基于自动编码器降维的 Cox 神经网络扩展模型在肺腺癌组学数据中的应用*

张永超[1]　兰　宁[1]　李　森[2]　张云飞[3]　赵晋芳[1]　罗天娥[1△]

【提　要】 **目的**　在自动编码器对肺腺癌基因表达组学数据进行降维的基础上，构建 Cox 的神经网络扩展模型，从而对肺腺癌患者预后进行预测。**方法**　首先通过两种无监督学习方法：自动编码器和主成分分析分别对肺腺癌的基因表达数据进行降维，然后构建 Cox-nnet 模型，并与 DeepSurv 模型进行比较，从中选择预测性能较好的方法来识别肺腺癌的高低危患者。**结果**　在 TCGA 与 GEO 两个数据集中，基于自动编码器降维后的 Cox-nnet 模型均有较好的一致性指数与 AUC 值，且高低预后两组患者的生存率都具有统计学差异。**结论**　自动编码器比主成分分析更适用于基因表达数据的无监督降维，且经自动编码器降维后的 Cox-nnet 模型拥有较好的预测性能，可以明显地区分肺腺癌的高低危患者，为肺腺癌的预后研究提供科学依据。

【关键词】 肺腺癌　主成分分析　自动编码器　Cox-nnet　预后预测
【中图分类号】 R195.1　　【文献标识码】 A　　**DOI**　10.11783/j.issn.1002-3674.2024.01.037

近年来我国肺癌的患病率持续增加，肺癌已成为我国社会的巨大负担[1]。在肺癌患者中，肺腺癌是一种最常见的亚型。由于癌症的异质性，肺癌患者的预后存在较大差异。因此肺癌患者的预后研究一直都很重要[2-3]。随着高通量测序技术的发展，全球产生了多种分子水平的组学数据，且大多容易获取[4]。通过分析组学数据，可以探索癌症的复杂的生物学机制。但由于组学数据具有高维、高噪、稀疏等特性，建模前常常需要对其降维或特征提取[5-6]。主成分分析（principal component analysis，PCA）是一种常用的无监督学习降维方法，但它只能对数据进行线性变换；自动编码器（autoencoders，AE）可以通过重构原始输入数据来产生新的特征达到降维目的[7]，并且能对复杂的非线性关系进行建模，近来有研究表明，经 AE 降维后的数据，在预测结直肠癌患者的生存率方面是可靠的[8]。

Cox 比例风险回归模型是常用的生存预测模型，但需要满足 PH 假定而且不适合分析高维数据。人工神经网络可用于非线性比例风险的建模，适用范围更广，遂本文采用 Cox 的神经网络扩展模型：Cox-nnet[9] 对肺腺癌患者进行预后预测。既往已有关于 Cox-nnet 模型的低维临床数据以及无降维的高维组学数据的研究应用[9-10]，本研究的目的是比较 PCA 与 AE 对肺腺癌基因表达数据的降维效果，并验证基于 AE 降维的 Cox-nnet 模型是否有更好的预测性能，从而为识别肺腺癌的高低危患者、改善患者预后提供理论依据。

* 基金项目：山西省自然科学基金（201801D121210）
1.山西医科大学卫生统计教研室（030001）
2.中国医学科学院血液病医院
3.亚利桑那州立大学
△通信作者：罗天娥，E-mail：luotiane1977@163.com

模型介绍

1.主成分分析

主成分分析（PCA）是从多个数值变量间的相互关系入手，将 X 个原始变量转化为少数几个独立的综合变量（即主成分）的方法，通常新的综合变量是原变量的线性组合。最终主成分保留的个数根据特征根和累积贡献率来确定，通常以大于 70% 为宜[11]。

2.自动编码器

自动编码器（AE）是一种无监督学习的神经网络模型，可用于高维数据的降维。网络由输入层、隐藏层和输出层组成，包括编码（encoder）和解码（decoder）两个功能（见图 1）。在编码过程中，AE 将原始的输入数据映射为一种隐式表达，从而学习输入数据所包含的信息，在解码过程中，将隐式表达映射到输出层，实现对输入数据的重构。

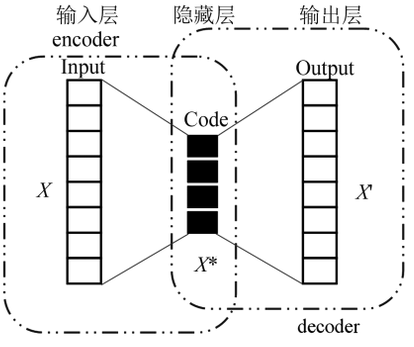


图 1　自动编码器框架

AE 的数学计算过程为：在编码部分，将输入层的高维变量 X 通过非线性激活函数转换为低维隐变量 X^*，即：

$$X^* = \sigma_1(W_1X + b_1) \quad (1)$$

式中 W_1 为编码权重矩阵，b_1 为编码偏置向量，σ_1 为编码激活函数。

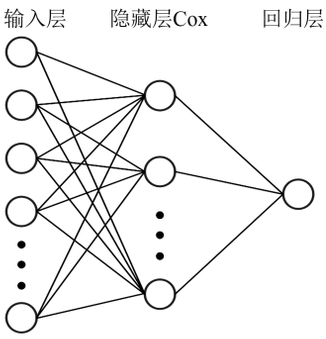


图 2 Cox-nnet 模型示意图

在解码部分，利用隐藏层的变量 X^*，对原始变量 X 进行重新构建，即：

$$X'=\sigma_2(W_2X^*+b_2) \quad (2)$$

式中 W_2 为解码权重矩阵，b_2 为解码偏置向量，σ_2 为解码激活函数。

由于 AE 的输出层是输入层的重构，所以输出层的节点数等于输入层的节点数，隐藏层是一种呈对称结构的全连接神经网络，基于深度学习的原理，可以有多层，进行多次的编码与解码，同时为了使 AE 达到降维的目的，会限制隐藏层中 X^* 的节点数，使其小于输入层的节点数。对于有多个隐藏层的 AE，隐藏层节点数在编码过程中逐步减少，在解码过程中逐步复原。AE 的目的就是为了使重构的变量 X' 与原始变量 X 之间的损失最小化，从而获得极具代表性的隐变量 X^*，这个过程是通过多次映射，对数据中的噪声进行识别和处理，并通过网络反向传播不断地迭代来实现的。

本研究中，AE 的损失函数定义为：$L=\frac{1}{2}\sum\|X'-X\|^2$，编码与解码过程都使用 Softplus 函数作为非线性激活函数[12]，其数学表达式为：$\sigma_{(x)}=\log(1+e^x)$。AE 的学习率设置为 0.001，最大迭代次数为 1000，并采用 Adam 优化算法来更新网络权重。

3.Cox-nnet

Cox-nnet 是传统 Cox 模型的神经网络扩展，它由输入层、一个全连接的隐藏层以及一个 Cox 比例风险输出层组成，在输出层根据隐藏层的激活水平进行 Cox 回归，模型的最终输出为预后指数 PI 值[9]。

该模型的风险函数为：

$$h(t|x_i)=h_0(t)\exp\theta_i$$

$$\theta_i=x_i^T\beta \quad (3)$$

损失函数定义为带有正则化的偏似然对数函数：

$$L(\beta,W)=pl(\beta,W)+\lambda(\|\beta\|_2+\|W\|_2)$$

$$pl(\beta)=\sum_{C(i)=1}(\theta_i-\log\sum_{t_i\geq t_j}\exp(\theta_j)) \quad (4)$$

Cox-nnet 中，隐藏层的节点数设定为输入层变量数的平方根，为了防止过拟合，模型中加入了 Dropout 正则化，模型的初始学习率设为 0.01，学习率衰减设为 0.9，同时运用了 Nesterov 加速梯度下降对学习速率进行优化，动量参数设置为 0.9，并通过 5 折交叉验证确定正则化参数的最优取值，模型的非线性激活函数为 tanh 函数[10]。

4.DeepSurv

DeepSurv 模型是一种深度前馈神经网络，输入层为患者的基线数据，隐藏层为全连接层，输出层是单个节点，它运用了批归一化、非线性激活函数、Dropout、Nesterov 动量、Adam 等方法对网络进行优化，进而来预测患者的协变量对其风险率的影响[13]。

5.模型评价

本研究采用一致性指数与时间依赖性 ROC 曲线下 1 年与 3 年 AUC 的值来评价模型的预测性能。一致性指数反映模型预测患者死亡时间顺序的准确程度，它的取值范围在 0.5~1 之间，越接近于 1 说明预测性能越好。时间依赖性 ROC 曲线是在传统 ROC 的基础上考虑了时间因素，因此可以绘制不同时点的 ROC 曲线。

6.软件实现

本研究所使用的 PCA、AE、Cox-nnet 与 DeepSurv 均由 Python 3.7 来实现，其中 PCA 由 sklearn 库中的 PCA 函数构建，AE 与 DeepSurv 由 tensorflow 库构建，Cox-nnet 由 theano 库构建。符号秩和检验采用 SPSS 24.0 完成。

实例分析

1.资料来源及整理

本研究采用 TCGA 数据库肺腺癌基因表达数据，该数据集通过 UCSC Xena（http://xena.ucsc.edu）平台获得。肺腺癌患者的生存时间与生存状态来源于 R 的 RTCGA.clinical 包，并去掉生存时间为 0 的患者，最终获得 390 例患者，19712 个基因特征。本研究利用 GEO 数据库中下载的 GSE72094 基因表达矩阵作为外部验证数据集，并从 easyGEO（https://easygeo.cn/）获取患者的生存时间与状态，去掉生存时间为 0 以及无生存状态的患者，数据集最终包含 398 例肺腺癌患者，22115 个基因特征。基因表达数据在降维前，先进行归一化处理，将所有的基因表达数据都缩放至（0，1）范围内。

2.实例分析结果

本研究中的基因数据集按照 8∶2 的比例随机划分为训练集与测试集，在训练集中构建模型，在测试集中进行性能评价。实验重复 20 次，通过符号秩和检验，按照 $\alpha=0.05$ 的检验水准，比较两种不同降维方式所构建的 Cox-nnet 模型的预测性能是否有统计学差异，同时为了验证 AE 联合 Cox-nnet 模型的预测能力，又将其与无监督学习降维下的 DeepSurv 模型，以及直接使用高维基因数据建立的 Cox-nnet 模型进行

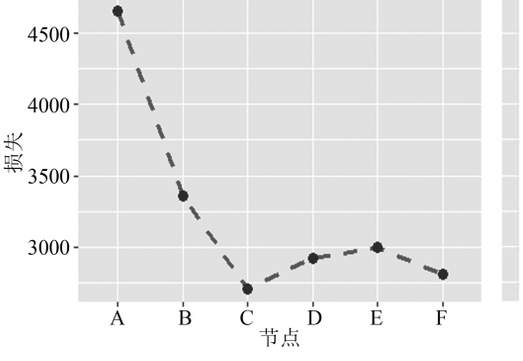
了比较。

随后选择预测性能最好的模型，并选取 20 次实验中一致性指数最大的一次对全部患者做出预后预测，按照模型输出的预后指数的中位数，将数据集分为高低预后指数两组，通过 Log-rank 检验比较两组患者的生存率。

(1) AE 隐藏层取不同的节点数所对应的损失

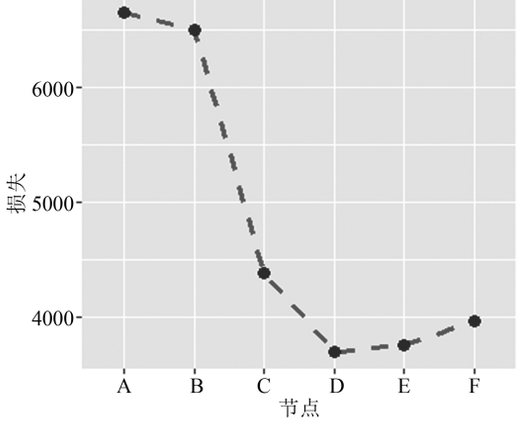
本研究中，AE 共进行了两次编码与解码，通过查阅已有文献[12,14]以及进行多次试验，不同的节点数所对应的损失见图 3 和图 4，将 TCGA 数据集隐藏层的节点数设置为(800,150,800)，即最终降维后的变量为 150 个；将 GEO 数据集隐藏层的节点数设置为(800,250,800)，即最终降维后的变量为 250 个。为了与 AE 进行比较，对于 TCGA 数据集，PCA 共选择了前 150 个主成分与前 200 个主成分，累积贡献率分别为 86%和 91%；对于 GEO 数据集，PCA 共选择了前 250 个主成分，累积贡献率为 87%，用于验证。

图 3 中 C 对应的损失最小，故隐藏层的节点数设置为(800,150,800)。



注：ABCDEF 分别表示在 TCGA 数据集中，AE 隐藏层的节点数为(600,150,600)、(600,200,600)、(800,150,800)、(800,200,800)、(1000,150,1000)、(1000,200,1000)。

图 3　TCGA 数据集不同隐藏层节点数对应的损失



注：ABCDEF 分别表示在 GEO 数据集中，AE 隐藏层的节点数为(500,250,500)、(600,250,600)、(700,250,700)(800,250,800)、(900,250,900)、(1000,250,1000)

图 4　GEO 数据集不同隐藏层节点数对应的损失

在 GEO 数据集中，AE 隐藏层的节点数选取了六种，分别对应 ABCDEF。其中 D 的损失最小，见图 4，即隐藏层的节点数设置为(800,250,800)。

(2) 不同降维方式下的模型性能比较

在 TCGA 数据集中，不同模型的符号秩和检验结果见表 1。结果表明，与 AE/PCA 联合 DeepSurv 模型、无降维的 Cox-nnet 模型以及 PCA 降维后的 Cox-nnet 模型相比，经 AE 降维后的 Cox-nnet 模型的一致性指数最高，差异具有统计学意义。同样，在 GEO 数据集中，经 AE 降维后的 Cox-nnet 模型拥有最高的一致性指数，差异有统计学意义，符号秩和检验结果见表 2。

表 1　TCGA 数据集一致性指数的比较

降维方法	$M(P_{25},P_{75})$	Z	P
AE150Cox-nnet	0.720(0.702,0.748)	-	-
PCA150Cox-nnet	0.697(0.665,0.721)	-3.059	0.002
PCA200Cox-nnet	0.700(0.690,0.724)	-2.147	0.032
AE150DeepSurv	0.696(0.654,0.710)	-2.449	0.014
PCA150DeepSurv	0.657(0.624,0.695)	-4.261	<0.001
PCA200DeepSurv	0.677(0.625,0.697)	-4.180	<0.001
Cox-nnet	0.685(0.660,0.709)	-3.287	0.001

表 2　GEO 数据集一致性指数的比较

降维方法	$M(P_{25},P_{75})$	Z	P
AE250Cox-nnet	0.701(0.682,0.718)	-	-
PCA250Cox-nnet	0.694(0.663,0.711)	-2.914	0.004
AE250DeepSurv	0.662(0.644,0.705)	-2.584	0.010
PCA250DeepSurv	0.655(0.610,0.693)	-3.341	0.001
Cox-nnet	0.673(0.654,0.704)	-2.340	0.019

不同模型的 1 年与 3 年 AUC 比较结果见表 3~表 6。在两个数据集中，经 AE 降维后的 Cox-nnet 模型拥有比其他模型更高或相近的 AUC 值。

表 3　TCGA 数据集的一年 AUC 值比较

降维方法	$M(P_{25},P_{75})$	Z	P
AE150Cox-nnet	0.797(0.732,0.817)	-	-
PCA150Cox-nnet	0.751(0.655,0.787)	-2.407	0.016
PCA200Cox-nnet	0.753(0.673,0.803)	-2.110	0.035
AE150DeepSurv	0.742(0.658,0.774)	-2.245	0.025
PCA150DeepSurv	0.704(0.646,0.748)	-3.449	0.001
PCA200DeepSurv	0.729(0.688,0.751)	-2.760	0.006
Cox-nnet	0.720(0.676,0.759)	-3.111	0.002

表 4　TCGA 数据集的三年 AUC 值比较

降维方法	$M(P_{25},P_{75})$	Z	P
AE150Cox-nnet	0.688(0.630,0.768)	-	-
PCA150Cox-nnet	0.643(0.601,0.693)	-2.164	0.030
PCA200Cox-nnet	0.658(0.613,0.749)	-1.055	0.291
AE150DeepSurv	0.633(0.538,0.739)	-2.151	0.032
PCA150DeepSurv	0.625(0.561,0.728)	-2.029	0.042
PCA200DeepSurv	0.636(0.567,0.757)	-1.650	0.099
Cox-nnet	0.684(0.639,0.736)	-1.477	0.148

表 5　GEO 数据集的一年 AUC 值比较

降维方法	$M(P_{25},P_{75})$	Z	P
AE250Cox-nnet	0.739(0.713,0.793)	-	-
PCA250Cox-nnet	0.710(0.676,0.745)	-2.394	0.017
AE250DeepSurv	0.697(0.667,0.752)	-2.191	0.028
PCA250DeepSurv	0.678(0.600,0.724)	-3.273	0.001
Cox-nnet	0.709(0.675,0.753)	-2.502	0.012

表 6 GEO 数据集的三年 AUC 值比较

降维方法	$M(P_{25}, P_{75})$	Z	P
AE250Cox-nnet	0.673(0.634,0.765)	-	-
PCA250Cox-nnet	0.700(0.633,0.762)	-0.663	0.507
AE250DeepSurv	0.658(0.604,0.709)	-1.380	0.168
PCA250DeepSurv	0.627(0.584,0.698)	-2.016	0.044
Cox-nnet	0.690(0.644,0.732)	-0.203	0.839

(3)Cox-nnet 模型运行时长

Cox-nnet 模型的运行时长见表 7。经 AE/PCA 降维后的 Cox-nnet 模型运行一次所需时间明显低于无降维的 Cox-nnet 模型。

表 7 Cox-nnet 模型运行时长

模型	时长/分钟
Cox-nnet	85
AE_Cox-nnet	20
PCA_Cox-nnet	15

(4)识别肺腺癌的高低危患者

利用 AE 降维后的 Cox-nnet 模型计算全部患者的预后指数。高预后指数与低预后指数两组患者的 Kaplan-Meier 生存曲线见图 5 和图 6。由 Log-rank 检验的结果可见,在 TCGA 数据集与 GEO 数据集中,高低预后组患者的生存率均有统计学差异(P 均<0.001),且高预后指数患者的生存率较低,低预后指数患者的生存率较高。

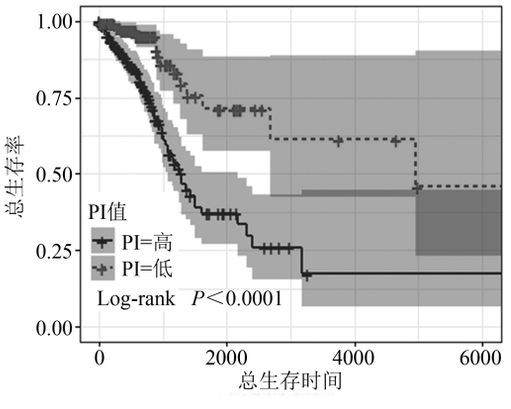


图 5 高低危患者生存曲线比较(TCGA)

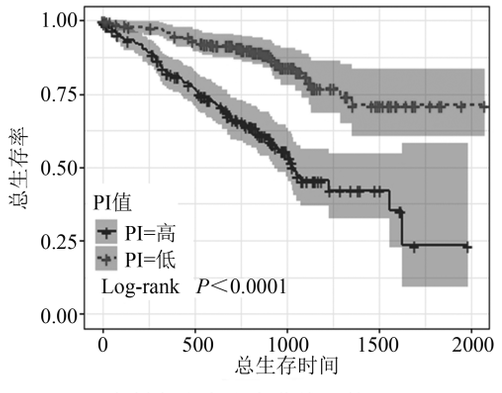


图 6 高低危患者生存曲线比较(GEO)

讨　论

已有研究表明,对于 RNA-SEQ 数据集的分析,Cox-nnet 拥有与其他方法相同甚至更好的预测精度,如 Cox 比例风险模型、随机生存森林等[9]。组学数据不仅有高维特征量和较小样本量的特性[15],而且数据中往往存在噪声和冗余,直接用于统计分析容易使模型产生过拟合[16]。本研究测试了将高维基因数据直接用于 Cox-nnet 模型进行预后分析。由于数据庞大,计算机运行较耗时,且预测性能低于联合 AE 后的模型预测性能,为了提取组学数据的有用信息,降维是一种有效的方法。因此本文对比了 AE 和 PCA 两种降维方式,利用降维后的数据构建 Cox-nnet 模型,结果表明,经 AE 降维后的 Cox-nnet 模型在两个数据集中都有较高的预测性能,可以很好地识别肺腺癌的高低危患者。同时又用降维后的数据建立了另一生存预测模型 DeepSurv,通过与 Cox-nnet 相比发现,基于 AE 降维的 Cox-nnet 模型仍有最佳的预测效果。

在本研究中,我们成功利用神经网络的自动编码器框架,从肺腺癌患者的高维基因表达数据中提取了重要特征。AE 与其他的无监督算法不同,如 PCA 是通过依次选择方差最大的原始数据点的线性组合,将观察数据转换到潜在空间的一种降维方法[17],它是在线性维度和数据近似正态分布的假设下进行的[18],但基因表达数据是不同基因间以及基因与环境间非线性相互作用的表征[19],PCA 不能捕捉非线性关系,可能不适合基因表达数据的无监督降维。而 AE 可从非线性空间中的观察值里捕获到更高水平的信息[20],换句话说,AE 可被认为是线性模型的非线性推广,可以从原始特征中掌握更复杂和高层次的关系[21],它通过对原始基因数据进行重构,提取有用信息,作为原始数据的估计。AE 不同于 PCA,PCA 降维后包含的信息量依赖于所保留的主成分个数,AE 没有此项限制,同时它对原始数据的分布不做要求,还能学习到原始数据的非线性特征,所以更适合基因表达数据的无监督降维。正如 Hinton 指出,在降低维度方面,AE 的性能要好于 PCA[7]。将 AE 降维后的特征用于 Cox-nnet 传播学习,完成对患者的生存预测,其预测效果优于 DeepSurv 模型,这可能由于 DeepSurv 模型深度增加,使得神经网络对训练集产生了过度拟合导致的。

本研究的不足之处在于只使用了一种组学数据,而单一的组学数据可能无法准确地预测出癌症患者的预后情况,结合多个组学研究的方法对于揭示疾病的复杂机制是有意义的[22],所以本研究后续会探讨 AE 整合多组学数据的生存预测性能,以及使用不同的癌症数据集予以验证。

参　考　文　献

[1] Wu F, Wang L, Zhou C. Lung cancer in China: Current and prospect. Curr Opin Oncol, 2021,33(1):40-46.

[2] 周琦,万亚平,左建宏,等.基于因果推断肺癌患者预后治疗影响因素研究.计算机技术与发展,2021,31(8):145-149.
[3] 李淼,罗天娥,郭强,等.随机生存森林模型在肺癌患者预后分析中的应用.中国卫生统计,2021,38(3):327-331.
[4] Canuel V, Rance B, Avillach P, et al. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. Brief Bioinform, 2015,16(2): 280-290.
[5] 童丹阳.基于多组学数据和临床所见的结肠癌预后分析方法研究.浙江大学,2021.
[6] Spirko-Burns L, Devarajan K. Supervised Dimension Reduction for Large-Scale "Omics" Data With Censored Survival Outcomes Under Possible Non-Proportional Hazards. IEEE/ACM Trans Comput Biol Bioinform, 2021,18(5): 2032-2044.
[7] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006,313(5786): 504-507.
[8] Song H, Ruan C, Xu Y, et al. Survival stratification for colorectal cancer via multi-omics integration using an autoencoder-based model. Exp Biol Med (Maywood), 2022,247(11): 898-909.
[9] Ching T, Zhu X, Garmire LX. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. PLoS computational biology, 2018,14(4): e1006076.
[10] 郑楚楚,张岩波,王蕾,等.基于 COX-NNET 的弥漫性大 B 细胞淋巴瘤预后预测模型.中国卫生统计,2021,38(1):119-123.
[11] Jolliffe, Ian T. Principal Component Analysis, 2nd, edn.2002.
[12] Yu B, Chen C, Qi R, et al. scGMAI: a Gaussian mixture model for clustering single-cell RNA-Seq data based on deep autoencoder. Briefings in Bioinformatics, 2021,22(4): bbaa316.
[13] Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol, 2018,18(1): 24.
[14] Seal DB, Das V, Goswami S, et al. Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration. Genomics, 2020,112(4): 2833-2841.
[15] Liu WB, Liang SN, Qin XW. A novel dimension reduction algorithm based on weighted kernel principal analysis for gene expression data. PloS one, 2021,16(10): e0258326.
[16] Wang H, van der Laan MJ. Dimension reduction with gene expression data using targeted variable importance measurement. BMC Bioinformatics, 2011,12(1): 1-12.
[17] Lin E, Mukherjee S, Kannan S. A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis. BMC Bioinformatics, 2020,21(1): 1-11.
[18] Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. Mol Aspects Med.2018,59: 114-122.
[19] Shi J, Luo Z. Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. Comput Biol Med, 2010,40(8): 723-732.
[20] Tan J, Ung M, Cheng C, et al. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. Pac Symp Biocomput, 2015,20: 132-143.
[21] Wang J, Xie X, Shi J, et al. Denoising autoencoder, a deep learning algorithm, aids the identification of a novel molecular signature of lung adenocarcinoma. Genomics Proteomics Bioinformatics,2020,18(4): 468-480.
[22] Hu W, Yang Y, Li X, et al. Multi-omics approach reveals distinct differences in left-and right-sided colon cancer. Molecular Cancer Research, 2018,16(3): 476-485.

（责任编辑:邓　妍）