

基于随机森林与模糊森林的精神障碍高维表型特征筛选与分类研究

王岩¹, 张子钰², 王复高^{3*}

1. 亚太科技大学, 马来西亚 吉隆坡 57000

2. 山东大学, 山东 济南 250100

3. 沂水县园林环卫保障服务中心, 山东 沂水 276400

DOI:10.61369/ASDS.2026050016

摘 要 : 精神障碍具有较强的临床异质性, 不同疾病类别之间常存在症状重叠和表型交叉。如何从高维表型数据中筛选具有分类意义的关键变量, 是精神障碍数据驱动研究中的重要问题。本文以 Consortium for Neuropsychiatric Phenomics 表型数据集为研究对象, 围绕精神分裂症、双相情感障碍和注意缺陷多动障碍三类精神障碍, 构建高维表型特征筛选与分类分析框架。针对表型变量数量多、变量间相关性较强等问题, 本文分别采用随机森林和模糊森林方法进行特征重要性排序与分类建模, 并设置三类疾病分类和单病种识别两类实验场景。结果显示, bprs_positive、adult_totalseverity、ns2、etotal 和 mpq_score 等变量在不同分类任务中具有较高重要性。随机森林和模糊森林在重要特征识别方面具有一定一致性, 但二者在不同疾病识别场景中的分类表现存在差异。研究结果表明, 基于随机森林和模糊森林的特征筛选方法能够为精神障碍高维表型数据分析提供有效工具, 也为后续开展精神障碍表型特征识别与分类研究提供了参考。

关 键 词 : 精神障碍; 随机森林; 模糊森林; 特征选择; 高维表型数据; 分类分析

Feature Selection and Classification of High-Dimensional Phenotypic Data for Mental Disorders Based on Random Forest and Fuzzy Forest

Wang Yan¹, Zhang Ziyu², Wang Fugao^{3*}

1.Asia Pacific University of Technology & Innovation, Kuala Lumpur, Malaysia 57000

2.Shandong University, Jinan, Shandong 250100

3.Yishui County Landscaping and Environmental Sanitation Support Service Center, Yishui, Shandong 276400

Abstract : Mental disorders are characterized by substantial clinical heterogeneity, with symptom overlap and phenotypic intersections often observed across different diagnostic categories. Identifying classification-relevant variables from high-dimensional phenotypic data is therefore an important issue in data-driven research on mental disorders. This study uses the phenotypic dataset from the Consortium for Neuropsychiatric Phenomics as the research object and focuses on three types of mental disorders: schizophrenia, bipolar disorder, and attention-deficit/hyperactivity disorder. A feature selection and classification framework for high-dimensional phenotypic data is constructed. To address the large number of phenotypic variables and the strong correlations among them, Random Forest and Fuzzy Forest are applied for feature importance ranking and classification modeling. Two types of experimental scenarios are designed, including three-class disease classification and single-disorder identification. The results show that bprs_positive, adult_totalseverity, ns2, etotal, and mpq_score have relatively high importance in different classification tasks. Random Forest and Fuzzy Forest show a certain degree of consistency in identifying important features, while their classification performance differs across disease identification scenarios. The findings suggest that feature selection methods based on Random Forest and Fuzzy Forest can provide useful tools for analyzing high-dimensional phenotypic data of mental disorders and offer a reference for further research on phenotypic feature identification and classification.

Keywords : mental disorders; Random Forest; Fuzzy Forest; feature selection; high-dimensional phenotypic data; classification analysis

作者简介:

王岩 (2000—), 男, 马来西亚亚太科技大学, 硕士研究生;

张子钰 (1998—), 女, 山东大学数学学院, 博士研究生;

通讯作者: 王复高 (1981—), 男, 沂水县园林环卫保障服务中心, 中级经济师。

引言

精神障碍具有较强的异质性，不同疾病类别之间常存在症状重叠、表型交叉和诊断边界模糊等问题。传统 DSM 诊断体系为精神疾病分类和临床判断提供了重要依据^[1]，但其以症状描述为核心的分类方式，仍难以充分解释患者在认知、行为、人格特质和神经心理表现等方面的复杂差异。已有研究指出，现有精神障碍分类体系在疾病边界、理论基础和神经科学解释方面仍存在一定局限^[2]。

随着数据驱动方法在精神健康研究中的应用不断深入，基于多维表型数据识别疾病相关特征和潜在分类模式，已成为精神障碍研究的重要方向^[3]。高维表型数据通常包含临床症状量表、人格特质、神经认知任务和神经心理评估等多类变量，具有变量数量多、类型复杂和相关性较强等特点。针对这一问题，本文采用随机森林和模糊森林方法，对精神分裂症、双相情感障碍和注意缺陷多动障碍进行特征筛选与分类分析，旨在比较两种方法在精神障碍高维表型数据中的应用效果，并识别具有分类意义的关键表型特征。

一、数据来源与预处理

（一）数据来源与样本构成

本文使用 Consortium for Neuropsychiatric Phenomics (CNP) 表型数据集作为研究对象^[4]。该数据集包含精神障碍患者和健康对照个体的多维表型信息，涵盖人口统计学资料、一般健康状况、临床症状量表、人格特质、神经认知任务和神经心理评估等内容。

本研究共纳入 272 名参与者，其中女性 117 名，男性 155 名；参与者平均年龄为 33.3 岁，标准差为 9.4。按照主要精神健康状况划分，样本包括精神分裂症患者 50 名、双相情感障碍患者 49 名、注意缺陷多动障碍患者 43 名，以及健康对照个体 130 名。该样本结构为比较不同精神障碍群体之间的表型差异提供了数据基础。

（二）表型变量构成

本文构建的多维表型数据共包含 1983 个候选特征。这些特征主要来自人口统计信息与一般健康信息、临床症状、人格特质、神经认知任务和神经心理评估等方面。其中，症状相关变量包括 Young Mania Rating Scale-C (YMRS-C)、Hamilton Psychiatric Rating Scale for Depression (HAM-D-28)、Scale for the Assessment of Negative Symptoms (SANS)、Scale for the Assessment of Positive Symptoms (SAPS) 和 Brief Psychiatric Rating Scale (BPRS) 等量表指标。

由于不同表型变量反映的心理和行为维度并不相同，本文在建模前将所有变量视为潜在候选特征，并通过后续特征筛选方法识别其中具有较高分类贡献的变量。这样可以避免在建模前过度依赖主观判断删除可能具有分类价值的表型信息。

（三）缺失值处理与变量初筛

原始表型数据中存在一定比例的缺失值。其中，部分临床症状量表主要面向精神障碍患者，因此在健康对照样本中缺失比例较高。考虑到本文重点关注精神障碍患者之间的表型差异及其分类特征，本文在相关分析中保留了这些能够反映疾病症状表现的变量，并在疾病分类任务中以精神分裂症、双相情感障碍和注意缺陷多动障碍三类患者为主要分析对象。

在变量初筛阶段，本文首先删除零方差变量、缺乏分析意义的二元变量以及明显冗余的特征，以减少无效信息对模型训练的

影响。对于保留变量中的缺失值，本文采用 K 近邻插补方法进行处理。该方法根据样本之间的相似性估计缺失观测值，能够在一定程度上保留变量之间的局部结构关系。

（四）高相关变量处理

精神障碍表型数据中存在较多相关性较强的变量。例如，不同症状量表可能反映相近的临床表现，不同人格或冲动性指标也可能对应相似的心理特征。高度相关变量可能影响随机森林模型中的变量重要性排序，使部分重要特征的重要性被分散或低估。

为降低共线性对特征筛选结果的影响，本文将相关系数大于 0.8 的变量提取为高相关变量子集，并在随机森林建模时暂时排除这些变量。与此同时，为避免遗漏具有实际解释价值的表型信息，本文在后续结果讨论中重新考察该高相关变量子集与模型筛选出的关键特征之间的关系，从而进一步分析重要表型变量的潜在含义。

二、随机森林和模糊森林特征选择

本文分别使用 R 语言中的 randomForest 包和 FuzzyForest 包构建随机森林模型与模糊森林模型，并在统一的数据划分和分类任务设置下比较两种方法的特征筛选结果与预测表现。测试集用于最终评价两类模型在不同分类任务中的预测误差，从而比较其分类效果和适用性。对于随机森林模型，本文主要利用袋外样本估计模型误差，因此未将验证集作为其主要调参依据；对于模糊森林模型，本文使用验证集评估模型表现并进行参数调节。

（一）随机森林

随机森林是一种基于集成学习思想的机器学习方法。该方法通过构建多棵决策树完成分类或回归任务，并利用多棵树的综合结果提高模型稳定性和预测能力^[5]。在高维表型数据分析中，随机森林不仅可以用于分类建模，还可以通过变量重要性指标对候选特征进行排序。由于该方法能够处理高维、非线性数据，因此在生物医学特征筛选和分类预测中具有较多应用^[6]。

本文采用 R 语言中的 randomForest 包构建随机森林模型^[7]。为确定合适的树数量参数，本文分别设置 250、500、1000 和 2000 棵树，并在每种设置下重复运行 10 次。实验结果显示，当树数量增加至约 1000 棵时，袋外误差趋于稳定。因此，本文最终采用 1000 棵树作为随机森林模型的主要参数设置。

在特征重要性评价方面，本文采用平均准确率下降和平均基

尼下降两个指标。平均准确率下降用于衡量某一变量被扰动或移除后模型预测准确率的下降程度。若删除某一变量后模型准确率明显下降,则说明该变量对分类结果具有较大贡献。平均基尼下降用于衡量变量在决策树节点分裂中对不纯度下降的平均贡献。本文综合两个指标对候选特征进行排序,并在模型简洁性和分类性能之间进行权衡,选择对分类任务贡献较高的关键变量。

(二) 模糊森林

模糊森林是在随机森林基础上发展起来的特征筛选方法,主要用于处理高维且变量相关性较强的数据^[8]。与传统随机森林直接在全部变量中计算重要性不同,模糊森林首先根据变量之间的相关结构将特征划分为若干模块,使同一模块内部变量之间具有较高相关性,而不同模块之间的相关性相对较低。随后,模型在每个模块内部使用递归特征消除随机森林进行筛选,并在保留下来的变量中进一步完成最终特征选择和重要性排序。

本文采用 R 语言中的 FuzzyForest 包构建模糊森林模型。由于模糊森林涉及多个参数,本文分别对树的数量、每轮迭代中删除特征的比例以及每个模块中保留特征的比例进行调节。具体而言,树的数量设置为 300、600 和 1000; drop_fraction 设置为 0.1、0.2 和 0.25; keep_fraction 设置为 0.05、0.1 和 0.15。每组参数组合重复运行 5 次,以比较不同设置下模型误差的变化情况。

实验结果显示,当树数量设置为 600、drop_fraction 设置为 0.2、keep_fraction 设置为 0.05 时,模型误差趋于稳定。因此,本文采用该参数组合作为模糊森林模型的主要设置。该方法能够在考虑变量相关结构的基础上筛选重要特征,有助于降低高维相关变量对特征排序结果的影响。

(三) 实验设计与评价方式

本文将数据集划分为训练集、测试集和验证集。其中,训练集占 60%,测试集占 20%,验证集占 20%。随机森林模型和模糊森林模型均在训练集上进行建模。对于随机森林模型,由于其内部能够通过袋外样本估计模型误差,因此本文主要使用袋外误差评价训练过程中的模型表现。对于模糊森林模型,本文使用验证集进行参数调节和模型表现评估。

为分析不同表型变量在精神障碍识别中的作用,本文设置两类实验场景。第一类为三类精神障碍分类任务,即同时区分精神分裂症、双相情感障碍和注意缺陷多动障碍。第二类为单病种识别任务,即分别将精神分裂症、双相情感障碍和注意缺陷多动障碍作为目标疾病,并将其与其他疾病进行二分类比较。通过这两类实验,可以同时考察模型在整体多分类任务中的表现,以及关键特征对不同疾病识别的贡献。

在模型比较方面,本文使用测试集进一步评价随机森林和模糊森林在不同分类任务中的预测误差。随机森林的袋外误差和模糊森林的验证误差用于反映模型在训练和调参阶段的表现;测试集预测误差则用于比较两种方法在独立数据上的分类效果。通过综合比较特征重要性排序、袋外误差、验证误差和测试误差,本文判断两种方法在不同精神障碍识别场景中的适用性。

三、实验结果

(一) 三类精神障碍分类结果

本文首先构建精神分裂症、双相情感障碍和注意缺陷多动障碍

三类分类任务,并分别使用随机森林和模糊森林进行特征筛选。表 1 和表 2 分别给出了两种模型在三分类任务中筛选出的重要特征。

表 1: 精神分裂症、双相情感障碍和 ADHD 分类结果(随机森林)

Importance Variables	Mean Decrease Accuracy	Mean Decrease Gini
bprs_positive	44.2	14.7
adult_totalseverity	35.9	13.6
ns2	22.3	8.12
etotal	20.4	9.12
mpq_score	19.9	9.54

表 2: 精神分裂症、双相情感障碍和 ADHD 分类结果(模糊森林)

Features	Variable Importance
bprs_positive	0.156
adult_totalseverity	0.137
ns2	0.073
etotal	0.060

由表 1 和表 2 可以看出,两种方法在三分类任务中筛选出的重要变量具有较高一致性。bprs_positive、adult_totalseverity、ns2 和 etotal 均同时出现在随机森林和模糊森林的重要特征列表中,说明这些变量在区分三类精神障碍时具有较稳定的分类贡献。其中,bprs_positive 和 adult_totalseverity 在两种模型中均排名靠前,表明症状相关变量在该分类任务中具有较强区分作用。

从模型误差看,随机森林在该任务中的袋外误差为 20.24%,模糊森林的验证误差为 8.823%。这一结果说明,在训练和验证阶段,模糊森林对三类精神障碍的分类表现较好。但由于袋外误差和验证误差的计算基础不同,仍需要结合测试结果进一步比较两种模型的预测能力。

(二) 精神分裂症识别结果

在精神分裂症识别任务中,本文将精神分裂症作为目标类别,并将双相情感障碍和 ADHD 归为其他疾病类别,构建二分类模型。表 3 和表 4 分别给出了随机森林和模糊森林筛选出的重要特征。

表 3: 精神分裂症识别中的重要特征和袋外误差估计(随机森林)

Importance Variables	Mean Decrease Accuracy	Mean Decrease Gini
bprs_positive	32.7	10.6
mpq_score	20.3	5.92
bprs_negative	18.1	4.45
adult_totalseverity	17.5	6.52
etotal	14.8	4.56
vmnm_main_acc	13.6	4.17

表 4: 精神分裂症识别结果(模糊森林)

Features	Variable Importance
bprs_positive	0.109
mpq_score	0.042
bprs_negative	0.034
etotal	0.029
adult_totalseverity	0.027

由表 3 和表 4 可以看出,bprs_positive 在两种模型中均排名第一,说明该变量对精神分裂症识别具有较强贡献。mpq_score、bprs_negative、etotal 和 adult_totalseverity 也同时出现在两种模型的重要特征列表中,表明这些变量在精神分裂症与其他精神

障碍的区分中具有一定稳定性。

从误差结果看，随机森林在精神分裂症识别中的袋外误差为7.14%，低于模糊森林的验证误差11.764%。这说明在该任务中，随机森林模型表现相对更好。结合特征排序结果可以认为，精神分裂症识别中的重要变量较为集中，两种方法筛选出的关键特征差异不大。

（三）双相情感障碍识别结果

在双相情感障碍识别任务中，本文将双相情感障碍作为目标类别，并将精神分裂症和ADHD归为其他疾病类别。表5和表6分别展示了随机森林和模糊森林的特征筛选结果。

表5：双相情感障碍识别中的重要特征和袋外误差估计（随机森林）

Importance Variables	Mean Decrease Accuracy	Mean Decrease Gini
ns2	25.1	10.5
bipollarii_mood	18.2	9.55
ymrs_score	16.3	9.08
hamd_28	12.8	8.77

表6：双相情感障碍识别结果（模糊森林）

Features	Variable Importance
ns2	0.051
ymrs_score	0.033
hal	0.030
mpq_score	0.024
bipollarii_sumscore	0.024
bprs_positive	0.020
bipollarii_mood	0.019

由表5和表6可以看出，ns2在随机森林和模糊森林中均排名第一，是双相情感障碍识别中最稳定的重要特征。ymrs_score和bipollarii_mood也同时出现在两种模型的结果中，说明这两个变量对双相情感障碍识别具有一定贡献。相比随机森林，模糊森林额外筛选出了hal、mpq_score、bipollarii_sumscore和bprs_positive等变量，提示在考虑变量相关结构后，部分特征可能表现出新的分类贡献。

从误差结果看，随机森林在该任务中的袋外误差为23.24%，模糊森林的验证误差为20.588%。两者差距较小，但模糊森林略低于随机森林，说明在双相情感障碍识别任务中，考虑变量相关结构可能有助于提升模型表现。

（四）注意缺陷多动障碍识别结果

在注意缺陷多动障碍识别任务中，本文将ADHD作为目标类别，并将精神分裂症和双相情感障碍归为其他疾病类别。表7和表8分别给出了随机森林和模糊森林筛选出的重要特征。

表7：ADHD识别中的重要特征和袋外误差估计（随机森林）

Importance Variables	Mean Decrease Accuracy	Mean Decrease Gini
adult_totalseverity	41.4	12.5
bprs_positive	33.8	9.74
scorev	11.1	4.22
mpq_score	10.8	5.44
asrs_score	8.48	4.20

表8：ADHD识别结果（模糊森林）

Features	Variable Importance
adult_totalseverity	0.137
bprs_positive	0.056
ant_mean_rtcon	0.018
dysfunc_pos	0.018
hamd_28	0.017
asrs_score	0.016
scorev	0.012

由表7和表8可以看出，adult_totalseverity在两种模型中均排名第一，说明该变量是ADHD识别中最主要的分类特征。bprs_positive、asrs_score和scorev也同时出现在两种模型的重要特征列表中，表明这些变量在ADHD与其他精神障碍的区分中具有一定稳定性。

从误差结果看，随机森林在ADHD识别中的袋外误差为13.1%，模糊森林的验证误差为2.94%。这一结果显示，模糊森林在该任务中的误差明显较低，说明其在ADHD识别场景中具有较好的分类表现。考虑到ADHD相关表型变量之间可能存在较强相关性，模糊森林对相关结构的处理可能是其表现较好的原因之一。

（五）随机森林与模糊森林分类表现比较

为进一步比较随机森林和模糊森林在独立测试集上的分类表现，本文分别计算了两种模型在四种分类设置下的预测误差率。表9展示了两种模型在测试集上的预测误差率，表10展示了随机森林袋外误差率和模糊森林验证误差率。

表9：四种设置下的预测误差率（随机森林与模糊森林）

Classification	Random Forest	Fuzzy Forest
Bipolar, ADHD, SCHZ	0.166	0.291
SCHZ, Others	0.0833	0.125
Bipolar, Others	0.333	0.25
ADHD, Others	0.125	0.0833

表10：四种设置下随机森林模型的袋外误差率和模糊森林模型的验证误差率

Classification	OOB error rate (RF)	validation error rate (FF)
Bipolar, ADHD, SCHZ	0.204	0.088
SCHZ, Others	0.071	0.118
Bipolar, Others	0.238	0.206
ADHD, Others	0.131	0.029

由表9可以看出，在测试集预测误差方面，随机森林在三类精神障碍分类任务和精神分裂症识别任务中表现更好，其误差率分别为0.166和0.0833，均低于模糊森林。模糊森林则在双相情感障碍识别和ADHD识别任务中表现更好，其误差率分别为0.25和0.0833，均低于随机森林。

结合表10可以发现，随机森林的袋外误差和模糊森林的验证误差并不总是与测试集误差完全一致。例如，在三分类任务中，模糊森林的验证误差低于随机森林的袋外误差，但测试集结果显示随机森林预测误差更低。这说明训练或验证阶段的误差不能完全代表模型在独立测试集上的表现。因此，在比较两种方法时，需要同时考虑特征筛选结果和测试集预测误差。

总体来看,随机森林和模糊森林在重要特征识别方面具有一定一致性,但二者在不同分类任务中的表现并不完全相同。随机森林在三分类任务和精神分裂症识别中表现较好,而模糊森林在双相情感障碍和 ADHD 识别中具有一定优势。这表明,模糊森林并非在所有任务中都优于随机森林,其优势可能更多体现在变量相关结构较明显的特定分类场景中。

四、实验结果讨论

(一) 关键表型特征的一致性

从实验结果来看,随机森林和模糊森林在不同分类任务中筛选出了一批具有较高一致性的表型变量。其中,adult_totalseverity 和 bprs_positive 在多个实验场景中反复出现,说明二者对精神障碍分类具有较强贡献。adult_totalseverity 主要反映成人 ADHD 相关症状的频率和严重程度,因此其在 ADHD 识别任务中排名靠前具有较强解释性。bprs_positive 反映精神症状评定中的阳性症状表现,该变量在精神分裂症识别任务中排名较高,也与其临床含义基本一致。

除上述两个变量外,ns2、etotal 和 mpq_score 也在不同模型或不同疾病识别任务中表现出一定重要性。ns2 在双相情感障碍识别任务中表现突出,并且在随机森林和模糊森林中均排名靠前,说明该变量可能与双相情感障碍和其他精神障碍之间的表型差异有关。etotal 与执行功能评估相关,mpq_score 与人格特质测量有关,二者的出现提示精神障碍分类不仅受到临床症状量表影响,也可能与人格、认知控制和执行功能等表型维度有关。

需要注意的是,变量重要性反映的是变量对模型分类结果的贡献,并不等同于因果关系或临床诊断依据。因此,本文更倾向于将这些变量理解为具有分类意义的候选表型特征,而不是直接将其解释为疾病诊断标志物。

(二) 不同疾病识别中的表型差异

在精神分裂症识别任务中,bprs_positive 在随机森林和模糊森林中均排名第一,说明阳性症状相关变量对区分精神分裂症和其他精神障碍具有较强贡献。mpq_score、bprs_negative、etotal 和 adult_totalseverity 也同时出现在两种模型的重要特征列表中,说明精神分裂症识别可能涉及症状表现、人格特质和执行功能等多个维度。尤其是 etotal 与执行功能表现相关,其出现提示认知控制、问题解决和冲动抑制等因素可能参与精神分裂症的表型区分。

在双相情感障碍识别任务中,ns2 是最稳定的重要变量。ymrs_score 和 bipollarii_mood 也同时出现在两种模型结果中。这说明双相情感障碍的识别不仅与躁狂症状和情绪状态相关,也可能与冲动性、新奇寻求或人格特质有关。模糊森林额外筛选出 mpq_score、bprs_positive 和 bipollarii_sumscore 等变量,提示在考虑变量相关结构后,部分与人格和症状表现相关的变量可能具有额外分类信息。

在 ADHD 识别任务中,adult_totalseverity 在两种模型中均排名第一,asrs_score 也同时出现在重要特征列表中。这与

ADHD 相关量表本身的测量目标一致。scorev 的出现说明外向性或人格特质可能也对 ADHD 与其他精神障碍的区分具有一定作用。已有研究指出,ADHD 群体内部可能存在不同的气质和表型亚型,这也说明人格和行为特质变量有助于理解 ADHD 的异质性^[9]。

(三) 高相关变量与冲动性相关特征

本文在预处理阶段将相关系数大于 0.8 的变量提取为高相关变量子集,并在后续分析中重新考察这些变量与模型筛选出的重要特征之间的关系。根据附录表 A.1,ns2、mpq_score、asrs_score 和 adult_totalseverity 与多个带有 bis 标识的变量存在较高相关性。这些变量主要与 Barratt 冲动性量表相关。Barratt 冲动性量表是常用的冲动性测量工具,能够从多个维度反映个体冲动水平^[10]。

这一结果提示,冲动性相关表型可能在精神障碍分类中具有一定作用。具体来看,adult_totalseverity 和 asrs_score 与 ADHD 症状评估相关,而 ADHD 本身常伴随注意控制、行为抑制和冲动控制方面的问题。因此,这些变量与 BIS 相关指标存在较高相关性具有一定合理性。ns2 在双相情感障碍识别中表现突出,也提示冲动性或新奇寻求相关特质可能有助于区分双相情感障碍与其他精神障碍。

此外,mpq_score 与人格特质测量有关。已有研究指出,冲动控制不足等外化精神病理特征与人格特质之间存在一定联系^[11]。因此,mpq_score 和 etotal 等变量的出现提示,精神障碍分类不仅依赖症状量表,也可能受到人格特质、冲动控制和执行功能等多维表型因素影响。

不过,这一结果仍需要谨慎解释。附录表 A.1 反映的是变量之间的相关关系,而不是因果机制。本文结果只能说明冲动性相关变量与模型筛选出的关键特征之间存在较强联系,后续仍需要结合更大样本、独立验证集和临床解释进一步检验其稳定性。

(四) 随机森林与模糊森林结果的比较

从模型比较结果来看,随机森林和模糊森林在重要特征识别方面具有一定一致性,但二者在不同分类场景中的预测表现并不完全相同。随机森林在三类精神障碍分类任务和精神分裂症识别任务中测试误差更低,而模糊森林在双相情感障碍和 ADHD 识别任务中表现更好。这说明模糊森林并不是在所有任务中都优于随机森林,其优势可能与具体数据结构、变量相关性和疾病分类场景有关。

模糊森林的主要特点是先根据变量相关结构划分模块,再在模块内部进行递归特征筛选。因此,当某一分类任务中存在较明显的相关变量组时,模糊森林可能更容易保留具有代表性的特征,并减少共线性对变量重要性排序的影响。ADHD 识别任务中,adult_totalseverity、asrs_score 和 BIS 相关变量之间可能存在较强关联,这可能是模糊森林在该任务中表现较好的原因之一。

相比之下,随机森林方法更直接,模型结构相对简单。在样本量有限的情况下,随机森林可能具有较稳定的预测表现。尤其是在精神分裂症识别任务中,bprs_positive 等变量与分类目标之间的关系较为明确,普通随机森林已经能够较好地捕捉主要分类

信息。因此，在精神障碍高维表型数据分析中，方法选择不应仅依据总体误差，而应结合具体疾病类别、变量相关结构和特征解释需求进行判断。

（五）研究局限

本文仍存在一定局限。首先，本文使用的数据样本量相对有限，尤其是三类精神障碍患者的样本规模较小，这可能影响模型训练和特征筛选结果的稳定性。其次，本文主要基于一次训练集、验证集和测试集划分进行分析，后续研究可采用重复交叉验证或 bootstrap 方法进一步评估结果的稳健性。再次，部分重要变量来自临床症状量表，这些变量本身可能与诊断类别存在较强关联，因此模型结果可能在一定程度上反映已有量表的信息。最后，本文主要使用误差率评价模型表现，后续研究可以进一步加入敏感度、特异度、F1值和 AUC 等指标，以更全面地评价分类模型性能。

总体而言，本文结果说明，随机森林和模糊森林均可用于精神障碍高维表型数据的特征筛选与分类分析。两种方法在关键特征识别方面具有一定一致性，但在不同疾病识别任务中的表现存在差异。后续研究可以在扩大样本规模、优化验证策略和引入更多评价指标的基础上，进一步检验这些候选表型特征的稳定性和解释价值。

五、结论

本文基于 Consortium for Neuropsychiatric Phenomics 表型数据集，围绕精神分裂症、双相情感障碍和注意缺陷多动障碍三类精神障碍，采用随机森林和模糊森林方法开展高维表型特征筛选与分类分析。研究分别设置三类精神障碍分类任务和三个单病种识别任务，以比较两种方法在不同分类场景下的特征筛选结果和模型表现。

实验结果表明，bprs_positive、adult_totalseverity、ns2、etotal 和 mpq_score 等变量在不同分类任务中表现出较高重要性。其中，adult_totalseverity 和 bprs_positive 在多个模型结果中反复出现，说明二者对精神障碍分类具有较稳定的贡献；ns2 在双相情感障碍识别中表现突出，提示人格特质和冲动性相关变量可能对部分精神障碍分类具有一定参考价值。随机森林和模糊森林在关键特征识别方面具有一定一致性，但二者在不同任务中的预测表现并不完全相同。随机森林在三类精神障碍分类和精神分裂症识别中表现较好，而模糊森林在双相情感障碍和 ADHD 识别中具有一定优势。

总体来看，随机森林和模糊森林均可用于精神障碍高维表型数据的特征筛选与分类建模。随机森林具有较好的模型稳定性和较强的直接建模能力，模糊森林则能够在一定程度上考虑高维变量之间的相关结构。二者结合使用，有助于从不同角度识别具有分类意义的候选表型特征。后续研究可在扩大样本规模、采用重复交叉验证、引入更多评价指标以及结合独立数据集验证的基础上，进一步检验相关特征的稳定性和解释价值。

附录

表 A.1：高相关变量子集中的变量

变量	ns2	mpq_score	asrs_score	adult_totalseverity
bis_factor2	0.465	-0.618	0.662	0.615
bis_factor1_ci	0.560	-0.624	0.568	0.463
bis_factor1	0.555	-0.639	0.549	0.417
bis_briefbis	0.629	-0.757	0.683	0.588
bis_2attimp	0.433	-0.585	0.759	0.647
bis_1mot	0.575	-0.684	0.458	0.495
bis_factor2_bi	0.505	-0.644	0.598	0.5489
bis_1sc	0.601	-0.698	0.556	0.432
bis_2motimp	0.599	-0.710	0.549	0.535
bis_latten	0.420	-0.542	0.721	0.635

表 A.2：重要特征简要介绍

Name	Actual Meaning	Comments
adult_totalseverity	Adult ADHD Clinical Diagnostic Scale (ACDS)	Rates the frequency and severity of ADHD symptoms
bprs_positive	scores of positive symptoms in Brief Psychiatric Rating (BPR)	thought of as the original schizophrenia scale but can be used for other disorders
ns2	Impulsiveness (NS2) in Temperament and Character Inventory (TCI)	attempts to explain the underlying causes of individual differences in personality traits
" bis " features	The Barratt Impulsiveness Scale (BIS)	is a widely used measure of impulsiveness
asrs_score	scores from Adult ADHD Self-Report Scale (ASRS)	Six of the eighteen questions were found to be the most predictive of symptoms consistent with ADHD.
mpq_score	Multidimensional Personality Questionnaire (MPQ) —Control subscale	has also documented relations between psychiatric conditions involving deficient impulse control
" dysfunc " features	Dickman' s Impulsivity Inventory (DII)	designed to assess the personality trait of impulsiveness
hamd_28	sum of scores of the Hamilton Rating Depression Scale for Depression (HAMD)	Has many anxiety symptoms and determine a patient' s level of depression
scorev	E score in Eysenck Personality Questionnaire (EPQ)	measures how much of an extrovert you are.

Name	Actual Meaning	Comments	Name	Actual Meaning	Comments
etotal	total correct of Delis–Kaplan executive function system (DKEFS) test	the assessment of executive functions including flexibility of thinking, inhibition, problem solving, planning, impulse control ect.	ymrs_score	The Young Mania Rating Scale (YMRS)	one of the most frequently utilized rating scales to assess manic symptoms

参考文献

[1] American Psychiatric Association. Diagnostic criteria from DSM–IV–TR[M]. American Psychiatric Pub, 2000.

[2] Frances A J, Widiger T. Psychiatric diagnosis: Lessons from the DSM–IV past and cautions for the DSM–5 future[J]. Annual Review of Clinical Psychology, 2012, 8: 109 – 130.

[3] Van Dam N T, et al. Data–driven phenotypic categorization for neurobiological analyses: Beyond DSM–5 labels[J]. Biological Psychiatry, 2017, 81(6): 484 – 494.

[4] Poldrack R A, et al. A phenome–wide examination of neural and cognitive function[J]. Scientific Data, 2016, 3(1): 1 – 12.

[5] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5 – 32.

[6] Qi Y. Random forest for bioinformatics[M]//Ensemble Machine Learning. Springer, 2012: 307 – 323.

[7] Liaw A, Wiener M. Classification and regression by randomForest[J]. R News, 2002, 2(3): 18 – 22.

[8] Conn D, et al. Fuzzy forests: Extending random forests for correlated, high–dimensional data[J]. 2015.

[9] Karalunas S L, et al. Subtyping attention–deficit/hyperactivity disorder using temperament dimensions: Toward biologically based nosologic criteria[J]. JAMA Psychiatry, 2014, 71(9): 1015 – 1024.

[10] Patton J H, Stanford M S, Barratt E S. Factor structure of the Barratt Impulsiveness Scale[J]. Journal of Clinical Psychology, 1995, 51(6): 768 – 774.

[11] Krueger R F, et al. Externalizing psychopathology in adulthood: A dimensional–spectrum conceptualization and its implications for DSM–V[J]. Journal of Abnormal Psychology, 2005, 114(4): 537.