

# 增强隐私保护度的数据混淆机制研究

邵必林<sup>1</sup>, 蔡 婷<sup>1</sup>, 边根庆<sup>1,2</sup>, 王小飞<sup>2</sup>

(1. 西安建筑科技大学管理学院, 陕西 西安 710055; 2. 西安建筑科技大学信息与控制工程学院, 陕西 西安 710055)

**摘要:** 针对数据隐私保护的安全问题, 提出了一种基于数据混淆的隐私数据保护机制。首先介绍了相关背景及理论基础, 然后论述了所采取的隐私保护二次数据混淆方法, 即先基于非固定位置置换的数据混淆进行第一次简易混淆, 再基于随机正交矩阵思想进行第二次数据混淆, 并通过混淆可逆变换, 准确地把原始数据提供给使用者。通过实验表明, 该机制在有效保护隐私数据的同时, 能明显提高其安全保护系数和等级。

**关键词:** 隐私保护; 数据混淆; 非固定位置置换; 随机正交矩阵

中图分类号: TP309

文献标志码: A

文章编号: 1006-7930(2016)01-0036-05

## Research of enhancing privacy protection on confused data

SHAO Bilin<sup>1</sup>, CAI Ting<sup>1</sup>, BIAN Genqing<sup>1,2</sup>, WANG Xiaofei<sup>2</sup>

(1. School of Management, Xi'an Univ. of Arch. & Tech., Xi'an 710055, China;

2. School of Information and Control Engineering, Xi'an Univ. of Arch. & Tech., Xi'an 710055, China)

**Abstract:** Aiming at the security of data privacy protection, the paper puts a protective mechanisms of private data based on the data confusion. Firstly, it introduces the relevant background and theoretical foundation. It then, discusses the adopted secondary data confusion method of privacy protection, which proceeds the first simple confusion on the basis of random location replacement. The second data confusion in view of the random orthogonal matrix idea. It provides the user with the accurate raw data through confusing reversible transformation. The experiment showed that the mechanism provides the effective protection for private data and increases the protective grades obviously.

**Key words:** privacy protection; data confusion; random location replacement; random orthogonal matrix

随着大数据时代的到来, 企业和个人越来越重视大数据技术带来的产业价值, 但是大数据的应用也增加了隐私数据保护的风险。同时, 计算机系统(软硬件)自身固有的脆弱性也易引起硬件设备在使用过程中产生故障; 软件系统由于无法做到穷尽测试, 也存在很多的漏洞、后门等潜在风险。非授权用户的攻击、授权用户的误操作、病毒的危害等, 都给隐私数据保护带来了严峻的挑战, 因此加强隐私数据的安全保护已成为当前学术界和实际应用领域亟待研究解决的热点问题。目前针对隐私数据的保护主要从法律和技术两方面进行<sup>[1]</sup>, 如图 1 所示。

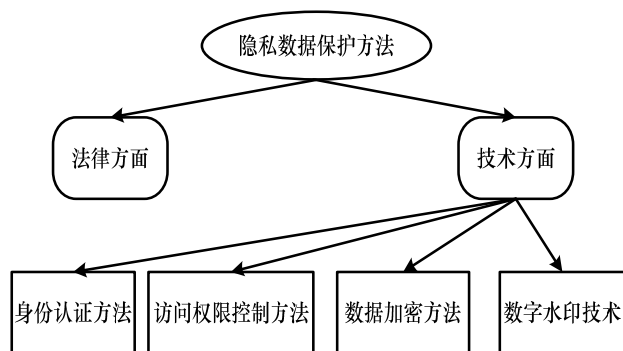


图 1 隐私数据保护方法

Fig.1 The method of privacy data protection

在法律方面, 主要是通过建立完善的隐私保护法律法规, 对侵犯与泄漏隐私的, 按隐私等级与获得的价值等严加惩处, 达到制度保护的目的。

在技术方面, 常用的隐私保护方法有身份认证、访问权限控制、加密机制、数字水印等。文献[2]分析了身份认证具有简单、灵活等优点, 但其验证服务分离降低了该方法的信任度, 使身份认证的风险增大, 攻击者通常可以找到一些方法以骗取用户的秘密, 而且在传统的身份认证技术中安全度越高的认证技术其复杂度也越高, 给用户造成的负担也就越重, 并且存在身份认证服务不完整等问题; 文献[3]研究了数据访问控制技术, 因为大数据的复杂性特点, 现有的数据访问控制技术不能较好地解决数据访问过程中的安全性问题, 亦很难满足数据访问控制的复杂性要求, 而且随着角色的增多, 访问控制系统的规模会不断扩大, 也增加了管理系统的开销; 文献[4]讨论了数据加密技术及其应用, 可以看出, 虽然常规性的加密技术不断得到广泛应用, 但数据信息的安全度并未能显著提高, 而安全性高的软件产品价格又非常高, 不适用于一般用户;

文献[5]中的水印技术目前处于快速发展阶段,但就现有水印算法而言,在理论上有许多相似之处,缺乏进一步研究的理论支持,并且其所有权的证明问题也未能得到解决。

对于个人来说,隐私数据一旦泄漏可能危害到自身的名誉、隐私合法权益等;对于企业而言,隐私数据泄漏,轻者会对企业的发展、利益等带来损失,重者若企业的关键技术被泄漏,很有可能会导致企业的倒闭;对于一个国家,机密数据被泄漏,可能会危及到整个国家的安全等<sup>[6]</sup>,然而就上述分析,目前主流的隐私保护方法在提高隐私数据保护度方面或多或少都存在一定不完善性。因此,进一步研究隐私数据的保护技术具有重要的学术价值和实际意义。基于此,本文提出了一种将数据混淆理论引入到隐私数据保护的安全机制,以进一步提高其保护的安全系数和安全等级。

## 1 数据混淆基本原理

### 1.1 混淆思想

混淆变换的目的是阻止攻击者破译原程序中代码或数据意义,因此混淆变换就是将代码或数据进行变换,尽可能地隐藏其中的意义,或者说以另一种形式表示其意义。

先定义以下几个概念:原始程序  $P$ ;混淆策略  $K$ ;混淆变换后的程序  $P'$ 。混淆变换的原理<sup>[7]</sup>如图2所示。

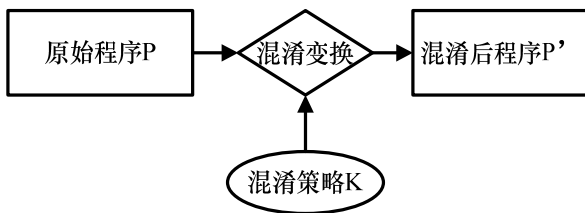


图2 混淆变换的原理示意图

Fig.2 The schematic diagram of confusion

混淆变换要保证  $P'$  与  $P$  具有同样的功能,即要满足下述两个条件:

- (1) 假设  $P$  发生意外终止或者未能终止,此时  $P'$  并不一定终止;
- (2) 当  $P$  和  $P'$  均正常终止,则  $P'$  和  $P$  输出的结果一定是一样的。

$P'$  与  $P$  相比,  $P'$  被反编译的难度更大,也更难被非法用户或一些相关工具破译。

根据数据混淆变换的思路及其不同混淆处理对象须对应不同的混淆方法,可将混淆技术分为布局

混淆、控制流混淆、预防混淆、数据混淆等几种<sup>[8]</sup>。布局混淆为了降低可用于攻击者阅读及理解的代码数据量,往往通过删除、改名等方法来实现,同时要保证不影响程序的执行,它是一种不可逆的变换方法,具有较高混淆强度,较高程序执行效率;控制流混淆是运用某种技术/方法来隐藏或者修改实际控制流程,以防止攻击者分析出其真正流程;预防混淆的目的是预防现有的反编译软件对程序进行破译,主要通过分析反编译软件漏洞及其自身缺陷,而后设计出相应的方案来进行预防。本文主要采用数据混淆方法,以提高隐私数据的保护度,该方法将在后文中给出详细论述。

### 1.2 数据混淆

数据混淆是在不影响软件功能的基础上,通过改变软件程序代码中数据、数据格式的方法来达到增加代码复杂度的目的。根据不同的混淆方式,可以把数据混淆分为次序变换、聚集变换、存储和编码变换等。次序变换如重新排列实际变量、变换排序方法等;聚集变换是把多个数据聚集从而构成新的数据结构,达到数据保护的目的,这种变换方法主要应用在混淆面向对象的高级语言,有数组聚集和对象聚集两种聚集方式,合并标量变量、重构数组等是常用的聚集方法;存储和编码变换是指改变软件代码中变量的编码方式与存储方式,这样可以改变变量的含义,使它们对于攻击者而言其操作和用途就变得更加难以理解<sup>[9]</sup>,常用的方法有分割变量、简单标量复杂化、改变变量的生命周期、把静态数据通过函数来表示、修改编码方式等。本文主要基于编码的方法,首先采用非固定位置置换的数据混淆方法进行第一次简易混淆,再利用正交矩阵的原理对一次混淆后的数据进行二次混淆,从而增强数据的隐私保护度。

## 2 基于数据混淆的隐私保护方法设计

本文提出的数据混淆技术对隐私数据进行保护主要分为四个阶段,首先对原始数据进行编码,编码结果以矩阵符号的形式输出;接着根据隐私保护需求对数据进行分块,将分块中具有特定意义的敏感数据字符串进行非固定位置置换,完成一次数据混淆;然后再用随机正交矩阵数据混淆算法实现二次数据混淆,增强其隐私保护度;最后,对数据进行逆向反混淆变换提供给使用者正确的数据信息。其方法如图3所示。

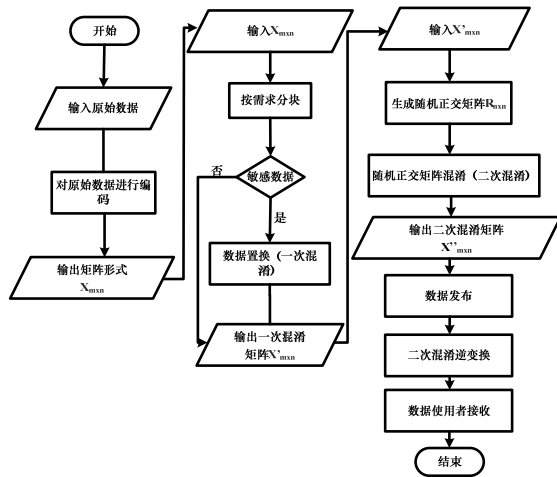


图3 基于数据混淆的隐私保护方法  
Fig.3 The privacy protection method based on data confusion

## 2.1 数据编码

分析现有数据信息的输入模式，主要有数字模式、字母数字模式、中国汉字模式等，按照文献[10]将输入的数据信息进行编码，而后将编码的数据构造成包含  $m$  条记录和  $n$  维属性的原始数据矩阵  $X_{m \times n}$ 。

## 2.2 基于非固定位置置换的数据混淆

对编码后的包含  $m$  条记录和  $n$  维属性的原始数据矩阵  $X_{m \times n}$  按隐私保护的实际需求分块后，便可对其中的敏感数据采用非固定位置置换的算法进行数据混淆，以隐藏原始数据<sup>[11]</sup>。

基于非固定位置置换的数据混淆方法的基本思路：依据数据位置信息从敏感数据字符串中随机置换位信息，从而达到掩盖信息原意的目的。该方法的实现以密钥分解理论为基础，即将一条信息分解成  $n$  个分块，完全具备其中任意至少  $k(k \leq n)$  个分块时，才能恢复原信息，从而提高了敏感信息的安全性。

现对改进的基于非固定位置置换具体算法步骤描述如下：

- ①根据敏感信息属性项长度设置随机数范围；
- ②计算置换起始位置，即对于数据矩阵中的每一条记录的敏感数据，产生随机数，符合第一步的数值范围，确定置换的起始位置；
- ③系统产生一组在设置的范围内的随机数作为置换位置增量；
- ④从起始位置开始置换位信息，原位信息取反，并将位置加上随机数得到下一个置换位置，直到置换位置大于待上传敏感信息大小为止；

⑤将对应位置的位信息从敏感属性值中置换出来，与顺序排列的随机数序列一起保存，混淆后数据保存于数据矩阵中，二者隔离保存。

算法实现过程如图4所示。

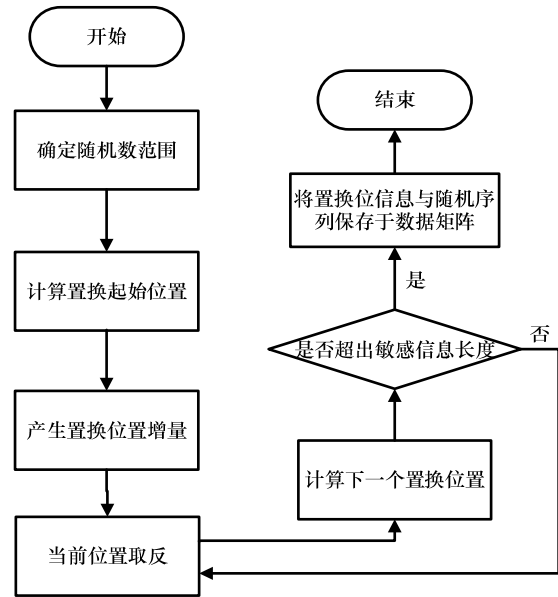


图4 非固定位置置换的数据混淆过程  
Fig.4 The process of data confusion by random character extraction

例如，有如表1的原始数据表

表1 原始数据表  
Tab.1 Original data

姓名	性别	年龄	国籍	疾病
Andy	男	23	英国	肿瘤
Hebe	女	18	美国	胃炎
Bart	男	35	加拿大	流感

此表中假设按安全等级分块，其中“姓名”、“年龄”、“疾病”属性为敏感属性，则将第一条记录敏感属性对应的二进制流进行非固定位置置换后的输出字符，变换过程如下：

Andy: 01000001 01101110 01100100 01111001  
 ↓ (起始位置: 5; 置换位置增量: 9)  
 01000101 01101100 01100101 01111001  
 (Eley)

23: 0001 0111  
 ↓ (起始位置: 2; 置换位置增量: 3)  
 0011 0011  
 (51)

肿瘤: 1101011011010111 1100000111110110  
 ↓ (起始位置: 2; 间隔长度: 4)  
 1111010011110101 1110001111010100  
 (笆摆)

其它记录敏感信息处理过程类似, 则采用非固定位置置换算法进行数据混淆产生的新数据, 如表2所示。

表2 经一次混淆后的数据表  
Tab.2 Obfuscated data

姓名	性别	年龄	国籍	疾病
Eley	男	51	英国	筳摆
hEBE	女	46	美国	甘根
FBzv	男	33	加拿大	恣藏

采用非固定位置置换算法处理后, 如输入明文为“Obfuscated”, 则在数据库中会将其显示为密文“ZjbgkhadUI”(起始位置: 3; 置换位置增量: 8), 从而便可起到有效保护敏感信息的目的。同时, 该算法是线性数据变换, 因而是可逆的, 即它是可以利用保存的随机位置序列和对应的置换信息重新逆构得到原始数据。因此, 非固定位置置换算法对原始数据起到了一次数据混淆变换的作用, 同时可以依据信息的敏感程度制定不同计算复杂度的置换规则, 从而增加攻击者的破译难度。

但是, 当待处理数据量较大时, 高计算复杂度的置换算法是不适用的, 而较简单置换算法又易于被破解。因此, 为了进一步增强隐私保护度, 本文结合随机正交矩阵的性质通过二次数据混淆来增强隐私数据的安全保护度。

### 2.3 随机正交矩阵数据混淆算法

在基于非固定位置置换算法对敏感数据进行一次数据混淆后, 此时原始数据矩阵  $X_{m \times n}$  变换为  $X'_{m \times n}$ , 然后据此进行基于随机正交矩阵数据混淆算法的二次混淆。

随机正交矩阵数据混淆算法的基本思路是: 假设矩阵  $X_{m \times n}$  是包含  $m$  条记录和  $n$  维属性的原始数据集, 矩阵  $R_{n \times n}$  表示一个随机生成的正交矩阵, 则依据  $Y = X \times R$  对原始矩阵  $X$  进行处理, 由于矩阵  $R$  的随机性, 通过发布处理后的矩阵  $Y$ , 即可实现真实数据的隐私保护处理。其中随机正交矩阵  $R$  的生成算法可参考文献[12]实现, 也可以由随机函数产生  $n \times n$  个数值生成, 其数值范围为 0~1 之间。在本文中用此思想即是对一次混淆后的数据矩阵  $X'_{m \times n}$ , 经  $X''_{m \times n} = X'_{m \times n} \times R_{n \times n}$  处理, 得到二次混淆数据矩阵  $X''_{m \times n}$ 。

具体算法描述如下:

①输入一次混淆后数据矩阵  $X'_{m \times n}$ ;

②生成随机正交矩阵  $R_{n \times n}$ ;

③利用正交矩阵的性质对数据矩阵  $X'_{m \times n}$  做变换  $X''_{m \times n} = X'_{m \times n} \times R_{n \times n}$  处理;

④变换结束放回矩阵  $X''_{m \times n}$ ;

⑤输出矩阵  $X''_{m \times n}$ , 结束。

算法实现过程如图5所示。

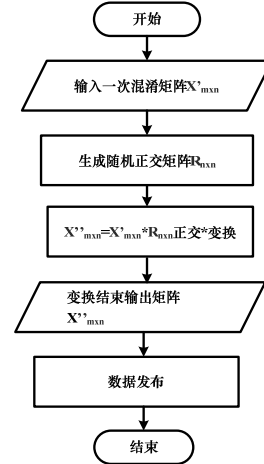


图5 随机正交矩阵数据混淆过程

Fig.5 The data obfuscation process by random orthogonal matrix

在一次数据混淆的基础上, 使用正交矩阵的性质对其进行二次数据混淆处理, 可进一步增强隐私数据保护的安全等级。

### 2.4 隐私数据混淆逆变换

在发布二次混淆后的数据矩阵集  $X''_{m \times n}$  后, 为了让数据使用者能接收到准确的数据, 需要对上述经混淆变换后的数据进行混淆逆变换重构。

逆变换过程与混淆过程相反, 本文所使用的两次混淆都是可逆变换且逆变换过程简单。其基本原理是先由矩阵  $R$  的正交性解得一次混淆后的矩阵  $X'_{m \times n} = X''_{m \times n} \times R_{n \times n}^T$ , 然后再利用保存的随机位置序列和对应的置换信息重新逆构得到原始数据。二次数据混淆可逆变换过程如图6所示。

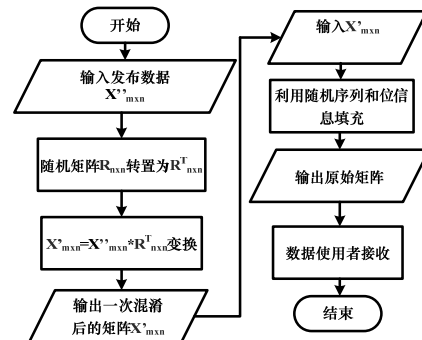


图6 混淆可逆变换过程

Fig.6 The inverse transformation for data confusion

### 3 实验论证

#### 3.1 实验环境和实验数据

CPU: i5-4200M 2.50GHz, 内存: 4GB

操作系统: Windows 8.1

MATLAB 版本: R2010a

实验中所采用的数据来源于作者科研项目。

#### 3.2 实验及分析

基于数据混淆的隐私保护方法包括原始数据序列化、隐私数据分块、重要信息数据混淆、随机正交矩阵二次混淆隐私数据等操作。实验采用本文中的基于数据混淆的隐私保护方法对数据进行处理, 分析其执行时间, 依次采取不同比例的数据进行数据隐私保护处理, 得到的执行时间如图 7 所示。从中可以看出, 由于基于数据混淆的隐私保护方法的计算复杂性低于传统的数据隐私保护算法, 随着测试数据量以及敏感属性比例的增加, 文中基于数据混淆的方法能更为有效地保护隐私数据, 降低了其信息泄露的风险, 与传统隐私数据保护方法相比, 显示出明显的优势。

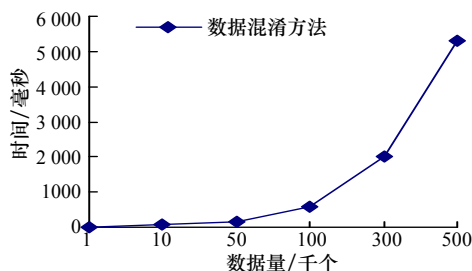


图 7 数据混淆方法中数据量与执行时间的关系

Fig.7 The relationship between data amount and execution time in data confusion

另一方面, 实验针对隐私数据中不同比例的敏感属性, 采取非固定位置置换进行数据隐私保护处理, 得到的敏感属性比例与隐私数据安全率的关系如图 8 所示。

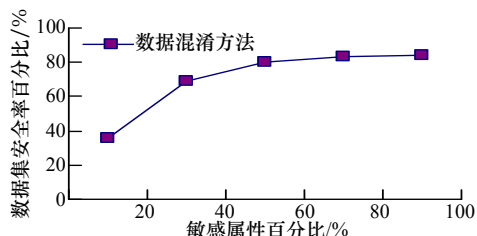


图 8 数据混淆方法中敏感属性比例与隐私数据安全率关系

Fig.8 The relationship between the proportion of sensitive attributes and privacy data security rate in data confusion

实验结果表明, 若隐私数据中敏感属性比例在 20%~60%之间, 数据安全率与敏感属性百分比接近正比例关系。当敏感属性比例超过 60%时, 数据集中数据安全度增长率渐趋于 0。

针对敏感属性, 基于数据混淆的隐私保护方法中采用非固定位置置换方法, 通过仿真实验验证了该方法的有效性、可行性和安全性。

### 4 结论

在大数据时代, 需要有更安全的保护措施才能使更多的企业和个人信任大数据应用, 进而敢于把信息交给大数据运营商, 促进大数据应用的发展。本文在分析现有隐私保护问题及方法的基础上, 提出了一种基于数据混淆的隐私数据保护方法, 该隐私数据保护方法对原始数据进行了二次混淆, 将非固定位置置换算法与随机正交矩阵混淆算法结合使用。在这种情况下, 攻击者必须同时知道置换规则和随机生成的矩阵  $R$  才能破译出原始数据, 因此加大了攻击者的攻击难度, 同时其可逆混淆过程又较为简单, 可以很容易地还原出原始数据提供给使用者, 并保证数据的准确性, 增强了隐私数据的隐私保护度, 通过实验论证也证实了该算法的有效性和可靠性。

#### 参考文献 References

- [1] KISSERLI Nessim, PRENEEL Bart. A taxonomy of self-modifying code for obfuscation[J]. Computers & Security, 2011, 30(8): 679-691.
- [2] 周晓斌, 许勇, 张凌. 一种开放式 PKI 身份认证模型的研究[J]. 国防科技大学学报, 2013, 35(1): 169-174.  
ZHOU Xiaobin, XU Yong, ZHANG Ling. Research on open identity authentication model for PKI[J]. Journal of National University of Defense Technology, 2013, 35(1): 169-174.
- [3] 刘莉苹. 基于属性的空间数据访问控制研究[J]. 计算机工程与设计, 2014, 35(3): 803-808.  
LIU Liping. Research of attributed based spatial data access control[J]. Computer Engineering and Design, 2014, 35 (3): 803-808.
- [4] GREESHMA Sarath, JAYAPRIYA R. Securing database server using homomorphism encryption and re-encryption[J]. Communications in Computer and Information Science, 2015, 536(8): 277-289.

(下转第 46 页)