

改进关联规则算法在医疗监控中的应用

田亚凯, 陈小惠

(南京邮电大学 自动化学院, 江苏 南京 210046)

摘要: 为了方便监护人员能够从监控中心的数据库中及时获取病人的生理参数和病情的关系, 文中提出一种改进的 Apriori 算法。该算法充分利用医疗数据的特点, 把整条事务当作一个属性, 避免了传统 Apriori 算法反复扫描事务数据库, 从而简化了生成频繁项集的过程。为了验证改进 Apriori 算法的有效性和可行性, 文中将此算法与传统的 Apriori 算法分别应用到监护系统中, 对这两种算法的运行结果进行了对比。结果表明, 改进 Apriori 算法在效率上有明显提高, 为监护人员针对一些突发性疾病做出及时诊断提供了良好的决策支持。

关键词: 数据挖掘; 关联规则; Apriori 算法; 医疗监护

中图分类号: TP39

文献标识码: A

文章编号: 1673-629X(2015)10-0183-04

doi: 10.3969/j.issn.1673-629X.2015.10.040

Application of an Improved Algorithm of Association Rules in Health Monitoring Center

TIAN Ya-kai, CHEN Xiao-hui

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

Abstract: In order to facilitate supervisors to get the relationship between physiological parameters and disease of patients from monitoring center database timely, put forward a kind of improved Apriori algorithm. The algorithm makes full use of the characteristics of medical data and regards the entire transaction as an attribute, avoiding the operation that traditional Apriori algorithm repeatedly scans the transaction database, so that simplifies the process while generating candidate sets. In order to verify the effectiveness and feasibility of the improved Apriori algorithm, both the Apriori algorithm and the new algorithm are applied to the monitoring system, also the results of the two algorithms are compared. The results show that the improved Apriori algorithm has obvious improvement in efficiency, providing a good decision support about some sudden illness for the care staff timely.

Key words: data mining; association rule; Apriori algorithm; medical monitoring

0 引言

数据挖掘是数据库领域研究的热点问题, 所谓数据挖掘是指从数据库的大量数据中提取出隐含的、先前未知的并有潜在价值的信息的过程^[1]。数据挖掘的方法有很多, 其中应用最广泛的方法之一就是发现数据中的关联规则^[2]。目前, 数据挖掘技术在医疗中被广泛应用, 同时经过许多国内外专家的努力, 一些关联规则也被很好地应用于各个医疗系统, 大大完善了目前的医疗诊断系统。

文中的医疗监控中心通过许多户外的采集前端, 利用它们自带的无线模块来获取它们采集到的病人的

生理数据(包括病人的体温和心率), 然后由医疗监控中心的监护人员做出医疗诊断。在此, 如果能及时获取病人的生理参数和病情的关系, 这样就能针对一些有突发性疾病的病人做出及时的诊断, 对保障病人的生命安全有重大意义。因此, 就需要准确快速地挖掘出这其中的关联规则。

目前主要的关联规则挖掘算法是 Apriori 算法^[3], 但是传统的 Apriori 算法要反复扫描事务数据库, 效率低。对此, 文中提出一种改进的 Apriori 算法应用于医疗系统中, 有效提高了 Apriori 算法的运行效率。

1 关联规则定义及 Apriori 算法

1.1 关联规则概述

关联规则的挖掘问题可形式化定义如下^[4]:

设 $I = \{i_1, i_2, \dots, i_m\}$ 是由 m 个不同的项组成的集合。给定一个事务数据库 D , 其中每一个事务 T 是 I 中一组项的集合, 每个事务 T 有唯一的标志符 TID。若项集 $A \subseteq I$ 且 $A \subseteq T$, 则事务 T 包含项集 A 。关联规则可表示为形如 $A \Rightarrow B$ 的表达式, 其中 $A \subset T, B \subset T$, 且 $A \cap B = \emptyset$ 。表示的意思是可以从事务中的某些项推导出另外的项。

关联规则有如下几个基本概念^[5]:

支持度 sup: 表示 D 中包含 $A \cup B$ 事务的百分比, 即

$$\text{sup}(A \Rightarrow B) = p(A \cup B)$$

置信度 conf: 表示 D 中包含 A 的事务中 B 出现的可能性。即

$$\text{conf}(A \Rightarrow B) = p(B | A) = \frac{\text{support_count}(A \cup B)}{\text{support_count}(A)}$$

其中, $\text{support_count}(A \cup B)$ 是包含项集 $A \cup B$ 的事务数; $\text{support_count}(A)$ 是包含项集 A 的事务数。

最小支持度: min_sup 表示项集在数学统计中的最小支持数。最小置信度 min_conf 表示规则的可靠性的最低值。最小支持度和最小置信度一般来自于用户的需求。

强规则: 同时满足最小支持度阈值 (min_sup) 和最小置信度阈值 (min_conf) 的规则称为强规则。

频繁项集: 当一个项集的支持度大于或等于给定的最小支持度时, 这个项集被称为频繁项集。

对于给定一个事务集 D , 挖掘关联规则就是找出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的关联规则。

1.2 Apriori 算法

Apriori 算法是关联规则中最重要的一种挖掘频繁项集的算法, 它的算法思想^[6]是逐层搜索的迭代方法, 利用已知的 $k-1$ 维频繁项集来生成 k 维频繁项集。

具体做法是: 首先找出频繁 1-项集, 记为 L_1 ; 然后利用 L_1 来挖掘 L_2 , 即频繁 2-项集; 不断如此循环下去直至无法发现更多的频繁 k -项集为止, 每挖掘一层就需要扫描整个数据库一遍。根据 Apriori 的性质, 由 L_{k-1} 寻求 $L_k (k \geq 2)$ 的过程主要分为以下两步:

(1) 连接。将 L_{k-1} 与自身连接产生候选 k -项集 L_k 。

(2) 剪枝。首先, 如果 L_k 的 $k-1$ 项子集不在 L_{k-1} 中, 则该候选项集是非频繁的, 把它从 L_k 中删除。然后结合最小支持度产生 L_k 。

Apriori 算法虽然是最常用最经典的算法, 但是其仍然存在不足之处, 如: 反复多次扫描数据库和产生大量候选项集等。同时算法在 I/O 上读取所消耗的时间很多^[7], 这使得算法的效率非常低。因此, 需要对算法的上述缺陷进行改进, 从而提高 Apriori 算法的运行效率。

2 改进的 Apriori 算法

通过对监护中心数据库进行分析之后, 发现病人的生理参数 (包括体温和心率) 和病情之间有着必然的联系。而整个医疗系统的主要功能是能够远程实时地监测病人的生理信息, 所以其采集前端是每隔 30 s 采集一次患者的生理参数, 因而其数据量又十分庞大。而传统的 Apriori 算法^[8]又要反复扫描如此大的数据库, 因此可以利用一种改进的 Apriori 算法, 只扫描一次事务数据库, 并且简化了产生频繁项集的过程, 从而提高了监护人员针对系统数据进行分析的效率。

2.1 改进算法的提出

众所周知, 病人的病情和许多因素有关, 包括年龄、性别、温度、心率, 而需要通过这些因素去挖掘出其与病情的关联规则。为此文中介绍并应用了一种改进的 Apriori 算法^[9]。该算法把数据库中的每个事务都看作一个属性值, 不同的事务就是不同的属性值, 在扫描数据库时, 先给每个项赋值, 通过比较事务集中事务的每个项的值是否一致来确定该事务的支持度, 然后计算出每个项的值都一致的事务的数量, 即为其支持度, 从而可以得到所需要的频繁项集。在计算其置信度时, 给每个事务再添加一个病情项, 将此病情项根据病人的病况分为 3 类, 分别为严重感染、中度感染和轻度感染, 在求出频繁项集时, 同时计算含有这 3 类病况的事务量, 再通过定义即可求出病人与每种病况之间的置信度, 从而最终确定病人与病情的关联规则。这样只需扫描一次事务数据库, 就可以求出所要的频繁项集, 简洁明了, 又十分适合医疗监护系统。

算法的具体思路如下:

(1) 读取数据库的第一条记录, 获取包含该记录和病情项组成的 2-项集, 并标记此时它们的支持计数为 1;

(2) 读取数据库的第二条记录, 同样获取含该记录和病情项组成的 2-项集, 如果在之前的记录和组合中均未出现, 则标记该记录和组合的支持计数为 1, 若已经出现, 则对它们的支持计数加 1;

(3) 与步骤 2 的处理方法相同, 依次处理数据库的所有记录, 并完成记录数和组合的相应支持计数的统计;

(4) 根据设置的最小支持度, 完成对它们的剪枝,

并计算出相应的置信度。再根据最小置信度,筛选出需要的强关联规则。

2.2 改进算法的设计与实现

2.2.1 数据挖掘的总体设计流程

利用关联规则进行数据挖掘,首先采集病人病情等生理参数信息,然后对数据进行预处理,生成原始信息表,再将原始信息表转化为挖掘所需的数据表,再对此数据表进行挖掘,得到挖掘模式或模型,最终得到所需要的强关联规则。数据挖掘总体设计流程见图 1。

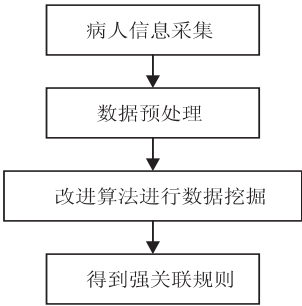


图 1 数据挖掘总体设计流程图

2.2.2 病人生理参数数据的预处理

由于文中算法是应用于远程医疗监控系统,其中病人的病情状态主要是根据所获得温度和心率来判别的,因此就需要从若干数据中挖掘出其与病情的关系,而温度和心率等患者生理参数是不能够直接进行数据挖掘的,因此需要依照医疗数据的特点,对它们进行形式的概化^[10]。

因为要挖掘的是温度和心率与疾病的关联规则,而采集到的温度和心率的值有一定偶然性,因此可以将一些变化范围较大的不正常数据先删除,再取一段时间的测量值作为参考数据,再取这些数据的平均值作为进行相关数据挖掘的原始数据^[11]。同时因为温度和心率必然是在一定范围的,而病人的病情又与其年龄和性别有一定的关系,因此就可以把温度、心率、年龄和性别这四个属性按照它们的数值类型大小划分成多个区间,划分的依据可以按照医生的专业经验,这样就能够便于挖掘出温度、心率和年龄与病情的关系。

具体做法是:将年龄分为 3 类,60<年龄用 A_3 表示,40≤年龄≤60 用 A_2 表示,20≤年龄<40 用 A_1 表示;将性别分为 2 类,男用 M_1 表示,女用 M_2 表示;将温度划分为三类,35.9<温度<36.8 用 T_1 表示、36.8≤温度<38.0 用 T_2 表示,38.0≤温度用 T_3 表示;将心率也分成三类,心率<80 用 X_1 表示,80≤心率≤110 用 X_2 表示,心率>110 用 X_3 表示,将这些新分成的类作为数据挖掘的项的值。

下面的例子是将采集到的温度和心率等原始数据进行转化,表 1 是原始生理参数表,表 2 是最终应用于数据挖掘的事务数据表。

表 1 原始生理参数表

姓名	性别	年龄	温度	心率
李明	男	26	37.3	90
晓余	女	24	37.5	86
小乐	女	25	37.5	88
小天	男	28	37.6	81

表 2 事务数据表

Age	Sex	Temp	Heart	State
A_1	M_1	T_2	X_2	I_1
A_1	M_2	T_2	X_2	I_2
A_1	M_2	T_2	X_2	I_2
A_1	M_1	T_2	X_2	I_1

2.2.3 根据事务属性值进行支持度统计

在对数据进行挖掘时,直接把包含四个项的一条事务看作最终的一个项集,如 $T_1 = \{A_1, M_1, T_2, X_2\}$,因此在扫描事务库时,直接找出本事务出现的次数即为其的度值,当其度值大于或者等于支持度^[12]时,即为频繁项集。通过上面的数据预处理,得到了适合挖掘的数据库,记录总数为 583 条。根据温度、心率、年龄和性别这四个属性,扫描数据库,得它们的支持度计数信息如表 3 所示。

表 3 事务支持度计数表

事务	属性值	支持度计数
T_1	$A_1 + M_1 + T_2 + X_2$	226
T_2	$A_1 + M_2 + T_2 + X_2$	189
T_3	$A_1 + M_2 + T_3 + X_3$	87
T_4	$A_1 + M_1 + T_3 + X_3$	81

2.2.4 求出强关联规则

已经找出所需要的频繁项集,而目的是要根据这些频繁项集推断出其与病情的关联规则。这就需要在数据库中添加病情这一项集,用 I 表示,从而使得每个事务形成一个新的项集,如 $TT_1 = \{A_1, M_2, T_3, X_3, I_2\}$,计算出每个不同事务中各类感染情况的计数值,具体如表 4 所示。

表 4 事务(含病情项)支持度计数表

事务	属性值	支持度计数
TT_1	$T_1 + I_1$	7
TT_2	$T_1 + I_2$	219
TT_3	$T_2 + I_1$	176
TT_4	$T_2 + I_2$	23
TT_5	$T_3 + I_2$	4
TT_6	$T_3 + I_3$	83
TT_7	$T_4 + I_3$	81

根据以上数据,可计算出相应的置信度,若大于设定的最小置信度,即为强关联规则^[13]。以 $T_1 \Rightarrow TT_2$ 为例,计算其置信度为:

$$\text{conf}(T_1 \Rightarrow TT_2) = 219/226 = 0.97$$

2.3 仿真算法与分析

为了验证改进 Apriori 算法的可行性和快速性,在

下面的环境下分别对改进前后的 Apriori 算法进行实验测试: Intel(R) Core (TM) 2 Duo CPU, 2.10 GHz 的主频, 2 GB 的内存, 320 GB 硬盘, Windows 7 操作系统, 数据库为 SQL-2005。实验的数据源为最近几个月医疗监控中心的数据, 改进 Apriori 算法和传统 Apriori 算法的运行时间的实验结果仿真图如图 2 所示。

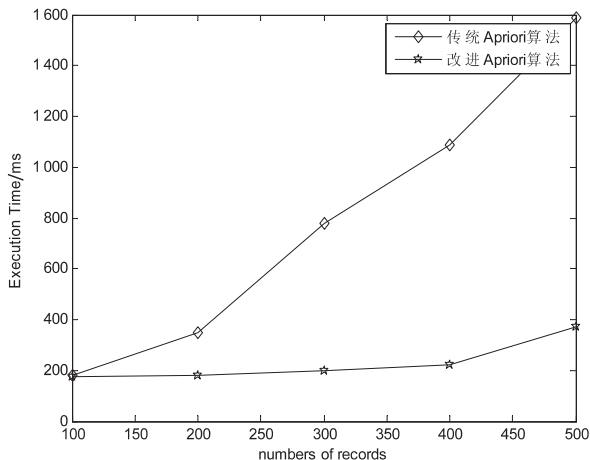


图 2 算法运行时间和记录数的关系

从图 2 可见, 改进 Apriori 算法所消耗的时间远小于传统 Apriori 算法, 并且随着记录数的增加, 改进 Apriori 算法的优势越来越明显, 这是因为改进 Apriori 算法在进行剪枝操作时, 只需要对数据库扫描一次, 大大节省了扫描时间, 从而提高了算法的运行效率。

3 改进 Apriori 算法的具体应用

将此改进的 Apriori 算法应用于医疗监护中心, 从系统登陆后进入关联规则界面。根据医护人员的需求, 输入相应的最小支持度和最小置信度, 如最小支持度为 0.05 和最小置信度为 0.8, 则点击“生成关联规则”按钮后, 系统所得的强关联规则为:

(1) 20-30 岁, 男, 37.1-37.5 度, 90-100 跳/分钟 → 中度感染—97%;

(2) 20-30 岁, 女, 36.7-37.0 度, 80-90 跳/分钟 → 轻度感染—93%;

(3) 20-30 岁, 女, 大于 38.5 度, 大于 110 跳/分钟 → 严重感染—95%;

(4) 20-30 岁, 男, 大于 38.5 度, 大于 110 跳/分钟 → 严重感染—100%。

比如第一个关联规则, 其表示 20-30 岁男性, 若体温在 37.1-37.5 度之间, 心跳在 90-100 之间, 则其中度感染的几率大概在 97% 左右。由此监护人员就可以根据所取得的强关联规则, 通过采集到的生理参数大致来判别病人的病情状况。

从以上实验可知, 改进 Apriori 算法已经能较好地应用到医疗监护系统中, 为医生及时做出诊断提供了

良好的条件。


4 结束语

文中应用了一种改进的 Apriori 算法, 该算法把整条事务当作一个属性, 减少了对不必要的事务和项目的处理, 并且随着算法的运行, 其整体优势明显提高, 从而避免了传统 Apriori 算法反复扫描数据库和产生大量候选项集的缺点^[14]。该算法充分利用了所获得的医疗数据的特点, 利用每条事务对应的属性, 只需要扫描一次事务数据库, 即可得到所需要的频繁项集, 从而大大提高了医疗监护系统的挖掘效率, 完善了当前的医疗诊断系统。

参考文献:

- [1] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001.
- [2] Agrawal R I, Nski T, Swam I A. Mining association rules between sets of items in large database[C]//Proc of the ACM SIGMOD international conference on management of data. Washington D C: ACM, 1993: 207-216.
- [3] 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.
- [4] 白东玲, 郭绍永, 王 晓, 等. 一种新型的改进 Apriori 算法研究[J]. 信息技术, 2013(7): 50-53.
- [5] 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.
- [6] 王 伟. 关联规则中的 Apriori 算法的研究与改进[D]. 青岛: 中国海洋大学, 2012.
- [7] Flenner R, Abbott M. Java P2P unleashed[M]. [s. l.]: Sams, 2003.
- [8] Han J J. Mining frequent patterns without candidate generation[C]//Proc of the 2000 ACM-SIGMOD international conference on management of data. Dallas, TX: ACM Press, 2001.
- [9] 刘兴涛, 石 冰, 解英文. 挖掘关联规则中 Apriori 算法的一种改进[J]. 山东大学学报: 理学版, 2008, 43(11): 67-71.
- [10] 龙冰莹, 陈小惠. 改进 Apriori 算法在医院监护中心的研究与应用[J]. 计算机技术与发展, 2013, 23(8): 137-140.
- [11] 尤 磊, 兰 洋, 熊 炎. 一种基于关系代数的 Apriori 优化方法[J]. 信阳师范学院学报: 自然科学版, 2010, 23(1): 156-160.
- [12] 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.
- [13] Thober M, Pendergrass J A, McDonell C D. Improving coherency of runtime integrity measurement[C]//Proc of the 3rd ACM workshop on scalable trusted computing. Alexandria, USA: ACM, 2008: 51-60.
- [14] 杨志刚, 何月顺. 基于压缩事务矩阵相乘的 Apriori 改进算法[J]. 中国新技术新产品, 2010(6): 57-58.

改进关联规则算法在医疗监控中的应用

作者：[田亚凯](#)，[陈小惠](#)，[TIAN Ya-kai](#)，[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)：2015(10)

引用本文格式：[田亚凯](#)，[陈小惠](#)，[TIAN Ya-kai](#)，[CHEN Xiao-hui](#) [改进关联规则算法在医疗监控中的应用](#)[期刊论文]-
[计算机技术与发展](#) 2015(10)