

# 面向精准医疗的多源异构数据融合技术研究\*

高景宏 赵 杰 李明原 曹明波

(1 郑州大学第一附属医院 郑州 450052 2 互联网医疗系统与应用国家工程实验室 郑州 450052)

翟运开

(1 郑州大学第一附属医院 郑州 450052 2 互联网医疗系统与应用国家工程实验室 郑州 450052

3 郑州大学管理工程学院 郑州 450001)

〔摘要〕 介绍数据融合概念、类型,从数据融合原理、层次步骤、关键技术几方面阐述面向精准医疗的多源异构数据融合技术,分析数据融合技术应用现状,为相关研究提供参考。

〔关键词〕 精准医疗;多源异构数据;数据融合;大数据分析与应用

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.05.013

**Study on Multi-source Heterogeneous Data Fusion Technology for Precision Medicine** GAO Jinghong, ZHAO Jie, LI Mingyuan, CAO Mingbo, 1The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, 2National Engineering Laboratory for Internet Medical Systems and Applications, Zhengzhou 450052, China; ZHAI Yunkai, 1The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, 2National Engineering Laboratory for Internet Medical Systems and Applications, Zhengzhou 450052, 3School of Management Engineering, Zhengzhou University, Zhengzhou 450001, China

〔Abstract〕 The paper introduces the concept and types of data fusion, elaborates the multi-source heterogeneous data fusion technology for precision medicine from the aspects of data fusion principle, hierarchical steps and key technologies, and analyzes the application status of data fusion technology, so as to provide references for relevant study.

〔Keywords〕 precision medicine; multi-source heterogeneous data; data fusion; big data analysis and application

〔收稿日期〕 2020-07-15

〔作者简介〕 高景宏,博士,助理研究员,发表论文 25 篇;通讯作者:翟运开,教授。

〔基金项目〕 国家重点研发计划“精准医学研究”重点专项“基于远程/移动医疗网络的精准医疗综合服务示范体系建设与推广”(项目编号:2017YFC0909900);河南省医学科技攻关计划联合共建项目“基于健康医疗大数据的郑州市雾霾污染人群健康风险评估”(项目编号:2018020120);河南省高校科技创新团队支持计划“医疗大数据分析与应用”(项目编号:2018IRTSTHN028);河南省自然科学基金青年项目“基于多源健康医疗数据的郑州市雾霾污染健康风险评估与机制研究”(项目编号:202300410409)。

## 1 引言

精准医学 (Precision Medicine) 是基于生物组学数据、临床诊疗、环境因素和生活习惯等, 结合现代生物信息、大数据分析和人工智能技术, 进行疾病精准诊治和个性化用药的一种新型医学模式<sup>[1]</sup>。精准医疗服务的开展离不开多源异构健康医疗大数据的支撑, 而数据融合是进行大数据集成分析与应用的必要前提。健康医疗数据来源广泛, 涉及电子病历系统 (Electronic Medical Record, EMR), 医院信息系统 (Hospital Information System, HIS), 实验室信息管理系统 (Laboratory Information Management System, LIS), 放射信息系统 (Radiology Information System, RIS), 医学影像存储与传输系统 (Picture Archiving and Communication Systems, PACS) 以及病理、心电、体检、公共卫生、慢病和死因监测、医疗保险等系统。不同医疗信息系统所遵循标准与协议各不相同, 数据类型各异, 结构化、半结构化与非结构化数据并存, 造成程度不一的数据孤岛现象, 严重影响大数据分析 with 精准医疗应用。针对多源异构数据进行有效融合成为亟待解决的问题。本研究对多源异构健康医疗数据融合技术进行初步探讨, 以期为进一步促进大数据分析应用与精准医疗服务提供参考。

## 2 数据融合概念

### 2.1 基本概念

数据融合概念最早出现在美国。1973 年美国首次将数据融合理念与技术应用于军事领域多源数据处理, 以基于多传感器信息进行自动决策<sup>[2-4]</sup>。近年来随着应用场景更趋广泛, 数据融合定义也得到丰富。现代数据融合技术是指为了实现及时完整与精准正确的状态判定、身份识别、态势评估等, 针对多源异构数据进行检测、抽取、预处理、关联、估计和整合等一系列操作的一种多层次多角度数据处理手段<sup>[2]</sup>。

### 2.2 基于精准医疗具体应用目标的数据融合

不同数据融合技术操作流程各异, 对于面向精准医疗应用的健康医疗大数据领域, 数据融合就是基于精准医疗具体目标, 对单一来源且不能完整描述研究对象特征、数据类型与结构各异的多源数据, 进行预处理与融合, 输出更为完善、可靠、精准、个体化的描述, 为后续大数据集成分析、辅助临床决策与精准医疗服务等提供基础支撑<sup>[2]</sup>。

## 3 数据融合类型<sup>[2-3,5]</sup>

### 3.1 数据级融合

通常比较依赖数据采集器类型, 是对采集的多源异构数据进行最底层的简单融合<sup>[6]</sup>。

### 3.2 特征级融合

将采集的多源异构数据进行特性提取, 然后整合提取的特征变量, 以更好、更准确地反映目标事物属性, 属于较为高级的数据融合<sup>[5-6]</sup>。

### 3.3 决策级融合

根据实际应用 (精准医疗服务) 的具体需求进行最高级融合与决策。

### 3.4 混合型融合

从系统整体角度出发, 兼顾各个数据融合层次之间的互联和反馈, 基于深度学习对数据级、特征级与决策级融合进行综合以提高数据融合系统效率与性能<sup>[5,7]</sup>。混合融合架构综合考虑各个层次信息融合处理结果, 避免单一层次数据融合信息不确定、不完整、易丢失等缺陷, 有效改善数据融合系统鲁棒性和灵活性<sup>[2,5]</sup>。

## 4 面向精准医疗的数据融合

### 4.1 原理

4.1.1 一般思路 数据融合技术目前已成为医疗大数据处理、疾病特征识别、健康风险评估和医疗

人工智能等领域的研究热点。其思路是：首先对采集的多源异构目标数据进行检测与配准，提取特征信息，接着对识别的特征进行选择或转换处理，然后关联或合并目标特征，融合成具有多模态特征的目标<sup>[2,8-9]</sup>。

4.1.2 数据融合原理 参考前人研究成果<sup>[6]</sup>，将面向精准医疗的数据融合原理概括如下：（1）收集和检测目标数据。根据精准医疗服务具体数据需求，针对性采集不同来源的健康医疗数据，对采集的数据进行初步检测与评估，辅助制定数据融合相关步骤与内容。（2）特征提取。选择最能代表和描述采集数据目标特征的变量进行特征指标数据提

取。（3）模式识别与处理。通过对特征变量的识别和转换，处理整合成关于目标的说明数据集。（4）关联。将关于目标说明的数据按照同类、同质要求进行聚类、分组。（5）合成。采用融合算法对具体目标进行整合，确保目标能够达到解释与描述的一致性，以适应不同精准医疗服务用户需求。

4.2 层次步骤

4.2.1 概述 面向精准医疗的数据融合系统由数据源层、计算层、数据层、分析层等架构而成，这也是数据融合的一般步骤流程<sup>[10]</sup>，见图 1。

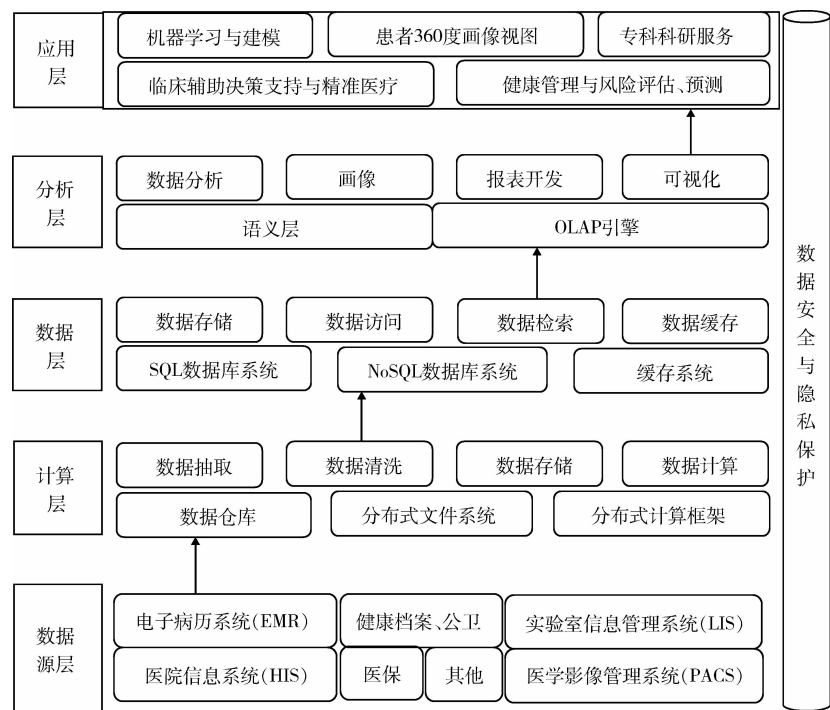


图 1 数据融合系统层次结构与实施步骤

4.2.2 数据源层 主要向上层模块提供多源异构的健康医疗原始数据。

4.2.3 计算层 主要对多源异构数据进行清洗、自然语言处理（Natural Language Processing, NLP）、字段标签分类和映射、数据聚类与标准化处理、多源信息整合与分析等后结构化提取，医学专家标注、无序抽样质控等医学指标标注，以及存储与必要计算，涉及数据仓库、分布式文件系统、分布式计算框架以及数据挖掘引擎等技术。

4.2.4 数据层 涵盖 SQL 系统、NoSQL 系统、缓存系统等，用于实现关系/非关系型数据库存储、访问与搜索，而缓存系统可以用来进行基于缓存的数据计算<sup>[10]</sup>。

4.2.5 分析层 包括在线分析处理引擎（On - Line Analysis Processing, OLAP）和语义层。OLAP 可对健康医疗大数据进行联机分析处理，而语义层可进行报表开发与可视化，实现数据分析与患者画像功能<sup>[8,10]</sup>。

4.2.6 应用层 基于对健康医疗大数据的融合与分析,用于支撑精准诊断、精准治疗、个体化用药、精准健康管理与风险评估等服务。在整个数据融合处理过程中,数据安全与患者隐私保护贯穿始终。

4.2.7 各层次数据融合 针对健康医疗大数据的数据融合在各个层次均可进行。首先在计算层,通过分布式数据库技术对数据源层的多源异构数据进行底层融合与初步筛选;其次在数据层,通过路由协议减少特征数据传输量,进行特征级数据融合;最后在分析层进行决策级融合,减少媒体访问控制 (Medium Access Control, MAC) 链路负荷,同时保障数据完整性<sup>[2,6,11-12]</sup>。健康医疗数据融合目的是为大数据集成分析与精准医疗应用做准备,应基于精准医疗服务目标进行数据针对性抽取、预处理及融合。

4.3 关键技术

4.3.1 数据预处理 健康医疗大数据来源离散而

广泛,分布于不同环境,数据量大,成分和类型复杂,包含结构化、半结构化和非结构化数据且质量参差不齐<sup>[10,13]</sup>。这些问题可造成后期数据分析处理成本增高、决策不准等,给数据融合带来挑战,也使数据预处理成为必然。精准医疗领域大数据融合前的预处理主要包括对数据源层数据的解析、抽取、转换、清洗、规约等操作,目的是利用多源异构数据融合技术对源数据进行相关性分析和归一化处理<sup>[14-15]</sup>。然后根据精准医疗具体应用目标,将数据按标准化、结构化格式抽取出来,进行一系列转化、集成,最后装载入面向精准医疗的专题、专病数据仓库。抽取是将多源异构数据转化为标准、结构化的构型后再进行提取,以便后续快速分析处理<sup>[11,15-16]</sup>。数据清洗是一种“去噪”过程,用以过滤无价值的干扰数据内容,聚焦有效、有用信息。针对不同数据类型需要采用不同处理方式进行融合前预处理,见表 1。

表 1 不同类型数据的处理

| 处理方法            | 数据类型                                           |                                                                     |                                                                                                                                       |
|-----------------|------------------------------------------------|---------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
|                 | 临床、心电数据                                        | 影像数据                                                                | 组学数据                                                                                                                                  |
| Hadoop + MPP 集群 | Hadoop: 数据整合、清洗、治理、转结构化<br>MPP: 应用数据集市,交互式数据分析 | 影像数据保存在 Hbase 集群,海量影像数据集中存储后的快速调阅深度学习集群提取出的影像特征和临床数据在 Hadoop 集群进行整合 | 基于 Spark 并行计算框架的处理加速通过 workflow 整合基因分析全流程 (Spark 并行 + HPC 胖节点计算),因不是所有分析步骤都能进行并行处理提取出的基因变异等信息和临床数据整合研究,构建临床与组学数据统一的数据仓库,生成疾病表型-基因数据队列 |
| 深度学习集群          | 半结构化文本数据转结构化                                   | 影像数据分析数据保存在 NAS 集群 (或 HBase 集群) 基于深度学习集群进行分析                        | N/A                                                                                                                                   |
| HPC 集群          | N/A                                            | N/A                                                                 | 初期: HPC 集群为主, HPC 集群安装及开源分析软件部署后即可快速开展工作,组学数据保存在 NAS 中,后续通过 workflow 整合基因分析全流程 (Spark 并行 + HPC 胖节点计算)                                 |

4.3.2 数据融合方法技术 包括统计、信号处理与估计理论等传统方法,以及人工智能、信息论等新兴技术<sup>[2]</sup>,见表 2。统计方法应用较多的是贝叶斯推理,常被用于多源异构数据融合,其原理是基于概率论原则对数据采集器信息进行组合,将相应的不确定性以概率形式予以呈现,然后估算在多源异构数据前提条件下某一特定假设为真的后验概

率<sup>[2,17]</sup>。信号处理与估计理论方法较为常用的是卡尔曼 (Kalman) 滤波法。首先对数据采集器节点收集的同质化信息进行分批,然后利用一系列加权公式予以处理,使计算出的融合信息更具真实性,还可减少数据融合过程中的信息传输量,从而延长整个数据融合系统使用周期<sup>[6,18]</sup>。传统多源异构数据融合多是基于机构本身数据中心进行整合,即先对

采集自不同环境的源数据进行集中管理，然后提供无差别数据服务<sup>[19]</sup>。此类方法聚焦于数据，未能很好地兼顾不同用户特定需求。随着海量健康医疗数据快速增长，传统数据融合方法面临存储代价大、融合成本高等局限，已无法满足基于大数据集成分析的精准医疗各类应用需求<sup>[14,19]</sup>。近年来，随着人工智能在健康医疗领域的快速发展，其在健康医疗大数据融合方面的应用热度提升。人工智能数据融合方法主要有神经网络、逻辑模糊理论、遗传算法等。神经网络方法可以模拟复杂的非线性映射且具有适应性强、运算速度快、联想性能高、自主学习和容错能力好等优点，能够有效适应多源异构健康医疗数据融合处理要求<sup>[2,20]</sup>。神经网络技术通过自主学习能力可以根据源数据之间的相似、相关性进行一定数据处理规则和融合算法制定，还可获得不确定推理机制<sup>[2,11,21]</sup>。这些均提示神经网络方法在健康医疗大数据融合领域的良好应用前景。

表 2  数据融合技术方法

| 方法分类 | 具体类别        | 关键技术    |
|------|-------------|---------|
| 传统方法 | 统计方法        | 支持向量机理论 |
|      |             | 贝叶斯推理   |
|      |             | 传统推理法   |
|      | 信号处理与估计理论方法 | 加权平均数法  |
| 新兴技术 | 人工智能方法      | 最小二乘法   |
|      |             | 卡尔曼滤波法  |
|      |             | 神经网络    |
|      |             | 逻辑模糊法   |
|      | 信息论方法       | 遗传算法    |
|      |             | 支持向量机   |
|      |             | 贝叶斯推理   |
|      |             | 传统推理法   |

4.4  数据融合技术应用

随着大数据分析的蓬勃发展与广泛应用，作为大数据分析应用“必由之径”的数据融合技术受到重视。姜建华等<sup>[22]</sup>在三角模糊数的多源异构数据统一量化表示方法基础上设计了一种数据融合算法，可以支持多用户决策，保证多用户决策结果的可靠程度。针对国内临床诊疗信息与组学数据的各自运

维、离散分布、信息孤岛严重、缺乏整合协作等问题，高东平等<sup>[11]</sup>设计了一种支持分布式采集、存储、索引和利用多中心临床诊疗与生物样本信息的协作网络系统，为大数据集成分析与精准医疗各类应用奠定了基础。结合深度学习模型和影像数据深度融合学习模型，针对多源异构影像数据的融合问题，惠国保<sup>[5]</sup>构建了一种泛化性强的深度学习模型，将深度特征学习技术运用到多源异构影像数据的提取、融合和挖掘。为提高数据的丰富度、准确性、置信度和可靠性，贺雅琪<sup>[2]</sup>提出了一种基于 D-S 证据理论（Dempster – Shafer Thoery）的加权模糊调节型数据融合框架。但是目前较少有研究涉及大数据融合技术在精准医疗领域的应用，随着多源异构健康医疗大数据在精准医疗中的深入应用，面向精准医疗具体服务的数据融合正成为亟须推进的研究课题。

5  结语

精准医疗已成为辅助临床决策、个性化精准治疗、减少误诊漏诊、节约医疗费用与资源的重要举措，而健康医疗大数据分析是精准医疗必不可少的支撑要素。大数据集成分析的前提是对多源异构健康医疗数据进行融合。面向精准医疗的数据融合是针对健康医疗数据的一系列处理手段，包括统计、信号处理与估计理论等传统方法以及人工智能、信息论等新兴技术。通过数据融合，可聚合多方信息源证据，获得相互协同兼顾、综合互补的推断与决策，有效提高数据利用效率、可靠性与置信度，减少大数据分析与精准医疗临床辅助决策的不确定性，提升精准医疗服务质量。本研究通过数据融合原理、步骤、关键技术等方面，对面向精准医疗的多源异构数据融合技术进行探讨，为提高健康医疗大数据整合、分析、挖掘和应用效率，以及精准医疗服务质量等提供信息与技术支撑。

参考文献

1  詹启敏, 张华, 陈柯羽, 等. 精准医学总论 [M]. 上海: 上海交通大学出版社, 2017.

- 2 贺雅琪. 多源异构数据融合关键技术研究及其应用 [D]. 成都: 电子科技大学, 2018.
- 3 Steinberg A N, Bowman C L, White F E. Revisions to the JDL Data Fusion Model [J]. Proceedings of SPIE - The International Society for Optical Engineering, 1999 (3719): 430 - 441.
- 4 Llinas J, Antony R T. Blackboard Concepts for Data Fusion Applications [J]. International Journal of Pattern Recognition & Artificial Intelligence, 1993, 7 (2): 285 - 308.
- 5 惠国保. 一种基于深度学习的多源异构数据融合方法 [J]. 现代导航, 2017, 8 (3): 218 - 223.
- 6 凌云. 基于物联网的异构传感数据融合方法研究 [J]. 计算机仿真, 2011, 28 (11): 138 - 140.
- 7 许海云, 岳增慧. 科学计量中多源数据融合方法研究述评 [J]. 情报学报, 2018 (3): 318 - 328.
- 8 李勇明, 肖洁, 王品, 等. 基于多模态深度迁移学习机制的多源异构数据融合系统: 中国, CN108734208A [EB/OL]. [2020 - 05 - 15]. [http://www.wanfangdata.com.cn/details/detail.do?\\_type=patent&id=CN201810464348.4](http://www.wanfangdata.com.cn/details/detail.do?_type=patent&id=CN201810464348.4).
- 9 陈敏, 刘宁. 医疗健康大数据发展现状研究 [J]. 中国医院管理, 2017, 37 (2): 46 - 48.
- 10 彭向晖, 黄文强, 卢春, 等. 多源异构数据融合系统及方法: 中国, CN108021670A [EB/OL]. [2020 - 12 - 06]. [http://www.wanfangdata.com.cn/details/detail.do?\\_type=patent&id=CN201711273802.X](http://www.wanfangdata.com.cn/details/detail.do?_type=patent&id=CN201711273802.X).
- 11 高东平, 王士泉, 戴阿咪. 融合临床与组学数据的重大疾病生命组学协作网络平台建设初探 [J]. 中国数字医学, 2017, 12 (8): 38 - 41, 84.
- 12 Lu G, Krishnamachari B, Raghavendra C S. An Adaptive Energy - Efficient and Low - Latency MAC for Data Gathering in Wireless Sensor Networks [J]. Wireless Communications & Mobile Computing, 2010, 7 (7): 863 - 875.
- 13 陈敏, 刘宁. 健康医疗大数据发展现状研究 [J]. 医学信息学杂志, 2017, 38 (7): 2 - 6.
- 14 秦爱民. 基于多源异构数据融合、机器学习及客服机器人的智能运维分析系统: 中国, CN109343995A [EB/OL]. [2020 - 10 - 25]. [http://www.wanfangdata.com.cn/details/detail.do?\\_type=patent&id=CN201811252658.6](http://www.wanfangdata.com.cn/details/detail.do?_type=patent&id=CN201811252658.6).
- 15 程运平. 医院数据分析与挖掘的研究与实现 [D]. 西安: 西安电子科技大学, 2015.
- 16 Kimball R, Caserta J. The Data Warehouse ETL Toolkit: practical techniques for extracting, cleaning, conforming and delivering data [M]. Hoboken: John Wiley & Sons, Inc., 2004.
- 17 赵汉青, 王志国. 论中医药多源异构大数据融合方法研究的意义 [J]. 中医学, 2018, 7 (5): 282 - 285.
- 18 Evensen G. The Ensemble Kalman Filter: theoretical formulation and practical implementation [J]. Ocean Dynamics, 2003, 53 (4): 343 - 367.
- 19 黄鑫, 张卓, 黄伟, 等. 一种多源异构数据的融合方法及装置: 中国, CN107545046A [EB/OL]. [2020 - 08 - 17]. [http://www.wanfangdata.com.cn/details/detail.do?\\_type=patent&id=CN201710707596.2](http://www.wanfangdata.com.cn/details/detail.do?_type=patent&id=CN201710707596.2).
- 20 Benke K, Benke G. Artificial Intelligence and Big Data in Public Health [J]. Int J Environ Res Public Health, 2018, 15 (12): 2796.
- 21 字云飞, 李业丽, 孙华艳. 基于深度神经网络的个性化推荐系统研究 [J]. 电子技术应用, 2019, 45 (1): 20 - 24, 28.
- 22 姜建华, 洪年松, 张广云. 一种多源异构数据融合方法及其应用研究 [J]. 电子设计工程, 2016, 24 (12): 33 - 36.

## 《医学信息学杂志》版权声明

(1) 作者所投稿件无“抄袭”、“剽窃”、“一稿两投或多投”等学术不端行为, 对于署名无异议, 不涉及保密与知识产权的侵权等问题, 文责自负。对于因上述问题引起的一切法律纠纷, 完全由全体署名作者负责, 无需编辑部承担连带责任。(2) 来稿刊用后, 该稿包括印刷出版和电子出版在内的版权、复制权、发行权、汇编权、翻译权及信息网络传播权已经转让给《医学信息学杂志》编辑部。除纸质载体形式出版外, 本刊有权以光盘、网络期刊等其他方式刊登文稿, 本刊已加入万方数据“数字化期刊群”、重庆维普“中文科技期刊数据库”、清华同方“中国期刊全文数据库”、中邮阅读网。(3) 作者著作权使用费与本刊稿酬一次性给付, 不再另行发放。作者如不同意文章入编, 投稿时敬请说明。

《医学信息学杂志》编辑部