水平分割数据的保护隐私聚类挖掘方法研究

袁 武,任勋益
(南京邮电大学 计算机学院,江苏 南京 210000)

摘 要:随着大数据时代的到来,数据共享在商业、政府和其他机构之间日渐频繁,如何保护各参与方的数据隐私成为亟待解决的问题。文中针对水平划分的数据容易产生的各参与方数据隐私泄露,共谋攻击和分布式、准诚信、大数据的挖掘环境中的新特点,提出了一种保护隐私的聚类挖掘算法。该方法结合 RSA 公钥加密技术和同态加密技术等密码学方法的优势,能够在不降低挖掘精度的前提下,保护各参与方的数据隐私。分析表明,它能够保证挖掘结果的安全性、有效性和正确性。

关键词:隐私保护;同态加密;水平分割数据;聚类挖掘; K -means 算法
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2015)05-0115-03
doi:10.3969/j.issn.1673-629X.2015.05.027

Research on Privacy Preserving Clustering Method for Horizontal Partitioned Data

YUAN Wu, REN Xun-yi
(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210000, China)

Abstract: Due to the advent of big data age, data sharing between business, governments and other parties is more and more frequent. Privacy preserving has become an important issue in data mining. In this paper, in view of horizontally partitioned data is easy to produce the parties data privacy, collusion attack and new features of distributed, semi-honest partner, big data mining in the environment, propose a clustering mining method to protect the privacy. Combined the advantages of RSA public key cryptosystem, homomorphic encryption scheme and other cryptograph methods, can preserve the privacy of all parties on the premise of not reducing the mining accuracy. The theoretical analysis shows that this method can guarantee the security, validity and correctness for result.

Key words: privacy preserving; homomorphic encryption; horizontally partitioned data; clustering mining; K -means algorithm

0 引 言

近年来,随着信息产业迈入大数据时代,不同应用领域的数据激增,数据挖掘 Data Mining (DM)理论及方法逐步成为研究的热点。数据挖掘的目的在于从海量数据中发现潜在的、未知的、隐含而有价值的规律和知识,从而指导人们更好地规划生产和销售活动。然而,在数据挖掘为客户提供有益知识的同时,也会直接或间接地泄露用户的数据隐私,这成为数据挖掘不能逃避的问题。具体来讲,隐私一般分为原始数据隐私和数据挖掘隐私两种。如何在保护用户隐私的前提下进行数据挖掘,成为数据挖掘领域的一个主要的研究方向。保护隐私的数据挖掘为这一问题提供了行之有效的解决方案。

保护隐私数据挖掘致力于在保证数据挖掘精度的前提下尽可能地保护数据隐私,使数据挖掘实施者在不精确访问原始数据的情况下,得到准确的模型和分析结果。到目前为止,学者们提出了很多保护隐私的数据挖掘方法:文献[1-4]提出了几种决策树分类中隐私保护的算法;文献[5-9]提出了几种保护隐私的聚类挖掘算法;文献[10-13]提出了几种保护隐私的关联规则挖掘算法;文献[14]提出了一种基于神经网络的保护隐私数据挖掘算法;文献[15]提出了一种基

收稿日期:2014-06-15 **修回日期**:2014-09-21 **网络出版时间**:2015-02-23
基金项目:国家自然科学基金资助项目(61073188)
作者简介:袁 武(1987-),男,硕士研究生,研究方向为保护隐私的数据挖掘;任勋益,副教授,博士,CCF 会员,研究方向为新型入侵检测技术、云端技术和虚拟化技术。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150223.1252.051.html

于贝叶斯网络的保护隐私数据挖掘算法。

参与挖掘的数据可能存储在单一数据库或是分布式的数据库中。分布式数据库的经典模式是数据仓库。在分布式环境中,由于数据存储在物理隔离的各个站点中,因此应该从以下几个方面进行数据隐私的保护:(1)保护各个站点的隐私数据安全,防止其他参与方获知不属于自身的数据。(2)保护传输过程中的数据安全,防止数据被恶意或半恶意的攻击者截获。(3)保护聚类挖掘过程中隐私数据的安全。

文中基于 RSA 公钥加密系统和同态加密技术提出一种基于水平划分数据的保护隐私数据挖掘方法。在分布式环境下,各参与方首先将数据进行同态加密,然后经由安全的信道进行数据共享,在对加密过的数据计算后,使用同态加密技术对计算结果解密,使得所有的计算均在加密后的数据上进行。准诚信的第三方在加密后的数据中进行聚类挖掘,避免直接使用用户的原始明文数据,同时在主站点和从站点之间传输密文数据,防止半诚信的参与方获取其他参与方的隐私数据。分析表明,此算法能在保护数据隐私的前提下得到正确的挖掘结果。

1 背景介绍

1.1 *K*–means 聚类算法

K –means 算法是基于距离的聚类挖掘算法,它使用距离作为相似度的评价标准,根据各簇中对象的距离及其均值计算相似度。距离越小,相似度越大。*K* –means 算法最终是要得到簇内相似度高且簇间相似度低的聚类结果。文中使用标准的欧拉距离来衡量各元素间的相似度,标准的欧拉距离计算公式如下:

$$D(x_i,x_j)=\sqrt{\sum_{q=1}^{r}(a_q(x_i)-a_q(x_j))^2}$$

1.2 同态加密和解密技术

秘密同态技术是对加密明文进行处理,得到一个输出,对输出结果进行解密后与对明文直接进行处理得到的结果相同。2009 年 Graig Gentry 提出基于理想格的完全同态加密技术,如果一种加密算法对加法和乘法都能找到其对应的同态操作,即满足 $e(m)*e(m')=e(m*m')$,则称该加密算法为全同态加密算法。在文中,提出一种新的同态加密和解密算法。

首先取一个大数 N 满足 $N=P\times Q$,其中 P 和 Q 为两个大素数。设 X 为待加密的明文,则加密计算如下:

$$Y=E_p(X)=(X+P\times R)\bmod N \tag{1}$$

其中,R 为$(1,Q)$ 中满足均匀分布的一个随机值。

密文 Y 传到接收端以后,使用密钥 p 解密 Y 得到明文,解密方法如下:

$$X=E^{-1}(Y)=D_p(Y)=(Y)\bmod P,Y=(X+P\times R)\bmod N \tag{2}$$

1.3 分布式环境中数据划分方法

分布式环境中数据有两种划分方法:一种是水平分割数据集;一种是垂直分割数据集。文中主要研究水平划分数据环境中的聚类挖掘算法。

假设分布式系统包含 n 个站点 $S_i(i=1,2,\cdots,n,n\geqslant 3)$,每个站点的数据集是 $D_i(i=1,2,\cdots,n,n\geqslant 3)$,每个数据集 $D_i(i=1,2,\cdots,n,n\geqslant 3)$ 包含的对象个数为 $m_i(i=1,2,\cdots,n,n\geqslant 3)$,则联合数据集 $D=\bigcup_{i=1}^{n}D_i(i=1,2,\cdots,n,n\geqslant 3)$。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结果挖掘出原始数据集 D_i,而且要保证联合数据集 D 挖掘的知识是真实有效的,即与直接挖掘 D_i 得到的结果相同。

2 分布式 *K*–means 聚类挖掘算法

分布式 *K* –means 聚类挖掘算法主要采取中心站点、局部站点两级架构。局部站点 $S_i(i=1,2,\cdots,n,n\geqslant 3)$ 计算本地聚簇数据后发送到中心站点,中心站点接收局部站点发来的聚簇结果,在云中计算联合数据集聚簇中心。

2.1 标准的分布式 *K*–means 聚类算法

输入:各站点的数据集 D_i,每个 D_i 对象个数 $m_i(i=1,2,\cdots,n,n\geqslant 3)$,聚簇个数 k;

输出:k 个最终聚簇。

分布式 *K* –means 算法—中心站点:

· 接收各局部站点聚类中心点 c_{ij} 及相应对象个数 $m_{ij}(i=1,2,\cdots,n,j=1,2,\cdots,k,n\geqslant 3)$;

· 计算全局数据集聚类中心点 $c_j(j=1,2,\cdots,k)$,$c_j=\sum_{i=1,j=1}^{i=n,j=k}(c_{ij}\times m_{ij})/\sum_{i=1,j=1}^{i=n,j=k}m_{ij}$;

· 中心站点随机产生 k 个初始聚簇中心,并发送到从站点 $S_i(i=1,2,\cdots,n,n\geqslant 3)$;

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

· 直到每个聚类不再发生变化。

分布式 *K* –means 算法—局部站点:

· 各局部站点 $S_i(i=1,2,\cdots,n,n\geqslant 3)$ 分别接收中心站点发来的 k 个初始聚簇中心;

· 局部站点 $S_i(i=1,2,\cdots,n,n\geqslant 3)$ 根据中心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 $m_i(i=1,2,\cdots,n,n\geqslant 3)$ 个对象的距离,确定每个 m_i 所属的类;

· 计算各局部站点的聚类中心点 c_{ij} 及相应的对象个数 $m_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,n,n\geqslant 3)$,并将结

果发送到中心站点;

· 直到每个聚类不再发生变化。

2.2 保护隐私的分布式 *K*-means 聚类挖掘算法

基于同态加密和 RSA 公钥加密技术,提出了分布式隐私保护数据挖掘算法。其基本思想是:局部站点 $S_i(i=1,2,\cdots,n,n\geqslant 3)$ 计算本地站点的聚簇数据,然后将计算结果加密后发送到中心站点,中心站点将局部站点结果发送到云中,由第三方对密文进行计算,然后采用完全同态解密算法进行解密。

输入:各局部站点数据集 D_i,每个 D_i 对象个数 $m_i(i=1,2,\cdots,n,n\geqslant 3)$,聚簇个数 k。

输出:k 个聚簇。

保护隐私的 K-means 聚类挖掘算法—中心站点:

· 中心站点使用 RSA 公钥加密算法生成密钥对 (e_i,d_i),将 e_i 发送到各参与方 $S_i(i=1,2,\cdots,n,n\geqslant 3)$;

· 中心站点随机产生 k 个初始聚类中心,并发送到各局部站点 $S_i(i=1,2,\cdots,n,n\geqslant 3)$;

· 接收各局部站点发来的加密数据 C_i,$c_{ij}^{'}$ 和 $m_{ij}^{'}$;

· 对 C_i 解密得到 P_i,在云中分别计算 $\sum_{i=1,j=1}^{i=n,j=k}(c_{ij}^{'}\times m_{ij}^{'})$ 和 $\sum_{i=1,j=1}^{i=n,j=k}m_{ij}^{'}$,根据解密算法(2)求出 $\sum_{i=1,j=1}^{i=n,j=k}(c_{ij}\times m_{ij})$ 和 $\sum_{i=1,j=1}^{i=n,j=k}m_{ij}$,然后计算联合数据聚类中心点 $c_j(j=1,2,\cdots,k)$;

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

· 直到每个聚类不再发生变化。

保护隐私的 K-means 聚类挖掘算法—局部站点:

· 各局部站点 $S_i(i=1,2,\cdots,n,n\geqslant 3)$ 分别接收中心站点发来的 k 个初始聚类中心和加密公钥 e_i;

· 局部站点 $S_i(i=1,2,\cdots,n,n\geqslant 3)$ 根据中心站点发来的初始聚类中心计算其与本站点数据集 D_i 包含的 $m_i(i=1,2,\cdots,n,n\geqslant 3)$ 个对象间的距离,确定每个 $m_i(i=1,2,\cdots,n,n\geqslant 3)$ 所属的类;

· 计算各局部站点聚类中心位置和其对象个数 $m_{ij}(i=1,2,\cdots,n,j=1,2,\cdots,k,n\geqslant 3)$;

· 使用(1)式中的方法对 c_{ij} 和 m_{ij} 进行完全同态加密得到 $m_{ij}^{'}=E_i(m_{ij})$ 和 $c_{ij}^{'}=E_i(c_{ij})$,使用 RSA 公钥加密算法加密 P_i 得到 $C_i=E_{e_i}(P_i)$,将 $m_{ij}^{'}$,$c_{ij}^{'}$ 和 C_i 发送回主站点;

· 直到每个聚类不再发生变化。

3 正确性和安全性分析

3.1 正确性分析

文中采用 RSA 公钥加密系统和同态加密系统对从站点的计算结果进行加密,以保证在半诚信的环境中,参与挖掘的各方数据隐私不被泄露。由于 RSA 公钥加密系统仅用于加密密钥,而同态加密系统的加密操作不影响最终的聚类结果,因此文中提出的算法能够获得准确的挖掘结果。由于加解密过程的存在,算法的时间复杂度会相应提高,但由于最为耗时的 RSA 公钥加密过程只用于加密特定参数而不是整个明文,可以减少大量的指数运算,因此挖掘过程时间的增加在可以容忍的范围之内。在挖掘过程中,中心站点可能出现大量的计算过程,将这个过程放在云端进行,这样能够有效减少挖掘过程的时间复杂度。

3.2 安全性分析

在该算法中,采用分层次的加密系统。首先使用同态加密过程对局部聚类结果进行加密,由于 R 是在 $(1,Q)$ 之间均匀分布的随机数,因此可以将它用作数字信封保存局部的聚类结果。将加解密参数 P 使用 RSA 公钥加密系统进行加密,连同加密后的聚类结果一起传输到主站点。由于实际应用中可能出现聚类结果较多、数据量较大的情况,因此主站点将收到的结果上传到云中进行聚类挖掘的过程,得到挖掘结果后返回主站点。RSA 公钥加密系统满足语义安全,因此各参与方都只能得到己方的输出和最终的结果,除此之外不能获得其他任何数据信息。由组合定理可知,如果某个协议的子协议和子协议的组织过程是语义安全的,则该协议就是语义安全的。因此,由于各参与方都能确保数据安全,在云中进行的聚类挖掘过程对同态加密过的数据进行挖掘,隐私数据的传输过程中无明文出现,所以文中提出的算法是安全的。

4 结束语

文中基于分布式环境提出了一种新的保护隐私的聚类挖掘算法。该算法采用语义安全的 RSA 公钥加密系统和同态加密系统来保护各参与方数据的安全。各参与方首先使用 K-means 算法计算局部的聚类过程,然后将计算结果进行加密,由中心站点接收局部聚类结果,并在云中完成剩余的挖掘工作。性能分析表明,该算法能够在有限增加时间复杂度和保持挖掘结果精确度的同时,防止各参与方隐私数据的泄露。

参考文献:

[1] Agrawal R, Srikant R K. Privacy-preseving data mining[C]// Proceedings of the ACM special interest group on management of data conference. Dallas, TX, USA: ACM, 2009: 439-450.

[2] Lindell Y, Pinkas B. Privacy-preserving data mining[C]// Proceedings of the 20th annual international cryptology. Santa

(下转第 122 页)

当 Domain Flux 失效后,启动 Random 协议实现命令控制,增加健壮性。

4 结束语

僵尸网络作为当今最具威胁的网络攻击计算平台,攻击手段多种多样,例如分布式拒绝服务攻击、垃圾邮件、网络钓鱼、点击欺诈或隐私窃取等等。它感染移动终端的最终目的就是获得利益,具有移动僵尸网络特色的攻击包含:恶意扣费、隐私信息窃取、垃圾短信干扰、电量消耗、网络流量消耗、DDOS 攻击等。由于现在的智能手机功能不断地增多,许多都涉及到个人的银行账户,多数功能都与经济挂钩,因此也诱导了不法分子不断地建立新的网络攻击行为。因此要警惕乱码的电话,不接收陌生短信,不接受陌生的网络连接请求,不在手机上浏览陌生网站和邮件,经常给手机杀毒,安装手机防火墙,等等。

参考文献:

[1] 王海龙,唐 勇,龚正虎. 僵尸网络命令与控制信道的特征提取模型研究[J]. 计算机工程与科学,2013,35(2):62-67.

[2] SymbOS. Exy. A[EB/OL]. 2009. http://www.symantec.com/security_response/writeup.jsp? docid=2009-022010-4100-99.

[3] Asrar I. Could sexy space be the birth of the sms botnet? [EB/OL]. 2009-07-13. http://www.symantec.com/connect/blogs/could sexy space be birth sms botnet.

[4] Porras P, Saidi H, Yegneswaran V. An analysis of the Ikee. B Iphone Botnet[C]//Proc of MOBISEC. Berlin: Springer, 2010:141-152.

[5] RootSmart[EB/OL]. 2012-02-03. http://www.csc.ncsu.edu/faculty/jiang/RootSmart/.

[6] 耿贵宁. 移动僵尸网络安全分析关键技术研究[D]. 北京:北京邮电大学,2012.

[7] 诸葛建伟,韩心慧,周勇林,等. 僵尸网络研究[J]. 软件学报,2008,19(3):702-715.

[8] 刘一静,孙 莹,蔺 洋. 基于手机病毒攻击方式的研究[J]. 信息安全与通信保密,2007(12):96-98.

[9] Su Jing, Chan K K W, Miklas A G, et al. A preliminary investigation of worm infections in a bluetooth environment[C]//Proc of ACM workshop on recurring malcode. Alexandria, VA: ACM, 2006.

[10] Singh K, Sangal S, Jain N, et al. Evaluating bluetooth as a medium for botnet command and control[C]//Proc of the 7th international conference on detection of intrusions and malware, and vulnerability assessment. Berlin: Springer, 2010:61-80.

[11] Knysz M, Hu Xin, Zeng Yuanyuan, et al. Open WiFi networks: lethal weapons for botnets? [C]//Proc of the 31st annual IEEE international conference on computer communications. [s.l.]: IEEE, 2012:2631-2635.

[12] 王 畅,戴 航,孙启禄. 智能手机上僵尸网络综述[J]. 微处理机,2012,33(2):39-44.

[13] 方滨兴,催 翔,王 威. 僵尸网络综述[J]. 计算机研究与发展,2011,48(8):1315-1331.

[14] 李 跃. 面向移动网络的僵尸网络关键技术研究[D]. 成都:西南交通大学,2013.

(上接第 117 页)

Barbara, CA, USA: [s.n.], 2005:36-54.

[3] 葛伟平,汪 卫,周皓峰,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.

[4] 路慧萍,童学锋. 保持隐私的决策树生成过程研究[J]. 计算机应用,2005,25(6):1382-1384.

[5] Jha S, Kruger L, Daniel P M. Privacy preserving clustering [C]//Proceedings of the 10th European symposium on research in computer security. Milan, Italy: [s.n.], 2005:397-417.

[6] Vaidya J, Clifton C. Privacy-preserving k-means clustering over vertically partitioned data[C]//Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining. Washington DC, USA: ACM, 2008:206-215.

[7] 张国荣,印 鉴. 分布式环境下保持隐私的聚类挖掘算法[J]. 计算机工程与应用,2007,43(18):165-167.

[8] 姚 瑶,吉根林. 一种基于隐私保护的分布式聚类算法[J]. 计算机科学,2009,36(3):100-102.

[9] 雷红艳,邹汉斌. 限制隐私泄露的隐私保护聚类算法[J]. 计算机工程与设计,2010,31(7):1444-1446.

[10] Clifton C, Kantarcioglu M, Vaidya J. Defining privacy for data mining[C]//Proceedings of the national science foundation workshop on next generation data mining. Baltimore, MD, USA: [s.n.], 2002:126-133.

[11] Kantarcioglu M, Clifton C. Privacy preserving distributed mining of association rules on horizontally partitioned data[J]. IEEE Transactions on Knowledge and Data Engineering, 2004,16(9):1026-1037.

[12] 宋宝莉,覃 征. 分布式环境下关联规则的安全挖掘算法[J]. 计算机工程,2006,32(21):35-37.

[13] 黄毅群,卢正鼎,胡和平,等. 分布式环境下保持隐私的关联规则挖掘算法[J]. 计算机工程,2006,32(13):12-14.

[14] Samet S, Miri A. Privacy preserving protocols for perception learning algorithm in neural networks[C]//Proceedings of the 4th IEEE international conference on intelligent systems. Varna, Bulgaria: IEEE, 2008:1065-1070.

[15] Yang Zhiqiang, Wright R N. Privacy-preserving computation of Bayesian networks on vertically partitioned data[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(9):1253-1264.

水平分割数据的保护隐私聚类挖掘方法研究

作者： 袁武， 任勋益， YUAN Wu， REN Xun-yi
作者单位： 南京邮电大学 计算机学院,江苏 南京,210000
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(5)