Adaptive Model Fusion Framework Driven by Data Similarity and Model Reliability

WANG Mei*, LI Yanpei, GAO Yatian

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Adaptive model fusion is particularly important for dynamically responding to the evolutionary characteristics of data and tasks. However, existing model fusion methods still have issues such as static weights being difficult to adapt to data similarity, dynamic fusion being driven by single factors, and being susceptible to data distribution drift. To address these shortcomings, this paper proposes an adaptive model fusion method driven by data similarity and model reliability. The method captures the similarity between samples through feature semantic alignment to obtain a similarity matrix, and further obtains the sample matching degree coefficient. Then, based on the base model selection algorithm of performance-diversity, the generalization ability and local performance of the base models are evaluated through multi-dimensional metrics to obtain the reliability coefficient of the base models. Finally, the fusion weight is calculated based on the data similarity coefficient and the reliability coefficient of the base models to obtain the final fusion model strategy. Experimental results on public datasets demonstrate the effectiveness of the proposed method.

Highlights:

1.Propose an adaptive model fusion method driven by data similarity and model reliability. By incorporating data distribution characteristics and model reliability into model fusion, the method maximizes the prediction performance of the model and realizes adaptive model fusion.

2.Propose a precise mixed data similarity measurement module.It achieves semantic alignment of numerical and categorical heterogeneous features through deep embedding, integrates improved K-Prototypes clustering to output sample-cluster similarity vectors, and underpins sample-level local dynamic adaptation.

3.Design a performance-diversity dual-goal optimization-based base model selection mechanism, leveraging multi-dimensional evaluation, diversity quantification, and dynamic decaying weights to automatically prune redundant models, adjust reliability weights, and boost fusion robustness.

4. Propose an adaptive model fusion framework that does not rely on scenario-specific prior distribution assumptions. It can flexibly adapt to data drift and heterogeneous model fusion requirements in different fields, providing an adaptive fusion solution for complex tasks.

Key words: adaptive model fusion; reliability assessment; sample compatibility; multi-dimensional metrics; dynamic weighting

---

Foundation items: National Natural Science Foundation of China (Nos.51774090, 62076234); Heilongjiang Science and Technology Innovation Base Project (No.JD24A009); Heilongjiang Natural Science Foundation (No.LH2024F005); Heilongjiang Postdoctoral Research Start-Up Fund (No.LBH-Q20080).

Received: 2025-06-15; **Revised**: 2025-07-03

***Corresponding author**, **E-mail**: wangmei@nepu.edu.cn.

ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 3,May 2026,pp. 767－779
DOI:10. 16337/j. 1004-9037. 2026. 03. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于数据相似性和模型可靠度驱动的自适应模型融合方法

王　梅，李艳培，高雅田

（东北石油大学计算机与信息技术学院，大庆 163318）

摘　要：自适应模型融合对于动态应对数据与任务的演化特性尤为重要。然而，现有模型融合方法仍存在静态权重难以适配数据相似性、动态融合受单因素驱动和易受数据分布漂移影响等问题。针对这些不足，本文提出基于数据相似性与模型可靠度驱动的自适应模型融合方法。该方法通过特征语义对齐，捕捉样本间的数据相似性，得到相似矩阵，进一步得到样本匹配度系数。然后，根据性能-多样性的基模型筛选算法，通过多维度度量评估基模型的泛化能力与局部性能，得到基模型的可靠度系数。最后，根据数据相似性系数和基模型的可靠度系数，进行融合权重计算，得到最终的融合模型策略。在公共数据集上的实验结果证明了本文所提方法的有效性。

关键词：自适应模型融合；可靠性评估；样本匹配度；多维度度量；动态加权

中图分类号：TP311.13;TP309　　**文献标志码：**A

引用格式：王梅，李艳培，高雅田．基于数据相似性和模型可靠度驱动的自适应模型融合方法[J]．数据采集与处理，2026，41(3)：767-779. WANG Mei，LI Yanpei，GAO Yatian. Adaptive model fusion framework driven by data similarity and model reliability[J]. Journal of Data Acquisition and Processing，2026，41(3)：767-779.

引　　言

模型融合作为突破单一模型性能瓶颈的关键技术，通过集成多基模型的预测优势，显著提升了机器学习系统在图像分类、自然语言处理及推荐系统等任务中的泛化能力与鲁棒性[1-3]。在模型融合研究中，集成学习[4]是最为核心的思想之一。集成学习通过结合多个基模型的预测结果，能够有效降低模型的方差或偏差，从而提升模型的性能。典型的集成学习方法包括Bagging、Boosting、Stacking和Blending[5-8]等。例如，随机森林通过Bagging方法结合多个决策树模型，显著提升了分类任务的准确性和鲁棒性；而XGBoost[9]和LightGBM[10]等Boosting算法则通过迭代训练多个弱学习器，进一步提升了模型的预测性能。

传统模型融合策略（如简单平均法、加权平均法等）的核心缺陷在于静态权重分配机制：权重在训练阶段通过验证集性能固定，推理时无法随输入数据的分布变化（数据漂移）或基模型的异构性调整[11]，这种方式忽视了实际应用中“动态适配性”的需求。传统静态融合作为模型融合的基础范式，其本质是训练阶段确定权重后，推理阶段保持不变。然而，这种机制难以应对复杂场景的动态需求。例如，在医疗影像诊断场景中，不同模态数据（如CT、MRI等）的动态输入需要针对性调整模型权重[12]；在金融风控场景中，欺诈模式演变要求实时更新基模型贡献度[13]。静态融合方法难以满足此类动态需求，成为制约实际应用的核心瓶颈。此外，动态集成学习方法[14]通常根据局部精度进行基模型的选择，进而增强模型的适应能力。主要的动态集成方法有OLA[15]、LCA[16]，该方法通过邻域驱动进行模型选

基金项目：国家自然科学基金(51774090，62076234)；黑龙江省科技创新基地项目(JD24A009)；黑龙江省自然科学基金(LH2024F005)；黑龙江省博士后科研启动基金(LBH-Q20080)。
收稿日期：2025-06-15；**修订日期：**2025-07-03

择机制,提升了复杂数据分布下模型的预测性能。随着机器学习和深度学习的持续发展,模型融合技术也在不断进步。自适应模型融合可以提升模型的敏感度,通过不同的数据输入动态地调整基模型来提升模型的的准确性和泛化性。Li等[17]提出深度模型融合这个新兴技术,将多个深度学习模型的参数预测合并为一个模型。该技术结合了不同模型的能力,弥补了单个模型的偏差和误差,以实现更好的性能。Lin等[18]提出了一种自适应多模型融合方法来预测建筑能耗,旨在为更好的能源控制提供有益的建议。自适应多模型融合方法在此研究中被证明优于基于两阶段聚类的回归方法和线性融合方法。得益于对聚类间模糊区样品的适当处理和融合过程中的筛选算法,本文提出的方法最终为建筑能源性能的分析和控制提供了更好的指导。Guo等[19]提出了一种简单灵活的相似性融合模型,用于整合多组学数据以识别癌症亚型,其中考虑每个组学数据中样本之间的相似性偏差,并使用广义线性模型预测样本之间的校正相似性,然后根据不同的数据视图权重从多组学数据中整合校正后的相似性信息。孙若凡[20]提出基于机器学习的模型融合方法,提出在模型中引入卷积神经网络(Convolutional neural network, CNN)深入挖掘互相关和自相关特征,以改善预测模型对数据特征分析能力,并以长短期记忆(Long short-term memory, LSTM)网络为时间序列预测的基础,建立5种不同的模型融合结构并预测未来天然铀价格,结合敏感性测试方法进行分析对比,发现采用CNN插入的LSTM-CNN-LSTM模型的预测效果最好,且受超前预测步和时间窗的影响较小。

综上,已有文献将数据分布和模型可靠度加入到模型融合中,以最大限度地提升模型的预测性能,从而实现自适应模型融合。但是,仍存在一些挑战和问题如下:(1) 传统加权策略依赖训练集整体分布,无法响应测试样本的局部特征;(2) 静态融合规则无法适应输入样本的实时变化或模型性能差异;(3) 数据-模型耦合性不足,多数方法聚焦单维度优化(如权重或邻域划分),缺乏对数据分布漂移与模型异构性的联合建模。针对上述挑战,本文设计了一种基于数据相似性和模型权重优化的自适应模型融合方法,主要贡献包括以下3方面:(1) 提出了一种基于数据相似性度量的动态融合机制,通过计算测试样本与训练数据的相似性,确定每个测试样本的局部特征与不同训练子集的关联程度;(2) 提出性能-多样性双目标优化的基模型选取与权重分配方法,以便解决冗余模型剔除、权重合理分配和可扩展性保障;(3) 在多个公共数据集上开展实验验证,实验结果能够证明此融合方法能够较好地进行回归预测,有效提升了回归预测的准确性和鲁棒性。

1 相关理论

1.1 K-Prototypes算法

K-Prototypes算法是由Huang[21]于1998年提出的混合数据类型聚类方法,旨在同时处理包含数值型和类别型特征的数据集,其核心思想是结合K-Means[22]和K-Modes[23]的优势,通过定义混合距离度量来优化聚类效果。K-Prototypes算法的相异度计算由数值型属性和分类型属性两部分构成,并引入权重因子调整其贡献度。样本为$\boldsymbol{X}_i=\{x_1,x_2,\cdots,x_m,x_{m+1},\cdots,x_{m+p}\}$,其中前$m$个样本是数值型,后面$p$个样本是分类型,$\boldsymbol{X}_i$到聚类中心$\boldsymbol{c}_j$的距离公式为

$$D(\boldsymbol{X}_i,\boldsymbol{c}_j)=\gamma\cdot D_{\text{num}}(\boldsymbol{X}_i,\boldsymbol{c}_j)+(1-\gamma)\cdot D_{\text{cat}}(\boldsymbol{X}_i,\boldsymbol{c}_j) \tag{1}$$

式中:γ为权重因子,D_{num}和D_{cat}分别代表数值型属性的相异度和分类型属性的相异度。D_{num}通常采用欧氏距离,表达式为

$$D_{\text{num}}(\boldsymbol{X}_i,\boldsymbol{c}_j)=\sum_{k=1}^{m}(x_{ik}-c_{jk})^2 \tag{2}$$

式中:m为数值型特征维度,x_{ik}为样本$\boldsymbol{X}_i$的第k个数值型属性的值,c_{jk}为聚类中心$\boldsymbol{c}_j$的第k个数值型属性的值。D_{cat}采用汉明距离(即类别不一致的计数),有

$$D_{\text{cat}}(\boldsymbol{X}_i, \boldsymbol{c}_j)=\sum_{k=m+1}^{p}\delta(x_{ik}, c_{jk}) \tag{3}$$

式中:若$x_{ik}=c_{jk}$,则$\delta=1$,否则为0;c_{jk}为第j类原型在分类属性上的众数。式(1)中$\gamma(0\leqslant\gamma\leqslant1)$用于平衡数值型和分类型属性的重要性。当$\gamma=0.5$时,两者的贡献相等;$\gamma>0.5$表示数值型属性更关键。$\gamma$可通过数据驱动方法自动计算或根据领域知识设定,其损失函数为

$$L=\sum_{k=1}^{K}\sum_{x_i\in C_k}D(\boldsymbol{X}_i, \boldsymbol{c}_k) \tag{4}$$

式中K为聚类个数。

1.2　模型融合

模型融合是一种通过集成多个基模型的预测结果或中间特征,构建更鲁棒、更准确复合模型的技术。其核心思想源于“多样性互补”原则:不同模型对数据的假设和捕捉模式的能力存在差异,融合这些差异化的信息可以降低单一模型的偏差和方差[24],提升整体泛化性能。

关于构建和集成子模型的标准,模型融合技术可以分为以下3种类型。(1)异构算法集成[25]:基于权重系数的动态耦合,通过引入不同算法或同算法不同超参配置的模型,构建多样化子模型池,并基于优化算法动态分配权重,实现预测结果的加权融合。(2)多阶段任务分解[26]:分步建模与级联融合,通过将复杂任务拆解为逻辑关联的子阶段,每个阶段构建专用子模型,通过级联式信息传递实现全局优化。(3)相似性驱动集成[27]:局部模型动态选择,基于新样本与历史数据的相似性度量,动态选择适配子模型并分配权重,实现局部最优预测。

2　基于数据相似性度量和基模型权重优化的模型融合算法

为了提高预测性能,本节将介绍一种通过考虑模型性能和多样性以及样本匹配度来集成子模型的新方法——CEDWF(Clustering-enhanced dynamic weighted fusion)。可靠性因子[28]是由基模型的性能和多样性决定,仅描述了子模型的性能,但当一个新的观测点出现时,还需要知道这个新点属于哪个聚类。在大多数情况下,不能期望数据有良好的标签,并且新点可以明确地分配到其中一个聚类。因此,引入一个新的系数匹配度来描述新点与某个聚类的匹配程度。通过考虑数据分布和基模型的适配度两个方面出发,选择最优的基模型组合。为了在融合过程中确定不同模型之间的关系与相似性,引入了相似性度量函数[29]。该函数基于评估结果的相似性来衡量各个模型之间的关联性。具体而言,通过计算模型输出的预测结果与实际标签之间的差异,结合多维度评估函数中的多个评分指标来度量模型之间的相似程度。通过这一过程,可以有效识别出在某些维度上表现相似或互补的模型,更加准确地为基模型提供可靠性计算。

最后,采用基于性能-多样性基模型筛选策略对各个基学习器的预测结果进行加权融合。加权系数的确定依据是各个基学习器的综合表现,特别是其在多维度评估函数中的得分。通过对不同评估维度的综合考量,可以使得表现较为优秀的基学习器在最终的融合结果中占据更高的权重,从而提高整个融合模型的性能。

2.1　基于聚类的数据相似性度量

在异构数据情况下,通过聚类的方式实现样本相似度的计算。具体地,采用深度嵌入的方法,将离散数据和连续数值数据映射到统一的低维连续向量空间,通过非线性变换捕捉特征间的关联,其本质可表示为

$$\boldsymbol{z}=f_\theta(\boldsymbol{X}_{\text{num}}, \boldsymbol{X}_{\text{cat}}) \tag{5}$$

式中：$\boldsymbol{X}_{\text{num}}\in\boldsymbol{R}^{m}$表示代表数值型特征，$\boldsymbol{X}_{\text{num}}$代表数值型特征向量，$\boldsymbol{R}^{m}$代表数值型特征空间；$\boldsymbol{X}_{\text{cat}}\in\boldsymbol{Z}^{p}$表示分类型特征，$\boldsymbol{X}_{\text{cat}}$代表分类型特征向量，$\boldsymbol{Z}^{p}$代表分类型特征空间；$f_{\theta}$代表深度神经网络；$\boldsymbol{z}\in\boldsymbol{R}^{d}$代表联合嵌入向量，其中$d\ll m+p$表示嵌入维度。

下面分别对数值特征和类别特征进行处理。其中数值型特征使用多层感知机处理，通过非线性激活函数LeakyReLU提取高阶特征交互，即

$$\boldsymbol{z}_{\text{num}}=\text{ReLU}(\boldsymbol{W}_{2}\cdot\text{ReLU}(\boldsymbol{W}_{1}\boldsymbol{X}_{\text{num}}+\boldsymbol{b}_{1})+\boldsymbol{b}_{2}) \tag{6}$$

式中：$\boldsymbol{W}_{1}$代表输入层到第一隐藏层的权重矩阵，$\boldsymbol{b}_{1}$代表第一隐藏层的偏置向量；$\boldsymbol{W}_{2}$代表隐藏层到输出层的权重矩阵，$\boldsymbol{b}_{2}$代表输出层的偏置向量；$\boldsymbol{z}_{\text{num}}$为数值型特征经多层感知机与非线性激活函数处理后得到的高阶特征交互向量。

类别型特征处理，在嵌入层使用全连接映射有

$$\boldsymbol{z}_{\text{cat}}^{(j)}=\boldsymbol{E}_{j}(\boldsymbol{X}_{\text{cat}}^{(j)})\quad \boldsymbol{E}_{j}\in\boldsymbol{R}^{n_{j}\times d} \tag{7}$$

式中：$\boldsymbol{X}_{\text{cat}}^{(j)}$表示第$j$个类别特征的离散标签；$\boldsymbol{E}_{j}$表示嵌入矩阵；$\boldsymbol{R}^{n_{j}\times d}$表示将离散样本映射为向量，其中$n_{j}$表示第$j$个类别特征的唯一值数量；$\boldsymbol{z}_{\text{cat}}^{(j)}$表示第$j$个类别特征的嵌入向量。

然后，使用拼接将特征拼接起来，有

$$\boldsymbol{z}=[\boldsymbol{z}_{\text{num}}\|\boldsymbol{z}_{\text{cat}}^{(1)}\|,\boldsymbol{z}_{\text{cat}}^{(2)},\cdots,\|\boldsymbol{z}_{\text{cat}}^{(p)}] \tag{8}$$

式中“‖”代表向量拼接操作符。

通过以上处理，深度嵌入通过将异构数据映射到统一的语义空间，解决了传统方法在混合属性处理中语义失配问题，实现语义对齐。然后使用K-Prototypes聚类对n个样本进行聚类计算，使用每个样本得到的联合嵌入向量计算距离，式(1)可以改进为

$$D(\boldsymbol{X}_{i},\boldsymbol{c}_{j})=D_{f_{\text{dis}}}(\boldsymbol{z}_{i},\boldsymbol{c}_{i}) \tag{9}$$

式中f_{dis}可以使用式(2)进行计算，求解最优的簇个数采用网格搜索法，最小化式(4)的损失，直至收敛。最后，可以把样本划分为$\boldsymbol{C}=\{\boldsymbol{C}_{1},\boldsymbol{C}_{2},\cdots,\boldsymbol{C}_{k}\}$，每个簇的中心表示为$\boldsymbol{\mu}=\{\boldsymbol{\mu}_{1},\boldsymbol{\mu}_{2},\cdots,\boldsymbol{\mu}_{k}\}$。

计算待预测样本$\boldsymbol{z}_{i}$与簇$\boldsymbol{C}_{j}$的相似性，确定其归属区域的隶属度，得到待测样本的属于簇$\boldsymbol{C}_{j}$的相似性系数，相似性系数计算函数为

$$s(\boldsymbol{z}_{i},\boldsymbol{C}_{j})=\frac{\left(\sum\limits_{h=1}^{d}(\boldsymbol{z}_{ih}-\boldsymbol{\mu}_{jh})^{2}\right)^{\frac{1}{2}}}{\sum\limits_{g=1}^{k}\left(\sum\limits_{h=1}^{d}(\boldsymbol{z}_{ih}-\boldsymbol{\mu}_{gh})^{2}\right)^{\frac{1}{2}}} \tag{10}$$

式中：$\boldsymbol{z}_{ih}$表示第i个样本的h维分量，$\boldsymbol{\mu}_{jh}$表示第j个聚类中心向量的h维分量。

根据相似性系数计算函数，总共n个样本，计算每个样本$\boldsymbol{z}_{i}$对每个簇$\{\boldsymbol{C}_{1},\boldsymbol{C}_{2},\cdots,\boldsymbol{C}_{k}\}$的相似系数$\{s(\boldsymbol{z}_{i},\boldsymbol{C}_{1}),s(\boldsymbol{z}_{i},\boldsymbol{C}_{2}),\cdots,s(\boldsymbol{z}_{i},\boldsymbol{C}_{k})\}$，其中样本$\boldsymbol{z}_{i}$对簇$\boldsymbol{C}_{k}$的相似性系数$s(\boldsymbol{z}_{i},\boldsymbol{C}_{k})$记为$\lambda_{ik}$，由此得到样本$\boldsymbol{z}_{i}$对于$k$个簇的相似性向量为$\boldsymbol{v}_{i}=(\lambda_{i1},\lambda_{i2},\cdots,\lambda_{ik})$，$n$个样本的相似性矩阵为$\boldsymbol{S}_{n\times k}$。

基于该数据相似性度量的计算过程如算法1所示。

算法1　改进K-prototypes聚类算法

输入：n个含有数值型和分类型特征的样本$\boldsymbol{X}=\{x_{1},x_{2},\cdots,x_{m},x_{m+1},\cdots,x_{m+p}\}$。

输出：相似性度量矩阵$\boldsymbol{S}$，簇心$\boldsymbol{\mu}$。

(1) 根据式(5～7)计算拼接得到特征对齐的向量$\boldsymbol{Z}$。

(2) 根据式(9)利用网格搜索法求解式(4)，得到$\{\boldsymbol{C}_{1},\boldsymbol{C}_{2},\cdots,\boldsymbol{C}_{k}\}$和$\boldsymbol{\mu}=\{\boldsymbol{\mu}_{1},\boldsymbol{\mu}_{2},\cdots,\boldsymbol{\mu}_{k}\}$。

(3) 利用式(10)计算得到相似性向量,进一步拼接求解得到$\boldsymbol{S}$。

(4) 输出相似性度量矩阵$\boldsymbol{S}$和簇心$\boldsymbol{\mu}$。

2.2　基于性能-多样性的基模型筛选优化算法

在模型融合中,基模型的选择及其权重分配直接影响融合效果。本节提出性能-多样性双目标优化的基模型筛选方法(Performance-diversity dual-goal optimization based base model selection, PDG-BMS),以便解决冗余模型剔除、权重合理分配和可扩展性保障。

(1)确定初始基模型集合$M=\{M_1, M_2, \cdots, M_N\}$,其中$M_i$表示基模型集合中的第$i$个模型,总共包含$N$个基模型。然后计算模型$M$分别在样本集合$U$的原始性能得分$\boldsymbol{\xi}=\{\xi_1, \xi_2, \cdots, \xi_N\}$,其中$\xi_i$表示基模型集合中的第$i$个模型在$U$的综合性能损失,其计算函数为

$$\mathrm{MSE}(M_i)=\frac{1}{n}\sum_{j=1}^{n}(y_j-\hat{y}_j)^2 \tag{11}$$

$$\mathrm{MAE}(M_i)=\frac{1}{n}\sum_{j=1}^{n}|y_j-\hat{y}_j| \tag{12}$$

$$\xi_i=\frac{1}{3}(\mathrm{MSE}(M_i)+\mathrm{MAE}(M_i)+(1-R^2(M_i))) \tag{13}$$

式中:y_j为第j个样本的真实值,$\hat{y}_j$为模型M_i的预测值;R^2代表模型在训练集合上的拟合程度。

(2)进行基模型的多样性度量,计算协方差矩阵$\boldsymbol{B}$,通过计算协方差矩阵的预测值差异,可衡量模型间的多样性,b_i为模型M_i与集合中其他所有模型的"平均差异度"(模型级多样性度量),用于量化单个模型对整个集合多样性的贡献,有

$$b_{ij}=\mathrm{Cov}(M_i, M_j) \tag{14}$$

$$b_i=\frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}b_{ij} \tag{15}$$

(3)通过预测分解计算多样性得分,有

$$\boldsymbol{H}=\eta_{\max}(\boldsymbol{B})-\eta_{\min}(\boldsymbol{B}) \tag{16}$$

式中:$\boldsymbol{H}$为模型集合的多样性得分,$\eta_{\max}(\boldsymbol{B})$为协方差矩阵$\boldsymbol{B}$的最大特征值,$\eta_{\min}(\boldsymbol{B})$为协方差矩阵$\boldsymbol{B}$的最小特征值。

(4)根据模型的初始性能和基模型多样性评估结果进行基模型筛选,目标函数为

$$w_i=\beta(t)\cdot\phi_i+(1-\beta(t))\cdot b_i \tag{17}$$

式中$\beta(t)$为t时刻的性能权重系数,取值范围$\beta(t)\in[0, 1]$,其核心作用是动态平衡性能与多样性的优先级,而非固定权重。

(5)通过动态权重机制通过指数衰减函数调整权重分配,根据式(16)判断当前集合的多样性是否足够,从而调整迭代进度t。从"探索多样性"到"探索性能",将性能权重从"0"逐渐调整到"1",实现"多样性优先"到"性能优先"的平滑过渡,避免固定权重的缺陷,参数可解释性增强,方便根据不同的任务需求调整,即

$$\beta(t)=1-\exp\left(-\beta\frac{t}{T}\right) \tag{18}$$

式中T为迭代总次数。

最后,根据式(16~17)进行目标函数优化,得到最优的基模型权重向量$\boldsymbol{w}=\{w_1, w_2, \cdots, w_n\}$。基于性能-多样性双目标优化过程如算法2所示。

算法2 基于性能-多样性基模型筛选优化算法

输入:基模型集合M和验证集V。

输出:基模型的权重向量$\boldsymbol{w}$。

(1) 根据每个基模型在集合V上的真实值和预测值,根据式(11～13)计算得到向量$\boldsymbol{\xi}$。

(2) 根据式(14～16)计算出多样分得分$\boldsymbol{H}$。

(3) 根据指数衰减函数不断调整权重β,根据式(17～18),直至收敛,得到最佳平衡权重因子β_{best}。

(4) 根据β_{best}通过式(17)计算得到基模型权重向量$\boldsymbol{w}$。

2.3 CEDWF算法框架

综上所述,融合模型通过第一阶段的数据相似性度量模块可以得到当前样本x的相似性向量$\boldsymbol{v}$,第二阶段基于性能-多样性的基模型筛选优化算法得到当前聚类中的最优基模型权重向量$\boldsymbol{w}$。

(1)根据前样本x的相似性向量$\boldsymbol{v}$,计算基模型在每个聚类中的“局部表现得分”。对于每个聚类$\boldsymbol{C}_k$,计算基模型M_j在该聚类内的局部适配度$\boldsymbol{a}_j$,有

$$\boldsymbol{a}_j(x)=\sum_{i=1}^{k}\boldsymbol{v}_i(x)\cdot(1-\boldsymbol{\xi}_{j,i}) \tag{19}$$

式中$\boldsymbol{\xi}_{j,i}$根据式(11～13)计算得到,为模型M_j在$\boldsymbol{C}_i$上的综合性能损失。

(2)权重融合。将基模型M_j的可靠度权重$\boldsymbol{w}_j$与当前样本的局部适配度$\boldsymbol{a}_j(x)$融合,得到最终融合权重$\boldsymbol{\gamma}_j$,有

$$\boldsymbol{\gamma}_j=\boldsymbol{w}_j^d\cdot\boldsymbol{a}_j^e \tag{20}$$

式中:$d>0$代表基模型的可靠度权重系数,重视模型整体性能;$e>0$代表局部适配度权重系数,重视样本特异性任务。

综上,本文提出了一种基于数据相似性度量和基模型权重优化的自适应模型融合算法CEDWF,算法的框架图如图1所示。CEDWF对输入数据先通过改进聚类算法进行数据相似性分析,划分数据相

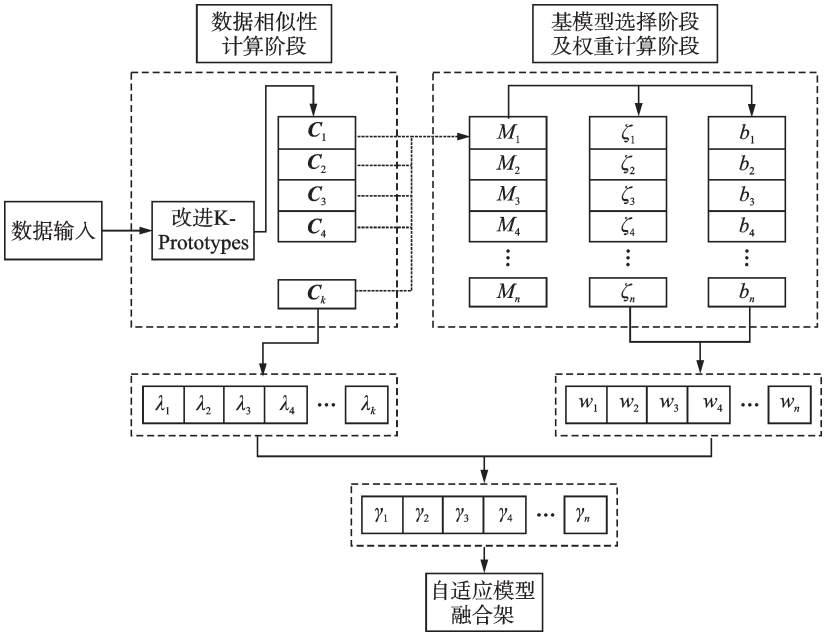


图1 CEDWF模型流程图

Fig.1 Flowchart of CEDWF model

似性高的域以及每个样本的相似性向量,并通过基模型的异构性和可靠性分析来动态集成模型,从而解决数据漂移问题并提高模型的预测能力,步骤如下。

(1) 将样本传入到改进的聚类模型中进行分簇预测,并输出每个样本的相似性向量。

(2) 进入到基模型的预测阶段,选择多个基模型在聚类完成后的每个簇上进行预测,得到基模型在每个簇上的性能向量,再对基模型之间进行预测相似性度量,剔除冗余模型,并调整基模型的性能向量,综合以上得到最终基模型在每个簇上的的可靠度向量。

(3) 根据数据相似性向量和基模型的可靠度向量,进行融合权重计算,得到最终的融合模型框架。

3　实验分析

3.1　实验数据集

本文采用UCI[30]里面的公开数据集BSD[31]和EED[32]全面地评估及验证提出的方法。数据集的详细信息如表1所示。BSD数据集总共包含17 379个样本,其中包含5个分类型特征:季节、月份、星期几、是否为休息日和天气状况类别;5个数值型特征:小时、标准化温度、标准化体感温度、标准化湿度和标准化风速。EED数据集总共包含768个样本,其中分类型特征有2个:建筑物取向和玻璃面积分布;数值型特征包含6个:相对紧凑度、表面积、墙面面积、屋顶面积、总高度和玻璃区。

表1　实验中使用的数据集详细信息

Table 1　Information of datasets used in experiments

数据集	样本数	分类型特征	数值型特征
BSD	17 379	5	5
EED	768	2	6

3.2　对比方法

将本文提出的方法与以下4类方法进行回归预测性能的比较实验:

(1) 单模型最优法:SOM[33]选择验证集上性能最优的单一模型,验证融合方法是否优于个体模型。

(2) 简单平均法:MA[34]选择在验证集上训练所有的基模型,对所有基模型预测结果取算术平均。

(3) 静态性能加权:SPW[35]根据基模型在验证集上的性能,根据准确率固定权重,对比动态权重与静态权重的优劣。

(4) Stacking集成法[7]:训练元模型学习基模型输出的组合关系,对比基于学习的融合策略。

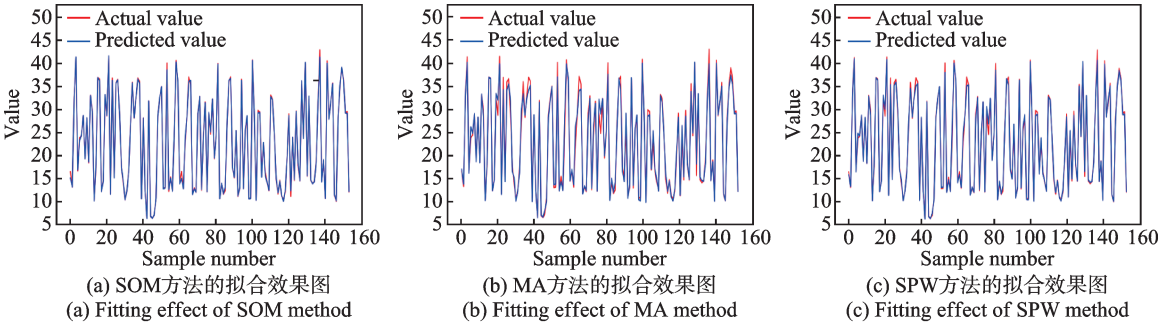
3.3　实验设置及评价指标

本文使用Windows10服务器进行实验,其配置为4核Intel Core i7-7700主频3.60 GHz处理器,RAM为32 GB;本次实验选择在Python语言编程环境以及PyTorch[36]实验环境下进行。选择的基模型有Random Forest[37]、XGBoost[9]、Linear Models[38]、SVR[39]、MLP[40]和Stacking[7]。对所有基模型进行网格搜索,确保公平比较。特征预处理统一,标准化都使用StandardScaler,分类型特征使用统一编码。

为了评估不同回归预测方法的性能,划分训练集和测试集为4:1。采用最广泛的4个评价指标评价最终的预测性能,包括精确度评估MSE、模型对数据的拟合能力R^2、平均绝对百分比误差MAPE以及相对误差MAE。

3.4　在数据集EED上的性能评估

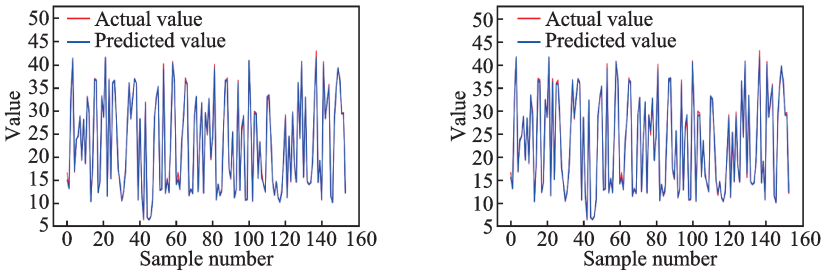
为了证明本文所提出的CEDWF算法预测模型的性能,根据第2节中提出的方法和第3.3节提出的评估标准将其与SOM、MA、SPW、Stacking典型的回归方法在公开数据集EED上进行了比较。这5种算法的拟合性能如图2所示。预测和实际值分别用蓝线和红线标记。显然,本文所提出的CEDWF融合模型优于其他4种方法模型。图2表明,基于CEDWF的融合模型的预测结果在大多数情况下能够准确拟合实际值。



(a) SOM方法的拟合效果图
(a) Fitting effect of SOM method

(b) MA方法的拟合效果图
(b) Fitting effect of MA method

(c) SPW方法的拟合效果图
(c) Fitting effect of SPW method



(d) Stacking方法的拟合效果图
(d) Fitting effect of Stacking method

(e) CEDWF方法的拟合效果图
(e) Fitting effect of CEDWF method

图2 基于相同测试集EED的5个模型的预测结果

Fig.2 Predicted results of five models based on the same test set EED

表2展示了5种方法的性能指标综合评估结果。由表2结果发现,CEDWF算法表现最佳,它的MSE值为0.162 5,较其他4种算法分别降低了25.46%、90.27%、66.69%和8.34%。CEDWF的R^2达0.998 4,高于其他基线模型,证明其能完整解释数据非线性特征。CEDWF的MAE为0.260 4,较Stacking(0.281 6)降低7.5%,较MA(0.888 3)降低70.7%,体现其对异常值的强鲁棒性。CEDWF的MAPE为1.249 5%,优于Stacking(1.262 9%)和SOM(1.456 9%),在需比例误差控制的场景(如资源调度)中实用性更优。

表2 不同方法下在EED数据集上的回归预测结果

Table 2 Regression prediction results of different methods on the EED dataset

方法	MSE	R^2	MAE	MAPE/%
SOM	0.218 0	0.997 9	0.337 3	1.456 9
MA	1.670 0	0.984 0	0.888 3	3.927 5
SPW	0.487 8	0.995 3	0.482 3	2.109 3
Stacking	0.177 3	0.998 3	0.281 6	1.262 9
CEDWF	**0.162 5**	**0.998 4**	**0.260 4**	**1.249 5**

3.5 在数据集BSD上的性能评估

5种方法在BSD数据集上的预测结果如图3所示,性能指标如表3所示。根据图3结果来看,SOM、Stacking和CEDWF三种模型拟合程度相差不是很大。根据数据集BSD本身的特点具有显著的线性关系以及表3指标可以发现:CEDWF较其他模型的精度有所提升。CEDWF的均方误差(3 864.39)较次优线性模型SOM(3 921.67)降低1.46%,证明其通过动态权重调整抑制了线性模型的过拟

合风险。CEDWF的R^2(0.878 0)高于SOM(0.876 2),表明其在线性主成分捕捉能力上更精准,尤其在高信噪比数据中能更完整地保留有效特征。CEDWF的MAE指标(41.17)较Stacking(41.83)提升1.6%,说明其对异常值的敏感性更低,该模型更具鲁棒性。CEDWF的MAPE(45.15%)较SOM(54.48%)提升17.1%,体现了CEDWF能够实现局部优化。

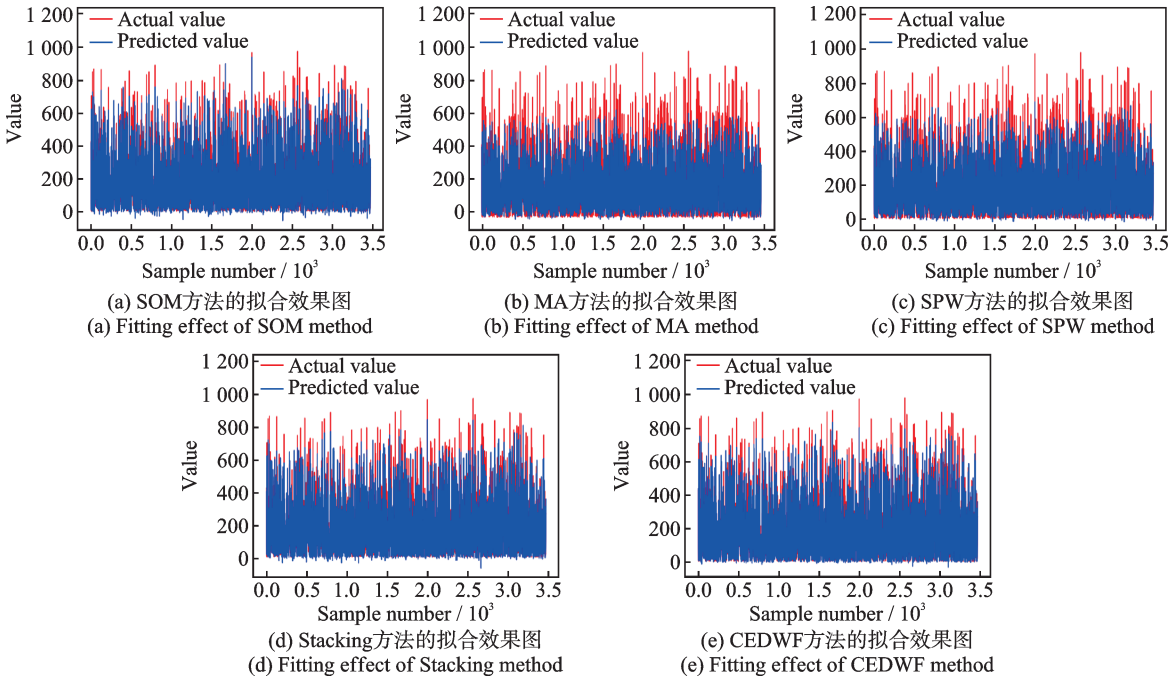


图3　基于相同测试集BSD的5个模型的预测结果

Fig.3　Predicted results of five models based on the same test set BSD

表3　不同方法下在BSD数据集上的回归预测结果

Table 3　Regression prediction results of different methods on the BSD dataset

方法	MSE	R^2	MAE	MAPE/%
SOM	3 921.673 3	0.876 2	42.340 2	54.482 0
MA	5 893.463 0	0.813 9	50.428 1	98.756 3
SPW	4 898.031 0	0.845 3	46.297 4	80.127 3
Stacking	3 957.750 0	0.875 0	41.833 8	50.288 3
CEDWF	**3 864.391 7**	**0.878 0**	**41.173 7**	**45.149 7**

3.6　实验结果

根据以上两组实验数据的系统对比,从预测精度和模型鲁棒性两个维度总结CEDWF的核心优势。在预测精度方面,3.4节实验中CEDWF的MSE(0.162 5)较次优模型Stacking(0.177 3)降低8.3%,较最差模型MA(1.670 0)降低90.3%,证明其通过动态残差加权机制有效抑制了非线性误差累积。在3.5节实验中,CEDWF的MSE(3 864.39)较传统线性模型SOM(3 921.67)降低1.46%,证明其弹性正则化策略在抑制过拟合的同时保留主成分信息。在鲁棒性方面,无论是在线性数据还是非线性数据中,根据预测评价指标值发现预测数据的适应性较好。

3.7 消融实验

为了验证基于数据相似性和模型可靠度驱动的自适应模型融合方法的有效性，本节分别设置3组对照实验：第1组Case 1使用数据相似性度量的策略进行回归预测；第2组Case 2使用基模型权重优化的方法进行回归预测；第3组Case 3使用CEDWF算法进行回归预测。为了保证实验的公正性，采用3.3节的实验设置和评价指标，使用EED[28]这个公开数据集进行验证，实验结果如图4所示。其次，为了评估不同策略对模型性能的影响，设置了不同的对比实验。不同实验模型的性能结果如表4所示。从表4可见，CEDWF方法获得了最好的回归预测性能。

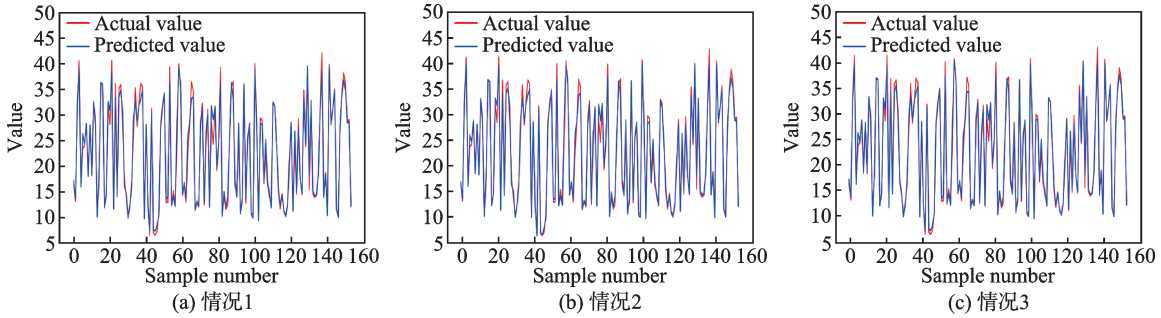


图4 基于相同测试集EED的消融实验预测结果

Fig.4 Ablation study results on same test set EED

表4 消融实验与完整模型性能对比

Table 4 Ablation study and performance comparison with the full model

Experiment	MSE	R^2	MAE	MAPE
Case 1	1.962 5	0.981 1	0.973 1	4.520 3
Case 2	1.670 0	0.983 9	0.888 3	3.927 4
Case 3	**1.384 9**	**0.986 7**	**0.821 9**	**3.839 0**

4 结束语

本文针对传统模型融合中数据-模型动态适配不足的核心问题，提出一种基于数据相似性度量和基模型权重优化的模型融合方法。该方法通过数据相似性度量与模型可靠度量化的双重优化机制，实现数据-模型的高效协同。实验结果和性能评估分析证明，本方法可以实现模型预测的性能。未来，研究将着重解决在面对高维度且数据分布复杂的数据集时，算法的计算复杂度将有所增加的问题，探索更高效的权重优化策略或模型融合规则，以降低在高维复杂数据上的计算成本，提高训练速度。

参 考 文 献：

[1] CHEN J, YE H, YING Z, et al. Dynamic trend fusion module for traffic flow prediction[J]. Applied Soft Computing, 2025, 174: 112979.

[2] WAN F, HUANG X, CAI D, et al. Knowledge fusion of large language models[EB/OL]. (2024-01-19). https://doi.org/10.48550/arXiv.2401.10491.

[3] LIU L, YU Y, WU Y, et al. Method for multi-task learning fusion network traffic classification to address small sample labels [J]. Scientific Reports, 2024, 14(1): 2518.

[4] DONG X, YU Z, CAO W, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14: 241-258.

[5] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24: 123-140.
[6] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.
[7] SHARMA N, DUTTA M. Designing a dynamic weighted stacking recommendation system[J]. International Research Journal of Multidisciplinary Scope, 2024, 5(4): 755-767.
[8] PICCIANO A G. Blending with purpose: The multimodal model[J]. Journal of Asynchronous Learning Networks, 2009, 13(1): 7-18.
[9] LIU Z, JIANG P, ZHANG L, et al. A combined forecasting model for time series: Application to short-term wind speed forecasting[J]. Applied Energy, 2020, 259: 114137.
[10] KE G, MENG Q, FINLEY T, et al. LightGBM: A highly efficient gradient boosting decision tree[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Red Hook, NY, USA: Curran Associates Inc., 2017: 3149-3157.
[11] MEHRAJ S, BANDAY M T. A dynamic weighted averaging technique for trust assessment in cloud computing[J]. International Journal of Cloud Applications and Computing (IJCAC), 2022, 12(1): 1-21.
[12] LI C, DING S, ZOU N, et al. Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling[J]. Journal of Biomedical Informatics, 2023, 143: 104399.
[13] ACHAKZAI M A K, PENG J. Detecting financial statement fraud using dynamic ensemble machine learning[J]. International Review of Financial Analysis, 2023, 89: 102827.
[14] CRUZ R M O, SABOURIN R, CAVALCANTI G D C. Analyzing dynamic ensemble selection techniques using dissimilarity analysis[C]//Proceedings of Artificial Neural Networks in Pattern Recognition: 6th IAPR International Workshop, ANNPR 2014. Montreal, QC, Canada: Springer International Publishing, 2014: 59-70.
[15] BRITTO JR A S, SABOURIN R, OLIVEIRA L E S. Dynamic selection of classifiers—A comprehensive review[J]. Pattern Recognition, 2014, 47(11): 3665-3680.
[16] VRIESMANN L M, BRITTO A S, OLIVEIRA L S, et al. Combining overall and local class accuracies in an oracle-based method for dynamic ensemble selection[C]//Proceedings of 2015 International Joint Conference on Neural Networks (IJCNN). [S.l.]: IEEE, 2015: 1-7.
[17] LI W, PENG Y, ZHANG M, et al. Deep model fusion: A survey[EB/OL]. (2023-09-27). https://doi. org/10.48550/arXiv.2309.15698.
[18] LIN P, ZHANG L, ZUO J. Data-driven prediction of building energy consumption using an adaptive multi-model fusion approach[J]. Applied Soft Computing, 2022, 129: 109616.
[19] GUO Y, ZHENG J, SHANG X, et al. A similarity regression fusion model for integrating multi-omics data to identify cancer subtypes[J]. Genes, 2018, 9(7): 314.
[20] 孙若凡.基于模型融合的国际短期天然铀价格预测研究[J].世界核地质科学,2024, 41(4): 712-719.
SUN Ruofan. Research on international short-term natural uranium price prediction based on model fusion[J]. World Nuclear Geoscience, 2024, 41(4): 712-719.
[21] HUANG Z. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304.
[22] AHMAD A, DEY L. A K-mean clustering algorithm for mixed numeric and categorical data[J]. Data & Knowledge Engineering, 2007, 63(2): 503-527.
[23] CHATURVEDI A, GREEN P E, CAROLL J D. K-Modes clustering[J]. Journal of Classification, 2001, 18: 35-55.
[24] WU Y, LIU L, XIE Z, et al. Promoting high diversity ensemble learning with ensemble bench[C]//Proceedings of 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI). [S.l.]: IEEE, 2020: 208-217.
[25] ZHANG Y, HONGLE D U. Imbalanced heterogeneous data ensemble classification based on HVDM-KNN[J]. CAAI Transactions on Intelligent Systems, 2019, 14(4): 733-742.
[26] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of International Conference on Machine Learning.[S.l.]: PMLR, 2017: 1126-1135.

[27] WANG B, LI Z, XU Z, et al. Casformer: Information popularity prediction with adaptive cascade sampling and graph transformer in social networks[J]. IEEE Transactions on Big Data, 2025, 11(4): 1652-1663.

[28] CARON M, MISRA I, MAIRAL J, et al. Unsupervised learning of visual features by contrasting cluster assignments[J]. Advances in Neural Information Processing Systems, 2020, 33: 9912-9924.

[29] REN Y, PU J, CUI C, et al. Dynamic weighted graph fusion for deep multi-view clustering[C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. California, USA: [s.n.], 2024: 4842-4850.

[30] DUA D, GRAFF C. UCI machine learning repository[EB/OL]. (1998-05-06). http://archive.ics.uci.edu/ml.

[31] DUA D, GRAFF C. UCI machine learning repository[EB/OL]. (2013-12-19). http://archive.ics.uci.edu/ml.

[32] DUA D, GRAFF C. UCI machine learning repository[EB/OL]. (2012-11-19). http://archive. ics. uci. edu/dataset/242/energy+efficiency.

[33] SCHWENKER F. Ensemble methods: Foundations and algorithms[J]. IEEE Computational Intelligence Magazine, 2013, 8(1): 77-79.

[34] CADE B S. Model averaging and muddled multimodel inferences[J]. Ecology, 2015, 96(9): 2370-2382.

[35] SHAYEGH B, LEE H H B, ZHU X, et al. Error diversity matters: An error-resistant ensemble method for unsupervised dependency parsing[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2025: 25119-25127.

[36] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An imperative style, high performance deep learning library[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2019: 8026-8037.

[37] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[38] CHAMBERS J M. Linear models[M]. Boca Raton: Chapman & Hall/CRC, 2017: 95-144.

[39] UKIL A. Support vector machine[M]. Berlin, Heidelberg: Springer, 2007: 161-226.

[40] SHAHREZA H O, HAHN V K, MARCEL S. MLP-hash: Protecting face templates via hashing of randomized multi-layer perceptron[C]//Proceedings of 2023 31st European Signal Processing Conference (EUSIPCO). [S.l.]: IEEE, 2023: 605-609.

作者简介:

王梅(1976-),**通信作者**,女,教授,研究方向:机器学习、模型选择和核方法,E-mail: wangmei@nepu.edu.cn。

李艳培(2000-),女,硕士研究生,研究方向:机器学习。

高雅田(1979-),女,副教授,研究方向:大数据和人工智能。

(编辑:刘彦东)