病历随访系统中数据挖掘的Apriori算法研究

王卫东[1],屈 洋[2]
(1.暨南大学 计算中心,广东 广州 510632;
2.暨南大学 医学院,广东 广州 510632)

摘 要:从存储成千上万份病历的病历随访数据库系统中挖掘出诊断所需的有价值数据,需要掌握有效的挖掘算法实现诊疗方面的数据挖掘。详细论述了数据挖掘的理念和如何根据病历随访数据库内庞大的数据群建立所需的关联规则方法。通过Apriori规则算法分析,建立起目的性极强的数据间的关联规则。通过讨论可以看出选择恰当的关联规则算法不仅可以提高在病历随访数据库中数据挖掘的效率,而且为建立某种疾病的诊疗信息库奠定了基础。

关键词:病历随访数据库系统;数据挖掘;Apriori规则算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2010)10-0004-04

Apriori Algorithm Research of Data Mining in Case History and Follow-Up System

WANG Wei-dong[1],QU Yang[2]
(1.Computer Center,Jinan University, Guangzhou 510632,China;
2.Medicine College,Jinan University, Guangzhou 510632,China)

Abstract:Picking up data needed by diagnosis from thousands and thousands of cases stored in case history and follow-up database system, hold a effective mining algorithm to realize data mining of disease treatment. In this study, discuss in detail the theory of data mining and the related practical method of constructing a rule of correlation depending on the mountain of data in the case history and follow-up database. Through Apriori algorithm analysis, constructed a highly purposive correlation rule among the data. Through discussion, see that a suitable Apriori algorithm for the construction of a correlation rule will not only increase the efficiency of data mining in case history and follow-up database, but also set foundation for building a disease treatment information database.

Key words:the case history and follow-up database system; data mining; Apriori algorithm

0 引 言

医院作为一个庞大的社会医疗保障体系,每年甚至每天都会积累海量的、不同疾病的医疗数据资料,其中大量存储的是各种病人的病历。作为数据挖掘就是要对由大量病历数据建立的随访信息库进行有价值的信息提取,如是既可对病人病情发展做后期追踪调查也可对大量同类病人的调查结果进行统计整理,进而可在早期对疾病的病理生理发展过程和预后作出准确的判断并相应地安排治疗方案,这必将为人类认识和抵御疾病产生重要的指导意义。

随访信息库中应包含病人的基本资料、医生信息、各项实验室检查、临床、影像和病理诊断、住院医嘱、诊疗操作、护理信息、病情转归等信息。如医生信息包括科室、姓名、教育程度和工作时间、职称、医疗特长等;诊断治疗信息至少包括接诊时间、治疗开始时间、主管医师情况、病程记录、治疗方案等。将以上信息纳入随访信息数据库中,是利用数据挖掘技术对病历随访数据库系统进行信息的科学分析的前提和保障。再运用适当的挖掘算法进行信息的清理,清除病患随访信息中无效的数据,提取有价值的数据。各类信息示例如表1所示。

1 关联规则的定义

在进行关联规则挖掘前先要明确这样一些有关数据的基本假设:设 $I=\{i1,i2,\cdots,im\}$ 为所有项目的集合[1],其中,每个元素 ik 称为项目(item),$k=1,2,$

收稿日期:2010-03-01;**修回日期**:2010-06-13
基金项目:教育部留学回国人员科研启动基金(教外司留[1999]363号)
作者简介:王卫东(1956-),男,山西人,教授,从事计算机的教学与软件开发应用;屈 洋,教授,从事临床医学和计算机在医学方面的应用研究。

…, m 。A 作为一个 I 中的项目的集合,将其称为项集。事务 T (transaction) 作为一个项目子集,每一个事务具有唯一的事务标识 TID。D 则作为事务 T 的数据库。它们之间的关系定义为:D 为各个特定的事务 T 的集合。事务 T 包含项集 A,且若项目集 A 中包含 k 个项目,则称其为 k 项集。仅以表 1 病历随访数据库信息表例的信息项为例说明上述定义的含义,I 作为所有病人病历的集合,即 $I = \{i1, i2, \cdots, im\}$ = { 病历编号,身份证号码、姓名、性别、年龄、身高、体重、血型、家族史、现病史、食物、药物过敏记录、经治医生姓名、疾病名称、发病时间、就诊前症状、目前症状、就诊前用药效果、现治疗方案记录、现处方记录、曾检查项目结果、病理检验结果、医生编码、医生姓名、……},设 A 是指向 I 中关于血脂高人群体特征考察项的集合即 k 项集,其中 A 暂且仅包括{ 性别、年龄、身高、体重、家族史 }5 项,也可写成 $A = \{I4, I5, I6, I7, I9\}$。$T$ 则是所有病历中关于血脂高人群体特征考察的 A 的集合,它的项目标识号必唯一,各类诊疗或医学研究所需的所有的特征事物 T 的集合则构成数据库 D。一个特定诊疗事务 T 的项目集的例子如表 2 所示。

表 1　病历随访数据库信息表例

病人信息表主要项目	诊疗信息主要项目	辅助检查主要信息项目	医生信息主要项目
病历编号	病历编号	病历编号	医生编码
身份证号码	疾病名称	疾病名称	姓名
姓名	发病时间	体温	出生年月
性别	就诊前症状	血压	性别
年龄	目前症状	心律	学历
身高	就诊前用药效果	血常规检查	职称
体重	现治疗方案记录	尿常规检查	工作年限
血型	医嘱	血生化检查	科室
家族史	外院检查结果	X 光 - 检查	科室电话
现病史	病理结果	超声波检查	移动电话
既往史	医生编码	心电图检查	医疗特长
食物、药物过敏记录	医生姓名	CT/MR 检查	备注
备注	备注	内窥镜检查	

表 2　一个特定诊疗事务 T 的项目集的例子

T 的项目标识号 TID	病历编号	关于血脂高人群体特征考察项 A 的各项记录
T0001	200801000156	I4, I5, I6, I7, I9
T0002	200810140156	I4, I5, I6, I7 (没有 I9 是因为没有血脂高的家族史)
T0003	200806090156	I4, I5, I9(没有 I7 是体重不超标)

所谓关联规则就是在上述各个集合的数据建立基础上找出所关心的那些属性之间是否存在的有趣的内在联系[2],譬如,关于血脂高的群体,总是试图找出性别、年龄、体重和家族病史这些体征对该病产生的影响。从概率统计的角度看,关联规则的研究就是讨论各集合间的逻辑蕴含关系,一个关联规则犹如 $X => Y$(意为:数据库中满足 X 中的条件的记录也必满足 Y 中的条件) 的蕴涵式。这里,$X \subseteq I$(I 集包含 X 项集),$Y \subseteq I$(同样 I 包含了 Y),并且 $X \cap Y = \varnothing$。那么若 D 中包含 X 的 $C\%$ 的事务同时也包含 Y,就称规则 XY 在事务集数据库 D 中具有可信度(Confidence) C。若 D 中有 $S\%$ 的事务包含 $X \cup Y$,那么称规则 XY 在事务集数据库 D 中具有支持度(Support) S。可信度是指 XY 蕴含的强度,它表现为规则的可信度。支持度是指模式在规则中所出现的频率。通常用下列表达式描述置信度和支持度:

Confidence($X => Y$) = $P(Y \mid X)$ X 条件下 Y 出现的概率 …… 可信度

Support($X => Y$) = $P(X \cup Y)$ X 与 Y 的并集 …… 支持度

如何理解这样的规则定义?仍以血脂高人群体特征为例,设取样人群数 1000 人,主要特征:男、年龄在 60 岁以上、体重超过 75kg,其中 60% 的人患有血脂高(即 600 名血脂高患者),在这 600 人中又有 30% 的人有血脂高家族史(即 200 人,占总人数的 20%),于是说在老年人群中六成的人患有血脂高,即可信度 Confidence = 60%,而在这些患者中有 30% 是必然的血脂高患者,而这些人占总人数的 20%,因此支持度 Support = 20%[3]。

由此可见,支持度和可信度在描述一个被成功挖掘的关联规则的可采用程度和可确认程度。于是作为一个给定的诊疗或医学研究事务集 D,挖掘关联规则[4]问题就转变为寻找支持度和可信度分别大于用户给定的最小支持度(min_support)和最小可信度(min_confidence)的关联规则的过程。

2　关联规则的 Apriori 算法

Apriori 算法始于 90 年代初在研究挖掘顾客交易数据库中项集间的关联规则时提出的算法,是目前最有影响的挖掘关联规则算法,其核心是基于频繁集(所有支持度大于最小支持度的项集称为频繁项集,简称频集)理论的递推方法[5]。它利用频繁项集性质的先验知识,使用逐层搜索的迭代方法 k - 项集,用于搜索 $(k+1)$ - 项集。首先,找出频繁 1 - 项集的集合。该集合记作 L1。L1 用于找频繁 2 - 项集的集合 L2,而 L2 用

于找 L3，如此下去，直到不能找到频繁 k - 项集。记住每个 Lk 均需要扫描数据库 D 一次。

例如：在病历随访数据库信息中对血脂高人群体特征{性别、年龄、身高、体重}进行关联规则的挖掘，第一次以 K 项集{性别}为基础找出频繁 1 - 项集的集合 L1，第二次以 L1 为基础找出{性别、年龄}的频繁 2 - 项集的集合 L2，第三次以 L2 为基础找出{性别、年龄、体重}的频繁 3 - 项集的集合 L3。再以年龄为首个单项逐次递推，实现次序为首次以 K 项集{年龄}为基础找出频繁 1 - 项集的集合 L1，再以 L1 为基础找出{年龄、体重}的频繁 2 - 项集的集合 L2。如此反复递进逐层查找，直到不能找到频繁 k - 项集为止。最终通过数据挖掘就会得到有关血脂高的 4 项特征指标间的有趣的关联，换句话说在这个方面的医学判断的数据挖掘任务结束。

通过上述关于 Apriori 算法的定义性说明和实例叙述，可以看出关联规则的挖掘可以分成两个步骤[6]：

1. 在大量事务 T 集合的数据库 D 中，根据最小的支持度，遍历所有的事务数据，用迭代法寻找高频率出现的频繁项集，这是一个建立连枝的过程。由实例也可见经典 Apriori 算法就是从一个一个的单项 A 开始记数，每次遍历完所有的事务后，建立一个对应某种疾病某个方向研究的相关项的频繁集，最终得出该疾病在指定研究方向上的所有相关项的频繁集，为下一步剪枝分析打下基础。

2. 根据最小的置信度，剪去其中置信度低于用户指定的最低置信度的频繁项集，最后保留下的频繁项集就是满足需要的关联规则——知识集。此步骤一般通过子集产生法就可以实现。

3 关联规则算法的深度研究——多层关联规则挖掘算法

对于蕴藏着海量数据的病历随访数据库系统来讲，挖掘其有用的诊疗信息是如何最大限度地发挥病历随访数据库系统作用的主要任务，进而研究病历随访数据库系统数据库的性质、数据间的关系，给出恰当的数据挖掘算法则成为能否成功实现数据挖掘的基础。

作为通用的 Apriori 算法来说，它的算法是基于产生频繁候选项集，并从候选项集中挖掘频繁闭项集。因此必然会存在如下一些缺点：

(1)重复扫描数据库多遍，如顶层项集中元素个数最多的为 N，则该算法扫描数据库至少 $N-1$ 遍；

(2)产生庞大的候选项集。这会在挖掘过程中大量开销计算机内存，严重时导致计算机运行失败；

(3)由于扫描以遍历整个数据库为前提，必增加搜索时间的开销，造成分析过程太慢，计算机工作效率过低[7]。

然而对病历随访数据库系统的数据作分析时，不难看出数据库中的相当多的项集成员 ID 具有分类层次性，如血脂高人群体特征{性别、年龄、体质指标(=体重÷身高)}就具有这类特点，譬如：性别分{男、女}，男性再分{老年组、中年组、青年组、少年组、儿童组}，老年组再分{体质指标超 23、体质指标在 18.5 到 23 之间、体质指标低于 18.5}，如此逐级、逐科目的分层势必简化 Apriori 算法[8]，更容易找到数据间的关联规则。此等方法使人由此联想到计算机理论中的一个重要组成部分——数据结构，在数据结构中计算机系统数据按节点逐层架构，为计算机运行的高效性和可靠性等诸方面提供了技术保障。将数据结构的概念移植在数据挖掘领域，称其为多层关联规则挖掘。

多层关联规则挖掘的基本方法：就病历随访数据库系统来说，由于数据的庞杂性，所以利用 Apriori 算法通过遍历在数据最细节的层次上或者需要较长的时间发现关联规则，或者很难发现关联规则。引入多层次关联规则挖掘概念后，首先将数据按节点分层，然后逐层次进行关联规则挖掘。多层关联规则可以分为同层关联规则和层间关联规则。多层关联规则的挖掘仍沿用“支持度 - 可信度”的框架。不过在支持度的使用上应考虑不同情行不同策略的原则[9]，譬如：

1) 同层间的相对独立性，以此尽可能地减少同层间信息的相互影响因子的存在，使诊疗方案的设计减少不必要的干扰。例如：{性别}的下级层次{男、女}的更进一步的分析中，男性的向下分支就应当与女性的分支延伸无关。

2) 层间使用递减的最小支持度。每个层次都有不同的最小支持度，较低层次的最小支持度相对较小。如是才能在细节处挖掘到非常有意思的关联规则，为诊疗方案的合理性或医学研究提供更多的有价值信息[10]。

仍以病历随访数据库系统的数据做血脂高病人群体特征分析时，多层次关联规则的挖掘算法显得非常明了见解，如先以{性别}建立最高层，然后建第二层{男、女}，再按男性分第三层{老年组、中年组、青年组、少年组、儿童组}，老年组再分第四层{体质指标超 23、体质指标在 18.5 到 23 之间、体质指标＜18.5}，如果此时看到在男性老年组体质指标(BMI)超出 23 的患血脂高的人数比例很高时，再细化分层次时就会得到许多意想不到的结论，譬如：在男性老年组 BMI 超出 23 基础上再增加 3 个项集{饮食习惯、作息习惯、体育

锻炼],并以此在男性老年组 BMI 超出 23 的节点下构成新的关联层,如此进行不断深入的分析,最终通过最小支持度和可信度的对比就可得出一套没有血脂高家族史的老年人群中预防血脂高发上的行之有效的医疗保健方案。如果将每一层的各个项称为节点,层与层之间的关系称为节点间的链接,显然多层次关联规则的挖掘过程与计算机领域内的数据结构理念如出一辙[11]。

4　结束语

针对病历随访数据库系统中海量的医疗数据进行数据挖掘,以此实现提供疾病诊断、管理决策分析是一种发展趋势[12]。建立符合医疗过程的数据库结构是基础,从海量的诊疗数据中挖掘数据间的关联关系则是知识性医疗数据挖掘能否正确实现的前提保证,选择快速、有效的关联规则算法是能否最终提供有价值的决策信息的关键,文中所介绍的关联规则挖掘 Apriori 算法针对医学数据所具有的多态性、不完整性、较强的时间性、复杂性、冗余性和不一致等特点的知识挖掘是成功的,能够满足病历随访数据库系统的数据挖掘需求。

参考文献:

[1] Luo Zhiping, Zhou Xinzhi, Gu Zhongbi. Time series hydrological based on data mining[J]. Computer Engineering and Applications, 2007, 43(30): 231 - 233.

[2] Agrawal R, Srikant R. Mining sequential patterns[C]// Proc. of the 11th Int Conf. on Data Engineering. Los Alamitos, CA: IEEE Computer Society, 1995: 3 - 14.

[3] Zaki M. SPADE: An efficient algorithm for mining frequent sequences[J]. Machine Learning, 2000, 42(3): 31 - 60.

[4] 季伟东. 一种 Apriori 算法的改进[J]. 计算机工程与科学, 2009(9): 68 - 70.

[5] 王　华. 医学数据挖掘中的数据预处理与 Apriori 算法改进[J]. 计算机系统应用, 2009(9): 94 - 97.

[6] 孙　明. 基于层次关联规则的日志本体事件领域关系学习[J]. 计算机应用研究, 2009(10): 3683 - 3686.

[7] 王　敏. Apriori 算法在税务系统中的应用[J]. 计算机技术与发展, 2009, 19(11): 175 - 178.

[8] 冯敬益. 基于 Web 的数据挖掘技术[J]. 中国科技信息, 2009(18): 92 - 93.

[9] 郑海波. 医疗信息的数据库管理模式[J]. 福建电脑, 2008, 24(9): 172 - 178.

[10] 邹长忠. 分布式数据库的关联规则更新算法[J]. 福州大学学报: 自然科学版, 2008, 36(5): 655 - 659.

[11] 孟凡荣. 基于云理论的煤矿安全监测数据关联规则挖掘[J]. 小型微型计算机系统, 2008(9): 1622 - 1626.

[12] 徐　刚. 数据挖掘及其在医学领域中的应用和展望[J]. 实用临床医学, 2006(11): 196 - 198.

(上接第 3 页)

完备的信息系统,很可能破坏原始数据信息系统,使得处理效果不理想,甚至出现完全与事实相悖离的重大错误。针对这类问题,在信息观下引入了粗糙模糊度度量,定义了一种新的知识熵,在此基础上,提出了一种基于信息观下粗糙模糊度的不完备信息系统属性约简算法。仿真实验结果说明了该算法的有效性和较好的时间优越性。对于算法的优化及其应用,有待进一步研究讨论。

参考文献:

[1] 覃伟荣,秦亮曦. 基于粗糙集理论的条件属性动态约简算法[J]. 计算机技术与发展, 2008, 18(8): 23 - 25.

[2] Kryszkiewicz M, Rybinski H. Finding Reducts in Composed Information Systems. Rough Sets, Fuzzy Sets and Knowledge Discovery[C]//Proceedings of the International Workshop on Rough Sets and Knowledge Discovery (RSKD'93). [s. l.]: [s. n.], 1994: 261 - 273.

[3] Xiao J M, Zhong T F. New Rough set Approach to knowledge Reduction in Decision Table[C]//proceeding of the International conference on Machine learning and Cybernetics. Shanghai, China: [s. n], 2004: 2208 - 2211.

[4] 苗夺谦,胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681 - 684.

[5] 汪小燕,杨思春. 一种基于分辨矩阵的新的属性约简算法[J]. 计算机技术与发展, 2008, 18(2): 77 - 79.

[6] 汪小燕,杨思春. 基于改进的二进制可分辨矩阵的核增量式更新方法[J]. 计算机技术与发展, 2009, 19(1): 97 - 99.

[7] Wang J, Wang J. Reduction Algorithms Based on Discernible Matrix : The ordered Attributes method[J]. Journal Computer Science Technology, 2001, 16(6): 498 - 504.

[8] 王国胤,于　洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759 - 766.

[9] 陈锦禾,沈　洁. 基于信息熵的主动学习半监督分类研究[J]. 计算机技术与发展, 2010, 20(2): 110 - 113.

[10] 纪　滨. 信息熵在粗糙集中衍生的几个概念[J]. 计算机技术与发展, 2008, 18(6): 73 - 75.

[11] 梁吉业,曲开社,徐宗本. 信息系统的属性约简[J]. 系统工程理论与实践, 2001, 21(12): 76 - 80.

[12] 付　昂,王国胤,胡　军. 基于信息熵的不完备信息系统属性约简算法[J]. 重庆邮电大学学报, 2008, 20(5): 586 - 592.