分布式异常检测中隐私保持问题研究

黄毅群,卢正鼎,胡和平,李瑞轩
(华中科技大学计算机学院,湖北武汉 430074)

摘 要: 隐私保持是目前数据挖掘领域的一个重要方向,其目标是研究如何在不共享原始数据的条件下,获取准确的数据关系.本文采用现实的多方安全计算模式,结合数据干扰技术,提出了一种保持隐私的异常检测算法.该算法选择那些超出局部阈值距离的两点间距离及其序号进行通讯,为了保持原始数据的隐私,随机抽取一些正常范围内的两点间距离及其序号,在加入干扰后分散在异常信息中.理论分析表明该算法既提供了现实的数据隐私又保障了算法的性能.

关键词: 隐私保持; 分布式数据挖掘; 异常检测; 多方安全计算
中图分类号: TP309.2 **文献标识码**: A **文章编号**: 0372-2112 (2006) 05-0796-04

Privacy Preserving Outlier Detection

HUANG Yi-qun, LU Zheng-ding, HU He-ping, LI Rui-xuan
(School of Computer Science & Technology, Huazhong University and Technology, Wuhan, Hubei 430074, China)

Abstract: Privacy preserving data mining has emerged to develop accurate models without sharing precise individual data records. Based on a practical secure multi-party computation model, an efficient algorithm for privacy preserving outlier detection with data perturbation techniques is proposed. The ID numbers of pairwise points whose distance exceeds the threshold is necessary to communicate among different sites. Besides, some pairwise points whose distance is within the threshold is chosen to hide private information. The algorithm maintains integrity and good privacy of the data sets of each party while keeping communication and computation cost low.

Key words: privacy preserving; distributed data mining; outlier detection; secure multi-party computation

1 引言

分布式环境中,与传统的集中式数据挖掘不同,保持隐私的数据挖掘需要解决如下矛盾:一方面,各数据持有方都希望保持自己的私有数据不为其他任何一方所知;另一方面,它们又希望通过合作获得全局数据模型.因此,需要研究新的算法使得各方在不共享原始数据的情况下进行正确的数据挖掘.

异常检测是数据挖掘的一个重要方面,它被用来发现数据集中显著不同于其他记录的一些个别记录.在信用卡诈骗、网络入侵检测等金融、电子商务领域都有广泛应用.

对于垂直划分的分布式数据,J Vaideep等提出了一种基于"安全和"方式的异常检测方法[1].由于"安全和"是一种半信赖的多方安全计算模式,在合谋的情况下,不论是异常点还是非异常点,其信息都会完全泄露;而且,该算法的计算和通讯开销是个严峻的问题[8].为此,本文基于现实的多方安全计算模式,结合数据干扰技术,提出来一种保持隐私的异常检测算法.算法中,各站点不需要将数据集中所有两点间的距离进行通讯,而是传递那些超出局部阈值距离的两点间距离及其序号;同时,为了保持隐私,随机选取一些正常范围内的两点间距离及其序号,在加入干扰后分散在异常信息中.由于各方均不能判断所得信息的准确性,因此,保障了现实的数据隐私;同时,干扰技术的采用和通讯量的减少大大提高了算法的性能.

2 一种保持隐私的分布式异常检测算法

2.1 基于距离的分布式异常点检测

定义 1 (Minkowski距离). 设 m 维空间中的点 $X=(x_1,\cdots,x_m)$和 $Y=(y_1,\cdots,y_m)$,两点之间的距离为[2]:

$$Distance(X,Y)=[(x_1-y_1)^k+\cdots+(x_m-y_m)^k]^{1/k} \quad (1)$$

为方便计算,我们采用距离的 k次方来替代距离.

定义 2 (基于距离的异常点). 如果数据集 DB中一

收稿日期: 2005-12-19;修回日期: 2006-03-09
基金项目:国家自然科学基金 (No. 60403027)

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

个点 q满足下列性质:数据集 DB中至少 $p*100\%$的点与 q的距离大于阈值距离 Dt,则称点 q为 $DB(p,Dt)$-$outlier$[3]. 其中,参数 p与 Dt由用户确定.

定义 3 对于垂直划分的分布式数据集,设位于 r个不同站点 $P_1,\cdots,P_r$的数据集分别为 $(DB_1,\cdots,DB_r)$,其中 DB_i $(1\leq i\leq r)$为局部数据库,全局数据库为 $DB=DB_1\cup\cdots\cup DB_r$. 若数据库 DB_i有 N条事务和 m_i个属性,则全局数据库 DB有 N条事务和 $m=\sum_{i=1}^{r-1}m_i$个属性.

由于在垂直划分的数据集中,属性被划分在不同的站点中,因此全局阈值和任意两点 O_i、O_j之间的距离也被分割在各站点中.

定义 4 在站点 $P_1,\cdots,P_r$上,全局数据集 DB中任意两点 O_i、O_j位于各站点的距离分别为 $d_1(o_i,o_j),\cdots,d_r(o_i,o_j)$,各站点的局部阈值距离为 $Dt_1,\cdots,Dt_r$.

那么,对于数据集 DB中的任意一点 O_i,存在如下 $(N-1)\times r$维的距离矩阵:

站点: P_1 $\cdots$ P_s $\cdots$ P_r

$$\begin{pmatrix} d_1(o_i,o_1)-Dt_1 & \cdots & \cdots & \cdots & d_r(o_i,o_1)-Dt_r \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \ddots & d_s(o_i,o_j)-Dt_s & \ddots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ d_1(o_i,o_{i-1})-Dt_1 & \cdots & \cdots & \cdots & d_r(o_i,o_{i+1})-Dt_r \\ d_1(o_i,o_{i+1})-Dt_1 & \cdots & \cdots & \cdots & d_1(o_i,o_{i+1})-Dt_r \\ \cdots & \ddots & \cdots & \ddots & \cdots \\ d_1(o_i,o_N)-Dt_1 & \cdots & \cdots & \cdots & d_r(o_i,o_N)-Dt_r \end{pmatrix}$$

其中第 j行第 s列的元素 $d_s(o_i,o_j)-Dt_s$表示点 O_i与点 O_j $(O_i\neq O_j)$之间位于站点 P_s相对于阈值的局部距离.

O_i、O_j两点之间的全局距离是 $\sum_{s=1}^{r}d_s(o_i,o_j)$,全局阈值是 $\sum_{s=1}^{r}Dt_s$,判断 $\sum_{s=1}^{r}(d_s(o_i,o_j)-Dt_s)\geq 0$可得知这两点的距离是否超出全局阀值 $\sum_{s=1}^{r}Dt_s$. 通过"安全和"方式既能计算每两点间地全局距离,又可以隐藏各站点中的保密数据,如 $d_s(o_i,o_j),Dt_s,(s\in[1,r],i,j\in[1,N],i\neq j)$.

定义 5 (安全和). 设存在三个以上的站点 $P_1,\cdots,P_r$,分别拥有数值 v_s, $s\in[1,r]$, $v_s\in[1,w]$;各方之间没有共谋现象. 为了计算 $v=\sum_{s=1}^{r}v_s$,令某一方为主站点 (设为 P_1), P_1产生随机数 R并将 $R+v_1 \bmod w$传给 P_2, P_2计算 $R+\sum_{s=1}^{2}v_s \bmod w$并将其传给 P_3. 如此类推, P_r最终得到 $v=R+\sum_{s=1}^{r}v_s \bmod w$并将其传给 P_1, P_1计算 $v-R$得到 $\sum_{s=1}^{r}v_s \bmod w$[4].

2.2 一种现实的多方安全计算模式

多方安全计算是一种为了完成某种计算任务而采用的分布式计算协议. 在协议运行前,参与计算的各方 (设有 r方)各自拥有一个保密的输入 $x_1,\cdots,x_r$;协议中,各方保持隐私输入不为它方 (包括任何的第三方)所知;协议运行后,各自获得输出 $f_1(x_1,\cdots,x_r),\cdots,f_r(x_1,\cdots,x_r)$,除此之外,各方不知道其他方输入的任何信息.

可以看出,保持隐私的数据挖掘是多方安全计算的一种特殊应用. 最初 YAO (1986)年提出了一种两方安全计算协议[5],随后 Goldreich等人将其推广为对于任何函数都成立的多方安全计算方法, Goldreich还指出"安全地计算即隐私地计算"[6].

就安全性而言,多方安全计算又分为可信赖的第三方模型、半信赖模型和恶意模型. 前两者都假设各方严格遵守协议的约定,然而,在现实中恶意的破坏行为是有必要考虑的. "安全和"就是一种半信赖模式,如果站点 P_{s-1}和 P_{s+1} $(s\in[1,r])$发生共谋,则站点 P_s中的数据 v_s就会泄漏,协议也就完全不安全.

另外,就算法性能来讲,理想的 SMC技术似乎是"零"信息泄露,但它们的计算和通讯开销都非常高[7],尤其应用在需要处理大量信息的数据挖掘工作时,算法会因为效率低而没有实际意义. 因此,我们提出一种现实的多方安全计算模式,如图 1所示. 它是基于如下实际情况:人们往往愿意接受那些更高效但安全性相对逊色的解决方案[15],泄漏一定的隐私来保障效率是可以接受的.

该模式中,协议应该同时保障以下几个方面:虽然用于挖掘的数据和算法都发生了改变,但新的算法仍然能获得与不采用信息隐藏时一样正确的数据关系;相应的信息隐藏方法能提供用户"可以接受"的数据隐私,并具备"可以接受"的性能,主要包括计算和通讯开销.

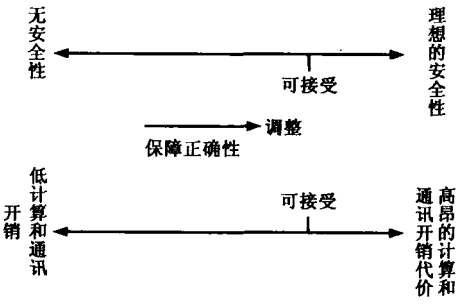


图 1　现实的多方安全计算模式

2.3 结合多方安全计算和干扰技术的算法

算法的基本思想是基于以下考虑: (1)数据集中绝大部分都是正常点,大多数的两点间的距离都会在正常范围以内; (2)通过减少通讯量和加入干扰来改善算法的安全性和效率.

基于定义 4,我们发现,局部异常点不一定是全局异常点,局部正常点也未必是全局正常点,因此,我们不能通过排除局部异常点来提高算法的性能. 然而,在距离矩阵中,若某行的所有元素 (即局部距离)都小于 0时,该行的和

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(即全局距离)也必然小于 0,那么,这两点的距离就可以不参与全局异常点的统计和通讯.这样,我们选择那些超出阈值的局部点间距离进行通讯,并通过"安全和"方式来计算两点间的全局距离;另外,为了避免因合谋而造成的信息完全泄漏,随机的选择一些属于正常范围的两点间距离及相关点的序号,在加入干扰后分散在异常信息之中.

算法描述如下:

输入:r个站点 $P_1, \cdots, P_r$,所有点的集合 O,站点 P_s 的局部阈值距离 Dt_s,任意两点 o_i, o_j 间的距离 $d_s(o_i, o_j)$.

假设 P_1 有权知道最终的结果.

Step1 for $s \leftarrow 1, \cdots, r$ do
Step2 At P_s: for all pairwise $(o_i, o_j) \in |O|, O_i \neq O_j$ do
Step3 $\Delta d_s(o_i, o_j) = d_s(o_i, o_j) - Dt_s$
Step4 if $\Delta d_s(o_i, o_j) > 0$ then
Step5 $o\_O_s.ID = \{(o_i.ID, o_j.ID)\}$
Step6 $o\_O_s.value = \Delta d_s(o_i, o_j)$
Step7 else
Step8 $n\_O_s.ID = \{(o_i.ID, o_j.ID)\}$
Step9 $n\_O_s.value = \Delta d_s(o_i, o_j) - R'_{i,j}$;
//$R_{i,j}$ is random number holding by P_s.
Step10 end if
Step11 Randomly choose $_s\ell$ pairwise ID from n_O_s into o_O_s as $o\_O'_s$.
Step12 end for
Step13 end for
Step14 for $s \leftarrow 2, \cdots, r$ do
Step15 Send $o\_O'_s.ID$ to P_1
Step16 end for
Step17 At P_1: $o\_O'.ID = \{o\_O'_1.ID \cup \cdots \cup o\_O'_r.ID\}$
Step18 for all pairwise points $(o_i, o_j) \in o\_O'$ do
Step19 $\Delta d(o_i, o_j) = R_{i,j} + \Delta d_1(o_i, o_j)$;
//$R_{i,j}$ is a random number holding by P_1.
Step20 $o\_O'.value = \Delta d(o_i, o_j)$
Step21 end for
Step22 Send $o\_O'.ID$ and $o\_O'.value$ to P_2.
Step23 for $s \leftarrow 2, \cdots, r-1$ do
Step24 At P_s: for all pairwise pointsdo $(o_i, o_j) \in o\_O'$ do
Step25 $\Delta d(o_i, o_j) = R_{i,j} + \Delta d_s(o_i, o_j)$
Step26 $o\_O'.value = \Delta d(o_i, o_j)$
Step27 end for
Step28 Send $o\_O'.ID$ and $o\_O'.value$ to P_{s+1}.
Step29 end for
Step30 At P_r: for all the pairwise points $(o_i, o_j) \in o\_O'$ do
Step31 $\Delta d(o_i, o_j) = R_{i,j} + \Delta d_s(o_i, o_j)$
Step32 $o\_O'.value = \Delta d(o_i, o_j)$
Step33 end for
Step34 for all objects $o_i \in o\_O'$ do
Step35 for all objects $o_j \in o\_O'$ do
Step36 At P_1 and P_r: if $\Delta d(o_i, o_j) - R_{i,j} > o$ then
(using secure comparison protocol same as Yao's Millionaire's problem.)
Step37 $m = m + 1$;
Step38 else
Step39 $m = m$
Step40 end if
Step41 end for
Step42 At P_1: if $n > p\% * N$ then
Step43 O_i is an outlier.
Step44 Send the result to any other parties authorized to know.
Step45 end if
Step46 end for

其中,各局部站点 P_s $(s \in [1, r])$先产生集合 o_O_s 和 n_O_s, $o_O_s.value$和 $o_D_s.ID$分别包含了所有超出局部阈值的两点间的距离及相关点的序号,而 $n_O_s.value$和 $n_O_s.ID$则分别包含了正常范围内的两点距离及其序号.为了隐藏 o_O_s 中的信息,各站点从 n_O_s 中随机地选取 $_s\ell$ 个(比例由用户确定,其数值小于或等于实际的真实值)与 o_O_s 合在一起产生新的集合 o_O_s'中,其中 $o\_O_s'.value$ 和 $o_O_s'.ID$分别是两点间的距离及这两点的序号.

随后,各方将集合 o_O_s'中每两点的序号传给有权获取最终结果的一方(设为 P_1)或第三方;P_1 汇集所选信息于集合 o_O 之中,并将 o_O 中每两点的序号(在集合 $o_O'.ID$中)及其数值(在集合 $o_O'.value$中)以"安全和"的方式向其他站点广播并计算全局相对距离,最后判断全局异常点.

3 算法的性能分析

3.1 安全性分析

SMC安全性的基本思想是:在协议运行后,不论有没有合谋,如果参与计算的任意一方所获得的(它方)数据随机地分布于一个统一范围内,那么,该方就不知道它方的任何信息,它只能获悉自己有权知道的协议运行结果.可以看出,多方安全计算的安全性发生在通讯中,根据中间通讯信息得到的真实信息越少,算法就越安全.

本文算法的通讯出现在第 15、22、28, 44步中.我们从这些步骤中得到的信息来分析算法的安全性.在第 15步,各方向 P_1 传递的是 $o_O_s'.ID$中的两两点的序号,由于它们既包含超出阈值距离的成对的点,又包含不超出阈值距离的成对的点,因此不能确定哪些是异常点或非异常点.在第 22和 28步,各方之间通讯 $o_O'.ID$和 $o_O'.value$中的信息,同样,由于加入了干扰信息 $(R'_{i,j})$以及安全和的存在,各方不能判断数据正确与否.在第 44步,只有权知道全局异常点的一方或几方才能获得最终结果.另外,由于各

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

方均知道所有点的数量和序号[8],因此各集合中所通讯的序号本身并不是隐私信息.

虽然不包含在 $o_i.O$ 中的点一定是正常点,然而,数据集中绝大多数的对象都应该是正常点.因此,算法泄漏的是极少数的数据,各方也很难判断这些信息的正确性.并且,通过 2.2节的分析,我们知道为了获得可实现的计算和通讯开销,揭示一定的隐私也是必要的和可以接受的.

文[1]仅仅采用"安全和"的方式来保持原始数据的隐私,在没有合谋的情况,算法很安全;如果出现合谋,某一方或几方的所有数据就会全部泄漏.与之相比,本文算法结合"安全和"与数据干扰技术,通过减少通讯量提高了算法的安全性.

3.2 计算和通讯开销分析

本文算法的计算开销主要依赖于数据集的大小 N,以及集合 $o_i.O$ 中成对的点的数量 n(与 $_s\ell$ 有关).各站点在计算每两点间的距离时,需要 $O(N^2)$ 的计算量;各方在产生集合 $o_i.O_s'$ 中(第 1～13步),每一步都有不超过 $O(N)$ 的计算量;在第 17～33步,各站点需要 $O(n)$ 的计算量来计算 $o_i.O$ 中的全局点间距离;在第 34～41步,还有 $O(n^2)$ 次的安全比较[5].因此,除安全比较外,各方的计算开销达到 $O(N^2)$,这是计算每两点之间的距离所必须的;而安全比较的多少 $O(n^2)$ 则根据用户的需要来调整.对于通讯开销,除安全比较之外,在第 15步中,各方只需向 P_1 通讯 1次;在第 22和 28步中,各方需向下一方通讯 1次;在第 44步中,P_1 只需向其他授权方各通讯 1次.

文[1]需要 $O(N^2)$ 的安全比较,这是不能调整的.而且,仅计算全局距离就需要通讯 $r * N^2$ 次.与之相比,本文算法的通讯量大大降低.

4 结论

为了适应分布式数据挖掘对数据的隐私性的要求,基于隐私保持的分布式数据挖掘正成为新的研究方向.本文采用一种现实的多方安全计算模式,结合数据干扰技术,提出了一种保持隐私的分布式异常检测算法.该算法不需要传递数据集中的所有点间距离,减少了通讯量,提高了算法的性能;由于加入了干扰,即使在合谋的情况下也可以隐藏数据的隐私.

参考文献:

[1] Vaidya J, et al. Privacy-Preserving Outlier Detection[M]. The Fourth IEEE International Conference on Data Mining. Brighton, UK, 2004. 9: 1 - 4.

[2] M. Kantardzic. 数据挖掘 ——概念、模型、方法和算法[D]. 闪四清,等,译. 北京:清华大学出版社, 2003. 104 - 105.

[3] Knorr E, et al. Algorithms for Mining Distance-Based Outliers in Large Datasets[M]. The VLDB Conference, New York, USA, 1998. 9: 392 - 403.

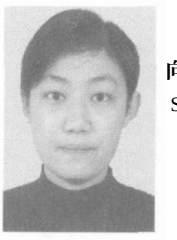
[4] Clifton C, et al. Tools for Privacy Preserving Distributed Data Mining[M]. SIGKDD Explorations, 2003, 4(2): 28 - 34.

[5] Yao A C. How to Generate and Exchange Secrets[M]. The Twenty-seventh IEEE Symposium on Foundations of Computer Science. Toronto, Ontario, Canada, 1986. 10: 162 - 167.

[6] Goldreich O. The Foundations of Cryptography[D]. volume 2, chapter 7: General cryptographic protocols. Cambridge University Press, 2004. 5.

[7] Du W, et al. A Practical Approach to Solve Secure Multi-Party Computation Problems[M]. New Security Paradigms Workshop 2002. Virginia Beach, Virginia, USA, 2002. 9: 127 - 135.

作者简介:



黄毅群　女,讲师,博士研究生,主要研究方向为数据挖掘、信息安全及隐私、应用密码学和SMC技术. E-mail: zfdu2000@163.com

卢正鼎　男,教授,博士生导师,主要研究方向为工程数据库、数据库系统安全、系统集成.

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net