一个保护私有信息的布尔关联规则挖掘算法

罗永龙,黄刘生,荆巍巍,姚亦飞,陈国良
(中国科学技术大学计算机科学技术系,安徽合肥 230027;国家高性能计算中心(合肥),安徽合肥 230027)

摘 要: 本文基于随机响应技术,提出了一种在保护隐私的关联规则挖掘中对数据进行伪装的方法;设计了在伪装的数据集上进行挖掘的算法;分析了算法的效率.实验结果表明,该算法在伪装的数据集上挖掘出的规则与原始规则相比,相对误差不超过 2 %,并给出了使得相对误差最小时相关参数的取值.

关键词: 数据挖掘; 关联规则; 随机响应
中图分类号: TP309 **文献标识码**: A **文章编号**: 0372-2112 (2005) 05-0900-04

An Algorithm for Privacy-preserving Boolean Association Rule Mining

LUO Yong-long, HUANG Liu-sheng, JING Wei-wei, YAO Yi-fei, CHEN Guo-liang
(*Department of Computer Science, University of Science and Technology of China, Hefei, Anhui* 230027, *China*; *National High Performance Computing Center at Hefei, Hefei, Anhui* 230027, *China*)

Abstract: In distributed systems, some traditional association rules mining algorithms have been developed with all original data being gathered into a centralized site. However, these algorithms are not fit for the situation where no user is willing to disclose his information. In the privacy preserving association rule mining problems, there are several participants engaged in the computation and the algorithms are run on the union of their databases. Currently, the secure union algorithm can be used to protect each user's privacy if all the user's databases have the same structure. However, in secure union algorithm, each participant should encrypt all the participants' data. So, if there are many participants engaged in the cooperative computation, this method is inefficient. Thus, in this paper, we introduce a data disguised method for privacy preserving association rule mining based on the randomized response techniques, present the mining algorithm on the disguised item set and analyze the complexity of this algorithm. The experiments show that the rule that this algorithm gets has fewer relative error which is less than 2 % compared with the original rules. We also give some values of the parameters which make the relative error is the lowest.

Key words: data mining; association rule; randomized response

1 引言

关联规则是由 Agrawal 等人 1993 年在文献[1]中首先提出的一个重要的数据挖掘(Data Mining)研究课题.目前已出现了许多有效的关联规则挖掘算法,这些算法大都集中于如何提高算法的效率.尽管关联规则挖掘的主要任务是如何在数据集上产生规则从而获得有效的知识,但在另一方面样本数据的准确性直接影响到关联规则的可信程度.

大多数传统的关联规则挖掘是由一个用户在本地的一个单一的数据库上进行操作.随着计算机网络的不断发展,产生规则的数据往往来自于网络中不同的用户,分布式关联规则挖掘也逐步得到研究.现有的分布式关联规则挖掘需要有一个算法执行中心来收集所有的原始数据,然后执行相应的挖掘算法[2].若用户对数据的隐私不太关心,目前的挖掘算法都能够很好地工作,但由于数据包含着用户大量的隐私信息,有时候用户不愿意提供相应数据或者只提供虚假数据,从而影响了产生的规则的有效性.例如我们希望作一项调查来推断某些生活习惯与某种疾病之间的关联程度,大多数用户都不愿意透露这类数据.

进行数据挖掘的同时保护用户数据的隐私是未来数据挖掘的一个极其重要而富有挑战性的课题[3],近年来已经有不少学者在这方面做了很多有益的工作[2~10].2000 年,文献[3,5]同时提出了两种不同的保护私有信息的数据挖掘(PPDM)问题,并分别采用数据扰动(data perturbation)技术和安全多方计算(secure multiparty computation)协议加以解决.随后推动了许多相关的研究:文献[6,7]讨论了在 *ID*3 算法中对数据的隐藏;文献[8]研究了对关联规则进行隐藏的问题;文献[9,10]讨论了保护私有信息的关联规则挖掘的一般方法.在保护隐私的关联规则挖掘中,按照用户的相互合作方式可以将共享的数据库分为垂直划分(vertically partitioned)[2]与水平划分(horizontally partitioned)[4]两类,文献[2]提出一种新的点积协议来求解垂直划分问题.文献[4]基于安全求并算法求解水平

收稿日期:2004-03-01;修回日期:2004-12-10
基金项目:国家 973 项目(No. 2003CB317000);安徽省教育厅重点科研项目(No. 2003kj049zd);安徽省教学研究项目(No. 2005166)

划分问题,若参与计算的用户数为 k,各用户记录数之和为 n,则安全求并算法需要 $O(kn)$ 次加密运算[4]. 因此,当 k 较大时,用安全求并算法解决水平划分问题效率很低. 特别地,当数据来自于网上调查时,若每个用户仅有一条记录,则 $k=n$,此时该方法性能极差. 为此,本文基于随机响应技术(Randomized Response)[6,7,11],介绍了一种布尔关联规则挖掘中隐私数据的伪装方法,提出了在伪装的数据集上进行规则挖掘的算法,通过对算法进行实验分析,验证了算法的有效性,同时研究了算法中若干参数的取值问题.

2　基本概念

2.1　关联规则挖掘

设 $I=\{i_1,i_2,\cdots,i_n\}$ 是项的集合. 设任务相关的数据 D 是数据库事务的集合,其中每个事务 T 是项的集合使得 $T\subseteq I$,每个事务有一个标识符,称作 TID. 设 A 是一个项集,事务 T 包含 A 当且仅当 $A\subseteq T$. 关联规则是形如 $A\Rightarrow B$ 的蕴涵式,其中 $A\subset I$, $B\subset I$,并且 $A\cap B=\varnothing$. 规则 $A\Rightarrow B$ 在事务集 D 中成立,具有支持度 s,其中 s 是 D 中包含 $A\cup B$ 的百分比,即它是概率 $P(A\cup B)$. 规则 $A\Rightarrow B$ 在事务集 D 中具有置信度 c,它是 D 中包含 A 的事务同时也包含 B 的百分比,这是条件概率 $P(B|A)$. 即 $\mathrm{Support}(A\Rightarrow B)=P(A\cup B)$, $\mathrm{Confidence}(A\Rightarrow B)=P(B|A)$. 挖掘关联规则就是产生那些支持度和置信度分别大于用户给定的最小支持度和最小置信度阈值的规则,阈值可以由用户或领域专家根据经验设定. 如果考虑的关联仅是项的在与不在,则是布尔关联规则挖掘[12].

2.2　随机响应技术

随机响应技术是在统计学中为了保护被调查者的隐私而设计的一种数据隐藏技术,于 1965 年首先由 Warner 在文献[11]中提出. 基本方法是:为了了解一群人中具有属性 A 的百分比,需要对这些人进行调查,由于属性 A 涉及到个人隐私,被调查者可能不愿意回答或做出错误的回答. 在相关问题模型(Related-Question Model)[6]中,首先对属性 A 设计两个互为否定的问题,例如:问题 1:具有属性 A?问题 2:不具有属性 A?对每个问题的回答均为是(yes)或否(no). 调查者首先确定一个实数 $\theta\in[0..1]$,被调查者通过随机函数产生一个 0 到 1 之间的随机实数 r,若 $r\leq\theta$,则被调查者回答问题 1,否则回答问题 2,即被调查者将以概率 θ 回答问题 1,以概率 $1-\theta$ 回答问题 2. 尽管调查者知道被调查者回答的是 yes 或 no,但前者并不知道后者到底回答的是哪个问题,从而保护了被调查者的隐私. 假设分别用 $P^*(A=\mathrm{yes})$ 和 $P^*(A=\mathrm{no})$ 来表示被调查者回答为 yes 和 no 的百分比. 则可以通过下述方程组求出被调查群体中具有属性 A 和不具有属性 A 的百分比的近似值 $P(A=\mathrm{yes})$ 和 $P(A=\mathrm{no})$:

$$\begin{cases} P^*(A=\mathrm{yes})=P(A=\mathrm{yes})\cdot\theta+P(A=\mathrm{no})\cdot(1-\theta) \\ P^*(A=\mathrm{no})=P(A=\mathrm{yes})\cdot(1-\theta)+P(A=\mathrm{no})\cdot\theta \end{cases} \tag{1}$$

显然,当 $\theta\neq0.5$ 并且被调查人数很多时,这两个近似值的误差就足够小.

3　保护隐私数据的关联规则挖掘

为描述方便,我们不妨先假设事务集中的每个事务都是 2-项集$\{A,B\}$,这里 A, B 均为布尔属性,然后将其扩展到多项集. 为挖掘关联规则 $A\Rightarrow B$,我们首先用随机响应技术对数据进行伪装,然后在伪装后的数据集上执行挖掘算法.

3.1　数据的伪装

假设有多个用户参与计算,每个用户有多条记录. 我们可以采用类似于第 2 节介绍的随机响应技术来伪装这些记录. 记(A,B)变换后为(A',B'),分别为 A、B 设置一个变换概率 θ_1、θ_2,即:用户对每条记录伪装时,先产生两个随机实数 $r_1,r_2\in[0,1]$,若 $r_1\leq\theta_1$,则 $A'_i=A_i$,否则 $A'_i=1-A_i$;若 $r_2\leq\theta_2$,则 $B'_i=B_i$,否则 $B'_i=1-B_i$. 由于变换后的属性值是随机的,算法执行中心无法知道每条记录的真实值. 假设有 3 个用户 $U1$, $U2$, $U3$ 共同参与计算, $U1$ 有记录 $T1$、$T2$、$T3$, $U2$ 有记录 $T4$、$T5$, $U3$ 有记录 $T6$、$T7$,如表 1. 不妨选取变换概率分别为 $\theta_1=0.3$, $\theta_2=0.6$,表 1 描述了每个用户的记录伪装过程.

表 1　数据的伪装示例

原始数据				随机实数		伪装数据		
TID		A	B	r_1	r_2	TID'	A'	B'
$U1$	$T1$	0	1	0.4	0.7	$T1$	1	0
	$T2$	1	0	0.7	0.3	$T2$	0	0
	$T3$	1	1	0.2	0.1	$T3$	1	1
$U2$	$T4$	0	1	0.1	0.8	$T4$	0	0
	$T5$	1	0	0.5	0.9	$T5$	0	1
$U3$	$T6$	0	1	0.4	0.2	$T6$	1	1
	$T7$	1	0	0.6	0.7	$T7$	0	1

3.2　数据的统计方法

下面我们用 P 来表示各类属性组合值的近似概率,并将 $P(A=1,B=1)$ 简记为 $P(11)$,它表示同时具有属性 A 和属性 B 的近似概率,其他组合值的近似概率依此类推.

我们用 P^* 表示伪装后的数据集中各类组合值的百分比,例如在表 1 中,用 $P^*(01)$ 表示伪装数据集中 $A'=0$ 且 $B'=1$ 的记录所占百分比, $P^*(01)=2/7$. 这一组值可以直接从伪装后的数据集中统计到,于是我们可以得到下面的方程组:

$$\begin{cases} P^*(11)=P(11)\theta_1\theta_2+P(10)\theta_1(1-\theta_2) \\ \qquad +P(01)(1-\theta_1)\theta_2+P(00)(1-\theta_1)(1-\theta_2) \\ P^*(10)=P(11)\theta_1(1-\theta_2)+P(10)\theta_1\theta_2 \\ \qquad +P(01)(1-\theta_1)(1-\theta_2)+P(00)(1-\theta_1)\theta_2 \\ P^*(01)=P(11)(1-\theta_1)\theta_2+P(10)(1-\theta_1)(1-\theta_2) \\ \qquad +P(01)\theta_1\theta_2+P(00)\theta_1(1-\theta_2) \\ P^*(00)=P(11)(1-\theta_1)(1-\theta_2)+P(10)(1-\theta_1)\theta_2 \\ \qquad +P(01)\theta_1(1-\theta_2)+P(00)\theta_1\theta_2 \end{cases} \tag{2}$$

若合理选择 θ_1 与 θ_2 的值,我们就可以从这个方程组中求解出 $P(11)$, $P(10)$, $P(01)$ 及 $P(00)$. 我们将在第 4 节讨论 θ_1 与 θ_2 的取值问题.

3.3　规则的产生

当数据集很大时,通过公式(2)求出的 $P(11)$, $P(10)$, $P(01)$, $P(00)$ 的值基本接近于从原始的数据集中统计到的结果,于是:

$$\mathrm{Support}(A)=P(A)=P(11)+P(10),$$

$$\text{Support}(A \Rightarrow B) = P(AB) = P(11)$$

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support}(A \Rightarrow B)}{\text{Support}(A)} = \frac{P(11)}{P(11) + P(10)}$$

当数据集中的每个事务为 m-项集 $(A_1, A_2, \cdots, A_m)$ 时,假设第 i 个属性的变换概率为 θ_i,我们记 $P(I_1, I_2, \cdots, I_m)$ 为各种组合情况的近似概率, $P^*(I_1^*, I_2^*, \cdots, I_m^*)$ 为从伪装后的数据集中统计到的各种组合值的百分比. 于是我们可以得到有 2^m 个方程的方程组:

即: $P^*(I_1^*, I_2^*, \cdots, I_m^*) = \sum P(I_1, I_2, \cdots, I_m)\theta_{k_1}\theta_{k_2}\cdots\theta_{k_m}$ (3)

其中 $\theta_{k_i} = \begin{cases} \theta_i, & \text{当 } I_i^* \oplus I_i = 0 \\ 1-\theta_i, & \text{当 } I_i^* \oplus I_i = 1 \end{cases}$ 这里 I_i^*, I_i 为 0 或 1, $i = 1, 2, \cdots, m$, $\oplus$为异或.

利用高斯消元法可以求出上述方程组的解,得到原始数据中各种组合的近似百分比,然后在此基础上利用已有的关联规则挖掘算法[13]求出所有的频繁项集并产生关联规则.

4 算法分析

基于上节的讨论,我们可以描述相应的算法如下:

Algorithm 1 {

// Mining rules on disguised dataset

// Input: k 个用户参与计算,第 i 个用户有 n_i 条记录且 $\sum_{i=1}^{k} n_i = n$,每条记录

// 均为 m-项集.

// Output: k 个用户共同在他们数据的并集上挖掘规则.

各用户用随机响应技术伪装自己的数据,并将其发送到中心节点; // Step1

for each $I_i^* \in \{0,1\}$, $i = 1, 2, \cdots, m$ do // Step2

中心节点在伪装的数据集上统计 $P^*(I_1^*, I_2^*, \cdots, I_m^*)$;

中心节点利用高斯消元法求解方程组(3),得到各组 $P(I_1, I_2, \cdots, I_m)$ 的值,这里

$I_i \in \{0,1\}$, $i = 1, 2, \cdots, m$; // Step3

结合具体算法产生规则,并将规则发送给各用户; // Step4

}// end

4.1 算法的安全性与复杂性

虽然本文中规则的挖掘仍集中在某个计算中心执行,但不同于传统计算方式,计算中心使用的是伪装后的记录,该节点不能推测到某个用户的单条记录信息. 当参与用户提供的记录数较少时,算法执行中心也不能从该用户伪装的数据集中挖掘出该用户的局部规则. 也就是说,利用随机响应技术,每个用户的单条记录信息及局部规则都可以得到保护.

对于计算复杂性,我们仅考虑为保护用户隐私所增加的代价[2,4]. 本文利用随机响应技术在对数据作变换时,仅需要对每条记录作一次线性变换,共需要 $O(n)$ 次线性变换. 此外算法需要求解一个线性方程组,对有 t 个变量的方程组,高斯消元法的时间复杂性为 $O(t^3)$,若考虑到规则挖掘中,记录数目 n 是变量,而属性数目是常量,则高斯消元法的时间 $O(t^3)$ 可以看作是一个较大的常数. 因此,虽然当项集中属性较多时,为保护用户隐私所增加的计算成本很大,但与文献[4]需要 $O(kn)$ 次加密运算相比,伪装的代价仍得到了很大改进,其原因是加密成本很高.

4.2 实验分析

为评价算法的准确性,我们首先产生一个数据集,在进行伪装前不考虑数据的隐私保护,按传统方法在原始数据集上挖掘出关联规则;然后采用第 3 节介绍的随机响应技术对数据集进行伪装,将在伪装后的数据集上产生的规则与原规则进行对比;再组合不同的参数值对规则的准确率进行评估. 本文中,我们仅对置信度进行分析,而支持度的分析与此类似. 假设在原数据集上产生的置信度为 C,由挖掘算法在伪装的数据集上所产生规则的置信度为 C^*,我们定义规则的相对误差为: $(|C^* - C|)/C$.

4.2.1 实验方法

Step1 随机产生有 8000 个事务的样本数据集 D,设第 i 个事务为 $T_i = (A_i, B_i)$,并求出原始的支持度与置信度

$\text{Support}(A)$, $\text{Support}(A \Rightarrow B)$, $\text{Confidence}(A \Rightarrow B)$

Step2 数据伪装:对每个样本数据 T_i 进行变换,映射到 $T_i' = (A_i', B_i')$,

产生一个随机实数 $r_1 \in [0,1]$,若 $r_1 \leq \theta_1$,则 $A_i' = A_i$,否则 $A_i' = 1 - A_i$;

产生一个随机实数 $r_2 \in [0,1]$,若 $r_2 \leq \theta_2$,则 $B_i' = B_i$,否则 $B_i' = 1 - B_i$.

Step3 从伪装后的数据集中计算 $P^*(11)$, $P^*(10)$, $P^*(01)$, $P^*(00)$.

Step4 据公式(2)求出 $P(11)$, $P(10)$, $P(01)$, $P(00)$,并求出在伪装后的数据集上产生的支持度 $\text{Support}(A \Rightarrow B)$ 及置信度 $\text{Confidence}(A \Rightarrow B)$,计算相对于 Step1 中规则相对误差.

Step5 分别取 θ_1, $\theta_2 = 0.05, 0.1, 0.15, 0.2, \cdots, 0.45, 0.55, \cdots, 0.9, 0.95$,重复 Step2～Step4,对比分析产生规则的相对误差,确定参数的合理取值.

Step6 改变样本数据集的规模,分别设定事务数为 1000, 2000, …, 14000,分析算法准确率及随数据量变化趋势.

Step7 产生不同的样本数据集,重复 Step1～Step6,分析不同的支持度与置信度对参数及样本数据量的依赖关系.

4.2.2 实验结论

(1) 图 1 描述了挖掘算法所产生规则的相对误差与参数 θ_1、θ_2 的关系,图中误差的大小用气泡的大小表示. 从图中我们可以看出当 $\theta_1, \theta_2 \in (0, 0.25) \cup (0.75, 1)$ 时,规则的误差很小. 若我们取 $\theta_1 = \theta_2 = \theta$,则误差随参数 θ 的变化趋势可由图

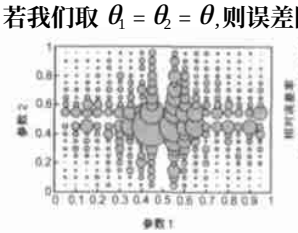

图 1 误差与参数的关系

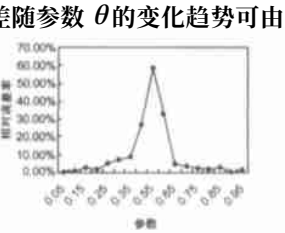

图 2 单参数平均相对误差

2 描述,当 $\theta \in (0, 0.25) \cup (0.75, 1)$ 时相对误差能够达到 3 % 以内;(2) 从图 3 可以看出,随着数据量的增加,规则的相对误差不断减小,当事务数达到 10000 条以上时,算法误差率基本可以降低到 2 % 以内;(3) 从图 3 可以看出,对于不同的支持度与置信度,误差率也不一样,图中 0.43/0.67 表示原始数据集中 Support$(A \Rightarrow B)$ = 0.43, Confidence$(A \Rightarrow B)$ = 0.67,其他类似. 当原始的支持度与置信度都比较高时,误差率相对比较低;而对支持度与置信度都比较低的规则,由于误差相对较大,而本身的值比较小,因此是不可靠的,我们可以通过增加数据量或改变概率参数提高规则的准确性. 因为关联规则挖掘是需要找到支持度与置信度都大于某个阈值的规则,因此较小的误差对算法的实用性并没有影响;(4) 图 4 表明当数据量比较小时,若选择 $\theta_1 = \theta_2$,规则的误差表现不稳定,而选用不同的参数,从图 3 可以看出误差能降低到一个合理的范围. $|\theta - 0.5|$ 越大时,信息的隐藏率越低,当 $\theta = 0$ 或 $\theta = 1$ 时,信息得不到任何保护[6]. 为能够使信息得到有效的保护,应使得 $|\theta - 0.5|$ 尽可能小,但 θ 离 0.5 越近,产生的误差也越大. 如何在两者之间进行平衡,我们将进一步进行研究.

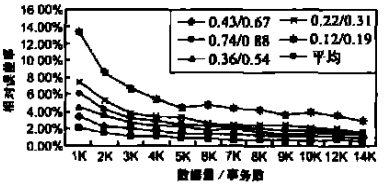

图 3　相对误差与数据量及支持度/置信度的关系(多参数)

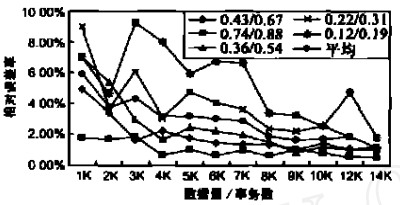

图 4　相对误差与数据量及支持度/置信度的关系(单参数)

5　结束语

本文中我们研究了在保护私有信息的条件下多个用户协作进行布尔关联规则挖掘的问题,算法首先使用随机响应技术对数据进行伪装,然后在伪装的数据集上进行挖掘,从而达到保护数据隐私的目的. 大量的实验结果表明当 $\theta_1, \theta_2 \in (0, 0.25) \cup (0.75, 1)$ 时,在大型数据集上挖掘出的规则与原始的规则相比,相对误差不超过 2 %,并且误差率随着数据量的增加而不断降低. 后面我们将进一步讨论规则的准确率与参数选择之间的关系,研究将随机响应技术与一些具体的挖掘算法相结合产生新的算法问题,并把随机响应技术应用到量化关联规则的挖掘上.

参考文献:

[1] Rakesh Agrawal, Tomasz Imieliński, et al. Mining association rules between sets of items in large databases[A]. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data[C]. Washion D C, USA, 1993. 207 - 216.

[2] J Vaidya, C Clifton. Privacy preserving association rule mining in vertically partitioned data[A]. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Edmonton, Canada, 2002. 639 - 644.

[3] R Agrawal, S Ramakrishnan. Privacy-preserving data mining[A]. In Proceedings of of the 2000 ACM SIGMOD International Conference on Management of Data[C]. Dallas, USA, 2000. 439 - 450.

[4] Murat Kantarcioglu, Chris Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data[A]. Transactions on Knowledge and Data Engineering[C]. IEEE Computer Society Press, Los Alamitos, CA, to appear. http://www.cs.purdue.edu/homes/clifton/document/Kantarcioglu.pdf.

[5] Y Lindell, B Pinkas. Privacy preserving data mining[A]. In Advances in Cryptology-CRYPTO '00, volume 1880 of Lecture Notes in Computer Science[C]. Springer-Verlag, 2000. 36 - 54.

[6] Wenliang Du, Zhijun Zhan. Using randomized response techniques for privacy preserving data mining[A]. In Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Washington D C, USA, 2003. 505 - 510.

[7] Zhijun Zhan, Wenliang Du. Privacy-Preserving Data Mining Using Multi-Group Randomized Response Techniques[R]. Technical Report, June 2003. http://www.cis.syr.edu/～wedu/Research/paper/multi_group.pdf.

[8] Y Saygin, V S Verykios, et al. Privacy preserving association rule mining[A]. In Proceedings of the 12th International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems[C]. 2002. 151 - 158.

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

[10] Stanley RM Oliveira, Osmar R Zaïane. Privacy preserving frequent itemset mining[A]. In Proceedings of IEEE ICDM Workshop on Privacy, Security and Data Mining[C]. Maebashi City, Japan, 2002. 43 - 54.

[11] S L Warner. Randomized response: A survey technique for eliminating evasive answer bias[J]. The American Statistical Association, 1965, 60(309): 63 - 69.

[12] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers, 2000. 225 - 278.

[13] Huang Liusheng, CHEN Huangping, et al. A fast algorithm for mining association rules[J]. Journal of Computer Science and Technology. 2000, 15(6): 619 - 624.

作者简介:

罗永龙　男,1972 年 4 月生,副教授,博士生,主要研究方向为信息安全、分布式算法. E-mail: ylluo@ustc.edu.