韦少钗, 杨晓帆, 洪程铭, 等. 基于蛋白质组学数据的小样本食管癌诊断方法[J]. 智能计算机与应用,2024,14(12):145-150. DOI:10.20169/j.issn.2095-2163.24101802

基于蛋白质组学数据的小样本食管癌诊断方法

韦少钗, 杨晓帆, 洪程铭, 李舜音
(广东潮州卫生健康职业学院, 广东 潮州 521000)

摘 要: 基于蛋白质组学数据的食管癌分析与诊断面临样本规模小、序列长度大等问题,影响分析方法的泛化性和准确性。针对该问题,本文提出一种面向小样本学习的食管癌诊断方法。该方法在 Transformer 的基础上,首先为其引入局部窗口机制,以缓解序列长度过大引起的组合爆炸,并利用特征洗牌操作增大窗口感受野;再借鉴 Masked Autoencoder 思想,设计非对称的特征编码器和解码器,同时在解码器中加入分类任务,提升模型所学习的特征质量;最后通过编码器创建食管癌和癌旁非肿瘤组织的原型表示,实现小样本分类。实验结果表明,所提方法在新数据集中仅需学习 6 对配对样本,占总样本的 10%,即可获得 93.45%和 95.19%的精确率和召回率,可为食管癌的智能化诊断提供思路。

关键词: 食管癌; 小样本; 蛋白质组学数据; Transformer; Masked Autoencoder
中图分类号: R735.1; TP183 **文献标志码**: A **文章编号**: 2095-2163(2024)12-0145-06

Proteomics-based diagnosis of esophageal cancer with a few samples

WEI Shaochai, YANG Xiaofan, HONG Chengming, LI Shunyin
(Guangdong Chaozhou Health Vocational College, Chaozhou 521000, Guangdong, China)

Abstract: The analysis and diagnosis of esophageal cancer based on proteomics data face the problems of small sample size and large sequence length, which affect the generalization and accuracy of the analysis method. To address this problem, a few-sample learning-oriented method for esophageal cancer diagnosis is proposed. Based on Transformer, the method firstly introduces a local window mechanism for it to alleviate the combinatorial explosion caused by the large sequence length, and uses the feature shuffling operation to increase the window receptive field. Then, drawing on the idea of Masked Autoencoder, the asymmetric feature encoder and decoder are designed, and the classification task is added into the decoder to improve the quality of feature learned by the encoder. Finally, the prototype representations of esophageal cancer and adjacent nontumor tissues are created by the encoder to achieve diagnosis with a few samples. The experimental results show that the proposed method only needs to learn 6 paired samples (10% of the total samples) in the new dataset to obtain a precision rate of 93.45% and a recall rate of 95.19%, which can provide ideas for intelligent diagnosis of esophageal cancer.

Key words: esophageal cancer; a few samples; proteomics data; Transformer; Masked Autoencoder

0 引 言

食管癌是最具侵袭性和致死性的恶性肿瘤之一[1]。目前,食管癌的诊断主要依靠病理检查,耗时费力且高度依赖医生经验。蛋白质组学数据为食管癌的诊断研究提供了重要信息[2],但该类数据的采集成本较高,较难开展大规模采集工作[3],加之不同数据集间存在批次效应、单个样本包含数千甚至数万种蛋白质的表达量,基于蛋白质组学数据的分析与诊断还存在一定挑战[4-5]。

深度学习具备强大的特征提取能力,近些年逐渐被应用至蛋白质组学分析领域[6],如 Kim 等学者[7]探究了深度学习中节点数量、随机失活比例等超参数对胰腺癌诊断的影响。Dong 等学者[8]使用深度神经网络区分肿瘤和非肿瘤样本。但上述方法存在 2 个问题:一是仅采用卷积层或全连接层捕获不同特征间的互作关系效率较低[9];二是模型难以通过一个或数个样本快速建立对新数据集的泛化能力。

针对第一个问题,Vaswani[10]提出了 Transformer

基金项目: 广东省高等职业院校医药卫生类专业教学指导委员会项目(2022LX048); 潮州市卫生健康局科研项目(2023075)。
作者简介: 韦少钗(1995—),女,助教,主要研究方向:生物信息学,深度学习技术。Email: shaochaiwei@163.com。
收稿日期: 2024-10-18

模块,以提高模型捕获不同特征间互作关系的能力,但该模块的计算复杂度为输入序列长度的平方级,难以应用于长度为数千、甚至数万的蛋白质组学数据。后续提出的 Swin Transformer[11] 虽能降低计算复杂度,但需借助较多训练样本才能获得与卷积神经网络相当的预测能力[12]。

针对第二个问题,研究学者设计了多种面向小样本学习的模型,诸如孪生网络[13]、原型网络[14]等。其中,孪生网络需要处理成对的输入数据,在训练和推理阶段计算复杂度较高;原型网络的应用前提是构建一个具备较强特征提取能力的编码器。目前,针对蛋白质组学数据专门设计的特征编码器仍相对较少。

基于上述问题,本文以 Liu 等学者[15]和 Zhao 等学者[16]公开的数据为例,提出一种基于蛋白质组学数据的小样本食管癌诊断方法。该方法在 Transformer 的基础上,首先为其引入局部窗口机制,以缓解序列长度过大引起的组合爆炸;再借鉴 Masked Autoencoder(MAE)[17]思想,设计非对称的特征编码器和解码器,同时在解码器中加入分类任务,辅助编码器学习不同特征间的互作信息和样本的类别信息;最后通过编码器创建食管癌和癌旁非肿瘤组织的原型表示,根据查询样本的编码表示和每个类别的原型表示的余弦相似度,确定查询样本所属类别。实验结果表明,所提方法的精确率和召回率分别为 93.45%和 95.19%。

1 实验数据

本文使用 Liu 等学者[15]和 Zhao 等学者[16]公开的数据进行实验(以下分别简称为 L 数据集和 Z 数据集)。2 个数据集共包含 184 对配对样本,其中 L 数据集包含 124 对,Z 数据集包含 60 对。每对配对样本包含食管癌和癌旁非肿瘤组织的蛋白质组学数据各 1 例,共 368 例,选取 L 数据集和 Z 数据集中共有的 5 687 种蛋白质进行实验。

本文采用随机划分的方式构建训练集、验证集、支持集和查询集:从 L 数据集中随机选取 87 对样本作为训练集,余下 37 对样本作为验证集;从 Z 数据集中随机选取 6 对样本作为支持集,余下 54 对样本作为查询集。为避免数据集划分的偶然性,对数据集进行了 5 次随机划分,形成 5 组训练集、验证集、支持集和查询集。其中,训练集用于调整编码器和解码器参数,验证集用于初步评估编码器和解码器性能,支持集用于创建食管癌和癌旁非肿瘤组织的原型表示,查询集用于评估本文方法最终的泛化能力。

2 食管癌诊断方法

2.1 方法流程

本文方法共包含 3 个步骤:

(1)用训练集调整编码器和解码器参数,并通过验证集初步评估编码器和解码器性能;

(2)将支持集样本输入编码器,创建食管癌和癌旁非肿瘤组织的原型表示;

(3)通过编码器获得查询样本的编码表示,根据该编码表示和每个类别的原型表示的余弦相似度,确定查询样本所属类别。

2.2 模型结构

模型结构如图 1 所示。由图 1 可知,模型共包含 12 个模块,特征自左向右传播,每个模块内部以 $L \times C$ 的形式表示输出特征的长度和通道数。考虑到模型中共有 2 次下采样,每次下采样后特征长度为原先的 1/4,输入数据的长度应为 16(4^2)的倍数,本文选择在输入尾部使用零值填充 9 个表达量,将输入长度增加至 5 696。

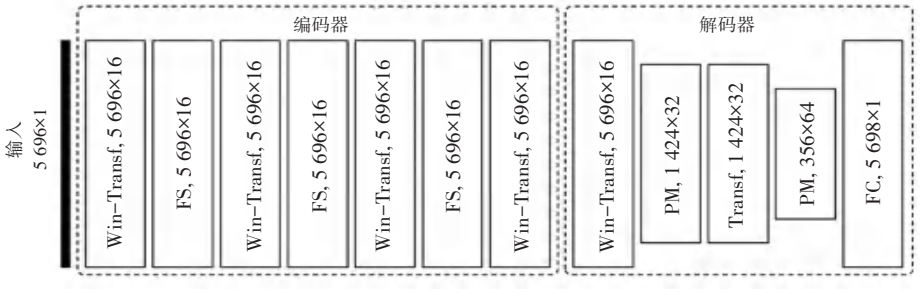


图 1 模型结构图

Fig. 1 Model architecture

模型中包含 5 种模块:基于局部窗口的 Transformer(Window based Transformer, Win-Transf)、特征洗牌(Feature Shuffling, FS)、区块合并(Patch Merging, PM)、标准的 Transformer(Transf)和

全连接(Fully Connected,FC)。其中,Win-Transf 模块用于获取每个窗口内不同特征的互作信息;FS 模块能将每个窗口所提取的互作信息分散至其它窗口;PM 模块的结构参照 Swin Transformer[11]设计,其作用是对输入特征进行下采样和通道数调整;Transf 模块用于获取所有特征的互作信息;FC 模块用于预测所有蛋白质的表达量和输入样本所属类别。

考虑到标准的 Transformer 的计算复杂度为特征长度的平方级,执行效率较低,本文借鉴 Swin Transformer 的思想,将注意力的计算限制在每个窗口内,即窗口内的特征仅能与该窗口内的其他特征进行内积。因窗口大小固定,Win-Transf 的计算复杂度与特征长度线性相关,执行效率得到较大幅度提高。Win-Transf 的结构参照 Swin Transformer 设计,设置注意力窗口大小为 712。

为增大窗口感受野,Swin Transformer 还对窗口进行了偏移操作,使其包含原本相邻窗口的信息,但偏移操作仅能进行相邻窗口间的信息交互。为进一步增大窗口感受野,本文参考 ShuffleNet[18]的通道洗牌方法,设计了 FS 模块,将每个窗口所提取的互作信息分散至其它窗口中。FS 模块的操作流程如图 2 所示。由图 2 可知,以大小为 12×2 的输入特征为例,假设窗口数量和大小分别为 3 和 4,该模块先在行方向上按位置间隔 3(该数值与窗口数量一致)选取元素,形成 3 个特征块,再将三者在行方向上拼接,供后续 Win-Transf 模块获取互作信息。

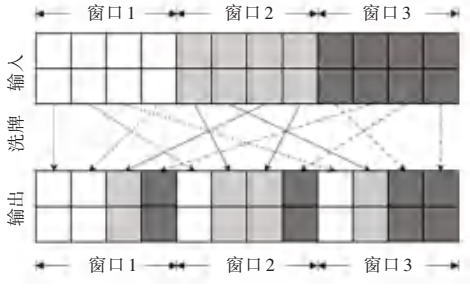


图 2　FS 模块操作流程

Fig. 2　Procedure of the FS block

由图 2 可知,FS 模块通过打乱不同窗口中的特征增强窗口间的信息交流,使得后续 Win-Transf 模块的每个窗口均能获得来自其他窗口的信息,进而提高模型表达能力。

2.3　MAE 自监督学习

为使模型具备较强的特征提取能力,研究人员通常采用大规模的人工标注数据集进行模型训练,但人工标注过于昂贵且费时。鉴于此,He 等学者[17]提出基于 MAE 的自监督学习方法,将数据本身的部分信息作为监督信号,并基于该监督信号训练模型。参考 MAE 的思想,本文设计了非对称的编码器和解码器结构,其中编码器仅能对可见的蛋白质进行特征提取,而解码器能对所有蛋白质进行特征提取。在模型训练阶段,每次迭代前随机选择 50%的蛋白质,将其表达量置零并设为不可见,且使用掩码机制确保编码器在提取特征时仅关注可见的蛋白质,而解码器则负责从编码表示和掩码标记中重建原始数据。解码器仅在训练阶段用于重建数据,训练完成后,弃置解码器。

考虑到 MAE 方法侧重于通过恢复不可见的输入数据进而学习数据中的关联信息,在没有类别监督的情况下,模型较难学习到样本的类别信息,本文继而参考了 Gidaris 等学者[19]提出的训练策略,将有监督的分类任务与自监督学习任务相结合,以进一步提高模型的特征提取能力,改善小样本分类性能。分类任务与自监督任务共用一个模型提取特征,仅将最后的全连接层的输出长度由 5 696 调整为 5 698,新增的 2 个输出用于预测样本所属类别。

3　模型训练

3.1　实验环境

实验环境具体如下:中央处理器为英特尔酷睿 i5-12500,运行内存 16 GB,GPU 为 NVIDIA RTX 3060,使用 PyTorch 1.12.0 作为深度学习框架,CUDA 版本为 11.3。

3.2　训练参数设置

用随机梯度下降法微调编码器和解码器参数,设置训练最大迭代次数为 6×10^3,其中前 4×10^3 次学习率为 10^{-3},接着 1×10^3 次学习率为 10^{-4},最后 1×10^3 次学习率为 10^{-5},冲量为 0.9,L2 正则化系数为 5×10^{-4},*mini - batch* 为 4,对全连接层设置 20%的随机失活。保留在验证集上损失最小的模型作为最佳模型,用于后续创建原型表示和评价泛化能力。

4　结果分析

4.1　评价指标

采用精确率(*Precision*)和召回率(*Recall*)[20]评价本文方法的泛化能力,用推理阶段处理单个样本所需的时间和模型文件大小评价计算复杂度,其中精确率和召回率的计算公式为:

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{1}$$

$$Recall = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{2}$$

其中，N_{TP} 表示肿瘤样本被正确分类的数量；N_{FP} 表示预测结果为肿瘤样本，实际为非肿瘤样本的数量；N_{FN} 表示预测结果为非肿瘤样本，实际为肿瘤样本的数量。

4.2 实验结果与分析

分类结果见表 1。由表 1 可知，5 次实验的精确率为 93.45%，召回率为 95.19%。模型文件（编码器）大小为 1.40 MB，处理单个样本所需的时间为 0.16 s。

表 1 本文方法在查询集上的分类结果

Table 1 Classification results of the proposed method on the query set

实验编号	N_{TP}	N_{FP}	N_{TN}	N_{FN}	精确率/%	召回率/%
1	51	2	52	3	96.23	94.44
2	52	4	50	2	92.86	96.30
3	51	5	49	3	91.07	94.44
4	50	4	50	4	92.59	92.59
5	53	3	51	1	94.64	98.15
合计	257	18	252	13	93.45	95.19

为验证本文改进的有效性，在最终提出的方案上，减少部分改进措施，以验证相应改进措施的必要性，实验结果见表 2。表 2 中，方案 5 对应本文所提方法，5 种方案的模型大小均为 1.40 MB。方案 1 表示使用标准的 Transformer 替换本文方法中的 Win-Transf 模块，因模型能直接学习序列内部所有特征的互作关系，该方案无需增大窗口感受野。方案 2 表示将本文方法中的特征洗牌操作替换为窗口偏移操作。方案 3 表示未将分类任务加入自监督学习中。方案 4 表示使用窗口偏移操作替换本文方法中的特征洗牌操作，且未将分类任务加入自监督学习中。

表 2 不同改进方案对分类性能的影响

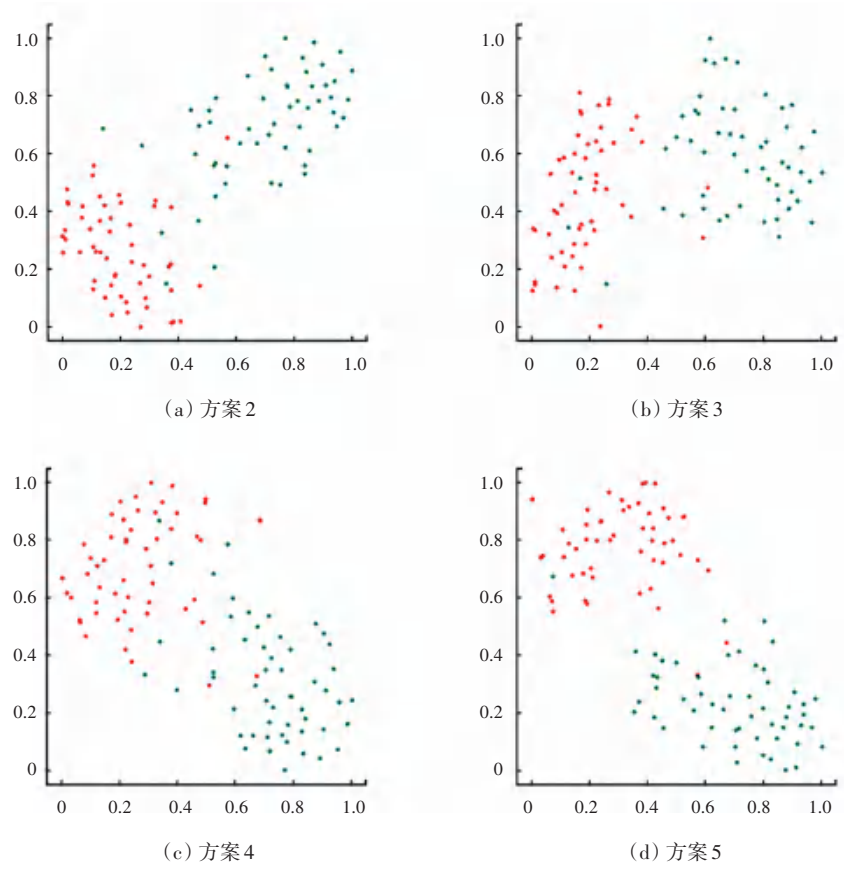
Table 2 Impact of different improvement schemes on classification performance

序号	局部窗口	增大感受野方法	分类任务	精确率/%	召回率/%	处理时间/s
1	×	—	√	83.21	82.59	1.18
2	√	Shift	√	92.00	93.70	**0.16**
3	√	Shuffle	×	91.14	91.48	**0.16**
4	√	Shift	×	86.59	88.52	**0.16**
5	√	Shuffle	√	**93.45**	**95.19**	**0.16**

表 2 表明，将窗口偏移改为特征洗牌和在自监督学习中引入分类任务两项改进，均能提升分类性能（对比方案 2、3 和 5）。两项改进均不会增加模型大小和处理时间，其中引入分类任务仅增加解码器全连接层参数量，但解码器并未应用于推理阶段。由方案 1 可知，基于全局的自注意力计算复杂度较高，与方案 5 相比单个样本的处理时间增加了 1.02 s。加之实验环境计算资源有限，方案 1 在训练阶段 *mini - batch* 仅能设为 1，模型较难达到收敛状态，故分类性能较差。

为进一步验证上述 2 项改进的有效性，本文利用 t-SNE 方法[21]对查询集所有样本的编码表示进行降维和可视化，结果如图 3 所示。因方案 1 所训练的模型较难达到收敛状态，可视化效果欠佳，故仅展示后 4 种方案。通过对比图 3（a）、（b）、（c）可知，2 项改进均可提升肿瘤样本和非肿瘤样本编码表示的区分度，且图 3（a）中 2 类样本的区分度略高于图 3（b），表明引入分类任务更有利于模型生成区分度更高的编码表示。图 3（d）中 2 类样本的区分度最高，表明 2 项改进同时应用可取得最高的分类性能。

表 3 为本文方法、Siamese Networks and Label Tuning（SNLT）[22]、Hybrid Attention-Based Prototypical Networks（HABPN）[23] 和 Simple Conditional Neural Adaptive Processes（SCNAPs）[24] 在查询集上的分类性能对比。为保持模型参数量接近，将 SNLT 原先使用的编码器 MPNet[25] 替换为本文所使用的编码器，HABPN、SCNAPs 使用 ResNet-20[26] 进行实验。



（a）方案2　（b）方案3　（c）方案4　（d）方案5

注：红色圆点表示非肿瘤样本，绿色圆点表示肿瘤样本。

图 3　查询样本的编码表示经降维后的可视化结果

Fig. 3　Visualization of the coded representation of the query sample after dimensionality reduction

实验表明，本文方法与其他方法相比精确率、召回率更高，但处理时间和模型大小高于 HABPN，其中处理时间比 HABPN 高 0.06 s，主要原因为 Transformer 涉及较多矩阵乘法和 Softmax 运算，而卷积神经网络通过局部连接和参数共享的方式降低计算量，在推理阶段效率更高[27]。SNLT 分类性能较低的主要原因为模型仅能使用样本的类别信息作为监督信号，限制了模型的表达能力；HABPN 和 SCNAPs 分类性能较低的原因为仅采用卷积层和全连接层难以捕获长距离的互作关系。

表 3　不同方法的分类性能对比

Table 3　Comparison of classification performance of different methods

方法	精确率/%	召回率/%	模型大小/MB	处理时间/s
本文方法	**93.45**	**95.19**	1.40	0.16
SNLT[22]	90.29	92.96	1.40	0.16
HABPN[23]	92.48	91.11	**1.08**	**0.10**
SCNAPs[24]	92.67	93.70	1.83	0.13

5　结束语

针对蛋白质组学数据样本规模小、序列长度大等问题，本文提出一种面向小样本学习的食管癌诊断方法。该方法首先在 Transformer 中引入局部窗口机制，并参考 ShuffleNet 设计了特征洗牌操作，以增大窗口感受野；再借鉴 MAE 思想，设计非对称的特征编码器和解码器，同时在解码器中加入分类任务，帮助模型生成更具代表性和区分度的特征表示；最后通过编码器创建食管癌和癌旁非肿瘤组织的原型表示，实现小样本分类。实验结果表明，所提方法具备较好的泛化性和准确性，可满足食管癌的智能化诊断需求。

参考文献

[1] RUSTGI A K, EL-SERAG H B. Esophageal carcinoma[J]. New England Journal of Medicine, 2014, 371(26): 2499-2509.

[2] LI L, JIANG D, ZHANG Q, et al. Integrative proteogenomic characterization of early esophageal cancer [J]. Nature Communications, 2023, 14(1): 1666.

[3] HOUFANI A A, FOSTER L J. Review of the real and sometimes

hidden costs in proteomics experimental workflows [M]// GEDDES _ MCALISTER J. Proteomics in Systems Biology: Methods and Protocols. New York: Humana Press, 2022: 1-14.
[4] CHANDRAMOULI K, QIAN Peiyuan. Proteomics: Challenges, techniques and possibilities to overcome biological sample complexity[J]. Human Genomics and Proteomics, 2009, 2009: 239204.
[5] 涂强强, 郭文静, 潘乔, 等. 基于小样本血浆蛋白质组学数据的抑郁症分类预测[J]. 智能计算机与应用, 2024, 14(8): 133-137.
[6] WEN Bo, ZENG Wenfeng, LIAO Yuxing, et al. Deep learning in proteomics[J]. Proteomics, 2020, 20(21-22): 1900335.
[7] KIM H, KIM Y, HAN B, et al. Clinically applicable deep learning algorithm using quantitative proteomic data[J]. Journal of Proteome Research, 2019, 18(8): 3195-3202.
[8] DONG Hao, LIU Yi, ZENG Wenfeng, et al. A deep learning-based tumor classifier directly using MS raw data[J]. Proteomics, 2020, 20(21-22): 1900344.
[9] WANG Xiaodong, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7794-7803.
[10] VASWANI A. Attention is all you need[J]. arXiv preprint arXiv, 1706.03762, 2017.
[11] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012-10022.
[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv, 2010.11929, 2020.
[13] KOCH G, ZEMEL R, SALAKHUTDINOV R. Siamese neural networks for one-shot image recognition[J]. ICML Deep Learning Workshop, 2015, 2(1): 1-30.
[14] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Advances in Neural Information Processing Systems. Long Beach, USA: NIPS Foundation, 2017, 30: 4077-4087.
[15] LIU Wei, XIE Lei, HE Yaohui, et al. Large-scale and high-resolution mass spectrometry-based proteomics profiling defines molecular subtypes of esophageal cancer for therapeutic targeting [J]. Nature Communications, 2021, 12(1): 4961.
[16] ZHAO D, GUO Y, WEI H, et al. Multi-omics characterization of esophageal squamous cell carcinoma identifies molecular subtypes and therapeutic targets[J]. JCI Insight, 2024, 9(10): e171916.
[17] HE Kaiming, CHEN Xinfei, XIE Saining, et al. Masked autoencoders are scalable vision learners[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16000-16009.
[18] ZHANG Xiangyu, ZHOU Xinyu, LIN Mengxiao, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6848-6856.
[19] GIDARIS S, BURSUC A, KOMODAKIS N, et al. Boosting few-shot visual learning with self-supervision[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 8059-8068.
[20] THARWAT A. Classification assessment methods[J]. Applied Computing and Informatics, 2021, 17(1): 168-192.
[2] LAURENS V D M, HINTON G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.
[22] MÜLLER T, PÉREZ-TORRÓ G, FRANCO-SALVADOR M. Few-shot learning with siamese networks and label tuning[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. ACL, 2022: 8532-8545.
[23] GAO Tianyu, HAN Xu, LIU Zhiyuan, et al. Hybrid attention-based prototypical networks for noisy few-shot relation classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6407-6414.
[24] BATENI P, GOYAL R, MASRANI V, et al. Improved few-shot visual classification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 14493-14502.
[25] SONG Kaitao, TAN Xu, QIN Tao, et al. MPNet: Masked and permuted pre-training for language understanding[J]. Advances in Neural Information Processing Systems, 2020, 33: 16857-16867.
[26] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.
[27] GUO Jianyuan, HAN Kai, WU Han, et al. CMT: Convolutional neural networks meet vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 12175-12185.