网络出版时间:2023-06-27 15:35:51 网络出版地址:https://kns.cnki.net/kcms2/detail/34.1065.R.20230626.1002.020.html

卷积神经网络在急性髓系白血病流式细胞术自动诊断中的应用

雷 伟[1],李智伟[2],芮东升[1],张 眉[1],郭玉娟[1],摆文丽[1],王 奎[1]

摘要 目的 建立卷积神经网络(CNN)模型对流式细胞术(FCM)数据进行自动分析,实现急性髓系白血病(AML)的初步诊断,探究将CNN模型应用于FCM数据分析中的可行性。**方法** 以FlowRepository数据库和新疆维吾尔自治区人民医院临床检测中心获得的骨髓FCM数据进行CNN应用的探索性研究,数据均已被临床确诊是否患有AML。其中,公开数据按照6∶2∶2划分训练集、验证集和测试集,本地数据作为外部测试集;为了使FCM数据能够适应CNN模型,提出一种基于图像矩阵原理的FCM数据结构,对原始数据进行预处理后,提取与AML初步诊断相关的变量,包括侧向散射光和CD45、CD13、CD33、HLA-DR、CD117、CD34的各抗原表达水平,将各变量写入矩阵;对训练集使用细胞抽样和数据增强方法增大样本量,在Python中使用keras软件包构建LeNet-5 CNN模型,将训练集和验证集分别用于模型的训练和调参,评价模型在测试集上的性能。**结果** CNN在两测试集上识别AML的准确率分别为0.931、0.851,灵敏度为0.667、0.636,特异度为0.968、0.940,受试者工作特征曲线下面积(AUC)为0.940和0.917。**结论** 基于提出的FCM数据结构,CNN模型能够实现对AML的初步诊断,表明CNN在FCM数据分析中具有一定的应用价值。

关键词 流式细胞术;急性髓系白血病;卷积神经网络;自动诊断

中图分类号 R 733.71

文献标志码 A **文章编号** 1000-1492(2023)07-1189-05

doi:10.19405/j.cnki.issn1000-1492.2023.07.021

急性髓系白血病(acute myeloid leukemia,AML)是成人最常见的白血病类型之一,在各类急性白血病中,AML患者生存率最低[1]。流式细胞术(flow cytometry,FCM)被广泛的应用于AML的诊断、免疫分型和微小残留病监测等方面[2]。在FCM应用过程中会产生高维数据,传统数据分析方法通常由分析者根据经验在可视化软件中进行设门操作,得到细胞亚群信息后结合相应标准诊断AML[3]。这种依靠人工设门的数据分析方法存在主观性强、效率低、分析维度局限等问题,已经成为FCM应用中的瓶颈[4],因此,提出FCM数据的自动分析方法辅助临床诊断AML具有实用价值。卷积神经网络(convolutional neural network,CNN)常用于医学图像的分类问题,可以代替医师进行重复的视觉工作,成为了某些疾病自动化诊断的最好方法[5]。该研究旨在提出一种基于图像矩阵原理的FCM数据结构,并建立CNN模型,实现对AML的自动识别,证明CNN应用在FCM数据分析中的可行性。

1 材料与方法

1.1 资料来源 本研究使用两组数据:数据1来源于FlowRepository数据库[6],编号为FR-FCM-ZZYA,包括359例骨髓FCM数据,其中正常人316例,AML患者43例。数据2来源于2016-2017年新疆维吾尔自治区人民医院临床检测中心存档数据,纳入的病例组为临床上按照MIC分型标准[7]确诊的AML初诊患者,对照组为非白血病贫血患者或健康志愿者,排除白血病以外所有患有与免疫系统相关疾病或其他重大疾病的参与者。数据为骨髓FCM数据,由专家分析后给予诊断结果,其中正常人50例,AML患者22例。两组数据中,每个数据均为8管,数据格式为FCS,并且在专家分析阶段完成了粘黏细胞、死细胞等非有效数据的清除。本项目已获得石河子大学医学院伦理委员会批准(批准号:2018-015-01),参与者已签署知情同意书。

1.2 数据读取 采用R中Bioconducter-flowcore工具包读取FCS文件和补偿矩阵,对各抗体荧光强度进行补偿。分别对侧向散射光(side scatter,SSC)和各抗原表达水平进行对数和双指数转化提高数据对称性[8]。提取SSC及各管中与AML初诊相关的抗原表达水平[9],包括CD45、CD34、CD117、HLA-DR、CD13和CD33,将各变量以CSV格式存储。

1.3 归一化 为了更好实现归一化,定义数据中各变量处于极端为异常值,设定异常值细胞占各管总细胞的0.1%,将异常值去除。对纳入的变量进行

2023-05-20 接收

基金项目:国家自然科学基金(编号:81860374)

作者单位:[1] 石河子大学医学院预防医学系,石河子 832002

[2] 新疆维吾尔自治区人民医院临床检测中心,乌鲁木齐 830001

作者简介:雷 伟,男,硕士研究生;

王 奎,男,副教授,硕士生导师,责任作者,E-mail:kwang_shzu@163.com

离差标准化,处理后变量会被映射到[0,1]之间(公式1)。

$$x^* = \frac{x - v_{min}}{v_{max} - v_{min}} \quad (公式1)$$

式中,v_{min}和v_{max}分别表示一组变量的最小值和最大值;x和x^*分别表示处理前后的变量值。

1.4 数据重构 在AML流式诊断中,对SSC和CD45的设门通常作为一种初始策略,为区分主要的造血细胞提供一个起点,再结合其他标志物进一步分析。基于以上原则,本研究提出一种数据结构(图1A),以图像数据结构作为参照,定义SSC和CD45作为图像矩阵像素的定位点,将数据中SSC和CD45的值分别乘以矩阵宽度(w)和高度(h)后取整,以SSC和CD45为坐标可将二维空间划分为w×h个区域,对其余5个抗原表达分别构建图像矩阵的颜色通道。本研究中矩阵的尺寸为32×32×5(对应深度学习经典图像数据集Cifar-10数据尺寸:32×32×3)。该数据结构的特点是在不改变原模型架构的情况下能够直接作为图像CNN的输入,因而可以兼容目前多数CNN模型。

1.5 数据集建立 将数据1中AML组和正常组按照6∶2∶2划分数据集,包括训练集215例(AML 25例,正常人190例),验证集72例(AML 9例,正常人63例),测试集72例(AML 9例,正常人63例),数据划分按照FlowRepository数据库中给定的顺序进行。数据2中的72例数据全部作为测试集(AML 22例,正常人50例)。在训练集上,从同一个人的各管数据中随机抽取细胞,各抗原表达水平按照SSC/CD45为定位写入矩阵的对应通道内(图1B),当矩阵中对应位置已被写入,则此次细胞抽取无效,进行下一次抽取,直至矩阵中的全部位置被写入,抽取将停止,该矩阵作为一个训练样本。重复上述过程,可以持续产生训练样本,以满足CNN对大样本的需求。为尽可能保证训练集中正负样本均衡,对25例AML和190例正常人数据分别进行190轮和25轮抽取,最终得到25×190+190×25=9 500个训练样本。为增强CNN模型的泛化能力,防止过拟合,对所有训练样本进行数据增强,随机进行各方向上20%的平移和缩放。在验证集和测试集上,从同一个人各管数据依次选择全部细胞,将抗原表达信息按照SSC/CD45为定位写入矩阵的对应通道内,对矩阵中相同位置的抗原表达水平取均值,该矩阵作为一个验证或测试样本。

1.6 模型的构建 本研究中CNN模型选择LeNet-5架构,该架构是现代CNN的起源架构之一,具有代表性(图1C)。为保证客观性,除调整模型的输入和输出尺寸外,不改变模型其他参数。训练时,用训练集和验证集分别进行模型的训练和调参,损失函数选择交叉熵函数(公式2)。

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i)] \quad (公式2)$$

式中,N代表样本总数;y_i表示样本i的标签,正类为1,负类为0;p_i表示样本i预测为正类的概率。

采用随机梯度下降法(stochastic gradient descent,SGD)作为优化器在训练中更新参数,训练完毕后模型以h5格式存储。测试时,将测试集输入模型,信号在模型中向前传播后通过sigmoid函数(公式3)计算得到二分类概率值,模型定义0.5为截断值以区分AML和正常人。

$$S(x) = \frac{1}{1 + e^{-x}} \quad (公式3)$$

式中,x为分类器前网络的输出,$S(x)$取值范围在[0,1]之间,当$x=0$时,$S(x)=0.5$。

对模型在测试集上的性能进行评价,评价指标包括准确率、灵敏度(查全率)、特异度、查准率和F1分数(公式4)。

$$F1 = \frac{2PR}{P + R} \quad (公式4)$$

式中,P表示查准率,R表示查全率,$F1$分数取值范围在[0,1]之间,是用来综合评价二分类模型精确度的指标。

1.7 统计学处理 采用R 4.0.2软件中Bioconducter-flowcore工具包实现FCS数据的信息提取。采用Python 3.7.1软件中sklearn、numpy工具包实现数据集的建立,模型框架的搭建及训练和测试的全过程均使用keras工具包实现,文中与模型相关但未说明的参数均为keras中的默认参数。采用matplotlib工具包绘制受试者工作曲线(receiver operator characteristic curve,ROC),计算曲线下面积(area under curve,AUC)评价模型的优劣。

2 结果

2.1 FCM数据各抗原表达水平单因素分析 对FCM数据各抗原表达水平进行单因素分析,见表1。在数据1中,正常人和AML患者在CD33、HLA-DR、CD117、CD34抗原表达水平差异有统计学意义;

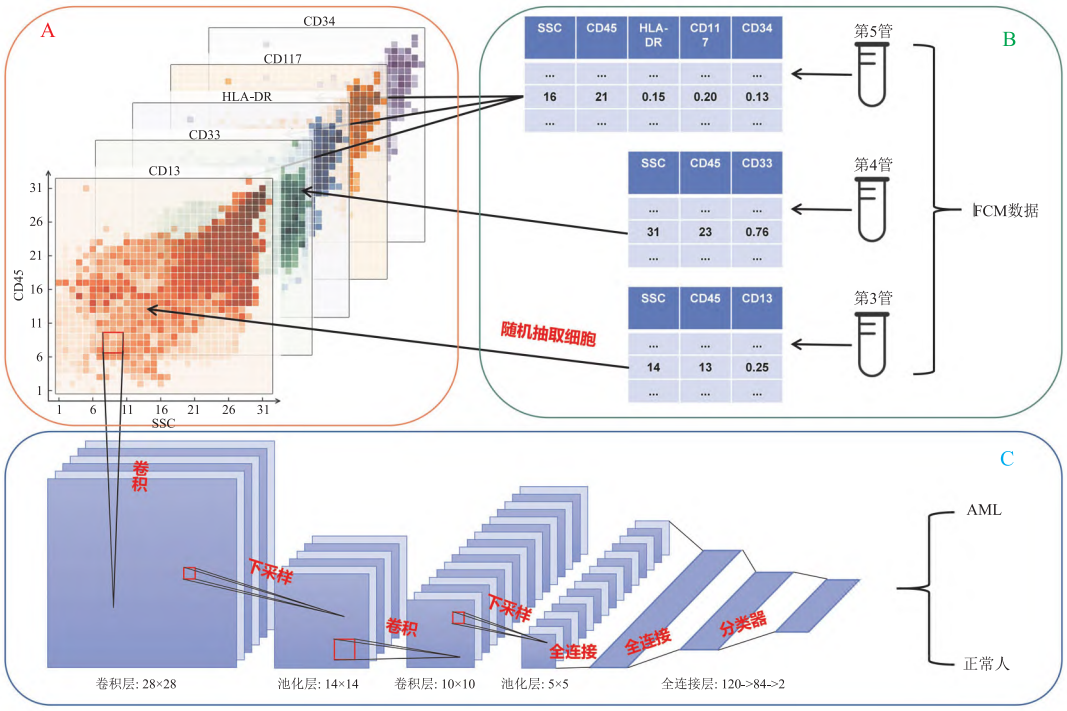

图 1 CNN 自动分析 FCM 数据流程图

A:数据结构示意图;B:FCM 数据写入过程;C:LeNet-5 架构

表 1 各抗原表达水平与 AML 患者的关系($\bar{x} \pm s$)

变量	正常人	AML 患者	t 值	P 值
数据 1				
CD45	0.591 ± 0.058	0.584 ± 0.097	0.675	0.500
CD13	0.435 ± 0.075	0.421 ± 0.115	1.067	0.287
CD33	0.310 ± 0.082	0.357 ± 0.167	−3.012	0.003
HLA-DR	0.189 ± 0.061	0.325 ± 0.137	−11.291	<0.001
CD117	0.232 ± 0.063	0.336 ± 0.111	−9.093	<0.001
CD34	0.222 ± 0.063	0.352 ± 0.131	−10.764	<0.001
数据 2				
CD45	0.580 ± 0.077	0.631 ± 0.114	−2.352	0.021
CD13	0.412 ± 0.089	0.383 ± 0.156	1.048	0.298
CD33	0.404 ± 0.081	0.440 ± 0.158	−1.333	0.187
HLA-DR	0.289 ± 0.069	0.435 ± 0.148	−5.938	<0.001
CD117	0.393 ± 0.077	0.468 ± 0.115	−3.442	<0.001
CD34	0.231 ± 0.048	0.346 ± 0.155	0.024	<0.001

在数据 2 中,正常人与 AML 患者在 CD45、HLA-DR、CD117、CD34 抗原表达水平差异有统计学意义。

2.2 CNN 模型训练过程 设置初始学习率为 0.005,训练轮次为 30,在 CNN 模型的训练过程中(图 2),训练集上损失函数不断下降并趋于平缓,表明模型能够从训练集上学习到特征。以验证集损失函数作为模型泛化性能评价指标,为避免模型出现过拟合,在验证集损失函数达到最低点时终止训练,训练时验证集损失函数在第 18 轮训练时达到最低点,将训练完成后的模型保存。

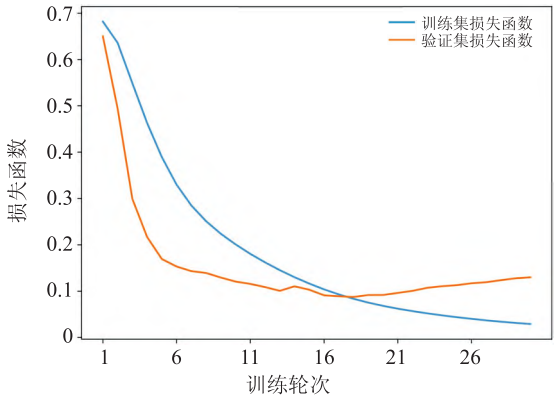

图 2 CNN 训练中损失函数变化图

2.3 CNN 模型性能评估 运用多个指标评价模型在测试集上的性能(表 2),在数据 1、数据 2 和合并后数据的 AUC 分别为 0.940(0.922 ~ 0.958)、0.917(0.885 ~ 0.949)和 0.932(0.916 ~ 0.948),见图 3。

表 2 CNN 在测试集上的性能

数据	准确率	灵敏度	特异度	查准率	F1 分数
数据 1	0.931	0.667	0.968	0.750	0.706
数据 2	0.851	0.636	0.940	0.824	0.718
合并	0.890	0.645	0.956	0.800	0.714

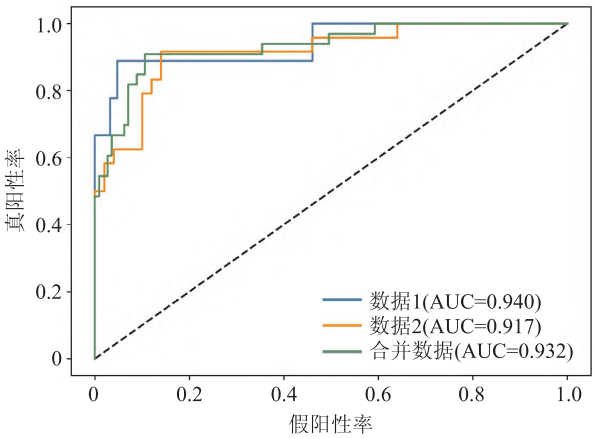

图 3 CNN 模型在测试集上的 ROC 曲线

3 讨论

AML 的精确诊断是其治疗和预后判断的关键，目前 MICM 分型依据是国际上通用的诊断方法，即细胞形态学、免疫学、细胞遗传学和分子生物学分型，其中免疫学分型是由 FCM 来实现的。随着对疾病的认识逐渐加深，越来越多的生物标志物被应用于 AML 的流式诊断中，给数据分析工作带来更大挑战，探究 FCM 数据自动分析成为近年来的研究热点[10-12]。Cheung et al[13]对现有的自动分析方法进行了使用调查后指出，虽然一些方法已经被证明有不错的效果，但仍然存在问题。例如许多软件通常只针对特定来源的数据进行自动分析，软件的跨平台使用问题依然难以解决，并且目前还没有针对白血病诊断的软件出现，已提出多数自动分析方法以无监督的机器学习为主，得到的结果需要人工进行二次分析，难以实现完全的自动化。因此尚没有任何一种自动分析方法能够被普遍接受，在实际临床工作中仍以人工分析为主。

传统的数据分析方法已经证明了将 FCM 数据转化为图像是一种切实可行的策略，其局限性产生原因是人类视觉的限制，而 CNN 模型已经成为代替人类视觉进行图像分析的最好方法，被广泛的应用于医学图像分类和识别方面，并证明在很多问题上与专业医师相当[14-15]。本研究针对 AML 提出了一种 FCM 数据结构，该结构参考了图像数据的存储方式，以常作为骨髓细胞类型判断依据的 SSC 和 CD45 作为像素的定位点，将其余与 AML 初步诊断相关抗原的表达水平写入图像数据的颜色通道。这种结构的优点是显而易见的，首先，该结构能够将 FCM 多管数据整合在同一矩阵中，可以通过调整矩阵的尺寸改变分辨率和纳入抗原的数量，有利于形成统一的标准；其次，该结构可以体现各抗原表达水平间的交互关系，有助于发现人工设门中可能遗漏的信息；另外，该结构可以在不改变模型参数的情况下兼容多数 CNN 架构，便于后续的自动化研究。本研究中选择了两组不同来源的数据，用公开数据进行建模后直接对本地数据进行测试，在本地数据上的测试结果与公开数据相似，均具有较高准确率，证明模型不仅能够准确识别 AML，还具有很强的鲁棒性，可以解决软件跨平台使用的问题，相较于其他算法更具有临床应用价值。

同时，本研究还存在一些局限性。由于本研究中使用的数据仅提供了 AML 患者和正常人的标签，缺乏更详细的疾病信息，因此仅探讨了 CNN 模型在 AML 初步诊断中的应用，对于免疫分型和微小残留病等问题并未提及。本研究仅纳入 7 个参数，且数据为二分类，因此选择了结构相对简单的 LeNet-5 架构，如果后续需要纳入更多变量或解决更加复杂的多分类问题，也可以选择深度更大的 CNN 模型。在本研究中对训练集采用了细胞随机抽样和数据增强，虽然一定程度上能弥补小样本对模型的不良影响，但是无法从根本上解决数据缺乏导致的模型训练中有效特征遗漏问题，如需进一步提高模型性能，扩大数据量是必须的途径。为了保证结果的客观性，本研究没有深入探讨模型参数的选择，而是尽可能的选择工具包默认参数，可能导致得到的模型并不是最优的，可在将来实用过程中进一步完善。

综上所述，本研究提出了一种 FCM 数据结构，并用 CNN 模型实现了 AML 的自动识别，表明 CNN 在 FCM 数据分析中具有一定的应用价值。

参考文献

[1] Shallis R M, Wang R, Davidoff A, et al. Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges[J]. Blood Rev, 2019, 36: 70-87.

[2] Weeda V, Mestrum S G C, Leers M P G. Flow cytometric identification of hematopoietic and leukemic blast cells for tailored clinical follow-up of acute myeloid leukemia[J]. Int J Mol Sci, 2022, 23(18): 10529.

[3] Chen X, Cherian S. Acute myeloid leukemia immunophenotyping by flow cytometric analysis[J]. Clin Lab Med, 2017, 37(4): 753-69.

[4] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE Trans Pattern Anal Mach Intell, 2020, 42(2): 318-27.

[5] Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare[J]. Nat Med, 2019, 25(1): 24 - 9.
[6] Spidlen J, Breuer K, Rosenberg C, et al. FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications[J]. Cytometry A, 2012, 81(9): 727 - 31.
[7] Arber D A, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia[J]. Blood, 2016, 127(20): 2391 - 405.
[8] Wang S, Brinkman R R. Data-driven flow cytometry analysis[J]. Methods Mol Biol, 2019, 1989: 245 - 65.
[9] 中国中西医结合学会检验医学专业委员会. 急性白血病系别判断的流式细胞免疫分型专家共识[J]. 中华检验医学杂志, 2021, 44(12): 1113 - 25.
[10] 孟晓辰, 王 玥, 祝连庆. 基于t分布邻域嵌入算法的流式数据自动分群方法[J]. 生物医学工程学杂志, 2018, 35(5): 697 - 704.
[11] 马闪闪, 董明利, 张 帆, 等. 基于核主成分分析的流式细胞数据分群方法研究[J]. 生物医学工程学杂志, 2017, 34(1): 115 - 22.
[12] 周丽娜, 苗林子, 龚 岩, 等. 人工智能辅助多参数流式细胞术诊断儿童急性B淋巴细胞白血病微小残留病[J]. 中华检验医学杂志, 2020, 43(12): 1196 - 204.
[13] Cheung M, Campbell J J, Whitby L, et al. Current trends in flow cytometry automated data analysis software[J]. Cytometry A, 2021, 99(10): 1007 - 21.
[14] Ting D S W, Cheung C Y, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes[J]. JAMA, 2017, 318(22): 2211 - 23.
[15] Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115 - 8.

Application of convolutional neural network in flow cytometry diagnosis of acute myeloid leukemia

Lei Wei[1], Li Zhiwei[2], Rui Dongsheng[1], Zhang Mei[1], Guo Yujuan[1], Bai Wenli[1], Wang Kui[1]
([1]*Dept of Preventive Medicine, School of Medicine, Shihezi University, Shihezi* 832002;
[2]*Clinical Testing Center, Xinjiang Uygur Autonomous Region People's Hospital, Urumqi* 830001)

Abstract ***Objective*** A convolutional neural network (CNN) model was established to automatically analyze flow cytometry (FCM) data to achieve the preliminary diagnosis of acute myeloid leukemia(AML), and explore the feasibility of applying CNN model to FCM data analysis. ***Methods*** The exploratory study of CNN application was carried out using the bone marrow FCM data obtained by the FlowRepository database and the Clinical Testing Center of Xinjiang Uygur Autonomous Region People's Hospital, and the data had been clinically confirmed whether AML was present. Among them, the public data was divided into training sets, validation sets and test sets according to 6∶2∶2, and local data was used for external test; In order to adapt the FCM data to the CNN model, an FCM data structure based on the image matrix principle was proposed, and after preprocessing the original data, the variables related to the preliminary diagnosis of AML were extracted, including sidescattered light and the expression levels of CD45, CD13, CD33, HLA-DR, CD117, CD34, and each variable was written into the matrix. Cell sampling and data augmentation methods were used to increase the sample size of the training set, the keras software package was used to build the LeNet-5 CNN model in Python, and the training set and the validation set were used for model training and parameter tuning respectively to evaluate the performance of the model on the test set. ***Results*** The accuracy of CNN to identify AML on the two test sets was 0.931, 0.851, the sensitivity was 0.667, 0.636, the specificity was 0.968, 0.940, and the area under the receiver operating characteristic curve was 0.940 and 0.917. ***Conclusion*** Based on the proposed FCM data structure, the CNN model can realize the preliminary diagnosis of AML, indicating that CNN has certain application value in FCM data analysis.

Key word flow cytometry; acute myeloid leukemia; convolutional neural networks; automated diagnosis