Hans Journal of Data Mining 数据挖掘, 2015, 5(4), 62-68
Published Online October 2015 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2015.54009

Hans汉斯

Association Rules' Application Study of the Electronic Medical Record Data Analysis in the Liver Disease of TCM

Yuwei Wang, Dan Xie*

College of Information Engineering, Hubei University of Traditional Chinese Medicine, Wuhan Hubei
Email: *tonghua123@sina.com

Received: Oct. 25th, 2015; accepted: Nov. 19th, 2015; published: Nov. 26th, 2015

Abstract

Objective: Based on analysis association rules mining model data for the liver disease of traditional Chinese medicine (TCM), to search for relation between checking index and the Chinese medicine dialectical. Method: By setting the minimum support and minimum confidence, we used association rules to analyze the patients' data for the liver disease of traditional Chinese medicine. According to the importance of generated rules, we screened out the rules which are positively correlated to the before and after rule, and evaluated the accuracy of the mining results with the lift chart. Result: Through the analysis of 317 samples, 48 rules are received, which reveals the relationship between the check index combination and TCM syndrome differentiation results. Conclusion: Using association rules in electronic medical record data analysis for the liver disease of TCM can reveal the influence of different examination indexes for TCM syndrome differentiation, and is advantageous to the auxiliary diagnosis.

Keywords

Association Rule, The Liver Disease of TCM, Data Analysis

关联规则在中医肝病电子病历数据分析中的应用研究

汪玉薇，解　丹*

*通讯作者。

文章引用: 汪玉薇, 解丹. 关联规则在中医肝病电子病历数据分析中的应用研究[J]. 数据挖掘, 2015, 5(4): 62-68.
http://dx.doi.org/10.12677/hjdm.2015.54009

湖北中医药大学信息工程学院，湖北 武汉
Email: *tonghua123@sina.com

收稿日期：2015年10月25日；录用日期：2015年11月19日；发布日期：2015年11月26日

摘　要

目的：基于关联规则挖掘模型分析中医肝病患者数据，探寻检查指标与中医辨证之间的关联。方法：通过设置最小支持度和最小可信度对中医肝病患者资料进行关联规则分析，根据生成规则的重要性筛选出前后件成正相关的规则，结合提升图评价挖掘结果准确性。结果：分析样本例数317例，共获得30条规则，揭示了检查指标组合与中医辨证结果间的关系。结论：在中医肝病电子病历数据中应用关联规则分析可以揭示不同检查指标对于中医辨证的影响，有利于辅助诊断。

关键词

关联规则，中医肝病，数据分析

1. 引言

自 1993 年 Agrawal 等人提出关联规则概念后，基于关联规则的挖掘算法被普遍应用于商场购物篮的分析，目的是发现顾客的购物习惯，其原理为“基于频繁项集生成规则并计算规则的贡献度”[1]。在中医诊断中，通常检查指标包括有患者基本信息、四诊信息和实验室检查指标等[2]，而肝病诊断指标常见的有肝病通用刻下症、舌脉诊、肝病专科检查等，中医肝病常见证型有肝郁脾虚、湿热蕴结、肝郁气滞、脾虚湿阻、肝肾阴虚等[3] [4]。借鉴关联规则挖掘商品数据的原理，探索中医诊断指标对证型判定的贡献度，可形成诊断量表[5]，利于中医临床辨证。本研究利用 Microsoft Visual Studio 2008 提供的挖掘工具，采用关联规则与决策树模型对中医肝病电子病历数据进行挖掘，寻找指标项与证型之间符合预设条件的规则，为中医肝病辨证提供参考[6]。

2. 资料与方法

2.1. 资料来源

本研究资料来自于某临床肝病研究所电子病历，共 317 例。资料涉及患者的年龄、性别、舌脉诊、肝病通用刻下症以及肝病专科检查，共计 27 项指标。该资料以定性资料为主，部分变量有少量缺失值。

2.2. 方法

本研究利用 SQL Server 2008 [7]进行挖掘分析，其中用于分析的数据存储在 SQL Server 数据库中，挖掘分析过程在 Microsoft Visual Studio 2008 平台上实现。该平台封装了常用挖掘模型，本研究选取其中关联规则模型(Microsoft_Association_Rules)与决策树模型(Microsoft_Decision_Trees)进行数据挖掘。其中关联规则模型采用的是经典的 Apriori 算法，其挖掘步骤如下：

1)扫描数据集，生成频繁项集。此过程一般比较耗时。这些项集的支持度必须要大于等于最小支持度(Minimum_Support)。

2) 基于第一步生成的频繁项集，产生关联规则。这些规则出现的概率(置信度)必须大于等于最小概率(Minimum_Probability)。

分析主要得到形如 A ≥ B 的规则，支持度表示同时满足规则前件 A 和规则后件 B 的例数占总例数的比例即概率 P(A∪B)，可信度表示在所有满足规则前件 A 的例数中满足规则后件 B 所占的比例即条件概率 P(B|A) [8] [9]。

3. 结果

3.1. 决策树模型挖掘结果

3.1.1. 生成决策树视图

将 27 项检查指标作为决策树模型的输入变量，证候作为可预测变量。在挖掘模型选项卡中设置挖掘模型的参数，将 COMPLEXITY_PENALTY (抑制决策树的生长，值越小，拆分的可能性越大，取值范围在 0 到 1 之间)设置为 0.01，其它参数保持默认值，运行挖掘模型后在“挖掘模型查看器”选项卡中查看挖掘结果。生成的决策树图形如图 1 所示。

3.1.2. 依赖关系网络

“依赖关系网络”选项卡显示决定挖掘模型预测能力的各个属性之间的关系。决策树模型挖掘结果的依赖关系网络显示，“胸部”及“巩膜”的望诊结果以及“是否有移动性浊音”对中医肝病证型的诊断具有重要参考价值，这与图 1 决策树相符。检查指标与证型间依赖关系网络如图 2 所示。

3.1.3. 挖掘模型准确性评估

建立好的数据挖掘模型并不能保证能够直接解决问题，还需要使用多种方法来评估和检验数据挖掘模型的质量和特征。可将数据分为定型集(训练集)和测试集来评估数据挖掘模型。通过将数据集分区为定型集和测试集时，定型集是取大多数数据，小部分数据用于测试。通过对全部数据的整体数据抽样，要保证定型集和测试集的相似。通过使用相似的数据来进行定型和测试，可以更好得验证数据挖掘模型。验证数据挖掘模型主要是从准确性、可靠性和有用性这三个方面入手。

在 SQL Server2008 中的挖掘模型验证方法可以用绘制模型准确性图表，挖掘模型的交叉验证等方法来进行模型验证。本研究选择提升图来对挖掘模型进行验证，决策树模型得分 0.55，总体正确率为 31.58% (理想模型为 51.00%)，预测概率为 63.62%，提升图见图 3。

3.2.关联规则模型挖掘结果

3.2.1. 频繁项集与生成规则

将 27 项检查指标作为关联规则挖掘模型的输入变量，证候作为可预测变量。在挖掘模型选项卡中设

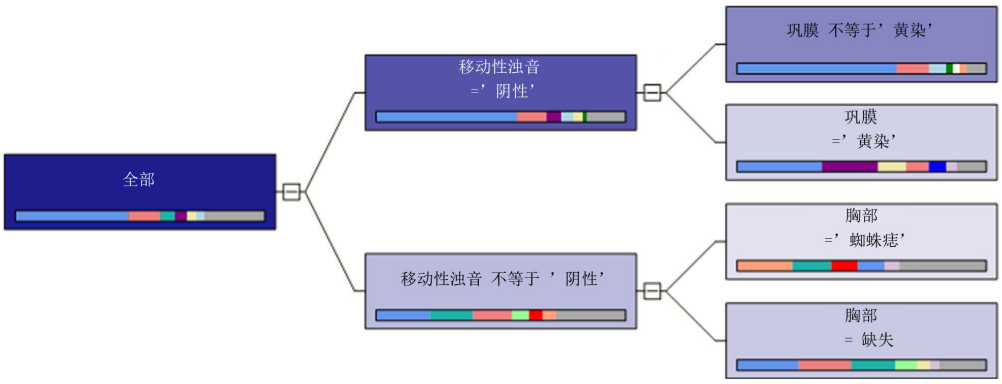

Figure 1. Decision tree with COMPLEXITY_PENALTY = 0.01
图 1. COMPLEXITY_PENALTY = 0.01 时的决策树

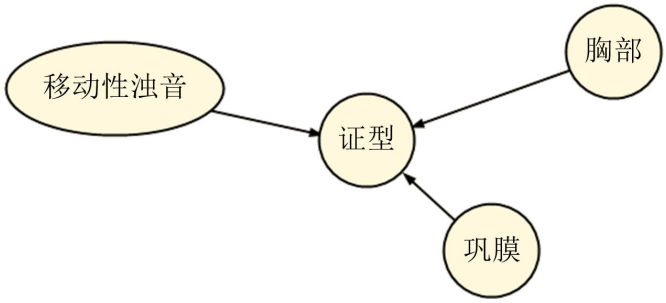

Figure 2. Schematic diagrams for the model of decision tree
图 2. 决策树依赖关系网络

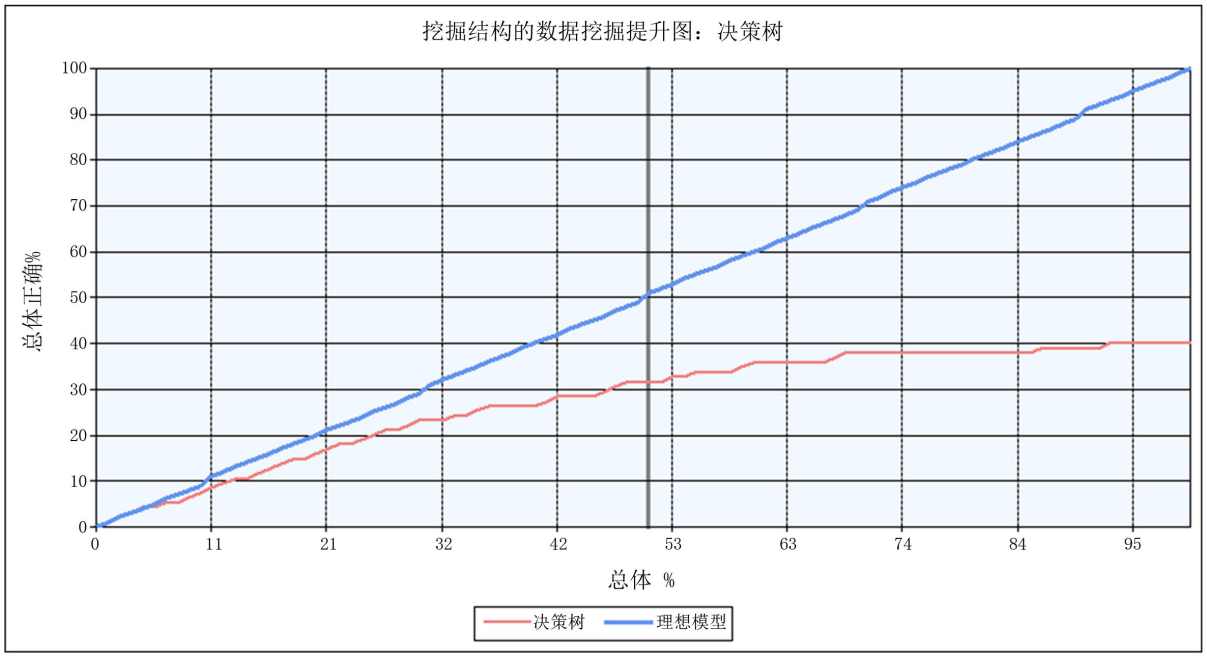

Figure 3. The lift chart for the model of decision tree
图 3. 决策树模型提升图

置挖掘模型的参数，将最小支持度设置为 0.2，其它参数保持默认值，运行挖掘模型在“挖掘模型查看器”选项卡中查看挖掘结果。

频繁项集最大长度默认为 3，本研究为与决策树模型挖掘结果形成对比，默认最大频繁项集长度为 3 不做改变。共挖掘到满足最小支持度的 3 项集 218 个、2 项集 103 个、1 项集 20 个，提取频繁 3 项集前 10 位如表 1 所示。

基于频繁项集，模型自动生成 30 条规则，设置最小概率(最小置信度)为 0.55，获得符合条件的规则 12 条，如表 2 所示。

3.2.2. 依赖关系网络

关联规则挖掘模型指标与证型间依赖关系网络如图 4 所示。

3.2.3. 挖掘模型准确性评估

关联规则模型得分 0.53，总体正确率为 29.47% (理想模型为 51.00%)，预测概率为 68.47%，提升图见图 5。

Table 1. The top 10 frequent 3 item sets with minimum support = 0.2
表 1. 支持度 = 0.2 排名前 10 位的频繁 3 项集

支持度	大小	项集
119	3	舌苔 = 薄白，舌色 = 淡红，脉象 = 弦
113	3	移动性浊音 = 阴性，舌色 = 淡红，脉象 = 弦
106	3	年龄 = 40 岁以上，舌色 = 淡红，脉象 = 弦
101	3	舌苔 = 薄白，年龄 = 40 岁以上，脉象 = 弦
100	3	舌苔 = 薄白，移动性浊音 = 阴性，脉象 = 弦
97	3	巩膜 = 正常，移动性浊音 = 阴性，脉象 = 弦
97	3	舌苔 = 薄白，移动性浊音 = 阴性，舌色 = 淡红
96	3	巩膜 = 正常，舌色 = 淡红，脉象 = 弦
96	3	舌苔 = 薄白，年龄 = 40 岁以上，舌色 = 淡红
95	3	巩膜 = 正常，移动性浊音 = 阴性，舌色 = 淡红

Table 2. Rule results with minimum confidence = 0.55
表 2. 最小置信度 = 0.55 的规则结果

概率	重要性	规则
0.685	0.471517	巩膜 = 正常，移动性浊音 = 阴性 -> 证型 = 肝郁脾虚
0.625	0.219495	精神 = 差，移动性浊音 = 阴性 -> 证型 = 肝郁脾虚
0.624	0.322745	巩膜 = 正常，舌色 = 淡红 -> 证型 = 肝郁脾虚
0.607	0.282102	舌苔 = 薄白，移动性浊音 = 阴性 -> 证型 = 肝郁脾虚
0.598	0.280754	巩膜 = 正常，脉象 = 弦 -> 证型 = 肝郁脾虚
0.598	0.202703	巩膜 = 正常，性别 = 男 -> 证型 = 肝郁脾虚
0.597	0.354607	移动性浊音 = 阴性，舌色 = 淡红 -> 证型 = 肝郁脾虚
0.592	0.346709	巩膜 = 正常 -> 证型 = 肝郁脾虚
0.588	0.218384	巩膜 = 正常，舌苔 = 薄白 -> 证型 = 肝郁脾虚
0.556	0.436964	移动性浊音 = 阴性 -> 证型 = 肝郁脾虚
0.555	0.268022	移动性浊音 = 阴性，脉象 = 弦 -> 证型 = 肝郁脾虚
0.554	0.170185	性别 = 男，移动性浊音 = 阴性 -> 证型 = 肝郁脾虚

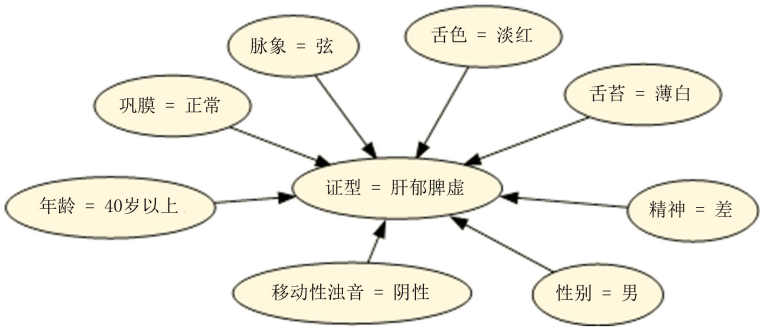

Figure 4. Schematic diagrams for the model of association rules
图 4. 关联规则依赖关系网络

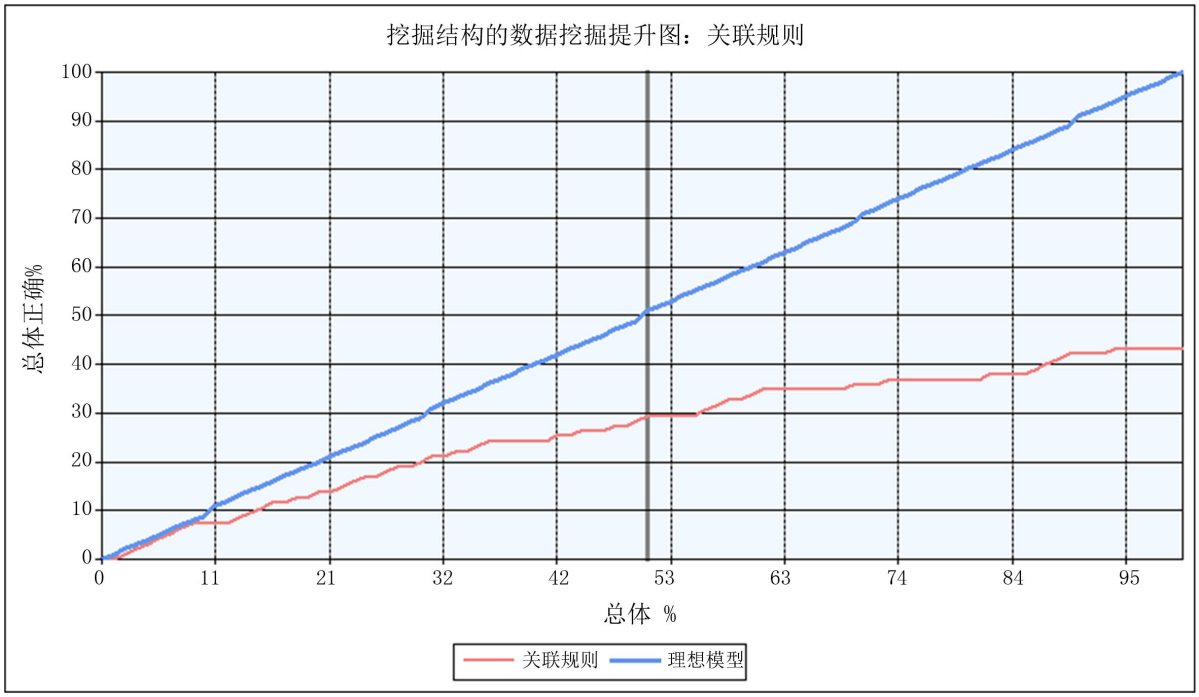

Figure 5. The lift chart for the model of association rules
图 5. 关联规则提升图

4. 讨论

从挖掘模型准确性来看，决策树模型与关联规则模型模拟中医肝病辨证近似，关联规则模型的提升度相比决策树模型稍高。从结果可读性来看，决策树模型优于关联规则模型，但决策树模型结果不及关联规则模型的挖掘结果详实，很多信息遗失了，而后者以规则的形式给出症状集与证型之间的关系，保留了原始数据的关系，更具临床分析价值。

此外，与其它统计学方法相比，基于关联规则的医学数据分析具有以下特点：(1) 不需要考虑变量间的复杂性，其分析结果不会因为资料中加入或去掉一个变量而影响已存在的结果。(2) 数据中的变量既是自变量又是目标变量，研究者不需要事先确定那个是目标变量，易获得某些意料之外、有意义的模式。(3) 关联规则分析通常只考虑某个或某几个变量，而不是针对全部变量，这使得分析结果与不同方法分析的结果在某些变量上有较大差异[10]。同时，传统关联规则分析要求变量必须是离散的，所以对于数据中计量型变量如何离散是应用中常见的一个问题。(4) 关联规则分析中所设置的一些参数直接影响了所获得规则的数量和价值。支持度越高规则的说服力越强，但设置较高有可能漏掉一些有意义的规则，而最小支持度阈值设置过低一方面增加了无效规则的产生影响分析进程。(5) 当例数较少时，存在偶然因素作用导致率的波动变化大，有可能产生一些假阳性规则。同样，最小可信度设置过高会使一些多分类变量不易出现在规则中。

综上，今后在对中医肝病电子病历数据进行分析时，要充分考虑到以上问题。随着数据规模不断增长，本应用研究成果将对肝病临床研究发挥更大指导作用。

基金项目

《基于中医特色的老年社区的健康监测与干预关键技术研究》(201307003)，中国中医科学院，2013年；《医学信息生“241”创新人才培养模式研究》(2014A08)，湖北中医药大学校级教学研究重点项目，

2014 年。

参考文献 (References)

[1] 范明, 孟小峰. 数据挖掘概念与技术[M]. 第 3 版. 北京: 机械工业出版社, 2012: 55.

[2] 唐伟, 周正光, 王欢欢. 胃脘痛中医辨证与胃镜表现的关联规则分析[J]. 中国中西医结合杂志, 2013, 33(3): 303-306.

[3] 崔树娜, 胡雪琴, 温先荣. 基于关联规则挖掘的白细胞减少症方药规律分析[J]. 中国中医药图书情报杂志, 2014, 38(1): 23-26.

[4] 吴嘉瑞, 童有健, 张晓朦, 张冰. 基于关联规则和复杂系统熵聚类的邓星伯治疗肺系病证用药规律研究[J]. 中国实验方剂学杂志, 2014, 20(7): 223-226.

[5] 车立娟, 马利庄, 胡义扬. 基于关联规则算法的慢性乙型肝炎证型诊断量表多中心研究[J]. 上海中医药杂志, 2014, 48(5):11-14.

[6] 杨霖, 洪菲, 杨华元. 针刺手法数据挖掘的关联规则与分类[J]. 上海针灸杂志, 33(11): 2014:1066.

[7] 汪明. SQL Server 2008 R2 关联规则研究[J]. 电脑知识与技术, 2011, 7(16): 3774-3776.

[8] 朱明. 数据挖掘导论[M]. 北京: 中国科学技术大学出版社, 2010: 5.

[9] 郭淑红. 基于 Apriori 算法的股票分析仿真系统[J]. 计算机仿真, 2010, 27(6): 334-337.

[10] 陈志泊. 数据仓库与数据挖掘[M]. 北京: 清华大学出版社, 2009: 116.