

基于深度聚类的高维临床数据用药规律研究

邓媛媛¹, 游海心¹, 梁 杨¹, 李力松¹, 刘 静¹, 曾 光², 丁长松^{1,3*}

1. 湖南中医药大学信息科学与工程学院, 湖南 长沙 410208

2. 湖南中医药大学中西医结合学院, 湖南 长沙 410208

3. 湖南省中医药大数据分析实验室, 湖南 长沙 410208

摘要: 目的 借助深度聚类在高维深层数据处理中的优势, 构建一种临床用药规律分析模型, 挖掘复杂多维的临床医学数据中的潜在用药规律。方法 收集真实世界临床脑卒中数据集进行筛选与规范, 利用图卷积神经网络、自动编码器和自监督机制等技术, 构建多源异构信息融合的深度聚类模型——多源融合卷积网络 (multi-source fusion convolutional network, MFCN)。通过 2 个数据集验证模型的性能, 并以脑卒中数据集为例, 进一步分析其临床用药规律, 从而验证模型挖掘高维临床数据的有效性。结果 MFCN 模型在 DBLP 和脑卒中数据集上的准确率分别为 79.32%、83.48%, 归一化互信息 (normalized mutual information, NMI) 分别为 0.490 8、0.531 6, 调整兰德系数 (average rand index, ARI) 分别为 0.541 9、0.581 7, F1 得分 (F1-score, F1) 分别为 0.787 7、0.833 9, 其性能指标均高于其他模型。在脑卒中数据集中, MFCN 模型成功挖掘出急性期的中药组合社团, 如甘草、茯苓和陈皮等, 并发现“症-药”关联组合, 如“甘草-脉弦”“甘草-舌红”“陈皮-肢体乏力”等。结论 深度聚类能有效处理复杂高维的临床数据, 并揭示临床用药的潜在规律, 为临床经验的提取和辅助决策提供了方法学参考。

关键词: 脑卒中; 深度聚类; 用药规律; 数据挖掘; 多源融合卷积网络

中图分类号: TP18; R285 **文献标志码:** A **文章编号:** 0253-2670(2025)05-1708-09

DOI: 10.7501/j.issn.0253-2670.2025.05.020

Research on medication rules in high-dimensional clinical data based on deep clustering

DENG Yuanyuan¹, YOU Haixin¹, LIANG Yang¹, LI Lisong¹, LIU Jing¹, ZENG Guang², DING Changsong^{1,3}

1. School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China

2. College of Integrated Chinese and Western Medicine, Hunan University of Chinese Medicine, Changsha 410208, China

3. Hunan Laboratory of TCM Big Data Analysis, Changsha 410208, China

Abstract: Objective To develop a clinical medication pattern analysis model to identify potential medication patterns within multi-dimensional clinical data by utilizing the advantages of deep clustering in high-dimensional and complex data processing. **Methods** The real-world clinical stroke data were collected for screening and specification. Then, techniques such as graph convolutional neural networks, autoencoders, and self supervised mechanisms are used to construct a deep clustering model—multi-source fusion convolutional network (MFCN) for multi-source information fusion. The performance of the model was validated on two datasets, with the stroke dataset used as a case study to further analyze its clinical medication patterns and demonstrate the effectiveness of the model in mining high-dimensional clinical data. **Results** The accuracy of the MFCN model on the DBLP and stroke datasets were 79.32% and 83.48%, respectively. The normalized mutual information (NMI) indicators were 0.490 8 and 0.531 6, the average rand index (ARI) indicators were 0.541 9 and 0.581 7, and the F1-score (F1) indicators were 0.787 7 and 0.833 9, respectively. Its

收稿日期: 2024-11-13

基金项目: 国家自然科学基金面上项目 (82474352); 湖南省中医药管理局重点项目 (A2024011, 2023-24); 湖南省自然科学基金项目 (2023JJ60124); 湖南省教育厅科学研究重点项目 (22A0255, 22A0281); 湖南省教育厅优秀青年项目 (22B0400); 湖南省教育厅科学研究一般项目 (22C0195); 长沙市自然科学基金项目 (kq2202265, kq2402172); 湖南中医药大学校级科研基金项目 (2021XJJ021)

作者简介: 邓媛媛, 女, 硕士研究生, 研究方向为中医药数据挖掘。E-mail: 514589246@qq.com

***通信作者:** 丁长松, 男, 教授, 博士研究生导师。E-mail: dingcs1975@hnu.edu.cn

performance indicators were higher than those of other models. In the stroke dataset, the MFCN model successfully identified Chinese herbal combinations in the acute phase, such as licorice, Fuling (*Poria*), and Chenpi (*Citri Reticulatae Pericarpium*), and discovered “symptom-drug” associations, such as “Gancao (*Glycyrrhizae Radix et Rhizoma*)-mai xian”, “*Glycyrrhizae Radix et Rhizoma*-tongue red”, and “Chenpi (*Citri Reticulatae Pericarpium*)-limb weakness”. **Conclusion** Deep clustering can effectively handle complex high-dimensional clinical data and reveal potential patterns of clinical medication, with view to providing methodological references for extracting clinical experience and assisting decision-making.

Key words: stroke; deep clustering; medication rules; data mining; multi-source fusion convolutional network

临床医案蕴藏了丰富的临床用药规律，对于临床药物治疗及疾病机制的探索具有重要的指导意义。然而，随着现代医学信息化的迅速发展，临床数据逐渐呈现出大规模、高维度、多样化的特点，同时还面临着噪声、描述不规范的问题，进一步增加了挖掘用药规律的难度。如何从这些复杂且高维的临床数据中提取出用药规律，是临床用药数据研究亟待解决的关键问题之一。

聚类算法作为中医的数据挖掘方法之一，能够在混乱的数据中发现其潜在的结构和模式^[1]。随着深度学习的发展，深度神经网络已具备深层次大规模的特征提取的能力^[2-3]。深度聚类将深度学习的特征提取能力与聚类算法的分类能力相结合，能够在无标签的条件下有效发现数据中的潜在模式。相比于传统的聚类方法，深度聚类在处理高维、非线性数据时表现出明显的优势^[4-5]，其能够自动学习和提取数据的潜在特征，使得对复杂数据的分群和模式发现更加准确。

在深度聚类研究中，学习有效的数据表示是深度聚类的关键先决条件。基于传统聚类方法，Yang 等^[6]利用自动编码器（auto-encoder, AE）提取原始样本特征并学习到一个具备良好判别性的特征表示，然后在特征向量上使用 K 均值（K-means）^[7]进行聚类分簇，最终获取样本的聚类结果。这种聚类方法只关注了数据的属性信息，忽略了数据之间的拓扑关系。受到图能反映数据间的拓扑关系的启发，Kipf 等^[8]提出了一种类似编码-解码结构的图自动编码器（graph convolutional auto-encoder, GAE）模型，其中将图卷积网络层（graph convolutional network, GCN）作为编码器学习节点特征，通过内积进行重构，采用交叉熵作为损失函数优化网络。为了同时捕获特征信息和拓扑结构信息，Bo 等^[9]提出了一种基于数据结构的聚类网络（structural deep clustering network, SDCN），该网络以双自监督的方式优化了 GCN 和深度嵌入式聚类（deep embedding clustering,

DEC）^[10]。为了多个网络生成的不同聚类分配的一致性，Yan 等^[11]提出了一种属性图聚类的双通道增强网络（mutual boost network, MBN），将节点特征和结构特征的异构信息同时传播，并通过对比聚类分配来提供相互指导，以自监督的方式进行同步优化。

为了进一步探索潜藏在海量临床病案的用药规律，本研究提出了一种深度聚类模型——多源融合卷积网络（multi-source fusion convolutional network, MFCN），融合了 GCN、AE 和聚类算法等技术，在脑卒中和计算机文献数据集（database systems and logic programming, DBLP）^[9]中验证其性能，并且以脑卒中数据集为例分析其临床用药规律，以期为临床用药提供依据，也为新药研发、中医创新发展提供新的研究范式。

1 方法

1.1 数据来源

DBLP 数据集是一个公开数据集，其描述了一个作者网络，在网络上的 2 个共同作者之间有 1 个链接。其涵盖了 4 个研究领域：机器学习、信息检索、数据挖掘和数据库。脑卒中数据集收集于湖南中医药大学附属第一医院信息科的信息数据库，就诊时间限定为 2015 年 11 月—2023 年 1 月，纳入经治疗且确有疗效的脑卒中患者 7 224 例，可分为急性期、恢复期和后遗症期 3 个阶段。

DBLP 为公共数据集，不需要进行数据筛选和预处理。脑卒中数据集为真实世界临床脑卒中病例数据集，其中涵盖了患者的病史记录、诊疗数据、药物使用情况等。临床病例数据均为人工表述，导致数据存在噪声及冗余，需进一步处理。

1.2 脑卒中数据纳入及排除标准

纳入标准：（1）就诊表中西医诊断描述为“脑卒中”“脑梗死”“脑梗塞”三者其中之一，或者中医诊断为“中风”的患者，年龄、性别不限；（2）使用中医或中西医结合治疗方法对脑卒中进行治疗，并且有明确的治疗效果；（3）病例内容详实完整，

包括临床表现、用药、具体剂量等信息。排除标准：剔除其他并发疾病过重的病例。

1.3 脑卒中数据预处理

1.3.1 中药正异名处理 依据《中国药典》2020 年版^[12]对脑卒中数据的中药方剂及中成药处方中的中药名称进行规范及统一，药典未收录的中药则参照《中华本草》^[13]、《中药大辞典》^[14]并结合《中药学》教材第 4 版^[15]进行核查。针对一药多名中药，如“生地”“生地黄”等统一为“生地黄”；将不同炮制方法的同种中药名称进行统一，如“燂桃仁”统一为“桃仁”。

1.3.2 症状术语规范 从《中医常见症状术语规范（2005 年版）》^[16]中收集 2 660 个症状术语，并按其部位进行分类，形成 42 个小类症状，举例如表 1 所示。基于这一标准术语，对脑卒中数据所记录的症状进行数据预处理。

1.3.3 中药、症状以及脑卒中病例向量构建 使用

表 1 症状术语分类

Table 1 Classification of symptom terminology

症状举例	小类名	大类名
头痛	疼痛部位类	疼痛症状
健忘	精神情志类	精神情志及睡眠症状
四肢乏力	颈、胸、腹、背、四肢 不适类	全身不适症状
大便秘结	二便症状	二便症状
口角歪斜	面部特征	头面体征

Word2vec^[17]词向量技术对规范后的症状进行表征初始化，从而生成 200 维度的具有语义的症状词嵌入。同时，使用 23 维向量来表征中药向量^[18]，将药物属性抽象描述数值化以构建初始特征。对于“四性”“五味”与“毒性”属性采用离散值量化，如“微寒”“寒”“大寒”分别量化为 0.5、1、2；“归经”属性采用二值量化，即有此属性用 1 表示，无此属性用 0 表示。药性量化示例数据如表 2 所示。

表 2 药性量化示例

Table 2 Quantification example of drug properties

药名	寒	热	温	凉	平	酸	苦	甘	辛	咸	肺	心包	心	肠	三焦	小肠	胃	胆	膀胱	脾	肝	肾	毒
麻黄	0	0	1	0	0	0	0.5	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0
桂枝	0	0	1	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0
紫苏	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
生姜	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0
香薷	0	0	0.5	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0

病例向量由患者的症状集表示和治疗的中药集表示构成。中药集由多味中药组成，其向量表示通过对各中药向量求和后取平均获得。同理，症状集向量由多个症状向量求和取平均得到。因此，病例节点向量通过 23 维中药集向量与 200 维症状集向量组合形成 223 维表示（图 1）。

1.3.4 脑卒中病例节点构图 以患者症状和治疗中药构建病例关系图。为了减弱常见症状（如发热）和中药（如甘草）对构图的影响，本研究设定一个 k 值， k 代表 2 个病例节点具有相同的症状或中药的数量，计算公式如式（1）所示。实验表明，当节点间出现 8 个相同的症状或者中药，则认为节点间存在边，即 k 取值为 8。

$$(\text{节点 } 1, \text{节点 } 2) = \begin{cases} 1, & k \leq \text{具有相同属性} \\ 0, & k > \text{具有相同属性} \end{cases} \quad (1)$$

1.4 聚类挖掘 MFCN 模型构建

1.4.1 模型构建 基于 GCN、AE 以及自监督机制

技术构建 MFCN 模型（图 2），以深入探讨复杂高维数据中的潜在规律。

模型构建过程可分为 2 部分：（1）AE 从图的属性中提取原始特征（ Y ）；GAE 从图的结构中获取结构特征（ G ）；借助增强自编码器（enhanced auto-encoder, EAE）将结构特征和属性特征进行逐层融合得到 Z 。（2）通过表示 Y 、 Z 与 G 三重特征，得到聚类软分配 Q 、 Q' 与 Q'' 。引入 Q' 与 Q'' 作为辅助目标分布，构建三重自监督机制，通过高置信度的指导来驱动模型训练。 Q' 生成目标分布 P 来指导自监督聚类。通过对比聚类分配，使 3 个聚类分配趋于一致，从而得到聚类结果。

1.4.2 模型性能评估 使用 4 种常见的聚类评估指标，即准确率（accuracy, ACC）、归一化互信息（normalized mutual information, NMI）、调整兰德系数（average rand index, ARI）和 F1 得分（F1-score, F1）评估模型的聚类性能。

NMI 是一种用于衡量聚类算法性能的指标，取

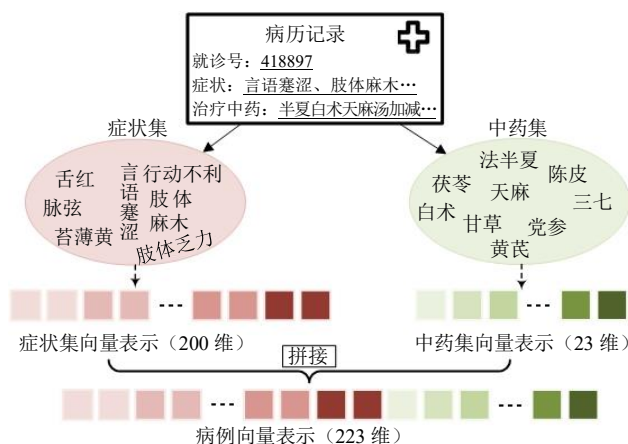


图 1 病例节点的向量表示

Fig. 1 Vector representation of case nodes

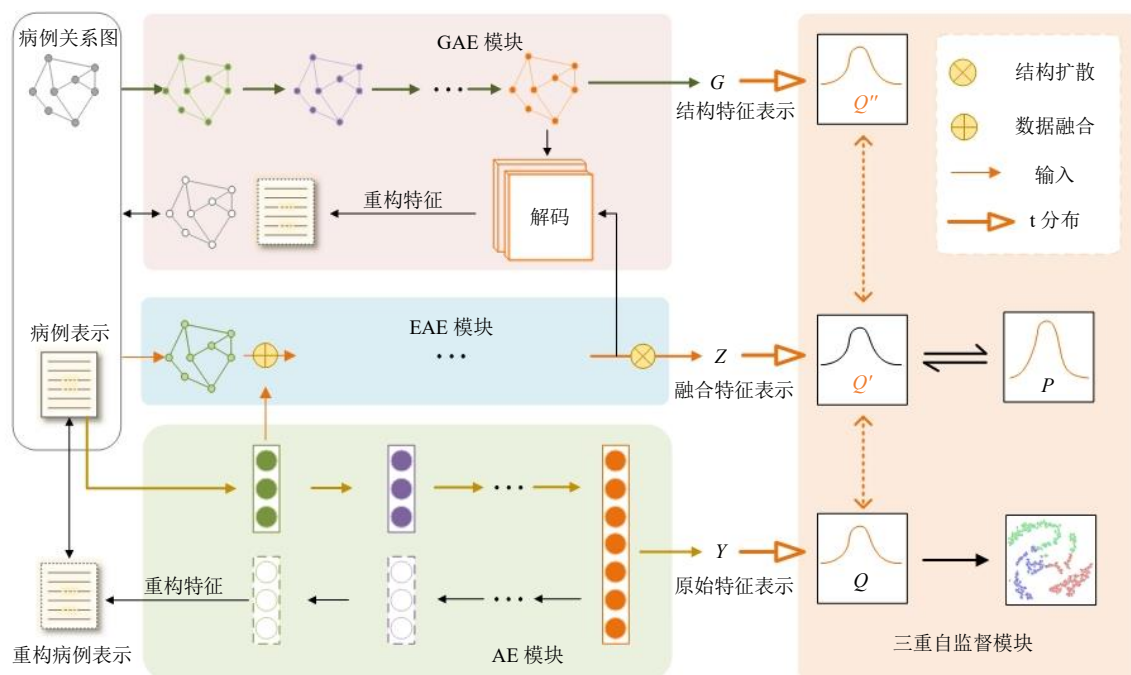


图 2 MFCN 模型结构图

Fig. 2 Structure diagram of MFCN model

值范围在 $[0, 1]$, 定义见式 (2)。其量化了聚类结果与真实标签之间的相似性, 取值为 1 表示聚类结果与真实标签完全一致。

$$NMI(Y, C) = I(Y, C) / 0.5[H(Y) + H(C)] \quad (2)$$

Y 代表真实标签, C 代表聚类标签, I 表示互信息, H 表示熵

ARI 是一种常见的聚类外部评价指标, 取值范围为 $[-1, 1]$, 值越大意味着聚类结果与真实情况越吻合。其通过计算在真实标签和聚类结果中被分配在相同或不同类簇的样本对的个数来进行聚类有效性的评价, 具体定义如式 (3) 所示。

$$ARI = (RI - E[RI]) / (\max(RI) - E[RI]) \quad (3)$$

RI 表示兰德指数, $E[RI]$ 表示 RI 的期望

1.5 脑卒中用药规律分析

随机选取脑卒中任一期型的病例作为研究对象, 进一步探讨其用药规律。首先, 使用 Python 工具库 mlxtend 中的 Apriori 算法, 对脑卒中急性期病例的中药与症状进行频次分析, 深入挖掘每个聚类团中药物使用频率、症状出现频率以及药物组合和“药物-症状”组合。随后, 应用 Cytoscape 3.8.2 软件绘制“药-症”共现网络, 以可视化分析

药物与症状之间的关系，从而揭示脑卒中急性期的用药规律。

2 结果

2.1 数据概况

DBLP 数据集包含 4 058 个节点和 3 528 条边，数据特征维度为 334，并被划分为 4 个类别。

脑卒中数据集中符合纳入标准的临床病例总计 2 572 例，分为急性期（924 例）、恢复期（837 例）和后遗症期（811 例）。经过正异名处理的治疗中药共 356 味，规范化后的患者症状为 558 种。脑卒中数据特征维度为 223，构建的节点图包含 2 572 个节点和 5 877 条边。

2.2 MFCN 模型构建及性能评估

本研究的实验数据集见“2.1”。实验基于 Python 3.7 编程语言，并使用深度学习框架 Pytorch1.13.1 搭建的环境，配置为 RTX4060 和 CUDA12.2。MFCN 的参数设置类似于经典深度聚类模型 SDCN^[11]，在训练之前，首先对自编码器进行预训练，学习率为 1×10^{-3} ，训练 50 个轮次。在 AE 和 GAE 中，每层的维数分别设计为 500、500、2 000 和 10。为了初始化聚类中心，使用预训练的自编码器得到数据表示，并通过 20 次 K 均值聚类来选择最佳结果。

为评估模型性能，将 MFCN 模型与以下 3 组基线模型进行对比。（1）传统的聚类方法（AE+K 均值^[6]）。利用自编码器提取特征，并采用 K 均值算法对聚类进行划分。（2）基于 GAE^[8]的聚类方法。使

用图卷积层作为编码器来学习节点特征表示的无监督图嵌入方法。（3）基于双网络的聚类方法（SDCN^[9]、MBN^[11]）。SDCN 设计了一个双通道网络，将图卷积网络集成到 DEC^[10]框架中，利用结构信息，在统一的框架中优化聚类分配和表示学习。MBN 改进了 SDCN，将节点特征和结构特征的异构信息同时传播，并通过对比聚类分配来相互指导，以自监督的方式进行同步优化。

本研究提出的 MFCN 模型与其他聚类模型的对比实验结果如表 3 所示，其中每列的最优结果以粗体显示。为减少随机性对结果的影响，将每个实验重复 10 次，并取平均值评估聚类结果。结果显示，MFCN 在 2 个数据集中都展现出最佳效果，均优于现有模型。在 DBLP 数据集中，MFCN 在 ACC、NMI、ARI 和 F1 指标上分别达到了 0.793 2、0.490 8、0.541 9 和 0.787 7。相较于最优的对比模型 MBN，MFCN 在各指标上分别提升了 1.35%、0.93%、1.81% 和 1.56%。在脑卒中数据集中，MFCN 在 ACC、NMI、ARI 和 F1 指标上分别达到了 0.834 8、0.531 6、0.581 7 和 0.833 9，相比 MBN 模型提升了 1.97%、4.8%、4.68% 和 2.04%。与经典 SDCN 模型相比，MFCN 在 ACC、NMI、ARI 和 F1 上则分别提升了 7.96%、6.45%、7.4% 和 20.54%。特别是 NMI、ARI 指标提升显著，表明 MFCN 在聚类效果上具有更优表现。此外，与传统的 AE+K 均值和 GAE 模型相比，MFCN 在 2 个数据集中同样显示出更好的性能。

表 3 模型性能评估结果

Table 3 Evaluation results of model performance

模型	DBLP 数据集				脑卒中数据集			
	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1
AE+K 均值	0.597 0	0.283 9	0.197 0	0.598 0	0.733 5	0.370 9	0.374 9	0.730 3
GAE	0.682 3	0.343 8	0.331 1	0.679 7	0.734 1	0.354 3	0.381 9	0.733 7
SDCN	0.701 9	0.357 1	0.384 4	0.690 9	0.755 2	0.467 1	0.507 7	0.628 5
MBN	0.779 7	0.481 5	0.523 8	0.772 1	0.815 1	0.483 6	0.534 9	0.813 5
MFCN	0.793 2	0.490 8	0.541 9	0.787 7	0.834 8	0.531 6	0.581 7	0.833 9

粗体表示最优结果。

Bold indicates optimal results.

2.3 用药规律分析

基于 MFCN 的实验结果，随机选取 1 个脑卒中的期型（急性期）为研究对象，进一步探讨脑卒中的用药规律。聚类结果显示，急性期患者共计 792 例，恢复期患者共计 880 例，后遗症期患者共计 900 例。通过 Pycharm 软件对脑卒中急性期进行关联规则分析，利用 Apriori 算法建模，深入剖析每个聚类

团中药物使用频率与症状出现频率、药物组合和“药物-症状”组合，构建“症-药”共现网络，从而揭示脑卒中急性期的用药偏好。

急性期脑卒中病例中涉及中药 286 味、症状 558 种。用药频率排名靠前的中药有甘草、茯苓、陈皮、天麻、法半夏等（表 4），症状出现频率排名靠前有脉弦、肢体乏力、头晕、舌红、行动不利等（图 3）。

表 4 急性期脑卒中高频中药分布 (用药频率排名前 10)

Table 4 Distribution of high frequency TCM in acute stroke (top 10 in medication frequency)

序号	中药	频次	频率/%	序号	中药	频次	频率/%
1	甘草	658	83.08	6	枳实	402	50.76
2	茯苓	605	76.39	7	鸡血藤	387	48.86
3	陈皮	595	75.13	8	地龙	377	47.60
4	天麻	438	55.30	9	石菖蒲	348	43.94
5	法半夏	434	54.80	10	竹茹	336	42.42

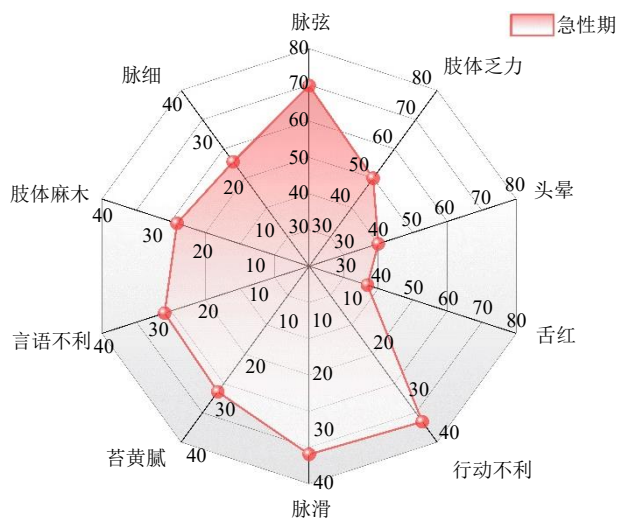


图 3 脑卒中症状频率分布

Fig. 3 Frequency distribution of stroke symptoms

通过关联规则对急性期脑卒中病例进行分析, 设置支持度 $\geq 30\%$, 置信度 $\geq 60\%$, 最大前项数为 2, 得到 550 条关联规则, 按照支持度排序得到“药-药”组合和“药-症”组合, 见表 5、6。表 5 展示了脑卒中急性期前 10 条药物组合规则。其中, 支持度最高的药组是“陈皮 $\rightarrow$ 茯苓”, 为 65.78%; 置信度最高的药组是“陈皮 $\rightarrow$ 枳实”, 为 92.52%; 增益最大的药组是“陈皮 $\rightarrow$ 枳实”, 为 1.23。表 6 列出了部分症状对应的“药物-症状”组合, 选取了每种症状的前 3 条规则。结果显示, 支持度最高的“药-症”组合为“甘草 $\rightarrow$ 脉弦”, 为 57.07%; 置信度最高的组合为“甘草 $\rightarrow$ 舌红”, 为 83.90%; 增益最大的组合是“陈皮 $\rightarrow$ 肢体乏力”中为 1.05。同时构建脑卒中急性期中药与症状的共现网络, 见图 4。

针对模型分类失效的典型病例, 本研究随机选取一个模型分类失效的病例, 并从分类正确的病例中筛选出与其最相似的病例进行对比展示, 见图 5。

3 讨论

本研究利用深度聚类技术构建了多源特征融合的 MFCN 模型, 并在 DBLP 与脑卒中数据集上进

表 5 急性期脑卒中药物组合关联规则分析

Table 5 Analysis of association rules for drug combination in acute stroke

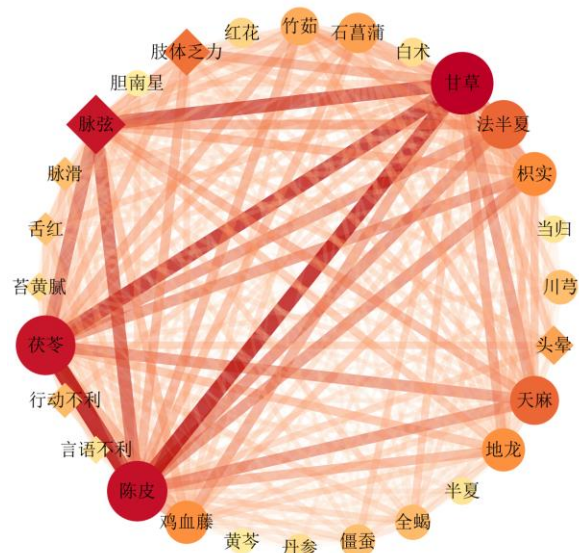
后项	前项	支持度/%	置信度/%	增益
陈皮	茯苓	65.78	87.86	1.17
甘草	陈皮	64.52	86.32	1.04
甘草	茯苓	62.75	83.81	1.01
甘草	茯苓+陈皮	56.31	85.60	1.03
陈皮	茯苓+甘草	56.31	89.74	1.19
茯苓	陈皮+甘草	56.31	87.28	1.14
陈皮	法半夏	47.98	87.96	1.17
陈皮	枳实	46.84	92.52	1.23
天麻	茯苓	46.72	62.40	1.13
天麻	陈皮	46.34	61.99	1.12

行了性能评估。以脑卒中数据集为例, 验证了模型在高维临床数据中探索用药规律的有效性, 归纳总结出脑卒中急性期的用药规律。

MFCN 模型在 2 个数据集中表现出了最佳的聚类性能, 优于其他模型。AE 模型和 GAE 模型分别基于数据的内容和结构特征进行聚类。SDCN 模型通过融合内容特征和结构特征, 因而在 2 个数据集

表 6 急性期脑卒中“药-症”组合关联规则分析
Table 6 Analysis of association rules for “drug symptom” combination in acute stroke

后项	前项	支持度/%	置信度/%	增益
甘草	脉弦	57.07	81.74	0.98
陈皮	脉弦	54.04	77.40	1.03
茯苓	脉弦	53.54	76.67	1.00
甘草	肢体乏力	41.54	83.08	1.00
陈皮	肢体乏力	39.39	78.79	1.05
茯苓	肢体乏力	36.99	73.99	0.97
甘草	头晕	33.08	82.65	0.99
茯苓	头晕	30.56	76.34	1.00
甘草	舌红	30.93	83.90	1.01



共现网络中节点间的连线越粗代表 2 个节点之间的关联度越高；圆形节点越大，则代表与之相关联的药物数量越多。
Thicker the connections between nodes in a co-occurrence network, the higher the degree of association between two nodes; larger the circular nodes, the more drugs are associated with them.

图 4 脑卒中急性期“症-药”共现网络
Fig. 4 Co-occurrence network of “symptom-drug” in acute stroke

中的聚类性能均优于 AE 模型和 GAE 模型。与 SDCN 相比，MBN 模型引入了数据表示增强机制，并通过内容与结构特征的聚类一致性约束提升了聚类效果，其 NMI 和 ARI 指标在 2 个数据集中分别提升了 12.44%、13.94% 及 1.65%、2.72%。基于此，MFCN 模型引入了多源融合策略，将内容特征、结构特征与增强特征进行了有效融合，并通过三重对比聚类分配实现了相互指导，从而进一步提升了聚类性能。具体而言，相较于 MBN、MFCN，MFCN 模型在 DBLP 数据集上的 NMI、ARI 分别提升了

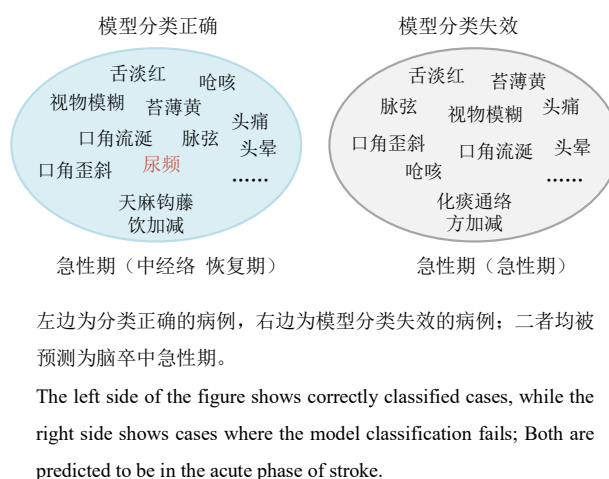


图 5 模型分类正确与失败的案例展示
Fig. 5 Case studies demonstrating correct and failed model classification

0.93%和 1.81%，在脑卒中数据集上的 NMI、ARI 分别提升了 4.8%和 4.68%。结果显示，MFCN 模型的聚类性能取得了显著优势，充分证明了其能够高效整合多源数据的信息，并具备处理复杂高维数据的出色能力，为后续研究提供了强有力的支持。

基于 MFCN 的实验结果，随机选取了脑卒中的急性期为研究对象。聚类结果显示，在急性期中，甘草、茯苓、陈皮等药物使用频率较高，这与脑卒中急性期治疗以清热解毒、化痰开窍为主的原则相符合^[19-20]。“陈皮-茯苓”组合在急性期有最高的支持度，二者合用具有燥湿化痰、醒神开窍。若出现痰湿蒙神证，可以使用以陈皮和茯苓为主的方药，如涤痰汤可改善神经功能、控制癫痫发作、减轻认知障碍、提高患者生活质量^[21-22]。“甘草-茯苓-陈皮”组合在脑卒中急性期有较高的置信度，三药合用，适用于痰湿阻络型患者^[23]。此外，构建“症-药”共现网络，挖掘出脑卒中急性期的潜在“药-症”组合。“甘草-脉弦”“甘草-舌红”及“陈皮-肢体乏力”等组合关联度较高。若患者出现脉弦、舌红和肢体乏力等症状，可用甘草缓解脉弦，同时配合清热药物治疗舌红，陈皮调理脾胃，促进气血生成，缓解肢体乏力。

为进一步完善模型的性能，本研究对模型分类失败的病例与分类正确的病例进行了对比分析（图 5）。2 个病例的症状相似度达到 90%以上，治疗方剂分别为天麻钩藤饮加减^[24]和化痰通络汤加减^[25]，但模型均预测为急性期。这一结果表明，尽管症状高度相似，但不同中药在临床应用中的具

体功效可能未完全体现在模型中。

4 展望

本研究所构建的 MFCN 模型在处理复杂的高维临床数据方面具有良好性能,成功地挖掘了脑卒中急性期的用药规律,为临床精准用药提供了科学依据。然而,本研究也存在一些局限性。(1)由于数据来源的限制,本研究仅涵盖了湖南中医药大学附属第一医院的脑卒中患者数据,可能存在地域性和人群选择的偏差。同时,用药规律分析主要基于历史数据,缺乏前瞻性的临床验证。(2)脑卒中作为一种临床异质性极高的疾病^[26],其临床表现和进程受到多种因素的影响。患者的性别、年龄及既往病史等因素都可能显著影响疾病的发生、发展及治疗效果,从而影响模型的分类结果。(3)本研究的用药规律分析主要基于病例数据中的症状与中药处方信息,未充分考虑中药的具体功效对模型分类结果的影响。中医方剂的选择不仅受到患者临床表现的影响,还与中药的功效特点、药物配伍的相互作用密切相关。

未来的研究可以通过扩大数据来源、开展多中心合作研究以及前瞻性临床研究来进一步验证和完善本研究的发现。同时,应将中药的功效机制、适应证以及患者的体质特点纳入模型考虑范围,进行更加细致的特征工程和数据预处理,提升模型的分类精准度。还可以探索将 MFCN 模型应用于其他疾病的临床数据集,如心血管疾病、糖尿病等,以评估其在不同疾病中的适用性和表现,从而为揭示疾病的作用机制提供新的思路 and 参考。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] Berahmand K, Li Y F, Xu Y. DAC-HPP: Deep attributed clustering with high-order proximity preserve [J]. *Neural Comput Appl*, 2023, 35(34): 24493-24511.
- [2] Cai J Y, Wang S P, Guo W Z. Unsupervised embedded feature learning for deep clustering with stacked sparse auto-encoder [J]. *Expert Syst Appl*, 2021, 186: 115729.
- [3] Cai J Y, Wang S P, Xu C Y, *et al.* Unsupervised deep clustering via contractive feature representation and focal loss [J]. *Pattern Recognit*, 2022, 123: 108386.
- [4] 孔星, 刘二小, 陈烽聚, 等. 深度聚类算法在 SuperDARN 雷达目标回波分类中的应用 [J]. *空间科学学报*, 2024, 44(5): 806-817.
- [5] 侯海薇, 丁世飞, 徐晓. 基于无监督表征学习的深度聚类研究进展 [J]. *模式识别与人工智能*, 2022, 35(11):

999-1014.

- [6] Yang B, Fu X, Sidiropoulos N D, *et al.* Towards k-means-friendly spaces: Simultaneous deep learning and clustering [A] // *Proceedings of International Conference on Machine Learning* [C]. Sydney: International Machine Learning Society (IMLS), 2017: 3861-3870.
- [7] Hartigan J A, Wong M A. Algorithm AS 136: A K-means clustering algorithm [J]. *Appl Stat*, 1979, 28(1): 100.
- [8] Kipf T N, Welling M. Variational graph auto-encoders [A] // *Proceedings of Advances in Neural Information Processing Systems Workshop on Bayesian Deep Learning* [C]. Barcelona: Curran Associates, 2016: 1-3.
- [9] Bo D Y, Wang X, Shi C, *et al.* Structural deep clustering network [A] // *Proceedings of The Web Conference 2020* [C]. Taipei: ACM, 2020: 1400-1410.
- [10] Xie J Y, Girshick R, Farhadi A, *et al.* Unsupervised deep embedding for clustering analysis [A] // *Proceedings of the 33rd International Conference on International Conference on Machine Learning* [C]. [2016-06-19]. New York: ACM, 2016: 478-487.
- [11] Yan X Q, Yu X Y, Hu S Z, *et al.* Mutual Boost Network for attributed graph clustering [J]. *Expert Syst Appl*, 2023, 229: 120479.
- [12] 中国药典 [S]. 一部. 2020.
- [13] 国家中医药管理局《中华本草》编委会. 中华本草 (第十卷) [M]. 上海: 上海科学技术出版社, 1999
- [14] 赵国平, 戴慎, 陈仁寿主编. 南京中医药大学编著. 中药大辞典 [M]. 第 2 版. 上海: 上海科学技术出版社, 2006.
- [15] 唐德才, 吴庆光. 中药学 [M]. 第 4 版. 北京: 人民卫生出版社, 2021.
- [16] 黎敬波, 马力主编. 中医临床常见症状术语规范 [M]. 北京: 中国医药科技出版社, 2005.
- [17] Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality [J]. *Adv Neural Inform Proce Syst*, 2013, 26: 3111-3119.
- [18] 邓乐, 丁长松, 黄辛迪, 等. 基于多层前馈神经网络的中药药性量化研究 [J]. *中草药*, 2020, 51(16): 4277-4283.
- [19] 高长玉, 吴成翰, 赵建国, 等. 中国脑梗死中西医结合诊治指南 (2017) [J]. *中国中西医结合杂志*, 2018, 38(2): 136-144.
- [20] 张锦鹏. 益气活血化瘀开窍法辅助治疗对急性缺血性脑卒中患者疗效、神经功能的影响研究 [J]. *中医临床研究*, 2022, 14(33): 65-67.

- [21] 周宇航, 曲东奇, 高松仙, 等. 涤痰汤加减治疗癫痫的 Meta 分析 [J]. 神经病学与神经康复学杂志, 2024, 20(2): 33-46.
- [22] 章薇, 娄必丹, 李金香, 等. 中医康复临床实践指南·缺血性脑卒中 (脑梗死) [J]. 康复学报, 2021, 31(6): 437-447.
- [23] 蔡绍杰, 胡丙成, 褚慧玲, 等. 胡丙成教授临床治疗中风病经验浅谈 [J]. 黑龙江中医药, 2021, 50(3): 2-3.
- [24] 马勇勇. 中药天麻钩藤饮加减配合针灸治疗颈性眩晕的临床研讨 [J]. 婚育与健康, 2024(14): 76-78.
- [25] 杨洋, 陈甦, 许成, 等. 化痰通络汤治疗痰瘀阻络型卒中后认知功能障碍的临床效果 [J]. 中外医学研究, 2023, 21(34): 5-10.
- [26] 陈晓荣, 颜流霞, 厚磊, 等. 2015—2019 年中国居民脑卒中发病与死亡特征分析 [J]. 中国循环杂志, 2024, 39(5): 470-476.

[责任编辑 潘明佳]