# 基于模糊集的隐私保护方法研究

王　茜, 杨传栋, 刘　泓
(重庆大学 计算机学院, 重庆 400044)

**摘　要**: 在数据发布的隐私保护研究中,针对 $k$-匿名方法的复杂性高、效率低及数据可用性差等问题,从基于模糊集的角度出发进行隐私保护的研究,重点是对数值型属性的处理,提出了基于模糊集的最大隶属度(MMD)算法。该算法对敏感数值型数据进行模糊化处理,把其变成语义型数据,结合隶属度一起发布以达到隐私保护的目的。并通过实验进行了验证,基于模糊集的隐私保护方法与 $k$-匿名方法相比,具有更高的效率,且信息损失要远远小得多,发布数据的可用性更好。

**关键词**: 隐私保护; 模糊集; 模糊化; 隶属函数; 隶属度; $k$-匿名
**中图分类号**: TP309　　**文献标志码**: A　　**文章编号**: 1001-3695(2013)02-0518-03
doi:10.3969/j.issn.1001-3695.2013.02.055

## Fuzzy-based methods for privacy preserving

WANG Qian, YANG Chuan-dong, LIU Hong
(*College of Computer Science*, *Chongqing University*, *Chongqing* 400044, *China*)

**Abstract**: This paper did research based on fuzzy sets to overcome the high complexity, low efficiency and poor data availability of $k$-anonymity in the research of privacy-preserving data publishing. It focused on the processing of numerical attributes, and proposed the maximal membership degree algorithm. It fuzzed sensitive numerical attributes to semantic data which was released combining with membership degree. Verified through experiments, compared with $k$-anonymity methods, the MMD has better efficiency, furthermore, its information losses will be far smaller than $k$-anonymity and the availability of released data is better.

**Key words**: privacy preserving; fuzzy sets; fuzzed; membership function; membership degree; $k$-anonymity

## 0　引言

随着信息技术的快速发展,越来越多的个人信息记录被不同的部门和机构大量收集,这些数据可能会被这些部门和机构用于数据发布或者数据挖掘,但是由于信息中可能包含个人隐私,如果将收集到的数据直接发布就会造成个人隐私泄漏。因此,为了保障个人敏感信息的安全,就要对发布的数据进行隐私保护。目前,已有多种隐私保护的方法被提出来,在隐私保护的技术方法中主要分成三类[1]:a)基于数据失真的技术,它是使敏感数据失真但同时保持某些数据或数据属性不变的方法,如采用添加噪声、交换等技术对原始数据进行扰动处理;b)基于数据加密的技术,它是采用加密技术在数据挖掘过程中隐藏敏感数据的方法,用于分布式环境及外包数据;c)基于限制发布的技术,它是根据具体情况有条件地发布数据,主要有泛化和隐匿技术。当然除此之外,还有一些结合多种技术的方法。

对于数据发布中所要处理的原始数据,如医疗数据、统计数据等,一般为数据表形式:表中每一条记录(或每一行)对应一个个人,包含多个属性值。这些属性可以分为三类:a)显式标志符(EID),能唯一标志单一个体的属性,如身份证号码、姓名等;b)准标志符(QID),联合起来能唯一标志一个人的多个属性,如邮编、生日、性别等联合起来则可能是准标志符;c)敏感属性(SA),包含隐私数据的属性,如疾病、薪资等。

Sweeney 和 Samarati[2~4] 提出的 $k$-匿名原则即是要求所发布的数据表中的每一条记录不能区分于其他 $k-1$ 条记录。本文称不能相互区分的 $k$ 条记录为一个等价类。这里的不能区分只对非敏感属性项而言。一般 $k$ 值越大,对隐私的保护效果越好,但丢失的信息越多。$k$-匿名的缺陷在于没有对敏感数据作任何约束,攻击者可以利用一致性攻击和背景知识攻击来确认敏感数据与个人的联系,导致隐私泄露。$(\alpha,k)$-匿名原则在此基础上进行了改进,其在保证发布的数据满足 $k$-匿名化原则[5] 的同时,还保证发布数据的每一个等价类中,与任一敏感属性值相关的记录的百分比不高于 $\alpha$。

针对 $k$-匿名的不足,$l$-diversity[6] 被提出来,$l$-diversity 保证每一个等价类的敏感属性至少有 $l$ 个不同的值。$l$-diversity 使得攻击者最多以 $1/l$ 的概率确认某个体的敏感信息。但是由于在一个组内某些敏感属性很自然地比其他敏感属性频繁,所以 $l$-diversity 不能阻止概率推理攻击。进而出现了 $l$-diversity 的另外两种形式,基于熵的 $l$-diversity 和递归$(c,l)$-diversity。

Li 等人[7] 在 2007 年针对全局隐私信息和单个统计个体隐私信息的保护问题分析了 $l$-diversity 方法的不足,提出了 $t$-closeness 隐私保护方法。若发布表每个等价类中敏感属性值的分布与该敏感属性取值在整个发布表中的分布差异不超过 $t$ 时,称该发布表满足 $t$-closeness 匿名要求。在 $k$-匿名基础上,进一步改进方法还有 $m$-Invariance[8]、支持多约束的 $k$-匿名

**收稿日期**: 2012-06-08; **修回日期**: 2012-07-21

**作者简介**:王茜(1964-),女,重庆人,教授,主要研究方向为信息安全、电子商务、远程教育课件工具;杨传栋(1987-),男,山东郯城人,硕士研究生,主要研究方向为隐私保护、信息安全(ycdjin@126.com);刘泓(1982-),男,重庆人,硕士研究生,主要研究方向为隐私保护、信息安全.

化[9]方法等。当然除此之外,还有一些别的隐私保护方法,如隐私保护关联规则的挖掘[10]、基于分布式的隐私保护研究[11]等。

从 2008 年开始 Kumari 和 Poovammal 等人[12~14]提出了应用模糊数学的隐私保护方法,是基于对敏感属性进行模糊化来达到隐私保护的目的。对敏感数值型属性通过模糊化来把数值型属性转换为语义型数据,然后根据隶属度来确定每个属性所属的语义属性集,以达到隐私保护的目的。基于模糊集的隐私保护方法为隐私保护提供了新思路,但是已有的基于模糊集的隐私保护方法没有提出合适的算法,而且不能完全保证隐私不被泄露。

本文中的隐私保护方法也是基于模糊集提出的,重点放在对数值型数据处理上,并且提出了基于模糊集的最大隶属度(MMD)隐私保护算法。

## 1　数学基础

### 1.1　模糊数学的基本概念

模糊数学,亦称弗晰数学或模糊性数学,是在模糊集合、模糊逻辑的基础上发展起来的模糊拓扑、模糊测度论等数学领域的统称,是研究现实世界中许多界限不分明甚至是很模糊的问题的数学工具。

**定义 1**　模糊子集、隶属度与隶属函数[15]。所谓给定论域 $U$ 上的一个模糊子集 $A$,就是给定由论域 $U$ 到区间[0,1]的一个映射:

$$u_A: U \to [0,1],$$
$$u \mapsto u_A(u) \in [0,1]$$

这个映射 $u_A$ 将任一 $u \in U$ 对应着一个确定的值 $u_A(u) \in [0,1]$,值 $u_A(u)$ 叫做 $u$ 对模糊子集 $A$ 的隶属度,映射 $u_A$ 叫做模糊子集 $A$ 的隶属函数。例如,设论域 $U = \{x_1(140), x_2(150), x_3(160), x_4(170), x_5(180), x_6(190)\}$(单位:cm)表示人的身高,那么 $U$ 上的一个模糊集"高个子"($A$)的隶属函数 $A(x)$ 可定义为 $A(x) = \dfrac{x-140}{190-140}$。

### 1.2　隶属函数的确定

模糊数学的基本思想是隶属程度的思想,应用模糊数学方法建立数学模型的关键是建立符合实际的隶属函数。下面介绍几种常用的确定隶属函数的方法。

1)模糊统计方法　它是一种比较客观的方法,主要是基于模糊统计实验的基础上,根据隶属度的客观存在性来确定。其步骤与概率统计随机实验是相对应的。与概率统计类似,但有区别:若把概率统计比喻为"变动的点"是否落在"不动的圈"内,则把模糊统计比喻为"变动的圈"是否盖住"不动的点"。

2)指派方法　它是一种主观的方法,主要依据人们的实践经验来确定某些模糊集隶属函数的一种方法。若模糊集定义在实数域 R 上,则模糊集的隶属函数称为模糊分布;指派方法就是根据问题的性质主观地选用某些模糊分布,再根据实际测量数据确定其中的参数。

3)借用已有的"客观"尺度　在经济管理、社会科学中可以直接借用已有的尺度作为模糊集的隶属度,如在论域 $U$ 上定义模糊集 $A$ = "设备完好",可以"设备完好率"作为隶属度来表示"设备完好"这个模糊集。

## 2　基于模糊集的隐私保护方法

### 2.1　最大隶属度隐私保护(MMD)算法

基于模糊集的隐私保护方法是对待发布表中的数值型数据和分类型数据分别进行处理。对非数值型数据采用分类树的方法,本文主要是针对敏感属性中数值型数据的处理研究。

**定义 2**　最大隶属度原则。设 $A_1, A_2, \cdots, A_n \in F(U)$ 构成了一个标准模型库,若对任意 $x_0 \in U$, $i \in \{1,2,\cdots,n\}$ 使得 $A_i(x_0) = \bigvee_{k=1}^{n} A_k(x_0)$,则认为 $x_0$ 相对隶属于 $A_i$。

引入阈值 $\lambda$,$\lambda$ 是接近于 1 的参数,对不同的属性取值会有所差异,当模糊度大于 $\lambda$,则用 $\lambda$ 来代替模糊度,这样设计是为了增加隐私的保护程度。算法的实现步骤:

a)对数值型数据进行模糊化,确定其模糊集为 $\{A_1, A_2, \cdots, A_n\}$,$n$ 为模糊集的个数。这里的模糊集通常是根据经验来确定。

b)确定模糊集的隶属函数。这里采用指派的方式,根据数据的统计特性来选择模糊分布确定隶属函数。

c)计算所有数值型数据对模糊集 $A_i$ 的隶属度,根据 $A_i(x_0) = \bigvee_{k=1}^{n} A_k(x_0)$ 来确定数值相对隶属的模糊集并记录隶属度。这样就得到每个数值所属的模糊集及隶属度。

d)对隶属度大于 $\lambda$,就用 $\lambda$ 来代替,作为最后要发布的隶属度。

e)进行上述处理后,数据值由一列变成了两列,这样就增加了发布表的复杂性,因此可以采用下面的处理方法,对模糊集进行标号,如 $n$ 个模糊集分别用序号 1,2,…,$n$ 来代替,发布时只需发布序号及序号所属模糊集的隶属度。例如模糊集 $A_i$ 为序号 $i$,隶属度为 0.8,则发布时只需发布为 $i.8$ 即可。

### 2.2　应用举例

有待发布如表 1 所示,敏感属性为数值型属性"收入",按照上述的算法对其进行模糊化处理。

表 1　初始微数据

| 姓名 | 年龄 | 性别 | 收入 | 姓名 | 年龄 | 性别 | 收入 |
|---|---|---|---|---|---|---|---|
| 张三 | 52 | 男 | 10 000 | 宋新 | 54 | 女 | 57 000 |
| 李四 | 63 | 男 | 48 000 | 刘丽 | 27 | 女 | 94 000 |
| 王五 | 34 | 男 | 15 000 | 吴发 | 60 | 男 | 51 000 |

a)对属性收入进行模糊化,确定模糊集为 $\{f_1$ = 低收入, $f_2$ = 中等收入, $f_3$ = 高收入$\}$。

b)再通过指派的方法确定隶属函数低、中、高收入的隶属函数分别为

$$f_1(x) = \begin{cases} 1 & x = \min \\ \dfrac{a_2 - x}{a_2 - \min} & \min < x < a_2 \\ 0 & x \geqslant a_2 \end{cases}$$

$$f_2(x) = \begin{cases} 0 & x \leqslant a_1 \\ \dfrac{x - a_1}{a_2 - a_1} & a_1 < x < a_2 \\ 1 & x = a_2 \\ \dfrac{a_3 - x}{a_3 - a_2} & a_2 < x < a_3 \\ 0 & x \geqslant a_3 \end{cases}$$

$$f_3(x) = \begin{cases} 0 & x \leqslant a_2 \\ \dfrac{\max - x}{\max - a_2} & a_2 < x < \max \\ 1 & x = \max \end{cases}$$

其中:min = 10000, max = 94000, $a_2$ = (min + max)/2, $a_1$ = (min +

$a_2)/2$, $a_3=(a_2+\max)/2$。

c)计算每一项的隶属度,取 $\lambda=0.99$,按照最大隶属度原则得到表2。

表2 中间过程的表

| 姓名 | 年龄 | 性别 | 收入 | 隶属度 |
|---|---|---|---|---|
| 张三 | 52 | 男 | 低 | 0.99 |
| 李四 | 63 | 男 | 中 | 0.80 |
| 王五 | 34 | 男 | 低 | 0.88 |
| 宋新 | 54 | 女 | 高 | 0.88 |
| 刘丽 | 27 | 女 | 高 | 0.99 |
| 吴发 | 60 | 男 | 中 | 0.95 |

d)为了进一步减少表的复杂性,本文对底、中、高收入分别用1、2、3来表示,进而把收入和隶属度两列合并为一列,得到表3,即为要发布的表。

表3 最终要发布的数据表

| 序号 | 年龄 | 性别 | 收入 | 序号 | 年龄 | 性别 | 收入 |
|---|---|---|---|---|---|---|---|
| 1 | 52 | 男 | 1.99 | 4 | 54 | 女 | 3.88 |
| 2 | 63 | 男 | 2.80 | 5 | 27 | 女 | 3.99 |
| 3 | 34 | 男 | 1.88 | 6 | 60 | 男 | 2.95 |

## 3 实验

本文采用来自UCI机器学习数据库中的成人数据集(adult data set)进行实验。选择其中的属性age、occupation、H_W(每周工作小时数)、country作为要发布的属性,把属性age作为敏感属性,运用第2章的算法进行实验。实验平台配置如下:AMD 2.71 GHz/2 GB,Windows XP,所涉及代码用Visual C++ 6.0实现。

### 3.1 时间效率

成人数据集中的数据总共有32 561个数据项,在实验中分别采用本文的基于模糊集的最大隶属度算法和一种基于$k$-匿名的算法对数据集处理,观察它们的执行时间,依次取5000、10000、15000、20000、25000、30000个数据进行分别进行实验,得到实验结果如图1所示。由图1可以看出,采用MDD算法的运行时间远远小于基于$k$-匿名的算法所运行时间,由此可以得出基于模糊集的最大隶属度法效率明显优于$k$-匿名的算法。

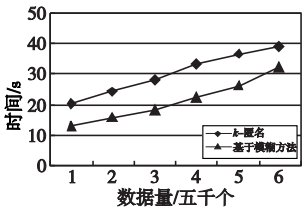


**图1 基于模糊集算法和$k$-匿名算法执行时间**

### 3.2 信息损失

基于$k$-匿名的隐私保护大部分采用的都是泛化和隐匿的方法,信息的损失程度取决于准标志符(QID)和敏感属性(SA)。泛化和隐匿程度越高,信息损失就越大。但是本文基于模糊集的隐私保护方法中没有对准标志符进行处理,只是对敏感属性由数值型数据转换为语义型数据,因此本文方法处理后的数据几乎没有信息损失,所发布数据的可用性也远远高于基于$k$-匿名的方法。

## 4 结束语

与以往隐私保护的方法不同,本文从数学中的模糊集出发,首先介绍了模糊数学中的模糊子集、隶属度与隶属函数的基本概念,然后根据最大隶属度原则提出了基于模糊集的最大隶属度算法,并在UCI数据集实验上取得了理想的效果,通过对实验结果的分析表明了该方法的优越性。本文中只是对单敏感属性基于模糊集的应用研究,在接下来的工作中,也会从多敏感属性进行研究,进一步扩展和完善基于模糊集的隐私保护方法。

**参考文献:**

[1] 周水庚,李丰,陶宇飞,等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报,2009,32(5):847-860.

[2] SWEENEY L. $k$-anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems,2002,10(5):557-570.

[3] SWEENEY L. Achieving $k$-anonymity privacy protection using generalization and suppression[J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 571-588.

[4] SAMARATI P. Protecting respondents' identities in micro data release[J]. IEEE Trans on Knowledge and Data Engineering, 2001,13(6):1010-1027.

[5] WONG R C, LI Jiu-yong, FU A W, *et al*. $(\alpha,k)$-anonymity: an enhanced $k$-anonymity model for privacy-preserving data publishing [C]//Proc of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006: 754-759.

[6] MACHANAVAJJHALA A, GEHRKE J, KIFER D, *et al*. L-diversity: privacy beyond $k$-anonymity[J]. ACM Trans on Knowledge Discovery From Data,2007,2(1):1-12.

[7] LI Ning-hui, LI Tian-cheng. VENKATASUBRAMANIAN S. Tcloseness: privacy beyond $k$-anonymity and $l$-diversity[C]//Proc of the 23rd IEEE International Conference on Data Engineering. 2007:106-115.

[8] XIAO Xiao-kui, TAO Yu-fei. M-invariance: towards privacy preserving re-publication of dynamic datasets[C]//Proc of ACM SIGMOD Conference on Management of Data. New York: ACM Press, 2007: 689-700.

[9] YANG Xiao-chun, LIU Xiang-yu, WANG Bin, *et al*. K-anonymization approaches for supporting multiple constraints[J]. Journal of Software,2006,17(5):1222-1231.

[10] 刘峰,薛安荣,王伟. 一种隐私保护关联规则挖掘的混合算法[J]. 计算机应用研究,2012,29(3):1107-1110.

[11] 刘英华,杨炳儒,马楠,等. 分布式隐私保护数据挖掘研究[J]. 计算机应用研究,2011,28(10):3606-3610.

[12] KUMARI V V, RAO S S, RAJU K, *et al*. Fuzzy based approach for privacy preserving publication of data[J]. International Journal of Computer Science and Network Security,2008,8(1):115-121.

[13] POOVAMMAL E, PONNAVAIKKO M. An improved method for privacy preserving data mining[C]//Proc of IEEE International Advance Computing Conference. [S.l.]:IEEE Press,2009:1453-1458.

[14] POOVAMMAL E, PONNAVAIKKO M. Preserving micro data release: categorical and numerical data[C]//Proc of the 5th International Conference: Sciences of Electronic, Technologies of Information and Telecommunication. [S.l.]:IEEE Press, 2009:5-7.

[15] 梁保松,曹殿立. 模糊数学及其应用[M]. 北京:科学出版社,2007:16-17,35-46.