

改进 Apriori 算法在医院监护中心的研究与应用

龙冰莹, 陈小惠

(南京邮电大学 自动化学院, 江苏 南京 210046)

摘要: 为了提高对医院监护中心历史数据的管理水平, 为监护人员提供有力的决策支持, 提出了一种针对该系统的改进 Apriori 算法。该算法引入了属性值度的概念, 减少了找出频繁项集所需要的时间, 也减少了扫描数据库的次数。为了验证改进 Apriori 算法的正确性、有效性和快速性, 文中将改进的 Apriori 算法与传统的 Apriori 算法分别应用到医院监护中心系统中去, 并对两种算法的效率进行了比较。结果表明, 改进 Apriori 算法能够得到所需要的强关联规则, 并在效率上有显著的提高, 为监护人员更好控制患者的病情提供了很好的决策支持。

关键词: 数据挖掘; 关联规则; Apriori 算法

中图分类号: TP39

文献标识码: A

文章编号: 1673-629X(2013)08-0137-04

doi: 10.3969/j.issn.1673-629X.2013.08.035

Research and Application of an Improved Apriori Algorithm in Hospital Monitoring Center

LONG Bing-ying, CHEN Xiao-hui

(College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: To increase the level of hospital historical data management and provide a strong support for supervisors, an improved Apriori algorithm is proposed specific to this system. The improved algorithm brings in a concept of the attribute-value-degree, reduce the need time to find frequent itemsets, and also reduce the times of scanning database. In order to validate the correctness, effectiveness and speediness of the new algorithm, both the new Apriori algorithm and the traditional Apriori algorithm are applied to this system, and the two algorithms are compared. The result shows that the new algorithm is able to dig useful and reasonable rules, has better performance in efficiency, and provides decision support of better control diseases to patients for supervisors.

Key words: data mining; association rule; Apriori algorithm

0 引言

医院监护中心将通过采集来的病人或疑似病人的体温和心率等生理参数, 传递到监护中心, 监护人员根据病人生理参数的变化来判断病人的病情, 对病情进行很好的判断和控制。在此监护中心, 若能知道各生理参数与病情的关系, 就能根据病人的生理参数值来判断病情的转变, 这样便能对病情有可能恶化的病人及时进行治疗, 控制病情。因此, 从已存的相关数据中挖掘出各生理参数与病情之间的规则是很有必要的。

关联规则 (Association Rule) 是 R. Agrawal 等人于 1993 年首先提出的, 是数据挖掘中一个重要研究内容。关联规则在医学领域的应用也非常广泛。国内外文献都有将关联规则的 Apriori 算法应用于挖掘各种

疾病之间的关系和医疗管理系统中的相关研究^[1]。目前, 常采用的主要的关联规则挖掘算法是 Apriori 算法和 FP-growth 算法^[2], 其他常用算法都是在这两种算法上进行的改进。其中, Apriori 算法是最经典的关联挖掘算法。

1 Apriori 算法

Apriori 算法是基于频繁项集理论的递推方法, 也就是一种逐层搜索的迭代方法, 即用 k _项集去探索 $(k+1)$ _项集。Apriori 算法使用频繁项集的先验知识, 首先找出频繁项集 1_项集的集合, 该集合作为 L_1 。再用 L_1 找出项集 2_项集集合 L_2 , L_2 找出 L_3 , 如此下去, 直到不能找出频繁 k _项集。寻找频繁项集可以分为两

收稿日期: 2012-11-01

修回日期: 2013-02-20

网络出版时间: 2013-04-22

基金项目: 国家自然科学基金资助项目 (61104216); 江苏省科技支撑计划项目 (BE2011843)

作者简介: 龙冰莹 (1988-), 女, 湖南衡阳人, 硕士, 主要从事数据挖掘方面的研究; 陈小惠, 教授, 主要从事智能仪器方面的研究。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130422.1727.052.html>

个过程:

(1) 连接。为了找出 L_k , 通过 L_{k-1} 与自己连接产生候选 k 维项集的集合。该候选项集的集合记作 C_k 。设 l_1 和 l_2 是 L_{k-1} 中的项集。记号 $l_i[j]$ 表示 l_i 的第 j 项。执行连接 $L_{k-1} \cup L_{k-1}$, 其中 L_{k-1} 的元素是可连接的, 如果它们的前 $k-2$ 个项相同, 连接 l_1 和 l_2 产生的结果项集是 $l_1[1]l_1[2]\cdots l_1[k-1]l_2[k-1]$ 。

(2) 剪枝。 C_k 是 L_k 的超集。根据性质: 任何非频繁的 $k-1$ 维项集都不可能是频繁 k 维项集的子集。因此, 如果一个候选 k 维项集的 $k-1$ 维子集不在 L_{k-1} 中, 则该候选也不可能是频繁的, 从而可以从 C_k 中删除。然后扫描数据库, 确定 C_k 中的每个候选项集的支持度, 从而确定 L_k 。

Apriori 算法作为经典的关联挖掘算法, 具有里程碑的作用。但在执行速度和效率上存在两个缺陷^[3]: 1) 产生大量的候选项集; 2) 多次扫描数据库。

目前, 国内外许多学者都为克服这两个缺点对 Apriori 算法进行了优化和改进。文献[4]提出了 DHP 算法采用 Hash 技术和数据库中的事务的缩减技术来提高 Apriori 算法的效率。文献[5]提出 TreeProjection 算法通过建立字典顺序的项集来产生频繁项集; 一些基于图、兴趣度的关联规则算法也在不断提出^[6,7]。文献[8]提出了基于分辨矩阵改进的 Apriori 算法, 用于提高关联规则挖掘的速度和效率。文献[9]利用“分割-整合”的思想改进了 Apriori 算法, 并将改进算法应用于图书推荐服务中, 证明了其有效性与快速性。关于多值属性关联规则的, 文献[10]中将多值分成多个布尔值转化成布尔形式来处理, 该方法简单, 但使得属性会大量增加, 效率很低。

2 改进的 Apriori 算法

文中在对要处理的事务数据库进行分析后, 发现该事务数据中存在着某些规律。因此, 文中在这些规律的基础上对传统的 Apriori 算法进行了改进, 该改进算法不需要多次扫描数据库和连接运算来产生频繁项集, 提高了效率, 而且更加适合该系统, 能针对系统的需要快速地挖掘出有用的规则。

2.1 改进算法的提出与相关定义

由于系统数据的来源是对某传染病瞬时采集的历史数据, 数据显示的是各病人每个采集时刻的生理参数与病情, 受文献[11]中加入时间判断来处理异常现象的启发, 文中在对这些数据进行关联规则算法前作了一些处理, 也加入了一个时间段的判断, 删除那些在这个时间段内病情跳跃比较大的数据, 这样所得的数据更具有可参考性。在求频繁项集时, 文中参考了文献[12]中设置事务度的方法, 对每个属性中的属性值

赋不同的值, 这样可以使得从每个属性中选取的不同属性值之和都不相同, 由于事务数据库中事务呈现的规律是每个事务都包含所有的属性, 而需要挖掘的关联规则是病人所有生理参数与病情的强关联规则, 因此, 只需要计算出相同事务属性值之和的个数大于最小支持度, 就可以得出所需的频繁项集。这样, 不仅删除了不需要的频繁项集, 而且大大减少了寻找频繁项集所耗费的时间。

为了更清楚表示给每个属性的赋值, 下面给出了两个定义。

定义 1 属性值的度。设属性 I_i 中包含 y_i 个属性 ($i=1, 2, \cdots, n$), 从第一个属性 I_1 的第一个属性值起, 到最后一个属性 I_n 的最后一个属性值, 分别赋值为 1 到 $y_1 + y_2 + \cdots + y_n$, 则称属性值对应的唯一的数值为属性值的度。

定义 2 项集的度。令项集 $I = \{I_1, I_2, \cdots, I_n\}$, 项集的度 $d(I)$ 为所包含属性值的度之和, 即 $d(I) = d(I_1) + d(I_2) + \cdots + d(I_n)$ 。

例: 若属性 I_1 中含有 3 个属性值, 属性 I_2 中含有 4 个属性值, 从 I_1 第一个属性值到 I_2 最后一个属性值分别赋值为: 1, 2, 3, 4, 5, 6, 7。则 $I = \{I_1, I_2\}$ (其中 I_1 取其第 2 个属性值, I_2 取其第 4 个属性值), 则 I 的度 $d(I) = 2+7=9$ 。

2.2 改进算法的设计

针对医院监护中心系统的特点对 Apriori 算法进行改进后, 算法的步骤包括:

步骤 1 事务数据库处理。对某医院或该系统中的历史数据进行处理, 抽取病人每诊断时刻的一些生理参数 (例如, 年龄、性别、体温、心率等) 和病情作为记录, 然后, 对数据记录进行处理, 删除那些在某个时间间隔内病情跳跃较大的数据, 将处理后的数据作为算法的输入。

For each $T_i \in D, T_j \in D (i \neq j)$ and T_i . People = T_j . People {

If (abs (T_i . time - T_j . time) <= interval && T_i . $I_{<1 \sim n-1>}$ == T_j . $I_{<1 \sim n-1>}$ && T_i . I_n != T_j . I_n)

Delete T_i, T_j ; }

Return $D' = \{T_1, T_2, \cdots, T_n\}$;

步骤 2 产生频繁项集。对每个属性的属性值赋值 (即获得其属性值度) 后, 计算所需形式的项集的度, 并统计相同项集度的数量, 若数量大于或等于最小支持度 (min_sup), 则为所需的频繁项集。

For each $I_i (i=1, 2, \cdots, m)$

$I_{i,j}$ (第 i 个属性, 第 j 个属性值) = (1, 2, \cdots , $y_1 + y_2 + \cdots + y_m$)

For each $T_i \in D' (i=1, 2, \cdots, n)$ {

```
if( Count( d(I = < I1, I2, ⋯, Im-1 > or < I1, I2, ⋯, Im > ) ) >= min_sup ) L. Add ( I ); }  
return L  
步骤 3 生成强关联规则。算法只生成形如这样的规则  $X \Rightarrow Y$ ,  $X$  是生理参数集合,  $Y$  是病情。若规则的置信度超过给定的最小置信度 (min_conf), 则认为是强关联规则。  
For each  $L_i \in L$  {  
   $X = \langle I_1, I_2, \dots, I_{m-1} \rangle$  ;  
   $Y = I_m$  ;  
  If( conf(  $X \Rightarrow Y$  )  $\geq$  min_conf ) AssoRule(  $X \Rightarrow Y$  );  
}
```

2.3 仿真实验与分析

为了验证改进 Apriori 算法的正确性、有效性和快速性,选取了某医院的某种传染病的历史数据作为样本。在相同的硬件配置条件:Genuine Intel(R) CPU、1.73GHz 的主频、1GB 的内存、160GB 硬盘、Windows XP sp3 操作系统环境下,对改进 Apriori 算法和传统 Apriori 算法效率进行测试,比较改进 Apriori 算法和传统 Apriori 算法的时间,在所有的仿真数据计算中,改进 Apriori 算法的挖掘结果是包含在传统 Apriori 算法的结果中的,这说明传统 Apriori 算法挖掘出了大量的无意义、无用的关联规则,结果冗长,不利于观看。而改进 Apriori 算法仅挖掘出了所需要的强关联规则,效率高,且计算时间远小于传统 Apriori 算法。实验结果仿真图如图 1 和图 2 所示。

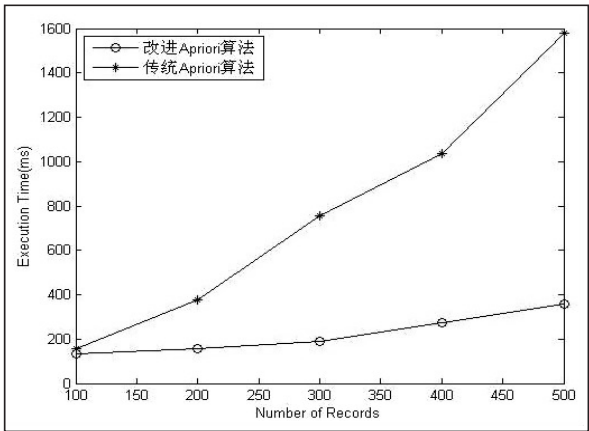


图 1 算法计算时间与记录数的关系

图 1 反映的是两种算法的计算时间与记录数的关系。从该图可见,改进 Apriori 算法和传统 Apriori 算法的计算时间随记录数目的增加而增加,但改进 Apriori 算法的增长幅度较小,且改进 Apriori 算法的计算时间远小于传统 Apriori 算法。

图 2 反映的是在相同记录数下两种算法的计算时间与最小支持度的关系。从该图可见,改进 Apriori 算法和传统 Apriori 算法的计算时间随最小支持度的增

加而减小,但改进 Apriori 算法基本保持一条与 X 轴平行的直线,且在最小支持度很小时它的计算时间远小于传统 Apriori 算法,这说明改进 Apriori 算法更适应最小支持度比较小时的关联规则挖掘。

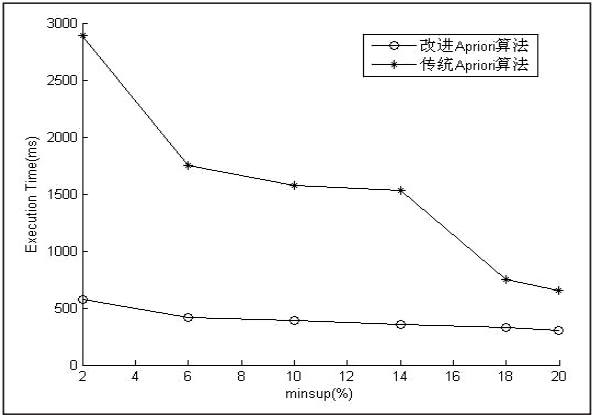


图 2 算法计算时间与最小支持度的关系

3 改进 Apriori 算法的应用

随着数据量越来越大,关联规则能在大量的数据中寻找它们之间的关系,在生活中的应用越来越广泛。文献[13]将关联规则算法应用于高校管理系统中,并达到了很好的效果;Apriori 算法在税务系统中的应用,对历史稽查数据中纳税人采用的主要违法违规手段之间的关系进行了数据挖掘,得到了一些合理的规则,对稽查工作有一定的指导意义。

关联规则在远程监护中心系统中的应用,主要是通过通过对历史数据挖掘出强关联规则,再通过强关联规则 and 该系统的病人的生理参数来判断系统中可能出现病情恶化的病人,并将信息反馈给监护人员,以至监护人员能及时地控制病人的病情。

该系统的框架图如图 3 所示。

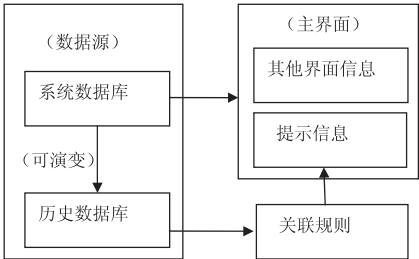


图 3 系统框架图

将改进 Apriori 算法加入该远程监护系统中,可以从系统登录后进入关联规则挖掘界面,监护人员在“最小支持度”和“最小置信度”中输入相应值,如“最小支持度为:0.06”和“最小置信度为:0.3”;首先点击“生成关联规则”,所得的强关联规则为:

20 ~ 30 岁,男,37.5 度,90 ~ 100 跳/分钟→中度感染——97%

20 ~ 30 岁,女,36.7 ~ 37.0 度,80 ~ 90 跳/分钟→

轻度感染——91%

20~30岁,女,大于38.5度,大于110跳/分钟→
严重感染——91%

20~30岁,男,大于38.5度,大于110跳/分钟→
严重感染——100%

再点击“重点监护者”,这时,“重点监护者”栏中显示的是由上面关联规则判断所得的病情有可能恶化的病人编号和姓名:

0001,李明

0002,晓余

监护人员可以对这两位病人进行重点监护,有效地控制病人的病情。

从以上实验中可知,改进Apriori算法在监护中心系统中的实际应用是正确的,并且是有效的、快速的。

4 结束语

传统的多值属性Apriori算法先将属性转化为布尔值再使用迭代自连接会产生大量的重复候选集和没必要扫描事务数据库,效率很低。文中提出了一种基于属性值度来找出频繁项集的改进算法。经实验表明,改进Apriori算法能够挖掘出持有特征属性形成的强关联规则,这些关联规则表明了生理参数与病情之间的关系,为监护人员更好控制病人病情提供了很好的决策支持。

参考文献:

- [1] 胡镜清,刘保延,王永炎.中医临床个体化诊疗信息特征与数据挖掘技术应用分析[J].世界科学技术:中医药现代

化,2004,6(1):14-16.

- [2] 刘步中.基于频繁项集挖掘算法的改进与研究[J].计算机应用研究,2010,29(2):475-477.
- [3] 刘维晓,陈俊丽,屈世富,等.一种改进的Apriori算法[J].计算机工程与应用,2011,47(11):149-151.
- [4] Gatos B, Mantzaris S, Perantonis S, et al. Automatic page analysis of a digital library from newspaper archives[J]. International Journal of Digital Libraries, 2003, 3(1): 77-84.
- [5] Aiello M, Monz C, Todoran L, et al. Document understanding for a broad class of documents[J]. International Journal on Document Analysis and Recognition, 2002, 5(1): 1-16.
- [6] 陈立宁,罗可. Apriori 算法用于频繁子图挖掘的改进方法[J]. 计算机工程与应用, 2011, 47(10): 113-117.
- [7] 王德兴,胡学钢,刘晓平,等.改进购物篮分析的关联规则挖掘算法[J].重庆大学学报:自然科学版,2006,29(4):105-107.
- [8] Wang Peiji, Shi Lin, Bai Jinniu, et al. Mining Association Rules Based on Apriori Algorithm and Application[C]//Proc. of 2009 International Forum on Computer Science-Technology and Applications. [s. l.]: [s. n.], 2009: 141-143.
- [9] 林郎碟,王灿辉. Apriori 算法在图书推荐服务中的应用与研究[J]. 计算机技术与发展, 2011, 21(5): 22-24.
- [10] 张朝晖,陆玉昌,张钺.发掘多值属性的关联规则[J].软件学报,1998,9(11):801-805.
- [11] 高琰,王台华,郭帆,等.应用非迭代Apriori算法检测分布式拒绝服务攻击[J].计算机应用,2011,31(6):1521-1524.
- [12] 汪维清,罗先文,胡继宽. Apriori-Sort 算法研究[J]. 计算机工程与应用, 2008, 44(36): 156-159.
- [13] 张宗郁,张亚平,张静远,等.改进关联规则算法在高校教学管理中的应用[J].计算机工程,2012,38(2):75-77.

(上接第136页)

- Anchorage, AK; [s. n.], 2007: 2045-2053.
- [2] Blaβ E O, Zitterbart M. An Efficient key establishment schema for secure aggregation sensor networks[C]//Proc. of the 2006 ACM Symposium on Information, Computer and Communications Security. New York, USA; [s. n.], 2006: 303-310.
- [3] 杨庚,王安琪,陈正宇,等.一种低能耗的数据融合隐私保护算法[J].计算机学报,2011,34(5):792-800.
- [4] Bista R, Kim H D, Chang J W. A new private data aggregation scheme for wireless sensor networks[C]//Proc. of 10th IEEE International Conference on Computer and Information Technology. Bradford, UK; [s. n.], 2010: 273-280.
- [5] Bista R, Yoo H K, Chang J W. A new sensitive data aggregation scheme for protecting integrity in wireless sensor networks[C]//Proc. of 10th IEEE International Conference on Computer and Information Technology. Bradford, UK; [s. n.], 2010: 2463-2470.
- [6] Groat M M, He W, Forrest S. KIPDA: K-Indistinguishable privacy-preserving data aggregation in wireless sensor networks

[C]//Proc. of the 30th IEEE International Conference on Computer Communications. Shanghai, China; [s. n.], 2011: 2024-2032.

- [7] 刘鑫芝.无线传感器网络安全数据融合算法研究[J].计算机与现代化,2010(5):151-155.
- [8] 唐慧,胡向东.无线传感器网络安全数据融合算法研究[J].通信技术,2007,40(12):290-293.
- [9] 罗蔚,胡向东.无线传感器网络中一种高效的安全数据融合协议[J].重庆邮电大学学报(自然科学版),2009,21(1):110-114.
- [10] 覃志松,黄延磊. Zigbee 无线传感器网络安全研究及改进[J].微计算机信息,2010,26(3-2):54-55.
- [11] 邓黎黎,刘才兴.基于信任的无线传感器网络安全路由研究[J].计算机技术与发展,2010,20(6):159-162.
- [12] Madden S, Franklin M J, Hellerstein J M. TAG: a tiny aggregation service for ad-hoc sensor networks[C]//Proc. of the 5th Symposium on Operating Systems Design and Implementation. New York, USA; [s. n.], 2002: 131-146.

改进Apriori算法在医院监护中心的研究与应用

作者：[龙冰莹](#)，[陈小惠](#)，[LONG Bing-ying](#)，[CHEN Xiao-hui](#)
作者单位：[南京邮电大学 自动化学院, 江苏 南京, 210046](#)
刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2013(8)

本文链接：http://d.wanfangdata.com.cn/Periodical_wjfz201308035.aspx