

基于决策树的异构特征空间学件组装方法

谭 鹏 周志华

(计算机软件新技术全国重点实验室(南京大学) 南京 210023)

(南京大学人工智能学院 南京 210023)

(tanp@lamda.nju.edu.cn)

Decision Tree-Based Assembly for Learnwares from Heterogeneous Feature Spaces

Tan Peng and Zhou Zhihua

(National Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023)

(School of Artificial Intelligence, Nanjing University, Nanjing 210023)

Abstract Machine learning techniques have been widely applied in various fields but still face numerous challenges, including heavy dependence on large-scale training data, limited adaptability to changing environments, concerns regarding data privacy and ownership, and the issue of catastrophic forgetting. The learnware paradigm provides a systematic solution to these challenges, focusing on reusing existing high-quality models to tackle new tasks instead of training models from scratch. The learnware comprises a model and a specification describing its capabilities. Various learnwares are managed by the learnware dock system. Existing studies have shown the effectiveness of this paradigm when learnwares and user tasks share the same feature space. However, in real-world scenarios, the system often lacks models that precisely match the feature space of user's task. Current work has preliminarily explored the identification and reuse of single learnware in heterogeneous feature spaces. However, the performance of a single heterogeneous learnware is limited and may fall short when addressing the entirety of a user task. While multiple heterogeneous learnwares can complement each other's capabilities, the key challenge lies in identifying and effectively integrating the localized strengths of different heterogeneous learnwares. We propose a decision tree-like learnware assembly algorithm, which effectively integrates the high-confidence regions of multiple heterogeneous learnwares. Furthermore, to facilitate the effective learnware recommendation, we introduce an identification mechanism based on reuse performance estimation. Experiments demonstrate that even in the absence of models that perfectly match the user's task, the proposed method can effectively assist the system in recommending potentially useful heterogeneous learnwares and significantly outperform models trained from scratch through tree-like assembly.

Key words learnware; heterogeneous feature space; model identification; model reuse; decision tree

摘 要 机器学习技术已广泛应用于许多领域,但仍面临诸多挑战,包括对海量训练数据的依赖、难以适应环境变化、数据隐私与所有权问题,以及灾难性遗忘等。学件范式为系统性解决这些问题提供了新思路,其核心在于有组织地复用已有的高质量模型解决新任务,而不必从头训练模型。学件由模型和刻画其能力的规约构成,各式各样的学件通过学件基座系统统一管理。已有研究验证了当学件与用户任务特征空间相同时该范式的有效性,然而在实际场景中,与用户任务特征空间完全匹配的学件往往稀缺,甚至不存在。目前已有研究初步探索了异构特征空间下的单学件查搜与复用,然而单一学件性能有限,难以

收稿日期: 2025-06-09; 修回日期: 2025-12-19

基金项目: 国家自然科学基金项目(62250069)

This work was supported by the National Natural Science Foundation of China (62250069).

通信作者: 周志华(zhouzh@lamda.nju.edu.cn)

在整个用户任务上表现出色。多个异构特征学件在性能上虽有互补潜力,但如何准确识别并整合推荐学件的局部优势区域成为关键挑战。为此,提出一种类决策树结构的多学件组装复用算法,能够有效整合多个来自不同特征空间学件的高置信度区域。为进一步提升学件查搜的准确性,还提出了基于学件复用性能预估的查搜准则。实验结果表明,即使系统中不存在与用户任务完全匹配的模型,系统仍可有效查搜出多个对用户任务有帮助的异构特征学件,并通过树形组装推荐学件显著超越用户本地从头训练模型的性能。

关键词 学件;异构特征空间;模型查搜;模型复用;决策树

中图法分类号 TP181

DOI: 10.7544/issn1000-1239.202550460 **CSTR:** 32373.14.issn1000-1239.202550460

机器学习技术已在医疗^[1]、气象^[2]、材料^[3]等多个领域取得显著成功。然而,其发展仍面临若干关键挑战:首先,从头构建具备专业级性能的模型十分困难,不仅需要大量高质量标注数据和充足的计算资源,还依赖复杂的特征工程与算法设计经验。其次,数据隐私与所有权问题、环境动态变化带来的模型适应挑战,以及灾难性遗忘等现象,限制了模型的分享与复用。尽管已有大量研究分别针对上述问题展开探索,但在实际应用中,这些问题往往同时出现并相互影响。

针对上述机器学习发展面临的若干问题,近期提出的“学件”(learnware)^[4-5]范式提供了系统性的解决方案。该范式旨在通过构建学件基座系统,帮助用户有组织地复用已有的高质量模型解决新任务。学件使得开发者提交的模型能够实现“超越原有用途的复用”,即模型可以被复用到开发者未曾预设的任务场景。学件由模型和规约2部分构成,其中规约描述了模型的专长与适用范围,使其能根据用户任务需求被高效准确地查搜和推荐。作为学件范式流程的核心,规约可通过刻画模型关联的数据分布实现^[5]。为推动学件范式研究,学件基座系统原型“北冥坞”^[6]已正式发布,该系统支持学件范式的主要流程,具备高度可扩展的架构,便于新算法的集成和验证。

先前关于学件范式的研究^[7-9]主要集中在同构场景,即模型与用户任务共享相同的特征空间。然而在现实应用中,即使处理相同任务,由于特征工程的差异,不同模型往往具有异构的特征空间。如图1所示,在医疗场景中,广泛使用的重症监护医疗数据库MIMIC^[10]将数据组织为26张标准表格,涵盖患者信息、ICU住院记录、实验室检测结果、用药记录等。针对疾病表型分类^[11]等典型医疗任务,开发者通常需聚合多张表格以构建预测模型。值得注意的是,由于先验知识不同或者对变量间因果关系建模^[12]存在差异,不同专家通常选择不同的表格组合进行差

异化的特征工程,导致构建模型的特征空间具有显著差异。此外,OMOP通用数据模型^[13]作为医疗数据整合的另一重要标准,同样采用标准化表格来整合多源医疗数据。在此类基于标准化数据库的应用场景中,特征空间异构模型普遍存在。近期,部分研究^[14-16]已开始探索此类异构特征空间模型的组织与复用方法。

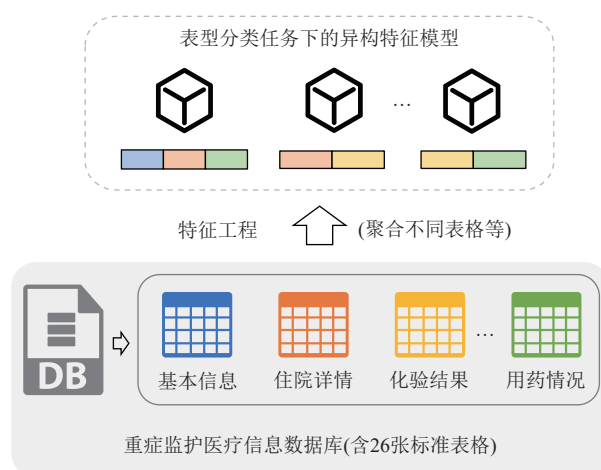


Fig. 1 Models generated from heterogeneous feature spaces in real-world medical scenarios

图1 现实医疗场景中的异构特征空间模型

为利用异构特征空间学件解决用户任务,需依次解决2个递进的核心问题:1)如何从学件基座系统中筛选出潜在有用的异构特征空间学件;2)用户如何高效复用这些学件解决新任务。针对异构特征学件的查搜问题,现有方法主要基于公共子空间的分布匹配进行学件推荐。然而,特征转换过程会导致分布信息失真,仅依赖分布相似度的查搜策略易导致推荐学件性能有限。为此,本文提出应结合学件的模型能力与规约中的分布信息,通过预估其在目标任务上的性能,实现更精准的推荐。优化查搜策略后,用户如何有效复用系统推荐的异构特征学件成为关键挑战。当前研究集中于通过特征空间转

换实现单学件复用,但实际应用中,单一学件的性能可能存在局限,仅在用户任务的部分区域上表现良好,如何根据用户任务识别并组装多个异构特征学件的优势预测区域成为重要问题。本文提出一种类决策树结构的学件组装算法,可自动识别并整合异构特征空间学件的高置信度区域,发挥多学件的性能互补优势。

本文的主要贡献可归纳为3个方面:

1)提出一种多学件树形组装复用算法。该算法通过类决策树结构组合学件的优势预测区域,使各学件仅处理其优势预测区域内的样本,有效提升多学件的复用性能。

2)设计了一种更精准的学件查搜机制。与现有的基于分布距离的查搜机制相比,该机制不仅考虑了任务间的分布相似性,还预估了学件模型在新任务上的预测性能。在此基础上,本文构建了完整的异构特征学件查搜与复用的流程。

3)实验结果表明,即使系统中不存在直接解决用户任务的学件,系统仍可有效查搜和组装多个潜在有用的异构特征学件,帮助用户有效解决自有任务并显著超越用户从头训练模型的性能。

1 相关工作

1.1 学件范式

在学件范式中,来自各种领域的大量机器学习模型可以被开发者提交至学件基座系统。该系统通过规约对学件进行统一组织管理。当用户面临新的机器学习任务时,可通过学件基座系统复用已有的高质量模型,而不必从头开始训练模型。

传统方法如迁移学习^[17-18]、领域自适应^[19-20]虽与学件研究存在一定关联,但其核心关注点存在本质差异,主要体现在2方面:其一,在任务假设方面,传统方法通常假设用于迁移学习的源域任务数量有限且均对目标域任务有帮助,因此直接基于少量源域任务来辅助目标域任务的学习。而在学件范式下,由于学件基座系统的开放特性,系统存储的海量学件中可能仅有少量与用户任务真正相关。因此,如何针对用户需求筛选出有帮助的学件成为该范式的关键挑战。其二,在数据可访问性方面,以往方法通常需要直接获取源域任务的原始训练数据。然而在实际应用中,受数据隐私保护要求和所有权限限制等因素影响,学件基座系统往往无法访问开发者或用户的原始数据。这一限制使得传统方法和技术难以

直接应用于学件范式。

学件范式^[4-5]在隐私保护的前提下提供了一种系统性的模型分享、组织和复用的框架。学件由模型和规约组成,其中规约作为核心组件,描述了模型的特性。在规约设计方面,近期被广泛使用的缩略核均值嵌入(reduced kernel mean embedding, RKME)规约^[5]通过核方法压缩训练数据特征和模型输出的联合分布构建。Lei等人^[21]则证明了该规约在不泄露原始数据的同时,能有效防御常见的推理攻击。Wu等人^[22]实现了基于任务特征数据的简化RKME规约构建方法。为在规约中进一步表征模型能力,Tan等人^[16]改进了RKME规约的生成机制,有效嵌入了模型的输出信息,使得规约能提供有效的压缩模型的输出。针对非表格数据(如图像和文本),谭志豪等人^[23]提出了基于神经切线核的RKME规约。此外,还有一些其他关于规约设计的探索,比如可复用性指标^[24]、LANE^[25]和SpecNet^[26]。在学件查搜复用方面,Wu等人^[21]首先提出通过分布匹配进行学件推荐,Zhang等人^[9]进一步将该方法扩展至用户任务有一部分未被系统容纳学件的任务所覆盖的场景。针对海量学件的高效查搜,Liu等人^[7]提出利用其他学件对规约进行演进以丰富规约对模型的刻画,并构建规约索引提升查搜效率;Xie等人^[8]提出使用少量代表性学件作为锚位学件加速学件查搜,无需遍历学件基座系统内所有学件。

现有研究主要考虑学件与用户任务特征空间一致的同构场景。然而在实际应用中,系统内学件往往有着不同的特征空间。针对这一挑战,Tan等人^[14]首次探索了异构特征空间下的学件组织、查搜和复用问题,但其方法依赖于原始训练数据和跨特征空间辅助数据。为消除这一限制,Tan等人^[15]提出了不依赖原始数据和辅助数据的学件基座系统构建利用方案。随后,Tan等人^[16]改进了规约的生成机制,有效嵌入标记信息,并在异构学件的组织和查搜过程中进行有效利用,进一步提升了用户任务上的性能。

此外,学件范式启发了模型复用^[27]、知识共享^[28]和模型库重用的研究^[29]。为了便于学件范式的研究,首个学件基座系统“北冥坞”^[6]已发布。该系统完整覆盖学件全流程,提供高度可扩展架构,为未来算法实现和实验研究提供了便利。

1.2 异构特征空间相关技术

现有关于异构特征空间的研究如异构领域自适应(heterogeneous domain adaptation)^[30-31]、异构迁移学习(heterogeneous transfer learning)^[32]和异构模型复用

(heterogeneous model reuse)^[33-34], 通常采用将不同特征空间的数据映射至公共子空间的策略。这一映射过程往往依赖于源域和目标域的原始数据, 或跨特征空间的共现数据, 以建立特征之间的关联。然而, 在学件范式框架下, 借助模型关联的 RKME 规约, 学件基座系统能够在无需原始数据和跨特征空间辅助数据的前提下挖掘特征关联, 实现异构特征空间学件的统一组织, 高效查搜和有效复用。值得注意的是, 现有研究主要集中于单一异构学件的复用, 而对多个异构特征空间学件的组装复用仍缺乏研究。

2 预备知识

规约是学件的核心, 能够在不泄露模型原始训练数据的前提下刻画模型的统计性质。本节介绍 RKME 规约^[5], 该规约采用核方法对训练数据特征与模型预测输出的联合概率分布进行压缩表示, 表征了模型的输入输出关系。

2.1 核均值嵌入

我们首先介绍核均值嵌入 (kernel mean embedding, KME)^[35]。KME 将分布映射到一个核函数对应的再生核希尔伯特空间 (reproducing kernel Hilbert space, RKHS)。给定在空间 $\mathcal{X}$ 上的分布 $\mathcal{D}$, 其 KME 定义为

$$\mu_k(\mathcal{D}) := \int_{\mathcal{X}} k(\mathbf{x}, \cdot) d\mathcal{D}(\mathbf{x}),$$

其中 $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 是对称正定核函数, 其对应的 RKHS 为 $\mathcal{H}_k$ 。当核函数 k 符合 $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} [\sqrt{k(\mathbf{x}, \mathbf{x})}] < \infty$, KME 存在且满足 $\mu_P \in \mathcal{H}_k$ 。对于从 $\mathcal{D}$ 采样的数据集 $\{\mathbf{x}_i\}_{i=1}^n$, 其 KME 的经验估计为

$$\hat{\mu}_k(\mathcal{D}) := \frac{1}{n} \sum_{i=1}^n k(\mathbf{x}_i, \cdot)。$$

KME 作为概率分布的表示, 具有优良的理论性质。当核函数为特征核 (characteristic kernel)^[36] 时, KME 能够完整保留原始分布的所有统计信息, 即原始概率分布与其 KME 表示之间存在双射关系。

2.2 RKME 规约

基于 KME, RKME 规约^[5] 提出用少量加权样本 $\{(\beta_j, \mathbf{t}_j)\}_{j=1}^m$ ($m \ll n$) 来近似原始数据特征及模型输出 $\{\mathbf{q}_i\}_{i=1}^n = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^n$ 的经验 KME, 其中 $\hat{y}_i = f(\mathbf{x}_i)$ 为模型预测值。缩略集通过式 (1) 生成:

$$\min_{\beta, \mathbf{t}} \left\| \sum_{i=1}^n \frac{1}{n} k(\mathbf{q}_i, \cdot) - \sum_{j=1}^m \beta_j k(\mathbf{t}_j, \cdot) \right\|_{\mathcal{H}_k}^2, \quad (1)$$

其中 $\{\beta_j\}_{j=1}^m$ 为非负权重系数。由此得到的 KME

$\Phi(\cdot) = \sum_{j=1}^m \beta_j k(\mathbf{t}_j, \cdot) \in \mathcal{H}$ 称为 RKME 规约, 其所在的 RKHS $\mathcal{H}_k$ 称为规约空间。该规约在不泄露原始数据的前提下捕获分布 $\mathcal{D}$ 的主要信息和模型的输入输出关系。

在一些简单的场景下, 比如基座系统中的模型所对应的任务分布仅存在协方差偏移 (covariate shift), 可以仅对模型的训练数据分布进行压缩来生成规约。为了缓解规约在生成过程中标记信息被特征信息淹没的问题, Tan 等人^[16] 提出同时压缩边缘分布和条件分布的规约生成机制。具体而言, 缩略集可以重写为 $\{\beta_j, \mathbf{z}_j, y_j\}_{j=1}^m$, 其中 $\mathbf{z}_j$ 和 y_j 分别为 $\mathbf{t}_j$ 的特征部分和标记部分。缩略集在生成时, 需要确保边际分布 $P(X)$ 和条件分布 $P(X|\hat{Y}=c)$ 都能得到良好近似。该缩略集的求解是通过在边际分布 $P(X)$ 和条件分布 $P(X|\hat{Y}=c)$ 上最小化最大均值差异 (maximum mean discrepancy, MMD) 实现。

$$F(\beta, \mathbf{Z}, \mathbf{Y}) = \left\| \sum_{i=1}^n \frac{1}{n} k(\mathbf{x}_i, \cdot) - \sum_{j=1}^m \beta_j k(\mathbf{z}_j, \cdot) \right\|_{\mathcal{H}_k}^2 + \theta \sum_{c=1}^C \left\| \sum_{i \in \mathcal{I}_c} \frac{1}{n} k(\mathbf{x}_i, \cdot) - \sum_{j \in \mathcal{I}'_c} \beta_j k(\mathbf{z}_j, \cdot) \right\|_{\mathcal{H}_k}^2, \quad (2)$$

其中 $\mathcal{I}_c$ 和 $\mathcal{I}'_c$ 分别表示有标记数据和缩略集中属于第 c 个类的样本索引集合。本文后续内容所用 RKME 规约均采用该生成方式。

3 问题形式化

本文研究在学件基座系统中, 如何查搜对用户任务潜在有用的若干异构特征学件, 并通过多学件组装复用, 有效帮助用户解决任务。

不失一般性, 假设完整的特征空间 $\mathcal{X}_{\text{all}}$ 由 Q 个不同的特征块组合而成, 即 $\mathcal{X}_{\text{all}} = \mathcal{X}_1 \times \mathcal{X}_2 \cdots \times \mathcal{X}_Q$ 。开发者的特征空间 $\mathcal{X}^{\text{dev}}$ 和用户的特征空间 $\mathcal{X}^{\text{user}}$ 为这些特征块的任意组合 $\times_{i \in C} \mathcal{X}_i$, 其中 C 为块的索引。学件任务和用户任务的标记空间相同, 均为 $\mathcal{Y} = \{1, 2, \dots, C\}$ 。

整个流程包括 2 个阶段: 提交阶段和部署阶段。在提交阶段, 开发者在数据集 $D_i := \{(\mathbf{x}_{ij}, y_{ij})\}_{j=1}^{n_i}$ 上训练一个表现良好的模型 f_i , 并生成 RKME 规约 S_i^{dev} , 该规约能够嵌入训练任务的分布信息和模型的输出信息而不暴露原始数据。在接收到所有异构特征空间模型及其规约后, 学件基座系统基于所有提交学件的规约 $\{S_i^{\text{dev}}\}_{i=1}^N$ 学习一个共同的子空间 $\mathcal{X}_{\text{sub}}$, 作为联系不同特征块的桥梁。

在部署阶段,用户拥有未标注数据 $D_0^u = \{\mathbf{x}_{0i}\}_{i=1}^{n_u}$ 和有限数量的标注数据 $D_0^l = \{(\hat{\mathbf{x}}_{0i}, y_{0i})\}_{i=1}^{n_l}$ 。用户生成 RKME 任务需求 S_0^{user} 并提交给学件基座系统。系统识别出潜在有用的多个异构特征空间模型 $\{f_i\}_{i \in I}$ (I 为推荐模型的索引集) 返回给用户。用户通过多学件复用解决任务。问题形式化如图 2 所示。

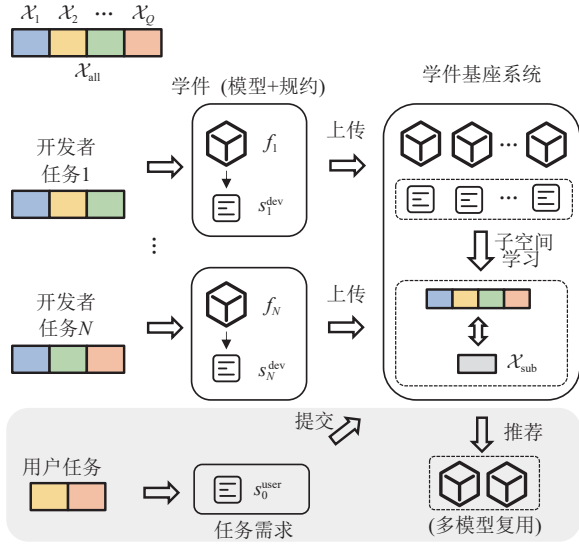


Fig. 2 Problem formulation

图 2 问题形式化

4 本文方法

本节介绍所提出的多学件查搜与复用算法。首先简要介绍系统如何组织异构特征学件;接着提出一种以复用性能预估为核心的查搜准则,该准则相比基于分布匹配的查搜准则,能够缓解因特征空间变换引起的分布信息失真问题,从而提升学件查搜效果;随后,本文提出一种类决策树结构的多学件组装算法,通过树形结构组合多个异构特征学件的高置信度区域,有效提升用户任务性能。

4.1 异构特征空间学件组织

为组织特征空间异构的学件,系统需要挖掘不同特征空间之间的关联。在不接触模型的原始训练数据和跨特征空间的辅助数据的前提下,系统可以利用所有容纳学件的规约挖掘特征关联并构建公共的子空间和相应的映射函数^[16]。

给定规约集合 $\{S_i^{\text{dev}} := \{(\beta_{ij}, z_{ij}, y_{ij})\}_{j=1}^{m_i}\}_{i=1}^N$, 学件基座系统可通过优化目标函数 $L = \alpha_1 L_{\text{reconstruction}} + \alpha_2 L_{\text{similar}} + \alpha_3 L_{\text{supervised}}$, 学习一个公共的子空间 X_{sub} 和相应的编码函数组 $\{h_k: X_k \mapsto X_{\text{sub}}\}_{k=1}^Q$ 及解码函数组 $\{g_k: X_{\text{sub}} \mapsto X_k\}_{k=1}^Q$ 。该目标函数包含 3 个部分: 重构

损失 $L_{\text{reconstruction}}$ 使得映射函数 (h_k, g_k) 实现特征块 X_k 上数据的投影与重构;相似性损失 L_{similar} 使得样本 z_{ij} 在不同特征块上的切片的子空间投影结果相近;监督损失 $L_{\text{supervised}}$ 利用标签信息提升子空间学习效果,使得在公共子空间中同类样本之间的距离小、不同类样本之间的距离大。

得到公共子空间后,系统可以通过映射函数组 $\{h_k, g_k\}_{k=1}^Q$ 将定义在任意特征块组合上的数据投影至子空间,具体而言,该过程通过对不同特征块切片的投影结果进行平均实现。此外,在复用异构特征学件时,映射函数可有效补全学件预测时的缺失特征。该框架具有普适性,可通过多种已有的子空间学习方法实现,如自监督学习^[37-39]、矩阵分解^[40-41]等。

4.2 基于复用性能预估的多学件查搜

学件基座系统实现异构特征学件的组织后,用户可以利用系统有效解决自有任务。用户首先基于未标注数据 $D_0^u = \{\mathbf{x}_{0i}\}_{i=1}^{n_u}$ 和有限数量的标注数据 $D_0^l = \{(\hat{\mathbf{x}}_{0i}, y_{0i})\}_{i=1}^{n_l}$ 生成 RKME 任务需求 $S_0^{\text{user}} = \{(\beta_j, z_j, y_j)\}_{j=1}^{m_u}$, 该缩略集可以通过式(3)生成,即要求生成的缩略集和待预测样本的边缘分布 P_X 相近,且和有标记样本的条件分布 $P_{X|Y}$ 相近^[16]。

$$F(S_0^{\text{user}}) = \left\| \sum_{i=1}^{n_u} \frac{1}{n_u} k(\mathbf{x}_{0i}, \cdot) - \sum_{j=1}^m \beta_j k(\mathbf{z}_j, \cdot) \right\|_{\mathcal{H}_k}^2 + \theta \sum_{c=1}^C \left\| \sum_{i \in \mathcal{I}_c} \frac{1}{n_l} k(\hat{\mathbf{x}}_{0i}, \cdot) - \sum_{j \in \mathcal{I}_c'} \beta_j k(\mathbf{z}_j, \cdot) \right\|_{\mathcal{H}_k}^2, \quad (3)$$

其中 $\mathcal{I}_c$ 和 $\mathcal{I}_c'$ 分别表示有标记数据和缩略集中属于第 c 个类的样本索引集合。

当用户向系统提交任务需求后,系统会评估每个学件的复用性能,并将潜在游泳的多个学件推荐给用户。现有研究^[15-16]通常将学件规约和用户需求投影至子空间,并计算其分布相似度实现学件查搜。然而由于这类方法未能充分考虑学件模型的实际性能,推荐学件复用性能往往有限。

本文提出了一种基于学件预测性能的异构特征学件查搜指标,即以 RKME 用户需求为验证数据,近似评估异构特征学件在用户任务上的表现。具体而言, RKME 用户需求表征为少量带权重的有标记样本点 $\{\beta_m, \mathbf{z}_m, y_m\}_{m=1}^M$, 其中的样本点 $\mathbf{z}_m$ 定义在部分特征块组合构成的特征空间上,基于学件基座系统在子空间学习过程中生成的编码函数组 $\{h_k\}_{k=1}^Q$ 和解码函数组 $\{g_k\}_{k=1}^Q$, 样本点的特征可扩充至完整的特征空间。特征扩充后的用户需求可以作为验证数据评估系统

中所有异构特征学件的性能,该预测性能即为异构特征学件的查搜指标。该评价指标不仅考虑了异构特征空间分布的相似性,更直接评估了学件在特定任务上的表现。

推荐指标的具体计算方式如下:当用户向基座系统提交任务需求 $S_0^{\text{user}} = \{(\beta_{0k}, z_{0k}, y_{0k})\}_{k=1}^{m_0}$ 后,基座系统首先通过编码函数组 $\{h_k: \mathcal{X}_k \mapsto \mathcal{X}_{\text{sub}}\}_{k=1}^Q$ 将特征 z_{0k} (定义在特征空间 $\times_{i \in C_0} \mathcal{X}_i$ 上) 按特征块进行切割并分别利用编码函数投影至子空间,取均值后得到 v_{0k} 。随后,系统通过解码函数组 $\{g_k: \mathcal{X}_{\text{sub}} \mapsto \mathcal{X}_k\}_{k=1}^Q$ 将 v_{0k} 投影到缺失的特征子块上,并将新旧特征拼接成定义在完整特征空间 $\mathcal{X}_{\text{all}} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_Q$ 上的样本 $\hat{x}_{0k}$ 。经过特征转换后,定义在部分特征块上的用户需求可被补齐成完整特征空间上的验证集 $\{(\beta_{0k}, \hat{x}_{0k}, y_{0k})\}_{k=1}^{m_0}$ 。该验证集可用于评估系统中所有异构特征学件的性能。系统中的每个学件可以截取验证集样本 $\hat{x}_{0k}$ 中对应特征空间的切片进行预测,得到模型输出 $\hat{y}_{0k}$ 。异构特征空间学件的复用得分 R 定义为学件在用户需求上的性能:

$$R = \sum_{k=1}^{m_0} \beta_{0k} l(y_{0k}, \hat{y}_{0k}), \quad (4)$$

其中 l 为效用函数,在分类任务中可选择正确率等评价指标。这种查搜指标近似反映了学件在特定用户任务上的性能。学件基座系统计算所有学件的预测性能 R 后,会将指标最高的 L 个学件推荐给用户。

我们将本文提出的查搜机制与传统的基于分布匹配的推荐机制进行对比。基于分布匹配的推荐准则的假设为:当用户任务分布与学件对应的训练任务分布相似时,学件在用户任务上表现良好。这一假设在同构场景下(即学件与用户特征空间一致时)得到了一定验证,但在异构场景下难以适用。以往研究^[15-16]通常通过学习公共子空间来关联不同特征子块,并建立相应的双向映射函数。然而,由于此类映射往往不可逆且伴随信息失真,即使学件规约与用户需求在子空间内的分布相近,也无法保证用户数据经投影与重构后,能与学件原始训练分布保持接近。而基于复用性能估计的方法直接模拟用户数据2次转换后的结果,并评测异构特征学件的性能,有效提升了学件查搜的准确性。在推荐速度方面,两者的差异主要在于:本文方法需要学件模型对少量样本(50个)进行预测,而传统方法则需计算分布距离。由于预测样本量少,2种方法在查搜速度上并无显著差异。

4.3 基于决策树的异构学件组装复用

当用户获得系统推荐的学件后,可通过复用这些学件有效解决自有任务。现有研究主要集中于单一学件的复用——用户仅需利用系统提供的映射函数补全缺失特征块,即可复用异构特征学件。然而,学件基座系统中通常存在多个对用户任务有益的学件,且这些学件在用户任务上有着不同的优势预测区域,仅复用单一学件难以在整个用户任务上表现良好。本文研究如何识别和组合多个学件的优势预测区域,利用多个学件的性能互补优势,有效提升用户任务上多学件的复用效果。

针对上述挑战,本文提出一种类决策树结构的多学件组装算法。该算法利用类似决策树的递归分割机制,在局部空间中动态匹配最优学件。然而,如何根据学件的局部优势信息有效指导特征空间的划分,构成了该方案的关键挑战。尽管基尼系数、信息增益及其比率等经典划分指标能够优化区域样本纯度,却无法识别学件的优势预测区域。为此,本文设计了一种启发式的划分机制:首先利用所有推荐学件对用户的少量有标记样本 $D_0^b = \{(\hat{x}_{0i}, y_{0i})\}_{i=1}^{n_a}$ 进行预测,基于预测置信度构建顺位矩阵;然后挑选最优的2个学件,使样本最优顺位的均值最小化;最后针对最优模型对在样本的顺位比较结果构建2分类数据集,并训练线性模型作为当前区域的划分边界。对于划分后生成的2个子区域,算法会递归地应用相同的划分流程,直至满足以下任一终止条件:1)达到预设的学件树最大深度;2)当前区域内的样本数量低于预设值。算法的具体细节详见算法1。

算法1. 异构学件的树形组装复用。

输入: M 个系统推荐的学件,用户少量有标记数据 $D_0^b = \{(\hat{x}_{0i}, y_{0i})\}_{i=1}^{n_a}$, 用户测试集数据 $D_0^a = \{x_{0i}\}_{i=1}^{n_a}$, 学件基座系统的编码函数组 $\{h_k: \mathcal{X}_k \mapsto \mathcal{X}_{\text{sub}}\}_{k=1}^Q$ 和解码函数组 $\{g_k: \mathcal{X}_{\text{sub}} \mapsto \mathcal{X}_k\}_{k=1}^Q$;

输出: 用户测试集数据预测结果 $\{\hat{y}_{0i}\}_{i=1}^{n_a}$ 。

- ① 特征空间对齐: 使用编码和解码函数组将 D_0^b 的输入特征补全至完整空间 $\mathcal{X}_{\text{all}}$, 得到扩充数据集 $\hat{D}_0^b = \{(\hat{x}_{0i}, y_{0i})\}_{i=1}^{n_a}$ 。
- ② 递归终止判定: 计算 $\hat{D}_0^b$ 上性能最优的学件 f^* ; 若当前如果区域样本数过少 $|\hat{D}_0^b| < N_{\min}$ 或者达到最大深度 $d_{\max}$, 则返回 f^* 。
- ③ 顺位矩阵构建: 各学件在 $\hat{D}_0^b$ 上预测并计算置信度矩阵 $C \in \mathbb{R}^{n_a \times M}$, 其中 C_{ij} 表示学件 j 在第 i 个样本真实标记的预测概率; 然后基于置

信度矩阵生成顺位矩阵 $\mathbf{R} \in \mathbb{R}^{n_l \times M}$, R_{ij} 表示学件 j 在第 i 个样本上的预测顺位。

- ④ 最优模型对搜索: 对于所有学件对 (f_i, f_j) , 计算其互补顺位均值, 即当前学件对在样本

$$\text{上的平均最小顺位 } \bar{R}_{ij} = \frac{1}{n_l} \sum_{k=1}^{n_l} \min(R_{ki}, R_{kj}),$$

选择最优模型对 $(f_a, f_b) = \arg \min_{i,j} \bar{R}_{ij}$ 。

- ⑤ 偏好数据集生成: 根据模型对的预测顺位构建二分类数据集 $D_{\text{pref}} = \{(\hat{\mathbf{x}}_{0i}, \mathbb{I}(R_{ia} \leq R_{ib}))\}_{i=1}^{n_l}$, 即模型 f_a 表现更好, 标记为 1; f_b 表现更好, 则记为 0。

- ⑥ 区域划分分类器训练: 在二分类数据集上训练线性分类器 $f_{\text{split}}: \mathbf{X}_{\text{all}} \mapsto \{0, 1\}$ 于 D_{pref} 。

- ⑦ 刷新最优模型对: 对于划分后的 2 个区域, 刷新 f_a, f_b 为实际划分区域的顺位均值最小的 2 个模型。

- ⑧ 划分性能验证: 单一最优模型性能为

$$P_{\text{single}} = \text{Acc}(f^*, \hat{D}_0), \text{ 组合模型性能为}$$

$$P_{\text{comb}} = \frac{1}{n_l} \sum_{i=1}^{n_l} \text{Acc}(\mathbb{I}(f_{\text{split}}(\hat{\mathbf{x}}_{0i}) = 1) f_a + \mathbb{I}(f_{\text{split}}(\hat{\mathbf{x}}_{0i}) = 0) f_b, y_{0i}).$$

如果 $P_{\text{comb}} \leq P_{\text{single}}$, 撤销划分, 返回 f^* , 且区域划分终止。

- ⑨ 递归子树构建: 分割数据集

$$S_0 = \{(\hat{\mathbf{x}}_{0i}, y_{0i}) \in \hat{D}_0^l \mid f_{\text{split}}(\hat{\mathbf{x}}_{0i}) = 0\},$$

$$S_1 = \{(\hat{\mathbf{x}}_{0i}, y_{0i}) \in \hat{D}_0^l \mid f_{\text{split}}(\hat{\mathbf{x}}_{0i}) = 1\},$$

并在数据集子集 S_0, S_1 执行步骤②~⑧递归生成子树。

- ⑩ 测试集预测: 对于每个测试样本 $\mathbf{x}_{0i} \in D_0^u$, 补全特征得到 $\hat{\mathbf{x}}_{0i}$, 从根节点开始递归选择子树直至叶节点, 基于对应模型 f_{leaf} 生成预测

$$\hat{y}_{0i} = f_{\text{leaf}}(\hat{\mathbf{x}}_{0i}).$$

算法 1 采用类决策树结构实现异构学件的自适组装, 其核心思想是识别并组合不同学件的高置信度区域。算法 1 有 2 处关键步骤: 其一, 基于学件预测置信度构建区域分割边界, 通过搜索最优学件对并训练线性分类器, 实现对特征空间的递归 2 分切割; 其二, 在树的生成过程中进行预剪枝, 每个节点分裂完后进行性能验证, 动态决定是否继续分割——仅当区域划分能提升预测性能时, 才会在该子区域递归构建子树, 否则保留当前最优单个学件作为叶节点。通过这种启发式组装机制, 确保每个学件仅处理其高置信度区域的样本。算法 1 既规避了单个

模型在局部区域表现不佳的问题, 又通过决策树的层级结构实现了模型优势区域的有效组合。相较于传统集成方法, 算法 1 适用于特征分布复杂、模型优势区域存在显著差异的场景。

我们的组装算法具有通用性, 系统端和用户端均可适用。学件组装一般由系统执行, 其依赖用户需求(少量带权有标记样本)提供的监督信息进行树形组装, 用户可直接基于树形组装学件预测数据或者结合自有数据进行进一步微调。用户也可基于推荐学件自行执行树形组装算法, 相比于前者可接触原始任务数据, 但组装时仅能使用系统推荐的学件。

4.4 完整流程

在学件基座系统的构建阶段, 开发者在自有任务上训练模型, 生成 RKME 规约。开发者将模型和规约打包成学件上传至系统。系统在接收多个异构特征空间的学件后, 会基于所有学件规约学习一个公共子空间, 以联系不同的特征子块, 并生成相应的双向映射函数。系统将上传学件的规约更新至统一嵌入空间, 实现异构特征学件的组织。

当用户需要利用学件基座系统解决特定任务时, 需基于无标记测试数据和少量有标记数据生成 RKME 任务需求, 并提交至系统。系统通过复用性能预估的查搜准则, 向用户推荐若干潜在可用的异构特征空间学件并帮助用户进行树形组装复用。最终, 用户可借助系统推荐的异构特征学件高效解决新任务。

5 实验结果与分析

5.1 实验设置

1) 数据集

本研究选用 15 个公开的分类任务数据集^[42]构建实验场景, 评估所提方法的性能。所选数据集具有 3 个特点: ①样本量跨度大(1 000~58 310 个样本); ②特征空间维度差异显著(7~7 200 维); ③类别数量覆盖广泛(2~10 类)。完整的 15 个数据集统计信息详见表 1。

2) 对比方法

我们将提出的方法与 3 类代表性方法进行系统对比, 包括从头训练方法、预训练模型方法和模型库复用方法。对于传统的从头训练方法, 我们选用表格任务最常用的梯度提升树算法, 选用 LightGBM^[43]进行测试, 其在大规模数据上具有训练速度快、内存占用低的优势。对于预训练模型方法, TabPFN v2^[44-45]为当前小规模数据集上的 SOTA 深度学习方法, 通过

Table 1 Basic Information of Datasets

表 1 数据集基本信息

数据集	类别数	特征数	样本数
credit-g	2	20	1 000
mfeat-karhunen	10	64	2 000
gina-agnostic	2	970	3 468
bioresponse	2	1 776	3 751
christine	2	1 636	5 418
theorem-proving	6	51	6 118
satimage	6	36	6 430
fabert	7	800	8 237
robert	10	7 200	10 000
artificial-characters	10	7	10 218
eye-movements	3	27	10 936
riccardo	2	4 296	20 000
nomao	2	118	34 465
click-prediction-small	2	11	39 948
volkert	10	180	58 310

人造数据集预训练先验拟合网络并在下游任务上通过一次前向传播实现预测;而 XTab^[46]在真实异构表格数据集上预训练公共主干网络并下游复用。对于模型库方法复用方法,一种基准策略是先直接获取全部模型进行异构迁移学习后选择最优。Align^[6]首先在特征空间中进行模型对齐,随后利用对齐后的模型生成预测结果;利用模型输出增广原始特征后,训练一个线性模型拟合增广后数据集。这种方法需要对齐所有模型的特征空间,计算复杂度过高,相比之下, Hetero^[16]通过训练一个公共的子空间和分布匹配实现单学件推荐,并通过子空间学习时产生的特征空间转化函数帮助用户进行便捷的异构特征空间模型复用,这种方法显著提升了异构特征空间学件查搜复用效率。

为了验证本文推荐机制的有效性,我们将 Hetero 的推荐机制进行替换,记为本文方法(单推荐)。而为了验证本文提出的树形组装算法的有效性,本文将流程中的多模型复用算法替换为以下 4 种集成算法进行对比: Average Ensemble^[47]直接平均所有推荐模型的预测结果, Best Validation^[47]基于少量用户标注数据筛选最优的 50% 模型进行平均集成, Selective Ensemble^[48]采用零阶优化通过同时优化模型数、验证集正确率、验证集上预测结果的间隔(margin)搜索最优模型组合, Job Selector^[15,21]通过训练一个学件选择器来判断每个测试集样本用什么学件进行预测。

3) 实验设置

所有数据集均按 4 : 1 比例划分为训练集与测试集^[42]。特征空间被随机划分为 4 个维度相近的特征块,其中开发者任务的特征空间通过任意 3 个特征块的组合构建 4 种特征空间,用户任务的特征空间则通过任意 2 个特征块的组合生成 6 种特征空间。规约缩略集的大小设置为 50,学件基座系统与子空间相关的编码器、解码器配置及优化目标、训练方式和 Hetero^[16]论文中实验设定保持一致。针对用户任务,我们从训练集中分层采样抽取 100 个标注样本。开发者与用户均使用搜参后的 LightGBM 进行模型训练。在进行学件查搜时,系统推荐 3 个模型给用户,树形组装在用户端进行。每种特征空间下的用户任务均进行 5 次重复实验,每次采样 100 个有标记样本,单个数据集上总计 30 次实验。

5.2 用户任务性能

表 2 展示了本文方法与对比方法的性能对比结果。实验结果表明,本文方法在绝大多数测试场景下均展现出显著优势。具体而言: 1) 在从头训练模型类方法中, LightGBM 作为表格数据模型训练的主流方法,其性能受限于从头开始的训练方式,在标注数据有限时表现欠佳。2) TabPFN V2 通过利用海量合成数据预训练先验网络,能够仅需 1 次前向传播即可完成下游任务的预测。实验表明,在标注样本数量有限的情况下,该方法性能优于 LightGBM。然而,由于该方法未考虑跨任务知识复用,其预测性能与本文方法相比仍存在显著差距。相比之下, XTab 尝试构建统一主干网络并在下游任务上微调,但由于表格数据具有显著的语义复杂性和分布异质性,其性能反而低于传统从头训练模型的算法 LightGBM。此外,该方法需要基于真实场景的原始数据进行训练,会暴露任务原始数据。与此不同,本文方法无需接触开发者任务和用户任务的原始数据,从而避免了数据泄露。3) 在模型库方法中, Hetero 通过训练公共子空间并结合分布匹配实现单学件推荐,已取得显著性能提升。然而,由于推荐机制存在缺陷,当我们将推荐机制替换为本文提出的基于复用性能预估的机制,即本文方法(单推荐),性能可获得进一步提升。需要指出的是,单个学件的复用性能仍存在局限。为此,本文提出多学件查搜复用方法,通过融合多个异构特征空间学件的局部高置信度预测结果,最终实现最优性能。

为验证树形组装方案的有效性,我们在表 3 中对比了多种学件集成方法,结果表明,本文提出的树

Table 2 Compared with Baseline Methods on Accuracy

表 2 与基线方法的准确率对比

%

数据集	LightGBM	TabPFN V2	XTab	Align	Hetero	本文方法（单推荐）	本文方法
credit-g	67.7±1.0	71.5±0.0	70.3±0.6	69.1±1.6	70.4±0.5	70.6±0.7	71.8±0.1
mfeat-karhunen	65.4±1.2	75.9±16.3	36.0±1.9	73.3±12.0	77.0±2.0	79.9±0.1	80.2±0.0
gina-agnostic	77.2±1.8	79.5±0.5	67.8±2.3	74.5±1.0	89.1±0.3	89.1±0.9	89.9±1.6
bioresponse	62.9±1.9	64.5±6.7	56.4±2.0	64.0±1.6	74.4±0.9	75.9±0.5	74.4±0.5
christine	62.0±1.4	62.2±0.5	61.8±1.2	66.9±0.8	72.7±0.1	72.7±0.1	72.7±0.1
theorem-proving	39.9±1.7	42.1±0.3	40.8±0.2	39.5±1.2	51.3±0.0	50.4±1.6	52.8±0.3
satimage	79.6±1.3	84.0±0.4	75.4±2.3	81.4±0.9	87.5±0.1	86.7±0.6	87.5±0.1
fabert	23.4±0.0	31.2±4.0	22.9±0.5	32.0±5.6	45.0±1.6	54.6±0.2	54.5±1.0
robert	21.9±0.6	22.7±0.4	13.4±1.3	24.6±0.8	41.7±2.0	45.9±0.1	45.6±0.1
artificial-characters	26.7±1.0	30.4±7.1	21.1±1.7	20.2±2.7	41.4±1.2	42.6±2.9	48.1±0.3
eye-movements	41.5±1.2	42.8±2.8	39.2±1.5	39.7±1.6	56.2±0.0	53.2±1.3	55.4±0.6
riccardo	74.4±1.9	75.9±0.1	75.0±0.0	72.3±2.0	82.1±0.4	82.8±0.2	81.7±0.1
nomao	85.4±1.4	87.4±2.4	82.2±1.0	86.7±3.4	90.8±0.2	90.9±0.4	90.9±0.5
click-prediction-small	79.5±1.0	83.2±0.0	82.6±0.4	83.0±0.4	83.2±0.0	83.2±0.0	83.2±0.0
volkert	39.8±0.8	42.8±1.8	37.4±1.0	42.1±2.1	46.2±0.0	49.9±1.4	51.2±0.3
平均性能	56.5	59.7	52.2	58.0	67.3	68.6	69.3
平均排名	5.9	3.9	6.5	5.4	2.3	2.1	1.9
赢/平/输	15/0/0	14/1/0	15/0/0	15/0/0	9/4/2	8/3/4	

注：“赢/平/输”表示本文方法超过/持平/弱于单个对比方法的次数；黑体值表示单个数据集上的最优结果；“±”前后数值表示准确率的均值和标准差。

Table 3 Accuracy of Different Ensemble Methods

表 3 不同集成方法的准确率

%

集成方法	准确率均值
Tree Assemble（本文）	69.33
Best Validation	69.00
Selective Ensemble	68.77
Average Ensemble	66.26
Job Selector	64.70

形组装(Tree Assemble)算法有着最优的性能。具体而言: 1)Average Ensemble 采用简单平均策略, 为所有学件分配相同权重。导致性能较弱学件的错误预测干扰性能较强学件的正确分类, 且不考虑学件在不同样本上的预测质量差异, 导致性能不佳; 2)Best Validation 和 Selective Ensemble 通过不同策略筛选推荐模型, 性能优于 Average Ensemble, 但仍采用平均集成方式, 未能有效挖掘模型的局部预测优势; 3)Job Selector 虽然为基于区域的组装方法, 但因未能识别学件的优势预测区域, 其性能最差且远不及 Tree Assemble。相比之下, Tree Assemble 通过自适应识别并组合不同模型的优势预测区域, 有效提升了多学件查搜复用效果。

5.3 不同用户数据量下的性能评估

如 5.2 节的用户任务实验所示, 在标注数据有限时, 复用异构特征空间学件的性能明显优于从头训练模型。本节进一步分析, 当用户采用不同规模的标注数据训练模型, 并将其预测结果与学件预测结果进行集成时, 用户任务性能的变化趋势。如图 3 所示, 当标注数据量为 100 时, 集成的性能显著优于用户本地从头训练模型; 即使标注数据量增加到 500, 集成性能在近 90% 的任务中依然保持优势; 当标注数据量达到 5 000 时, 将学件与模型模型进行集成仍能提升约 33% 任务的性能。综上, 实验结果表明, 在不同规模的标注数据下, 将学件预测结果集成到用户任务预测中能够持续带来稳定的性能提升, 且这一性能优势随着用户有标记数据量的增加虽然有所减弱, 但依然保持显著。

5.4 树形组装复用

为更直观地展示用户任务的类决策树的区域划分结果, 我们采用主成分分析将用户高维特征划分映射到 2 维平面进行可视化。基于 15 个数据集共计 450 次用户任务实验, 我们从中选取了若干具有代表性的划分结果进行展示, 如图 4 所示。典型的划分模式如图 4(a)所示, 当启发式区域分割仅需 1 次

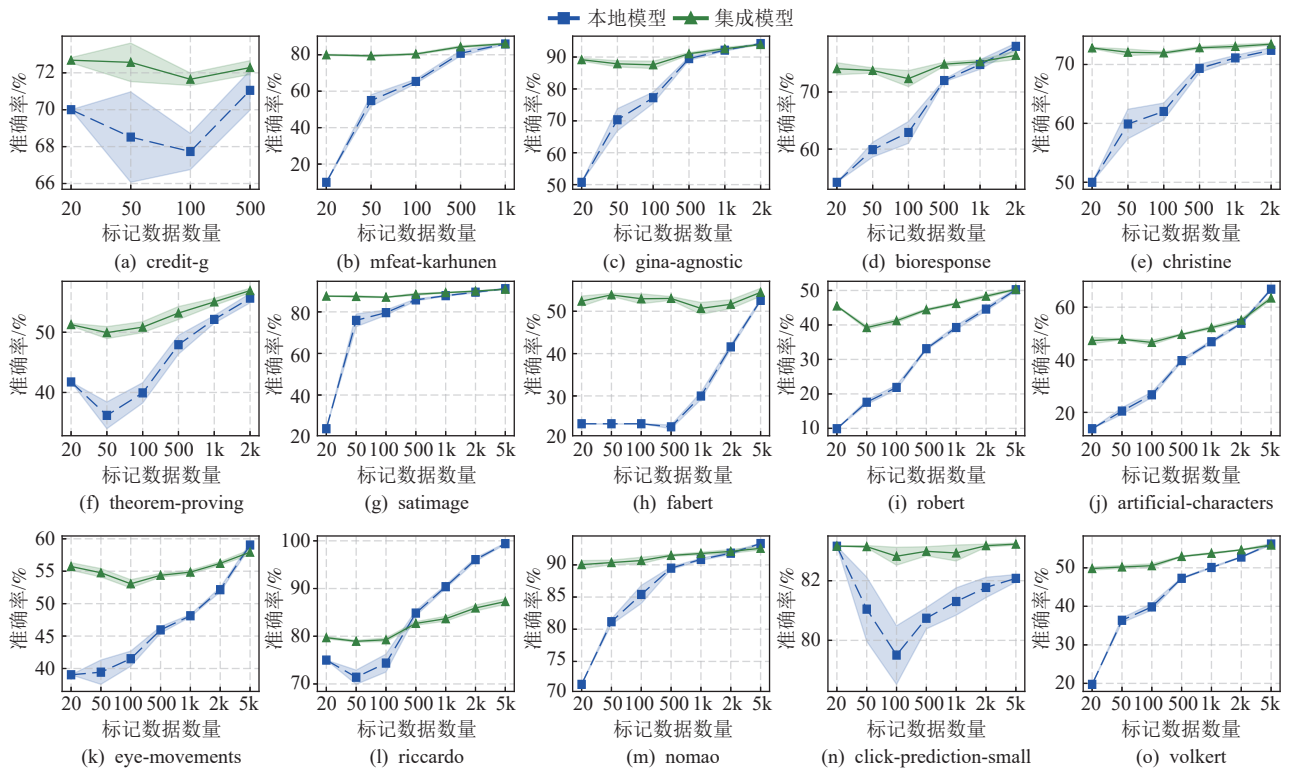


Fig. 3 The curve of user task performance varies with the amount of labeled data

图3 用户任务性能随着标记数据数量变化的曲线

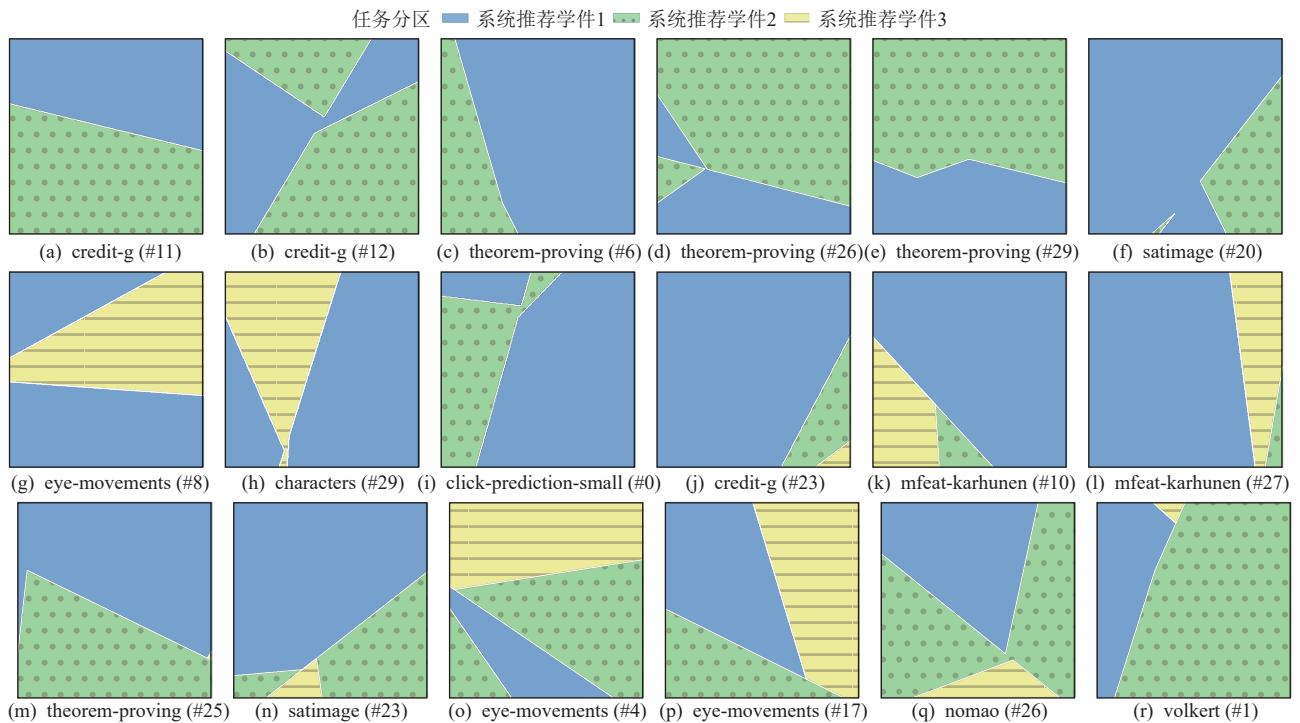


Fig. 4 Regional division diagram of tree-structured learnware assembly

图4 学件树形组装的区域划分图

即可达到最优性能时,生成的2个子区域无法通过划分进一步提升性能。相反,如果子区域仍有优化空间,则会出现如图4(b)~(i)所示的多级划分结果。

在上述情形下,树形组装算法仅从推荐模型中选择2个最优模型进行组合。而当系统推荐的所有模型的优势预测区域均不太一样时,则会出现如图4

(j)~(r)所示的复杂分割情况。在用户任务自适应划分出的4个区域中,每个区域对应着某个学件的高置信度预测区间。这种基于异构特征空间学件的自适应高置信度区域组合策略,能够显著提升学件复用性能。

6 总 结

学件范式帮助用户有组织地复用已有高质量模型解决新任务,而不必从头开始训练模型。本文提出了一种多学件树形组装复用算法,该算法通过类决策树结构对特征空间进行递归划分,使各学件仅需处理其优势预测区域的样本,有效整合了推荐学件的局部优势,有效提升了多学件的复用性能。此外,本文设计了一种更精准的学件推荐机制,通过预估异构特征学件在目标任务上的复用性能进行学件查搜。基于上述方法,本文构建了完整的异构特征学件查搜与复用流程。实验结果表明,即使学件基座系统中不存在专门针对用户任务的模型,系统仍能查搜出对用户任务具有显著帮助的异构特征空间学件,并帮助用户通过树形组装快速构建新模型,有效解决自有机器学习任务。

作者贡献声明: 谭鹏负责算法设计、实验验证以及论文撰写;周志华在研究问题、求解思路和论文写作上给予指导。

参 考 文 献

- [1] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. *Nature*, 2021, 596(7873): 583–589
- [2] Kochkov D, Yuval J, Langmore I, et al. Neural general circulation models for weather and climate[J]. *Nature*, 2024, 632(8027): 1060–1066
- [3] Merchant A, Batzner S, Schoenholz S S, et al. Scaling deep learning for materials discovery[J]. *Nature*, 2023, 624(7990): 80–85
- [4] Zhou Zhihua. Learnware: On the future of machine learning[J]. *Frontiers of Computer Science*, 2016, 10(4): 589–590
- [5] Zhou Zhihua, Tan Zhihao. Learnware: Small models do big[J]. *Science China Information Sciences*, 2024, 67(1): 112102
- [6] Tan Zhihao, Liu Jiandong, Bi Xiaodong, et al. Beimingwu: A learnware dock system [C]//Proc of the 30th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2024: 5773–5782
- [7] Liu Jiandong, Tan Zhihao, Zhou Zhihua. Towards making learnware specification and market evolvable [C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024, 38(12): 13909–13917
- [8] Xie Yi, Tan Zhihao, Jiang Yuan, et al. Identifying helpful learnwares without examining the whole market [C]//Proc of the 26th European Conf on Artificial Intelligence. Ohmsha: IOS, 2023: 2752–2759
- [9] Zhang Yujie, Yan Yuhu, Zhao Peng, et al. Towards enabling learnware to handle unseen jobs [C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 10964–10972
- [10] Johnson Alistair E W, Pollard Tom J, Shen Lu, et al. MIMIC-III, a freely accessible critical care database[J]. *Scientific Data*, 2016, 3(1): 1–9
- [11] Harutyunyan H, Khachatrian H, Kale D C, et al. Multitask learning and benchmarking with clinical time series data[J]. *Scientific Data*, 2019, 6(1): 96
- [12] Wang Tianzuo, Zhou Zhihua. Active causal effect identification with expert knowledge[J]. *SCIENTIA SINICA Informationis*, 2023, 53(12): 2341–2354 (in Chinese)
(王天佑, 周志华. 基于专家知识的主动因果效应辨识[J]. *中国科学: 信息科学*, 2023, 53(12): 2341–2354)
- [13] Biedermann P, Ong R, Davydov A, et al. Standardizing registry data to the OMOP common data model: Experience from three pulmonary hypertension databases[J]. *BMC Medical Research Methodology*, 2021, 21(1): 238
- [14] Tan Peng, Tan Zhihao, Jiang Yuan, et al. Towards enabling learnware to handle heterogeneous feature spaces[J]. *Machine Learning*, 2024, 113(4): 1839–1860
- [15] Tan Peng, Tan Zhihao, Jiang Yuan, et al. Handling learnwares developed from heterogeneous feature spaces without auxiliary data [C]// Proc of the 32nd Int Joint Conf on Artificial Intelligence. New York: Curran Associates, 2023: 4235–4243
- [16] Tan Peng, Liu Haitian, Tan Zhihao, et al. Handling learnwares from heterogeneous feature spaces with explicit label exploitation [C]//Advances in Neural Information Processing Systems 37. Cambridge, MA: MIT, 2024: 12767–12795
- [17] Pan S J, Yang Qiang. A survey on transfer learning[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(10): 1345–1359
- [18] Zhuang Fuzhen, Qi Zhiyuan, Duan Keyu, et al. A comprehensive survey on transfer learning[J]. *Proceedings of the IEEE*, 2020, 109(1): 43–76
- [19] Mansour Y, Mohri M, Rostamizadeh A. Domain adaptation: Learning bounds and algorithms [C]//Proc of the 22nd Annual Conf on Learning Theory. New York: PMLR, 2009: 20–35
- [20] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation [C]//Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 1180–1189
- [21] Lei Haoyi, Tan Zhihao, Zhou Zhihua. On the ability of developers' training data preservation of learnware [C]//Advances in Neural Information Processing Systems 37. Cambridge, MA: MIT, 2024: 36471–36513
- [22] Wu Xizhu, Xu Wenkai, Liu Song, et al. Model reuse with reduced kernel mean embedding specification[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2021, 35(1): 699–710

- [23] Tan Zhihao, Shi Haoyu, Chen Zixuan, et al. Learnware reduced kernel mean embedding specification based on neural tangent kernel[J]. *Chinese Journal of Computers*, 2024, 47(6): 1232–1243 (in Chinese) (谭志豪, 史浩宇, 陈梓轩, 等. 基于神经切线核的学件 RKME 规约[J]. *计算机学报*, 2024, 47(6): 1232–1243)
- [24] Ding Yaoxiang, Zhou Zhihua. Boosting-based reliable model reuse [C]// Proc of the 12th Asian Conf on Machine Learning. New York: PMLR, 2020: 145–160
- [25] Chen Wei, Mao Junxiang, Zhang Minling. Learnware specification via label-aware neural embedding [C]// Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 15857–15865
- [26] Shen Zhiyu, Li Ming. Identifying useful learnwares via learnable specification[J]. *Frontiers of Computer Science*, 2025, 19(9): 199344
- [27] Wu Xizhu, Liu Song, Zhou Zhihua. Heterogeneous model reuse via optimizing multiparty multiclass margin [C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 6840–6849
- [28] Mohammed T, Naas S A, Sigg S, et al. Knowledge sharing in AI services: A market-based approach[J]. *IEEE Internet of Things Journal*, 2022, 10(2): 1320–1331
- [29] Gong Lina, Zhang Junxuan, Wei Mingqiang, et al. What is the intended usage context of this model? An exploratory study of pre-trained models on various model repositories[J]. *ACM Transactions on Software Engineering and Methodology*, 2023, 32(3): 1–57
- [30] Duan Lixing, Xu Dong, Tsang I W. Learning with augmented features for heterogeneous domain adaptation [C]//Proc of the 29th Int Conf on Machine Learning. New York: ACM, 2012: 667–674
- [31] Wang Chang, Mahadevan S. Heterogeneous domain adaptation using manifold alignment [C]// Proc of the 22nd Int Joint Conf on Artificial Intelligence. New York: Curran Associates, 2011: 1541–1546
- [32] Day O, Khoshgoftaar T M. A survey on heterogeneous transfer learning[J]. *Journal of Big Data*, 2017, 4(1): 29
- [33] Ye Hanjia, Zhan Dechuang, Jiang Yuan, et al. Rectify heterogeneous models with semantic mapping [C]// Proc of the 35th Int Conf on Machine Learning. New York: ACM, 2018: 1904–1913
- [34] Ye Hanjia, Zhan Dechuang, Jiang Yuan, et al. Heterogeneous few-shot model rectification with semantic mapping[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 43(11): 3878–3891
- [35] Schölkopf B, Smola A J. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond[M]. Cambridge, MA: MIT Press, 2002
- [36] Fukumizu K, Gretton A, Sun Xiaohai, et al. Kernel measures of conditional dependence [C]// Advances in Neural Information Processing Systems 20. Cambridge, MA: MIT, 2008: 489–496
- [37] Uçar T, Hajiramezanali E, Edwards L. SubTab: Subsetting features of tabular data for self-supervised representation learning [C]// Advances in Neural Information Processing Systems 34. Cambridge, MA: MIT, 2021: 18853–18865
- [38] Bahri D, Jiang H, Tay Y, et al. Scarf: Self-supervised contrastive learning using random feature corruption [C/OL]//Proc of the 9th Int Conf on Learning Representations. 2022 [2022-04-25]. https://openreview.net/forum?id=CuV_qYkmKb3
- [39] Jiao Pengfei, Liu Huan, Lü Le, et al. Globally enhanced heterogeneous temporal graph neural networks based on contrastive learning[J]. *Journal of Computer Research and Development*, 2023, 60(8): 1808–1821 (in Chinese) (焦鹏飞, 刘欢, 吕乐, 等. 基于对比学习的全局增强动态异质图神经网络[J]. *计算机研究与发展*, 2023, 60(8): 1808–1821)
- [40] Xu Wei, Gong Yihong. Document clustering by concept factorization [C]//Proc of the 27th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2004: 202–209
- [41] Wang Hao, Yang Yan, Li Tianrui. Multi-view clustering via concept factorization with local manifold regularization [C]//Proc of the 16th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2016: 1245–1250
- [42] McElfresh D, Khandagale S, Valverde J, et al. When do neural nets outperform boosted trees on tabular data? [C]// Advances in Neural Information Processing Systems 36. Cambridge, MA: MIT, 2023: 76336–76369
- [43] Ke Guolin, Meng Qi, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree [C]// Advances in Neural Information Processing Systems 30. Cambridge, MA: MIT, 2017: 3149–3157
- [44] Hollmann N, Müller S, Eggensperger K, et al. TabPFN: A transformer that solves small tabular classification problems in a second [C/OL]//Proc of the 11th Int Conf on Learning Representations. 2023 [2023-05-01]. <https://openreview.net/forum?id=cp5PvcI6w8>
- [45] Hollmann N, Müller S, Purucker L, et al. Accurate predictions on small data with a tabular foundation model[J]. *Nature*, 2025, 637(8045): 319–326
- [46] Zhu Bingzhao, Shi Xingjiang, Erickson N, et al. XTab: Cross-table pretraining for tabular transformers [C]//Proc of the 40th Int Conf on Machine Learning. New York: ACM, 2023: 43181–43204
- [47] Zhou Zhihua. Ensemble Methods: Foundations and Algorithms [M]. London: Chapman & Hall, 2021
- [48] Wu Yuchang, He Yixiao, Qian Chao, et al. Multi-objective evolutionary ensemble pruning guided by margin distribution [C]//Proc of the 17th Int Conf on Parallel Problem Solving from Nature. Berlin: Springer, 2022: 427–441



Tan Peng, born in 1997. PhD. His main research interests include learnware and machine learning.

谭鹏, 1997年生。博士。主要研究方向为学件、机器学习。



Zhou Zhihua, born in 1973. PhD, professor, PhD supervisor. Member of the Chinese Academy of Sciences. Fellow of ACM, AAAI, AAAS, IEEE, IAPR, IET/IEE, CCF, CAAI and CSIAM. His main research interests include artificial intelligence, machine learning, and data mining.

周志华, 1973年生。博士, 教授, 博士生导师。中国科学院院士。ACM, AAAI, AAAS, IEEE, IAPR, IET/IEE, CCF, CAAI, CSIAM 会士。主要研究方向为人工智能、机器学习、数据挖掘。