doi：10.13413/j.cnki.jdxblxb.2023193

基于模糊数学的多源异构数据融合模型

李 鑫[1,2]，梁永玲[2]
（1. 中国科学院大学 沈阳计算机技术研究所，沈阳 110168；
2. 中国刑事警察学院，沈阳 110854）

摘要：针对多源异构型数据来源复杂、结构特殊，导致其融合难度较大的问题，为提升数据融合效率与准确性，提出一种基于模糊数学的多源异构数据融合模型. 首先，利用联邦加权平均融合策略，整合从各传感器传输至数据级融合层中的元数据，得到数据级融合结果；其次，结合主成分分析法与典型相关分析法，提取出由网络本体语言统一后数据的特征，完成特征级数据融合；再次，根据模糊数学理论建立、更新模糊规则库，通过决策融合算法，得到决策级融合结果；最后，将上述不同层级的数据融合结果结合建立数据融合模型，得到最终的数据融合结果. 实验结果表明，该方法的最大协方差值和绝对误差值不超过 0.15，最短融合时间仅为 12.6 ms. 该方法的融合精度和稳定性较好，时效性与抗扰性均具有显著的优越性.

关键词：模糊数学；典型相关分析法；模糊规则；多源异构数据；数据融合
中图分类号：TP391 **文献标志码**：A **文章编号**：1671-5489(2024)03-0691-06

Multi-source Heterogeneous Data Fusion Model Based on Fuzzy Mathematics

LI Xin[1,2], LIANG Yongling[2]
(1. *Shenyang Institute of Computing Technology, University of Chinese Academy of Sciences, Shenyang* 110168, *China*;
2. *Criminal Investigation Police University of China, Shenyang* 110854, *China*)

Abstract: Aiming at the problem that multi-source heterogeneous data had the complex sources and unique structure, resulting in a greater difficulty in its fusion. In order to improve the efficiency and accuracy of data fusion, we proposed a multi-source heterogeneous data fusion model based on fuzzy mathematics. Firstly, by utilizing a federated weighted average fusion strategy, the metadata transmitted from various sensors to the data level fusion layer was integrated to obtain the data level fusion results. Secondly, combined with the principal component analysis method and canonical correlation analysis method, the features of data unified by Web Ontology Language were extracted to complete the feature level data fusion. Thirdly, a fuzzy rule library established and updated based on fuzzy mathematics theory was used to obtain decision level fusion results through decision fusion algorithms. Finally, we combined the data fusion results of above different levels to establish a data fusion model, and obtained the final data fusion result. The experimental results show that the maximum covariance value and absolute error value of the proposed method do not exceed 0.15, and

收稿日期：2023-05-15.

第一作者简介：李 鑫(1985—)，男，汉族，硕士，讲师，从事大数据挖掘、计算机应用和公安情报分析的研究，E-mail：xjxylx@126.com.

基金项目：国家重点研发计划“公共安全风险防控与应急技术装备”专项基金(批准号：21405).

the shortest fusion time is only 12.6 ms. The fusion accuracy and stability of this method are good, and both timeliness and anti-interference have significant advantages.

Keywords: fuzzy mathematics; typical correlation analysis method; fuzzy rule; multi-source heterogeneous data; data fusion

随着互联网信息技术与云计算技术的不断升级换代，使目前社会已全面进入大数据时代[1]. 在传感技术与物联网的共同催化下，数据感知技术已打破了以往的界限，改变着人们的生活方式、行业的运作模式及科研的探讨方向[2]. 随着感知数据规模越来越大，格式、内容、关系逐渐呈现出多样化的发展趋势，增长动态化、分布范围广、来源差异明显、结构不等价等特征愈加显著[3-5]. 在数据处理的不同阶段中，以感知数据为支撑的融合操作尤为关键. 该处理过程不仅是数据质量与决策制定的有效保证，也是完成数据分析与挖掘的前提. 数据融合技术能赋予数据更高的应用价值，是信息社会发展方向的决定性因素.

作为数据处理方向的研究热点，数据融合方法受到广泛关注. 如夏伟等[6]面向电力行业，采用最小二乘法、Lagrange 插值法、联合 Kalman 滤波法，融合了配电网中的多源异构数据；张巧灵等[7]和莫慧凌等[8]以互联网络为研究对象，分别采用混合型神经网络与 Tucker 分解理论设计了数据融合算法.

多源异构型数据因其来源复杂、结构特殊，需要综合性更强、鲁棒性更高的融合技术. 模糊数学是解决模糊性问题的主流方法. 因此，将模糊数学作为理论基础，可应对存在模糊现象的多源异构数据融合应用场景. 基于模糊数学理论建立模糊规则库，抽象且准确地描述出多源异构数据间特殊的逻辑规则，可减缓数据来源、结构等冲突. 利用主成分分析法提取数据特征，降低特征维度，可使数据在保留有用信息的同时，仍具有强关联性.

1 多源异构数据融合

多源异构数据具有数据来源差异明显、结构不等价等特征. 为使数据间的连通与反馈更优，更好地补偿模糊信息导致的鲁棒性能衰减，可通过数据级融合、特征级融合和决策级融合方法，利用不同的融合层次，从多角度全方位地完成多源异构数据的融合[9-11]. 图 1 为多源异构数据融合流程.

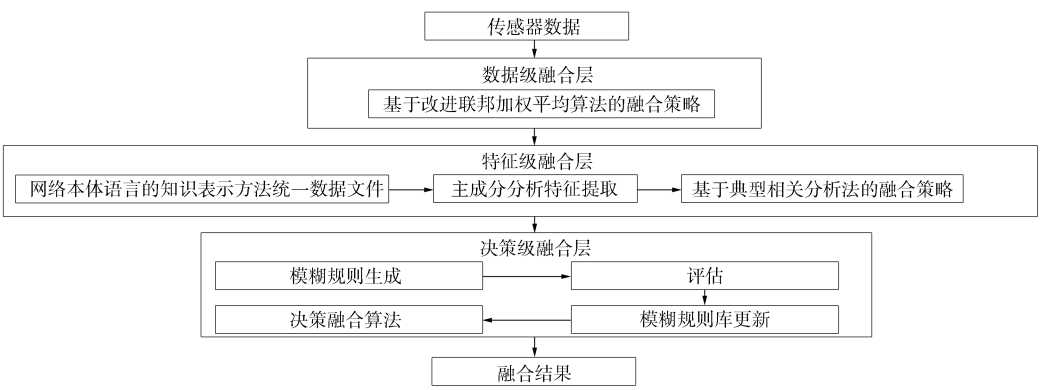


图 1 多源异构数据融合流程

Fig. 1 Flow chart of multi-source heterogeneous data fusion

由图 1 可见，带有传感器物理特性的元数据传输至数据级融合层后，凭借联邦加权平均融合策略的自适应搜寻能力，找到不同传感器的最佳加权值. 通过最小化多源异构数据的总均方误差值，取得最好的数据级融合效果.

采用网络本体语言(ontology wed language, OWL)的知识表示方法[12]，统一、规范化地描述完成数据级融合后的结构化数据文件，以便于建立数据间的强关联性. 根据主成分分析法提取的数据属性特征，利用基于典型相关分析法的融合策略，完成特征级融合层的融合操作.

将特征级融合层的融合结果作为决策级融合层的输入，基于模糊数学理论建立模糊规则库，抽象且准确地描述多源异构数据间特殊的逻辑规则，缓解数据来源、结构等冲突，避免发生异常现象. 为确保模糊规则可用且非过拟合，利用支持度与置信度指标进行评估，更新模糊规则库. 通过提出的决策融合算法，获取融合结果.

1.1　数据级融合

首先，基于改进的联邦加权平均算法[13]，建立数据级融合策略. 假设传感器数量为 n，采集到的元数据集合为$\{x_1,x_2,\cdots,x_n\}$，各数据的对应权重集合为$\{\omega_1,\omega_2,\cdots,\omega_n\}$. 若 X 是数据集$\{x_1,x_2,\cdots,x_n\}$的无偏估计值，则传感器元数据的总均方误差为

$$\Delta S^2=\sum_{i=1,\ j=1,\ i\neq j}^{n}\left[\omega_i(X-x_i)^2-\omega_j(X-x_j)^2\right], \tag{1}$$

其中 ω_i,ω_j 为元数据 x_i,x_j 的对应权重.

其次，通过多元函数极值求解法，获取各传感器的最佳加权因子：

$$\omega_i'=\frac{\omega_i(X-x_i)^2-\omega_j(X-x_j)^2}{\Delta S^2}. \tag{2}$$

最后，结合第 i 个传感器的历史数据均值 $\overline{x}_i'$ 及最小化均方误差 $\Delta S^2_{\min}=\min\ \Delta S^2$，得到数据融合结果：

$$\hat{x}=\sum_{i=1}^{n}\frac{\omega_i'(x_i+\overline{x}_i')}{\Delta S^2_{\min}}. \tag{3}$$

1.2　特征级融合

利用主成分分析法提取数据特征，降低特征维度，使数据在保留有用信息的同时，仍具有强关联性. 基于方差最大化准则，利用主成分分析法，线性组合由数据 p 维特征向量所组成的特征矩阵$(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_p)$，取得数据初始特征的 p 个主成分：

$$\boldsymbol{f}_p=\sum_{k=1}^{p}a_{kp}\boldsymbol{y}_p,\qquad k\in p, \tag{4}$$

其中 a_{kp}是特征矩阵的系数，各系数的平方和值为 1.

针对提取出的数据特征(即 p 个主成分)，以典型相关分析法为基础，设计出特征级数据融合策略. 假设数据第 k 个和第 l 个主成分 $\boldsymbol{y}_k,\boldsymbol{y}_l$ 在空间中的投影分别是 $\boldsymbol{y}_k',\boldsymbol{y}_l'$(即典型相关变量)，对应的向量为 $\boldsymbol{\alpha},\boldsymbol{\beta}$，则投影与向量满足下列关系：

$$\begin{cases}\boldsymbol{y}_k'=\boldsymbol{\alpha}\boldsymbol{f}_p\boldsymbol{y}_k,\\ \boldsymbol{y}_l'=\boldsymbol{\beta}\boldsymbol{f}_p\boldsymbol{y}_l,\end{cases}\qquad k,l\in p. \tag{5}$$

通过典型相关分析，令特征投影 $\boldsymbol{y}_k'$ 与 $\boldsymbol{y}_l'$ 间拥有最强的关联性，以此得到主成分 $\boldsymbol{y}_k,\boldsymbol{y}_l$ 的全部典型相关特征. 若两成分的典型特征分量为 $d_{\boldsymbol{y}_k}$和 $d_{\boldsymbol{y}_l}$，则通过下列融合策略的数学表达式即可完成数据的特征级融合，得到融合结果 $\hat{\boldsymbol{y}}$：

$$\hat{\boldsymbol{y}}=\begin{pmatrix}d_{\boldsymbol{y}_k} & \boldsymbol{y}_k\\ d_{\boldsymbol{y}_l} & \boldsymbol{y}_l\end{pmatrix}. \tag{6}$$

1.3　决策级融合

基于模糊数学理论，找到融合特征中决策代表性最强的部分，生成模糊规则，定义多源异构数据的权限区域，并指定有效的决策者[14]. 模糊规则的生成条件是：如果强代表性的决策部分 A 中有数据特征向量 $\hat{\boldsymbol{y}}$，则该部分中决策者 D 做出正确决策的概率为 $\overline{P}_D^A$. 概率计算公式为

$$\overline{P}_D^A=\frac{P(C\mid D)}{\sum_{M=1}^{n}P_{D_M}^A}, \tag{7}$$

其中 $P(C\mid D)$表示决策者 D 正确标记 C 类规则的次数，$P_{D_M}^A$表示 M 位决策者在决策部分 A 中给出正确决策的次数.

利用支持度和置信度评估在部分区域内是否存在大规模数据和高概率数据，保证规则的可用性，

避免过拟合现象. 根据评估结果删除多余的不可用规则，更新模糊规则库[15-16]. 将向量 $\boldsymbol{\varphi}$ 赋予对应规则的数据源，则规则库中的第 h 条规则为 $\varphi_{P_M h}$. 由此推导出该规则对应数据源权值的计算公式为

$$w_{C_h}=\frac{\overline{P}_D^A}{\varphi_{P_M h}}. \tag{8}$$

令规则库权重均值为各类别权值，则拥有最高权值的类即得到最终决策信息. 决策级融合结果为

$$\hat{z}=\frac{\sum_{h=1}^{n}(w_{C_h}-\overline{w})}{w_{C_h}-\overline{w}}>0, \tag{9}$$

其中 $\overline{w}$ 表示规则库权重均值，w_{C_h} 表示第 h 条规则所属类别 C_h 的权重.

最后将多源异构数据的数据级、特征级和决策级融合结果进行整合，建立数据融合模型[17-19]，得到最终的多源异构数据融合结果为

$$U=\sum_{i,j=1}^{n}(\hat{x}+\hat{\boldsymbol{y}}+\hat{z})^{ij}. \tag{10}$$

2 实验与结果分析

2.1 实验数据

利用 Python 仿真软件和随机数发生器，得到 9 组含有 3 个随机数的数组，用作多个传感器的模拟采集数据. 将 3 组数据分别作为温度传感器、压力传感器、液体流量传感器的监测数据，得到 3 个传感器的多源异构数据集合，作为实验所需数据集. 数据集信息列于表 1. 为验证模型的可行性和优越性，分别采用基于联合 Kalman 滤波、基于时间序列、基于联邦学习 3 种数据融合方法与本文方法进行对比.

表 1 实验测试数据集信息

Table 1 Information of experimental test dataset

实验组别	温度传感器/℃	压力传感器/Pa	液体流量传感器/mL
1	62.1	912.2	17.8
	90.7	455.4	28.0
	73.0	531.9	63.8
2	76.1	248.3	33.9
	66.6	870.4	17.7
	69.3	845.6	94.6
3	82.2	359.8	23.6
	95.4	621.9	46.2
	68.0	812.9	38.4

2.2 数据融合准确度

采用协方差值检验数据融合准确度，不同数据融合方法的数据融合准确度测试结果如图 2 所示. 由图 2 可见，基于联合 Kalman 滤波、基于时间序列、基于联邦学习 3 种数据融合方法的融合协方差最大值分别是 0.69,0.80,0.67，而本文方法的协方差最大值不超过 0.15. 且 3 种对比方法的波动幅度较大，其中最不稳定的是联合 Kalman 滤波融合法，而本文方法的变化幅度较小. 实验结果表明，本文方法在融合精度和稳定性方面性能良好.

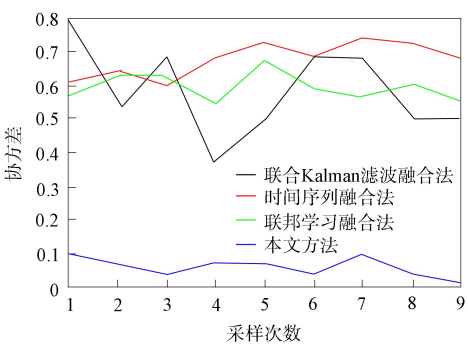


图 2 不同方法的数据融合协方差值对比

Fig. 2 Comparison of data fusion covariance values of different methods

2.3 数据融合效率

不同方法的数据融合效率对比结果如图 3 所

示. 由图 3 可见，前 4 次采样结束时，4 种融合方法的处理时长偏差较小，之后差距逐渐明显. 在 9 次采样中，基于联合 Kalman 滤波、基于时间序列、基于联邦学习 3 种数据融合方法的平均融合时间分别是 33.14，33.62，36.08 ms，且随着数据规模、冗余数据量及冲突的增加，融合时间越来越长. 而本文方法极大缩短了多源异构数据的融合处理时间，具有较优的融合速度. 最长融合时间仅为 20.04 ms，最短融合时间仅为 12.6 ms，表明其数据融合效率较高.

2.4　数据融合抗扰性

在实际使用中，传感器会出现失效问题，因此，在压力传感器的实验数据中添加固定的偏置量，通过绝对误差对数据融合效果进行评价，侧面检验各种融合方法对失效情况的抗干扰能力，结果如图 4 所示. 由图 4 可见，本文方法每次采样数据融合后的绝对误差数值均低于对比方法. 因为该方法基于模糊数学理论建立模糊规则库，准确地描述了多源异构数据间特殊的逻辑规则，即使传感器失效，仍能高质量地完成数据融合.

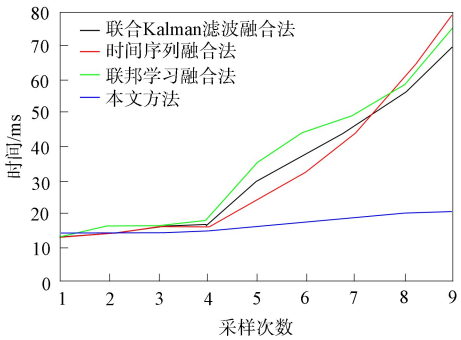


图 3　不同方法的数据融合时间对比

Fig. 3　Comparison of data fusion time of different methods

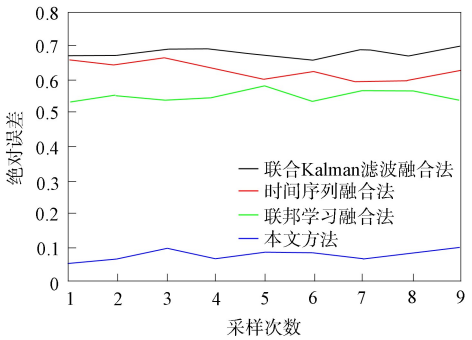


图 4　不同方法的数据融合绝对误差值对比

Fig. 4　Comparison of absolute error values of data fusion of different methods

综上所述，针对海量多源异构数据的融合处理难题，本文将模糊数学理论应用于多源异构数据融合处理中. 实验结果表明：该方法的协方差最大值不超过 0.15，说明其数据融合精度较高；最长融合时间仅为 20.04 ms，最短融合时间仅为 12.6 ms，说明数据融合效率较高. 实验结果验证了本文方法的实际应用效果，说明其具有可行性.

参　考　文　献

[1] 宋涛，李秀华，李辉，等. 大数据时代下车联网安全加密认证技术研究综述 [J]. 计算机科学，2022，49(4)：340-353.（SONG T，LI X H，LI H，et al. Overview of Research on Security Encryption Authentication Technology of IoV in Big Data Era [J]. Computer Science，2022，49(4)：340-353.）

[2] 陈广，宋志伟，陈少兵，等. 数据感知技术在电力物资供应链数据质量管理中的应用 [J]. 科技管理研究，2021，41(18)：182-191.（CHEN G，SONG Z W，CHEN S B，et al. Application of Data Perception Technology in Data Quality Management of Power Material Supply Chain [J]. Science and Technology Management Research，2021，41(18)：182-191.）

[3] 曾丽，曾玉林. 基于节点信誉度的传感器数据安全融合方法 [J]. 计算机仿真，2021，38(7)：290-293.（ZENG L，ZENG Y L. Sensor Data Security Fusion Method Based on Node Reputation [J]. Computer Simulation，2021，38(7)：290-293.）

[4] 许小媛，李海波，黄黎. 云存储多异构文件联合延迟尾概率凸优化分析 [J]. 计算机工程与应用，2021，57(5)：88-94.（XU X Y，LI H B，HUANG L. Convex Optimization Analysis of Joint Delay Tail Probability of Multi-heterogeneous Files in Cloud Storage [J]. Computer Engineering and Applications，2021，57(5)：88-94.）

[5] 孟嘉，厉文婕，于广荣，等. 面向效用最大化的数据中心动态资源分配 [J]. 计算机应用研究，2021，38(6)：1728-1733.（MENG J，LI W J，YU G R，et al. Maximizing Utilities in Data Center with Dynamic Resource Allocation [J]. Application Research of Computers，2021，38(6)：1728-1733.）

[6] 夏伟，蔡文婷，刘阳，等. 基于联合 Kalman 滤波的配电网多源异构数据融合 [J]. 电力系统保护与控制，2022，

50(10): 180-187. (XIA W, CAI W T, LIU Y, et al. Multi-source Heterogeneous Data Fusion of a Distribution Network Based on a Joint Kalman Filter [J]. Power System Protection and Control, 2022, 50(10): 180-187.)

[7] 张巧灵, 高淑萍, 何迪, 等. 基于时间序列的混合神经网络数据融合算法 [J]. 应用数学和力学, 2021, 42(1): 82-91. (ZHANG Q L, GAO S P, HE D, et al. A Hybrid Neural Network Data Fusion Algorithm Based on Time Series [J]. Applied Mathematics and Mechanics, 2021, 42(1): 82-91.)

[8] 莫慧凌, 郑海峰, 高敏, 等. 基于联邦学习的多源异构数据融合算法 [J]. 计算机研究与发展, 2022, 59(2): 478-487. (MO H L, ZHENG H F, GAO M, et al. Multi-source Heterogeneous Data Fusion Based on Federated Learning [J]. Journal of Computer Research and Development, 2022, 59(2): 478-487.)

[9] HAN Y, BAO W X, ZHANG X W, et al. Hyperspectral and Multispectral Data Fusion via Nonlocal Low-Rank Learning [J]. Journal of Applied Remote Sensing, 2022, 16(1): 016508-1-016508-18.

[10] LIU Y, ZHANG Y. A Weighted Evidence Combination Method for Multisensor Data Fusion [J]. Journal of Internet Technology, 2022, 23(3): 553-560.

[11] RAIMONDI D, SIMM J, ARANY A, et al. A Novel Method for Data Fusion over Entity-Relation Graphs and Its Application to Protein-Protein Interaction Prediction [J]. Bioinformatics, 2021, 37(16): 2275-2281.

[12] 唐成华, 侯梦迪, 高庆泽, 等. 多类软件本体的恶意软件语义描述模型 [J]. 小型微型计算机系统, 2021, 42(11): 2433-2439. (TANG C H, HOU M D, GAO Q Z, et al. Malware Semantic Description Model Based on Multi-class Software Ontology [J]. Journal of Chinese Computer Systems, 2021, 42(11): 2433-2439.)

[13] 罗长银, 王君宇, 陈学斌, 等. 改进的联邦加权平均算法 [J]. 计算机应用, 2022, 42(4): 1131-1136. (LUO C Y, WANG J Y, CHEN X B, et al. Improved Federated Weighted Average Algorithm [J]. Journal of Computer Applications, 2022, 42(4): 1131-1136.)

[14] 曹文欣, 赵文, 路博, 等. 基于模糊数学的STS管幕结构的连接参数优化 [J]. 东北大学学报(自然科学版), 2022, 43(2): 258-265. (CAO W X, ZHAO W, LU B, et al. Optimization of Connection Parameters of Steel Tube Slab Structures Based on Fuzzy Mathematics [J]. Journal of Northeastern University (Natural Science), 2022, 43(2): 258-265.)

[15] 王璐, 李根梓, 戴亚飞, 等. 基于模糊集合的评审决策信息融合 [J]. 物理化学学报, 2023, 39(3): 104-108. (WANG L, LI G Z, DAI Y F, et al. On Information Fusion by Fuzzy Sets for Decision Making in Peer Review [J]. Acta Physico-Chimica Sinica, 2023, 39(3): 104-108.)

[16] 杨阳, 王鹏江, 吉晓冬, 等. 基于模糊数学的掘进机截割头转速分档预测方法 [J]. 煤炭工程, 2022, 54(2): 172-176. (YANG Y, WANG P J, JI X D, et al. Prediction for Roadheader Cutting Head Speed of Different Gears Based on Fuzzy Mathematics [J]. Coal Engineering, 2022, 54(2): 172-176.)

[17] 马亚彤, 王松, 刘英芳. 融合多模态数据的人体动作识别方法研究 [J]. 计算机工程, 2022, 48(9): 180-188. (MA Y T, WANG S, LIU Y F. Research on Human Action Recognition Method by Fusing Multimodal Data [J]. Computer Engineering, 2022, 48(9): 180-188.)

[18] 郝刚, 金涛. 基于多尺度数据融合的锂电池健康状态评估 [J]. 江苏大学学报(自然科学版), 2023, 44(5): 524-529. (HAO G, JIN T. Lithium Battery Health Evaluation Based on Multi-scale Data Fusion [J]. Journal of Jiangsu University (Natural Science Edition), 2023, 44(5): 524-529.)

[19] 徐伟华, 黄旭东, 蔡可. 基于粒计算的多源信息融合方法综述 [J]. 数据采集与处理, 2023, 38(2): 245-261. (XU W H, HUANG X D, CAI K. Review of Multi-source Information Fusion Methods Based on Granular Computing [J]. Journal of Data Acquisition and Processing, 2023, 38(2): 245-261.)

(责任编辑: 韩 啸)