

**本科毕业设计（论文）开题报告**

|  |  |
| --- | --- |
| **题目：** | **阿兹海默症生物标志物挖掘系统设计** |
|  |  |

|  |  |  |
| --- | --- | --- |
| **学号** | ： | 2101620112 |
| **姓名** | ： | 刘朝 |
| **学院** | ： | 人工智能学院 |
| **专业** | ： | 机器人工程 |
| **指导教师** | ： | 蓝朝旺 |
| **指导教师职称** | ： | 讲师 |

开题报告填写要求

1．开题报告作为毕业设计（论文）答辩委员会对学生答辩资格审查的依据材料之一。此报告应在指导教师指导下，由学生在毕业设计（论文）工作前期内完成，经指导教师签署意见审查后生效。

 2．开题报告内容必须用黑墨水笔工整书写，或按教务处统一设计的电子文档标准格式打印，禁止打印在其它纸上后剪贴，完成后应及时交给指导教师签署意见。

 3．学生查阅资料的参考文献应在10篇及以上（不包括辞典、手册）。

 4．有关年月日等日期的填写，应当按照国标GB/T 7408—94《数据元和交换格式、信息交换、日期和时间表示法》规定的要求，一律用阿拉伯数字书写。如“2012年12月25日”或“2012-12-25”。

|  |
| --- |
| 1．本课题的研究内容、重点及难点 |
| 一．主要内容：  在阿尔茨海默病（Alzheimer's Disease, AD）的研究中，随着技术的进步，特别是单细胞RNA测序技术的应用，研究者们能够更深入地了解疾病的分子机制。然而，由于miRNA和CSF组织复杂且数据维度庞大，如何有效地从众多基因和特征中筛选出与疾病相关的关键信息，成为研究的一个重要挑战。因此，特征选择（Feature Selection）作为一种数据预处理和分析方法，能够帮助研究者从大量的基因表达数据中识别出与阿尔茨海默病相关的最具信息量的特征，为疾病的早期诊断、机制研究和预后治疗提供支持。  特征选择通过评估不同特征的相关性、重要性和冗余性，选择出对疾病预测或分类最有用的特征。对于阿尔茨海默病来说，这一过程尤其重要，因为病理进程通常表现为大脑中多个细胞群体之间复杂的相互作用和基因表达的微妙变化。传统的特征选择方法包括过滤法（Filter Methods）、包裹法（Wrapper Methods）和嵌入法（Embedded Methods），每种方法都有其优缺点。近年来，集成特征选择方法逐渐成为研究的热点，因为它结合了多种特征选择策略，能够提高选择准确性和稳定性。在阿尔茨海默病的单细胞RNA测序数据分析中，集成特征选择方法尤为重要。通过集成不同的特征选择算法，研究人员能够综合多个模型的优点，增强特征选择的鲁棒性和精度，进一步加深对疾病的理解。例如，通过结合不同来源的特征选择方法，可以在单细胞数据中挖掘出潜在的生物标志物，为早期诊断提供支持，同时也有助于为个体化医疗提供精准的特征信息。通过识别和验证这些重要的基因或细胞亚群，我们可以更清楚地了解阿尔茨海默病的发病机制，进而推动治疗策略的发展。  本课题旨在通过对阿尔茨海默病患者大脑组织单细胞测序数据进行集成特征选择，探索与疾病发生密切相关的基因表达模式和细胞亚群。这一过程将为阿尔茨海默病的早期诊断、精准治疗以及个性化医疗提供新的视角和方法，推动该领域的研究进展。   1. 本选题的重点如下：   (1)疾病机制的探索：通过单细胞RNA测序数据结合集成特征选择方法，深入挖掘阿尔茨海默病相关的基因表达模式和细胞亚群，寻找疾病的有效生物标志物，并尝试推广至类似领域。  (2)算法的实现与创新：采用集成特征选择方法，使用三种不同的集成策略，提高特征筛选的鲁棒性，精度以及有效性，为阿尔兹海默症预测和分子机制研究提供支持。  (3)医学实践的推动：识别关键基因在阿尔茨海默病中的差异表达，为阿尔茨海默病的早期诊断、精准治疗和预后诊断提供相关依据。  三．本选题的难点如下：  (1)高维数据处理：阿尔茨海默病单细胞RNA测序数据维度庞大，数据复杂度高，从中筛选出与疾病相关的关键信息具有挑战性。  (2)特征选择模型的优化：如何有效地集成多种特征选择算法，同时平衡准确性与稳定性，减少过拟合和特征冗余。  (3)生物学验证与解释：筛选出的基因和细胞亚群如何在生物学层面进行验证，如KEGG通路富集分析，查阅相关资料及文献，为研究结果转化为对疾病机制的具体解释。 |
| 2．准备情况（已查阅的参考文献或进行的调研） |
| 一．调研情况  阿尔茨海默病（Alzheimer's Disease, AD）是一种复杂的神经退行性疾病，其发病机制涉及多种细胞类型和分子途径。随着单细胞RNA测序（scRNA-seq）技术的应用，研究人员能够在单细胞水平上解析AD患者大脑组织的基因表达谱，揭示疾病相关的细胞异质性和分子机制[1]。然而，scRNA-seq数据通常具有高维度和高噪声的特点，特别是在miRNA和脑脊液数据样本中，数据复杂性进一步增加。因此，从这些数据中有效提取与AD相关的关键特征，成为研究中的重要挑战。特征选择（Feature Selection）作为一种数据预处理和分析方法，旨在从大量基因表达数据中识别出最具信息量的特征，为疾病的早期诊断、机制研究和预后治疗提供支持[2][3]。  **特征选择方法在AD研究中的应用**  特征选择方法主要包括过滤法（Filter Methods）、包裹法（Wrapper Methods）和嵌入法（Embedded Methods）。每种方法都有其优缺点，研究者常根据具体数据特点选择合适的方法[4]。近年来，集成特征选择方法逐渐成为研究热点。通过结合多种特征选择策略，集成方法能够提高选择的准确性和稳定性[5][9][10][11][12][13]。例如，结合三种不同的特征选择方法（BSPC、随机森林、DUBStepR），通过分析单细胞RNA测序数据（GSE103334），研究了神经退行性疾病模型（CK-p25）中微胶细胞活化过程中基因表达的动态变化[6]。  **单细胞RNA测序在AD研究中的进展**  单细胞RNA测序技术在AD研究中取得了显著进展。通过该技术，研究人员能够分析多种脑细胞类型，包括神经元、少突胶质细胞、星形胶质细胞和小胶质细胞，发现这些细胞类型在AD患者中表现出明显的基因表达差异。 此外，单细胞分析有助于对AD大脑中受到扰动的各种细胞过程进行细致入微的描述，揭示健康衰老和认知能力下降之间的差异，并突出显示与AD相关的细胞类型特异性分子程序[7]。  尽管取得了上述进展，基于深度学习的特征选择方法在AD单细胞RNA测序数据分析中仍面临挑战[8]。例如，如何选择适当的网络结构和参数设置，以及如何解决数据稀疏性和噪声干扰等问题，仍需进一步研究。未来的研究需要优化和改进这些方法，以满足更高精度、更准确的AD单细胞数据分析需求。  二．现有设备和实验条件：  现有设备：计算机一台  三．已查阅的文献资料：   1. Neff R A, Wang M, Vatansever S, et al. Molecular subtyping of Alzheimer’s disease using RNA sequencing data reveals novel mechanisms and targets[J]. Science advances, 2021, 7(2): eabb5398. 2. Alshamlan H, Omar S, Aljurayyad R, et al. Identifying effective feature selection methods for Alzheimer’s disease biomarker gene detection using machine learning[J]. Diagnostics, 2023, 13(10): 1771. 3. Paplomatas P, Krokidis M G, Vlamos P, et al. An ensemble feature selection approach for analysis and modeling of transcriptome data in alzheimer’s disease[J]. Applied Sciences, 2023, 13(4): 2353. 4. Stańczyk U. Feature evaluation by filter, wrapper, and embedded approaches[J]. Feature selection for data and pattern recognition, 2015: 29-44. 5. Bolón-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: A review and future trends[J]. Information fusion, 2019, 52: 1-12. 6. Paplomatas P, Krokidis M G, Vlamos P, et al. An ensemble feature selection approach for analysis and modeling of transcriptome data in alzheimer’s disease[J]. Applied Sciences, 2023, 13(4): 2353. 7. Jin W, Pei J J, Roy J R, et al. Comprehensive review on single-cell RNA sequencing: A new frontier in Alzheimer's disease research[J]. Ageing Research Reviews, 2024: 102454. 8. Neff R A, Wang M, Vatansever S, et al. Molecular subtyping of Alzheimer’s disease using RNA sequencing data reveals novel mechanisms and targets[J]. Science advances, 2021, 7(2): eabb5398. 9. HK R, HA D, MS P K, et al. A robust framework for Alzheimer’s disease detection and staging: incorporating multi-feature integration, MRMR feature selection, and Random Forest classification[J]. Multimedia Tools and Applications, 2024: 1-29. 10. Faisal F U R, Khatri U, Kwon G R. Diagnosis of Alzheimer's disease using combined feature selection method[J]. Journal of korea multimedia society, 2021, 24(5): 667-675. 11. Osipowicz M, Wilczynski B, Machnicka M A, et al. Careful feature selection is key in classification of Alzheimer’s disease patients based on whole-genome sequencing data[J]. NAR Genomics and Bioinformatics, 2021, 3(3): lqab069. 12. Mahendran N, PM D R V. A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease[J]. Computers in Biology and Medicine, 2022, 141: 105056. 13. El-Gawady A, Tawfik B B S, Makhlouf M A. Hybrid Feature Selection Method for Predicting Alzheimer’s Disease Using Gene Expression Data[J]. CMC-COMPUTERS MATERIALS & CONTINUA, 2023, 74(3): 5559-5572. |
| 3、实施方案、进度实施计划及预期提交的毕业设计资料 |
| 一.实施方案:  （1）数据预处理  针对开源数据库（如NCBI、GEO、ADNI，ADSP）获取的原始数据进行预处理，包括  缺失值处理，标准化，非数值形数据编码，平衡数据标签。经过这些步骤可以将复  杂、异构的原始数据转化为高质量的、统一格式的标准化数据集，为后续特征选择、  模型训练和评估提供坚实的基础。  （2）特征提取与集成方法  特征选择分为三种主要方法：过滤式：通过统计指标筛选对目标变量影响较大的特  征（CFS，MRMR），形成特征子集1。包裹式：结合特定模型的性能（RFE）来选择特  征，形成特征子集2。嵌入式：在模型训练过程中通过正则化方法选择特征（如LASSO  回归等），形成特征子集3。将三个特征子集按照不同的集成策略得到一个新的集成  后的子集，再通过RF，SVM，GBDT做分类评估。  （3）分类模型构建与评估  模型性能评估采用多种指标，包括 F1-Score、AUC和 MCC。通过特征评估和模型性  能的综合分析，进一步优化特征选择过程，提升模型的泛化能力。  （4）功能验证与应用  为了验证最终特征集的生物学意义或实际应用价值，采用文献验证，KEGG通路分析  和泛化能力测试，尝试为结果提供解释性和可靠性支持。  总体架构如图一所示    图 1 项目总体组织结构  二．进度实施计划：  （1）第1～2周：查阅各种文献和相关资料，了解课题发展方向及成果，编写毕业设计开题报告；  （2）第3～4周：大量阅读相关文献资料，学习总结已有的方法并探索可用于解决问题的方法；  （3）第5～7周：搜集阿尔兹海默症生物标志物数据集以及后续可能用到的泛化性验证数据集合，进行前期数据预处理，利用已有算法做预实验并分析结果；  （4）第8～12周：算法研究与设计，设计并实现三类传统的特征选择算法中的各一种，完成特征选择算法算法的实现和调试，评估特征选择算法的性能，并进行参数调优，实现三种ensemble策略，做一个不同领域的泛化性实验，并记录实验过程，分析结果，作为毕业论文的实验部分的参考。  （5）第13～15周：毕业论文撰写及润色；  （6）第16周：毕业设计答辩；  三．预期提交的毕业设计资料：  （1）毕业设计开题报告1份；  （2）英文翻译材料1份(包括不少于2万字符的英文原文和译文);  （3）完成相关软件系统一套(包含源程序清单,用户使用说明书);  （4）毕业设计说明书1份(不少于1.5万字，附中英文摘要，其中英文摘要300~500个英文单词)。 |
| 指导教师意见 |
| 指导教师： |
| 开题小组意见 |
| 开题小组组长签字： |
| 院系审核意见 |
| 院系主管领导签字： |