·医学信息研究·

数据挖掘技术在健康数据分析中的应用*

尚 岑 王东雨 宇文姝丽

（河北大学管理学院 保定 071002）

〔摘要〕 结合健康数据自身的特点，阐述数据挖掘技术用于疾病诊断、治疗及预后评估的优势，探讨现有的健康数据挖掘应用情况以及发展趋势，提出所面临的问题和挑战，为促进数据挖掘技术在医学相关领域中的更广泛应用提供借鉴。

〔关键词〕 数据挖掘；健康数据；医学信息；医学应用

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.05.013

Application of Data Mining Technology in Health Data Analysis *SHANG Cen, WANG Dong－yu, YUWEN Shu－li, School of Management, Hebei University, Baoding 071002, China*

〔**Abstract**〕 Combining with features of health data, the paper describes the advantages of data mining technology in disease diagnosis, treatment and prognosis evaluation. It explores the current situation of applying health data mining and its development tendency, proposes existing problems and challenges and provides reference for wider application of data mining technology in medicine－related fields.

〔**Keywords**〕 Data mining; Health data; Medical information; Medical application

1 引言

随着数字化医疗和医疗信息化在医疗卫生领域的应用，特别是电子病历的广泛使用，数据共享成为可能，也使得健康医疗机构积累起海量健康数据。如何从这些复杂的信息中提取出有价值的信息成为必须要解决的问题之一。健康数据具有冗余性、不完整性、模糊性且带有噪声等特点，决定了健康数据挖掘和其他数据挖掘之间的差异和特殊性[1]。笔者结合健康数据的众多特点，探讨利用数据挖掘技术分析健康信息的优劣势，对健康信息数据挖掘的热点方向和医疗相关领域的应用进行阐述，以期促进数据挖掘技术在医疗领域更广泛的应用和借鉴，使其有效应用于诊断、治疗及预后评估等医疗实践中的各环节，有助于从海量健康数据中提取有价值的知识和规则，从而为疾病的诊断和治疗、医院的决策管理和科研服务提供科学合理的依据，为高效分析、利用健康数据提供新的方法。

2 健康数据分析的必要性

2.1 健康数据的特点[1-2]

2.1.1 异质性且噪声高 医院信息系统中充斥着各种未经处理的数据，所以进行数据挖掘之前需要进行预处理，但是如何保证降噪过程中数据的完整性

〔修回日期〕 2016－03－21

〔作者简介〕 尚岑，硕士研究生；王东雨，硕士研究生。

〔基金项目〕 2016 年河北省社科基金项目“基于行动者网络的河北省医疗卫生信息共享研究”（项目编号：HB16tq0054）。

和结果的可信度，是进行挖掘之前面临的首要难题。

2.1.2 形式和格式多种多样 例如各种文本、结构化的数据表、非（半）结构化文本文档、医疗影像等，需要进行数据格式的处理，不同数据格式需要对应不同的处理方式，增加了数据分析、处理的复杂性，使数据的再利用难度增加。

2.1.3 变量多、专业性强 临床检测数据、医学影像数据中包含大量与疾病相关的属性变量，而且不同患者的属性变量存在诸多差别，要充分挖掘这些专业性强的数据，需要医学知识和挖掘技术都精通的复合型人才，但是数据科学和医学知识学科交融的优势还没有明显的效果。

2.1.4 冗余化 健康数据总是在不断增长的，但是数据量和数据价值并非呈正比的增长关系，由于数据库中存在大量重复的记录，必须通过一定的方法对其进行分析、整合，才能挖掘出对患者诊治、科研、医院管理等有价值的信息。

2.1.5 隐私性 对大规模的健康数据进行挖掘获得知识，帮助进行疾病诊断、药物开发、管理决策分析是一种发展趋势，但是大规模挖掘健康数据也会涉及患者的个人隐私，在隐私保护的前提下，如何保证数据挖掘结果的准确性及其高效算法的实现是值得关注的研究方向。

2.1.6 不完整性 病案和病例是针对某一位具体的患者，而并未包含某种疾病的全部信息，同时记录的信息本身就具有不确定性和模糊性的特点，这些不完整性都会对治疗过程、结果产生重要的影响。

基于以上特点，健康数据挖掘不是一蹴而就的，因为最终结果关乎人的生命，所以必须针对数据特点采取相应的技术手段，排除对最终结果不利的影响，这还有很长的一段路要走，需要学科间的合作交流，同时这些特点也表明自动化、大规模的数据挖掘工具对深度解读蕴藏在这些数据后的新知识非常必要，通过数据挖掘获得的新知识将有力地支撑医疗实践各个环节的决策和判定，降低诊断的误差，提高诊疗效果。

2.2 数据挖掘的优势

2.2.1 提高疾病诊断准确率 复杂疾病的诊断，往往没有一个单一、明确的生理化学指标，医生在多项医学检验数据的辅助下，根据直觉和经验的判断难以避免地出现偏差、错误和过度检查，影响医疗资源合理配置及医疗服务质量的提高等[4]。随着电子健康记录的普及，数据挖掘技术对健康和临床数据的解读不再停留于建立患者症状与疾病的简单联系，而是在挖掘以往大量病历和患者现有临床表现的基础上，建立疾病与大量健康数据的内在关系，根据以往大量同类型疾病的健康和临床数据建立模型，据此模型对新患者做出疾病预测[5]。医疗个性化推荐技术作为医疗数据挖掘的新方向，可以大大地降低医疗误判率，实现精准诊断，提高患者生命安全保障的同时提高医护工作者的诊疗水平，还可以降低相同病例在不同医疗机构由于人员水平差异而导致的诊断差异，从而提高治疗效果。

2.2.2 提高治疗效果和预后评估准确率 复杂疾病和慢性病如糖尿病和癌症等治疗需要漫长的过程，与此相关的数据集可以看作是一个时序数据。在这时序数据中，存在着若干关键的节点，在每个节点中医疗机构需要对下一步的治疗方案包括使用何种药物、给药剂量和用药时间等做出判断，保证每一个阶段的治疗都能够达到最优效果[6]，和疾病诊断类似，判断过程同样掺杂着人为因素，不能保证最终治疗效果理想。数据挖掘技术的介入，某种程度上改善了这种情况，可以根据成千上万患者的病历和治疗历史，建立模型，然后依据患者的目前状态，给出最佳的治疗意见。与提高疾病诊断准确率和提高治疗效果类似，数据挖掘也可以用以提高预后评估准确率，基本思想和实现方法也和前两者相似，即先用海量数据建立模型，然后结合患者现有数据对预后做出趋势预测。

3 数据挖掘技术在健康数据分析中的应用

3.1 公共卫生领域

随着科学技术的进步，数据挖掘技术在公共卫生预测领域的作用正在日益凸显[7]。Google 比美国疾病控制与预防中心提前 1 ~ 2 周预测到了甲型 H1N1 流感爆发，正是数据预测在公共卫生预测应

用中的典型案例；以用户在 Twitter 上的推文以及英国健康保健局发布的城市流感样病例率为数据源，追踪人口接触信息以及人口位置信息，将有助于了解流行病的行为[8]。国内“百度预测”中的疾病预测，依靠百度强大的数据优势，通过主动收集和被动收集两种方式，借助用户搜索对相关疾病数据进行可视化预测。借助百度搜索数据对流感疫情进行监测，是数据挖掘技术在中国公共卫生预测实践中的典型运用[9]，将百度和 Google 疾病预测的不同进行分析对比，能够更加清晰地看出 Google 预测和百度预测在疾病预测中的优劣势[10]。

3.2 精准诊疗

由于医学数据存在异质性、隐私性、多样性和冗余性等特点，不同数据结构、不同目的，需要选择不同的数据挖掘方法。最常用的数据挖掘方法有关联规则、决策树、人工神经网络和聚类分析[11]，根据实际情况选择适当的数据挖掘方法能够更好地揭示医疗数据间的关系和规律，通过疾病间的关联规则、药物的相互作用、基因测序和患者的病史等上下文环境对疾病的发展过程、阶段以及治疗效果进行预测，同时也可根据现实情况进行治疗方案的及时调整，达到疾病的精准诊断、精准治疗，实现患者病情的个性化治疗的目的，有助于降低患者的医疗成本，缓解看病贵问题。就此有人提出以医疗账单为数据源，建立治疗费用、住院时间等数据的预测模型，使用数据挖掘技术发现账单中的异常数据，结合领域专家建立的规则库分析异常账单数据，发现其中可能存在的问题并给出警告[12]。

3.3 远程医疗

随着生物传感器技术的发展，信号处理和通讯技术的广泛应用，远程医疗系统利用网络通讯系统进行异地诊疗的功能得到极好发挥[13]，广大居民足不出户就可以进行健康检测、病情数据采集，方便地享受到专家的诊断和建议，可降低社会运行成本，缓解就医难等状况。当前中国 60 岁以上老年人达到了 2.12 亿，而独居老人的比例约占 1/3 ~ 1/2[14]，远程医疗系统结合现代通讯技术可以实现对老人的身体状态的实时监控，一旦血压、脉搏等触发条件，系统自动发出警报，以便监护中心及时救护，这也是其最新的运用方向[15]。数据挖掘在远程医疗的具体应用有多种不同的表现形式，都可以归纳为 3 类：（1）预警：对突发性疾病进行监控，采集和分析病人的生理参数，对危险状态进行报警。（2）预测：主要用于根据历史数据和现在观察的状况，推测病人在近期或者中远期的状况。（3）智能诊断：当病人出现病症或者不舒服时，根据历史病历和监护数据，诊断相关病症。

3.4 医院信息系统

在提高医疗服务质量中数据挖掘技术需要处理多种多样的健康医疗数据。因此，对这些数据的挖掘必须建立在有效的数据收集、记录和保存的基础上。医院信息系统是医疗数据挖掘技术的重要组成部分，健康数据的复杂性使得开发适用医疗数据与医院信息系统相关的挖掘技术成为紧急任务。医院内部往往有多个不同的应用目标，因此要有与其相对应的挖掘任务。不同的挖掘任务，可能会形成多个对应任务信息组成的定制数据库，针对这些数据库应当采取不同数据挖掘算法，对于影像、信号或者其他非标准化的临床数据，需要进行预处理，但是由于医学信息涉及患者隐私，在进行挖掘时需要进行特别的数据处理，保护其隐私，只有系统的布局思考才能更有效地找出潜在的数据关联，更加科学地进行医疗数据挖掘，增加数据的价值性和有用性[16]。

3.5 医疗健康相关行业

有人认为数据挖掘技术中基于决策树的分类方法适用于我国医疗设备行业的管理情况[17]。通过数据挖掘技术、预警和预测分析可以合理规划企业和员工的规模，提高运行效率，把握和发现市场机遇，规避风险，为企业的经营决策提供支持，增加企业行业竞争力，促进我国医疗设备管理领域综合实力的提高。在医药行业中，可以通过数据挖掘开发更有效的医疗产品，同时还能根据市场需求开发出新的产品；最重要的是可以帮助医药企业发现药物副作用。有作者认为 94% 的不良反应没有被报

告，提出主动检测的方式，其原理主要是利用文本挖掘技术和数据挖掘技术从电子健康档案、电子病历、社交网络、搜索引擎中发掘潜在药品导致不良反应事件来发现药物副作用[18]。利用药品不良反应存在时间先后顺序，挖掘电子病例中可能存在的药物不良反应[19]，另外数据挖掘技术还可以用于医疗保险企业筛分保险索赔，以发现欺诈索赔等。

4 健康数据挖掘的发展趋势

4.1 引入数据挖掘技术的意义

医疗健康领域产生海量的数据，和其他领域相比，相应的数据挖掘工作却显得严重滞后。为提高医疗健康水平，医疗领域需要找到一个能够有效处理大数据的方法。数据挖掘技术在医疗实践中的应用是一个过程，而不是一次性的任务，随着时间的推移，健康数据会越积越多，数据间的关联性也会更加复杂，因此行之有效的医疗数据挖掘技术，对于医学领域的发展有着重大的现实意义。

4.2 个性化医疗推荐系统

个性化医疗推荐系统可以为患者生成专属的治疗方案，充分挖掘患者间、疾病间和医生间的关系，为疾病的科学诊断保驾护航。由于患者的个性差异和疾病间的共性共存，同时在许多层面还存在着某种联系，个性化医疗推荐系统就是运用数据挖掘的方法找出并建立其中的联系，根据关联建立患者、疾病和临床数据间的模型，针对病人的治疗经历、基因、遗传、环境、生活方式等信息挖掘出适合该患者的个性化治疗方案[20]，个性化推荐流程，见图1。

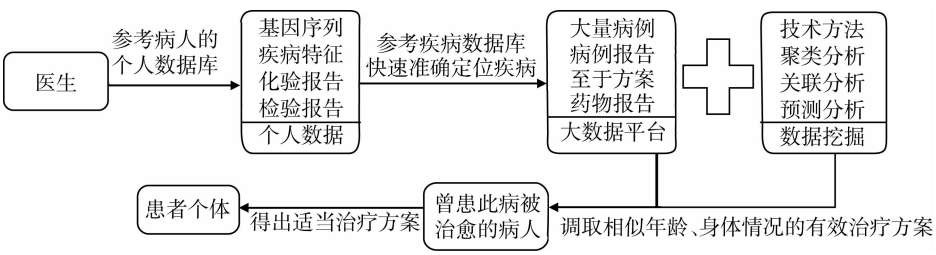

图1 个性化医疗推荐流程

但是个性化医疗方案的实现还面临着诸多挑战，主要在于部分用于疾病预测、疗效预测的数据源难以获得，医院不愿意公开自身独有的数据，信息孤岛现象严重；其次，个人基因分析技术有待突破，基因检测费用昂贵，基因多态性的特质可能导致评估错误及预测错误，致使通过基因检测提供个性化治疗难以获得较高的性价比；政府部门没有相关保障性方案用以推动相关技术的发展，还有资金问题、人才问题，这些都是制约个性化医疗的关键；再者，用户不愿意提交个人医疗数据的部分原因是担心隐私泄露，这就对医疗数据提供商的安全和隐私保护提出了要求。

5 结语

综上所述，数据挖掘技术在健康数据和医疗上的应用有着广阔的前景，但同样也面临着诸多挑战：（1）医院间数据共享共建机制不健全，造成数据孤岛，结果也可能存在很大的偏差，所以打破信息壁垒，建立健全共享机制尤为重要。（2）由于用健康数据挖掘的结果指导医疗实践直接与患者生命相关，所以需要更高的预测精度，开发更准确的算法。（3）医疗数据的复杂性和专业性是个严峻的挑战。（4）大数据的动态变化使得处理的难度增加。（5）界面友好、全自动化和统一的挖掘工具的开发，需要医疗人员和计算机相关专业人紧密配合。（6）数据安全和隐私保护是医疗健康数据处理过程中必须重视且不可回避的。（7）医学和数据科学复合型人才的匮乏，制约着健康数据价值的开发。

总之，数据挖掘技术在健康数据和医疗上的应用还处在起步期，医疗个性化推荐技术也在不断进步完善，国家政策的支持无疑是重大利好，但是如

何解决相关技术应用过程中的各种难题和挑战，笔者认为或许会成为未来医疗健康领域和计算机领域研究的焦点。

参考文献

1 蔡佳慧，张涛，宗文红. 医疗大数据面临的挑战及思考［J］. 中国卫生信息管理杂志，2013，(4)：292－295.

2 颜延，秦兴彬，樊建平，等. 医疗健康大数据研究综述［J］. 科研信息化技术与应用，2014，5（6）：3－16.

3 胡新平. 医疗数据挖掘中的隐私保护［J］. 医学信息学杂志，2009，30（8）：1－4.

4 谢红美，蒋春娣. 提高门诊、急诊病人分诊准确率的方法研究［J］. 特别健康，2013，(12)：14－17.

5 Zou WB，Yang F，Li ZS. How to Improve the Diagnosis Rate of Early Gastric Cancer in China［J］浙江大学学报医学版，2015，44（1）：9－14.

6 孙磊. 健康管理中时序数据挖掘相关问题研究与应用［D］. 北京：清华大学，2012：28－35.

7 Davidson M W，Haim D A，Radin J M. Using Networks to Combine“Big Data” and Traditional Surveillance to Improve Influenza Predictions［J］. Scientific Reports，2015，(5)：1－5.

8 Lampos V，Bie T D，Cristianini N. Flu Detector－tracking Epidemics on Twitter［J］. Machine Learning and Knowledge Discovery in Databases，2010，(6323)：599－602.

9 张洪龙. 基于百度搜索数据的中国流感疫情监测研究［J］. 中华预防医学杂志，2014，48（4）：310－311.

10 百度和 Google 疾病预测有不同［J］. 健康管理，2015，(3)：28－29.

11 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.

12 董诚，林立，金海. 医疗健康大数据：应用实例与系统分析［J］. 大数据，2015，(21)：1－9.

13 王欣. 基于数据挖掘的远程医疗诊断辅助系统的开发［D］. 成都：电子科技大学，2013：14－16.

14 张香梅、盛卉. 截至去年底中国 60 岁以上老年人口已达 2.12 亿［EB/OL］.［2015－6－12］. http://people.com.cn/2015/0612/c1001－27143289.html.

15 肖果平，肖霖. 智能家居在老人医疗中的应用［J］. 现代电子技术，2013，36（9）：23－25.

16 韩煌. 数据挖掘技术在医院信息系统中的应用［J］. 医学信息学杂志，2010，31（10）：28－31.

17 张和华，向华，吴旋，等. 数据挖掘技术在医疗设备行业中的应用研究［J］. 中国医学装备，2015，12（1）：48－50.

18 Karimi S，Wang C，Jimenez A M，et al. Text and Data Mining Techniques in Adverse Drug Reaction Detection［J］. ACM Computing Surveys，2015，47（4）：56－58.

19 Jin H D，Chen J，He H X，et al. Mining Unexpected Temporal Associations：applications in detecting adverse drug reactions［J］. IEEE Transactions on Information Technology in Biomedicine，2008，12（4）：488－500.

20 高汉松，肖凌，许德玮，等. 基于云计算的医疗大数据挖掘平台［J］. 医学信息学杂志,2013,34(5):7－12.

（上接第 53 页）

透明，可以有效地执行现有的平台，建立一个共同的平台，语言无关的技术层。基于 HRP 和排程系统的各种各样的平台依靠这个技术层可实现彼此的连接和集成，为医院创造更多的价值。

参考文献

1 Hartmut Stadtler，Christoph Kilger. Supply Chain Management and Advanced Planning：Concepts，Models，Sofeware and case studies［M］. Bereil：Springer，2010：81－86.

2 陈宁江，李昌武，俞闽敏. 一种融合负载感知和检测点的 Web 服务适应性失效检测机制［J］. 微电子学与计算机，2010，(8)：61－65.

3 武亚琴，闫华，郝梅，等. 借助 HRP 系统优化低值物资管理流程［J］. 医学信息学杂志，2014，35（3）：38－41.

4 王木林. 面向 Web 的网格服务互相调用研究［J］. 电脑知识与技术，2010，(12)：316－318.