基于多源数据融合的缺血性脑卒中用药风险预警研究

王文卓 秦秋莉

（北京交通大学 北京 100091）

〔摘要〕 **目的/意义** 利用基于多源数据融合的机器学习算法，预测缺血性脑卒中患者的临床药物治疗风险。**方法/过程** 基于国际脑卒中试验数据集，融合患者人口统计学、生命体征检查及临床药物治疗数据，利用随机森林、逻辑回归和梯度提升决策树算法预测用药风险。**结果/结论** 3 种算法在预测性能方面都表现较好，其中梯度提升决策树的召回率达到 91.6%，曲线下面积为 0.832，效果最佳。多源数据融合的机器学习算法在缺血性脑卒中用药风险预警中具有良好适用性。

〔关键词〕 多源数据融合；缺血性脑卒中；风险预测；智慧医疗

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.10.009

Study on Predicting the Risk of Ischemic Stroke Medication Treatment Based on Multi－source Data Fusion

WANG Wenzhuo, QIN Qiuli

Beijing Jiaotong University, Beijing 100091, China

〔**Abstract**〕 **Purpose/Significance** To predict the risk of medication for ischemic stroke patients using a machine learning algorithm based on multi－source data fusion. **Method/Process** The study is based on the international stroke trial datasets. By fusing features of patient demographics, vital sign examination and medication data, it predicts medication risks using random forest, logistic regression and gradient boosting decision tree (GBDT) algorithms. **Result/Conclusion** The results show that three algorithms performe well, with the best recall of 91.6% and area under the curve is 0.832 for GBDT algorithm. The machine learning algorithms with multi－source data fusion has good applicability in ischemic stroke medication risk prediction.

〔**Keywords**〕 multi－source data fusion; ischemic stroke; risk prediction; smart medical care

1 引言

随着公众的医疗需求不断增强，医疗质量逐渐成为焦点。为了更好地保障医疗质量，必须控制医疗风险。在众多医疗风险中，药品相关的安全性问题在临床及诊疗中最突出。国家药品监督管理局数据显示，我国自 1999 年累计接收药品不良反应事件报告高达 1 883 万份，且数量呈不断上升趋势，其中严重药品不良反应事件占比可达同期总数的 11.0%[1]。如何帮助医务工作者及时识别用药风险正逐渐成为临床治疗中的一大挑战。

脑卒中作为一种常见的急性脑血管疾病，近 10 年在中国的患病率增长惊人，是导致成人死亡和残疾的主要原因。数据显示，尽管全球脑卒中死亡率

〔修回日期〕 2023－05－04

〔作者简介〕 王文卓，硕士研究生；通信作者：秦秋莉，博士，副教授，硕士生导师。

下降，但极高的发病率和致残率带来的疾病负担依然居高不下[2]。随着人口老龄化加剧，脑卒中的发病率和死亡率可能会急剧上升，特别是在低收入和中等收入国家。由于脑卒中没有特效药，医生往往会在使用抗血小板药物的基础上，联合多种药品治疗，增加了不良反应发生率。相关研究结果显示，如果联合使用氯吡格雷与质子泵抑制剂，患者急性心肌梗死发生风险会增加 27%[3]。在针对老年脑卒中患者的多重用药调查中发现，150 例患者中发生过药物不良反应的占比达到 70.67%（106/150）[4]，脑卒中患者迫切需要有效的用药风险预警策略。

近年来，机器学习技术飞速发展，广泛应用于各个领域。已有许多研究集中于机器学习在脑卒中预防、预后和康复方面的应用。随着医疗行业信息技术的不断发展，数据融合技术逐渐被引入并应用于医疗领域。目前，各医疗机构产生和收集的医疗数据碎片化严重，且没有得到充分有效的利用。作为处理多源数据的重要手段，数据融合被定义为一种可对来自多个不同来源的数据自动检测、关联和组合，实现更加精确的信息估计技术[5]。脑卒中临床用药风险相关的研究往往聚焦于某一种药物的不良反应[6-7]，但在混合用药和结合患者个体情况方面仍有欠缺，未能做到最大限度利用相关医疗数据。因此，利用多源数据融合的机器学习技术解决缺血性脑卒中临床用药风险问题仍存在一定局限性，患者的个人数据和临床用药数据尚未较好融合，临床治疗中的医疗事故无法得到有效控制。

鉴于缺血性脑卒中治疗的复杂性和高风险，有必要对其开展风险预测研究以保障患者生命安全。因此，本文将采用多源数据融合的机器学习建模方法，结合患者个人数据、医生用药数据、药物不良反应数据，建立缺血性脑卒中临床用药风险预测模型。该模型可以根据预测结果及时发出预警，为防控脑卒中用药风险提供技术支持，对建立用药风险预警机制具有一定参考意义。

2 研究现状

2.1 多源数据融合的机器学习研究现状

大数据时代数据呈现海量、异质性等特点，多来源会导致收集的数据异构，为数据融合带来机遇和挑战[8]。数据融合技术在医疗领域主要应用于远程医疗、智慧养老以及病情诊断等方面。杨杰等[9]基于数据融合和数据挖掘技术，提出一种能够自动监测分析数据并识别异常情况的远程医疗监护系统，可以实现网络报警、远程实时诊断。

机器学习和深度学习技术已经在医疗领域得到广泛应用，为临床治疗提供辅助。Sung S M 等[10]利用各种机器学习算法预测急性轻微脑卒中患者的早期神经功能恶化。Liu J 等[11]利用机器学习模型评估导致脑卒中的显著危险因素，结果显示患者的人口统计学和生活方式信息与脑卒中的发病率和治疗效果密切相关。

2.2 临床用药风险管理

在控制临床用药风险方面，中国学者针对用药的相关问题（如药物滥用错用、药物间相互作用、药物不良反应以及与患者自身病理状况不匹配等），开发用药风险预警系统。这些系统普遍体现出针对性强、普适性差的特点，有待进一步完善。董作军等[12]提出，当下风险预警工作主要还存在风险相关数据尚未标准化、横向信息流通不畅、风险评估指标体系不够健全等问题。冯红云等[13]则指出目前针对药品不良反应的预警系统须要改善系统数据的标准性、预警规则的合理性，否则预警信息的准确度将很难提升。综上所述，用药风险预警工作仍待进一步探索和努力，以形成完善的预警机制，多源数据融合使用将在很大程度上改进传统的药物预警方式。

3 研究方法

3.1 多源异构数据融合方法

融合使用 IST－1、IST－3 等数据集中 3 035 例缺血性脑卒中患者的随访数据和药物不良反应数据，包括人口统计学信息和药物治疗记录等。数据中的患者分别来自 12 个国家的 156 家医院，年龄 40～95 岁。数据内包含患者的药物治疗情况，如阿司匹林服用、抗血小板治疗等，同时记录患者服药

时间，如：入院后1 日、入院后2 ~7 日。数据融合方法，见图1。

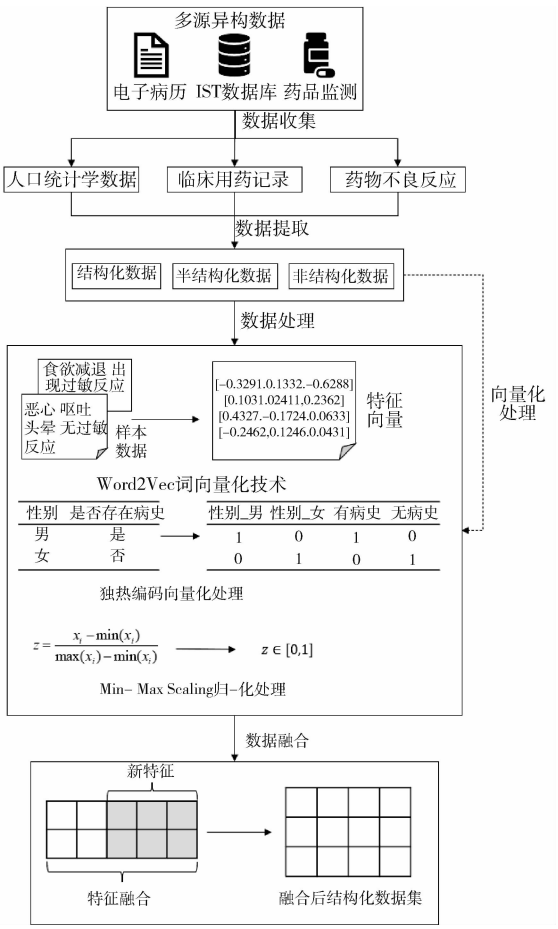

图1 数据融合流程

来自不同数据源的数据存在异构问题，含有非结构化的文本数据和半结构化数据。未融合的异构数据无法直接作为机器学习算法的数据输入，有效信息往往会被忽略，无法实现数据价值最大化利用。本文通过数据收集、数据提取、异构数据向量化处理等步骤将多源异构数据进行数据融合，最终输出一个包含全部有效特性的融合后数据集，用于风险预测。在异构数据向量化处理环节，主要使用 Word2Vec、one - hot 独热编码和数据归一化技术。处理后，非结构化数据将被转化为结构化向量，实现异构数据特征提取，融合形成全新的结构化数据集。

3.1.1 Word2Vec 采用 Word2Vec 词向量化技术实现语义空间信息到向量空间的映射。编码数据中的部分语料实现文本数据到向量的转化，将非结构化数据转化成结构化数据。

3.1.2 独热编码 由于数据集中存在部分文本标签数据，如患者性别、是否存在病史等。这类数据需要将文本标签转化为数值。为了规避数值大小对该特征值的影响，采用独热编码，借助 Python sklearn 中的 replace 函数替换标签。

3.1.3 数据归一化处理 使用线性函数归一化方法，使数据映射到［0，1］，实现对原始数据的等比缩放，消除数据量纲影响：

$$z = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (1)$$

选取患者年龄、性别、体重、血糖、脑卒中类型等特征。患者入院前的部分病史也对临床治疗有一定影响，因此选取患者入院前是否患有高血压、糖尿病、短暂性脑缺血发作（transient ischemic attack，TIA），共3 个病史特征。在药物治疗方面，使用18 个临床治疗用药特征进行风险分析，分别为阿司匹林、降压药、降糖药、抗血小板药物、溶栓药物、抗生素共6 类药物，按其用药时间分为入院前用药、入院24 小时内用药和入院2 ~7 日用药共3 个时间段。部分特征，见表1。

表1 风险预警特征（部分）

特征	平均数	数据表示
性别		男性（1，0）；女性（0，1）
年龄（岁）	77.31	数值型
血糖（mmol/L）	7.26	数值型
脑卒中类型		出血性：1；缺血性：2
既往是否有脑卒中或短暂性脑缺血发作病史	0.85	Y：1；N：0
入院前是否有高血压	0.40	Y：1；N：0
在24 小时内是否服用阿司匹林	0.62	Y：1；N：0
在24 小时内是否进行降压治疗	0.67	Y：1；N：0
入院2 ~7 日内是否服用阿司匹林	0.27	Y：1；N：0

3.2 模型构建流程

风险预警模型的整体架构将采用数据融合模型

的经典 3 层架构[14]，包括数据层、特征层和应用层。模型构建流程，见图 2。在数据层构建中，将采集的多源数据输入 MySQL 数据库，并借助 Python 3.10 对数据预处理，处理缺失值、重复值、错误值等问题，并融合异构数据。数据层经过预处理的数据将进行特征属性的确定和提取。根据《中国急性缺血性脑卒中诊治指南》[15]和相关量表，最终确定并提取相关特征 25 个，包含患者人口学特征、病史特征和临床药物诊疗特征。同时对特征标准化，将多源数据转化成可供预测的特征层数据。利用特征层中提取的特征作为应用层输入，借助机器学习的随机森林、逻辑回归和梯度提升决策树算法，对用药处方进行风险预测，并输出最终结果。结合预测结果，模型依据预警规则进行判断，若判断结果为高风险，则及时发出预警。

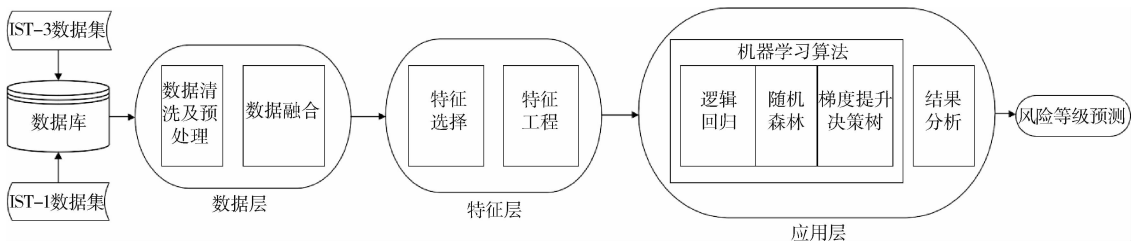

图 2 模型构建流程

3.3 模型风险预警算法

在已有脑卒中机器学习预测研究[10-11]的基础上，根据多特征分类预测的问题性质（二分类预测）选择机器学习算法进行风险预警。逻辑回归、随机森林和梯度提升决策树算法在二分类问题中表现很好，且与其他机器学习算法相比，对本文的研究问题适用性更好，因此选择这 3 种机器学习算法。随机森林可以很好地处理高维数据，具有良好性能[16]。逻辑回归是一种传统的统计方法，尤其适用于二分类因变量建模。梯度提升决策树是一种基于树集合的可加模型，经过多轮迭代，通过减少训练过程中产生的剩余误差分类数据：

$$c_{mj} = \arg\min_{c} \sum L(y_i, f_{m-1}(x_i) + c) \quad (2)$$

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I(x \in R_{\mathrm{mj}}) \quad (3)$$

利用单一算法实验可能存在一定的偶然性和随机性，因此采用 3 种机器学习算法进行实验，以证明机器学习方法解决该问题的可行性。通过对比找到最适用于脑卒中用药风险预警的算法。算法以融合后的结构化数据集作为输入，数据集被随机分为训练集、验证集和测试集（8∶1∶1）。模型首先利用训练集训练算法分类器，然后使用验证集调节模型中的超参数，最后利用训练好的模型对测试集数据输出预测结果，为评估模型性能提供参考依据。

3.4 模型评估指标

使用机器学习算法的准确率、召回率、*F*1 分数和曲线下面积（area under the curve，AUC）评价药物治疗风险预测模型的性能。准确率是在预测为正样本的实例中预测正确的频率值。召回率是在标签为正样本的实例中预测正确的频率。*F*1 分数是准确率和召回率的调和均值，计算方式如下。其可以平衡准确率和召回率的结果，范围为 0 ~ 1，值越大表示模型性能越好。

$$F1 - \mathrm{score} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \quad (4)$$

4 研究结果与讨论

以患者诊疗用药后 7 日内是否死亡作为结局变量，判定患者临床诊疗用药的疗效是否与个人身体状况相适应，并预测实际风险水平的指标。若某次临床用药导致该患者 7 日内死亡的结局，则此次用药存在很大风险，应实时向医生发出预警。风险等级判别及预警规则，见表 2。

表 2 风险等级判别及预警规则

特征	标签值	风险等级	预测规则
入院治疗后 7 日内死亡结局	1	高风险	发出预警
入院治疗后 7 日内无死亡结局	0	较低风险	不预警

4.1 不同模型预测结果

针对 3 035 例缺血性脑卒中患者记录，通过融合多源数据集，进行特征工程，开展风险预测，评估 3 种机器学习算法的效果，见表 3。3 种机器学习算法都有很好的性能，能够在高风险的结果中找到正确的样本；梯度提升决策树算法具有最高的准确率。

表 3 模型预测结果

算法	准确率	召回率	*F*1	AUC
随机森林	0. 894	0. 914	0. 896	0. 817
逻辑回归	0. 872	0. 908	0. 877	0. 801
梯度提升决策树	0. 899	0. 916	0. 902	0. 832

受试者工作特征曲线，见图 3。3 种算法对缺血性脑卒中患者用药治疗风险预测具有一定的适用性。相同条件下，梯度提升决策树算法的受试者操作特征曲线灵敏度最高，表明其具有更好的风险预测性能。逻辑回归与随机森林相比，结果无显著差异，两者性能在一定程度上相似。综合各项指标分析，梯度提升决策树算法在预测缺血性脑卒中药物治疗风险方面表现最好，实验证实基于数据融合的缺血性脑卒中用药风险预警模型具备可行性，能够为临床用药风险管控提供较好的技术支持，为患者生命安全带来更大保障。

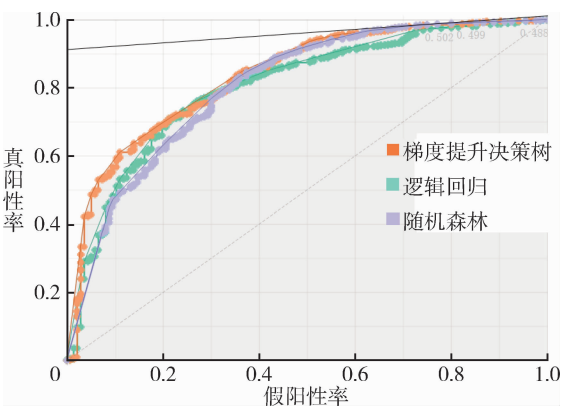

图 3 受试者工作特征曲线

4.2 特征重要性分析

特征重要性比例排序，见图 4。在所有特征中，年龄对疾病用药风险的重要性排序最高，说明年龄对缺血性脑卒中患者的治疗方案至关重要。从药物诊疗特征看，首日使用阿司匹林和降压药，2 ~ 7 日内使用抗生素，以及首日使用降糖药对预测结果都有较为明显的贡献度。此外，降压、溶栓、抗血小板药物的使用和治疗方式在所有影响因素中的重要程度位于前 50% 分位的位置，说明对结果有一定程度的影响，治疗过程中也需要一定考量。

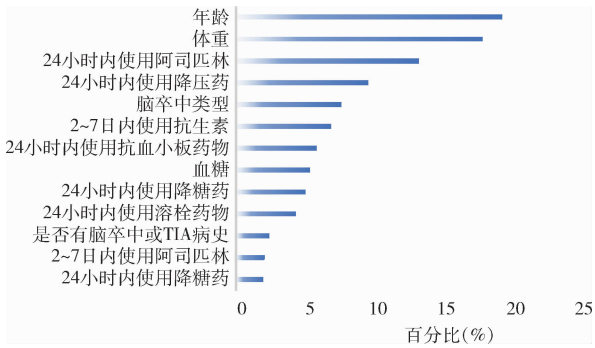

图 4 特征重要性排序（部分）

该风险预警模型使用的特征可在临床治疗过程中即时获取，通过预警患者临床用药处方风险，快速判断治疗方案的风险等级，及时调整高风险等级患者措施，具有耗时短、准确性高的特点，为医护人员合理提供合理诊断依据。经过实验验证，该模型的预测准确率达到 89. 9%。综上，该模型对于早期判断脑卒中患者用药是否存在高危风险有较好的预警能力，可以帮助医护人员较早识别高危患者，并为医生及时调整临床用药方案提供理论依据。

5 结语

本文基于多源数据集，以及数据融合模型架构，使用缺血性脑卒中患者 25 项指标预测临床药物治疗风险。验证了数据融合的机器学习算法在药物治疗风险预测问题上的适用性，结果表明 3 种算法在一定程度上均具有可行性。本文提出的模型，充分将药物治疗时患者个人的人口统计学信息和相

关数据纳入实验预测。结果显示，通过结合患者的性别、年龄、体重、血糖和相关病史等数据，可以提高预测的准确性和有效性。阿司匹林、降压药、抗生素、降糖药的药物治疗以及抗凝、溶栓的治疗方式是影响缺血性脑卒中患者病情的关键。患者药物治疗的时机也非常重要，基于不同时间段的综合用药记录和患者个体情况，利用数据融合的机器学习算法进行缺血性脑卒中用药治疗风险预测，可以有效降低医疗风险，提高临床治疗效果。

参考文献

1 国家药品监督管理局. 国家药品不良反应监测年度报告(2021 年)[EB/OL]. [2022 - 03 - 30]. https://www.nmpa.gov.cn/xxgk/yjjsh/ypblfytb/20220329161925106.html.

2 NI X J. Evidence - based practice guideline on integrative medicine for stroke 2019 [J]. Journal of evidence - based medicine, 2020, 13 (2): 137 - 152.

3 招远祺, 朱根福, 谢雁鸣, 等. “中风—慢病管理系统”监控中风患者药物安全性探索 [J]. 广州中医药大学学报, 2013, 30 (6): 909 - 911.

4 刘彤云, 胡松, 贾黎, 等. 老年脑卒中病人的多重用药调查 [J]. 青岛大学学报 (医学版), 2021, 57 (5): 708 - 711.

5 BLEIHOLDER J, NAUMANN F. Data fusion [J]. ACM computing surveys (CSUR), 2009, 41 (1): 1 - 41.

6 王德光. 丁苯酞软胶囊治疗老年缺血性脑卒中的安全性及有效性分析 [J]. 中国现代药物应用, 2022, 16 (21): 65 - 67.

7 陈肖. 注射用丹参多酚酸治疗缺血性脑卒中的疗效与安全性 [J]. 齐齐哈尔医学院学报, 2021, 42 (4): 303 - 305.

8 ZHANG L, XIE Y. Multi - source heterogeneous data fusion [C]. Chengdu: International Conference on Artificial Intelligence and Big Data, 2018.

9 杨杰, 沈利, 胡英. 结合数据融合和数据挖掘的医疗监护报警 [J]. 计算机仿真, 2000, 17 (6): 39 - 41.

10 SUNG S M, KANG J Y, CHO H J, et al. Prediction of early neurological de - terioration in acute minor ischemic stroke by machine learning algorithms [J]. Clinical neurology and neurosurgery, 2020, 195 (1): 105892.

11 LIU J, SUN Y, MA J, et al. Analysis of main risk fac - tors causing stroke in Shanxi province based on machine learning models [EB/OL]. [2023 - 04 - 02]. https://doi.org/10.1016/j.imu.2021.100712.

12 董作军, 邱琼, 钟元华, 等. 基于故障模式、影响与危害性分析的食品药品监管风险预警系统构建 [J]. 中国新药与临床杂志, 2016, 35 (10): 713 - 717.

13 冯红云, 刘佳, 董铎, 等. 影响聚集性药品不良事件预警系统运行的风险因素分析及思考 [J]. 中国药物警戒, 2016, 13 (4): 219 - 222.

14 MENG T, JING X Y, YAN Z, et al. A survey on machine learning for data fusion [J]. Information fusion, 2020, 57 (5): 115 - 119.

15 中华医学会神经病学分会, 中华医学会神经病学分会脑血管病学组. 中国急性缺血性脑卒中诊治指南 2018 [J]. 中华神经科杂志, 2018, 51 (9): 666 - 682.

16 LI H, LIN J, LEI X, et al. Compressive strength prediction of basalt fiber reinforced concrete via random forest algorithm [J]. Materials today communications, 2022, 30 (3): 103117.