数据挖掘中关联规则算法及其在医学数据中的应用

柳青

广州中山大学肿瘤防治中心,510060, 电话:87343050, Email:liuqing5@mail.sysu.edu.cn

摘要：本文介绍了关联规则挖掘算法的基本概念、思想、步骤及当今改进的一些算法，分析了医疗数据的主要特点，同时描述了关联规则挖掘在医学应用中的发展情况及存在的问题。

关键词：数据挖掘　关联规则　医学信息

关联规则挖掘（Association Rules Mining）是为了在数据库中发现关联关系，它是数据挖掘（Data mining，DM）最先研究的问题之一，也是数据挖掘的一个主要研究方向[1]。关联规则可直观的表达数据中项集（变量的各种取值）间的联系。这种联系并不是基于某种特定的分布，依靠数据在特定模型中的多次迭代拟和而来，而是根据项集在数据资料中出现的概率来构建。因而，这种方法有异于传统的统计学方法，其优势在于结果明确，容易解释。在实际应用中，当变量类型比较复杂，变量取值的分布不定并难于转换，或者各变量不独立，不能满足传统统计学方法的要求时，通过关联规则的挖掘，可以得到数据中隐含于变量取值中的信息。

1 关联规则的基本原理

1.1 关联规则的基本概念

关联规则挖掘的目的是找出数据库中不同数据项集之间隐藏的关联关系。在挖掘关联规则时首先要把数据库资料中的各种事件作为数据项，多个数据项组成某一特定事物的项集。如在医学数据库中，对于确诊发病这一事件，各种症状、体征及需要研究的危险因素就构成了它的数据项集。设I={i1,i2,..,im}是全体数据项集，其中ik(k=1,2,…,m)是各数据项。设任务相关的数据D是全体事务集，其中每个事务T是项集，使得$T \subseteq I$。设A是一个项集，且$A \subset T$。则项集A的支持数（suppor count）为该项集在事务D中出现的次数。支持数与D中事务总数的比值为项集A的支持

度（support），即A在D中出现的概率。如果给定一个最小支持度minsup，项集A的支持度>minsup，则称这个项集为大项集或频繁项集（frequent itemset）。关联规则的逻辑蕴涵为：$A \Rightarrow B$，$A \subset I$，$B \subset I$，且$A \cap B = \varnothing$。如此定义的关联规则具有如下两个重要的属性：①支持度（support）：P(A∪B)，即A和B这两个项集在事务集D中同时出现的概率。②置信度（confidence）：P(B | A)，即在出现项集A的事务集D中，项集B也同时出现的概率。同时满足最小支持度阈值和最小置信度阈值的规则称为强规则，即关联规则[2]。给定一个事务集D，挖掘关联规则的问题就是产生支持度和置信度分别大于用户给定的最小支持度和最小可信度的关联规则，也就是产生强规则的问题[3]。在很多情况下，只靠支持度和置信度还不能找到有用的关联规则，这时需要利用其他的特征对发现的规则的价值进行评估。对于关联规则$A \Rightarrow B$，项集B的支持度称为规则的期望置信度[4, 5]（expected confidence），即在没有任何条件影响下B出现的概率：P(B)。置信度与期望置信度的比值称为作用度（Lift），它描述的是A的出现对B的影响程度，反映的是关联规则的有效性。

1.2 关联规则分类

可以从多个角度对关联规则进行分类[6]

（1）基于规则中处理的变量类型，关联规则可以分为布尔型和数值型。布尔型关联规则处理的数据都是离散的、分类化的，它显示了这些变量间的关系。数值型关联规则处理的变量包含有数量信息（数值型变量），它表示的是属性值之间的关联关系。在挖掘数值型关联规则时要先对数值型变量进行离散化处理，再对处理后的数据进行挖掘。

（2）基于规则中的抽象层次，可以分为单层关联规则和多层关联规则。单层关联规则中，所有变量都不考虑现实中多个层次的区分，多层关联规则就能够体现实际生活中概念的层次性。

（3）基于规则中涉及到的数据的维数，可分为单维关联规则和多维关联规则。单维关联规则只涉及数据表中的单个维（字段）间的关系，多维关联规则表示的是多个维之间的关系。根据是否允许同一字段在规则中重复出现，多维关联规则又可以分为维间关联规则和混合关联规则。

（4）通过对关联规则进行一定的约束和限制，可以生成某些具有针对性的特殊类型的关联规则，这样得到的规则通常是实际工作中人们最感兴趣，也是对实践活动最具指导意义的规则。

1.3 关联规则挖掘算法

关联规则的挖掘一般可分成两个步骤：（1）找出所有支持度大于等于最小支持度阈值的频繁项集；（2）由频繁模式生成满足置信度阈值的关联规则。第一步的工作是相当费时的，而第二步在第一步的基础上很容易实现，因此关联规则挖掘算法的性能主要由第一步决定[7]。

Apriori算法是关联规则挖掘的基本算法。针对单维、单层、布尔型关联规则挖掘的Apriori算法[8, 9]的核心是候选项集的生成。它基于以下性质（Apriori性质）：如果一个项集是频繁项集，那么它的所有子集都是频繁项集。反之，如果一个项集的某个子集不是频繁项集，那么这个项集也不是频繁项集。算法的第一步是找出所有频繁项集；第二步再由频繁项集产生强关联规则。首先产生一阶频繁项集L_1，然后是二阶频繁项集L_2，直到有某一阶的频繁项集L_R为空，这时算法停止。这里在第k次循环中，过程先产生候选k-项集的集合C_k，C_k中的每一个项集是对两个只有一个项不同的属于L_{k-1}的频繁项集做一个(k-2)-连接来产生的。C_k中的项集是用来产生频繁项集的候选集，最终的频繁项集L_k必须是C_k的一个子集。最后再由前述定义判断那些频繁项集的关联规则。

目前国内外专家学者针对不同关联规则类型的挖掘和数据含量不同的关系型数据库提出多种关联规则挖掘算法，并在减少数据读取和内存占用的目标下对Apriori算法进行了改进。主要有： 采用杂凑技术改进候选集生成过程的DHP（Direct Hashing and Pruning）算法；采用分而治之的思想来解决内存不足问题的分块挖掘算法（Partition）；抽样算法（Sampling）；动态项集计数算法DIC（Dynamic Itemset Counting）[10, 11]等。由于数据挖掘的许多方法主要面向成熟的商业数据库，如何处理一般数据库中的类别和数值类型数据，尤其是医学数据库中的复杂数据，仍有待进一步研究。因此，目前尚未有上述改进方法应用于医学数据库的相关报道。

2 关联规则在医学中的应用

在关联规则挖掘算法的应用方面，一些研究者将其应用在医疗数据库中，以考察该算法在处理实际医学问题时采取的策略和存在的问题。文献[20]中给出了一个在关于心脏病的医疗数据集（表1）中Apriori算法应用的尝试。研究者的目的是通过此数据集，考察心脏病患者现有记录指

标中隐藏的关系。研究者需要知道的是数据集中各指标间的事先未知的关系，而非仅仅是各变量间的线性依存关系，变量取值之间的相互影响有可能会是我们在专业和常识上尚未阐明的一些原因。如果能直接利用数据的信息，通过数据集本身记录的各种事件发生的概率来确定变量间的关联规则，那么就绕开了先入为主的专业思维，获得客观的结论或提示。

2.1 数据转换

表1　原始医疗数据表

ID	Gender	Age	Smoker	Operation	Prognosis	LAD%	RCA%
001	F	73	Y	intervention	relief	85	60
002	M	68	Y	intervention	relief	60	100
003	M	43	N	nonintervention	recovery	75	45
004	M	59	N	intervention	ineffective	80	99
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

此数据集包含8个变量（属性），有类别型也有数值型。分别为病案号（ID），性别（Gender），年龄（Age），是否吸烟（Smoker），手术方式（Operation），冠状动脉左前降支异常率（LAD%），右冠状动脉异常率（RCA%），预后效果（Prognosis）。病案数为425例。在进行关联规则挖掘前，需要将各变量的取值转化为一系列的整数或代码，映射为事务的项。此例各种变量的取值范围相对较小。在详细的医学记录数据中，映射的项集可包含十数项。数值型变量（包括图像资料和时间资料）则采用划分区间的方式转换为整数或代码。病案号不是分析变量，所以可以去除。为方便关联规则的表述，原来的变量名也转换为代码。表2为各变量取值转换的字符映射表，其中各数值变量的分类区间依照专业知识进行划分。表3为转换后的医疗数据表。

表2　映射表

原变量取值水平	代码	原变量取值水平	代码	原变量取值水平	代码
M	S1	Operation=intervention	S7	50%≦LAD%＜70%	S13

F	S2	Operation=nonintervention	S8	LAD%≧70%	S14
Age < 60	S3	Prognosis=relief	S9	RCA% < 50%	S15
Age≧60	S4	Prognosis=recovery	S10	RCA%≧50%	S16
Smoker=Y	S5	Prognosis=ineffective	S11		
Smoker=N	S6	LAD% < 50%	S12		

表 3　映射为项的原医疗数据集

ID	A1	A2	A3	A4	A5	A6	A7
001	S1	S4	S5	S7	S9	S14	S16
002	S2	S4	S5	S7	S9	S13	S16
003	S2	S3	S6	S8	S10	S14	S15
004	S2	S3	S6	S7	S11	S14	S16
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.2 产生频繁项集

表 3 中的 S1,S2,…,S16 即事务中的项集。根据 Apriori 算法，从专业知识方面考虑，设置最小支持度为 20%，最小置信度为 75%，以期获得较高关联程度。由此会产生满足最小支持度的若干频繁 1-项集。将各频繁 1-项集合并可得满足最小支持度的频繁 2-项集。以此类推，当产生的频繁 4-项集为空时，即没有满足最小支持度的 4-项集，算法停止。

2.3 建立关联规则

根据关联规则建立的原则从获得的各阶频繁项集中产生关联规则。如频繁 3-项集{S3，S7，S9}的支持度为 73%，S3 ，S7 的支持度分别为 66% ，58%，故可计算得 S3 S7⇒S9 的置信度为 90% 。 此处 S3，S7 位于规则左侧而 S9 位于规则右侧，符合最小置信度的要求,故关联规则成立。依上述原则可建立如表 4 的关联规则表。

表 4　关联规则表

可形成的关联规则	置信度（%）	关联规则是否成立
S3 S6 ⇒ S9	90	是

S13	⇒	S16	80	是
S4	⇒	S10	50	否
	⋮		⋮	⋮

2.4 关联规则的评价和解释

下面沿用支持度-置信度框架对结果进行解释。如表 4 所示第一个关联规则，年龄小于 60 且采用介入手术治疗的患者有 90%的可能出现手术后症状的缓解。这在医学上属于有趣的关联规则，它描述了手术有效的适应患者的基本特征，在临床上有一定指导意义。而下一个成立的规则为冠状动脉左前降支异常率 LAD%在 50%和 70%之间的患者有 80%的可能出现右冠状动脉异常率大于 50%，这一规则虽然置信度大于预先设定的 75%最小置信度，但显然在医学专业中是无趣的。因为医生从专业角度就能够解释身体病变部位之间存在有这样的关联，而且这样的规则左右两侧均为疾病的症状，对临床工作指导意义不大。此文献中未引入客观兴趣度来筛选和评价规则，因而会产生无实际意义的规则。

3 关联规则在应用中出现的问题

3.1 有趣关联规则的界定和衡量[12]

在实际医学工作中，我们希望通过医学数据得到的关联规则能对实践有指导意义，称之为有趣的规则。这样的规则才能完备描述所研究问题特征或预测问题的发展情况。一种普遍的做法是在关联规则生成前预先设定一个作用度，即前述置信度与期望置信度的比值。规则的作用度高于设定值时认为其有趣。文献中的应用实例采用的是传统的作用度-置信度框架对结果进行评价，并通过专业知识来选择有趣的规则，缺乏客观的度量，也加大研究人员的工作量。目前研究者多主张使用兴趣度来衡量规则的有效性[13]。在进一步的研究中，可以尝试将兴趣度和置信度、支持度一起作为生成关联规则的测试条件，综合衡量规则的有趣性[14]。

3.2 项的位置

Apriori算法通过多阶项集间的多次连接产生满足最小支持度和置信度的关联规则时，各频繁项出现的位置不同会造成不同的支持度，位于规则的左侧或右侧的项是由各自本身的支持度和项集的支持度以及规则的置信度联合决定的。因而在挖掘结果出现前，所研究事务的项集出现在

规则的哪一侧是研究者未知的。这样在大量事务构成的数据集中就会产生许多与研究问题无关的关联规则，这些规则满足最小支持度，置信度，也会满足所设置的作用度，而规则中的项的左右位置很可能是混乱的，我们不能依据这样的规则回答诸如“哪些事件导致哪些结局？”之类的实际问题。而且在产生这样的规则时，会对数据库进行重复的扫描，加大计算系统的负荷[15, 16]。因此，在对医学数据集进行关联规则挖掘时，应该通过设定规则的限定条件对规则进行筛选，规定一些项出现在规则的某侧。这些设定应该从医学专业知识的角度考虑，类似统计学中的模型，将研究因素置于规则左侧，研究结局置于右侧。同时，参考以往研究的经验和本次研究的目的，这样就可以即减少产生规则的数量又能快速发现有趣的规则。

3.3 关联规模的控制[17-19]

在复杂的医学数据库中，会产生大量的关联规则。即使在控制上述两个问题后，关联规则的规模也是相当庞大的，其中不乏许多重复的规则。如当产生规则 X1⇒Y，X2⇒Y，X1⊂X2 时，规则 X1⇒Y 比 X2⇒Y 简单，并会有更高的支持度，结果中就应该去掉后者以缩减规则的规模。

3.4 关联规则挖掘在医学中的应用

国外最早是在商业数据库中挖掘关联规则，并用所得的规则指导实际商业运作，取得理想的效果，由此检验了Apriori算法及其挖掘所得关联规则的可靠性。近年来，在医学信息领域应用Apriori算法挖掘关联规则的研究已广泛开展。国外文献报道了研究者采用此算法挖掘基因表达数据库中的关联规则[20, 21]，也有人尝试将此算法应用于门诊病例数据的分析，利用产生的多个症状与疾病之间关联规则合理分流病人，做出相应的医疗决策，以减少错误就诊的发生[22-24]。不少专家提出应用此算法的基本思想，结合统计学方法研究疾病的并发症及提取更有价值的隐藏的医疗信息[25-28]。在此类研究中，数据中的连续型变量经过转换成为分类或等级变量，从统计学角度而言是损失了样本信息的，因而挖掘所得的关联规则尚需进一步验证。

参考文献

1.Jiawei Han ,Micheline Kamber ， 范明,孟小峰等译. Data Mining:Concepts and Techniques. 1版. 北京:机械工业出版社,2001.

2.陈安，陈宁 等. 数据挖掘技术及应用. 北京:科学出版社,2006. 50-110.

3.Agrawal R，Imielinski T,Swami A. Mining association rules between sets of items in large databeses. Proceedings of the ACM SIGMOD Conference on MAnagement of Data.

4.Konias S, Chouvarda I, Vlahavas I, et al. A novel approach for incremental uncertainty rule generation from databases with missing values handling: application to dynamic medical databases. Med Inform Internet Med, 2005,30(3):211-25.

5.Mitsuhashi N, Hishigaki H, Takagi T. Applying Association Rule Discovery Algorithm to Multipoint Linkage Analysis. Genome Inform Ser Workshop Genome Inform, 1997,8:100-109.

6.夏火松. 数据仓库与数据挖掘技术. 北京:科学出版社,2004. 143-160.

7.秦亮曦, 史忠植. 关联规则研究综述. 广西大学学报(自然科学版), 2005(04).

8.Agrawal R，Skirant R. Fast Algorithms for Mining Association Rules in Large Databases. California: IBM Almaden Research Center, 1994.

9.谈恒贵, 王文杰, 李克双. 频繁项集挖掘算法综述. 计算机仿真, 2005(11).

10.Wolff R, Schuster A. Association rule mining in peer-to-peer systems. IEEE Trans Syst Man Cybern B Cybern, 2004,34(6):2426-38.

11.姚卫新, 黄丽华. 智能数据分析在医学领域的应用综述. 计算机工程, 2004(07).

12.Bodenreider O, Aubry M, Burgun A. Non-lexical approaches to identifying associative relations in the gene ontology. Pac Symp Biocomput, 2005:91-102.

13.周皓峰, 朱扬勇, 施伯乐. 一个基于兴趣度的关联规则的采掘算法. 计算机研究与发展, 2002(4):450-457.

14.Horng JT, Huang HD, Huang SL, et al. Mining putative regulatory elements in promoter regions of Saccharomyces cerevisiae. In Silico Biol, 2002,2(3):263-73.

15.Artamonova II, Frishman G, Gelfand MS, et al. Mining sequence annotation databanks

for association patterns. Bioinformatics, 2005,21(Suppl_3):iii49-iii57.

16.Laxminarayan P, Alvarez SA, Ruiz C, et al. Mining statistically significant associations for exploratory analysis of human sleep data. IEEE Trans Inf Technol Biomed, 2006,10(3):440-50.

17.Becquet C, Blachon S, Jeudy B, et al. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. Genome Biol, 2002,3(12):RESEARCH0067.

18.Hristovski D, Stare J, Peterlin B, et al. Supporting discovery in medicine by association rule mining in Medline and UMLS. Medinfo, 2001,10(Pt 2):1344-8.

19.Georgii E, Richter L, Ruckert U, et al. Analyzing microarray data using quantitative association rules. Bioinformatics, 2005,21 Suppl 2:ii123-ii129.

20.Creighton C, Hanash S. Mining gene expression databases for association rules. Bioinformatics, 2003,19(1):79-86.

21.Carmona-Saez P, Chagoyen M, Rodriguez A, et al. Integrated analysis of gene expression by Association Rules Discovery. BMC Bioinformatics, 2006,7:54.

22.Mullins IM, Siadaty MS, Lyman J, et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Comput Biol Med, 2006,36(12):1351-77.

23.Doddi S, Marathe A, Ravi SS, et al. Discovery of association rules in medical data. Med Inform Internet Med, 2001,26(1):25-33.

24.曲哲, 林国庆, 余奎. 数据挖掘技术在医学影像中的应用. 医疗设备信息, 2004(06).