基于堆叠降噪自编码器的肝癌亚型分类

张甜甜 赵庶旭 王小龙
(兰州交通大学电子与信息工程学院 甘肃 兰州 730070)

摘 要 肝癌是威胁人类健康的常见恶性肿瘤之一。通过对基因数据使用深度学习方法进行整合来系统地获取对肝癌的认知,使用多组学的疾病分析方法来探究各组学之间的相互关系,有助于更准确的临床决策。然而,由于多组学数据具有高维稀疏性,存在大量的冗余特征和较少的可用临床标签样本。堆叠降噪编码器(SDAE)是能够从海量数据中获取有效特征的高效模型,因此基于 SDAE 模型提出一种层次式堆叠降噪编码器,来学习肝癌的 RNA 表达、miRNA 表达和 DNA 甲基化数据的特征并进行整合和识别。实验结果表明:Hi-SDAE 方法提高了对肝癌亚型分类的准确度,为肝癌针对性治疗提供了更有价值的参考依据。

关键词 堆叠降噪 自动编码器 数据降维 多组学整合 肝癌亚型

中图分类号 TP311.13 R730.2 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.06.012

CLASSIFICATION OF LIVER CANCER SUBTYPES BASED ON STACKED DENOISING AUTOENCODER

Zhang Tiantian Zhao Shuxu Wang Xiaolong
(*School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, Gansu, China*)

Abstract Liver cancer is a common malignant tumor that threatens human health. To systematically acquire the knowledge of liver cancer by integrating genetic data using deep learning methods, we use a multi-omics disease analysis approach to explore the interrelationships between the groups and to obtain more accurate clinical decisions. However, due to the high dimensional sparsity of multi-omics data, there are a large number of redundant features and fewer available clinical label samples. Stacked denoising autoencoder (SDAE) is an efficient model that can obtain effective features from massive data. Therefore, based on the SDAE model, a hierarchical stacking denoising encoder was proposed to learn and integrate the characteristics of RNA expression, miRNA expression and DNA methylation data of liver cancer. The results show that the Hi-SDAE method improves the accuracy of the classification of liver cancer subtypes, and provides a more valuable reference for the targeted treatment of liver cancer.

Keywords Stacked denoising Autoencoders Data dimensionality reduction Multi-omics integration Liver cancer subtypes

0 引 言

在全球范围内,原发性肝癌是癌症相关死亡的第四大原因,而 85% ~90% 以上的原发性肝癌是属于肝细胞癌(HCC)。中国是肝癌患病率最高、生存率却很低的国家,我国肝癌患者的平均年龄在 55 ~59 岁之间,要比肝癌低发率国家早 20 年左右[1-2]。为了提高肝癌诊治与预防的效益,将 2015 年美国政府提出的“精准医学”治疗模式应用到肝癌诊断中,即对肝癌患者的所处阶段和不同的状态对应下的基因表达进行分子层面的分类,即亚型。医生根据其分子层面共同特性,可找到有价值的肝癌标志物以及为探讨治疗肝癌相关的药物与方案提供新的思路,而对肝癌患者而言,肝癌亚型分类为肝癌患者实现针对性治疗作了铺垫,也帮助人们更加全面系统地认识肝癌的发生发展

收稿日期:2021 -01 -18。国家自然科学基金项目(6206020135)。**张甜甜**,硕士,主研领域:生物信息学。**赵庶旭**,教授。**王小龙**,博士。

机制[3]。

目前,现有的肝癌亚型识别主要是基于统计法对单组学数据进行建模分类。Lee 等[4]利用无监督方法对肝癌的基因表达数据进行亚型分类,完成肝癌不同生存期对应的基因表达特征。Hoshida 等[5]利用无监督聚类方法对转录组数据来实现肝癌亚型分类,并发现临床特征与亚型之间的关联。然而这些方法都是针对单组学数据进行肝癌亚型的识别并没有考虑到不同组学数据之间的联系和相互作用,缺少生物系统间的完整性,不易察觉肝癌的生物学行为。因此,刘刚等[6]利用聚类方法整合 DNA 甲基化、拷贝数变异、miRNA 表达和 RNA 表达进行肝癌亚型的分类。Kuar 等[7]利用传统的机器学习方法结合基因组学和表观组学实现肝癌患者早期和晚期分类。Chaudary 等[8]利用 VE 方法将 RNA 表达、miRNA 表达和 DNA 甲基化数据进行整合之后,利用 K-means、SVM 模型对整合数据进行肝癌亚型的鉴定。同时,研究表明将基因组学、转录组学和表观组学等多组学进行整合分析,可以多层级全方位的了解肝癌的产生过程。但多组学数据往往由不同的测量平台而获得,并且多组学数据之间是不独立的,致使数据难以融合。这削弱了模型预测的能力,降低了模型的鲁棒性,从而导致产生不可靠的结论。

为了提高亚型分类任务中模型的泛化能力,前人在下游分析之前使用传统的降维算法和主成分分析(PCA)方法进行特征降维。尽管这些方法各有其优势,但忽略了数据之间的非线性关系[9]。近年来,基于深度学习(DL)的模型在处理高维数据和提取数据的线性和非线性关系方面取得了优势,自动编码器框架(AE)已经成功应用于分析高维基因表达数据和整合组学数据[10-11],但是单一的自动编码器模型无法从嘈杂复杂的高维数据中提取出所有有用的特征表示。Danaee 等[12]使用堆叠降噪自编码器(SDAE)将高维的单组学数据压缩到低维并从中提取了有效的特征信息,为深度学习技术应用到生物医学数据中提供基础。为了研究多组学数据之间的相互关系和内在联系,本文基于以上的模型和研究提出了一种层次式深度学习网络堆叠降噪编码器(Hi-SDAE)的来整合肝癌的 RNA 表达,DNA 甲基化和 miRNA 表达数据。

1 数据及数据预处理

1.1 肝癌组学数据集及特征分析

本文使用 UCSCxena 网站[13]从癌症基因组图谱(TCGA)数据库中获取肝癌相关的组学数据,使用 RNA 表达、miRNA 表达和 DNA 甲基化数据及临床数据作为研究数据。TCGA 计划是以帮助改善诊断方法、治疗标准和有效地预防癌症为目标而构建的公开数据集[14-15]。TCGA 数据库不仅含有丰富的癌症类型,还含有多种与癌症相关的多组学数据以及相应的临床数据,本文下载的肝癌数据主要分为以下三种公开数据:

1) RNA 表达数据:使用北卡罗来纳大学 TCGA 基因表征中心的 IlluminaHiseq2000RNA 测序平台测量 RNA 表达谱,由转换的 RSEM 归一化计数,此数据集显示基因水平的转录估计。

2) DNA 甲基化数据:使用 Illumina Infinium HumanMethylation450 平台通过实验测量了 DNA 甲基化谱,DNA 甲基化值是介于 0 与 1 之间的连续变量,表示从无甲基化到完全甲基化的状态转换。基因异常甲基化会导致肝细胞癌的无限繁殖及扩散。

3) miRNA 表达数据:使用 BCGSC Illumina Hiseq_miRNASeq 平台通过 RPM 估算表达值后获得。使用此数据集作为肿瘤启动子或抑制因子来控制肝癌细胞的增殖、迁移、侵袭和发展。

1.2 数据的预处理

多组学数据的预处理是为了从海量基因数据中提取出有研究价值的数据,主要包括数据过滤、缺失值填补和数据的归一化。

1) 数据过滤:目的是过滤掉明显的噪声数据或仅有少量表达值和表达值极小的数据。由于生物医学数据往往不完整和有偏差,即测量技术的局限性和一些其他条件的约束(如:自然条件)导致数据有残缺。组学数据的缺失值太多会造成其分析结果不可信,还有一部分组学数据仅在少数样本中表达,也不具有后续分析的统计价值。文中下载的肝癌多组学数据同样含有缺失值,因此假定患者的生物学特征缺失值超过 20%,将删除该患者的此条数据。

2) 缺失值填补:有些组学数据仅在少数样本中发生缺失或者其缺失值未达到预先设定的过滤值,因此这一部分组学数据需要进行缺失值的填补,目的是为了提高输入数据的有用性,确保实验结果的准确性和真实性。KNN 插值法[16]是一种广泛使用的缺失值插补方法。它使用相邻观测值的完整值来估计缺失值,被广泛认为是优于传统插补技术的方法。本文使用 KNN 插值法来填补缺失值。

3) 数据归一化:是为了将不同平台获取的多组学数据进行整合分析时,消除不同的统计属性和表示形式对后续分析造成的不利影响。文中采用标准分数进

行归一化[17],计算公式如下:

$$\tilde{f} = \frac{f - E(f)}{\sqrt{Var(f)}} \tag{1}$$

式中:f 为输入的基因特征,$\tilde{f}$ 为归一化之后的数值,$E(f)$ 和 Var(f) 分别为 f 的均值和方差。表 1 即为获得到的多组学数据在经过预处理方法前后的数据变化情况。

表 1　组学数据预处理前后的变化

数据类型	原始数据		预处理后数据	
	特征	样本	特征	样本
DNA 甲基化	485 578	429	67 369	389
RNA 表达	20 530	420	6 287	389
miRNA 表达	2 137	423	261	389

2　层次式堆叠降噪自编码器的构建

层次式堆叠降噪自编码器的构建首先利用了堆叠降噪自编码器重构单组学数据的特征表示,然后使用自动编码器集成学习多个组学数据,最后将集成的高级表达输入到分类器进行亚型分类。为了获得最佳的分类效果,模型采用预训练模型权重和微调的训练方式。

2.1　自编码器

传统的自编码器(AutoEncoder, AE)[18]是一种无监督的神经网络模型,一般用来对原始数据集进行分布式的、压缩的学习特征表达,同时达到很好的降维效果。自编码器由输入层、隐藏层和输出层三层网络构成,从输入层到隐藏层称之为编码过程,编码是指学习原始数据的隐藏表达特征。从隐藏层到输出层称之为解码过程,解码是指重构原始数据得到新的特征表示,以此达到特征提取。自动编码器是通过定义重构误差函数来评定其学习效果。

若 AE 的输入向量 $\boldsymbol{x} \in [0,1]^d$,则其编码函数为:

$$\boldsymbol{y} = f_\theta(\boldsymbol{x}) = s(\boldsymbol{wx} + \boldsymbol{b}) \tag{2}$$

式中:中间的隐藏层变量 $\boldsymbol{y} \in [0,1]^{d'}$,参数为 $\theta = \{\boldsymbol{w}, \boldsymbol{b}\}$,$\boldsymbol{w}$ 为 $d \times d'$ 的权重矩阵,$\boldsymbol{b}$ 为偏置。则其解码函数为:

$$\boldsymbol{z} = g_{\theta'}(\boldsymbol{y}) = s(\boldsymbol{w}'\boldsymbol{y} + \boldsymbol{b}') \tag{3}$$

式中:重构的原始特征为 $\boldsymbol{z} \in [0,1]^d$,参数为 $\theta' = \{\boldsymbol{w}', \boldsymbol{b}'\}$,且设置 $\boldsymbol{w} = \boldsymbol{w}'$。

为了使原始输入数据 $\boldsymbol{x}$ 和重构出的特征数据 $\boldsymbol{z}$ 之间的误差非常小,通过梯度下降法优化模型的参数以最大程度地减少重构误差。

$$\theta, \theta' = \arg\min \frac{1}{n} \sum_{i=1}^{n} L(x^{(i)}, z^{(i)}) = \arg\min \frac{1}{n} \sum_{i=1}^{n} L(x^{(i)}, g_{\theta'}(f_\theta(x^{(i)}))) \tag{4}$$

式中:$L(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|^2$ 为重构误差函数。而由于网络在处理分类问题时含有 sigmoid 或 softmax 等函数,会产生梯度消失,从而导致模型训练学习缓慢的现象产生。并且交叉熵更适用于二分类或多分类的情况下用作损失函数。因此,本文选定交叉熵为损失函数,表示为:

$$L(\boldsymbol{x}, \boldsymbol{z}) = -\sum_{k=1}^{d} [x_k \log z_k + (1 - x_k) \log(1 - z_k)] \tag{5}$$

2.2　降噪自编码器

降噪自编码器(Denoising AutoEncoder, DAE)[19-20]是在 AE 的基础上改进的,目的是能够更稳定地捕捉到更深层次的特征表示,让其在识别任务中获得更好的泛化性能。DAE 模型进行训练时,首先将完整的原始数据进行腐蚀破坏产生一个损坏的输入,然后将损坏之后的数据通过 AE 的编码函数映射到隐藏层,得到隐含特征表示。再通过 AE 的解码函数将隐含特征表示映射到输出层重构出原始数据的特征表示。DAE 算法的核心思想是在学习特征表示时,对输入模式进行局部腐蚀具有鲁棒性。因此输入模型进行训练的数据为腐蚀过的损坏数据 $\bar{\boldsymbol{x}}$,而其隐藏变量表示为 $\boldsymbol{y} = f_\theta(\bar{\boldsymbol{x}}) = s(\boldsymbol{w}\bar{\boldsymbol{x}} + \boldsymbol{b})$,重构数据表示为 $z = g_{\theta'}(\boldsymbol{y}) = s(\boldsymbol{w}'\boldsymbol{y} + \boldsymbol{b}')$,通过梯度下降法优化参数以缩小重构误差。

$$\tilde{\boldsymbol{x}} \sim q_D(\tilde{\boldsymbol{x}} \mid \boldsymbol{x}) \tag{6}$$

$$\theta, \theta' = \arg\min E_{q^0(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})}[\boldsymbol{x}, g_{\theta'}(f_\theta(\tilde{\boldsymbol{x}}))] \tag{7}$$

式中:$q_D^0(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})$ 为噪声分布,$\tilde{x}$ 为加入噪声后的数据输入;$\boldsymbol{w}$、$\boldsymbol{b}$ 为编码函数的权重和偏置值;$\boldsymbol{w}'$、$\boldsymbol{b}'$ 为解码函数的权重和偏置值。

2.3　堆叠降噪自编码器

堆叠降噪自编码器错误:引用源未找到是将多个降噪编码器逐个堆叠连接而成。模型首先采用贪婪式分层预训练权重逐层进行初始化,然后通过微调的训练方式训练整个网络架构,SDAE 模型网络结构如图 1 所示。具体来说,贪婪式分层方法是将原始数据输入到第一层 DAE 上,经过训练得到第一层的隐含特征,然后将第一层的隐含特征作为输入到第二层的 DAE 训练得到下一级的隐含特征,直到最后一层的 DAE 得到信息充分的重构数据,如此将很多的 DAE 模型堆叠起来形成一个深度的学习架构。SDAE 模型是在加深网络深度的同时利用经过腐蚀处理的原始输入实现了鲁棒性更好的特征学习表示。而由于其本身并不

能完成分类任务,因此需要在 SDAE 模型后添加一个 softmax 分类器以达到预测分类的目的。

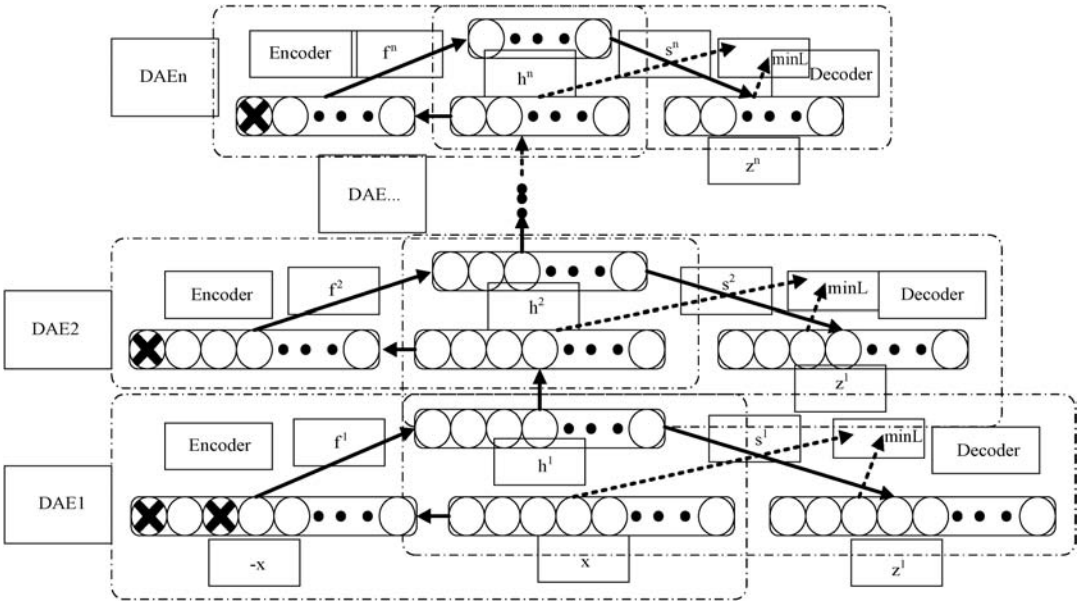

图 1 堆叠降噪编码器结构图

2.4 层次式堆叠降噪自编码器

由于 RNA 表达数据属于基因组,miRNA 表达数据属于转录组,而 DNA 甲基化数据是 DNA 化学修饰的一种形式。这三者处于不同层次的分子数据,为了避免不同类型测序数据的平台差异的影响和克服来自不同测量数据的整合偏差。文中提出了针对层次化数据的堆叠降噪编码器(Hierarchical Stacked Denoising Autoencoder,Hi-SDAE)方法来整合 RNA 表达、miRNA 表达和 DNA 甲基化数据来进行肝癌的亚型分类。首先,将 RNA 表达、miRNA 表达和 DNA 甲基化数据分别输入 SDAE 模型中,得到单组学数据的中间表示形式,然后在 SDAE 模型的输出层额外增加一层 AE,并将三个 SDAE 的输出作为 AE 的输入,再次通过编码函数得到其中间表达,然后经过编码函数得到高级表达,从而整合单组学数据的中间表示形式,最后整合成的高级表达特征输入 softmax 分类器进行肝癌亚型的分类识别。其整体框架如图 2 所示。S1、S2、S3、S4 分别代表一种由其肝癌临床信息所得类别。)

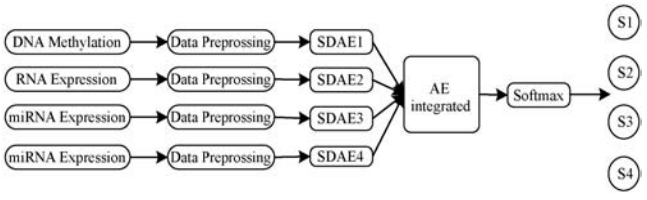

图 2 分层的 SDAE 总体框架

本文中 SDAE 模型进行堆叠的隐藏层数和隐藏变量可根据实际需要进行自定义,而且每个组学数据的特征学习的 SDAE 是不一样的。一般来讲,数据量大时需要设置的隐藏层数和隐藏节点数相对较多,数据量小则设置相对较少,其核心目的是使得误差能够最小。因此该模型可灵活地为单组学数据设计训练模型框架,即考虑了各种类型数据的固有统计属性又保持了不同组学数据的相关性,这在保留了组学数据多样性的前提下得到数据的重构表达,并最终提高了分类任务的分类精确度。

3 实验与结果分析

3.1 实验结果

SDAE 模型训练中采用的噪声分布为高斯噪声分布,其值设为 0.5,并使用合成过采样技术(SMOTE)将数据转换成更平衡的表示形式进行训练。图 3 即为 SDAE 模型分别训练 DNA 甲基化、RNA 表达和 miRNA 表达数据时对应的损失率。

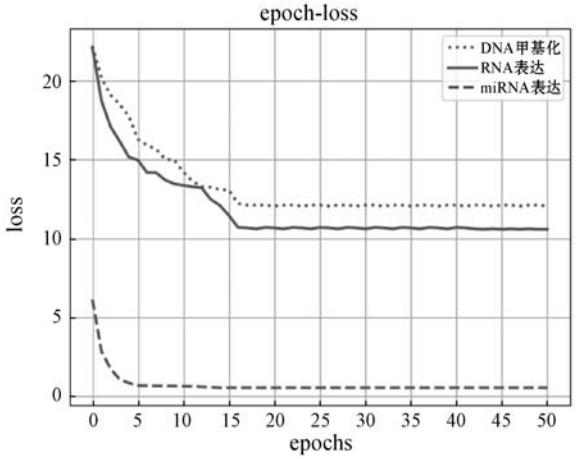

图 3 SDAE 模型的损失率

由于 SDAE 模型的工作是对原始数据进行特征提取,本身并不具备分类识别功能,因此在 Hi-SDAE 模型的输出层后加 softmax 网络分类器,实现了亚型分类的任务。依据下载的临床数据将肝癌分为 4 个亚型,将实验数据以 3∶1的比例划分测试集与训练集,并采用十折交叉验证法来提高模型的泛化能力。

评估公式分别使用准确率 A(accuracy),其计算式为:

$$A = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{8}$$

式中:T_P 表示实际值为正,预测值为正;F_P 表示实际值为负,预测值为正;T_N 表示实际值为负,预测值为负;F_N 表示实际值为正,预测值为负。

该模型在测试数据集上分类精确度为 0.861,分类结果如图 4 所示。该实验结果较好,符合预期。

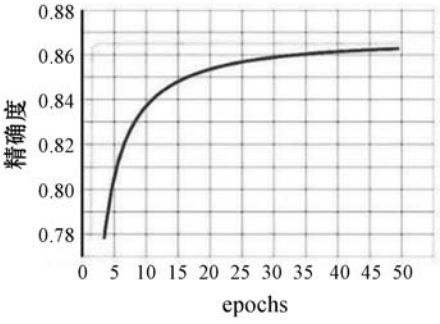

图 4　亚型分类结果图

3.2　性能评估与比较

为了测试整合多组学数据是否有助于肝癌亚型分类以及评估 Hi-SDAE 模型在学习表征方面的能力,将分别利用传统的非负矩阵分解(NMF)方法和主成分分析(PCA)方法在肝癌的 RNA 表达、miRNA 表达、DNA 甲基化以及整合数据进行特征降维分析,分类精确度作为其性能的评判标准,其结果如表 2 所示。

表 2　不同方法的性能比较

数据类型	PCA	NMF	SDAE
RNA 表达	0.769	0.731	0.746
miRNA 表达	0.731	0.692	0.784
DNA 甲基化	0.692	0.654	0.823
整合后数据	0.808	0.769	0.861

实验结果表明,与仅使用单组学数据进行亚型分类相比,整合多组学数据有助于提高亚型分类的准确度。与传统的降维方法相比较,本文提出层次式的 SDAE 模型的分类准确度明显高于传统的 PCA 和 NMF 方法,模型具有更优的鲁棒性。

4　结　语

多组学数据的整合分析有助于为其临床诊断、疗效评估及预后判断提供更全面的肝癌亚型基因信息。然而肝癌的组学数据存在大量噪声和无关特征,导致传统的数据分析方法难以直接应用于肝癌组学数据。因此文中提出一种基于 SDAE 模型的层次集成深度学习框架(Hi-SDAE)模型,将肝癌的多个组学数据整合并识别肝癌亚型。首先使用 SDAE 网络学习每种数据类型的表示,然后利用 AE 网络整合每种类型的学习表示以形成高级表达,将高级表达输入到 softmax 分类器中进行亚型分类。同时也通过实验证明了使用该模型整合多组学数据最终提高了亚型分类的准确率,同时表明层次深度学习框架为整合多种组学数据为研究亚型分类提供了一种新的选择。但是深度学习方法的局限性在于训练时需要大量的数据集,对于有限的肝癌样本,需要采用更强大的特征学习能力的模型来处理小型生物学数据。而且为了能够更加全面完整地了解肝癌的发展机制,如何整合更多类别的组学相关数据,进一步提高亚型分类的准确率将是接下来我们需要解决的问题。

参　考　文　献

[1] Siegel R L, Miller K D, Jemal A. Cancer statistics[J]. CA: A Cancer Journal for Clinicians,2020,70(1):7 - 30.

[2] Zucman-Rossi J, Villanueva A, Nault J C, et al. Genetic landscape and biomarkers of hepatocellular carcinoma[J]. Gastroenterology,2015,149(5):1226 - 1239.

[3] Wang K, Lim H Y, Shi S, et al. Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma[J]. Hepatology,2013,58(2):706 - 717.

[4] Lee J S, Chu I S, Heo J, et al. Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling[J]. Hepatology,2004,40(3):667 - 676.

[5] Hoshida Y, Nijman S M, Kobayashi M, et al. Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma[J]. Cancer Research,2009,69(18):7385 - 7392.

[6] Liu G, Dong C P, Liu L. Integrated multiple "-omics" data reveal subtypes of hepatocellular carcinoma[J]. PloS One,2016,11(11):e0165457

[7] Kaur H, Bhalla S, Raghava G P S. Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles[J]. PLoS One,2019,14(9):e0221476.

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

[9] Gupta A, Wang H, Ganapathiraju M. Learning structure in gene expression data using deep architectures, with an application to gene clustering[C]//IEEE International Conference on Bioinformatics and Biomedicine, 2015: 1328 - 1335.

[10] Chen L J, Cai C H, Chen V, et al. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model[J]. BMC Bioinformatics, 2016, 17(1): 7703 - 7729.

[11] Miotto R, Li L, Kidd B A, et al. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records[J]. Scientific Reports, 2016, 6: 26094.

[12] Danaee P, Ghaeini R, Hendrix D A. A deep learning approach for cancer detection and relevant gene identification [C]//Pacific Symposium on Biocomputing, 2019.

[13] Goldman M J, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform[J]. Nature Biotechnology, 2020, 38(6): 675 - 678.

[14] Akbani R, Ng K S, Werner H M, et al. Abstract 4262: A pan-cancer proteomic analysis of the cancer genome atlas (TCGA) project[J]. Cancer Research, 2014, 74(19): 4262.

[15] Chang K, Creighton C J, Davis C, et al. The cancer genome atlas pan-cancer analysis project[J]. Nature Genetics, 2013, 45(10): 1113 - 1120.

[16] Xiang Q, Dai X H, Deng Y, et al. Missing value imputation for microarray gene expression data using histone acetylation information[J]. BMC Bioinformatics, 2008, 9(34): 252 - 259.

[17] Wang B, Mezlini A M, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale[J]. Nature Methods, 2014, 11(3): 333 - 337.

[18] Ng A. Sparse autoencoder[EB/OL]. [2021 - 01 - 18]. https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf.

[19] Gehring J, Miao Y J, Metze F, et al. Extracting deep bottleneck features using stacked auto-encoders[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 3377 - 3381.

[20] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11(12): 3371 - 3408.

[21] Xu J, Xiang L, Liu Q, et al. Stacked sparse autoencoder (SSAE) for Nuclei detection on breast cancer histopathology images[J]. IEEE Transactions on Medical Imaging, 2016, 35(1): 119 - 130.

(上接第 54 页)

[4] 张晓峰,李昭,陈鹏. 高性能计算的发展现状分析[J]. 信息通信,2019(1):41 - 43.

[5] 和荣,王小宁,卢莎莎,等. 高性能计算环境通用计算平台[J]. 计算机系统应用,2019,28(12):55 - 62.

[6] 苏旬阳. 基于高性能计算云的作业调度系统的设计与实现[D]. 呼和浩特:内蒙古大学,2019.

[7] Ramakrishnan L, Jackson K, Canon S, et al. Defining future platform requirements for e-Science clouds[C]//ACM Symposium on Cloud Computing. ACM, 2010: 101 - 106.

[8] Wang G, Ng T. The impact of virtualization on network performance of amazon EC2 data center[C]//29th Conference on Information Communications. ACM, 2010: 1163 - 1171.

[9] 王超,曹继军,罗章,等. 面向 HPC 互连网络的低延迟前向纠错编码研究与实现[J]. 计算机工程与科学,2020,42(11):1965 - 1972.

[10] Rad P, Chronopoulos A, Lama P, et al. Benchmarking bare metal cloud servers for HPC applications[C]//2015 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM). IEEE, 2015.

[11] Liavas A, Kostoulas G, Lourakis G, et al. Nesterov-based alternating optimization for nonnegative tensor factorization: Algorithm and parallel implementation[J]. IEEE Transactions on Signal Processing, 2018, 66(4): 944 - 953.

[12] Sliwinski T, Kang S. Applying parallel computing techniques to analyze terabyte atmospheric boundary layer model outputs[J]. Big Data Research, 2017, 7: 31 - 41.

[13] Carreno E, Diener M, Cruz E, et al. Automatic communication optimization of parallel applications in public clouds [C]//IEEE/ACM International Symposium on Cluster. ACM, 2016: 1 - 10.

[14] Zhu C, Giorgi G, Christoph G. 2D relative pose and scale estimation with monocular cameras and ranging[J]. Navigation, 2018, 65(1): 25 - 33.

[15] Zhang R, Rossi F, Pavone M. Analysis, control, and evaluation of mobility-on-demand systems: A queueing-theoretical approach[J]. IEEE Transactions on Control of Network Systems, 2018, 6(1): 115 - 126.

[16] Zhang W, Cheng A, Subhlok J. DwarfCode: A performance prediction tool for parallel applications[J]. IEEE Transactions on Computers, 2016, 65(2): 495 - 507.

[17] Netto M, Calheiros R, Rodrigues E, et al. HPC cloud for scientific and business applications: Taxonomy, vision, and research challenges[J]. ACM Computing Surveys, 2018, 51(1): 1 - 29.

[18] Shi J, Taifi M, Pradeep A, et al. Program scalability analysis for HPC cloud: Applying Amdahl's law to NAS benchmarks[C]//2012 SC Companion: High Performance Computing, Networking Storage and Analysis. IEEE, 2013: 1215 - 1225.