基于机器学习的基因表达数据分析的研究进展

一、 基因表达数据分析（15 分）

基因表达数据分析是生物信息学中的一个重要领域，涉及对基因在不同条件下的表达水平进行测量和比较。基因表达数据中蕴含着基因活动的信息，可以反映细胞当前的生理状态。

基因表达数据的分析通常分为三个层次，第一个层次是分析单个基因的表达水平；第二个层次是考虑基因组合，将基因进行分组，研究基因的共同功能、相互作用以及协同调控；第三个层次就是推断基因调控网络。

基因表达数据通常通过高通量技术如微阵列[6]（microarray）或 RNA 测序（RNA-seq）获得。其中，RNA-Seq 用于测量基因表达，检查基因表达随时间或应用治疗的变化，发现和注释完整的转录本，检查转录后修饰，并表征选择性剪接和聚腺苷化。RNA-Seq 的其他重要特性包括其高分辨率和大动态范围，这导致了大量的获取数据，并促进了转录组学研究的显著进展[17]。由于上述优点，RNA-Seq 已经取代微阵列进行基因表达分析[1]。

目前对基因表达数据的处理主要是进行聚类或分类分析，将表达模式相似的基因划分到一类中，在此基础上寻找相关基因。由于基因表达数据是通过测量不同基因在特定条件或样本中的表达水平而得到的，这些数据对于理解生物学过程和疾病机制具有重要意义。然而，基因表达数据通常具有高维度和复杂性，传统的分析方法难以有效提取有用的信息，因此聚类和分类分析成为处理和解释基因表达数据的重要工具。

聚类分析是一种无监督学习方法，不需要任何先验知识，根据数学特征提取分类标准，对数据进行分类，特别适用于模式分类数不确定的情况。常见的聚类方法包括层次聚类、K 均值聚类和自组织映射等。分类分析则是一种有监督学习方法，用于根据已知类别标签对基因或样本进行分类。常见的分类方法包括支持向量机、随机森林、神经网络等。

对生物学学者而言，对基因表达数据进行聚类或者分类的操作，可以发现潜在的生物学模式，比如哪些基因在特定条件下会一同表达，这有助于理解基因功能和相互作用。对医护人员而言，通过聚类或分类操作，还可以进行疾病诊断和分类，基因表达分类分析可用于预测特定类型的癌症或其他遗传疾病，通过分析患者的基因表达数据与已知疾病状态的数据进行匹配，有助于疾病的早期诊断和治疗。对于癌症患者或其他病患而言，分类分析有助于确定患者对特定治疗方案的响应，从而实现个性化精确治疗[12]。例如，通过基因表达数据，可以预测患者对某种药物的敏感性，从而制定个体化的治疗策略。

二、 基因表达数据分析中的机器学习方法（20 分）

在基因表达数据分析中，机器学习方法可以分为传统机器学习算法和非传统机器学习算法。传统机器学习算法包括经典的统计学习方法和一些较早发展的算法，主要用于数据的分类和回归。非传统机器学习算法主要指近年来迅速发展的深度学习或其他学习方法，这些方法能够处理更复杂、更大规模的数据，且在许多应用中表现出色。

（一）、传统机器学习算法

1. 支持向量机（SVM）

SVM 通过寻找数据点之间的最优超平面来实现分类。它特别适用于高维数据，可以有效处理基因表达数据中的复杂模式。常用于癌症分类、疾病诊断等诸多场景。

2. 决策树及其集成方法（随机森林 RF、梯度提升树）

决策树通过分裂数据来构建树状模型，集成方法如随机森林（Random Forest）和梯度提升树（Gradient Boosting Trees），通过组合多个树模型提高预测性能。用于基因选择、生物标志物发现和分类任务。例如随机森林可以识别与特定疾病相关的重要基因。

3. 朴素贝叶斯（Naive Bayes）

基于贝叶斯定理，假设特征之间相互独立。该方法简单且高效，适用于高维数据。用于基因表达数据的快速分类。例如，朴素贝叶斯可以根据基因表达数据快速识别疾病样本。

（二）、非传统机器学习算法

1.深度神经网络（Deep Neural Networks, DNN）

DNN 由多个隐藏层组成，每一层包含许多神经元。每个神经元接收来自前一层的输入，进行加权求和和非线性激活，生成输出传递给下一层。用于基因表达数据的复杂模式识别和分类。例如深度神经网络可以用于多分类任务，如不同癌症类型的分类。

2.卷积神经网络（Convolutional Neural Networks, CNN）

CNN 主要用于处理图像数据，但也可以用于基因表达数据的分析，主要包含卷积层、池化层和全连接层。CNN 通过卷积操作提取局部特征，并通过池化层减少特征维度，最后通过全连接层对数据进行分类。用于基因表达数据的特征提取和分类。例如 CNN 可以用于分析基因表达谱图。

3.循环神经网络（Recurrent Neural Networks, RNN）

RNN 通过循环单元处理序列数据，每个时间步的输出不仅依赖于当前输入，

还依赖于前一时间步的状态。LSTM 是一种特殊的 RNN， 引入了记忆单元和门机制，解决传统 RNN 的长程依赖问题。LSTM 单元包括输入门、遗忘门和输出门。LSTM 模型用于分析基因表达的时间序列数据，捕捉动态变化规律。此外，利用 LSTM 模型，研究人员可以预测疾病的进展和发展趋势。

4. 自组织映射（Self-Organizing Maps, SOM）

SOM 是一种无监督学习的神经网络，通过自组织过程将高维数据映射到低维空间，同时保持数据的拓扑结构。SOM 由输入层和二维输出层（网格）组成，每个网格单元对应一个原型向量，用于将高维基因表达数据映射到二维平面，便于数据的可视化和模式发现。通过 SOM 模型，可以将基因表达数据分组，发现具有相似表达模式的基因群。

三、 基于机器学习的基因表达数据分析综述（50 分）

随着高通量测序技术的发展，基因表达数据的生成速度和规模不断增加，传统的数据分析方法已难以有效处理如此庞大的数据集。机器学习技术因其强大的数据处理和模式识别能力，逐渐成为基因表达数据分析的重要工具。本文将基于近年来的相关文献，以传统算法和非传统算法分类，总结机器学习在基因表达数据分析中的应用。

（一）、传统机器学习算法在基因表达数据分析中的应用

传统的机器学习方法，如支持向量机、KNN 算法、朴素贝叶斯、随机森林以及相关方法被用于早期癌症检测等应用。王艺任、王会等人，通过支持向量机递归特征消除算法和人工神经网络算法筛选构建一种基于 mRNA 基因表达的鼻咽癌诊断预测模型,为临床早期筛查、干预以及分子机制的研究提供参考[18]。Phimmarin Keerin 等人向 KNN 模型引入了一种新的汇总方法，提出了两种有序加权平均（OWA）聚合变体，提高了排除缺失信息的准确性[5]。Stephan Seifert 等人将有关结构和功能关系的外部知识整合到预测疾病状态的模型中，比较了随机森林引导通路选择的各种方法，提出针对不同预期的相关通路应采用不用的策略[13]。Jérémie Breda 等人推导出一种称为 Sanity（采样噪声校正的转录活性推断）的贝叶斯标准化程序，在执行查找最近邻细胞和将细胞聚类为亚型时，优于其他标准算法[3]。

（二）、非传统机器学习算法在基因表达数据分析中的应用

随着机器学习技术的不断发展，传统的机器学习算法如支持向量机、随机森林和朴素贝叶斯已经在基因表达数据分析中取得了显著成果。然而，近年来一些非传统的机器学习算法逐渐被引入并展示了其在处理复杂、高维和非线性数据方面的优势。

卷积神经网络(CNN)是一种深度学习架构，最初主要用于图像分析和处理，

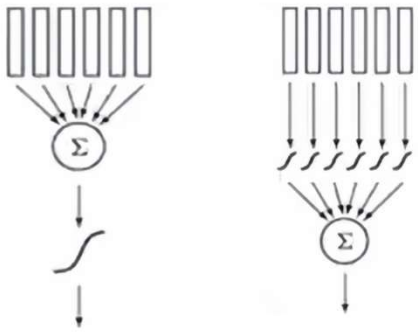
在基因表达数据的分析方面也有着出色的表现。Reinel Tabares-Soto 等人使用深度学习中的 CNN 算法对肿瘤数据进行分类，k 倍交叉验证后识别准确率达到了 94.43%[16]。AA Joshi 等人先将布谷鸟搜索(CS)和蜘蛛猴优化(SMO)相结合，再加入 CNN 算法，应用于大型基因表达数据集时，显著提高了预测准确率[4]。Prabhuraj Metipatil 等人提出了一个基于微阵列基因表达预测癌症类型的有效框架，使用混合 CNN + BiLSTM 方法识别不同形式的癌症，实验结果优于现有的 CNN 和 LSTM 分类器[10]。Jung Hun Oh 等人提出了一种名为 PathCNN 的新方法，该方法使用新定义的通路图像，在集成多组学数据上构建可解释的 CNN 模型，产生了良好的预测能力[11]。

循环神经网络(RNN) 是一种使用序列数据或时序数据的人工神经网络，用于分析基因表达的时间序列数据，捕捉动态变化规律。Noopur Singh 等人提出了一种基于双向长短期记忆循环神经网络的深度学习模型，通过增加训练时的时期数，提高模型的准确率，且该模型可以与完整基因组等庞大的序列数据兼容[14]。Sergii Babichev 等人研究了 LSTM 和 GRU 这两种类型的 RNN，提出了一种优化 RNN 架构和超参数值的算法，有效提升样本分类的准确性和分布质量[2]。Ramachandro Majji 等人基于松鼠搜索猎鹿的深度循环神经网络 (SSDH-based DRNN)，利用基因表达数据预测癌症存活率，拥有更高的预测准确率和更低的预测误差[9]。

图神经网络利用图结构数据进行学习，适用于基因共表达网络和基因相互作用网络的分析。Tianyu Liu, Yuge Wang 等人通过引入模态相似性学习图神经网络的新模型，很好的解决了由于生物医学背景不同，相似功能的基因数据具有异质性这一难题[7]。Shoujia Zhang 等人提出了一种用于微阵列数据中癌症分类的马尔可夫毯排序方法的图神经网络模型，通过基因属性和基因间的多类型关系网络来考虑微阵列数据中的癌症分类问题，有效提高了准确率和 f1 分数[19]。Tianci Song, Eric Cosatto 等人提出了一个基于图神经网络的框架来预测组织组织学图像中高表达基因的空间表达，与其他方法比提高了预测性能，且可用于更好的描述生物学上的空间特征[15]。

四、 基于机器学习的基因表达数据分析的未来趋势（15 分）

基于机器学习的基因表达数据分析在未来具有广阔的发展前景，首要的就是机器学习本身的快速发展。机器学习的前景非常广泛，涉及到物联网、生物信息学、自然语言处理等许多行业和领域。机器学习的各种算法也在不断的迭代，最早的多层感知机（Multi-Layer Perceptron，MLP），加入了空间相关性后，逐步发展成为了 DNN；加入了时间相关性后，发展成了 RNN，MLP 可以说是当前机器学习技术的基石。然而在 2024.05，Ziming Liu 团队受 Kolmogorov-Arnold 表示定理的启发，提出了 KAN（Kolmogorov-Arnold Networks）[8]。该研



究一经发布，就引起了机器学习界的广泛关注与讨论。传统的 MLP 是将各输入先进行线性组合，然后再使用 Sigmoid 或 ReLU 函数进行非线性激活。而 KAN 则是先将每个输入进行非线性激活，然后再进行线性组合。此外，与 MLP 最大不同的点是，KAN 的激活函数被参数化，可进行学习，从根本上消除了对线性权重矩阵的依赖。有学者认为，KAN 在未来将很有可能取代 MLP。如果 KAN 方案被证明有效，那么它将会极大地推动机器学习的发展。

除了机器学习技术的发展外，基因表达数据分析本身也令人充满期待。随着高通量测序技术的发展，不仅基因表达数据，其他类型的组学数据（如蛋白质组学、代谢组学、表观遗传学等）也越来越丰富。未来，整合多种组学数据进行分析，能够提供更全面的生物学理解。这种多组学数据的融合分析将有助于揭示复杂的生物机制。基于基因表达数据分析在个性化医疗中具有巨大潜力。通过分析患者的基因表达谱，可以实现许多重大疾病（如文中多次提到的癌症）的早期诊断、预测疾病进展和制定个性化治疗方案。

此外，在查阅文献的过程中，笔者发现虽然参与基因数据表达研究的中国学者有很多，但是相关的高质量的中文文献却较少。这需要所有中国学者的共同努力，才能提高中国的学术地位。

参考文献

[1] ALHARBI F, VAKANSKI A, 2023. Machine learning methods for cancer classification using gene expression data: a review[J]. Bioengineering, 10: 173.

[2] BABICHEV S, LIAKH I, KALININA I, 2023. Applying a recurrent neural network-based deep learning model for gene expression data classification[J]. Applied Sciences, 13: 11823.

[3] BREDA J, ZAVOLAN M, NIMWEGEN van, 2021. Bayesian inference of gene expression states from single-cell RNA-seq data[J]. Nature Biotechnology, 39: 1008-1016.

[4] JOSHI A A, AZIZ R M, 2024. A two-phase cuckoo search based approach for gene selection and deep learning classification of cancer disease using gene expression data with a novel fitness function[R]//Multimedia Tools and Applications. Springer: 1-32.

[5] KEERIN P, BOONGOEN T, 2021. Improved knn imputation for missing values in gene expression data[J]. Computers, Materials and Continua, 70: 4009-4025.

[6] 李坤鹏, 王泽朋, 周玉, 等, 2024. 人工智能在肿瘤基因表达数据中的应用研究进展[J]. 中国医学物理学杂志, 41: 389-396.

[7] LIU T, WANG Y, YING R, 等, 2024. MuSe-GNN: Learning unified gene representation from multimodal biological graph data[J]. Advances in Neural Information Processing Systems, 36.
[8] LIU Z, WANG Y, VAIDYA S, 等, 2024. Kan: Kolmogorov-arnold networks[R]//arXiv preprint arXiv:2404.19756.
[9] MAJJI R, RAJESWARI R, VIDYADHARI C, 等, 2023. Squirrel search deer hunting-based deep recurrent neural network for survival prediction using pan-cancer gene expression data[J]. The Computer Journal, 66: 245-266.
[10] METIPATIL P, BHUVANESHWARI P, BASHA S M, 等, 2023. An efficient framework for predicting cancer type based on microarray gene expressions using CNN-BiLSTM technique[J]. SN Computer Science, 4: 381.
[11] OH J H, CHOI W, KO E, 等, 2021. PathCNN: interpretable convolutional neural networks for survival prediction and pathway analysis applied to glioblastoma[J]. Bioinformatics, 37: i443-i450.
[12] RAVINDRAN U, GUNAVATHI C, 2023. A survey on gene expression data analysis using deep learning methods for cancer diagnosis[J]. Progress in Biophysics and Molecular Biology, 177: 1-13.
[13] SEIFERT S, GUNDLACH S, JUNGE O, 等, 2020. Integrating biological knowledge and gene expression data using pathway-guided random forests: a benchmarking study[J]. Bioinformatics, 36: 4301-4308.
[14] SINGH N, NATH R, SINGH D B, 2022. Splice-site identification for exon prediction using bidirectional LSTM-RNN approach[J]. Biochemistry and Biophysics Reports, 30: 101285.
[15] SONG T, COSATTO E, WANG G, 等, 2024. Predicting spatially resolved gene expression via tissue morphology using adaptive spatial GNNs[R]//bioRxiv. Cold Spring Harbor Laboratory: 2024-06.
[16] TABARES-SOTO R, OROZCO-ARIAS S, ROMERO-CANO V, 等, 2020. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data[J]. PeerJ Computer Science, 6: e270.
[17] WANG Y, MASHOCK M, TONG Z, 等, 2020. Changing technologies of RNA sequencing and their applications in clinical oncology[J]. Frontiers in oncology, 10: 447.
[18] 王艺任, 王会, 向红俐, 等, 2023. 支持向量机递归特征消除算法和人工神经网络算法筛选构建基于 mRNA 基因表达的鼻咽癌诊断预测模型[J]. 中华生物医学工程杂志, 29: 375-381.
[19] ZHANG S, XIE W, LI W, 等, 2023. GAMB-GNN: Graph Neural Networks learning from gene structure relations and Markov Blanket ranking for cancer classification in microarray data[J]. Chemometrics and Intelligent Laboratory Systems, 232: 104713.