软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(9):2293−2308 [doi: 10.13328/j.cnki.jos.005183]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于格的隐私保护聚类数据挖掘方法*

崔一辉[1,2], 宋 伟[1,2], 王占兵[1,2], 史成良[1,2], 程芳权[1,2]

[1](软件工程国家重点实验室(武汉大学),湖北 武汉 430072)

[2](武汉大学 计算机学院,湖北 武汉 430072)

通讯作者: 宋伟, E-mail: songwei@whu.edu.cn

摘 要: 由于云计算的诸多优势,用户倾向于将数据挖掘和数据分析等业务外包到专业的云服务提供商,然而随之而来的是用户的隐私不能得到保证.目前,众多学者关注云环境下敏感数据存储的隐私保护问题,而隐私保护数据分析的相关研究还比较少.但是如果仅仅为了保护数据隐私,而不对大数据进行挖掘分析,大数据也就失去了其潜在的巨大价值.提出了一种云计算环境下基于格的隐私保护数据挖掘方法,利用格加密构建隐私数据的安全同态运算方法,并且在此基础上实现了支持隐私保护的云端密文数据聚类分析数据挖掘服务.为保护用户数据隐私,用户将数据加密之后发布给云服务提供商,云服务提供商利用基于格的同态加密算法实现隐私保护的 *k*-means、隐私保护层次聚类以及隐私保护DBSCAN数据挖掘服务,但云服务提供商并不能直接访问用户数据破坏用户隐私.与现有的隐私数据发布方法相比,隐私数据发布基于格的最接近向量困难问题(CVP)和最短向量困难问题(SVP)具有很高的安全性.同时,有效保持了密文数据间距离的精确性.与现有研究相比,挖掘结果也具有更高的精确性和可用性.对方法的安全性进行了理论分析,并设计实验对提出的隐私保护数据挖掘方法效率进行评估,实验结果表明,提出的基于格的隐私保护数据挖掘算法与现有的方法相比具有更高的数据分析精确性和计算效率.

关键词: 数据挖掘;隐私保护;隐私保护的数据挖掘;基于格的加密

中图法分类号: TP311

中文引用格式: 崔一辉,宋伟,王占兵,史成良,程芳权.一种基于格的隐私保护聚类数据挖掘方法.软件学报,2017,28(9):2293−2308. http://www.jos.org.cn/1000-9825/5183.htm

英文引用格式: Cui YH, Song W, Wang ZB, Shi CL, Cheng FQ. Privacy preserving cluster mining method based on lattice. Ruan Jian Xue Bao/Journal of Software, 2017,28(9):2293−2308 (in Chinese). http://www.jos.org.cn/1000-9825/5183.htm

Privacy Preserving Cluster Mining Method Based on Lattice

CUI Yi-HUI[1,2], SONG Wei[1,2], WANG Zhan-Bing[1,2], SHI Cheng-Liang[1,2], CHENG Fang-Quan[1,2]

[1](State Key Laboratory of Software Engineering (Wuhan University), Wuhan 430072, China)

[2](Computer School, Wuhan University, Wuhan 430072, China)

Abstract: Due to the various advantages of cloud computing, users tend to outsource data mining task to professional cloud service providers. However, user's privacy cannot be guaranteed. Currently, while many scholars are concerned about how to protect sensitive data from unauthorized access, few scholars engage research on data analysis. But if potential knowledge cannot be mined, the value of big data may not be fully utilized. This paper proposes a privacy preserving data mining (PPDM) method based on lattice, which support

* 基金项目: 国家自然科学基金(61232002, 61572378, 61202034); CCF 中文信息技术开放课题(CCF2014-01-02); 武汉市创新团队项目(2014070504020237); 武汉大学自主科研项目(2042016gf0020, 2016-2017)

Foundation item: National Natural Science Foundation of China (61232002, 61572378, 61202034); CCF Chinese information technology open topic (CCF2014-01-02); Wuhan Innovation Team Project (2014070504020237); Wuhan University independent research project(2042016gf0020, 2016-2017)

收稿时间: 2016-07-10; 修改时间: 2016-09-04; 采用时间: 2016-11-10; jos 在线出版时间: 2017-02-20

CNKI 网络优先出版: 2017-02-20 14:02:19, http://www.cnki.net/kcms/detail/11.2560.TP.20170220.1402.011.html

ciphertext intermediate point and distance homomorphic computing. Meanwhile, it builds a privacy preserving cloud ciphertext data clustering data mining Method. Users encrypt privacy data before outsource the data to cloud service providers, cloud service providers use homomorphic encryption to achieve privacy protection mining algorithms including k-means, hierarchical clustering and DBSCAN. Compared with the existing PPDM method, the presented method with high security is based on shortest vector difficulties (SVP) and the closest vector problem (CVP). Meanwhile, it maintains the accuracy of distance between two data, providing mining results with high accuracy and availability. Experiments are designed for the privacy preserving cluster mining (PPCM) with cardiac arrhythmia datasets of machine learning, and the experimental results show that the method based on lattice ensure not only security but also accuracy and performance.

Key words: data mining; privacy preserving; privacy preserving data mining (PPDM); lattice-based cryptography

随着云计算的发展和普及,用户越来越倾向于将数据存储到云上,租用云计算服务中心提供的丰富计算和存储资源,能够为用户提供更加高效、专业的数据分析服务.然而,云服务提供商往往并不完全可信,用户在享用云计算高效服务的同时,用户隐私数据也直接暴露给云服务中心,因此,数据隐私安全问题就成为用户使用云计算不得不考虑的首要问题.为解决云计算外包服务模式下的用户数据安全问题,现有研究更多地致力于解决用户隐私数据的存储安全,如密文可搜索机制、数据验证等.这些研究大多适用于查询、授权访问等简单应用场景,而并不适用于数据聚类挖掘、关联分析等复杂应用场景.而数据需要隐藏发布的根本原因在于数据的潜在价值,隐私数据发布后,数据的深层可用性对数据隐藏技术的发展和成熟至关重要[1].

用户将数据挖掘分析任务外包到云上,不可避免地会涉及到敏感数据,如何操作、分析这些数据的同时保护用户隐私,成为必须要解决的问题[2,3].目前,众多学者主要致力于服务提供商可信前提下的数据分析和处理研究,但这种假设在云计算的外包服务环境下并不成立,比如云服务提供商出于商业利益的驱使对用户隐私数据的滥用、云服务管理员出于好奇窥视用户隐私信息、云服务提供商遭遇黑客攻击等,都会导致用户隐私泄露.

2000年,Agrawal首次提出了隐私保护数据挖掘(privacy preserving data mining,简称PPDM)的概念[4].如图1所示,隐私保护数据挖掘是指用户将数据外包给云服务提供商存储管理,由云服务提供商在确保用户隐私的前提下进行数据挖掘服务,并将有价值的挖掘结果返回给用户.为保护数据隐私,PPDM 中隐私数据本身和数据挖掘结果对于任何非授权第三方都不可见.例如:医疗机构获得了大量第一手的医疗大数据,其需要对医疗大数据进行聚类分析,进而得到潜在的疾病聚簇以及疾病治疗方案之间的相互关联性.显然,各个医疗机构数据挖掘的能力相对有限,往往需要将原始医疗数据发布给专业的云服务提供商,并委托其按照需要对大数据进行挖掘和分析,这时,病患的隐私成为不得不考虑的问题.保证用户数据的安全性和保障数据挖掘结果的准确性往往相互矛盾,现有数据挖掘算法和研究往往没有考虑云环境下云服务提供商不可信的情况,因此,如何在保证用户隐私的前提条件下提高挖掘结果的准确性,是个非常具有挑战性的问题.

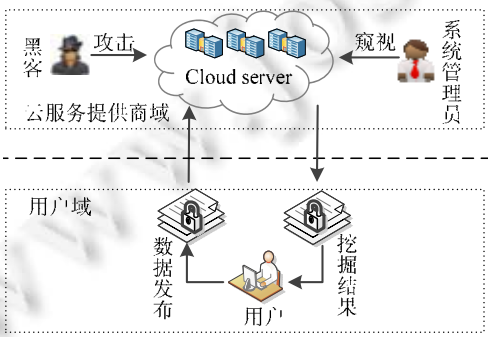

Fig.1 Data mining problem of privacy preserving

图1 隐私保护数据挖掘问题

没有任何一种隐私保护技术适用于所有应用场景[5].针对不同的应用场景,众多学者提出了很多高效的方法[6−8].目前,关联规则和决策树挖掘等隐私保护挖掘研究相对较多[9−11],而聚类挖掘由于涉及计算量较大,支持

隐私保护的聚类挖掘研究还相对较少.本文主要研究支持隐私保护的高精确性聚类挖掘问题,在保护用户数据隐私的前提下,提供高精度的聚类挖掘服务.Zhang 等人认为:现有研究在云端进行数据处理过程中,产生的中间临时数据存在着安全隐患[12].本文提出的算法很好地解决了这个问题.针对现有研究安全性难以保障或者数据挖掘精确性不足的问题,本文提出了一种基于格的保持距离加密算法,该加密算法本身具有计算中心点和计算密文距离的同态性质,并在此基础上设计了云服务环境下隐私保护的聚类挖掘算法,使得云服务提供商在不解密的前提下实现数据挖掘操作,这样可以在保护用户隐私的同时提升数据的可用性,在密文环境下,可以更加精确地进行数据分析,从而获得更多有价值的信息.论文提出的隐私保护数据挖掘方法可以容易地扩展到其他数据挖掘应用场景,文中基于提出的方法实现了 PPK-means(privacy preserving *K*-means)、隐私保护层级聚类以及 PPDBSCAN(privacy preserving DBSCAN)方法,实现云端密文数据环境下的精确数据挖掘服务.

本文的主要贡献包含以下 4 方面的内容.

(1) 在 GGH[13]公钥密码体制的基础上设计了一种基于格的保持距离加密算法,算法安全性基于格的最近向量求解困难问题和最短向量困难问题,安全性得到有效保障.基于格的困难问题与大素数难分解等问题相比性能较好,而且加密方式属于批处理加密方式,更加适合大数据的元组存储形式,每次加密以行为单位而非将元组中的元素逐列进行处理,可以提高隐私数据的加密发布效率;

(2) 在提出的保持距离加密算法基础上,设计了密文状态下的数据均值计算方法和密文数据之间的距离计算方法.与现有的聚类方法相比,论文提出的加密算法保持密文隐私数据计算与明文计算保持同态特性,使得隐私数据挖掘的精确性大大提升;

(3) 本文设计的隐私保护聚类挖掘框架可以容易地扩展实现多种聚类挖掘算法,如 *k*-means、层次挖掘以及 DBSCAN 等.而现有的隐私保护挖掘算法大多数只能支持某一种特定挖掘算法,因此,本文提出的方法具有更好的可扩展性,可以更好地支持复杂应用需求,从多个角度挖掘出潜在的聚类关系;

(4) 本文进一步构建了隐私保护的 kd-tree 结构,在保证数据隐私的前提下,高效地完成最近邻检索、KNN 查询和范围查询,实现密文环境下高效的隐私保护 DBscan 等聚类挖掘算法.

1 相关工作

隐私保护的聚类挖掘从数据存储方式上可以分为云服务提供商存储和用户存储,从隐私模型的角度包括差分隐私、基于平移变换、基于水印、基于匿名和基于加密等,在挖掘方法上面支持的挖掘算法也不尽相同,数据类型主要包括字符型和数值型两种.随着云存储的快速发展,本文主要致力于解决数据存储在云服务提供商场景的隐私保护数据挖掘问题研究.隐私保护数据挖掘研究现状见表 1.

Table 1 Research status of privacy preserving data mining

表 **1** 隐私保护数据挖掘研究现状

典型算法			Allard[7]	Xun Yi[14]	Vlachos[8]	*k*-匿名	Oliveira[15]
数据存储方式	云服务提供商				√	√	√
	用户存储	分布式		√			
		集中式	√				
隐私模型	差分隐私		√				
	平移变换						√
	基于水印				√		
	基于匿名					√	
	加密			√			
	基于噪音		√				
挖掘方法	*k*-means			√		√	√
	DBSCAN					√	√
	层次聚类				√	√	√

隐私保护的数据挖掘过程分为两个阶段:首先,用户需要对原始数据进行隐私保护的数据发布;然后,云服务提供商需要在经过隐私保护处理的数据上进行数据挖掘.隐私保护的数据挖掘算法大致分为两类:一类是对

数据本身进行泛化扰动或者添加噪音进而保护数据隐私[16−18,19],该类方法的优点在于避免了复杂的处理过程,方法性能相对较好;另一类通过对数据本身进行平移变换到另一个向量空间,以确保数据隐私[10],此类方法的优点是挖掘的精度往往比较高.

第 1 类主要使用泛化和添加噪音的方式进行隐私保护的数据挖掘.

k-匿名方案[20]和 *l*-多样化方案[21]主要是通过数据的泛化来隐藏用户的隐私,其数据发布前后见表 2、表 3,抵御了连接攻击,在一定程度上可以很好地解决数据发布中的数据隐私保护问题,当数据仅仅为了简单应用时,这两种方案简单有效.但是进行聚类数据挖掘时,由于数据被泛化,挖掘分析的结果不是很精确.例如在原始数据上进行数据挖掘,挖掘出的聚簇数目为 3,但是由于数据的泛化,挖掘出的聚簇数目为 2.文献[22]提出了基于差分隐私的聚类技术,有效地在保证用户隐私的条件下进行挖掘.Mohan 认为:尽管差分隐私提出了一种理论上可以在保护用户隐私数据记录前提条件下数据处理,但是其实际使用的输出结果存在精度丢失问题[23].现实中,精确挖掘的需求不容忽视,尤其是在医学生物以及银行商业金融领域.文献[16]提出了一种通过增加噪音和构建决策分类树来进行隐私保护聚类挖掘的方法,因为辅助信息决策分类树的存在,挖掘性能有较大提升,但是用户的隐私也会有一定程度的泄露.文献[18]提出的 NeSDO 算法采用合成数据替换真实数据,发布合成数据进行数据挖掘,数据发布的性能较好,挖掘精度由于合成数据的原因受到一定影响.且合成数据是邻域数据的均值,也会泄露一定的数据隐私.文献[24]提出的基于组内数据交换方法的挖掘方法,其优势主要体现在操作相对简单性能最佳,但是挖掘精确性和安全性会受到一定影响.文献[19]通过扰动原始数据从而达到隐私保护的目的,并且在此技术上进行数据挖掘.

Table 2 The raw data

表 **2** 原始数据

Id	Age	Weight	Heart_rate	Disease
1237	78	80	76	心脏病
3420	79	78	79	心脏病
2543	69	66	67	心脏病
4461	68	67	66	心脏病
2543	61	64	64	心脏病
4461	62	63	63	心脏病

Table 3 The release data

表 **3** 发布数据

Id	Age	Weight	Heart_rate	Disease
1237	71~80	71~80	71~80	心脏病
3420	71~80	71~80	71~80	心脏病
2543	61~70	61~70	61~70	心脏病
4461	61~70	61~70	61~70	心脏病
2543	61~70	61~70	61~70	心脏病
4461	61~70	61~70	61~70	心脏病

第 2 类主要使用多维向量空间映射方法.

RBT 算法[15]是目前隐私保护的聚类挖掘精度最高的算法,该算法的隐私保护数据发布实质是将元组数据进行平移旋转,平移旋转保证了数据的精度,但是其安全性相对较低,且明文和密文一一对应,会遭遇统计攻击.文献[25]结合算法 *k*-means 设计了 Pk-means 方法进行聚类挖掘,另外,Pk-means 支持的聚类算法比较单一,其实现了一个隐私保护的 *k*-means 算法.事实上,每一种聚类挖掘算法都有该算法适合的聚簇类型,单一的挖掘算法 Pk-means 不能满足所有类型的聚类挖掘需求.全同态加密为动态数据隐私保护提供了一种理论支持,但其在解决任意计算方面的实用性还存在很大的差距.因此,数据隐私保护的研究主要还是针对特定的计算类型,应用同态加密方法研究代价可以被实际应用所接受的隐私保护机制[2].

综上所述,现有的算法往往存在以下问题:第 1 类方法在性能上较好,挖掘精度存在一定问题;第 2 类方法在精确性上优于第 1 类方法,但是安全性上存在一定问题.另外,部分算法只能针对一种挖掘算法进行挖掘,云计算

环境下,亟需一种能够适应多种挖掘算法的隐私保护数据发布方式和数据挖掘方法,并且兼顾安全性、精确性和挖掘性能.

2 准备知识

本节介绍论文中使用的密码学工具和基本定义.论文中使用的符号见表 4.

Table 4 Notations

表 **4** 符号定义

符号	符号定义
K_{pub}	用户公钥,用户公开自己的公钥给云服务提供商,后者利用 K_{pub} 进行隐私保护的同态计算
K_{pri}	用户私钥,用户使用自己的私钥处理云服务提供商返回的聚类结果,得到明文结果
$M_{\otimes}$	扩展高维向量积运算矩阵
$\vec{p}$	云服务提供商求解距离的辅助向量
ψ	衡量优质基的阈值参数
dis_{e_1,e_2}	密文状态下数据向量 e_1 和 e_2 之间的欧式距离
m_i	数据集中的第 i 明文向量
e_i	密文元组(s_i,t_i)
∂	隐私保护的 k-means 算法中迭代结束条件阈值
$\wedge$	为单位矩阵
$code$	向量 m 经过混淆编码生成的向量

定义 **1**. 线性独立空间上有向量集合 $v_1,v_2,\ldots,v_n\subset R^m$,格(lattices)就是这些向量的线性组合,格 L 的维数(n)等于构成格的线性不相关向量的个数[13]:

$$L=\{a_1v_1+a_2v_2+\ldots+a_nv_n|a_i\in Z,v_i\subset R^m,1\leq i\leq n\}.$$

定义 **2**. 可以采用 Hadamard 比率来衡量格中基向量的正交程度,对于格 L 的一组基向量$\{v_1,v_2,\ldots,v_n\}$,两两正交程度较高,即 $Hadamard(v_1,v_2,\ldots,v_n)$值大于阈值$\theta$时,称$\{v_1,v_2,\ldots,v_n\}$就是格 L 的一组优质基[13].

定义 **3**. 当格 L 的一组基向量$\{v_1,v_2,\ldots,v_n\}$两两正交程度较低,即 $Hadamard\{v_1,v_2,\ldots,v_n\}$值小于阈值$\theta$时,称$\{v_1,v_2,\ldots,v_n\}$就是格 L 的一组劣质基[13].

根据格基的等价变换原理,由一个优质的基转化为一个劣质的基很容易,相反则很困难.本文利用格加密的这一性质,由用户掌握格的优质基作为密钥,生成劣质基供云端对数据进行聚类挖掘计算,但是云端并不能直接解密访问用户隐私数据,从而在隐私保护前提下实现聚类数据挖掘服务.

定义 **4**(扩展的高维向量积). 向量积通常存在于三维向量,论文扩展三维向量积到多维向量积.对于任意 n 维向量 $\vec{a}=(a_1,a_2,...,a_n),\vec{b}=(b_1,b_2,...,b_n)$,记 $\vec{a}\otimes\vec{b}$ 为向量 $\vec{a},\vec{b}$ 的扩展向量积,其中,

$$\vec{a}\otimes\vec{b}=M_{\otimes}\begin{bmatrix}b_1\\b_2\\\ldots\\b_n\end{bmatrix},M_{\otimes}=\begin{bmatrix}-a_2 & a_1 & 0 & \ldots & \ldots & \ldots & 0\\ 0 & -a_3 & a_2 & \ldots & \ldots & \ldots & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \ldots & \ldots & \ldots & -a_{i+1} & a_i & \ldots & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \ldots & \ldots & \ldots & \ldots & -a_n & a_{n-1} & \ldots\\ -a_1 & \ldots & \ldots & \ldots & \ldots & \ldots & a_n\end{bmatrix}.$$

由上述定义可知:当两个高维向量 $\vec{a},\vec{b}$ 方向相同时,其高维向量积 $\vec{a}\otimes\vec{b}=0$.文中利用这一性质实现云端加密隐私数据的同态精确计算.

定义 **5**(隐私保护 ***k*-*d*** 树). 每个节点都为 k 维点的二叉树.所有非叶子节点可以视作用一个超平面把空间分区成两个半空间(half-space).节点左边的子树代表在超平面左边的点,节点右边的子树代表在超平面右边的点.本文在混淆编码的基础上由数据拥有者构建的 k-d tree 在保证用户隐私的前提下提高数据挖掘的性能.

定义 **6**. 吉文斯旋转表示为如下形式的矩阵(乘积 $G(i,j,\theta)\cdot x$ 表示向量 x 在 $(\vec{i},\vec{j})$ 平面中的逆时针旋转θ弧度.这里的 c=cos(θ)和 s=sin(θ)出现在第 i 行和第 k 行与第 i 列和第 k 列的交叉点上.$G(i,j,\theta)$中,非零元素的描述为

$g_{ii}=c,g_{jj}=c,g_{ij}=-s,g_{ji}=s$):

$$G(i,j,\theta)=\begin{bmatrix}1 & \cdots & 0 & \cdots & 0 & \cdots & 0\\ \vdots & \ddots & \vdots & & \vdots & & \vdots\\ 0 & \cdots & c & \cdots & -s & \cdots & 0\\ \vdots & & \vdots & \ddots & \vdots & & \vdots\\ 0 & \cdots & s & \cdots & c & \cdots & 0\\ \vdots & & \vdots & & \vdots & \ddots & \vdots\\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1\end{bmatrix}.$$

3 基于格的保持相对距离数据加密方法

本文提出的隐私保护聚类挖掘方法(privacy preserving cluster mining,简称 PPCM)系统架构如图 2 所示.用户为保护数据隐私,利用加密算法对数据进行加密.为了支持云端的聚类挖掘分析,论文在 GGH 的基础上设计了一种基于格的保持相对距离加密算法,确保云端对密文进行距离运算与明文的距离保持同态特性,云服务提供商按照用户的聚类挖掘请求对云端的加密隐私数据进行聚类分析,并返回密文分析结果,用户利用个人私钥对分析结果进行解密.

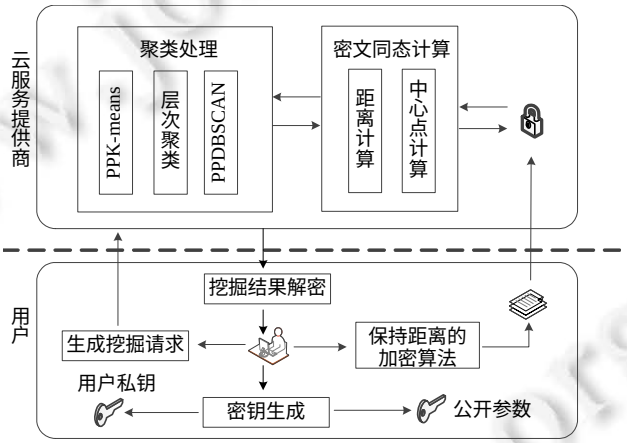

Fig.2 PPCM architecture

图 2 PPCM 系统架构

PPCM 利用基于格的距离同态加密算法实现,基于格的距离同态加密算法由以下 5 部分组成:第 3.1 节介绍 setup 过程、系统公开参数和用户密钥的生成方法,每个用户与云服务提供商使用同一公开参数;第 3.2 节介绍向量的混淆编码(confusion coding),并在此基础上构建隐私保护 *k-d* tree;第 3.3 节基于 GGH 公钥密码体制设计了基于格的保持距离的隐私保护数据处理算法(Enc),将数据外包存储在云服务提供商;第 3.4 节介绍用户获得云端密文挖掘结果后的明文数据解密过程(Dec);第 3.5 节和第 3.6 节介绍云服务提供商密文状态下的同态运算.

3.1 用户密钥和公开参数生成(**setup**)

用户需要生成用户密钥 K_{pri},K_{pub} 以及公开参数 $\tilde{p}$.用户选择优质基 $K_{pri}=v_1,v_2,\ldots,v_n$ 以及随机整数矩阵 U,满足 $\det(U)=\pm1$.由于使用随机算法求解满足 $\det(U)=\pm1$ 的矩阵时间开销很大,本文采用对单位矩阵进行初等变换的方法生成转换矩阵 U.对单位矩阵进行交换行或者交换列的初等变换矩阵的行列式等于原行列式乘以–1.将某一行或者某一列的倍数加到其他的行或者列,行列式不变.左乘 A 为对单位矩阵∧做随机的初等行变换,右乘矩阵 B 对矩阵单位矩阵∧做随机的初等列变换.本算法涉及的初等变换不包括对于矩阵某一行(列)乘以 k.我们得到一组劣质基 $K_{pub}=UK_{pri}=w_1,w_2,\ldots,w_n.K_{pub}$ 即为用户的公钥.用户将私钥 K_{pri} 安全存储在本地,公钥 K_{pub} 提交

给云服务提供商.安全性与矩阵中元素的取值范围有关,矩阵中元素取值范围越大算法越安全可靠,但是随着矩阵中元素大小的增加,也会导致性能的下降.

算法 **1**. 密钥生成 *keyGenerate*(ψ,n).

输入:ψ是判断优质基的阈值,n 是密钥维度;

输出:K_{pub},K_{pri}.

1: While (true) {

2: 随机选取 $v_1,v_2,\ldots,v_n$;

3: if (*Hadamard*(v_1,v_2,v_3)>阈值 ψ) {

4: $K_{ri}=v_1,v_2,\ldots,v_{n-1},v_n$;

5: break;

6: }

7: $U=A\wedge B$ //$\wedge$为单位矩阵.

8: $K_{pub}=UK_{pri}$

9: returnK_{pri} 和 K_{pub}.

云服务提供商在聚类挖掘的过程中需要求解向量之间的距离,加密的密文包含随机数.如何在计算距离的过程中抵消掉随机数的影响,获得一个相对准确的距离用于聚类挖掘,是数据处理过程中需要解决的重要问题.三维空间中具有如图 3 所示性质:$(r_1-r_2)(l_1,l_2,l_3)=\{0,0,0\}$,即,选定直线上的随机向量之差和直线的方向向量的向量积为 0.本文将三维空间的向量积性质扩展高维,利用本文定义的高维向量积在云服务商求解距离的过程中抵消掉随机数,进而计算出密文向量之间的相对距离,此距离在混淆编码过程中进行了缩放.

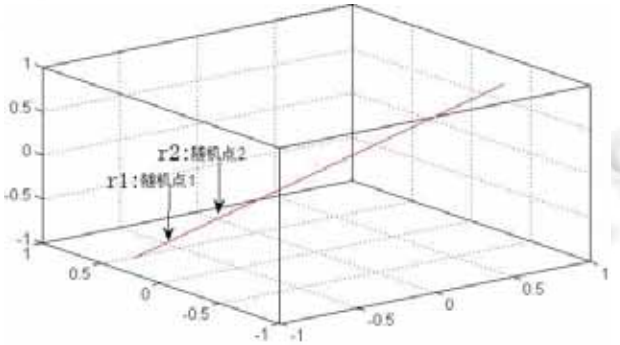

Fig.3 Three dimensional vector space

图 3 三维向量空间情形

随机数生成公式如下所示(初始化过程中生成 $\vec{p}=\{l_1,l_2,...,l_{n-1},l_n\}$ 为生成随机数所需向量,此向量作为公开参数上传云服务提供商,云服务提供商借助其完成聚类挖掘所需各类同态运算).

高维随机向量生成公式:

$$(r_1,r_2,...,r_n)=\begin{Bmatrix} r_1=r_1'+l_1t \\ r_2=r_2'+l_2t \\ ... \\ r_{n-1}=r_{n-1}'+l_{n-1}t \\ r_n=r_n'+l_nt \end{Bmatrix},$$

其中, $r_1',r_2',...,r_{n-1}',r_n'\neq 0$.

3.2 混淆编码和隐私保护*k-d* tree生成

首先,我们对原数据进行混淆编码,混淆编码本质上是一个保距映射(保持相对距离),经过混淆的原始数据实际上进行了变换和缩放,其目的是抵御背景知识攻击,经过混淆编码的向量数据相对位置不发生改变,真实距离转换成为相对距离,因此不改变聚类特性.云服务提供商完成聚类挖掘,将挖掘结果返回给用户,用户首先解密聚类结果,然后进行逆混淆编码得到聚类结果.数据发布的过程中,列名同样是需要隐藏的元素,以降低暴力破解的攻击风险.但攻击者确定某一列是血压时,随机选择某一密钥进行解密,当其将数据解密后发现在正常的血压范围里面,随机选择的密钥就有可能是解密密钥;但是如果无法确定选择的列的属性,就很难确定当前选择密钥是否正确.本质上讲,混淆编码过程首先进行一系列的吉文斯旋转,然后进行缩放,进而进行向量的交换混淆,最后进行平移变换.T_{pq} 将一单位矩阵的第 p 行的所有元素与第 q 行互换,这一变换 $T_i(k_i)$,将第 i 行的所有元素乘以非零常数 k_i.

混淆编码过程 $code=key_1 m+key_2$,其中, $key_1=\prod T_{pq}\prod_{i=1}^{n}T_i(k)\prod G(i,j,\theta)$,$key_2=[k_1,k_2,\ldots,k_{n-1},k_n]^T$.

在混淆编码的数据上面构建的 *k-d* tree,需要证明经过混淆编码向量向量仍旧保持了了其相对位置.显而易见,缩放变换、元素交换以及平移变换不改变元素之间的相对位置,下面我们证明一系列的吉文斯变换不会改变向量之间的相对位置.

性质 **1**. 经过一系列吉文斯变换,向量之间的距离不变.

证明:设向量 $\vec{x}=(x_1,x_2,...,x_n),\vec{y}=(y_1,y_2,...,y_n)$,将向量 $\vec{x},\vec{y}$ 分别执行一轮吉文斯变换,得到 $\vec{x}',\vec{y}'$,即:

$$\begin{cases}\vec{x}'=G(i,j,\theta)\cdot\vec{x}=(x_1,...,x_i\cos\theta-x_j\sin\theta,...,x_i\sin\theta+x_j\cos\theta,...,x_n)\\ \vec{y}'=G(i,j,\theta)\cdot\vec{y}=(y_1,...,y_i\cos\theta-y_j\sin\theta,...,y_i\sin\theta+y_j\cos\theta,...,y_n)\end{cases},$$

则转换过后二者终点的距离:

$$\begin{aligned}d_1(\vec{x}',\vec{y}')&=\sqrt{\sum_{k=1}^{n}(x_k'-y_k')^2}\\&=\sqrt{\sum_{k\neq i,j}^{n}(x_k'-y_k')^2+(x_i'-y_i')^2+(x_j'-y_j')^2}\\&=\sqrt{\sum_{k\neq i,j}^{n}(x_k'-y_k')^2+[(x_i-y_i)\cos\theta-(x_j-y_j)\sin\theta]^2+[(x_i-y_i)\sin\theta+(x_j-y_j)\cos\theta]^2}\\&=\sqrt{\sum_{k\neq i,j}^{n}(x_k-y_k)^2+(x_i-y_i)^2+(x_j-y_j)^2}\\&=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}=d_1(\vec{x},\vec{y}).\end{aligned}$$

综上证明:经过一轮吉文斯变换,向量之间的距离不会发生变换.同理可证:经过一系列吉文斯变换,向量的距离不会发生变化.所以在混淆编码上面构建的 *k-d* tree,在保证安全的同时保持了其索引特性. □

3.3 用户加密发布数据(**Enc**)

本文中,明文多维数据以向量的形式存储,对于明文高维向量 m,用户通过高维随机向量生成 r,r 的模应该小于格中向量之间距离的一半.用户使用自己的公钥 K_{pub} 加密 m,生成的密文二元组(s_i,t_i)由两部分组成(第 1 部分 s_i 用于解密过程还原明文数据;第 2 部分 t_i 为 $K_{pub}m$ 的哈希值,主要用于密文状态下的辅助同态计算过程):

$$e_i=(s_i,t_i)=(k_{pub}\cdot m_i+r_i,hash_M(k_{pub}\cdot m_i)) \quad (1)$$

M 为哈希矩阵,矩阵的秩为 1,见公式(2).$M_{hash}(K_{pub}\cdot m)$的哈希过程是单向的,对于该哈希等式的求解,满足条件的解有无数多个.对于同一个数据集,M_{hash} 的取值应满足公式(3),云服务提供商才可以计算密文距离:

$$M_{hash}=\begin{bmatrix} h_1 & h_2 & ... & h_{n-1} & h_n \\ 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{n\times n} \quad (h_i \subset Z) \tag{2}$$

$$rank([M_{hash};M_{\otimes}])==n \tag{3}$$

3.4 用户解密还原挖掘结果(Dec)

云服务提供商将挖掘结果返回给用户,此时,挖掘结果是密文状态的,防止了云服务提供商的非授权访问.用户得到挖掘结果需要进行如下操作,最终获得明文挖掘结果:

$$v = round(S_i K_{pri}^{-1}) \cdot K_{pri},$$

$$code = v \cdot K_{pub}^{-1}.$$

用 Babai 算法计算最近向量 v 最接近 s_i,再次计算 $v \cdot K_{pub}^{-1}$,即得到明文编码 $code$.

逆混淆编码过程还原混淆编码数据为原始数据:

$$m = key_1^T (code - key_2).$$

3.5 云服务提供商距离的密文计算

在隐私保护聚类挖掘的过程中,云服务提供商需要计算密文之间的距离.对于一对密文向量(s_1,t_1)及(s_2,t_2),密文的计算过程如下:

$$s_1=K_{pub}\cdot m_1+r_1, t_1=hash_M(K_{pub}\cdot m_1) \tag{4}$$

$$s_2=K_{pub}\cdot m_2+r_2, t_2=hash_M(K_{pub}\cdot m_2) \tag{5}$$

等式(4)减等式(5),得:

$$s_1-s_2=K_{pub}\cdot(m_1-m_2)+r_1-r_2 \tag{6}$$

$$t_1-t_2=M_{hash}\cdot(K_{pub}\cdot m_1)-M_{hash}\cdot(K_{pub}\cdot m_2)=M_{hash}\cdot(K_{pub}\cdot(m_1-m_2)) \tag{7}$$

对公式(6)两边分别乘以 $\vec{p}=\{l_1,l_2,...,l_{n-1},l_n\}$,由定义 4 扩展的高维向量积的性质可得等式(8):

$$\vec{p}\cdot(S_1-S_2)=\vec{p}\otimes(K_{pub}\cdot(m_1-m_2))=M_{\otimes}\cdot(K_{pub}\cdot(m_1-m_2)) \tag{8}$$

$\vec{p}$ 与 K_{pub} 对于云服务提供商已知,又由等式(8)结合等式(7)可以求出 $K_{pub}\cdot(m_1-m_2)$,K_{pub} 对于云端已知,进而可以得到(m_1-m_2),最终可以求出 m_1 和 m_2 的欧式距离.

$\vec{p}\times(S_1-S_2)$对于云服务提供商已知,且云服务提供商掌握 $\vec{p}$,但是在等式(8)中,$\vec{p}$ 生成的矩阵的值为 0,所以依靠密文二元组的第 1 部分无法求解出 $K_{pub}\cdot(m_1-m_2)$.当直线取值满足等式(3)时,由等式(7)和等式(8)可以求解出唯一的 $K_{pub}\cdot(m_1-m_2)$,进而求解出唯一的(m_1-m_2),最终求得欧氏距离.

算法 **2**. *computeDistance*.

Input:密文元组 e_1,e_2;

Output:元组之间的欧式距离 dis_{e_1,e_2}.

1: (S_1-S_2)

2: $\vec{p}\times(S_1-S_2)$

3: 等式(7)和等式(8)可求解出 $K_{pub}\cdot(m_1-m_2)$

4: 左乘 K_{pub}^{-1} 即可得到 m_1-m_2

5: $dis_{e_1,e_2}=computeEuclidean(m_1-m_2)$

6: return dis_{e_1,e_2};

3.6 均值密文计算

对于明文 $m_1,m_2,\ldots,m_n$,加密处理后如下所示:

$$(e_1,e_2,...,e_n)=\begin{Bmatrix} e_1=(K_{pub}\cdot m_1+r_1\ ,hash_M(K_{pub}\cdot m_1)) \\ e_2=(K_{pub}\cdot m_2+r_2,hash_M(K_{pub}\cdot m_2)) \\ ... \\ e_n=(K_{pub}\cdot m_n+r_n,hash_M(K_{pub}\cdot m_n)) \end{Bmatrix}.$$

密文均值计算公式:

$$e_{mid}=\left\{\sum_{i=1}^{n}s_i\Big/n,\sum_{i=1}^{n}t_i\Big/n\right\}.$$

性质 **2**. 密文状态下均值运算具有同态性质.

证明:明文的中心点为 $m_{mid}=(\sum_{i=1}^{n}m_i)/n$,由于 $(\sum_{i=1}^{n}r_i)/n$ 仍为直线上的点,所以中心点的计算具有同态性质.密文状态下的计算结果与直接对均值进行隐私保护的数据发布处理,结果是一致的.

$$s_{mid}=(s_1+s_2+...+s_n)/n=K_{pub}\left(\sum_{i=1}^{n}m_i\Big/n\right)+\sum_{i=1}^{n}r_i\Big/n=K_{pub}\left(\sum_{i=1}^{n}m_i\Big/n\right)+r=K_{pub}\cdot m_{mid}+r,$$

$$t_{mid}=(t_1+t_2+...+t_n)/n=hash_M\left(K_{pub}\left(\sum_{i=1}^{n}m_i\Big/n\right)\right)=hash_M(K_{pub}\cdot m_{mid}).$$ □

3.7 性能分析和安全分析

与现有的方法相比,由于本文采用基于格的困难问题的批处理方式进行数据发布,与基于大素数难分解等困难问题进行数据发布的方式相比性能较好.密文 *k*-mean 算法的时间复杂度是 $O(n\log(n))$,密文层次聚类的时间复杂度为 $O(n^2)$,密文 DBSCAN 时间复杂度 $O(n)$.由于本文构建了隐私保护的 kd-tree,密文挖掘过程中处理性能有很大的提升.

最短向量问题(the shortest vector problem,简称 SVP)是寻找一个格 L 中最短的非零向量.即,寻找一个 $v\in L$ 满足其欧几里德范数$||v||$最小.最接近向量问题(the closest vector problem,简称 CVP)是对于一个非格 L 中的向量 w,在格中寻找一个向量 v,使得$||w-v||$最小.CVP 和 SVP 都是 NP 完备问题,因此求解起来十分困难.因此,这两个问题都是可以作为密钥体制的基础的.论文中对于明文 m,其密文 $S_i=K_{pub}\cdot m_i+r_i$,其中,公钥 K_{pub} 是一组劣质级,云端掌握 S_i 和 K_{pub},求解明文 m_i 必须计算出 $K_{pub}\cdot m_i$.$K_{pub}\cdot m_i$ 求解是一个最接近向量求解问题,这是一个 CVP 问题,可以被证明是 NP 问题.

其次,算法每次加密元组使用的 r_i 都是随机,因此,即使是同一明文,两次不同的加密结果也不一样,从而避免了统计攻击.由于加密过程中随机数 r 的存在,攻击者即使获得若干组明文密文对,也无法通过其对应关系获取用户的隐私数据明文信息,或者依靠对应关系计算出用户私钥.因此,本算法可以抵御选择明文攻击(chosen-plaintext attack).本文算法与 RBT[15]的平移旋转相比安全性更高,PBT 中,数据的安全性依赖于θ,明文与密文有一一对应关系,当泄露一组明文密文对时,θ就会泄露,所以 PBT 中的隐私保护处理方法不能抵御选择明文攻击.

当敌手 A 通过公钥无法计算得到算法私钥,该算法即具有不可攻破性.而本文方法对于敌手 A 掌握公开参数 K_{pub},求解 K_{pri} 需要求解最短向量问题,该问题是 NP 问题,因此,本文的算法具有不可攻破性.密钥取值位数在生成密钥的时候可以调节,当密钥长度是 1 024 位,暴力破解需要 2^{1024} 次枚举,安全性得到了一定的保证.

4 基于 *k-d* tree 的隐私保护聚类挖掘方法(PPCM)

4.1 基于密文的*k-d* tree

隐私保护的聚类挖掘过程中,图 4 所示的 *k*-mean 需要求解出指定范围中的元素,图 5 所示的 dbscan 需要求解核心点.该问题可以转化为求解 KNN 问题,即,k 最近邻是否在指定半径内.图 6 所示的层级聚类需要求解最近邻问题,事实上就是 KNN 问题的特殊情况.构建高效的高维密文索引,可以高效地进行范围查询和 KNN 查询,是个亟需解决的困难问题.本文在混淆编码的基础上构建了密文 *k-d* tree,用于聚类挖掘中的性能优化.用户将数据上传到云服务提供商之前,构建密文 *k-d* tree(如图 7 所示).

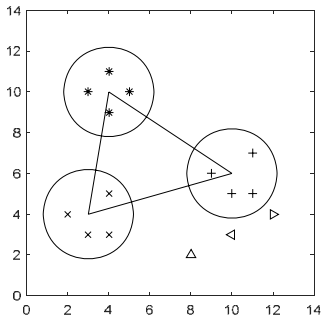

Fig.4 Range query problem of *k*-means

图 4 *k*-means 范围查询问题

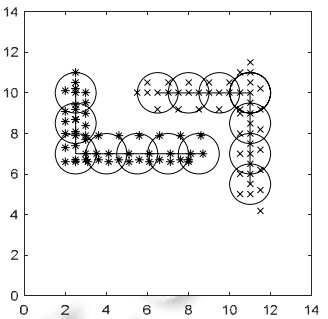

Fig.5 KNN problem of Dbscan

图 5 Dbscan 的 KNN 问题

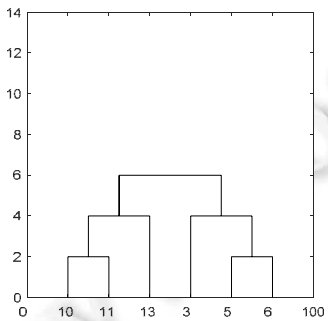

Fig.6 Nearest neighbor problem of hierarchical clustering

图 6 层次聚类最近邻问题

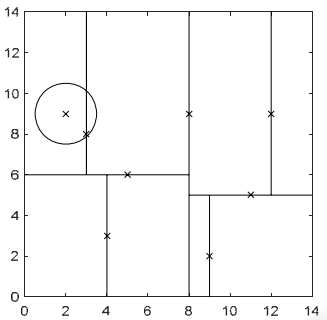

Fig.7 Two-dimensional ciphertext *k-d* tree example

图 7 二维密文 *k-d* tree 示例

k-d 树最邻近搜索的过程如下:从根节点开始,递归地往下移.往左还是往右的决定方法与插入元素的方法一样(如果输入点在分区面的左边则进入左子节点,在右边则进入右子节点).一旦移动到叶节点,将该节点当做目前最佳点.解开递归,并对每个经过的节点运行下列步骤:如果目前所在点比目前最佳点更靠近输入点,则将其变为目前最佳点.检查另一边子树有没有更近的点,如果有,则从该节点往下找.当根节点搜索完毕后,完成最邻近搜索.

4.2 隐私保护聚类挖掘

首先,用户生成自己的公私钥对 K_{pri},K_{pub} 以及公开参数向量 $\vec{p}=\{l_1,l_2,...,l_{n-1},l_n\}$;然后,对数据进行隐私保护的数据处理,并将处理结果上传到云端,云端调用密文中心点计算算法和密文距离计算算法进行数据挖掘;最终,将多种挖掘算法的结果返回给用户,避免使用单一算法挖掘难以发现潜在的挖掘聚簇的问题.用户端使用保持距离的解密算法还原挖掘的最终明文结果.

用户预处理过程中,用户将元组使用第 3.3 节中的算法将隐私数据进行加密处理后,结果见表 5.

Table 5 Post-release data

表 **5** 发布数据

Id	Age	Weight	Heartrate
1237	*S*(75),*T*(75)	*S*(80),*T*(80)	*S*(63),*T*(63)
3420	*S*(79),*T*(79)	*S*(64),*T*(64)	*S*(55),*T*(55)
2543	*S*(67),*T*(67)	*S*(52),*T*(52)	*S*(70),*T*(70)
4461	*S*(62),*T*(62)	*S*(58),*T*(58)	*S*(76),*T*(76)

上传到云服务提供商,同时还需要上传的是加密过程中产生随机数向量的生成参数 $\vec{p}$.对于云端隐私保护的数据挖掘结果,用户可以使用第 3.4 节的算法得到明文数据.当用户可以判断聚类类型时,用户可以直接指定

挖掘算法;否则,云端使用多种聚类挖掘算法进行处理,返回聚类结果给用户.

云端收到用户的聚类分析请求后,根据聚类请求选择合适的聚类算法进行处理.隐私保护的聚类处理核心问题是密文数据点距离计算问题和均值密文计算问题.PPK-means 反复迭代计算密文均值和密文距离,直至簇的均值不发生变化,即:两次迭代计算均值之间的距离小于某一阈值∂,见算法 4.求解过程中,试用 *k-d* tree 进行优化,以质心为中心、两个质心的连线距离的一般为半径,寻找此范围里面的向量,从而大幅度减少了距离计算的时间开销.基于层级的聚类需要求解距离最近的元组,该问题是 KNN 问题的特例,使用 *k-d* tree 减少了距离运算,每个元组查找最近邻的时间复杂度由 $O(n)$降低为 $O(\log n)$.dbscan 需要求解向量是否为核心点,然后连接核心点,此时,利用隐私保护 *k-d* tree 时间复杂度为 $O(\log n)$.

算法 **4**. 隐私保护的聚类算法 PPK-means.

输入:数据集 C、簇的个数 K 和用户指定的阈值 、迭代此处 *num*;

输出:聚簇 C_i.

1: choose $c_1,c_2,\ldots,c_n$ as initial centroid

2: Do {

3: *generateSearchPath*();

4: *rangeQuery*();

5: 调用第 3.5 节距离的密文计算方法,将每个点指派到最近的中间点,形成 K 个簇.

6: 调用第 3.6 节密文的中心点计算方法,重新计算每个簇的中间点

7: i++;

8: }while (质心发生变化大于指定的阈值∂ or $i<num$)

基于格的保持距离的加密算法可以精确地计算出向量之间的相对距离以及中心点与各个向量值间的相对距离.对于 *k*-means 而言,本文的发布方法可以准确地求解均值和密文条件下数据之间的距离,因此可以保证挖掘的正确性.对于层次聚类而言,本文算法求解的距离具有保持距离相对大小的特性,因此可以选择出当前相距最小的两个元素.对于 DNSCAN 而言,本文的隐私保护数据发布方法可以准确地求解ε邻域中的核心点.

综上所述,数据经过基于格的保持距离数据处理方法处理,上传到云端可以进行聚类挖掘过程中涉及的所有同态密文计算,因此可以保证挖掘的正确性.

5 实验结果与分析

5.1 实验环境

实验环境:Window 7 操作系统;CPU intel B950 2.10GHz;4.00GB 内存.编程环境:Matlab7.0.11 编译器.

实验数据:本文的实验数据选用加州大学尔湾分校的机器学习数据库中的心脏心律失常数据集的真实数据[26],针对真实数据集规模有限的问题,本文利用正态分布生成了模拟数据集,使用模拟数据集进行较大规模的实验.参照 RBT 方法[15],论文选择数据集中的 3 列数据:年龄(age)、体重(weight)和心率(heart rate).实验中,随机选择 10 000,20 000,30 000 和 40 000 条记录规模分别进行了实验分析.衡量优质基的阈值参数采用 Ψ=0.9,加密过程批处理维度 3.

论文研究了一种隐私保护的数据聚类挖掘方法,引入挖掘准确率用于度量隐私保护数据挖掘算法与明文条件下数据挖掘算法的一致性,见公式(9)(隐私保护的聚类方法准确率与对应的普通明文条件下的数据挖掘结果进行比较,准确率值越高,表明隐私保护的聚类数据分析方法分析结果越精确):

$$rate_{accuracy}=\sum_{i=1}^{k}\frac{|C_i\cap S_i|}{|C_i|}\times\frac{|C_i|}{|C|} \tag{9}$$

其中,C 为数据集,C_i 为明文条件下得到的第 i 个数据聚类,S_i 为与之对应的隐私保护数据挖掘聚类方法得到的第 i 个密文数据聚类.本文设计的准确率度量方法中,考虑了聚类大小权重因素($|C_i|/|C|$)的影响.

本文设计实验,从两个方面对提出的隐私保护聚类数据挖掘方法进行评估,即,挖掘算法的性能和精确度.

5.2 实验结果分析

本文针对基于格的保持距离的加密算法进行了密钥生成、加密解密性能以及密文计算的时间开销的实验测试,实验结果证明,本文提供的方法切实可行.与现有的方法相比,聚类挖掘的精度优于现有方法,用户发布数据时在时间开销有所增加,但是在可接受范围内.实验分别针对基于格的保持距离的加密算法和聚类挖掘算法进行了分析比较.

5.2.1 基于格的保持距离的加密算法性能分析

密钥生成性能实验结果如图 8 所示,用户生成密钥的时间开销主要是生成公钥的时间,私钥的生成时间忽略不计.一次公钥私钥对生成时间约为 8ms.对于初始化密钥过程中的转换矩阵计算取值,实验结果表明,进行 30 次初等变换即可.由图 9 可以看出,加密的时间要远远超过解密的时间.这是因为加密过程中符合条件的随机数的生成占用了一定的时间.平均一次加密的时间为 0.4ms.如实验数据所示,本文提出的算法与现有的方法相比性能更好.主要是由于一次可以加密或解密一个元组中多列.由图 10 可以看出:40 000 次的密文距离计算只需要 67s,40 000 次的中间点求解计算只需要 67s,聚类挖掘所需要的云端的密文计算开销在可接收范围内.考虑到实际应用中云计算服务器性能能强,实际应用中可以获得更好的隐私数据挖掘效率.

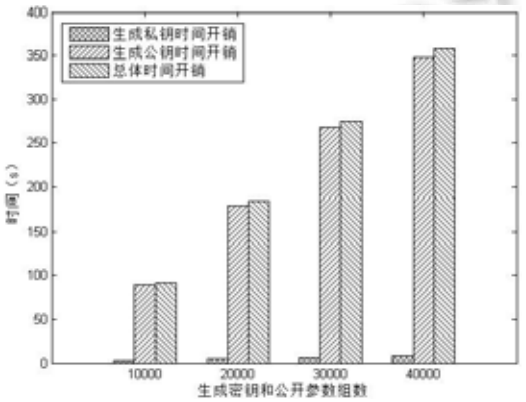

Fig.8 Key generation time overhead

图 8 密钥生成时间开销

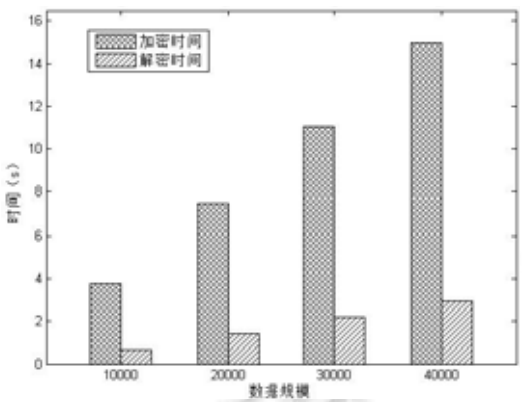

Fig.9 Encryption and decryption time overhead

图 9 加密解密时间开销

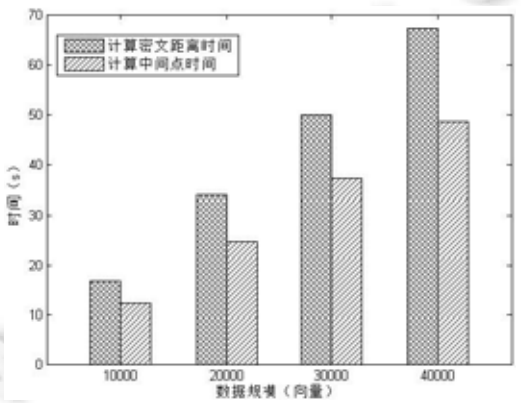

Fig.10 CSP calculation time overhead

图 10 云端密文计算时间开销

5.2.2 隐私保护的聚类挖掘方法性能与准确率分析

本文提出的聚类挖掘方法包括用户预处理部分时间开销、云服务提供商挖掘部分时间开销以及与现有方

法挖掘准确率的比较.

从图 11 可以看出:与现有的方法相比,本文 PPCM 发布数据的时间较高,但是在可以接受的范围内.主要原因是发布算法基于格的困难为题,提升了隐私保护数据的安全性而带来的时间开销.

从图 12 可以看出:密文 k-mean 算法的时间复杂度是 $O(n\log(n))$,密文层次聚类的时间复杂度为 $O(n^2)$,密文 DBSCAN 时间复杂度 $O(n)$.实验结果表明:实际情况中,k-mean 并非严格按照理论分析的时间复杂度产生时间开销,由于数据集的差异以及初始中心点的选择不同,PPk-mean 算法的时间开销会与理论分析有所差异.从结果可以看到:隐私保护的层次挖掘算法性能最差,PPDBscan 性能最好.一方面,这与选取的数据及有关系;另一方面,3 种算法针对不同聚类类型的数据集各有优缺点,并无可比性.很多情况下,用户需要执行多种挖掘方法,以便发现数据集中潜在的聚类.

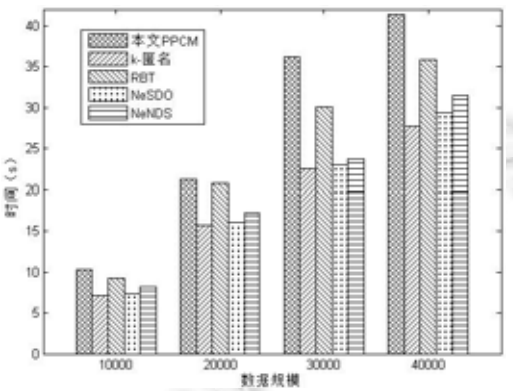

Fig.11 User preprocessing time overhead

图 11 用户预处理时间开销

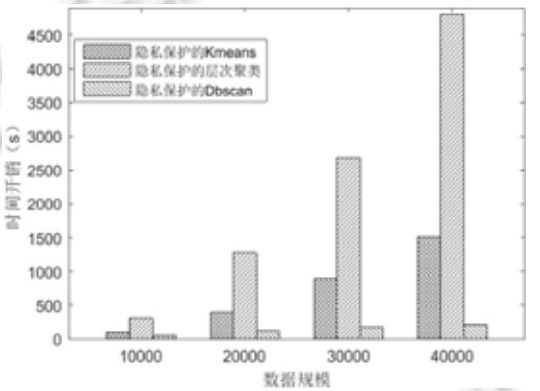

Fig.12 Privacy protection clustering mining time overhead

图 12 隐私保护的聚类挖掘时间开销

从图 13 可以看出:本文提出的算法与目前准确率最高的 RBT 算法的准确率相当,但是更加安全.在极端情况下,由于 k 取值的不同,k 匿名挖掘甚至于不能达到实验中的正确率.由于不同的挖掘算法侧重点不同,对于同一数据返回的结果相差较大,本文的准确率针对 3 种方法分别进行比较,即:原始数据的 k-means 结果与密文环境下 PPk-means 的结果进行比对,计算精确率.

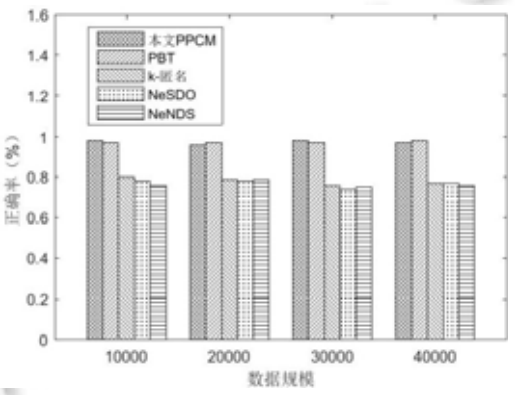

Fig.13 Privacy protection mining algorithm accuracy

图 13 隐私保护挖掘算法准确率

与现有的方法相比,本文的方法在增强安全的同时,很好地保证了挖掘结果的精度和挖掘的效率.虽然牺牲了一定效率,但保证了隐私数据的精度和安全性.对于用户而言,挖掘精度对于用户至关重要,挖掘结果的精度

出现偏差,很可能对于用户决策造成不可弥补的损失.

6 总结与展望

本文主要研究了云计算环境下支持隐私保护的隐私数据加密发布方法以及密文隐私数据的挖掘方法.首先,本文提出了一种用于隐私数据发布的基于格的保持距离的加密算法,在该方法的基础上,设计了聚类挖掘过程中需要的密文计算方法,可以对密文数据进行处理,处理的结果保持同态特性.云服务提供商在保障用户隐私的条件下,有能力进行聚类挖掘,最终将挖掘结果返回给用户.隐私保护的聚类挖掘方法提供了 3 种聚类挖掘机制,用户可以根据自己的数据特征,选择有效的聚类挖掘算法进行云端隐私数据挖掘操作.

References:

[1] Ni WW, Chen G, Chong ZH, Wu YJ. Privacy-Preserving data publication for clustering. Journal of Computer Research and Development, 2012,49(5):1095−1104. (in Chinese with English abstract)

[2] Ding Y, Wang HM, Shi PC, Wu QB, Dai HD, Fu HY. Trusted Cloud Service. Chinese Journal of Computers, 2015,38(1):133−149 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2015.00133]

[3] Wu XD, Zhu XQ, Wu GQ, Ding W. Data mining with big data. IEEE Trans. on Knowledge and Data Engineering, 2014,26(1): 97−107. [doi: 10.1109/TKDE.2013.109]

[4] Agrawal R, Srikant R. Privacy-Preserving data mining. ACM Sigmod Record, 2000,29(2):439−450. [doi: 10.1145/342009.335438]

[5] Zhou SG, Li F, Tao YF, Xiao XK. Privacy preservation in data applications: A survey. Chinese Journal of Computers, 2009,32(5): 847−861 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2009.00847]

[6] Mohammed N, Chen R, Fung BCM, Yu PS. Differentially private data release for data mining. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. 493−501. [doi: 10.1145/2020408.2020487]

[7] Allard T, Hébrail G, Masseglia F, Pacitti E. Chiaroscuro: Transparency and privacy for massive personal time-series clustering. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2015. 779−794. [doi: 10.1145/2723372.2749453]

[8] Vlachos M, Schneider J, Vassiliadis VG. On data publishing with clustering preservation. ACM Trans. on Knowledge Discovery from Data (TKDD), 2015,9(3):23. [doi: 10.1145/2700403]

[9] Giannotti F, Lakshmanan LVS, Monreale A, Pedreschi D, Wang H. Privacy-Preserving mining of association rules from outsourced transaction databases. Systems Journal, IEEE, 2013,7(3):385−395. [doi: 10.1109/JSYST.2012.2221854]

[10] Fong PK, Weber-Jahnke JH. Privacy preserving decision tree learning using unrealized data sets. IEEE Trans. on Knowledge and Data Engineering, 2012,24(2):353−364. [doi: 10.1109/TKDE.2010.226]

[11] Johnson A, Shmatikov V. Privacy-Preserving data exploration in genome-wide association studies. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2013. 1079−1087. [doi: 10.1145/2487575.2487687]

[12] Zhang XY, Liu C, Nepal S, Pandey S, Chen JJ. A privacy leakage upper bound constraint-based approach for cost-effective privacy preserving of intermediate data sets in cloud. IEEE Trans. on Parallel and Distributed Systems, 2013,24(6):1192−1202. [doi: 10.1109/TPDS.2012.238]

[13] Nguyen P. Cryptanalysis of the goldreich-goldwasser-halevi cryptosystem from crypto'97. In: Proc. of the Annual Int'l Cryptology Conf. Berlin, Heidelberg: Springer-Verlag, 1999. 288−304. [doi: 10.1007/3-540-48405-1_18]

[14] Yi X, Zhang Y. Equally contributory privacy-preserving *k*-means clustering over vertically partitioned data. Information Systems, 2013,38(1):97−107. [doi: 10.1016/j.is.2012.06.001]

[15] Oliveira SRM, Zaïane OR. Achieving privacy preservation when sharing data for clustering. In: Proc. of the Workshop on Secure Data Management in Conjunction with VLDB. Berlin, Heidelberg: Springer-Verlag, 2004. 67−82. [doi: 10.1007/978-3-540-30073-1_6]

[16] Islam MZ, Brankovic L. Privacy preserving data mining: A noise addition framework using a novel clustering technique. Knowledge-Based Systems, 2011,24(8):1214−1223. [doi: 10.1016/j.knosys.2011.05.011]

[17] Zhu Y, Liu L. Optimal randomization for privacy preserving data mining. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2004. 761−766. [doi: 10.1145/1014052.1014153]

[18] Chong ZH, Ni WW, Liu TT, Zhang Y. A privacy-preserving data publishing algorithm for clustering application. Journal of Computer Research and Development, 2010,47(12):2083−2089 (in Chinese with English abstract).

[19] Li YP, Chen MH, Li QW, Zhang W. Enabling multilevel trust in privacy preserving data mining. IEEE Trans. on Knowledge and Data Engineering, 2012,24(9):1598−1612. [doi: 10.1109/TKDE.2011.124]

[20] Navarro-Arribas G, Torra V, Erola A, Castella-Roca J. User *k*-anonymity for privacy preserving data mining of query logs. Information Processing & Management, 2012,48(3):476−487. [doi: 10.1016/j.ipm.2011.01.004]

[21] Xiao X, Yi K, Tao Y. The hardness and approximation algorithms for l-diversity. In: Proc. of the 13th Int'l Conf. on Extending Database Technology. ACM Press, 2010. 135−146. [doi: 10.1145/1739041.1739060]

[22] Dwork C. A firm foundation for private data analysis. Communications of the ACM, 2011,54(1):86−95. [doi: 10.1145/1866739.1866758]

[23] Mohan P, Thakurta A, Shi E, Song D, Culler DE. GUPT: Privacy preserving data analysis made easy. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2012. 349−360. [doi: 10.1145/2213836.2213876]

[24] Parameswaran R, Blough DM. Privacy preserving data obfuscation for inherently clustered data. Int'l Journal of Information and Computer Security, 2008,2(1):4−26. [doi: 10.1504/IJICS.2008.016819]

[25] Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis. In: Proc. of the 39th Annual ACM Symp. on Theory of Computing. ACM Press, 2007. 75−84. [doi: 10.1145/1250790.1250803]

[26] Blake CL, Merz CJ. UCI repository of machine learning databases. Irvine: University of California, Department of Information and Computer Science, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html

附中文参考文献:

[1] 倪巍伟,陈耿,崇志宏,吴英杰.面向聚类的数据隐藏发布研究.计算机研究与发展,2012,49(5):1095−1104.

[2] 丁滟,王怀民,史佩昌,吴庆波,戴华东,富弘毅.可信云服务.计算机学报,2015,38(1):133−149. [doi: 10.3724/SP.J.1016.2015.00133]

[5] 周水庚,李丰,陶宇飞,肖小奎.面向数据库应用的隐私保护研究综述.计算机学报,2009,32(5):847−861. [doi: 10.3724/SP.J.1016.2009.00847]

[18] 崇志宏,倪巍伟,刘腾腾,张勇.一种面向聚类的隐私保护数据发布方法.计算机研究与发展,2010,47(12):2083−2089.

崔一辉(1981－),男,陕西西安人,硕士,主要研究领域为应用密码学,云安全,隐私保护.

史成良(1994－),男,学士,主要研究领域为应用密码学,云安全,隐私保护.

宋伟(1978－),男,博士,副教授,主要研究领域为应用密码学,云安全,隐私保护.

程芳权(1985－),男,博士,主要研究领域为应用密码学,云安全,隐私保护.

王占兵(1992－),男,硕士,主要研究领域为应用密码学,云安全,隐私保护.