面向敏感值的层次化多源数据融合隐私保护

杨月平 王 箭 薛明富
(南京航空航天大学计算机科学与技术学院 南京 211106)

摘 要 数据融合技术能够使用户得到更全面的数据以提供更有效的服务。然而现有的多源数据融合隐私保护模型没有考虑数据提供者的重要程度,以及数据不同属性和属性值的敏感度。针对上述问题,提出了一种面向敏感值层次化的隐私模型,该模型通过数据提供者对数据的匿名程度要求来设置数据属性以及属性值的敏感度以实现敏感值的个性化隐私保护。同时结合k-匿名隐私模型以及自顶向下特殊化TDS的思想提出了一种面向敏感值的多源数据融合隐私保护算法。实验表明,该算法既能实现数据的安全融合,又能获得更好的隐私保护。

关键词 数据融合,敏感度,层次化隐私模型,k-匿名
中图法分类号 TP292.2 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.030

Hierarchical Privacy Protection of Multi-source Data Fusion for Sensitive Value

YANG Yue-ping WANG Jian XUE Ming-fu
(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract Data fusion technology enables users to get more comprehensive data to provide more effective service. However, the existing multi-source data fusion privacy protection models do not consider the importance of the data providers, and the sensitivity of different attributes and attribute values. According to the above problems, this paper proposed a hierarchical privacy model for sensitive value. The model enables data providers to set sensitive value of data attributes and attribute values by anonymous degree requirements to realize the individual privacy protection of sensitive values. At the same time, this paper proposed a multi-source data fusion privacy protection algorithm for sensitive value combining with k-anonymous privacy model and the top-to-down specialization TDS. Experiments show that the proposed algorithm can not only realize data security fusion, but also obtain better privacy protection.

Keywords Data integration, Sensitivity, Hierarchical privacy mode, k-anonymous

1 引言

随着大数据时代的来临,大量数据被存储在不同的存储系统中,例如:医院存储患者的医疗数据,银行存储财产收入数据,统计机构拥有户口调查数据等。通常这些数据拥有者想通过融合数据做更好的决策分析或者为顾客提供更好的定制服务。例如:医疗数据的融合可以帮助医生对病情做出更好的决策,金融数据的融合可以让银行为顾客提供更合理的定制服务。尽管数据共享可以帮助顾客获得想要的信息,但是在半诚实模型下共享的数据存在着隐私泄露的风险。

这个研究问题是在瑞士的一个金融合作项目中被发现的,问题描述如下:贷款公司A和银行B分别拥有相同个体的不同属性集,这些数据集有相同的ID标识,A拥有 $T_A(ID, Age, Balance)$,B拥有 $T_B(ID, Job, Salary)$,这些公司想通过融合他们的数据来提供更好的决策,例如:银行是否要贷款给公司。若简单地将数据进行融合,A可以得到B中的敏感数据,B也可以得到A中的敏感数据,或者通过融合后的数据可以推断出某个具体的个体信息,该情况对于A和B而言都是不愿看到的。

在现有的隐私模型中,k-匿名[1]作为一种有效的隐私保护模型,一直受到广泛的关注,它要求发布的数据中至少存在 $k-1$ 个相同的等价组,使攻击不能标识出指定的个体。使用k-匿名实现多源数据的融合一直备受青睐。文献[2]提出DkA框架来实现k-匿名的数据融合,它主要通过自底向上泛化的思想融合两方数据表,但是每次进行泛化时需要使用安全交叉集判断是否满足准标示符k-匿名条件,当数据量过大时,使用的比较协议过多,时间花费过长,且只能实现两个数据表的融合。为了解决上述问题,文献[3]提出了一种利用k-匿名实现的多源数据融合模型,该模型解决了文献[2]中存在的不足,但是在融合多源数据表时没有考虑到数据提供者

到稿日期:2016-08-03 返修日期:2016-12-26 本文受中国博士后科学基金(2014M561644),江苏省博士后科学基金(1402034C)资助。
杨月平(1992-),男,硕士生,主要研究方向为信息安全隐私保护,E-mail:yangyueping0720l@163.com;**王 箭**(1968-),男,博士,教授,主要研究方向为信息安全,E-mail:xuesl@nuaa.edu.cn;**薛明富**(1986-),男,博士,讲师,主要研究方向为硬件木马检测。

的重要程度，例如：一定程度上，市级政府机构存储的数据相比于县级政府机构存储的数据的安全级别更高；二级医院相比于一级医院的数据的权威程度更高，融合时数据提供者的相对重要程度不同。同时在每个数据表中不同属性的敏感程度也不相同，例如：salary 属性和 health condition 属性相比于 sex 属性更加敏感，需要的隐私级别更高。

为了解决上述问题，本文提出了一种层次化的隐私模型，对不同的数据提供者以及数据属性和属性敏感值进行不同的定义，同时基于所给定义提出一种面向敏感值的多源数据融合算法以进行数据的安全融合。

2　基本知识

本节主要阐述了一些基本定义：k-匿名、准标示符、TDS 以及融合条件。

2.1　k-匿名

垂直分区数据表的属性大致分 3 类：1）能够直接标识具体个体的信息，如名字、身份证号、SSN 等，这类属性在融合表发布时直接被删除；2）准标示符属性（QI），多张表中不同属性融合在一起可以推断出具体的个体信息，如属性组 race，birth，gender；3）敏感属性，包含个人隐私信息，如个人薪水、身体状况等。

定义 1（k-匿名）　存在数据表 $T(T_1,T_2,T_3,\cdots,T_n)$，$QI$ 代表 T 上的准标示符，$T(QI)=\{T_1(QI),T_2(QI),\cdots,T_m(QI)\}$ 代表 T 在 QI 上投影的取值，当且仅当 $T(QI)$ 上每个等价组至少存在 $k-1$ 个相同的值时，数据表 T 满足 k-匿名。

定义 2（准标示符）　存在数据表 $T(T_1,T_2,T_3,\cdots,T_n)$，存在 $QI(T_i,T_j,\cdots,T_m)$ 的属性值仅此推断出一个具体的个体，其中 $T_i\neq T_j\neq\cdots\neq T_m$，$i,j,\cdots,m\in(1,2,\cdots,n)$，则 QI 为准标示符。

2.2　匿名条件

这里阐述了多源数据融合需要达到的隐私条件。

定义 3（匿名条件）　考虑融合后数据表 T 中存在 P 个准标示符 $QID_1,QID_2,\cdots,QID_p$，代表准标示符 QID_j 在 T 上取 qid_j 的记录数；$A(QID_j)$ 代表准标示符在 QID_j 上记录数最少的个数，融合表 T 满足匿名条件当且仅当 $\forall A(QID_i)(1\leqslant i\leqslant p)$ 不小于 k。

定义 3 表明当融合表存在多个准标示符时，为了达到匿名条件，需要对不同的准标示符设定不同的 k 值，只有在每个准标示符上达到相应的隐私效果才能决定整个数据表的隐私效果。同时文献[4]特别指出，如果存在多个准标示符 QID_1，$QID_2,\cdots,QID_p$，以及存在 QID_i 是 QID 的子集且 $k_i\leqslant k_j$，那么 $\langle QID_{j,k_j}\rangle$ 覆盖 $\langle QID_{i,k_i}\rangle$，因为当 QID_j 满足匿名条件时，QID_i 也会满足匿名条件。

例如，$QID_1=\langle\{Sex,job\},4\rangle$ 代表融合表在匿名条件 $\langle Sex,Job\rangle$ 上每个等价组的取值至少有 4 条记录，则表 1 中 $\langle Male,Lawyer\rangle$，$\langle Male,Accountant\rangle$，$\langle Female,Accountant\rangle$，$\langle Female,Lawyer\rangle$，$\langle Male,Janitor\rangle$ 不能满足匿名要求，因此表 1 不满足匿名条件。

表 1　融合数据表

共有属性		TA	TB	
ID	Class	Sex	Job	Salary(K)
1—3	0Y3N	Male	Janitor	30
4—7	0Y4N	Male	Mover	32
8—12	2Y3N	Male	Carpenter	35
13—16	3Y1N	Female	Technician	37
17—22	4Y2N	Female	Manager	42
23—25	3Y0N	Female	Manager	44
26—28	3Y0N	Male	Accountant	44
29—31	3Y0N	Female	Accountant	44
32—33	2Y0N	Male	Lawyer	44
34	1Y0N	Female	Lawyer	44

2.3　TDS 思想

文献[5]描述了 TDS 的大致思想：首先将所有属性值泛化到属性分类树[6]的最顶层，然后每次对候选者进行相应的打分操作，找到分数最高的候选者，依据属性分类树进行向下特殊化操作，依次迭代，直到不满足 k-匿名条件时结束。下文提出的多源数据融合算法通过借鉴此思想实现了多源数据的融合，达到了 k-匿名的隐私保护。

3　层次化模型

本节主要阐述了一种层次化模型，包含数据提供者的重要程度、数据属性的重要度以及属性值的敏感度。

3.1　层次化模型

层次化模型在多源数据融合时主要分为 3 层：第一层，数据提供者的重要程度不同，由于每个数据源提供完整数据的部分数据，这些提供商来自不同的机构，不同机构的保密程度不同，因此数据提供者的重要程度不尽相同。第二层，不同数据表中属性的重要程度不同，虽然这些属性来自同一个数据源，但是不同属性要求的保密程度是不同的，例如 Salary 属性相对于 Job 属性要求的保密程度更高，那么它的隐私保密程度需被设置更高。第三层，属性值的敏感程度不同，例通过如 Salary 属性中工资＞40K 的属性值很容易推断出个人敏感信息，因此其保密程度需要被设置得更高。

根据上文描述，图 1 给出了表 1 的层次化模型。首先，对 TA 和 TB 数据提供方的重要程度进行相应的设置；然后，对数据表中的不同属性依据需要达到的隐私程度进行不同的设置；最后根据敏感性对属性中的不同属性值进行不同的设置，完成整个层次化模型的隐私程度的设置。

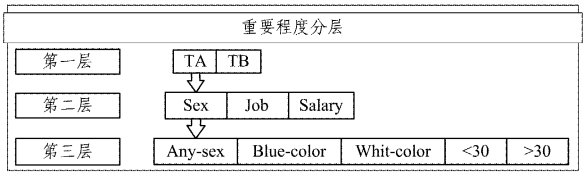


图 1　层次化模型

3.2　重要度

重要度表现为数据提供方需要保护的隐私程度。对于数据源而言，重要度体现了数据提供方是否来自保密机构，或者若数据相对于整体数据而言更加有用，则数据需要被保护的程度等级更高。对于数据属性而言，重要度反映了同一张表

中不同属性的相对数据有用的程度以及需要被保护的程度。数据属性反映的敏感信息越多，那么属性相对越重要，隐私保护程度越高。对于敏感值而言，重要度体现了不同敏感值需要达到的隐私保护水平，例如 Salary 属性中，大于 30K 的敏感值相对比较少，推断出个人信息的可能性比较大，因此其需要的隐私保护水平更高。

3.3 重要度划分

重要度主要通过敏感度因子决定，层次模型不同，层次敏感度因子的决定方式也不尽相同，下文阐述了不同层敏感程度的划分问题。

3.3.1 第一层敏感度划分

层次化模型的第一层由数据的提供者组成，数据存储在不同的机构，保密程度也不相同。例如：几家医院之间的垂直数据融合，A 是省级重点医院，B 是普通医院，则 A 的数据信息更加权威，因此 A 需要更高的保密程度。数据提供者的重要度由公共执行机构决定，重要度由敏感度因子(定义 4)确定，比如：医院的重要度由公共卫生局决定，政府机构的重要度由更高层机构决定等。敏感度因子越小，数据提供者的重要度越高，需要的隐私保护效果越强。

定义 4(第一层敏感度因子) 存在 n 个数据提供者，P_1，P_2，…，P_n 相应的敏感度因子为 f_1，f_2，…，f_n，则敏感度因子满足 $\sum_{i=1}^{n} f_i=1$，$0<f_i<1$，隐私程度越高，f_i 越小。

3.3.2 第二层敏感度划分

层次化模型第二层反映了数据属性之间的重要程度。为了满足不同提供者对各自拥有数据的属性进行划分，通过属性敏感度因子确定不同属性的重要程度。数据拥有者对各自属性使用定义 5 和定义 6 进行设置可以满足属性的相对隐私保护，例如：A 数据表拥有 Sex，Salary 属性，使用定义 5 计算出属性的敏感度即运用属性特殊化对结果造成的影响为(0.3，0.7)，由于 Salary 属性的敏感度相对于 Sex 属性更高，因此用定义 6 设置(0.7，0.3)的属性敏感度因子来表示 Salary 属性的更高敏感度，其需要更高的隐私保护。

定义 5(第二层属性敏感度) 存在数据表 $T(Att_1,Att_2,\cdots,Class)$，其中 $Class$ 表示数据表的类属性，Att_i 表示第 i 个基础属性，那么 Att_i 的属性敏感度为 $gain(Att_i)=info(Class)-info_{Atti}(Class)$，$0<gain(Att_i)<1$。

$$info(Class)=-\sum_{i=1}^{m} p_i * \log_2 p_i$$

$$info_{Atti}(Class)=\sum_{j=1}^{v}\frac{|Class_j|}{|Class|}info(Class_j)$$

其中，p_i 表示第 i 个类别在整个训练元组中出现的概率，可以用属于此类别元素的数量除以训练元组元素总数量作为估计；v 表示属性中值的个数。

定义 6(属性敏感度因子) 存在数据表 $T(Att_1,Att_2,\cdots,Att_n)$，相应的属性敏感度为 $gain(Att_1)$，$gain(Att_2)$，…，$gain(Att_n)$，那么属性敏感度因子表示为 $g(Att_i)=1-gain(Att_i)/\sum_{j}^{n}gain(Att_j)$，满足 $\sum_{i}^{n}g(Att_i)=1$，$0<g(Att_i)<1$，属性对类属性造成的影响程度越低，相应的属性敏感度因子值越小。

3.3.3 第三层敏感度划分

层次化模型第三层反映了属性中属性值的敏感度问题，对于一些相对敏感的属性中的敏感值，如 health condition 属性中的 cancer，HIV 值相对于 flu 值更敏感，那么这些敏感值需要得到相对的隐私保护。表 2 列出了敏感程度及对应的敏感值范围。

表 2 敏感值设定表

ID	敏感程度	敏感值
1	最严重	1
2	较严重	2
3	轻微严重	3
4	中等	4
5	普通	5

根据表 2 中敏感值的设定，对属性中的敏感值进行相应的设置以满足敏感值的隐私保护。属性分为分类属性和连续属性，分类属性指可以根据敏感值的要求以及属性分类树进行设置；连续属性指可以利用文献[3]的方法对区间进行划分，然后依据属性值的敏感要求对区间进行设置。属性值敏感度因子定义如下。

定义 7(第三层敏感度因子) 存在数据表 $T(Att_1,Att_2,\cdots,Att_n)$，其中属性 Att_i 有值$(v_1,v_2,\cdots,v_j)$以及相应的属性分类树 $Tree_{Att_i}$，对应的敏感值为$(value_1,value_2,\cdots,value_j)$，属性分类树中支点的敏感度因子为相应的叶子敏感值之和，且敏感度因子的值越大表示该值的敏感性越低；敏感度因子的值越小表明该值的敏感性越高。

Job 属性中不同属性值依据敏感要求进行了不同的设置，如图 2 所示。属性分类树的上一层的泛化节点的敏感值由子节点的敏感值进行相加。敏感值越大，其敏感程度越低。

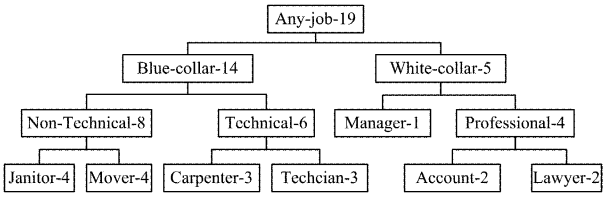


图 2 Job 属性中敏感值的设定

4 多源数据融合算法

4.1 问题定义

对多个提供者的多张数据表进行安全融合，每张数据表代表完整数据的部分属性，由共同的 ID 进行连接，由于提供者的安全系数不同，利用公共权威机构先对提供者进行重要度划分，然后每个提供者对各自属性以及敏感值进行安全设置，利用准标示符以及 k 值决定融合达到的匿名程度，同时保证融合后的数据满足两个条件：1)数据是有效的，可以实现数据挖掘操作；2)数据满足数据拥有者的安全设置。

4.2 算法思想

数据表首先将数据泛化到分类树的最顶层，然后利用 ID 进行数据连接，形成一个融合数据表 Tg，同时指出初始候选者。每一轮迭代时对候选者进行相应的打分操作，然后与其他候选者的分数进行比较以确定最终候选者；同时更新候选者，利用最终候选者使融合表向下进行特殊化，并且指示其他

方进行相应更新，整个过程的终止条件为：没有有效的候选者。

4.3 融合算法

采用上述算法思想设计了多源数据融合算法，算法过程如下。

输入：属性分类树 AttrTree，TA，$\cup Cut_i$，k

输出：匿名融合表 Tg

过程：

1. 初始化数据进行融合得到泛化融合表 Tg；
2. 对本地候选者打分 Score(v)，v 是本地候选者；
3. while $\cup Cut_i$ 存在有效候选者 do
4. 　找到本地分值最高的候选者 HighScore(V_A)；
5. 　if 本地存在有效的候选者 then
6. 　与其他的方候选者进行比较找到 winner；
7. 　else
8. 　没有有效候选者；
9. 　end if
10. 　if winner 是本地候选者 then
11. 　v→child(v)更新$\cup Cut_i$，Tg；
12. 　使用指令(id，c)发送给其他数据提供者来更新它们的 $\cup Cut_i$，Tg；
13. 　else
14. 　等待通知指令，更新$\cup Cut_i$，Tg；
15. 　end if
16. end while
17. 输出 Tg；

其中，算法的一些符号解释如下。

$\cup Cut_i$ 表示候选者的集合，由多源数据表中的候选者融合而成，例如表 1 中的 $T_A(Sex)$，$T_B(Salary, Job)$。数据表首先进行泛化处理，*Sex* 属性只有 *any_sex* 值，(*Salary*，*Job*)属性只有(18－99，*any_job*)值，假设第一轮选择了 *any_job*→{*professional*，*artist*}，则 $\cup Cut_i$ 更新为 $\cup Cut_i$＝{*any_sex*，*professional*，*artist*，18－99}。

HighScore(*VA*)是选择出的本地候选者，主要是对本地存在的候选者进行各自的打分操作，然后通过简单的 max 函数进行比较来确定分数最高的候选者，如 TB 中存在{*professional*，*artist*，18－99}候选者。首先对每个候选者进行打分，然后用 max(*score*(*professional*)，*score*(*artist*)，*score*(18－99))来确定分值最高的候选者，但是在与其他数据表候选者进行比较时使用安全最大值协议(secure max protocol)以避免候选者真实分数的隐私泄露。

当本地候选者在所有的候选者中胜出时，通过候选者所在的属性分类树 *AttrTree* 和本地原始数据表 TA 按照 *id* 对本地的融合表 *Tg* 进行向下的特殊化更新数据处理。当本地的融合表更新完成后，发送(*id*，*v*)更新指令给其他数据提供者以更新相应的 *Tg*。这里，*v* 为候选者向下特殊化的值。例如在{*white-color*，*blue-color*}中候选者选择 *blue-color*→{*non-technica*，*technical*}进行向下特殊化操作，根据 TB 和 Job 属性分类树来确定对应的 *id*，然后在 B 的 *Tg* 中进行向下特殊化操作，B 更新完成后再将对应的(*id*，*v*)发送给 A，更新 A 中的 *Tg*，完成指令工作。

算法描述：1)算法执行前，每一方的数据提供者先对数据进行重要度设置，然后通过文献[8]的方法找到准标示符，并且确定匿名 *k* 值；2)每一方的数据先泛化为属性分类树的最顶层值，然后通过共同的 ID 融合形成一个泛化融合表 *Tg*；3)每一方拥有的候选者进行打分并且找到本地分值最高的候选者，与其他方的候选者进行分数比较，确定 *winner*；4)利用优胜候选者以及分类树对 *Tg* 进行特殊化操作，同时更新 $\cup Cut_i$，指示其他方进行相应的更新。

从 $\cup Cut_i$ 选择分支的核心环节对候选者进行打分操作。打分函数 *Score* 的公式[9]如下：

$$Score(v)=GainRadio(v)=\frac{Gain(v)}{SplitInfo(v)}$$

由于本文主要对敏感值进行隐私保护，因此在计算分数时需要加入层次化敏感度，得到新的计算分数公式：

$$newScore(v)=GainRadio(v)\times num(1)\times num(2)\times num(3)$$

其中，*num*(1)代表该候选者所属的提供者的第一层重要度因子；*num*(2)代表该候选者所属的属性的重要度因子；*num*(3)代表该候选者所在的分类树的重要度(从分类树顶层算起)。

例如：表 1 中 TA 拥有(*ID*，*Class*，*Sex*)属性，TB 拥有(*ID*，*Class*，*Job*，*Salary*)属性，A，B 同时执行上述算法。首先，A 和 B 先确定匿名条件〈QID_1＝{(*Sex*，*Job*)，4}，QID_2＝{(*Sex*，*Salary*)，5}〉，初始化数据表得到值为{*any_sex*，*any_job*，1－99}的数据表，以及{*any_sex*，*any_job*，1－99}候选者，A 对候选者 *any_sex* 打分并与 B 中的候选者采用安全最大值进行比较以找到 *winner*，假设 *winner* 是 B 中的 *any_job*，则利用分类树 *any_job*→{*blue-collor*，*white-collor*}更新融合表及利用候选者〈*any_sex*，*blue-collor*，*white-collor*，1－99〉指令(*id*，*v*)更新 A 中的数据，判断是否满足匿名条件。依次迭代执行算法，直至没有有效的候选者。

4.4 正确性与复杂性分析

正确性：1)对于信息需要[3]而言，融合后的数据表是有用的，可以实现数据挖掘等操作任务；2)对于隐私而言，算法实现过程中首先需要比较打分函数为了不透露具体分数，采用安全多方最大值协议进行比较，不会透露具体的分数信息；3)指令(*id*，*v*)中 *id* 代表标号，*v* 的值相对泛化，不会违反隐私需要[3]的条件；4)在打分操作时利用新的打分函数可以将敏感度问题联系在一起，一定程度上保护了敏感数据。

复杂性：算法每一轮迭代主要表现在以下 4 个方面：1)首先对每一个候选者进行打分，比较并找出本地分数最高的候选者；2)与其他方候选者依次进行比较以找到每一轮中分数最高的候选者；3)找到候选者时需要依次访问本地数据并且更新，然后指示其他方进行相应的指令更新；4)统计记录数目并判断是否满足匿名条件。其中，1)的时间复杂度为 O(|*T*|)，|*T*|代表候选者的个数；在 2)中由于采用了安全最大值比较协议，时间复杂度为 O(|*N*|)，|*N*|代表实体个数，因此完成 1)和 2)的时间复杂度为 O(|*NT*|)；在 3)中更新融合表数据时需要更新每一条记录，同时需要指示其他方做出相应的更新，时间复杂度为 O(*n*|(*N*－1)|)，*n* 代表融合表的记录数；在 4)中对融合表进行匿名条件判断时，需要找到准标示符 *a*(*qid*)，时间复杂度为 O(n^2)。

5 实验与分析

本文算法由 Java 语言和 Weka 开源软件实现。本实验硬件环境为 Windows10 i3-3220CUP 3.30GHz 处理器，内存为 8G。操作系统平台为 Windows 10。

5.1 实验数据集和参数设置

为了体现本文模型的特点和作用，实验数据集和文献[3]的数据集保持一致，利于比较分析。实验数据集使用 UCI Adult 数据集(包含 data 和 test 数据集)，该数据集有 14 个基本属性以及一个 class 分类属性。

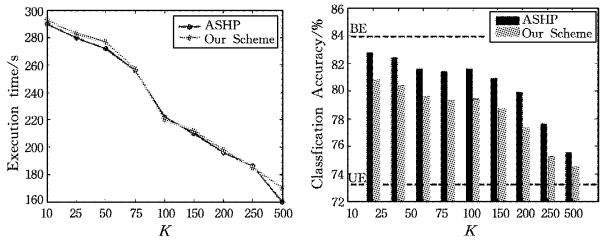
本实验将 Adult 数据集中前 7 个属性(age，Work-class，Fnlwgt，education，Education-num，Marital-status，occupation)作为数据表 1，将后 7 个属性(relationship，race，sex，Capital-gain，Capital-loss，Hours-per-week，Native-country)作为数据表 2 进行实验数据融合。实验数据的准标示符为：age，work-class，marital-status，race，sex，native-country。

图 3 中第一层敏感度因子设置为(0.5，0.5)，在数据表 1 中第二层属性敏感度因子通过定义 5 和定义 6 得到：education (0.1)，occupation(0.1)。其他 5 个属性敏感度值大约为 0.16；数据表 2 中第二层属性敏感度因子为 relationship (0.1)，race (0.1)，native-country (0.1)，其他属性敏感度值为 0.175；第三层 education，occupation，relationship，race，native-country 中属性值敏感度因子设为 3，其他属性值敏感度因子设为 4。

图 4 在图 3 的基础上对敏感度因子进行改进，将第三层 education，occupation，relationship，race，native-country 中的属性值敏感度因子设为 1，其他属性值敏感度因子设为 5。

5.2 实验数据分析

图 3 和图 4 反映了匿名程度 k 与融合花费的时间以及融合后数据的准确度的情况。



(a) $|Q|=6$，$|D|=45000$　　(b) $|QI|=6$，$|D|=45000$

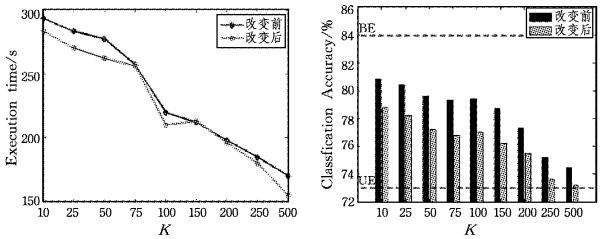
图 3 时间花费和分类精确度

图 3(a)反映了所提算法与文献[3]中 ASHP 算法在融合数据时的时间花费情况。从图 3 中可以看出，所提算法在融合数据的时间花费上与 ASHP 算法相差不大。随着匿名 k 值的增大，算法的时间花费减少，匿名程度相对较弱。当 k 值在 75～100 内时融合时间花费下降的坡度最大，主要原因是数据表中大部分数据值在向下特殊化时集中在该区间内变化，时间花费在整体上满足数据融合时的特殊化趋势。

图 3(b)示出了数据融合后对 *Class* 属性的分类精确度的对比，其中 BE 线代表原始数据融合后的分类精确度，UE 线代表融合数据的有用性。首先，从整体上可以看出本文算法最终所得的融合数据是有用的，数据精确度大于 UE 线；且本文算法由于在数据上对数据值进行了相应的敏感度设置，对数据值的隐私保护相对有所提高，因此在数据精确度上比文献[3]的 ASHP 算法更低。其次，随着 k 值的增加，数据的隐私保护程度降低，导致了所提算法的精确度降低，数据的有用性降低，但是当 k 值为 75 时，数据精确度出现了小幅上升的趋势。因为当 k 值为 75 时，Class 属性值的数据融合更充分，并且没有加入一些其他抽象数据值，所以精确度表现出上升的趋势。

图 4(a)反映了敏感度因子改变前后数据融合所花费的时间对比。从图中总体趋势来看，随着 k 值的增加，时间花费都呈现出降低趋势，且没有什么时间花费显著的差异。但是改变敏感度因子后实验融合的时间花费相对较少，由于降低了敏感属性值的敏感度因子，导致该候选者不易被选中，在相对不敏感的属性上进行多次特殊化操作致使实验更快结束，因此实验的时间花费得到了降低。

图 4(b)反映了敏感度因子改变前后数据的分类精确度对比情况。从图中可以看出两种情况下的数据精确度在 UE 线之上，融合后的数据是有效的。但是随着 k 值的增加，分类的精确度出现了明显降低，并且改变敏感度因子后的分类精确度下降得更加明显，这说明改变敏感度因子后数据在融合过程中的有用性得到了降低。



(a)　　(b)

图 4 敏感度因子改变前后的时间花费和分类精确度

由数据挖掘分类器规则可知，数据的分类精确度与准标示符属性、敏感属性以及敏感属性的敏感值都有重要的联系。第二层属性敏感度因子就是根据不同属性对结果造成的影响程度向下进行划分而得到的，第三层敏感度因子的设置将相对敏感的属性值进行不同程度的设置，使相对敏感的值的敏感度因子更小。根据上面的设置，每次在对候选者进行打分时，敏感度因子将会降低分值，使敏感属性被选中的概率降低。即使敏感属性被选中而进行向下特殊化，但由于属性树中相对敏感的候选者的敏感度因子更小，因此被选中的概率依旧会被降低。本文的模型通过设置敏感度因子降低了敏感值候选者的分数，从而降低了敏感值候选者被选中的概率，进而达到了保护敏感值的层次化隐私的目的。

结束语 本文提出了一种面向敏感值的层次化模型，进而提出了一种面向敏感值的层次化多源数据融合隐私保护算法，实验表明本文算法既能实现数据的安全融合且融合后的数据能保持有效性，又能给敏感数据提供相对有效的保护。但是 k-匿名隐私模型不能阻止背景知识的攻击。近年来差分隐私保护模型得到了学术界的关注，它可以提供有效的隐私保护，而不关注攻击者是否拥有背景知识，所以我们下一步将研究面向敏感值的差分隐私安全数据融合方法。

参 考 文 献

[1] SAMURAI P，SWEENEY L. Generalizing data to provide anonymity when disclosing information [C]// The 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems. IEEE Press，1998：188.

[2] JIANG W, CLIFTON C. A secure distributed framework for achieving k-anonymity [J]. Journal of Very Large Data Bases J, 2006, 15(4): 316-333.

[3] MOHAMMED N, FUNG B C M, DEBBABI M, et al. Anonymity meets game theory: secure data integration with malicious participants [J]. Journal of Very Large Data Bases, 2011, 20(4): 567-588.

[4] SONG J L, HUANG L M, LIU G H. Algorithm for Finding Quasi-identifiers in the k-anonymity Method [J]. Journal of Chinese Mini-Micro Computer Systems, 2008, 29(9): 1688-1693. (in Chinese)
宋金玲，黄立明，刘国华. k-匿名方法中准标示符的求解算法[J]. 小型微型计算机系统，2008，29(9)：1688-1693.

[5] FUNG B C M, WANG K, YU P S, et al. Anonymizing Classification Data for Privacy Preservation [J]. IEEE Transaction on Data Engineering, 2007, 19(5): 711-725.

[6] WANG P S, MA Q J. Research on k-anonymity algorithm for privacy preservation [J]. Computer Engineering and Applications, 2011, 47(28): 117-119. (in Chinese)
王平水，马钦娟. 隐私保护 k-匿名算法研究[J]. 计算机工程与应用，2011，47(28)：117-119.

[7] YANG X C, WANG Y Z, WANG B, et al. Privacy Preserving Approaches for Multiple Sensitive Attributes in Data Publishing [J]. Chinese Journal of Computers, 2008, 31(4): 574-587. (in Chinese)
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.

[8] MACHANAVAJJHALA A, GEHRKE J, KIFER D. l-diversity: privacy beyond anonymity [C]// The 22nd International Conference on Data Engineering. New York, ACM Press, 2006: 24-35.

[9] TRAIAN T M, BINDU V. Privacy protection: p-sensitive k-anonymity property [C]// The 22nd International Conference on Data Engineering. New York: ACM Press, 2006: 94.

[10] XIAO X K, TAO Y E. Personalized privacy preservation [C]// Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. Chicago, Illinois, USA: ACM Press, 2006: 229-240.

[11] SWEENEY L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal of Uncertainty Fuzziness, 2012, 10(5): 571-588.

[12] JURCZYK P, XIONG L. Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers [C]// LFIP Wag 11.3 Working Conference on Data& Applications Security XXiii. 2009: 191-207.

[13] TAKENOUCHI T, KAWAMURA T, OSUNA A. Distributed Anonymization Method with Hiding the Presence of Individuals [J]. Ibices Transactions on Information & Systems, 2013, 96(3): 596-610.

[14] HAN J M, YU J, YU H Q, et al. Individuation Privacy Preservation Oriented to Sensitive Values[J]. Acta Electronica Sinica, 2010, 38(7): 1723-1728. (in Chinese)
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.

(上接第 141 页)

[14] YANG Y H, HUANG H Z, SHEN Q N, et al. Intrusion detection based on incremental GHSOM neural network model[J]. Journal of Computer Science, 2014(5): 1216-1224. (in Chinese)
杨雅辉，黄海珍，沈晴霓，等. 基于增量式 GHSOM 神经网络模型的入侵检测研究[J]. 计算机学报，2014(5)：1216-1224.

[15] CHEN X, TAO J, et al. Intrusion detection algorithm based on Bias game model in wireless networks [J]. Journal of Communication, 2010, 31(2): 107-112(in Chinese)
陈行，陶军，等. 无线网络中基于贝叶斯博弈模型的入侵检测算法研究[J]. 通信学报，2010，31(2)：107-112.

[16] WANG H, CHEN H Y, LIU S F, et al. Intrusion detection system based on improved naive Bayes algorithm [J]. Computer Science, 2014, 41(4): 111-115. (in Chinese)
王辉，陈泓予，刘淑芬，等. 基于改进朴素贝叶斯算法的入侵检测系统[J]. 计算机科学，2014，41(4)：111-115.

[17] DUAN X T, JIA C F, LIU C B. Detection method of hierarchical hidden Markov model and variable length semantic model based on Intrusion [J]. Journal of Communication, 2010, 31(3): 109-114. (in Chinese)
段雪涛，贾春福，刘春波. 基于层次隐马尔科夫模型和变长语义模式的入侵检测方法[J]. 通信学报，2010，31(3)：109-114.

[18] ZHANG Y, TAN X B, CUI X L, et al. Network security situation awareness method based on Markov game model [J]. Chinese Journal of Software, 2011, 22(3): 495-508. (in Chinese)
张勇，谭小彬，崔孝林，等. 基于 Markov 博弈模型的网络安全态势感知方法[J]. 软件学报，2011，22(3)：495-508.

[19] XI R R, YUN X C, ZHANG Y Z, et al. An improved quantitative evaluation method of network security situation [J]. Chinese Journal of Computers, 2015, 38(4): 749-758. (in Chinese)
席荣荣，云晓春，张永铮，等. 一种改进的网络安全态势量化评估方法[J]. 计算机学报，2015，38(4)：749-758.

[20] FENG X W, WANG D X, HUANG M H, et al. A method of causal knowledge mining based on Markov [J]. Computer Research and Development, 2014, 51(11): 2493-2504. (in Chinese)
冯学伟，王东霞，黄敏桓，等. 一种基于马尔科夫性质的因果知识挖掘方法[J]. 计算机研究与发展，2014，51(11)：2493-2504.

[21] DENG X Y, DENG Y, ZHANG Y J, et al. A Markov reliability model and application [J]. Journal of Automation, 2012, 38(4): 666-672. (in Chinese)
邓鑫洋，邓勇，章雅娟，等. 一种信度马尔科夫模型及应用[J]. 自动化学报，2012，38(4)：666-672.

[22] LI F W, DENG W, ZHU J. A network security situation prediction mechanism based on complex network [J]. Computer Application Research, 2015, 32(4): 1141-1144. (in Chinese)
李方伟，邓武，朱江. 一种基于复杂网络的网络安全态势预测机制[J]. 计算机应用研究，2015，32(4)：1141-1144.

[23] DONG J. Research on improved HMM network security risk assessment method [D]. Wuhan: Huazhong University of Science and Technology, 2008. (in Chinese)
董静. 改进的 HMM 网络安全风险评估方法研究[D]. 武汉：华中科技大学，2008.

[24] LEI J. Research on network security threat and situation assessment [D]. Wuhan: Huazhong University of Science and Technology, 2008. (in Chinese)
雷杰. 网络安全威胁与态势评估方法研究[D]. 武汉：华中科技大学，2008.