposed. Secondly, a new sensitive groups constructing method which based on sensitive attributes decomposition was proposed, and a keyword weight evaluation method called IDF was used to measure the sensitivity of the sensitive values. At the same time, a multi sensitive attributes ($l_1,l_2,\cdots,l_d$)-diversity privacy preserving method for sensitivity attack which called MICD was proposed, which guaranteed the sensitivity difference between sensitive values in sensitive groups by sensitivity inverse clustering. Experiment results demonstrated that the MICD algorithm could better protect sensitive attributes against sensitivity attack, and more effective on information loss.

Key words: privacy preserving; sensitivity attack; inverse clustering; multi sensitive attribute; ($l_1,l_2,\cdots,l_d$)-diversity; sensitivity degree difference

1 引言

随着信息技术的发展,数据发布中的隐私泄露问题越来越被人们所关注,如何有效保护个体的隐私安全并维持发布数据的可用性已成为信息安全领域的研究热点与难点之一.最初的隐私保护方法是删除数据表中的标识符属性,但攻击者仍可通过准标识符属性与背景知识获取目标个体的隐私信息.2002 年 Sweeney 等[1]提出了 k-匿名模型以解决数据表难以抵御攻击者通过逻辑推理获取个体敏感信息的链接攻击[2]的问题.为克服 k-匿名模型难以抵御同质性攻击的缺陷,Machanavajjhala 等[3]提出了 l-多样性模型以保证等价类中敏感值的多样化.在 k-匿名模型和 l-多样性模型的基础上,研究者们又提出了多种改进模型[4~6]以提高隐私保护方法的效用.

然而,以上方法通常假设数据表中仅存在唯一的敏感属性,但现实中的数据表往往存在多个敏感属性,以上方法不足以保护多敏感属性数据表的隐私安全.针

收稿日期:2013-05-07;修回日期:2013-09-02;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61370083,No.61073043,No.61073041);高等学校博士学科点专项科研基金(No.20112304110011,No.20121204110012);哈尔滨市科技创新人才研究专项资金(优秀学科带头人)(No.2011RFXXG015)

对此问题,杨晓春等[9]基于有损连接的思想,提出了一种针对多敏感属性数据发布的多维桶分组技术.Yang等[10]分析了在多敏感属性数据中应用传统泛化技术的局限性,并提出了一种基于分解技术实现敏感组多敏感属性$(l_1,_2,\cdots,l_d)$-多样性的隐私保护方法.以上研究都是 l-多样性模型的扩展[7,8],仅考虑等价类中敏感值的形式差异,忽略了敏感值的敏感度差异,容易将高敏感性的敏感值分配到同一等价类中,产生较大的隐私安全威胁.据此,本文发现了一种新的隐私攻击方式—敏感性攻击;针对多敏感属性数据发布中的隐私泄露及攻击者的敏感性攻击,本文在 l-多样性的基础上,提出了一种基于分解,以逆聚类方法构造敏感组的多敏感属性$(l_1,l_2,\cdots,l_d)$-多样性的隐私保护方法,在满足数据表多敏感属性值多样性的同时,有效的降低了数据表的敏感性,增强了数据的可用性.

2　多样性模型

2.1　单敏感属性 l-多样性模型

定义 1　等价类　设数据表 T' 为表 T 泛化后的匿名表,表 T' 中在准标识符属性上具有相同属性值的记录的集合称为表 T' 的等价类.

定义 2　l-多样性　设数据表 T' 为表 T 泛化后的匿名表,EC 为 T' 中的一个等价类,若 EC 中不同敏感值的个数不少于 $l(l\geqslant 2)$,则称等价类 EC 是满足 l-多样性的;如果表 T' 中的所有等价类都满足 l-多样性,则称表 T' 是满足 l-多样性的.

表 2 是表 1 原始医疗信息表满足 2-多样性的匿名表.其中,属性 Name 为标识符属性,属性 Age、Gender 与 Zipe Code 为准标识符属性,属性 Disease 为敏感属性.

表 1　医疗数据表

Name	Age	Gender	Zipe Code	Disease
Alice	24	F	13000	Flu
Betty	25	F	13009	Cancer
Carl	27	M	16000	HIV
Diana	29	F	16001	Gastritis
Ella	30	F	16004	H1N1

表 2　医疗数据 2-多样性表

ID	Age	Gender	Zipe Code	Salary
1	20 – 25	F	1300 *	Flu
2	20 – 25	F	1300 *	Cancer
3	26 – 30	G	1600 *	HIV
4	26 – 30	G	1600 *	SARS
5	26 – 30	G	1600 *	H1N1

2.2　多敏感属性$(l_1,l_2,\cdots,l_d)$-多样性模型

为实现表 T 的发布表 T' 的多敏感属性 l-多样性,需保证 T' 的等价类中每个敏感属性的不同敏感值都不少于 l.若以泛化的思想生成匿名表 T',会增加等价类规模和发布数据的信息损失;同时,不同敏感属性间的关联也会增加目标个体隐私泄露的风险.因此,本文使用属性分解的方法将准标识符属性与敏感属性分割开,并切断敏感属性间关联,实现信息损失的最小化.

定义 3　敏感组.　设表 T 中准标识符属性为$\{QI_1,QI_2,\cdots,QI_n\}$,敏感属性为$\{S_1,S_2,\cdots,S_d\}(d\geqslant 2)$,由 $n(n\geqslant 2)$条元组构成的敏感组 SA 定义为

$$SA=\{\{t_1^{QI},t_2^{QI},\cdots,t_n^{QI}\},dif\_S_1,dif\_S_2,\cdots,dif\_S_d\}$$

其中,t_i^{QI} 为第 i 条元组的 QI 属性值集合,dif_S_j为敏感组中的 n 条元组在敏感属性 S_j 上的不同敏感值的集合.

如表 3 所示,ID = 1 的敏感组 $SA_1=\{\{\{24,F,13000\},\{25,F,13009\}\},\{Flu,Cancer\},\{1500,4000\}\}$.基于属性分解构造敏感组的方法切断了敏感属性间的关联,攻击者无法获得目标对象的全部敏感信息.例如,攻击者得知目标对象为 ID = 1 的敏感组中第二条元组,也仅能推测目标对象的疾病为 Flu 或 Cancer,工资为 1500 或 4000,无法知道目标对象的具体疾病和工资.

表 3　医疗数据(2,2)-多样性表

ID	Age	Gender	Zipe Code	Disease	Salary
1	24	F	13000	Flu	1500
	25	F	13009	Cancer	4000
2	27	M	16000	HIV	2500
	29	F	16001	Gastritis	3000
	30	F	16004	SARS	5000

对于多敏感属性数据表中不同敏感属性的敏感值数目不同,若为全部敏感属性设置相同的多样性约束 l 会产生较大的信息损失.因此,需为不同敏感属性设置不同的多样性约束.

定义 4　$(l_1,l_2,\cdots,l_d)$-多样性.　设数据表 T 的敏感属性为$\{S_1,S_2,\cdots,S_d\}(d\geqslant 2)$,$T'$ 为表 T 的发布表,SG 为 T' 中的一个敏感组,若 SG 中第 i 个敏感属性的不同敏感值个数不少于 $l_i(l_i\geqslant 2)$,则称敏感组 SG 是满足$(l_1,l_2,\cdots,l_d)$-多样性的;如果表 T' 中的所有敏感组都满足$(l_1,l_2,\cdots,l_d)$-多样性,则称表 T' 是满足$(l_1,l_2,\cdots,l_d)$-多样性的.

例如,表 3 中每个敏感组在敏感属性 Disease 和 Salary 上的不同敏感值数目都不少于 2,是满足(2,2)-多样性的.

3 敏感性攻击

l-多样性模型要求发布数据表的每个等价类中不同敏感值的个数不少于 $l(l\geqslant 2)$,以抵御攻击者的同质性攻击带来的威胁.目前,已产生了许多针对 l-多样性模型的改进算法[11~15].攻击者得到发布数据表后信息增益的大小取决于数据隐私保护的程度高低.l-多样性模型虽然能够保证每个等价类中不同敏感值的形式多样化,但却无法确保敏感值的敏感度多样性.一旦攻击者确定目标个体属于大多敏感值为高度敏感的等价类中,获得的信息增益要远远高于目标个体属于普通等价类时所获得的.即使发布表实现了敏感值的 l-多样性,但数据隐私保护的程度却很低.本文将这种由攻击者的背景知识和发布的数据确定目标个体的归属等价类,通过等价类中敏感值的敏感程度高低分析目标个体的敏感信息的攻击方式称为敏感性攻击.

4 多敏感属性隐私保护算法

4.1 敏感度与距离度量

数据表中的敏感属性具有多种属性值,不同的属性值对应的敏感等级不同,所需的保护程度也不同.IDF(Inverse Document Frequency,逆文档频率)是信息检索中常用的关键词权重评估技术,它的思想是:文档集或语料库中,出现频率高的关键词比出现频率低的关键词传达的信息需求低,应被分配较低的权重.同样,在数据表中出现频率高的敏感值的敏感程度要低于出现频率低的敏感值.因此,本文引进 IDF 技术来评估敏感值的敏感度,将数值型属性映射到相应的区间转化成分类型属性进行处理.

定义 5 敏感值的敏感度. 设 n 为数据表 T 的元组总数,$\mathrm{F}(S,w)$为表 T 中敏感属性 S 上属性值为 w 的元组数目,敏感属性 S 上敏感值 w 的敏感度为

$$S\_\mathrm{degree}(S,w)=\log(n/\mathrm{F}(S,w))$$

为抵御攻击者的敏感性攻击,使每个敏感组中的敏感值具有较大的敏感度差异,本文提出一种基于逆聚类的思想实现数据表多敏感属性$(l_1,l_2,\cdots,l_d)$-多样性的方法,首先给出以敏感属性的敏感度判定元组间距离的度量方法.

定义 6 敏感度距离. 设数据表 T 的敏感属性为$\{S_1,S_2,\cdots,S_d\}$,t_i、t_j为表 T 中的两个不同元组,t_i、t_j间的敏感度距离 $S_dis(t_i,t_j)$定义为

$$S\_\mathrm{dis}(t_i,t_j)=\sum_{m=1}^{d}|S\_\mathrm{degree}(S_m,t_i)-S\_\mathrm{degree}(S_m,t_j)|$$

其中,$S_\mathrm{degree}(S_m,t_i)$为元组 t_i 在敏感属性 S_m 上的属性值的敏感度.

敏感组中心由敏感组中每个敏感属性的不同敏感值的平均敏感度获得.

定义 7 敏感组中心. 设数据表 T 的敏感属性为$\{S_1,S_2,\cdots,S_d\}$,敏感组 SG 由元组$\{t_1,t_2,\cdots,t_m\}$组成,则 SG 的敏感组中心 SG_center 定义为

$$SG\_\mathrm{center}=(\overline{S\_\mathrm{degree}(S_1)},\overline{S\_\mathrm{degree}(S_2)},\cdots,\overline{S\_\mathrm{degree}(S_d)})$$

其中,$\overline{S_\mathrm{degree}(S_i)}$为 SG 的第 i 个敏感属性 S_i 的不同敏感值的平均敏感度.

4.2 敏感组构造原则

最大敏感组构造原则[10]是以某一敏感属性为基准,l 为该敏感属性的多样性约束,将该敏感属性的不同敏感值划分到不同桶中,不断从桶容量最大的 l 个桶中各抽取一条元组构成初始敏感组.文献[10]证明了遵循最大敏感组构造过程能够构造最多的初始敏感组.

基于属性值分布更加均匀的敏感属性构造初始敏感组,能够生成更多的初始敏感组.敏感属性的信息熵值大小能够反映出该敏感属性的属性值分布的不确定性,因此,本文将信息熵值最大的敏感属性定义为主敏感属性,并以主敏感属性为基准构造初始敏感组.

定义 8 主敏感属性. 将多敏感属性数据表 T 的敏感属性$\{S_1,S_2,\cdots,S_d\}$按信息熵值降序排列,得到新的排序集合$\{S_1',S_2',\cdots,S_d'\}$,信息熵值最大的敏感属性 S_1' 定义为表 T 的主敏感属性.

本文在构建初始敏感组时,以元组分配惩罚最大的原则分配元组.

定义 9 元组分配惩罚. 设表 T 中的元组 t 与敏感组 SG,$t\notin SG$,元组 t 加入到敏感组 SG 中的分配惩罚 $P_{\mathrm{all}}(t,SG)$定义为 t 与 SG_center 在非主敏感属性上的距离,即

$$P_{\mathrm{all}}(t,SG)=S\_dis_{nonS_{key}}(t,SG\_\mathrm{center})$$

由于本文在以主敏感属性为基准构建初始敏感组的过程中,已经实现了初始敏感组中主敏感属性的敏感值间敏感度差异,同时为更好地实现敏感组中非主敏感属性的属性值间敏感度差异,计算 $P_{all}(t,SG)$时不考虑主敏感属性间的敏感度距离.

为保证敏感组中敏感值的敏感度的差异性,本文在最大敏感组构造原则基础上提出最低敏感度敏感组构造原则.

定义 10 最低敏感度敏感组构造原则. 以主敏感属性为基准,l_{key}为主敏感属性的多样性约束,将主敏感属性的不同敏感值划分到不同桶中,不断从桶容量最大的 l_{key}个桶中抽取元组分配惩罚最大的元组构成初始敏感组,此过程称为最低敏感度敏感组构造原则.

对初始敏感组中未满足$(l_1,l_2,\cdots,l_d)$-多样性的敏感组,通过敏感组合并的方式来实现敏感组的$(l_1,l_2,\cdots,l_d)$-多样性.

定义 11　敏感组合并惩罚.　设表 T 中的两个敏感组 SG_i与 SG_j, SG_i与 SG_j合并产生的敏感组合并惩罚定义为

$$P_{\text{com}}(SG_i,SG_j)=S\_\text{dis}(SG_i\_\text{center},SG_j\_\text{center})$$

其中, SG_i_center为敏感组 SG_i 的中心.

对于经合并未能满足$(l_1,l_2,\cdots,l_d)$-多样性的敏感组,通过添加噪声的方式实现敏感组的$(l_1,l_2,\cdots,l_d)$-多样性多样性.

定义 12　噪声添加惩罚.　设表 T 中敏感属性$\{S_1,S_2,\cdots,S_d\}$的多样性约束分别为$\{l_1,l_2,\cdots,l_d\}$,对于在非主敏感属性 S_i上未满足 l_i-多样性的敏感组 SG, $\forall s\in S_i$且 $s\notin SG.S_i$,向 $SG.S_i$中添加噪声值 s 的噪声添加惩罚定义为

$$P_{\text{add}}(s,SG.S_i)=S\_\text{dis}(s,SG.S_i\_\text{center})$$

其中, $SG.S_i_\text{center}$为敏感组 SG 在敏感属性 S_i上的中心.

4.3　信息损失度量

为了满足发布表多敏感属性$(l_1,l_2,\cdots,l_d)$-多样性的要求,在构建敏感组时可能会适当添加噪声以保证发布数据的安全性.本文用噪声比率来衡量噪声属性值在发布数据表 T'中的比例,并用该度量衡量表 T'的信息损失.噪声比率定义为:

$$\text{NoiseRatio}=n_{\text{Noise}}/n$$

其中, n_{Noise}为添加的噪声总数目, n 为表 T 中元组的数目.噪声比率反映了在发布的所有敏感值中噪声数据所占的比例.理想情况下,当所有敏感组都能通过合并实现$(l_1,l_2,\cdots,l_d)$-多样性时,此时噪声比率值最小为0;当所有敏感组都需要添加噪声实现$(l_1,l_2,\cdots,l_d)$-多样性时,噪声比率达到最大值$(\sum l_i-d)/l_{\text{key}}$.显然,噪声比率越小,向发布数据表 T'中添加的噪声越少,信息损失也越小,表 T'的数据可用性也越高.

4.4　基于属性分解的多敏感属性逆聚类算法

本文针对敏感性攻击,提出一种基于属性分解的多敏感属性逆聚类隐私保护算法(Multi Sensitive Attribute Inverse Clustering based on Attribute Decomposition, MICD),算法的具体思想是:首先,按信息熵值降序排列敏感属性并确定主敏感属性,将表 T 中元组按主敏感属性的属性值划分为多个桶;其次,以主敏感属性为基准,按照最低敏感度敏感组构造原则,每次分别从桶容量最高的前 l_{key}个桶中抽取分配惩罚最大的元组对表 T 逆聚类,形成主敏感属性 l_{key}-多样性的初始敏感组;再次,将表 T 中剩余元组分配到敏感度距离最远的初始敏感组中;最后,分别以每个非主敏感属性 S_i为基准,通过合并敏感度距离最大的敏感组和添加噪声两种途径实现敏感组中每个非主敏感属性的多样性.MICD 算法如下所示.

算法 1　基于属性分解的多敏感属性逆聚类算法(MICD)

输入:待发布数据表 T,敏感属性$\{S_1,S_2,\cdots,S_d\}$,多样性约束$(l_1,l_2,\cdots,l_d)$

输出:发布表 T'

//初始化

1.1　while($i=1,i<d,i++$)do

1.1.1　{计算 S_i的不同敏感属性个数 n_i;

1.1.2　　if $n_i<l_i$ then 重新设定 l_i;}

1.2　计算敏感属性$\{S_1,S_2,\cdots,S_d\}$的信息熵值;

1.3　按信息熵值降序排列敏感属性,确定主敏感属性和主敏感属性的多样性约束 l_{key};

1.4　确定每个敏感值的敏感度;

1.5　按主敏感属性的属性值将表 T 划分为多个桶$\{B\}$;

1.6　$G=\varnothing$;

//初始敏感组的构建

2.　while($\{B\}$中非空桶的数量$\geqslant l_{\text{key}}$)do

2.1　{计算$\{B\}$中每个桶的容量;

2.2　计算桶容量的降序排列$\{B_1,B_2,\cdots,B_n\}$;

2.3　在 B_1中随机抽取元组 t;

2.4　$B_1=B_1-\{t\}$, $SG=\{t\}$;

2.5　while($i=2,i<l_{\text{key}},i++$)do

2.5.1　{在 B_i中抽取 $P_{\text{all}}(t',SG)$最大的元组 t';

2.5.2　　$B_i=B_i-\{t'\}$, $SG=SG\cup\{t'\}$;}

2.6　$G=G\cup SG$;}

//剩余元组的分配

3.　for(每个$\{B\}$中的剩余元组 t)do

3.1　{寻找 $P_{\text{all}}(t,SG)$最大的敏感组 SG;

3.2　$SG=SG\cup\{t\}$;}

//非主敏感属性多样性的构建

4.　for(每个非主敏感属性 S_i)do

4.1　{将 G 中未满足 l_i-多样性的敏感组加入 G_1中;

4.2　for(G_1中每个敏感组 SG)do

4.2.1　{if G_1中存在与 SG 合并后满足 l_i-多样性的敏感组

4.2.1.1　{在这些敏感组中选择 $P_{\text{com}}(SG,SG')$最大的敏感组 SG';

4.2.1.2　　$SG=SG\cup SG'$;}

4.2.2　　else

4.2.2.1　　{在 SG 中添加 $P_{\text{add}}(s,SG.S_i)$最大的 $l_i-|SG.S_i|$个敏感值;}

4.2.3　　将 SG 加入 G 中;}}

5.　将 G 中所有敏感组加入 T'中,发布 T';

4.5　正确性与复杂性分析

(1)正确性分析　　从整个算法的执行过程看,首先,以主敏感属性为基准并划分桶,不断在容量最大的桶中分别抽取分配惩罚最大的 l_{key}条元组逆聚类构建敏感组,实现了主敏感属性的 l_{key}-多样性,且逆聚类的

方式保证了敏感组中敏感值的敏感度差异.然后,通过不断合并合并惩罚最大的敏感组或向敏感组中添加噪声,以实现非主敏感属性的多样性.以上过程不但保证了每个敏感组在敏感属性上敏感值的多样性,也保证了敏感值的敏感度的差异性,算法是正确可行的.

(2)复杂性分析 设表 T 为含有 d 个敏感属性、n 条元组的数据表.算法的步骤 1 为初始化工作,可在 $O(n)$内完成;步骤 2 为初始敏感组的构建,每构建一个初始敏感组至多需时间 $O(n)$,设步骤 2 共构建 m 个初始敏感组,则需时间 $mO(n)$.步骤 3 为剩余元组的分配,每分配一个元组需计算 m 次元组分配惩罚,因此步骤 3 至多需时间 $mO(n)$.步骤 4 为非主敏感属性多样性的敏感组构建,每次至多需计算 $m-1$ 次敏感组合并惩罚或添加噪声实现非主敏感属性 S_i 上的 l_i-多样性,步骤 4 至多需时间 $(d-1)(m-1)O(n)$.步骤 5 需将构建的所有敏感组加入发布表 T' 中,需时间 $O(n)$.由于 $m \leqslant n/l_{\text{key}}$ 且 d 为常数,因此,MICD 算法可在时间 $O(n^2)$内实现多敏感属性数据表的隐私保护.

5 实验及结果分析

5.1 实验数据及参数设定

实验采用 UCI Machine Learning Repository 中的 Adult 数据集作为实验数据,删除存在缺省值的记录,共包含 30162 条记录.本文选取 Adult 数据集中的 9 个属性作为实验对象,选取{Age, Sex, Salary, Native-country}4 个属性作为准标识符属性,选取{Education, Occupation, Marital-status, Race, Work class}5 个属性作为敏感属性,并以敏感属性数目的不同分为 4 种情况,具体描述如表 4 所示.其中属性 Education 的信息熵值最大,被确定为主敏感属性.

表 4 敏感属性组成描述

$\lvert S\rvert$	**Sensitive attributes**
2	Education, Occupation
3	Education, Occupation, Marital-status
4	Education, Occupation, Marital-status, Race
5	Education, Occupation, Marital-status, Race, Work class

本实验将本文所提的 MICD 算法与文献[10]中所提的 Decomposition 算法、文献[11]中所提 Decomposition + 算法从敏感性攻击抵御能力、信息损失度与执行时间三方面进行比较分析.实验的运行环境为:硬件环境为 Inter Pentium(R) 4 CPU 3.00GHz 处理器,2.00GB 内存,Microsoft Windows XP 操作系统,算法均在 VC + +6.0 与 Matlab 7.0 混合编程环境下实现.

5.2 实验结果分析

实验 1 敏感性攻击抵御能力分析 本部分实验按表 4 确定敏感属性的组成.将每个敏感属性中敏感度最高的前 30% 及与其他敏感值差异最大的多个敏感值设定为该属性的高敏感属性值,各属性的高敏感属性值集合如表 5 所示.任何敏感组中若存在某一敏感属性 70% 的敏感值都为高敏感的,则意味着该敏感组易受敏感性攻击.

表 5 高敏感属性值组成描述

Race	Amer-Indian-Eskimo, Other
Education	Preschool, 1st-4th, 5th-6th, Doctorate, 12th
Marital-status	Married-AF-spouse, Married-spouse-absent
Work class	Without-pay, Federal-gov
Occupation	Armed-Forces, Priv-house-serv, Protective-serv

图 1 给出了随多样性约束 l 增大三种算法产生易受敏感性攻击的敏感组的概率对比,其中敏感属性个数 $|S|$ 为 3.敏感属性的 l 值都从 2 开始,每次实验增加 1.图 1 可知三种算法的被攻击概率都随 l 值增大呈先增加后降低的趋势.当 $l<6$ 时,三种算法的被攻击概率都随 l 值增大而增加.由于随 l 值增大,MICD 算法主要以合并的方式实现非主敏感属性的多样性,高敏感属性值被划分到同一敏感组中的概率增加,MICD 算法的被攻击概率随 l 值增大而增加;而另两种算法都仅采用添加噪声的方式达到非主敏感属性的多样性,l 值较小且逐渐增大时,两种算法添加的噪声增多,敏感组中存在多个高敏感属性值的概率增大,Decomposition 算法与 Decomposition + 算法的被攻击概率随 l 值增大.当 $l \geqslant 6$ 时,三种算法的被攻击概率都随 l 值的增大而降低.因为,此时敏感组中敏感属性的多样性增加,且在 l 值增大过程中逐渐接近每个敏感属性的属性值基数,l 值越接近该基数,敏感组中高敏感属性值所占比例越低,因此,三种算法的被攻击概率都有所降低.由于 MICD 算法在敏感组合并与噪声添加的过程中限制了敏感组的敏感度,MICD 算法的被敏感性攻击概率要低于另两种算法.

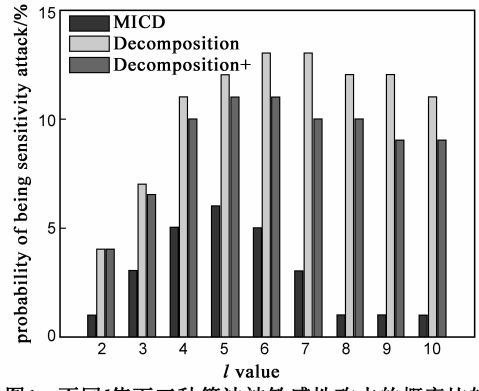

图1 不同l值下三种算法被敏感性攻击的概率比较

图 2 给出了随敏感属性个数 $|S|$ 增多三种算法生

成的易被敏感性攻击的敏感组的概率对比,其中 l 值为 3.图 2 可知三种算法的被攻击概率都随 $|S|$ 增多而增加.由于 $|S|$ 值的增加,MICD 算法在生成满足 $(l_1, l_2, \cdots, l_d)$-多样性的敏感组时所涉及的敏感属性个数增多,敏感组的合并次数增多,高敏感属性值被划分到同一敏感组中的概率增加,因此 MICD 算法的被攻击概率随 $|S|$ 值增大而增加.而随 $|S|$ 值增大,Decomposition 算法和 Decomposition + 算法添加的噪声也增多,敏感组中存在多个高敏感属性值的概率增大,算法被敏感性攻击的概率随 $|S|$ 值增大.由于 MICD 算法在敏感组合并与噪声添加的过程中限制了敏感组的敏感度,因此,MICD 算法被敏感性攻击的概率要低于另两种算法.

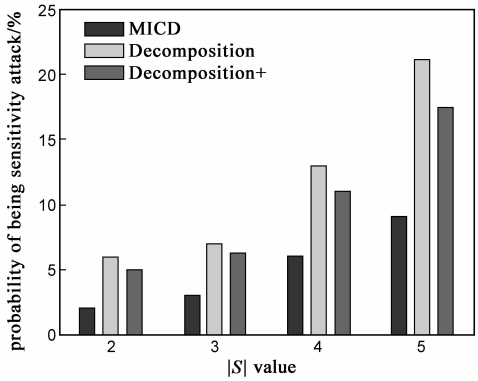

图2　不同$|S|$值下三种算法被敏感性攻击的概率比较

实验 2　噪声比率分析　　图 3 给出了随多样性约束 l 值增大三种算法的噪声比率对比,其中敏感属性个数 $|S|$ 为 3.图 3 可知 MICD 算法的噪声比率随 l 值变化呈先增大后减小的趋势.这是由于 l 较小且逐渐增大时,非主敏感属性未满足多样性需求的初始敏感组增多,需添加的噪声也逐渐增多;当 l 值较大时,初始敏感组中非主敏感属性的多样性逐渐增强,敏感组大多通过合并方式实现非主敏感属性的多样性需求,噪声比率逐渐降低.而 Decomposition 算法和 Decomposition + 算法的噪声比率随 l 值增大而增大.l 值较小时,另两种算法所生成的初始敏感组基本全部满足 $(l_1, l_2, \cdots, l_d)$-多样性需求,此时添加噪声较少;随 l 值增大,敏感组的非主敏感属性的属性值的多样化增高,算法需添加的噪声逐渐增多但增长趋势变缓.而相对另两种算法,MICD 算法采用合并敏感组和添加噪声两种方式达到非主敏感属性的多样性,因此 MICD 算法的噪声比率要远低于 Decomposition 算法和 Decomposition + 算法.

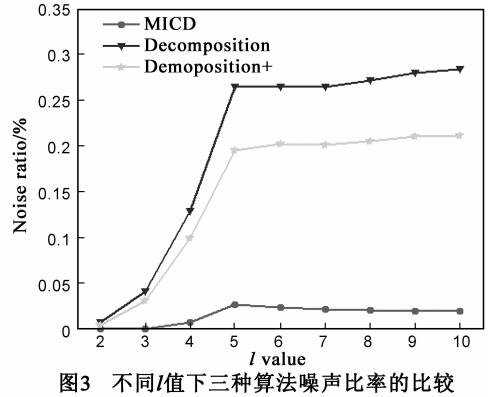

图3　不同l值下三种算法噪声比率的比较

图 4 给出了随敏感属性个数 $|S|$ 增多三种算法的噪声比率对比,其中 l 值都为 3.图 4 可知三种算法的噪声比率都会随 $|S|$ 增多而增加.由于 $|S|$ 增加,三种算法需处理的敏感属性个数增多,添加的噪声也增多.由于 l 值较小,在 $|S|<4$ 时,三种算法所生成的初始敏感组基本全部满足 $(l_1, l_2, \cdots, l_d)$-多样性需求,此时噪声比率接近 0.由于表 T 结构不变,每次生成的初始敏感组相同,在构建非主敏感属性多样性时,$|S|$ 值每增加 1,噪声增量呈线性增长且增长较少.而相对另两种算法,MICD 算法首先采取敏感组合并的方式实现非主敏感属性的多样性,以降低噪声添加量,因此 MICD 算法的噪声比率要低于另两种算法.

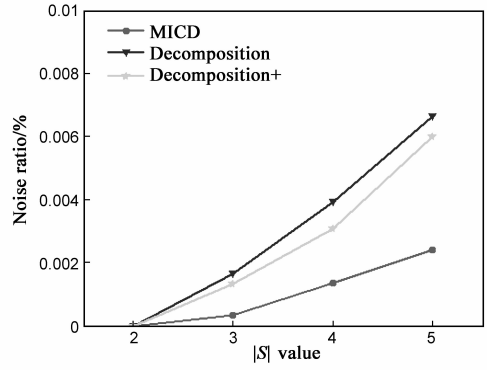

图4　不同$|S|$值下三种算法噪声比率的比较

实验 3　执行时间分析　　图 5 给出了随多样性约束 l 增大三种算法的执行时间对比,其中敏感属性个数 $|S|$ 为 3.图 5 可知三种算法的执行时间随 l 值增大而降低.由于随 l 值的增大,MICD 算法主要以合并的方式实现非主敏感属性的多样性,且合并惩罚的计算次数逐渐降低,减少了敏感组合并与添加噪声的时间开

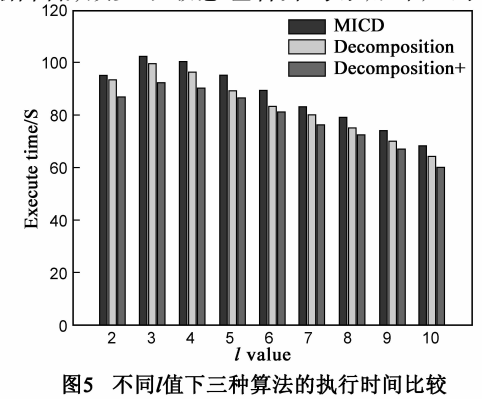

图5　不同l值下三种算法的执行时间比较

销.同理,随 l 值增大,另两种算法生成的初始敏感组中元组增多,向敏感组中添加的噪声减少,算法的执行时间降低.相比 Decomposition 算法,Decomposition + 算法少了候选噪声集合的构建过程,所需执行时间要少于 Decomposition 算法.由于 MICD 算法较另两种算法多了敏感组的合并过程,相同条件下的执行时间要高于另外两种算法,但三种算法的执行时间差异在 10s 以内.MICD 算法以较少的时间开销换取了数据的精确性,是可以被接受的.

图 6 给出了随敏感属性个数 $|S|$ 增多三种算法的执行时间对比,其中 l 值都为 3.图 6 可知三种算法的执行时间都会随 $|S|$ 值的增多而增加.这是由于敏感属性个数 $|S|$ 增多时,生成满足 $(l_1, l_2, \cdots, l_d)$-多样性的敏感组所涉及的敏感属性个数增多,因此需要更大的时间开销.与另两种算法相比,MICD 算法多了敏感组合并的过程,相同条件下的执行时间要略高于另两种算法,但三种算法的执行时间差异在 10s 以内.

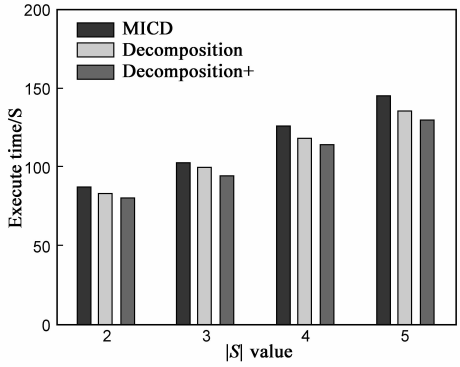

图6 不同|S|值下三种算法的执行时间比较

6 结论

本文针对 l-多样性模型仅考虑等价类中敏感值形式上的差异性,而忽略敏感值的敏感度差异性问题,发现了一种新的隐私攻击方式——敏感性攻击.针对攻击者的敏感性攻击及多敏感属性数据发布的隐私保护,提出了一种基于属性分解的多敏感属性隐私保护算法 MICD,该算法以敏感度逆聚类的思想构造敏感组,在保证数据表满足 $(l_1, l_2, \cdots, l_d)$-多样性的同时,增加了敏感组中敏感值的敏感度差异性.实验结果表明,MICD 算法不仅能够有效抵御敏感性攻击,且具有较低的噪声比率,能够更加有效地保护多敏感属性数据发布中的隐私安全.今后的工作将考虑不同敏感属性的个性化需求,研究多敏感属性数据发布中的个性化隐私保护问题.

参考文献

[1] Sweeney L. k-anonymity: A model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge Based Systems, 2002, 10(5): 557 – 570.

[2] Sweeney L. Computational disclosure control: A primer on data privacy protection[D]. Massachusetts Institute of Technology, 2001.

[3] Machanavajjhala A, Kifer D, Gehrke J. l-diversity: Privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1 – 52.

[4] 杨高明,杨静,张健沛.聚类的(α, k) – 匿名数据发布[J].电子学报,2011,39(8):1941 – 1946.
Yang Gaoming, Yang Jing, Zhang Jianpei. Achieving (α, k)-anonymity via clustering in data publishing[J]. Acta Electronica Sinica, 2011, 39(8): 1941 – 1946. (in Chinese)

[5] Batya Kenig, Tamir Tassa. A practical approximation algorithm for optimal k-anonymity[J]. Data Mining and Knowledge Discovery, 2012, 1(25): 134 – 168.

[6] 韩建民,于娟,虞慧群,等.面向敏感值的个性化隐私保护[J].电子学报,2010,38(7):1723 – 1728.
Han Jianmin, Yu Juan, Yu Huiqun, et al. Individuation privacy preservation oriented to sensitive values[J]. Acta Electronica Sinica, 2010, 38(7): 1723 – 1728. (in Chinese)

[7] Xiaoxun Sun, Min Li, Hua Wang. A family of enhanced (L, α)-diversity models for privacy preserving data publishing[J]. Future Generation Computer Systems, 2011, 27(3): 348 – 356.

[8] 王波,杨静.一种基于逆聚类的个性化隐私匿名方法[J].电子学报,2012,40(5):883 – 890.
Wang Bo, Yang Jing. A personalized privacy anonymous method based on inverse clustering[J]. Acta Electronica Sinica, 2012, 40(5): 883 – 890. (in Chinese)

[9] 杨晓春,王雅哲,王斌,等.数据发布中面向多敏感属性的隐私保护方法[J].计算机学报,2008,31(4):574 – 587.
Yang Xiaochun, Wang Yazhe Wang bin, et al. Privacy preserving approaches for multiple sensitive attributes in datapublishing[J]. Chinese Journal of Computers, 2008, 31(4): 574 – 587. (in Chinese)

[10] Ye Y, Liu Y, Wang C. Decomposition: privacy preservation for multiple sensitive attributes[A]. Proc of the 14th International Conf on Database Systems for Advanced Applications[C]. Berlin: Springer, 2009. 486 – 490.

[11] Devayon D, Dhruba K. Decomposition + : Improving l-diversity for multiple sensitive attributes[A]. Proc of 2nd International Conf on Advances in Computer Science and Information Technology[C]. Berlin: Springer, 2012. 403 – 412.

[12] Xiaokui Xiao, Ke Yi, Yufei Tao. The hardness and approximation algorithms for l-diversity[A]. Proc of the 13th International Conf on Extending Database Technology[C]. New York: ACM, 2010. 135 – 146.

[13] Ninghui Li, Jian Zhang, Molloy I. Slicing: A new approach for privacy preserving data publishing[J]. IEEE Transactions on

Knowledge and Data Engineering,2009,24(3):561 – 574.

[14] 王智慧,许俭,汪卫,等.一种基于聚类的数据匿名方法[J].软件学报.2010,21(4):680 – 693.
Wang Zhihui, Xu Jian, Wang Wei, et al. Clustering-based approach for data anonymization[J]. Journal of Software, 2010, 21(4):680 – 693. (in Chinese)

[15] J Liu, K Wang. On optimal anonymization for l + -diversity [A]. Proc of IEEE 26th International Conf on Data Engineering[C]. New York: IEEE Computer Society, 2010. 213 – 224.

作者简介

张　冰　女,1986 年生于黑龙江哈尔滨.哈尔滨工程大学计算机科学与技术学院博士研究生.主要研究方向为数据挖掘、隐私保护.
E-mail: zhangbing006@hrbeu.edu.cn

杨　静　女,1962 年生于黑龙江哈尔滨.哈尔滨工程大学计算机科学与技术学院教授、博士生导师.主要研究方向为数据库与知识工程、数据挖掘、隐私保护、软件理论等.
E-mail: yangjing@hrbeu.edu.cn