分布式聚类算法的隐私保护研究

刘英华[1,2] 杨炳儒[1] 曹丹阳[1] 马 楠[1]

(北京科技大学信息工程学院 北京 100083)[1] (中国青年政治学院 北京 100089)[2]

摘 要 隐私保护数据挖掘是在不精确访问原始数据的基础上,挖掘出准确的规则和知识。针对分布式环境下聚类挖掘算法的隐私保护问题,提出了一种基于完全同态加密的分布式聚类挖掘算法(FHE-DK-MEANS算法)。理论分析和实验结果表明,FHE-DK-MEANS算法不仅具有很好的数据隐私性,而且保持了聚类精度。

关键词 数据挖掘,隐私保护,聚类,分布式数据

中图法分类号 TP311 **文献标识码** A

Research on Privacy Preserving Distributed Clustering Algorithm

LIU Ying-hua[1,2] YANG Bing-ru[1] CAO Dan-yang[1] MA Nan[1]

(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)[1]

(China Youth University for Political Science, Beijing 100089, China)[2]

Abstract Privacy preserving data mining is to discover accurate rules and knowledge without precise access to the raw data. This paper focused on privacy preserving clustering algorithms mining in a distributed environment, and presented a fully homomorphic encryption algorithm based on distributed k-means (FHE-DK-MEANS algorithm). Theoretical analysis and experimental results show that FHE-DK-MEANS algorithm can provide better privacy and accuracy.

Keywords Data mining, Privacy preserving, Clustering, Distributed data

数据挖掘就是从海量数据中发现潜在的、隐含的、未知而有价值的知识。然而,随着数据挖掘在为某些客户提供有价值的知识的同时,一些隐私也被泄露了。隐私一般分为原始数据隐私和数据挖掘隐私两种。2000年Agrawal等人首次提出隐私数据挖掘(PPDM, Privacy Preserving Data Mining)的概念,现在PPDM已经成为数据挖掘的一个重要研究方向。

隐私保护数据挖掘的最终目的就是在保证数据挖掘的同时尽可能保护敏感数据,在不精确访问原始真实数据的条件下,得到准确的模型和分析结果。隐私保护技术主要分为数据扰动技术、数据加密和查询限制3种。数据扰动技术是采用数据交换、添加噪声等方法扰动原始数据,使敏感数据失真,但能保证通过数据挖掘工具挖掘出的知识真实有效;数据加密是基于加密技术对原始数据加密以保护隐私。查询限制是通过限制数据的查询,避免数据挖掘者获取完整真实的原始数据,以实现数据隐私的保护。

分布式隐私保护数据挖掘是指在存储于多个站点的数据集中完成数据挖掘,侧重于如何通过各站点共享的数据挖掘出有价值的知识,同时避免各站点的原始数据被其他站点获取。

聚类是将数据划分成不同的类或簇的过程,相同类或簇中的数据具有很大的相似性,不同的类或簇中的数据则相似度低。聚类分析是实际挖掘的主要任务之一,它可以作为一个独立的工具获得数据的分布状况,通过各组类或簇的数据特征进一步地分析和挖掘数据。

本文提出了一种基于完全同态加密的分布式聚类挖掘算法。在分布式环境中,各参与站点将数据通过同态加密技术加密后共享,在加密后的数据计算后,通过同态加密技术对计算结果解密,即所有操作均在加密后的数据上进行,半可信第三方数据挖掘者在加密后的数据中进行聚类数据挖掘,避免使用原始明文数据,从而实现了数据挖掘中的隐私保护。理论证明和实验结果充分表明,此算法在隐私保护数据挖掘方面是有效的。

1 问题描述

分布式环境下的数据挖掘主要分为两种:一种是水平划分数据集,另一种是垂直划分数据集。文本主要研究在多个水平划分数据站点上的隐私保护聚类挖掘算法。

设水平分布式系统中包含 n 个站点 $S_i(i=1,\cdots,n,n\geqslant 3)$,每个站点的数据集是 $D_i(i=1,\cdots,n,n\geqslant 3)$,每个数据集 D_i 包含的对象个数是 $m_i(i=1,\cdots,n,n\geqslant 3)$,则联合数据集为 $D=\bigcup_{i=1}^{n}D_i(i=1,\cdots,n,n\geqslant 3)$。在对联合数据集 D 进行聚类挖掘时,要保证每个站点 S_i 的数据集 D_i 的数据安全,即其他站点不能通过最终的聚类挖掘结果挖掘出原始的数据集 D_i,而且要确保通过联合数据集 D 挖掘的知识是真实有效的,即与

到稿日期:2011-04-23 返修日期:2011-07-14 本文受国家自然科学基金项目(60875029),北京市科技计划专项课题资助。

刘英华(1975—),女,博士生,主要研究方向为隐私保护数据挖掘、知识发现,E-mail:yinghliu@163.com;杨炳儒(1943—),男,教授,主要研究方向为知识发现与智能系统、柔性建模与集成技术;曹丹阳(1978—),男,博士生,主要研究方向为数据挖掘;马 楠(1978—),女,博士生,主要研究方向为数据挖掘。

真实访问原始数据集 D_i 挖掘出的结果相同。

2 隐私保护的分布式聚类算法研究

在分布式环境中实现隐私保护数据挖掘，必须解决站点间安全数据传输的问题，加密是一种常见的方法。

2.1 完全同态加密技术

1978 年 Rivest 等人在文献[1]中提出了秘密同态(PM, Privacy Homomorphism)的概念，1996 年 Domingo 等人在文献[2]中对文献[1]存在的已知明文攻击不安全问题做了改进。秘密同态技术是对加密的明文数据进行处理(例如加、减、乘、除、模等运算)，得到一个输出，解密后的输出结果与用同一方法处理未加密的明文数据得到的输出结果是一样的，这些算法都属于部分同态。2009 年 Craig Gentry 在文献[3]中提出完全同态加密(Fully Homomorphic Encryption)技术。如果一种加密算法对于乘法和加法都能找到对应的操作，则称其为全同态加密算法。

定义 1 设加密操作为 E，解密操作为 D，明文为 m_i，符号⊕和⊗代表运算符。若有等式 $D(E(m_1)\oplus E(m_2)\oplus\cdots\oplus E(m_i))=D(E(m_1\otimes m_2\otimes\cdots\otimes m_i))$ 成立，则 $(E,M,\oplus,\otimes)$ 满足同态加密，也称为部分同态。

定义 2 设加密操作为 E，解密操作为 D，明文为 m_i，符号⊕和⊗代表运算符。若有等式 $D(E(m_1)\oplus E(m_2)\oplus\cdots\oplus E(m_i))=D(E(m_1+m_2+\cdots+m_i))$ 成立，且 $D(E(m_1)\otimes E(m_2)\otimes\cdots\otimes E(m_i))=D(E(m_1\times m_2\times\cdots\times m_i))$，则 $(E,M,\oplus,\otimes)$ 满足完全同态加密。亦即一种同态加密算法对于乘法运算和加法运算都能找到对应的操作，则称其为完全同态加密。

本文提出了一种全新的完全同态加密算法。若加密数据 m，则选择一个大数 p 作为密钥，然后随机选择一个大数 k，且 k 满足 $m<<k<<p$(符号<<表示远远小于)，随机产生两个小整数 r 和 q。

加密算法：$E(m)=qp+kr+m$

解密算法：$D(E(m))=(E(m) \bmod p) \bmod k$

设 $E(m_i)=q_ip+kr_i+m_i$，则 $E(m_1)+E(m_2)+\cdots+E(m_i)=(q_1+q_2+\cdots+q_i)p+(r_1+r_2+\cdots+r_i)k+(m_1+m_2+\cdots+m_i)$，因为 $(r_1+r_2+\cdots+r_i)k+(m_1+m_2+\cdots+m_i)$ 依旧远远小于 p，$(m_1+m_2+\cdots+m_i)$ 依旧远远小于 k，所以有

$D(E(m_1)+E(m_2)+\cdots+E(m_i))=((E(m_1)+E(m_2)+\cdots+E(m_i)) \bmod p) \bmod k=(k(r_1+r_2+\cdots+r_i)+(m_1+m_2+\cdots+m_i)) \bmod k=m_1+m_2+\cdots+m_i$

则 $E(m_1)\times E(m_2)\times\cdots\times E(m_i)=(A)p+(B)k+(m_1\times m_2\times\cdots\times m_i)$，其中 $(A)p$ 表示计算后具有 p 因子的表达式，$(B)k$ 表示计算后具有 k 因子的表达式，因为 $(B)k+(m_1\times m_2\times\cdots\times m_i)$ 依旧远远小于 p，$m_1\times m_2\times\cdots\times m_i$ 依旧远远小于 k，所以有

$D(E(m_1)\times E(m_2)\times\cdots\times E(m_i))=((E(m_1)\times E(m_2)\times\cdots\times E(m_i)) \bmod p) \bmod k=((B)k+(m_1\times m_2\times\cdots\times m_i)) \bmod k=m_1\times m_2\times\cdots\times m_i$

2.2 分布式聚类算法 K-means

K-means 算法是基于距离的聚类算法，使用距离作为相似度的评价指标，即根据各簇中对象的距离均值计算相似度。距离越近，相似度就越大。所以 K-means 算法的最终目标是得到簇内相似度高且与其他簇相似度低的簇。

设联合数据集 D 初始分成 k 个聚簇 $w_j(j=1,\cdots,k)$，簇中对象个数为 $t_j(j=1,\cdots,k)$，初始中心点分别为 $c_j(j=1,\cdots,k)$，联合数据集 D 的聚簇 $w_j(j=1,\cdots,k)$ 对应的站点 $S_i(i=1,\cdots,n,n\geqslant 3)$ 的局部聚类中心是 $c_{ij}(i=1,\cdots,n,j=1,\cdots,k,n\geqslant 3)$，联合数据集 D 的聚簇 $w_j(j=1,\cdots,k)$ 的数据对象个数是 m_{ij} $(i=1,\cdots,n,j=1,\cdots,k,n\geqslant 3)$，则计算聚类函数 $E=\sum_{i=1}^{k}\sum_{j=1}^{t_i}d_{ij}(x_j,c_i)$ 是此算法的关键。

分布式 K-MEANS(DK-MEANS)算法的基本思想是采用主站点、从站点两级架构。从站点 $S_i(i=1,\cdots,n,n\geqslant 3)$ 计算本地站点聚簇数据后发送到主站点，主站点根据从站点发来的数据计算联合数据集聚簇中心。

输入：各站点数据集 D_i，每个 D_i 对象个数 $m_i(i=1,\cdots,n,n\geqslant 3)$，聚簇个数 k。

输出：k 个聚簇。

分布式 K-MEANS 算法——主站点：

1) repeat;

2) 接收从站点的聚类中心点 c_{ij} 及相应的对象个数 $m_{ij}(i=1,\cdots,n,j=1,\cdots,k,n\geqslant 3)$；

3) 计算联合数据集聚类中心点 $c_j(j=1,\cdots,k)$，$c_j=\sum_{i=1,j=1}^{i=n,j=k}(c_{ij}\times m_{ij})/\sum_{i=1,j=1}^{i=n,j=k}m_{ij}$；

4) 主站点随机产生 k 个初始聚簇中心，并发送到从站点 $S_i(i=1,\cdots,n,n\geqslant 3)$；

5) 计算 $\sum_{i=1}^{k}\sum_{j=1}^{t_i}d_{ij}(x_j,c_i)$；

6) until 每个聚类不再发生变化。

分布式 K-MEANS 算法——从站点：

1) 从站点 $S_i(i=1,\cdots,n,n\geqslant 3)$ 分别接收主站点发送的 k 个初始聚簇中心；

2) repeat

3) 从站点 $S_i(i=1,\cdots,n,n\geqslant 3)$ 根据主站点发来的初始聚簇中心计算其与本站点数据集 D_i 包含的 $m_i(i=1,\cdots,n,n\geqslant 3)$ 个对象的距离，确定每个 m_i 的所属聚类；

4) 计算各从站点的聚类中心点 c_{ij} 及相应的对象个数 $m_{ij}(i=1,\cdots,n,j=1,\cdots,k,n\geqslant 3)$，并发送到主站点；

5) until 每个聚类不再发生变化。

2.3 隐私保护算法

基于完全同态加密的分布式隐私保护数据挖掘算法(FHE-DK-MEANS)的基本思想是：从站点 $S_i(i=1,\cdots,n,n\geqslant 3)$ 计算本地站点聚簇数据，采用本文设计的完全同态加密算法将其加密后发送到主站点，主站点对密文进行计算，然后采用完全同态解密算法解密。

输入：各站点数据集 D_i，每个 D_i 对象个数 $m_i(i=1,\cdots,n,n\geqslant 3)$，聚簇个数 k。

输出：k 个聚簇。

FHE-DK-MEANS 算法——主站点：

1) 主站点随机产生 k 个初始聚簇中心，并发送到从站点 $S_i(i=1,\cdots,n,n\geqslant 3)$；

2) repeat;

3) 接收从站点发来的加密数据 c'_{ij} 和 m'_{ij}；

4) 分别计算 $\sum_{i=1,j=1}^{i=n,j=k}(c'_{ij}\times m'_{ij})$ 和 $\sum_{i=1,j=1}^{i=n,j=k}m'_{ij}$ 后，根据解密算法求出 $\sum_{i=1,j=1}^{i=n,j=k}(c_{ij}\times m_{ij})$ 和 $\sum_{i=1,j=1}^{i=n,j=k}m_{ij}$，然后计算联合数据集聚类中心点 c_j

$(j=1,\cdots,k)$；

5) 计算$\sum_{i=1}^{k}\sum_{j=1}^{ti} d_{ij}(x_j,c_i)$；

6) until 每个聚类不再发生变化。

FHE-DK-MEANS 算法——从站点：

1) 从站点 $S_i(i=1,\cdots,n,n\geqslant 3)$分别接收主站点发送的 k 个初始聚簇中心；

2) repeat；

3) 从站点 $S_i(i=1,\cdots,n,n\geqslant 3)$根据主站点发来的初始聚簇中心计算其与本站点数据集 D_i 包含的 $m_i(i=1\cdots n,n\geqslant 3)$ 个对象的距离，确定每个 $m_i(i=1,\cdots,n,n\geqslant 3)$的所属聚类；

4)计算各从站点的聚类中心点及相应的对象个数 $m_{ij}(i=1,\cdots,n,j=1,\cdots,k,n\geqslant 3)$，完全同态加密后 c'_{ij} 和 m'_{ij} 发送到主站点；

5) until 每个聚类不再发生变化。

3 实验结果和分析

本文基于 Weka 平台封装了 FHE-DK-MEANS 算法，对 Weka 的 2 个数据集和 UCI 机器学习数据集中的 3 个数据集进行了 FHE-DK-MEANS 算法和 DK-MEANS 算法的对比实验，实验结果见图 1 和图 2。FHE-DK-MEANS 算法的聚类精度与不加隐私保护的聚类算法聚类精度一致。隐私保护度最高为 91%，最低为 82%。从实验结果可以得出结论，FHE-DK-MEANS 算法在保持较高分类精度的同时也能取得较好的隐私保护效果。

FHE-DK-MEANS 算法中数据在主、从站点间收发，收发的数据均为加密数据，整个过程中任何参与站点均无法获取其他站点的原始真实信息，所以 FHE-DK-MEANS 算法很好地保护了分布式各站点信息的隐私。从实验结果得出 FHE-DK-MEANS 算法的执行时间略高，这是因为加密过程增加了算法的执行时间。

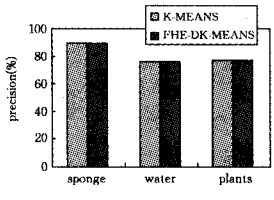

图 1 聚类精度比较

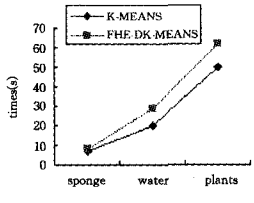

图 2 执行时间比较

结束语 本文针对面向分布式聚类数据挖掘算法中的隐私保护问题，提出了保护效果较好的 FHE-DK-MEANS 算法。算法对从站点数据集采用完全加密算法加密，加密后的数据发送到主站点。主站点对加密数据计算后，根据完全加密算法解密得到结果。主站点得到的结果与原始真实数据集明文传送得到的效果是一致的。理论证明和实验分析均表明该算法能较好地保护隐私数据，同时又能有效保持分布式数据集聚类挖掘的可用性。

参 考 文 献

[1] Rivest R L, Adleman L, Detrouzos M L. On Data Banks and Privacy Homomorphism[C]// Foundations of Secure Computation. New York: Academic Press, 1978: 169-179

[2] Domingo F J, Herrear J I J. A new privacy homomorphism and applications[J]. Information Processing Letters, 1996, 60(5): 277-282

[3] Gentry C. Fully Homomorphic Encryption Using Ideal Lattices [C]// Proc of the 41st Annual ACM Symposium on Theory of Computing (STOC'09). Bethesda, USA. New York, USA: ACM, 2009: 169-178

[4] Kamakshi P, Babu A V. Preserving Privacy and Sharing the Data in Distributed Environment Using Cryptographic Technique on Perturbed data [J]. Journal of Computing, 2010, 2(4): 115-119

[5] Lindell Y, Pinkas B. Privacy preserving data mining[J]. Jorunal of Cryptology, 2002, 15: 177-206

[6] Vaidya J, Lifton C. Privacy-preserving K-Means Clustering over Vertically Partitioned Data[C]// Proc the SIGKDD '03. Washington DC, USA, 2003: 24-27

[7] Kumar K A, Rangan C P. Privacy Preserving DBSCAN Algorithm for Clustering [J]. Advanced Data Mining and Applications, 2007, 4632: 57-68

[8] Li Xiong, Jurczyk P, Liu Ling. Mining Distributed Private Databases Using Random Response Protocols[C]// National Science Foundation Symposium on Next Generation of Data Mining and Cyber-enabled Discovery for Innovation. Baltimore, USA, 2007

[9] Ali İnan, Kaya S V, Saygln Y. Privacy preserving clustering on horizontally partitioned data[J]. Data & Knowledge Engineering, 2007, 63(3): 646-666

[10] 童云海，陶有东，唐世渭，等. 隐私保护数据发布中身份保持的匿名方法[J]. 软件学报，2010，21(4)：771-781

(上接第 130 页)

[3] Aeronautical Radio, Inc. ARINC Specification 653—2006 Avionics application software standard interface[S]

[4] 褚文奎，张凤鸣，樊晓光. 综合模块化航空电子软件体系结构综述[J]. 航空学报，2009，30(10)：1912-1917

[5] NASA. NASA-STD-8719. 13B—2004 Software safety standard [S]

[6] Johnson W G. MORT safety assurance systems[M]. Marcel Dekker, Inc., 1980

[7] 樊晓光，褚文奎，张凤鸣. 软件安全性研究综述[J]. 计算机科学，2011，38(5)：8-13，27

[8] Kazman R, Klein M, Barbacci M, et al. The architecture tradeoff analysis method[C]// 4th International Conference on Engineering of Complex Computer Systems (ICECCS'98). 1998: 35-45

[9] Conmy P M. Safety Analysis of Computer Resource Management of Software [D]. Heslington (England): University of York, 2005

[10] SAE. ARP4761—1996 Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment [S]

[11] Bate I, Hawkins R, McDermid J. A contract-based approach to designing safe systems[C]// 8th Australian Workshop on Safety Critical Systems and Software. Canberra, 2003, CRPIT 33: 25-36

[12] 刘安. 基于模型驱动开发方法的开放式结构悬挂物管理系统研究[D]. 西安：空军工程大学，2010

[13] 郑人杰，殷人昆，陶永雷. 实用软件工程(第二版)[M]. 北京：清华大学出版社，1997