广西科学院学报
Journal of Guangxi Academy of Sciences

2009,25(4):264～266
Vol. 25,No. 4 November 2009

Apriori算法的改进及其在单病种信息挖掘中的应用*
The Application of Improved Apriori Algorithm in Data Mining of Single Sort Disease

黄肇明
HUANG Zhao-ming

(广西医科大学第一附属医院教务部,广西南宁 530022)
(Guangxi Medical University First Affiliate Hospital,Nanning,Guangxi,530022,China)

摘要:采取二维数组方法一次性将数据全部读进内存的方法改进Apriori算法,并用改进的Apriori算法对一种单病种信息进行挖掘,得出和管理指标相关的信息,这些信息对今后预防和控制疾病有帮助。

关键词:数据挖掘 关联规则 单病种管理
中图法分类号:TP311.13 **文献标识码**:A **文章编号**:1002-7378(2009)04-0264-03

Abstract: Apriori algorithm is improved by using 2-dimension array. Once data is input directly into array of main memory, then improved apriori algorithm is used to mine data of single sort disease. Information related to management indicators is obtained, which is helpful to disease control and prevention.

Key words: data minining, association rule, single sort disease

数据挖掘技术已经在商业及工业领域中得到广泛的应用,取得了显著的经济和社会效益。然而数据挖掘技术在医学领域的应用尚处于起步阶段,目前只是将其作为一种辅助工具来使用,如辅助肺癌诊断[1],对医院管理的决策支持[2],中医和临床药学中的辅助应用等。单病种管理又称临床路径,是医疗卫生机构的一组成员共同制定的一种照顾病人模式,它使病人从入院到出院按一定模式接受治疗护理,以控制诊疗质量和经费。加强和改进单病种管理模式最重要的一个步骤是找出和管理指标相关的信息。1993年Agrawal等人提出关联规则的概念,应用关联规则挖掘查找大量数据库中项集之间的关联关系[3]。Apriori算法是关联规则挖掘技术中的主要算法。目前针对经典Apriori算法的改进已有不少,如利用项集的有序特性来进行改进[3],对数据库在Ck中不包含频繁项集的事务直接删除[4],采用树作为存贮结构来免多次扫描数据库[5],利用基于FP-tree来进行改进[6]等。本文在分析Apriori算法的不足的基础上,采用二维数组方法对Apriori算法进行适当的改进,并探讨改进的Apriori算法在单病种管理中的应用。

1 改进的Apriori算法

Apriori算法是一种逐层搜索的迭代方法,它的基本思想是利用已知的$k-1$项集来生成k项集,再扫描一次数据库来判断候选频繁项目集是否是频繁项目集。该算法主要存在2点不足:(1)需多次扫描交易数据库,由L_{K-1}产生L_K的过程中,需要扫描数据库一次,如果k很大的话,则需要多次扫描数据库,另外又由于这些数据库记录非常多,所以这样势必影响算法的效率,要提高效率关键是减少数据库遍历的次数。(2)识别频繁项目集时算法采用模式匹配,效率较低。在产生C_k的过程中,要检查k-项集的$(k-1)$子集在不在L_{K-1}中,另外在1-项集很大的情况下,利用Apriori算法会产生很大的候选2-项集,再产生很大的候选3-项集,如此下去。所以算法的效率仍需提高。

基于Apriori算法的不足,采取二维数组方法一次性将数据全部读进内存,这样以后扫描数据库的

收稿日期:2009-08-22
作者简介:黄肇明(1970-),男,硕士,高级工程师,主要从事数据库技术、数据挖掘、医院信息系统开发与应用。
*广西卫生厅自筹经费科研项目(桂卫Z2007089)资助。

操作就可以在速度较快的内存中进行,避免多次扫描物理数据库,并且采取事务压缩的方法,改进算法性能。

2 改进的 Apriori 算法在单病种信息挖掘中的应用

采用改进的 Apriori 算法对2005年以来在广西医科大学第一附属医院出院的脑梗塞病人信息进行挖掘,并用 Microsoft studio 2005 c#.NET 开发工具改进算法编程。

2.1 数据预处理

为了使该数据表的数据适于用关联规则 Apriori 算法来处理,有必要对其进行一些预处理。数据预处理主要包括数据清理、数据集成、数据变换和数据归约。

2.1.1 数据清理

通过填写空缺值来平滑噪声数据,识别删除孤立点,解决不一致,最终达到数据清理目的。它主要包括重复数据处理、空缺值数据处理、噪声数据处理等。海量的医院数据总存在冗余信息,为了高效率地挖掘信息,必须处理好这些数据。对空缺值的处理,根据不同情况有不同的方法。当数量不大时常采用忽略或人工填写的方法;否则使用编写软件的方式来处理,例如采用该字段属性平均值或最有可能的值填充空缺值。

2.1.2 数据集成

将多个数据库、数据立方体或一般文件中的异构数据合并处理,然后将其存放在一个一致的数据存储中,以解决语义的模糊性。该部分主要涉及数据的冲突、数据类型的选择、不一致数据处理以及因数据可能来自多个实际系统而造成的异构数据转换等问题。数据集成并非简单的数据合并,而是数据统一规范化的复杂过程。它需要统一原始数据中的所有矛盾,如字段的同名异义、异名同义、单位不统一、字长不一致等,从而把原始数据在最低层次上加以转换、提炼和聚集,形成最初始的数据信息。病历挖掘中的原始数据从医院数据库中取出,含众多的字段。首要任务就是识别实体,把相同属性字段合并归一,如"年龄"和"出生年月"两个字段可相互导出,只保留一个即可。

依据上述思想,我们随机抽取部分病历信息(表1),通过数据预处理对其进行清理、集成、转换和规约操作,为后面挖掘关联规则,如季节与易发病因的关系、年龄段常见病、不同专业和不同性别的常见病等做好有效的铺垫。

2.1.3 数据变换

基于支持度—信任度框架理论的关联规则合适于交易类型的数据库,对于医院信息数据库中的一些连续数据,如数值数据已经不适用。为此要对病历数据项进行距离划分,依据变量的取值范围将其分成若干数据段(表2)。显然,划分的尺度要适中,尺度太小则会加大工作量,缺乏足够支持度而丢失部分有用规则;尺度太大则会失去代表性,而隐藏有义的规则。

表1 原始数据

序号	性别	年龄	民族	职业	住院天数	治疗效果	入院情况
1	男	70	汉族	退休人员	15	好转	急
2	女	45	汉族	退休人员	20	好转	一般
3	男	73	汉族	退休人员	13	好转	一般
4	男	80	汉族	离休人员	12	好转	急
5	男	79	汉族	退休人员	15	好转	一般
6	女	60	壮族	个体户	12	好转	一般
7	男	65	壮族	退休人员	12	好转	一般
8	男	65	壮族	农民	5	好转	一般
9	男	56	壮族	农民	16	好转	一般
10	男	70	壮族	教师	4	好转	一般
11	女	43	汉族	工人	34	好转	一般
12	男	47	汉族	干部	11	治愈	一般

表2 变换规则

字段名称	变换规则
性别	A1 男,A2 女
年龄	1～10 B1,11～20 B2,21～30 B3,31～40 B4,41～50 B5,51～60 B6,61～70 B7,71～80 B8,81～90 B9
民族	汉族 C1,壮族 C2,回族 C3,蒙古族 C4,苗族 C5,维吾尔族 C6,叙佬族 C7,满族 C8
职业	退休 D1,离休 D2,个体户 D3,农民 D4,干部 D5,教师 D6,工人 D7
住院天数	1～10 E1,11～20 E2,21～30 E3,31～40 E4,41～50,E5
治疗效果	治愈 H1,好转 H2,有效 H3,无效 H4
入院情况	入院情况:急 G1,一般 G2

通过对原始脑梗塞数据进行清理、集成、转换等预处理(表3),可以达到数据格式一致、数据类型相同、数据存储集中以及数据信息精练的目的。

2.2 信息挖掘结果

挖掘得出的频繁项集如表4所示。由频繁项集生成的关联规则如表5所示。

从满足最小支持度和最小置信度得出的3条规则分析以上结果,可以得出结论:(1)如果病人在入院时不是危险急诊情况,也就是说病情不是很重的情况,经过住院治疗一般能治愈。(2)如入院情况是

急的情况，可能已经比较危险，病情比较长或比较重，经过住院治疗，一般只能达到治疗有效，控制病情，较难治愈。(3)50岁以上人群，是脑梗塞易发年龄。这些信息对今后脑梗塞疾病的预防和控制有一定的帮助。

表3 预处理后的数据

序号	性别	年龄	民族	职业	住院天数	治疗效果	入院情况
1	A1	B7	C1	D1	E2	H2	G1
2	A2	B5	C1	D1	E2	H2	G2
3	A1	B8	C1	D1	E2	H2	G2
4	A1	B8	C1	D2	E2	H2	G1
5	A1	B8	C1	D1	E2	H2	G2
6	A2	B6	C2	D3	E2	H2	G2
7	A1	B7	C2	D1	E2	H2	G2
8	A1	B7	C2	D4	E1	H2	G2
9	A1	B6	C2	D4	E2	H2	G2
10	A1	B7	C2	D6	E1	H2	G2
11	A2	B5	C1	D7	E4	H2	G2
12	A1	B5	C1	D5	E1	H1	G2

表4 频繁项集

频繁项集			支持度计数
B5	G2	H1	358
B5	G2	H2	691
B7	G1	H3	310
支持度30%			

表5 生成的关联规则

关联规则	置信度(%)
G2⇒H1	69.12
G1⇒H3	78.8
B5⇒G2	65.66
挖掘最小置信度50%	

3 结束语

关联规则挖掘是近年来数据挖掘研究中一个相当活跃的领域。实践证明，没有一种挖掘算法对所有类型的数据都优于其它算法，每种相对较优的算法都有它具体的适用环境[7]。本文改进了关联规则挖掘 Apriori 算法，使其适用于医院单病种数据信息的挖掘，并且用该算法对具体的单病种信息进行挖掘，得出了一些有意义的规则，其他病种是否也存在有意义的规则，是今后研究的工作。

参考文献：

[1] 陈卉，王晓华. 数据挖掘技术在计算机辅助肺癌诊断中的应用[J]. 中国组织工程研究与临床康复，2007，11(5)：879-881，885.

[2] 柳毅，潘瑞芳，叶春明. 数据挖掘技术对医院管理的决策支持[J]. 浙江传媒学院学报，2006，5：56-68.

[3] 刘美玲，徐章艳，卢景丽，等. 利用项集有序特性改进 Apriori 算法[J]. 广西师范大学学报：自然科学版，2004，22(1)：33-37.

[4] 朱祥玉，侯德文，陈希. 对关联规则挖掘 Apriori 算法的进一步改进[J]. 信息技术与信息化，2005，6：81-83.

[5] 杨健兵. 数据挖掘中关联规则的改进算法及其实现[J]. 微计算机信息，2006，7：195-197.

[6] 冯志新，钟诚. 基于 FP-tree 的最大频繁模式挖掘算法[J]. 计算机工程，2004，30(11)：123-124.

[7] 覃亮曦，史忠植. 关联规则研究综述[J]. 广西大学学报：自然科学版，2005，30(4)：310-317.

(责任编辑：尹 闯)

国家计算机病毒中心发现木马程序新变种

最近，国家计算机病毒应急处理中心通过对互联网的监测发现一种木马程序新变种 Trojan _ Sasfis. LU。该变种会通过移动存储设备进行自我复制传播，一旦发现新的移动存储设备接入操作系统，就会在系统的每个磁盘分区的根目录下创建一个自动运行的配置文件和木马主程序文件，以达到双击盘符使变种被激活的目的。该变种还会将其自身图标伪装成“Windows Media Player”样式，诱使计算机用户点击运行。该变种运行后，会将自身复制到受感染操作系统的系统目录中并重新给文件命名，其属性设置为“系统、隐藏、只读”中的一种。该变种会创建新的浏览器 IE 进程(进程名：iexplore. ex)，并将恶意程序代码注入其中隐藏运行。该变种还会将其自身注册为受感染操作系统的系统服务而被运行，随后不断尝试与远程控制端进行链接，一旦链接成功，恶意攻击者就可以通过某些端口进行监听、任意数据包的交换、远程恶意代码指令的发送等操作，最终导致用户计算机系统被远程控制，系统中的文件被恶意删除，系统自动远程下载上传恶意程序文件等。计算机用户要注意及时升级系统中的防病毒软件，打开系统中防病毒软件的“系统监控”功能，从注册表、系统进程、内存、网络等多方面对各种操作进行主动防御。

(据科学网)