

基于数据挖掘的院内临床诊断智能分析

罗慧中 赵 光

(郑州大学人民医院信息中心 郑州 450003)

〔摘要〕 随着医院信息化建设的发展,关注重点从信息系统本身的功能越来越多地转向信息系统中的数据利用。医疗数据的利用对健康管理、疾病分析、治疗经验总结、新型技术开发都有重要的意义。借助数据挖掘技术,提取本院相关数据,以智能化诊断为切入点,模拟院内临床诊断知识库的建立,经验证系统判定的关联性与临床实际基本相符,提出可扩展至院内临床路径的可能性。

〔关键词〕 数据利用;数据挖掘;诊断;知识库

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2016.06.013

Intelligent Analysis of Hospital Clinical Diagnosis Based on Data Mining LUO Hui-zhong, ZHAO Guang, People's Hospital of Zhengzhou University, Zhengzhou 450003, China

〔Abstract〕 With the development of hospital information construction, the focus has gradually changed from functions of the information system to data utilization. The utilization of medical data has great significance in health management, disease analysis, summary of treatment experience, and development of new technologies. By data mining technology, the paper extracts data related to the author's hospital, starts from intelligent diagnosis and simulates the establishment of the knowledge base for hospital clinical diagnosis. Upon verification, the relevance judged by the system is basically consistent with the clinical practice. Thus, it proposes the possibility to extend the application to hospital clinical pathways.

〔Keywords〕 Data efficiency; Data mining; Diagnosis; Knowledge base

1 引言

医院的信息化建设已经经历了几十年,现在大多数医院都已经完全依赖于信息系统完成日常工作,大型医院尤其如此。2012 年国家“十二五”医改规划中提出要“建立全国统一的电子健康档案、电子病历、药品器械、医疗服务、医保信息等数据标准体系,加快推进医疗卫生信息技术标准化建

设”。任务层层分解,具体到单个医疗机构就是要围绕电子病历系统,加强信息系统之间的数据集成度,打造以病人为中心的院内电子病历。电子病历系统近几年才开始全国范围的建设,但大多数医院的电子病历只是改变了病历的书写方式,与其他系统之间的交互还远没有达到以病人为中心的程度。

2 郑州大学人民医院信息化建设现状

郑州大学人民医院建院历史有 100 多年,信息化建设始于 20 世纪 90 年代,起初以收费系统为主,2011 年根据医院的规模和发展速度,重建了医院信

〔收稿日期〕 2016-01-18

〔作者简介〕 罗慧中,硕士,发表论文 2 篇。

息系统,升级了检验信息系统,并且引进了电子病历(Electronic Medical Records, EMR)系统和影像存储与传输系统,2013 年出于业务流程需要又陆续引进了移动查房、手术麻醉、重症监护等临床信息管理系统,目前医院信息化依旧处在引进增加系统的建设阶段。和其他大多数医院一样,在选购信息系统的过程中,本着同类当中选最好的思路,各系统采用的都是不同厂商的产品,不同的产品意味着不同的数据结构、开发环境。同时,由于建设早期缺乏基础数据标准的概念,所以不同的系统中还存在着不同的基础数据字典。信息系统积累的数据为医院提供的服务基本上就是报表查询,基于医院信息系统产出,报表内容围绕财务和医院的主要运营指标展开,用于临床科研的数据贫乏,主要原因除了电子病历是非结构化的形式以外,更主要的是各个系统之间缺乏数据上的连通、功能上的整合,这不仅需要技术上的改造,也需要医院业务流程管理的整顿。

3 背景及相关工作

3.1 研究背景

企业服务总线 and 临床数据中心被认为是解决目前医疗信息化烟囱林立的有效手段。企业服务总线解决不同应用服务器的连接问题,消除不同技术架构之间的差异。这两项工作都需要对医院的业务流程和信息系统服务方式进行整合,短时间内还不具备实现的可能性。所以工作在这种环境下,利用现有的数据和 SQL Server 的 Business Intelligence 工具,对医院积累的病历资料进行梳理,建立院内小型的诊断库,并且设计信息系统的插件,保证后续数据的积累,提高判断的正确率。

3.2 数据挖掘概述

数据挖掘(Data Mining, DM)是为了解决数据丰富,知识贫乏状况应运而生的,是从海量数据中获取知识的可靠技术。现在人们已经不再满足对数据库进行简单的查询,而是希望借助现代信息处理技术,能够得到隐藏在数据中反映事物本质和预测

事物发展趋势的有用知识,以这些知识为基础辅助科学决策^[1]。数据挖掘由 3 个步骤组成:准备数据、挖掘规律和规律表示。准备数据是指对原始数据进行初始梳理,提取需要挖掘的数据集;挖掘规律是指用某种算法(聚类、分类、关联规则算法等)对数据进行分析,找出其隐藏的规律;规律表示简单说是以某种用户容易理解的方式展示挖掘结果。数据挖掘算法主要有分类、聚类、关联规则、预测等功能。分类是指从训练数据集中提出一个模型(也可称为分类器),基于此模型将所有数据划分到不同类中。聚类是指在不清楚分类的情况下,根据相似度将数据进行划分。关联规则是指从数据中找出具有价值的数据项关系。预测是指通过历史数据找出数据变化的规律,生成预测模型,根据模型对未来数据的发展趋势进行预测。

3.3 测试原理及步骤

以现有信息系统的数据为数据源,以 C16.000 贲门恶性肿瘤作为研究对象,K25.903 胃溃疡作为参考对照组,按照住院号、姓名、年龄、性别、居住地、住院天数、诊断编码、诊断名称的顺序制表。在医院信息系统中抽取出院主诊断为贲门恶性肿瘤和胃溃疡的病历各 100 份,年龄需要按照联合国卫生组织对于年龄划分的最新标准,将零散的数据化为按段划分横向对比的参数:44 岁以下为青年,45~59 岁为中年,60~74 岁为年轻的老年人,75~89 岁为老年人,90 岁以上为长寿老年人;住院天数划分为 7 天及以下,8~15 天,16~30 天,31 天及以上。依照医院信息系统中的住院号分别在检验信息系统和影像存储与传输系统中抽取阳性的检验结果制表,表头项目和顺序为:住院号、姓名、项目名称、阳性指标。影像结果制表,表头项目和顺序为:住院号、检查部位、影像发现、特征 1、特征 2。从影像存储与传输系统中查询所有病人的影像结果,影像结果与检验结果不同,都是文字的描述,需要对查询到的数据进行二次处理,将描述特征提取出来,如斑块、肿物、强回声等,去掉个性化的字词,如表示位置的上、下,或者度量大小的,如 3mm*4mm 等,使得同一种阳性发现尽可

能用同样的词语描述，方便后期归类。最后从电子病历系统中抽取相应病人的病历文件，查看病人的入院志和首次病程，将查体过程中的异常发现摘取出来制表。表头项目和顺序为：住院号、姓名、病史、查体1、查体2。同时考虑随机病人只患一种病症的可能性较小，所以在查看病历的过程中，也需要借机排除检验信息系统和影像存储与传输系统中不相关的阳性发现，如此次研究的病症为贲门癌，那么病人的心电图、脑血管、血脂、血糖等方面的异常结果就需要作为噪声数据进行排除。将以上数据导入数据库后，还不能直接利用于下一步的分析，描述性的语句作为列值的时候不能直接对比，所以都需要进行语意的转换，将语句转为列，用是否判断性的符号填充进去作为列值，所以要将检查部位和影像发现列合并为一列，然后有异常则将列

值赋值1，无此发现则赋值0^[2]。检验结果导入数据库时为一个检验项目一行，与其他表的一个病人一行数据的格式不符，需要经过行转列的处理，处理后的格式，见表1、表2。

表 1 行转列处理后的检验结果项目及格式

住院号	姓名	项目名称 1	项目名称 2	…	项目名称 N
-	-	阳性指标 1	阳性指标 2	…	阳性指标 N

表 2 合并转换处理后的影响结果项目及格式

住院号	影像特征 1	影像特征 2	…	部位发现 N
-	1/0	1/0	1/0	1/0

此时再进行合并查询得到表3。最终数据库中呈现的数据排列，见表4。

表 3 合并查询表头项目及格式

住院号	年龄	性别	居住地	家族史	…史	查体 1	查体 2	影像特征 1	特征 2	检验项目 1	检验项目 2
-	-	-	-	-	-	-	-	1/0	1/0	阳性结果	阳性指标

表 4 贲门恶性肿瘤数据挖掘合并查询结果

住院号	性别	年龄阶段	住院天数	上腹疼痛	上腹胀	体重下降	反酸	呃逆	呕吐	暖气	纳差	胸骨后痛	腹部压痛
ZY010001350524	M	老年前期	21	0	1	0	0	0	0	0	0	0	0
ZY010001351923	M	中年	24	0	1	0	0	0	1	0	1	0	1
ZY010001350971	M	老年前期	14	1	1	0	0	0	0	0	0	0	1
ZY010001353806	F	老年前期	12	0	1	0	0	0	0	0	0	0	0
ZY010001351265	M	老年前期	18	0	0	0	0	0	0	0	0	0	0
ZY010001349263	F	老年前期	17	1	0	0	0	0	0	0	0	0	0
ZY010001355861	M	老年前期	13	1	1	0	0	1	0	0	0	0	1
ZY010001355790	M	中年	20	0	1	0	0	1	0	0	0	0	0
ZY010001360050	M	中年	7	0	0	0	0	0	0	0	0	0	0
ZY010001362179	M	老年期	19	0	1	1	0	0	0	0	0	0	0

进食 哽咽	食管 肿块	黑便	出血性 体窦炎	糜烂性 胃炎	红斑性 胃窦炎	胃小 弯病变	胃溃疡	贲门 占位	贲门 壁增厚	贲门 糜烂	食管炎	胃幽门 螺旋	白细胞	癌胚 抗原
1	1	0	1	0	0	0	0	1	1	0	0	1	H	H
0	0	0	0	0	1	0	0	0	0	0	0	1	H	L
0	0	0	0	0	1	0	0	0	1	0	0	0	H	L
1	0	0	0	0	0	1	0	0	0	1	0	1	L	L
1	0	0	0	0	0	0	0	1	0	0	0	0	H	L
1	0	0	0	0	0	0	0	1	0	0	0	0	H	H
0	0	0	0	0	1	0	0	0	1	1	1	1	L	H
0	0	0	0	0	0	0	0	0	0	0	0	1	L	L
0	0	0	0	1	1	0	0	1	1	0	0	0	L	L
1	1	0	1	0	1	0	0	0	1	0	0	1	H	L

3.4 试验结果

[illegible]

图1 神经网络算法对贲门癌、胃溃疡的预测结果

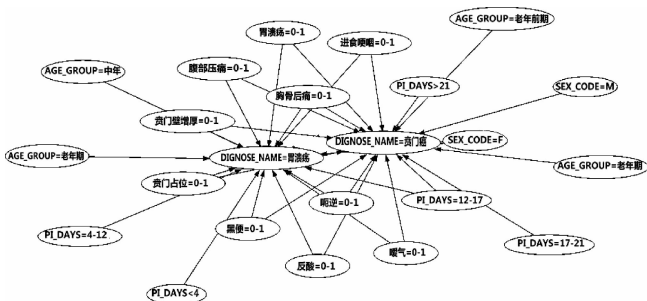


图2 贲门癌、胃溃疡与病症的关联

研究最大的不足是样本数据较小,各属性值在算法分析的结果中所占比重不够精确;其次受医院信息化建设段水平制约,数据提取存在较大的难度,相信这一问题将来可以通过提高电子病历的录入质量和提取效率加以解决,同时也寄希望于院内临床数据中心,通过中心的建立对现有数据重新组织利用;最后受笔者本身对 SQL Server 的掌握程度影响,在选择最优算法时存在盲目性,同时对医学的了解不多,在选择病症和参照对象时也存在不足。

4 结语

以此分析思路可以设计开发院内的诊断特征搜集库作为插件嵌入医生站, 医生通过词组式描述患者的症状, 在获取患者的检验检查报告后录入系统中, 在最终确诊时, 提取这些词组和阳性发现, 加入该诊断所对应的特征列, 这个特征库可用于门户网站和掌上 APP 的智能导诊功能, 方便患者自检。此种方法还可进行扩展, 将从电子病历系统中提取数据的表加入医嘱信息列, 分为用药、检验、检查, 进一步细分病症的等级, 加强与住院天数的关系对照, 分析出本院治疗该类病症的住院天数和治疗方案并探讨优化^[3], 所得结论可以作为医院院内的临床路径, 为临床提供更规范、高效的治疗方法, 也更加方便管理层规范临床治疗行为, 和谐医患关系, 提升医院的品牌形象。

参考文献

- 郭茜, 张淑娟. 数据挖掘和分析在医院管理中应用探讨 [J]. 医学信息, 2010, (6): 1776-1777.
- 罗小楠. 电子病历系统的研究: 信息录入与质量监控电子化实现 [D]. 西安: 第四军医大学, 2005.
- 刘尚辉, 王露, 郑德禄. Apriori 关联规则在甲状腺结节病案分析中的应用 [J]. 中国卫生统计, 2011, 28(2): 178-179.