智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于多教师自适应知识蒸馏的TSK模糊分类器

张雄涛, 陈天宇, 赵康, 李水苗, 申情

引用本文:
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.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202410028

DOI: 10.11992/tis.202410028
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250625.1928.005

基于多教师自适应知识蒸馏的 TSK 模糊分类器

张雄涛[1,2], 陈天宇[1,2], 赵康[1,2], 李水苗[2,3], 申情[1,2]
（1. 湖州师范学院 信息工程学院, 浙江 湖州 313000; 2. 浙江省现代农业资源智慧管理与应用研究重点实验室, 浙江 湖州 313000; 3. 湖州师范学院 信息技术中心, 浙江 湖州 313000）

摘　要: 目前层次型或深度模糊系统性能优异, 但是模型复杂度较高; 而基于蒸馏学习的轻量型 TSK(Takagi-Sugeno-Kang) 模糊分类器主要以单教师知识蒸馏为主, 若教师模型表现不佳, 则会影响蒸馏效果和模型的整体性能; 此外, 传统的多教师蒸馏通常使用无标签策略分配教师模型输出的权重, 容易使低质量教师误导学生。对此, 本文提出了一种基于多教师自适应知识蒸馏的 TSK 模糊分类器 (TSK fuzzy classifier based on multi-teacher adaptive knowledge distillation, TSK-MTAKD), 以多个具有不同神经表达能力的深度神经网络为教师模型, 利用本文提出的多教师知识蒸馏框架从多个深度学习模型中提取隐藏知识, 并传递给具有强大不确定处理能力的 TSK 模糊系统。同时设计自适应权重分配器, 将教师模型的输出与真实标签做交叉熵处理, 更接近真实值的输出将被赋予更高权重, 提高了模型的鲁棒性与隐藏知识的有效性。在 13 个 UCI 数据集上的实验结果充分验证了 TSK-MTAKD 的优势。

关键词: TSK 模糊分类器; 知识蒸馏; 多教师网络; 自适应权重分配; 隐藏知识; 模糊系统; 不同视角; 深度学习
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)05−1136−12

中文引用格式: 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.
英文引用格式: 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.

TSK fuzzy classifier based on multi-teacher adaptive knowledge distillation

ZHANG Xiongtao[1,2], CHEN Tianyu[1,2], ZHAO Kang[1,2], LI Shuimiao[2,3], SHEN Qing[1,2]
(1. School of Information Engineering, Huzhou University, Huzhou 313000, China; 2. Zhejiang Province Key Laboratory of Smart Management and Application of Modern Agricultural Resources, Huzhou 313000, China; 3. Information Technology Center, Huzhou University, Huzhou 313000, China)

Abstract: Currently, hierarchical and deep fuzzy systems demonstrate excellent performance, but they often suffer from high model complexity. Lightweight Takagi-Sugeno-Kang (TSK) fuzzy classifiers based on distillation learning typically rely on single-teacher knowledge distillation. However, if the teacher model underperforms, then the distillation effect and the overall model performance can be compromised. Furthermore, traditional multiteacher distillation approaches often assign weights to teacher model outputs using label-free strategies, which may allow low-quality teachers to mislead the student model. Aiming to address these issues, this paper introduces a TSK fuzzy classifier based on multiteacher adaptive knowledge distillation (TSK-MTAKD). The method employs multiple deep neural networks, each with different neural expression capabilities, as teacher models. The proposed distillation framework extracts dark knowledge from these models and transfers it to a TSK fuzzy system, leveraging its strong capability to handle uncertainty. Additionally, an adaptive weight allocator is introduced, which performs cross-entropy calculations between the output of the teacher model and the true label. Outputs that are closer to the true label are assigned higher weights, thereby improving model robustness and the quality of dark knowledge. Experimental results on 13 UCI benchmark datasets validate the advantages of the TSK-MTAKD approach.

Keywords: TSK fuzzy classifier; knowledge distillation; multiple teacher networks; adaptive allocation of weights; dark knowledge; fuzzy system; different perspectives; deep learning

深度学习[1]通过深层神经网络来模拟和学习数据中的复杂特征和表示, 以实现各种机器学习任务, 如图像识别、自然语言处理、语音识别、推

收稿日期: 2024−10−22.　**网络出版日期:** 2025−06−26.
基金项目: 国家自然科学基金项目 (62376094, U22A201856).
通信作者: 申情. E-mail: sq@zjhu.edu.cn.

荐系统等。它已经成为人工智能领域的主要技术, 并在解决复杂问题和处理大规模数据方面具有广泛的应用[2]。

除了深度学习外, 在机器学习中, 模糊系统是一种基于模糊逻辑理论的建模和控制方法, 旨在处理那些难以准确建模的复杂系统, 特别是存在不确定性和模糊性数据的系统。TSK(Takagi-Sugeno-Kang) 模糊系统[3] 是一种特定类型的模糊逻辑系统, 用于进行模糊推理和建模, 在模糊逻辑领域中被广泛应用, 其通常用于建立基于 IF-THEN 模糊规则的推理系统、控制系统, 以及模糊数据建模等。TSK 模糊系统的一个关键特点是它能够将模糊输入数据映射到确定性的输出, 这使得它在需要准确控制和建模的应用中非常有效。此外, 它通常通过模糊推理和模糊学习方法进行建模和优化, 以适应特定应用的需求。不仅如此, 其也因为直观的模糊规则, 从而拥有较高的可解释性。

除此以外, 知识蒸馏[4] 也是一种机器学习技术, 旨在通过将一个复杂模型 (通常称为“教师模型”) 的知识传递给一个较简单的模型 (通常称为“学生模型”), 来提高学生模型的性能。这个概念提出后便被广泛应用于深度学习领域。知识蒸馏的基本思想是, 通过教师模型的预测结果和中间表示 (通常是模型的软标签或特征表示) 来指导学生模型的训练。这有助于学生模型更好地学习复杂任务, 特别是在数据有限的情况下。

近年来, 随着知识蒸馏方法的提出, 研究者们也开始探索将 TSK 模糊逻辑系统与知识蒸馏相结合的创新方法。如 Jiang 等[5] 提出了一种基于卷积神经网络 (convolutional neural network, CNN) 的 CNNBaTSK(CNN-based born-again Takagi-Sugeno-Kang fuzzy classifier) 模型, 采用非迭代学习方法来求解模糊规则的后向参数, 为知识蒸馏方法提供了新的视角。Júnior 等[6] 通过知识蒸馏技术, 将复杂模型的知识传递给 TSK 模糊系统, 提高了模型的可解释性和性能。Gu 等[7] 提出了一种将深度神经网络 (deep neural network, DNN) 的知识传递给 TSK 模糊推理系统的方法, 其通过知识蒸馏技术, 将 DNN 的知识表达为模糊规则, 从而更容易解释特定的决策。Zhang 等[8] 提出了 HTSK-LLM-DKD(high-order TSK to low-order TSK fuzzy classifier with least learning machine based decoupling knowledge distillation), 其将高阶 TSK 模糊分类器作为教师模型, 低阶 TSK 模糊分类器作为学生模型, 利用解耦知识蒸馏将高阶 TSK 模糊分类器中的隐藏知识迁移到低阶 TSK 模糊分类器中。Erdem 等[9] 提出了一种基于深度学习的知识蒸馏方法, 将深度模型的泛化能力迁移到区间二型模糊逻辑系统中, 显著提高了模糊逻辑系统的学习性能。

目前知识蒸馏与 TSK 模糊逻辑系统相结合的研究大多采用单教师知识蒸馏的方法, 但是在教师模型表现不佳的情况下, 教师会误导学生, 从而影响模型的性能。除此以外, 现有的多教师知识蒸馏通常对教师的输出取平均或采用其他无标签策略[10-11], 这对于不同能力的教师模型来说显然是不合理的。为此, 本文提出了基于多教师自适应知识蒸馏的 TSK 模糊系统算法 (TSK fuzzy classifier based on multi-teacher adaptive knowledge distillation, TSK-MTAKD), 从多个教师模型中提取隐藏知识传递给学生模型, 即利用多个教师模型来提升 TSK 模糊逻辑系统的性能, 提高模型的鲁棒性。除此以外, 本文还设计了自适应权重分配器, 使教师模型面对不同的数据集能根据准确率自适应地分配输出权重。本文的主要贡献如下:

1) 本文将图卷积网络 (graph convolutional network, GCN)、CNN、Transformer 3 个神经网络作为教师模型, TSK 模糊逻辑系统作为学生模型, 利用多教师知识蒸馏, 从 3 个教师模型中提取隐藏知识传递给学生模型, 从而提升 TSK 模糊分类器的性能。与传统的单教师知识蒸馏模型相比, 本文的 TSK-MTAKD 在保证模型可解释性的同时, 提升了模型的性能, 增加了模型的鲁棒性。

2) 本文设计了一种自适应权重分配器, 计算多个教师模型的输出与真实标签之间的交叉熵损失, 使蒸馏模型能根据不同规模的数据集下教师模型的性能差距自适应地分配教师模型的权重, 与传统的多教师权重分配方法相比, 本文的自适应权重分配器能根据实际情况合理地分配权重, 从而更有利于隐藏知识的提取。

3) TSK-MTAKD 在 13 个 UCI 数据集上充分体现了其有效性, 在大多数数据集上都取得了最好的性能, 其性能优于高阶 TSK 模糊分类器, 具有更强大的泛化能力。

1　相关工作

1.1　TSK 模糊逻辑系统

TSK 模糊逻辑系统是最为著名的模糊分类器之一, 其最为著名的是它的模糊规则, 以第 k 条规则为例, TSK 模糊逻辑系统的模糊规则可以表示为

$$\begin{aligned}&\text{IF } x_1 \text{ is } A_1^k \text{ and } x_2 \text{ is } A_2^k \text{ and } \cdots\ x_m \text{ is } A_m^k,\\&\text{THEN } y_k = f^k(\boldsymbol{X})\ k = 1,2,\cdots,K\end{aligned} \tag{1}$$

式中：x_i代表输入向量$\boldsymbol{X}$的第i个特征；A_i^k代表第k条规则在第i个输入特征x_i上的前件模糊集；$f^k(\boldsymbol{X})$代表第k条规则的后件；K代表模糊规则的数量。这些模糊规则构成了一个模糊知识库，每个模糊规则都可以被看成一块模糊知识。式 (1) 中的 IF-THEN 模糊规则和人类的语言非常接近，因此可以说模糊规则能被人类读懂和理解，即 TSK 模糊系统具有较高的可解释性。其前件首先将输入数据划分为多个模糊区域，这通常用到模糊聚类[12]、K-means 聚类[13] 或等间距划分等算法，然后应用高斯隶属度函数计算得到每个输入数据的隶属度，最后将其规范化，这就得到了模糊规则的“IF”部分，这一部分描述了输入数据和各模糊区域之间的关联关系。其后件是对前件部分进行求解和优化，通常使用最小学习机[14] 或全连接层[15] 等算法，求解过程中得到的系数就是后件参数，这也就是模糊规则的“THEN”部分，这一部分描述的是隶属度和结果的关系。

近年来，TSK 模糊系统有迅速的发展和改进，Qin 等[16] 以特殊方式堆叠组合 0 阶 TSK 模糊分类器，显著提升了其分类性能，并增强了对环境变化的适应性。Xue 等[17] 引入了一种独特的门函数，通过在训练阶段同时进行特征选择和规则提取，有效提升了模糊规则的质量。Cui 等[18] 用层归一化的方法缓解了梯度消失的问题，提高了 TSK 的泛化性能。

1.2　深度学习模型

CNN[19] 是一类专门设计用于处理网格结构数据的深度学习模型，最初主要应用于计算机视觉任务。CNN 通过卷积层[20]、池化层和全连接层等组件，以端到端的方式学习输入数据的特征表示，通过卷积操作来有效地提取输入数据中的特征。这些层通常包括 softmax 激活函数，用于多类别分类问题。近年来，CNN 也发展出了许多变体，如 ResNet[21]、MobileNet[22]、EfficentNet[23] 等。

GCN[24] 是另一种深度学习模型，GCN 的核心思想是通过卷积操作来学习图数据中节点之间的关系，使得模型能够有效地捕捉图的结构信息。GCN 主要用于处理图结构数据，其中节点表示图中的实体，边表示节点之间的关系。这种图结构的表示可以用邻接矩阵来描述，邻接矩阵的矩阵元素表示节点之间的连接关系。近年来，GCN 被广泛用于交通流预测等领域，并在原有的基础上产生了许多新的变体，如 Zhao 等[25] 将时空交通数据建模为图结构，并利用图卷积操作来学习时空特征，利用时空邻近性以及节点之间的交互关系，有效地捕获交通数据的复杂时空动态。Abu-Ei-Haija 等[26] 结合多个尺度的图卷积操作，在不同的尺度下捕获图数据的局部和全局特征，提高了节点分类任务的性能和泛化能力。

Transformer[27] 是一种基于自注意力机制[28] 的深度学习模型架构，于 2017 年被提出，其最初用于自然语言处理 (natural language processing, NLP) 领域。Transformer 使用自注意力机制，通过对输入序列中不同位置的元素赋予不同的权重，实现对序列内元素之间关系的建模。这使得模型能够处理长距离依赖关系，而无需依赖固定大小的局部窗口。近年来，越来越多的学者在 Transformer 的基础上研究出新的深度学习模型。Devlin 等[29] 提出了一种基于 Transformer 的预训练语言模型 BERT(bidirectional encoder representations from transformers)，通过双向 Transformer 编码器来学习句子表示。Yang 等[30] 通过组合自回归和自编码的方式来进行训练模型，在多项 NLP 任务上取得了最佳性能，超越了之前的许多模型。除此以外，最近热门的 GPT(generative pre-trained transformer) 系列模型[31] 也是基于 Transformer 的预训练语言模型，GPT 模型使用了 Transformer 的解码器结构，并采用了自回归生成的方式来生成文本。

深度学习模型因其强大的学习能力被各大领域的学者广泛使用，其支持端到端的学习并且能适应大规模数据，因此在许多领域取得了巨大的成功。但是深度学习模型一般是黑盒模型，其内部的决策过程难以解释，并且在调优和调试时由于模型内部结构复杂，其行为和性能很难被完全理解和解释。

1.3　知识蒸馏

在机器学习中，大模型能从大量的复杂数据中进行学习和泛化特征，学习能力强；相比之下，小模型结构简单，计算消耗更少，但其泛化能力比不过大模型。因此可以通过知识蒸馏将大模型学到的隐藏知识，即映射关系传递给小模型，以此来提高小模型的性能。首先将数据集同时输入教师模型和学生模型，通过带有温度参数τ的 softmax 函数将两者的输出$\boldsymbol{Z}_i$转化为软标签β_i：

$$\beta_i = \frac{\exp(\boldsymbol{Z}_i/\tau)}{\sum_{j=1}^{C}\exp(\boldsymbol{Z}_j/\tau)} \tag{2}$$

随后再计算两者的蒸馏损失，然后计算学生模型与真实标签的硬损失。两个损失构成了学生模型的损失函数，知识蒸馏的目标就是最小化学生模型的损失函数，以提高模型的性能。

知识蒸馏可以从性能较好但运行时间较长的

大模型中提取隐藏知识, 传递给运行时间短但性能较差的小模型, 在保证模型运行时间短的同时提升了模型的性能。近年来, 许多学者利用知识蒸馏来提升小模型的性能表现[32-34]。知识蒸馏模型如图 1 所示。

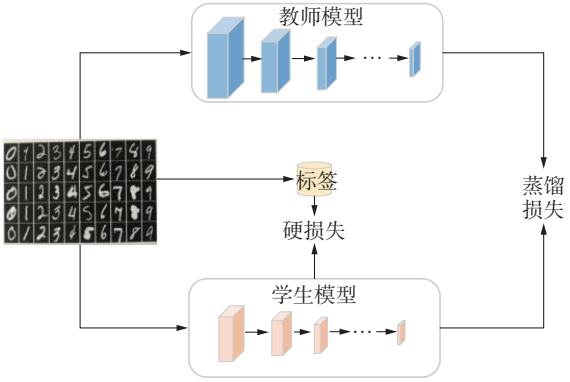

图 1　知识蒸馏模型

Fig. 1　Knowledge distillation model

2　模型具体介绍

2.1　整体模型框架

本文采用多教师自适应知识蒸馏, 选用 GCN、CNN、Transformer 作为教师模型, 一阶 TSK 模糊逻辑系统作为学生模型, 将隐藏知识从教师网络提取出来, 并传递给学生模型, 在保持学生模型高解释性的同时, 提高学生模型的性能。模型的整体架构如图 2 所示。从图 2 中可以看到, 将输入同时输进学生模型和 3 个教师模型, 可以得到各自的概率输出。3 个教师模型的概率输出通过设计的自适应权重分配器整合成一个新的概率输出, 再计算其与学生模型输出的软标签, 计算两者的蒸馏损失。将学生模型的输出与真实标签作交叉熵, 就可以得到两者的硬损失。将硬损失与蒸馏损失各自加权, 即可得到模型的总损失。

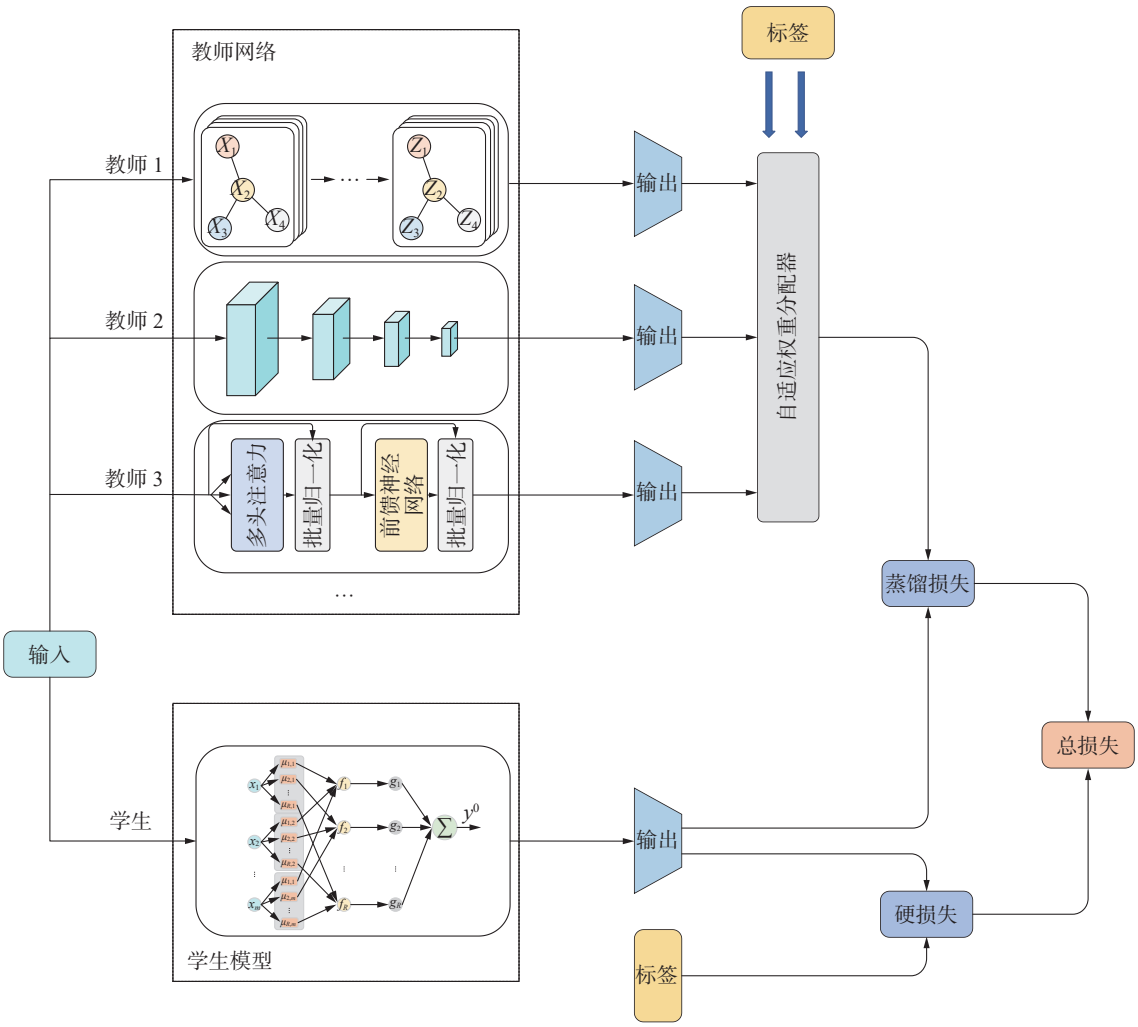

图 2　TSK-MTAKD 整体模型

Fig. 2　Overall model of TSK-MTAKD

2.2　学生模型和教师模型的处理

2.2.1　教师模型

GCN 能根据节点间的关系创建邻接矩阵, 具有良好的图结构处理能力; CNN 因其层次化结构而在图像等高维数据处理上性能强大并得到了广泛的应用; Transformer 则因其采用注意力机制而

使系统拥有强大的序列数据处理能力。因此，为了提高模型面对不同数据的处理能力，本研究选择 GCN、CNN、Transformer 作为教师网络，假设输入的特征为 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_m)^{\mathrm{T}}\in\mathbb{R}^{N\times m}$：

对于 GCN，假设有 N 个样本，本文针对输入的特征 $\boldsymbol{X}$ 创建一个邻接矩阵 $\boldsymbol{A}$，$\boldsymbol{A}\in\mathbb{R}^{m\times m}$，然后通过捕获特征之间的内在联系来更新邻接矩阵 $\boldsymbol{A}$，对于整体的 GCN 框架，正向传播公式为

$$\boldsymbol{Z}_{\mathrm{GCN}}=f(\boldsymbol{X},\boldsymbol{A})=\mathrm{softmax}(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}^{(0)}) \tag{3}$$

式中：$\boldsymbol{Z}_{\mathrm{GCN}}$ 是 GCN 的输出；$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$，$\boldsymbol{I}$ 是单位矩阵；$\tilde{\boldsymbol{D}}$ 是 $\tilde{\boldsymbol{A}}$ 的度矩阵，$\tilde{\boldsymbol{D}}_{ii}=\sum_j\tilde{\boldsymbol{A}}_{ij}$；$\boldsymbol{W}^{(0)}$ 是第 0 层的权重矩阵。

对于 CNN，本文选择使用含有 4 层卷积层的 CNN。将输入的特征 $\boldsymbol{X}$ 输入到教师模型 CNN 中，首先在卷积层中，采用卷积核进行卷积运算，随后加入偏置，再利用激活函数进行非线性变换，得到的输出作为下一层卷积层的输入再次卷积，具体公式为

$$\begin{aligned}\boldsymbol{Z}_0&=f(\boldsymbol{W}_0\boldsymbol{X}+\boldsymbol{b}_0)\\ \boldsymbol{Z}_1&=f(\boldsymbol{W}_1\boldsymbol{Z}_0+\boldsymbol{b}_1)\\ &\vdots\\ \boldsymbol{Z}_r&=f(\boldsymbol{W}_r\boldsymbol{Z}_{r-1}+\boldsymbol{b}_r)\end{aligned} \tag{4}$$

式中：$\boldsymbol{Z}_r$ 表示第 r 层卷积层的输出，$f(\cdot)$ 代表激活函数，$\boldsymbol{W}_r$ 代表第 r 层卷积层的卷积核，b_r 表示第 r 层卷积层的偏置。

卷积后利用全连接层将深度特征映射到新的特征空间，随后得到的输出利用 softmax 激活函数转化为概率输出，具体公式为

$$\boldsymbol{Z}_{\mathrm{a}}=f(\boldsymbol{W}_{\mathrm{a}}\boldsymbol{Z}_r+\boldsymbol{b}_{\mathrm{a}}) \tag{5}$$

$$\boldsymbol{Z}_{\mathrm{CNN}}=\mathrm{softmax}(\boldsymbol{Z}_{\mathrm{a}}) \tag{6}$$

式中：a 表示全连接层，$\boldsymbol{Z}_{\mathrm{a}}$ 表示全连接层的输出，$f(\cdot)$ 代表激活函数，$\boldsymbol{W}_{\mathrm{a}}$ 代表全连接层的权值，b_{a} 表示全连接层的偏置。由此经过逐层的神经表达，获得最后的输出 $\boldsymbol{Z}_{\mathrm{CNN}}$，$\boldsymbol{Z}_{\mathrm{CNN}}\in\mathbb{R}^{N\times C}$，其中 N 是样本数，C 是类别数。

对于 Transformer，首先对输入特征 $\boldsymbol{X}$ 作 3 种不同的线性变换得到 $\boldsymbol{Q}$(query)，$\boldsymbol{K}$(key)，$\boldsymbol{V}$(value)：

$$\boldsymbol{Q}=\boldsymbol{W}^{\boldsymbol{Q}}\boldsymbol{X}\in\mathbb{R}^{m\times h} \tag{7}$$

$$\boldsymbol{K}=\boldsymbol{W}^{\boldsymbol{K}}\boldsymbol{X}\in\mathbb{R}^{m\times h} \tag{8}$$

$$\boldsymbol{V}=\boldsymbol{W}^{\boldsymbol{V}}\boldsymbol{X}\in\mathbb{R}^{m\times g} \tag{9}$$

式中：$\boldsymbol{W}^{\boldsymbol{Q}}$、$\boldsymbol{W}^{\boldsymbol{K}}$、$\boldsymbol{W}^{\boldsymbol{V}}$ 都是可学习的参数矩阵，g、h 是各自的隐藏层维度。随后进行注意力加权，得到加权后的特征：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{m}}\right)V \tag{10}$$

将上述过程重复 G 次，然后把输出拼接起来，公式描述为

$$\boldsymbol{T}_i=\mathrm{Attention}(\boldsymbol{X}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{X}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{X}\boldsymbol{W}_i^{\boldsymbol{V}}) \tag{11}$$

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_G)\boldsymbol{W}^{\mathrm{O}} \tag{12}$$

式中：$\boldsymbol{W}^{\mathrm{O}}$ 是输出变换矩阵，$\boldsymbol{W}_i^{\boldsymbol{Q}}\in\mathbb{R}^{m\times h}$，$\boldsymbol{W}_i^{\boldsymbol{K}}\in\mathbb{R}^{m\times h}$，$\boldsymbol{W}_i^{\boldsymbol{V}}\in\mathbb{R}^{m\times g}$。然后利用残差连接，防止梯度消失或梯度爆炸。最后利用 softmax 得到输出结果，具体公式为

$$\boldsymbol{X}_{\mathrm{attention}}=\boldsymbol{X}+\boldsymbol{X}_{\mathrm{attention}} \tag{13}$$

$$\boldsymbol{Z}_{\mathrm{Tr}}=\sigma(\boldsymbol{W}^1(\boldsymbol{W}^0\boldsymbol{X}_{\mathrm{attention}}+\boldsymbol{b}^0)+\boldsymbol{b}^1) \tag{14}$$

式中：$\boldsymbol{Z}_{\mathrm{Tr}}$ 表示 Transformer 的输出，$\boldsymbol{W}^0$ 和 $\boldsymbol{W}^1$ 是参数矩阵，$\boldsymbol{b}^0$ 和 $\boldsymbol{b}^1$ 是偏置参数，$\sigma(\cdot)$ 是 softmax 函数。

2.2.2　学生模型

TSK 模糊逻辑系统具有模型小、运行时间快和可解释性较强等特点[35]，被广泛运用于处理模糊问题，因此本研究选择一阶 TSK 模糊逻辑系统作为学生模型。

对于输入的特征 $\boldsymbol{X}$，IF-THEN 规则中的 $f^k(\boldsymbol{X})$ 可以表示为

$$\begin{aligned}f^k(\boldsymbol{X})=P_0^k+x_1^kP_1^k+x_2^kP_2^k+\cdots+x_m^kP_m^k,\\ k=1,2,\cdots,K\end{aligned} \tag{15}$$

式中：$P_i^k(i=1,2,\cdots,m)$ 是第 k 条模糊规则的后件参数，K 是模糊系统中模糊规则的总数。第 k 条模糊规则将输入向量 $\boldsymbol{X}$ 映射到输出 $f(\boldsymbol{X})$ 可以定义为

$$f(\boldsymbol{X})=\sum_{k=1}^{K}\frac{\mu^k(\boldsymbol{X})}{\sum_{k'=1}^{K}\mu^{k'}(\boldsymbol{X})}f^k(\boldsymbol{X})=\sum_{k=1}^{K}\tilde{\mu}^k(\boldsymbol{X})f^k(\boldsymbol{X}) \tag{16}$$

式中：$\mu^k(\boldsymbol{X})$ 是第 k 条模糊规则的模糊隶属程度，规范化后得到 $\tilde{\mu}^k(\boldsymbol{X})$。

$$\mu^k(\boldsymbol{X})=\prod_{i=1}^{b}\mu_{A_i^k}(x_i) \tag{17}$$

式中：$\mu_{A_i^k}(x_i)$ 是 x_i 在模糊集 A_i^k 上的隶属程度，一般使用高斯隶属度函数来计算模糊隶属度程度：

$$\mu_{A_i^k}(x_i)=\exp\left(\frac{-(x_i-v_i^k)^2}{2\delta_i^k}\right) \tag{18}$$

式中：δ_i^k 是核宽；v_i^k 是中心参数，是每个模糊规则的中心。v_i^k 和 δ_i^k 被称为前件参数，v_i^k 从 $\{0,0.25,0.5,0.75,1\}$ 中随机选择，可以被自然语言解释为{非常低，低，中，高，非常高}，这也保证了前件的可解释性。因此，学生模型的输出 $\boldsymbol{Z}_{\mathrm{s}}$ 可以线性表示为

$$\boldsymbol{Z}_{\mathrm{s}}=f(\boldsymbol{X})=\boldsymbol{P}_{\mathrm{d}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{d}} \tag{19}$$

式中：$\boldsymbol{P}_{\mathrm{d}}$ 是模糊规则的后件参数，相关的参数计算公式为

$$\boldsymbol{x}_{\mathrm{d}}=\left[(\tilde{\boldsymbol{x}}^1)^{\mathrm{T}},(\tilde{\boldsymbol{x}}^2)^{\mathrm{T}},\cdots,(\tilde{\boldsymbol{x}}^K)^{\mathrm{T}}\right]^{\mathrm{T}}\in\mathbb{R}^{K(m+1)} \tag{20}$$

$$\tilde{\boldsymbol{x}}^K = \tilde{\mu}^k(\boldsymbol{X})\boldsymbol{X}_{\mathrm{e}} \in \mathbb{R}^{K(m+1)} \tag{21}$$

$$\boldsymbol{X}_{\mathrm{e}} = (1, \boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{K(m+1)} \tag{22}$$

$$\boldsymbol{P}_{\mathrm{d}} = [(\boldsymbol{P}^1)^{\mathrm{T}}, (\boldsymbol{P}^2)^{\mathrm{T}}, \cdots, (\boldsymbol{P}^K)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{K(m+1)} \tag{23}$$

$$\boldsymbol{P}^k = (p_0^k, p_1^k, \cdots, p_m^k)^{\mathrm{T}} \in \mathbb{R}^{K(m+1)} \tag{24}$$

对于学生模型的后件参数, 本文将用梯度下降算法进行更新, 相关公式为

$$H = -\sum_{i=1}^{N}\sum_{t=1}^{C} Y_{i,t} \log(\boldsymbol{Z}_{\mathrm{s}_{i,t}}) \tag{25}$$

$$\boldsymbol{P}_{\mathrm{d}}(q+1) = \boldsymbol{P}_{\mathrm{d}}(q) - \eta \frac{\partial H}{\partial \boldsymbol{P}_{\mathrm{d}}(q)} \tag{26}$$

式中: H是交叉熵损失, η是给定学习率。

2.3 自适应权重分配器

在多教师知识蒸馏中, 每个教师都会有一个输出, 整合每个教师的输出就会得到一个新的输出, 此时就会涉及权重分配的问题。以往的多教师模型一般是对多个教师的输出取平均, 或是类似于设置固定参数值的无标签策略[9,11]。但这往往是不合理的, 不同的教师对于不同的数据集会有不一样的效果, 因此, 本文设计了一种能自适应分配教师模型输出权重的方法, 让不同的教师模型根据不同的应用场景来自适应地分配权重。

在现实中, 准确率更高的教师模型往往更被信任, 更被认为“能教出好的学生”。因此, 本文以此设计自适应权重分配器, 计算每个教师模型的输出与真实标签的交叉熵损失, 损失更小的被认为准确率更高, 其对应的教师模型则被赋予更高的权重。以本文中 3 个教师模型为例, 对于 3 个教师的输出 $\boldsymbol{Z}_{\mathrm{Total}} = [\boldsymbol{Z}_{\mathrm{GCN}}, \boldsymbol{Z}_{\mathrm{CNN}}, \boldsymbol{Z}_{\mathrm{Tr}}]$, 分配权重的具体公式为

$$L_t^q = -\sum_{c=1}^{C} Y^c \log(\boldsymbol{Z}_{t_q}^c) \tag{27}$$

$$w_t^q = 1 - \frac{\exp(L_t^q)}{\sum\limits_j \exp(L_t^j)} \tag{28}$$

式中: t_q表示第q个教师模型, L_t^q表示第q个教师模型与真实标签的交叉熵损失, C是类别数, Y是真实标签, w_t^q表示第q个教师模型对应的权重。从公式可以看出, 更高的损失对应着更低的权重。由此, 经过自适应权重分配器整合后的教师模型总输出$\boldsymbol{Z}_{\mathrm{Teacher}}$可以表示为

$$\boldsymbol{Z}_{\mathrm{Teacher}} = \sum_{q=1}^{Q} w_t^q \boldsymbol{Z}_{t_q} \in \mathbb{R}^{N\times C} \tag{29}$$

2.4 多教师知识蒸馏

知识蒸馏是一种利用大模型提升小模型性能的方式, 它能在保持小模型运行速度与可解释性的同时, 提高小模型的泛化性能。传统的单教师模型在教师模型表现不佳的情况下, 会影响整体的蒸馏效果, 从而影响模型的整体性能。本文采用了 3 个深度学习网络: GCN、CNN、Transformer 来作为教师模型。

多教师知识蒸馏与单教师知识蒸馏相同点在于, 在将多个教师网络的输出整合成一个输出后, 利用带有温度参数τ的 softmax 函数将教师与学生的输出转化为软标签$\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_i, \cdots, \beta_C] \in \mathbb{R}^{1\times C}$, 教师与学生的软标签都可以用公式算得:

$$\beta_i = \frac{\exp(\boldsymbol{Z}_i/\tau)}{\sum\limits_{j=1}^{C} \exp(\boldsymbol{Z}_j/\tau)} \tag{30}$$

通过计算 KL(Kullback-Leibler divergence) 散度来比较两者的差异, 即蒸馏损失L_{kd}。再将学生模型的输出与真实标签作交叉熵, 计算出硬损失L_{CE}。最后 2 个损失加权, 就得到了模型的总损失函数L_{Total}, 具体公式为

$$L_{\mathrm{kd}} = \mathrm{KL}(\boldsymbol{\beta}^{\mathrm{t}} \| \boldsymbol{\beta}^{\mathrm{s}}) = \sum_{c=1}^{C} \beta_c^{\mathrm{t}} \log\left(\frac{\beta_c^{\mathrm{t}}}{\beta_c^{\mathrm{s}}}\right) \tag{31}$$

$$L_{\mathrm{CE}} = -\sum_{c=1}^{C} Y^c \log(Z_{\mathrm{s}}^c) \tag{32}$$

$$L_{\mathrm{Total}} = \alpha L_{\mathrm{kd}} + (1-\alpha) L_{\mathrm{CE}} \tag{33}$$

式中: C是类别数, Z^c表示第c类的概率输出, t代表整合后的教师, s代表学生。

2.5 TSK-MTAKD 学习算法

TSK-MTAKD 算法包括教师模型和学生模型的构建、教师模型输出的整合以及知识蒸馏, 具体流程见算法 1~2。

算法 1 构建教师模型和学生模型

输入 数据集$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N)^{\mathrm{T}} \in \mathbb{R}^{N\times m}$; 真实标签$\boldsymbol{Y} = (y_1, y_2, \cdots, y_i, \cdots, y_N)^{\mathrm{T}} \in \mathbb{R}^{N\times 1}$; 模糊规则数$K$; 最大迭代次数$\theta$; 阈值参数$\xi$; 学习率$\eta$。

输出 3 个教师模型的输出和学生模型的输出。

1) 以随机的方式从由{0,0.25,0.5,0.75,1}构成的固定模糊划分中选择高斯函数的中心v_i^k, 设置核宽δ_i^k为正值, 利用式 (15)~(17) 计算得归一化的模糊隶属度;

2) 利用式 (19)~(23) 计算学生模型的前件参数矩阵;

3) 利用梯度下降算法计算学生模型的后件参数;

4) 初始化后件参数$\boldsymbol{P}_{\mathrm{d}}$并设定$q = 1$;

Repeat

利用式 (24)~(25) 计算 $\boldsymbol{P}_{\mathrm{d}}(q+1)$;

$$q=q+1;$$

Until $H(q)-H(q-1)\leqslant\xi$ or $q\geqslant\theta$

5) 利用式 (2) 计算教师模型 GCN 的输出 $\boldsymbol{Z}_{\mathrm{GCN}}$;

6) 利用式 (3)~(5) 计算教师模型 CNN 的输出 $\boldsymbol{Z}_{\mathrm{CNN}}$;

7) 利用式 (6)~(13) 计算教师模型 Transformer 的输出 $\boldsymbol{Z}_{\mathrm{Tr}}$;

8) 计算得到学生模型的概率输出 $\boldsymbol{Z}_{\mathrm{s}}=\boldsymbol{P}_{\mathrm{d}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{d}}$;

算法 2　基于多教师自适应知识蒸馏的 TSK 模糊分类器 (TSK-MTAKD)

输入　数据集 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_N)^{\mathrm{T}}\in\mathbb{R}^{N\times m}$; 真实标签 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_i,\cdots,y_N)^{\mathrm{T}}\in\mathbb{R}^{N\times 1}$; 3 个教师模型的输出 $\boldsymbol{Z}_{\mathrm{GCN}}\in\mathbb{R}^{N\times C}$, $\boldsymbol{Z}_{\mathrm{CNN}}\in\mathbb{R}^{N\times C}$, $\boldsymbol{Z}_{\mathrm{Tr}}\in\mathbb{R}^{N\times C}$; 学生模型的输出 $\boldsymbol{Z}_{\mathrm{s}}\in\mathbb{R}^{N\times C}$; 最大迭代次数 ε; 阈值参数 ξ; 学习率 η; 温度参数 τ; 蒸馏参数 α。

输出　TSK-MTAKD 的输出。

1) 利用式 (26)~(28) 计算各教师模型与真实标签的交叉熵, 并以此分配各自所占的权重, 求出最终的教师模型总输出 $\boldsymbol{Z}_{\mathrm{Teacher}}\in\mathbb{R}^{N\times C}$;

2) 通过式 (29) 计算得到教师模型总输出和学生模型输出的软标签 $\boldsymbol{\beta}^{\mathrm{t}}$ 和 $\boldsymbol{\beta}^{\mathrm{s}}$;

3) 利用梯度下降算法更新 TSK-MTAKD 的后件参数:

4) 初始化后件参数 $\boldsymbol{P}_{\mathrm{d}}$ 并设定 $q=1$;

Repeat

$$\boldsymbol{P}_{\mathrm{d}}(q+1)=\boldsymbol{P}_{\mathrm{d}}(q)-\eta\frac{\partial L_{\mathrm{Total}}}{\partial\boldsymbol{P}_{\mathrm{d}}(q)};$$

$$q=q+1;$$

Until $L_{\mathrm{Total}}(q)-L_{\mathrm{Total}}(q-1)\leqslant\xi$ or $q\geqslant\varepsilon$

5) 计算 TSK-MTAKD 的输出。

TSK-MTAKD 的时间复杂度主要由其学生模型决定, 包括: 建立中心矩阵的时间复杂度为 $O(5mK)$, 建立宽度矩阵的时间复杂度为 $O(mK)$, 生成前件参数矩阵的时间复杂度为 $O(5Nm^3K)$, 生成后件参数矩阵的时间复杂度为 $O(\varepsilon NmKC)$。其中, ε 是最高迭代次数, N 是样本数, m 是特征数, K 是模糊规则数, C 是类别数, 因此 TSK-MTAKD 的总时间复杂度为 $O(5mK+mK+5Nm^3K+\varepsilon NmKC)\approx O(5Nm^3K+\varepsilon NmKC)$。

3　实验

本文选择使用 4 个模糊分类器和 4 个模糊知识蒸馏模型与 TSK-MTAKD 进行对比实验, 本文实验运行的硬件环境为: Intel(R) Core(TM) i9-12900H 2.5 GHz 搭载 16GB RAM 与 Microsoft Windows 11 系统, 编程环境为: Python 3.7.16 配备 Torch 1.8.1 库。

3.1　数据集

由于本文主要应用于分类任务, 因此本节的实验从 UCI 数据库中随机选择了 13 个不同规模不同场景的分类数据集, 本实验对数据集采用归一化处理, 并将分类特征转化为数值特征, 数据集特征数与样本数具体介绍如表 1 所示。

表 1　数据集简介

Table 1　Introduction to the dataset

数据集	特征数	样本数
Iris(IRI)	4	150
Wine(WIN)	13	178
Sonar(SON)	60	208
Seeds(SEE)	7	210
Ionosphere(ION)	32	351
Vote(VOT)	16	435
Wisconsin(WIS)	9	683
QSAR-Biodegradation(QSA)	41	1 055
Cardiotocography(CAR)	21	2 126
Titanic(TIT)	3	2 201
Segmentation(SEG)	18	2 310
Brainweb(BRA)	3	20 000
Adult(ADU)	14	48 841

3.2　实验环境与参数设置

本文的 TSK-MTAKD 是从 3 个教师模型中提取隐藏知识的新型 TSK 模糊蒸馏分类器, 因此, 为评价此模型的性能, 本文选择将 n 阶 TSK 模糊分类器 ($n=1,2$) 作为对比模型, 本文将两种利用不同方法求解后件参数的 TSK 模糊分类器区分开来, 其中, $\mathrm{TSK}_{\mathrm{v1}}^{n}$ (n 为阶次, $n=1,2$) 的后件参数使用最小学习机求解, $\mathrm{TSK}_{\mathrm{v2}}^{n}$ (n 为阶次, $n=1,2$) 的后件参数使用梯度下降法进行更新。同时, 本文针对多教师知识蒸馏模型, 将 3 个单教师知识蒸馏模型加入到对比模型中, 在准确率和加权 F1 分数上与本文的 TSK-MTAKD 进行对比。其中, GKD、CKD、TrKD 的教师模型分别采用 GCN、CNN、Transformer, 而这 3 个单教师模型的学生模型都选择使用 TSK 模糊逻辑系统, 以此比较各自对学生模型的提升效果。除此以外, 本文设计 TSK-MTKD 模型作为对比模型, TSK-MTKD 采用与 TSK-MTAKD 相同的教师模型网络, 但采用取平均的权重分配策略。对比模型的具体信息如表 2 所示。

表 2　对比模型简介

Table 2　Introduction to comparative model

模糊分类器			
名称	前件	后件	阶次
TSK^1_{v1}	等间距划分	最小学习机	1
TSK^2_{v1}			2
TSK^1_{v2}		梯度下降	1
TSK^2_{v2}			2

知识蒸馏模型		
名称	教师模型	学生模型
GKD	GCN	TSK
CKD	CNN	TSK
TrKD	Transformer	TSK
TSK-MTKD	多教师	TSK

本文采用五折交叉在选取的数据集上验证所有模型性能, 对于模型所有可调参数采用网格搜索来选择。其中, 模糊规则数K的寻优范围为{1,2,⋯,20}; 温度参数τ的寻优范围为{1,5,10,20,100}; 蒸馏参数α的寻优范围为{0,0.25,0.50,0.75,1}; 最大迭代次数θ和ε都设置为 30; 阈值参数ξ设置为10^{-5}; 学习率η设置为 0.01; 其余参数设置为默认值[36]。

本文选择 2 个常用的性能指标来评判所有模型的分类性能, 分别是准确率 (accuracy, ACC) 和加权 F1 分数 (weighted F1 score, W-F)。对于每个数据集的最佳结果, 本文将用粗体进行标记, "—" 表示所采用的方法不能在 3 h 内计算出其结果。

3.3　实验结果与分析

表 3 给出了 TSK-MTAKD 与其他 8 个模糊分类器的性能对比结果, 针对这个表格, 本文得出以下结论。

表 3　所有模型在 UCI 数据集上的平均准确率和平均加权 F1 分数对比

Table 3　Comparison of average accuracy and average weighted F1 score of all models on UCI datasets

数据集	指标	IRI	WIN	SEE	TIT	ION	WIS	QSA	SON	SEG	VOT	CAR	BRA	ADU
TSK^1_{v1}	ACC	86.33	88.78	92.41	78.99	85.90	96.85	86.63	78.57	99.41	94.13	87.72	95.76	84.06
	W-F	86.43	88.84	92.42	78.96	85.11	96.82	86.56	78.43	99.40	94.18	87.76	95.75	83.92
TSK^2_{v1}	ACC	87.00	84.29	92.95	**79.11**	84.05	96.92	80.88	78.82	99.43	92.64	88.61	**95.77**	84.63
	W-F	87.11	84.31	92.99	**79.02**	83.43	96.86	80.82	77.80	99.42	92.66	88.87	**95.79**	**84.50**
TSK^1_{v2}	ACC	96.66	97.83	92.79	78.00	92.58	96.92	87.48	84.99	99.58	94.59	90.92	95.44	84.29
	W-F	96.72	97.85	92.84	77.64	91.72	96.64	86.93	84.63	99.43	94.29	91.04	95.45	84.26
TSK^2_{v2}	ACC	96.66	98.88	92.42	78.25	91.44	96.99	87.06	83.65	99.61	92.06	90.79	95.55	—
	W-F	96.67	98.84	92.39	78.20	90.43	96.69	86.39	83.20	99.56	91.58	90.45	95.45	—
GKD	ACC	97.22	98.32	92.86	78.55	92.77	97.07	88.23	85.03	99.60	94.60	91.11	95.57	84.36
	W-F	97.38	98.33	92.90	78.45	92.01	96.76	88.33	84.86	99.57	94.44	91.06	95.47	84.35
CKD	ACC	96.98	98.85	92.85	78.55	92.63	97.08	88.26	85.01	99.61	94.62	91.26	95.52	84.37
	W-F	97.12	98.88	92.86	78.24	92.24	96.88	88.25	84.72	99.60	94.32	91.08	95.46	84.38
TrKD	ACC	97.13	98.33	92.80	78.46	92.70	97.08	88.22	85.01	99.61	94.67	91.17	95.52	84.33
	W-F	97.41	98.86	92.89	78.26	92.36	96.86	88.16	84.88	99.59	94.35	91.12	95.45	84.31
TSK-MTKD	ACC	97.25	98.87	92.97	78.51	92.86	97.12	88.32	85.24	99.61	94.77	91.28	95.57	84.40
	W-F	97.42	98.88	92.99	78.41	92.44	96.98	88.36	84.92	99.58	94.56	91.12	95.47	84.38
TSK-MTAKD	ACC	**97.43**	**98.99**	**93.74**	78.76	**93.26**	**97.42**	**88.48**	**85.82**	**99.63**	**95.16**	**91.36**	95.59	**84.65**
	W-F	**97.45**	**98.92**	**93.77**	78.46	**92.59**	**97.32**	**88.46**	**85.60**	**99.61**	**94.92**	**91.20**	95.49	84.41

注: "—"表示所采用的方法不能在3 h内计算出其结果, 加粗表示本列最优结果。

1) 在准确率和加权 F1 分数方面, TSK-MTAKD 在绝大多数数据集上都取得了最好的性能表现, 特别是对于低阶 TSK 与高阶 TSK, TSK-MTAKD 都有显著的性能提升, 本文分析这是由于教师模型提取了隐藏知识传递给学生模型, 从而提升了模型的整体性能。

2)TSK-MTKD 与 3 个单教师的蒸馏模型相比, 都具有一定的提升, 本文分析这是由于多教师的蒸馏模型相对于单教师蒸馏模型而言, 能从多个教师模型中提取到更多的隐藏知识, 从而帮助学生模型提高了分类性能。

3)TSK-MTAKD 相较于使用平均分配策略的 TSK-MTKD 模型而言, 有更优异的性能表现, 本文认为, 这是由于 TSK-MTAKD 的自适应权重分

配器能主动分配权重，提取到 3 个教师中的隐藏知识，整合成最利于学生的隐藏知识，提高隐藏知识的利用率，从而稳定地提升模型性能。

表 4~5 给出了 TSK-MTAKD 和其余 4 个蒸馏模型与其相应的学生模型在准确率和加权 F1 分数上的实验结果对比。可以看出，TSK-MTAKD 展现出最高的性能提升，在准确率方面，TSK-MTAKD 在 13 个数据集的平均提升幅度比单教师蒸馏模型多了约 0.35 百分点，比 TSK-MTKD 多了约 0.25 百分点；在加权 F1 分数方面，TSK-MTAKD 在 13 个数据集的平均提升幅度比单教师蒸馏模型多了约 0.33 百分点，比 TSK-MTKD 多了约 0.21 百分点。由此可以得出，TSK-MTAKD 通过从多个教师提取隐藏知识并自适应地分配权重传递给学生模型，比单教师蒸馏模型更加有效地提高了 TSK 模糊分类器的性能；此外，利用本文设计的自适应权重分配器能较好地整合教师模型的隐藏知识，从而提高模型的鲁棒性。

表 4　GKD、CKD、TrKD、TSK-MTKD 和 TSK-MTAKD 与相应学生模型在 UCI 数据集上的准确率以及提升幅度对比

Table 4　Comparison of accuracy and improvement of GKD, CKD, TrKD, TSK-MTKD, and TSK-MTAKD with corresponding student models on UCI datasets　%

数据集	GKD			CKD			TrKD			TSK-TAKD			TSK-MTAKD		
	蒸馏	学生	提升	蒸馏	学生	提升	蒸馏	学生	提升	蒸馏	学生	提升	蒸馏	学生	提升
IRI	97.22	96.32	0.90	96.98	96.35	0.63	97.13	96.26	0.87	97.25	96.35	0.90	97.43	96.33	1.10
WIN	98.32	97.77	0.55	98.85	97.80	1.05	98.33	97.75	0.58	98.87	97.78	1.09	98.99	97.72	1.27
SEE	92.86	91.71	1.15	92.85	91.70	1.15	92.80	91.67	1.13	92.97	91.72	1.25	93.74	91.77	1.97
TIT	78.55	77.66	0.89	78.55	77.77	0.78	78.46	77.68	0.78	78.51	77.67	0.84	78.76	77.81	0.95
ION	92.77	92.45	0.32	92.63	92.28	0.35	92.70	92.37	0.33	92.86	92.44	0.42	93.26	92.48	0.78
WIS	97.07	96.87	0.20	97.08	96.87	0.21	97.08	96.84	0.24	97.12	96.85	0.27	97.42	96.86	0.56
QSA	88.23	87.35	0.88	88.26	87.37	0.89	88.22	87.31	0.91	88.32	87.36	0.96	88.48	87.33	1.15
SON	85.03	84.61	0.42	85.01	84.71	0.30	85.01	84.68	0.33	85.24	84.75	0.49	85.82	84.77	1.05
SEG	99.60	99.46	0.14	99.61	99.46	0.15	99.61	99.48	0.13	99.61	99.47	0.14	99.63	99.47	0.16
VOT	94.60	94.12	0.48	94.62	94.08	0.54	94.67	94.15	0.52	94.77	94.19	0.58	95.16	94.21	0.95
CAR	91.11	90.81	0.30	91.26	90.86	0.40	91.17	90.85	0.32	91.28	90.78	0.50	91.36	90.74	0.62
BRA	95.57	95.43	0.14	95.52	95.40	0.12	95.52	95.38	0.14	95.57	95.43	0.14	95.59	95.42	0.17
ADU	84.36	84.21	0.15	84.37	84.25	0.12	84.33	84.24	0.09	84.40	84.25	0.15	84.65	84.28	0.37
平均	91.94	91.44	0.50	91.96	91.45	0.51	91.92	91.43	0.49	92.06	91.46	0.60	92.33	91.48	0.85

表 5　GKD、CKD、TrKD、TSK-MTKD 和 TSK-MTAKD 与相应学生模型在 UCI 数据集上的加权 F1 分数以及提升幅度对比

Table 5　Comparison of weighted F1 scores and improvement of GKD, CKD, TrKD, TSK-MTKD, and TSK-MTAKD with corresponding student models on UCI datasets　%

数据集	GKD			CKD			TrKD			TSK-MTKD			TSK-MTAKD		
	蒸馏	学生	提升	蒸馏	学生	提升	蒸馏	学生	提升	蒸馏	学生	提升	蒸馏	学生	提升
IRI	97.38	96.45	0.93	97.12	96.58	0.54	97.41	96.57	0.84	97.42	96.50	0.92	97.45	96.48	0.97
WIN	98.33	97.58	0.45	98.88	97.84	1.04	98.86	97.78	1.08	98.88	97.77	1.11	98.92	97.72	1.20
SEE	92.90	91.81	1.09	92.86	91.77	1.09	92.89	91.68	1.21	92.99	91.72	1.27	93.77	91.81	1.96
TIT	78.45	77.33	1.12	78.24	77.32	0.92	78.26	77.38	0.88	78.41	77.35	1.06	78.46	77.27	1.19
ION	92.01	91.52	0.49	92.24	91.67	0.57	92.36	91.65	0.71	92.44	91.66	0.78	92.59	91.68	0.91
WIS	96.76	96.58	0.18	96.88	96.62	0.26	96.86	96.60	0.26	96.98	96.61	0.37	97.32	96.61	0.71
QSA	88.33	86.90	1.43	88.25	86.88	1.37	88.16	86.77	1.39	88.36	86.85	1.51	88.46	86.81	1.65
SON	84.86	84.24	0.62	84.72	84.52	0.20	84.88	84.56	0.32	84.92	84.31	0.61	85.60	84.51	1.09
SEG	99.57	99.38	0.19	99.60	99.39	0.21	99.59	99.40	0.19	99.58	99.39	0.19	99.61	99.38	0.23
VOT	94.44	93.72	0.72	94.32	93.71	0.61	94.35	93.77	0.58	94.56	93.76	0.80	94.92	93.90	1.02
CAR	91.06	90.74	0.32	91.08	90.59	0.49	91.12	90.86	0.26	91.12	90.73	0.39	91.20	90.53	0.67
BRA	95.47	95.34	0.13	95.46	95.32	0.14	95.45	95.33	0.12	95.47	95.33	0.14	95.49	95.31	0.18
ADU	84.35	84.12	0.23	84.38	84.15	0.23	84.31	84.16	0.15	84.38	84.15	0.23	84.41	84.15	0.26
平均	91.82	91.21	0.61	91.84	91.25	0.59	91.88	91.27	0.61	91.96	91.24	0.72	92.17	91.24	0.93

图 3~4 给出了 TSK-MTAKD 和其余 4 个蒸馏模型在准确率和加权 F1 分数上提升的对比, 可以发现, 在面对同一个数据集时, 3 个单教师的蒸馏模型体现出不一样的性能, 如在 IRI 数据集上, GCN 作为教师模型的模糊蒸馏分类器能有较大的性能提升, CNN 作为教师模型的模糊蒸馏分类器性能提升较小; 而在 WIN 这一数据集上的效果则恰恰相反; 在 WIS 这一数据集上, 3 个单教师模糊蒸馏分类器的提升效果都不太明显, 但采用多教师知识蒸馏的 TSK-MTKD 和 TSK-MTAKD 对几乎每个数据集都有显著的提升效果。针对这一结果, 本文分析这是由于教师模型对于不同类别的数据集会有性能的差异, 从而影响蒸馏效果, 而多教师知识蒸馏能从不同的教师模型中学到知识, 从而减小数据集类型对模型的影响。除此以外本文发现, 使用平均分配权重策略的 TSK-MTKD 在部分数据集上对学生模型的提升并不是很明显, 如在 IRI、BRA 等数据集上, TSK-MTKD 对学生模型的提升效果几乎等价于某个单教师模型, 而在 TIT 这一数据集上 TSK-MTKD 甚至不如其中一个单教师模型, 而使用自适应权重分配器的 TSK-MTAKD 却能始终拥有较好的性能提升效果。这是由于在面对不同效果的教师模型时, 平均分配策略使性能较好的教师模型无法拥有更多的"话语权", 从而减少了有效知识的传递, 而自适应权重分配器能根据教师模型的表现情况自适应分配权重, 进而提高模型的性能。

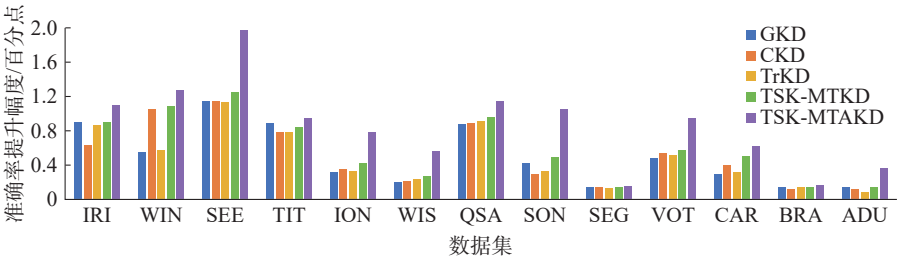

图 3　GKD、CKD、TrKD、TSK-MTKD 和 TSK-MTAKD 在准确率提升幅度上的对比

Fig. 3　Comparison of GKD, CKD, TrKD, TSK-MTKD, and TSK-MTAKD in terms of accuracy improvement

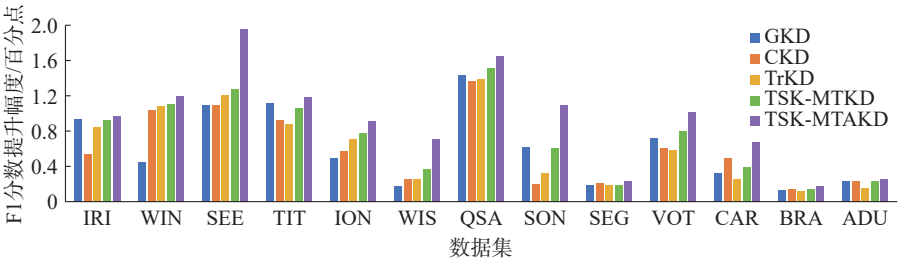

图 4　GKD、CKD、TrKD、TSK-MTKD 和 TSK-MTAKD 在加权 F1 分数提升幅度上的对比

Fig. 4　Comparison of GKD, CKD, TrKD, TSK-MTKD, and TSK-MTAKD in terms of weighted F1 score improvement

除此以外, 为观察本文提出的 TSK-MTAKD 与其他几个对比算法之间是否存在显著性差异, 本文引入 Frideman Ranking 测试。由于 TSK^{2}_{v2} 在 ADU 这一数据集上运行时间超过 3 h, 因此本文对除 ADU 数据集外的 12 个数据集进行统计分析, 置信度设置为 0.05。图 5 给出了排序结果, TSK-MTAKD 获得了最佳等级, 这说明 TSK-MTAKD 与其他 8 个对比算法之间具有显著差异。

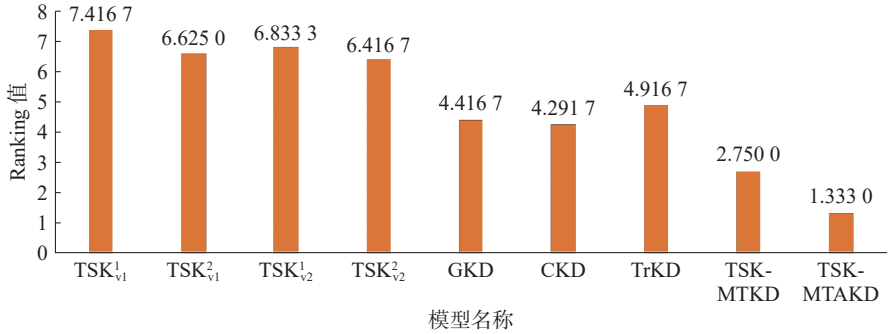

图 5　各模型 Ranking 值

Fig. 5　Ranking values of each model

4　结束语

本文主要将多教师知识蒸馏和 TSK 模糊分类器相结合，从而提升 TSK 模糊分类器的性能表现。本文提出了一种新型的 TSK 模糊蒸馏分类器，称为 TSK-MTAKD，从多个教师模型中提取隐藏知识，通过自适应权重分配器将 3 个教师的隐藏知识进行整合，得到对应的软标签，从而传递给 TSK 模糊分类器，实现性能的提升。在 13 个 UCI 数据集上的实验证明了 TSK-MTAKD 的性能优势。

除此以外，TSK-MTAKD 还有更多的地方值得研究。首先，将深入研究此模型在癫痫检测和运动预测等领域的实际应用。其次，目前出现了许多新的蒸馏方式，除了本文的多教师蒸馏外，改进蒸馏方式，利用如自蒸馏、解耦蒸馏等可能可以获得更好的蒸馏效果，因此蒸馏方式改进也是后续值得研究的方向。

参考文献：

[1] 苏丽, 孙雨鑫, 苑守正. 基于深度学习的实例分割研究综述[J]. 智能系统学报, 2022, 17(1): 16−31.
SU Li, SUN Yuxin, YUAN Shouzheng. A survey of instance segmentation research based on deep learning[J]. CAAI transactions on intelligent systems, 2022, 17(1): 16−31.

[2] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1−9.

[3] 赵壮壮, 王骏, 潘祥, 等. 任务间共享和特有结构分解的多任务 TSK 模糊系统建模[J]. 智能系统学报, 2021, 16(4): 622−629.
ZHAO Zhuangzhuang, WANG Jun, PAN Xiang, et al. Multi-task TSK fuzzy system modeling based on inter-task common and special structure decomposition[J]. CAAI transactions on intelligent systems, 2021, 16(4): 622−629.

[4] 黄震华, 杨顺志, 林威, 等. 知识蒸馏研究综述[J]. 计算机学报, 2022, 45(3): 624−653.
HUANG Zhenhua, YANG Shunzhi, LIN Wei, et al. Knowledge distillation: a survey[J]. Chinese journal of computers, 2022, 45(3): 624−653.

[5] 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.

[6] JÚNIOR J S S, MENDES J, SOUZA F, et al. Distilling complex knowledge into explainable T–S fuzzy systems [J]. IEEE transactions on fuzzy systems, 2025, 33(3): 1037−1048.

[7] GU Xiangming, CHENG Xiang. Distilling a deep neural network into a Takagi-Sugeno-Kang fuzzy inference system[EB/OL]. (2020−10−10)[2024−10−22]. https://arxiv.org/abs/2010.04974v1.

[8] ZHANG Xiongtao, YIN Zezong, JIANG Yunliang, et al. Fuzzy knowledge distillation from high-order TSK to low-order TSK[EB/OL]. (2023−02−16)[2024−10−22]. https://arxiv.org/abs/2302.08038v1.

[9] ERDEM D, KUMBASAR T. Enhancing the learning of interval type-2 fuzzy classifiers with knowledge distillation[C]//2021 IEEE International Conference on Fuzzy Systems. Luxembourg: IEEE, 2021: 1−6.

[10] YOU Shan, XU Chang, XU Chao, et al. Learning from multiple teacher networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 1285−1294.

[11] WU M C, CHIU C T, WU K H. Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019: 2202−2206.

[12] PAL N R, PAL K, KELLER J M, et al. A possibilistic fuzzy c-means clustering algorithm[J]. IEEE transactions on fuzzy systems, 2005, 13(4): 517−530.

[13] VENKATACHALAM K, REDDY V P, AMUDHAN M, et al. An implementation of K-means clustering for efficient image segmentation[C]//2021 10th IEEE International Conference on Communication Systems and Network Technologies. Bhopal: IEEE, 2021: 224−229.

[14] ZHOU Ta, CHUNG F L, WANG Shitong. Deep TSK fuzzy classifier with stacked generalization and triplely concise interpretability guarantee for large data[J]. IEEE transactions on fuzzy systems, 2017, 25(5): 1207−1221.

[15] ALBAWI S, ABED MOHAMMED T, AL-ZAWI S. Understanding of a convolutional neural network[C]//2017 International Conference on Engineering and Technology. Antalya: IEEE, 2017: 1−6.

[16] QIN Bin, NOJIMA Y, ISHIBUCHI H, et al. Realizing deep high-order TSK fuzzy classifier by ensembling interpretable zero-order TSK fuzzy subclassifiers[J]. IEEE transactions on fuzzy systems, 2021, 29(11): 3441−3455.

[17] XUE Guangdong, WANG Jian, ZHANG Bingjie, et al. Double groups of gates based Takagi-Sugeno-Kang (DG-TSK) fuzzy system for simultaneous feature selection and rule extraction[J]. Fuzzy sets and systems, 2023, 469: 108627.

[18] CUI Yuqi, XU Yifan, PENG Ruimin, et al. Layer normalization for TSK fuzzy system optimization in regression problems[J]. IEEE transactions on fuzzy systems, 2023, 31(1): 254−264.

[19] 刘万军, 姜岚, 曲海成, 等. 融合 CNN 与 Transformer 的 MRI 脑肿瘤图像分割[J]. 智能系统学报, 2024, 19(4): 1007−1015.
LIU Wanjun, JIANG Lan, QU Haicheng, et al. MRI brain tumor image segmentation by fusing CNN and Transformer[J]. CAAI transactions on intelligent systems, 2024, 19(4): 1007−1015.

[20] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278−2324.

[21] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[22] WANG Wei, LI Yutao, ZOU Ting, et al. A novel image classification approach via dense-MobileNet models[J]. Mobile information systems, 2020, 2020(1): 7602384.

[23] TAN Mingxing, LE Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. California: PMLR, 2019: 6105−6114.

[24] 赵文竹, 袁冠, 张艳梅, 等. 多视角融合的时空动态 GCN 城市交通流量预测[J]. 软件学报, 2024, 35(4): 1751−1773.
ZHAO Wenzhu, YUAN Guan, ZHANG Yanmei, et al. Multi-view fused spatial-temporal dynamic GCN for urban traffic flow prediction[J]. Journal of software, 2024, 35(4): 1751−1773.

[25] ZHAO Ling, SONG Yujiao, ZHANG Chao, et al. T-GCN: a temporal graph convolutional network for traffic prediction[J]. IEEE transactions on intelligent transportation systems, 2020, 21(9): 3848−3858.

[26] ABU-EL-HAIJA S, KAPOOR A, PEROZZI B, et al. N-GCN: multi-scale graph convolution for semi-supervised node classification[C]//Uncertainty in Artificial Intelligence. Tel Aviv: PMLR, 2020: 841−851.

[27] SUBAKAN C, RAVANELLI M, CORNELL S, et al. Attention is all you need in speech separation[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 21−25.

[28] 任欢, 王旭光. 注意力机制综述[J]. 计算机应用, 2021, 41(S1): 1−6.
REN Huan, WANG Xuguang. Review of attention mechanism[J]. Journal of computer applications, 2021, 41(S1): 1−6.

[29] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019: 4171−4186.

[30] YANG Zhilin, DAI Zihang, YANG Yiming, et al. XL-Net: generalized autoregressive pretraining for language understanding[C]//Proceedings of the 32nd Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2018: 5754−5764.

[31] HAUPT C E, MARKS M. AI-generated medical advice-GPT and beyond[J]. JAMA, 2023, 329(16): 1349−1350.

[32] CUI Kaiwen, YU Yingchen, ZHAN Fangneng, et al. KD-DLGAN: data limited image generation *via* knowledge distillation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 3872−3882.

[33] GUO Ziyao, YAN Haonan, LI Hui, et al. Class attention transfer based knowledge distillation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 11868−11877.

[34] GOU Jianping, XIONG Xiangshuo, YU Baosheng, et al. Multi-target knowledge distillation via student self-reflection[J]. International journal of computer vision, 2023, 131(7): 1857−1874.

[35] ZHANG Pu, SHANG Changjing, SHEN Qiang. Fuzzy rule interpolation with K-neighbors for TSK models[J]. IEEE transactions on fuzzy systems, 2022, 30(10): 4031−4043.

[36] TIAN Xiaobin, DENG Zhaohong, YING Wenhao, et al. Deep multi-view feature learning for EEG-based epileptic seizure detection[J]. IEEE transactions on neural systems and rehabilitation engineering, 2019, 27(10): 1962−1972.

作者简介:

张雄涛, 副教授, 博士, 主要研究方向为人工智能与模式识别、机器学习。E-mail: 1047897965@qq.com。

陈天宇, 硕士研究生, 主要研究方向为模糊系统、深度学习。E-mail: 2529935825@qq.com。

申情, 教授, 博士, 主要研究方向为智能信息处理、智慧交通。E-mail: sq@zjhu.edu.cn。