*k*-匿名数据的世系及其应用

张三××××××

(东华大学 计算机科学与技术学院 信安××××××，学号：××××××)

摘 要 数据的世系描述了数据的产生及变化的过程，它应用于数据挖掘、数据核查、数据恢复和引用等很多领域。*k*-匿名数据是一种特殊的不确定数据，它是由确定值根据相应的泛化树派生出来的，因此，*k*-匿名数据的世系包含泛化树及派生规则。通过对*k*-匿名数据生成过程的分析，提出了泛化树的形式化定义，在此基础上，从数据接收者角度提出了它的构造算法，使接收者能够更方便有效的对数据进行挖掘分析。现有的对不确定性数据的挖掘算法并不适用于*k*-匿名数据，为了解决这个问题，把*k*-匿名数据的世系应用到挖掘中去，基于泛化树提出了针对*k*-匿名数据的关联规则挖掘算法，最后，通过与原始数据挖掘实验对比，验证了新算法的可行性，并且通过理论分析证明了新算法与传统算法相比降低了挖掘的时间复杂度，提高了挖掘效率。

**关键词** *k*-匿名；不确定性数据；数据挖掘；关联规则；泛化树

数据的世系描述了数据产生、随时间推移而演变的整个过程，它的应用领域广泛，包括数据质量评价、数据核查、数据恢复和数据引用等。因此，世系在越来越多的应用中扮演着关键的角色。但是，现在对不确定数据的世系的研究工作很有限，对*k*-匿名数据的世系研究工作更是还没有开始。根据数据世系的定义，数据世系应包含静态的源数据信息和动态的数据演化过程。*k*-匿名数据的世系描述了静态数据源（即原始表）通过泛化树进行演化，最终得到*k*-匿名表的过程。本文据此提出了*k*-匿名数据世系（provenance of *k*-anonymous data）的定义。

为了进行隐私保护，我们现在得到的很多数据是以*k*-匿名数据的形式出现，但是要对这些数据挖掘分析就存在了很多困难。*k*-匿名数据是一种特殊的不确定性数据，它的元组还原成确定的数据后，每个可能世界出现的概率是相等的。现有的对不确定性数据的挖掘方法是建立在概率不等的基础上，它们大多根据概率大小判定不确定性数据属于哪一类或者哪一组。如果依然用这些方法对*k*-匿名数据进行分析，产生的可能世界将会随着数据量的增大呈指数级增长。如：使用U-Apriori[1]挖掘算法进行关联规则挖掘，一条*k*-匿名数据中的元组就可能产生几十种等概率出现的可能世界，当元组较多时，所需的计算量是巨大的，并不适用于*k*-匿名数据。*k*-匿名数据与不完全数据也不同，它是通过人为的经验来进行数据的隐私保护。接收者无法得到它的原始数据，也无法通过基于贝叶斯理论、决策树等处理不完全数据的方法来对其进行处理。因此，找出一种适用于*k*-匿名数据的挖掘方法是*k*-匿名隐私保护模型亟待解决的问题。

由于*k*-匿名数据的世系描述了原始表通过泛化树进行演化，得到*k*-匿名表的过程，所以我们可以根据世系找出原始表与*k*-匿名表之间的关系，并把它应用到*k*-匿名数据的挖掘中。从*k*-匿名数据世系的定义中我们知道*k*-匿名表是原始表根据特定的泛化树泛化而成的[2]，因此泛化树是世系中最重要的组成部分，不过作为数据接收者，我们仅能得到经过处理后发布的*k*-匿名表，而无法得到泛化树，但是我们可以通过对*k*-匿名数据生成过程的分析推测出泛化树，进而把它应用到挖掘中去。本文首先提出泛化树和世系的定义，根据定义推出泛化树的构造算法，然后基于所构造的泛化树提出新的关联规则挖掘算法，取代了列举所有可能世界的传统算法，极大的减少了计算量，降低了挖掘分析的时间复杂度，最后，通过与原始数据的挖掘结果对比，验证了新算法的可行性。

**1 相关工作**

**1.1 数据的世系**

近年来，数据的世系得到了越来越多的关注[3]。从1990年Y. Richard. Wang等研究异构数据库系统中数据的世系以来[4]，很多研究者从不同角度定义了数据的世系，影响较大的有以下几种：

1991年，D. P. Lanter定义了GIS中数据的世系（data lineage）是有关产生这个数据项的原始素材和演化过程的信息[5]。1997年，A. Woodruff等定义数据的世系（data lineage）是有关数据处理历史的信息集合，包括数据起源（源数据的标识符、所属文件、文件的操作信息）和数据演化过程（运用的算法和相应参数）[6]。 2000 年，Y. Cui等定义数据的世系（data lineage or provenance）是关于数据演化信息的集合[7]。2001年，P. Buneman等限定在数据库应用中定义数据的世系（data provenance, lineage or pedigree）是数据库中数据项的源数据和处理过程[8]。2005年，Y. L. Simmhan等定义数据的世系（data provenance）是用于确定输出数据的演化历史和源数据信息[9]。2007年，B. Glavic等定义数据的世系（data provenance, lineage or pedigree）包括数据项的产生和具有当前的表现形式所经历的处理过程和源数据信息[10]。但是，目前还没有人对*k*-匿名数据这种特殊的不确定性数据提出它的世系。

**1.2 数据挖掘**

针对面向不确定数据的挖掘方法，文献[1]提出了不确定数据的关联规则挖掘算法U-Apriori，它是经典关联规则挖掘算法的改进，该文中提出以期望支持度取代传统的项集支持度记数，并且为了提高效率，U-Apriori还对原始数据集采用LGS-trimming技术修剪掉那些低概率出现的项，文献[11]在分析移动数据聚类时首次提出了基于K-means的不确定聚类算法UK-means，文献[12]对UK-means算法进一步改进，在数据建模过程中采用取样技术选取满足一定PDF的数据对象，并引入了最小边界矩形MBR（minimum bounding rectangle）来描述数据点可能出现的区域，并提出了多种剪枝策略来降低聚类过程中所计算的距离期望的数量，文献[13]提出了不确定性数据的近似聚类算法，包括不确定的K-means，K-median，K-center，这些算法仅是伪近似算法，S. Guha等在文献[14]中对此问题进行了进一步研究，提出了一个真正意义上的近似聚类算法。虽然不确定数据挖掘工作已经取得了很大的成效，对于*k*-匿名这种特殊的不确定数据仍然未见有效的方法。

Arik Friedman等在文献[15]中提出了在数据挖掘的同时对数据进行*k*-匿名隐私保护的方法，它是作为发布者对原始数据进行挖掘并将其*k*-匿名后发布挖掘结果。与文献[15]不同，本文从*k*-匿名数据接收者的角度出发，以发掘*k*-匿名数据中隐含的知识为目的，提出了对*k*-匿名数据的关联规则挖掘算法。

**2 基本定义**

设*T*(*U*)为一个关系表，其中 *T* 为表名, *U*={*U1*, *U2*,…,*Un*}为属性集,为了方便，将*U*分为以下三类[16]：(1)身份属性(identity attribute,记做*Ui*)，可以唯一地标识一个人，如姓名、地址等，这些属性在信息发布之前就必须移除；(2)敏感属性(sensitive attribute，记做*US*)，如疾病等，这些属性值是隐私必须受到保护；(3)准标识符(quasi-identifier，记做*UQI*)这些属性可能与外界已经发布的信息相连接从而泄露隐私(如年龄、区号等属性)，必须对这些属性值进行处理(如泛化操作)才可以发布。

原始数据 泛化树 k-匿名数据

**表 1 原始数据表 *T***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Gender | Age | Zip-code | Disease |
| 1 | M | 25 | 90210 | AIDS |
| 2 | F | 43 | 90211 | AIDS |
| 3 | M | 29 | 90212 | Cancer |
| 4 | M | 41 | 90213 | AIDS |
| 5 | F | 41 | 07620 | Cancer |
| 6 | F | 40 | 07620 | Flu |
| 7 | M | 48 | 07620 | None |
| 8 | F | 40 | 07620 | Flu |
| 9 | F | 40 | 33109 | Cancer |
| 10 | F | 24 | 33109 | None |
| 11 | M | 48 | 33109 | Flu |
| 12 | M | 49 | 33109 | None |

表1 给出了一个待发布原始数据表*T*，其中Gender，Age，Zip-code是准标识符，Disease是敏感属性。

在*k*-匿名数据中，为了进行隐私保护，很多属性的属性值是经过泛化(generalization)的，我们得到的*k*-匿名数据往往都是泛化后的数据，这些泛化的数据存在着与其对应的泛化树。

**定义1 泛化树(G-tree)：**给定一个属性的属性域*D*(*D*为有限集)，一个树的结点的集合*S*={*T*，*v*1，*v*2，……，*vn*，*s*1，*s*2，……，*sn*} (其中*T*为根节点，*s*为叶节点，*v*为除根和叶以外的中间节点)，函数*f*是*S*到*D*的幂集的映射，*S*中存在父子关系的结点*a*和*b*，满足*f*(*b*)*f*(*a*)，对根节点和叶节点有：|*f*(*si*)|=1(1≤*i*≤*n*)，*f*(*s*1)∪*f*(*s*2)∪……∪*f*(*sn*)=*f*(*T*)且*f*(*T*)*D*。图1为年龄属性的一棵泛化树。

10-60

10-25

20-60

10

……

25

20

……

60

**图1 年龄属性的一棵泛化树**

**定义2 *k*-匿名数据的世系：**我们把*k*-匿名数据的产生，即静态的原始数据通过泛化树演化成为*k*-匿名数据的整个过程的信息叫做*k*-匿名数据的世系。它包含了原始数据、泛化树和*k*-匿名数据。如图2所示。

**图2 *k*-匿名数据的世系**

泛化树是*k*-匿名数据世系的核心部分，对研究*k*-匿名数据具有至关重要的作用。但是作为数据接收者，我们能够得到的仅仅是泛化后的数据，因此，如何得到泛化树并把它应用到挖掘中就是下一步工作的重点。

**3 泛化树的构造方法**

根据泛化树的定义，泛化树 G-Tree 具有以下特点：

(1) 任意中间节点是对以其为根的子树上节点的一个概括;

(2) 根节点是对树中所有节点的一个概括;

(3) 泛化树描述了属性域上属性值与其对应的概括值之间的关系。

由泛化树的这些特点，我们来构造一棵泛化树：

**算法1** 构造泛化树

输入：*k*-匿名表中属性*U*的所有属性值*U*=｛*x*1，*x*2，……，*xn*｝；

输出：泛化树G-Tree



**算法2** 在泛化树中插入新的节点（在算法1中被调用）



算法分析：通过得到的*k*-匿名表中属性*U*的属性值｛*x*1，*x*2，……，*xn*｝来构造该属性的泛化树。在算法1中，先构造一个树形结构，第一条元组的值作为其根节点，依次向树中插入余下每一条属性值，形成新的节点，直到所有属性值都插入完毕。在插入新节点的算法2中，首先分析新节点与根节点的关系，当新节点与根节点相等时，不插入新节点；当新节点包含根节点或两者互不包含时，执行算法形成新的泛化树；当根节点包含新节点时，利用递归调用方法本身，把新节点插入到根节点的子树中去，即将根节点的子节点视为新的根节点，再将其与要插入的节点进行比较，以此向下进行，直到把新节点插入到树中。最后，在算法1中，把每个现有的叶节点拆为它的基数为1的子集并插入为它的子节点，如：对{1-3}这个叶节点插入{1}，{2}，{3}3个子节点。

利用算法1和算法2，我们就可以通过得到的*k*-匿名表中的数据来构造每个属性的泛化树。

例：表2为表1的4-匿名表，并对表2的属性Age构造其泛化树。

**表 2 *T*的4-匿名表*TA***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Gender | Age | Zip-code | Disease |
| 1 | \* | 25-45 | 9021\* | AIDS |
| 2 | \* | 25-45 | 9021\* | AIDS |
| 3 | \* | 25-45 | 9021\* | Cancer |
| 4 | \* | 25-45 | 9021\* | AIDS |
| 5 | \* | 40-50 | 07620 | Cancer |
| 6 | \* | 40-50 | 07620 | Flu |
| 7 | \* | 40-50 | 07620 | None |
| 8 | \* | 40-50 | 07620 | Flu |
| 9 | \* | 20-50 | 33109 | Cancer |
| 10 | \* | 20-50 | 33109 | None |
| 11 | \* | 20-50 | 33109 | Flu |
| 12 | \* | 20-50 | 33109 | None |

20-50

25-45

40-50

25

……

45

40

……

50

**图3 表2中属性Age的泛化树**

根据构造好的泛化树，我们可以把它应用到挖掘中去，对*k*-匿名表来进行挖掘分析。

**4 *k*-匿名数据的关联规则挖掘算法**

**4.1 不确定数据的关联规则挖掘算法U-Apriori**

利用不确定性数据的关联规则挖掘算法U-Apriori，我们很容易想到利用可能世界模型，各元组的任意合法组合均能构成一个可能世界实例，实例的概率可以通过相关元组的概率计算得到。在不确定性数据模型下，由于无法获得项集的确定支持度计数，所以传统的支持度定义无法应用，为此，本文引入期望支持度的概念来计算频繁项集。

例：要从表2中挖掘年龄段为40-50的人与疾病Flu之间的关联规则。我们需要得到的结果是置信度*P*：



要挖掘年龄段为40-50的人与疾病Flu之间的关联规则，就需要分别计算出年龄为40-50的人的期望支持度（即*P*的分母）与这个年龄段中得到疾病Flu的人的期望支持度（即*P*的分子），这需要对每个元组分别进行计算。

首先，分析表2可产生的可能世界实例：对于*ID*为1的元组来说，各个属性可取的值如下表：

**表3 属性Gender可取值表**

|  |  |
| --- | --- |
| 可取值 | 概率 |
| F | 1/2 |
| M | 1/2 |

**表4 属性Age可取值表**

|  |  |
| --- | --- |
| 可取值 | 概率 |
| 25 | 1/21 |
| 26 | 1/21 |
| 27 | 1/21 |
| …… | …… |
| 43 | 1/21 |
| 44 | 1/21 |
| 45 | 1/21 |

**表5 属性Zip-code可取值表**

|  |  |
| --- | --- |
| 可取值 | 概率 |
| 90210 | 1/10 |
| 90211 | 1/10 |
| 90212 | 1/10 |
| …… | …… |
| 90217 | 1/10 |
| 90218 | 1/10 |
| 90219 | 1/10 |

由以上表格可计算*ID*为1的元组可产生的可能世界实例个数为2\*21\*10=420个。表4中12条元组一共可产生的可能世界实例为4204\*224\*624个。并且由于每个元组中可能世界实例出现的概率是完全相同的，也无法使用剪枝策略[1]对其进行修剪简化，再对每个可能世界求出所要计算的项的支持度来进行关联规则挖掘，这将是一个无比繁重的工作。当有更大的数据量时，计算所需的时间是以指数及增长的，它的时间复杂度为*O*(*nn*)。因此，传统的不确定性数据的关联规则挖掘算法是不适用于这种*k*-匿名数据表的。

**4.2 基于泛化树的关联规则挖掘算法**

从4.1的计算中可知，由于*k*-匿名数据表的每条元组中每个可能世界实例的产生是等概率的，因此，用传统的不确定数据发掘算法会有极大的时间复杂度，是不可取的。我们需要尝试一种新的挖掘算法，即基于泛化树的关联规则挖掘算法。

首先，我们可以通过泛化树的构造算法从*k*-匿名表中得到每个属性的泛化树。表2中属性Age的泛化树如图3。然后，通过泛化树来计算待挖掘属性值的期望支持度。

**算法3** 计算待挖掘属性值在每条元组中的期望支持度

输入：待挖掘属性*U*的泛化树*GTU*，待挖掘属性值*xU*，第*i*条元组*U*的属性值*xi*

输出：待挖掘属性值在该元组中的期望支持度*ESupUi*



算法分析：由于在*k*-匿名数据表中，每条元组还原成泛化前确定元组的概率是相等的，就可以通过计算待挖掘属性值在每个元组中所占比例来得到它在每条元组中的期望支持度。把每条元组的期望支持度加起来就可以得到待挖掘属性在所有元组中的期望支持度并可以据此来进行挖掘。算法3就是计算待挖掘数据在每条元组中的期望支持度*ESupUi*的方法。则属性值*xu*在数据集中的期望支持度为：



当*support*(*xu*)大于用户所给的最小支持度阈值时，*xu*即为频繁项集。构造泛化树的时间复杂度为*O*(*n*2)，挖掘期望支持度算法的时间复杂度也为*O*(*n*2)，因此整个算法的时间复杂度为*O*(*n*2)。

通过以上分析，可以得到对关联规则挖掘算法中置信度的计算方法：



**算法4** 基于泛化树的*k*-匿名数据集的关联规则挖掘算法

输入：待挖掘项集*A*、*B*

输出：项集*A*、*AB*的支持度*supA*、*supAB*，*A**B*的置信度*P*



当*supAB*大于用户给定的阈值（即*A*和*AB*均为频繁项集）并且*P*大于用户所给最小置信度时，即可得到强关联规则*A**B*。

**5 实验数据及结果**

**5.1 实验数据来源**

研究采用的数据源是UCI机器学习库中的Adult数据库，该数据库中有32561条数据记录，每条记录有15个属性。这些数据源自美国人口普查库，我们精简了相关属性，保留了其中的6个属性，其中1个属性为敏感值，将age，sex，[final weight]，education，work-class作为准标识符，salary作为敏感属性值。

**5.2 实验结果**

我们对这组数据随机抽取其中的5290条数据作为训练样本，将其5-匿名后对样本数据进行关联规则挖掘。需要挖掘年龄段为22-27与工资为50k以上之间的关联规则。

很明显，对于泛化过的5290条数据来说，如果使用不确定性数据的关联规则挖掘算法，计算量将是巨大的，我们无法通过实验来得到结果进行对比。因此，我们采取基于泛化树的关联规则挖掘算法。

首先，通过泛化树的构造算法求出其年龄的泛化树，如图4：

0-95

0-14

15-19

20-24

……

70-74

75-84

85-95

……

23

24

25-29

25

26

……

**图4 年龄的泛化树**

由基于泛化树的关联规则挖掘算法得到：



通过对原始数据进行统计，在随机抽取的5290人中，年龄段22-27的人数为818人，在这些人中，工资为50k以上的人数为139人，所占比例为0.169926。图5为原始数据统计与挖掘结果的对比直方图。



**图5 结果对比直方图**

可以看出，实验结果和真实数据的差距并不大，可以将这些得到的结果与用户给定的阈值和最小置信度相比较来判断它们之间的关联规则。因此，基于*k*-匿名数据世系中泛化树的关联规则挖掘算法可以看作是对*k*-匿名数据挖掘的一种新的尝试。

**6 总结与展望**

对*k*-匿名数据的挖掘是一个亟待解决的问题。为此，本文首先针对*k*-匿名数据表提出了*k*-匿名数据的世系，然后，从中提出泛化树的构造算法，并把它应用到了*k*-匿名数据的挖掘中，即利用所构造的泛化树来产生新的关联规则挖掘算法，最后，给出了在该算法下生成的泛化树和挖掘结果，实验验证该算法降低了挖掘的时间复杂度，提高了挖掘效率。

后续工作中，我们在进一步提高该算法的同时，还将结合确定和不确定数据的世系及挖掘算法，对*k*-匿名数据的世系和挖掘算法进行进一步的研究，使我们在保护隐私的同时挖掘出更多有用的信息，提高*k*-匿名数据的可用性。

**参 考 文 献**

1. Chun-Kit Chui, Ben Kao and Edward Hung. Mining frequent itemsets from uncertain data [J]. In: PADDK 2007,47-58
2. Xiaoxun Sun, Lili Sun and Hua Wang. Extended k-anonymity models against sensitive attribute disclosure [J]. Computer Communications 34(4):526-535 ,2011
3. 高明,金澈清,王晓玲,田秀霞,周傲英. 数据世系管理技术研究综述[J]. 计算机学报. 2010,33(3): 373-389.
4. Y. Richard Wang, Stuart E. Madnick. A polygen model for heterogeneous database systems: the source tagging perspective// Proceedings of the 16th International Conference on Very Large Data Bases, Brisbane, Queensland, Australia, February 5-9,1990,16:519-538.
5. D. P. Lanter. Design of a lineage-based meta-data base for GIS. Cartography and Geographic Information Systems, 1991, 18:255-261.
6. A. Woodruff, M. Stonebraker. Supporting fine-grained data lineage in a database visualization environment// Proceedings of the 13rd IEEE International Conference on Data Engineering. Birmingham, England, April 1997: 91-102.
7. Y. Cui, J. Widom, J. L. Wiener. Tracing the lineage of view data in a warehousing environment. The ACM Transactions on Database Systems, 2000, 25(2):179-227.
8. P. Buneman, S. Khanna, W. C. Tan. Why and where: a characterization of data provenance// Proceedings of the 17th International Conference on Data Engineering, London, UK, April 2001:316-330.
9. Yogesh L. Simmhan, Beth Plale, Dennis Gannon. A survey of data provenance techniques. Computer Science Department: Indiana University, Bloomington IN 47405 Technical Report IUB-CS-TR618 2001.
10. Boris Glavic, Klaus Dittrich. Data provenance: A categorization of existing approaches// Proceeding of the [6th MMC Workshop of BTW 2007](http://www.multimedia-metadata.info/Events/Slides%20and%20Presenations/6thWorkshopSlides), Aachen, Germany, March 2007: 227-241.
11. Y.F. Li, J.W. Han, J. Yang. Clustering moving objects [R]. Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.
12. W. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, K.Y. Yip. Efficient clustering of uncertain data [R]. Proc. of the 6th International Conference on Data Mining, 2006.
13. G. Cormode, A. McGregor. Approximation algorithms for clustering uncertain data [J]. In: Proc. of PODS(2008).
14. S. Guha, K. Munagala. Exceeding expectations and clustering uncertain data [J]. In: Proc. of PODS (2009).
15. Arik Friedman, Ran Wolff, Assaf Schuster. Providing k-anonymity in data mining [J]. The VLDB Journal (2008)17:789-804.
16. Wang Danli, Liu Guohua, Song Jinling, Li Fangling. Problem of finding the optimal value on quasi-identifier for k-anonymity model [J]. Journal of Frontiers of Computer Science and Technology, 2010, 4(11): 1010-1018.(王丹丽,刘国华,宋金玲,李芳玲. k-匿名模型中准标识符最佳值的求解问题[J]. 计算机科学与探索. 2010, 4(11): 1010-1018.)