

一种用于隐私保护关联规则挖掘的 平均信息分布聚类混合算法

刘 伟, 崔永锋, 周丁丁

(周口师范学院 计算机科学与技术学院, 河南 周口 466001)

摘 要: 针对现有的隐私保护关联规则挖掘算法无法满足效率与精度之间较好折中的问题, 提出了一种平均信息分布聚类混合算法 AIDCH(The average information distributed clustering hybrid algorithm). 算法建立了关联规则向量, 在其中用到了信息论方面的内容. 计算信息源各个特征的次数积累关联, 提取一种潜在的明显特征, 以邻域潜在的特征作为聚类对象进行聚类, 引入数据挖掘关联本体概念, 在非单调性约束的条件下进行挖掘, 克服由隐私保护带来的关联空间数据弱化的弊端. 实验表明, 该算法在保护隐私的情况下, 能够获得精度和效率之间较好的折中, 具有一定的实用价值.

关键词: 隐私保护; 关联规则挖掘; 关联本体; 潜在特征提取; 聚类

中图分类号: TM769

文献标识码: A

文章编号: 1000-7180(2014)02-0168-05

Hybrid Algorithm of Average Information Distributed Clustering for Privacy Protection Association Rules Mining

LIU Wei, CUI Yong-feng, ZHOU Ding-ding

(Department of Computer Science and Technology, Zhoukou Normal University, Zhoukou 466001, China)

Abstract: To solve the problem that the existing privacy preserving association rule mining algorithm cannot meet better trade-off between efficiency and accuracy, the paper proposes average information distributed clustering hybrid algorithm. The algorithm creates a vector of association rules, which uses content of information theory. Accumulation of calculation information source the number of times and extraction obvious features of a potential, the potential characteristics of the neighborhood as clustering object clustering, and the introduction of data mining association ontology concept, digging under the conditions of the non-monotonicity constraint, to overcome the weakening drawbacks associated space data by the Privacy. The experiments show that the algorithm can obtain a good tradeoff between accuracy and efficiency in the case of the protection of privacy.

Key words: privacy protection; association rules mining; correlation ontology; potential feature extraction; cluster

1 引言

信息时代的到来, 用户的隐私信息保护受到越来越多的关注, 基于此, 文献[1]中提出了分布式隐私保护挖掘技术. 它的目标是以用户隐私信息为限定条件, 在此限定条件下挖掘出符合要求的信息. 这种技术的方法主要有随机干扰和安全多方计算.

1993 年, Agrawal 等人提出了关联规则, 它的作用是找出事物的时常发生的属性的所有子集和这些子集的关联性. 关联规则挖掘主要针对大型事务数据库, 目的是发现数据间的关系, 最近几年在隐私保护方面取得了一定的成就^[2-4].

数据挖掘可以获得先前未知模式和信息, 比如我们每天的上网, 用百度搜索数据, 我们输入的关键

收稿日期: 2013-04-23; **修回日期:** 2013-06-12

基金项目: 河南省重点科技攻关项目(102102210265); 河南省基础与前沿研究项目(132300410400); 河南省信息技术教育研究规划项目(ITE12064)

词,仅对我们的搜索可能有帮助,但百度公司可以从成千上万的用户搜索输入关键词中,挖掘出哪些关键词出现的频率高等有用信息^[5-6].但不能无限制地挖掘,因为无限制的挖掘会侵犯用户隐私数据.而隐私保护的限制,又降低了数据挖掘的效率和准确度.为此,文献[7]中引入了大量的方法.在分布式环境下,安全多方计算策略被采用.文献[8]中把安全多方计算运用到隐私保护挖掘中,并在此基础上,研究了通信量和协议优化.在安全多方计算中,为了得到准确的挖掘结果,常采用密码协议.但用密码协议,就涉及到加密,解密运算.当数据量很大时,会带来巨大的通信开销与计算量,从而使效率下降^[9].在数据集中环境中,常采用随机干扰法.近来,文献[10]中提出了分布式环境下的干扰方法,这种方法主要采用分布重构与数据干扰,实现了关联规则下的隐私保护挖掘.缺点是属性间的相关性由于每个属性独立进行干扰而遭到破坏.这种方法给用户的效率可以满足其需求,但数据挖掘的准确度不高^[11-12].

以上方法数据挖掘的效率和精确度的折衷度不能取得令人满意的结果,大多顾此失彼.有的算法牺牲用户隐私信息的前提下取得了效率和准确度都高的结果.本文提的平均信息分布聚类混合算法,利用隐私保护下的潜在的关联特征,很好解决了由于约束存在,挖掘效率不高的问题.

2 AIDCH 算法

2.1 数据关联规则向量模型

AIDCH(The average information distributed clustering hybrid algorithm)算法将数据看成一个信息源,从信息论观点来分析数据空间向量,将领域中潜在的特征作为聚类对象,用层次平均信息量提取数据特征.为了便于模型化,在不影响数据类别的情况下,对所有数据进行归一,并设其长度为 N , d 为数据矢量,用来表示潜在特征数据出现的频率.公式如下:

$$d = (f_1, f_2, \dots, f_i, \dots, f_M), \sum_i f_i = 1 \quad (1)$$

为了获取目标函数,要对数据进行归类,在目标函数中,用 c 表示一个完整的数据集合.根据得到的数据集合中对应的划分粗细特征,判断集合中数据的共同属性,为了准确描述集合中数据的异同,需要设定一个可对其进行优化的目标函数.在优化函数设定之前,可把需要操作的数据当成一个完整的信息源,为了描述数据之间的特征淡化关联,这种信源看成是一个无记忆离散集合,设第 i 个特征出现的频率是 $f_i^{(c)} = \frac{1}{|c|} \sum_{d \in c} f_i^{(d)}$, c 的数据数包含了 $|c|$. 可以将 c 的特征用公式表示如下:

$$c = c = (f_1^{(c)}, f_2^{(c)}, \dots, f_i^{(c)}, \dots, f_M^{(c)}), \sum_i f_i^{(c)} = 1 \quad (2)$$

采用特征的频数表示为

$$c = (t_1^{(c)}, t_2^{(c)}, \dots, t_i^{(c)}, t_i^{(c)}, M), \sum_i t_i^{(c)} = N |c| \quad (3)$$

2.2 潜在特征提取

数据关联规则向量模型建立完毕之后,需要对数据的潜在特征进行提取.步骤如下:任意 m 维概率矢量 $P_i = (p_{1i}, p_{2i}, \dots, p_{mi}), 1 \leq i \leq n$, 以及任意的 n 维概率矢量 $Q = (q_1, q_2, \dots, q_i, \dots, q_n)$, 有式(4)成立.

$$\sum_{i=1}^n q_i H_m(p_{1i}, p_{2i}, \dots, p_{mi}) \leq H_m\left(\sum_{i=1}^n q_i p_{1i}, \sum_{i=1}^n q_i p_{2i}, \dots, \sum_{i=1}^n q_i p_{mi}\right) \quad (4)$$

在上式中,数据 i 的频率矢量是 P_i ; 每个数据的潜在特征提取数为 N ; 数据 i 所包含的总信息量为 $NH_m(p_{1i}, p_{2i}, \dots, p_{mi})$. N 个数据的数据类 $c = \{d_i, \dots, d_n\}$ 的总信息量用不等式右边来表示.

数据组成的信息源在确定性上没有单独的数据强,信息量的变化在数据类之间可以体现数据之间,数据类中潜在特征分布的异同.为了表示反馈所属类别与数据间的共性和个性,潜在特征目标函数公式可表示如下:

$$\arg \min_{\Omega} N \sum_{c \in \Omega} \sum_{d \in c} \{H(c) - H(d)\} \quad (5)$$

在上面公式中,最好的分类方案用 Ω 表示.潜在特征层次聚类中,优化方案的局部阶段可以用下面公式表示:

$$\hat{c} = \arg \min_c \frac{N \sum \{H(c) - H(d)\}}{|c|} \quad (6)$$

在上式中,设计方案着重在数据类中数据数目的异同,在潜在特征层次聚类中,聚类路径的选择是通过每一步选择信息丢失的最小方案来实现的.接下来,着重研究潜在特征关联性聚类.

2.3 潜在特征关联性聚类

在对数据信息进行潜在特征提取后,下一步是找出数据分类方法 Ψ , 这个数据分类方法必须满足样本数据 D 上有最大的后验概率.

$$\Psi = \arg \min_{\Omega} P(\Omega | D) \quad (7)$$

在公式 $P(\Omega | D) = \frac{P(\Omega | D)P(\Omega)}{P(D)}$ 中, $P(D)$

是一个常数,表示的是先验概率.我们假定所有的 Ω 都具有一样的先验概率, $P(\Omega)$ 表示无信息先验概率.我们可以推出以下关系:

$$P(\Omega | D) \propto P(D | \Omega)$$

$$P(D | \Omega) = \prod_{k=1}^K P(c_k | \Omega) \quad (8)$$

在以上公式中, $P(d | \Omega)$ 在聚类划分方案 Ω 中, 聚类数据 d 产生的概率, 同样的, 数据聚类划分方法 Ω 中数据集合 D 的概率为 $P(D | \Omega)$; 每一个潜在特征数据看成一个信源, 潜在特征聚类信息作为所属的性质, 这就会出现一个随机符号序列。

这是因为我们事先假定

$$P(d | \Omega) = P(d | c(d))$$

来判定的, 也就是每个数据都要通过所属的类别, 对应数据类 c 的信息源, 我们用 $\hat{c}$ 来表示。另外, 在无记忆离散信息源中, 潜在特征聚类符号频率分布可以用以下公式表示, 公式如下:

$$\hat{c} = (f_1, f_2, \dots, f_i, \dots, f_M), \sum_i f_i = 1 \quad (9)$$

在以上公式中, $\hat{c}$ 发生潜在特征聚类符号 i 的概率为 f_i , 如果我们假定一个数据 d , 那么它的概率, 可以用下式表示:

$$\begin{aligned} p(d | \hat{c}) &= p(w_1 | \hat{c}) p(w_2 | \hat{c}) \cdots p(w_N | \hat{c}) \\ &= \prod_{i=1}^N p(w_i | \hat{c}) \end{aligned} \quad (10)$$

式中, w_i 表示为数据 d 的第 i 个潜在特征聚类词。已知当数据矢量 $d = (t_1, t_2, \dots, t_i, \dots, t_M)$, $\sum_i t_i = N$, 数据 d 由潜在特征信息源发生的概率如下公式表示:

$$p(d | \hat{c}) = \prod_{i=1}^M f_i^{t_i} \quad (11)$$

由于潜在特征聚类数据之间的相互无关联性, 在数据类中, 潜在特征聚类数据 c 被信息源 $\hat{c}$ 产生的概率可用公式表示如下:

$$P(c | \hat{c}) = \prod_{d \in c} P(d | \hat{c}) \quad (12)$$

2.4 非单调约束条件下数据挖掘

完成了潜在信息特征关联性聚类分析, 下面我们对隐私保护下的潜在特征关联聚类数据进行挖掘, 主要步骤如下:

本文使用最大似然估计来处理一个已知的潜在特征关联聚类数据类别 c , 最大似然估计, 信息源 $\hat{c}$ 可以用各符号的频率可以借助最大似然估计来获得是, 公式如下:

$$\hat{c} = \left[\frac{t_1^{(c)}}{N | c}, \frac{t_2^{(c)}}{N | c}, \dots, \frac{t_M^{(c)}}{N | c} \right] \quad (13)$$

取对数后得到对数似然比 (LLR) 为

$$\begin{aligned} LLR &= \sum_{k=1}^K \sum_{d \in c_k} \sum_{m=1}^M t_m^{(d)} \log \left[\frac{t_m^{(c)}}{N | c} \right] - t_m^{(d)} \log \left[\frac{t_m^{(d)}}{N} \right] \\ &= N \sum_{k=1}^K \sum_{d \in c_k} [H(d) - H(c_k)] \end{aligned} \quad (14)$$

所以, 本文可以给出特定类别数下的优化目标

函数:

$$\arg \max_{\Omega} N \sum_{c \in \Omega} \sum_{d \in c} \{H(d) - H(c)\} \quad (15)$$

以上公式表示的是平均信息量聚类模型中得到的改进函数, 这个函数的平均信息量模型表示的是似然比, 这个似然比是平均信息量潜在特征聚类数据分类方案对 Ω_0 的对数, 挖掘出的信息是符合条件的, 在保护用户隐私的情况下, 能得到效率与准备度很好的折衷。下面通过实验来验证以上结论的正确性。

3 实验

3.1 实验方案介绍

本实验用一个参数生成装置来产生潜在特征聚类数据的值, 为了精确的表示生成的数据, 我们设定一个参数, 并引入一个随机函数, 能产生 $[0-0.5)$ 的随机数, 并用产生的随机数, 作为聚类数据的测试值。再根据矩阵的阶和隐私数据的范围, 来找出满足条件的测试矩阵。

本实验设置最小支持度设为 0.04, 隐私保护度设为 73%。将文献[7]所提的 PPARSCR 混合算法本文实验算法, 分别从约束选择, 通信量, 效率与准确率的折中程度上进行测试比较。

3.2 约束选择实验

本文是在隐私保护的情况下, 进行的聚类特征的提取, 因此必须建立约束条件, 分析事务数, 支持度阈值三方面进行测试。对于约束条件而言, 引入约束选择率的概念, 约束选择率就是数据集中频繁不满足约束与全部频繁项集之间的比值。实验结果见图 1~图 3。

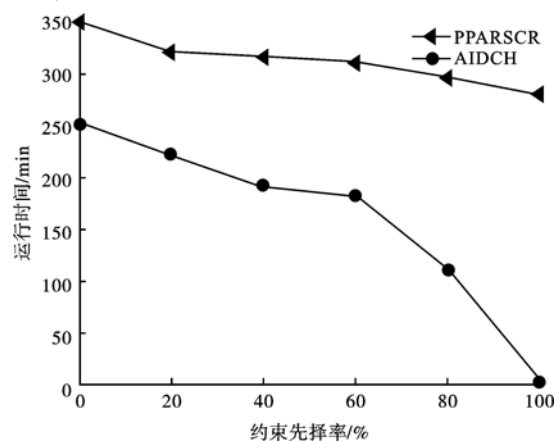


图 1 约束选择率与运行时间对比

实验结果表明: AIDCH 算法将非单调约束性与剪枝方面进行了完善的协调, 并且对多种约束关联是有作用的, 能充分的保护用户的隐私, 并且有很好的扩展性, 可以大量减少数据扫描的冗余次数, 在扩展性和运行时间上都强于 PPARSCR 算法。

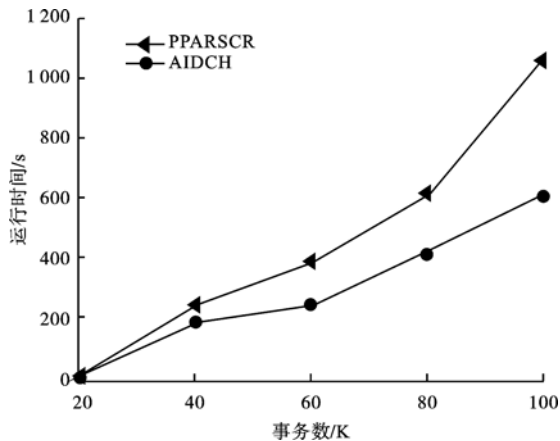


图2 事务数与可扩展性的对应

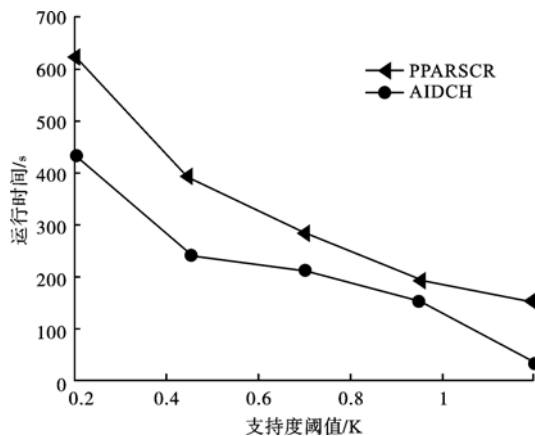


图3 支持度阈值运行结果

3.3 通信量与执行时间实验

通信量与数据方数量对应两算法的比较结果,如图4所示,PPARSCR算法的通信量要明显高于文中AIDCH算法,主要原因是PPARSCR算法是将局部频繁项加密后进行传输,使数据通信量加大,而本文算法AIDCH只传输的选项集干扰后的支持数矩阵,减少了通信量。

对算法的运行速度进行测试,结果如图5所示,从图中可以看出:算法PPARSCR的运行时间明显高于本文算法AIDCH,主要原因在于:文中算法在执行候选频繁项集支持数运算时,只涉及到加法运算,而PPARSCR算法还要涉及到加密、解密运算,从而增加了运算量。

3.4 效率与准确率的折衷度实验

效率和准确率的折衷是本文讨论的重点,与文献[7]中所提算法PPARSCR相比,运行结果如图6所示。

结果表明:文中算法AIDCH和文献[13]中的算法PPARSCR,都随着数据量的增加,相应的折衷度在下降,但从图中可以看出,当数据量很大时,比如超过100 MB, AIDCH算法仍保持很高的折衷度,而PPARSCR算法折衷度极具下降。

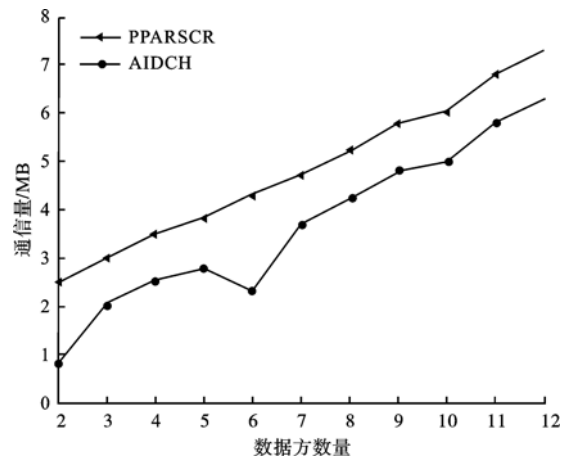


图4 算法通信量比较

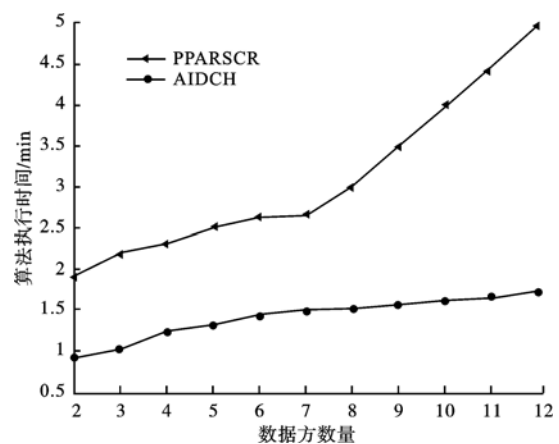


图5 算法执行时间比较

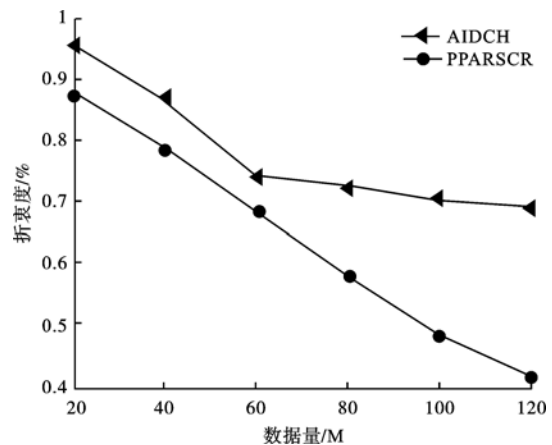


图6 算法折衷度比较

4 结束语

本文通过求得该信息源的各个特征的次数来积累关联规则的信息量,解决了半诚实模型下水平分布数据的隐私数据保护问题,提出针对隐私保护约束条件下的AIDCH算法,算法分别对隐私保护下的数据进行了潜在关联特征的提取,之后又对提取的特征进行聚类,根据聚类后的特征进行数据挖掘,从实验结果来看,该算法减少了大量的计算与通信

开销,提高了挖掘效率.无论是从约束选择,通信量,还是从效率和准确率的折衷度上,本文算法均于传统算法.

参考文献:

- [1] Huang Liming, Zhang Chao, Song Jinling, et al. How to select optimal K-value for K-anonymity model[J]. JCIT, 2013, 8(1): 811-819.
- [2] Zhang Haitao, Xu Liang, Huang Huihui, et al. Mining sequential patterns from anonymous datasets for LBS users privacies protection[J]. JCIT, 2013, 8(2): 77-86.
- [3] 刘峰,薛安荣,王伟.一种隐私保护关联规则挖掘的混和算法[J]. 计算机应用研究, 2009, 29(3): 1109-1111.
- [4] Zhu Bin, Liao Junguo, He Yong, et al. Selectively disclosing sensitive attributes in digital certificate for privacy protection[J]. IJACT, 2013, 5(2): 755-736.
- [5] Luo Wenjun, Liu Lu, Yuan Shumei, et al. Batch verification of the data integrity in cloud computing[J]. JDCTA, 2012, 6(15): 373-380.
- [6] Deng Zhijuan, Zhong Shaojun. A kind of text classification design on the basis of natural language processing[J]. IJACT, 2013, 5(1): 668-677.
- [7] Wu Fangjun. Scale-free characteristic in open source software: an empirical case study[J]. IJACT, 2013, 5(1): 792-799.

- [8] Andy S Y Lai, POON Y C. i-Questionnaire - A Software Service Tool for Data Analysis in e-Business[J]. IJIPM, 2013, 4(1): 1-9.
- [9] 金菁. 基于改进的聚类平均信息量文本数据挖掘算法研究[J]. 计算机应用研究, 2009, 29(3): 982-984.
- [10] Somboon Anekritmongkol, Kulthorn Kasemsan. SQL model in language encapsulation and compression technique for association rules mining[J]. IJIPM, 2013, 4(1): 65-75.
- [11] 胡悦,王伟民. 物联网无线传感器能耗优化 SIFT 算法的研究[J]. 科技通报, 2013(9): 98-101.
- [12] Jin Haiyan, Li Shuai, Wang Bingbo. Algorithm of wavelet texture segmentation based on geometrically regularity[J]. JDCTA, 2013, 7(1): 99-106.
- [13] Deng Xiaoyi. An enhanced artificial bee colony approach for customer segmentation in mobile E-commerce environment[J]. IJACT, 2013, 5(1): 139-148.

作者简介:

刘伟男, (1975—), 副教授, 硕士. 研究方向为计算机网络技术、智能算法研究、移动互联网.

崔永锋男, (1978—), 副教授, 硕士. 研究方向为计算机应用技术、工程优化算法、移动互联网.

周丁丁女, (1980—), 实验师. 研究方向为计算机应用技术.

(上接第 167 页)

签,设计中给出了有源 RFID 低功耗标签的总体构架,对标签所用主要模块低功耗特性进行了简要分析.介绍了休眠、就绪、工作等状态的转换的低功耗方法,以及标签从低功耗中唤醒的机制,提出了动态倍频的低功耗方案,实现了从高频的 ECC 校验到低频的休眠状态之间的动态频率调整.

对于低功耗机制还可以结合 CC2520 射频芯片的特性进行优化,本文设计低功耗机制时只考虑到对 CPU 部分的能耗降低,而未考虑射频模块的低功耗特性,可以作为下一步工作研究方向.

参考文献:

- [1] Klaus Finkenzeller. RFID handbook: fundamentals and applications in contactless smart cards and identification[M]. New York: Wiley, 2003.
- [2] 廖志鹏,黄俊,雷杏. 基于 RFID 超低功耗射频标签的研究与实现[J]. 压电与声光, 2012, 34(6): 951-954.
- [3] 宋彦,郭兵,沈艳. 嵌入式系统低功耗的软件实现方案[J]. 单片机与嵌入式系统应用, 2011(9): 4-7.
- [4] VTiwari, S Malik, A Wolfe. Power analysis of em-

bedded software: a first step towards software power minimization[J]. IEEE Transactions on VLSI Systems, 1994, 2(4): 437-445.

- [5] GJB7377-2. 2011. 军用射频识别空中接口协议: 2.45GHz 参数[S]. 2011.
- [6] 张伟,韩进. 低功耗嵌入式系统的分析与应用[J]. 计算机系统应用, 2009(7): 109-113.
- [7] Luca B. A survey of design techniques for system-level dynamic power management[J]. IEEE Transactions on VLSI Systems, 2000, 8(3): 299-316.
- [8] 王永超,郭瑞,包贵浩,等. 低功耗有源 RFID 标签设计与实现[J]. 电子测量技术, 2010, 33(7): 31-33.

作者简介:



胡翔男, (1988—), 硕士. 研究方向为嵌入式与物联网.