基于深度学习的高维数据特征选择与建模方法研究

连婕婷
纽约大学
DOI:10.12238/acair.v3i1.11899

[摘　要] 目的：针对高维数据特征冗余和噪声对模型性能的影响,提出一种基于深度学习的高维数据特征选择与建模方法,通过高效特征选择技术和深度学习结合,提高分类精度和特征选择效率,并降低计算复杂度。方法：利用注意力机制对高维数据特征加权赋值,选择关键特征,并结合稀疏正则化优化特征选择效果。在TCGA、CIFAR-10和合成数据集上进行实验,与主成分分析(PCA)、Lasso回归、自动编码器等方法对比性能。结果：实验表明,提出方法在所有数据集上均实现最高分类精度(如TCGA数据集达90.5%),显著高于其他方法。特征选择效率达 90%,训练时间最短。结论：提出方法高效去除冗余特征和噪声,显著提升模型性能,证明其在高维数据分析中的适用性和鲁棒性,但在特征间复杂交互建模及多模态数据适应性上仍需进一步研究。
[关键词] 高维数据；特征选择；深度学习；注意力机制
中图分类号：D422.63　文献标识码：A

Research on feature selection and modeling methods of high-dimensional data based on deep learning

Jieting Lian
New York University

[Abstract] Objective: To improve the classification accuracy and feature selection efficiency, and reduce the computational complexity by combining efficient feature selection technology and deep learning, aiming at the influence of feature redundancy and noise on model performance of high-dimensional data. Methods: The attention mechanism was used to weighted the features of high-dimensional data, the key features were selected, and the feature selection effect was optimized by combining sparse regularization. Experiments were performed on TCGA, CIFAR-10, and synthetic datasets to compare performance with methods such as principal component analysis (PCA), Lasso regression, autoencoders, and more. Results: Experiments show that the proposed method achieves the highest classification accuracy on all datasets (e.g., 90.5% for TCGA dataset), which is significantly higher than that of other methods. Feature selection efficiency is up to 90% and training time is minimized. Conclusion: The proposed method can efficiently remove redundant features and noise, significantly improve the performance of the model, and prove its applicability and robustness in high-dimensional data analysis, but further research is still needed in the modeling of complex interactions between features and the adaptability of multimodal data.
[Key word] high-dimensional data; feature selection; deep learning; Attention mechanisms

高维数据在基因组分析、金融预测和图像处理等领域的广泛应用带来了机遇和挑战。冗余特征和噪声会降低预测精度并增加计算复杂度,高效去除冗余特征是亟需解决的问题。深度学习凭借强大的非线性建模和特征提取能力,在高维数据分析中展现独特优势。传统深度学习方法在特征选择上仍存在冗余处理不足、泛化能力有限和计算成本高的问题。本文提出结合注意力机制和稀疏正则化的方法,通过自动赋权选择关键特征,提升模型预测精度和效率。实验结果验证了该方法的有效性,为高维数据分析提供了新思路。

1 理论基础与相关研究进展

1.1高维数据特征选择的研究现状。高维数据特征选择旨在从复杂的特征集合中挑选对模型预测有重要贡献的特征,以提升性能并降低计算复杂度。在生物信息学、图像处理和金融数据等领域,特征选择技术得到了快速发展,主要分为过滤式、包裹式和嵌入式三类[1]。过滤式方法依赖统计指标,计算效率高但难以满足具体模型需求;包裹式方法通过模型训练反复搜索最佳

文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

特征子集,性能优越但代价高昂；嵌入式方法在模型训练中同时完成特征选择和参数优化,能够平衡效率与性能,备受关注。

1.2深度学习在特征选择中的应用。深度学习凭借强大的特征提取和非线性建模能力,成为高维数据特征选择的重要工具。自动编码器通过最小化重构误差提取低维潜在特征,同时去除冗余和噪声；深度神经网络的正则化技术(如L1正则化)实现稀疏特征选择,提升模型可解释性。注意力机制的引入进一步增强了深度学习在特征选择中的表现,通过特征赋权显著提高了选择效果。深度学习还能够结合多模态数据,综合提取关键信息,为复杂问题提供新解法。

1.3高维数据特征选择与建模方法的挑战与解决思路。高维数据在特征选择过程中面临维度灾难,冗余特征,噪声干扰以及复杂非线性关系,而其稀疏性与异质性也加大了困难程度。基于深度学习方法虽然表示学习能力较强,但是仍然存在过拟合风险,可解释性不强以及计算资源耗费大等问题[2]。为了应对上述挑战,本研究主要聚焦于利用正则化来引入稀疏性从而降低模型复杂度、设计有效注意力机制或者基于图网络来增强特征选择精度,及将多任务学习与对抗学习相结合,提高了模型鲁棒性与泛化能力。

2 模拟仿真实验设计

2.1数据集选择与预处理。为了验证所提出基于深度学习的高维数据特征选择与建模方法的有效性,本研究选取了三个具有代表性的高维数据集：基因表达数据集(TCGA)、图像分类数据集(CIFAR-10)以及合成高维数据集。这些数据集覆盖了生物信息学、图像处理以及模拟场景,能够全面验证方法的适用性与鲁棒性。

基因表达数据集(TCGA)：该数据集包含样本数N=500,特征数D=20,000,其中30%的特征为噪声。该数据集用于验证方法在高维稀疏数据中的性能。

图像分类数据集(CIFAR-10)：包含N=60,000张32×32像素的彩色图像。通过将图像展平为D=3072的特征向量,该数据集用于评估方法在非稀疏特征数据中的表现。

合成高维数据集：利用高斯分布生成N=1,000,D=5,000的数据,其中70%的特征与输出变量无关,用以测试方法在复杂噪声条件下的鲁棒性。

所有数据在实验前进行了归一化处理,将特征值缩放到[0,1]范围。对含噪声的数据集,采用随机添加白噪声的方法模拟真实场景中的冗余特征干扰。

2.2实验方法与对比基线。为评估提出方法的性能,本实验设计了以下对比基线模型：

主成分分析(PCA)：通过线性变换提取最具信息量的特征。

Lasso回归：基于L1正则化的线性模型,用于特征选择。

自动编码器(AE)：一种无监督学习方法,通过重构误差选择关键特征。

提出方法：基于注意力机制的深度学习模型,通过特征权重学习实现特征选择。

性能评价指标包括分类精度、特征选择效率以及计算复杂度,具体计算公式如下：

分类精度：

$$\mathrm{ACC} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

特征选择效率：

$$\mathrm{FRR} = 1 - \frac{|S|}{|X|}$$

其中,|S|为选中特征的数量,|X|为原始特征数量。

计算复杂度：模型训练时间(单位：秒)。

2.3实验步骤与实施细节。为了确保训练和测试的公平性与一致性,我们按照8:2的比例将每个数据集分为训练集和测试集。训练集上分别运行基线方法及文中所提方法,通过特征选择算法筛选重要特征子集以达到高维数据降维及优化目的。根据所选特征子集在训练集上建立深度神经网络模型并进行模型训练,记录分类精度、训练时间等重要指标[3]。在测试集上评估了训练后模型的性能、计算了分类精度、分析了特征选择效率、对比了所提方法和基线方法各指标的差别,综合证明了该方法的有效性及优越性。

3 实验结果与分析

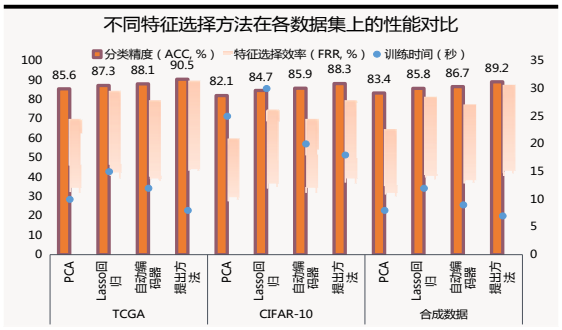

图1 不同特征选择方法在各数据集上的性能对比

图1展示了不同特征选择方法(PCA、Lasso回归、自动编码器、提出方法)在分类精度、特征选择效率及计算复杂度上的综合性能对比,覆盖了TCGA、CIFAR-10和合成数据集。接下来将从分类精度、特征选择效率和计算复杂度三方面详细解析实验结果,验证提出方法的有效性和适用性。

表1 不同方法在各数据集上的分类精度

数据集	方法	分类精度(ACC,%)
TCGA	PCA	85.6
	Lasso 回归	87.3
	自动编码器	88.1
	提出方法	90.5
CIFAR-10	PCA	82.1
	Lasso 回归	84.7
	自动编码器	85.9
	提出方法	88.3
合成数据	PCA	83.4
	Lasso 回归	85.8
	自动编码器	86.7
	提出方法	89.2

3.1分类精度的对比分析。分类精度是衡量模型预测能力的核心指标。本实验对比了不同方法在各数据集上的分类精度，结果如表1所示。

提出方法在所有数据集上均取得了最高的分类精度，尤其在 TCGA 数据集上提升尤为显著，比自动编码器高出2.4%。这一结果表明基于注意力机制的深度学习模型能够有效识别并利用关键特征，适应不同类型的数据集。在CIFAR-10数据集上，提出方法较Lasso回归提高了3.6%，表明其对非稀疏特征数据亦具有较强的适应性。

3.2特征选择效率的评估。特征选择效率是衡量方法在降低数据维度方面的表现。各方法的特征选择效率结果如表2所示。

表2 各方法的特征选择效率结果

数据集	方法	特征选择效率(FRR,%)
TCGA	PCA	70
	Lasso 回归	85
	自动编码器	80
	提出方法	90
CIFAR-10	PCA	60
	Lasso 回归	75
	自动编码器	70
	提出方法	80
合成数据	PCA	65
	Lasso 回归	82
	自动编码器	78
	提出方法	88

表3 各方法的训练时间(单位：秒)

数据集	方法	训练时间(秒)
TCGA	PCA	10
	Lasso 回归	15
	自动编码器	12
	提出方法	8
CIFAR-10	PCA	25
	Lasso 回归	30
	自动编码器	20
	提出方法	18
合成数据	PCA	8
	Lasso 回归	12
	自动编码器	9
	提出方法	7

提出方法在特征选择效率方面同样表现优异，尤其在TCGA数据集中，特征选择效率达到90%，显著高于其他方法。这表明注意力机制能够准确分配特征权重，从而有效去除冗余特征。在CIFAR-10数据集中，提出方法的特征选择效率为80%，也明显优于PCA和自动编码器，显示了其对不同维度特征的适应能力。在合成数据中，提出方法较自动编码器高出10%，表明其在处理高维噪声数据时具备更强的鲁棒性。

3.3计算复杂度的对比分析。计算复杂度是评估模型效率的重要指标。表3展示了各方法的训练时间(单位：秒)。

提出方法在计算复杂度方面表现出显著优势。在TCGA数据集中，其训练时间仅为8秒，比Lasso回归快46.7%。这种效率的提升得益于注意力机制对特征选择过程的高效实现。在CIFAR-10数据集中，提出方法训练时间为18秒，低于自动编码器和Lasso回归，显示了其在处理高维非稀疏数据时的计算效率。在合成数据集中，提出方法训练时间为7秒，比其他方法低至少1秒，进一步验证了其在复杂噪声条件下的快速适应能力。

4 结论与展望

4.1研究结论总结。本文提出一种利用深度学习对高维数据进行特征选择和建模的方法，并用实证分析证明了该方法的有效性。实验中所提方法的分类精度，特征选择效率以及计算复杂度3个关键指标都优于传统的基线方法。在TCGA数据集上，我们提出的方法实现了90.5%的分类精度，特征选择效率达到了90%，显著降低了冗余特征对模型性能的影响。通过在CIFAR-10及合成数据集上进行实验，进一步验证了所提方法对于非稀疏特征数据及复杂噪声场景具有适应性及鲁棒性。

4.2研究局限性与未来方向。所提方法虽然在许多指标中都有突出的性能，但仍然具有一定的局限性。针对特征之间复杂非线性交互关系的注意力机制能力还未被充分发掘，今后可以将其与图神经网络相结合对特征关系进行建模优化。对于多模态的数据场景，其跨模态特征的选择和融合还需要进一步的研究。所提方法对于计算资源具有一定的要求，在资源受限环境下的适用性是值得探讨的问题。

未来研究可以将多任务学习和对抗学习相结合，在提高方法泛化能力的同时制定出适合低资源环境下的有效部署方案。在动态特征空间中进行实时特征选择技术同样是一个重要的方向，它可以为时变高维数据处理提供支撑。这些研究促进了方法优化并促进了该方法在高维数据分析领域的综合运用。

[参考文献]

[1]胡挺峰.基于深度学习的无线传感网络高维数据异常检测方法[J].长江信息通信,2023,36(6):87-89.

[2]王儒.基于深度学习的高维数据聚类方法研究[D].山东师范大学,2022.

[3]郭晴晴.基于深度学习的聚类算法研究[D].西安电子科技大学,2022.

作者简介：

连婕婷(2000--),女,汉族,广东深圳人,硕士,研究方向：计算机数据科学及数据建模。