·论著·

FCM 数据细胞亚群分类和标注的自动化研究

摆文丽[1],农卫霞[2],李智伟[3],雷 伟[1],郭玉娟[1],张向辉[1],芮东升[1],王 奎[1]

(1.石河子大学医学院预防医学系,新疆 石河子 832000;
2.石河子大学医学院第一附属医院血液风湿科,新疆 石河子 832000;
3.新疆维吾尔自治区人民医院临床检验中心,新疆 乌鲁木齐 830001)

摘要:目的 研究自动化分析方法用于流式细胞术数据分析、解决细胞亚群自动分类和标注问题的价值,以期为疾病诊断提供参考。**方法** 收集 2021 年急性白血病骨髓流式检测数据 528 例,通过补偿、转换以及去粘连完成流式细胞术原始数据预处理,对预处理后的数据使用无监督聚类方法进行聚类分析,利用生成的细胞亚群的中心位置,即宏细胞的分布规律来训练有监督分类模型,将亚群进一步分类,最后通过人工识别与标注,将细胞亚群标注为已知的细胞类型。**结果** 无监督聚类方法与有监督分类方法共同用于流式细胞术数据分析能够实现细胞亚群的自动分类与标注,且准确度达到或基本达到手工分析水平。**结论** 该研究提出的流式细胞术数据自动分类和标注方法,解决了目前流式细胞自动化分析存在的细胞聚类方法和病人分类方法之间不相关问题,为全程自动化提供了解决方案;且提供的临床诊断中所需的中间结果,可用于疾病诊断的质量控制。

关键词:流式细胞术;聚类分析;亚群分类;亚群标注;自动化分析

中图分类号:R319 **文献标识码:**A **DOI:**10.3969/j.issn.1006-1959.2024.06.013

文章编号:1006-1959(2024)06-0078-06

Automatic Classification and Labeling of Cell Subsets in FCM Data

BAI Wen-li[1],NONG Wei-xia[2],LI Zhi-wei[3],LEI Wei[1],GUO Yu-juan[1],ZHANG Xiang-hui[1],RUI Dong-sheng[1],WANG Kui[1]

(1.Department of Preventive Medicine,Shihezi University School of Medicine,Shihezi 832000,Xinjiang,China;
2.Department of Hematology and Rheumatology,the First Affiliated Hospital of Shihezi University School of Medicine,Shihezi 832000,Xinjiang,China;
3.Clinical Laboratory Center,Xinjiang Uygur Autonomous Region People's Hospital,Urumqi 830001,Xinjiang,China)

Abstract:Objective To study the value of automatic analysis method for flow cytometry data analysis and solving the problem of automatic classification and labeling of cell subsets, so as to provide reference for disease diagnosis.**Methods** The data of bone marrow flow cytometry from 528 cases of acute leukemia in 2021 were collected, and the original flow cytometry data were preprocessed by compensation, conversion and deadhesion. The preprocessed data were analyzed by unsupervised clustering method, and the supervised classification model was trained by using the central location of the generated cell subsets, namely the distribution rule of macro cells, to further classify the subsets. Finally, the cell subsets were labeled as known cell types by manual recognition and labeling.**Results** Unsupervised clustering method and supervised classification method could be used in flow cytometry data analysis, which can realize automatic classification and labeling of cell subsets, and the accuracy can reach or almost reach the level of manual analysis.**Conclusion** The method of automatic classification and labeling of flow cytometry data proposed in this study bridge the gap between cell clustering and patient classification existing in current flow cytometry automation, and provid a solution for the whole process automation. The intermediate results required for clinical diagnosis can be used for quality control of disease diagnosis.

Key words: Flow cytometry;Cluster analysis;Subpopulation classification;Subpopulation registration;Automatic analysis

流式细胞术(flow cytometry,FCM)是一种能够精确、快速地对生物细胞或微粒的理化特性和生物学特性进行定量分析的技术[1]。随着精准医疗和基因生物学的发展,FCM 已经成为恶性血液病诊断的重要依据[2]。FCM 数据在人工分析中最关键和最耗时的步骤是识别数据中的同质细胞群,这个过程为“设门”[3]。数据传统的分析方法是通过不同参数组合进行人工设门,随着检测参数成倍增加,产生了多组合、高维度的流式数据,而 FCM 数据分析成为 FCM 中最具挑战性和最耗时的诊断步骤[4-7]。自动设门是基于细胞群荧光强度分布的数学建模,可以使用有监督和无监督的方法来执行,用于解决人工设

作者简介:摆文丽(1997.7-),女,甘肃白银人,硕士研究生,主要从事卫生统计学和流式细胞术的研究

通讯作者:王奎(1967.3-),男,重庆人,博士,副教授,主要从事卫生统计学、生物信息学以及流式细胞术的研究

门所面临的问题。目前常见的自动化分析方法包括 FlowMeans[8]、SPADE[9]、Citrus[10]、FlowSOM[11]以及 PCA[12]等，其中最常用的是 FlowMeans，其是一种无监督聚类方法，通过合并多个聚类以获得最终细胞亚群[13,14]，但只能将 FCM 数据中相似的细胞聚成亚群[15,16]，不能实现亚群的标注，因此需要工作人员去一一识别，存在一定局限性。基于此，本研究旨在分析 FlowSOM 与有监督分类模型[17]（混合正态分布模型）联合应用于 FCM 数据自动化分析中的效果，现报道如下。

1 资料与方法

1.1 数据来源 数据来源于实验室 2021 年 1 月-12 月同一面板急性白血病骨髓检测数据，共 528例，包括 412 例正常人、68 例 AML、9 例 T-ALL 以及 39例 B-AL。本研究经当地政府伦理委员会批准。

1.2 数据分析 FCM 数据细胞亚群的自动分类和自动标注可以分成 4 个阶段进行：①预处理：通过读取数据、补偿和转换、去粘连完成 FCM 数据预处理；②细胞聚类：使用 FlowSOM 方法对预处理的数据进行细胞聚类，聚类的结果以宏细胞的方式可视化；③亚群分类：利用混合正态分布模型，训练有监督分类模型对细胞亚群进行分类；④亚群标准：对③得到的有限个数的细胞亚群类进行识别和标注建立多对多映射，完成细胞亚群的标注。

1.2.1 数据预处理 通过补偿、转换和去粘连完成 FCM 数据的预处理。①首先应用补偿矩阵对数据进行补偿，补偿矩阵采用流式 fcs 格式数据自带的补偿矩阵，通过读取荧光抗体名称与提取荧光通道的数据矩阵，对荧光抗体做补偿[5]；②接着对 FCM 数据做转换，对前向散射光 FSC 进行线性变换（除以 100 k），侧向散射光 SSC 进行 Log_{10} 对数转换，对抗体做双指数变换；③最后使用百分位法在 FSC-A 和 FSC-H 平面对数据做去粘连处理，具体步骤如下：首先选取 FSC-H 大于 0.5 且 FSC-A 小于 2 的细胞子集，计算其在全体细胞中的占比；当子集占比小于等于 0.75 时，使用子集计算变量 FSC-A 与 FSC-H 的百分位点 P_5 和 P_{75}，否则计算 P_5 和 P_{90}；以两个对子为端点做基准线段，将连线垂直上移和下移 0.225 单位做两条平行线；两条平行线之外的点即为粘连细胞；FSC-H 小于 0.2 的点对应于细胞碎片，其余的为进入后续分析的细胞，包括正常细胞和凋亡细胞。上述切割点的选择用试错法确定。

1.2.2 细胞聚类 细胞聚类采用无监督分析方法，在操作中不需要任何标签，任何预定义的类作为引用。聚类算法识别同一聚类中的事件，将相似的细胞保留在同一个集群中，不同的细胞保留在不同的集群中。FlowSOM 具有节点网格，每个节点代表多维空间中的点[17]。自组织映射（the self-organizing map，SOM）将数据中的单元格分配给最近的节点，该节点以及周围的节点向新单元格更新，以此类推，节点被分配到数据空间中的高密度区域，节点网格中相近的节点比较远的节点更相似[18]。因此，所有的单元格将会分配到距离他们最近的节点，从而将 FCM 数据中相同的细胞聚类在一起形成细胞亚群。为便于观察聚类结果，FlowSOM 聚类结果以亚群中心点展示，下文中把亚群中心点称为宏细胞。聚类的目标是将 FCM 数据分为若干个类群，并保证类群内的样本尽可能密集，不同类群之间尽可能离散。FlowSOM 将 FCM 数据中相似的节点聚在一起形成无标签的细胞亚群，以宏细胞的形式展示。当比较 5×5、10×10 和 15×15 网格时，发现节点数量越多对应的纯度越高，但是聚类结果很混乱；根据经验，前 4 管使用 12×12 网格，第 5 管使用 10×10 网格，因此前 4 管的每管有 144 个宏细胞，第 5 管有 100 个宏细胞。

1.2.3 亚群分类 聚类分析后得到细胞聚类结果，但由于 FlowSOM 是无监督学习方法，不同抗体组合的样本得到的亚群构成不一致，导致亚群次序混乱缺乏统一标签，需要对细胞亚群进行分类[19]。把标本分为训练集和测试集，训练基于混合正态分布的有监督分类模型对所有的宏细胞进行分类，也就是对细胞亚群进行统一分类，混合正态分布模型的类别数设置为 20。有监督的混合正态分布模型对 FlowSOM 生成的宏细胞结果进行分类。具体步骤如下：为了避免数据过少导致训练集分类结果代表性差，选择 60%~70%的数据作为训练集，30%~40%作为验证集，因此从 AML、T-ALL、B-ALL 数据中分别随机挑选 41、9、39 例数据作为训练集；正常人数据有 412例，如果随机选择 60%的数据作为训练集，这样使得训练集中正常人数据远远多于患者数据，正常人细胞亚群特征覆盖异常细胞亚群，造成分类不准

确，因此选择 100 例正常数据作为训练集。训练集 170 例数据，共 97 920 个宏细胞；测试集 358 例数据，共 206 208 个宏细胞，为了使分类结果清晰明了，从两个数据中随机选取 25 000 个宏细胞来显示。

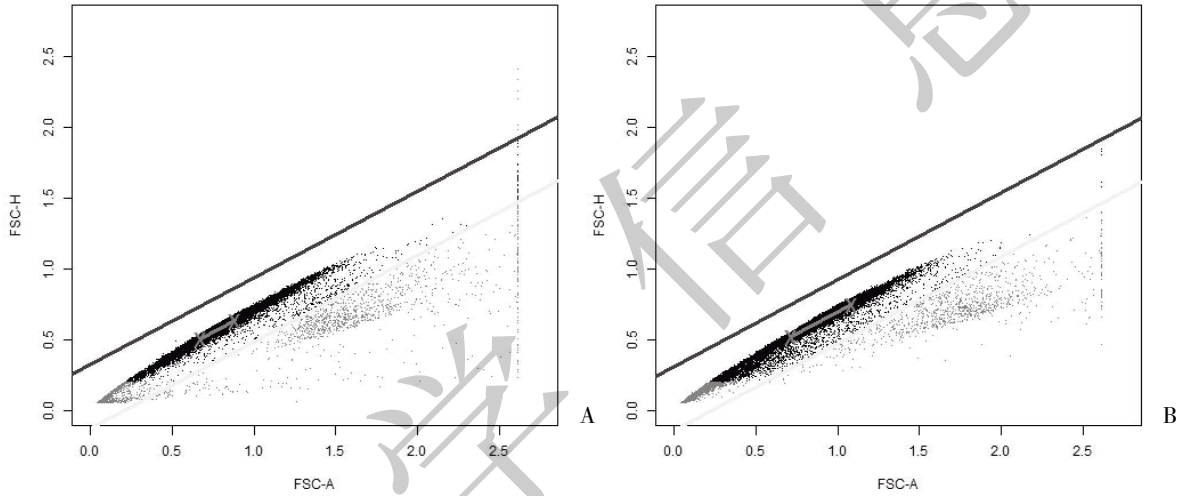
1.2.4 亚群标注 为使细胞亚群分类更加精确，分类模型中亚群数目的设置通常高于常规使用的细胞类型数。因此在亚群标注过程中，通过提取细胞聚类信息以及各类细胞的细胞数，将宏细胞映射到 9 个细胞类别并进行命名标注。

2 结果

2.1 粘连细胞的识别去除 以 FSC-A 和 FSC-H 为坐标绘制散点图，基准线上下移动 0.225 个单位产生两条平行线将粘连细胞去除，见图 1，经检查去粘连结果，发现粘连细胞划分均合理。

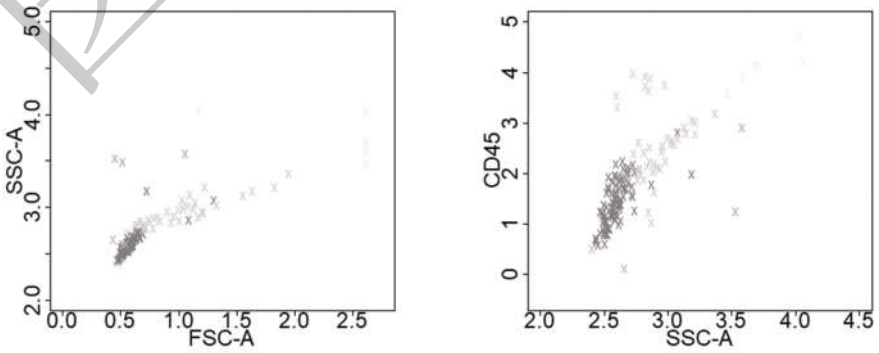
2.2 聚类分析 各类细胞的宏细胞分布是有规律可循，服从特定的概率分布，见图 2。

2.3 亚群分类与标注 共有 20 个类别，且各类宏细胞位置合理，未见异常，见图 3；另对 20 个细胞类别进行一一识别和标注，得到 9 种已知细胞类，分别是淋巴细胞、单核细胞、中性粒细胞、嗜酸粒细胞、原始细胞、幼稚细胞、有核红细胞、凋亡细胞、其他细胞，见图 4。



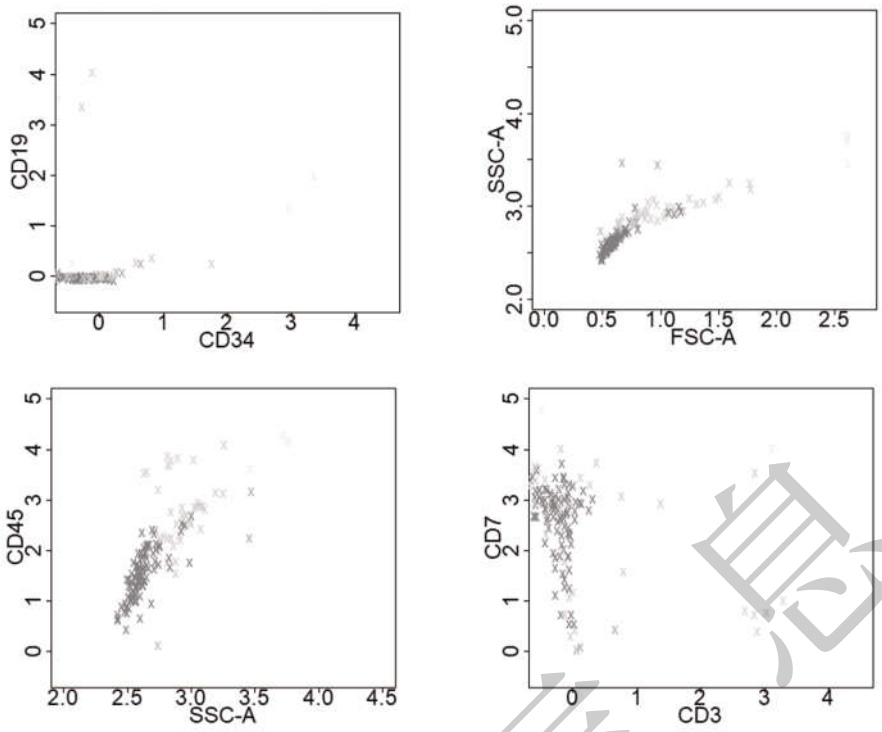
注：A：对应子集在全部细胞的占比小于等于 75%的案例，B：对应占比大于 75%的案例

图 1 预处理结果



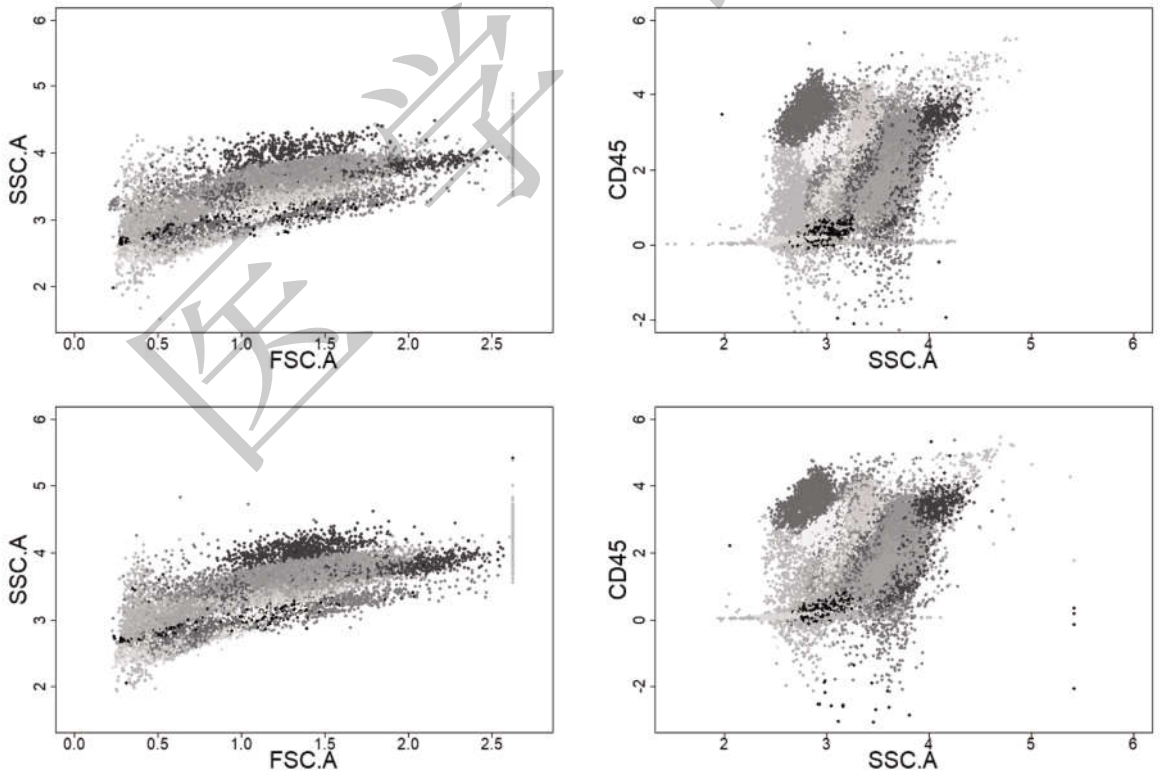
注：第一行为第一管宏细胞的散点图，第二行为第二管宏细胞的散点图

图 2 FlowSOM 聚类结果



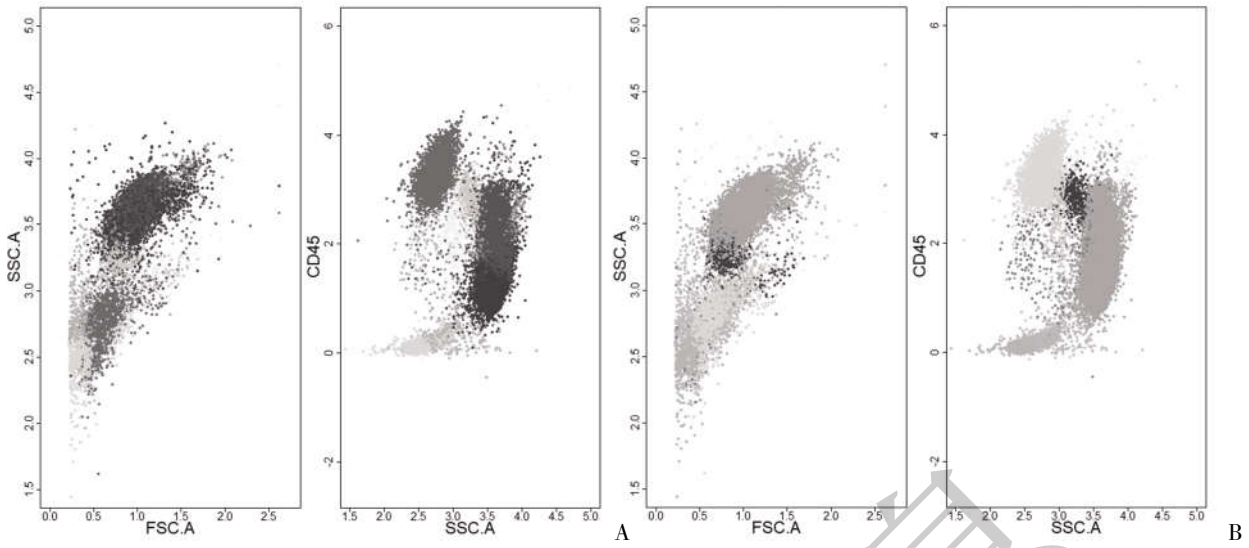
注:第一行为第一管宏细胞的散点图,第二行为第二管宏细胞的散点图

图 2 FlowSOM 聚类结果(续)



注:第一行为训练集的分类结果,第二行为测试集的分类结果

图 3 模型测试与验证结果



注：A：亚群标注前，有监督分类模型得到 20 个亚群的结果，同一亚群为相同的颜色，标注前这 20 个细胞类没有名字；B：亚群标注后的结果，第一行的 20 个类被标注为 9 个已知细胞类，其中相同颜色为同一种亚群

图 4 细胞亚群标注前后比较

3 讨论

由于 FCM 具有高通量、高灵敏度、高精确度以及多参数检验的特点[20]，被广泛的应用于生物学研究及临床诊断中[21-23]，同时会产生高维度、多组合的 FCM 数据。而传统人工分析具有分析效率低、主观性高的问题。近年来不断有学者提出 FCM 数据的分析需要自动化分析方法的帮助[5,24]。

针对以上问题，本研究提出无监督聚类方法与有监督分类方法共同用于 FCM 数据分析，模拟人工分析过程，获取临床流式实验室的原始检测数据，预处理过程通过补偿、转换、粘连细胞以及细胞碎片的去除，使得 FCM 数据规范化，检查发现每例数据的粘连细胞去除均合理；之后将无监督聚类方法与有监督分类方法结合起来用于 FCM 数据聚类、亚群分类与标注，显著优点是其既能够快速分类又能够提高分类数目的准确度。

无监督聚类方法 FlowSOM 作为分析的起点，将 FCM 数据中相似的细胞聚在一起形成无标签的细胞亚群，通过设定的参数，FlowSOM 将 FCM 数据中相似的细胞聚在一起形成无标签的细胞亚群，以宏细胞的形式展示。从聚类结果看出，FlowSOM 具有良好的性能以及快速的运行时间，是对 FCM 数据进行快速探索性分析的理想工具。但是将宏细胞进一步聚类时会出现不同细胞类型合并的现象，不能通过 FlowSOM 模型的元聚类对亚群进一步聚类与特征提取。因此，使用有监督分类模型混合正态分布模型对 FlowSOM 生成的宏细胞进行分类，有监督学习算法可以达到这样一种状态：在提供足够的信息数据前提下，它能够预测未见数据的正确标签；混合正态分布模型对亚群进行分类时，首先将数据集分为训练集和测试集，使用训练集训练有监督分类模型过程中，对亚群类别参数进行设定，发现随着亚群数的增加，分类精确度会提高，但是不利于对亚群进行标注；反之，亚群数减少，精确度降低，但是会出现将不同细胞亚群分到一起的现象。故根据经验，将细胞亚群设置为 20 个，接下来使用测试集对模型进行测试，检查训练集与测试集的分类结果，未见异常，可以认为有监督分类模型能够准确地对训练集和测试集进行分类。最后通过设定标签的形式将 20 个类别依次识别并用已知的细胞类别进行标注，即将宏细胞映射到 9 个细胞类别，对这 9 个细胞类别进行命名标注，检查所有数据标注前与标注后的可视化结果图，亚群标注结果清晰，未见异常。

总之，通过将基于本研究方法的亚群分类与标注结果与传统人工分析结果进行对比，成功验证了自动化分析方法在 FCM 数据分类与标注中的可行性和高准确性，具有较好的应用前景，可以为下游 FCM 数据自动化诊断提供参考，并且能够保留原始

数据更多的特征信息，为下游诊断结果的质量控制提供依据。本研究也有不足之处：作为流式数据全程自动化分析的重要组成，而且分类结果较难用评价指标进行评价，因此利用分类结果进行特征提取和疾病诊断，诊断结果与专家人工分类结果基本相同，从而反推证明本研究提出的 FCM 数据自动化分类方法可靠；自动化分析 FCM 数据时假设流式实验室在样本准备、荧光染色、仪器校准和调整阶段均正常，在实际情况中，可能出现数据大幅度偏移，建立在分布规律基础上的亚群标注结果可能会出现偏差。目前，本研究提出的自动化分析方法已经在公共数据库 Flowrepository.orgAML 项目提供的数据以及本地实验室急性白血病骨髓检测数据进行过测试，效果良好。

参考文献：

[1]Farmer JR,DeLelys M.Flow Cytometry as a Diagnostic Tool in Primary and Secondary Immune Deficiencies [J].Clin Lab Med,2019,39(4):591-607.

[2]李宗儒,江倩.《中国成人急性淋巴细胞白血病诊断与治疗指南（2021 年版）》——Ph 阳性急性淋巴细胞白血病治疗的解读与思考[J].临床血液学杂志,2022,35(3):159-164.

[3]Li Y,Mahjoubfar A,Chen CL,et al.Deep Cytometry: Deep learning with Real-time Inference in Cell Sorting and Flow Cytometry[J].Sci Rep,2019,9(1):11088.

[4]Bashashati A,Brinkman RR.A survey of flow cytometry data analysis methods[J].Adv Bioinformatics,2009,2009:584603.

[5]Cheung M,Campbell JJ,Whitby L,et al.Current trends in flow cytometry automated data analysis software [J].Cytometry A,2021,99(10):1007-1021.

[6]Kumanovics A,Sadighi AA.Flow cytometry for B-cell subset analysis in immunodeficiencies[J].J Immunol Methods,2022,509:113327.

[7]Aghaeepour N,Finak G,Hoos H,et al.Critical assessment of automated flow cytometry data analysis techniques[J].Nat Methods,2013,10(3):228-238.

[8]Fuda F,Chen M,Chen W,et al.Artificial intelligence in clinical multiparameter flow cytometry and mass cytometry-key tools and progress[J].Semin Diagn Pathol,2023,40(2):120-128.

[9]Devine RD,Behbehani GK.Mass Cytometry, Imaging Mass Cytometry,and Multiplexed Ion Beam Imaging Use in a Clinical Setting[J].Clin Lab Med,2021,41(2):297-308.

[10]Paproski RJ,Pink D,Sosnowski DL,et al.Building predictive disease models using extracellular vesicle microscale flow cytometry and machine learning[J].Mol Oncol,2023,17(3):407-421.

[11]Lacombe F,Lechevalier N,Vial JP,et al.An R -Derived FlowSOM Process to Analyze Unsupervised Clustering of Normal and Malignant Human Bone Marrow Classical Flow Cytometry Data[J].Cytometry A,2019,95(11):1191-1197.

[12]Novikova NI,Matthews H,Williams I,et al.Detecting Phytoplankton Cell Viability Using NIR Raman Spectroscopy and PCA[J].ACS Omega,2022,7(7):5962-5971.

[13]Liu B,Zhang T,Li Y,et al.Kernel Probabilistic K-Means Clustering[J].Sensors (Basel),2021,21(5):1892.

[14]Cheung M,Campbell JJ,Thomas RJ,et al.Assessment of Automated Flow Cytometry Data Analysis Tools within Cell and Gene Therapy Manufacturing[J].Int J Mol Sci,2022,23(6):3224.

[15]Lacombe F,Lechevalier N,Vial JP,et al.An R -Derived FlowSOM Process to Analyze Unsupervised Clustering of Normal and Malignant Human Bone Marrow Classical Flow Cytometry Data[J].Cytometry A,2019,95(11):1191-1197.

[16]Bene MC,Axler O,Violidaki D,et al.Definition of Erythroid Differentiation Subsets in Normal Human Bone Marrow Using FlowSOM Unsupervised Cluster Analysis of Flow Cytometry Data[J].Hemasphere,2021,5(1):e512.

[17]Lo K,Brinkman RR,Gottardo R.Automated gating of flow cytometry data via robust model-based clustering[J].Cytometry A,2008,73(4):321-332.

[18]Van Gassen S,Callebaut B,Van Helden MJ,et al.FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data[J].Cytometry A,2015,87(7):636-645.

[19]Bene MC,Lacombe F,Porwit A.Unsupervised flow cytometry analysis in hematological malignancies: A new paradigm[J].Int J Lab Hematol,2021,43 Suppl 1:54-64.

[20]Mair F,Prlic M.OMIP-44: 28-Color Immunophenotyping of the Human Dendritic Cell Compartment [J].Cytometry A,2019,95(8):925-926.

[21]Suo Y,Gu Z,Wei X.Advances of in vivo flow cytometry on cancer studies[J].Cytometry Part A,2020,97(1):15-23.

[22]McKinnon KM.Flow Cytometry: An Overview[J].Curr Protoc Immunol,2018,120(1):1-11.

[23]Jaye DL,Bray RA,Gebel HM,et al.Translational applications of flow cytometry in clinical practice[J].JImmunol,2012,188(10):4715-4719.

[24]Van Nguyen T,Alfaro AC.Applications of flow cytometry in molluscan immunology: Current status and trends [J].Fish Shellfish Immunol,2019,94:239-248.

收稿日期：2023-02-18；修回日期：2023-04-19
编辑/杜帆