数据挖掘理念在医院病历随访系统中的应用

王卫东,屈 洋
(暨南大学,广东 广州 510632)

摘 要:通常在医院里每天都会记录大量的医疗信息,数据挖掘技术就是从记录成千上万份病历的病历随访数据库系统中找出患某种疾病病人的共同特征,从而为诊断这种疾病提供依据,或利用数据挖掘技术筛选出的信息来建立具有一定参考价值的某种疾病的医疗方案数据库,以此提高医院的诊断效率和对突发性疾病的爆发及时提供有效的应急措施。详细论述了数据挖掘的理念和如何利用数据挖掘技术实现病历随访数据库有价值的医疗信息的获取的方法。通过数据挖掘技术的使用,大幅提高医疗信息数据库的有价值数据的集成与分析,改变医院病历随访数据库系统的"数据丰富,有效信息缺乏整理"的现象。通过讨论可以看出选择数据挖掘技术不仅可以提高病历随访数据库系统的应用效率而且对突发性流行病的控制与预防能提供极具参考价值的信息。

关键词:病历随访数据管理系统;数据挖掘;知识发现
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2010)07-0199-04

Application of Data Mining Thought in Case History Random Visiting System

WANG Wei-dong, QU Yang
(Jinan University, Guangzhou 510632, China)

Abstract: There are a large of information recorded in the hospital every day. Data mining technique is to find common character about a illness from thousands of case history in the case history random visiting database, and to offer the method of diagnosing this illness, or to build a having reference worth medical treatment pattern database with the effective information filtered by data mining for some illness. Dependent data mining, advances diagnoses efficiency of the hospital and offer an effective emergency measure for the paroxysmal illness. Detailed discuss data mining theory and how to utilize data mining technique to realize the method of getting worth medical treatment information from the case history random visiting database. Through using data mining technique, there is more advance about the analysis and integration of the worth information in the medical treatment information database and change the phenomenon about "data richness, absence settle" in the case history random visiting database system of the hospital. Through discussion, see that the data mining technique is not only to advance the application efficiency of the case history random visiting database system, but also to offer very worth information for control and prevention of the paroxysmal epidemic.

Key words: case history random visiting database management system; data mining; knowledge discover

0 引 言

作为医务人员总是希望从已有的成千上万份病历中找出患某种疾病的病人的共同特征,从而为治愈这种疾病提供有价值的技术支撑。但现有数据库管理系统中的数据分析工具不论是查询、统计还是报表,其处理方式都是对指定数据进行一些简单的或具有一定筛选条件的数据处理,而不能对这些数据所包含的内在信息进行提取。随着病历随访数据库管理系统的数据量激增,科研人员总是希望能够从中提供更高层次的数据分析功能,从而更好地对医疗决策或科研工作提供强有力的支持[1,2]。正是为了满足这种要求,应用于大型数据库的数据挖掘(Data Mining)技术脱颖而出。数据挖掘称为数据库的知识发现(Knowledge Discover Database,简称 KDD),就是从大量数据中提取出可信的、有效的数据的高级处理过程。

收稿日期:2009-12-05;**修回日期**:2010-02-20
基金项目:教育部留学回国人员科研启动基金(教外司留[1999]363号)

作者简介:王卫东(1956-),男,山西人,教授,研究方向为计算机的教学与软件开发应用;屈 洋,教授,从事临床医学和计算机在医学方面的应用研究。

1 数据挖掘的任务

数据挖掘所涉及的学科领域很多,数据挖掘的方

法也在逐步发展与完善。就医学研究而言首先关注的是如何利用数据挖掘发现有价值的数据信息。因此医学领域内数据挖掘的主要任务有：

(1)数据表的关联分析。

关联规则[3]是指两个或两个以上变量的取值之间存在某种规律性，就称为关联，作为数据库的关联则是指两个或两个以上字段的取值之间存在某种规律性。数据关联是数据库中存在的一类非常重要的、可被发现的知识，是数据挖掘的首要任务。关联有简单关联和因果关联等关联类型。数据挖掘的关联分析的目的就是找出数据库中各数据表或数据之间的关联网。在建立关联分析的同时，应当对关联关系引入加权因子，譬如：支持度、可信度、兴趣度和相关程度等参数，使得所挖掘的关联规则更符合需求。

(2)数据的聚类分析。

聚类分析在医学数据库中的应用显得尤为重要，这是因为此手段可把病历随访数据库中的数据按照疾病的相似性归纳成若干疾病类别，通过数据挖掘的聚类分析可以建立各类疾病特征的宏观的概念，找出数据的分布模式，获取可能的数据属性之间的相互关系，为疾病的预防与治疗提供极具价值的数据。

(3)数据变化趋势的预测分析。

预测分析是利用病历随访数据库中的历史数据找出某种疾病的变化规律，建立治疗或预防模型，并由此模型对对应疾病提出可实施性比较高的医疗方案。

(4)数据挖掘的算法。

随着数据挖掘研究的不断深入而产生了各种数据挖掘的算法[4~6]，但大体上可分成：神经网络算法、遗传算法、决策树方法、统计分析算法和模糊集算法等。这里仅简单说明模糊集算法。所谓模糊集算法就是利用模糊集合理论对病历随访数据库中的数据反映出的实际医疗问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析。由于病历随访数据库中数据之间的关联程度高，因此所需解决的问题的复杂性也高，数据的模糊性也越强，目前模糊集合理论多使用云理论(模糊数学模型)对数据进行不精确性、不确定性、不完整性的研究与分析。

2 数据挖掘在医院医疗中的应用

2.1 建立医院病历随访系统原始资料库

医院作为一个庞大的社会医疗保障体系，每年甚至每天都会积累海量的、不同疾病的医疗数据资料，其中大量存储的是各种病人的病历。作为随访系统就要根据大量的电子病历建立随访信息库并以此为基础进行有价值的数据挖掘，既可对病人病情发展做后期追踪调查，也可对同类疾病作预防性治疗方案，这对日后同类疾病的医疗方案的实施有重要的指导意义。

随访信息库中应包含病人的基本资料、医生信息、疾病检验、医疗诊断、住院医嘱、手术情况、护理信息、病情转归、病人出院情况记录、出院后的用药记录、病人的复诊情况等信息。譬如医生信息包括科室、姓名、职称、医疗特长、诊疗时间等；医疗诊断信息包括治疗日期、主治医师姓名、病情记录、医嘱、治疗方案等(见表1)。将以上信息纳入随访信息数据库中，是利用数据挖掘技术对病历随访数据库系统进行信息的科学分析的前提和保障。即根据病历随访数据管理系统的原始资料，运用适当的挖掘算法进行信息的清理，清除病患随访信息中无效的数据，提取有价值的数据。

表1 病历随访数据库信息表例

病人信息表主要项目	治疗信息主要项	病理检查主要信息项	医生信息主要项
病历编号	病历编号	病历编号	医生编码
身份证号码	疾病名称	疾病名称	姓名
姓名	发病时间	体温	出生年月
性别	就诊前症状	血压	性别
年龄	目前症状	心律	学历
身高	就诊前用药效果	血常规检查	职称
体重	现治疗方案记录	尿常规检查	工作年限
血型	现处方记录	血生化检查	科室
家族病史	曾检查项目结果	X－检查	科室电话
本人病史	病理检验结果	超声波检查	移动电话
药物过敏记录	医生编码	心电图检查	医疗特长
经治医生姓名	医生姓名	CT检查	备注
备注	备注	内窥镜检查	

2.2 基于病历随访数据库系统的数据挖掘

如前所述病历随访数据库系统蕴藏着海量数据，其间随使用目的的不同，必存在相当数量的无效数据，譬如：不做病例追踪时，病人的联系电话就可能成为无效数据。所以通过现代的数据挖掘技术对病历随访数据库中的原始数据库进行数据分析、清理和挖掘，剔除无效数据，筛选有价值的信息，这些信息的逐渐积累就会形成一个大规模的、有价值的疾病治疗信息库，最终形成医疗信息的共享，这种共享不仅使各种疾病在治疗过程中有所借鉴，同时对于突发性流行病的治疗与预防也可有所借鉴。

数据挖掘与传统数据分析方法不同的是数据挖掘使用的是基于知识发现的方法，运用模式匹配和其它算法决定数据之间的重要联系。

知识发现的过程通常由以下一些步骤构成：

(1)数据清理；

(2)数据集成；

(3)数据选择；

(4)数据变换；

(5)数据挖掘;

(6)模式评估;

(7)知识表示。

在实现上述数据挖掘进程中数据的关联规则分析[7]十分重要,例如数据清理时数据的选用与噪声分析均与关联规则的挖掘密切相关,以表 1 所描述的病人信息表的项目为例,分析糖尿病病人群体分类的数据挖掘,首先将病历随访数据库中病人信息表内性别、年龄、体重、家族病史项作为一个关系组 $r1$,其中性别、年龄、体重、家族病史作为这个关系组的属性,即 $r1\{a11,a12,a13,a14\}$,此时已将病人信息表中的无关信息(如:身份证号码、姓名、血型、药物过敏记录、主治医生姓名)剔除,而对于保留的四项属性做噪声分析则可认为性别项对糖尿病病人群体分类影响因素的权重最小,可以剔除,最终以对糖尿病病人群体分类影响因素最大的三项属性(年龄、体重、家族病史)构成所需数据挖掘关联规则 $r1\{a11,a12,a13\}$,其中属性 $a11$、$a12$、$a13$ 分别对应年龄、体重、家族病史,据此得出病历随访数据库所有病人信息表中关于此关系的关联规则的集合。设病历随访数据库中有关病人信息表的集合为:$D=\{t1,t2,\cdots,ti,\cdots\}$,每个 $ti(i=1,2,\cdots,n)$ 都是 D 上的一个子集(即每个数据表 ti 都是数据库 D 的内表),$R=\{r1,r2,\cdots,ri,\cdots\}$ 是一个关系规则项的集合,每个 ri $(i=1,2,\cdots,n)$ 都是 R 上的一个子集,此时每个 ri 的描述为 $ri\{ai1,ai2,ai3\}$ 或用 $ri\{aij\}$ 描述,其中 $j=(1,2,3)$,显然每个 aij 代表每个 ri 子集中的年龄、体重、家族病史三项属性之一,具体属性项则由 j 的值决定,于是有关糖尿病病人群体分类的数据库就由数据挖掘得出的这一系列具有唯一标识 TID 的关系规则所关联的数据构成,为下一步糖尿病病情的预防与治疗提供了有价值的信息。

进行病历随访数据库数据挖掘的主要目的是对常见疾病提供可行的治疗方案和对突发性流行病的预防和治疗提供有价值的信息,因此在对数据库数据作相关分析的同时必须采用适当的算法,使提供的数据的可信度达到预期值以满足医疗层面的决策需求和医疗质量的有效保证。常用的算法有决策树、贝叶斯、神经网络和模糊集等。

决策树是一个类似节点流程图的树型结构,其中每个节点表示在一个属性上的测试,从节点流出的每个分支代表一个测试输出。仍以糖尿病病人群体分类为例,决策树算法中以年龄为节点,其测试输出按传统医学分析习惯只考虑三个分支:老年、中年、青年,而当今随着肥胖儿的不断增多次分支必增加儿童分支。决策树上的节点处每个树叶代表该属性的类或类分布。决策树是通过预先准备好的、已知的历史数据建立起来的,然后利用病历随访数据库提供的数据对建好的决策树进行测试,通过关联规则中各加权因子的权重调整使决策树符合在病历随访数据中数据挖掘的需要。

贝叶斯分类是统计学分类算法[8]。它通过大量数据的样板空间预测病历随访数据中数据成员之间存在关联规则关系的可能性,显然贝叶斯算法在遗传学中研究遗传对高发性疾病的影响分析中起着很大的作用。仍以糖尿病病人群体分类为例,给定样本属于一个特定类(糖尿病的家族史)的概率,现代医学据此早已给出了糖尿病的家族史对家族后代成员的影响的结论。

神经网络算法[9,10]是在对人脑神经网络的基本研究的基础上,采用数理方法和信息处理的角度对人脑神经网络进行抽象,并建立的某种简化模型。由于它具有自适应性、并行处理能力和非线性处理的优点,所以在医学领域的信号处理、特征提取、模式识别等方面被广泛应用。例如非线性处理是指在对已经挖掘出的病历随访数据中的数据进行分析时,数据出现异变的处理决策。

模糊集[11]是针对具有不精确性、不确定性、不完整性的现存随访数据的数学研究工具。基于模糊集理论的数据挖掘技术,通过大量经过验证的历史数据得到的模糊算法模型能够有效用于疾病诊断,尤其在早期肿瘤的影像分析中根据模糊集数据挖掘分析所得的结论可以给可信度极高的诊断。

通过上述分析与介绍,不难看出数据挖掘技术的应用几乎涵盖了涉及医疗各个领域的全部数据资源,譬如临床医疗信息和医院管理信息。

3　医学领域内数据挖掘发展趋势

医学领域内的数据挖掘是面向整个医学数据库或医学信息集合提供知识和决策,它已经是医疗决策支持系统的重要组成部分[12]。病历随访数据库系统所构架的数据仓库完成了各类数据的采集、储存和管理等工作,数据挖掘技术则面对经过初步加工的数据事先跟为专注的知识发现。

在知识发现这个层面上数据挖掘研究发展方向集中到以下几个方面:

(1) 基于约束的数据挖掘方法,以此提高数据挖掘的总体效率。

(2)加强对各种非结构化数据的挖掘,譬如对多媒体数据的提取。

(3) 发现知识的数据挖掘过程可视化处理,如此

有利于在数据挖掘过程中的人机交互。

(4)数据挖掘中的隐私保护,由于医疗信息对于病人来讲属于个人的隐私,因此在数据挖掘过程和共享分析结果的同时必须保证病人的信息的安全性。

总之医学领域内的数据挖掘是一门涉及面广、技术难度大、数据繁杂的交叉学科,需要从事该项研究的科研人员通力合作才能实现数据挖掘在算法的高效性和准确性等关键技术方面有所突破。

4 结束语

针对病历随访数据库系统中海量的医疗数据进行数据挖掘,以此实现提供疾病诊断、管理决策分析是一种发展趋势。建立符合医疗过程的数据库结构是基础,采集、存储所有与医疗相关的信息是数据挖掘能否正确实现的前提保证,选择适合医疗数据类型的数据挖掘工具及挖掘技术是能否最终提供有价值的决策信息的关键,文中所介绍的数据挖掘理念针对医学数据所具有多态性、不完整性、较强的时间性、复杂性、冗余性和不一致性等特点的数据挖掘是成功的,能够满足病历随访数据库系统的数据挖掘需求。

参考文献:

[1] Kamber M, Han J, Chang J. Metarule-guided mining of multi-dimensional association rules using data cubes[C]//In Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining(KDD'97). [s.l.]:[s.n.], 1997:207-210.

[2] 王 华. 数据挖掘在医学上的应用[J]. 安徽医药,2008,13(8):746-748.

[3] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]// Proceedings of the ACM SIGMOD Conference on Management of Data. [s.l.]:[s.n.],1993:207-216.

[4] 孟凡荣. 基于云理论的煤矿安全监测数据关联规则挖掘[J]. 小型微型计算机系统,2008(9):1622-1626.

[5] 程舒通. 关联规则挖掘技术研究进展[J]. 计算机应用研究,2009(9):3210-3213.

[6] 王 华. 医学数据挖掘中的数据预处理与 Apriori 算法改进[J]. 计算机系统应用,2009(9):94-97.

[7] 孙 明. 基于层次关联规则的日志本体事件领域关系学习[J]. 计算机应用研究,2009(10):3683-3686.

[8] Cho Sungbin. A linear Bayesian stochastic approximation to update project duration estimates[J]. European Journal of Operational Research,2009,196(2):585-593.

[9] 柳炳祥. 基于云理论与神经网络集成的模糊系统[J]. 计算机应用,2008(2):305-306.

[10] 刘海燕. 一个基于神经网络的信息系统安全性综合评估模型[J]. 计算机工程与科学,2008(11):16-18.

[11] 朱金伟. 基于数据挖掘的中医药数据预处理方法[J]. 计算机工程,2006,32(15):280-282.

[12] 徐 刚. 数据挖掘及其在医学领域中的应用和展望[J]. 实用临床医学,2006(11):196-198.

(上接第 198 页)

利用了 Google Maps 提供的地图服务,灵活地实现了用户对于个人简历的在线浏览、修改和发布的需求,也使招聘单位和网上好友对你有更直观、更立体的了解。

由于 Google 地图服务在互联网上获得极大的成功,各大公司也争相模仿[11],相应推出了各自的地图服务。国内的知名网站也陆续推出了自己的地图服务,其中有搜狐旗下的搜狗和百度的百度地图。Google 地图服务的成功不仅取决其创新性,还在于其开放性。广大的开发者使用 Google Maps 应用到网络的方方面面[12],相信 Google 地图服务在今后的网络应用中还将继续大放异彩。

参考文献:

[1] 巫细波,胡伟平. Google Maps 运行机制以及应用研究[J]. 华南师范大学学报,2009(2):106-110.

[2] 一帘幽梦. 巧用 Google Maps 做地图导航[J]. 电脑迷,2007(6):88-89.

[3] 戴 兵. 基于 Google Maps API 的校园地图设计[J]. 电脑知识与技术,2008(2):184-185.

[4] 王天亮,陈 刚,徐宏炳. 基于共享数据库的数据共享技术[J]. 计算机工程与设计,2007,28(8):1923-1926.

[5] 汪明申,王 强. Mashup 系统构建研究[J]. 现代图书情报技术,2009(5):34-38.

[6] Jackson C, Wang H. Subspace: Secure cross-domain communication for web Mashups[C]//Proceedings of the 16th International Conference on World Wide Web WWW'07. New York:ACM,2007:611-620.

[7] Merrll D. Mashups: 应用程序新成员[EB/OL]. 2006. http://www-128.ibm.com/developerworks/cn/xml/x-mashups.html.

[8] 成 富. 使用开放 API 和工具快速开发情景式 mashup 应用[EB/OL]. 2009. http://www.ibm.com/developerworks/cn/web/0910-chengfu-mashup/.

[9] Yee R. Web 2.0 Mashup 开发实践[M]. 唐扬斌,译. 北京:人民邮电出版社,2009:306-314.

[10] 刘治国,王育坚. 浅谈向 Google Earth 发布 3D 模型的方法[J]. 信息技术,2009(6):25-27.

[11] 吴永杰,晏金成. 基于 Google maps 的租房信息网站[J]. 软件导刊,2009,8(3):75-76.

[12] Norstrom M. Geographic information system(GIS) as a tool in surveillance and monitoring of animal diseases[J]. Acta Veterinaria Scandinavica Supplementum,2001,94:79-85.