医学数据挖掘中的数据预处理与 Apriori 算法改进①

Data Preprocess and Algorithm Apriori Improvement of the Medical Data Mining

王　华　(安徽医科大学 计算机中心 安徽 合肥 230032)

胡学钢　(合肥工业大学 计算机与信息学院 安徽 合肥 230009)

摘　要： 医学数据挖掘是提高医院信息管理水平，为疾病的诊断和治疗提供科学的、准确的决策的需要。分析了医疗数据的特点，并以慢肺阻疾病诊断的数据集为例，阐述了把医疗数据转换成事务数据格式的方法，描述了关联规则挖掘在医疗数据分析中应用所遇到的难题，针对这些难题给出了一些算法的改进措施，并用数据进行测试。结果表明，此算法优于原来算法，它可以减少产生的规则的数量，从而能快速发现有趣的医疗关联规则。

关键词： 数据挖掘 算法 关联规则 数据预处理 频繁项集

1　引言

计算机信息管理系统在医疗机构的广泛应用,同时电子病历的大量应用,医疗设备的数字化，促进了医学信息的数字化，这些宝贵的医学信息资源对于疾病的诊断、治疗和医学研究都是非常有价值的。然而目前大多数医院缺乏数据的集成和分析，更谈不上医学决策和知识的自动获取，大量的数据被描述为“数据丰富，但信息贫乏”，如何利用数据挖掘[1]技术从这些海量的数据中找出有价值的知识和规则，挖掘数据中所隐藏的规律利用这些来为疾病的诊断和治疗提供科学的决策总结各种医治方案的疗效，更好地为医院的决策管理、医疗、科研和教学服务，已成为一个非常重要的研究课题[2]以及对海量的数据进行自动获取，都需要新的技术来实现。

2　医学信息的特点

医学信息[3]包括纯数据(如体征参数,化验结果)，信号(如肌电信号,脑电信号等)，图像(如 B 超,EF 等医学成像设备的检测结果)，文字(如病人的身份记录,症状描述,检测和诊断结果的文字表述),以及用于科普、咨询的动画、语音和视频信息。医学信息的多模式特性是它区分其它领域数据的最显著特征，这种多属性模式并存加大了医学数据挖掘的难度。

病例和病案的有限性使医学数据库不可能对任何一种疾病信息都能全面地反映，表现为医学信息的不完全性。许多医学信息的表达本身就具有不确定和模糊性的特点。疾病信息所体现出的客观不完整和描述疾病的主观不确切，形成了医学信息的不完整性。医学数据库是一个庞大的数据资源，每天都会有大量相同的或部分相同的信息存储在其中。比如，对于某些疾病，病人所表现的症状、化验的结果、采取的治疗措施都可能完全一样。医学信息的所具有的这些特点，使得医学数据挖掘与普通的数据挖掘存在较大的差异，决定了医学数据挖掘的特殊性。医学数据库中含有海量的原始信息，其中包括大量模糊的、不完整的、带有噪声的信息。在数据挖掘之前，必须对这些信息

① 基金项目:安徽省自然科学基金(050420207);安徽医科大学科研项目(2008kjzj05,2006kj28);安徽省计生委基金(07028)
收稿时间:2009-01-10

进行清理和过滤，确保数据一致性，将其变成适合挖掘的形式。

数据挖掘技术在分析医学数据的研究中被广泛地采用并取得了许多有价值的成果。数据挖掘的方法也有很多，其中应用最广泛的方法之一就是发现数据中的关联规则[4]。目前，有很多专家致力于探索在医疗数据中发现关联规则的方法，并已证明了一些运用于医疗专家系统的有效的规则[5]，这些规则能够辅助疾病诊断，发现了一些新的医疗规则，丰富了专家诊断系统。

3　医学信息的预处理

数据在关联规则挖掘前要经过选择和转换以适合挖掘的形式，那么根据医疗数据的特点，医疗数据集的关联规则挖掘要转换成包含项的事务数据格式。

3.1 属性的选择和噪声数据的处理

这里只讨论类别类型和数值类型的属性，这样的话问题就会得到简化，设 $A_1,A_2,\cdots,A_p$ 是所有属性，设 $R=\{r_1,r_2,\cdots,r_n\}$是有 n 个元组的一个关系，这些元组的值来自于 $A_1\times A_2\times,\cdots,\times A_p$，这里 A_i 是类别类型或者是数值类型，数据集的大小是 n，它的维度是 p。以表 1 原始医疗数据表为例：数据集的大小是 5，共有 5 个元组：$r_1,r_2,\cdots,r_5$，维度即属性是 5：$A_1\times A_2\times,\cdots,\times A_5$(性别×年龄×吸烟×肺功能×咳嗽)。对于病历中象姓名，地址等这样的属性没任何意义，必须删除。根据医疗知识进行数据选择，以病历为记录，症状为条件属性，疾病为决策属性。对于噪声数据可以采用分箱、聚类、人机结合等方法去除。

3.2 空缺值的处理

数据挖掘算法只适用于无空缺值的的数据库，填充空缺值的方法有：去除具有空缺值的样本，或用平均值填充。

3.3 连续属性的离散化

离散化问题可以定义：A 是数据的所有属性中的一个连续属性，A 的值域为区间[a,b]，[a,b]上的 K 个划分 πk 是由 k 个区间构成的集合。$\pi k=\{[a_0,a_1],[a_1,a_2],\cdots,[a_{k-1},a_k]\}$，式中 $a_0=a$，$a_k=b$。离散化就是在[a,b]上产生的 π_k 过程，我们一次处理一个属性，把它映射成一系列连续的整数，我们把这些整数看作项。(1)类别属性：它的值是分类的，一个属性有多个不同的值，通过把每个不同的类别类型属性值和一个整数联系，用这个整数代替，这样可以很容易把类别类型映射成为我们所需要的项。(2)数值类型：这是属性的第二个重要类型。为了运用数值类型属性数据来发现它们中的关联规则，必须把它分成不同的区间，叫特征区间，这些区间形成一个索引，大大方便了发现关联规则，等深的分割方法可以最小化信息的丢失，在所用的例子里，采用医师已经分割好的区间，因为这与医疗专业知识是有很大关系的，往往医生对划分区间有专业经验，有一些已经约定俗成的分割点一直被采用，例如：划分体重为偏瘦、正常、超重，分割点是人们都知道的，它已经成为标准来衡量一个成年病人是否超重；还有年龄的划分，法律规定 18 岁为未成年与成年的分割点，还有中年、老年都有分割点。因此，在文中例子的数据属性已经经过参考，划分了一定的准确的区间。

经过前几步处理，数据的形式有数值与布尔型，必须全部转为布尔型，对于离散型数据，每个不同的值用一个整数与之对应。对于数据是连续的，根据每个不同的区间用一个整数与之对应。例如就慢肺阻说，50-60,70-80 的老人发病率高，因此[50，60],[70，80]是两个与决策属性有较高相关度的特征区间。肺功能是记录病人是否有慢肺阻的重要指标，所以被分成三个区间。下面的例子是关于慢肺阻疾病的简单数据集，有 5 个属性(p=5)，5 个病人(n=5)，表 1 是原始医疗数据表，表 2、表 3 把所有属性映射成项的映射表，表 4 是转换后的事务数据表。

表 1　原始医疗数据表

	性别	年龄	吸烟	肺功能	咳嗽
R_1	F	52	Y	严重	Y
R_2	M	55	N	严重	N
R_3	F	76	Y	严重	Y
R_4	M	73	Y	正常	Y
R_5	M	60	N	一般	N

表 2　映射表 1

M	F	age<70	70≤age	Smoke=Y	Smoke=N	肺功能=正常
1	2	3	4	5	6	7

表 3　映射表 2

肺功能=一般	肺功能=严重	咳嗽=N	咳嗽=Y
8	9	10	11

表 4　事务数据表

A1	A2	A3	A4	A5
2	3	5	9	11
1	3	6	9	10
1	4	5	9	11
1	4	5	7	11
1	3	6	8	10

4　关联规则的基本概念

一个事务数据库中的关联规则挖掘可以描述如下：

设 I={i_1，i_2，…，i_m}是一个项集集合，事务数据库 D={t_1，t_2，…，t_n}是由一系列具有唯一标识 TID 的事务成，每个事务 t_i(i=1，2，…，n)都对应 I 上的一个子集。设 I1⊆I，项集集 I1 在 D 上的支持度(support)是包含 I1 的事务在 D 中所占的百分比，即 support(I1)= ||{t∈D|I1⊆t}||/||D||。一个定义在 I 和 D 上的形如 I1==>I2 的关联规则通过满足一定的可信度(confidence)来给出。所谓规则的可信度是指包含 I1 和 I2 的事务数与包含 I1 的事务数之比，即 Confidence(I1⇒I2)=support(I1∪I2)/support(I1)，其中 I1，I2⊆I，I1∩I2=Φ。

给定一个事务数据库，关联规则挖掘问题就是通过用户指定最小支持度(minsupport)和最小可信度(minconfidence)来寻找合适关联规则的过程。

Agrawal 等在 1993 年设计了一个基本算法 apriori[6]，提出了挖掘关联规则的一个重要方法，这是一个基于两阶段频集思想的方法，将关联规则挖掘算法的设计可以分解为两个子问题：

(1) 发现频繁项集集

通过用户给定的 minsupport，寻找所有频繁项集(Frequent Itemset)，即满足 support 不小于 minsupport 的项集集。事实上，这些频繁项集集可能具有包含关系。一般地，我们只关心那些不被其它频繁项集集所包含的所谓频繁大项集(Frequent Large Itemset)的集合。这些频繁大项集是形成关联规则基础。

(2) 生成关联规则

通过用户给定的 minconfidence，在每个最大频繁项集项集集中，寻找 confidence 不小于 minconfidence 的关联规则。

5　医学数据挖掘关联规则的问题与解决

在 Apriori 算法中，最小支持度和最小置信度是关键的阈值，也是算法的限制条件，在处理医疗数据时这两个条件是不够的，如人们通常是想把多个因素同疾病联系起来，而不是颠倒过来，如上例但挖掘时常出现恰恰相反的情况，这些规则都不是有趣的。因此如果要快速发现有趣的医疗规则，Apriori 算法还存在一些困难[7]。

5.1 医学数据挖掘关联规则的问题

① 项出现位置问题：所给的有趣规则出现在关联中的项必须是一个频繁项集，但项的出现会推出许多无趣的规则。换句话说，支持度仍要删去无趣的规则，但置信度不足以删除无趣的规则，因为有许多具有高置信度的规则包含在左部或右部禁用的项，因此有些项需要被限制出现在规则的特定部分，限定项仅出现在左部，或仅出现在右部，或在两边都可以出现，这样既可以减少产生规则的数量有可以快速发现有趣的规则。

② 关联规模：包含许多项的关联和规则很难解释并且潜在地产生大量的规则，而且，它们会降低用户交互作用的过程。因此应当为关联规则设一个缺省的阈值，被发现的关联的规模较大是一个实际算法性能瓶颈。

5.2 算法改进

① 项出现位置的约束，添加项出现位置的约束标记。设 I={i_1,i_2,…,i_n}是一给定项的集合，由 A_1,A_2,…,A_p通过映像得到，将其作为处理对象，设C={c_1, c_2,…, c_p}是对应的每一个属性的出现位置约束标记。每一个约束标记可能有三个值：c_i 值为 1：项仅出现在一个规则的左部；c_i 值为 2：项仅出现在规则的右部；c_i 值为 o：项能够出现在一个规则的左部或右部。

② 最大支持度选取的约束。在医学数据关联规则挖掘时，经常处理的是长模式或富模式，在处理这些高维数据时会大量保留一类无关规则，这类规则中的

一些项具有非常高的支持度。 因此，为了有效消除这类无关规则对算法性能的影响，算法中加入决策属性中属性值的最大支持度 maxsup 约束。

③ 最小支持度选取的约束。最小支持度直接决定了满足要求的项集、模式集合和规则集合的大小。 最小支持度的值选取得较高则会消去大量不满足最小支持度的项集、模式和规则，但同时存在两个主要问题。其一，许多有用的，甚至是有重要意义的信息的支持度很低，因此若最小支持度选择得过大会使这些模式或规则被消去。频繁集缺少某个类别的属性值。其二为了获得最佳挖掘结果需多次运行挖掘算法，并通过逐次调低最小支持度来达到目的， 因此，需预先选择一个适当小的最小支持度。 若输入的最小支持度 minsup 大于决策属性中属性值频繁项的最小的支持度，则以决策属性中属性值频繁项的最小支持度作为 minsup。

5.3 实验分析

对于表 4 进行规则挖掘程序的参数设置：最小支持度=20%，最大支持度=40%，最小置信度=80%，项(7，8，9，10，11)被限制出现在规则的右部，有关年龄、性别、吸烟习惯的项(2，3，4，5，6)被限制出现在规则的左部。

利用改进算法进行实验可以的到两个(9,11)出现在右部的所有规则，无用规则减少，例如得到 4，5，1=>11(40%,100%)和 2，3，5=>9(20%,100%)

从实验结果也可以看到，如果设 X=>Y 是一个合法的规则，则添加右部一项限制条件以后，会有 $O(2^{|x|+|y|-1})$个无用规则被删除。

6 结语

本文分析了医学数据的特点，利用一个医疗数据集进行了数据的预处理，针对算法 Apriori 在挖掘医学关联规则的挖掘时出现的问题提出了其改进，并用实验验证了它的有效性。因此下一步的工作需要用大型医疗数据集进行测试，从而进一步改进算法。

参考文献

1 Fayyad U, et al.Knowledge discovery and data mining towards a unifying framework. KDD'96 Proc. 2nd Int. Conf. on Knowledge Discovery & Data Mining. AAAI Press, 1996:3－8.

2 王华,江启成,胡学钢.数据挖掘在医学上的应用.安徽医药, 2008,13(8):746－748.

3 屈景辉,廖琪梅,许卫中.医学信息数据库的建立与数据挖掘.第四军医大学学报, 2001,22(1):88－89.

4 包剑.关联规则挖掘研究.计算机系统应用, 2005,14(11):56－58.

5 王华,胡学钢.基于关联规则的数据挖掘在临床上的应用.安徽大学学报, 2006,30(2):21－24.

6 Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases Proceedings of the ACM SIGMOD International Conference on Management of Data. Washington, USA: [s. n.], ACM Press, 1993－05.

7 李虹,蔡之华.关联规则在医疗数据分析中的应用.微机发展, 2003,13(6):94－97.