涂强强, 郭文静, 潘乔,等. 基于小样本血浆蛋白质组学数据的抑郁症分类预测[J]. 智能计算机与应用,2024,14(8):133-137. DOI:10.20169/j.issn.2095-2163.240822

基于小样本血浆蛋白质组学数据的抑郁症分类预测

涂强强, 郭文静, 潘 乔, 陈德华
(东华大学 计算机科学与技术学院, 上海 201620)

摘 要: 抑郁症是一种常见的精神障碍,约 27%的人在一生中会出现类似症状,早期诊断对治疗至关重要,但传统诊断方法存在主观局限性,易误诊或漏诊,因此需要一种客观的诊断方法来提高诊断率。蛋白质组学技术研究蛋白质表达水平变化,可以帮助理解疾病机制,有助于开发临床诊断工具。蛋白质组学数据通常具有特征维度高,样本量少的特点,本文提出了一种基于小样本学习的抑郁症分类预测模型,相比于传统机器学习模型,该模型对抑郁症的分类预测能力显著提升。

关键词: 抑郁症; 小样本学习; 蛋白质组学技术; 临床诊断
中图分类号: TP311.5 **文献标志码**: A **文章编号**: 2095-2163(2024)08-0133-05

Prediction of depression classification based on small sample plasma mass spectrometry data

TU Qiangqiang, GUO Wenjing, PAN Qiao, CHEN Dehua
(College of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: Depression is a common mental disorder, with about 27% of people experiencing similar symptoms during their lifetime. Early diagnosis is essential for treatment, but traditional diagnostic methods have subjective limitations that make them prone to misdiagnosis or omission, so an objective diagnostic method is needed to improve diagnosis rates. Proteomics technology studies changes in protein expression levels, which can help understand disease mechanisms and contribute to the development of clinical diagnostic tools. Proteomics data are usually characterised by high feature dimensions and small sample sizes. In this paper, we propose a classification prediction model for depression based on small sample learning, which significantly improves the classification prediction ability of depression compared to traditional machine learning models.

Key words: depression; small sample learning; proteomics technology; clinical diagnosis

0 引 言

抑郁症作为最常见的精神疾病之一,致残率高,会影响一个人的情绪、思维和行为。抑郁症的主要症状包括长期低落的情绪、失去兴趣和快乐感、自责和无助感、疲劳、睡眠障碍、食欲改变和注意力难以集中。据统计,27%的人在一生中会出现抑郁症或与抑郁症发作类似的症状[1]。

目前,抑郁症的诊断主要依赖临床医生通过患者的症状和体征进行评估,并参考标准化的诊断工具,例如美国精神障碍诊断与统计手册和国际疾病分类等,该诊断方式依赖于临床医生的主观判断,包括对患者自陈报告的解释和对症状的评估,具有主观性,存在误诊和漏诊的情况。目前一些新兴的诊断方式,如基于生物标志物的血浆、尿液、脑脊液等含有的蛋白质的诊断方式有助于提供更客观和准确的抑郁症诊断。

蛋白质组学在蛋白质组水平上研究蛋白质表达水平的变化,以提供相关蛋白的糖基化、磷酸化,蛋白信号转导通路,疾病机制或蛋白-药物之间的相互作用的重要信息[2]。蛋白质组学数据通常具有“大 P,小 N”的特点,即特征数量多,而样本数量少,使用传统机器学习方式处理此类型数据往往效果不佳。因此本文从小样本学习的角度出发,提出了一

作者简介: 涂强强(2000-),男,硕士研究生,主要研究方向:智慧医疗; 潘 乔(1977-),男,博士,副教授,主要研究方向:机器学习,人工智能,大数据与云计算; 陈德华(1976-),男,博士,教授,主要研究方向:智慧医疗,数据科学,深度学习可解释性等。
通讯作者: 郭文静(1986-),女,博士,讲师,主要研究方向:传感器网络,体域网,物联网等。Email:wjguo@dhu.edu.cn

收稿日期: 2023-05-03

种基于小样本血浆质谱数据的抑郁症分类预测模型。

1 相关研究

Wesseling 等[3]提出了一种标记的多重选择反应监测分析方法,分析 56 种之前已知的与主要精神疾病相关的蛋白质,发现 Wnt-signalling 和谷氨酸受体丰度的改变主要发生在双相情感障碍和整个神经精神疾病谱的能量代谢异常中,钙信号主要在精神分裂症和情感性精神病中受到影响,锚蛋白 3 与情感性精神病相关联,22q11.2 缺失综合征相关蛋白 septin 5 与精神分裂症相关联;Lee 等[4]使用蛋白质组学分析 25 名未用药女性 MDD(Major Depressive Disorder)患者以及 25 名健康对照者的血液样本,使用多参数统计分析,最终得到潜在的生物标志物组,包括载脂蛋白 D、载脂蛋白 B、维生素 D 结合蛋白、血浆铜蓝蛋白、hornerin 和 profilin 1,诊断准确率为 68%;Han 等[5]调查了 130 情绪低落的对照组,53 名长期患有 MDD 的患者、40 名新患上 MDD 的患者以及 72 名当前未患有 MDD 的健康人,使用重复嵌套交叉验证方法来评估模型选择中的变异,并确保模型的重复性,最后得出 DBS(Dried Blood Spots)蛋白 α-1-酸性糖蛋白、α-2-球蛋白、醛脱氢酶 1 家族成员 A1、胆固醇酯转移蛋白 E 和补体因子 H 可以用于预测 MDD。

2 模型介绍

目前,对基于血浆蛋白质组学数据进行抑郁症分类预测的研究通常使用机器学习进行分类训练,但由于特征维度较高且样本数量较少,这些传统方法的预测准确率往往不高。因此本文提出基于小样本血浆蛋白质组学数据的抑郁症分类预测模型,模型整体框架图如图 1 所示。模型主要由 3 个部分组成:

(1)基于图卷积的特征提取层 HIGCL(Hierarchical Graph Convolution Layer):用于提取特征;

(2)全连接层:用于分类;

(3)因果样本权重模块:用于消除特征间虚假相关性。其中 HIGCL 由特征图卷积层和样本图卷积层组成,采用多领域聚合的方式提取特征信息,并在训练模型时引入 DropEdge 思想,在每轮训练时随机删除一定比例的边,以缓解模型在小样本上的过拟合问题[6-7]。

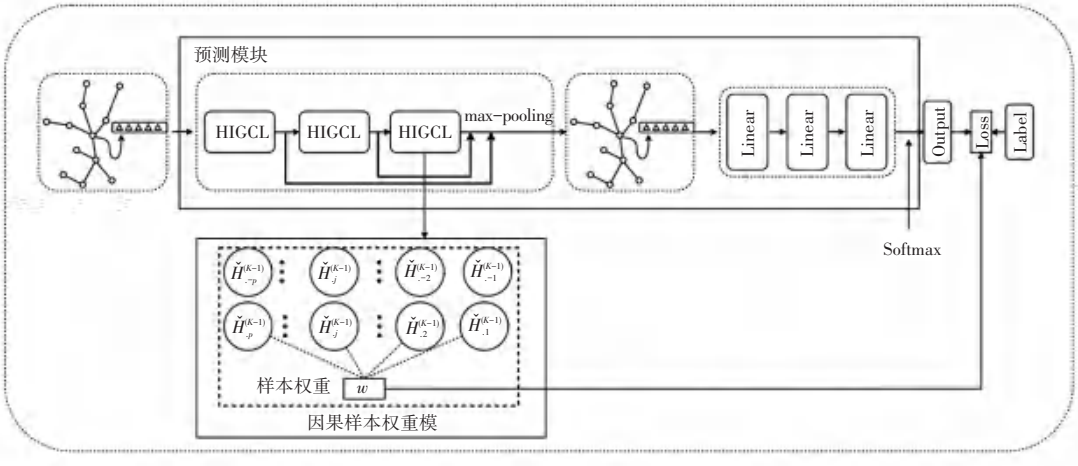

图 1 基于小样本学习的抑郁症分类方法总体框架

Fig. 1 Overall framework of depression diagnosis method based on small sample learning

2.1 蛋白质特征空间图卷积层

使用 HintDB 数据库获取蛋白质关系图 G^p,相互存在联系的蛋白质在 G^p 用 1 表示,不存在联系的用 0 表示,蛋白质与其自身的关系设置为 1。使用稀疏连接的图卷积操作在蛋白质特征空间来聚合邻居节点的蛋白质特征信息,可表达为公式(1):

$$H^{(1)} = \sigma(X(\boldsymbol{A}^{(p)} \odot W^{(0)})) \quad (1)$$

其中,$\boldsymbol{A}^{(p)}$ 表示蛋白质关系矩阵,由 $G^{(p)}$ 计算而来,其中的 p 表示蛋白质特征,$\boldsymbol{A}^{(P)}$ 表示蛋白质之间的关系权重,而 $W^{(0)}$ 为可训练参数,形状为 $n \times n$,n 表示蛋白质特征的数量,$\odot$ 是逐元素乘法运算。

$\boldsymbol{A}^{(p)}$ 与 $W^{(0)}$ 逐元素相乘得到一个新的权重矩阵,只有具有相互作用关系的蛋白质之间才会有权重信息,因此输入层和第一层之间的连接是稀疏的。经过训练之后的 $W^{(0)}$ 最后被用于筛选对模型影响较大的特征。

2.2　样本空间图卷积层

样本空间图卷积层的作用是在样本空间聚合特征信息。在样本空间 $G^{(s)}$ 中，每个节点代表一个样本，边代表样本之间的关系，并具有权重值。为了得到样本空间 $G^{(s)}$，首先计算每个样本之间的相似度；其次，引入注意力机制重新计算每个样本与其 K 个邻居的关系权重，得到样本相似度图；最后，以该图为依据，聚合样本特征信息。

为了学习到样本更好的表征以用于分类，一个常见的假设是属于同一个集群\类别的样本应该具有相似的表征，根据这个假设，模型基于样本相似度图 $G^{(s)}$ 在样本空间聚合邻近节点特征信息。可表达：

$$h_i^{(2)} = \boldsymbol{w} \odot h_i^{(1)} \tag{2}$$

$$H^{(3)} = \sigma(\boldsymbol{A}^{(s)} H^{(2)}) \tag{3}$$

其中，$h_i^{(1)}$ 和 $h_i^{(2)}$ 分别是 $H^{(1)}$ 和 $H^{(2)}$ 中 i^{th} 样本；$\boldsymbol{w} \in R^{(1\times p)}$ 是特征权重向量；$H^{(1)}$，$H^{(2)}$，$H^{(3)}$ 形状相同；$\boldsymbol{A}^{(s)}$ 是由样本相似图 $G^{(s)}$ 计算得到的邻接矩阵。

为了获得样本相似度图 $G^{(s)}$，需要得到样本之间相似度，本文使用缩放指数相似核（Scaled Exponential Similarity Kernel），来计算两个样本值之间的相似度 S_{ij}，计算公式如下：

$$\varepsilon_{ij} = \frac{\text{mean}(\rho(x_i, N_i)) + \text{mean}(\rho(x_j, N_j)) + \rho(x_i, x_j)}{3} \tag{4}$$

$$S_{ij} = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu\, \varepsilon_{ij}}\right) \tag{5}$$

其中，$\mu \in [0.3, 0.8]$ 是一个超参数；ε_{ij} 是一个缩放参数；$\rho(.,.)$ 为基于余弦相似度计算得到的样本相似度；N_i 表示 x_i 的邻居；N_j 表示 x_j 的邻居。

现在的 $G^{(s)}$ 包括所有的样本之间的相似度，为每个样本选取前 K 个 最相似的样本作为其邻居，使用 Softmax 计算每个样本与其 K 个邻居之间的注意力系数，得到样本之间关系权重：

$$\alpha_{ij} = \text{Softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in N_i} \exp(e_{ik})} \tag{6}$$

其中，N_i 表示样本 i 的邻居集合。

在 $G^{(s)}$ 中，每个样本只与跟其相似度前 K 个样本相连接，连接两个样本的边的权重由相似度经过注意力机制计算而来，$\boldsymbol{A}^{(s)}$ 为样本相似度图 $G^{(s)}$ 变换而来的样本图邻接矩阵，对角线的值是样本与其自身的权重为 1。其他位置的元素为对应的两个样本之间的注意力系数。通过这种方式，可以明确区分不同的邻居样本对中心样本的影响权重。

2.3　因果样本权重模块

在理想条件下，存在一组样本权重值，使得原始的特征矩阵接近正交，并最小化变量之间的相关性。DGCN（Dual Graph Convolutional Networks）模型在图神经网络中添加了微分去相关正则化器来估计每个标记节点的样本权重，在损失计算时考虑样本权重进行样本损失加权来消除学习到的蛋白质特征之间的虚假相关性[8]。

对于特征变量与输出结果的因果关系可以使用 *MTEF*（Marginal Treatment Effect Function）进行估计，见公式（7）：

$$MTEF = \frac{E[Y_i(t)] - E[Y_i(t-\Delta t)])}{\Delta t} \tag{7}$$

其中，t 代表要计算的蛋白质特征；$Y_i(t)$ 代表样本 i 在蛋白质特征 $T=t$ 时的输出结果；Δt 表示蛋白质特征的增加值。

使用样本权重对基于 *MTEF* 的因果关系估计进行重加权，作为正则项 DVD（Differentiated Variable Decorrelation），如式（8）所示：

$$\min_{w} \mathcal{L}_{DVD}(H) = \sum_{j=1}^{p} \left(\alpha^{\mathrm{T}} \cdot abs\left(\frac{H_{\cdot j}^{\mathrm{T}} \Lambda_w H_{\cdot -j}}{n} - \frac{H_{\cdot j}^{\mathrm{T}} w}{n} \cdot \frac{H_{\cdot -j}^{\mathrm{T}} w}{n}\right)\right)^2 + \frac{\lambda_1}{n}\sum_{i=1}^{n} w_i^2 + \lambda_2 \left(\frac{1}{n}\sum_{i=1}^{n} w_i - 1\right)^2 \tag{8}$$

其中，$abs(.)$ 表示为每个元素取绝对值，添加 $\frac{\lambda_1}{n}\sum_{i=1}^{n} w_i^2$ 正则化项来减少样本权重方差以实现预测稳定性，添加 $\lambda_2 \left(\frac{1}{n}\sum_{i=1}^{n} wi - 1\right)^2$ 正则化项来避免样本权重全为 0，$w \geqslant 0$ 保证样本的不权重会变为负数。

本文提出的模型基于图卷积神经网络，需要进行去相关的特征为经过图卷积聚合之后的特征 H，即为经过 $\sigma(X, \boldsymbol{A}, \theta_g)$ 输出后的特征，X 为输入的蛋白质特征，$\boldsymbol{A}$ 代表特征邻接矩阵，θ_g 为图卷积模型参数，$H_{\cdot -j}$ 表示将特征 j 置为 0 后其余特征不变的特征集合，变量权重 α 实际等于 H 的线性回归系数。

本文将 DVD 正则项集成至模型中，对最后一层图卷积层的输入特征去相关，初始样本权重 w 为 1，α 的计算方式为 $Var(W, axis=1)$，W 为线性变化层。引入因果样本权重后的新的损失计算见公式（9）：

$$\min_{\theta} \mathcal{L}_G = \sum_{l \in Y_L} w_l \cdot \ln(q(\tilde{H}_l^{(K)}) \cdot Y_l) \quad (9)$$

其中，$q(.)$ 为概率分布计算函数；Y_l 为节点真实标签；θ 为图卷积神经模型中的参数。

使用 $\boldsymbol{w}$ 重新加权每个样本计算出来的损失值，累加之后得到最终的损失值，根据该损失优化模型参数。

3 血浆质谱数据预处理

本文使用的数据集来自质谱数据共享网站 Pride，数据集编号“PXD028841”，使用了该数据集中的部分数据作为实验数据集，包括 88 个受试者的血浆样本经过液相色谱串联质谱法采集之后得到的原始质谱数据。受试者可以分为两类，患有抑郁症的患者，以及作为对照组的健康人。每个类别的样本数量皆为 44 例。

鉴定实验使用的数据为使用 LC/MS2 采集得到的原始质谱数据，其以 raw 格式存储。为了得到每个样本的蛋白质组学信息，本文使用了蛋白质定量软件 MaxQuant 对原始质谱数据进行了鉴定，鉴定过程主要包括对原始质谱数据的解析、数据库匹配和质量控制等步骤，最终得到了每个样本的原始蛋白质组学数据。MaxQuant 相关搜索参数见表 1，由于搜索得到的蛋白质信息存在蛋白质污染、噪声、缺失值等问题，本文使用 Perseus 软件对 MaxQuant 输出的原始蛋白质组学数据依次进行删除鉴定诱饵蛋白质、删除反序蛋白质、删除环境蛋白质、删除低表达蛋白质、基于正态分布对缺失值进行填充，并对蛋白质定量结果进行对数变换，处理完毕之后的数据集总共有 88 条记录，每条记录有 408 个蛋白质 LFQ（Label-Free Quantification）定量信息，将该数据作为后续的实验数据集。

表 1 MaxQuant 搜索参数

Table 1 MaxQuant search parameters

参数名	参数值
搜索引擎	Andromeda
数据库	Uniprot(20,404)
初级搜索容差	6×10^{-6}
MS/MS 离子容差	20×10^{-6}
可变修饰	蛋白质的 N-乙酰化和甲硫氨酸的氧化
固定修饰	半胱氨酸氨基甲基化
酶	完全胰蛋白酶消化
最小长度	6
最多损失裂解肽	2
错误发现率	1%

4 实 验

本文使用准确率（*Accuracy*）、精确率（*Precision*）、召回率（*Recall*）、*F*1 分数（*F*1-Score）来评估模型在二元分类时的准确性，见公式（10）~公式（13）：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F1\text{-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (13)$$

其中，*TP* 表示真正例，指分类器将正类样本预测为正类的数量；*TN* 表示真负例，指分类器将负类样本正确预测为负类的数量；*FP* 表示假正例，指分类器将负类样本错误预测为正类的数量；*FN* 表示假负例，指分类器将正类样本错误预测为负类的数量。

将本文提出的模型与逻辑回归、随机森林、支持向量机、多层感知机、AdaBoost、AffinityNet[9]、HIGCN 7 种基线模型进行对比实验，结果见表 2。在测试集上表现，传统的机器学习模型中，逻辑回归的表现最佳，小样本神经网络 HIGCN 在全部 5 个评价指标上都优于所有的机器学习，而本文提出的模型在 *Accuracy* 上达到了 0.87，*Precision* 达到了 0.89，*Recall* 达到了 0.84，*AUC* 达到了 0.86，*F*1-Score 达到 0.86，优于其他所有模型，验证了该模型的分类性能。

表 2 多模型对比实验结果

Table2 Experimental comparison result of multiple models

模型名	*Accuracy*	*Precision*	*Recall*	*F*1 - Score
逻辑回归	0.77	0.78	0.84	0.81
随机森林	0.67	0.73	0.73	0.73
支持向量机	0.72	0.82	0.57	0.67
多层感知机	0.50	0.43	0.70	0.54
AdaBoost	0.67	0.63	0.73	0.68
AffinityNet	0.75	0.78	0.73	0.75
HIGCN	0.83	0.83	0.84	0.83
Our	**0.87**	**0.89**	**0.84**	**0.86**

本文对于模型的关键组成部分进行了消融实验，以验证不同部分对模型的影响：（1）DropEdge；（2）因果样本权重模块（CausalSampleWeight）。消

融实验使用数据集的 20%作为训练集，其余 80%作为测试集，学习率设为 0.01，权重衰减设为 1e-4，优化器选择 Adam，重复十次实验取平均值作为最终结果，消融实验结果见表 3。可以发现各个模型获得一定程度的提升，引入 DropEdge 思想后，*Accuracy* 从 0.83 提升到了 0.85，*Precision* 从基础模型的 0.83 提升至 0.86，*AUC* 从 0.83 提升至 0.84，*F*1 - Score 从 0.83 提升到了 0.85。加入因果样本权重模块后，*Accuracy* 从 0.83 提升到了 0.85，*Precision* 从 0.83 提升到了 0.86，*AUC* 从 0.83 提升到 0.84，*F*1 - Score 从 0.83 提升到0.85。而本文最终模型的各个指标也都优于基础模型。*Accuracy* 从 0.83 提升到了 0.87，*Precision* 从 0.83 提升到了 0.89，*AUC* 和 *F*1 - Score 从 0.83 提升到了0.86。

表 3　消融实验结果

Table 3　Results of ablation experiment

模型名	*Accuracy*	*Precision*	*Recall*	*F*1 - Score
HIGCN	0.83	0.83	0.84	0.83
HIGCN+DropEdge	0.85	0.86	0.84	0.85
HIGCN+因果样本权重	0.85	0.86	0.84	0.85
Our	**0.87**	**0.89**	**0.84**	**0.86**

5　结束语

本文提出了一种基于血浆蛋白质组学数据的抑郁症分类预测模型，该模型利用图卷积网络在蛋白质特征空间和样本空间中提取潜在特征。为解决小样本数据特征共线引起的问题，本文引入了因果样本权重模块，尽可能消除特征与预测结果之间的虚假相关性；为了缓解过拟合问题，引入了 DropEdge 思想，在模型的训练过程中对图的边进行随机删除，增加了模型的泛化能力。

通过一系列实验证明本文提出的基于小样本血浆蛋白质组学数据的抑郁症分类预测模型在小样本数据上具有较好的性能，可以有效地用于抑郁症的分类预测任务，这一研究为解决小样本数据中的抑郁症分类预测问题提供了一种新的方法。

参考文献

[1] 冯洁洁，马来阳，徐莉力，等. 磁共振成像在重度抑郁症合并失眠中的研究进展[J]. 磁共振成像，2022，13(12)：141-145.

[2] 尹稳，伏旭，李平. 蛋白质组学的应用研究进展[J]. 生物技术通报，2014，258(1)：32-38.

[3] WESSELING H，GOTTSCHALK M G，BAHN S. Targeted multiplexed selected reaction monitoring analysis evaluates protein expression changes of molecular risk factors for major psychiatric disorders[J]. International Journal of Neuropsychopharmacology，2015，18(1)：109-131.

[4] LEE M Y，KIM E Y，KIM S H，et al. Discovery of serum protein biomarkers in drug-free patients with major depressive disorder [J]. Progress in Neuro-Psychopharmacology and Biological Psychiatry，2016，69：60-68.

[5] HAN S Y S，TOMASIK J，RUSTOGI N，et al. Diagnostic prediction model development using data from dried blood spot proteomics and a digital mental health assessment to identify major depressive disorder among individuals presenting with low mood [J]. Brain，Behavior，and Immunity，2020，90：184-195.

[6] XU K，LI C，TIAN Y，et al. Representation learning on graphs with jumping knowledge networks [C]//Proceedings of International Conference on Machine Learning. PMLR，2018：5453-5462.

[7] RONG Y，HUANG W，XU T，et al. Dropedge：Towards deep graph convolutional networks on node classification[J]. arXiv preprint arXiv：1907.10903，2019.

[8] FAN S，WANG X，SHI C，et al. Debiased graph neural networks with agnostic label selection bias[J]. IEEE Transactions on Neural Networks and Learning Systems，2022，35(4)：4411-4422.

[9] MA Tianle，ZHANG Aidong. Affinitynet：Semi-supervised few-shot learning for disease type prediction[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019：1069-1076.