智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于深度模糊知识蒸馏的多变量时间序列预测模型

蒋云良, 余梅丽, 金森洋, 申情, 张雄涛

引用本文:
蒋云良, 余梅丽, 金森洋, 等. 基于深度模糊知识蒸馏的多变量时间序列预测模型[J]. 智能系统学报, 2026, 21(3): 639-650.
JIANG Yunliang, YU Meili, JIN Senyang, et al. Multivariate time series forecasting model based on deep fuzzy knowledge distillation[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(3): 639-650.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202508031

DOI: 10.11992/tis.202508031
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20260305.1645.004

基于深度模糊知识蒸馏的多变量时间序列预测模型

蒋云良[1,2,3], 余梅丽[1,2], 金森洋[1,2], 申情[1,2], 张雄涛[1,2]
（1. 湖州师范大学 信息工程学院, 浙江 湖州 313000; 2. 浙江省全省智能教育技术与应用重点实验室, 浙江 金华, 321004; 3. 浙江师范大学 计算机科学与技术学院, 浙江 金华 321004）

摘　要: 多变量时间序列在交通流量、气象监测等领域广泛存在, 其特征间存在复杂的时空依赖关系和高度不确定性, 传统机器学习模型难以有效捕获潜在模式。尽管近年来的深度学习方法在预测精度上取得了显著提升, 但往往依赖于庞大的网络结构与高计算开销, 限制了在实时或资源受限场景中的应用。为解决以上问题, 提出了一种新的用于时序预测的轻量级深度模糊知识蒸馏模型 (Takagi-Sugeno-Kang with deep fuzzy knowledge distillation, TSK-DFKD)。具有强大表达能力的教师模型将深度暗知识迁移到轻量级的学生模型, 从而降低预测成本。在学生模型中, 首次利用具有不确定知识处理能力的模糊推理网络, 有效应对时序数据的不确定性。在蒸馏过程中, 引入了教师有界损失来替代传统的交叉熵损失, 以实现知识蒸馏下的高效时序数据预测。在 5 个公开数据集上进行了实验, 与 9 个最先进的基线模型相比, 本文所提方法预测性能更优, 效率更高。

关键词: 注意力机制; 多变量时间序列; 知识蒸馏; 时序预测; TSK 模糊系统; 教师有界损失; 深度学习; 时间注意力; 空间注意力
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2026)03−0639−12

中文引用格式: 蒋云良，余梅丽，金森洋，等．基于深度模糊知识蒸馏的多变量时间序列预测模型 [J]．智能系统学报，2026, 21(3): 639–650.
英文引用格式: JIANG Yunliang, YU Meili, JIN Senyang, et al. Multivariate time series forecasting model based on deep fuzzy knowledge distillation[J]. CAAI transactions on intelligent systems, 2026, 21(3): 639–650.

Multivariate time series forecasting model based on deep fuzzy knowledge distillation

JIANG Yunliang[1,2,3], YU Meili[1,2], JIN Senyang[1,2], SHEN Qing[1,2], ZHANG Xiongtao[1,2]
(1. School of Information Engineering, Huzhou Normal University, Huzhou 313000, China; 2. Zhejiang Key Laboratory of Intelligent Education Technology and Application, Jinhua 321004, China; 3. School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China)

Abstract: Multivariate time series are common in areas such as traffic flow and weather monitoring. Their features show complex spatiotemporal dependencies and high uncertainty. Traditional machine learning models cannot effectively capture these hidden patterns. In recent years, deep learning methods have improved prediction accuracy, but they often rely on large network structures and high computational costs, which limit their use in real-time or resource-limited settings. To address these issues, a novel lightweight deep fuzzy knowledge distillation framework (TSK-DFKD, Takagi-Sugeno-Kang with deep fuzzy knowledge distillation) is proposed for time series forecasting. By transferring deep dark knowledge from a powerful representative teacher model to a lightweight student model, prediction costs can be reduced. The student model utilizes a fuzzy reasoning network, which has strong capabilities in handling uncertain knowledge, to effectively tackle uncertainties in time series data. During distillation, a teacher's bounded loss is introduced in place of traditional cross-entropy loss, enabling efficient knowledge distillation in time series data prediction. Experiments conducted on five open datasets demonstrate that TSK-DFKD outperforms nine state-of-the-art baselines in prediction performance and efficiency.

Keywords: attention mechanism; multivariate time series; knowledge distillation; time series forecasting; TSK fuzzy system; teacher-bounded loss; deep learning; temporal attention; spatial attention

时间序列预测是通过历史时序数据来识别模式或趋势, 并利用这些信息预测未来的表现[1]。时序数据具有高复杂性和不确定特性[2], 给准确预测带来了巨大挑战。首先, 时序数据往往表现

收稿日期: 2025−08−26.　**网络出版日期:** 2026−03−06.
基金项目: 国家自然科学基金项目 (62376094); 国家自然科学基金区域创新发展联合基金重点支持项目 (U22A20102); 浙江全省智能教育技术与应用重点实验室开放研究基金项目 (2025ZNJYKF003).
通信作者: 张雄涛. E-mail: 1047897965@qq.com.

出强烈的非平稳性，其统计特性 (如均值、方差) 会随时间剧烈变化，导致传统预测模型难以胜任；其次，时序数据通常存在多变量相关性，不同时间序列之间可能存在复杂的相互作用，而这些相互作用可能对预测结果产生重要影响。因此，与单变量时序预测不同，多变量时间序列预测需综合考虑时间序列内的时间相关性和多维变量间的相关性[3]。此外，时序数据还易受如噪声等外部因素的干扰，进一步增加了预测的不确定性[4-5]。针对以上挑战，本文首次利用模糊系统来解决时序数据预测中的不确定性问题。TSK(Takagi-Sugeno-Kang) 模糊系统是一种基于模糊逻辑的系统建模方法，具有良好的非线性逼近能力和适应性[6-7]，能有效处理现实世界中的不确定性问题[8-9]。利用“IF-THEN”的规则来定义模糊系统规则，每个模糊规则由前件和后件组成，前件描述输入变量的模糊条件，后件是一个线性函数，描述输出与输入变量的关系[10]。通过模糊逻辑和模糊推理机制，将复杂的时间序列数据转化为更容易理解和建模的模糊规则形式。因此，利用模糊系统进行时间序列预测具有得天独厚的优势。

随着深度学习技术的发展，越来越多的研究者采用深度网络模型方法解决时序数据复杂建模问题。尤其是 Transformer，凭借其强大的全局相关性建模能力在时间序列预测得到了广泛应用。为了进一步提高模型性能，研究者不断开发并设计了各种 Transformer 变体。如文献 [11] 通过采用动态学习图结构，模型能够适应时空关系的变化，从而增强模型动态建模能力。这些变体使模型具有了强大的特征提取和时空动态建模功能，但也使得网络架构变得越发复杂。深度堆叠的网络结构和复杂的模块化设计不仅增加模型训练和推理时间的成本，也使得其在实际场景中部署变得困难[12-13]。知识蒸馏技术为以上问题提供了一种有效解决途径，它通过将复杂大模型的知识迁移到轻量级模型中，从而降低模型复杂度。例如，在文献 [14] 中，使用知识蒸馏技术来进行目标识别，通过使用一个加权的交叉熵函数来减少噪声的干扰，引入温度参数来软化教师模型的输出，使学生能更好地学习到教师的知识；文献 [15] 采用生成对抗网络来实现不同模型架构之间的知识转移，将 CNN（convolutional neural network）的特征提取器作为生成器，通过判别器最大化教师和学生特征提取器之间的相似性，学生模型能生成与教师相似的特征图来提高模型的性能；文献 [16] 引入自蒸馏和相互学习提取特征并进行知识传递，自蒸馏通过让深层知识指导浅层知识，利用最小化 Kullback-Leibler[17] 散度实现两个网络之间相互学习。综上，现有基于知识蒸馏的模型框架缓解了模型复杂度的问题，但这些模型主要还是用于分类任务中。用于时序数据预测的回归任务中的知识蒸馏框架目前研究甚少。

因此，本文提出了一种全新的多变量时序预测框架 TSK-DFKD(Takagi-Sugeno-Kang with deep fuzzy knowledge distillation)。该框架由教师模型和学生模型组成，其中教师模型采用深度全局时序建模网络，通过在时间和空间维度分别应用注意力机制来捕获长期时空依赖；学生模型则为轻量级模糊推理网络，利用 TSK 模糊系统进行快速预测，实现对复杂时序数据的高效建模。该框架首次在多变量时序预测模型中将深度网络的特征提取能力与模糊系统的可解释性相结合，提出了一种结构上新颖且高效的解决方案。结合知识蒸馏技术与 TSK 模糊系统来提升模型性能。通过知识蒸馏将教师模型中深层时空依赖知识迁移至学生模型，显著增强轻量级学生模型对复杂动态关系的表征能力。在此基础上，本文引入教师有界损失替代传统交叉熵损失，以提升蒸馏过程的稳定性与效率。此外，首次将 TSK 模糊系统融入蒸馏框架，用以处理时序数据中的不确定性，从而在保证轻量化的同时进一步提高模型的预测精度和速度。在 5 个真实公开的数据集上进行了实验，以验证模型的有效性。结果表明，所提出的模型 TSK-DFKD 在推理速度和预测准确性等方面均超过了最新的时间序列预测模型。

1　问题定义

多变量时间序列预测广泛应用于交通、气象等领域，是利用历史观测数据推断多个相互关联变量的未来变化趋势，可以定义为：给定历史时间步长为 M_0 的时间序列 $\boldsymbol{X}_{m-M_0+1:m}$，训练一个模型 $\theta(\cdot)$ 来推断未来时间步长 M 的预测值。可以表示为

$$[\boldsymbol{X}_{m-M_0+1},\boldsymbol{X}_{m-M_0+2},\cdots,\boldsymbol{X}_m]\xrightarrow{\theta(\cdot)}[\boldsymbol{X}_{m+1},\boldsymbol{X}_{m+2},\cdots,\boldsymbol{X}_{m+M}] \tag{1}$$

式中：$\boldsymbol{X}_m\in\mathbf{R}^{N\times C}$，$N$ 为节点数量，C 为数据的特征数量。

然而，真实场景下的多变量时序数据通常具有高复杂性与不确定性。不同变量间存在复杂的时空依赖关系与非线性交互，导致模型在捕获长期动态特征时面临困难；此外，数据中存在噪声、缺失值，进一步增加了预测的不确定性。因此，多变量时间序列预测的核心挑战在于如何在复杂的时空相关性下有效建模，并在不确定性中实现

稳健的预测。

2　TSK-DFKD 模型

所提出的深度模糊知识蒸馏模型 TSK-DFKD，如图 1 所示，由 3 部分组成：数据嵌入层、教师模型θ_T和学生模型θ_S。在数据嵌入层中使用空间嵌入矩阵$\boldsymbol{S}$、时间嵌入矩阵$\boldsymbol{T}_D$和$\boldsymbol{T}_W$、原始映射矩阵$\boldsymbol{F}$，将嵌入的数据融合再作为教师模型和学生模型的输入。然后，全域时序建模网络通过时间注意力和空间注意力机制来学习时间依赖性和空间相关性。最后，模糊推理模型根据教师模型的知识来进行时间预测。

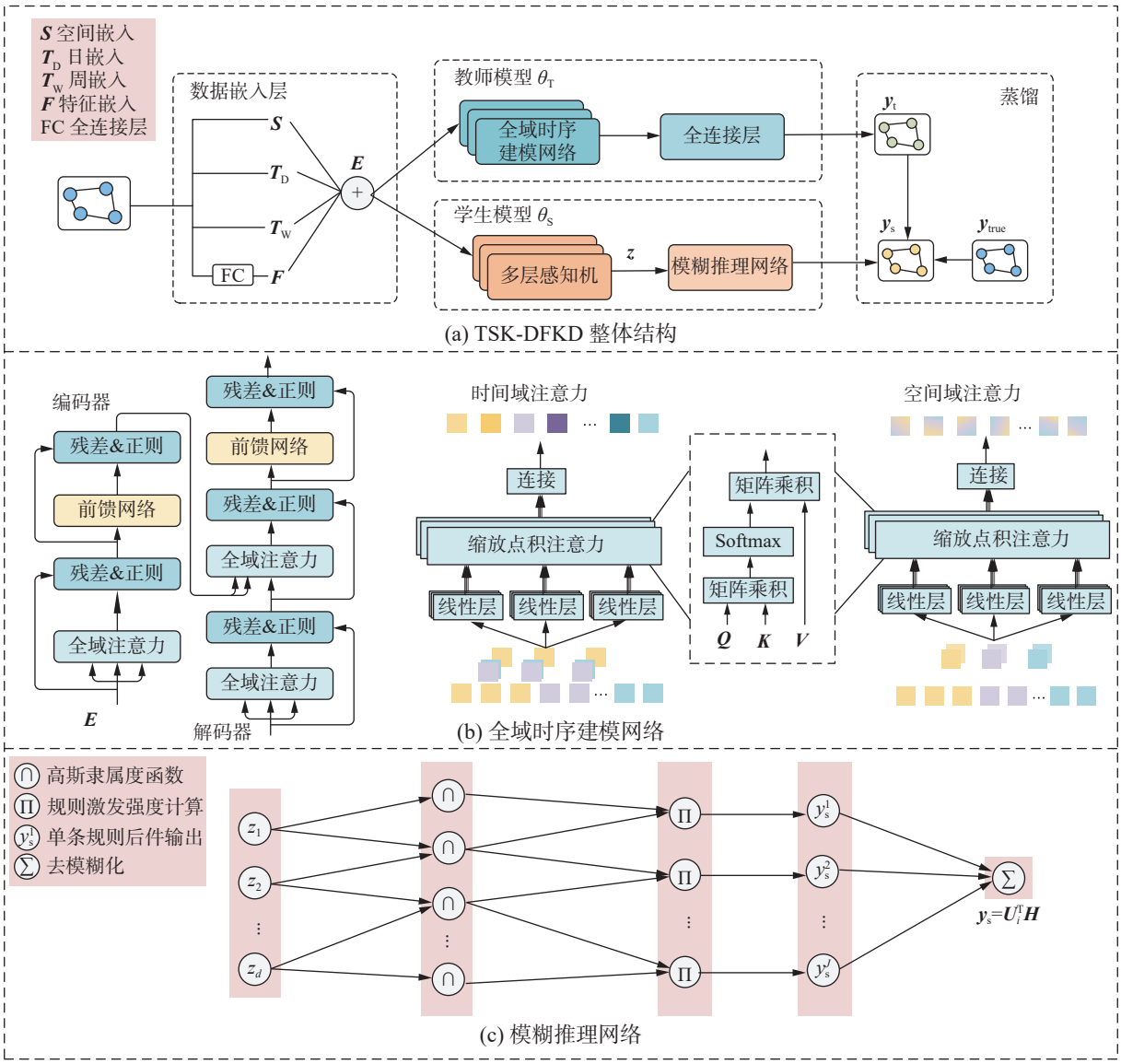


图 1　TSK-DFKD 模型框架

Fig. 1　Framework of TSK-DFKD

2.1　数据嵌入层

在时序预测中，预测结果的准确性会受到多种因素的影响，包括空间特征、一天的时间步以及一周的有效天数。本文使用嵌入矩阵来对原始的数据进行处理。如图 1(a) 所示，原始输入由 3 部分组成：空间信息 (如 GPS 传感器)、日周期的短期信息以及周周期的长期信息。对此使用一个空间嵌入矩阵$\boldsymbol{S}\in\mathbf{R}^{M_0\times N\times f}$，和两个时间嵌入矩阵$\boldsymbol{T}_D\in\mathbf{R}^{M_0\times N_D\times f}$、$\boldsymbol{T}_W\in\mathbf{R}^{M_0\times N_W\times f}$对数据进行处理，$\boldsymbol{T}_D$为一天中的时间步，$\boldsymbol{T}_W$为一周的天数。利用空间嵌入矩阵将原始时间序列数据中的空间信息转换为空间向量，同时通过两个时间嵌入矩阵分别提取日周期性和周周期性特征，生成对应的时间向量。并且通过使用全连接层来保留原始时间序列的隐藏特征$\boldsymbol{F}\in\mathbf{R}^{M_0\times N\times f}$。

$$\boldsymbol{F}=\mathrm{FC}(\boldsymbol{X}_{m-M_0+1:m}) \tag{2}$$

式中：f为特征嵌入的数量，FC(·)表示全连接层的映射函数。随后，将这些嵌入连接在一起，形成一个包含空间和时间信息的综合特征表示$\boldsymbol{E}\in\mathbf{R}^{M_0\times N\times d}$：

$$\boldsymbol{E}=\boldsymbol{S}\oplus\boldsymbol{T}_{\mathrm{D}}\oplus\boldsymbol{T}_{\mathrm{W}}\oplus\boldsymbol{F} \tag{3}$$

式中：时空综合特征维度$d=4f$，“⊕”表示连接操作。

2.2　教师全域时序建模网络

全域时序建模网络θ_{T}通过注意力机制来捕获全局特征进行建模。使用数据嵌入层的输出$\boldsymbol{E}$作为输入，在时间和空间维度上应用编码器和解码器来捕捉序列中的上下文信息，每个编码器和解码器由多个核心模块构成，主要包括多头注意力机制、前馈网络、残差连接以及层归一化。

2.2.1　编码器

编码器通常采用多层结构，每一层均由两个核心组件构成：多头自注意力机制和前馈神经网络。多头自注意力机制使得模型能够同时在不同的表示子空间中并行地处理信息，在分析时序数据时，能够有效地识别上下文联系。对于输入序列$\boldsymbol{E}$，首先通过 3 个不同的线性变换 (分别对应查询 (Query)、键 (Key) 和值 (Value)) 进行处理，然后计算注意力得分并进行加权求和。

$$\boldsymbol{Q}=\boldsymbol{E}\boldsymbol{W}^{Q} \tag{4}$$

$$\boldsymbol{K}=\boldsymbol{E}\boldsymbol{W}^{K} \tag{5}$$

$$\boldsymbol{V}=\boldsymbol{E}\boldsymbol{W}^{V} \tag{6}$$

式中：$\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$、$\boldsymbol{W}^{V}$为权重矩阵，$\boldsymbol{E}$为输入向量。

之后计算每个头的注意力分数，然后通过 Softmax 函数转换为权重，这些权重用于加权求和值向量。$\sqrt{d_K}$为键向量的维度。

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V} \tag{7}$$

再将所有注意力头的输出连接起来，通过另一个线性变换来整合不同头的信息。其中h是头的数量，$\boldsymbol{W}^{O}$是用于整合头信息的权重矩阵。

$$h_i=\mathrm{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i) \tag{8}$$

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(h_1,h_2,\cdots,h_h)\boldsymbol{W}^{O} \tag{9}$$

通过残差连接来缓解网络中的梯度消失问题[18]。经过残差连接后，输出会进行层归一化[19]操作，来确保每一层的输出都保持稳定，增强训练过程的稳定性和收敛速度。最后第一子层的输出为$\boldsymbol{R}_{\boldsymbol{C}_1}=\mathrm{LayerNorm}(\boldsymbol{x}+\mathrm{MultiHead}(\boldsymbol{x}))$，其中，$\boldsymbol{x}$为多头注意力的输出。在每个编码器层中，第 2 个子层是前馈神经网络，它对自注意力层的输出进行进一步的处理。这个网络可以表示为

$$\mathrm{FFN}(\boldsymbol{R}_{\boldsymbol{C}_1})=\mathrm{Softmax}(\boldsymbol{R}_{\boldsymbol{C}_1}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{10}$$

式中：$\boldsymbol{R}_{\boldsymbol{C}_1}$为第一个子层的输出，$\boldsymbol{W}_1$为权重和偏置。最终编码器的输出为$\boldsymbol{E}^{\mathrm{en}}=\mathrm{LayerNorm}(\boldsymbol{R}_{\boldsymbol{C}_1}+\mathrm{FFN}(\boldsymbol{R}_{\boldsymbol{C}_1}))$。

2.2.2　解码器

解码器与编码器的结构相似，都包含两个子层：多头自注意力和前馈网络。与编码器相同，解码器的每个子层也包含残差连接和层归一化操作。但解码器在多头注意力中有两个不同之处：解码器中有一个遮掩机制，用于保证解码器每个时间步的预测仅依赖于当前以及之前的时间步，防止信息泄露。

$$\mathrm{mask\_att}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}+M_{\mathrm{mask}}\right)\boldsymbol{V} \tag{11}$$

解码器中的第 2 个注意力子层将编码器的输出$\boldsymbol{E}^{\mathrm{en}}$与解码器的当前输入结合。解码器的$\boldsymbol{Q}$来自前一层的输出，而$\boldsymbol{K}$和$\boldsymbol{V}$则来自编码器的输出。解码器能够利用编码器提供的源序列的表征，帮助生成与源序列相关的目标序列。解码器的第 2 个子层是前馈网络。与编码器类似，前馈神经网络对解码器的输入进行进一步的非线性变换，最后解码器输出标记为$\boldsymbol{E}^{\mathrm{de}}$。通过回归线性层得到最终教师模型的输出$y_t=\mathrm{FC}_{\mathrm{regression}}(\boldsymbol{E}^{\mathrm{de}})$。为了评估预测的准确性并指导模型的学习过程，目标函数$\mathcal{L}_t$为

$$\mathcal{L}_t=\mathcal{L}_{\mathrm{sL1}}(y_t,y_{\mathrm{true}})=\begin{cases}\dfrac{1}{2}(y_t-y_{\mathrm{true}})^2, & |y_t-y_{\mathrm{true}}|<1\\ |y_t-y_{\mathrm{true}}|-\dfrac{1}{2}, & \text{其他}\end{cases} \tag{12}$$

2.3　学生模糊推理网络

学生模型θ_{S}建立在 TSK 模糊推理网络上。结合深度学习的特征提取能力与模糊逻辑的推理优势，来实现高效且精准的时空信息建模。通过这种结合，模型能够在处理复杂时空数据时提供更强的表达能力和更高的预测精度，同时保留了模糊系统的可解释性和深度学习的自动学习能力。与教师模型相比，学生模型的体积更小，运行速度更快。学生模型将数据嵌入层的输出$\boldsymbol{E}$作为输入，通过多层感知机 (multilayer perceptron, MLP) 进行特征提取和转换。第l层的 MLP 层可以表示为

$$\boldsymbol{E}^{l+1}=\mathrm{FC}_2^{l}(\varphi(\mathrm{FC}_1^{l}(\boldsymbol{E}^{l})))+\boldsymbol{E}^{l} \tag{13}$$

其中φ为激活函数。$\boldsymbol{z}=\boldsymbol{E}^{L}$作为模糊规则的输入向量，$L$为最后一层 MLP 层，第$j$条规则可以用以下一般形式表示[20-21]：

$$\begin{aligned}&\text{IF } z_1 \text{ is } A_1^j \text{ and } z_2 \text{ is } A_2^j \text{ and } \cdots z_d \text{ is } A_d^j,\\ &\text{THEN } y^j(\boldsymbol{z})=h_0^j+h_1^j z_1+\cdots+h_d^j z_d \quad j=1,2,\cdots,J\end{aligned} \tag{14}$$

式中：J表示模糊规则总数，$\boldsymbol{z}=(z_1,z_2,\cdots,z_d)^{\mathrm{T}}$是前件部分的$d$维输入向量，$A_i^j(i=1,2,\cdots,d)$为模糊集合。$y^j(\boldsymbol{z})$表示第$j$条规则的输出。$\boldsymbol{h}^j=(h_0^j,h_1^j,\cdots,h_d^j)^{\mathrm{T}}$为对应规则的后件参数。

首先，计算模糊规则隶属度$\mu_j(\boldsymbol{z})$以及核宽σ_i^j计算公式为

$$\mu_{A_i^j}(z_i)=\exp\left[-\frac{(z_i-\psi_i^j)^2}{2\sigma_i^j}\right] \tag{15}$$

$$\sigma_i^j=\frac{\sum_{g=1}^{N}u_g^j(z_{ig}-\psi_i^j)}{\sum_{g=1}^{N}u_g^j} \tag{16}$$

式中：输入z_i是当前的数据点，ψ_i^j为 FCM(fuzzy C-means) 聚类[22] 的中心，σ_i^j为核宽。当有多条规则时，系统通过乘积计算它们的激活值，假设有j条规则：

$$\mu_j(\boldsymbol{z})=\prod_{i=1}^{d}\mu_{A_i^j}(z_i) \tag{17}$$

为了对不同规则的输出进行加权，需要对激活值进行归一化。归一化后的激活值$\hat{\mu}_j(\boldsymbol{z})$计算公式为

$$\hat{\mu}_j(\boldsymbol{z})=\frac{\mu_j(\boldsymbol{z})}{\sum_{j=1}^{J}\mu_j(\boldsymbol{z})} \tag{18}$$

最终输出通过对所有规则输出进行加权求和得到：

$$\boldsymbol{y}_{\mathrm{s}}=\sum_{j=1}^{J}\hat{\mu}_j(\boldsymbol{z})\cdot(h_0^j+h_1^jz_1+h_2^jz_2\cdots+h_d^jz_d) \tag{19}$$

令：

$$\boldsymbol{h}^j=(h_0^j,h_1^j,\cdots,h_d^j)^{\mathrm{T}} \tag{20}$$

$$\boldsymbol{H}=\left((\boldsymbol{h}^1)^{\mathrm{T}},(\boldsymbol{h}^2)^{\mathrm{T}},\cdots,(\boldsymbol{h}^J)^{\mathrm{T}}\right)^{\mathrm{T}} \tag{21}$$

$$\boldsymbol{u}_i^j=\hat{\mu}_j(\boldsymbol{z})(1,\boldsymbol{z}^{\mathrm{T}})^{\mathrm{T}} \tag{22}$$

$$\boldsymbol{U}_i=\left((\boldsymbol{u}_i^1)^{\mathrm{T}},(\boldsymbol{u}_i^2)^{\mathrm{T}},\cdots,(\boldsymbol{u}_i^J)^{\mathrm{T}}\right)^{\mathrm{T}} \tag{23}$$

式 (19) 可以转换成线性回归问题来求解：

$$\boldsymbol{y}_{\mathrm{s}}=\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{H} \tag{24}$$

在离散的分类问题中，常采用交叉熵损失[23] 作为损失函数。但在回归问题中，往往无法直接判断教师模型的哪些信息对学生模型有效。教师模型的回归输出可能会给学生模型提供错误的指导，甚至可能会提供与真实值不一致的回归值。因此本文将教师有界损失作为蒸馏损失。教师有界损失避免了直接将教师的回归输出作为目标，而是将其作为学生模型学习的上限。学生模型的回归值应尽可能接近真实值，但当学生模型的表现优于教师模型时，则不再进一步增加额外的惩罚项$\mathcal{L}_{\mathrm{b}}$。只有当学生模型的误差大于教师模型的误差时，教师有界损失会惩罚学生模型。这意味着，学生模型的回归值应尽量接近或优于教师模型的回归输出。

$$\mathcal{L}_{\mathrm{b}}(y_{\mathrm{s}},y_{\mathrm{t}},y_{\mathrm{true}})=\begin{cases}\mathcal{L}_{\mathrm{sL1}}(y_{\mathrm{s}},y_{\mathrm{true}}),\ \mathcal{L}_{\mathrm{sL1}}(y_{\mathrm{s}},y_{\mathrm{true}})+\\ \qquad m>\mathcal{L}_{\mathrm{sL1}}(y_{\mathrm{t}},y_{\mathrm{true}})\\ 0,\quad \text{其他}\end{cases} \tag{25}$$

$$\mathcal{L}_{\mathrm{KD}}=\mathcal{L}_{\mathrm{sL1}}(y_{\mathrm{s}},y_{\mathrm{t}})+\beta\mathcal{L}_{\mathrm{b}}(y_{\mathrm{s}},y_{\mathrm{t}},y_{\mathrm{true}}) \tag{26}$$

式中：y_{s}为学生网络θ_{S}的预测值，y_{t}为教师网络θ_{T}的预测值，y_{true}为真实标签。m为一个阈值，β为一个超参数。$\mathcal{L}_{\mathrm{sL1}}$为SmoothL1损失[24]，在回归任务中，当误差较小时采用L2损失的形式，这使得梯度在误差较小时不会消失，从而提高了数值稳定性；当误差较大时，采用L1损失的形式，这使得损失函数对异常值有较好的鲁棒性。在$\mathcal{L}_{\mathrm{b}}$中，可以根据需要选择其他损失函数。

在模糊推理网络中，后件参数是模型决策过程中的关键组成部分，总损失函数表示为

$$\mathcal{L}_{\mathrm{GD}}=\frac{\lambda}{2}\|\boldsymbol{H}\|_1+\mathcal{L}_{\mathrm{KD}} \tag{27}$$

式中：λ为正则化参数，$\mathcal{L}_{\mathrm{KD}}$为教师有界损失，$\boldsymbol{H}$为模糊规则的后件参数，采用L1范数能够保证模型训练的稳定性，并且稀疏的后件规则可以应对模型对训练数据的过拟合。

2.4　TSK-DFKD 算法

教师模型通过梯度下降[25] 来更新参数最小化损失$\mathcal{L}_t$，学生模型的损失函数$\mathcal{L}_{\mathrm{GD}}$中由于包含$\boldsymbol{H}$的L1范数不可微，不能直接使用梯度下降来更新参数。本文采用近端梯度下降法 (proximal gradient descent, PGD)[26] 来最小化带有L1正则化项的损失函数。PGD 的每一步更新：

$$\boldsymbol{H}^{(t+1)}=\mathrm{prox}_{\varepsilon h}(\boldsymbol{H}^{(t)}-\varepsilon\nabla_{\xi}(\boldsymbol{H}^{(t)})) \tag{28}$$

式中：ε为学习率，$\nabla_{\xi}(\boldsymbol{H})=\nabla_{\boldsymbol{H}}L_{\mathrm{KD}}(\boldsymbol{H})$为可导部分的梯度，$\boldsymbol{H}^{(t)}$为第$t$次迭代的参数向量，$\mathrm{prox}_{\varepsilon h}(\tau)$是对非光滑函数$h$的近端算子，定义为

$$\mathrm{prox}_{\varepsilon h}(\boldsymbol{\tau})=\arg\min_{x}\left(\frac{1}{2\varepsilon}\|\boldsymbol{x}-\boldsymbol{\tau}\|_2^2+h(\boldsymbol{x})\right) \tag{29}$$

当$h(\boldsymbol{H})=\frac{\lambda}{2}\|\boldsymbol{H}\|_1$时，近端映射是软阈值操作$\mathrm{prox}_{\varepsilon\frac{\lambda}{2}\|\cdot\|_1}(\boldsymbol{\tau})=\mathrm{SoftThresh}_{\varepsilon\frac{\lambda}{2}}(\boldsymbol{\tau})$，软阈值对向量$\boldsymbol{\tau}$的每个向量$\boldsymbol{\tau}_i$都会进行计算：

$$\mathrm{SoftThresh}_{\alpha}(\boldsymbol{\tau}_i)=\mathrm{sign}(\boldsymbol{\tau}_i)\max(|\boldsymbol{\tau}_i|-\alpha,0) \tag{30}$$

令$\alpha=\varepsilon\lambda/2$，后件参数$\boldsymbol{H}$的迭代公式为

$$\boldsymbol{H}^{(t+1)}=\mathrm{SoftThresh}_{\varepsilon\lambda/2}(\boldsymbol{H}^{(t)}-\varepsilon\nabla_{\boldsymbol{H}}L_{\mathrm{KD}}(\boldsymbol{H}^{(t)})) \tag{31}$$

整个学习过程的伪代码可以使用算法 1 描

述。其主要包含两个阶段：

教师训练阶段。首先通过数据嵌入层获得数据特征$\boldsymbol{E}$，将$\boldsymbol{E}$作为教师模型θ_{T}的输入，教师模型通过时间注意力和空间注意力进行建模，训练完成之后冻结参数。

学生蒸馏训练阶段。将数据嵌入矩阵$\boldsymbol{E}$分别作为教师模型θ_{T}和学生模型θ_{S}的输入，用训练完的教师模型计算得到教师输出。学生模型通过多层感知机将输入特征$\boldsymbol{E}$转换到更深层特征$\boldsymbol{z}$。在模糊规则中利用 FCM 聚类算法获得模糊前件参数，使用近端梯度下降法对后件参数$\boldsymbol{H}$进行求解。

算法 1　TSK-DFKD 模型训练

输入　历史时间序列$X_{m-M_0+1:m}$，未来时间序列$X_{m+1:m+M}$，正则化参数λ，学习率ε。

输出　训练完成的 TSK-DFKD 模型。

教师训练阶段：

1) 由历史序列$X_{m-M_0+1:m}$并根据式 (2) 和 (3) 计算得到数据嵌入表示$\boldsymbol{E}$；

2) 根据式 (4)~(6) 计算嵌入表示的注意力得分$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$；

3) 根据式 (7) 计算得到注意力头得分Attention，并通过式 (8) 和 (9) 加权整合多个注意力头的得分，得到MultiHead；

4) 通过式 (10) 中的残差连接与式 (11) 中的层归一化，得到编码器最终输出$\boldsymbol{E}^{\mathrm{en}}$；

5) 根据式 (12) 计算得到解码器输出y_{t}；

6) 使用 Adam 算法迭代优化式 (12) 中的目标函数得到训练完成的教师模型；

学生蒸馏训练阶段：

1) 固定教师模型的学习参数，正则化参数λ，初始化学习率ε；

2) 由历史序列$X_{m-M_0+1:m}$根据式 (2) 和 (3) 计算得到数据嵌入表示$\boldsymbol{E}$；

3) 通过已训练的教师模型得到教师输出y_{t}；

4) 根据式 (13) 计算得到模糊规则的输入向量$\boldsymbol{z}$；

5) 由 FCM 算法计算得到核中心，并使用式 (16) 计算核宽，得到前件参数；

6) 根据式 (17) 计算各个规则的隶属度并通过式 (18) 进行归一化；

7) 根据式 (19)~(24) 计算得到模糊系统的输出y_{s}；

8) 使用近端梯度下降法根据式 (28)~(31) 迭代优化式 (27) 得到训练完成的 TSK-DFKD 模型。

3　实验

3.1　数据集

为了评估模型的有效性，本文使用 5 个真实数据集进行实验 (数据集下载地址：https://drive.google.com/drive/folders/14EJVODCU48fGK0Fkye-Vom_9lETh80Yjp)，这些数据集涵盖了气候变化、电力消耗和交通流量等多个领域的时序数据，可以从不同角度对模型预测性能进行评估。详细信息如表 1 所示。

表 1　数据集信息

Table 1　Details of datasets

任务	数据集	节点	比例	频率/min
单步预测	Electricity	321	7∶1∶2	60
	Weather	21	7∶1∶2	10
多步预测	PEMS03	358	6∶2∶2	5
	PEMS04	307	6∶2∶2	5
	PEMS08	170	6∶2∶2	5

3.2　评价指标

本文使用 4 个评估指标评估时序预测的准确性。对于单步预测的数据集，本文使用平均绝对误差E_{MA}(mean absolute error, MAE) 和均方误差E_{MS} (mean-square error, MSE) 来衡量模型的预测性能。在多步预测实验上，使用平均绝对误差E_{MA} (MAE)、平均绝对百分比误差E_{MAP}(mean absolute percentage error, MAPE) 和均方根误差E_{RMS}(root mean square error, RMSE) 来衡量模型的预测性能。指数越小说明性能越好，4 个评估指标的表达式分别为

$$E_{\mathrm{MA}}=\frac{1}{n}\sum_{i=1}^{n}|y_i-f(x_i)| \tag{32}$$

$$E_{\mathrm{MS}}=\frac{1}{n}\sum_{i=1}^{n}(y_i-f(x_i))^2 \tag{33}$$

$$E_{\mathrm{MAP}}=\frac{100\%}{n}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-f(x_i)}{y_i}\right| \tag{34}$$

$$E_{\mathrm{RMS}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-f(x_i))^2} \tag{35}$$

式中：y_i为真实值，$f(x_i)$为函数预测值。

3.3　实验设置

为了评估模型的预测性能，本文将 TSK-DFKD 模型与当前先进的时间序列预测模型进行对比。本文选取了目前最新的基于 GCN（graph convolutional network）、Transformer 和 MLP 的时间序列预测模型。

DLinear[27]：基于线性模型的时间序列预测方法。它仅使用简单的单层线性模型就取得了与基于 Transformer 的模型相当的效果，同时大大降低

计算复杂度。

TiDE[28]: 一个基于多层感知器的编码器−解码器模型, 它结合了线性模型的简单性和速度, 同时能够处理协变量和非线性依赖关系。

TimesNet[29]: 基于卷积神经网络的时间序列预测模型。它通过设计的多尺度卷积神经网络捕捉时间序列不同时间尺度的变化模式。

Informer[30]: 通过一次前向传播生成完整预测序列, 采用自注意力机制和变分自编码器结构来捕捉长序列时间序列的复杂依赖关系。

FEDformer[31]: 利用频率域的稀疏表示来捕捉时间序列数据的全局趋势和季节性模式, 通过随机选择傅里叶分量实现多尺度特征提取。

PatchTST[32]: 基于分 Patch 策略的多变量时间序列预测方法, 将时间序列数据分割成 Patch 序列并输入 Transformer 中以捕获时间序列数据的长期依赖关系, 从而取得较好的预测性能。

Crossformer[33]: 针对多变量时间序列预测的 Transformer 模型, 通过维度分段嵌入和两阶段注意力层, 主动利用跨维度依赖关系, 建立层次化编码器−解码器结构, 以捕捉不同时间尺度的信息, 从而提高预测精度。

iTransformer[34]: 一种针对多变量时间序列预测的新型 Transformer 模型, 通过将时间序列的每个变量独立嵌入为变量子标记, 利用自注意力机制捕捉变量间的相关性, 并通过前馈网络学习每个变量的时间序列表示, 解决了传统 Transformer 在处理多变量时间序列时因时间步嵌入导致的变量间信息丢失和注意力图无意义的问题。

CNNBaTSK[21]: 是从 CNN 中蒸馏知识到模糊系统中的模型, 其中深度特征被用作前件的训练参数, 借助 FCM 聚类算法生成模糊规则的前件参数, 用原始数据训练模糊规则的后件参数。

为了确保实验的公平性, 所有模型均在相同的数据预处理条件下进行: 1) 5 个数据集都经过表 1 中的比例划分为训练集、验证集和测试集。2) 数据集都经过“Z-score”进行归一化处理。3) 对于单步预测数据集, 历史长度为 96, 预测长度为 $M \in \{96,192,336,720\}$; 对于多步预测数据集, 历史长度为 12, 预测长度为 12。

本文在带有两个 GeForce RTX 4090 显卡的服务器上使用 PyTorch 框架进行训练。并使用“Z-score”分别对 5 个数据集进行标准化, 来提高梯度下降过程中的收敛速度。此外, 使用“Adam”优化器进行端到端训练, 单步预测学习率为 0.000 1, 多步预测学习率为 0.001。批量大小设置为 32, 注意力头数量为 4。采用网格搜索+交叉验证的方式, 模糊规则从 {1, 2, 3, 4, 5, 8, 10, 15, 20} 中搜索。为防止过拟合, 在训练过程中引入了“Early Stopping”机制, 最大训练轮数设置为 100, 当验证集性能在连续 20 个 epoch 内未提升时提前终止训练。

3.4　实验结果

表 2 记录了单步预测任务的结果。历史长度 M_0 为 96, 预测长度 $M \in \{96,192,336,720\}$。与基准模型相比, TSK-DFKD 模型在数据集 Electricity 上优于模型 iTransformer, MSE 降低了 1.1%, MAE 降低了 1.4%。同样, 在数据集 Weather 上的均方误差和平均绝对误差相较于 9 个基准模型均有所降低。这表明引入模糊规则知识蒸馏不仅能够有效提升学生模型的泛化能力, 还能够在时序建模任务中提供更强的可解释性。

表 2　单步预测任务中对比模型的实验结果

Table 2　Experimental results of a comparison model in a single-step prediction task

模型	预测长度	Electricity		Weather	
		MSE	MAE	MSE	MAE
DLinear	96	0.210	0.302	0.195	0.252
	192	0.210	0.305	0.237	0.295
	336	0.223	0.319	0.282	0.331
	720	0.258	0.350	0.345	0.382
TiDE	96	0.200	0.271	0.202	0.261
	192	0.201	0.274	0.242	0.298
	336	0.214	0.289	0.287	0.335
	720	0.257	0.335	0.351	0.386
TimesNet	96	0.168	0.272	0.172	0.220
	192	0.184	0.322	0.219	0.261
	336	0.198	0.300	0.280	0.306
	720	0.220	0.320	0.365	0.359

续表 2

模型	预测长度	Electricity		Weather	
		MSE	MAE	MSE	MAE
Informer	96	0.274	0.368	0.300	0.384
	192	0.296	0.386	0.598	0.544
	336	0.300	0.394	0.578	0.523
	720	0.373	0.439	1.059	0.741
FEDformer	96	0.193	0.308	0.217	0.296
	192	0.201	0.315	0.276	0.336
	336	0.214	0.329	0.339	0.380
	720	0.246	0.355	0.403	0.428
PatchTST	96	0.174	0.259	0.177	0.218
	192	0.178	0.265	0.225	0.259
	336	0.196	0.282	0.278	0.297
	720	0.237	0.316	0.354	0.348
Crossformer	96	0.254	0.347	0.164	0.232
	192	0.261	0.353	0.211	0.276
	336	0.273	0.364	0.269	0.327
	720	0.303	0.388	0.355	0.404
CNNBaTSK	96	0.212	0.300	0.262	0.365
	192	0.275	0.351	0.274	0.373
	336	0.335	0.395	0.289	0.387
	720	0.448	0.460	0.342	0.420
iTransformer	96	0.148	0.239	0.174	0.214
	192	0.167	0.258	0.221	0.254
	336	0.178	0.271	0.278	0.296
	720	0.210	0.299	0.344	0.358
TSK-DFKD	96	**0.137**	**0.230**	**0.151**	**0.203**
	192	**0.150**	**0.244**	**0.204**	**0.250**
	336	**0.167**	**0.263**	**0.253**	**0.295**
	720	**0.208**	**0.297**	**0.339**	**0.355**

注：最佳结果用粗体突出显示，次佳结果用下划线显示。

多步预测任务实验结果如表 3 所示。可以得出以下结论，以 PEMS08 数据集为例：1) 与基于 GCN 的模型 TimesNet 的预测结果相比，TSK-DFKD 同样展现出了显著的优势。MAE、MAPE 和 RMSE 平均改进量分别为 4.02、2.17% 和 5.86。同时，与基于 MLP 的模型 DLinear 相比，TSK-DFKD 模型在 MAE、MAPE 和 RMSE 上的平均提升幅度更大，分别为 8.21、5.49% 和 12.13。相比较单独使用 GCN 或 MLP 的方法，TSK-DFKD 模型能够更好地建立全局相关性。2) 与能挖掘全局信息、局部信息和变量间相关性的模型 Crossformer 相比较时，TSK-DFKD 模型在 MAE、MAPE 和 RMSE 上平均改进量分别为 1.51、1.0% 和 1.56，这表明 TSK-DFKD 模型不仅能够有效地建模局部和全局相关性，还能将这些因素融合在一起，从而显著提高预测的准确性。3) 与模糊模型 CNNBaTSK 相比，TSK-DFKD 模型在 PEMS 数据集上 MAE、MAPE 和 RMSE 指标上均有明显提升。

表 3　多步预测任务中对比模型的实验结果

Table 3　Experimental results of comparing models in multi-step prediction tasks

模型	PEMS03			PEMS04			PEMS08		
	MAE	MAPE/%	RMSE	MAE	MAPE/%	RMSE	MAE	MAPE/%	RMSE
DLinear	21.36	21.32	34.53	27.98	19.11	43.91	22.43	14.64	35.42
TiDE	21.58	19.67	34.96	28.26	18.56	44.41	22.57	13.61	35.81

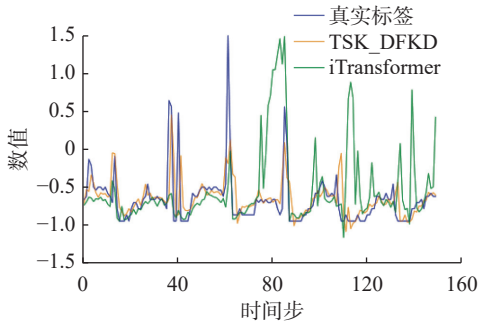
续表 3

模型	PEMS03			PEMS04			PEMS08		
	MAE	MAPE/%	RMSE	MAE	MAPE/%	RMSE	MAE	MAPE/%	RMSE
TimesNet	16.84	16.17	27.22	22.63	15.28	35.89	18.24	11.32	29.15
Informer	19.41	19.05	33.67	23.30	16.30	37.25	23.26	13.87	36.02
FEDformer	15.50	16.51	25.53	20.11	14.13	32.17	16.55	10.52	25.99
PatchTST	20.98	19.15	33.57	27.43	18.10	42.92	21.75	13.21	34.50
Crossformer	18.90	18.06	31.07	19.28	13.02	30.74	15.73	10.15	24.85
CNNBaTSK	21.21	25.21	35.28	24.86	18.40	38.54	22.20	14.37	32.98
iTransformer	18.36	17.34	29.47	23.16	15.33	38.05	16.61	10.74	28.23
TSK-DFKD	**15.21**	**16.16**	**25.20**	**18.47**	**12.37**	**30.05**	**14.22**	**9.15**	**23.29**

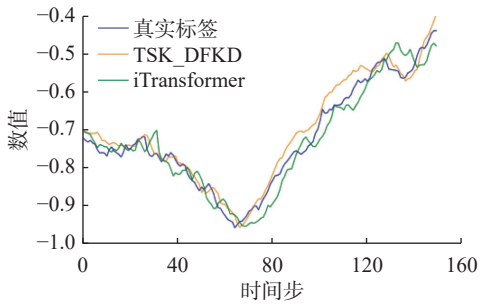
注：最佳结果用粗体突出显示，次佳结果用下划线显示。

3.5　可视化分析

为了直观地评估模型的预测效果，对模型 iTransformer 和 TSK-DFKD 进行了长期预测结果可视化对比。实验数据选取了 Weather 和 Electricity 两个长期预测数据集。图 2 给出了预测准确度。蓝色、橙色和绿色分别对应真实标签、TSK-DFKD 的预测和 iTransformer 的预测。从图 2 中可以看到 TSK-DFKD 模型的预测与实际值更为接近，并且对时间序列的波动具有更好的拟合能力，证明其具有较好的时序预测能力。相比之下，iTransformer 模型的预测表现出显著的滞后。从图 2(a) 中可以看出，当峰值发生变化时，iTransformerr 模型的预测存在明显的滞后，而 TSK-DFKD 模型的预测更接近真实值。



(a) 在数据集 Electricity 的预测与真实值对比

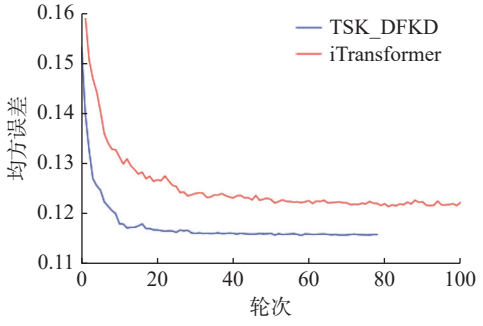


(b) 在数据集 Weather 的预测与真实值对比

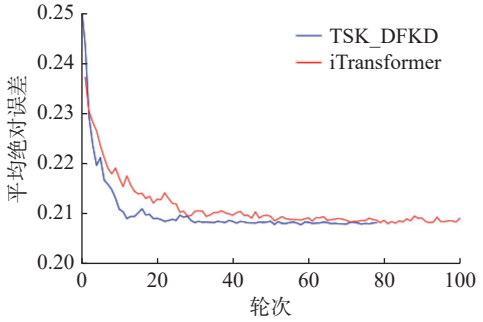
图 2　预测结果可视化

Fig. 2　Visualization of prediction results

以数据集 Electricity 的验证子集为例，使用实验结果的两个指标来比较 TSK-DFKD 模型和 iTransformer 的性能。在图 3(a) 中，iTransformer 在大约 20 轮后开始出现震荡，并逐渐收敛，在第 81 轮时达到最低值的均方误差。相比之下，TSK-DFKD 模型从一开始就快速收敛至较低水平，并在该水平附近波动，直至满足早停条件。在图 3(b) 中，iTransformer 的波动很大，在后期依旧震荡，相比之下，TSK-DFKD 模型的波动相对较小。这表明 TSK-DFKD 模型不仅在训练初期学习得更快，而且在训练过程中持续优化，最终实现了更优的预测性能。可以将这一成就归功于 TSK-DFKD 模型的高效学习能力和强大的推理能力，这使得它在处理时间序列数据时更为精确和可靠。



(a) 在 Electricity 验证集上模型的均方误差对比



(b) 在 Electricity 验证集上模型的平均绝对误差对比

图 3　训练过程中指标变化可视化

Fig. 3　Visualization of indicator changes during training

3.6　消融实验

为验证模型每个模块的有效性，本文在 Electricity 和 Weather 数据集上进行消融实验，并设置了 4 种变体模型：w/o Emb，删除数据嵌入；w/o Stu，采用学生模型和模糊系统来进行预测，检验蒸馏有效性；w/o Stu_TSK，删除学生模型中模糊系统部分，检验模糊系统有效性；w/o KD_TSK，删除蒸馏和模糊系统，仅包含 MLP 的学生模型，检验两个模块的整体贡献。结果如表 4 所示。从表 4 可以看出：与去掉蒸馏的学生模型相比，Electricity 数据集上 MSE 上升 0.4%，MAE 提升 0.6%；Weather 数据集上 MSE 提升 0.5%，MAE 提升 1.6%。与去掉模糊系统的学生模型相比，Electricity 数据集上 MSE 上升 0.5%，MAE 提升 0.5%；Weather 数据集上 MSE 提升 0.3%，MAE 提升 0.7%。与去掉蒸馏和模糊系统的学生模型相比，Electricity 数据集上 MSE 上升 2.7%，MAE 提升 1.5%；Weather 数据集上 MSE 提升 1.6%，MAE 提升 2.2%。在消融掉不同模块的时候，各数据集上的性能下降幅度存在差异，这表明蒸馏模块和模糊学习模块在不同数据集中的作用程度各异，从而验证了聚合设计的有效性。

表 4　消融实验结果

Table 4　Ablation experiments

方法	Electricity		Weather	
	MSE	MAE	MSE	MAE
w/o Emb	0.188	0.256	0.171	0.227
w/o Stu	0.140	0.236	0.155	0.220
w/o Stu_TSK	0.138	0.235	0.153	0.210
w/o KD_TSK	0.163	0.245	0.166	0.225
TSK_DFKD	**0.136**	**0.230**	**0.150**	**0.203**

注：最佳结果用粗体突出显示。

3.7　参数灵敏性分析

对模型中的超参数进行了实验，以确定最佳参数，修改了 5 个关键超参数：学习率ε、MLP 层数、蒸馏损失中超参数m、超参数β和模糊规则数量。实验结果如图 4 所示。从图 4(a) 中看到学习率从ε=0.01 降低到ε=0.000 1 时；从图 4(b) 中看到将 MLP 层数从 2 提高到 5 时；从图 4(c) 中看到m值从m=0.5 提高到m=3 时；从图 4(d) 中看到β值从β=0.1 提高到β=1 时；从图 4(e) 中看到模糊规则数量从 1 提高到 8 时，MAE 和 MAPE 的实验结果先降低后增加。这些结果表明，在 PEMS08 数据集上模型的最佳超参数是ε=0.001、3 层 MLP 层、m=2、β=0.8 以及模糊规则数量为 3。

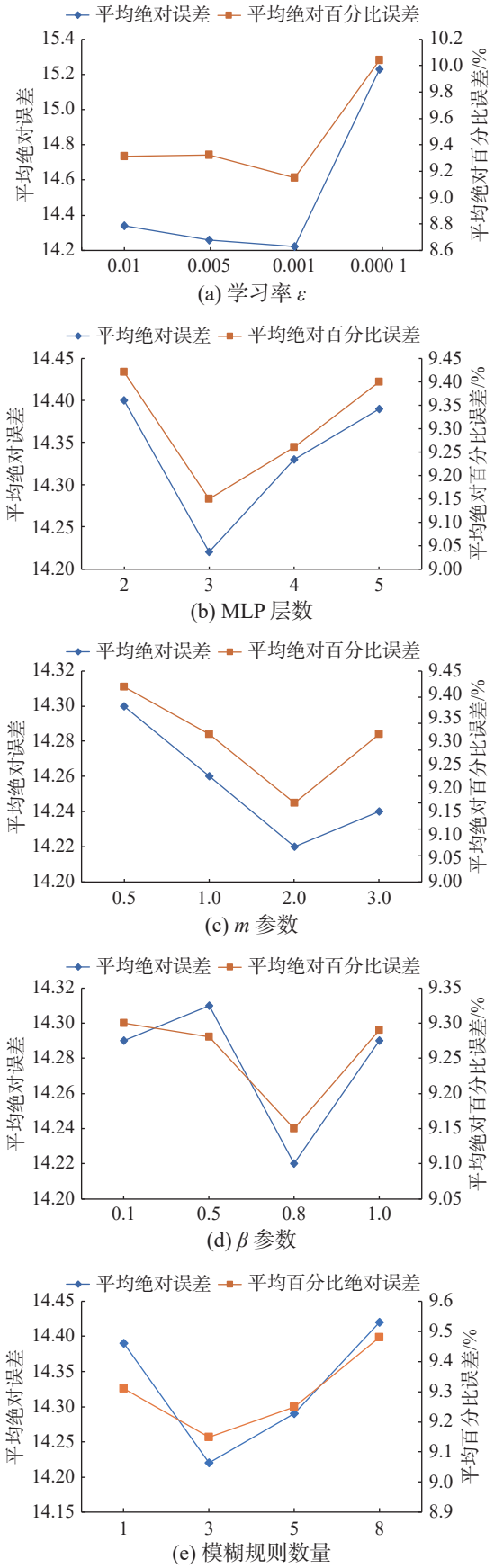


图 4　在 PEMS08 数据集上进行参数灵敏性分析对比

Fig. 4　Parametric sensitivity analysis on the PEMS08

3.8　效率分析

推理速度的快慢决定了模型的实用性，因此

最大限度缩短推理时间成为关键目标。与此同时, 边缘设备的部署场景还对模型提出了轻量化要求在本节中将推理时间和模型参数量作为核心对比指标来评估不同模型的效率, 为保证不同模型之间比较的公平性, 在推理测试中统一采用固定的批次大小 (batch size 为 32) 并以总处理时间作为推理速度。在 Weather 数据集上对各模型进行预测长度为 96 的平均效率作比较。图中气泡的大小代表模型参数数量。如图 5 所示, TSK_DFKD 在效率方面优于基于 Transformer 的模型, 也优于基于 MLP(TiDE) 和 GCN(TimesNet) 的模型, 展现出较强的轻量化和实际部署能力。

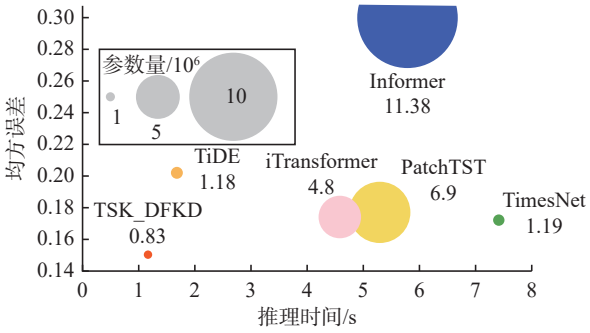


图 5　模型效率对比

Fig. 5　Comparison of models efficiency

4　结束语

为了更好地捕捉时空特征, 本文提出一种基于深度模糊知识蒸馏的多变量时间序列预测模型。通过将具有强大时空建模能力的教师模型的暗知识迁移到具有不确定性处理能力的轻量级 TSK 学生模型中, 能够快速实现对未来时间序列的精准预测, 同时显著降低模型的复杂度和提高运行效率。最后, 实验证明了所提方法的创新性和在多种任务场景下的优异性能表现。未来计划将进一步优化模型, 探索更先进的知识蒸馏技术, 并结合数据增强和特征工程方法提升模型性能。

参考文献:

[1] 陈梅, 柳博雅, 王钰, 等. 基于时间序列形态的模糊聚类算法[J]. 控制与决策, 2025, 40(4): 1116−1126.
CHEN Mei, LIU Boya, WANG Yu, et al. Fuzzy clustering algorithm based on time series morphology[J]. Control and decision, 2025, 40(4): 1116−1126.

[2] 胡磊, 韩敏. 基于核共轭梯度演化模糊系统的混沌时间序列在线预测[J]. 控制与决策, 2024, 39(9): 3099−3107.
HU Lei, HAN Min. Online prediction of chaotic time series based on kernel conjugate gradient evolving fuzzy system[J]. Control and decision, 2024, 39(9): 3099−3107.

[3] SHIH S Y, SUN Fankeng, LEE H Y. Temporal pattern attention for multivariate time series forecasting[J]. Machine learning, 2019, 108(8): 1421−1441.

[4] ZHANG Yuxin, CHEN Yiqiang, WANG Jindong, et al. Unsupervised deep anomaly detection for multi-sensor time-series signals[J]. IEEE transactions on knowledge and data engineering, 2023, 35(2): 2118−2132.

[5] STANKEVIČIŪTĖ K, ALAA A M, VAN DER SCHAAR M. Conformal time-series forecasting[C]//Neural Information Processing Systems. online: NeurIPS. 2021: 6216−6228.

[6] ZHOU Erhao, VONG C M, NOJIMA Y, et al. A fully interpretable first-order TSK fuzzy system and its training with negative entropic and rule-stability-based regularization[J]. IEEE transactions on fuzzy systems, 2023, 31(7): 2305−2319.

[7] BIAN Zekang, ZHANG Jin, NOJIMA Y, et al. Hybrid-ensemble-based interpretable TSK fuzzy classifier for imbalanced data[J]. Information fusion, 2023, 98: 101845.

[8] 张雄涛, 李水苗, 翁江玮, 等. 基于视角−规则的深度 TSK 模糊分类器及其在多元癫痫脑电信号识别中的应用[J]. 控制与决策, 2024, 39(4): 1315−1324.
ZHANG Xiongtao, LI Shuimiao, WENG Jiangwei, et al. Recognition of multivariate epilepsy EEG signals based on view-to-rule deep TSK fuzzy classifier[J]. Control and decision, 2024, 39(4): 1315−1324.

[9] 蒋云良, 印泽宗, 张雄涛, 等. 高阶 Takagi-Sugeno-Kang 模糊知识蒸馏分类器及其在脑电信号分类中的应用[J]. 智能系统学报, 2024, 19(6): 1419−1427.
JIANG Yunliang, YIN Zezong, ZHANG Xiongtao, et al. TSK fuzzy distillation classifier with negative Euclidean probability and High-order fuzzy dark knowledge transfer and its application on EEG signals classification[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1419−1427.

[10] 张雄涛, 陈天宇, 赵康, 等. 基于多教师自适应知识蒸馏的 TSK 模糊分类器[J]. 智能系统学报, 2025, 20(5): 1136−1147.
ZHANG Xiongtao, CHEN Tianyu, ZHAO Kang, et al. TSK fuzzy classifier based on multi-teacher adaptive knowledge distillation[J]. CAAI transactions on intelligent systems, 2025, 20(5): 1136−1147.

[11] JIANG Jiawei, HAN Chengkai, ZHAO W X, et al. PDFormer: propagation delay-aware dynamic long-range transformer for traffic flow prediction[J]. Proceedings of the AAAI conference on artificial intelligence, 2023, 37(4): 4365−4373.

[12] PALEYES A, URMA R G, LAWRENCE N D. Challenges in deploying machine learning: a survey of case studies[J]. ACM computing surveys, 2022, 55(6): 1−29.

[13] QIU Yaner, MA Liyun, PRIYADARSHI R. Deep learning challenges and prospects in wireless sensor network deployment[J]. Archives of computational methods in engineering, 2024, 31(6): 3231−3254.

[14] CHEN Guobin, CHOI W, YU Xiang, et al. Learning efficient object detection models with knowledge distillation[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM,

2017: 742−751.
[15] XU Qing, CHEN Zhenghua, WU Keyu, et al. KDnet-RUL: a knowledge distillation framework to compress deep neural networks for machine remaining useful life prediction[J]. IEEE transactions on industrial electronics, 2022, 69(2): 2022−2032.
[16] LI Ying, LI Ping, YAN Doudou, et al. Deep knowledge distillation: a self-mutual learning framework for traffic prediction[J]. Expert systems with applications, 2024, 252: 124138.
[17] YANG Xue, YANG Xiaojiang, YANG Jirui, et al. Learning high-precision bounding box for rotated object detection *via* kullback-leibler divergence[C]//Neural Information Processing Systems. online: NeurIPS. 2021: 18381−19394.
[18] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770−778.
[19] XU Jingjing, SUN Xu, ZHANG Zhiyuan, et al. Understanding and improving layer normalization[C]//Neural Information Processing Systems.Vancouver: NeurIPS. 2019: 32.
[20] ZHANG Yuanpeng, WANG Guanjin, HUANG Xiuyu, et al. TSK fuzzy system fusion at sensitivity-ensemble-level for imbalanced data classification[J]. Information fusion, 2023, 92: 350−362.
[21] JIANG Yunliang, WENG Jiangwei, ZHANG Xiongtao, et al. A CNN-based born-again TSK fuzzy classifier integrating soft label information and knowledge distillation[J]. IEEE transactions on fuzzy systems, 2023, 31(6): 1843−1854.
[22] DENG Zhaohong, CHOI K S, CHUNG F L, et al. Scalable TSK fuzzy modeling for very large datasets using minimal-enclosing-ball approximation[J]. IEEE transactions on fuzzy systems, 2011, 19(2): 210−226.
[23] BOTEV Z I, KROESE D P, RUBINSTEIN R Y, et al. The cross-entropy method for optimization[M]//Handbook of Statistics-Machine Learning: Theory and Applications. Amsterdam: Elsevier, 2013: 35−59.
[24] WU Wei, FAN Qinwei, ZURADA J M, et al. Batch gradient method with smoothing regularization for training of feedforward neural networks[J]. Neural networks, 2014, 50: 72−78.
[25] REYAD M, SARHAN A M, ARAFA M. A modified Adam algorithm for deep neural network optimization[J]. Neural computing and applications, 2023, 35(23): 17095−17112.
[26] MARDANI M, SUN Qingyun, VASAWANALA S, et al. Neural proximal gradient descent for compressive imaging[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 9596−9606.
[27] ZENG Ailing, CHEN Muxi, ZHANG Lei, et al. Are transformers effective for time series forecasting?[J]. Proceedings of the AAAI conference on artificial intelligence, 2023, 37(9): 11121−11128.
[28] DAS A, KONG Weihao, LEACH A, et al. Long-term forecasting with TiDE: time-series dense encoder[EB/OL]. (2023−04−17)[2025−01−01]. https://arxiv.org/abs/2304.08424.
[29] WU H, HU T, LIU Y, et al. Timesnet: Temporal 2d-variation modeling for general time series analysis[C]//Proceedings of the International Conference on Learning Representations. Kigali, 2023.
[30] ZHOU Haoyi, ZHANG Shanghang, PENG Jieqi, et al. Informer: beyond efficient transformer for long sequence time-series forecasting[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(12): 11106−11115.
[31] ZHOU Tian, MA Ziqing, WANG Xue, et al. Fedformer: frequency enhanced decomposed transformer for long-term series forecasting[M]//AI for Time Series. Boca Raton: CRC Press, 2026: 10−34.
[32] NIE Yuqi, NGUYEN N H, SINTHONG P, et al. A time series is worth 64 words: long-term forecasting with transformers[EB/OL]. (2022−11−27)[2025−01−01]. https://arxiv.org/abs/2211.14730.
[33] ZHANG Yunhao, YAN Junchi. Crossformer: transformer utilizing cross-dimension dependency for multivariate time series forecasting[C]//International Conference on Learning Representations, 2023.
[34] LIU Yong, HU Tengge, ZHANG Haoran, et al. iTransformer: inverted transformers are effective forecasting[C]//Proceedings of International Conference on Learning Representations.Vienna：ICLR, 2024.

作者简介：

蒋云良，教授，博士生导师，博士，主要研究方向为深度学习、智慧交通、智慧医疗和智能教育。先后主持和参与国家和省部级科研项目 13 项。发表学术论文 63 篇，出版学术著作 2 部，授权发明专利 26 项。E-mail: jyl@zjhu.edu.cn。

余梅丽，硕士研究生，主要研究方向为模糊系统、深度学习。E-mail: 1937365006@qq.com。

张雄涛，副教授，博士，主要研究方向为深度学习、模糊系统、智慧交通和智慧医疗。E-mail: 1047897965@qq.com。