面向隐私保护的数据挖掘技术研究

吕 品[1],陈年生[2],董武世[2]
(1.武汉工程大学 计算机科学与工程学院,湖北 武汉 430073;
2.湖北师范学院 计算机系,湖北 黄石 435002)

摘 要:隐私与安全是数据挖掘中一个越来越重要的问题。隐私与安全问题的解决能破坏图谋不轨的挖掘工程。文中研究了数据挖掘中隐私保护技术的发展现状,总结出了隐私保护技术的分类,详细讨论了隐私保护技术中最重要的隐私保持技术,最后得出了隐私保护技术算法的评估指标。

关键词:隐私保持;探索技术;数据实用;不确定性水平
中图分类号:TP309.2 **文献标识码**:A **文章编号**:1673-629X(2006)07-0147-03

Study of Data Mining Technique in Presence of Privacy Preserving

LÜ Pin[1], CHEN Nian-sheng[2], DONG Wu-shi[2]
(1. School of Computer Science, Wuhan Institute of Technology, Wuhan 430073, China;
2. Department of Computer Science, Hubei Normal University, Huangshi 435002, China)

Abstract: Privacy and security concerns have become an increasingly important issue in data mining. Privacy and security concerns can constrain such access, threatening to derail data mining projects. This paper researches the data mining situations where these privacy and security issues arise, summaries the classification of privacy preserving techniques, discusses the most technique of privacy preservation among the privacy preserving techniques in detail, and addresses the evaluation of privacy preserving algorithms.

Key words: privacy preservation; heuristic-based techniques; data utility; uncertainty level

0 引 言

随着计算机技术和网络信息技术的发展,人们产生和搜集的数据和信息量急剧增加。因而敏感数据的收集、机构的合作以及跨国公司的经营运作给数据挖掘提出了新的挑战,即数据挖掘中的隐私和安全性问题。隐私保护挖掘在数据挖掘中是一个新兴的研究方向。已有的数据挖掘算法对数据所作的分析会产生副作用,即导致隐私暴露。因此,对已存在的数据挖掘算法应当从隐私保护的角度重新考虑。数据挖掘中的隐私保护主要考虑两个方面的问题:首先是敏感的原始数据,其次是从数据库中提取的敏感知识应当删除,因为它同样可能危及他人隐私。隐私保护挖掘的主要目的就是用某种技术改进已有的数据挖掘算法来修改原始数据,使得敏感的数据和知识不被泄露。

收稿日期:2005-10-26
基金项目:湖北省自然科学基金资助项目(2004ADA023)
作者简介:吕 品(1973-),女,湖北鄂州人,硕士,讲师,研究方向为数据挖掘、算法分析与设计、软件工程。

1 隐私保护技术的分类

数据挖掘中采用了很多方法实现数据保密。隐私保护技术主要是基于以下几个方面来分类的:数据的分布、数据的修改、数据挖掘算法、数据或规则的隐藏以及隐私保持[1]。

1.1 数据分布

根据数据的分布情况,隐私保护技术可以分为针对集中式数据的隐私保护技术和分布式的隐私保护技术。分布式的隐私保护技术又分为数据水平分割的隐私保护技术和数据垂直分割的隐私保护技术。数据的水平分割主要原因是多个机构或组织对于不同的个体收集了相似的信息;数据的垂直分割主要原因是多个机构或组织收集了同样的个体的不同信息。

1.2 数据修改

主要将数据库中的某些敏感信息作一定的修改以确保原来数据库中的隐私信息不被泄露。常用的方案主要有如下几种:值替代方法,就是用一个新的值替代原有的值;分组方法,就是用一个问号替代一个已存在的属性值;聚集方法,就是将几个值进行合并为一个粗糙类;交换方法,就是单个记录间值的交换;取样方法,主要指的是用于

挖掘的数据只是总样本中的一个样本。

1.3 数据挖掘算法

在现阶段,数据隐藏技术都是在每一个数据挖掘算法中单独考虑。例如:决策树算法、关联规则算法、粗糙集以及贝叶斯网络等数据挖掘算法中已经有了许多隐私保护的重要思想。V. S. Verykios 等人提出了研究组合数据挖掘算法中的数据隐藏技术,即将两个或更多的数据挖掘算法合并,在这些组合的数据挖掘算法中研究数据隐藏技术[1]。

1.4 数据或规则的隐藏

这种方法主要是针对原始数据或聚集数据的隐藏。以规则的形式隐藏聚集数据的复杂度是相当高的。因此,针对这种形式的数据绝大多数的隐藏技术是探索式的。

1.5 隐私保持

这是一种最重要的隐私保护技术,它的主要特点是有选择性地修改原始数据。这种修改能使隐私在不受危害的情况下获得较高的技术实用性。这种实用性主要体现在使用隐私保护技术以后有用信息的丢失量。隐私保持技术主要有三种:基于探索式的隐私保持技术、基于密码学的隐私保持技术和基于重构的隐私保持技术。

2 隐私保持技术

2.1 基于探索式的隐私保持技术

基于探索式的隐私保持技术所针对的数据对象是集中式的。由于在这些算法中应用基于探索式的隐私保持技术进行有选择性地修改数据是一个 NP 难度问题,因此,应用该技术时常会涉及到复杂度问题。

基于探索式的隐私保持技术中用到的数据修改方法主要有:值替代、数据分组。以基于值替代的关联规则为例。问题的描述是:假设 D 是源数据库,R 是从 D 中挖掘出的重要的关联规则,Rh 是 R 中要隐藏的规则。如何将 D 转换为向外界公开的 D,也能从 D 中挖掘出除了 Rh 以外的所有规则 R?为了达到这个目的,必须有选择性地修改数据,使得敏感规则的支持度降低。M. Atallah 等人提出了在上述关联规则中隐藏大量敏感项集是一个 NP 难度问题,并给出了它的形式化证明[2]。LiWu Chang 和 I. S. Moskowitz 提出了一种在关联规则中用数据分组的方法修改数据,这种技术在医学应用中特别有用[3],Y. Saygin 等人应用了这种技术,它的方法是强行改变支持度和置信度的定义,并将最小的支持度和置信度变化到一个各自对应的区间上[4]。只要敏感规则的支持度和置信度低于这两个区间的中点,就认为敏感信息不会被泄露。

2.2 基于密码学的隐私保持技术

基于密码学的隐私保持技术所针对的数据对象是分布式的。它涉及到分布式数据的垂直分割与水平分割。数据挖掘算法中研究了很多有关密码技术解决实际的隐私问题。例如:安全两方或多方计算问题。Wenliang Du 等人提出了一个系统转换结构允许将一个计算转换为安全的多方计算[5]。B. Pinkas 提出了将密码学理论的研究应用于数据挖掘中的隐私保护,并且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[6]。C. Clifton 等人提出了支持隐私保持的四种安全多方计算的方法。它们分别是:安全和,安全并集,安全交集大小以及标量积方法[7]。

M. Kantarcioglu 等人利用安全和的方法建立了一个水平数据分割的朴素贝叶斯分类模型实现隐私保护[8]。安全和的方法是假定一个已知值 $v = \sum_{i=1}^{k} v_i$ 且位于区间 $[0,\cdots,n]$ 上,其中 v_i 是第 i 个站点的值。在 k 个站点中选择其中一个为主站点并标记为 1 号站点,其余站点标记为 $2,\cdots,k$。站点 1 从均匀分布的区间 $[0,\cdots,n]$ 上产生一个随机数 R。站点 1 将值 $R + v_1 \bmod n$ 传送给站点 2。由于 R 均匀分布于区间 $[0,\cdots,n]$,所以值也均匀分布于区间 $[0,\cdots,n]$。于是站点 2 对于站点 1 的局部值 v_1 一无所知。其余站点 $i = 2,\cdots,k$ 所执行的算法是:站点 i 接收 $v = R + \sum_{j=1}^{i-1} v_j \bmod n$,由于这个值均匀分布于区间 $[0,\cdots,n]$,所以站点 i 对其余站点的信息也一无所知。站点 i 计算出 $R + \sum_{j=1}^{i} v_i \bmod n = (v_i + V) \bmod n$ 并且将这个计算出的值传送给站点 $i+1$。站点 k 执行以上步骤并且将计算出的结果传给站点 1。站点 1 由于知道 R 和 $\sum_{i=2}^{k} v_i$,所以减去 R 和 $\sum_{i=2}^{k} v_i$ 的值就可以得到实际的结果。如果忽略整个计算过程,只考虑最终的结果,可以得出站 1 在整个计算过程中没有得到任何其它站点的信息。这种方法也面临着一个问题,如果两个或多个站点互通信息,那么这些站点可以得到其它站点的信息。

2.3 基于重构技术的隐私保持技术

重构技术主要分为数值型数据的重构技术以及二进制数据与分类数据的重构技术。对于数值型数据的重构技术:R. Agrawal 和 R. Srikant 提出了用离散化的方法与值变形的方法修改原始数据,然后用重构算法构造原始数据的分布[9]。D. Agrawal 和 C. C. Aggarwal 利用期望最大化算法 EM(Expectation Maximization)提出了一个改进的基于贝叶斯的重构方法[10]。对于二进制数据与分类数据的重构技术:S. J. Rizvi 等人和 A. Evfimievski 等人在他们的文章中都研究了关联规则中的二进制数据和分类数据的处理方法,都利用了随机化技术进行数据的修改,既保证了数据的使用率又达到了隐私保护的目的[11,12]。

3 隐私保持技术的评估

关于隐私保持技术的评估,最重要的工作就是要建立合适的评价标准和相关的参考标准。笔者在阅读有关文献后得出,在实际的应用中,不可能用一个标准衡量所用的隐私保持技术,或者说,没有一个隐私保护技术所采用

的算法在一个标准上比所有其它的隐私保持技术算法性能更好,而是一个算法可能在某一个特定的应用中在这个标准上比其它算法好。因而,应该向用户提供一套度量准则,让用户能根据自己的需要选择最合适的隐私保持技术。通常,隐私保持技术的评价指标有:性能、数据实用、不确定性水平和耐久性。

评价性能的方法是估计该算法的时间复杂度或算法中基本操作的平均次数。数据实用指的是在应用中使用了隐私保护技术后信息的丢失量。尽管不同的隐私保持策略可以对信息进行隐蔽,但由于不确定性的存在,这些隐蔽的信息仍然能被推理出来。所以,从操作的观点,信息量的修改应达到最大。隐私保护算法的最终目的是反对信息的未授权者获取该信息。这些侵犯者往往会利用各种各样的数据挖掘算法危害隐私。因此,一个针对具体的挖掘技术而研制的隐私保护算法是不可能适用于所有其它的挖掘算法的。所以,耐久性指的是某一隐私保护算法应能运用到不同的数据挖掘技术。

4　结束语

通过对数据挖掘中的隐私与安全问题的研究,提出了隐私保护技术的分类。通过对基于探索式的隐私保持技术、基于密码学的隐私保持技术、基于重构的隐私保持技术的详细论述,表明了研究者们对敏感数据和规则的保护这一领域的重视。当前,数据挖掘中的隐私与安全问题只能在某些特定的数据挖掘算法中有一定效率,而不能将其推广到一般。通用的隐私保护技术必然是未来的研究趋势。

参考文献:

[1] Verykios V S, Bertino E, Fovino I N, et al. State - of - the - art in Privacy Preserving data mining[J]. ACM SIGMOD Record, 2004, 33:50 - 57.

[2] Atallah M, Elmagarmid A, Ibrahim M, et al. Disclosure Limitation of Sensitive Rules[A]. Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange. IEEE Conference Proceeding[C]. Chicago, Illinois: [s. n.], 1999. 45 - 52.

[3] Chang LiWu, Moskowitz I S. Parsimonious downgrading and decision trees applied to the inference problem[A]. Proceedings of the 1998 workshop on New security paradigms[C]. Charlottesville, Virginia, United States: ACM Press, 1998. 82 - 89.

[4] Saygin Y, Verykios V S, Elmagarmid A K. Privacy Preserving Association Rule Mining[A]. Proceedings of the 12th International Workshop on Research Issues in Data Engineering (RIDE'2002)[C]. San Jose, USA: IEEE Computer Society Press, 2002. 151 - 158.

[5] Du Wenliang, Attalah M J. Secure multiproblem computation problems and their applications: A review and open problems [R]. CERIAS Tech Report 2001 - 51, Center for Education and Research in Information Assurance and Security and Department of Computer Sciences, Purdue University, West Lafayette, IN, 2001.

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

[7] Cliffton C, Kantarcioglu M, Vaidya J, et al. Tools for privacy preserving distributed data mining[J]. ACM SIGKDD Explorations Newsletter, 2002, 4(2):28 - 34.

[8] Kantarcioglu M, Clifton C. Privacy - preserving distributed mining of association rules on horizontally partitioned data [A]. The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02). ACM SIGMOD'2002[C]. Madison, Wisconsin: [s. n.], 2002. 24 - 31.

[9] Agrawal R, Srikant R. Privacy - preserving data mining[A]. Proceedings of the 2000 ACM SIGMOD international conference on Management of data[C]. Dallas, Texas, United States: ACM, 2000. 439 - 450.

[10] Agrawal D, Aggarwal C C. On the design and quantification of privacy preserving data mining algorithms[A]. Proceedings of the twentieth ACM SIGMOD - SIGACT - SIGART symposium on Principles of database systems[C]. Santa Barbara, California, United States: ACM Press, 2001. 247 - 255.

[11] Rizvi S J, Haritsa J R. Maintaining data privacy in association rule mining[A]. In Proceedings of the 28th International Conference on Very Large Databases(VLD)[C]. Hong Kong, China: [s. n.], 2002. 682 - 693.

[12] Evfimievski A, Srikant R, Agrawal R, et al. Privacy preserving mining of association rules[A]. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining[C]. Edmonton, Alberta, Canada: ACM Press, 2002, 217 - 228.

(上接第 146 页)

大学学报, 2005, 45(增刊):214 - 218.

[3] Rosenberg J, Schulzrinne H, Camarill G, et al. SIP: Session Initiation Protocol[S]. RFC3261, 2002.

[4] Krawczyk H, Bellare M, Canetti R. HMAC: Keyed - Hashing for Message Authentication[S]. RFC2104, 1997.

[5] Gennaro R, Rohatgi P. How to Sign Digital Streams[A]. Advances in Cryptology - CRYPTO '97, 17th Annual International Cryptology Conference[C]. Santa Barbara, California, USA: Springer, 1997. 180 - 197.