DOI:10.3969/j.issn.1005-202X.2024.03.018

医学人工智能

人工智能在肿瘤基因表达数据中的应用研究进展

李坤鹏，王泽朋，周玉，李四海
甘肃中医药大学信息工程学院，甘肃 兰州 730000

【摘要】肿瘤是影响人类健康的严重疾病，早期诊断对提高治疗成功率和患者生存率至关重要。肿瘤基因表达数据的研究已经成为揭示肿瘤疾病机制的主要工具，人工智能在肿瘤基因表达数据分析中扮演着重要角色。本文从机器学习方法的角度，探讨监督式学习、无监督式学习和深度学习在肿瘤预测和分类中的潜在优势，特别关注特征选择算法对基因筛选的影响及其在高维度基因表达数据中的重要性。通过全面综述人工智能在肿瘤基因表达数据分析中的应用与发展，旨在为未来的研究方向提供参考，促进进一步发展。

【关键词】基因表达数据；人工智能；机器学习；特征选择；综述
【中图分类号】R318；TP18 【文献标志码】A 【文章编号】1005-202X(2024)03-0389-08

Review on application of artificial intelligence in tumor gene expression data analysis

LI Kunpeng, WANG Zepeng, ZHOU Yu, LI Sihai
School of Information Engineering, Gansu University of Chinese Medicine, Lanzhou 730000, China

Abstract: Tumors are serious diseases threatening human health, and the early diagnosis is essential to improve treatment success and patient survival. The study of tumor gene expression data has become a major tool for revealing tumor disease mechanisms, in which artificial intelligence plays an important role. The potential advantages of supervised learning, unsupervised learning and deep learning in tumor prediction and classification are explored from the perspective of machine learning methods. Special attention is paid to the impact of feature selection algorithms on gene screening and their importance in high-dimensional gene expression data. By providing a comprehensive overview of the application and development of artificial intelligence in the analysis of tumor gene expression data, the study aims to provide an outlook for future research directions and promote further development.

Keywords: gene expression data; artificial intelligence; machine learning; feature selection; review

前 言

肿瘤不仅是全球主要死因之一，也是制约人类期望寿命延长的重要因素，具有潜伏期长、复发率高以及早期难以检测等特点[1]。随着基因芯片和高通量测序技术的快速发展，获取肿瘤基因表达数据愈发便捷，这些数据已经成为肿瘤预测诊断、亚型分类、基因筛选和生物信息学分析等领域的主要工具之一[2]。然而，肿瘤基因表达数据的分析面临着小样本、高维度和多噪声等挑战，直接利用这些数据可能导致分析结果不可靠并增加时间成本[3]。在过去的几年里，人工智能(Artificial Intelligence, AI)技术的快速发展为生物医学领域带来许多新的机会和挑战[4-5]。特别是在肿瘤基因表达数据分析方面，AI的应用逐渐成为一个备受关注的领域。

通过文献回顾可以发现，截至目前尚未有一份系统性的综述文章涵盖AI在肿瘤基因表达数据分析中的应用。因此，本文旨在提供一份详尽的综述，系统总结AI在肿瘤基因表达数据分析中的最新研究进展，将特别关注AI技术在肿瘤分类预测和肿瘤关键基因筛选中的应用，以期为未来的肿瘤研究提供有价值的参考和启示。

1 基本概念和发展

在肿瘤基因表达数据分析领域，AI的应用主要涵盖临床应用、生物信息学分析等领域。在临床应

【收稿日期】2023-11-22
【基金项目】甘肃省科技计划项目(21JR1RA272)；甘肃省教育厅-高校教师创新基金(2023B-105)；甘肃省自然科学基金(22JR5RA606)
【作者简介】李坤鹏，硕士，研究方向：生物信息学、机器学习，E-mail: kunpeng@gszy.edu.cn
【通信作者】李四海，副教授，研究生导师，研究方向：数据挖掘、机器学习、光谱分析，E-mail: lisihai@gszy.edu.cn

用方面，重点体现在肿瘤预测与诊断、肿瘤亚型分类等方向。这些研究往往采用机器学习分类器，通过对肿瘤基因表达数据的分析，实现对肿瘤的预测和亚型分类。另一方面，生物信息学分析主要集中在肿瘤关键基因的筛选。采用特征选择算法对肿瘤基因表达数据进行降维和筛选，以便精准地鉴定和确定肿瘤的关键基因。这些方法的应用有助于深化对肿瘤分子机制的理解，并为个体化治疗和精准医学提供支持。肿瘤基因表达数据的分析流程如图1所示。

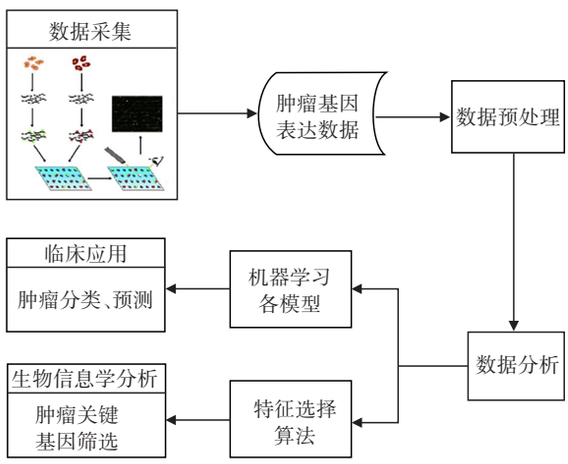

图1 肿瘤基因表达数据分析流程

Figure 1 Tumor gene expression data analysis process

基因芯片又称微阵列，由Fodor等[6]于1991年提出，这一技术的应用十分广泛。基因芯片能够制备成千上万个基因片段，通过仪器将这些基因片段固定在微小的芯片上，这些芯片上的DNA片段被称为探针。研究人员可以从感兴趣的样本中提取、分离和纯化RNA，然后使用聚合酶链反应等技术制备带有荧光标记的互补DNA（Complementary DNA, cDNA）。制备的cDNA与微阵列芯片上的探针进行DNA杂交，通过让cDNA与探针中的互补序列结合来实现。利用扫描设备等工具，测量杂交后的微阵列芯片的荧光信号，并将其转化为数值数据，这些数据构成了基因表达数据。基因表达数据能够有效地解释癌症在基因层面的运作机制[7]。基于上述技术，许多数据存储机构提供基因组信息数据资源。癌症基因组图谱数据库是一个由美国国家癌症研究所支持的大型协作计划，目标是通过全面的分子生物学分析，深入了解多种癌症的分子特征，促进癌症的分类、诊断和治疗[8]。基因表达综合数据库是美国国家生物技术信息中心开发的基因表达数据库，除了基因表达数据，该数据库还包含质谱分析等数据[9]。除上述数据库，常用的数据库还包括ArrayExpress、ENCODE、KEGG等[10-12]。

通过基因表达数据对肿瘤分类预测最早由Golub等[13]于1999年提出，此后各项相关研究便层出不穷。该领域研究通常使用机器学习技术实现。机器学习是计算机科学与统计学的交叉领域，旨在通过学习经验数据来改进计算机系统的性能[14]。它致力于开发算法和模型，使计算机能够自动从数据中学习、识别模式和规律，以进行预测和决策[15]。机器学习的核心思想是利用数据来训练模型以执行各种任务，包括分类、回归和聚类等。在机器学习中，数据的质量和多样性对模型的性能和泛化能力至关重要。深度学习是机器学习的一个分支，能够在更大规模的数据中进行学习并建立模型，至今已有多种深度学习模型应用在肿瘤基因数据分析领域并取得较好的效果[16]。随着数据量的不断增加和算法的不断改进，机器学习将继续成为解决各种现实问题的关键技术。

癌症关键基因筛选是从数万个基因中鉴别出与癌症发展密切相关的特征基因的过程。因此需要对基因数据降维，约简基因子集，这通常使用到AI中的特征选择技术。早期基因筛选通常采用统计学方法，如倍数变化法、统计检验法，但这些方法通常需要人为指定阈值，容易导致假阳性率过高，且未充分考虑基因之间的相互关联[17]。特征选择算法在处理高维小样本数据方面表现出强大的能力，其采用的距离度量和互信息度量也能够全面考虑基因与类别以及基因之间的关联[18]。特征选择的核心理念在于通过消除冗余、噪声或不相关的特征，聚焦于数据中最关键的信息，从而简化模型的复杂性，提高泛化能力。特征选择的好处是显而易见的，它能够简化模型，提高训练和预测速度，减小过拟合的风险，并提高模型的可解释性。然而，特征选择需要谨慎处理，不当的特征筛选可能导致信息损失和模型性能下降。因此，合适的特征选择方法应根据具体问题、数据集和学习算法来选择，以获得最佳效果。在肿瘤基因表达数据处理中，特征选择有助于数据降维，筛选出影响肿瘤的关键基因，并为后续分析提供更优质的数据。

2 AI在肿瘤基因表达数据中的应用

2.1 机器学习在肿瘤分类预测中的应用

在机器学习领域，构建分类预测模型是一项关键任务，其准确性和鲁棒性直接影响着模型的实际应用。这些模型通常分为监督式学习器和无监督式学习器两大类。深度学习是机器学习的一个特定分

支,它使用深度神经网络来学习复杂的表示,从而使计算机系统能够更高效地执行各种任务。

2.1.1 监督式学习器在肿瘤分类预测中的应用　监督式学习器用于肿瘤基因表达数据的常见示例包括K-最近邻、贝叶斯方法、支持向量机(Support Vector Machine, SVM)等。以SVM为例,Furey等[19]使用SVM对卵巢癌组织和正常卵巢组织数据集进行分析,虽然存在置信度不高的问题,但其分类预测性能较高,其他的分类模型也取得了类似的实验结果。通常置信度不高可能是由于特征维度过高、类别不平衡、高噪声以及不合理的超参数设置等原因引起的。为了应对这些问题并进一步提高模型性能,近年来涌现了各种研究。例如,Gao等[20]先使用快速相关性过滤(Fast Correlation Based Filter, FCBF)算法对肿瘤基因数据进行降维,再使用粒子群和人工蜂群两种优化算法改进SVM,提出PA-SVM模型,在9个癌症数据集上均取得最优性能。Huynh等[21]使用少数类别过采样算法针对基因表达数据中的类别不平衡问题进行处理,该方法成功将多个机器学习模型的准确率提高2%~7%。Prabhakar等[22]使用5种不同分类器对卵巢癌基因表达数据进行分类预测,结果显示使用径向基核函数的SVM取得最佳结果,准确率达到99.48%,这一结果也验证了超参数对模型性能的影响。除了对这些方面作出更改,各种新的模型也被应用于肿瘤分类预测中。Ludwig等[23]提出一种模糊决策树算法,该算法在结肠癌基因数据上取得80.28%的准确率并获得较高的置信度。Tabares-Soto等[24]对比机器学习的经典方法逻辑回归和深度学习模型中的卷积神经网络(Convolutional Neural Network, CNN)在11个癌症数据集上的训练表现,结果显示在这些数据集上,逻辑回归和CNN的平均准确率分别为90.6%和94.43%。从以上研究可知,在面对高噪声数据时,深度学习的性能表现更为出色。

监督式学习器在肿瘤分类预测中通常能取得较好的实验结果,尽管有一些局限性,也能够通过引入各种算法和处理方法进行解决。但监督式学习器训练时需要提供完整的数据信息,而面对没有标签的数据,监督式方法无法适用。

2.1.2 非监督式学习器在肿瘤分类预测中的应用　非监督式学习是机器学习的一种范式,其特点是训练数据无需包含标签或目标输出。与监督式学习不同,非监督式学习旨在从数据中发现结构、模式和关系,而不是预测特定的目标变量。Golub等[13]提出一种非监督聚类学习器,通过对急性白血病DNA微阵列表达数据的学习,能够自动识别急性髓性白血病和急性淋巴细胞白血病之间的区别,无需事先了解这些类别。Alagukumar等[25]使用一种基于关联规则的方法来分类乳腺癌基因芯片数据,这种方法将关联规则挖掘与分类技术相结合,算法包括统计基因过滤、离散化、类关联规则和预测或类分配等4个阶段,与线性分析、SVM和决策树模型进行比较,在交叉验证中获得最高准确率。Nguyen等[26]提出一种改进的层次分析法(Analytic Hierarchy Process, AHP),用于选择信息最丰富的基因,作为癌症分类2型模糊系统(IT2FLS)的输入,采用模糊C均值聚类的无监督学习策略来初始化IT2FLS的参数,该方法在淋巴瘤、白血病和前列腺癌数据集上均获得更好的效果。值得注意的是,改进的AHP方法在其他分类器应用时也提高了它们的性能,这凸显了特征选择的关键性。

非监督式学习能够对缺失标签的数据进行训练,但其也存在很多不足。例如,通常缺乏明确的评估指标,因为没有预先定义的目标变量,很难量化模型的性能和成功度。由于没有标签进行指导,非监督式学习容易受到输入数据中的噪声和异常值的影响,可能会导致模型学到的模式不准确或不理想。

2.1.3 深度学习在肿瘤分类预测中的应用　随着基因芯片制备技术的不断完善和各类基因数据库的丰富,肿瘤基因数据的规模不断增大。然而,传统机器学习在处理大规模数据方面存在一定的局限性。随着深度学习的发展,深度学习模型凭借其出色的大规模数据处理能力逐渐成为肿瘤基因表达数据分析的有力工具。在近年来的研究中,神经网络模型开始广泛应用于肿瘤基因表达数据的分析。深度学习模型采用非监督特征选择的高效算法,逐层学习训练,极大地提升模型的性能。深度学习模型主要包括人工神经网络、CNN、循环神经网络(Recurrent Neural Network, RNN)等。Nguyen等[27]将概率神经网络(Probabilistic Neural Network, PNN)引入到前文提到的AHP方法中,使得模型在处理复杂数据时具有更高的准确率和效率,该方法在结肠癌数据诊断中获得88.89%的准确率。Zeebaree等[28]使用一种简单的CNN模型对10个肿瘤数据集分别进行分类预测,该模型包括一个卷积层、一个池化层和一个全连接层,与SVM和随机森林进行对比实验,在大多数数据集上表现更为出色,并根据方差分析,CNN模型在准确率上表现最优。深度学习模型的应用除了直接替换机器学习模型,还可以对各个机器学习模型进行集成,对集成后的模型参数进行优化。Xiao等[29]采用深度学习对多个机器学习模型进行集成,对肺腺癌、胃腺癌和乳腺癌进行实验,与单一机器学习分

类器相比，该方法在受试者工作特征曲线下的曲线面积均达到最大，分别为0.988、0.988和0.979，准确率也显著提高，分别为99.20%、98.78%和98.41%。

监督式、非监督式及深度学习方法分类模型总结归纳见表1。

表1 机器学习模型总结

Table 1 Summary on machine learning models

分类方法	文献	应用方式	实验结果
监督式	文献[19]	使用SVM对卵巢癌组织进行分类	准确率较高，但存在置信度不高的问题
	文献[22]	对在卵巢癌数据集上使用5种模型进行对比分析	使用径向核函数的SVM准确率提高，验证了超参数对模型性能的影响
	文献[23]	使用模糊决策树方法对结肠癌数据进行预测	模糊决策树相较于对比的方法获得了更好的性能
	文献[20]	使用粒子群和人工蜂群算法优化SVM	在多个数据集上获得最好性能，证实了数据降维对模型性能的影响
非监督式	文献[13]	使用聚类学习器对白血病进行分类	聚类学习器无需了解数据类别，能够自动化训练，且有较高的分类准确率
	文献[25]	结合关联规则与分类技术在乳腺癌数据上进行实验	关联规则能够明显提高个分类器的性能
	文献[26]	结合层次分析法和聚类在多个数据集上进行实验	在各数据集上分类性能均提高，并验证了特征选择的重要性
深度学习	文献[24]	使用深度学习在多个肿瘤数据集上进行测试	深度学习表现出更高性能
	文献[28]	使用CNN对多个数据集进行分析	CNN表现出更高性能
	文献[27]	将PNN引入层次分析法	PNN使AHP获得了更高性能
	文献[29]	使用深度学习模型集成多个机器学习模型对肿瘤数据进行预测	相较于集成前，集成后的模型性能均得到了提升，且深度学习模型能够更好地调整各集成模型的参数

2.2 特征选择方法在肿瘤关键基因筛选中的应用

特征选择方法通常可分为3种类型：过滤式、包裹式和嵌入式[30]。过滤式方法独立于具体学习算法，它主要关注特征的排序而不是选择特征子集，并以其较好的泛化性能而闻名。包裹式方法考虑特征之间的相关性，但相对较耗时，因为它们通常需要重新运行特征选择过程。嵌入式方法将特征选择嵌入到学习算法中，减少冗余的特征选择步骤，但常常缺乏确定性阈值。

2.2.1 统计学方法在肿瘤关键基因筛选中的应用 早期对肿瘤基因表达数据的特征选择主要基于统计学方法，这类方法属于典型的过滤式算法。例如，Alagukumar等[25]在使用关联分类对微阵列基因表达数据进行分类的第一阶段中，首先使用T检验，根据自由度为n-2的t分布计算P值，按照P<0.05的标准，从上万个基因中筛选出2 701个基因，这也是传统生物信息学常用的差异分析方法。Maniruzzaman等[31]采用非参数检验方法Wilcoxon，将大量基因筛选至30~200个，并与随机森林分类器结合，取得令人满意的分类结果。微阵列显著分析法（Significance Analysis of Microarrays, SAM）是一种用于分析基因表达数据的统计方法，旨在识别差异表达的基因。SAM方法由Tusher等[32]于2001年首次提出，它强调微阵列数据中的特征选择和基因差异表达的重要性。任雨冬等[33]使用高斯核函数和欧氏距离函数改进SAM方法，并在白血病数据集上对比支持向量机递归特征消除法（Support Vector Machine- Recursive Feature Elimination, SVM-RFE）、SAM和Relief算法进行对比实验，通过和数据集官方给出的差异基因进行验证，该方法筛选关键基因的能力最好。

2.2.2 递归特征消除法（RFE）在肿瘤关键基因筛选中的应用 RFE属于包裹式选择方法，算法核心思想是反复构造模型，每轮训练删去拥有最小绝对值权重的特征，在下一轮训练中选取新的特征值继续进行，直到剩余特征数量达到设定的阈值。秦传东等[34]在L1-SVM和L2-SVM的基础上提出一种双正则化的支持向量机，并利用巴氏距离剔除对分类无关的基因，从而筛选出肿瘤关键基因，该算法在两个公共肿瘤基因表达数据集上表现出良好性能。谢志伟等[35]提出RD-SVM算法，使用随机矩阵替换算法构建多组随机向量来表示基因子集，以SVM的准确率评价基因子集的优劣，该方法能够考虑特征之间的相互

作用，并筛选出最优基因子集。张世芝等[36]以线性核SVM为分类器，构建一种稳健的特征排序及后向剔除方法，首先通过Monte Carlo采样方法构建多个数据子集，并利用各子集上建立的线性SVM模型对变量排序，然后对所有变量排序进行整合，最后按后向特征剔除方法进行特征选择，通过对白血病和前列腺癌数据的分析，该方法能有效选择特征基因且得到的特征基因对样本的识别能力稳健。

2.2.3 优化算法在肿瘤关键基因筛选中的应用　统计学方法和递归特征消除法的共同缺点是需要人为设置筛选阈值，这通常是不好把控的。优化算法是一类用于寻找问题最优解或最优近似解的数学方法，在特征选择领域常用于对目标的搜索和优化。粒子群优化（Particle Swarm Optimization, PSO）是一种启发式优化算法，灵感来自于鸟群或鱼群等自然群体的行为，最初由Kennedy等[37]提出，旨在解决优化问题，特别是连续空间中的优化问题。PSO算法的基本思想是模拟鸟群中的粒子，这些粒子在搜索空间中移动，每个粒子都有一个位置和速度，它们通过与群体中其他粒子的交互来调整自己的位置和速度，以寻找问题的最优解。关健等[38]提出一种基于PSO的特征选择算法，在PSO算法的适应值计算中，通过分析特征基因之间的判别熵信息，剔除冗余特征以搜索最佳特征子集，该算法在白血病和小蓝圆细胞瘤数据集上取得较好的效果。刘金勇等[39]在使用PSO算法前，先对肿瘤基因数据进行聚类，并对聚类结果进行选择，将被选中的簇的中心作为PSO的初始值，每个被选中的簇作为一个搜索空间，克服传统PSO算法收敛速度慢的特点。杜洪波等[40]提出PSO融合改进的K-means算法，除了不同类簇中最优的基因，也找到相对应类簇中的次重要的基因，该算法在结肠癌和白血病数据集上表现优于大部分算法。熊颖[41]提出一种结合LASSO算法和二进制PSO（BPSO）的基因选择方法，借助LASSO算法的噪声过滤和模型泛化优势进行关键基因筛选，并同时计算基因对分类的贡献值，最后利用BPSO算法进行最佳信息基因的选择。叶超超等[42]提出一种结合Relief-F和决策树的适应性PSO算法（APSO），该方法首先利用Relief-F快速过滤大量无关基因和噪声，缩小基因选择范围，然后以分类回归树作为适应度函数，用APSO算法对基因进行最终搜索。遗传算法（Genetic Algorithm, GA）是一种受到生物遗传进化过程启发的优化算法，用于寻找函数的最优解或解决组合优化问题。GA模拟了生物进化中的遗传和自然选择过程，将这些过程应用于求解问题的搜索和优化。Salem等[43]先使用信息增益（Information Gain, IG）进行特征提取，接着使用GA进行子集约简，再对癌症进行分类。魏莎莎等[44]提出一种结合互信息特征评估和GA的基因选择方法，使用GA搜索最佳特征子集，而互信息被用作特征之间的相关性度量，以帮助选择与目标变量相关性最高的特征子集，该算法在多个分类器上均取得较高的性能。萧秋兰等[45]提出一种融合GA和学习自动机的算法，适应度函数为染色体分配分数，而染色体上每个基因位置是绝对随机的，该方法运用学习自动机的特点，以寻找最优染色体个体。除了PSO和GA，还有多种优化算法被用于特征选择领域，Medjahed等[46]提出基于Binary Dragonfly群体优化算法的优化器来确定SVM-RFE之后的最佳基因子集，该算法受到自然界蜻蜓静态和动态群体行为的启发，能够解决SVM-RFE没有考虑基因冗余，并在某些值上变得不稳定等问题。Abdelnabi等[47]使用两阶段特征选择算法对乳腺癌和结肠癌进行特征选择，首先使用信息增益从原始数据集中选择重要基因，然后应用灰狼优化算法进一步剔除冗余基因，通过SVM验证特征基因子集的分类性能，结果显示该方法能够提高分类精度和特征选择的稳定性。

2.2.4 混合式和集成式方法在肿瘤关键基因筛选中的应用　随着特征选择技术不断发展，混合式和集成式方法也逐渐成为研究热点，这些方法结合不同的特征选择策略以提高效率和性能。混合式方法组合不同的特征选择技术，通过充分发挥各自的优势，以达到更好的特征选择效果。集成式方法则将多个特征选择算法集成在一起，以综合它们的选择结果。这种方法可以采用投票、加权平均等策略，将不同算法的输出融合，从而减小个别算法的偏差，提高鲁棒性。任雨冬等[33]使用一种典型的混合式特征选择方法，通过结合不同算法的优点，能够更精确地筛选出关键基因。张靖等[48]在通过信噪比去除无关基因后，使用迭代LASSO算法在5个肿瘤基因数据集上进行特征选择，其在4个分类器上的性能对比信噪比和标准LASSO算法均达到最高精度。陈涛等[49]提出一种Relief-F算法结合改进的邻域粗糙集算法的混合式特征选择算法，对比混合前的两种算法，能够以更少的基因获得更好的分类效果。叶明全等[50]提出一种结合FCBF和SVM-RFE的特征选择算法，使用对称不确定性作为特征权重的衡量标准，使用FCBF算法快速剔除无关基因，然后使用SVM-RFE方法进一步减少冗余基因，该算法将5个肿瘤基因数据集的关键基因筛选至个位数，且在4个分类器上均取得较好性能。Prabhakar等[22]先以相关系数、T检验和Kruskal-Wallis检验等标准的统计学方法筛选特征基

因，将特征基因数量从15 154个筛选到5 000个，再使用4种合适的随机优化算法对特征基因子集进一步优化筛选，最后特征子集精简到50~150个。Momenzadeh等[51]使用隐式马尔科夫模型集成基于巴氏距离、熵、受试者特征曲线、*T*检验和Wilcoxon 5种算法，通过这些算法计算综合基因排名进行特征选择，最终在3个数据集上减少特征到5~20个，并保持高分类精度。表2总结了近年来肿瘤基因筛选采用的特征选择算法。

表2 特征选择算法总结

Table 2 Summary on feature selection algorithms

方式	文献	算法描述
过滤式	文献[25]	统计学方法
	文献[31]	非参数检验方法
	文献[33]	使用高斯核函数和欧氏距离函数改进SAM的方法
包裹式	文献[34]	双正则化改进SVM
	文献[35]	随机矩阵替换法改进SVM
	文献[38]	判别熵改进PSO
	文献[39]	聚类结合PSO
	文献[42]	结合Relief-F和决策树的APSO算法
	文献[46]	使用Binary Dragonfly优化SVM-RFE后的子集
嵌入式	文献[36]	Monte Carlo采样方法结合线性核SVM
	文献[40]	PSO融合改进K-means
	文献[41]	LASSO结合BPSO算法
	文献[44]	使用互信息做度量的遗传算法
	文献[45]	GA嵌入学习自动机
混合式	文献[43]	IG混合遗传算法
	文献[47]	IG混合灰狼优化算法
	文献[48]	信噪比混合迭代LASSO
	文献[49]	Relief-F混合邻域粗糙集
	文献[50]	FCBF混合SVM-RFE
集成式	文献[22]	统计学方法集成多种随机优化算法
	文献[51]	隐式马尔科夫模型集成多种特征选择算法

2.3 AI在肿瘤基因表达数据中的其他应用

2.3.1 AI在肿瘤基因调控网络中的应用 AI在揭示肿瘤关键基因的相互作用方面发挥着关键作用，通过构建基因调控网络（Gene Regulatory Network, GRN），有助于深入理解肿瘤发生和发展的分子机制，为癌症生物标志物的鉴定提供重要支持[52]。Wang等[53]将机器学习应用于GRN重新布线的工作中，以识别乳腺癌生物标志物，该模型使用差异基因调控网络检测功能模块，并使用递归特征消除逻辑回归分类器来单独选择每个模块中的生物标志物基因。Zhang等[54]提出TFmeta模型，根据特征重要性和梯度提升树模型推断代谢酶的基因调控网络，以发现控制癌症代谢重编程的转录因子。该模型在实验中对75对肺癌组织和正常肺组织数据进行测试，预测出19个关键转录因子，可能是糖酵解中心代谢途径代谢酶基因表达变化的主要调节因子，也可能是非小细胞肺癌患者糖酵解失调的基础。此外，还可以使用深度学习方法从空间结构或图形关系角度进行基因分析。Ressom等[55]利用RNN和两种群体智能算法从时序基因表达数据推断GRN，并得到良好的结果。Yang等[56]利用深度神经网络在图像处理上的优势，对GRN进行重建，相较于传统方法，新方法获得更好的性能。Yuan等[57]提出一种共表达卷积神经网络，对GRN进行深度学习分析，在预测转录因子目标、识别疾病相关基因、因果推理等任务中都获得更好性能，且可以轻松扩展以集成其他类型的基因数据。

2.3.2 AI在肿瘤药物反应研究中的应用 通过分析肿瘤基因表达数据，AI可用于预测患者对特定药物的反应，从而实现药物研究和个体化治疗。Iorio等[58]将29个组织中的11 289个肿瘤分子数据与256种药物的敏感性进行相关性分析，揭示能够增加对药物敏感性的基因变化组合，机器学习方法展示在预测药物反应时不同数据类型的相对重要性。Sharifi-Noghabi等[59]以肿瘤基因数据作为输入，以预测剂量反应曲线和最大抑制浓度作为输出，在多个泛癌数据集上进行实验，实验结果表明预测剂量反应曲线比最大抑制浓度更适合作为药物敏感性测量方法。Yadav等[60]提出一种定量药物敏感性评分系统，数学模型估计法和连续插值法使得该方法对于不同来源的数据具有鲁棒性，该系统能够用于个体化抗癌治疗的药物反应预测。Smirnov等[61]开发一个用于分析大规模药物基因组学数据集的R包PharmacoGx，可以轻松高效地对药物敏感性反应进行预测。

3 总结与展望

本文讨论肿瘤基因表达数据研究中常用的机器学习方法和特征选择算法并做了简要梳理。肿瘤基因表达数据的分析在现代医学研究和临床实践中扮演着重要的角色，但这些数据的高维度、多噪声和小样本特点给分析带来挑战。AI技术为克服这些挑战提供有效的工具，特别是机器学习和特征选择算法，在肿瘤分类预测、关键基因筛选等方面的应用取得显著的进展，能够从基因表达数据中挖掘有用的信

息，帮助临床医生更好地了解肿瘤特性。目前该领域研究仍存在一些亟待解决的问题。

3.1 预处理方法的完善

随着基因芯片制备技术的进步和各个数据库的丰富，多个基因芯片的联合分析已经成为一种趋势。然而，由于不同来源的基因芯片制备方法通常遵循不同的标准，因此针对各芯片的预处理方法标准尚未达成一致，这导致了批次效应的存在，成为需要重点解决的问题。

3.2 更多深度学习模型的应用

目前，大多数应用于肿瘤基因数据进行肿瘤分类预测的模型仍以传统机器学习为主，深度学习模型的应用相对较少且局限于简单的神经网络。近年来，深度学习技术迅猛发展，具有处理高维复杂数据的能力，尤其是基于时序信息的长短期记忆网络和引入了注意力机制的复杂模型。这些模型展现出更大的潜力，研究人员有待充分挖掘和探索其在肿瘤基因表达数据分析中的应用。

3.3 特征选择算法的效率提高

现今的特征选择算法主要采用优化算法替代人为指定阈值的方式。然而，优化算法的寻优策略通常需要多次迭代，每次迭代都伴随着模型的训练，对计算机性能提出一定的要求。因此，探索更有效的方法以提高算法效率、降低时间复杂度至关重要。

3.4 丰富其他领域的应用

除了肿瘤分类预测和肿瘤关键基因筛选，AI在肿瘤基因表达数据中的研究还包括GRN研究和药物反应研究，但应用并不广泛，技术较落后，丰富这些领域的研究将是未来的趋势。此外，未来的研究方向还应包括生存分析、多数据类型融合等。例如，将AI技术应用于生存分析模型，比如Cox比例风险模型，结合基因表达数据，对患者的生存时间进行预测，这有助于了解不同基因表达模式与患者生存状况的关系。将基因表达数据与其他生物信息数据（如基因突变、蛋白质互作、病理图像等）结合起来，通过AI方法挖掘更深层次的信息，为肿瘤研究提供更全面的视角。

【参考文献】

[1] 刘宗超, 李哲轩, 张阳, 等. 2020全球癌症统计报告解读[J]. 肿瘤综合治疗电子杂志, 2021, 7(2): 1-13.
Liu ZC, Li ZX, Zhang Y, et al. Interpretation on the report of Global Cancer Statistics 2020[J]. Journal of Multidisciplinary Cancer Management (Electronic Version), 2021, 7(2): 1-13.

[2] Miller MB, Tang YW. Basic concepts of microarrays and potential applications in clinical microbiology[J]. Clin Microbiol Rev, 2009, 22(4): 611-633.

[3] Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data[J]. Stat Appl Genet Mol Biol, 2009, 8(1): 28.

[4] Topol EJ. Welcoming new guidelines for AI clinical research[J]. Nat Med, 2020, 26(9): 1318-1320.

[5] Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine[J]. Nat Med, 2022, 28(1): 31-38.

[6] Fodor SP, Read JL, Pirrung MC, et al. Light-directed, spatially addressable parallel chemical synthesis[J]. Science, 1991, 251(4995): 767-773.

[7] 杨耀. 肿瘤基因数据的特征选择算法研究[D]. 兰州: 甘肃中医药大学, 2022.
Yang Y. Research on feature selection algorithm of tumor gene data[D]. Lanzhou: Gansu University of Chinese Medicine, 2022.

[8] The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EN, et al. The cancer genome atlas pan-cancer analysis project[J]. Nat Genet, 2013, 45(10): 1113-1120.

[9] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets-update[J]. Nucleic Acids Res, 2013, 41(D1): D991-D995.

[10] Sarkans U, Füllgrabe A, Ali A, et al. From arrayexpress to biostudies[J]. Nucleic Acids Res, 2021, 49(D1): D1502-D1506.

[11] Luo YH, Hitz BC, Gabdank I, et al. New developments on the encyclopedia of DNA elements (ENCODE) data portal[J]. Nucleic Acids Res, 2020, 48(D1): D882-D889.

[12] Kanehisa M, Furumichi M, Sato Y, et al. KEGG for taxonomy-based analysis of pathways and genomes[J]. Nucleic Acids Res, 2023, 51(D1): D587-D592.

[13] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(5439): 531-537.

[14] Obermeyer Z, Emanuel EJ. Predicting the future-big data, machine learning, and clinical medicine[J]. N Engl J Med, 2016, 375(13): 1216-1219.

[15] Paixão GM, Santos BC, Araujo RM, et al. Machine learning in medicine: review and applicability[J]. Arq Bras Cardiol, 2022, 118(1): 95-102.

[16] Kriegeskorte N, Golan T. Neural network models and deep learning[J]. Curr Biol, 2019, 29(7): R231-R236.

[17] 张敏. 乳腺癌特征基因的筛选及预测[D]. 蚌埠: 安徽财经大学, 2020.
Zhang M. Screening of characteristic genes and prediction in breast cancer[D]. Bengbu: Anhui University of Finance and Economics, 2020.

[18] 王连喜, 蒋盛益. 一种基于特征聚类的特征选择方法[J]. 计算机应用研究, 2015, 32(5): 1305-1308.
Wang LX, Jiang SY. Novel feature selection method based on feature clustering[J]. Application Research of Computers, 2015, 32(5): 1305-1308.

[19] Furey TS, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10): 906-914.

[20] Gao LY, Ye MQ, Wu CR. Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony[J]. Molecules, 2017, 22(12): 2086.

[21] Huynh PH, Nguyen VH, Do TN. A combined enhancing and feature extraction algorithm to improve learning accuracy for gene expression classification[C]//Future Data and Security Engineering. Cham: Springer International Publishing, 2019: 255-273.

[22] Prabhakar SK, Lee SW. An integrated approach for ovarian cancer classification with the application of stochastic optimization[J]. IEEE Access, 2020, 8: 127866-127882.

[23] Ludwig SA, Picek S, Jakobovic D. Classification of cancer data: analyzing gene expression data using a fuzzy decision tree algorithm[M]//Kahraman C, Topcu YI. Operations Research Applications in Health Care Management. Cham: Springer International Publishing, 2018: 327-347.

[24] Tabares-Soto R, Orozco-Arias S, Romero-Cano V, et al. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data[J]. PeerJ Comput Sci, 2020, 6: e270.

[25] Alagukumar S, Lawrance R. Classification of microarray gene expression data using associative classification[C]//2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16). Piscataway, NJ, USA: IEEE, 2016: 1-8.

[26] Nguyen T, Nahavandi S. Modified AHP for gene selection and cancer classification using type-2 fuzzy logic[J]. IEEE T Fuzzy Syst, 2016, 24(2): 273-287.
[27] Nguyen T, Khosravi A, Creighton D, et al. A novel aggregate gene selection method for microarray data classification[J]. Pattern Recognit Lett, 2015, 60-61: 16-23.
[28] Zeebaree DQ, Haron H, Abdulazeez AM. Gene selection and classification of microarray data using convolutional neural network[C]//2018 International Conference on Advanced Science and Engineering (ICOASE). Piscataway, NJ, USA: IEEE, 2018: 145-150.
[29] Xiao YW, Wu J, Lin ZL, et al. A deep learning-based multi-model ensemble method for cancer prediction[J]. Comput Methods Programs Biomed, 2018, 153: 1-9.
[30] Kang CZ, Huo YH, Xin LH, et al. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine[J]. J Theor Biol, 2019, 463: 77-91.
[31] Maniruzzaman M, Jahanur Rahman M, Ahammed B, et al. Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms[J]. Comput Methods Programs Biomed, 2019, 176: 173-193.
[32] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response[J]. Proc Natl Acad Sci U S A, 2001, 98(9): 5116-5121.
[33] 任雨冬, 陆震, 李婧惟, 等. 基因表达数据中加权SAM法的基因选择和分类预测研究[J]. 实用预防医学, 2020, 27(12): 1537-1540.
Ren YD, Lu Z, Li JW, et al. Gene selection and classification prediction of weighted SAM method in gene expression data[J]. Practical Preventive Medicine, 2020, 27(12): 1537-1540.
[34] 秦传东, 刘三阳. 基于双重正则化支持向量机的肿瘤基因选择[J]. 吉林大学学报(工学版), 2013, 43(1): 192-197.
Qin CD, Liu SY. Tumor gene selection based on double regularized support vector machine[J]. Journal of Jilin University (Engineering and Technology Edition), 2013, 43(1): 192-197.
[35] 谢志伟, 王志明, 骆剑锋. 基于RD-SVM的肿瘤信息基因选择算法[J]. 计算机应用与软件, 2015, 32(5): 310-313.
Xie ZW, Wang ZM, Luo JF. Tumour informative gene selection algorithm based on RD-SVM[J]. Computer Applications and Software, 2015, 32(5): 310-313.
[36] 张世芝, 张明锦. 基于SVM的嵌入式特征基因选择方法研究[J]. 计算机与应用化学, 2016, 33(1): 85-88.
Zhang SZ, Zhang MJ. Study on SVM-based embedded feature selection method[J]. Computers and Applied Chemistry, 2016, 33(1): 85-88.
[37] Kennedy J, Eberhart R. Particle swarm optimization[C]//Proceedings of ICNN'95-International Conference on Neural Networks. Piscataway, NJ, USA: IEEE, 1995: 1942-1948.
[38] 关健, 韩飞, 杨善秀. 基于粒子群优化和判别熵信息的基因选择算法[J]. 计算机工程, 2013, 39(11): 187-190.
Guan J, Han F, Yang SX. Gene selection algorithm based on particle swarm optimization and J-divergence entropy information[J]. Computer Engineering, 2013, 39(11): 187-190.
[39] 刘金勇, 郑恩辉, 陆慧娟. 基于聚类和微粒群优化的基因选择方法[J]. 数据采集与处理, 2014, 29(1): 83-89.
Liu JY, Zheng EH, Lu HJ. Gene selection based on clustering method and particle swarm optimization[J]. Journal of Data Acquisition and Processing, 2014, 29(1): 83-89.
[40] 杜洪波, 白阿珍, 朱立军. 改进的K-means融合微粒群优化的基因选择方法[J]. 沈阳工程学院学报(自然科学版), 2018, 14(1): 66-70.
Du HB, Bai AZ, Zhu LJ. Gene selection method based on the fusion of improved K-means algorithm and particle swarm optimization[J]. Journal of Shenyang Institute of Engineering (Natural Sciences), 2018, 14(1): 66-70.
[41] 熊颖. 基于Lasso和二进制粒子群优化的基因选择方法研究[D]. 镇江: 江苏大学, 2019.
Xiong Y. A study of gene selection method based on Lasso and binary particle swarm optimization[D]. Zhenjiang: Jiangsu University, 2019.
[42] 叶超超, 潘巨龙. 面向基因选择的结合Relief-F和决策树的APSO算法[J]. 计算机应用研究, 2019, 36(2): 395-398.
Ye CC, Pan JL. New APSO algorithm for gene selection combined with Relief-F and decision tree[J]. Application Research of Computers, 2019, 36(2): 395-398.
[43] Salem H, Attiya G, El-Fishawy N. Classification of human cancer diseases by gene expression profiles[J]. Appl Soft Comput, 2017, 50: 124-134.
[44] 魏莎莎, 陆慧娟, 安春霖, 等. 一种基于互信息最大化的模型无关基因选择方法[J]. 计算机科学, 2014, 41(9): 243-247.
Wei SS, Lu HJ, An CL, et al. Model-free gene selection method based on maximum mutual information[J]. Computer Science, 2014, 41(9): 243-247.
[45] 萧秋兰, 郑虹. 微阵列肿瘤分类混合基因选择算法[J]. 长春工业大学学报, 2019, 40(6): 540-546.
Xiao QL, Zheng H. A hybrid gene selection algorithm for microarray cancer classification[J]. Journal of Changchun University of Technology, 2019, 40(6): 540-546.
[46] Medjahed SA, Saadi TA, Benyettou A, et al. Kernel-based learning and feature selection analysis for cancer diagnosis[J]. Appl Soft Comput, 2017, 51: 39-48.
[47] Abdelnabi ML, Jasim MW, El-Bakry HM, et al. Breast and colon cancer classification from gene expression profiles using data mining techniques[J]. Symmetry (Basel), 2020, 12(3): 408.
[48] 张靖, 胡学钢, 李培培, 等. 基于迭代Lasso的肿瘤分类信息基因选择方法研究[J]. 模式识别与人工智能, 2014, 27(1): 49-59.
Zhang J, Hu XG, Li PP, et al. Informative gene selection for tumor classification based on iterative Lasso[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(1): 49-59.
[49] 陈涛, 洪增林, 邓方安. 基于优化的邻域粗糙集的混合基因选择算法[J]. 计算机科学, 2014, 41(10): 291-294.
Chen T, Hong ZL, Deng FA. Hybrid gene selection algorithm based on optimized neighborhood rough set[J]. Computer Science, 2014, 41(10): 291-294.
[50] 叶明全, 高凌云, 伍长荣, 等. 基于对称不确定性和SVM递归特征消除的信息基因选择方法[J]. 模式识别与人工智能, 2017, 30(5): 429-438.
Ye MQ, Gao LY, Wu CR, et al. Informative gene selection method based on symmetric uncertainty and SVM recursive feature elimination[J]. Pattern Recognition and Artificial Intelligence, 2017, 30(5): 429-438.
[51] Momenzadeh M, Sehhati M, Rabbani H. A novel feature selection method for microarray data classification based on hidden Markov model[J]. J Biomed Inform, 2019, 95: 103213.
[52] Yan WY, Xue WJ, Chen JJ, et al. Biological networks for cancer candidate biomarkers discovery[J]. Cancer Inform, 2016, 15(Suppl 3): 1-7.
[53] Wang YJ, Liu ZP. Identifying biomarkers for breast cancer by gene regulatory network rewiring[J]. BMC Bioinformatics, 2022, 22(Suppl 12): 308.
[54] Zhang Y, Zhang XF, Lane AN, et al. Inferring gene regulatory networks of metabolic enzymes using gradient boosted trees[J]. IEEE J Biomed Health Inform, 2020, 24(5): 1528-1536.
[55] Ressom HW, Zhang Y, Xuan J, et al. Inference of gene regulatory networks from time course gene expression data using neural networks and swarm intelligence[C]//proceedings of the 2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology. Piscataway, NJ, USA: IEEE, 2006: 1-8.
[56] Yang Y, Fang QW, Shen HB. Predicting gene regulatory interactions based on spatial gene expression data and deep learning[J]. PLoS Comput Biol, 2019, 15(9): e1007324.
[57] Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data[J]. Proc Natl Acad Sci U S A, 2019, 116(52): 27151-27158.
[58] Iorio F, Knijnenburg TA, Vis DJ, et al. A landscape of pharmacogenomic interactions in cancer[J]. Cell, 2016, 166(3): 740-754.
[59] Sharifi-Noghabi H, Jahangiri-Tazehkand S, Smirnov P, et al. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models[J]. Brief Bioinform, 2021, 22(6): bbab294.
[60] Yadav B, Pemovska T, Szwajda A, et al. Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies[J]. Sci Rep, 2014, 4(1): 5193.
[61] Smirnov P, Safikhani Z, El-Hachem N, et al. PharmacoGx: an R package for analysis of large pharmacogenomic datasets[J]. Bioinformatics, 2016, 32(8): 1244-1246.

（编辑：陈丽霞）