稀疏自动编码器在甲状腺乳头状癌分型中的应用*

秦 阳[1] 任 月[1] 兰 宁[2] 张永超[3] 刘毅蓉[1] 赵晋芳[1] 罗天娥[1,4Δ]

【提 要】 **目的** 探索稀疏自动编码器(sparse autoencoder,SAE)在甲状腺乳头状癌(papillary thyroid carcinoma,PTC)分子分型中的应用价值。**方法** 整合PTC中分型不明确型(not otherwise specified,NOS)患者的mRNA、miRNA、DNA甲基化数据。采用自动编码器(autoencoder,AE)和SAE进行特征提取,用k-means聚类并获取对应标签,用轮廓系数(silhouette coefficient,SC)、方差比准则(calinski-harabasz index,CH)和Davies-Bouldin指数(Davies-Bouldin index,DBI)来评估模型的聚类效果。构建随机生存森林模型获取影响预后的重要基因。**结果** 在SAE-k-means的结果中,最佳聚类数为2,将PTC中的NOS患者分成两个亚型,即subtype1和subtype2。两个亚型之间的生存差异具有统计学意义(χ^2=5.36,P=0.021),subtype2预后较差。在subtype1中,*CXCL5*和*DGKI*等基因有重要的临床意义;在subtype2中,*ATP2C2*、*EHF*、*PON2*等基因有重要的临床意义。**结论** SAE在聚类和预后方面的性能都优于AE,有效识别了不同的亚型,寻找到了相对重要的影响预后的基因,为PTC分子分型提供了新思路,为治疗和预后评估提供了理论支持。

【关键词】 稀疏自动编码器 多组学整合 分子分型 随机生存森林
【中图分类号】 R195.1 【文献标志码】 A **DOI** 10.11783/j.issn.1002-3674.2026.05.009

Application of Sparse Autoencoder in the Subtype Classification of Papillary Thyroid Carcinoma

Qin Yang, Ren Yue, Lan Ning, et al (*Division of Health Statistics*, *School of Public Health*, *Shanxi Medical University*, *Taiyuan* 030001)

【**Abstract**】 **Objective** To explore the application value of sparse autoencoder (SAE) in molecular subtyping of papillary thyroid carcinoma (PTC). **Methods** To integrate the data of mRNA, miRNA, and DNA methylation of patients with not otherwise specified (NOS) in PTC. Autoencoder (AE) and SAE are used to extract features, and k-means is used to cluster and obtain corresponding labels. The clustering effect of the model is evaluated by the silhouette coefficient (SC), calinski-harabasz index (CH), and Davies-Bouldin index (DBI). A random survival forest model is constructed to obtain the important genes affecting the prognosis. **Results** In the results of SAE-k-means, the optimal number of clusters is 2, which divides NOS patients in PTC into two subtypes, subtype1, and subtype2. The survival difference between the two subtypes is statistically significant (χ^2=5.36, P=0.021), with subtype2 having a poor prognosis. In subtype1, *CXCL5* and *DGKI* have important clinical significance. In subtype2, genes such as *ATP2C2*, *EHF*, and *PON2* have important clinical significance. **Conclusion** SAE is superior to AE in clustering and prognosis, effectively identifying different subtypes and finding relatively important genes that affect prognosis, which provides a new idea for PTC molecular typing and theoretical support for treatment and prognosis evaluation.

【**Key words**】 Sparse autoencoder; Multi-omics integration; Molecular subtyping; Random survival forest

甲状腺癌(thyroid cancer,TC)发病率在全球范围内持续上升[1]。甲状腺癌最多见的亚型是甲状腺乳头状癌(papillary thyroid carcinoma,PTC),占甲状腺癌的88%[2]。PTC通常生长缓慢,但某些情况下可能侵入周边组织并形成转移灶,影响患者的预后和生存率。因此,识别甲状腺癌亚型进行个性化治疗在临床上至关重要。由于高通量测序技术的发展,可以通过整合DNA、RNA和蛋白质等多组学数据来实现精细的分子分型,有助于全面深入了解癌症的分子机制,对癌症的诊断、治疗和预后研究有重大的价值[3],从而为患者提供更加个性化的治疗方案,提高预后和生存率[4]。

自动编码器(autoencoder,AE)是一种深度学习模型,旨在使用非线性函数的组合,将输入数据压缩到低维空间中,并尽可能准确地重构原始数据,可以用于降维、特征提取、数据去噪等任务。它成功地应用于分析高维组学数据和整合异质数据。AE倾向于聚合相似路径的基因,在解释生物学功能方面有很大优势[5]。稀疏自动编码器(sparse autoencoder,SAE)是在AE的基础上,对隐藏层的神经元添加稀疏性惩罚。与AE相比,SAE产生的特征表示更加紧凑、具有更好的可解释性和鲁棒性,并且可以更好地适用于数据降维、特征提取和分类任务等应用场景。

本研究通过整合PTC患者的mRNA、miRNA、DNA甲基化和临床数据,利用AE、SAE进行特征提取,k-means进行聚类,利用轮廓系数(silhouette coefficient,SC)、方差比准则(calinski-harabasz index,CH)和Davies-Bouldin指数(Davies-Bouldin index,DBI)指标评估不同模型的聚类效果,构建随机生存森林模型分析影响PTC患

* 基金项目:山西省自然科学基金(201801D121210)
1. 山西医科大学公共卫生学院卫生统计教研室(030001)
2. 西安市疾病预防控制中心
3. 北京市丰台区疾病预防控制中心
4. 重大疾病风险评估山西省重点实验室
△通信作者:罗天娥,E-mail:luotiane1977@163.com

者预后重要的基因，来探索SAE在PTC分子分型中的应用价值，以期为癌症的精准医疗提供理论支持。

资料与方法

1. 数据来源与处理

资料来源于GDC-TCGA数据集（http://xena.ucsc.edu/），下载counts数据，共568个样本和60489个特征；miRNA数据，共573个样本和1882个特征；DNA甲基化数据，共571个样本和485578个特征；临床数据615个样本和生存数据614个样本。

通过数据清理，对每个组学数据分别进行样本ID匹配、样本去重复和基因去重复。剔除样本中缺失比例超过10%的特征，缺失值用中位数填补。标准化处理后选取具有mRNA、miRNA、DNA甲基化、生存数据和临床数据的PTC患者，共492个样本。其中组学数据包含19943个mRNAs，1562个miRNAs以及18360个DNA甲基化位点，结局为生存时间及生存状态。临床数据包括年龄、性别、病理分期。492例PTC患者中分型不明确型（not otherwise specified，NOS）351例、柱状细胞变异型（columnar-cell）37例、滤泡变异型（follicular-variant）104例。

特征筛选分两种情况，一是不进行Cox的特征筛选，直接将数据清理后的三种组学数据进行整合，共492个样本和39865个特征。二是进行Cox的特征筛选，每个组学数据分别用单因素Cox回归模型进行特征筛选（筛选标准P<0.1）。其中，包含2576个mRNAs，229个miRNAs以及2278个DNA甲基化位点，整合后共492个样本和5083个特征。将所得的整合数据作为AE和SAE的输入进行对比分析。

2. 模型方法

（1）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。该神经网络的基本结构由输入层、隐藏层和输出层组成，核心部分包括编码（encoder）和解码（decoder）两个功能。在编码过程中，AE将多种来源构成的原始高维输入数据映射为隐藏层的低维数据，即一种隐式向量表达，从而可以充分地学习输入数据所包含的信息。在解码过程中，将隐藏层中的隐式向量表达重新映射到输出层的高维数据，实现对输入数据的重构，尽可能地使输入数据可以在输出层中复现。

AE的目的就是为了使重构的变量与原始变量之间的损失最小化，从而获得极具代表性的隐变量，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。通过联合简单的神经元来创建AE架构，使得来自一层神经元的信息充当另一层的输入。以这种方式，堆叠网络架构被设计成蝴蝶形状，其中输入层的数量与输出层的数量保持相等，从而在中心产生瓶颈层。这种形式的架构迫使网络开发压缩数据表示，同时保持输入数据的关键特征。为了训练AE，使用具有均方误差（mean square error，MSE）的重建误差作为其目标函数。自动编码器架构见图1。

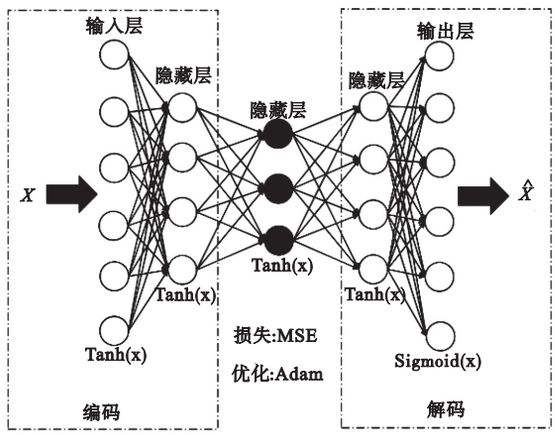


图1 自动编码器架构

（2）稀疏自动编码器

SAE是AE的一种变体，它在压缩输入数据时更加倾向于产生稀疏的特征表示，优化目标不仅包括重构误差，还包括最大化编码器中间层的稀疏性。在AE的基础上，对隐藏层的神经元添加稀疏性惩罚，即通过使隐藏层大部分的神经元输出值为0，仅保留少数神经元输出值为1，这意味着编码器会被强制让大部分输入数据在中间层对应的节点上得到较小的响应，进而使得提取到的极少数特征能够更好地表示输入数据。稀疏自动编码器架构见图2。

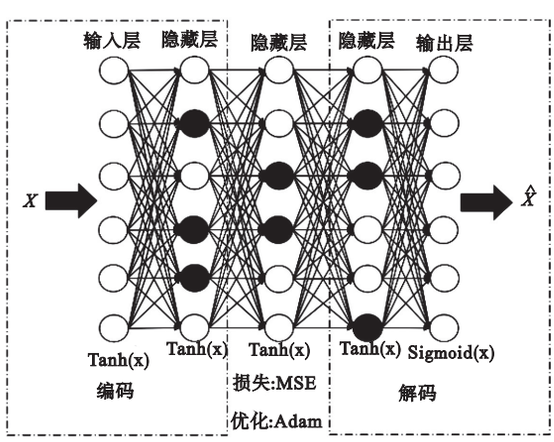


图2 稀疏自动编码器架构

在本研究中，输入层使用tanh激活函数进行激活，输出层使用sigmoid函数进行激活，并在损失函数中添加L1正则项来惩罚过度激活。使用Adam算法进行优化。AE和SAE都使用256的批次大小和100个时期进行训练。

（3）k-means聚类

通过SC和CH来确定最佳聚类数k，然后用k-means进行聚类分型，最大迭代次数默认为10。获得聚类标签之后，根据生存数据（时间和状态）绘制不同分组的生存曲线，评估所确定分组之间的生存差异，P

<0.05被认为差异具有统计学意义。

（4）随机生存森林模型

从 k-means 聚类中获得标签后，根据方差分析 F 值选择与聚类标签最相关的前 N 个基因。本文将 mRNA 的 N 值设置为 1000，甲基化为 1000，miRNA 为 500，最终将多组学数据处理成有 2500 个基因的数据。整合临床数据（年龄、性别、病理分期）、生存数据（时间、状态）、最优模型的聚类标签和方差分析处理后的多组学数据来构建随机生存森林模型[6]，通过变量重要性评分，来了解各影响因素在建立模型时的作用。

3. 评价指标

（1）聚类指标

使用 SC、CH 和 DBI 指标评价 AE 和 SAE 的 k-means 聚类效果。SC 值范围为-1~1，该值越接近 1，聚类效果越好；CH 值大于零，值越大代表类自身越紧密，类之间越分散，表示该模型的聚类效果越好；DBI 值大于零，指数越小，表示聚类效果越好。

（2）重要性评分

每个变量有一个重要性得分，即 VIMP（variable importance），该值越大就意味着此变量对于模型的准确度影响越大，重要性就越高，表明该变量对于 PTC 患者的生存结局有着重要影响。

4. 软件实现

基于 R 4.2.1 统计软件，数据清洗和生存分析采用 tidyverse 包和 survival 包；计算 SC、CH、DBI 指标以及聚类采用 cluster 包、vegan 包和 kmeans 包，随机生存森林模型的构建以及变量重要性评分采用 ranger 包；基于 Python 3.9.12 统计软件，实现 AE 和 SAE 对多组学数据的特征提取。

结　果

1. PTC 患者的生存评价

原始资料 PTC 患者包括 NOS 型、columnar-cell 型、follicular-variant 型。比较三组 PTC 患者的生存率，并绘制总体生存曲线，可知三组 PTC 患者生存差异不具有统计学意义（χ^2=3.14，P=0.208），见图 3。由于 NOS 的分型不清楚，并且生存率较低。因此，本文针对 PTC 中的 351 例 NOS 型患者，整合临床数据和多组学数据来寻找合适的分子亚型。

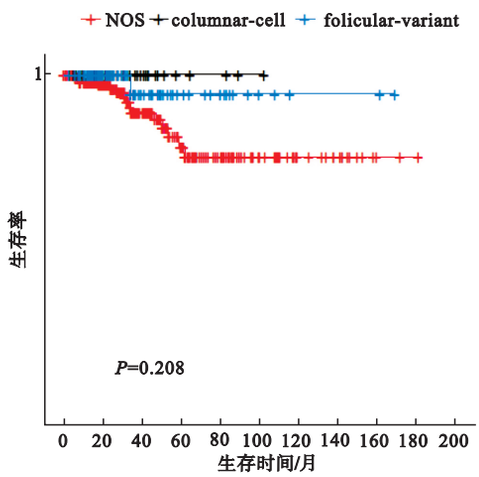


图 3　三组 PTC 患者的生存曲线图

2. AE 和 SAE 参数选择

在 AE 和 SAE 中，瓶颈层的节点选择范围为（50、100、150 和 200），其他隐藏层的节点选择范围为（500、1000、1500、2000 和 2500），L1 正则化参数选择范围为（0.01、0.001、0.0001、0.00001 和 0.000001）。最终选择损失函数最小的组合，AE 选择瓶颈层节点为 150 和其他隐藏层节点为 2500。SAE 选择瓶颈层节点为 200、其他隐藏层节点为 500 和 L1 正则化参数为 0.000001。

在 Cox-AE 和 Cox-SAE 中，其他隐藏层的节点选择范围为（2542、3050），最终 Cox-AE 选择瓶颈层节点为 200 和其他隐藏层节点为 3050。Cox-SAE 选择瓶颈层节点为 200、其他隐藏层节点为 2542 和 L1 正则化参数为 0.000001，见表 1。

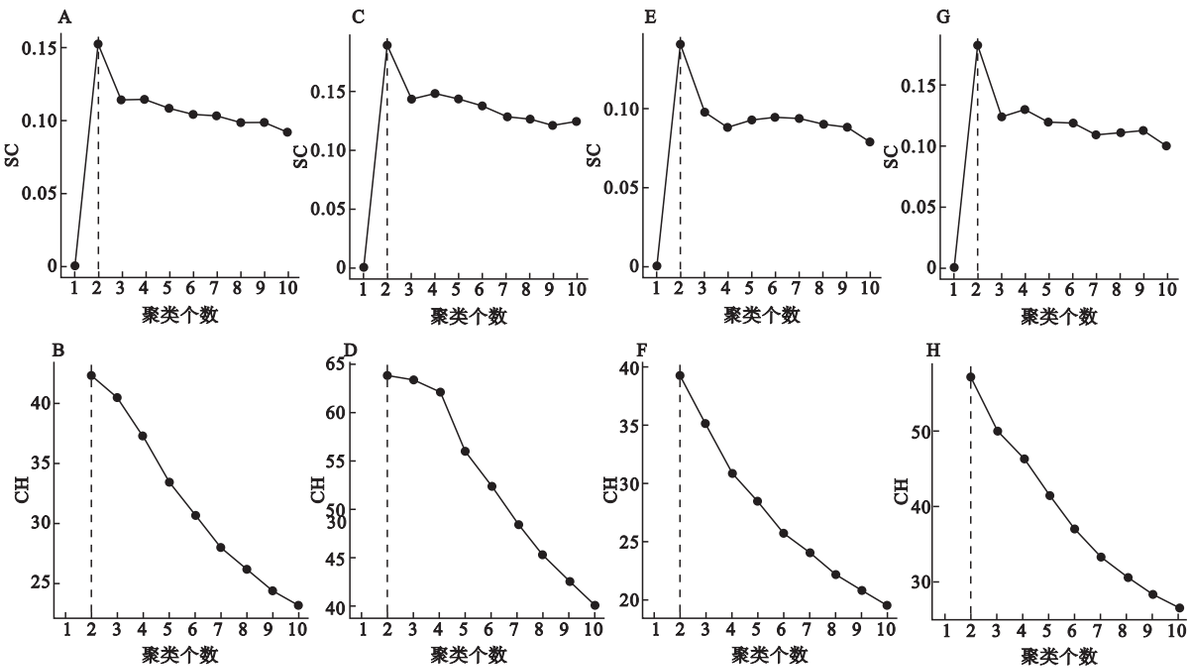
表 1　AE 和 SAE 参数选择情况

方法		单组学	特征数	Cox 模型过滤	输入层	瓶颈层	隐藏层	L1 正则化
No-Cox*	AE	mRNA	19943	-*				
		miRNA	1562	-	39865	150	2500	-**
		甲基化	18360	-				
	SAE	mRNA	19943	-				
		miRNA	1562	-	39865	200	500	0.000001
		甲基化	18360	-				
Cox	Cox-AE	mRNA	19943	2576				
		miRNA	1562	229	5083	200	3050	-
		甲基化	18360	2278				
	Cox-SAE	mRNA	19943	2576				
		miRNA	1562	229	5083	200	2542	0.000001
		甲基化	18360	2278				

*：No-Cox 表示在 AE 和 SAE 对数据降维之前不进行 Cox 模型的特征筛选，**：-表示在损失函数中不添加 L1 正则化惩罚。

3. 聚类效果评价

通过绘制SC指标和CH指标与聚类个数k之间的折线图来确定最佳聚类数，k的范围为1~10。纵坐标代表SC和CH的值，横坐标代表聚类个数，当纵坐标最大时所对应的横坐标为最佳聚类数，所有方法的最佳聚类数均为2，见图4。在SAE-k-means聚类结果中，SC和CH值均最大，DBI值最小，表示该模型的聚类效果最好，代表类自身越紧密，类之间越分散，即更优的聚类结果，见表2。



注：A、B为AE方法；C、D为SAE方法；E、F为Cox-AE方法；G、H为Cox-SAE方法

图4　SC和CH随k的折线变化图

表2　不同方法的聚类效果比较情况

方法	SC	CH	DBI
AE-k-means	0.15	40.56	2.08
SAE-k-means	**0.19**	**63.88**	**1.82**
Cox-AE-k-means	0.14	39.29	2.32
Cox-SAE-k-means	0.18	57.13	1.86

注：加粗代表最佳值。

4. NOS患者分型结果评价

在AE和SAE对数据降维之前不进行Cox模型的特征筛选的结果中，经k-means聚类得到两个亚型标签，AE将351例NOS患者分成了292例subtype1和59例subtype2；SAE分成了280例subtype1和71例subtype2。以不同亚型来分组绘制生存曲线，比较两组之间的生存情况。在AE-k-means和SAE-k-means中，两组之间的生存差异均具有统计学意义（χ^2=4.11，P=0.043；χ^2=5.36，P=0.021），见表3、图5A和5B。

在AE和SAE对数据降维之前进行Cox模型的特征筛选的结果中，Cox-AE-k-means和Cox-SAE-k-means两组之间的生存差异均具有统计学意义（χ^2=6.53，P=0.011；χ^2=7.27，P=0.007），见表3、图5C和5D。

表3　不同方法的生存分析结果

方法	样本量		χ^2	P
	subtype1组	subtype2组		
AE-k-means	292	59	4.11	0.043*
SAE-k-means	280	71	5.36	0.021*
Cox-AE-k-means	284	67	6.53	0.011*
Cox-SAE-k-means	285	66	7.27	0.007*

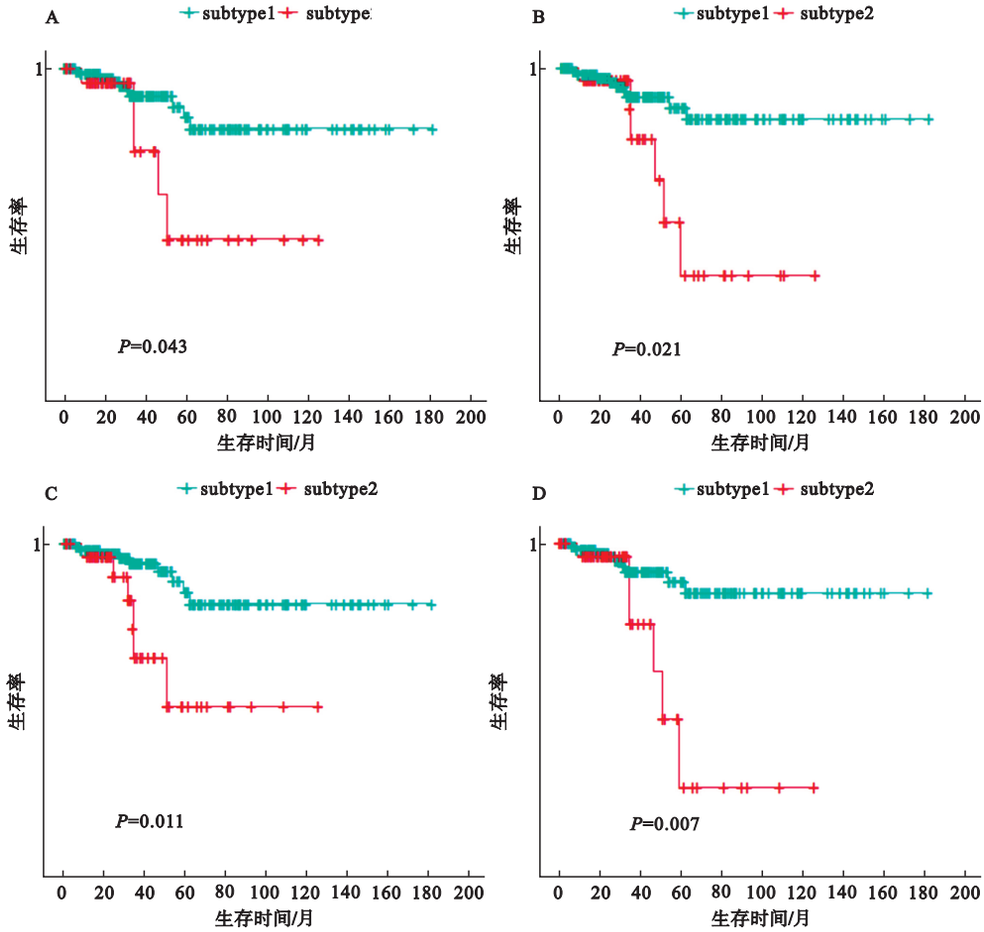
*：P<0.05。

5. 变量重要性评分

综上所述，SAE-k-means为最优模型，图6为该模型NOS患者可视化聚类图。以NOS患者两个亚型分组，分别来构建随机生存森林模型，从中评估各变量在两组生存结局中的重要性。VIMP的值越大，表明该变量的重要性越大，纵轴表示VIMP值，横轴表示选取的前30个变量，subtype1的情况见图7A、subtype2的情况见图7B。据相关文献报道，在subtype1中，*CXCL5*和*DGKI*基因有重要的临床意义；在subtype2中，*ATP2C2*、*EHF*、*PON2*基因有重要的临床意义。

讨　论

通过文献学习发现利用多组学数据对PTC进行分型的研究较少[7-8]。本研究从多组学角度，整合mRNA表达、miRNA表达和DNA甲基化数据对PTC中NOS型患者进行分型研究。多组学数据整合时降维是分析中的关键问题[9]。本文主要研究稀疏自动编码器的降维效果，并与AE进行比较，文中还对比分析自动编码器降维前是否需要采用单因素Cox模型进行



注：A，AE方法；B，SAE方法；C，Cox-AE方法；D，Cox-SAE方法。

图5 NOS分为两型患者的生存曲线比较

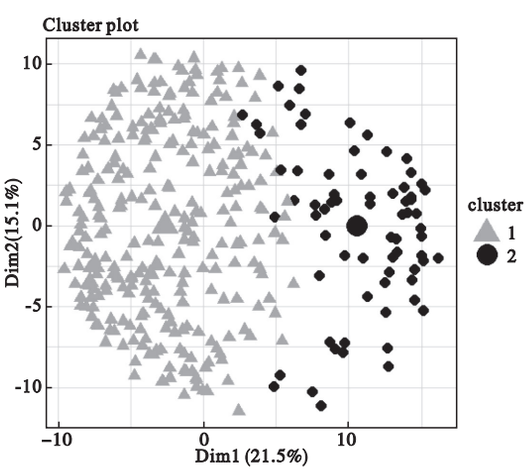


图6 SAE-k-means可视化聚类图

特征筛选，使用SC、CH和DBI指标评价四种降维模型（AE-*k*-means、SAE-*k*-means、Cox-AE-*k*-means、Cox-SAE-*k*-means）的*k*-means聚类效果，发现SAE-*k*-means的聚类效果最好。因此，本文利用SAE对多组学数据直接进行降维，通过*k*-means聚类找到了具有生存差异的两个亚型，即subtype1和subtype2。并与AE对比，SAE在聚类效果和预后分析方面有较好表现。

AE通过减少隐藏层中的节点数量来限制网络中的信息流量，确保模型对输入数据不产生记忆性，以此来达到瓶颈的目的并产生新的特征。由于AE仍可能存在一些输入记忆的机会，会限制其特征提取的能

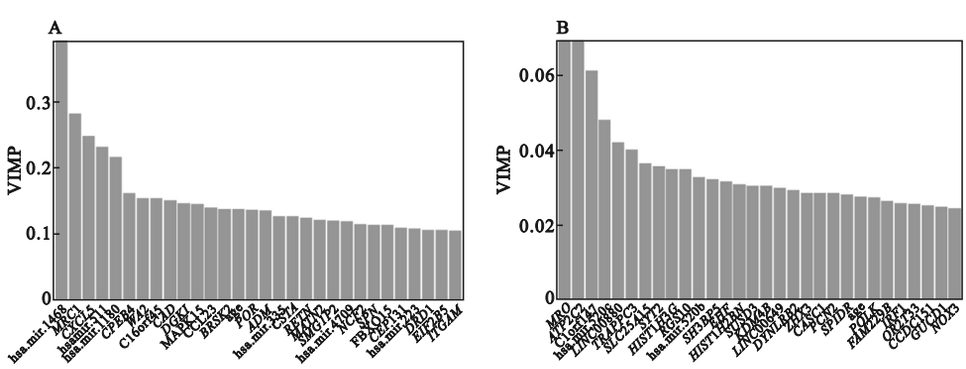


图7 subtype1（A）和subtype2（B）的变量重要性评分图

力。故为了进一步改善信息瓶颈的产生，通过构建SAE在网络上添加稀疏性约束，通过测量每个训练批次的隐藏层的激活，并在损失函数中添加L1正则项来惩罚过度激活，从而实现网络选择较少的隐藏层节点数[10]。本文AE选择的瓶颈节点数为150，其他隐藏层节点数为2500，SAE选择的瓶颈节点数为200，其他隐藏层节点数为500，L1正则化参数选择为0.000001。

在subtype1中，*CXCL5*可以通过激活JNK和p38MAPK途径来促进PTC细胞的增殖[11]。活化的CXCL5−CXCR2通过调节Akt/GSK−3β/β−连环蛋白途径，有助于PTC细胞的转移表型[12]。Bai等[13]发现了*RGS*8、*DGKI*和*OCA*2共表达基因可能是甲状腺癌患者有前途的预后分子标志物。在subtype2中，Zhao等[14]通过构建预后基因风险预测模型，表明了*ATP*2*C*2可作为预后因素，对甲状腺癌患者的生存有较好的预测。Lyu等[15]发现*EHF*的异位表达有助于甲状腺癌的发生，可能是甲状腺癌的潜在治疗靶点。Xie等[16]发现LINC00488−miR−376a−3p−PON2轴可以作为治疗甲状腺癌的新型生物标志物或潜在靶点。由于subtype2组预后较差，需着重关注这些基因的预后预测价值，它们有望作为PTC未来药物治疗的潜在靶点，基因间的调控作用需进一步研究。

研究的局限性：第一，在模型中仅对比了AE和SAE的降维性能，后续研究将探讨更多种自动编码器方法。第二，自动编码器和机器学习方法由于训练时间过长和硬件要求过高，需要设置合适的参数。

综上所述，SAE可用作对多组学数据进行特征提取的降维工具，不仅解决了维度灾难的问题，而且还提取到了关键特征，并表现出了较好的聚类效果和预后效果。有效识别了不同的亚型，寻找到了与生存密切相关的基因，为PTC治疗和预后评估提供了理论依据和临床指导。

参　考　文　献

[1] 魏佳. 基于循环肿瘤DNA的甲状腺癌分子诊断技术研究[D]. 长春：吉林大学, 2022.

[2] 赵海月, 宋旸, 沈迪寒, 等. 甲状腺癌的分型及其分子发病机制研究进展[J]. 中国体视学与图像分析, 2017, 22(4): 465-471.

[3] 徐昭阳, 刘改琴, 王彤, 等. 基于CiteFuse的肺腺癌单细胞多组学数据整合细胞类型识别[J]. 中国卫生统计, 2026, 43(3):351-358.

[4] 高振博. 基于多组学数据的聚类方法研究[D]. 大连：大连理工大学, 2020.

[5] Chaudhary K, Poirion OB, Lu L, et al. Deep learning–based multi-omics integration robustly predicts survival in liver Cancer[J]. Clinical Cancer Research, 2018, 24(6): 1248-1259.

[6] 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[J]. 中国卫生统计, 2021, 38(3): 327-331.

[7] Titov SE, Ivanov MK, Karpinskaya E V, et al. miRNA profiling, detection of BRAF V600E mutation and RET-PTC1 translocation in patients from Novosibirsk oblast (Russia) with different types of thyroid tumors[J]. BMC Cancer, 2016, 16(1): 201.

[8] Cao ZX, Weng X, Huang JS, et al. Receptor–ligand pair typing and prognostic risk model for papillary thyroid carcinoma based on single-cell sequencing[J]. Frontiers in Immunology, 2022, 13: 902550.

[9] Feldner-Busztin D, Firbas Nisantzis P, Edmunds SJ, et al. Dealing with dimensionality: the application of machine learning to multi-omics data[J]. Bioinformatics, 2023, 39(2): btad021.

[10] Madhumita, Paul S. Capturing the latent space of an Autoencoder for multi-omics integration and cancer subtyping[J]. Computers in Biology and Medicine, 2022, 148: 105832.

[11] Cui D, Zhao Y, Xu J. Activation of CXCL5-CXCR2 axis promotes proliferation and accelerates G1 to S phase transition of papillary thyroid carcinoma cells and activates JNK and p38 pathways[J]. Cancer Biol Ther, 2019, 20(5): 608-616.

[12] Cui D, Zhao Y, Xu J. Activated CXCL5-CXCR2 axis promotes the migration, invasion and EMT of papillary thyroid carcinoma cells via modulation of β-catenin pathway[J]. Biochimie, 2018, 148: 1-11.

[13] Bai M, Ke S, Yu H, et al. Key molecules associated with thyroid carcinoma prognosis: A study based on transcriptome sequencing and GEO datasets[J]. Front Immunol, 2022, 13: 964891.

[14] Zhao H, Zhang S, Shao S, et al. Identification of a Prognostic 3-Gene Risk Prediction Model for Thyroid Cancer[J]. Front Endocrinol (Lausanne), 2020, 11: 510.

[15] Lyu Y, Sui F, Ma J, et al. Increased expression of EHF contributes to thyroid tumorigenesis through transcriptionally regulating HER2 and HER3[J]. Oncotarget, 2016, 7(36): 57978-57990.

[16] Xie F, Li L, Luo Y, et al. Long non-coding RNA LINC00488 facilitates thyroid cancer cell progression through miR-376a-3p/PON2[J]. Biosci Rep, 2021, 41(3): BSR 20201603.

（责任编辑：张　悦）