·新技术与新方法·

基于无监督深度学习的纳米孔测序 O^6-甲基鸟嘌呤识别

关晓宇[1]，王宇[2,3]，张金月[2,3]，邵伟[1]，黄硕[2,3]，张道强[1]

1. 南京航空航天大学 计算机科学与技术学院 工信部模式分析与机器智能重点实验室（南京 211106）
2. 南京大学 化学化工学院 生命科学分析化学国家重点实验室（南京 210023）
3. 南京大学 化学与生物医学创新中心（南京 210023）

【摘要】 O^6-甲基鸟嘌呤（O^6-CMG）是 DNA 中的一种高致突变烷基化产物，它会导致生命体罹患胃肠道肿瘤。现有的研究主要是利用耻垢分枝杆菌膜蛋白（MspA）纳米孔技术，借助枯草芽孢杆菌噬菌体 Phi29 DNA 多聚酶（Phi29 DNA polymerase）对突变进行精确定位。近年来，机器学习技术被广泛应用于纳米孔测序数据的分析，但是机器学习往往需要大量的数据标记，这给研究者们带来了额外的工作负担，大大影响了其实用性。因此，本文提出了一种纳米无监督深度学习（nano-UDL）方法，该方法能自动识别含有突变段的纳米孔数据。nano-UDL 方法利用深度自动编码器从纳米孔数据中提取特征，然后通过均值漂移（MeanShift）聚类算法对特征数据进行分类。此外，该方法还联合优化了聚类损失和重构损失，从而提取最优的特征用于聚类。实验结果表明，nano-UDL 方法在 O^6-CMG 数据集上具有较高的识别精度，能准确识别出所有包含 O^6-CMG 的序列段。为了进一步验证 nano-UDL 方法的鲁棒性，本文进行了超参数敏感性验证和消融实验。利用 nano-UDL 方法分析纳米孔数据不但可以有效降低人工分析数据带来的额外成本，而且对包括基因组测序在内的诸多生物研究具有重要意义。

【关键词】 甲基鸟嘌呤；纳米孔测序；DNA 损伤；胃肠道肿瘤；深度学习；无监督学习

Unsupervised deep learning for identifying the O^6-carboxymethyl guanine by nanopore sequencing

GUAN Xiaoyu[1], WANG Yu[2,3], ZHANG Jinyue[2,3], SHAO Wei[1], HUANG Shuo[2,3], ZHANG Daoqiang[1]

1. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, P. R. China

2. State Key Laboratory of Analytical Chemistry for Life Sciences, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, P. R. China

3. Chemistry and Biomedicine Innovation Center, Nanjing University, Nanjing 210023, P. R. China

Corresponding author: ZHANG Daoqiang, Email: dqzhang@nuaa.edu.cn; HUANG Shuo, Email: shuo.huang@nju.edu.cn

【Abstract】 O^6-carboxymethyl guanine(O^6-CMG) is a highly mutagenic alkylation product of DNA that causes gastrointestinal cancer in organisms. Existing studies used mutant *Mycobacterium smegmatis* porin A (MspA) nanopore assisted by Phi29 DNA polymerase to localize it. Recently, machine learning technology has been widely used in the analysis of nanopore sequencing data. But the machine learning always need a large number of data labels that have brought extra work burden to researchers, which greatly affects its practicability. Accordingly, this paper proposes a nano-Unsupervised-Deep-Learning method (nano-UDL) based on an unsupervised clustering algorithm to identify methylation events in nanopore data automatically. Specially, nano-UDL first uses the deep AutoEncoder to extract features from the nanopore dataset and then applies the MeanShift clustering algorithm to classify data. Besides, nano-UDL can extract the optimal features for clustering by joint optimizing the clustering loss and reconstruction loss. Experimental results demonstrate that nano-UDL has relatively accurate recognition accuracy on the O^6-CMG dataset and can accurately identify all sequence segments containing O^6-CMG. In order to further verify the robustness of nano-UDL,

DOI：10.7507/1001-5515.202104068
基金项目：国家自然科学基金（61861130366，61876082，61732006，62136004）；国家重点研发计划（2018YFC2001600，2018YFC2001602）
通信作者：张道强，Email：dqzhang@nuaa.edu.cn；黄硕，Email：shuo.huang@nju.edu.cn

hyperparameter sensitivity verification and ablation experiments were carried out in this paper. Using machine learning to analyze nanopore data can effectively reduce the additional cost of manual data analysis, which is significant for many biological studies, including genome sequencing.

【Key words】 Carboxymethyl guanine; Nanopore sequencing; DNA lesion; Gastrointestinal cancer; Deep Learning; Unsupervised learning

引言

基因组测序有助于提高人类对疾病的认识，如何辨别与人类疾病相关的遗传危险因素至关重要。O^6-甲基鸟嘌呤（O^6-carboxymethyl guanine，O^6-CMG）在一定情况下会触发突变，引起 DNA 复制时的编码错误，这种错误往往与胃肠道肿瘤相关。近年来，纳米孔测序技术已经成为了一种新兴的大分子感知识别技术，可用于 DNA 测序[1-3]。在之前的研究工作中，Wang 等[4]使用枯草芽孢杆菌噬菌体 Phi29 DNA 多聚酶（Phi29 DNA polymerase）辅助的耻垢分枝杆菌膜蛋白（mycobacterium smegmatis porin A，MspA）纳米孔作为识别 O^6-CMG 的测序工具。纳米孔装置由两个充满液体的储层组成，由纳米级的孔道相连接[5]。储层上下两侧的带电粒子产生正负电位差，带电粒子由于电位差的驱动而通过中间的孔道，因而产生了可以反映分子特性的电流信号[6-10]。

由于产生的这种电流信号主要是时间序列，这正是机器学习方法擅于处理的数据形式，因而最近越来越多的研究将机器学习方法用于处理纳米孔数据。例如，Carral 等[11]提出了一个无监督机器学习模型，可以通过分析基于二维纳米孔原始数据来识别不同的分子事件。Farshad 等[12]设计了一个具有列文伯格-马夸尔特（Levenberg-Marquardt，LM）传递函数的两层神经网络来分析氮化硅（Si_3N_4）纳米孔数据。对于纳米孔甲基化数据，Jia 等[13]训练了自适应增强（adaptive boosting，AdaBoost）分类器，从牛津纳米孔技术（oxford nanopore technology，ONT）原始数据中检测 DNA 甲基化事件。Schreiber 等[14]提出了一个隐马尔可夫模型用来分割和整合纳米孔数据，实验结果表明，与以往的人工操作方法相比，该方法能显著降低错误率。此外，Liu 等[15]训练了支持向量机（support vector machine，SVM）分类器可以实现高精度检测纳米孔 N6-甲基腺嘌呤（N6-methyladenosine，m6A）RNA 修饰。Ni 等[16]提出了深度信号（DeepSignal）方法，该方法可以检测基因组位点的 5-甲基胞嘧啶（5-methylcytosine，5mC）和 m6A 甲基化信号，在基因组水平的精度（accuracy，ACC）为 0.9。Simpson 等[10]训练了一个隐马尔可夫模型来区分 5mC 和非 5mC。Stoiber 等[17]提出了一种新的算法纳米源（Nanoraw），该算法可以有效识别 4-甲基胞嘧啶（4-methylcytosine，4mC）、5mC 和 m6A 三种不同的甲基化标记。

但是，上述大多数研究都是基于监督学习框架展开[18-20]，而监督学习需要大量精确标注的样本用于训练模型，这就需要领域内的专家耗费大量的时间和精力对样本进行精确标注，因此将监督学习应用到纳米孔领域成本很高[21]。为此，Xu 等[22]提出了一个方案，利用无监督聚类方法可以找出纳米孔数据中的潜在关系，由于其学习过程是不需要任何标签的，这将大大减少专家的工作量。基于上述研究，本文提出了一种纳米无监督深度学习方法（nano-unsupervised-deep-learning，nano-UDL）用于纳米孔数据分析，并在 O^6-CMG 纳米孔数据集上验证了方法的有效性。本文提出的 nano-UDL 方法可以自动识别突变序列，且无需人工标注标签。nano-UDL 方法首先采用自动编码器从原始纳米孔数据中提取序列特征；然后，对提取的特征进行聚类；最后，实现了对 O^6-CMG 的高精度识别。本文拟定的对比方法有 K 均值聚类（K-Means）[23]、基于密度的噪声应用空间聚类（density-based spatial clustering of applications with noise，DBSCAN）[24]、均值漂移（MeanShift）[25]、凝聚层次聚类（Agglomerative）[26]、谱聚类（SpectralClustering）[27]等。期望通过与上述方法进行对比研究，能够证实本文方法的有效性，为今后纳米孔突变检测研究奠定理论研究基础，为更多类似的生物信息相关研究问题提供研究思路。

本文的主要创新思路如下：

（1）nano-UDL 方法是一种无监督学习方法，使用深度自动编码器提取特征[28-29]，随后使用 MeanShift 聚类算法对提取的特征进行分类。

（2）对于 O^6-CMG 纳米孔数据集，本文拟通过超参数敏感性验证和消融实验证实 nano-UDL 方法的稳定性，并通过对比实验的结果验证了 nano-UDL 方法可以达到提升精度的目的。

1 方法

nano-UDL 方法用于精确定位 O⁶-CMG 纳米孔数据流程示意图如图 1 所示，整体的流程分为以下三个阶段：

（1）首先，对纳米孔采集的原始序列执行分割算法，切割出若干子序列，如图 1（a）所示；

（2）其次，nano-UDL 方法模型的构建阶段，包含了自动编码器提取特征用于聚类模型 MeanShift 聚类，然后联合优化聚类损失与重构损失以实现自动编码器参数调优，如图 1（b）所示；

（3）最后，将调优后的 nano-UDL 模型用于 O⁶-CMG 数据集聚类，分别形成了包含 O⁶-CMG 位点的序列簇（定义为类别 ***A***）和不包含 O⁶-CMG 位点的序列簇（定义为类别 ***B***）。只需对类别 ***A*** 中的序列进行计算即可实现对 O⁶-CMG 位点的精准识别，如图 1（c）所示。

1.1 数据生成

首先需要做的是切割原始的纳米孔数据（本文所用数据来自 Wang 等[4]的工作，得到了作者的授权，如需引用此数据请联系原作者）。O⁶-CMG 的原始数据是长序列，包含所有突变序列和非突变序列。测序波形往往开始于一个非常高的位置，这个位置所对应的电流称为开孔电流。当样品通过纳米孔时，电流下降。当样品完全通过时，电流上升到开孔电流位置。在此期间产生的电流就是一个有效序列。换句话说，每一个事件都可以通过开孔电流的幅值来区分，低于开孔电流的序列段就是有效序列。每个小时间序列段就反映了分子的具体特性。O⁶-CMG 数据的具体的分割图如图 2 所示，图 2 中横坐标为时间，单位是 s，纵坐标为电流，单位是 pA，实体三角为 O⁶-CMG 位点。本文基于切割信号基线对长原始序列进行切割，生成 1 010 个子序列，从而形成了 O⁶-CMG 纳米孔数据集，其中包括存在突变段序列 55 个（类别 ***A***），非突变等噪声序列 955 个（类别 ***B***）。

1.2 nano-UDL 方法的构建

nano-UDL 方法的思路是自编码器提取特征之

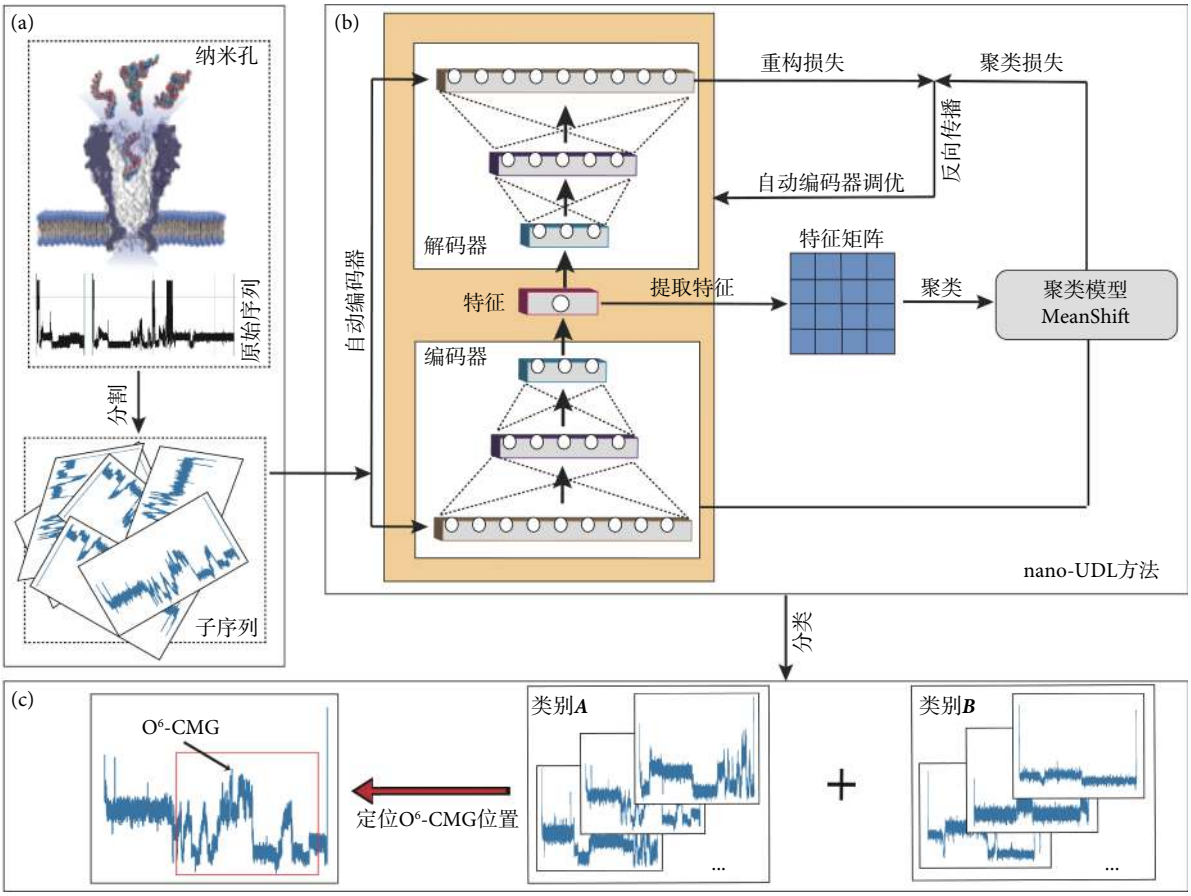

图 1 nano-UDL 用于精确定位 O⁶-CMG 纳米孔数据示意图 （a）数据生成过程；（b）nano-UDL 方法的构建过程；（c）对数据进行分类并精准定位 O⁶-CMG 位点

Fig.1 Diagram of nano-UDL for accurate location of O⁶-CMG nanopore data (a)The process of the data generate; (b) The construction process of nano-UDL method; (c) The data were classified and the O⁶-CMG locus are accurately located

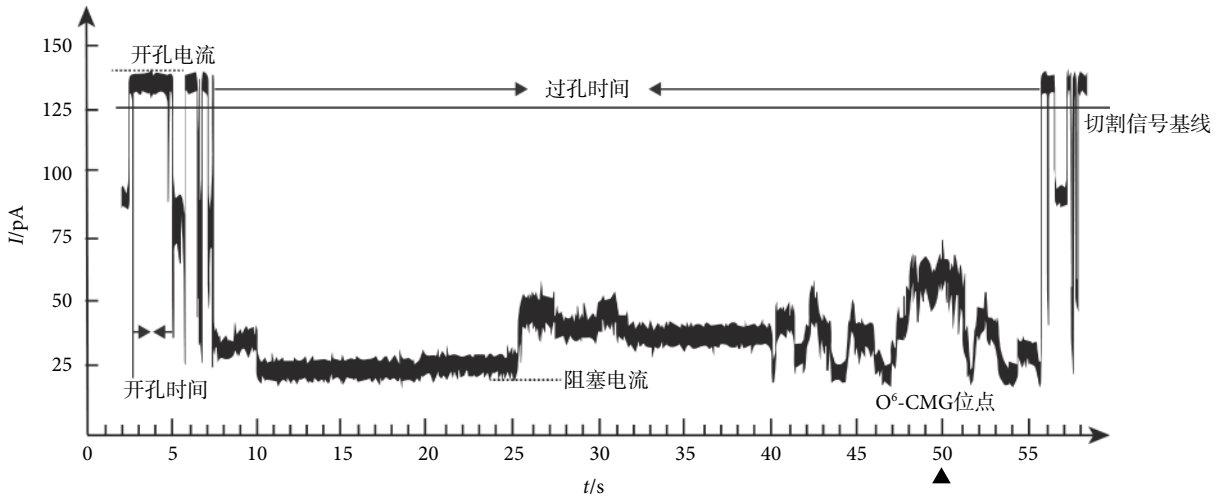

图 2　分割数据示意图

Fig.2　Schematic diagram of data segmentation

后用聚类算法进行聚类。核心功能是聚类，聚类问题可以视为将 n 个点的$\{x_i \in \boldsymbol{X}\}$被聚成 k 个簇，并且每个簇被表示成为μ_j, j=1, ⋯, k。

非线性映射 $f_\theta : \boldsymbol{X} \to \boldsymbol{Z}$ 应用于数据变换，而不是直接在数据空间 $\boldsymbol{X}$ 中进行聚类，其中 θ 是可学习的参数，$\boldsymbol{Z}$ 是潜在特征空间。

通过同时学习特征空间 $\boldsymbol{Z}$ 中的 k 个聚类中心 $\{\mu_j \in \boldsymbol{Z}\}_{j=1}^{k}$ 和将数据点映射到 $\boldsymbol{Z}$ 的深度神经网络（deep neural network，DNN）参数来对数据进行聚类。nano-UDL 方法有两个阶段：

（1）对深度自动编码器进行参数初始化，提取特征用于聚类模型聚类，给定初始参数为 θ 和 $\{\mu_j \in \boldsymbol{Z}\}_{j=1}^{k}$。本文应用 MeanShift 聚类算法对 O⁶-CMG 数据进行聚类[25]。

（2）参数优化，其中聚类损失以最小化库尔贝克·莱布勒（Kullback-Leibler，KL）散度逼近原始分布[30]。解码器的重构损失用均方误差（mean squared error，MSE）来度量。联合优化聚类损失和重构损失实现对参数的优化。

1.2.1　*自动编码器提取特征*　本文选择堆叠式自动编码器来初始化网络参数 θ 和簇质心$\{\mu_j\}$，因为最近的研究表明，它们的表现要明显优于其他自动编码器[28-29, 31-33]，已经普遍应用于诸多真实场景。

堆叠式自动编码器网络的运行是逐层进行的，每一层都经过训练，可以在随机损坏后重建上一层的输出[29]。自动编码器是两层神经网络，如式（1）所示：

$$y = g_2(W_2 \cdot f(g_1(W_1 \cdot f(x) + b_1)) + b_2) \tag{1}$$

其中，x 是模型输入，y 是模型输出，$f(\cdot)$ 是随机映射[34]，它随机将一部分输入的维度更改为 0，g_1 是编码器层的激活函数，g_2 是解码器层的激活函数，$\theta = \{W_1, b_1, W_2, b_2\}$是网络参数。为了训练模型，将损失函数设置为最小二乘损失 $\|x - y\|_2^2$，将其最小化以实现收敛。当训练完一层后，将当前层输出作为下一层的输入，以训练下一层。在本文中所有采用的编码器/解码器网络中都有整流线性单位（rectified linear units，ReLU）[35]。进行逐层训练之后，将所有编码器层与所有解码器层连接在一起，反向顺序训练以构建深度自动编码器，以最大程度地减少重构损失来对其进行微调。为了初始化聚类中心，把数据输入到初始化的自动编码器中以获取嵌入的初始数据点，并在特征空间 $\boldsymbol{Z}$ 中执行 MeanShift 聚类算法以获得 k 个初始质心 $\{\mu_j \in \boldsymbol{Z}\}_{j=1}^{k}$。

1.2.2　*参数优化*　给定初始估计的非线性映射 f_θ 和初始簇$\{\mu_j \in \boldsymbol{Z}\}_{j=1}^{k}$，为了提高聚类性能，首先通过 t 分布来计算嵌入点与聚类质心之间的软分配。接下来更新非线性映射 f_θ，并通过最小化软分配分布和目标分布之间的 KL 散度来微调聚类中心。所有过程都是迭代的，直到收敛到最小值。用 t 分布作为核心来度量嵌入点 z_i 与聚类中心μ_j 相似性，如式（2）所示：

$$q_{ij} = \frac{\left(1 + \left\|z_i - \mu_j\right\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'} \left(1 + \left\|z_i - \mu_{j'}\right\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}} \tag{2}$$

其中，$z_i = f_\theta(x_i) \in \boldsymbol{Z}$ 对应嵌入后的 $x_i \in \boldsymbol{X}$，α 为学生 t 分布的自由度，q_{ij} 可以解释为将样本 i 分配到 j 簇的概率（即软分配），j' 是聚类形成簇的标

号，$\mu_{j'}$ 是 j' 所对应的聚类中心。由于在无监督设置下对验证集执行交叉验证是不合理的，所以在所有实验中设置 $\alpha=1$。

本文所建立的模型是通过匹配目标分布的软分配来训练的。为此，将 KL 散度作为目标损失 L_c，散度计算的对象是软分配 q_{ij} 与辅助分布 p_{ij}，如式（3）所示：

$$L_c=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{3}$$

其中，如何选择目标分布 P 是决定方法性能的关键，目标就是让分布更加靠近聚类中心，所以本文需要目标分布 P 尽量接近于分布 Q，以期待最终样本有着更好的聚类结果。由于 q_{ij} 是软分配，使用软概率目标更自然、更灵活。在实验中，首先将 q_{ij} 提高到 2 次方，然后通过每一簇的概率归一化来计算 p_{ij}，如式（4）所示：

$$p_{ij}=\frac{q_{ij}^2/f_j}{\sum_{j'}q_{ij'}{}^2/f_{j'}} \tag{4}$$

其中，j' 是聚类形成簇的标号，$q_{ij'}$ 是对应的针对于 j' 的软分配，$f_j=\sum_i q_{ij}$ 是针对于 q_{ij} 软聚类概率，同理 $f_{j'}=\sum_i q_{ij'}$ 是针对于 $q_{ij'}$ 软聚类概率。

解码器重构损失 L_r 用 MSE 来度量，如式（5）所示：

$$L_r=\sum_{i=0}^{n}\|x_i-g_{W'}(z_i)\|_2^2 \tag{5}$$

其中，$z_i=f_W(x_i)$，x_i 是输入，f_W 和 $g_{W'}$ 分别为编码器映射和解码器映射。自动编码器可以保留数据生成分布的局部结构[36-37]。本文将聚类损失和重构损失合并为 nano-UDL 模型的损失函数 L 如式（6）所示：

$$L=L_r+\gamma L_c \tag{6}$$

其中，γ 是调节两个损失的超参数调节因子，在这种情况下，使用整合后的损失不会对嵌入空间造成破坏。在本文 2.6 节将讨论 γ 的取值。

为了计算最优损失函数，本文使用小批量随机梯度下降（stochastic gradient decent，SGD）来优化聚类损失和重构损失。特别地，有两种参数需要优化：自动编码器的权值和聚类中心。L_c 相对于嵌入点 z_i 和聚类中心 μ_j 的梯度如式（7）~式（8）所示：

$$\frac{\partial L_r}{\partial z_i}=2\sum_{j=1}^{K}\left(1+\left\|z_i-\mu_j\right\|^2\right)^{-1}(p_{ij}-q_{ij})(z_i-\mu_j) \tag{7}$$

$$\frac{\partial L_c}{\partial \mu_j}=2\sum_{j=1}^{n}\left(1+\left\|z_i-\mu_j\right\|^2\right)^{-1}(q_{ij}-p_{ij})(z_i-\mu_j) \tag{8}$$

其中，K 是最大的聚类簇数，n 为样本总数，则对于小批量样本 m 和学习率 λ，更新后的 μ_j 如式（9）所示：

$$\mu_j=\mu_j-\frac{\lambda}{m}\sum_{i=1}^{m}\frac{\partial L_c}{\partial \mu_j} \tag{9}$$

解码器的权重 W' 由下式进行更新，如式（10）所示：

$$W'=W'-\frac{\lambda}{m}\sum_{i=1}^{m}\frac{\partial L_r}{\partial W'} \tag{10}$$

因此，编码器的权重 W 通过下式进行更新，如式（11）所示：

$$W=W-\frac{\lambda}{m}\sum_{i=1}^{m}\left(\frac{\partial L_r}{\partial W}+\gamma\frac{\partial L_c}{\partial W}\right) \tag{11}$$

1.3 O^6-CMG 位点的定位

上述策略通过自动编码器和聚类算法的联合调优，构建了 nano-UDL 模型。接下来则需要 nano-UDL 模型对 O^6-CMG 数据集进行聚类以实现对包含突变位点序列的筛选。由于 O^6-CMG 数据集包含了两个类别（类别 ***A*** 和类别 ***B***），因此 O^6-CMG 数据集被聚类成为两个簇分别对应类别 ***A*** 和类别 ***B***，如图 1（c）所示。

nano-UDL 方法初步实现了对 O^6-CMG 子序列的分类，而为了精确定位子序列中 O^6-CMG 的位置，进一步需要对分类好的类别 ***A*** 中的序列执行分段计算。取类别 ***A*** 所在簇的其中一个序列 ***S***，将其分成 n 段等长小段序列$[s_1, s_2, \cdots, s_n]$，对这 n 个小段计算幅值变化率，取幅值变化最快的段即为 O^6-CMG 位点所在的位置。具体计算方法如式（12）所示：

$$loc=\text{argmax}\left(grad(s_1), grad(s_2), \cdots, grad(s_n)\right) \tag{12}$$

其中，$grad(\cdot)$ 是计算幅值变化率函数，本文统一取 $n=100$，通过上式计算最大的梯度所在的段即为当前序列 ***S*** 的 O^6-CMG 位点（loc）所在的位置。

通过以上三个步骤（数据生成、nano-UDL 方法的构建、O^6-CMG 位点的定位）可以实现在不需要人工干预的情况下，完成对 O^6-CMG 数据集中的 O^6-CMG 突变位点的精准定位。

2 实验

本文在 O^6-CMG 数据集上评估 nano-UDL 方

法的性能，并与其他传统的聚类算法进行比较。在本节中，不仅通过四个评价指标来展示 nano-UDL 方法的性能，还进行超参数敏感度验证和消融实验，以此验证模型的鲁棒性。

2.1　评价指标

评价指标是评价算法性能的指标。由于本文主要探讨的是无监督学习，因此本文给出了四种标准的无监督评价指标，用于与其他算法进行评估和比较。基于数据集的类型，聚类簇的数量设置为 2，本文用无监督聚类精度来评估性能，即准确率 (accuracy，ACC)，其计算如式（13）所示：

$$\mathrm{ACC}=\max_{m}\frac{\sum_{i=1}^{n}1\{y_i=m(c_i)\}}{n} \tag{13}$$

其中，y_i 为真实标签，c_i 为算法产生的聚类分配，m 为聚类与标签之间可能的映射，n 为样本总数，$1\{\cdot\}$ 是非 0 即 1 判断器，如果映射后的样本标签和真实标签一致则为 1，否则为 0。

归一化互信息（normalized mutual information，NMI）（符号为：NMI）用以计算来自同一数据集的两个标签之间相似度的归一化度量，定义如式（14）所示：

$$NMI=\frac{I(l;c)}{\frac{1}{2}(H(l)+H(c))} \tag{14}$$

其中，$I(l;c)$为真实标签 l 与预测聚类标签 c 之间的互信息，H 为其信息熵。NMI 的结果不因簇（类）的排列而改变，它们被归一化到[0, 1]，0 表示不相关，1 表示完全相关。

另外一种评价指标是调整后的兰德指数（adjusted rand index，ARI）[38]，它经常用于聚类验证，因为它是两个类别之间的一致性的度量。

卡帕（Kappa）系数是新的聚类性能度量指标[39]。Kappa 系数是用于一致性检验的指标，也可以用来衡量分级效果。因为对于分类问题来说，所谓一致性就是模型预测结果是否与实际分类结果一致。Kappa 系数的计算是基于混淆矩阵的，混淆矩阵的范围在−1 ~ 1 之间，通常大于 0。

基于混淆矩阵的 Kappa 系数计算公式如式（15）所示：

$$K=\frac{p_o-p_e}{1-p_e} \tag{15}$$

其中，K 表示为 Kappa 系数；p_o 为观测到的符合率，即 ACC；p_e 表示随机匹配率，即对应于所有类别的“实际数量与预测数量的乘积”之和除以“样本总数的平方”。

2.2　实验细节

一般监督学习使用验证集上的交叉验证来确定超参数，而非监督学习则不需要设置验证集，因此本文使用默认超参数，尽量避免调优。本文将 O^6-CMG 数据集的网络尺寸依次设置为 256、256、512 和 4，所有层都是全连接。

在预训练阶段，从均值为 0 和标准差为 0.01 的高斯分布中选取随机数作为初始值。通过 1 000 次迭代对每个层进行预训练。为了对整个自动编码器进行参数调优，网络执行 2 000 次的迭代。全部层都进行预训练和调优，批次大小设置为 256，开始学习率设置为 0.1，每 200 次迭代衰减 10%。参数 γ 将在第 2.6 节中进行讨论。通过设置上述参数，使重构损失达到较好的效果。

为了初始化聚类中心，执行 Meanshift 聚类算法 20 次，然后选择最佳的聚类结果。为了使 KL 散度最小，设置学习率为 0.01，收敛阈值设置为 0.001。对于所有的基线算法，本文对参数进行调整，并选择效果最好的结果进行比较。

本文主要目标是能够最大限度地将突变段序列从整个分割后的序列中找出来，同时尽量降低噪声序列被判断为突变序列的情况，所以在最终的评价指标中，选择使用最基本的二分类指标 ACC 作为初步评价指标，进一步增加了包括 NMI、ARI、Kappa 三种应用于衡量聚类效果的度量指标。

2.3　实验结果

本文对不同的算法的性能进行了定量和定性的评估，如图 3 所示，其中前 5 种对比方法的输入特征是通过人工手动获取，提取的特征包括诸如均值、方差等统计信息特征。对比方法的输入特征是相同的，而本文所提出的 nano-UDL 方法输入的特征是通过如图 1 所示的自动编码器优化获取。通过观察发现，在 NMI、ARI、Kappa、ACC 等所有评价指标上，nano-UDL 方法均明显高于其他方法，并在 O^6-CMG 数据集上表现出良好的聚类性能。这是因为 nano-UDL 方法使用多层自动编码器作为特征提取器，能够捕捉到非深度学习模型无法获得的局部特征。此外，图 3 中前 5 种方法都是非深度学习模型，nano-UDL 方法是基于深度学习模型的。由此可以观察到基于深度学习模型的方法明显比非深度学习的方法能发挥更好的性能。

自动编码器和不同聚类算法结合的实验结果对比如图 4 所示。图 4 中前 4 个柱状图都是自动编

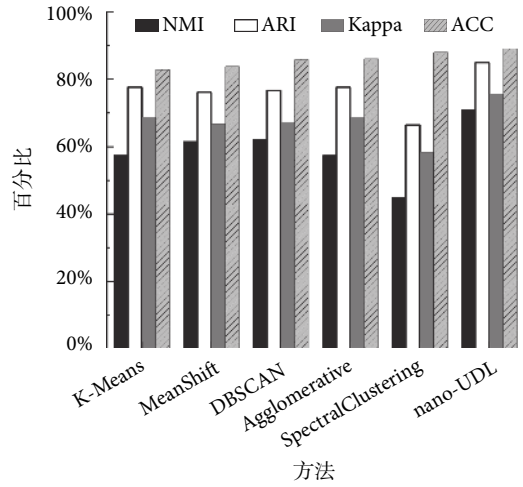

图 3 O^6-CMG 纳米孔数据集聚类性能比较

Fig.3 Comparison of clustering performance on O^6-CMG nanopore datasets

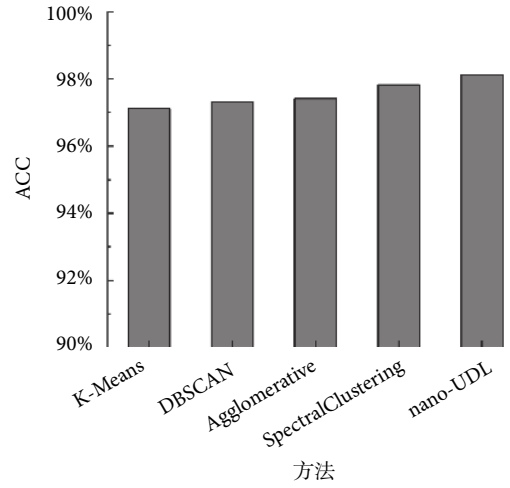

图 4 自动编码器和不同聚类方法的组合的聚类精度

Fig.4 The clustering accuracy for the combination of autoencoder and different clustering methods

码器与对应聚类算法结合后的结果，结果表明，nano-UDL 方法优于其他组合。本文选用的 MeanShift 聚类算法针对于 O^6-CMG 数据集有着良好的聚类效果，其得益于 O^6-CMG 数据在通过自动编码器特征提取后的特征分布更适合这种基于密度类型的聚类方法。

2.4 超参数敏感性

为了研究 nano-UDL 方法在 O^6-CMG 数据集上的参数敏感性，本文对不同参数下的 ACC 进行柱状图直观对比，如图 5 所示。结果表明，nano-UDL 方法在大多数参数组合下都保持了良好的结果，并且相对稳定。进一步地，对其他几种度量指标也同样做了类似的实验，实验结果同图 5 类似，同样证实了 nano-UDL 稳定性。

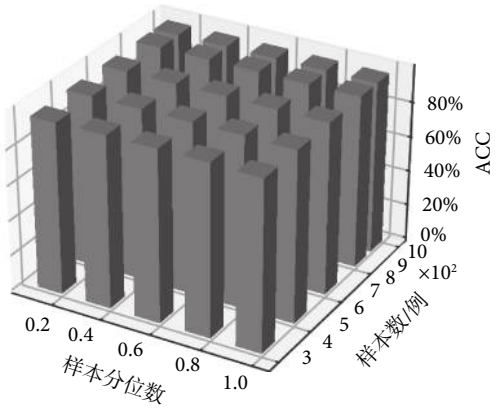

图 5 超参数选择的实验结果

Fig.5 Experimental result of the hyperparameter selection

2.5 消融实验

本文方法所采用的损失函数是聚类损失和解码器重构损失的线性加权和。因此本文主要是通过对聚类损失和解码器重构损失进行的联合优化，使 nano-UDL 方法的自动编码器模型得到最优的聚类结果。为验证重构损失添加之前和之后对实验结果的影响，本文执行了一个消融实验。其中在“无重构损失”的情况下，ACC 为 98.12%，聚类时间为 1.735 s，相对而言在“有重构损失”的情况下 ACC 为 98.32%，聚类时间为 2.228 s。实验结果表示，“有重构损失”比“无重构损失”发挥出了更好的性能。对于聚类时间而言，“有重构损失”聚类时间略长于“无重构损失”，原因是计算解码器重构损失需要额外时间。总的来说，将聚类损失和重构损失相结合时 nano-UDL 方法性能最好。

2.6 参数 γ 的确定

为了解如式（6）所示的聚类损失参数 γ 对 nano-UDL 方法性能的影响，本文在 [0.01, 100] 范围中采样，选取 13 个值。优化器设置为随机梯度下降法（stochastic gradient descent, SGD），动量 β=0.9，学习速率 λ 设为 0.1。结果如图 6 所示，可以观察到当 γ 值过小时，聚类损失项的积极作用被消除，而值过大则会导致潜在特征空间的扭曲从而影响性能，其中横坐标值做了对数拉伸处理，以便横坐标可以等比例显示。当 γ 接近 0.5 时，聚类结果可以达到最优。因此，本文在所有的实验中设置 γ=0.5。

2.7 与监督方法对比

在本节中，将 nano-UDL 方法与其他监督学习算法进行比较，验证的算法有分类回归树[40]（classification and regression trees, CART）、SVM[41]、AdaBoost[42]、随机森林（random forest,

RF）[43]、K 近邻（k-nearest neighbor，KNN）和卷积神经网络（convolutional neural networks，CNN）。这些监督算法均具有不同的监督率，为此，本文在常规监督学习算法上使用 0.1 ~ 0.9 比例的训练集用于模型训练。除此之外，本文利用 CNN 模型来验证传统深度学习模型的性能。实验结果如图 7 所示，随着监督率的增加，模型的分类效果也在逐步提高。虽然在完全监督情况下，nano-UDL 方法不如某些全监督算法。但是，在没有监督或者监督信息很少的情况下，本文所提出的 nano-UDL 方法具有显著的优势。除 CNN 和 nano-UDL 方法这两种深度学习模型外，其余模型输入采用相同的特征，从图 7 中可以看出 CART、SVM、AdaBoost 与 KNN 在全监督情况下不如 nano-UDL 方法。也就是说，手工提取的特征用上述这四种模型训练并不能达到预期的效果，但例外的是 RF 模型在同样的实验设置下却可以表现得较好。这也从另一个侧面验证了深度学习相比于手工提取特征技术，可以针对性地对特定任务进行特征调整。与传统机器学习相比不同的是，在某些特征情况下，模型间可能存在较大的性能偏差。而本文提出的 nano-UDL 方法在无监督的背景条件下，也能达到部分有监督信息训练算法的性能，这也证实了无监督深度学习在纳米孔数据分析领域具有一定的应用价值。

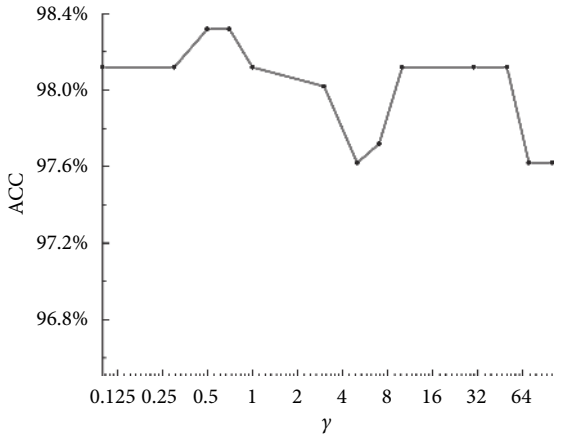

图 6　聚类参数 γ 对式（6）的影响

Fig.6　The effect of clustering coefficient γ on equation (6)

3　讨论

为了证明本文所提出框架的有效性，本文采用了消融实验、精度对比等机器学习模型验证策略对 nano-UDL 方法进行了验证。通过对实验输出结果进行分析，本文从以下两个方面对 nano-UDL 方法进行讨论：

可靠性：nano-UDL 方法有效地保证了突变序列检测的高精度和高置信度，这是构建可靠且稳定系统的必然要求。如图 3 所示的 nano-UDL 方法的结果比一些传统的聚类算法具有更高精度和置信度。从图 5 的结果可以看出超参数对聚类结果的影响很小，充分说明了本文提出模型的稳定性。通过改变监督样本的大小，从图 7 可以看出，nano-UDL 方法的结果相对较好而且相对稳定。综上所述，实验结果充分说明了 nano-UDL 方法的可靠性。

训练代价：虽然 nano-UDL 方法相比于其他聚类方法有着较好的性能，但由于这种方法策略是基于多梯度进行计算的，其中聚类损失和重构损失，增加了训练的时间复杂度。在训练阶段，本文使用少量的样本对模型进行预训练，以减少模型训练的成本。通过使用预先训练的模型初始化网络，可以

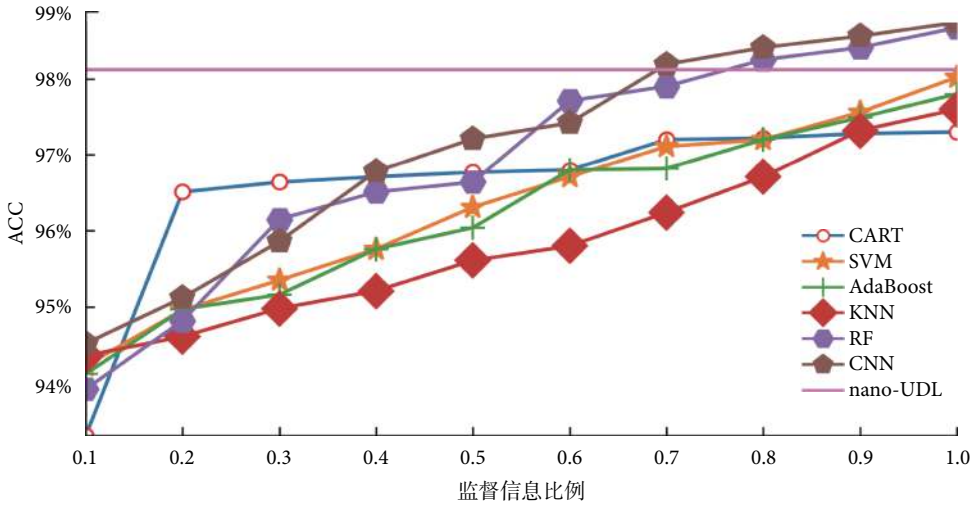

图 7　不同监督率下的不同算法精度变化曲线

Fig.7　The change of classification accuracy of different algorithms under different supervision rates

有效地保证模型在训练过程中时刻处于最佳。由于本文的训练数据集样本较少(1 010 例),所以消耗的系统内存、GPU 资源等计算资源都在有效范围内。

目前的纳米孔数据分析技术都是基于统计学习方法或有监督的机器学习算法,不能保证时间的有效性,因为其需要大量的分析计算成本。本文提出的方法是一种可以降低学习代价的无监督学习算法,该方法对纳米孔 DNA 突变的研究有较好的效果,将对今后更多纳米孔 DNA 突变的研究提供指导思路。本文提出的方法是尝试将无监督深度学习方法应用到纳米孔数据的首次探索,希望这种在生物信息学应用中利用先进的机器学习技术的思想在未来能够逐渐走向成熟。

4 结论

针对 O^6-CMG 纳米孔数据,本文提出了一种新的无监督深度聚类模型 nano-UDL,该模型将数据点聚类到一个可以通过联合优化的特征空间中,最终实现精准识别突变位点的目的。为了逼近目标分布,该方法通过最小化聚类损失和重构损失进行迭代训练。nano-UDL 方法可以看作是一种不需要任何人工干预的启发式工具,它可以用来处理原始纳米孔数据。通过实验分析,nano-UDL 方法在提高性能的同时,对超参数设置具有鲁棒性。nano-UDL 这种无监督深度学习策略可以在减少人工计算成本的情况直接检测出 O^6-CMG 突变位点。期望本文提出的方法将对未来更多类似的从 DNA 测序中精准定位突变位点的任务提供一个理论依据。

重要声明

利益冲突声明:本文全体作者均声明不存在利益冲突。

作者贡献声明:关晓宇主要负责项目主持、算法程序设计以及论文编写;王宇和张金月主要负责实验数据的采集和分析;邵伟,黄硕,张道强主要负责提供实验指导,数据分析指导,论文审阅修订。

参考文献

1 Kasianowicz J J, Brandin E, Branton D, *et al*. Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci U S A, 1996, 93(24): 13770-13773.

2 Venkatesan B M, Bashir R. Nanopore sensors for nucleic acid analysis. Nat Nanotechnol, 2011, 6(10): 615-624.

3 Ying Y L, Cao C, Long Y T. Single molecule analysis by biological nanopore sensors. Analyst, 2014, 139(16): 3826-3835.

4 Wang Y, Patil K M, Yan S, *et al*. Nanopore sequencing accurately identifies the mutagenic DNA lesion O^6—carboxymethyl guanine and reveals its behavior in replication. Angewandte Chemie International Edition, 2019, 58(25): 8432-8436.

5 Henley R Y, Ashcroft B, Farrell I, *et al*. Electrophoretic deformation of individual transfer RNA molecules reveals their identity. Nano Lett, 2016, 16(1): 138-144.

6 Smith A M, Abu-Shumays R, Akeson M, *et al*. Capture, unfolding, and detection of individual tRNA molecules using a nanopore device. Front Bioeng Biotechnol, 2015, 3: 91.

7 Zhang X, Xu X, Yang Z, *et al*. Mimicking ribosomal unfolding of RNA pseudoknot in a protein channel. J Am Chem Soc, 2015, 137(50): 15742-15752.

8 Zhang X, Zhang D, Zhao C, *et al*. Nanopore electric snapshots of an RNA tertiary folding pathway. Nat Commun, 2017, 8(1): 1458.

9 Krause M, Niazi A M, Labun K, *et al*. Tailfindr: alignment-free poly(A) length measurement for Oxford nanopore RNA and DNA sequencing. RNA, 2019, 25(10): 1229-1241.

10 Simpson J T, Workman R E, Zuzarte P C, *et al*. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods, 2017, 14(4): 407-410.

11 Carral A D, Sarap C S, Liu K, *et al*. 2D MoS2 nanopores: ionic current blockade height for clustering DNA events. 2D Mater, 2019, 6(4): 045011.

12 Farshad M, Rasaiah J C. Molecular dynamics simulation study of transverse and longitudinal ionic currents in solid-state nanopore DNA sequencing. ACS Appl Nano Mater, 2020, 3(2): 1438-1447.

13 Jia Shen, Luo Haochen, Gao Qiheng, *et al*. Detection of m6A RNA methylation in nanopore sequencing data using support vector machine//2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Suzhou: IEEE, 2019: 1-5.

14 Schreiber J, Karplus K. Analysis of nanopore data using hidden Markov models. Bioinformatics, 2015, 31(12): 1897-1903.

15 Liu H, Begik O, Lucas M C, *et al*. Accurate detection of m(6)A RNA modifications in native RNA sequences. Nat Commun, 2019, 10(1): 4079.

16 Ni P, Huang N, Zhang Z, *et al*. DeepSignal: detecting DNA methylation state from nanopore sequencing reads using deep-learning. Bioinformatics, 2019, 35(22): 4586-4595.

17 Stoiber M, Quick J, Egan R, *et al*. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. BioRxiv, 2016: 094672.

18 Alpaydin E, Bishop C M. Introduction to machine learning. MIT press, 2014.

19 Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms//Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh: ACM, 2006: 161-168.

20 Michie D, Spiegelhalter D J, Taylor C C. Machine learning, neural and statistical classification. Journal of the American Statistical Association, 1994, 91(433): 2291432.

21 Riedmiller M. Advanced supervised learning in multi-layer perceptrons-from backpropagation to adaptive learning algorithms. Computer Standards & Interfaces, 1994, 16(3): 265-278.

22 Xu R, Wunsch II D C. Clustering. IEEE Computational Intelligence Magazine, 2009, 4(3): 92-95.

23 Macqueen J. Some methods for classification and analysis of multivariate observations//Proceedings of the Fifth Berkeley

Symposium on Mathematical Statistics and Probability. Berkeley: Univ of California Press, 1967, 1(14): 281-297.

24 Ester M, Kriegel H P, Sander J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise// Proceedings of the second international conference on knowledge discovery and data mining, Oregon: ACM, 1996, 96(34): 226-231.

25 Cheng Y. Mean shift, mode seeking, and clustering. IEEE Trans Pattern Anal Mach Intell, 1995, 17(8): 790-799.

26 Kurita T. An efficient agglomerative clustering algorithm using a heap. Pattern Recognit, 1991, 24(3): 205-209.

27 Ng A Y, Jordan M I, Weiss Y. On spectral clustering: analysis and an algorithm. Advances in Neural Information Processing Systems, 2002, 2: 849-856.

28 Le Q V. Building high-level features using large scale unsupervised learning//2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver: IEEE, 2013: 8595-8598.

29 Vincent P, Larochelle H, Lajoie I, *et al*. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 2010, 11: 3371-3408.

30 Johnson D H, Sinanovic S. Symmetrizing the kullback-leibler distance. IEEE Transactions on Information Theory, 2001: 14941762.

31 Lv Yisheng, Duan Yanjie, Kang Wenwen, *et al*. Traffic flow prediction with big data: a deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(2): 865-873.

32 Bengio Y, Lamblin P, Popovici D, *et al*. Greedy layer-wise training of deep networks//Advances in Neural Information Processing Systems, Vancouver: MIT Press, 2007: 153-160.

33 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504-507.

34 Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout:a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

35 Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines//Proceedings of the 27rd international conference on machine learning, Haifa: ACM, 2010.

36 Peng X, Xiao S, Feng J, *et al*. Deep subspace clustering with sparsity prior//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York: Morgan Kaufmann, 2016: 1925-1931.

37 Lecun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444.

38 Steinley D. Properties of the Hubert-Arable adjusted rand index. Psychol Methods, 2004, 9(3): 386.

39 Fleiss J L, Cohen J, Everitt B S. Large sample standard errors of kappa and weighted kappa. Psychol Bull, 1969, 72(5): 323.

40 Timofeev R. Classification and regression trees (CART) theory and applications. Humboldt University, Berlin, 2004: 1-40.

41 Suykens J , Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett, 1999, 9(3): 293-300.

42 Freund Y, Schapire R E. Experiments with a new boosting algorithm//Proceedings of the 13rd international conference on machine learning, Bari: ACM, 1996, 96: 148-156.

43 Liaw A, Wiener M. Classification and regression by randomForest. R news, 2002, 2(3): 18-22.

收稿日期：2021-04-21　修回日期：2021-11-22
本文编辑：陈咏竹