文章编号:1001-506X(2012)07-1505-06

基于 SMC 的隐私保护聚类模型

方炜炜[1,2], 杨炳儒[2], 夏红科[1,2]
(1. 北京信息科技大学计算中心,北京 100192; 2. 北京科技大学信息工程学院,北京 100083)

摘 要: 隐私保护数据挖掘指在实现准确挖掘知识的同时确保敏感数据不泄露。针对垂直分布式数据存储结构的聚类隐私保护问题,提出基于全同态加密协议和数据扰乱方法的隐私保护聚类模型。该模型通过采用安全比较协议解决了垂直分布式聚类的两个隐私保护关键步骤:求解最近簇和判断质心变化,从而实现了数据的有效保护。理论证明了该模型的安全性并分析了其时间复杂度和通信耗量,实验结果表明该隐私保护聚类模型是安全有效的。

关键词: 安全多方计算; 同态加密; 聚类; 隐私保护数据挖掘
中图分类号: TP 309 **文献标志码**: A **DOI**:10.3969/j.issn.1001-506X.2012.07.36

Privacy-preserving clustering modeling based on SMC

FANG Wei-wei[1,2], YANG Bing-ru[2], XIA Hong-ke[1,2]
(*1. Computer Center, Beijing Information Science and Technology University, Beijing 100192, China;*
2. School of Information Engineering, Beijing University of Science and Technology, Beijing 100083, China)

Abstract: Privacy-preserving data mining aims to accurately mine knowledge while unrevealing sensitive data. For solving the privacy-preserving clustering problem in vertical distribution, a privacy-preserving clustering model based on full homomorphous encryption protocols and data perturbation technology is proposed. The model protects original data effectively by using secure comparison protocols to compute the nearest cluster and estimate the updating of the cluster center, which are two key steps in clustering process. Theory argument demonstrates the security of the privacy-preserving clustering model and analyzes computation complexity and communication costs. Experiment results prove that the privacy-preserving clustering model is secure and effective.

Keywords: secure multi-party computation(SMC); homomorphous encryption; clustering; privacy preserving data mining

0 引 言

随着越来越多的信息以电子形式出现并在 Web 上可以访问,伴随着越来越强大的数据挖掘工具的开发和投入使用,人们越来越关注数据挖掘可能威胁我们的隐私和数据安全。在涉及个人隐私数据或企业敏感信息的数据挖掘应用中,如何提高数据安全增强技术已经引起了学术界的广泛关注。隐私保护数据挖掘(privacy-preserving data mining, PPDM)由文献[1]提出,其任务是旨在解决数据挖掘过程中"知识规则的准确挖掘"与"原始信息的隐私保护"的冲突问题,目前已成为数据挖掘领域的一个研究热点。

PPDM 技术关注于一般模式发现,而不是关于个人的特定信息,在不涉及底层数据值的同时获取有效的数据挖掘结果。通常有两种解决途径:①数据扰乱。文献[2]提出布尔关联规则挖掘算法,采用正态分布或高斯分布的随机数据对原始数据库进行数据扰乱,然后基于 Warner 模型对原始数据进行特征重构;文献[3-5]基于随机响应技术分别处理决策树分类、贝叶斯分类和关联规则的挖掘任务,敏感数据的属性值通过一种应答特定问题的方式,以概率 θ 间接提供给外界。文献[6]采用数据交换和添加噪音相结合的方法实现分类属性数据的隐私保护。文献[7]中指出该类方法在应用中的众多缺陷,如添加噪音而造成挖掘精度的损失、新的构建数据模型可推导部分原始隐私信息等等,因而近年来该类技术发展缓慢。②安全多方计算(se-

收稿日期:2011-07-04; **修回日期**:2012-01-12。
基金项目:国家自然科学基金重点项目(61175048)资助课题
作者简介:方炜炜(1979-),女,讲师,博士,主要研究方向为隐私保护数据挖掘。E-mail: liveinbetter@163.com

cure multiparty computation, SMC)。文献[8]中证明了不同种类的数据挖掘问题都可以转化为安全的多方计算问题;文献[9]中归纳了常应用于分布式数据挖掘过程中的安全求和、安全求并、安全求交集大小以及安全求标量积方法;文献[10]中则指出常规安全多方计算协议在具体背景下并不一定适用,要根据应用需求设计相应的协议。

聚类是数据挖掘领域中一个重要的研究分支,它基于数据的相似性,将对象分成若干簇,使同一簇中的对象具有很高的相似度,而不同簇的对象高度相异。目前,基于SMC协议的隐私保护数据挖掘技术主要应用于关联规则[11-12]、分类[13-14]、序列模式[15]和水平分布式聚类[16-17],而垂直分布式的隐私保护聚类模型的研究较少。文献[18]中提出了垂直分布式的隐私保护聚类模型,通过采用安全求和协议来统计数据信息。但其中间计算结果均未加密且被各参与方分享,一旦两方合谋利用安全求和协议的破绽便可求解出各方的原始数据信息。本文基于全同态公钥加密协议[19]和数据扰乱方法设计了安全比较协议,实现隐私保护的垂直分布式聚类模型。

1 传统垂直分布式聚类

定义 1 垂直分布式聚类:数据集 D 包含 N 个对象 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$,每个对象 $\boldsymbol{x}_i(1\leqslant i\leqslant N)$ 由 r 个属性($\boldsymbol{x}_{i1}$, $\boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{ir}$)组成。数据集 D 分布式存储在 r 个站点 P_1, $P_2, \cdots, P_r$,且满足如下条件:

(1) $D_{P1} \cup D_{P2} \cup \cdots \cup D_{Pr} = D$;

(2) $D_{Pj} = \{\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}, \cdots, \boldsymbol{x}_{Nj}\}$, $1\leqslant j\leqslant r$。

数据集 D 聚类,即是 D 基于一种划分准则生成 k 个簇,使得同一簇中的对象属性集是相似的,而不同簇间的对象属性集是相异的。K-means 聚类是由 MacQueen J B[17] 提出的一种基于质心的典型划分方法,其基本步骤如下:

步骤 1 随机选择 k 个对象,每个对象代表一个簇的质心;

步骤 2 应用欧几里得距离计算 D 中每个对象与各个簇质心的距离,并将每个对象划分到与其最近的簇;

步骤 3 重新计算各个簇的质心;

步骤 4 反复执行步骤 2 和步骤 3,直到质心变化低于设定阈值。

当数据拥有方 $P_1, P_2, \cdots, P_r$ 共同参与聚类任务,其目标是不提供本地原始数据信息$\{x_{1j}, x_{2j}, \cdots, x_{Nj}\}$ $(1\leqslant j\leqslant r)$的前提条件下,将 D 划分成 k 个簇。

2 分布式隐私保护聚类模型

垂直分布式隐私保护聚类模型步骤如下:

步骤 1 随机选择$\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k\}$作为 k 个簇的质心,各参与方仅拥有 k 个簇的本地属性值,如 P_j 拥有数据信息 $\{\mu_{1j}, \mu_{2j}, \cdots, \mu_{kj}\}(1\leqslant j\leqslant r)$;

步骤 2 各参与方 P_j 计算局部统计数据 $M_{isj} = (\mu_{sj} - x_{ij})^2(1\leqslant s\leqslant k, 1\leqslant i\leqslant N, 1\leqslant j\leqslant r)$,采用同态加密协议和添加扰乱数据思想生成 $E(M_{isj} + l_{isj})$,然后发送给半可信挖掘者;

步骤 3 半可信挖掘者利用安全比较协议,求解出每个对象的最近簇,并用索引数组 $C[s](1\leqslant s\leqslant k)$录各簇的相关对象集;

步骤 4 各参与方 Pj 依据索引数组 $C[s]$中涉及的对象,在本地计算$\{\mu'_{1j}, \mu'_{2j}, \cdots, \mu'_{kj}\}$实现质心更新;

步骤 5 各参与方 P_j 计算局部统计数据 $O_{sj} = (\mu_{sj} - \mu'_{sj})^2(1\leqslant s\leqslant k, 1\leqslant j\leqslant r)$,并利用同态加密协议和添加扰乱数据思想生成 $E(O_{sj} + l_{sj})$,然后发送给半可信挖掘者;

步骤 6 半可信挖掘者计算 $\sum_{j=1}^{r} E(O_{sj} + l_{sj})$,采用同态加密协议进行解密得到 $\sum_{j=1}^{r} O_{sj}$,判断其值是否小于阈值 h,如果小于则停止迭代;否则执行步骤 2。

垂直分布式数据集 D 应用传统 K-means 方法进行聚类时,步骤 2 求解欧几里得距离、距离比较,步骤 4 判断更新前后的质心变化是否低于阈值,均需要汇总各地数据信息进行计算。如何不提供原始本地信息,而又能正确地解决聚类任务是本文的研究重点。本文基于文献[19]提出的全同态加密协议和数据扰乱方法设计了安全比较协议,实现了分布式隐私保护的聚类模型。

在如上隐私保护聚类模型中,隐私保护的关键点在于:①步骤 2 和步骤 3 采用全同态加密协议和安全比较协议求解每个对象的最近簇;②步骤 5 和步骤 6 采用全同态加密协议和数据扰乱方法判断 k 个簇更新前后的质心变化是否小于阈值。

2.1 求解最近簇

采用欧几里得距离公式计算每个对象和 k 个簇的质心距离,即

$$W(\boldsymbol{x}_i, \boldsymbol{\mu}_s) = \sum_{j=1}^{r} (\mu_{sj} - x_{ij})^2 \tag{1}$$

式(1)表示第 i 个对象和第 s 个簇(1≤i≤N, 1≤s≤k)的距离。

寻找第 i 个对象的最近簇,即求解$\min\limits_{1\leqslant s\leqslant k}(W(\boldsymbol{x}_i, \boldsymbol{\mu}_s))$。各参与方 P_j 能够计算局部统计数据($M_{i1j}, M_{i2j}, \cdots, M_{isj}, \cdots, M_{ikj}$),其中 $M_{isj} = (\mu_{sj} - x_{ij})^2(1\leqslant s\leqslant k, 1\leqslant j\leqslant r)$;如果不考虑原始数据的隐私保护,半可信挖掘者汇总如图 1 所示的各方局部统计数据然后进行计算。

求解第 i 个对象的最近簇,即计算等式:

$$\min_{1\leqslant s\leqslant k}(W(\boldsymbol{x}_i, \boldsymbol{\mu}_s)) = \min_{1\leqslant s\leqslant k}\left(\sum_{j=1}^{r} M_{isj}\right) \tag{2}$$

$W(\boldsymbol{x}_i,\boldsymbol{\mu}_s)$

图 1　各参与方提供的局部统计数据

为保护局部统计数据 M_{isj} 和 $\sum_{j=1}^{r} M_{isj}$ 不泄露，采用了全同态加密协议（详细内容请参阅参考文献[19]），各参与方将 M_{isj} 进行数据扰乱然后加密发送给半可信挖掘者；半可信挖掘者采用安全比较协议在不获取任何局部统计信息 M_{isj} 和 $\sum_{j=1}^{r} M_{isj}$ 的前提条件下，在不求解出最小值 $\min_{1\leqslant s\leqslant k}(\sum_{j=1}^{r} M_{isj})$ 数据信息的情况下，找到最近簇 $\min_{1\leqslant s\leqslant k}(C(\boldsymbol{x}_i))$。

协议 1　多方安全比较最小值协议

输入　各参与方 $P_j(1\leqslant j\leqslant r)$ 拥有局部统计数据 $(M_{i1j},M_{i2j},\cdots,M_{isj},\cdots,M_{ikj})$；

输出　$\min_{1\leqslant s\leqslant k}(C(\boldsymbol{x}_i))$，即第 i 个对象的最近簇。

步骤 1　密钥中心机构（Key-Center）生成随机向量 $(\boldsymbol{l}_{i1},\boldsymbol{l}_{i2},\cdots,\boldsymbol{l}_{ir})$，其中 $\boldsymbol{l}_{ij}=(l_{i1j},l_{i2j},\cdots,l_{ikj})(1\leqslant i\leqslant N,1\leqslant j\leqslant r)$ 并满足条件 $\sum_{j=1}^{r} l_{isj}=0(1\leqslant s\leqslant k)$，将 $\boldsymbol{l}_{ij}$、加密公钥 $\{\upsilon,p,n\}$ 和加密私钥 key_j 分别发送给各参与方 $P_j(1\leqslant j\leqslant r)$。

步骤 2　参与方 P_j 按协议 1 执行 $E(M_{isj}+l_{isj})=(M_{isj}+l_{isj}+\upsilon p)key_j^{\upsilon}\bmod n$，然后发送 $E(M_{isj}+l_{isj})$ 给半可信挖掘者（Semi-Honest Miner）。

步骤 3　密钥中心机构选取随机数 s，计算 $K_j=s\cdot lcm(key_1^{\upsilon},key_2^{\upsilon},\cdots,key_r^{\upsilon})/key_j^{\upsilon}(1\leqslant j\leqslant r)$，并发送给半可信挖掘者。

步骤 4　半可信挖掘者计算 $E(W(\boldsymbol{x}_i,\boldsymbol{\mu}_s))=\sum_{j=1}^{r}E(M_{isj}+l_{isj})(1\leqslant s\leqslant k)$，初始化 $\min_{(1\leqslant s\leqslant k)}(\sum_{j=1}^{r}M_{isj})=E(W(\boldsymbol{x}_i,\boldsymbol{\mu}_1))$。

步骤 5　从 $s=2$ 执行如下操作，直到 $s=k$ 结束：

（1）若 $E(W(\boldsymbol{x}_i,\boldsymbol{\mu}_s))-\min_{1\leqslant s\leqslant k}(\sum_{j=1}^{r}M_{isj})<0$，则 $\min(\sum_{j=1}^{r}M_{isj})=E(W(\boldsymbol{x}_i,\boldsymbol{\mu}_s))$，并记录 $\min_{1\leqslant s\leqslant k}(C(\boldsymbol{x}_i))=s$；

（2）$s=s+1$。

求解最近簇算法如下：

算法 1　寻找最近簇（find nearest cluster，FNC）

输入　$P_j(1\leqslant j\leqslant r)$ 拥有数据 $\{x_{1j},x_{2j},\cdots,x_{Nj}\}$ 和 $\{\mu_{1j},\mu_{2j},\cdots,\mu_{kj}\}$；

密钥中心机构拥有数据 $(\boldsymbol{l}_{i1},\boldsymbol{l}_{i2},\cdots,\boldsymbol{l}_{ir})$、加密钥 $\{\upsilon,p,n\}$，key_j 和 K_j；

输出　$C[\min_{1\leqslant s\leqslant k}(C(\boldsymbol{x}_i))]$。

At Key-Center：

(1) Sends $\boldsymbol{l}_{ij}$ and $\{\upsilon,p,n\}$, key_j to P_j；

(2) Sends K_j to Semi-Honest Miner；

(3) For all $j=1$ to r　　**At $\boldsymbol{P_j}$**：

(4) For $i=1$ to n

(5) For $s=1$ to k

(6) $M_{isj}=(\mu_{sj}-x_{ij})^2$

(7) $E(M_{isj}+l_{isj})=(M_{isj}+l_{isj}+\upsilon p)key_j^{\upsilon}\bmod n$

(8) End for

(9) End for

(10) Sends $E(M_{isj}+l_{isj})$ to Semi-Honest Miner

(11) End for

At Semi-Honest Miner：

(12) For $i=1$ to n

(13) For $s=1$ to k

(14) $E(W(\boldsymbol{x}_i,\boldsymbol{\mu}_s))=\sum_{j=1}^{r}E(M_{isj}+l_{isj})$

(15) End for

(16) $\min_{1\leqslant s\leqslant k}(\sum_{j=1}^{r}M_{isj})=E(W(\boldsymbol{x}_i,\boldsymbol{\mu}_1))$

(17) For $s=2$ to k

(18) If $E(W(\boldsymbol{x}_i,\boldsymbol{\mu}_s))-\min_{1\leqslant s\leqslant k}(\sum_{j=1}^{r}M_{isj})<0$

(19) $\min_{1\leqslant s\leqslant k}(\sum_{j=1}^{r}M_{isj})=E(W(\boldsymbol{x}_i,\boldsymbol{\mu}_s))$

(20) $\min_{1\leqslant s\leqslant k}(C(\boldsymbol{x}_i))=s$

(21) End if

(22) End for

(23) $C[\min_{1\leqslant s\leqslant k}(C(\boldsymbol{x}_i))]=\boldsymbol{x}_i$

(24) End for

2.2　判断质心变化

各参与方 $P_j(1\leqslant j\leqslant r)$ 通过算法 1 求解出每个对象的最近簇，并用索引 $C[s](1\leqslant s\leqslant k)$ 记录各簇的相关对象；然后在本地按照式(3)更新质心，z 表示簇 s 的相关对象个数，x_{ij} 表示与簇 s 相关的对象。

$$\mu'_{sj}=\sum_{i=1}^{z}\boldsymbol{x}_{ij}/\boldsymbol{z} \tag{3}$$

判断质心更新后的位置变化是否小于阈值，如果小于阈值则停止迭代，否则重新求解最近簇，其算法思想如图 2 所示。

图 2 判断质心变化的算法思想

算法 2 检查门限算法(check threshold algorithm, CKA)

输入:$P_j(1\leqslant j\leqslant r)$拥有数据$\{\mu'_{1j},\mu'_{2j},\cdots,\mu'_{kj}\}$和$\{\mu_{1j},\mu_{2j},\cdots,\mu_{kj}\}$;

密钥中心机构拥有数据$(\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_r)$(满足条件$\sum_{j=1}^{r} l_{sj}=0,1\leqslant s\leqslant k$)、加密钥□,$v,p,n$, key_j□解密私钥 t 和 K_j;

输出:执行算法 1 或停止迭代。

At Key-Center:

(1) Sends $\boldsymbol{l}_j$ and □,v,p,n, key_j□to P_j;

(2) Sends K_j and □□to Semi-Honest Miner;

(3) For all $j=1$ to r **At P_j**:

(4) For $s=1$ to k

(5) $O_{sj}=(\mu_{sj}-\mu'_{sj})^2$

(6) $E(O_{sj}+l_{sj})=(O_{sj}+l_{sj}+vp)key_j^v \bmod n$

(7) End for

(8) Sends $E(O_{sj}+l_{sj})$ to Semi-Honest Miner

(9) End for

At Semi-Honest Miner:

(10) Compute $\sum_{j=1}^{r} E(O_{sj}+l_{sj})$

(11) $\sum_{j=1}^{r} O_{sj}=(b^{tv})^{-1}\cdot\sum_{j=1}^{r} E(O_{sj}+l_{sj})\%p$

(12) If $\sum_{j=1}^{r} O_{sj}>h$ then

(13) Run FNC algorithm

(14) Else program end

3 安全和效率分析

本文提出的隐私保护分布式聚类模型是由数据源提供方$\{P_1,P_2,\cdots,P_r\}$、半可信挖掘者和密钥中心机构三种角色构成的网络通信模型,如图 3 所示,每对参与方之间不仅拥有安全的信道,而且参与方之间还拥有 1 个认证的广播信道。

图 3 隐私保护分布式聚类通信模型

(1) 数据源提供方:正确地执行协议,但可能会试图通过本地计算的中间结果或联合其他数据源提供方推导诚实方的隐私输入。

(2) 密钥中心机构:不折不扣地执行协议,不能获取任何参与方的原始数据和计算结果,并且不与任何参与方合谋。

(3) 半可信挖掘者:不折不扣地执行协议,不能获取任何数据提供方的原始数据,并且不与任何参与方合谋。

在半可信模型下,数据源提供方 $P_j(1\leqslant j\leqslant r)$拥有隐私数据$\{x_{1j},x_{2j},\cdots,x_{Nj}\}$,在密钥中心机构和半可信挖掘者的协助下,多方计算 $f(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_r,\boldsymbol{l})=(Y_1,Y_2,\cdots,Y_r)$,其中 $\boldsymbol{l}$ 是随机扰乱数据,$Y_1,Y_2,\cdots,Y_r$ 是各方获取的计算结果。当满足如下条件时,即可认为该模型是安全有效的:

① 隐私性:对于各数据源提供方而言,通过执行 f 后除计算结果 Y_j 外无法获取或推导任何其他参与方拥有的数据;对于密钥中心机构和半可信挖掘者而言,无法获取或推导任何通过安全信道所传输的数据以外的信息。

② 公平性:各数据源提供方同时获取最终计算结果。

③ 准确性:数据源提供方在密钥中心机构和半可信挖掘者的协议下,能准确地实现聚类任务。

3.1 全同态加密协议安全性

定义 2 公钥加密方案的单向性安全[20-21]:对于公钥加密方案(K,E,D),敌手 A 仅知道公开信息的情况下,在概率空间 $M\times B$ 上成功地对 E 求逆的概率是可忽略的,即

$$Succ_A=\Pr[(k_p,k_s)\leftarrow K(1^k):A(k_p,m;r)]=m$$

是可忽略的,称公钥加密方案是单向性安全。其中,M 是消息空间;B 是加密方案的随机抛币空间;K 是密钥生成算法;E 是加密算法;k_p 是公钥;k_s 是私钥;r 是随机输入;m 是明文。

定义 3 公钥加密方案的语义安全性/多项式时间不

可区分安全性[20-21]：对于公钥加密方案(K,E,D)，敌手$A=(A_1,A_2)$是一个二阶段攻击者，如果满足

$$adv_A = 2\times \Pr[(k_p,k_s)\leftarrow K(1^k),(m_0,m_1)\leftarrow A_1(k_p),$$
$$c = E(kp,\mathrm{m}_b;r) : A_2(m_0,m_1,c) = b] - 1$$

是可忽略的，称公钥加密方案是语义安全的或多项式时间不可区分的。其中，m_0，m_1 是长度相等的明文消息；c 是密文；$b\in\{0,1\}$是通过随机抛币得到比特。

定理 1　本文所采用的全同态加密协议[19]具备极微本原可靠、单向性安全和语义性安全。

证明　极微本原是安全协议的最基本组成构件，是协议可证明安全性的必要条件。本文所采用的全同态加密协议的极微本原是离散对数的计算困难问题，人们思考这个问题已经有几百年历史，由于不能破译而广泛接受其安全性。然而，极微本原可靠并不是协议安全的充分条件。因此，还需要考虑协议的单向性安全和语义安全性。

协议的单向性安全是指当敌手知道公开信息的情况下能否对一个给定的密文 c 恢复其对应的明文 m。由于协议中使用非对称密码，加密公钥$\langle v,p,n\rangle$和解密私钥$\langle t\rangle$之间没有本质联系，$b'\equiv s\cdot lcm(key_1^v,key_2^v,\cdots,key_r^v)\bmod n$，对于给定模数 n、整数 b 和 $s\cdot lcm(key_1^v,key_2^v,\cdots,key_r^v)$，计算指数 t 是很困难的，即无法从加密公钥中获取有关解密私钥的信息。这样，即便参与方获知存储在半可信挖掘第三方的密文 c，因为无法获取解密私钥，所以不能破译其他参与方的原始信息，从而抵御了已知密文攻击。因此全同态加密协议是单向性安全的。

协议的语义安全性是指敌手了解明文的某些信息后仍不能从密文得到除明文长度以外的任何信息。由密钥中心机构发送给各参与方相同的加密公钥$\langle v,p,n\rangle$和不同的加密私钥$\langle key_j\rangle$即相同的明文 m 经不同参与方加密后的密文 c_1 和 c_2 不相同的，解密 c_1 的私钥 t_1 满足 $b'\equiv key_1^v \bmod n$，解密 c_2 的私钥 t_2 满足 $b'\equiv key_2^v \bmod n$。因此，敌手在已知明密文对(m,c_1)和(m,c_2)的情况下，也无法破译新密文 c_3 的对应明文。　　证毕

3.2　隐私保护聚类模型安全性

定义 4　半可信模型下多方安全计算[23]：设定 n 代表参与方的个数，有确定性 n 元函数 f：$(\{0,1\}^*)^n\rightarrow(\{0,1\}^*)^n$，对于 $I=\{i_1,\cdots,i_t\}\subseteq\{1,\cdots,n\}$，令 f_I 代表 $f_{i1}(x_1,\cdots,x_n),\cdots,f_{it}(x_1,\cdots,x_n)$组成的序列。设 Π 是计算 f 的有 n 成员参与的协议。在 Π 的一次以$(x_1,\cdots,x_n)$为输入的运行中，第 i 个成员的所见记为 $VIEW_i^{\Pi}(\bar{x})$，被定义为$(x_i,r_i,m_1,\cdots,m_{Ni})$，其中 r_i 代表第 i 个成员的随机串，m_j 代表他接受到的第 j 个消息，N_i 代表他接受到的消息数目。对于 $I=\{i_1,\cdots,i_t\}$，令 $VIEW_I^{\Pi}(\bar{x})=(I,VIEW_{i1}^{\Pi}(\bar{x}),\cdots,VIEW_{it}^{\Pi}(\bar{x}))$，如果存在多项式时间算法 S，使得对于每个如上所述的 I，有如下等式成立：

$$S(I,(x_{i1},\cdots,x_{it}),f_I(\bar{x}))_{\bar{x}\in(\{0,1\}^*)^n}\equiv$$
$$VIEW_I^{\Pi}(\bar{x})_{\bar{x}\in(\{0,1\}^*)^n}$$

则 Π 安全计算了 f。

基于定义 4，在半可信模型下，如果每个参与方都可以多项式的时间通过该方的输入和输出在多项式时间内模拟出与真实视图计算不可区分的模拟视图，那么这个协议就是安全的，详细证明可参考文献[10]。

定理 2　在半可信模型假设中，在安全多方计算理论条件下，本文提出的隐私保护聚类模型是安全的。

证明　如第 2 节所示的隐私保护聚类模型，协议在步骤 2、步骤 4 和步骤 5 进行了数据交换。我们只需要证明这 3 个步骤中，如果参与方根据本地输入和全局输出多项式时间内得到的模拟视图和真实视图是计算不可区分的，那么该隐私保护聚类模型是安全的[10]。我们基于仿真的证明方法[10]来解决该问题，通过构造一个仿真器 Simulator 来仿真协议的运行，一旦仿真结果与实际协议运行中成员的所见计算不可区分，则意味着协议满足安全性。

如果 Pa 是敌手，仿真器 Simulator 以 Pa 的输入$\{\mu_{1a},\mu_{2a},\cdots,\mu_{ka}\}$为输入，Simulator 的隐式输入为 Pa 的加密公钥和私钥。在步骤 2 中，Pa 计算局部统计数据 $M_{isa}(1\leqslant s\leqslant k,1\leqslant i\leqslant N)$，通过添加扰乱数据 l_{isa} 和全同态加密协议生成 $E(M_{isa}+l_{isa})$，然后发送给半可信挖掘者；在步骤 4 中，半可信挖掘者将记录各簇的相关对象集索引数组 cluster[s]发送给各参与方，Pa 依据 cluster[s]中涉及的对象，本地计算$\{\mu'_a,\mu'_{2a},\cdots,\mu'_{ka}\}$实现质心更新；在步骤 5 中，$Pa$ 计算局部统计数据 $O_{sa}=(\mu_{sa}-\mu'_{sa})^2(1\leqslant s\leqslant k)$，并利用同态加密协议和添加扰乱数据思想生成 $E(O_{sa}+l_{sa})$，然后发送给半可信挖掘者。很明显，仿真器 Simulator 和 Pa 在实际协议运行中的输入、输出相同；又由于全同态加密协议的语义安全性，仿真中每一步中计算的密文与实际协议运行中每一步产生的密文是计算不可区分的，因此仿真结果与 Pa 在实际协议运行中看到的消息是计算不可区分的，这也就意味着除了输入和输出的信息外，敌手不能从协议运行中获得更多的信息。故本文提出的隐私保护模型是安全的。　证毕

3.3　效率分析

按照第 2 节所示的隐私保护聚类模型，在步骤 2 中本地计算局部统计数据 M_{isj} 的时间复杂度为 $O(krN)$，采用数据扰乱及同态加密生成密文的时间复杂度为 $O(krN)$，发送给半可信挖掘者的通信量为 $O(krN)$；在步骤 3 中，半可信挖掘者利用安全比较协议求解 cluster[s]的时间复杂度为 $O(krN)$；步骤 4 中各参与方本地更新质心的时间复杂度为 $O(|\ \mathrm{cluster}[s]\ |\ r)$；步骤 5 中本地计算局部统计数据 O_{sj} 的时间复杂度为 $O(kr)$，采用数据扰乱和全同态加密生成

密文的时间复杂度为 $O(kr)$，发送给半可信挖掘者的通信量为 $O(kr)$；步骤 6 中半可信挖掘者统计数据的时间复杂度为 $O(kr)$，发送给密钥中心机构的通信量为 $O(k)$。

4 结束语

本文讨论了垂直分布式聚类挖掘的隐私安全问题，基于全同态加密协议和数据扰乱方法设计了安全比较协议，从而实现了隐私保护的垂直分布式聚类模型。该模型通过添加随机向量扰乱原始数据和全同态加密协议实现密文上的统计分析。由于随机向量的选取特性保证了统计数据的扰乱前后的一致性，但有效防范了原始数据的泄露；采用全同态加密协议实现了半可信挖掘者在不获取明文的情况下可对统计数据的分析，从而既保护了原始数据也保护了中间计算结果。文章证明了所采用的全同态加密协议具备极微本原可靠、单向性安全和语义性安全；并采用多方安全计算定义和基于仿真的证明方法论证隐私保护聚类模型的安全性。

参考文献：

[1] Agrawal R, Srikant R. Privacy -preserving data mining[C]// *Proc. of the ACM SIGMOD International Conference on Management of Data*, 2000:439 - 450.

[2] Rizvi S J, Haritsa J R. Maintaining data privacy in association rule mining[C]// *Proc. of the 28th International Conference on Very Large Databases*, 2002:682 - 693.

[3] Du W L, Zhang Z J. Building decision tree classifier on private data[C]// *Proc. of the IEEE ICDM Workshop on Privacy, Security and Data Mining*, 2002:1 - 8.

[4] 葛伟平，汪卫，周皓锋，等. 基于隐私保护的分类挖掘[J]. 计算机研究与发展, 2006, 43(1):39-45. (Ge W P, Wang W, Zhou H F, et al. Classification mining based on privacy preserving [J]. *Journal of Computer Research and Development*, 2006, 43(1): 39 - 45.)

[5] 张鹏，唐世渭. 一种有效的隐私保护关联规则挖掘方法[J]. 软件学报，2006，17(8)：1764 - 1774. (Zhang P, Tang S W. One effective privacy preserving association rule mining method[J]. *Journal of Software*, 2006, 17(8):1764 - 1774.)

[6] Islam M, Brankovic L. Privacy preserving data mining: a noise addition framework using a novel clustering technique [J]. *Knowledge-based Systems*, 2011, 24(1): 1214 - 1223.

[7] Charu C, Yu P. *Privacy preserving data mining models and algorithms*[M]. New York:Springer Science+Business Media, 2007.

[8] Pinkas B. Cryptographic techniques for privacy-preserving data mining [J]. *ACM SIGKDD Explorations Newsletter*, 2002, 4(2): 12 - 19.

[9] Cliffton C, Kantarcioglu M, Vaidya J. Tools for privacy preserving distributed data mining [J]. *ACM SIGKDD Explorations Newsletter*, 2004, 4(2):28 - 34.

[10] Goldreich O. *Foundations of cryptography, basic applications*[M]. Cambridge: Cambridge University Press, 2004:233 - 278.

[11] Vaidya J, Clifton C. Privacy preserving association rule mining in vertically partitioned data[C]// *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002:639 - 644.

[12] Zhan J, Matwin S, Chang L. Privacy-preserving collaborative association rule mining [J]. *Journal of Network and Computer Applications*, 2007, l30(1):1216 - 1227.

[13] Emekci F. Privacy preserving decision tree over multiple parties [J]. *Data & Knowledge Engineering*, 2007, 63(1): 348 - 361.

[14] 张鹏，唐世渭. 朴素贝叶斯分类中的隐私保护方法研究[J]. 计算机学报，2007，30(8)：1267 - 1276. (Zhang P, Tang S W, Bayesian classification privacy preserving method research [J]. *Journal of Computer*, 2007, 30(8): 1267 - 1276.)

[15] Kim S W, Park S, Won J I, et al. Privacy preserving data mining of sequential patterns for network traffic data [J]. *Information Sciences*, 2008, 178(3) :694 - 713.

[16] Ali I, Yucel S. Privacy preserving clustering on horizontally partitioned data [J]. *Data &Knowledge Engineering*, 2007, 63 (2): 646 - 666.

[17] Jagannathan G, Pillaipakkamnatt K. A new privacy preserving distributed K-clustering algorithm [C]// *Proc. of the SIAM International Conference on Data Mining*, 2006:1213 - 1223.

[18] Vaidya J, Clifton C. Privacy preserving K-means clustering over vertically partitioned data [C]// *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003:490 - 510.

[19] 方炜炜，任江，夏红科. 异构分布的多元线性回归隐私保护模型[J]. 计算机研究与发展，2011，48(9)：1685 - 1692. (Fang W W, Ren J, Xia H K. Heterogeneous distributed linear regression privacy-preserving modeling [J]. *Journal of Computer Research and Development*, 2011, 48(9): 1685 - 1692.)

[20] 王克，戴一奇. 统计分布的多方保密计算[J]. 计算机研究与发展，2010，47(2)：201 - 206. (Wang K, Dai Y Q. Secure multiparty computation of statistical distribution[J]. *Journal of Computer Research and Development*, 2010, 47(2):201 - 206.)

[21] 冯登国. 可证明安全性理论与方法研究[J]. 软件学报，2005，16(1)：1743－1755. (Feng D G. Proved security theory and method research[J]. *Journal of Software*, 2005, 16(1):1743－1755.)

[22] 沈昌祥，张焕国，冯登国，等. 信息安全综述[J]. 中国科学，2007，37(2)：129 - 150. (Shen C X, Zhang H G, Feng D G, et al. Information security review [J]. *Chinese Science*, 2007, 37 (2): 129 - 150.)

[23] Yao A C. Protocols for secure computation[C]// *Proc. of the 23rd Annual IEEE Symposium on Foundations of Computer Science*, 1982:160 - 164.