数据挖掘在合理用药信息分析中的应用

傅　翔[1], 杨樟 卫[2], 陈盛新[1] (第二军医大学: [1]药学院; [2]长海医院, 上海 200433)

摘要　介绍了数据挖掘的基本概念、过程、常用方法, 并从治疗方案、处方模式、药物监测、不良事件监测和成本控制等方面, 说明数据挖掘在合理用药信息分析中的应用。

关键词　数据挖掘; 合理用药; 信息分析
中图分类号: R952　　文献标识码: A　　文章编号: 1006- 0111(2009) 06- 0411- 04

在公共卫生领域, 合理用药已成为一个令人关注的问题。不合理的用药引起药物不良反应、耐药、延误病情、甚至造成病人死亡。从决策论的角度看, 用药实质上是一个在人 (患者、医药人员)、疾病和药物之间的信息运动过程[1]。信息技术的兴起和发展已使卫生保健发生彻底的变革, 临床用药过程每一刻都在产生着大量的数据信息。通过对这些数据信息的分析、挖掘和利用, 应该能发现新的知识、新的问题, 推动药物安全、有效、经济的使用。

由于卫生领域数据的特点 (海量、非线性、高维度、干扰强等)[2], 有必要采用新的信息技术, 自动、智能地对临床用药过程产生的信息进行分析。数据挖掘 (Data mining)就是一种很有前景的智能信息技术。

1　数据挖掘技术概述

1 1基本概念　数据挖掘, 也称为数据库中的知识发现。1989年, 在美国人工智能协会会议上提出了数据库知识发现 (KDD)的概念[3], KDD是指从数据库中发现知识的全部过程, 即识别出存在于数据库中有效的、新颖的、具有潜在效用的数据和信息, 加工成可理解的概念、模式、规则[4]等的高级处理过程。严格地说, 数据挖掘只是知识发现的一个步骤, 但在实际使用中, 对两者往往不加区别。

1 2 知识发现的基本过程　知识发现的过程如图 1[5]表示, 可描述为: ①数据准备; ②数据选择: 检索和分析任务相关数据; ③数据处理: 清理和集成; ④数据变换: 数据变换成适合挖掘的形式; ⑤数据挖掘: 使用智能方法发现或提取数据模式; ⑥模式解释与评估: 模式的价值体现在对分析者的有趣度和未知度, 根据某种度量, 识别表示知识的真正有趣的模式, 评价其结果, 解释其价值; ⑦知识表示: 使用可视化等形式向用户提供挖掘到的知识。

值得注意的是, 知识发现并不是完全单向的, 包含了大量的反馈, 多个步骤之间相互影响、反复调整, 形成一种螺旋式上升过程; 并与使用者的决策步骤密切相关[6]。

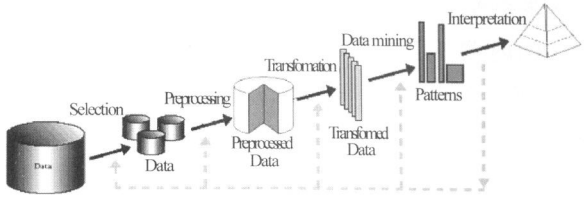

图 1　知识发现的过程

1 3 数据挖掘常用模式及方法[7]　数据挖掘技术从一开始就是面向应用的, 它不仅是面向特定数据库的简单检索、查询、调用, 而且要对这些数据结构和内容调整, 进行微观、中观乃至宏观的统计、分析、综合和推理, 以指导实际问题的求解, 形成知识模式。模式是对客观事物的一种抽象描述, 是整个数据集的全局性描述, 常用模式有关联、分类、聚类等。各种具体挖掘方法是数据挖掘功能实现的前提和基础。

1 3 1　关联 (association)规则　关联规则反映了一个变量与其他变量之间的相互依存性和关联性, 一般用 4个参数来描述: 可信度 (confidence)、支持度 (support)、期望可信度 (expected confidence)、提高度 (lift)。关联规则的典型方法如"购物篮分析", 因最初用于分析顾客购物篮中所买商品的关联规则, 从中发现顾客购买习惯而得名。

1 3 2　分类 (classification)　在数据挖掘中, 分类是指利用恰当的算法, 对分析对象进行类型区分。分类方法有统计法 (贝叶斯法等)、机器学习法 (决策树等)、神经网络法 (BP算法等)。

1 3 3　聚类 (cluster)分析　聚类是对一个数据对

基金项目: 上海市重点学科建设项目资助(B907).
作者简介: 傅翔 (1972-), 男, 博士生 . E-mail: fqj2000@ hotmail. com.
通讯作者: 陈盛新. E-mail: sxchen@ smmu. edu. cn

象的集合进行分析,将数据集划分为多个类或簇,使同一类中数据对象具有较高相似度,而不同类数据对象具有较大差异度。聚类方法包括划分法、层次法、基于密度方法、孤立点分析等。

1 3 4 预测(prediction) 预测是指利用过去和当前数据对未来数据状态进行预测。时间序列分析和回归分析就是对未来值进行预测。

2 数据挖掘在临床用药决策中的应用

信息的数字化和自动化已成为促进合理用药的重要工具[8]。随着医学信息技术的不断成熟,数据挖掘与现代医疗的结合日益紧密。国内外学者已经应用数据挖掘技术,从治疗方案、处方模式、药物监测、不良事件监测和成本控制等临床用药的多角度开展信息分析,促进药物的合理使用。

2 1 患者治疗方案的制定

2 1 1基于信息论的决策树分类算法属于从特殊到一般的归纳学习方法,即用决策树来表示分类的规则。国内学者介绍了决策树技术在药物治疗中的应用,收集一组患有同一疾病的患者的数据,在治疗过程中,每位患者均对5种药物中的1种有明显的反应。通过研究,血液中钠和钾的比例以及血压都会影响药品的选择。通过决策树算法建模,得到决策树规则。结果表明,对于钠钾比例小于14 642的高血压患者,年龄将决定如何选择药品;对于低血压患者,胆固醇含量是最有效的预测变量[9]。

2 1 2 长期接受药物治疗慢性病(如心血管病)的患者,由于个体差异,其病情的进展情况可能被忽视,因而从临床数据观察病情趋势对治疗有效性至关重要。虽然药物利用评价关注药物相互作用、禁忌证、重复治疗和处方再调剂(refill),但很少同时评价实验室数据和治疗结果。以自组织映射神经网络(SOM)和粗集理论(RST)的联合应用为基础建立模型,如对某教学医院医学数据库中病例的检验结果、治疗药物和给药频率数据进行挖掘[10]。SOM算法用距离函数(distance function)测量数据集间的相似性,数据集包含患者生日、性别、生化检验值(高、低密度脂蛋白,甘油三酯,血糖,糖基化血红蛋白)、药物以及给药频率;药物治疗指标用具有相同药理作用的药物数量表示。该过程能有效地检测出调查期间诊断发生变化的患者,及时提醒医生重新评估患者健康状况,制定治疗方案。

2 2 处方行为模式调查

2. 2 1 为调查抗生素类药物的处方行为,收集2 100名心脏手术患者的感染情况和抗生素处方,将医生按所负责病人的感染发生率分为低感染组和高感染组。采用购物篮分析法和Kernel密度函数估测法,进行处方行为调查。发现医生不同处方偏好(抗生素的品种、数量)会引起患者感染发生率的差异。在抗生素选择中,低感染组使用环丙沙星更多,高感染组的医生更倾向于使用左氧氟沙星。此分析结果有助于医生处方决策和患者治疗结果的改进[11]。

2 2 2 中药在我国和东南亚地区的使用很普遍,从临床角度而言,中药处方行为依赖于医生的经验,难以建立较为一致的中药处方的通用规则。实际使用中也经常出现单味中药、经典配方(复方)协同使用的情形。台湾学者对台湾慢性肝炎患者的中药处方模式进行了挖掘[12]:以2002年台湾国民健康保险(NHI)数据库门诊病人中药偿付要求为对象,按国际疾病诊断分类,鉴别出慢性肝炎患者,对他们的处方进行分析,应用关联性规则评价中药协同处方,分析中药配方之间或配方与单味中药之间的联系。结果表明最常用两种中药组合是加味逍遥散与丹参;最常用三种中药组合是加味逍遥散、丹参和茵陈蒿。这些组合的发现为进一步进行临床试验验证提供了研究方向。

2 2 3 在另一项研究中[13],从2004年的台湾NHI数据库中抽取20万名患者作为研究队列,发现其中46 938名患者使用过中药。用购物篮分析法发掘这些病人中药处方之中单味中药(single herb)之间、与中药配方(herbal formula)之间、或中药配方之间特殊关联。结果表明可信度和提高度最佳的中药组合分别为乳香-没药、夜交藤-酸枣仁汤、当归拈痛汤-舒筋活血汤。该研究提供了台湾中药处方的利用概况,便于进一步验证中药使用的安全性和有效性。

2 3 治疗药物监测

2 3 1 药物引起QT延长并可能导致潜在的威胁生命的心律失常(如尖端扭转型室性心动过速),部分抗菌药物(如格雷沙星和司帕沙星)因此受到监管部门干预而退出市场。为深入了解其他抗菌药物的潜在心律失常作用,避免潜在的风险。研究人员首先通过文献数据挖掘,基于可获取的证据,按引起QT延长的临床相关性证据强度将抗菌药物分为5类,结果共有21种抗菌药物有引起QT延长的风险;其中6种氟喹诺酮和3种大环内酯类抗菌药物归入临床证据强度最高类。利用ESAC(欧洲抗菌药物消耗监测)方案提供的14个欧洲国家长达8年(1998~2005年)的抗菌药物使用数据(以DID即每1000个居民日的DDDs表示),挖掘并分析人群对这些抗菌药物的暴露情况和利用趋势,推进了抗菌药物使用的风险收益评估和合理使用[14]。

2.3.2 抗菌药物的使用与细菌耐药性之间存在着关联。在特定的生态系统中(如医院)除了存在着各种复杂因素(如多种抗菌药物联合应用、选择偏倚等)外,要证实两者之间的关联性,还必须考虑序列观察点间可能的联系,尤其是随时间变化(如延迟效应)的联系。时间序列分析是在历史行为基础上,预测未来表现,解释特征和影响因素,适用于固定间隔重复数据测量,且间隔远短于整个观察期的研究。研究者运用数据挖掘技术中的时间序列分析,采用自回归移动平均模型(ARMA)和传递函数模型(transfer function models)分析抗菌药物使用(每 1000住院日的 DDDs)和耐药性(耐药菌月百分比)数据,证实其时间的关联性,量化抗生素使用对耐药的作用,估计抗菌药物使用变化造成耐药变化的延迟时间,并在以往抗生素使用和耐药数据的基础上预测未来耐药水平[15]。该研究发现,医院生态系统对抗生素的变化趋势比以往想象的快。通过加强对抗菌药物使用系统性监测,采取干预措施调节处方行为能够降低院内感染的发生[16]。

2.3.3 通过数据挖掘,提出实验室检验数据与药物利用数据间的关联假设,并加以检验,有可能及时发现和监测[17]尚未被认识的药物毒性作用、临床延迟作用等。应用数据挖掘技术可以推进药物警戒(pharmacovigilance)的发展,解决对日益膨胀的药物安全数据库的筛选难题。回顾性调查表明,数据挖掘能先于传统的"信号"发送警示一些显著的医学关联。发达国家和国际监控中心已利用数据挖掘从药物和不良反应的组合数据库中分辨因果关联,进而开展个体病例或人群研究的分析。如 WHO不良反应监控中心的 BCPNN(贝叶斯判别可信区间递进神经网络),美国 FDA的 MGPS(多项伽玛泊松分布缩减法),英国 MHRA的 PRRs(比例报告比之比法)等[18]。

2.4 药物不良事件(ADE)的监测

2.4.1 据统计,45%的药物不良事件是由用药差错引起的,25%的用药差错发生在药品配制和发放阶段[19];而药物名称形似或音似是造成配方错误的重要原因。研究者通过对美国医疗补助申领者数据库(Medicaid claims database)的回顾性分析,评估药名相似引起的配方工作中潜在差错,分析差错发生频率和药品名称相似性之间的关联规则。以编辑距离(信息技术中两个字符串之间相似度的一个度量方法,编辑距离越小,相似性越高)和标准编辑距离算法测量,结果证实,发药差错发生率与编辑距离值呈负相关[20]。该法有助于用药差错的预防,加强对药名相似的药物引起用药差错的深入研究;尤其与诊断信息联合应用后,将对可疑的发药行为发出警报。

2.4.2 为验证标准化的计算机监测能否从门诊的电子医疗记录中发现 ADE的敏感性和特异性,研究人员提取、合并了门诊病人人口特征、医疗记录、诊断名称、过敏史、药历和诊所记录等数据,采用 4种搜索规则,比较其相对作用:①ICD-9诊断名称;②过敏患者的处方开药;③检验数据与已知事件的联系;④医学术语搜索。经临床病例审查验证检索的质量和准确性,结果表明术语搜索对 ADE的检测最为有力[21]。

2.5 医疗成本的控制

2.5.1 在美国,政府医保或商业医保方采用数据挖掘新技术(如 Aetna's claim system),可以在配药环节,即药物发放前,进行前瞻性的评价,提醒药师治疗重复可能发生,便于药师判断取舍,减少同类药物(比如他汀类)重复治疗,已取得良好成本效果[22]。国内学者提议将医保患者配药时间、医保卡号、年龄、性别、配药时间间隔、配药品种或类别、配药数量、药物剂量、疾病诊断、就医科别等数据形成数据仓库,利用数据挖掘中的神经网络技术或决策树分析法对这些变量的权重训练和测试后,建立医保患者配药行为监控模型,及时发现和制止恶意配药的行为[22]。

2.5.2 医疗机构面临着医疗需求和财政支出双重压力。建立卫生保健数据仓库,并应用商业化的联机分析处理(OLAP)和数据挖掘技术,有助于将临床数据和财政数据分析相结合,做出合理的决策。例如,诊断相关组(DRG)是美国医院诊疗保险中采用的定额化支付制度。通过对支出超过定额的诊断组进行数据挖掘,运用患者规则归纳法,以亏损作为输出变量,以住院患者各项属性为输入变量,成功发现了平均成本较高区域,并进一步分析病人年龄、财政情况、身份等引起住院日期延长和医院亏损的因素,促进医疗机构对内部政策进行再评估和调整[24]。

2.5.3以色列学者对哮喘药物的利用建立预测模型[25],联合运用聚类算法、聚类效力评测和决策树分类规则在多年哮喘药物利用数据库中进行知识发现,结果有 3.6%的患者被认为对哮喘药物资源的使用模式为异常或不当。分类结果显示皮质激素类药物(口服和吸入)的使用和患者年龄可作为归纳模型的主要预测因子。

3 结语

合理用药的实现必须基于正确的决策,而正确

(下转第 433页)

扫描法测定[J]. 药学学报, 1990, 25(7): 530.
[6] 吴　波, 马长清, 张寒俊. 反相高效液相色谱法测定蛇床子中蛇床子素的含量[J]. 药物分析杂志, 2005, 25(7): 843.
[7] 翁德新, 李天傲. 蛇床子中 5种成分的 HPLC测定法[J]. 中国中药杂志, 2007, 32(8): 1883.
[8] 杨　宪, 杨水平, 张　雪. 蛇床子药材的高效液相色谱指纹图谱[J]. 药学学报, 2007, 42(8): 877.
[9] 蔡金娜, 徐国钧, 金蓉鸾, 等. 中药蛇床子中香豆素类成分的毛细管气相色谱分析[J]. 中国药科大学学报, 1991, 22(6): 345.
[10] 岳美娥, 牛　翔, 张书圣. 蛇床子及其制剂中香豆素类活性成分的胶束电动毛细管色谱含量测定[J]. 分析测试技术与仪器, 2007, 13(4): 272
[11] 秦路平, 张卫东, 张汉明. 蛇床子生物学及其应用[M]. 成都: 成都科技大学出版社, 1996. 4

收稿日期: 2009-02-26

(上接第 413页)

的决策则依靠于数据的连续分析和对信息及信息背后知识的获取。数据挖掘是认知背后的科学和隐藏的模式, 关联和顺序的发现。相比商业等其他领域, 数据挖掘技术在卫生领域, 如药物利用合理性方面的应用还不普遍。医学研究的数字化增强了卫生人员获取卫生信息的能力, 伴随着数据仓库技术和挖掘技术的发展, 数据挖掘这一融合了数据库、人工智能、及其学习和传统统计学的方法, 在以患者为中心的医学信息发展过程中, 必将显现出巨大的应用潜力。

参考文献:

[1] 汤　韧, 易　涛, 张　宜. 信息技术在合理用药中的应用[J]. 医药导报, 2005, 24(9): 853.
[2] Villmann T. Neuralmaps for faithful data modelling in medicine-state-of-the-art and exemplary applications[J]. Neurocomputing, 2002, 48(1-4), 229.
[3] 于长春, 贺　佳, 范思昌, 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.
[4] 卢启程, 邹　平: 数据挖掘的研究与应用进展[J]. 昆明理工大学学报, 2002, 27(5): 62.
[5] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases[J]. AI Magazine. 1996, 17(3): 37.
[6] Han J, Kamber M. Data mining: concepts and techniques[M]. Morgan Kaufman, 2001, Zuruck.
[7] 张承江主编. 医学数据仓库与数据挖掘[M]. 北京: 中国中医药出版社(第一版), 2008, 北京.
[8] 屈　建. 医院药学中的合理用药[J]. 药学服务与研究, 2004, 4(1): 5.
[9] 靳淑敏, 张翠肖, 孙珊珊. 决策树技术及其在药物治疗中的应用[J]. 科技情报开发与经济, 2008, 18(22): 164.
[10] Chou HC, Cheng CH, Chang JR. Extracting drug utilization knowledge using self-organizing map and rough set theory[J]. Expert Systems with Applications, 2007, 33: 499.
[11] Patricia BC. Choice of antibiotic in open heart surgery[J]. Intelligent Decision Technologies, 2007, 1: 63.
[12] Fang PC, Yen YK, Yu CC, *et al*. Frequency and pattern of Chinese herbal medicine prescriptions for chronic hepatitis in Taiwan[J]. Journal of Ethnopharmacology, 2008, 117: 84.
[13] Shu CH, Jung NL, Chuan FL, *et al*. The prescribing of Chinese herbal products in Taiwan: a cross-sectional analysis of the national health insurance reimbursement database[J]. Pharmacoepridemiology and Drug Safety, 2008, 17: 609.
[14] Emanud R, Elusabetta P, Chlavra Z, *et al*. Exposure to antibacterial agents with QT liability in 14 European countries: trends over an 8-year period[J]. British Journal of Clinical Pharmacology, 2008, 1: 88.
[15] Monnet DL, Lopez JM, Campillos P, *et al*. Making sense of antimicrobial use and resistance surveillance data: application of ARIMA and transfer function models[J]. Clin Microbiol Infect, 2001; 7(Suppl 5): 29.
[16] Fridkin SK, Lawton R, Edwards JR, *et al*. Monitoring antimicrobial use and resistance: comparison with a national benchmark on reducing vancomycin use and vancomycin-resistant enterococci[J]. Emerg Infect Dis, 2002, 8(7): 702.
[17] Gordon D, David K, Josh P, *et al*. Linking laboratory and pharmacy, opportunities for reducing errors and improving care[J]. Arch Intern Med. 2003, 163: 893.
[18] Manfred H, David M, Charles M, *et al*. The role of data mining in pharmacovigilance[J]. Expert Opinion On Drug Safety, 2005, 4(5): 929.
[19] 傅　翔, 栾智鹏, 陈盛新. 用药安全[J]. 药学实践杂志, 2009, 27(3): 236.
[20] Hemant MP, Paul SC, Cathy A, *et al*. Retrospective detection of potential medication errors involving drugs with similar names[J]. J Am Pharm Assoc, 2005, 45: 616.
[21] Benjamin H, David WB. Computerized data mining for adverse drug events in an outpatient setting[EB/OL] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232160/?page=1.
[22] Martin S. Reducing therapeutic duplication successful at the dispensing level[EB/OL] http://www.managedcaremag.com/archives/0808/0808.medmgmt.html
[23] 沈小庆, 盛炳义, 方　曙, 等. 数据挖掘技术在医院药学中的应用[J]. 中华医院管理杂志, 2006, 22(8): 549.
[24] Michael S, Taiki S, Hua CS, *et al*. Case study: how to apply data mining tchniques in a healthcare dta wrehouse[J]. Journal of Healthcare Information Management, 2001, 15(2): 155.
[25] Last M, Carel R, Barak D. Utilization of data-mining tchniques for ealuation of ptterns of athma dugs ue by abulatory ptients in a lrge halth mintenance oganization[EB/OL]. http://portal.acm.org/citation.cfm?id=1336075

收稿日期: 2009-11-13