关联规则在儿童门诊信息分析中的应用*

李 准 严少彪 余 颖 曾 姝

宋 萍 赵文龙

(重庆医科大学附属儿童医院病案统计科 重庆 400014)

(重庆医科大学医学信息学院 重庆 400016)

[摘要] 以重庆市大型儿童综合医院 2015 年门诊病人信息为数据源, 运用 SQL Server 2008 R2 对数据进行预处理, 使用关联规则技术挖掘出症状体征与挂号科室之间的关联规则, 为门诊病人选择挂号科室提供决策依据, 为后续研发门诊病人自助分科系统奠定基础。

[关键词] 关联规则; 数据挖掘; 门诊信息; 症状体征

[中图分类号] R-056 **[文献标识码]** A **[DOI]** 10.3969/j.issn.1673-6036.2016.08.017

Application of the Association Rule in the Analysis of Children Outpatient Information LI Zhun, YAN Shao-biao, SHE Ying, ZENG Shu, Medical Records Department of the Children's Hospital of Chongqing Medical University, Chongqing 400014, China; SONG Ping, ZHAO Wen-long, College of Medicine Information of Chongqing Medical University, Chongqing 400016, China

[Abstract] Taking the outpatient information of large comprehensive children hospitals in Chongqing as data sources, the paper conducts data preprocessing with SQL Server 2008 R2, uses the technology of association rule to excavate the rule of association between the symptoms and signs and the registered departments. It provides a decision-making basis for outpatients to select the department to be registered and lays a foundation for subsequent research and development of the self-help department diversion system for outpatients.

[Keywords] Association rule; Data mining; Outpatient information; Signs or symptoms

1 引言

数据挖掘是致力于数据分析和理解, 揭示数据内部蕴藏知识的技术, 已成为未来信息技术应用的

重要目标之一。近年来, 随着医院信息化水平的不断提高, 门诊信息资源不断增加, 门诊信息数据挖掘的重要性日益凸显。关联规则是数据挖掘领域中成果颇丰而且比较活跃分支, 其概率、支持度和重要性是判断规则有效性、实用性、关联程度的常用指标, 三者缺一不可^[1]。在医学领域目前国内外主要将关联规则技术应用于探索住院病人疾病与症状之间的关联关系^[2], 而有关门诊信息的相关研究极少^[3-4], 尤其是儿童门诊信息。因此, 本文以儿童门诊病人的症状体征和挂号科室为研究对象, 采用数据挖掘中的关联规则技术挖掘出症状体征与挂号科室之间的关联规则, 为门诊病人选择挂号科室提供决策依据, 为后续研发门诊病人自助分科系统奠定基础。

[收稿日期] 2016-03-28

[作者简介] 李准, 助理统计师, 发表论文 4 篇; 通讯作者: 宋萍, 高级统计师。

[基金项目] 重庆医科大学医学信息学院院级课题“儿童门诊信息数据挖掘与应用研究”(项目编号: 2014A001); 重庆市科委社会事业与民生保障科技创新专项“基于大数据和 DRGS 的疾病编码辅助支持系统研发与应用研究”(项目编号: cstc2016shmxzx120011)。

2 资料与方法

2.1 资料来源

资料来源于重庆市某大型儿童综合医院 2015 年 2 479 773 条门诊病人信息, 包括病人住院号、年龄、性别、就诊时间、挂号科室、门诊诊断名称、主诉、简要病史、查体 9 个属性。经过数据预处理后得到 2 334 697 条有效记录。

2.2 研究方法

(1) 数据收集。运用 SQL Server 2008 R2 备份 2015 年门诊病人信息, 将备份文件还原到本地数据库中, 以备后续数据处理。(2) 数据属性。根据《疾病和有关健康问题的国际统计分类第十次修订本 (ICD-10)》^[5] 规定选取 394 个常见症状体征建立症状体征标准库; 在原有 36 个挂号科室的基础上梳理出 14 个科室建立挂号科室标准库, 关联规则数据模型, 见表 1。(3) 数据预处理。利用 SQL Server 2008 R2 对原始数据进行预处理, 包括数据清理、数据集成、数据规约和数据转换。主要步骤为: 对挂号科室进行标准化处理; 将每个症状体征与每一条记录的简要病史进行模糊匹配, 获得所需结果; 转换为适合关联分析的数据模型。共获得 752 117 条记录, 见表 2。

表 1 关联规则数据模型

变量名	属性	域	属性类别
id	住院号	分 2 334 697 类	numeric
ghks	挂号科室	分 14 类	varchar
zztz	症状体征	分 394 类	varchar

表 2 转换后的数据集

住院号 (id)	挂号科室 (ghks)	症状体征 (zztz)
9388271	急诊科	头痛
9388271	急诊科	头痛
9388265	内科	发热
9388265	内科	咳嗽
9388262	内科	咳嗽
.....		
9388252	特需门诊	喷嚏

2.3 关联规则分析

结合本研究, 支持度 (Support) 是指症状体征 A 和挂号科室 B 同时出现的频率, 用来度量一个项集的出现频率。支持度高, 说明规则普遍性较高, 反之则较低。概率 (Probability), 也称置信度 (Confidence), 是指出现症状体征 A 的患者中挂号科室为 B 的概率。概率高, 说明症状体征 A 出现挂号科室为 B 的可能性高、可信度好; 反之则可能性低、可信度差。重要性 (Importance), 也称作用度、提升度、兴趣度分数, 是指症状体征 A 出现对挂号科室 B 出现的影响程度, 也是可能性提高的倍数^[6-7]。若重要性大于 1, 说明二者为正相关, 症状体征 A 出现则挂号科室 B 刚好可能出现; 若重要性等于 1, 说明是两个独立事件 (规则无意义); 若重要性小于 1, 说明二者为负相关, 症状体征 A 出现则挂号科室 B 不太可能出现, 本文仅研究正相关的情况。运行 SQL Server Business Intelligence Development Studio 9.0, 创建新项目, 新建数据源及数据源视图, 创建关联规则挖掘结构, 将住院号勾选为键列, 症状体征和挂号科室勾选为输入列, 挂号科室勾选为可预测列, 设置测试数据百分比为 30%, 测试数据集中的最大事例数为 1 000; 设置规则最小概率为 0.4, 最低重要性为 1; 设置项集支持为 7 521 (即支持度为 1%), 最小项集大小为 2, 最大行数为 2 000。

3 结果与分析

3.1 频繁项集

从表 3 看出咳嗽、发热、喷嚏、腹痛、皮疹 5 个症状体征与内科、急诊科、特需门诊、知名专家 4 个挂号科室关联性较强。出现咳嗽、挂号科室为内科的病人有 181 185 人, 同理可推测其他频繁项集含义。

3.2 关联规则

通过关联规则分析结果获得概率大于 40%、重要性大于 1 的规则共 9 条。第 1、2 条规则概率大于

80%，说明可信度非常好；第 4 条规则的支持度最高，为 1.89，说明普遍性高；第 1、4 条规则重要

性较高，均大于 2，说明关联程度较高，规则发生的可能性均提高了 2 倍以上，见表 4。

表 3 2 项频繁项集

支持	项集大小	项集	支持	项集大小	项集
181 185	2	zztz = 咳嗽, ghks = 内科	16 576	2	zztz = 腹痛, ghks = 内科
72 967	2	zztz = 发热, ghks = 内科	13 742	2	zztz = 咳嗽, ghks = 特需门诊
52 790	2	zztz = 喷嚏, ghks = 内科	8 204	2	zztz = 皮疹, ghks = 内科
42 400	2	zztz = 咳嗽, ghks = 急诊科	8 026	2	zztz = 咳嗽, ghks = 知名专家
22 432	2	zztz = 发热, ghks = 急诊科	7 814	2	zztz = 喷嚏, ghks = 急诊科

表 4 症状体征与挂号科室关联规则

序号	概率 (%)	支持度 (%)	重要性	关联规则
1	100.00	0.04	3.31	zztz = 不愉快 -> ghks = 心理科
2	94.20	0.24	1.61	zztz = 行走困难 -> ghks = 外科
3	77.20	0.25	1.58	zztz = 口呼吸 -> ghks = 耳鼻喉科
4	62.70	1.89	2.25	zztz = 发红 -> ghks = 眼科
5	58.70	0.12	1.40	zztz = 昏迷 -> ghks = 外科
6	57.20	0.33	1.45	zztz = 鼻出血 -> ghks = 耳鼻喉科
7	56.00	0.02	1.38	zztz = 耳后淋巴结肿大 -> ghks = 外科
8	44.80	0.02	1.34	zztz = 吞咽困难 -> ghks = 耳鼻喉科
9	40.50	0.25	1.25	zztz = 心脏杂音 -> ghks = 外科

4 讨论

4.1 强规则

从 9 条规则中找出两条症状体征与挂号科室之间的强关联规则。规则不愉快 -> 心理科、行走困难 -> 外科的概率和重要性均较高，概率均大于 80%，重要性均大于 1.5，说明有效性和关联程度较高；规则不愉快 -> 心理科支持度非常低，可能是因为心理科本身门诊病人量（占总量 0.48%）较少的原因，与相关研究结果一致^[8]。规则不愉快 -> 心理科表明症状体征为不愉快的病人全部挂心理科，规则行走困难 -> 外科表明症状体征为行走困难的病人中有 94.2% 的病人挂外科，与实际诊疗挂号情况相符。因此，将规则不愉快 -> 心理科、行走困难 -> 外科归为被证实的关联规则。

4.2 一般规则

除两条强规则外，还发现 7 条一般规则。规则

口呼吸 -> 耳鼻喉科、发红 -> 眼科的支持度和重要性均较高，概率均介于 60% ~ 80% 之间，重要性均大于 1.5，说明有效性和关联程度高；规则发红 -> 眼科支持度极高，说明普遍性极高，非常实用。因此，将规则口呼吸 -> 耳鼻喉科、发红 -> 眼科归为被证实的关联规则。规则昏迷 -> 外科、鼻出血 -> 耳鼻喉科、耳后淋巴结肿大 -> 外科、吞咽困难 -> 耳鼻喉科、心脏杂音 -> 外科的支持度和重要性均一般，概率均介于 40% ~ 60% 之间，重要性均介于 1 ~ 1.5 之间，说明有效性和关联程度一般。规则耳后淋巴结肿大 -> 外科、吞咽困难 -> 耳鼻喉科支持度均为 0.02%，说明普遍性极低，实用性差。因此，将昏迷 -> 外科、鼻出血 -> 耳鼻喉科、心脏杂音 -> 外科 3 条规则归为潜在价值的关联规则，将耳后淋巴结肿大 -> 外科、吞咽困难 -> 耳鼻喉科两条规则归为无用的关联规则。

(下转第 88 页)

使用,促进不同卫生相关专业机构和业务部门更加有效地信息共享和协作。此外,还应建立政府和社会互动的大数据采集形成和共享机制,通过政府数据公开共享,引导公众、企业、行业协会、科研机构、社会组织等主动采集并开放数据,鼓励公众、企事业单位发掘利用开放数据资源,开展公共卫生信息创新应用研究,为推进健康中国建设和提升城乡居民健康水平服务。

参考文献

- 董佩,严晓玲,邱五七,等. 公共卫生信息相关群体需求与责任分析 [J]. 医学信息学杂志, 2016, 37 (1): 63-67.
- 马家奇. 公共卫生信息管理现象与策略 [J]. 中国公共卫生管理, 2006, 22 (5): 377-378.
- Jydstrup R A, Gross M J, Cost of Information Handling in Hospitals [J]. Health Serv Res, 1966, 1 (3): 235-271.
- Scott J N, Koo D, Pinner R, et al. Informatics and Public Health at CDC [J]. Morbidity and Mortality Weekly Report, 2006, 55 (Suppl 2): 25-28.

- Haux R. Health Information systems - Past, Present, Future [J]. International Journal of Medical Informatics, 2006, (75): 268-281.
- 代涛. 医学信息学的发展和思考 [J]. 医学信息学杂志, 2011, 32 (6): 2-16.
- 饶克勤,王明亮. 知识管理理论、方法与实践: 知识管理与卫生循证决策 [M]. 北京: 科学出版社, 2010: 165-166.
- 加拿大 Inforway [EB/OL]. [2016-01-10]. <http://www.inforway-inforoute.ca>.
- Heitmueller A, Henderson S, Warburton. Developing Public Policy to Advance the Use of Big Data in Health Care [J]. Health Affairs, 2014, (33): 1523-1530.
- 喻国明,何睿. 健康信息的大数据应用: 内容、影响与挑战 [J]. 编辑之友, 2013, (6): 20-22.
- Khoury M J, Ioannidis JPA. Big Data Meets Public Health: human well-being benefit from large-scale data if large-scale noise is minimized [J]. Science, 2014, 346 (6213): 1054-1055.
- 涂子沛. 大数据: 正在到来的数据革命, 以及它如何改变政府、商业与我们的生活 [M]. 桂林: 广西师范大学出版社, 2012: 276-280.

(上接第 78 页)

5 结语

本研究基于数据挖掘中的关联规则技术共发现 9 条规则, 其中 4 条为已被证实的关联规则, 3 条为潜在价值的关联规则, 两条为无用的关联规则。因此, 除两条无用规则外, 其余 7 条规则将为门诊病人选择挂号科室提供有力的决策依据, 为后续研发门诊病人自助分科系统奠定坚实的基础。

参考文献

- 任仙龙,胡冬梅,王文娟,等. 关联规则在社区居民慢性病患者率分析中的应用 [J]. 中国卫生统计, 2013, 30 (6): 818-820.
- 牟冬梅,冯超,王萍. 数据挖掘方法在医学领域的应用

及 SWOT 分析 [J]. 医学信息学杂志, 2015, 36 (1): 53-57.

- 高馨,田杰,宋萍. 门诊信息数据挖掘研究进展 [J]. 现代医院管理, 2016, 14 (1): 75-78.
- 王帅,林晓东,沈明辉,等. 基于 R 语言的基层门诊用药大数据分析 [J]. 中华医学图书情报杂志, 2015, 24 (3): 54-58.
- 董景五,薛欣. 疾病和有关健康问题的国际统计分类第十次修订本 (ICD-10) 第一卷 [M]. 北京: 人民卫生出版社, 2015. 669-698.
- 李准,冯思佳,赵文龙,等. 关联规则技术在冠心病电子病历中的应用 [J]. 医学信息学杂志, 2015, 36 (1): 58-62.
- 李准. 基于冠心病电子病历的数据挖掘研究 [D]. 重庆: 重庆医科大学, 2013.
- 刘涛. 2007—2014 年沈阳市儿童医院门诊人次构成及变化分析 [J]. 中国卫生统计, 2015, 32 (1): 183.