隐私保护的关联规则挖掘研究

耿 波,仲 红,徐 杰,闫娜娜
(安徽大学 计算机信息学院,安徽 合肥 230039)

摘 要:数据挖掘在各个方面都极大地方便了人们的生产、生活,并在很大程度上提高了工作的效率。然而,人们发现了它的最致命的弊端,那就是对隐私信息的暴露。如何保护私有信息或敏感数据在挖掘过程中不被泄露,同时又能得到较为准确的挖掘结果,已经成为数据挖掘领域中的一个很有意义的研究课题。文中针对关联规则模式挖掘中的隐私保护问题,提出了新的算法,不仅成功地隐藏了指定隐私规则集,同时使得对原始数据的改动达到最小,从而有效地降低了意外产生新规则比率和意外隐藏规则比率。

关键词:关联规则;隐私保护;敏感规则
中图分类号:TP309 **文献标识码**:A **文章编号**:1673-629X(2009)02-0081-02

Study on Privacy-Preserving Association Rule Mining

GENG Bo, ZHONG Hong, XU Jie, YAN Na-na
(School of Computer and Information, Anhui University, Hefei 230039, China)

Abstract: Data mining facilitates people in all fields, and improves the efficiency of our work. Nevertheless, people have gradually seen its most lethal defect: the exposure of private information. How to protect private or sensitive information in the data mining process now become a meaningful study topics of data mining. For privacy-preserving in association rule mining, a new algorithm has been proposed in this paper. It reduces modifying ratio, and consequently reduces probability of producing new rule and the probability of hidden accident.

Key words: association rule; privacy-preserving; sensitive rules

0 引 言

数据挖掘的任务,就是要从海量的数据中,发现其中的有用的信息。这在各个方面都极大地方便了人们的生产、生活,并且在很大程度上提高了工作的效率。

尽管如此,人们也逐渐发现了数据挖掘的最致命的弊端,那就是在利用数据挖掘技术和工具给人们提供知识和信息的同时,也严重暴露了隐私信息。

2000年,Agrawal. R等人在文献[1]中首先提出了在数据挖掘过程中考虑隐私信息的保护是未来数据挖掘领域中的一个非常重要的研究重点。如何保护私有数据或敏感信息在数据挖掘过程中不被泄露,同时又能够得到较为准确的挖掘结果,已经成为数据挖掘研究中的一个很有意义的研究课题。随后,很多文献都研究了对关联规则挖掘中的隐私保护问题[2~6]。

收稿日期:2008-05-14
基金项目:国家自然科学基金项目(60773114);安徽省自然科学基金项目(070412051);安徽高校省级重点自然科学研究项目(KJ2007A43)
作者简介:耿 波(1982-),男,硕士研究生,研究方向为数据挖掘;仲 红,硕士生导师,研究方向为数据挖掘。

文中针对关联规则挖掘中敏感规则的隐藏问题提出了新的算法,不仅成功地隐藏了指定隐私规则"$X \Rightarrow Y$",同时使得对原始数据的改动尽可能最小,从而有效控制了意外产生新规则比率和意外隐藏规则比率。

1 预备知识

1.1 关联规则

关联规则是Agrawal. R等人1993年首次提出的重要的数据挖掘研究课题[7]。定义一个项集 $I=(i_1, i_2, \cdots, i_m)$。定义事务集 D,其中每个事务 t 都是项量集 I 中的项,即:$t \subseteq I$。一个关联规则定义为 $x \Rightarrow y$,其中 $x \subseteq I, y \subseteq I$, and $x \cap y = \varnothing$。x 称为规则的前量,y 称为后量。置信度定义为:$|x \cup y| / |x|$,其中 $|x|$ 表示事务集中包含 x 的事务的数量。$|x \cup y|$ 表示事务集中既包含 x 又包含 y 的事务的数量。规则的支持度计算为 $|x \cup y| / n$。其中 n 是规则集 D 中所有事务的总数。

1.2 关联规则模式中的隐私数据隐藏问题

数据挖掘的目标是从数据库中挖掘出隐藏的潜在

的未被发现的规则或模型。然而,隐私保护的目的是隐藏某些敏感信息,使得它们不能被通过数据挖掘技术发现。在关联规则挖掘的过程中,假设给定了一个被认为是隐私的关联规则集合,在挖掘的过程中,希望通过对数据库的少量的改动使这个集合中的每条关联规则在挖掘过程中不被挖掘出来,这就是"关联规则模式中的隐私数据隐藏"问题的阐述。

2 提出的问题及算法设计

2.1 提出的问题及设计思想

在隐藏的过程中,不仅希望达到隐藏的效果,而且希望对数据库的改动能够最小。并且希望在修改数据库时,尽量不要产生新的规则或意外隐藏了本不该隐藏的规则,这样在算法中就需要注意对所要修改的事务的选择。

例如,对于表 1,假设最小支持度设为 33%,最小置信度设为 70%。可以发现以下两条关联规则:B→A(37.5%,75%),B→C(37.5%75%),括号内是支持度和置信度。假设规则 B→A 是所希望隐藏的规则,可以通过修改数据库来达到降低 B→A 的置信度或支持度,以达到隐藏的目的,例如可以将事务 T3:ABC 修改为 BC,这样就使得 T3 不再支持规则 B→A,使得规则 B→A 的置信度降到 50%,从而达到隐藏的效果。但是,我们发现,事务 T3:ABC 还包含了 A 与 C 的关联,如果将其改为 BC,将会影响这种关联,而造成意外的隐藏,也就是说,在改动数据库的同时,可能会隐藏了本不希望隐藏的规则。然而,如果修改事务 T8:AB 为 B,同样也能达到隐藏规则 B→A 的目的,并且由于 T8 只包含 A 和 B 的关联,所以不会影响的其他的关联规则的置信度。

表 1 事务表

Tid	Items
T1	AC
T2	ABC
T3	ABC
T4	A
T5	AC
T6	BC
T7	AC
T8	AB

另外还有一种情况,如果对同一个事务的修改,可以使该事务失去对两个或多个隐私规则的支持,那么优先选择修改这样的事务,因为,这样可以在对数据库做很小的改动的情况下隐藏更多的隐私规则。总之在笔者设计的算法中不仅考虑了达到隐藏的目的,而且尽可能地将对数据库的改动和意外隐藏率和产生新规则的比率都降到最低。

2.2 算法简单描述:

输入:原始数据库 D,包含 n 条敏感规则的集合 $M\{G_1,G_2,G_3,\cdots,G_n\}$,最小支持度阈值 s,最小置信度阈值 c。

输出:经过转换后的数据库 D',使得 M 中的每条规则被隐藏起来。

首先,从原始数据库 D 中抽取客户事务集合 CRS $=\{T_1,T_2,T_3,\cdots,T_m\}$,使其中每一事务 T_i 都至少支持敏感规则集 M 中的一条敏感规则,CRS 即为候选处理事务集合。

For($i=1\cdots m$)

{

在 M 中找到事务 T_i 所支持的所有的敏感规则集;

计算 T_i 所支持的敏感规则的个数 N_i;

计算 T_i 所包含的项数 M_i;

}

根据 N_i 的大小对 CRS 做降序排列;

对于 N_i 相同的按照 M_i 做升序排列;

然后,对 CRS 中的候选处理事务逐个处理(而对每一个事务 T_i 的处理见过程 2.2),直到发现一条敏感规则的置信度小于置信度阈值 c;

将这条敏感规则从敏感规则集合 M 中删除,得到包含 $n-1$ 条敏感规则的集合 M_2,以及新的数据库 D_2;

再将这个新的数据库 D_2,敏感规则集合 M_2 代入此算法做相同的处理得到数据库 D_3,以及包含 $n-2$ 条敏感序列的敏感规则集合 M_3;

如此递归下去,直到敏感规则集合中的规则被删除完,这时得到的新数据库即为输出。

2.3 事务 T 的处理过程

事务 T 支持敏感关联规则 $G_1,G_2,\cdots,G_L$。我们对事务 T 处理如下:

比较 $G_1,G_2,\cdots,G_L$ 的置信度,找到置信度最低的 G_i,然后对 T 进行修改,使其不再支持 G_i。

3 实 验

为了说明算法的效果,设计了一系列的实验做测试。测试的方面有:隐藏失败率,产生新规则比率和丢失原有规则比率等各方面的效果。

实验进行的环境如下,CPU:Athlon1.5G,内存 256G,操作系统:Windows XP。用 IBM 随机数据产生器产生数据集。数据集的大小为从 10k 到 20k 个事务。每个事务平均有 5 个项,整个数据库包含 100 个

(下转第 86 页)

释过程、执行计划、绑定变量的使用以及会话中发生的等待事件等。通过分析跟踪文件的信息，可以找到有问题的 SQL。

(3)TKPROF。

TKPROF 是一个对 SQL_TRACE 所产生的跟踪文件信息进行分析并产生一个更加清晰、合理的输出结果的工具。如果一个系统执行效率比较低，一个比较好的方法是跟踪用户的会话并使用 TKPROF 工具格式化输出，从而找出问题所在。

(4)DBMS_PROFILER。

DBMS_PROFILE PACKAGE 主要用于优化服务器端 PL/SQL 程序，包括存储过程、函数、触发器等，用于对各种 PL/SQL 程序进行测试，找出程序中性能不佳的地方，然后进行改进。此外，该工具还可以用于数据库产品的测试。

(5)ADDM。

ADDM (AUTOMATIC DATABASE DIAGNOSTIC MONITOR)是 Oracle 10g 新增的性能优化工具，是构建到 Oracle 10g 数据库内核中的自诊断引擎。它包括统计信息的自动收集、数据库诊断监控以及 SQL 语句的自动优化等。通过对数据库状态的定时检查，ADDM 能够自动确定数据库性能瓶颈，给出解决性能瓶颈的措施建议。同时，报告系统潜在的性能问题。

4 结束语

Oracle10g 数据库系统性能优化与调整是一个复杂、繁琐的系统工程，贯穿于数据库系统开发的整个过程。数据库系统配置的调整，包括内存结构调整、磁盘 I/O 调整以及磁盘碎片调整等，直接决定了整个数据库系统的性能，应该利用各种性能优化与调整工具进行反复的调整、比较以消除瓶颈，获得系统最优性能。

参考文献：

[1] 腾永昌. Oracle9i 数据库管理员使用大全[M]. 北京：清华大学出版社，2005：766－778.

[2] Ault M. Oracle 数据库管理与维护技术手册[M]. 北京：清华大学出版社，2003：595－663.

[3] Lawson C. Oracle 性能优化科学与艺术[M]. 北京：清华大学出版社，2006.

[4] 盖国强，冯春培. Oracle 数据库性能优化[M]. 北京：人民邮电出版社，2006.

[5] 柳 丹. Oracle PL/SQL 面向对象特性 Web 应用研究[J]. 计算机技术与发展，2006，16(1)：234－237.

[6] 苏淑文，翁敬农. Oracle 系统异构环境下的信息集成[J]. 计算机技术与发展，2007，17(3)：128－131.

[7] 杜志源，刘 刚，王永智. 高校教务管理系统数据库性能优化的研究[J]. 计算机工程与设计，2007，28(20)：5066－5068.

(上接第 82 页)

项。对于每个数据集，最小支持度和最大置信度都随机产生，分别在 3%～10%，和 30%～50%。关联规则总数为 10 到 240 不等。隐藏的规则在 5 到 20 个。试验结果显示：DSR 的隐藏失败率(0%)，产生新规则(0%)，意外隐藏规则(7%)。

从实验结果可以看出所设计的算法在隐藏隐私规则的同时也很好地达到了控制意外隐藏率和产生新规则比率的效果。

4 结束语

针对数据挖掘中的关联规则挖掘中隐藏敏感规则的问题提出了新的算法，不仅成功隐藏了指定隐私规则“$X \Rightarrow Y$”，同时使得对原始数据的改动尽可能得达到最小，从而有效地控制了意外产生新规则比率和意外隐藏规则比率。

参考文献：

[1] Agrawal R, Ramakrishnan S. Privacy－preserving data mining [C]//In Proceedings of the 2000 ACMSIGMOD International Conference on Management of Data. Dallas, USA: [s. n.], 2000：439－450.

[2] Stanley R M Oliveira, Osmar R Za ane. Privacy preserving frequent itemset mining[C]//In Proceedings of IEEE ICDM Workshop on Privacy, Security and Data Mining. Maebashi City, Japan: [s. n.], 2002：43－54.

[3] Evmievski A, Srikant R. Privacy preserving mining of association rules[C]//In Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada: [s. n.], 2002：217－228.

[4] Vaidya J, Clifton C. Privacy preserving association rule mining in vertcally partitioned data[C]//In Proceedings of the 8th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada: [s. n.], 2002：639－644.

[5] 罗永龙，黄刘生，荆巍巍，等. 一个保护私有信息的布尔关联规则挖掘算法[J]. 电子学报 2005，33(5)：900－903.

[6] 仲 波，张远平. 一个保护私有信息的序数型关联规则挖掘算法[J]. 科学技术与工程，2006，16(24)：3863－3866.

[7] Agrawal R. Mining association rules between sets of items in large databases [C]//In Proc. of the ACM SIGMOD Intl Conf. on Management of Data. Washington, D. C.: [s. n.], 1993：207－216.