

● 余 辉, 梁镇涛, 鄢宇晨 (武汉大学信息资源研究中心, 湖北 武汉 430072)

多来源多模态数据融合与集成研究进展*

摘 要: [目的/意义] 在国内研究的基础上, 对国外数据融合与集成相关文献进行系统梳理, 旨在为数据融合与集成领域工作提供参考。[方法/过程] 主要从相关研究、研究路径与方法以及应用与挑战三个方面进行回顾, 并从概念、路径和应用三个方面指出未来研究方向。[结果/结论] 目前研究路径与方法包括数据处理模型、基于阶段的策略、基于特征表示的方法以及基于语义的应用 4 个层面; 从多传感器系统、生物医学、环境研究和城市建设 4 个领域研究发现, 每个领域都存在着共同或者特殊的应用挑战。

关键词: 多源数据; 多模态数据; 数据融合; 数据集成; 研究进展

DOI: 10.16353/j.cnki.1000-7490.2020.11.027

引用格式: 余辉, 梁镇涛, 鄢宇晨. 多来源多模态数据融合与集成研究进展 [J]. 情报理论与实践, 2020, 43 (11): 169-178.

Review on Multi-source and Multi-modal Data Fusion and Integration

Abstract [Purpose/significance] On the basis of domestic research, this paper systematically sorts out related literatures on data fusion and integration abroad, aiming at providing references for the work in the field of data fusion and integration. [Method/processes] This paper mainly reviews relevant research, research approaches and methods, and applications and challenges, and points out future research directions from three aspects: concept, path and application. [Result/conclusions] Research approaches and methods include data fusion and integration model and framework, phase-based strategy, feature-based representation method and semantic-based application. And it is found that there are common or special application challenges in each field.

Keywords: multi-source data; multi-modal data; data fusion; data integration; research progress

在大数据时代, 数据分析工作通常需要处理不同来源、不同领域的的数据, 这些数据呈现出不同的模态^[1]。多源多模态数据能比单一数据提供更多信息, 通过相互之间支持、补充、修正, 能提供更准确的信息。D. L. Hall 将多传感器与单一传感器数据融合进行了比较, 认为多传感器数据融合在数据的准确性和实际应用方面会更有优势^[2]。此外, 在信息呈现和表达上, 多来源、多模态数据还增加了鲁棒性^[3], 如音频不需要视线, 视频不怕环境噪声。因此在对信息质量要求较高时, 如医学诊断对准确性的高要求, 气象预测对数据的全面性要求等, 多源数据的融合与集成是提高数据分析质量的重要基础。

物联网提供大量数据, 云计算提供大数据处理能力, 在此背景下, 数据融合与数据集成的研究热度一直在增加。目前国内数据融合与集成以具体领域应用型论文居

多, 涉及地质、测震、遥感、船舶、瓦斯浓度、新闻、科学评价、版权、元数据集成框架等。多模态研究除了常见的几种数据类型的融合和集成之外, 多与检索领域相关。相比国外, 国内关于数据融合和集成的路径方法研究起步较晚。本文以 Web of Science 的外文文献为主要来源, 以“Multi-source” “multi-modal” “multimodal” “Data Fusion” “Data Integration” 组合为篇名进行初始检索, 并由参考文献进行扩展, 对数据融合和集成的概念、相互关系和研究路径、方法及应用做一个研究回顾与展望。

1 相关研究

1.1 关键概念界定

1.1.1 多来源 三元世界理论把世界分为物理世界、人类社会和信息空间, 它的基本逻辑为前两个世界产生数据并传输到信息空间中, 信息空间是对信息融合、集成与分析, 然后指导前两个世界的决策行为^[4], 如图 1 所示。物理世界中主要由物联网产生数据, 人类社会有社会网络和政府部门的数据, 这些数据包括但不限于传感器、官方网站、社会新闻、社交媒体等众多来源的数据。

* 本文为国家自然科学基金重大项目“国家安全大数据综合信息集成与分析方法”(项目编号: 71790612) 和国家自然科学基金项目“大数据环境下基于特征本体学习的无监督文本分类方法研究”(项目编号: 71571064) 的成果。

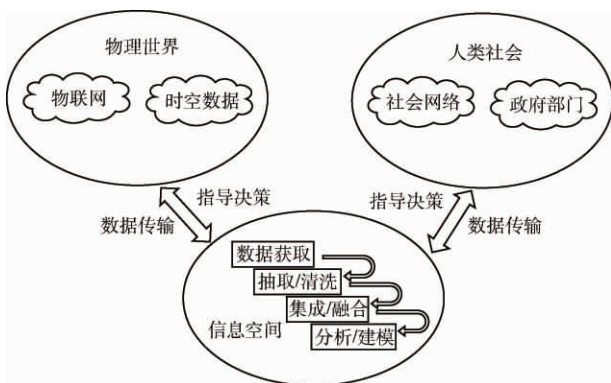


图1 三元世界数据传输与信息集成逻辑

J. B. Daher 等认为在某些情况下,数据并非来自同一信息源,而是来自不同的来源,并被收集在一个数据集中,这种类型的数据称为多源数据^[5]。总体来说多来源数据就是取自多个端口的数据集。

1.1.2 多模态 多模态可以以人的感官为基础,包括视觉看到的模态、听觉听到的模态以及触觉感到的模态等,每个模态都可以表示人的感知。多个模态可以相互组合起来形成完整的人体模态感知。在数据领域,多模态用来表示不同形态的数据形式,或者同种形态不同的格式,一般表示成,文本、图片、音频、视频、混合数据^[6]。D. Lahat 等将多模态定义为一种相比于单模态来说可以加强使用的资质^[3]。赵亮认为多模态数据是指对于同一个描述对象,通过不同领域或视角获取到的数据,并且把描述这些数据的每一个领域或视角叫做一个模态 (Modality)^[7]。总体来说,多模态数据是不能直接进行处理的、能产生相互联系的不同数据集。

1.1.3 数据融合 数据融合 (Data Fusion) 也称为信息融合 (Information Fusion), 起源于 1973 年美国国防部资助开发的声纳信号处理系统, 具有更广义化的概念在 20 世纪 90 年代被提出。Joint Directors of Laboratories (JDL) 对数据融合给出的定义是: 处理来自单一和多个来源的数据和信息的关联的多层次过程, 以实现重新定位, 并完善及时地对其形势、风险及重要性进行评估^[8]。D. L. Hall 和 J. Llinas 给出的定义是: 数据融合技术将来自于多个传感器的数据和来自相关数据库的相关数据相结合, 以得出比单个传感器更准确更具体的推论^[9]。国内学者祁友杰等提出数据融合是对多个传感器和信息源所提供的关于某一环境特征的不完整信息加以综合, 以形成相对完整、一致的感知描述, 从而实现更加准确的识别和判断功能^[10]。Zhou Jingren 等认为多源数据融合是认知、综合和判断各种数据的过程。融合过程中涉及的数据通常具有

诸如多源、异质性和不完整性之类的属性^[11]。总结学者们的研究, 可以发现数据融合是对多来源数据分析处理的过程, 目的是得出更准确、统一的信息, 通常用来增强决策过程^[12]。

1.1.4 数据集成 数据集成目前没有一个标准的定义。M. Lenzerini 认为数据集成是将存在于不同数据源中的数据集中组合, 并将这些数据统一提供给用户^[13]。Wang Ping 等提出数据集成并不是简单的叠加, 而是两个或多个数据集通过格式转换、结构重组、语义匹配、尺度转换和数据融合等有机和虚拟的组合^[14]。Zhou Jingren 等认为数据集成是在统一平台上集成重要数据源, 数据消歧和数据存储的过程^[11]。张朝阳认为数据集成是将若干个分散的数据源中的数据, 逻辑或者物理地集中在一个数据集中的过程, 目标是实现数据共享和信息交流, 核心任务是要将互相关联的分布式异构数据源集成到一起^[15]。从学者们的定义中, 可以知道数据集成主要目标是集成多个信息源数据, 多从逻辑和物理以及规范上进行集中, 通过集成, 具有不同特征的数据集将成为一个新的数据集, 新的数据集应该表达与以前相同的含义, 甚至挖掘出一些潜在的规则 and 知识^[14]。

1.2 概念间的联系

1.2.1 多来源与多模态的关系 多来源和多模态之间没有绝对的关系, 根据是否为多来源和是否为多模态两两相互组合, 可以形成 4 对关系, 即单来源单模态、单来源多模态、多来源单模态和多来源多模态。徐晓滨等研究的分布式异类多传感器系统图中包括了不同类型的传感器, 以及同类型的多个传感器 (用 Sensor 表示), 如图 2 所示^[16]。其中同一类型传感器中不同的 Sensor 采集到的数据为多来源单模态数据, 如田鸽等用无人机多镜头取景用多源数据融合方法高效地生成立体三维建模^[17]; 不同类型的传感器产生的数据为多来源多模态数据; 从多个同类型的传感器收集的数据是多来源单模态数据, 如制作全国各地温度地图, 各地采集的温度数据为同一模态。

这些组合也不一定是绝对的, 如单一设备收集的数据, 可以视为单来源数据, 但在此设备中可能包括多个相同或不同传感器, 又可以认为是多来源数据。郭丽娟在多领域研究中的多来源文本^[18], 为多来源单模态。这里以

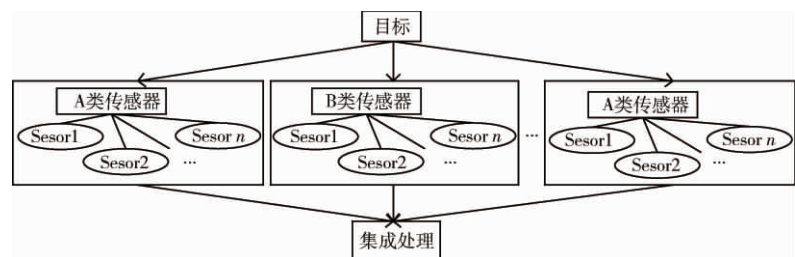


图2 分布式异类多传感器系统

文本为模态,但文本也可以分为不同的结构、格式等,所以一般来说不同来源的数据,多涉及多模态处理问题。在处理不同粒度的数据时,是否为多来源和是否为多模态的判断都不是绝对的。

1.2.2 数据融合与数据集成的联系 对于数据融合与数据集成的关系,不同的研究中表述不一样。有学者认为是相同的,也有认为融合是集成的一部分,也有认为集成是融合的一阶段。Zhou Jingren 等指出数据集成和数据融合之间的差异很难描述,一些研究人员承认这两个概念在大多数情况下几乎是相同的,还有许多研究人员也将数据集成视为数据融合的主要阶段^[11]。Zheng Yu 则指出传统的数据融合被认为是数据集成的一部分^[19]。F. Naumann 等也认为数据融合是数据集成的一部分,如图3所示^[20]。

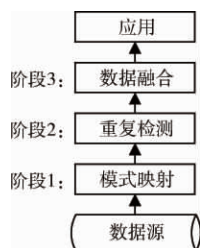


图3 数据集成过程

本文在总结学者们多种定义和方法基础上,认为虚拟化为数据集成,物质化为数据融合,并认为此过程中包含上层和底层的两阶段数据集成,数据融合是上层数据集成的一部分,底层数据集成是数据融合的数据来源前提。集成多为位置上的集中,或者是顺序上的排列组合,生成一个较大数据集或有序的信息,数据总信息是不变的。融合是对不同数据集进行处理后生成新数据集,可能会有信息损失或增强。融合更侧重于多个数据对象合并的过程,集成侧重于多个数据对象集中在一起的呈现。融合和集成是多源多模态数据处理中不可分割的两种方法,交替或者同时存在于数据处理过程中。融合和集成共同服务于信息集成平台,在信息集成平台中,数据融合与集成协作关系如图4所示。

2 数据融合与集成路径与方法

本文总结学者们对数据融合与集成的路径和方法研究分为三个模块:模型研究、策略研究和方法研究。模型研究主要从较为宏观角度解决数据融合与集成的整体过程;策略研究是对数据融合与集成的经典划分方案,从策略角度研究融合与集成的路径;方法研究中介绍在模型框架下,按照所属策略,如何具体地进行数据融合与集成。本文对数据融合与集成的具体研究路径和方法做了整理分类,如图5所示。

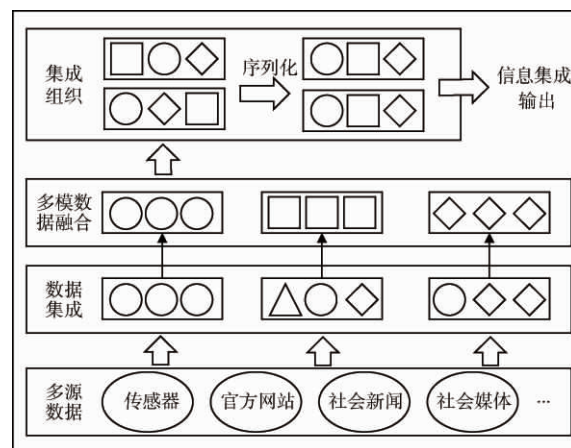


图4 信息集成平台下数据融合和数据集成的关系

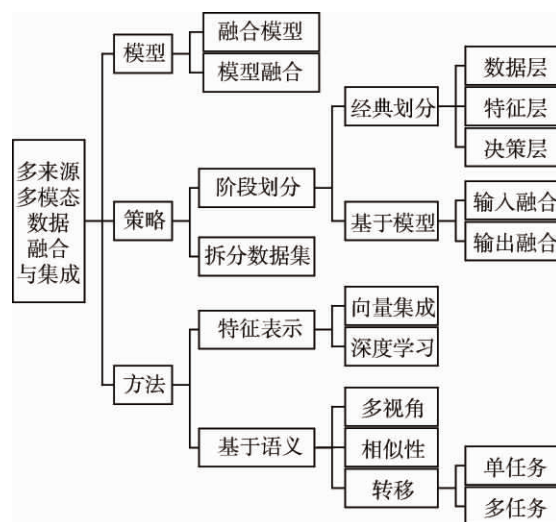


图5 数据融合与集成路径

2.1 数据处理模型

2.1.1 过程建模 数据融合模型对指导数据融合与集成有重要意义, J. Llinas 等提出了4级融合模型: 目标提炼、态势分析、威胁估计、过程精炼^[21], 是较早的数据融合模型。后来有学者进行了扩展和延伸, 如 Tang Jianzhi 等构建了由空间数据层、数据访问层、应用程序服务器层和用户层组成的4层结构层模块^[22]。国外最新的数据融合模型是基于 Joint Directions of Laboratories (JDL) 构建的模型^[21]改进的^[10], 数据融合模型如图6所示。

2.1.2 模型融合 在数据融合与集成研究上,除了直接对数据融合模型做出研究外,还有学者通过对已有模型进行加工,结合数据融合要求的特殊性,提出了对多个已有模型进行融合来满足数据融合的需求。Liu 在基于城市大数据不同粒度之间特征的相关关系和共享的研究中表示,通过模型融合(即模型组合或跨模型组合,模型可以是CNN, RNN, LSTM, RBM等)来实现多任务集成的深度

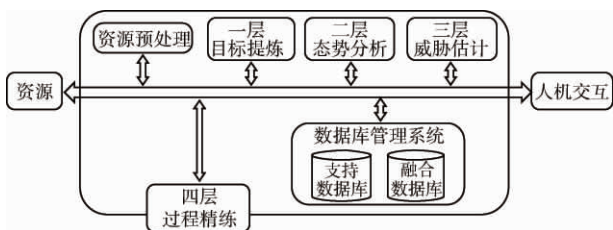


图6 数据融合模型

学习,并同时完成数据融合。例如,通过组合 CNN 和 RNN 模型,可以使数据的时空属性很好地融合在一起^[6]。Zhang Weishan 等在用多源数据融合进行智能冰箱研究时,把三个训练模型融合,并用这三个模型融合后的产出,作为神经网络的输入^[23],模型融合结构如图7所示。

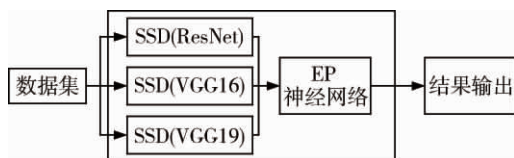


图7 三模型融合结构

除了对多个模型进行融合的研究外,还有学者对数据同化系统进行了研究,同化系统在系统内部对多种模态数据的处理进行了融合,相关成果已经在各个国家进行了实施应用,如英国气象局用于分析对流层和平流层观测资料组合的英国数据同化系统、欧洲旱涝预报陆地数据同化系统和中国西部陆地数据同化系统等^[24]。通过数据同化,将不同分辨率的多源数据进行融合,挖掘出更多潜在的信息。

2.2 基于阶段的数据融合与集成策略

基于阶段来划分数据融合与集成,是从整体融合视角出发的一种融合策略。通过对数据融合工作进行目标分解,使得融合过程的可执行度更高。

2.2.1 分阶段融合 分阶段进行数据融合,就是按照流程顺序对不同阶段的数据进行融合,可以对所有阶段都分别进行融合,也可以只对某一阶段进行融合。数据融合最常见的三个模式即按三个阶段进行划分。如动作单元、特征融合、决策融合^[25],以及像素级融合、特征级融合和决策级融合^[21]等,其中像素级融合也称为数据级融合,通常把数据分析工作分为数据层、特征层和决策层三个典型的阶段。数据层融合可以看作直接把数据统一格式,转换为同模态数据进行统一处理分析,如Liu等把不同来源的异构数据进行结构化集成,用多源数据融合框架把数据都转换为 RDF (资源描述框架) 的统一数据格式,再通过数据融合算法进行信息融合^[26];特征层融合,如表面材料识别中,将特征融合映射到子空间就属于特征层的融

合^[27];决策层融合,如Yu Yuhai等利用 Logistic 回归分别对文本与相关图像进行情感预测,最后将两个预测概率进行加权平均,得出最终结果^[28]。三层数据融合模型如图8所示。

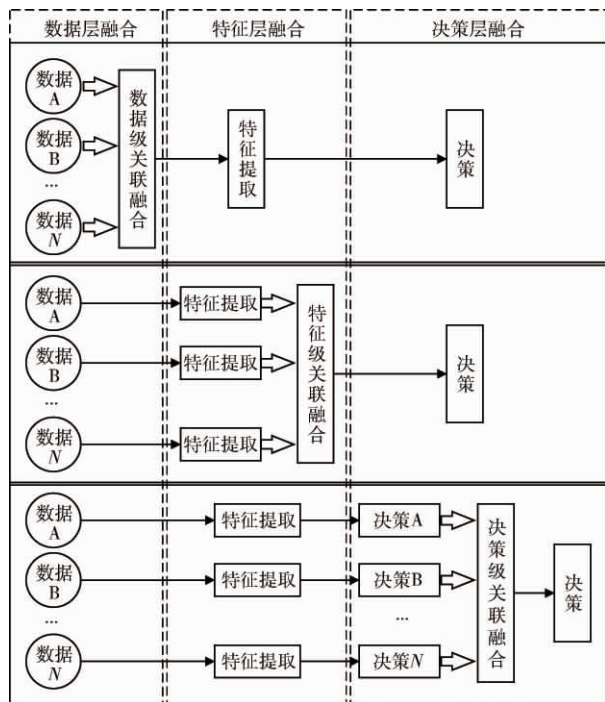


图8 三层数据融合模型

除了经典三层数据融合模型外,也有学者按系统和模型(如深度学习模型)的输入和输出对把数据融合分为两个阶段,如Liu等在研究城市大数据融合时就分了输入融合、输出融合和两个阶段都融合这三种融合模式^[6],如图9所示。第一种,融合通过深度学习模型训练时空数据,基于特征层融合所有模型的产出^[1];第二种,训练时空数据时通过深度学习网络融合数据,然后再融合产出;第三种,同时考虑早期和晚期两个阶段的融合。可以通过两个阶段分为三级,实际应用也可以扩展为多级融合。

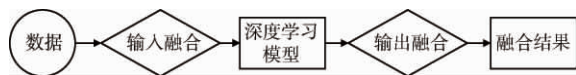


图9 基于模型的融合阶段划分

2.2.2 基于阶段拆分数据集 在数据融合与集成分析中,多源多模态数据使得分析更准确和完整,但也使分析过程变得复杂,为了简化分析过程,有学者把数据分析目标分解成独立的阶段目标,这些阶段目标可以各自用数据集的一部分进行分析,即在不同的数据处理阶段,利用不同的数据集。与决策层融合不同,这里通常以数据分析过程的先后顺序划分阶段,把数据融合问题分解为独立的小问题,然后分先后分别进行分析,因此对数据集之间的模式

是否一致没有要求,各阶段完成不同的阶段目标,逐步实现数据分析的总目标。

基于阶段划分,不同的阶段使用不同的数据集(并不需要融合),如在城市交通网设计研究中,先用路线网络数据把城市分为不相连的区域,再用人口流动数据来判断区域内路线设计是否较好。Zheng Yu 等^[29]将工作分为两个阶段,阶段1使用地图分割方法将城市通过主要道路划分为区域,然后,将出租车的GPS轨迹映射到这些区域上,以绘制区域图,如图10所示。区域图实际上融合了来自道路网络和出租车轨迹的知识,通过分析区域图能实现识别道路网络的不当设计,检测和诊断交通异常等功能,即在数据分析过程中,通过不同阶段使用不同的数据集,对多源数据进行了利用。

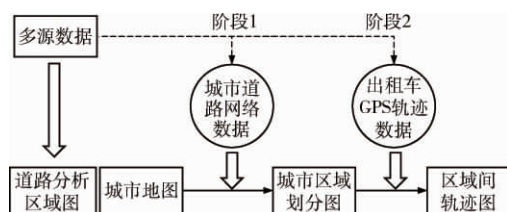


图10 城市交通道路合理性分析

2.3 基于特征表示的数据融合与集成

基于特征表示的数据融合和集成方法是多源多模态数据处理的重要方法,早期数据融合通过把数据转换为单一的基于特征的数据集进行处理。现在从不同角度去进行数据融合,如最简单的数据融合方法是直接组合两个一维数据集中具有相同含义的一维数据集,或者从不同的数据集中抽取相同维度的数据进行直接连接。还可以提取不同维度的数据进行融合分析,使用深度学习方法进行研究,如使用深度神经网络(DNN)从不同数据集提取的原始特征中学习的新表示形式,然后导入模型用于分类或者预测等数据分析工作中。

2.3.1 特征向量集成 特征向量集成指直接串联各个特征,把各特征都均一融入到一个向量^[30]。这种方法不考虑向量间的权重或者相关关系,直接用所有的向量数据生成一个更高维的向量。在信息维度不大信息量较少时,可以直接串连成信息,并且具有可读性,如用来描述形容对象时,在数据融合后的集成阶段,可以采取直接串联融合后的数据。

但在数据处理过程中,数据维度较高,来源较多时,不宜采用。因为不同的数据集差距很大,有可能造成过拟合问题,也可能忽略了各自的特征。此外,不同数据集之间的特征可能并不独立,而是存在相关关系。这些都让特征向量直接融合的方法无法用于数据处理工作,但也有学者在此方法的基础上进行了数据集成方法的改进和研究。

V. Walter 和 D. Fritsch 基于关系匹配原理提出了一种自动匹配线性数据元素的方法,这种方法在不考虑权重因子、阈值或启动参数的情况下,仅通过数据集之间的统计调查构建了该方法,并在 ATKIS 模型和 GDF 模型中对道路网数据进行了检验。通过该方法,可以正确匹配出高比例的路网要素,并可用于其他线路要素的匹配^[31]。

2.3.2 基于深度学习的融合 在基于特征表示的数据融合的前期,一般很少采用直接串联各特征的方法。即特征向量需要进行处理后再进行融合,基于机器学习和深度学习的数据融合,是利用多个模式之间的特征层来实现融合,这也是目前多来源多模态数据融合的主要处理路径。

多模态深度学习融合和经典深度学习基本原理相同,只是用来学习和训练的数据集有所区别,从单一数据集变为多源多模态数据集,或者是交叉使用不同模态的数据集。J. Ngiam 等提出了一种深度自动编码器架构来捕获两种模式(如音频和视频)之间的“中间层”特征表示,三种学习设置(包括跨模态学习、共享表征学习和多模态融合)与经典深度学习模型对比如表1^[32]所示。

表1 多模态学习设置对比

对比	特征学习	监督训练	测试
经典深度学习	音频	音频	音频
	视频	视频	视频
跨模态学习	音频+视频	视频	视频
	音频+视频	音频	音频
共享表征学习	音频+视频	音频	视频
	音频+视频	视频	音频
多模态融合	音频+视频	音频+视频	音频+视频

在设置好学习数据集后,在学习方法上基于深度学习的多源数据融合模式可以分为基于分类的方法和基于规则的方法^[25],也同样可以按阶段分为早期融合和晚期融合。其中基于分类器的融合,早期融合即用特征向量构建分类器,一个分类器可以由多模态特征向量构建,在方法上可以用 SVMs 或神经网络等分类器,可以被同特征同维度向量训练;或者用生成模型,如贝叶斯网络可以从多维特征向量中学习,生成模型能处理可变长向量。分类器也可以用于晚期决策层,从单一分类器结果中学习。这种情况下,不同分类器的得分融合到同一个向量中,然后再用二级分类器为每一个概念学习这个向量。而基于规则的融合模式更适合在决策级别进行,通过在不同模态分类器中的得分(均或最大值)、投票,或先验知识自定义规则进行决策层融合。

在深度学习应用的算法上,和经典机器学习一样,使用特征表示的深度学习进行数据融合受模型和参数的影响较大、可解释性差,是目前深度学习领域需要解决的难题。目前多源数据融合的算法可以分为经典估计统计方法

和现代信息论与人工智能方法^[33],也可以分为物理模型、参数分类和认知模型三类^[10]等,总体都包括估计、推理和人工智能的一些方法,数据融合算法总结如图11所示。

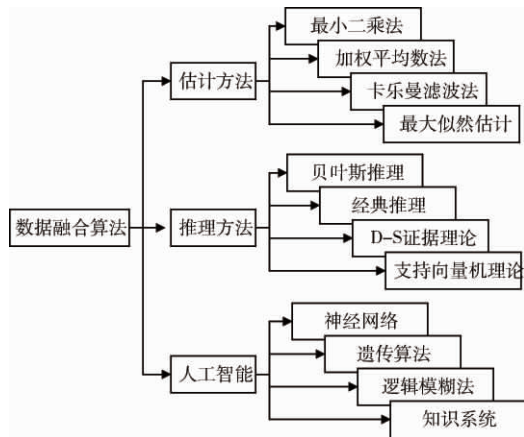


图11 数据融合算法

2.4 基于语义的数据融合与集成

特征学习不关心每一特征的含义,而每一特征在现实中都是真实有意义的值,而且这些特征之间还存在着联系。基于语义的数据融合在特征表示学习方法的基础上,结合语义上的关系对数据集对象进行进一步的处理,以实现数据信息的增强、补充和知识迁移。

2.4.1 多视角信息增强 多视角信息是典型的多来源多模态数据融合的产物,在多视角下每个示例可以分为两个不同的视图,它们分别从不同角度对对象进行描述。不同视角产生不同的数据集用来表示同一对象,这些数据集具有潜在的一致性和互补性。基于语义的多视角数据融合方法并不是独立于特征表示外,而是建立在特征表示和阶段划分的基础上,把不同的特征放进不同模型,从不同的方面描述对象,最后整合结果或者相互加强补充结果,如协同训练方法、随机森林、多核学习、子空间学习等都属于这种方法。

空气质量预测是基于多视角融合集成多源数据的经典例子。空气质量问题可以分为时间和空间两个不同视角。在时间视角上,通常认为空气质量不会出急剧变化,而应该是平滑的过渡,即当前时间的空气质量应该与过去几个小时的空气质量有关;在空间视角上,通常认为同一个地区空气质量不会相差太大,也应该是平滑的过渡,即当前地区的空气质量与其他相邻地区的空气质量相关^[34]。空气质量中可以把时间和空间作为两个内核,分别训练数据得出预测结果,然后通过聚合器动态调整生成一个结果^[35],如图12所示。通过结合时间和空间两个视角的多核学习产出能比单一视角更加准确地对空气质量指数进行预测。



图12 基于多视角的多核学习预测空气质量流程

2.4.2 相似性信息补充 多来源多模态数据融合除了在加强准确性方面有应用外,还由于多种数据集可以相互补充信息而应用到各个领域。如多模态融合在内容分析和检索领域流行起来是因为不同媒体间内容形成了互补^[25,36],而随着大量异构生物医学数据的出现,人们开发了许多集成这些数据集的方法,以从多个源域提取互补的知识^[37]。多源信息互补在日常生活中较为常见,如在会议上,通常有录音和笔录两种形式进行会议记录,不仅可以相互补充和扩展记录,而且当录音不清或笔记不清时可以相互补充缺失的那部分信息。

基于相似度的数据融合与集成,就是利用不同对象之间的潜在相关性(或相似度)融合不同的数据集,以形成较为完整的数据集。如果我们知道两个对象(X, Y)在某些度量方面是相似的,那么Y在缺少数据时可以利用X的信息。如一致性回归融合模型的主要思想是认为不同模态在描述同一事物时所表达的情感应该是一致的^[38];协同过滤中,用共同维度的不同矩阵训练不同的数据集,在研究相似用户相似的兴趣爱好和行为中,对不同用户数据进行学习,预测用户数据;矩阵分解中的奇异值分解(SVD)^[39]和非负矩阵分解(NMF)^[40]也被用来做相似度的数据融合与集成, Yang等介绍了一种基于经典矩阵分解技术的新型数据集成方法,其主要原理是将源之间的同质和异质效应分开,以从外部噪声中提取协调信号^[41]。对齐是另一种基于相似性的数据融合补充信息的方法,对齐利用每个数据集中实例的关系来加强数据集之间关系的知识,从而最终将最初不同的数据集映射到一个联合潜在空间,其中每个数据集中的局部相似实例和跨数据集的相应实例在该空间中接近或相同。

2.4.3 关联对象转移学习 机器学习的前提是学习数据和目标具有相同特征空间和分布,但有时候并不能获取到这样的条件的训练集。基于转移的机器学习将知识从源域转移到另一个目标域,以处理目标域中的数据稀疏性问题(包括要素结构缺失或观察缺失)。

根据转移学习的研究目标,又可以分为单一目标的转移学习和多目标协同学习。单一目标如从图书推荐到旅游推荐的转移,通过看桌子样式来预测椅子样式,学习骑自行车的经验帮助学习骑摩托车,以及看用户的交易记录推测用户的兴趣,然后也可以进行旅游推荐等,这些都只有一个研究目标,而数据集并不直接与研究目标相关,需要通过联系进行转移学习。多目标的转移学习则是拥有多个

数据集, 并且需要进行多项任务的预测, 不同的数据集通过共同学习来共同完成不同的目标, 如有用户图书借阅数据和旅游数据, 需要同时给用户进行图书和旅游地的推荐。两种转移学习图示如图 13 所示。

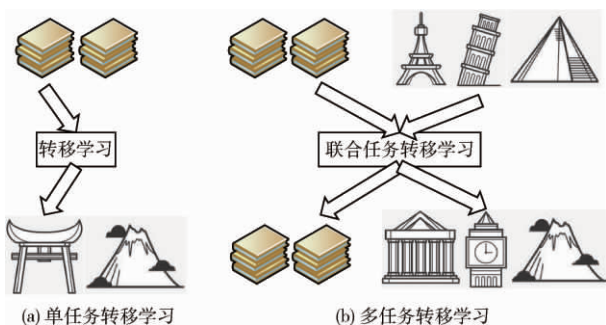


图 13 两种转移学习图示

不同领域的多任务转移学习有重要的现实意义, 把不同领域的数据集放在一起学习, 如空气质量与交通情况一起学习, 同时预测空气质量和交通情况会更有效, 因为这两者本身就有一定的联系, 如空气好人们更愿意出行, 出行多又影响空气质量。多任务转移学习不仅借用了其他领域数据集来实现本领域的预测目标, 也从多视角提高了预测任务的准确性。

3 应用与挑战

数据融合与集成的应用范围非常广, 在数据融合与集成过程中, 各领域有一些共同的挑战。在数据源上存在多种问题: 数据质量问题, 低质量的数据, 表现为不准确、错误、格式错误、不完整等^[42]; 噪音问题, 噪音不是多源多模态数据特有的, 但在多源多模态中产生了新的问题, 因为每种方法都可会产生噪音, 并且有可能相互影响; 还有数据量较大、数据不一致等问题, 在方法研究上难以寻找与目标领域完全相同的特征空间和分布; 机器学习的参数控制, 可解释性问题; 以及对融合效果没有较好评估方法等。

学者们也对这些问题进行研究, 如在硬件和设备上, 对于时空数据中广泛使用的大量无线传感器, 有学者提出一种带有数据融合的协议机会路由, 在源节点数量增加时, 能大幅度降低网络延迟, 并延长网络寿命^[43]; 在数据层面, 有学者通过建立时空多视角学习模型, 填补缺失数据^[44]; 在方法上, J. Tekli 等提出需要扩展 SemIndex 的功能, 以处理更具表达性的半结构化数据集以及关联数据集^[45], 以及降低维度的神经网络方法^[46]等。

除了通用挑战外, 由于各领域的特殊性, 数据融合与集成在各个领域都有自身难题需要解决, 如对硬件性能的要求, 对设备耐久度的要求等, 以下从多传感交互系统、

生物医学、环境与遥感以及城市建设 4 个研究应用领域进行论述。

3.1 多传感器系统

多传感器系统较为典型的有人机交互系统, 也包括需要多传感器采集数据分析的系统。视频录制通常包括声音和图像两种传感器的使用, 来形成目前的 2D 或 3D 效果, 在一些影院中还会加入其他传感器以形成更多维度上的传感, 如嗅觉、触觉等加入的 5D 影院。人机交互在计算机普及后, 目标更高, 通常不仅要能完成交互, 还要求交互过程是自然、有效率的, 这些交互建立在视听、触觉、味道、移动(手势感知)和语言命令等的传感器感知上。而且在人机交互中还实时性有较高要求。跨模态检索是多模态数据融合与集成的一个重要应用, 也属于人机交互的一种, 它基于不同模态数据间的特征建立联系, 从而完成不同模态之间的相互检索。J. Tekli 等提出了使用具有语义索引功能的知识库进行数据集成的方法就是利用语义进行多模态数据关联实现跨模态检索任务^[45]。

在多传感器系统应用领域中, 如何设置合适的传感器并自适应地选择足够的传感器仍然是一个挑战; 如何在快速变化的环境中(例如, 智能交通)实现隐私保护仍然是一个问题; 适用于多种情况的混合数据融合仍需要进一步的研究^[47]。

3.2 生物医学

在生物医学领域, 数据融合与集成用于了解大脑功能、医学诊断以及智能监控患者健康状态。脑成像是了解大脑活动的重要途径, 有多少元素参与到大脑感知活动, 目前已经可以从脑波动图像同步中记录数据。脑成像的挑战是 EEG 逆问题, 它常被表述为一个线性逆问题, 但颅骨内许多不同的空间电流模式可以产生相同的测量结果, 目前主要是通过大量假设和脑电图中加入大量的先验信息^[48], 但这显然是低效的。医学诊断即通过多方面数据对病情进行诊断, 医学上单一症状一般只作为怀疑是某种病症, 确诊需要做多项检查, 这些不同的检查指标共同进行医学诊断。如学者们用阿尔茨海默病的临床和多模态神经成像数据集来预测疾病的进展, 集成多种形式的数值数据, 包括时间序列和非时间序列数据, 以从多模态数据集中提取互补特征^[37]。智能监控患者是通过各种智能穿戴设备对患者用户进行各项健康数据的监测, 及时发现病情安全隐患。

然而使用移动和可穿戴设备收集用于人类活动识别的多种模式的大型数据集具有挑战性; 当前的人类活动检测和健康监控的互操作性和可伸缩性较为缺乏, 并且难以维持全面的辅助生活和健康监控; 重要的研究方向包括通过超梯度提升和证据推理分类器组合等增强方法对超参数调

整和分类器意见的融合进行设计和评估^[49]。

3.3 环境研究

环境研究分为两个级别,一个是太空级别,另一个是地球内部。太空研究主要包括宇宙天体的形成和机理探究,地球内部则是气候与自然环境的研究。宇宙和地球观测中最主要的应用就是研究外太空以及宇宙和地球的形成与发展。主要用到遥感技术,形成外空成像,如黑洞被证实是存在的,就是利用全球多个基地的数据进行融合与集成,再经过几年时间的复杂计算得出的结论。气象预测方面,基于遥感技术观测云图,多源数据融合也被应用于自然灾害的破坏图绘制,基于集成学习的数据融合能实现快速绘图^[50]。

在多源卫星图像融合方面,对相同输入数据下的融合水平和融合方法的性能分析比较研究仍然缺失,并且所有的研究都在相对较小的地图上进行,大面积的地图并没有纳入过研究范围^[51]。在多源遥感数据融合的处理上,不同的空间分辨率使得不同卫星传感器得到的图像存在差异,图像之间的地理配准存在困难,Zhu等提出未来的研究可以考虑开发对地理配准误差不敏感的时空融合方法:在数据融合期间可能发生突然的土地覆盖变化(比如洪水事件),这种情况会使得地理区域的边界变得模糊,Zhu等提到未来可以用深度学习方法来解决这一问题;此外,当前对融合结果没有一个广泛接受的评估方法,也没有标准的数据集,并且融合的计算效率有待提升,Zhu等指出有两种方法可以提高效率:基于融合图像的均匀性,在像素方式和区块方式计算之间取得平衡,以及使用更高级的编程策略^[52]。

3.4 城市建设

城市建设中,学者们研究较多包括交通问题和定位问题,以及各行业的生产动作的优化。Song等用多源数据融合进行交通拥堵情况的因素分析,提出了城市建设应该考虑微环境^[53]。还有学者通过结合手机信号和出租车全球定位数据,模拟交通路网中的交通流量,定位拥挤路段的车辆来源来选择控制方案^[54]。多源融合定位系统(MFPS)是使用数据融合技术的定位方案,其中不同的定位源在融合中心(FC)处合并在一起,以提高定位精度并获得更好的可靠性和鲁棒性^[55]。在生产优化上,Zhou Jingren等基于材料数据的融合对物品材料进行判断,并提出材料数据融合技术的优化应该集中在基于深度学习模型的优化、优化数据融合中使用的算法、从知识发现和知识使用的角度进行优化等方面。此外还需要注重对未知数据的处理和对数据融合中数据属性的保护^[11]。

城市建设中多源数据融合用到的数据量比单一来源更大,在收集、传输和处理上都降低了效率。在多源定位系

统中可以用大传感器数据预处理方法在数据提取、收集过程中对数据进行压缩,提高传输效率^[56]。L. Romeo等运用CSIL工具作为一种附加的时空分析解决方案,它简化了不同空间数据的处理、管理、分析和可视化。在基于时间的性能评估中显示,CSIL是一种节省时间的解决方案,特别是在处理多个完全不同的数据集时。在CSIL工具的开发和改进上,还需要进行整合不确定性分析,加权方案和增加工具可用性等工作^[57]。

4 结束语

本文从相关概念、路径与方法以及应用和挑战三个方面对数据融合与集成进行研究回顾与梳理。在相关概念上学者们给出了自己的见解,并未完全形成统一的概念,但并不影响研究者们对具体任务特征的掌控;在路径和方法研究上,多数学者都结合具体的领域给出了当前领域的数据融合与集成方法,并证明具有一定的成效;在实际应用中,各领域在不同程度上利用数据融合与集成去解决实际任务。结合现有研究成果,未来研究也可以从这三个方面出发:①在概念层面,随着研究的深入应该形成较统一的概念,它包括数据融合与集成的主要特征,并且各领域可以有自己的附加特征;②在路径研究上,可以进一步对数据融合与集成的通用结构模型进行改进与完善,并提供多样化接口以适应更多的人工智能方法;③在实际应用中,不断应对出现的难点与挑战,实现对研究工作的反馈,理论与实践共同促进数据融合与集成工作的发展。□

参考文献

- [1] ZHENG Y. Methodologies for cross-domain data fusion: an overview [J]. IEEE Transactions on Big Data, 2015, 1 (1): 16-34.
- [2] HALL D L, MCMULLEN S A H. Mathematical techniques in multisensor data fusion [M]. Artech House, 2004.
- [3] LAHAT D, ADALI T, JUTTEN C. Multimodal data fusion: an overview of methods, challenges, and prospects [J]. Proceedings of the IEEE, 2015, 103 (9): 1449-1477.
- [4] 巴志超,李纲,安璐,等. 国家安全大数据综合信息集成: 应用架构与实现路径 [J]. 中国软科学, 2018 (7): 9-20.
- [5] DAHER J B, BRUN A, BOYER A. A review on heterogeneous, multi-source and multi-dimensional data mining [EB/OL]. [2020-02-28]. <https://hal-lirmm.ccsd.cnrs.fr/INRIA/hal-01811232v1>.
- [6] LIU J, LI T, XIE P, et al. Urban big data fusion based on deep learning: an overview [J]. Information Fusion, 2020, 53: 123-133.
- [7] 赵亮. 多模态数据融合算法研究 [D]. 大连: 大连理工

- 大学, 2018.
- [8] WHITE F. Data fusion lexicon [M]. Virginia USA: Defense Technical Information Center, 1991: 16.
- [9] HALL D L, LLINAS J. An introduction to multisensor data fusion [J]. Proceedings of the IEEE, 1997, 85 (1): 6-23.
- [10] 祁友杰, 王琦. 多源数据融合算法综述 [J]. 航天电子对抗, 2017, 33 (6): 37-41.
- [11] ZHOU Jingren, HONG Xin, JIN Peiquan. Information fusion for multi-source material data: progress and challenges [J]. Applied Sciences, 2019, 9 (17): 3473.
- [12] DE S, GUPTA K, STANLEY R J, et al. A comprehensive multi-modal data fusion approach for failure assessment in aircraft lap-joint mimics [J]. IEEE Transactions on Instrumentation and Measurement, 2013, 62 (4): 814-827.
- [13] LENZERINI M. Tutorial-data integration: a theoretical perspective [J]. Symposium on Principles of Database Systems, 2003: 233-246.
- [14] WANG Ping, HUANG Fang, GUAN Li, et al. Framework for multi-sources spatial data integration analysis [C]. Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on. IEEE, 2006.
- [15] 张朝阳. 基于语义元数据的分布式异构数据库集成研究 [D]. 长沙: 国防科学技术大学, 2009.
- [16] 徐晓滨, 陈丽, 文成林. 一种基于多源异类信息统一表示的多传感器数据融合算法 [J]. 河南大学学报: 自然科学版, 2005 (3): 67-71.
- [17] 田鸽, 韩磊, 赵永华. 多源数据融合的实景三维建模在土地整治中的应用 [J]. 生态学杂志, 2019, 38 (7): 2236-2242.
- [18] 郭丽娟, 彭雪, 李正华, 等. 面向多领域多来源文本的汉语依存句法树库构建 [J]. 中文信息学报, 2019, 33 (2): 34-42.
- [19] ZHENG Yu. Methodologies for cross-domain data fusion: an overview [J]. IEEE Transactions on Big Data, 2015, 1 (1): 16-34.
- [20] NAUMANN F, BILKE A, BLEIHOLDER J, HERSCHEL M. Data fusion in three steps: resolving inconsistencies at schema, tuple-, and value-level [J]. IEEE Data Eng. Bull., 2006, 29 (1).
- [21] LLINAS J, HALL D L. An introduction to multi-sensor data fusion [C]. IEEE International Symposium on Circuits & Systems. IEEE, 1998.
- [22] TANG Jianzhi, REN Yingchao, YANG Chongjun, et al. A WebGIS for sharing and integration of multi-source heterogeneous spatial data [C] //2011 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2011, Vancouver, BC, Canada, July 24-29, 2011. IEEE, 2011.
- [23] ZHANG Weishan, ZHANG Yuanjie, ZHAI Jia, et al. Multi-source data fusion using deep learning for smart refrigerators [J]. Computers in Industry, 2018, 95: 15-21.
- [24] HUANG Chunlin, LI Xin. A review of land data assimilation system [J]. Remote Sensing Technology and Application, 2011, 19 (5).
- [25] ATREY P K, HOSSAIN M A, EL SADDIK A, et al. Multi-modal fusion for multimedia analysis: a survey [J]. Multimedia Systems, 2010, 16 (6): 345-379.
- [26] LIU W, ZHANG C, YU B, et al. A general multi-source data fusion framework [C] //Proceedings of the 2019 11th International Conference on Machine Learning and Computing, 2019.
- [27] 刘卓锟, 刘华平, 黄文美, 等. 视听觉跨模态表面材质检索 [J]. 智能系统学报, 2019, 14 (3): 423-429.
- [28] YU Yuhai, LIN Hongfei, MENG Jiana, ZHAO Zhehuan. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks [J]. Algorithms, 2016, 9 (2).
- [29] ZHENG Yu, LIU Yanchi, YUAN Jing, XIE Xing. Urban computing with taxicabs [C] //Proceedings of the 13th ACM International Conference on Ubiquitous Computing, 2011.
- [30] WANG Z, ZHANG D, ZHOU X, et al. Discovering and profiling overlapping communities in location-based social networks [J]. IEEE Transactions on Systems Man Cybernetics-Systems, 2014, 44 (4): 499-509.
- [31] WALTER V, FRITSCH D. Matching spatial data sets: a statistical approach [J]. International Journal of Geographical Information Science, 1999, 13 (5): 445-473.
- [32] NGIAM J, KHOSLA A, KIM M, et al. Multimodal deep learning [C]. Proceedings of International Conference on Machine Learning, Washington, 2011: 689-696.
- [33] 张素智, 陈小妮, 李鹏辉, 等. 食品安全大数据的融合及分类技术综述 [J]. 计算机技术与发展, 2020 (1): 1-11.
- [34] YUAN N J, ZHENG Y, XIE X, et al. Discovering urban functional zones using latent activity trajectories [J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27 (3): 712-725.
- [35] ZHENG Y, YI X, LI M, et al. Forecasting fine-grained air quality based on big data [C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.
- [36] SNOEK C G M, WORRING M. Multimodal video indexing: a review of the state-of-the-art [J]. Multimedia Tools and Applications, 2005, 25 (1): 5-35.
- [37] LEE G, KANG B, NHO K, et al. MildInt: deep learning-based multimodal longitudinal data integration framework [J]. Frontiers in Genetics, 2019, 10: 617.
- [38] YOU Q Z, LUO J B, JIN H L, et al. Cross-modality consist-

- ent regression for joint visual-textual sentiment analysis of social multimedia [M]. NEW YORK: Assoc Computing Machinery, 2016: 13-22.
- [39] GOLUB G H, REINSCH C. Singular value decomposition and least squares solutions [J]. *Numerische Mathematik*, 1970, 14 (5): 403-420.
- [40] HOYER P. Nonnegative matrix factorization with sparseness constraints [J]. *Journal of Machine Learning Research (JMLR)*, 2004, 5: 1457-1469.
- [41] YANG Z, MICHAELIDIS G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data [J]. *Bioinformatics*, 2016, 32 (1): 1-8.
- [42] GOUVEIA J, SEIXAS J, GIANNAKIDIS G. Smart city energy planning: integrating data and tools [C] // *Proceedings of the 25th International Conference Companion on World Wide Web*, Canada: International World Wide Web Conferences Steering Committee, 2016.
- [43] LI J, JIA X, LV X, et al. Opportunistic routing with data fusion for multi-source wireless sensor networks [J]. *Wireless Networks*, 2019, 25 (6): 3103-3113.
- [44] YI X, ZHENG Y, ZHANG J, et al. ST-MVL: filling missing values in geo-sensory time series data, 2016 [C] // *Proceedings of IJCAI*, 2016.
- [45] TEKLI J, CHBEIR R, TRAINA A J M, et al. Full-fledged semantic indexing and querying model designed for seamless integration in legacy RDBMS [J]. *Data & Knowledge Engineering*, 2018, 117: 133-173.
- [46] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. *Science*, 2006, 313 (5786): 504-507.
- [47] DING W, JING X, YAN Z, et al. A survey on data fusion in internet of things: towards secure and privacy-preserving fusion [J]. *Information Fusion*, 2019, 51: 129-144.
- [48] BIESSMANN F, PLIS S, MEINECKE F C, et al. Analysis of multimodal neuroimaging data [J]. *IEEE Reviews in Biomedical Engineering*, 2011, 4.
- [49] NWEKE H F, TEH Y W, MUJTABA G, et al. Data fusion and multiple classifier systems for human activity detection and health monitoring: review and open research directions [J]. *Information Fusion*, 2019, 46: 147-170.
- [50] ADRIANO B, XIA J, BAIER G, et al. Multi-source data fusion based on ensemble learning for rapid building damage mapping during the 2018 Sulawesi Earthquake and Tsunami in Palu, Indonesia [J]. *Remote Sensing*, 2019: 11 (7).
- [51] ORYNBAIKYZY A, GESSNER U, CONRAD C. Crop type classification using a combination of optical and radar remote sensing data: a review [J]. *International Journal of Remote Sensing*, 2019, 40 (17): 6553-6595.
- [52] ZHU X, CAI F, TIAN J, et al. Spatiotemporal fusion of multi-source remote sensing data: literature survey, taxonomy, principles, applications, and future directions [J]. *Remote Sensing*, 2018, 10.
- [53] SONG J, ZHAO C, ZHONG S, et al. Mapping spatio-temporal patterns and detecting the factors of traffic congestion with multi-source data fusion and mining techniques [J]. *Computers, Environment and Urban Systems*, 2019, 77: 101364.
- [54] PU W, CHENGCHENG W, JIYE L, et al. Traffic control approach based on multi-source data fusion [J]. *IET Intelligent Transport Systems*, 2019, 13 (5): 764-772.
- [55] ZHAO W, MENG W, CHI Y, et al. Factor graph based multi-source data fusion for wireless localization [C]. 2016 IEEE Wireless Communications and Networking Conference (WCNC), 2016.
- [56] WANLONG Z, SHUAI H, WEIXIAO M, et al. BSDP: big sensor data preprocessing in multi-source fusion positioning system using compressive sensing [J]. *IEEE Transactions on Vehicular Technology*, 2019, 68 (9): 8866-8880.
- [57] ROMEO L, NELSON J, WINGO P, et al. Cumulative spatial impact layers: a novel multivariate spatio-temporal analytical summarization tool [J]. *Transactions in GIS*, 2019, 23 (5): 908-936.

作者简介: 余辉, 男, 1993年生, 博士生。研究方向: 网络信息智能处理。梁镇涛, 男, 1996年生, 硕士生。研究方向: 信息分析与知识发现。鄢宇晨, 女, 1997年生, 硕士生。研究方向: 科学计量与知识图谱。

作者贡献声明: 余辉, 文献调研和撰写。梁镇涛, 文献修改和整理。鄢宇晨, 文献资料收集。

录用日期: 2020-06-04