关联规则挖掘中的隐私保护研究

张 瑞[1],郑 诚[2],陈娟娟[1]
(1.安徽大学 计算机科学与技术学院,安徽 合肥 230039;
2.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:数据挖掘中的关联规则反映一个事件和其他事件之间依赖或相互关联的知识。随着大量数据不停地收集和存储积累,人们希望从中发现感兴趣的数据关联关系,从而帮助他们进行决策。随着信息技术的发展,数据挖掘在一些深层次的应用中发挥了积极的作用。但与此同时,也带来隐私保护方面的问题。隐私保护是当前数据挖掘领域中一个十分重要的研究问题,其目标是要在不精确访问真实原始数据的条件下,得到准确的模型和分析结果。为了提高对隐私数据的保护程度和挖掘结果的准确性,提出一种有效的隐私保护关联规则挖掘方法。针对关联规则挖掘中需预先给出最小支持度和最小置信度这一条件,提出了一种简单的事务数据库中事务的处理方法,即隐藏那些包含敏感项目的关联规则的方法,以对相关事务作处理,达到隐藏包含敏感项目的关联规则的目的。理论分析和实验结果均表明,基于事务处理的隐私保护关联规则挖掘方法具有很好的隐私性、简单性和适用性。

关键词:隐私保护;联规则;敏感项目
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1673-629X(2008)10-0013-03

Research on Privacy Preserving in Association Rules Mining

ZHANG Rui[1], ZHENG Cheng[2], CHEN Juan-juan[1]
(1.School of Computer Science and Technology of Anhui University, Hefei 230039, China;
2.Ministry of Edu. Key Lab. of Intelligent Computing & Signal Processing, Anhui Univ., Hefei 230039, China)

Abstract: Association rules mining in data mining reflects relations between events. With the large amounts of data collection and storage continuously accumulated, people want to find data associations which they are interesting in, and to assist them in decision-making. With the developments of information technology, data mining plays an active role in applications. But at the same time, it has brought some problems of privacy preserving. Privacy preserving is currently a very important issue in the field of data mining. The object is to get veracious model and analyze the results with imprecise data acess. In order to raise the level of protection of data privacy and the accuracy of mining results, propose an effective privacy preserving method. The minimum support and confidence should be given in associations mining, against this, a simple transactions handling method has been given. Can hide the associations which contain sensitive items by the way of dealing with transactions. Theoretical analysis and experimental results show that, this method based on tansaction processing has got good privacy, simplicity and applicability.

Key words: privacy preserving; association rules; sensitive item

0 引 言

数据挖掘中的关联规则反映一个事件和其他事件之间依赖或相互关联的知识。如果两项或多项属性之间存在关联,那么其中一项的属性值就可以依据其他属性值进行预测。关联规则挖掘发现大量数据中项集之间有趣的关联或相关联系[1]。随着大量数据不停地收集和存储积累,人们希望从他们的交易数据中发现感兴趣的数据关联关系,从而帮助商家进行商务决策的制定,如分类设计、交叉购物等。

最为著名的关联规则提取方法是R.Agrawal提出的Apriori算法,该算法是一种挖掘布尔关联规则频繁集的算法。关联规则的提取可分为两步:第一步,迭代识别所有的频繁项目集,要求频繁项目集的支持度不低于用户给定的最低值;第二步,从频繁项目集中构造置信度不低于用户审定的最低值的规则。识别或发现所有频繁项目集是关联规则提取算法的核心,也是计

收稿日期:2008-01-27
基金项目:国家自然科学基金(60475017);安徽省高等学校自然科学研究项目(2006kj055B)
作者简介:张 瑞(1982-),男,安徽滁州人,硕士研究生,主要从事数据挖掘方向研究;郑 诚,副教授,硕士生导师,主要从事数据挖掘、机器学习研究。

算量最大的部分。

关联规则挖掘的一个典型例子是购物篮分析，该过程通过发现顾客放入其购物篮中不同商品之间的关系，分析顾客的购买习惯，了解哪些商品频繁地被顾客同时购买，这种关联的发现可以帮助零售商制定营销策略[2]。例如，顾客在同一次去超级市场购物时，购买牛奶的同时也购买面包的可能性有多大？通过帮助零售商有选择地经销和安排货架，这种信息也可以引导销售。例如，将牛奶和面包尽可能放近一些，可以进一步刺激一次顾客去商店时同时购买这些商品。

随着信息技术的发展，数据挖掘在一些深层次的应用中发挥了积极的作用。但与此同时，也带来隐私保护方面的问题。例如，通过一般的方法对银行卡客户的交易行为等信息的关联分析，可以发现用户在交易行为上的特点，但不可避免地会造成用户的隐私泄漏。所以在数据挖掘过程中解决好隐私保护的问题，成为数据挖掘的一个研究热点[3]。

数据挖掘的目标是从数据库中提取隐藏的或者是潜在的有用规则或者模式，然而，数据挖掘中隐私保护的目标是把特定的敏感信息隐藏起来，而不被数据挖掘技术发现。对于给定需要隐藏的项目集，提出了相应的修改事务数据库中事务的算法，在较小的修改开销下，解决了关联规则提取中的隐私保护问题，同时保证处理后的关联规则在随后的关联规则挖掘中不被发现。

1 基本概念

关联规则：

令 $I=\{i_1,i_2,\cdots,i_m\}$ 是事务数据库的项集，$T=\{t_1,t_2,\cdots,t_N\}$ 是库中所有事务的集合。每个事务 t_i 包含的项集都是 I 的子集。在关联分析中，包含 0 个或多个项的集合被称为项集(itemset)。如果一个项集包含 k 个项，则称为 k - 项集。事务的宽度定义为事务中出现项的个数[4]。

关联规则(Association Rule)是行如 $x\rightarrow y$ 的蕴涵表达式，其中 x 和 y 是不相交的项集，即 $x\cap y=\varnothing$。关联规则的强度可用支持度(support)和置信度(confidence)度量。支持度描述给定项集的频繁程度，即项的重要性。置信度确定 y 在包含 x 的事务中出现的频繁程度，即项间关系。计算公式如下：

规则的支持度 $\text{Support}(x\rightarrow y)=|x\cup y|/N$

规则的置信度 $\text{Confidence}(x\rightarrow y)=|x\cup y|/|x|$

其中，$x\subseteq I$，$y\subseteq I$，N 为事务数据库包含事务的个数，$|x\cup y|$ 为同时包含 x 和 y 的事务个数，$|x|$ 为 D 中包含 x 的事务个数。关联规则从本质上说是条件概率：A 发生时，B 同时也出现的概率有多大。挖掘关联规则是找出那些支持度和置信度大于预先定义的 minSupp、minConf 的规则[3]。

2 问题描述

给定事务数据库 D、最小支持度 minSupp、最小置信度 minConf，敏感项目集 S，若使用传统的关联规则挖掘方法对事务数据库处理，有规则 $x\rightarrow y$ 被发现($y\in S$)，即项目 y 是需要被隐藏的敏感项目，则挖掘出的规则是不满足隐私保护要求的，这样的挖掘会造成隐私信息的泄漏。因此，提出如下保护隐私的关联规则提取方法。

由上面定义部分中规则的置信度计算公式 $\text{Confidence}(x\rightarrow y)=|x\cup y|/|x|$ 知，在保持 $|x\cup y|$ 的值不变的前提下，若适当地增加 $|x|$ 的值，则可以使得规则 $x\rightarrow y$ 的置信度降低到小于给定的最小置信度 minConf。具体内容是：对于需要隐藏的规则，如 $x\rightarrow y(y\in S)$，若把项 x 加入到不包含 x 或 y 的事务 t 中，则 $|x\cup y|$ 值不变，而 $|x|=|x|+1$，重复这个动作，直到 $\text{Confidence}(x\rightarrow y)=|x\cup y|/|x|<\text{minConf}$。这样，就可以达到隐藏该规则的目的。

3 具体算法

由前面的问题描述，给出隐私保持的关联规则挖掘算法如下：

输入：源数据库 D，给定 minSupp，minConf，敏感项集 S。

输出：事务库 D' //对事务库中包含敏感项目的事务做了处理后的库。

1. For i:=1 to N do //N 为事务库的事务数

2. if $|T_i|=1$ then $D'=D-T_i$//消除原始库中宽度为 1 的事务

3. end for

4. New database D_1//建立一个与事务库 D 结构相同的库 D_1

5. For i:=1 to N' do //N'为事务库 D'包含的事务数

6. if $S\cap T_i=\varnothing$ then put T_i into D_1//D_1 中存放不包含项 x 或 y 的事务

7. end for

8. For i:=1 to |S| do //|S|为敏感项集包含的项目数

9. begin

10. Extract association rules like $x\rightarrow y_i$

11. For each rule
12. If rule.Conf≥minConf then
13. begin
14. NumNeeded= mod(N'(Supp(xy)/minConf - Supp(x))) + 1
15. K=N' - count(x∩y_i)
16. end
17. end if
18. if NumNeeded > K then return(can not hide rule)
19. else
20. begin
21. select NumNeeded transactions from D1//随机选取 NumNeeded 个无关事务
22. for i:=1 to NumNeeded do
23. begin
24. insert x into T_i
25. update the corresponding T_i in D'
26. end
27. end for
28. end
29. Clean database D_1//清除中间库中存放的事务
30. end for
31. end
32. Output the updated database D as D'

算法开始时,给定 minSupp,minConf。对数据库所包含的事务,顺序地进行扫描,去除那些宽度为 1 的事务,即仅包含一个项的事务,因为这些单项事务对于关联规则的提取影响不大。得到对原始事务库处理后的中间库 D',接着建立数据库 D_1,用来存放 D' 中不包含敏感项目的事务。然后对敏感项目集中的每个项目,依次作如下处理:提取该项目的可能的关联规则,对每一个规则 $x \rightarrow y_i$,如果该规则的置信度 rule.Conf 满足 rule.Conf < minConf,则该规则在给定的 minConf 下,不会被发现,因此不需要作隐藏处理。如果 rule.Conf ≥ minConf,那么,该规则在给定的置信度 minConf 下可能被发现,因此需要作相应的处理。为了使该规则的置信度小于给定的阈值 minConf,至少需要 $mod(N'(Supp(xy)/minConf - Supp(x))) + 1$ 个不包含 x 或 y_i 的事务,用于达到隐藏规则 $x \rightarrow y_i$ 的目的,如果不包含 x 或 y_i 的事务数 $K <$ NumNeeded,则不能实现隐藏规则 $x \rightarrow y_i$ 的目标,反之,则能够通过在 $mod(N'(Supp(xy)/minConf - Supp(x))) + 1$ 个事务插入项 x,使得 rule.Conf < minConf,从而,达到隐藏关联规则 $x \rightarrow y_i$ 的目的。

4　分析证明

算法中 NumNeeded = $mod(N'(Supp(xy)/minConf - Supp(x))) + 1$ 部分证明如下:

对于需要隐藏的关联规则 $x \rightarrow y$,$Confidence(x \rightarrow y) = |x \cup y| / |x|$,假定需要添加 m 个 x 项到事务库 D' 中不包含 x 或 y 的事务中,才能使 $Confidence(x \rightarrow y) = |x \cup y| / |x| <$ minConf,此不等式等价于 $|x \cup y| - |x| minConf <$ minConf * m,即 $m > |x \cup y| /$minConf $- |x|$,对此式右边提取事务库 D' 包含的事务数 $|D'|$,得到 $m > |D'| (Supp(x \cup y)/$minConf $- Supp(x))$,因此,满足此不等式的 m 最小值为 $mod(N'(Supp(xy)/minConf - Supp(x))) + 1$ = NumNeeded,证毕。

举例说明文中所提出的算法,有事务数据库如 Table1 所示,项集 $I = \{A,B,C,D\}$,如果某一事务的项在事务中出现,则它的值为 1,否则为 0。如事务 $T_1 = \{A,B,C,D\}$,则事务 T_1 等价于事务 $T_1 = \{1,1,1,1\}$,这样在对事务处理时,显得更为直观。

对表 1 中 7 个事务,给定 minSupp = 50%,minConf = 70%,使用经典的关联规则提取算法 Apriori 得到一些规则,以规则 $A \rightarrow B$ 为例,此时 Supp1 = 57%,minConf1 = 80%,用文中给出的算法处理,先清理 T_7,对于剩下的 6 个事务,把 A 添加到事务 T_5 中去,即表 1 中 T_5 行,A 列的 0 被替换为 1,这样,Conf2 = 67%,从而实现了规则 $A \rightarrow B$ 的隐藏。

表 1　事务库样例

事务标签	A	B	C	D
T_1	1	1	1	0
T_2	1	1	1	1
T_3	1	1	0	0
T_4	1	1	1	0
T_5	0	0	1	1
T_6	1	0	1	1
T_7	0	0	0	1

5　实验分析

实验机器 CPU 为 Celeron1.6GHz,512M 内存,使用超市数据 basket 作为实验样本,实验数据量为 1000,项目总数 $|I| = 11$,平均事务长度 ATL=2.8,对于由该数据集抽取的事务组成的事务库,给定不同的最小支持度和最小置信度,生成一系列的关联规则集合。如表 2 所示。

(下转第 19 页)

至少要 40ms,理想下每帧检测时间不得高于 10ms。普通 OpenCV 函数库下 scale=1.0 的检测时间基本都在 10ms 以上,无法满足系统要求。采用文中所述方法,当检测窗口 scale=1.3 时基本满足系统实时检测要求,在检测窗口 scale > 2.0 时其检测时间基本保持在 0.06(ms)至 1.2(ms)之间。

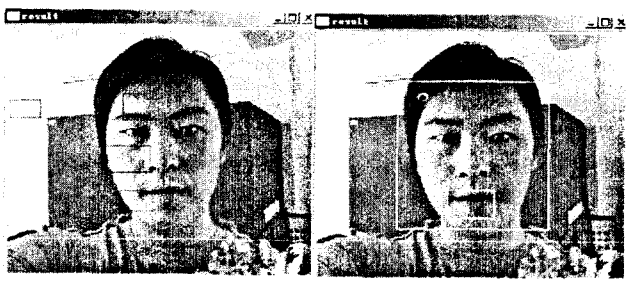

图 5　使用文中方法前后的嘴部检测效果图

4　结束语

文中针对 AVSR 系统下可视化前端设计中人脸嘴部快速检测的要求,采用了基于 Harr-like 特征推进级联分类的方法,有效降低了其目标检测的计算时间并提高了准确性。此方法首先引入了积分图像的概念,通过查找表的方法对大量垂直、旋转 Harr-like 特征值进行了快速的计算。在基于 AdaBoost 学习算法上通过推进选择单值分类的基础分类器,以级联的方式合并强分类器从而丢弃了大量检测中无用的子窗口,最后通过人脸嘴部的定位进一步降低了其检测的目标区域。实验表明,此方法在 AVSR 系统上的检测效果基本上可以达到实时应用的目地,为进一步的视-音融合及识别提供了有效的视觉特征。

参考文献:

[1] Gong. Speech recognition in noisy environments: a survey[J]. Speech Communication, 1995, 16: 261-291.

[2] Potamianos G, Luettin J. Audio-visual speech recognition [R]. Final Workshop 2000 Report, Center for Language and Speech Processing. Baltimore, MD: The Johns Hopkins University, 2000.

[3] Liang H, Liu X X, Zhao Y B, et al. Speaker independent audio-visual continuous speech recognition[C]//In Proc. of IEEE ICME. Lausanne, Switzerland: [s. n.], 2002.

[4] Viola P, Jones M J. Rapid Object Detection using a Boosted Cascade of Simple Features[J]. IEEE CVPR, 2001(1): 511-518.

[5] Papageorgiou C, Oren M, Poggio T. A general framework for Object Detection[C]//In International Conference on Computer Vision. [s. l.]: [s. n.], 1998.

[6] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[C]//In Computational Learning Theory: Eurocolt '95. [s. l.]: Springer-Verlag, 1995: 23-37.

[7] Amit Y, Geman D, Wilder K. Joint induction of shape features and tree classifiers[J]. IEEE Transactions on Pattern Ananlysis and Machine Intelligence, 1997, 19(11): 1300-1305.

[8] Cristinacce D, Cootes T. Facial feature detection using AdaBoost with shape constraints[C]//British Machine Vision Conference. [s. l.]: [s. n.], 2003.

(上接第 15 页)

表 2　支持度、置信度与规则数对应表

minSupp	minConf	Rules
10%	15%	169
10%	20%	126
10%	25%	89
10%	30%	37
10%	35%	16

包含项 wine 的事务有 287 条,若隐藏项 wine,由于最小置信度和最小支持度的不同组合,更改事务的数量占总事务数的 20%～45%,当修改更多的事务时,相应的时间开销也将增加。文中,研究了由数据挖掘技术引起的事务数据库隐私保护问题,提出了隐藏那些包含敏感项目的关联规则的算法。

6　结束语

目前处理含有隐私项目的关联规则的方法大都是基于支持度或置信度的减少的。文中提出的算法能够隐藏包含敏感项目的关联规则,而且不需要预知关联规则的种类,并在随后给予了证明。对于给定的事务数据库、敏感项目集,在将来,还需要改善算法的效率,如:减少扫描事务数据库的次数。此外,如果数据库频繁更新,如何在这种情况下,充分保持已作隐藏处理的那些关联规则的不可见性,是有待研究的问题。

参考文献:

[1] 陈子阳,马朝虹,李宇佳,等.量化关联规则的隐私保持挖掘方法[J].计算机工程,2005,31(11):74-76.

[2] 罗永龙,黄刘生.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.

[3] Evfimievski A, Srikant R, Agrawal R. Privacy preserving mining of association rules[J]. Information Systems, 2004, 29: 343-364.

[4] Seifert J W. Data mining and the search for security[J]. Government Information Quarterly, 2004, 21: 461-480.