第 34 卷　第 23 期　　计　算　机　工　程　　2008 年 12 月
Vol.34　No.23　　Computer Engineering　　December 2008
·软件技术与数据库·　文章编号：1000—3428(2008)23—0034—02　文献标识码：A　中图分类号：TP311

基于中医疗效评价的交互式数据挖掘框架

印　莹，张　斌，赵宇海，张晓红，张明卫
(东北大学信息科学与工程学院，沈阳　110004)

摘　要：设计并实现了基于小儿肺炎中医疗效评价的交互式数据挖掘框架。该框架采用数据挖掘、数理统计和逻辑分析相结合的方法，通过回顾性和前瞻性多角度的验证与比较研究，揭示各证和所属症状之间的关联性，不仅突破了传统的疗效评价方法的限制，而且优化了疗效规范，建立了客观的、人机交互可度量的病证结合的疗效评价体系。

关键词：疗效评价；数据挖掘；关联规则；伴随关联规则；量表

Interactive Data Mining Framework Based on Chinese Medicine Therapeutic Evaluation

YIN Ying, ZHANG Bin, ZHAO Yu-hai, ZHANG Xiao-hong, ZHANG Ming-wei
(School of Information Science and Engineering, Northeast University, Shenyang 110004)

【Abstract】The paper designs and implements interactive data mining framework based on therapeutic evaluation of childrens' pneumonia. Combining data mining, mathematical statistic with logic analysis framework, the discovery the associations between zheng with symptoms by retrospective analysis and forward looking study. The paper not only breaks through the limitation of traditional therapeutic evaluation method, but also optimizes therapeutic criteria by building the objective, interactive, metrizable therapeutic evaluation system.

【Key words】therapeutic evaluation; data mining, association rules; co-occur association rules; scale

1　概述

数据挖掘[1]作为一门新兴的交叉学科，其基本目标就是从大量的数据中提取隐藏的、潜在的知识和信息。该技术自 20 世纪末提出以来，引起了许多专家学者的广泛关注，已应用到金融业、零售业、医疗保健和政府决策等多个领域。数据挖掘在医疗领域的研究大多集中在辨证挖掘方面[2]，然而将数据挖掘技术应用到临床疗效实现客观的疗效评价体系还是一个空白。临床疗效是中医学赖以生存和发展的根本。以前的中医疗效评价[3]多采用单凭整体症状改善的个案记载和回顾性总结的方式，这种经验式总结方法曾经对中医学发展起到了巨大的推动作用。然而，随着科学技术的发展和社会进步，传统的中医临床疗效评价结果缺乏严格规范和标准的问题逐渐显露出来，已成为制约中医理论发展的“瓶颈”。本项目改变了以往疗效评价体系的研究方式，通过数据挖掘技术对小儿肺炎数据进行研究，融合了数据库、人工智能[4]和数理统计等相关技术，挖掘出中医临床数据上客观存在的诊治规律，并由这些规律出发，经专家审核确认并加以验证建立了客观的中医疗效判定规范。运用此方法建立的疗效判定模型，可明显提高疗效评价的客观化和科学化水平。

2　数据的特点

病历数是各个分中心收集的几千份病历，有以下特点：

(1)数据具有时间序列变化的特点。病程 10 天，第 0 天是病人在治疗前的病程记录，每一个分片表示病人的一天病程，每个病人共计 11 个分片。

(2)症状的多极层次结构。中医数据的特点是属性之间有很强的关联关系，这种关系说明了它们之间相互联系、依赖的程度。

(3)症状的时序伴随特性。属性和属性之间存在着时序伴随关系。即一种症状的上升(下降)可以影响另一种症状的上升(下降)，这种伴随关系形如：$A\uparrow\rightarrow B\uparrow$, $A\downarrow\rightarrow B\downarrow$，即 A 的正变化或负变化引起 B 的正变化或负变化，A 对症状的变化起主要的作用，B 作为伴随 A 的症状。

3　基于数据挖掘的交互式疗效评价框架

交互式疗效评价体系如图 1 所示。

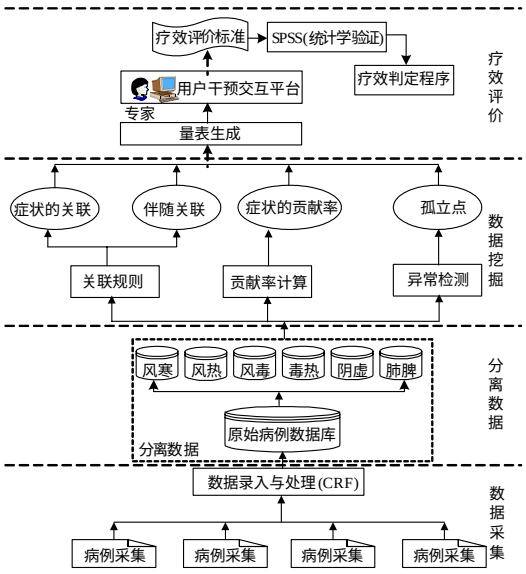

图 1　交互式疗效评价框架

基金项目：国家自然科学基金资助项目(60773218)
作者简介：印　莹(1980－)，女，博士研究生，主研究向：数据库，数据挖掘；张　斌，教授、博士生导师；赵宇海，讲师、博士；张晓红、张明卫，博士研究生
收稿日期：2008-03-15　**E-mail：**yy_00000000@163.com

基于上述特有的数据特性，本文以小儿肺炎为示范，采用数据挖掘、数理统计和逻辑分析相结合的方法，通过回顾性和前瞻性多角度的验证与比较研究，揭示各证和所属症状之间的关联性，优化辨证规范和建立公认的、可度量的病证结合的疗效评价方法体系。

3.1 数据采集与预处理

数据采集指从各采集点采集数据，消除噪声数据，最终以 CRF 表的形式录入到数据库中。预处理主要删除不合格数据：(1)数据和分离数据。删除不合格数据是指：有些数据在录入时并没有在病人的分类证一栏中作标识，或者发病记录只有一天，而一天的记录无法得到疗效判定结果，需要删除。(2)分离数据。前期采集由多种类型混在一起的 11 天肺炎病历数据，根据数据集的字段信息进行分离数据可以将数据集分成风寒闭肺、风热闭肺、痰热闭肺、毒热闭肺、肺脾气虚等 7 张表。

3.2 疗效评价中用到的数据挖掘关键技术

预处理之后的数据是干净的数据，它保证不含丢失值和缺失值，数据也不存在歧义。本节重点介绍影响疗效评价的几个因素及解决办法。

3.2.1 静态关联规则

在中医疗效挖掘的过程中，关联挖掘[5]讨论的问题并不仅仅是数据集中属性间的关联关系，它还要解决如何从证属症状中获得证的特异症、主症和次症的问题以及症状的贡献率。然而对中医数据高维的特点，关联规则的结果很冗余，对大量的结果来说，并不是症状的所有程度都需要给出程度关联的挖掘，专家可以选择感兴趣的规则，如：只对重程度的症状变化感兴趣，只对中等程度的症状变化感兴趣，或只对轻程度的症状感兴趣，框架允许对结果进行筛选，用户可以只看到感兴趣的规则。

3.2.2 时序伴随关联关系

疗效评价中关联挖据的目的是不仅要得到症状间的关联关系，而且要得到更加详细的症状的同步关系，即一种症状的上升(下降)可以影响另一种症状的上升(下降)，得到这样的规则便于用户调整症状的贡献率权值，即$A\uparrow\rightarrow B\uparrow$, $A\downarrow\rightarrow B\downarrow$，$A$的正变化或负变化引起$B$的正变化或负变化，即$A$对症状的变化起到很主要的作用，$B$作为伴随$A$的症状。这样既可以挖掘出中医数据库中具有伴随症状的关联规则[6]，又能充分利用中医学者的行业背景知识，调整症状的贡献率权值。

3.2.3 疗效评价量表

目前的证候量化大多是症状的量化，如果能做到真正的证候量化，对临床的指导意义可能更大。目前的中医量表针对具体的疾病，包含的症状指标众多，症状的量值采用 0, 1, 2, 3 这 4 级标准来表示症状的正常、轻、中、重程度，由于不同的证中包含的症状可能不同，即使是相同的症状在不同的证中对疗效的重要程度也是不同的，因此不同证型使用相同的疗效量表，无法体现症状在不同证中的疗效作用。因此，本文采用数据挖掘技术得到真正有临床意义的动态量表。

算法 1 疗效评价量表算法

输入：小儿肺炎数据表

输出：症状的贡献率

1 数据预处理

2 读取病例数据，对数据进行筛选，得到疗效评价数据

3 辩证分类疗效数据

4 For 每一个类病人

5 统计每类的病人数 N，每个病人每天的证的程度

6 For 每一个病人 i

7 计算每个病人的症状变化情况，既病程变化差值V_k-V_{k+1}

8 计算每个病人证的变化程度，S_k-S_{k+1}

9 计算变化程度$X_i=(S_k-S_{k+1})/(V_k-V_{k+1})$

10 End For

11 考虑个体差异和标准化，并映射得到症状对该症的贡献率

$S=\sqrt{\sum_{i=1}^{N}(X_i-\overline{X})^2/N}$，其中，$\overline{X}=\sum_{i=1}^{N}(X_1+X_2+\cdots+X_n)/N$

$W_1(A)=S/\max(X_i)$

12 把每个症状的权重映射到原始 CRF 表，计算得到该症状在该证中的量表

$POWER(A)=W(A)*W_1(A)$.//POWER(A)表示新的量表值

13 End For

14 转到步骤 4，循环步骤 4~步骤 13，得到所有证的量表

贡献率是针对证内的数据进行关联挖掘，发现症状对证的贡献程度。关联挖掘最重要的作用是计算症状的加权值，即贡献率。贡献率是衡量量表的最重要的指标。以前针对属性的贡献率都是临床医师指定哪些症状重要，哪些症状不重要。然而，症状的权重不是绝对的，而是相对的，不能人为指定，要根据领域知识确定相对重要性并不是人为确定。疗效数据库是动态数据库，动态贡献率则讨论症在证的转变过程中的贡献程度，这是一个随时间变化的复杂过程。贡献率可以分别构造贡献矩阵，再通过目标函数形成综合贡献率。综合贡献率与证的构成比最终形成症状的量表。

获得了症状的动态贡献率，此时，人工交互是必不可少的一个环节，专家交互对结果进行审查。系统根据贡献率表，给出小儿肺炎各症的权值表，根据症状的伴随关联情况来调整症状的权重。算法 1 就是获得疗效评价量表的生成过程。

4 实验结果

实验数据来自几个采集中心的肺炎数据集，包括：上海，广州，沈阳，山东，黑龙江，成都等几大城市的科研协作单位采集，每个分中心的病历都由亚型 1 到亚型 7 等多个亚型的数据集和一些正常人的数据采集。

由于空间所限，因此本文不能对所发现的实验结果一一列举。表 1 展示的是关联规则挖掘的结果。体现的重点是症状之间的关联关系。前一症状表现基本可替代后一症状表现在专家干预生成量表时考虑其等效性的前提下，适当减少后一症状表的权值，并增加前一症状表现的权值，有助于生成更符合客观要求且更为科学的辩证规范量表。

表 1 静态关联挖掘结果

关联度	关联关系
0.89	肺部听诊/程度(呼吸音粗)====>脉象(数)
0.99	咽部症状(轻微咽赤)=====>脉象(浮)
0.97	脉象(滑)=====>舌色(红)
0.92	肺部听诊(呼吸音粗)==>舌色/淡红(淡红)
0.96	咳嗽 =====>舌苔/程度(腻苔)
0.91	脉象(浮) ====>咳嗽(阵咳)

表 2 展示的是伴随关联规则挖掘结果。伴随关联挖掘是动态关联结果的体现，得到的不仅仅是关联关系，而是症状随着时序变化的伴随关系，伴随关系可以得到症状的同步或者异步的伴随情况，得到了专家的高度认可，更加有助于调整贡献率的权值。表 3 展示了疗效评价症状贡献率和量表的挖掘结果。即小儿肺炎中风热闭肺证和痰热闭肺证中各症状的不同表现对小儿肺炎疗效评价的贡献率。大量的结果无法一一列出，可以看到症状在不同证的贡献率是不同的，因此，

（下转第 46 页）