A Research on the Traditional Chinese Medicine Assistant Diagnosis and Treatment Information System Based on Relationship Rule Of the Data Mining

Li Xiaohua[1],Chenqian[1] Liang Zhiwei[1], Luo Yunjian[1], Lu Yubo[1]

(1. The 2nd Hospital Affilited to Guangzhou University of Traditional Chinese Medicine, Guangzhou 510405)

Target: By the computer technology of data mining in Guangdong Province Hospital of Traditional Chinese Medicine's Source Material of Electronic Medical Record , Discuss and found a research methodology and operable technology platform by the same with Traditional Chinese Medicine Assistant Diagnosis and Treatment. **Method:** Establish a Traditional Chinese Medicine's data warehouse using oracle database production. Through data mining or data conversion or data load on digital Traditional Chinese Medicine's Source Material of Electronic Medical Record. Summarize usable rule based on relation regular analysis. **Result:** Found the relationship of disease's factor such as symptom, sensitive history, etc. and developed a software system which could integrate isomerous database system and high efficiency data query and analysis.

Conclusion: Through data mining or data conversion or data load on digital Traditional Chinese Medicine's Source Material of Electronic Medical Record, discovering the latency or apparent diagnosis and treat rule by data mining which can help to science trimming and mining Electronic Medical Record ,and offer the hold out argument for clinical diagnosis and treatment.

Key Words: database warehouse; data mining; relation regular rule; Assistant Diagnosis and Treatment.

基于关联规则的数据挖掘技术在中医医学辅助诊疗系统的应用研究

李小华[1] 陈倩[1] 梁志伟[1] 罗云坚[1] 吕玉波[1]

(1. 广州中医药大学第二附属医院, 广州 510120)

摘 要：目的 运用计算机数据处理技术，对广东省中医院住院临床病历为原始资料，进行数据挖掘整理，探讨并建立一种适用于中医临床诊疗规律辅助发现的有效研究方法和可操作的技术平台。**方法** 选用 oracle 数据仓库产品,建立中医临床病历数据仓库，对已经数字化的中医临床病历数据进行数据抽取、数据转换、数据加载。在此基础上，应用数据挖掘中的关联分析，从而总结出有用的规律。**结果** 发现疾病与症状、过敏史之间存在一定的关联，并研制出针对不同数据库数据采集的数据，实现有机集成和高效率查询分析的软件系统。**结论** 利用数据仓库技术对数字化病历原始数据进行抽取、转换、加载，并在此基础上进行数据挖掘，可以发现显在或潜在临床诊疗规律，有助于科学地整理和挖掘临床电子病历信息，为临床诊疗提供有力的支持论据。

关键词： 数据仓库 数据挖掘 关联规则 辅助诊疗

删除的内容: 已经数字化的中医

删除的内容: 医学

删除的内容: 了相关的,可从

删除的内容: 及

删除的内容: 的

删除的内容: 有效的

删除的内容: 并

删除的内容: 从中发现其中的有意义的

删除的内容: 决策

随着医院管理的不断深入和计算机信息技术的不断发展，当前以电子病历系统为核心的新一代“以病人临床诊疗信息”为中心的医院信息系统正逐步取代以往以“医院财务管理为核心”的系统。电子病历是以信息技术为手段，实时采集病人在整个医疗过程中所产生的各类原始记录，它包含病历文书、各种医嘱、检查与检验结果等，且涵盖文字、数字、图像、声音、医学影像等以多种电子介质为载体的临床诊疗信息。如何从电子病历采集的海量的临床诊疗信息中，多角度、高效率地分析研究，从中发现隐藏的、潜在的和有意义的知识和信息，为临床诊疗决策提供有力支持，是当前医学信息学的热点。因此，在电子病历系统的基础上进行医学辅助诊疗系统的研究使电子病历系统同时具有辅助临床决策支持功能具有实际意义。

本研究从医院电子病历系统中采集、保存的中医临床病历信息中，运用计算机信息处理技术，对临床诊疗信息进行数据处理，供临床科研人员进行多层次、多角度的查询检索。同时运用数据挖掘技术，对定义主题的数据集进行关联规则的分析、归纳、处理，旨在研究和探索与诊疗规律相关信息，建立一种适用于中医医学诊疗规律辅助发现的有效研究方法和可操作的技术平台[1，2]。

1 研究资料和方法

1.1 研究资料来源

本研究选取广东省中医院 2005 年 1 月至 12 月的收入院的哮喘病的数字化临床病历 115 例，作为研究的样本数据集。病历的诊断标准 ICD10 编码为 J45.903[5]

1.2 研究方法及其过程

1.2.1 创建临床病历数据仓库：针对电子病历系统是一个庞大的涉及多学科的系统，数据往往存放于同构或异构数据库，且信息量巨大，要进行有效的面向决策支持分析的数据处理难度大的特点，我们开发了《中医临床诊疗辅助系统》。系统选用 oracle 数据仓库产品（建模工具 Oralce Warehouse Builder，数据加载工具 Oralce Internet Developer Suite，数据存储系统 Oralce 9i Database Enterprise Edition）对已经数字化中医临床诊疗数据进行包括抽取、转换、清洗、加载等处理，把分布在各中医临床业务数据库的数据转化为数据仓库中的合法、格式适合查询操作的数据，并以此建立中医临床病历数据仓库。系统为用户提供了可以自由的、快速的定制各种复杂组合条件的数据提取操作环境，提供了数据挖掘中的关联分析模块，并提供了与 SPSS 软件的外部数据接口，可以利用 SPSS 软件进行聚类分析。

系统的总体框架见图 1

删除的内容: 、研

删除的内容: 及过程

删除的内容: 1

带格式的: 项目符号和编号

删除的内容: 某

删除的内容: 研究

删除的内容: 2. 1

删除的内容: 研究方法

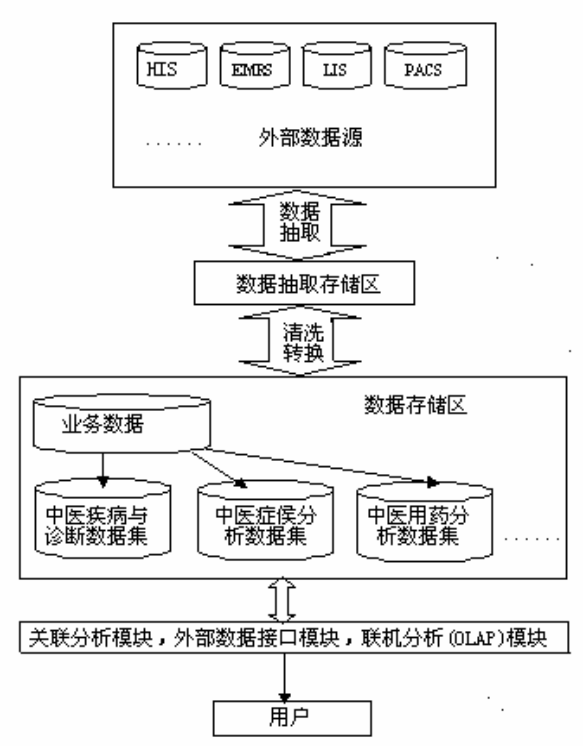

图 1　系统总体框图

1.2.2 数据挖掘的过程

数据挖掘的过程是多步骤的处理过程，在处理中可能有多次反复。主要是以下处理步骤。

（1）、数据处理：利用领域专门知识对目标数据集中的“脏数据”进行清洗。

（2）、数据降维和转换：降维和转换是指在考虑了数据不变表示或发现了数据的不变表示的情况下，减少变量的实际数目并设法转换到一个更易找到解的空间上，目的是将数据转换到适合数据挖掘处理的形式。

（3）、选择挖掘算法：运用选定的算法，如神经元网络、决策树、聚类分析技术、关联分析技术、序列发现等数据挖掘算法，从数据中找到最佳的知识发现模式。本研究中选用的是关联分析技术。

（4）、模式评价：根据最终用户的决策目的对数据发现的模式进行评价，以决定所得模式是否存入知识库。评价主要依靠领域专家的经验来完成。如果结果不能令决策者满意，还需要反复以上步骤。

数据挖掘的过程见流程图 2

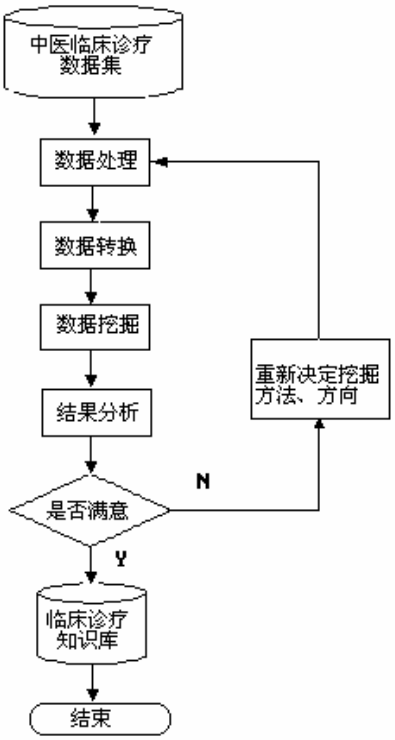

图 2 数据挖掘流程图

在数据挖掘中，关联规则挖掘是其中最常用的方法之一。关联规则就是描述在一个事物中，元组之间同时出现规律的知识模式[8]。

关联规则是形如 $X \to Y$ 的蕴涵式，这里 $X \subseteq I, Y \subseteq I$ ，并且 $X \cap Y = \Phi$ 。一般用下面两个参数描述关联规则的属性。

（1）可信度（ Suport)。规则 $X \to Y$ 在交易数据库 D 中的支持度是事务集中包含 X 和 Y 的事物数与所有事物数之比，记为 $\mathbf{support(X \to Y)}$，即

$$SUPPORT(X \to Y) = \frac{|\{T : X \cup Y \subseteq T, T \in D\}|}{|D|}$$

（2）可信度，又称置信度（Confidence）。规则 $X \to Y$ 在事物集中的可信度是包含 X 和 Y 的事物数与包含 X 的事物数之比，记为 $confidence(X \to Y)$ ，即

$$CONFIDENCE(X \rightarrow Y) = \frac{|\{T : X \cup Y \subseteq T, T \in D\}|}{|T : X \subseteq T, T \in D|}$$

挖掘关联规则问题，就是产生支持度和可信度分别大于用户给定的最小支持度(minsupp)和最小可信度(minconf)的关联规则。[8]

在本研究中，通过病人的疾病、年龄、症状、过敏史之间关系进行数据挖掘，以期发现一些潜在、有意义的关联规则[6]。

在本研究中采用 Apriori 算法，利用 k-项集来探索(k+1)-项集,然后再根据预先设定的最小支持度和可信度产生规则。具体做法如下：

每一份符合纳入标准的病历作为一个事务，用 TID 标记，每个事务由年龄，症状，过敏史等属性组成；为了满足分析需要，按照表 1、表 2、表 3 对年龄、症状、过敏史进行代码的转换。

表 1　年龄分组规则表

代码	年龄组	代码	年龄组	代码	年龄组
A1	不满周岁	**A2**	1-4 周岁	**A3**	5-9 周岁
A4	10-14 周岁	**A5**	15-19 周岁	**A6**	20-29 周岁
A7	30-39 周岁	**A8**	40-49 周岁	**A9**	50-59 周岁
A10	60-69 周岁	**A11**	70-79 周岁	**A12**	80-84 周岁
A13	85 岁以上				

表 2 症状代码表

代码	症状
Z1	咳嗽
Z2	发热
Z3	气促气喘
Z4	胸痛
…	…

表 3 过敏史

代码	过敏史
B1	有食物过敏
B2	有药物过敏
B3	有接触、气味等过敏
B0	无过敏史

经过数据分组及代码转换处理后，得到满足条件的样本数据集如表 4 示，以此作为挖掘数据库 D。

表 4　符合条件的原始数据集

代码	年龄分组	症状	过敏史
T01	A10	Z4,	B2
T02	A9	Z1，Z3	B2
T03	A2	Z1，Z3	B0
T04	A7	Z3	B2
T05	A7	Z1	B2
T06	A3	Z1	B2
T07	A2	Z4	B0
T08	A6	Z1，Z3	B0
T09	A10	Z1，Z3	B2
T10	A8	Z1	B0
…	…	…	…

按照 Apriori 算法处理,设定最小支持度为 20、最小可信度为 75%，扫描数据库 D，得出表 5

表 5 1-项集数据集合

1-项集	支持度(%)
A2	44.74
Z1	89.47
Z3	81.58
B0	20.18
B2	67.00
B3	41.70

表 5 表示在最小支持度为 20%时，产生的 1-项集数据集合。在此基础上，不考虑其它项集，扫描数据库 D，得出表 6

表 6 2-项集数据集合

2-项集	支持度(%)
A2 Z1	44.74
A2 Z3	39.47
A2 B3	35.09
Z1 Z3	66.67
Z1 B3	52.63
Z3 B2	23.68
Z3 B3	46.49

表 6 表示在最小支持度为 20%时，产生的 2-项集数据集合。在此基础上，不考虑其它项集，扫描数据库 D，得出表 7

表 7 3-项集数据集合

3-项集	支持度(%)
A2 Z1 Z3	35.09
A2 Z1 B3	35.09
A2 Z3 B3	29.82

表 7 表示在最小支持度为 20%时，产生的 3-项集数据集合。

产生频繁项集工作到此结束。由频繁 3-项集计算符合最小可信度 75%的关联规则如表 8 所示

表 8 关联规则表

可能的形成的关联规则	可信度(%)
A2->Z1 Z3	78.43
A2->Z3 B3	78.43

2、结果与分析

分析表 8 的结果，我们得到这样的一些强规则：在支气管哮喘病人中以年龄为 1-9 周岁为主，症状表现为咳嗽、气促，且有气味等接触性物体过敏史。

对数据挖掘得出的规则，结合国内外有关哮喘病病因的研究分析，可以分析如下：

支气管哮喘，是一种以嗜酸粒细胞、肥大细胞反应为主的气道变应性炎症和气道高反应性为特征的疾病。 有过敏体质的人接触致敏原，使平滑肌立即发生痉挛，引起哮喘。[9]

据有关资料显示,我国成人哮喘发病率为

删除的内容:

1、 3 讨论

充分利用电子病历系统中的宝贵的临床医学信息资源，建立临床诊疗信息采集平台，进而开展数据挖掘是大有可为的[10]。

本文提出了运用计算机数据仓库处理技术，对已经数字化的中医临床病历原始资料进行数据挖掘整理，旨在通过大量临床数据分析归纳出辅助诊疗决策的结果[1]，探讨并建立一种适用于中医医学诊疗规律辅助发现的有效研究方法和可操作的技术平台。

本文演示了用数据挖掘的方法对支气管哮喘的病人的年龄、症状、过敏史等关系进行关联分析，结果发现某种疾病中，年龄与过敏史、症状之间存在一定的相关关系。如能将这些关联规则置于临床诊疗知识库，医生在临床实践中，遇到相关的病种符合关联规则时，系统可提示医生，然后医生可根据专业素质来判断是否正确[11]，对临床辅助诊疗有一定的现实意义。

本研究的方法如应用于医院的其他学科，对更多的病种的病因、病机进行分析研究，可拓展临床诊疗知识库的内容，使其对临床辅助诊疗的支持力度更大。

本研究与其他的中医临床诊疗研究比较具有几个方面的优越性：①数据直接来源于受法律约束的、海量的、临床病历数据，不需要人工二次录入，避免了人工资料整理工作时的不慎，因此能比较忠实反应临床工作者的诊疗实践；②课题组开发的《中医辅助诊疗系统》能够满足自由地、快速地定制各种复杂组合条件的数据提取要求，系统可以多层次、高效地整理、归纳临床病历的信息，为比较客观地、全面地发现临床诊疗规律提供操作平台。③本研究将数据挖掘方法用于从疾病、年龄、症状、过敏史之间的分析，分析结果发现某种疾病中，年龄与过敏史、症状之间存在一定的相关关系。系统可为日后发现临床诊疗规律，提供了新的思路和可操作技术。它可以使海量的临床诊疗数据变为知识仓库，为医院管理和辅助诊疗决策提供有力的支持。

参考文献：

[1] 梁志伟 . 从临床诊疗术语发现诊疗规律的方法学研究 [J] 广州中医药大学学报 2006 2(1):34-39

[2] 周雪忠，吴朝晖，刘保延．生物医学文献知识发现研究探讨及展望[J]．复杂系统与复杂性科学，2004，1(3)：45-55.

[3] 刘晋平，黄宇虹，陆小左．数据挖掘在中医脉诊中的应用[J]． 天津中医学院学报，2003，22(3)：9-10.

[4] 贺宪民，吴骋 ，于长春，等．数据挖掘技术在医学领域中的应用[J]．第二军医大学学报，

2003， 24(11)：1250-1252.

[5] 世界卫生组织(北京协和医院世界了卫生组织疾病分类合作中心编译). 疾病和有关健康问题的国际统计分类(ICD-10) [M]. 人民卫生出版社，第一版, 1996.

[6] 石义芳，孔令人，于芳，陈培正. 数据挖掘和知识发现技术在病人流量分析中的应用 现代预防医学，2006 33(2): 237,243

[7] 王波，张斌，魏伟杰，马玉慧、 面向中医辨证规范的交互式数据挖掘框架 世界科学技术—中医药现代化*思路与方法，2006 8(1): 24-30

[8] 包昌火,谢新洲主编 信息分析从书_数据仓库和数据挖掘 [M] 北京 清华大学出版社 2006.4

[9] 李明华,段凯生,朱桂全 哮喘病学[M] 北京 人民卫生出版社，第一版，1998，15-18

[10] 胡镜清，刘保延，王永炎. 中医临床个体化诊疗信息特征与数据挖掘技术应用分析[J] 世界科学技术：中医药现代化，2004，6(1)：14-16.

[11] 王华，胡学钢. 基于关联规则的数据挖掘在临床上的应用 [J] 安徽大学学报（自然科学版），2006.3，30（2）21-25

[12] 李晓毅，徐兆棣. 关联规则挖掘在医疗诊断上的应用 [J] 辽宁师范大学学报（自然科学版），2006.6，29（2）133-135

本文属广东省科技厅资助课题项目（编号 C10115）