Software Engineering and Applications 软件工程与应用, 2015, 4(5), 96-100
Published Online October 2015 in Hans. http://www.hanspub.org/journal/sea
http://dx.doi.org/10.12677/sea.2015.45013

Hans汉斯

Exploratory Research on Association Rules about Electronic Medical Records of TCM Liver Disease

Dan Xie, Siqi Wang, Yuwei Wang

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

Received: Sept. 30th, 2015; accepted: Oct. 14th, 2015; published: Oct. 21st, 2015

Abstract

Objective: To investigate the application of the association rule and decision tree classification of data mining technology on diagnosis of TCM liver diseases. Method: To collate and screen the data about hospitalization medical records of TCM liver disease, and use Weka to carry on the data mining analysis, which mainly uses Apriori algorithm in association rules and C4.5 algorithm in decision tree classification. Result: Ten rules were found and the decision tree of the common TCM liver disease was generated. The relationship between lab examination information and TCM dialectical treatment has been given. Conclusion: The results obtained from the data mining are generally consistent with the results from the doctors. This method can be used in disease diagnosis based on the electronic medical record, which has certain practical value.

Keywords

TCM Liver Disease, Electronic Medical Record, Data Mining

中医肝病电子病历病证关联分析探索性研究

解　丹，王斯琪，汪玉薇

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9月30日；录用日期：2015年10月14日；发布日期：2015年10月21日

文章引用: 解丹, 王斯琪, 汪玉薇. 中医肝病电子病历病证关联分析探索性研究[J]. 软件工程与应用, 2015, 4(5): 96-100. http://dx.doi.org/10.12677/sea.2015.45013

摘　要

目的：探讨数据挖掘技术中的关联规则与决策树分类在中医肝病诊断中的应用。方法：整理筛选来自某中医医院肝病科电子病历的数据，利用Weka数据挖掘软件对其进行数据分析，主要采用关联规则中的Apriori算法和决策树分类中的C4.5算法。结果：经算法分析产生十条相关规则，并生成中医常见肝病的决策树，揭示了检查指标与中医辨证间的关系。结论：数据挖掘结果与医生实际诊断结果基本一致，可将该方法用于基于电子病历的疾病诊断，具有一定实用价值。

关键词

中医肝病，电子病历，数据挖掘

1. 引言

中医肝病是指肝脏的正常生理功能受损而出现的各种病理表现，根据患者不同的临床表现辨病为黄疸、积聚、肝癌、肝着、臌胀等[1] [2]。肝脏疾病大多属于慢性难治性疾病，严重影响了患者的身心健康，因此提高肝病的诊疗水平将减轻患者痛苦，提高其生存质量。本文选用某中医院的肝病科电子病历数据，通过整理筛选建立中医肝病数据库，并运用数据挖掘技术中的关联规则和决策树分类对数据进行分析与探索性研究，探讨中医肝病与症状之间的联系，以期为中医肝病诊断研究提供借鉴作用。

2. 研究方法

2.1. 数据来源及预处理

2.1.1. 数据来源

研究对象来自就诊于湖北省中医院门诊及住院患者，病例的采样时间为 2013 年 1 月~12 月，原始病例 210 例。

2.1.2. 纳入标准

① 符合西医诊断标准。② 符合黄疸、肝着、鼓胀、积聚、肝厥和肝癌六大中医证型诊断标准者。③ 愿意接受研究者。④ 年龄 18~75 岁。

2.1.3. 排除标准

① 不符合病例纳人标准者。② 重叠患有除肝病以外的重大疾病患者，如精神病患者、心血管患者等。③ 妊娠或哺乳期的女性患者。

2.1.4. 数据预处理

对原始病例按患者 ID 和就诊时间进行预处理，主要工作是提取与中医肝病相关的数据属性，并进行适当的筛选，并进行了规一化处理，最终保留中医诊断病名、病位、病性、舌色、舌苔等 61 个属性，纳入病例 98 例。

2.2. 数据挖掘方法

数据挖掘具有自动预测趋势和行为、概念描述、关联分析、分类、聚类、偏差检测等功能，常用的方法有关联规则、决策树分类、朴素贝叶斯分类、K-means 聚类等，为研究中医肝病与症状之间的关联性，本文主要采用关联规则及决策树分类对电子病历数据进行分析。

2.3. 数据挖掘软件

本文使用的数据挖掘软件是 Weka (Waikato environment for knowledge analysis)，由新西兰怀卡托大学开发，主要用于数据挖掘和知识发现的数据挖掘软件。它集合了大量机器学习算法，包括对数据进行预处理、分类、回归、聚类、关联规则，并可将结果在交互式机器上进行可视化演示[3]。

2.4. 数据挖掘算法

2.4.1. 关联规则

关联规则是从数据库或数据仓库中抽取频繁出现的模式，以寻找两个或多个属性取值之间存在的某种规律性。常用的关联规则分析算法包括 Apriori 算法、频繁模式增长(FP-增长)方法、多维关联规则挖掘等。本文采用的是经典 Apriori 算法，该算法是一种基于水平数据表示、广度优先搜索的挖掘算法，结构简单，易于理解，没有复杂的推导[4]。

2.4.2. 决策树分类

决策树方法是解决分类问题的一个有效方法，它能很容易地构造出规则，而且规则通常易于理解和解释。决策树是一个类似于流程图的树状结构，其中每个内部节点表示在一个属性上的测试，每一个分枝代表一个测试输出，而每个树叶结点代表类或类分布，决策树的生成则是一个从上而下、分而治之的过程[5]。决策树分类的算法有 ID3 算法、C4.5 算法、C5.0 算法等，本文采用经典 C4.5 算法，该算法可以处理具有连续值的属性，并使用信息增益比来作为选择根节点和各内部节点中分枝属性的评价标准，克服了 ID3 算法在选择属性时偏向于取值较多的属性的不足[6]。

3. 研究结果

3.1. 中医肝病分布情况

纳入的 98 例肝病患者经中医辨病，病名分布由高到低依次为肝着(37 例)、积聚(18 例)、黄疸(11 例)、肝癌(9 例)、臌胀(6 例)、肝积(4 例)、肝厥(3 例)、血证(3 例)、肝癖(3 例)、水肿(2 例)、胁痛(1 例)、心悸(1 例)，其分布情况如图 1 所示。

3.2. 关联规则结果分析

在将数据进行关联规则分析计算之前，对数据进行了规范化处理，并在 Weka 软件中设置好最小支持度与最小置信度。支持度是指事务集 D 中包含 X∪Y 的事物所占的百分比，置信度是指事务集 D 中包含 X 的事务同时也包含 Y 的百分比，即 support (X→Y)=P(X∪Y)，confidence (X→Y)=P(X|Y)。这里设置最小支持度为 0.15，最小置信度为 0.7，得到的相关关联规则结果如表 1 所示。

本研究是基于肝病住院病历的相关数据，故所有记录的病位都包含肝。从图 1 中医肝病分布图中可以发现肝着所占的比例较大，在关联规则数据挖掘的结果中也体现了该点。肝着一般是因肝热病、肝瘟等之后，肝脏气血瘀滞，着而不行。若证型为肝郁脾虚证，则是肝胆气郁，横犯脾胃，脾失健运所致。由此可见，诊断为肝着，且病位包含脾时，证型极大可能为肝郁脾虚，即表 1 第 2 条规则所示。当肝细胞损伤时，血液中胆红素会升高，此时巩膜会出现黄染，随之皮肤粘膜也会出现黄染，对应表 1 的第 4 条规则。如黄疸患者的皮肤、巩膜等组织会出现黄染，严重时尿、痰、泪液及汗液也会出现黄染。

3.3. 决策树分类结果分析

由前面中医肝病分布图可以发现，肝厥、血证、肝癖、水肿、胁痛、心悸的记录数量较少，所以进行决策树分类时只针对记录数为前六位的症状。将处理好的数据通过 Weka 软件计算分析后，可以得到如图 2

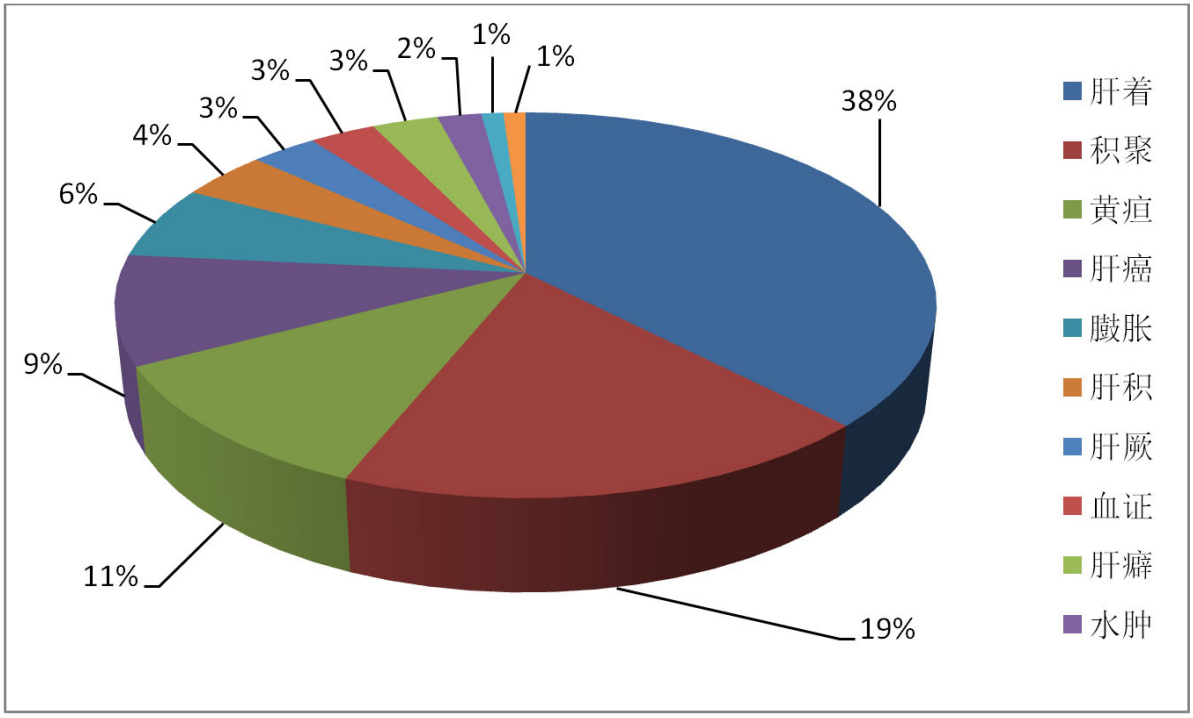

Figure 1. The distribution diagram of liver disease of TCM
图 1. 中医肝病分布图

Table 1. The calculation results of association rules
表 1. 关联规则计算结果

序号	属性 1	属性 2	出现频次	属性 3	出现频次	置信度
1	中医诊断 = 肝着	病性 = 本虚标实	21	证型 = 肝郁脾虚	21	0.97
2	中医诊断 = 肝着	病位 = 脾	24	证型 = 肝郁脾虚	23	0.96
3	中医诊断 = 肝着		37	证型 = 肝郁脾虚	35	0.95
4	皮肤黄染 = 中度		21	眼巩膜 = 黄染	17	0.81
5	眼巩膜 = 黄染		21	皮肤黄染 = 中度	17	0.81
6	小便颜色 = 较黄		21	病位 = 脾	17	0.81
7	病性 = 本虚标实	病位 = 脾	27	证型 = 肝郁脾虚	20	0.8
8	小便颜色 = 较黄		21	证型 = 肝郁脾虚	15	0.71
9	腹部移动性浊音 = 阳性		24	刻下症 = 腹胀	17	0.71
10	病性 = 本虚标实		47	证型 = 肝郁脾虚	33	0.7

所示的决策树，其中正确分类为 63.53%。

参与分类的属性很多，但经过算法剪枝后，进行分类的属性为身目尿黄染、皮肤粘膜黄染、胁痛等。积聚是腹内结块，或痛或胀的病证，黄疸、胁痛病后，湿浊留恋，气血蕴结，可导致其发生。图 2 中对积聚进行分类时，就涉及到了胁痛、皮肤粘膜轻度黄染等情况，与前面所述相吻合。

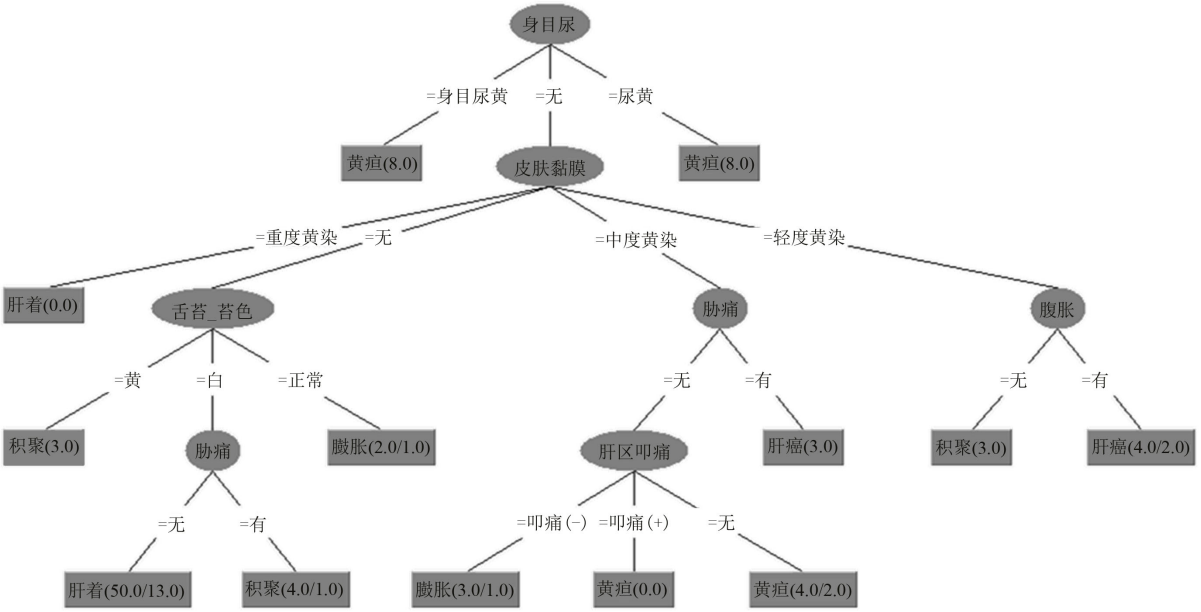

Figure 2. The classification result diagram of decision tree
图 2. 决策树分类结果图

4. 结语

正确的诊断对于确立治疗原则、指导用药以及调理康复等具有重要意义，医院数据库中含有大量有用的诊断信息，利用数据挖掘的相关技术可用来寻找病与症之间的联系、病与病之间的差异等，从而提高诊断水平，规范诊断过程。本文以肝病电子病历中的信息为数据来源，采用数据挖掘技术中关联规则与决策树分类，进行了探索性研究，找到了十条相关规则，并对记录数较多的中医肝病进行分类，形成决策树。算法分析的结果与事实医生诊断结果基本吻合，因此患者可以通过对比自己的临床症状，进行初步的疾病辅助预测，具有一定现实价值。此外，在以后的相关研究中需增加符合纳入标准的病例，提高算法的精度，使得预测结果更准确。

基金项目

湖北省省级教学研究项目“医学信息生科研数据处理实战能力教科临三位一体培养模式研究”，项目编号：2012305；湖北中医药大学校级教学研究重点项目“医学信息生‘241’创新人才培养模式研究”，项目编号：2014A08。

参考文献 (References)

[1] 余世峰, 刘凤斌, 罗仕娟 (2010) 中医肝病临床疗效评价量表理论结构模型构建的探讨. *中药新药与临床药理*, **4**, 449-450.

[2] 王融冰 (2015) 肝病的中医药治疗. *临床肝胆病杂志*, **1**, 2-6.

[3] 郑文娟, 王会青, 陈俊杰 (2013) 基于 Weka 平台的 FCM 算法的研究与实现. *计算机应用与软件*, **10**, 41-44.

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

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

[6] 陈志泊 (2009) 数据仓库与数据挖掘. 清华大学出版社, 北京, 116.