文章编号：2096-1472(2017)-11-01-02

医学大数据可视分析研究

降 惠
(长治医学院计算机教学部，山西 长治 046000)

摘 要：可视化分析可以提高科研人员对数据隐藏信息的洞察力。本文将医学数据源分为普通数据、高维数据、公共卫生数据、管理数据、评估数据五种类型，分析了每种类型数据的集成转换方法和交互呈现方式。但不容忽视的是，医学数据可视分析还面临数据不一致、人机交互普及性较差和可视技术不丰富等问题，有待进一步研究和探讨。

关键词：医学；大数据；可视分析；研究
中图分类号：TP391.1 **文献标识码**：A

The Visual Analysis Study of Healthcare Big Data

JIANG Hui
(*The Department of Computer Teaching,Changzhi Medical College,Changzhi* 046000,*China*)

Abstract: With visualized analysis,scientific researchers can improve the insight to hiding information in data.The medical data sources are classified into five different types,including general data,high dimensional data,public health data,management data and evaluation data.The paper analyzes the integration and conversion method and the interaction and presentation method of each type of data.But some noticeable problems still need further studying and probing,such as the inconsistency of data,the low popularity of human-computer interaction,and the inadequacy of visualized technology.

Keywords:medicine;big data;visualized analysis;study

1 引言(Introduction)

随着互联网、云计算、物联网等信息技术的飞速发展，我国省、市、县、乡、村各级医疗机构不断推行信息化、智能化，互联网中的医学数据正在以井喷的方式急剧增加。医学数据正在变得无处不在、触手可及。然而，创造数据的最高价值是发现其蕴含的潜在知识，这个发现的过程就是数据分析[1]。在“互联网+医学”大背景下，医学数据是一种无形资源，只有通过深入分析才能帮助人们认识新知识，掌握新规律，发现新理论，功克新难关。但医学大数据与其他领域大数据相比，具有其独特的特性，每个实体具有上百个临床或诊断变量。心理学家研究又表明人类只可以同时正确分析处理四种变量[2]。在临床诊断分析中，研究人员通常会使用降维的方式来处理数据，将问题分割成人类可以认知的维度或将相同的概念进行组合，但降维技术往往会忽视临床数据中可以突破和理解医学数据中有价值的内容。研究发现，当分析过程中呈现出直观的数据图形时，分析人员可以增强对数据背后隐藏信息的洞察力[3]。因此，研究人员尝试采用人机交互的方式——可视分析技术来改善医学大数据分析。

2 可视分析(Visualized analysis)

可视化分析技术是一种综合利用可视化界面和分析理论来辅助用户对复杂数据进行解释和推理的技术。可视分析是信息可视化、认知科学、人机交互、数据挖掘、数据处理、图像、统计等多领域融合的研究方法。可视化是利用人眼感知能力和人类智慧，对数据进行交互的可视表达，以增强认知的一门学科[4]，是将难以直接显示或不可见的数据映射为可感知的图形、颜色、符号等，以提高数据识别效率并高效传递有用信息[5]。可视分析包括数据集成、呈现和交互。可视化是用户与数据的接口。

3 医学数据源及其类型(Medical data sources and types)

3.1 医学数据来源

医疗“大数据”来源广泛，内容丰富。它可以来源于电子医疗记录、医学检测、家庭监测、社交媒体、零售药房、公共卫生控制中心和医疗保险。

电子医疗记录是患者医疗就诊全过程的数字化记录。它记录了患者人口统计信息、病史、病症、药物治疗、影像检测、病程记录和账单数据等信息，是最有价值的数据来

源[6]。目前国内外医院基本实现了患者病史、就医全过程和康复随访的电子记录，各级医疗机构可以提供患者医疗全过程数据。

医学检测数据包括检验实验室的实验仪器报告数据和影像诊断中心的诊断报告数据，包括生理数据、生化数据和生命体征数据。影像数据包括核磁、CT、超声、X光检测数据。国外医学检测数据来源于医院外的独立医学实验室，如美国的Quest、LabCorp实验室，加拿大的MDS实验室和日本的BML实验室。我国2016年印发了《关于医学检验实验室基本标准和管理规范(试行)的通知》和《关于医学影像诊断中心基本标准和管理规范(试行)的通知》，今后医学检验实验室和医学影像诊断中心将作为独立的法人单位，相应的医学检测数据将来源于医院外的独立部门。

家庭医疗检测数据来源于体温计、体温贴、制氧机、血糖仪、血压计、多功能治疗仪、脂肪测量仪、洗鼻器、按摩椅等。家庭医疗检测使得数据的获取精确到秒。

零售药房是指依法取得《药品经营许可证》的单一门店的药品零售经营企业。零售药房主要服务于附近的居民。零售药房的销售记录，是医学大数据的一个主要来源。

公共卫生控制中心数据主要收集了地方各种流行病的发病情况，包括发病人数、患者年龄、发病日期、发病天数和最终诊断治疗结果等。

医疗保险数据来源于各医疗保险公司。在我国医疗保险数据包括患者使用一类、二类、三类药品费用，处置费、手术费、检查费、医学检验费、医学影像诊断费、护理费用等。

3.2 医学数据类型

医学数据可以分为普通数据、高维数据、公共卫生数据、管理数据、评估数据五种。普通数据包括电子病历、临床设备和临床软件等产生的数据，如血液检测数据、心电图数据、病情描述文本数据等，数据量较小。高维数据包括患者多维度的个人数据，如家族史、患病史等。公共卫生数据包括患者的家庭住址、发病天数、发病日期等信息，往往具有时间和空间特性。管理数据主要包括医疗保险数据、药品安全数据、患者治疗效果数据、患者候诊时间等。评估数据指患者对自身健康信息的评估，包括患者家庭医疗监测、自我评估测验数据等。

4 医学数据可视分析技术(Visualized analysis technology of medical data)

针对以上医学数据源和医学数据类型，医学数据可视分析主要包括普通数据可视分析、高维数据可视分析、公共卫生数据可视分析、管理数据可视分析、评估数据可视分析等。

4.1 普通数据可视分析

普通数据集成通过R语言、Python、Excel、SPSS、Matlab、SAS、Tableau、Spotfire等实现。普通医学数据一般为结构化数据，类型单一，数据集成计算较为容易。普通数据可视化呈现方式包括线图、直方图、饼图、散点图、热点图、心电图、脑电图等。基于这些动态交互式界面，医生可以分类患者，可以直观观测患者个人体征和病情。但对于患者病情的介绍一般为文本数据，为非结构化数据。数据集成分析可以采用Python中的NLTK(自然语言处理)包。数据呈现比较好的形式是标签云。标签云技术是一种将关键词根据词频或其他规则，将不同关键词用不同大小颜色等呈现出来的一种可视化效果。在临床病历病情描述中应用最为广泛。

4.2 高维数据可视分析

高维数据可视分析可以通过R语言、Python、SAS等实现。数据集成可以采用层次聚类的方法。高维医疗数据可视化分析呈现方式有平行坐标、树图、依赖图、时序分析。平行坐标将患者的生命特征表示为等距离的多个垂直平行轴，其中每条曲线表示一个患者个体。平行坐标可以观测每位患者各生命特征之间的关系。具有层次特性的数据通常采用树图来进行分析。家族史通常采用依赖图来进行分析。个人患病史通常采用时序分析法进行分析，主要关注患者个体随时间推移的患病过程。

4.3 公共卫生数据可视分析

公共卫生数据可视分析可以通过Python、R语言、Geoda、OpenGeoda、ArcGIS等实现。数据转换采用计算空间权重矩阵，通过全局空间相关性和局部空间相关性等进行分析。公共卫生数据通常采用地理空间分析方法，可视化呈现方式有统计点图、二维散点图、分级地图、时序分析、时空探索分析等。对于公共卫生数据分析主要从时间和空间两个维度分析病例数据的传播和蔓延。

4.4 管理数据可视分析

管理数据通常采用主控制台(Dashboard)技术。主控制台技术将不同的可分析技术集成到一个平台上，使管理者可以一目了然地分析数据、汇总信息并作出科学决策，如Brown[7]等用主控制台技术监视和快速分析与护士相关的多维数据。

4.5 评估数据可视分析

评估数据通常采用手机应用软件来实现可视分析，患者可以了解自身健康状态，合理安排作息和饮食，配合医生开展更好的治疗。

5 问题与挑战(Problems and challenges)

5.1 医学数据格式、结构、标准的不一致性

医疗数据来源广泛，除了具有了其他大数据的一般特性外，还具有几种不一致性。(1)格式：由于医疗数据来源于不

同的医疗系统，产生于不同的医疗软件，所以生成的数据格式往往不同。一方面，数据格式丰富，包含文本、数字、图像、声音、多媒体等。另一方面，相同的数据可能在不同的软件中重复记录，但不同的软件数据的记录方式可能存在很大的差异，同一属性有的可能标记为文本，而有的软件中则标记为数字，使得数据分析时数据间的连接具有了一定的挑战性。(2)结构：医院信息化管理仍不健全，医院信息录入者输入的数据形式多样，有结构化表格数据，也存在一些非结构化的病历、医学影像检测数据。(3)标准：药房或药品研究人员可能会以药品的化学成分来标记对象，而医院医护工作者往往采用药品的通用名称或商品名称进行标记。

这些不一致性使得数据的质量无法保证，数据集成困难很大，而这些又恰恰是数据可视分析的基础和前提，将直接影响到数据可视分析的科学性和准确性。

5.2 人机交互的普及性有待提高

目前，医学数据可视化主要针对医学数据分析人员，对于患者、医生和护理人员的人机交互分析并未完善。未来任何领域的普通个体均有大数据分析的需求。“人人都懂大数据，人人都能可视化”已成为大数据发展的目标之一。因此，提供自助式大数据可视分析技术有待进一步研究。

5.3 可视技术有待丰富

针对不同的医学数据，虽然已经涌现出很多不同的可视分析方法，但可视技术以直方图、散点图、树图、空间分布图、时间序列图为主，可视技术仍有很大丰富空间。探索更多符合人类认知的可视分析技术仍是今后努力的一个方向。

6 结论(Conclusion)

医学数据可视分析将大量医学普通、高维、公共卫生、管理和评估数据转换成直观形式。在符合人类认知和感知规律的基础上，通过计算机应用软件实现数据集成和转换，通过不同的可视化呈现方式，实现医学数据分析的“增值”效果。但不容忽视的是，医学数据可视分析还面临数据不一致、人机交互普及性较差和可视技术不丰富等问题。基于数据挖掘的医学数据可视分析有待进一步研究和探讨。

参考文献(References)

[1] Cohen J,Dolan B,Dunlap M,et al.MAD skills:New analysis practices for big data[J].PVLDB,2009,2(2):1481-1492.

[2] Graeme S Halford,Rosemary Baker,Julie E McCredden,et al. How many variablescan humans process[J].Psychological Science,2005,16(1):70-76.

[3] 任磊,杜一,马帅,等.大数据可视分析综述[J].软件学报,2014,25(9):1909-1936.

[4] MunzneR.T.WileyInterdisciplinaryR eviews Computational Statistics[J].Visualization analysis and design,2015,2(4):387-403.

[5] Charles D.H,ChrisJ.The Visualization Handbook[M].2004:76-85.

[6] Trivedi,Shrawan Kumar,Deynil,et al.Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence[M].IGI Global,2017:242.

[7] Diane Storer Brown,Carolyn E Aydin,Nancy Donaldson. Quartile dashboards:Translatinglarge data sets into performance improvement priorities[J].Journal for Healthcare Quality,2008,30(6):18-30.

作者简介：

降 惠(1983-)，女，硕士，讲师.研究领域：数据挖掘，医学计算机应用.

(上接第6页)

[9] Yequan Wang,Minlie Huang,Xiaoyan Zhu,et al.Attention-based LSTM for Aspect-level Sentiment Classification[J]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing(EMNL),2016:606-615.

[10] Sepp Hochreiter,J ü rgen Schmidhuber.Long short-term memory[J].Neural computation,1997,9(8):1735-1780.

[11] Kelvin Xu,Jimmy Ba,Ryan Kiros,et al.Show,attend and tell:Neural image caption generation with visual attention[C]. Proceedings of the 32nd International Conference on Machine Learning(ICML),2015:2048-2057.

[12] Volodymyr Mnih,Nicolas Heess,Alex Graves,et al.Recurrent models of visual attention[C].Advances in Neural Information Processing Systems 27(NIPS),2014:2204-2212.

[13] Zichao Yang,Diyi Yang,Chris Dyer,et al.Hierarchical Attention Networks for Document Classification[C].Proceedings of Human Language Technologies.The Annual Conference of the North American Chapter of the Association for Computational Linguistics(NAACL-HLT),2016:1480-489.

[14] Dzmitry Bahdanau,Kyunghyun Cho,Yoshua Bengio. Neural machine translation by jointly learning to align and translate[C].International Conference on Learning Representations(ICLR),2015.

作者简介：

成 璐(1988-)，女，硕士，助教.研究领域：人工智能，自然语言处理，无线传感网络.