# 基于 $k$-匿名的多源数据融合算法研究

杨月平,王 箭
(南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

**摘 要**:数据在当今的网络环境下变得越来越重要,融合技术能够使不同数据提供者有效地融合他们的数据,并且提供给顾客可定制且有效的服务。数据融合技术通常采用每轮自顶向下选择候选者,并进行数据更新的方法,而这种方法随着数据量的增加使得数据融合的时间花费巨大,难以满足数据融合的时间需求。为了减少融合数据过程中的花费,提高多源数据融合的精度,结合自顶向下分类树算法 TDS,属性分类树,提出了一种基于 $k$-匿名的多源数据融合算法。利用 GUI 的 Adult 数据集进行仿真实验,并比较了数据融合的复杂度以及融合精度的差异。实验结果表明,所提出的基于 $k$-匿名多源数据融合算法融合过程时间花费更少,可以达到理想的数据融合精度,同时还实现了多源数据的融合。

**关键词**:数据融合; $k$-匿名;自顶向下分类树;属性分类树
**中图分类号**:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2017)05-0102-06
doi:10.3969/j.issn.1673-629X.2017.05.022

# Research on Data Fusion Algorithm for Multi-party Based on $k$-anonymity

YANG Yue-ping, WANG Jian
(Institute of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

**Abstract**: In today's network environment, data has become more and more important. Data integration technology can make the effective data integration for different data providers, and provide customized service for the customers. Data fusion technology usually adopts the top-down to choose candidates for updating data in each round, and with the increase of amount of data, this kind of method costs a lot of time, which is difficult to meet the time requirements of data fusion. In order to reduce the cost in the process of data fusion and improve the accuracy of data integration for multi-party, a multi-party data fusion algorithm based on $k$-anonymous combining with the top-to-down TDS algorithm and the attribute classification tree has been proposed. Simulation experiments have been conducted with Adult set of GUI as well as comparison of accuracy of data fusion with complexity. The experimental results show that the proposed algorithm has taken less time and effectively achieve ideal accuracy of data fusion.

**Key words**: data integration; $k$-anonymous; top-to-down TDS; attribute classification tree

## 0 引 言

随着大数据时代的来临,大量的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。通常这些数据拥有者想要融合它们的数据,从而进行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。

这个研究问题是在一个瑞士的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),这些公司想要融合他们的数据提供更好的决策。例如:银行是否要贷款给公司。但是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,或者融合后的数据可以推断出某个具体的个体信息,这个问题对于 A 和 B 都是不愿看

**收稿日期**:2016-05-31 **修回日期**:2016-09-09 **网络出版时间**:2017-03-13
**基金项目**:中国博士后科学基金(2014M561644);江苏省博士后科学基金(1402034C)
**作者简介**:杨月平(1992-),男,硕士研究生,研究方向为隐私保护;王 箭,教授,研究方向为信息安全。

到的。

目前已经提出了一些框架来融合数据表。2006 年,Jiang 和 Clifon 提出了 DkA(安全分布式框架)[1] 来融合两个不同的数据表,形成一个满足 $k$-匿名[2]的融合表。但是 DkA 框架存在两个缺点:一是不适合大规模数据集,DkA 随着数据集的增加,其加密花费随之增加;二是 DkA 仅适用于两个数据表之间的融合。为了克服上述问题,2011 年 Fung 等提出了一种安全融合两方数据的 $k$-匿名框架,适合两方大规模的数据融合,但是该框架存在一个问题,每次进行特殊化时都要进行两方安全最大值计算,整个算法完成时间花费较大。

针对上述问题,提出了一种基于 $k$-匿名的多源数据融合算法[3]。该算法减少了数据融合的时间花费,得到的融合数据具有数据挖掘的价值,同时满足多源数据之间的融合。

## 1　融合模型框架

数据融合主要由多个数据拥有者组成(在半诚实模型下,数据拥有者想要在融合过程中尽可能地得到其他信息),数据流程大概分为下面几步:首先,多个数据拥有者将各自的数据表进行融合。其次,融合后的数据需要进行匿名化处理,在匿名化过程中数据拥有者不能得到除自身之外的其他敏感信息。最后,匿名化的融合表可以进行数据分类分析(比如:进行分类),或者直接发送给数据的接收者,这里数据的接收者可以是自身或者其他接收者。数据融合模型如图 1 所示。

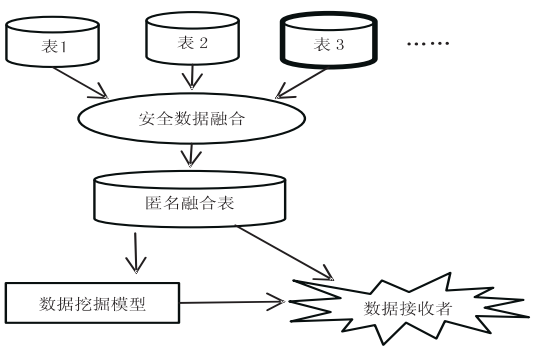


图 1　数据融合模型框架图

图 1 阐述了数据融合的基本框架,数据拥有者有两个关心的问题:

(1)直接将数据表融合在一起会暴露表中敏感信息给对方;

(2)融合后的表可能推断出某个个体的具体信息,存在着隐私泄露的风险。

例如,表 1 展示了一个融合的原始数据表 $D$,Ta(信用卡公司)表有属性(ID,Class,Sex…),Tb(贷款公司)表有属性(ID,Class,Job…),Tc(银行)有属性(ID,Class,Salary…),这些表拥有共同的 ID,Class 属性,Class 的值代表是否已经被贷款,Y 表示已经贷款,N 表示拒绝贷款。

表 1　原始数据表

| 共有属性 | | Ta | Tb | Tc |
|---|---|---|---|---|
| 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 | Account | 44 |
| 29-31 | 3Y0N | Female | Account | 44 |
| 32-33 | 2Y0N | Male | Lawyer | 44 |
| 34 | 1Y0N | Female | Lawyer | 44 |

从表中第三行可以看出,共有 5 条记录<Male,Carpenter,35>,其中 2 条 Y 表示被贷款,3 条 N 表示拒绝贷款。融合后的数据表,Ta 可以得到 Tb 中的 Job 信息,可以得到 Tc 中的 Salary 敏感信息。融合后的数据表(Female,Lawyer)是唯一的,通过联合攻击可以推断出某个个体的具体信息。这样的融合数据表存在着隐私泄露的风险,可以采取泛化处理,将 Account 和 Lawyer 泛化为 Professional,这样(Female,Lawyer)将不是唯一的,不能通过联合攻击推断出个体信息。

## 2　隐私和信息需要

这节主要阐述了多源 $k$-匿名融合后所要满足的需要条件:隐私需要和信息需要。

### 2.1　隐私需要

融合后的数据需要满足 $k$-匿名:融合后的匿名数据表的数据中存在一定数量(至少为 $k$)的在准标识符上不可区分的记录,使攻击者不能判别出隐私信息所属的具体个体,从而保护了个人隐私。$k$-匿名通过参数 $k$ 指定用户可承受的最大信息泄露风险,$k$ 值越大,代表隐私保护程度越高,相反 $k$ 值越小,数据值接近真实值,隐私保护程度越低。$k$-匿名化在一定程度上保护了个人隐私,但同时会降低数据的可用性。因此,$k$-匿名化的研究工作主要集中在保护私有信息的同时提高数据的可用性。

此外,在数据匿名化的过程中,数据提供者 A 不能得到比最后匿名化融合数据表更加详细的信息。例如:表 1 中的 Account,Lawyer 是比 Professional 更加详细的信息,那么在匿名化的过程中 A 不可以确定 Job 的值是 Account 或者 Lawyer。

### 2.2 信息需要

多源数据融合发布的匿名数据是有效且有用的，可以用来进行数据挖掘操作，比如可以用来进行分类分析。那么为什么不直接的发布分类器或者统计数据给数据接收者？在商业环境的项目合作中，数据接收者（信用卡公司）想要接收的是融合的金融数据，可用这些有用的数据在其它工作中进行分析，挖掘相应的数据结果，而不是想得到具体的统计数据或者某一分类器，这些数据接收者想要对数据进行更加灵活的操作。为不同的项目需要向数据提供商的 IT 部门请求不同参数的分类器在显示应用场景下是不切实际的。

## 3 问题定义

这节定义了匿名需要条件，介绍了属性分类树，阐述了安全数据融合问题，陈述了匿名化技术。

### 3.1 匿名需要

假设有一张融合表 $T(\mathrm{ID}, \mathrm{Att}_1, \mathrm{Att}_2, \cdots, \mathrm{Class})$。其中，ID 代表一个个体的标识符（如 SSN），这个标识符在数据发布之前已经被移除；$\mathrm{Att}_i$ 代表个体的属性，包括连续属性和分类属性；Class 代表类标签。数据提供者想要减少隐私泄露的风险，防止联合攻击。联合攻击可能发生在两个属性和多个属性的联合问题上，联合属性可以潜在地确定某个个体的信息称为准标识符 QID（quasi-identifiers）[4-6]，准标识符属性上的值记录数越小，隐私泄露风险越大。问题定义如下：

定义 1（匿名需要）：设在表 $T$ 中有 $p$ 个准标识符，分别为 $\mathrm{QID}_1, \mathrm{QID}_2, \cdots, \mathrm{QID}_p$。$a(\mathrm{qid}_j)$ 代表在标识符 $\mathrm{QID}_j$ 上取 $\mathrm{qid}_j$ 的记录数，$A(\mathrm{qid}_j)$ 表示 $a(\mathrm{qid}_j)$ 中的最小值，$\mathrm{QID}_j$ 的匿名由 $A(\mathrm{qid}_j)$ 决定。融合表 $T$ 满足匿名需要 $\{ < \mathrm{QID}_1, k_1 >, < \mathrm{QID}_2, k_2 >, \cdots, < \mathrm{QID}_p, k_p > \}$，必须使得 $A(\mathrm{QID}_j) \geqslant k_j, 1 \leqslant j \leqslant p$。$k_j$ 的值决定了匿名的程度，$k_j$ 值越大，匿名程度越高，隐私保护程度越高；相反，$k_j$ 值越小，匿名程度越小，隐私保护程度越小。

定义 1：通过多个数据提供者指出 $\mathrm{QID}_i$ 概括了传统的 $k$-匿名方案，文献[7]详细指出了 $\mathrm{QID}_i$ 具体信息。定义 1 同时指出，如果 $\mathrm{QID}_i$ 是 $\mathrm{QID}_j$ 的子集，并且 $k_i \leqslant k_j$，那么 $< \mathrm{QID}_j, k_j >$ 包含 $< \mathrm{QID}_i, k_i >$，在 QID 中可以移除 $< \mathrm{QID}_i, k_i >$。

通过上述定义，$< \mathrm{QID}_1 = \{\mathrm{sex}, \mathrm{job}\}, 4 >$ 指出 sex，job 属性联合的准标识符 qid 的记录数不少于 4，则表 1 中下面的 qid{sex，job} 不满足匿名需要条件：<Male，Janitor>，<Female，Manager>，<Male，Account>，<Female，Account>，<Male，Lawyer>，<Female，Lawyer>。

### 3.2 属性分类树

泛化技术[8]形成属性分类树的核心技术，泛化技术表现为继承或实现关系，具体形式为类与类之间的继承关系，接口与接口之间的继承关系，类对接口的实现关系。泛化是将具有相同特征的值抽取出来泛化为更高层次的值，表现为 $\mathrm{child}(v) \rightarrow v$。图 2 给出了 $\cup \mathrm{Cut}_i = \{\mathrm{sex}, \mathrm{job}, \mathrm{salary}\}$ 中 job 属性分类树。

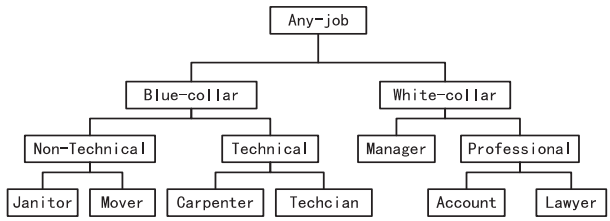


图 2 job 属性分类树

从图中可以看出，Lawyer，Account 泛化为 Professional，在数据融合时每个数据提供者提供私有数据表的属性树。

### 3.3 安全数据融合

多源数据融合[9]首先需要多个数据提供者的多张数据表，数据提供者想要融合这些表同时希望释放最少的信息进行数据融合，融合的数据表具有分类分析的价值。最少信息的程度由联合匿名需要 $\{ < \mathrm{QID}_1, k_1 >, < \mathrm{QID}_2, k_2 >, \cdots, < \mathrm{QID}_p, k_p > \}$ 决定。

定义 2（安全数据融合）：给出多个数据表 $T_1, T_2, \cdots, T_n$，联合的匿名需要条件 $\{ < \mathrm{QID}_1, k_1 >, < \mathrm{QID}_2, k_2 >, \cdots, < \mathrm{QID}_p, k_p > \}$，以及 $\cup \mathrm{QID}_j$ 中属性的分类树，安全数据融合问题是产生一个泛化的融合表 $T$，同时 $T$ 满足下面几个条件：$T$ 要满足联合匿名需要条件；$T$ 中的信息是有用的，可以满足数据挖掘分类分析的需要；在匿名化的过程中，一方不能得到其他方比最终融合信息更加具体的信息。

例如：在融合表 $T$ 中，sex，job 属性有值 Female，Professional，并且 A 知道 Professional 来自 Lawyer，那么条件 3 被违反了，所采用的模型确保数据融合的过程中一方不能得到比最终融合信息更加详细的信息。

### 3.4 匿名技术

可以匿名一个表 $T$ 从表的最顶端值（每个属性分类树的最顶端值）进行特殊化操作，特殊化过程写成 $v \rightarrow \mathrm{child}(v)$。例如：White - collar → {Manager，Professional} 特殊化时根据原始数据表 Tb 中 Job 属性的值泛化表 $T$ 决定特殊化为 Manager 或是 Professional。特殊化时有效指的是特殊化后的融合表满足匿名需要条件。

特殊化的过程可以看作是从最顶端选择分裂值向下分支的过程，$\cup \mathrm{Cut}_i$ 代表要分支的选择路径，其中每个 $\cup \mathrm{Cut}_i$ 代表一种分支。特殊化过程就是依次地选择路径进行向下分支，直到不满足匿名条件或者没有路径分支。

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[10]如下:

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

$Gain(v)$:信息增益定义为原来的信息需求(即仅基于类比例)与新需求(即对 $v$ 划分之后得到的)之间的差,即

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j)$$

$$Gain(v) = Info(D) - Info_A(D)$$

一般说来,对于一个具有多个属性的元组,用一个属性就将它们完全分开几乎不可能,否则的话,树的深度就只能是 2 了。从这里可以看出,一旦选择一个属性 $A$,假设将元组分成了两个部分 $A_1$ 和 $A_2$,由于 $A_1$ 和 $A_2$ 还可以用其他属性接着再分,所以又引出一个新的问题:接下来要选择哪个属性来分类? 对 $D$ 中元组分类所需的期望信息是 $Info(D)$,那么同理,当通过 $A$ 将 $D$ 划分成 $v$ 个子集 $D_j(j=1,2,\cdots,v)$ 之后,要对 $D_j$ 的元组进行分类,需要的期望信息就是 $Info(D_j)$,而一共有 $v$ 个类,所以对 $v$ 个集合再分类,需要的信息就是 $Info_A(D)$ 了。

$SplitInfo(v)$:表示属性 $A$ 将 $D$ 划分为 $v$ 个子类所需的信息,$|D_j|$ 表示划分的子类 $j$ 的记录数。

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * \log_2(\frac{|D_j|}{|D|})$$

使用信息增益有一个缺点,那就是它偏向于具有大量值的属性,信息增益率使用“分裂信息”值将信息增益规范化,解决了这一问题。

对于连续属性,使用信息增益 $Gain(v)$ 确定分界值,由于连续属性分裂只有两种分支,采用信息增益完全可以选择出最好的分支结果。

例如:对于连续属性 Salary,最顶层值(1,99),为了确定分支点,评估 5 个值 30,32,35,37,42 所获得信息增益 $Gain(v)$,经计算在分裂值 37 所获得信息增益最大。则将最顶端值划分为(1,37),[37,99)。

## 4　多源数据融合算法

### 4.1　算法思想

从 TDS[7,11-12]最顶层树[13-14]向下分支的过程中,每次多方之间候选者进行打分比较,然后选出分最高的候选者进行向下分支并且通知其他方进行向下划分,这样的过程花费较大。倘若首先已经选择了分数最高的候选者,对分数最高的候选者向下划分得到新的候选者,再去和其他候选者进行比较,依次向下直到没有候选者或者分数小于其他候选者。更新完该表之后通知其他表进行更新,该过程不需要每次都通知更新,减少了花费情况。

### 4.2　多源融合算法

Fung 提出了 TDS 算法[15-16]泛化一张表 $T$,该算法不满足隐私需要条件(见 2.1 节)。首先,TDS 将各个数据表进行融合,然后对融合后的数据表进行泛化处理满足 $k$-匿名,但是将数据表融合时,数据表中的隐私信息已经泄露,不满足隐私需要。

可以采用 TDS 的泛化思想,在数据表融合前将每个属性泛化为相应属性分类树的最顶层值,然后将数据表进行融合,在每一轮迭代中选择 $\cup$ $Cut_i$ 中 Score 最高的值,依靠相应的属性分类树进行分支。算法结束的条件为 $\cup$ $Cut_i$ 没有有效的候选者,或者说算法结束与当前状态的分支将不满足匿名需要条件(见定义 1)。算法过程如下:

1. 初始化融合表 Tg,属性分类树 AttrTree,Ta, $\cup$ $Cut_i$;
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. 与 $HighScore(V_B)$, $HighScore(V_C)$ 等进行比较;
6. if 本地候选者分最高
7. 更新 $\cup$ $Cut_i$,对新加入的候选者打分 $Score(new_1)\cdots$;
8. while $Score(new_i) \geqslant \max(HighScore(V_B), HighScore(V_C))$ do
9. 更新 $\cup$ $Cut_i$,对新加入的候选者打分 $Score(new_2)\cdots$;
10. end while
11. 确定候选者 winner;
12. end if
13. if winner 是本地候选者;
14. 更新 Tg, $v \rightarrow child(v)$;
15. 使用指令 (id,$c$) 通知 B,C$\cdots$. 更新 Tg;
16. else
17. 等待通知指令, $\cup$ $Cut_i$ 更新 Tg;
18. end if
19. 更新 $\cup$ $Cut_i$,检查有效性;
20. end while
21. 输出 Tg

算法描述如下:

(1)每一方数据初始化时将自己的数据表泛化为最顶层值然后发送给 B,C$\cdots$ 形成融合数据表 Tg,融合的前提可以通过文献[7]中提到的交换加密技术将拥

有相同 ID 值的数据进行融合。A 中拥有私有表 Ta,所有的候选者 $\cup$ $Cut_i$ ,sex 属性分类树 AttrTree。

(2)选择候选者:在每轮迭代中,首先每一方对自己的候选者进行打分,选择本地分数最高的候选者,依次与其他方的候选者进行比较。倘若分数最高的 winner 是本地的候选者,可以根据属性树向下分支更新本地的 Tg 以及本地的 $\cup$ $Cut_i$ ,然后对新的候选者进行打分比较新的候选者是否依然分值最高,直到新的候选者不满足匿名需要条件或者分数不是最高,这一轮迭代结束。

(3)通过指令 (id,$c$) 通知其他方更新 Tg 以及 $\cup$ $Cut_i$ 。

例如:表 1 中,Tb 拥有(ID,Class,Job)属性,融合时,Ta,Tc 将会把自己要融合的数据泛化为最高值(any_sex,any_salary)发给 Tb,这样初始化时 Tb 将会得到(any_sex,any_job,any_salary),并且得到 $\cup$ $Cut_i$ = {any_sex,any_job,any_salary} 的候选者。Tb 对 any_job 打分并且与其他候选者进行比较,若 any_job 分最高,选为候选者,这时通过 job 属性树可以将{blue-collar,white-collar}替换 any_job 候选者,再对新的候选者进行打分判断新的候选者分数是否最高。依次执行下去,每轮迭代时,先特殊化一张表满足条件,再去通知其他表的更新。

### 4.3 正确性与复杂性分析

正确性:

(1)对于信息需要来说,匿名化融合表在满足匿名条件后信息是有用的,相比于 TDS[11] 在算法融合过程中不会透露敏感信息给对方。

(2)对于隐私来说,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,并且打分函数代表的是该属性对分类的增益大小,不会透露具体信息;指令 (id,$c$) 中 id 代表标识,$c$ 的值相对泛化,不会违反隐私需要条件。

复杂性(算法主要花费在四个方面):

(1)该算法首先要对本地的候选者进行打分,选择本地候选者。

(2)与其他候选者进行比较,得到分数最高的候选者并用新的候选者分数进行比较。

(3)更新本地 Tg。

(4)通过指令通知其他方更新。

步骤(1)需要遍历每个候选者,时间复杂度为 $o(|T|)$ 。步骤(2)在比较分数时由于采用的是冒泡法找到最大值,时间复杂度为 $o(|T|)$ ,由于每个候选者的下个分支可能还是候选者,所以时间复杂度为 $o(n|T|)$ ,$n$ 代表向下分支次数。步骤(3)中更新 Tg 依次遍历特殊化,时间复杂度为 $o(|T|)$ 。步骤(4)中更新其他方的表需要依次访问,时间复杂度为 $o(|T|)$,需要更新 $m$ 张表,时间复杂度为 $o(m|T|)$ 。

对于文献[17]提出的算法,由于在每一轮迭代时需要选择候选者然后进行指令更新,花费相当大。文中提出的算法尽量选择出下一轮的候选者然后进行指令更新,避免文献[17]中每一次进行指令更新,花费相对减少。

## 5 实验与分析

实验数据融合算法采用 Java 开发工具多线程机制实现,分类器测试和训练部分使用开源机器学习工具 Weka3.6。实验环境为 Windows10 i3-3220CUP 3.30 GHz,内存 4 G。实验数据集采用 UCI Adult 数据集(包含 data 和 test 数据集)。该数据集有 14 个属性以及 1 个 class 属性(≥50K 和≤50K)。

QID 匿名条件采用每一方提供属性组合形式完成,$k$ 值手动定义。实验 1 主要是测试针对不同的 $k$ 值匿名条件下,在不同的 QID 情况下,完成匿名数据融合的时间花费情况。为了比较提出算法的融合时间问题,首先将 Adult 数据集分成两方数据进行融合,如图 3 所示。然后将 Adult 数据集分成三方进行数据融合,如图 4 所示,分析时间花费问题。$QID_5$ 代表匿名条件总共由 5 个属性决定,相同的 $QID_7$ 代表匿名条件由 7 个属性共同决定。

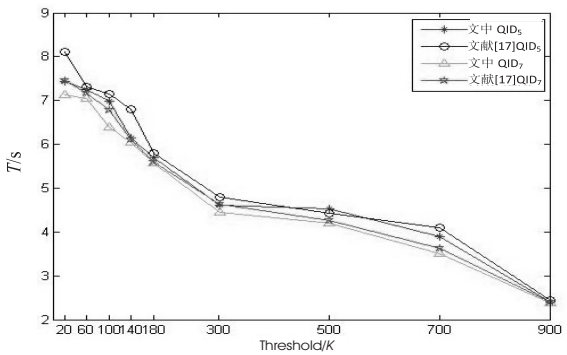


图 3 两方数据融合时间花费图

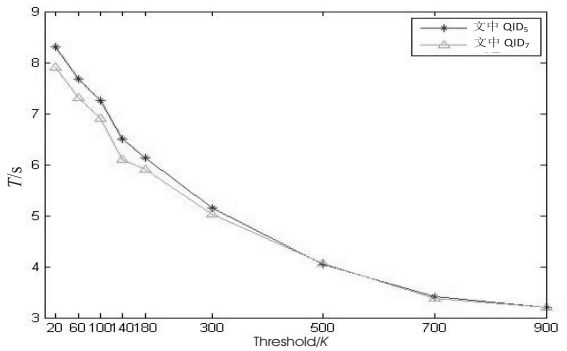


图 4 三方数据融合时间花费图

从图 3 中可以看出,当 QID 的属性增加时,即匿名条件更加严格时,花费的时间减少,这是因为融合的数据在判断匿名条件时更快地满足匿名条件,融合的表更容易生成融合数据。同时可以看出,在相同的 $k$ 值,相同的 QID 情况下,提出算法融合时间花费比文献[17]算法花费更少。

从图 4 中可以看出,当 QID 的属性增加时,花费的时间越来越少,需要特殊化的次数也越来越少;当 $k$ 值增加时,匿名条件更加严格,融合表生成时间花费随之降低。

实验 2 主要是测试实验 1 情况下生成的匿名融合表做 C4.5 分类分析,如图 5 所示。

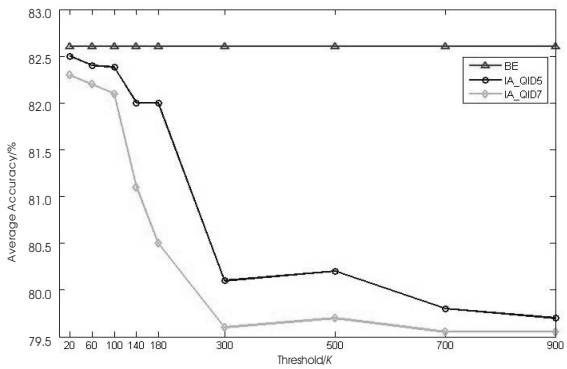


图 5　三方数据融合分类精确度图

图中,BE 代表原始数据分类器的精度, $IA\_QID_5$ 代表在 $QID_5$ 匿名条件下融合数据表的分类精度,同样的 $IA\_QID_7$ 代表在 $QID_7$ 匿名条件下融合数据表的分类精度。可以看出,随着 $k$ 值的增加,分类精度随之降低,数据的有效性越低, $k$ 值越小,分类的精度越高。在某些特殊的情况下, $k$ 值越大,分类精度却增加,出现这样的原因是因为在某些属性层次上 $k$ 值的增加可以降低泛化带来的噪声干扰,使得在一定的 $k$ 值范围内增加了精确度。从实验的精确度可以看出,数据是有效的。

## 6　结束语

为了减少融合数据过程中的花费,提高多源数据融合的精度,在分析现有多源数据融合算法存在不足的基础上,提出了一种基于 $k$ -匿名的多源数据融合算法。与文献[17]算法进行了比较,提出算法采用子节点分数与父节点分数进行比较的方法,减少了数据的更新次数和时间花费。实验结果表明,该算法在融合多源数据过程中花费时间较少,并且匿名后的融合数据是有效的,可以进行分类分析。

**参考文献:**

[1] Wei J, Clifton C. A secure distributed framework for achieving k-anonymity[J]. The VLDB Journal, 2006, 15(4): 316-333.

[2] 岑婷婷,韩建民,王基一,等. 隐私保护中 K-匿名模型的综述[J]. 计算机工程与应用,2008,44(4):130-134.

[3] 吴　艳. 多传感器数据融合算法研究[M]. 西安:西安电子科技大学,2003.

[4] 宋金玲,黄立明,刘国华. k-匿名方法中准标识符的求解算法[J]. 小型微型计算机系统,2008,29(9):1688-1693.

[5] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal on Uncertainty Fuzziness and Knowledge-based System, 2011, 10(5): 571-588.

[6] 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] 刘　明,叶晓俊. 个性化 K-匿名模型[J]. 计算机工程与设计,2008,29(2):282-286.

[8] 吕　品,钟　珞,于文兵,等. MA-Datafly:一种支持多属性泛化的 k-匿名方法[J]. 计算机工程与应用,2013,49(4):138-140.

[9] Dayal U, Hwang H Y. View definition and generalization for database integration in a multidatabase system[J]. IEEE Transactions on Software Engineering, 1984, 10(6): 628-645.

[10] Mohammed N, Fung B C M, Yu P S. Differentially private data release for data mining[C]//International conference on knowledge discovery& data mining. [s. l.]: [s. n.], 2011: 493-501.

[11] Fung B C M, Wang K, Yu P S. Anonymizing classification data for privacy preservation[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(5): 711-725.

[12] 吴步祺,白小明,张　乐. 医疗信息发布中 k-匿名模型的分析与改进[J]. 计算机与现代化,2009(10):182-184.

[13] 晏　华,刘贵松. 采用熵的多维 K-匿名划分方法[J]. 电子科技大学学报,2007,36(6):1228-1231.

[14] 兰丽辉,鞠时光,金　华. 社会网络数据的 k-匿名发布[J]. 计算机科学,2011,38(11):156-160.

[15] Inan A, Kantarcioglu M, Bertino E, et al. A hybrid approach to private record linkage[C]//Proceedings of the international conference on data engineering. [s. l.]: IEEE, 2008.

[16] Zhang N, Zhao W. Distributed privacy preserving information sharing[C]//Proceedings of the VLDB. [s. l.]: [s. n.], 2005: 889-900.

[17] Mohammed N, Fung B C M, Debbabi M. Anonymity meets game theory: secure data integration with malicious participants[J]. The VLDB Journal, 2011, 20(4): 567-588.