智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

用于高维小样本特征选择的超网络设计

魏俊伊, 董红斌, 余紫康

引用本文:
魏俊伊, 董红斌, 余紫康. 用于高维小样本特征选择的超网络设计[J]. 智能系统学报, 2025, 20(2): 465-474.
WEI Junyi, DONG Hongbin, YU Zikang. Hypernetwork design for feature selection of high-dimensional small samples[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(2): 465-474.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202402018

DOI: 10.11992/tis.202402018
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250107.1643.008

用于高维小样本特征选择的超网络设计

魏俊伊, 董红斌, 余紫康
(哈尔滨工程大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001)

摘　要: 特征选择是受各行业广泛关注的问题。特征选择针对的数据集通常是高维的, 且样本数较少, 例如生物、医学领域的数据集。虽然很多的正则化网络在这种数据集上的表现能够优于复杂的网络, 但是在小数据量上许多潜在的特征关系仍然会被过度挖掘, 从而出现过拟合的情况。为了解决此类问题, 提出了端到端的稀疏重构网络, 模型先对特征进行稀有增强和奇异值嵌入, 之后通过并行辅助网络对嵌入矩阵进行训练, 重构预测权重, 实现了削减参数的超网络学习方式。参数较少的网络受过拟合的影响也会随之减少, 有效降低了无效参数对网络的影响。对生物、医学领域的 12 种高维小样本数据集进行了实验, 并通过对比实验发现在 8 种特征选择网络中降维后, 本网络的分类准确率平均提升了 3.26 百分点。另外, 通过消融实验分别证明了分解层、重构层、关联层的作用, 最后分析权重结果, 进一步阐述了模型的扩展应用。

关键词: 特征选择; 正则化网络; 过拟合; 端到端; 稀疏重构; 奇异值; 辅助网络; 超网络; 高维小样本
中图分类号: TP311　　**文献标志码**: A　　**文章编号**: 1673−4785(2025)02−0465−10

中文引用格式: 魏俊伊, 董红斌, 余紫康. 用于高维小样本特征选择的超网络设计 [J]. 智能系统学报, 2025, 20(2): 465–474.
英文引用格式: WEI Junyi, DONG Hongbin, YU Zikang. Hypernetwork design for feature selection of high-dimensional small samples[J]. CAAI transactions on intelligent systems, 2025, 20(2): 465–474.

Hypernetwork design for feature selection of high-dimensional small samples

WEI Junyi, DONG Hongbin, YU Zikang
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract: Feature selection is a widely recognized challenge across various industries. They typically target high-dimensional datasets with fewer samples, such as those in biology and medicine field. Many regularization networks outperform complex network structures on such datasets. However, numerous underlying feature relationships can still be overfitted, particularly with limited data. This study proposes an end-to-end sparse reconstruction network to address this issue. First, the model enhances features through sparsity and singular value embedding. Then, it trains the embedding matrix through a parallel auxiliary network to reconstruct prediction weights, which implements a parameter-reducing super-network learning approach. This approach reduces the impact of overfitting on networks with fewer parameters, which effectively mitigates the influence of ineffective parameters on the network. Experiments conducted on 12 high-dimensional small-sample datasets in biology and medicine field reveal an average improvement of 3.26 percentage point in classification accuracy after dimensionality reduction in eight feature selection networks. Furthermore, the roles of the disintegration layer, reconstruction, and correlation layer are separately validated through ablation experiments, followed by weight result analysis, which further elucidates the extended applications of the model.

Keywords: feature selection; regularization network; overfitting; end-to-end; sparse reconstruction; singular value; auxiliary network; hypernetwork; high-dimensional small sample

近年来, 医学、海洋等领域产生了海量的特征数据。这种数据集往往数量很少, 特征中存在潜在的相关信息, 如果不能从这些特征中挖掘出更有价值的特征信息, 那么样本中的噪声可能会导致数据无法利用, 甚至与现实意义产生偏差。因此在各个技术领域, 都需要高效的小样本数据

收稿日期: 2024−02−20.　**网络出版日期**: 2025−01−08.
基金项目: 黑龙江自然科学基金项目 (LH2020F023).
通信作者: 董红斌. E-mail: donghongbin@hrbeu.edu.cn.

的特征选择算法来挖掘出更重要的关键信息[1]。

对于高维小样本数据，如果仅凭简单的线性网络，则无法找出特征之间蕴藏的关联信息；由于样本水平的稀缺，搭建复杂的特征过滤网络又会导致过拟合，不能产生较好的泛化能力[2]。因此网络既需要有良好的学习能力，又要通过正则化等方式来增强模型的泛化能力。已经有学者通过研究，提出了专用的正则化特征选择网络来进行高维特征的筛选，但由于传统的正则化网络具有庞大的学习参数，在第一层包含了 90% 以上的参数[3]，这显然会导致神经网络的过拟合。尤其在有限的样本数量下，第一层大量的参数和某些神经元较高的权重水平会导致深层的网络过度依赖于浅层的网络层中的神经元，从而产生高方差以及过拟合的问题。而更换为较复杂的特征选择网络依然会产生以上问题，因此如何平衡网络模型的复杂度，提高模型泛化能力的同时还要具有较强的学习能力，这是神经网络进行特征过滤较为关键性的问题，也是本文关注的重点问题。

1　相关工作

本文所讨论的工作是针对于小样本高维特征数据集进行特征选择，这种数据集的特征数量会远远大于样本数量[4]。因此在网络的训练过程中，往往会加入正则化机制，以便选择更具相关性的特征。例如，将 L1 正则化机制加入到神经网络结构[5]。但是在参数过多的情况下，优化 L1 约束会产生陷入局部最优、过拟合等问题。深度神经追踪 (deep neural persuit, DNP)[6] 网络为了解决这个问题，采用了对每个特征的近似 L1 正则化网络，基于多 Dropout，调整平均梯度，从而进行高维特征的过滤。然而 DNP 没有考虑到特征之间的潜在关联，因此可能产生较弱的学习水平。

多层感知群 (multilayer perceptions, MLPs)[7] 证明了在小样本高维数据集下，由多个良好正则化的普通 MLP 建立的架构显著优于先进的神经网络架构。甚至产生的效果要优于传统的机器学习方法，如极限梯度提升 (extreme gradient boosting, XGBoost)[8]。事实上在神经网络的特征选择问题中，诸多重要的发现开始并不是用于对小样本数据的研究，而是在计算机视觉领域引起广泛关注的模型。hypernetworks[9] 的提出不仅推动了视觉网络的发展，也为特征选择 (feature selection, FS) 架构指明了一个新的方向，它提出了超网络的概念：使用一个网络为另一个网络生成权重的方法。这种方法原本是用于提高网络对于图像中特征的挖掘能力，现在被广泛应用于特征选择网络[10]。主网络的结构由普通的神经网络结构组成，而超网络则接受一组包含其权重结构的信息，并负责生成主网络的权重。这样做能够增加网络的高效性和可拓展性。首先将其应用在小样本高维特征数据的是营养网络 (dietnet)[11]，模型由 1 个主网络和 2 个辅助网络组成。主网络采用了多 MLP 结构，辅助网络用来预测脂肪层的参数和重构脂肪层的参数，通过这种方式来减少主网络自由参数的数量，从而提高模型的泛化能力。但模型在迭代过程中仍然具有较高的方差和梯度。

从正则化网络的角度出发，混凝土自编码器 (concrete autoencoders, CAE)[12] 引入了自编码器，它能够识别信息量划分出的特征子集，同时从特征重构中学习，能够减小重构误差，提高特征分类精度。这种方式处理简单图像任务十分有效，但在小样本的特征选择问题上仍需要改进。特征选择网络 (feature selection net, Fsnet)[13] 在 CAE 的基础上设计了重建损失的正则化网络，它能有效减小参数数量，增强了模型的稳定性和可扩展性，适用于小样本高维特征的训练，但 Fsnet 使用了直方图嵌入方式，并使用了解码器优化重建损失，导致在连续的特征值下无法具有良好表现。权重预测网络 (weight predictor network, WPN)[14] 扩展了 dietnet 网络，使用 2 个辅助网络减少脂肪层和重建层的权重，并采用了非负矩阵分解 (non-negative matrix factorization, NMF)[15] 方式进行特征的嵌入，在连续特征值的数据集下产生了良好的效果。但在诸多医学数据集上的损失下降过程表明简单的矩阵分解嵌入和网络的架构也有优化的可能。lassonet[16] 网络通过将特征选择和参数学习相结合探索一个更为理想的正则化路径。但由于其模型的复杂性，本身具有大量的超参数，在高维小样本特征中学习无法产生良好的效果。表 1 给出了本文提出的稀疏重构网络 (sparse reconstruction network, SRnet) 和相关模型的信息。

表 1　网络的相关特征选择模型对比

Table 1　Comparison of network feature selection models

模型	数据性质	小样本高维	超网络架构	特征嵌入方法	编码−解码	损失函数性质
DNP[6]	连续	√	×	—	×	正则可微
MLPs[7]	离散	√	×	—	×	可微

续表 1

模型	数据性质	小样本高维	超网络架构	特征嵌入方法	编码–解码	损失函数性质
dietnet[11]	离散	√	√	直方图	√	正则可微
CAE[12]	连续	×	×	—	×	可微
Fsnet[13]	连续	×	×	直方图	√	正则可微
lassonet[16]	连续	×	×	—	×	正则不可微
WPN[14]	连续	√	√	矩阵分解	×	正则可微
SRnet	连续	√	√	奇异值分解	×	正则可微

因此, 本文以简化网络模型、降低网络模型参数、提高网络泛化能力为目标, 提出了一种能够稀疏权重端到端的网络架构, 核心内容如下:

1) 超网络对主网络的参数进行削减, 降低网络复杂度, 提高网络的泛化能力, 采取多重失活策略, 防止模型过拟合。

2) 将超网络细分为分解层、重构层和关联层, 并行训练, 辅助网络负责对主网络的权重预生成。

3) 为了进一步提高特征选择网络对较独立特征的选择能力, 设计了稀有增强算法, 这种方法能够探索出不常出现的特征组合的潜在相关性, 应用在超网络的关联层。

4) 超网络还采用了矩阵奇异值的特征嵌入方法, 保证特征信息合理嵌入, 应用于超网络的分解层和重构层。

5) 设计了基于重构因子和关联因子的损失函数, 从而使小网络不仅能够学习到大网络的知识, 同时能够降低大网络的参数。

2　问题定义

2.1　问题相关定义

将小样本高维特征数据集定义为 $\{\boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}\}$, $i \in \{1,2,\cdots,n\}$。其中, $\boldsymbol{x}^{(i)} \in \mathbb{R}^D$ 是一个样本, $y_i \in \boldsymbol{\gamma}$ 是该样本所携带的标签, $\boldsymbol{\gamma}$ 表示数据集中标签的类空间。定义 $\boldsymbol{X} = [\boldsymbol{x}^{(1)}\ \boldsymbol{x}^{(2)}\ \cdots\ \boldsymbol{x}^{(N)}] \in \mathbb{R}^{N\times D}$ 是不含标签的特征矩阵, $\boldsymbol{Y} = [y^{(1)}\ y^{(2)}\ \cdots\ y^{(N)}] \in \mathbb{R}^N$ 是标签向量[17]。对于第 j 个特征, 对应的嵌入为 $\boldsymbol{e}^{(j)} \in \mathbb{R}^M$, M 是嵌入向量的维度空间, 嵌入后的维度 $M < D$。特征选择的目标是最大化目标函数 $\max F(x)$, 选择特征集合 s.t.$T = \{t^{(1)}, t^{(2)}, \cdots, t^{(d)}\}$, 满足 $\forall d \in \{1, 2, \cdots, D\}$。

2.2　网络相关定义

定义 $f_W : \mathbb{R}^D \to \gamma$ 表示神经网络中数据的映射关系, $\boldsymbol{W}$ 为神经网络的参数矩阵:

$$\boldsymbol{W} = \begin{bmatrix} f_{11} & \cdots & f_{1D} \\ \vdots & \ddots & \vdots \\ f_{N1} & \cdots & f_{ND} \end{bmatrix}$$

式中: f_{ij} 表示了权重层的一个权重值。该网络的输入是 $\boldsymbol{x}^{(i)}$, 输出一个预测后的标签 y_i。对于一个线性层网络, 设第一层的神经元个数为 K, 则 $\boldsymbol{W} = [\boldsymbol{f}^{(1)}\ \boldsymbol{f}^{(2)}\ \cdots\ \boldsymbol{f}^{(D)}] \in \mathbb{R}^{K\times D}$。在超网络中, 经过特征嵌入[18]后, 网络输入的嵌入向量可以表示为 $\boldsymbol{e} = [\boldsymbol{e}^{(1)}\ \boldsymbol{e}^{(2)}\ \cdots\ \boldsymbol{e}^{(D)}] \in \mathbb{R}^{K\times D}$, 经过训练后, 网络输出稀疏重建后的权重矩阵。

3　模型与方法

3.1　模型结构

模型由主网络和 3 层辅助网络构成, 主网络由多个 MLP 组成, 如图 1 所示。模型不会直接学习权重矩阵 $\boldsymbol{W}_l$, 而是通过辅助网络计算权重向量, 绕过首层的学习, 这样能够削减模型约 92% 的学习参数, 降低过拟合的风险。本文提出的稀疏重构网络如图 2(a) 所示, 辅助网络结构如图 2(b) 所示, 内部网络并行训练, 结合后的整体是一个端到端[19]的稀疏网络。图 3 给出了训练的整体流程。

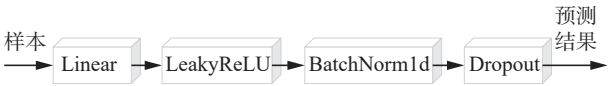
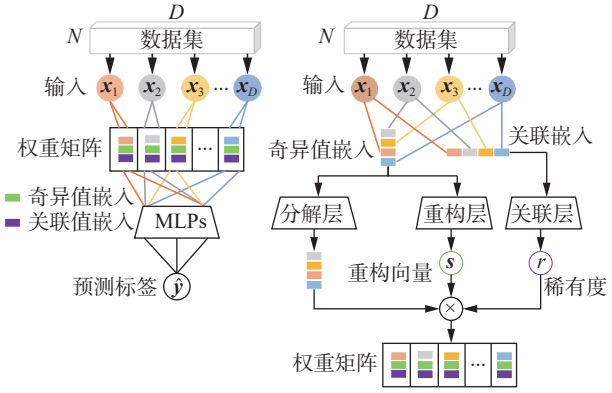


图 1　MLPs 内部结构

Fig. 1　MLPs internal structure



(a) 预测标签的标准网络　　(b) 处理特征权重的稀疏网络

图 2　SRnet 结构

Fig. 2　SRnet structure

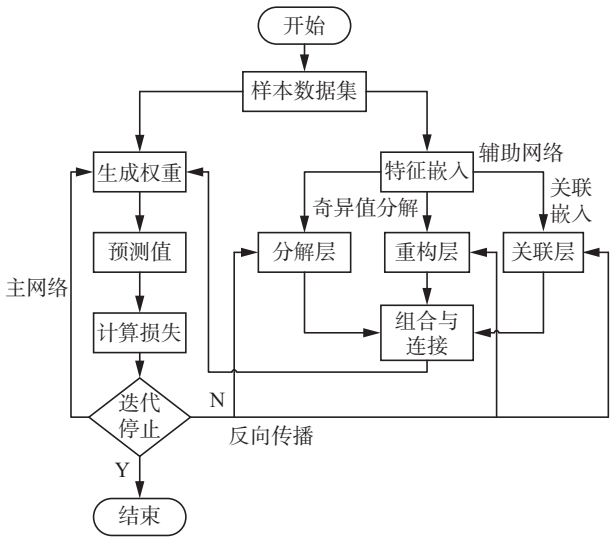


图 3　SRnet 工作流程

Fig. 3　SRnet workflow

3.1.1　主网络

主网络 MLPs 的输入是处理后的矩阵，隐藏层中的神经元数由于经过了削减处理，因此可以调节为 30 左右。由于首层超参数的自适应降低，网络的中间层参数数目会对数级降低，整体提高模型的泛化能力。图 2(a) 中，将权重矩阵输入 MLP 进行训练，训练后得到预测值，计算损失函数后反向传播，更新模型相关的参数。

3.1.2　分解层

权重分解网络 (weight decomposition network, WDN) 表示为 $f_{\mathrm{WDN}}:\mathbb{R}^M\to\mathbb{R}^K$，输入经过奇异值分解嵌入后的特征向量 $\boldsymbol{e}=[\boldsymbol{e}^{(1)}\ \boldsymbol{e}^{(2)}\ \cdots\ \boldsymbol{e}^{(M)}]$，进行特征嵌入能够表征特征更多的信息。WDN 网络层中设置了 5 层 MLP，加入了批量标准化 (batch normalization, BN) 层、Dropout 层，并把最后一层设置为 tanh 激活层，使结果在 $(-1,1)$ 区间内。网络中的可学习参数 θ_{WDN} 不断更新，输出的权重矩阵将特征与首个隐藏层中 K 个神经元来连接起来。

3.1.3　重构层

稀疏重构层 $f_{\mathrm{SRN}}:\mathbb{R}^M\to\mathbb{R}$ 对于每个特征 $\boldsymbol{l}$，奇异值分解嵌入从 $\boldsymbol{X}$ 映射到不同于权重分解层的另一部分，$\boldsymbol{g}=[g^{(1)}\ g^{(2)}\ \cdots\ g^{(M)}]\in\mathbb{R}^M$ 是奇异值向量，作为稀疏重构层的输入。网络输出 $s_l\in(0,1)$，它代表了特征重构分数。s_l 的值越大代表特征的重要性越高，则需要对特征进行小幅重构，反之则说明特征的重要性较低，需要进行大幅重构。从整体来看，重构层能对大量权重进行稀释，而依然保证特征重要性的高低排序。稀疏重构网络设置为 5 层 MLP 结构，包含可学习参数 θ_{SRN}，最后一层为 Sigmoid 激活层。

3.1.4　关联层

增强关联网络 (enhanced relation network, ERN) 表示为 $f_{\mathrm{ERN}}:\mathbb{R}^M\to\mathbb{R}^K$，包含可学习参数 θ_{ERN}，将 $\boldsymbol{X}=[\boldsymbol{x}^{(1)}\ \boldsymbol{x}^{(2)}\ \cdots\ \boldsymbol{x}^{(N)}]\in\mathbb{R}^{N\times D}$ 经过稀有增强后，映射为矩阵 $\boldsymbol{H}=[h^{(1)}\ h^{(2)}\ \cdots\ h^{(M)}]\in\mathbb{R}^M$ 后作为网络的输入。$\boldsymbol{R}^{\mathrm{T}}=[r^{(1)}\ r^{(2)}\ \cdots\ r^{(K)}]\in\mathbb{R}^K$ 为关联网络的输出，r 表示每个特征对应的稀有度。稀有度越高，则代表该特征对应的权重需要被小幅增强。以肺炎数据集为例，假设特征 $\boldsymbol{a}$ 和特征 $\boldsymbol{b}$ 的重要性是大致相同的，与 $\boldsymbol{a}$ 共同出现的特征数量是少的，而与 $\boldsymbol{b}$ 出现的特征数量是较多的，$\boldsymbol{R}$ 向量经过连接形成的矩阵和 WPN 计算出的权重 $\boldsymbol{W}$ 矩阵相乘后特征 $\boldsymbol{a}$ 的值会高于特征 $\boldsymbol{b}$。这能够有效缩减特征的维度，对于分类效果近似的特征组合，独立特征往往在特征规模较小的情况下对标签的影响较大，本网络能够小范围内稀疏权重的同时，进行特征降维。

3.2　**特征嵌入方法**

3.2.1　随机投影嵌入

随机投影嵌入是一种降维技术，用于处理高维数据，它通过随机映射将高维空间中的数据点投影到低维空间，同时保留原始数据的结构信息，以实现降维。先随机生成投影矩阵，矩阵是由独立同分布的随机变量填充而成[20]。然后将原始高维数据进行线性映射，将其投影到低维空间，这里使用 MLP 进行映射，最后得到低维空间的数据表示。

3.2.2　点直方图嵌入

点直方图嵌入法是由 Kaur 等 [21] 于 2020 年提出的一种特征嵌入方法，它用于嵌入特征的分布信息。对于任一特征 $\boldsymbol{l}$，它计算特征值 F 的归一化直方图。设某一直方图柱的高度和中心分别是 $\boldsymbol{h}$ 和 $\boldsymbol{c}$，则原特征值是这 2 个向量的每个元素相乘得到的值，嵌入后的向量能很好地保留特征值的统计信息。

3.2.3　非负矩阵变换嵌入

非负矩阵嵌入法[22] 是将一个非负矩阵 $\boldsymbol{X}$ 分解为 $\boldsymbol{e}^i$ 和 $\boldsymbol{e}^j$，它们分别表示特征矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 的列空间。$\boldsymbol{W}$ 的列空间表示样本的特征空间，$\boldsymbol{H}$ 的列空间表示特征在空间中的坐标，特征经过嵌入后能够发现数据的潜在结构。

3.2.4　奇异值分解嵌入

奇异值分解[23] 通过对样本矩阵进行奇异值分解实现，本文中的模型先将矩阵 $\boldsymbol{X}$ 分解为奇异矩阵和对角矩阵，保留右奇异矩阵 $\boldsymbol{V}$ 和奇异值向量 $\boldsymbol{S}$。模型设置了截断级别，最后得到截断后的

右奇异矩阵 $\boldsymbol{V}$ 和奇异值矩阵 $\boldsymbol{S}$, $\boldsymbol{V}$ 作为分解层的嵌入, $\boldsymbol{S}$ 作为重构层的嵌入。

3.2.5 关联嵌入

特征选择网络普遍关注的是特征组合的分类能力而缺乏对冗余特征的处理。本文提出了关联嵌入法, 对于冗余相关的特征, 它会赋予独立的特征更高的权重。计算特征的统计学信息量作为嵌入结果[24], 根据活跃系数ε, 记录每个特征单独出现的频率π, 然后计算每个特征与其他特征共同出现的频率φ, 最后计算稀有因子r, $S(x)$表示 Sigmoid 激活函数α。经过首层 softmax 函数激活后, r能够表征特征的稀有度。稀有增强算法实现流程见算法 1。

算法 1　稀有增强算法

输入　样本矩阵 $\boldsymbol{X}=[\boldsymbol{x}^{(1)}\ \boldsymbol{x}^{(2)}\ \cdots\ \boldsymbol{x}^{(N)}]\in\mathbb{R}^{N\times D}$, 敏感系数$\alpha$, 活跃系数$\varepsilon$。

输出　稀有系数r。

1)for each feature $i=1,\cdots,$D do
2)if $\boldsymbol{x}[i]>f(i,\boldsymbol{x},\varepsilon)$:
3)update π
4)for each feature $j=i+1,\cdots,D$ do
5)if $\boldsymbol{x}[j]>$ g$(i,j,\boldsymbol{x},\varepsilon)$:
6)update φ
7)$r=1-S(\pi/\varphi)$
8)end for
9)end for
10)return r

3.3 训练过程

网络从 3 层辅助网络开始, 由辅助网络生成的权重矩阵输入主网络, 经过训练后计算损失函数, 并进行梯度下降、反向传播更新参数, 因此 4 层网络同时进行训练。模型的目标是进行特征选择, 选择出能够提高分类准确率的特征, 因此损失函数即要降低维度, 还要提高分类能力。

$$L(\theta_{\mathrm{WDN}},\theta_{\mathrm{SRN}},\theta_{\mathrm{ERN}},\boldsymbol{W})=\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\theta_{\mathrm{WDN}}},f_{\theta_{\mathrm{SRN}}},f_{\theta_{\mathrm{ERN}}};f_{W}(\boldsymbol{x}^{(i)}),y^{i})+\lambda\sum_{l=1}^{D}f_{\theta_{\mathrm{SRN}}}(\boldsymbol{e}^{l})+\sigma^{2}\sum_{m=1}^{D}f_{\theta_{\mathrm{ERN}}}(\boldsymbol{e}^{m})\tag{1}$$

算法 2　SRnet 模型

输入　样本矩阵 $\boldsymbol{X}=[\boldsymbol{x}^{(1)}\ \boldsymbol{x}^{(2)}\ \cdots\ \boldsymbol{x}^{(N)}]\in\mathbb{R}^{N\times D}$, 训练标签 $\boldsymbol{Y}=[y^{(1)}\ y^{(2)}\ \cdots\ y^{(N)}]\in\mathbb{R}^{N}$, 敏感系数$\alpha$, 活跃系数$\varepsilon$, 主网络$f_{\theta}$, 权重分解网络$f_{\mathrm{WDN}}$, 稀疏重构网络$f_{\mathrm{SRN}}$, 关联网络$f_{\mathrm{ERN}}$, 稀疏因子$\lambda$, 关联因子$\sigma$, 学习率$\rho$。

输出　训练模型。

1)update embedded vector $\boldsymbol{e}$, $\boldsymbol{h}$
2)for each minibatch $\boldsymbol{B}=\{(\boldsymbol{x}^{(i)},y_i)\}_{i=1}^{N}$ do
3)for each feature $i=1,\cdots,b$ do
4)$w^{(l)}=f_{\theta_{\mathrm{WDN}}}(\boldsymbol{e}^{(l)})$
5)$s^{(l)}=f_{\theta_{\mathrm{SRN}}}(\boldsymbol{e}^{(l)})$
6)$h_l=f_{\theta_{\mathrm{ERN}}}(h^{(l)},\alpha,\varepsilon)$
7)end for
8)$\boldsymbol{W}^{[1]}=[f^{(1)}\ f^{(2)}\ \cdots\ f^{(N)}]\in\mathbb{R}^{K\times D}$
9)for each sample $j=1,\cdots,c$ do
10)$\hat{y}_j=f_{\theta}(\boldsymbol{x}^{(j)})$
11)end for
12)update L with formula (2)
13)$\theta_{\mathrm{WDN}}\xleftarrow{\text{梯度下降}}\theta_{\mathrm{WDN}}-\alpha\nabla_{\theta_{\mathrm{WDN}}}L$
14)$\theta_{\mathrm{SRN}}\xleftarrow{\text{梯度下降}}\theta_{\mathrm{SRN}}-\alpha\nabla_{\theta_{\mathrm{SRN}}}L$
15)$\theta_{\mathrm{ERN}}\xleftarrow{\text{梯度下降}}\theta_{\mathrm{ERN}}-\alpha\nabla_{\theta_{\mathrm{ERN}}}L$
16)end for

网络的整体损失函数由标签与预测值的平均交叉熵损失 (包含主网络和辅助网络部分损失)、重构损失和稀有度损失 (正则损失) 组成, 如式 (1)。θ_{WDN}、θ_{SRN}、θ_{ERN}均为可微损失, 由于辅助网络均用来计算主网络中的权重矩阵, 因此可以计算出参数θ_{WDN}、θ_{SRN}、θ_{ERN}的梯度。交叉熵损失部分保证特征集合在训练的过程中, 分类准确率升高, 而正则损失使特征集合变得稀疏。从特征削减的角度来看, $\lambda\sum f_{\theta_{\mathrm{SRN}}}(\boldsymbol{e}^{l})$减少学习参数的同时, 进行特征的稀疏。交叉熵损失部分的降低能够保证筛选后特征的分类能力。$\sigma^{2}\sum f_{\theta_{\mathrm{ERN}}}(\boldsymbol{e}^{m})$促进选择独立特征, 使分类水平大致相同的情况下, 少量特征的作用更占优势。

网络采用基于梯度的标准优化算法 Adamw[25], 由于网络中的超参数经过神经网络训练和矩阵计算, 所以可以计算梯度, 并使用梯度下降法更新参数。模型计算 4 个网络中参数的梯度, 4 个网络同时进行端到端训练, 不断降低目标函数值。经过长期训练, 没有发现优化层面的问题, 参数能够被很好地更新。

4 实验

4.1 实验内容

4.1.1 实验简述

本章先评估模型的特征选择性能, 并从整体对多种小样本高维数据集的不同模型进行效果的比较。通过 6 个实验进行分析, 首先比较 8 种分类算法的分类准确率, 之后分析 5 种特征嵌入方

法对模型产生的效果，进行消融实验并分析了稀疏重构、关联网络对于 SRnet 性能的影响，分析比较最新的网络 WPFS(weight predictor network with feature selection)[14]、本网络和 MLPs 的损失函数下降曲线，通过本算法产生的特征重要性分布反映特征数据集的分类难度，探究正则化参数对模型效果和特征的影响。

4.1.2　数据集和设置

本文针对小样本高维特征选择问题，必须选择样本数少、特征数量高的数据集，因此本文选择了 12 个生物医学数据集[26]。样本数量范围为 50 ~ 203，每个样本的特征数量高于 2 000，标签数量范围为 2 ~ 9。

每个数据集执行 5 次交叉验证，重复 10 次，模型运行 50 次，选择 15% 作为训练数据，恰好模拟小型数据集的场景。主网络设置为 4 层神经网络，中间层设置为 100 个神经元，最后一层作为激活层，使用 softmax 激活。WDN、SRN、ERN 都是 5 层神经网络，WDN 采用 tanh 激活，SRN 采用 Sigmoid 激活，ERN 采用 softmax 激活。每个网络都采用了 BN、Dropout 层 (p=0.25) 和 LeakyReLU 进行映射。以 8 为 batch 大小进行训练。使用 AdamW 优化器进行优化，学习率范围在 $[2\times10^{-1},2\times10^{-3}]$，初始选择0.2，不断衰减到 2×10^{-3} 之后固定使用此学习率。

4.2　算法性能实验

4.2.1　嵌入方法对比实验

首先进行不同嵌入方法的对比实验，实验在多个数据集下进行，从 dietnet、WPN 等模型中选取高效的嵌入方法进行实验，结果如图 4 所示。原始特征嵌入和投影嵌入方法由于无法进一步探究特征数据间的潜在影响，因此实验产生了较差的分类准确率。点直方图在一些离散的数据集下能够产生较好的效果，但从总体来看，仍然是奇异值分解嵌入和非负矩阵嵌入能够产生更好的效果，因为它们在连续型数据上也产生了不错的效果。经过对比，奇异值分解嵌入的实验模型表现出了最好的效果。

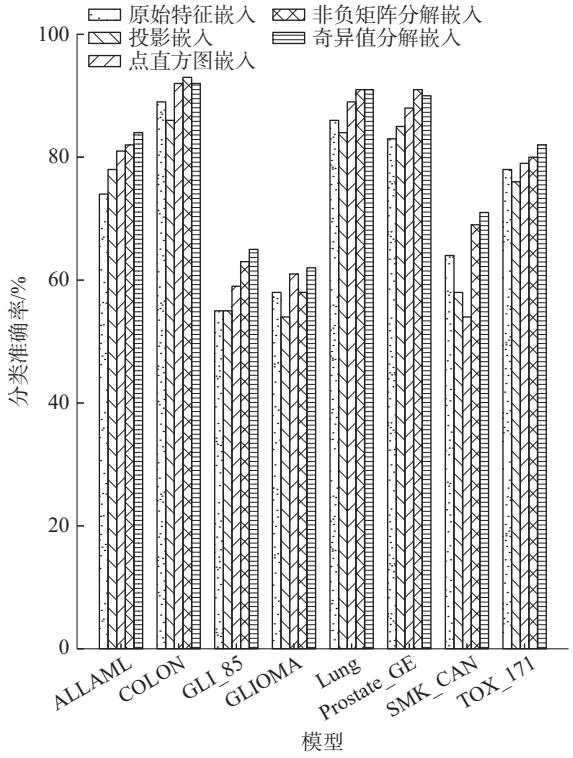


图 4　采用不同嵌入方法的模型准确率

Fig. 4　Model accuracy using different embedding methods

4.2.2　多数据集下的算法对比实验

表 2 给出了在 12 种数据集下，8 种算法的性能对比，考察分类模型运行 50 次的平均准确率。从表中可以看出，SRnet 模型平均每个数据集下都要高出平均得分 2% 以上，证明了模型良好的特征选择能力。

表 2　8 个特征选择模型分类效果评分

Table 2　Classification score of 8 feature selection model

数据集	SRnet	WPFS[14]	FsNet[13]	dietnet[11]	lassonet[16]	DNP[6]	LGBM[8]	MLP[7]
ALLAML	+6.93	+5.33	−2.58	+2.63	−4.08	−1.18	−4.28	−2.77
CLL_SUB	+4.42	+5.34	+2.29	+3.55	−7.91	−5.05	−0.88	−1.76
COLON	+3.72	+5.18	−4.14	+3.30	−3.86	−4.79	−1.97	+2.56
GLI_85	+2.47	+3.16	−1.07	+7.01	−5.22	−5.26	−3.24	+2.15
GLIOMA	+2.59	+2.08	−1.41	−1.23	+1.09	−1.25	−3.30	+1.43
Leukemia	+2.52	−0.22	+0.41	+2.86	−3.19	+3.30	−4.98	−0.70
Lung	+3.64	+1.74	−0.84	−0.31	+2.16	−2.30	−0.26	−3.83
Lymphoma	+4.60	+2.66	+2.23	−2.09	−0.81	+1.35	−6.16	−1.78
nci9	+5.17	−1.29	+3.92	+1.02	−2.06	−5.57	+0.62	−1.81
Prostate_GE	−0.12	+1.19	+4.31	+1.21	−1.66	−2.28	+2.42	−5.07
SMK_SCAN	+1.87	−0.66	+4.29	+1.92	+1.20	−3.06	−2.05	−3.51

续表 2

数据集	SRnet	WPFS[14]	FsNet[13]	dietnet[11]	lassonet[16]	DNP[6]	LGBM[8]	MLP[7]
TOX_171	+1.31	+2.62	+3.79	+2.87	+1.36	−3.30	−5.96	−2.69
超出和	+39.12	+27.13	+11.20	+22.74	−22.98	−29.39	−30.04	−17.78
平均得分	+3.26	+2.26	+0.93	+1.90	−1.92	−2.45	−2.50	−1.48

4.3　消融实验

在多个数据集下, 分别进行了加入分解层、重构层、关联层、以及加入全部稀疏重构网络的实验, 实验数据各取运行 50 次的平均分类准确率。

对于单独加入分解层的实验, 输入进入分解层, 不进入其他层, 不与权重因子进行积运算, 而是直接输出到主网络。当实验不加入分解层时, 重构层和关联层的输出直接和主网络的权重矩阵求点积。如图 5 所示, 图中加入整个网络后分类准确率提升了约 5.9 ~ 7.9 百分点, 其中最重要的是超网络的分解层部分, 单独使用它时分类准确率能够提升 3.8 ~ 7.0 百分点。关联层和重构层对分类准确率的提升起到小幅作用, 大约提升幅度在 0.2 ~ 2.3 百分点, 这可以归结为其对于主网络权重的稀疏作用, 以及稀疏因子和关联因子通过学习特征的重要程度对于选出特征的分类能力的优化。

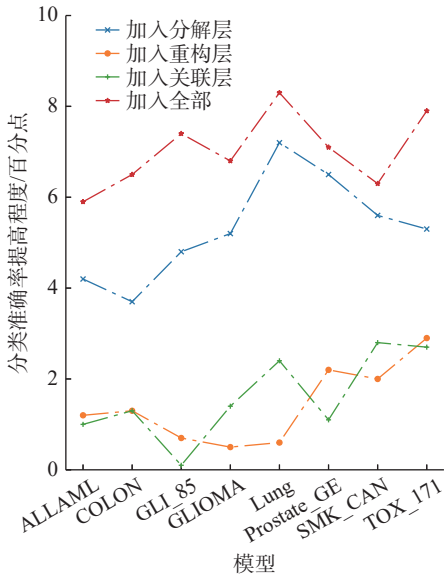


图 5　多数据集下的消融实验

Fig. 5　Ablation experiments with multiple datasets

4.4　迭代实验

图 6 给出了在 Lung 数据集下, SRnet 和 MLPs 的损失下降曲线, 可以看出 MLPs 有十分严重的过拟合问题。MLPs 的损失在下降的过程中有时并没有向验证曲线靠拢, 而是表现出了很大的波动, 这会产生很大的方差。

图 7 给出了 Lung 数据集下 SRnet 与 WPFS 的损失下降曲线, 可以看出虽然 WPFS 在学习的过程中, 能够在某一个点学习到较好的损失值, 但整体还是有过拟合的倾向, 并没有模型的下降曲线稳定。

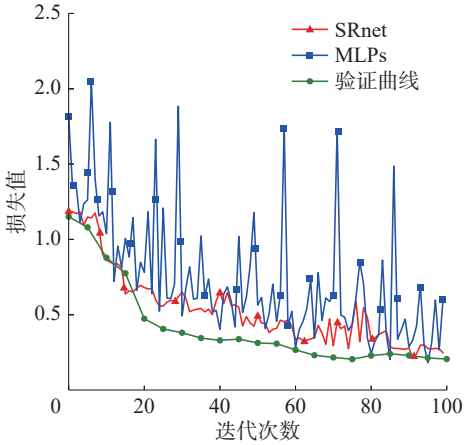


图 6　Lung 数据集下损失下降曲线

Fig. 6　Loss decline curve on Lung dataset

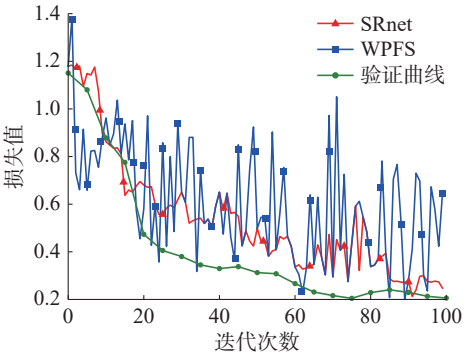


图 7　模型与 WPFS 的损失下降对比

Fig. 7　Loss reduction comparison between model and WPFS

SRnet 的曲线较贴近于验证曲线, 降低了训练过程的整体方差, 能够剔除分类水平较差的特征组合。并且 SRnet 更贴合验证曲线, 减小了过拟合的影响, 学习效果更好, 这表明了改进后的模型具有稳定性和高效性。

4.5　探究超参数的影响

本文对损失函数中λ和σ进行了实验调节, 当损失值趋于稳定, 实验表明$\lambda = 10^{-3}$, $\sigma = 10^{-2}$保证了参数稳定的同时, 能够将特征降维到原来的 0.6% ~ 0.8%。为了进一步比较重构层和关联层的效果, 分别进行λ和σ取值范围的 50 次对比实验, 结果取平均值。图 8 体现了 2 个子网络对于提高

分类准确率给予的帮助，可以发现在不同数据集下两者都产生了相应的效果。

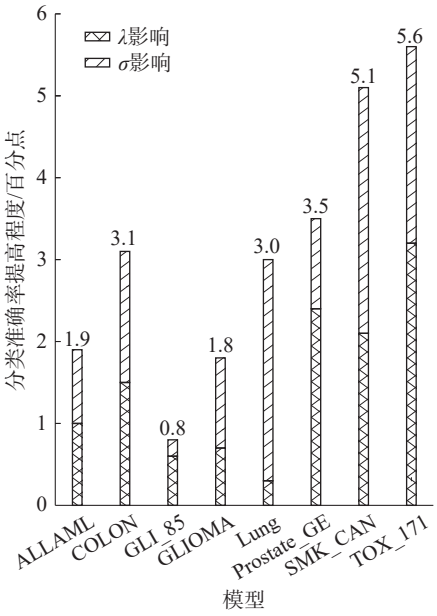


图 8　多数据集下λ和σ的影响水平

Fig. 8　Influence level of multiple datasets

经过对比，关联层对于更高维的特征数据集产生的效果更明显，而重构层对更多标签的特征数据集产生了更好的效果。这与两者使用了不同的特征嵌入方法有关，关联层能够挖掘少量的高效特征，而重构层对特征数据本身的特点进行了挖掘，通过此实验再次证明了重构层和关联层产生的效果都能不同程度地提高分类准确率。

4.6　模型的应用分析

4.6.1　应用于特征选择

本模型具有原生的特征选择能力，在迭代后期阶段，权重矩阵中 96% 以上的权重均为 0 或小于10^{-3}，那么对应特征被剔除，其余特征可根据特征规模、s_l、r_l的排序进行选择。模型中设计的损失函数具有自适应过滤的作用，正则化损失能够使无效特征的权重下降，而交叉熵损失保证有效特征的权重不会被下降到较低的水平。交叉熵损失具有决定性作用，为了使损失函数下降，有效特征不会被剔除。而因为无效特征不会影响标签的预测准确率，因此其权重将会被下降到接近于 0。综上所述模型能够应用于特征选择领域，进行特征降维工作。

4.6.2　应用于特征的语义推测

特征选择往往与挖掘特征原本的含义相关，在实验的过程中发现，$s_l \cdot r_l$能够表征特征的重要性，当λ与σ增大时，特征数量会减少。通过逐步缩减这 2 个参数能探究在分类准确率近似的情况下，哪部分特征逐步被剔除或替换掉。通过结合$s_l \cdot r_l$与领域知识分析这部分特征产生的作用，能够对部分特征进行语义推测，从而应用于医学、声学等领域。

4.6.3　应用于任务分类

图 9 分析了 12 个数据集下划分的 3 种特征重要性$s_l \cdot r_l$分布类型，$s_l \cdot r_l$经过了归一化，柱状图中每个分布的和均为 1。图 9(a) 体现了简单任务的重要性分布。本模型能保留特征重要性较高的部分，剔除重要性较低的部分。图 9(b) 的中等任务不容易直接划分出特征，而图 9(c) 中体现的困难任务几乎都是中等重要性的特征，很难直接进行划分。因此实验中采取了一种折中的保留策略：剔除重要性分数在 0.1 以下的特征，并尽可能的保留前 1.8% 特征维度的特征。图 9(d) 中分析了这 3 种任务曲线产生的方差比率，简单任务的方差较高，中等任务较低，复杂任务最低。因此可以通过实验设定一个方差阈值，将其作为划分任务难度的标准。实验按此方法将 12 个数据集划分出 3 种类别，简单任务 (COLON、Lung、Leukemia、Prostate_GE)，中等任务 (GLIOMA、TOX_171、CLL_SUB、Lymphoma)，困难任务 (SMK_SCAN、ALLAML、nci9)。

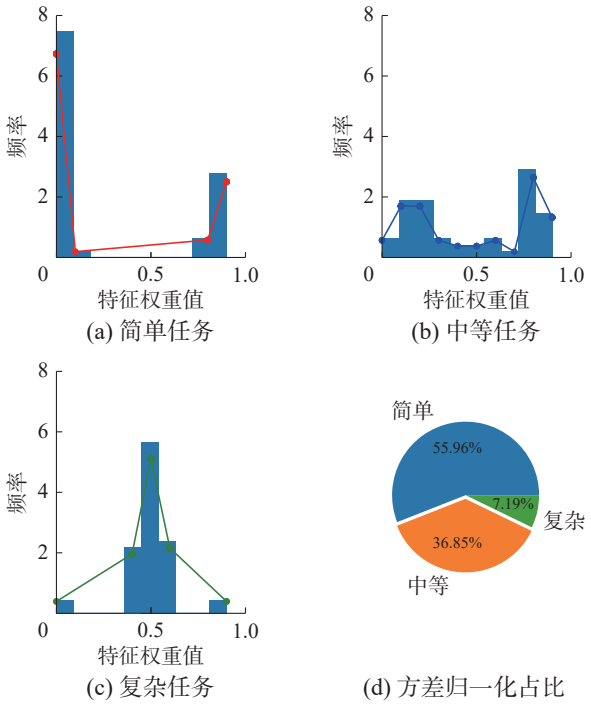


图 9　特征重要性对任务分类的贡献

Fig. 9　Contribution of feature importance to task classification

5　结束语

本文提出了对高维小样本数据进行特征选择的稀疏重构网络。受到 MLPs 论文的启发，在小

样本领域复杂网络的不适用性会导致严重的过拟合倾向; 受超网络的启发, 使用辅助网络对主网络进行权重削减可以有效降低过拟合的可能。因此本文采用了 4 个微型网络进行并行训练, 主网络的权重参数由 3 个辅助网络计算而来, 端到端地进行特征选择。分解层能够削减网络权重, 重构层能够进一步稀疏权重, 并且作为判定特征重要性评分的一部分, 另一部分由关联层训练输出。在 12 个现实生物医学数据集下进行了一系列实验证明了模型的效果, 对于多个生物领域数据集在 8 种模型的对比实验中优于其他 7 种模型。通过实验验证了模型的每一部分对分类准确率的提升效能, 并且在实验的过程中发现了很多可能的模型应用, 它能够帮助学者划分出不同研究难度的任务, 通过对模型调参, 结合专业知识可能挖掘出特征的部分意义。同时也说明小样本高维特征数据中还有很多值得挖掘的信息, 这或许对专业知识进一步的探索与发展有很大的帮助。未来也将继续研究特征选择网络的不同模型结构, 致力于探索小样本数据集的高维特征中所蕴含的更多信息。

参考文献:

[1] 余紫康, 董红斌. 具有混合策略的樽海鞘群特征选择算法[J]. 智能系统学报, 2024, 19(3): 757−765.
YU Zikang, DONG Hongbin. Salp swarm feature selection algorithm with a hybrid strategy[J]. CAAI transactions on intelligent systems, 2024, 19(3): 757−765.

[2] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of machine learning research, 2014, 15: 1929−1958.

[3] 金红, 胡智群. 基于非负矩阵分解的稀疏网络社区发现算法[J]. 电子学报, 2023, 51(10): 2950−2959.
JIN Hong, HU Zhiqun. The non-negative matrix factorization based algorithm for community detection in sparse networks[J]. Acta electronica sinica, 2023, 51(10): 2950−2959.

[4] 谢承旺, 郭华, 韦伟, 等. MaOEA/d^2: 一种基于双距离构造的高维多目标进化算法[J]. 软件学报, 2023, 34(4): 1523−1542.
XIE Chengwang, GUO Hua, WEI Wei, et al. MaOEA/d^2: many-objective evolutionary algorithm based on double distances[J]. Journal of software, 2023, 34(4): 1523−1542.

[5] 胡艳梅, 杨波, 多滨. 基于网络结构的正则化逻辑回归[J]. 计算机科学, 2021, 48(7): 281−291.
HU Yanmei, YANG Bo, DUO Bin. Logistic regression with regularization based on network structure[J]. Computer science, 2021, 48(7): 281−291.

[6] LIU Bo, WEI Ying, ZHANG Yu, et al. Deep neural networks for high dimension, low sample size data[C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne: IJCAI, 2017: 2287−2293.

[7] KADRA A, LINDAUER M, HUTTER F, et al. Well-tuned simple nets excel on tabular datasets[EB/OL]. (2021−06−21)[2024−02−20]. https://arxiv.org/abs/2106.11189.

[8] CHEN Tianqi, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016: 785−794.

[9] VON OSWALD J, HENNING C, GREWE B F, et al. Continual learning with hypernetworks[EB/OL]. (2019−06−03)[2024−02−20]. https://arxiv.org/abs/1906.00695.

[10] ZHANG Lin, LI Xin, HE Dongliang, et al. RRSR: reciprocal reference-based image super-resolution with progressive feature alignment and selection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 648−664.

[11] ROMERO A, CARRIER P L, ERRAQABI A, et al. Diet networks: thin parameters for fat genomics[C]//International Conference on Learning Representations. Toulon: OpenReview.net, 2017: 1−11.

[12] BALIN M F, ABID A, ZOU J Y. Concrete autoencoders: differentiable feature selection and reconstruction[C]//International Conference on Machine Learning. Los Angeles: PMLR, 2019: 444−453.

[13] SINGH D, CLIMENTE-GONZALEZ H, PETROVICH M, et al. FsNet: feature selection network on high-dimensional biological data[C]//2023 International Joint Conference on Neural Networks. Gold Coast: IEEE, 2023: 1−9.

[14] MARGELOIU A, SIMIDJIEVSKI N, LIÒ P, et al. Weight predictor network with feature selection for small sample tabular biomedical data[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023: 9081−9089.

[15] 黄路路, 唐舒宇, 张伟, 等. 基于 Lp 范数的非负矩阵分解并行优化算法[J]. 计算机科学, 2024, 51(2): 100−106.
HUANG Lulu, TANG Shuyu, ZHANG Wei, et al. Non-negative matrix factorization parallel optimization algorithm based on Lp-norm[J]. Computer science, 2024, 51(2): 100−106.

[16] LEMHADRI I, RUAN Feng, TIBSHIRANI R. LassoNet:

neural networks with feature sparsity[J]. The journal of machine learning research, 2021, 22(1): 5633−5661.

[17] OBOZINSKI G, TASKAR B, JORDAN M. Multi-task feature selection[J]. Statistics department, UC Berkeley, Tech. Rep, 2006, 2(2.2): 2.

[18] HUANG Xiao, SONG Qingquan, YANG Fan, et al. Large-scale heterogeneous feature embedding[C]// Proceedings of the AAAI conference on artificial intelligence. Honolulu: AAAI, 2019: 3878−3885.

[19] 刘宇宸, 宗成庆. 跨模态信息融合的端到端语音翻译[J]. 软件学报, 2023, 34(4): 1837−1849.
LIU Yuchen, ZONG Chengqing. End-to-end speech translation by integrating cross-modal information[J]. Journal of software, 2023, 34(4): 1837−1849.

[20] CHEN Haochen, SULTAN S F, TIAN Yingtao, et al. Fast and accurate network embeddings *via* very sparse random projection[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing: ACM, 2019: 399−408.

[21] KAUR G, SINGH S, RANI R, et al. A comprehensive study of reversible data hiding (RDH) schemes based on pixel value ordering (PVO)[J]. Archives of computational methods in engineering, 2021, 28(5): 3517−3568.

[22] KHAN Z, ILTAF N, AFZAL H, et al. Enriching non-negative matrix factorization with contextual embeddings for recommender systems[J]. Neurocomputing, 2020, 380: 246−258.

[23] HUANG Tianlin, ZHAO Rujie, BI Lvqing, et al. Neural embedding singular value decomposition for collaborative filtering[J]. IEEE transactions on neural networks and learning systems, 2022, 33(10): 6021−6029.

[24] LIU Haoyue, ZHOU Mengchu, LIU Qing. An embedded feature selection method for imbalanced data classification[J]. IEEE/CAA journal of automatica sinica, 2019, 6(3): 703−715.

[25] YAO Zhewei, GHOLAMI A, SHEN Sheng, et al. ADAHESSIAN: an adaptive second order optimizer for machine learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2021: 10665−10673.

[26] YANG Junchen, LINDENBAUM O, KLUGER Y. Locally sparse neural networks for tabular biomedical data[C]// International Conference on Machine Learning. Baltimore: PMLR, 2022: 25123−25153.

作者简介：

魏俊伊，硕士研究生，主要研究方向为群智能算法和深度学习。E-mail：weijunyi@hrbeu.edu.cn。

董红斌，教授，博士生导师，中国计算机学会高级会员，主要研究方向为多智能体系统、机器学习。主持和完成国家自然科学基金项目、工信部基础研究项目、黑龙江省自然科学基金项目，荣获黑龙江省高校科学技术奖和黑龙江省优秀高等教育科学成果奖。发表学术论文 90 余篇，出版教材 2 部。E-mail：donghongbin@hrbeu.edu.cn。

余紫康，硕士研究生，主要研究方向为群智能算法、数据挖掘。E-mail：yuzk6@mail2.sysu.edu.cn。