中南大学研究生创新项目
（自主探索类）

申　报　书

项 目 名 称： 基于数据包络分析的选择性集成学习方法及应用研究

所 属 学 科： 管理科学与工程

项目关键词： 集成学习；数据包络分析；效率评价

二、立项依据

（项目的研究目的、意义；国内外研究现状分析和发展趋势；项目应用前景和学术价值；现有研究基础、条件、手段等）

1 研究目的及意义

人工智能技术的快速发展为当前人类社会带来了巨大变革。习近平总书记指出，“把新一代人工智能作为推动科技跨越发展、产业优化升级、生产力整体跃升的驱动力量，努力实现高质量发展”，将人工智能的发展上升到国家战略层面[1]。机器学习作为人工智能的共性基础与关键技术，近年来，各种机器学习理论与算法都得到阶跃式的发展与革新。其中，集成学习（Ensemble Learning）由于其出色的泛化能力和预测准确度，成为了机器学习领域的热点，成功应用于智能交通中的车辆检测，图像和视频处理中的目标识别，生物信息学中的基因组功能预测等多个领域[2, 3]。其优秀的预测性能源于集成学习利用多个学习器获得比仅使用单一学习器更强的泛化能力的思想。但学习器的数量并非越多越好，一是可能造成过量的储存与计算开销，二是纳入性能较差的学习器可能导致预测准确度下降。基于此，2002 年 Zhou 等[4]首先提出“选择性集成（Selective Ensemble）”方法，其理论分析和实验结果均表明，仅选择部分准确率高、差异大的学习器进行集成能够取得比一般集成学习技术更优的泛化性能。随着国内外研究的发展，选择性集成技术已经成为集成学习领域中效果最好的学习范式[5-7]。**然而，现有选择性集成方法的选择准则大多沿用 Zhou 的思想，即以准确度和多样性为依据，鲜有考虑模型的训练与使用成本。**

随着人工智能算力的爆炸式增长，模型训练与使用的高能耗问题日益突出。Strubell 等[8]研究结果表明，包含约 1.1 亿参数的 BERT-base 模型[9]需要在 64 个 NVIDIA Tesla V100 GPU 上训练约 79.2 个小时，其训练过程的能耗约排放 1,400 磅二氧化碳，大致相当于一次横跨美洲的飞行。此外，为量化研发一个新模型需投入的计算资源，Strubell 还对其取得 EMNLP 2018（自然语言处理顶级会议）最佳论文的研究，基于语言信息的自注意力（Linguistically Informed Self-Attention，LISA）模型[10]的训练过程进行案例分析。结果表明，在预训练阶段，模型能耗成本与模型的大小呈正比，而在根据应用场景对模型参数和结构进行微调（fine-tune）以保证预测精度时，**能耗成本呈指数级增长，性能收益却微乎其微。**这仅仅是训练与使用模型，若需要完整地实施机器学习项目并取得良好的效果，除算力相关的成本外，还需考虑数据获取与标注成本，算法开发、部署与维护成本等一系列费用[11]。根据美国市场研究机构 Dimensional Research 对来自 20 个不同行业的数据科学家或人工智能专家的调查结果表明[12]，企业每年需为每个机器学习项目支付的费用约为 51,750 美元至 136,750 美元，这对于个人、小型团队或初创企业而言可能是难以承受的。

我国已进入人工智能与行业发展深度融合的阶段，泛化能力强、应用范围广的集成学习技术无疑成为促进我国政府与企业数智化转型、推动我国经济高质量发展的首选方案。

然而，开发与实施集成学习项目需要投入的数据资源、计算资源与能耗成本尚未得到广泛关注。**因此，如何以尽量少的投入取得尽量优的性能，进而构建整体高效的集成学习技术，成为选择性集成学习领域亟需解决的问题。**

为解决上述问题，本项目拟从学习器的选择与学习器的组合两个方面改进现有的选择性集成学习方法。一方面，项目将基于减少计算资源、提高预测精度等多个目标**构建全面、客观的评价指标体系，并采用数据包络分析（Data Envelopment Analysis，DEA）方法[13]对学习器的效率进行评价**，作为选择性集成学习方法的选择依据。另一方面，项目将充分利用学习器的评价结果，**设计高效、合理的组合策略，在多源、异构、动态海量数据的应用场景下，保证集成学习器的整体效率。**

在解决“如何选择”以及“如何组合”这两个关键问题后，“如何应用”成为选择性集成方法研究的必要拓展。虽然集成学习最早是为了解决有监督分类预测问题所提出的，但随着国内外研究的进展，集成学习方法逐渐与其他机器学习范式结合起来。因此，本项目拟**在改进选择性集成学习方法的基础上，分别探讨其如何在三种常见的机器学习的范式，即有监督学习、半监督学习和无监督学习的情形下进行拓展**，并提出相应的算法框架，扩大选择性集成方法的适用范围，以解决不同的现实问题。

综上所述，面向我国发展新一代人工智能、推动经济高质量发展的重大国家战略需求，本项目聚焦集成学习方法，首先，充分考虑计算资源投入与算法性能等多个目标，结合 DEA 方法，从选择与组合两个角度完善选择性集成学习方法；其次，结合不同机器学习范式的特点将选择性集成学习方法进行拓展，并提出有监督集成、半监督集成与无监督集成算法框架；最后，以湖南省财信金融大数据集为依托，对约 50 万家湖南省中小企业风险智能决策问题进行应用研究，**力求理论联系实践，充分为我国发展新一代人工智能提供科学有效的理论参考和方法支撑。**

2 国内外研究现状分析和发展趋势

围绕本项目主题，以下将从选择性集成学习方法、DEA 及 DEA 与集成学习方法的结合、不同机器学习范式下的集成学习方法三个方面对国内外研究现状进行总结。

2.1 选择性集成学习方法研究现状

选择性集成学习的主要思想是基于某种评价准则，从已有的学习器中选择一些用于构建集成学习器，这个过程可细分为选择和组合两个步骤[5]。为便于区分，本项目将已生成的基础学习器称为基学习器，基学习器经过选择与组合后最终得到的学习器称为集成学习器。现有的选择性集成方法主要在选择步骤上存在差异，即通过使用不同的评价准则或评价方法来选择表现相对较好的学习器，而组合步骤普遍采用平均法、投票法、加权投票法等。以下将从基学习器的选择与基学习器的组合两个方面进行文献回顾。

2.1.1 基学习器的选择

基学习器的选择方法可以分为三类。一是基于类簇的选择，这类方法通过一定的准则将相似的基学习器划分为同一类簇，并从中选择具有代表性的基学习器。如 Giacinto 等[14]通过基学习器同时预测错误的概率定义基学习器间的距离，采用层次聚类方法将其划分至不同类簇，再选择每个类簇中距离其他类簇平均距离最远的基学习器作为代表。二是基于排序的选择，这类方法按照评价标准对基学习器的性能进行排序，仅选择排序靠前的基学习器。如 Martinez-Munoz 等[15]总结了可用于对基学习器排序的指标，大致包括准确度、误差、互补性、边缘距离等。三是基于优化的选择，这类方法旨在以最大化或最小化与集成学习器泛化能力相关的目标找到基学习器的子集。如 Aksela 和 Laaksonen[16]定义了指数误差数（Exponential Error Count, EEC）用于度量基学习器误差之间的多样性，通过最小化 EEC 来选择最优的基分类器子集。可以发现，目前大多选择性集成方法所考虑的评价准则都可以划分为准确度和多样性两大类，**忽略了计算资源、能耗成本等在发展新一代人工智能时代决策者难以忽视的现实因素。**

2.1.2 基学习器的组合

常见的组合方法有平均法、投票法等[2]。其中，平均法是最基础的组合方法，它将基学习器的输出结果的算术平均值作为最终预测结果。投票法可分为多数投票法、加权投票法、软投票法等，多数投票法将超过半数的基学习器的预测结果作为最终预测结果，加权投票法在此基础上为不同基学习器赋予不同权重，软投票法则考虑了不同基学习器的置信度。这些方法各有优劣，正如“天下没有免费的午餐”，也没有任何一种组合方法适用于所有的情况和数据集[17]。然而，对于一组给定的基学习器集合，任何组合方法都可以看作对基学习器的预测结果赋予权重的过程。尽管实现权重分配的方式有所不同，但它们内在的逻辑都是相同的：权重代表了基学习器的性能表现（performance），基学习器的表现越好，对集成学习器预测结果的贡献就应该越大，在组合中所占的权重就应该越高。**如何准确地评价基学习器的性能表现，进而合理地分配基学习器的组合权重，是选择性集成学习方法的研究重点。**

此外，基学习器的性能表现不是一直不变的。在动态的大规模数据背景下，待预测的样本可能随时间的推移出现“概念漂移”的现象[18, 19]，即模型预测的目标变量的统计特性随时间以不可预见的方式发生变化。在此情况下，若集成学习模型仍保持不变，预测性能会明显下降。动态选择性集成方法为每个待预测的样本选择最合适的基分类器子集进行组合预测[20]，能够在一定程度上能够克服性能下降的问题，但其计算成本相较于静态选择性集成方法大幅上升，在实际应用中很难推广。**如何设计更加高效的动态组合策略，自适应地为基学习器分配组合权重，也是选择性集成学习领域待解决的难题。**

【小结】根据文献回顾可知，目前选择性集成方法的评价准则较为单一，大多只关注基学习器的准确度与多样性，未考虑资源投入等现实因素，且现有的组合方法难以适用于复杂的、动态的、大规模数据下的应用场景。因此，面向发展新一代人工智能的战略需求，**亟需提出更加全面的评价体系与更加高效的组合策略用于改进选择性集成学习方法。**

2.2 DEA 及 DEA 与集成学习方法的结合研究现状

DEA 是由 Charnes，Cooper 和 Rhodes 于 1978 年提出的用于评价一组具有多个投入产出指标的决策单元（Decision-Making Unit，DMU）相对效率的非参数方法。相较于其它评价方法，DEA 具有较强的客观性，仅通过各决策单元的观测数据判断 DMU 是否有效。根据投入和产出的径向或非径向、不同类型的规模回报假设、生产可能集的凸性或非凸性，DEA 衍生出多种评价模型，如由 Charnes 等[21]提出的基于规模收益不变的 CCR 模型，Banker 等[13]提出的基于规模收益可变的 BCC 模型，基于松弛变量的 SBM 模型[22]，超效率模型[23]，交叉效率模型[23]、Free Disposal Hull（FDH）模型[24]等。目前 DEA 已被广泛应用到各个领域，如环境[25-27]、金融[28, 29]、管理[30-32]、教育[33-35]等现实效率评价问题中。从 Web of Science 核心期刊库中检索 2003 至 2022 年以 DEA 为主题的相关文献共 25468 篇，近 20 年来发表篇数统计如图 1 所示，这些统计数据能够反映出 DEA 的研究热度逐年上升。其中，DEA 与机器学习领域的交叉研究约 200 篇，包括与 SVM、决策树、神经网络等不同机器学习方法的结合[36-38]。

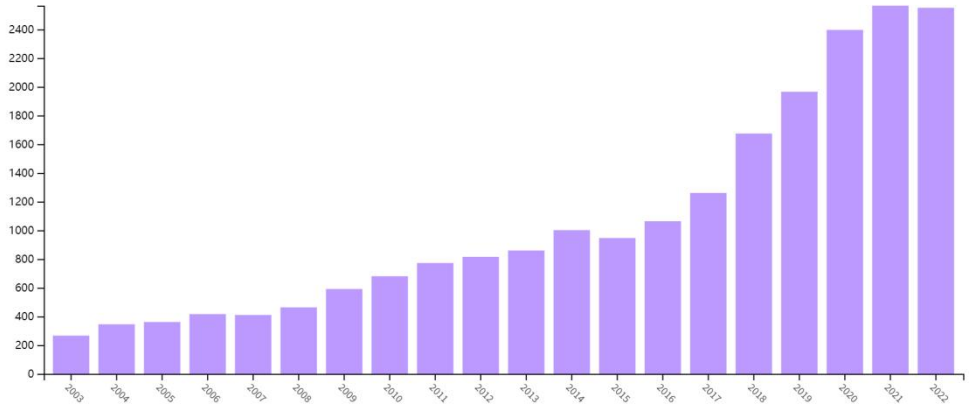

图 1 2003 至 2022 年 DEA 相关文献刊发统计（源自 Web of Science 核心检索）

本项目聚焦 DEA 与集成学习方法的结合研究。通过文献回顾发现，多数研究将 DEA 与集成学习预测用于方法的不同阶段，如 Thaker 等[39]先通过 DEA 测算出效率，再将其作为随机森林的训练样本中的变量进行预测，或 Anouze 等[40]利用随机森林进行特征选择后，再使用 DEA 计算效率。**仅有少数研究使用 DEA 指导集成学习器的构建过程，以提高模型的预测性能。**Zheng 等[41]使用 BCC 模型构建的生产前沿面代替机器学习中的 ROC 曲线，用于评价基学习器的预测性能并对基学习器进行组合。进一步地，Zhu[42]将该思想结合集成

学习中 Stacking 技术构建集成学习器，并证明了方法的有效性。最近，Chao 等[43]引入 BCC 模型评价不平衡分类器的预测效果，具体而言，该研究通过 DEA 效率曲线指导不平衡分类器在训练时，如何权衡提高少数类准确率的收益与降低多数类准确率的成本，并通过实证研究发现集成学习方法对于处理不平衡分类问题具有显著优势。在上述研究中，DEA 不仅可以判断出基学习器是否有效，还可以根据评价结果对基学习器进行组合，即通过效率刻画基学习器的性能表现，并基于此为其赋予权重。然而，**上述研究中使用的 BCC 模型以自评效率为基础，可能在评价时夸大自身的优势，导致产生表面有效的现象**。此外，**BCC 模型假设生产可能集为凸集，这个假设在评价集成学习背景下的基学习器时可能难以满足**。为更好地反映 DMU 之间效率的相对关系，可以采用交叉效率评价、FDH 模型等方法评价基学习器的效率，并以此为依据使基学习器组合权重的分配更加客观合理，充分达到通过 DEA 评价方法指导集成学习器选择与组合的目的。

【小结】根据文献回顾可知，近年来，DEA 在理论与应用方面均得到了较为充分的研究，而鲜有 DEA 与集成学习领域的交叉研究。仅有的文献中使用的 DEA 模型都从自评效率的角度入手，且都基于凸性假设，适用性有限。因此，在已有研究的基础上，为保证集成学习中基学习器选择与组合结果的合理性，**亟需根据实际应用场景与评价指标体系，提出更加合适的 DEA 评价方法用于指导集成学习器的构建**。

2.3 不同机器学习范式下集成学习方法研究现状

机器学习按照对于数据样本的学习形式进行分类，主要可分为监督学习、无监督学习、半监督学习，其区别在于监督学习需要提供标注的样本集，无监督学习不需要提供标注的样本集，半监督学习需要提供少量标注的样本。虽然集成学习方法最早是为解决有监督分类问题而提出的，但随着国内外研究的发展，也逐渐与其余学习范式结合起来。以下从有监督集成分类、半监督集成和无监督集成三个方面进行文献回顾。

2.3.1 有监督集成方法

集成学习方法首先在监督学习领域被引入并获得巨大成功。常见的集成学习方法主要可分为 Bagging、Boosting 和 Stacking[3, 44, 45]。Bagging 是最早提出的集成学习方法之一，该方法通过有放回地从数据集中采样，形成不同的子集对基学习器进行训练，基学习器训练完成后通过平均法、投票法、加权投票法等方式进行集成，其优势是通过并行训练加快集成模型的速度。随机森林[46]是 Bagging 方法的重要的进改形式，它不仅使用 Bagging 方法抽取样本，还使用随机特征子空间对样本特征进行随机选择，对多个决策树进行集成进而构造泛化能力强的集成学习器。Boosting 是一类增强集成学习方法，从初始训练集中训练一个基模型，对于预测错误的样本，增加其权重后再训练下一个基模型。重复上述过程，直至基模型的数量达到上限，最后再将所有基模型进行加权集成。Adaboost[47]算法是 Boosting

的典型代表，随后国内外学者又在此基础上提出了 XGBoost[48]、LightGBM[49]等算法，对集成学习的发展也产生了深远影响。Stacking 方法一般由两层预测模型结构组成，首先将原始数据集划分成若干子数据集，输入到第一层预测模型的各个基学习器中，再将每个基学习器输出作为第二层元学习器的输入，由元学习器输出最终预测结果。该方法一般用于对不同的模型组件（异质基学习器）进行集成，适用于处理多源异构数据[50]。

2.3.2 半监督集成方法

半监督学习是指学习器不依赖外界交互、自动地利用未标记样本和已标记样本来提升学习性能的学习方式[51, 52]。这种学习方式非常适用于已标记样本稀少、获取困难，人工标注费时费力，但未标记样本却充足易得的客观现实条件。半监督集成学习则是一种同时利用半监督学习与集成学习的新学习方式，主要研究如何将半监督学习与集成学习有机结合。2002 年 Bennett 和 Demiriz[53]提出的 ASSEMBLE 方法首次使用了“半监督集成（Semi-supervised Ensemble）”的概念。随后，Zhou[54]以基于分歧的半监督方法为例，从理论和实践角度论证了集成学习和半监督学习之间的互益性。对集成学习而言，可通过利用**半监督学习引入无标记样本来解决集成学习有标记样本量不足的问题，且半监督学习还可以增加集成学习中学习器的多样性**；对于半监督学习而言，集成学习能进一步降低半监督学习的泛化误差，且使得半监督学习的收敛速度更快[52]。从直观上看，将半监督学习和集成学习结合的动机是合理且可行的。随着国内外研究的不断发展，现有研究可根据集成学习的类别大致分为半监督 Boosting、半监督 Bagging、半监督随机子空间三类，其中代表性方法如表 1 所示。

表 1 半监督集成学习代表性方法

类别	代表方法	描述
半监督 Boosting	ASSEMBLE[53]	利用无标记样本最大化 Boosting 分类间隔。
	SemiBoost[55]	通过 Boosting 方式划分并集成多个半监督分类器。
	SemiBoost-CR[56]	在 SemiBoost 基础上，同时选取置信度较高和置信度较低的未标注样本，以对训练集进行扰动，提高集成学习器多样性。
半监督 Bagging	Semi-Bagging[57]	通过 Bagging 方式划分并集成多个半监督分类器。
半监督随机子空间	Extends RF to SSL[58]	在随机森林中引入无标记样本提升多样性。
	SSC-RSDR[59]	利用随机子空间方法对基于图的半监督方法进行降维，最终进行集成输出。

各类方法存在较大的差异，需要根据应用场景的特点选择合适的半监督集成方法才能发挥其优势与作用。此外，现有研究通常使用无标记样本提升基学习器的多样性，但**缺乏关于无标记样本对于多样性提升效果的条件与边界的讨论，以及对于多样性如何影响集成学习器准确性的理论分析和实验证明**。

2.3.3 无监督集成方法

本项目主要关注无监督学习中的聚类任务。聚类的目的是通过将未标记数据分组为类簇来找到它们的固有结构，在文本挖掘、模式识别、图像分割等领域具有广泛的应用[60, 61]。然而，不同的算法从同一数据集中发现的分布结构可能有所不同，具有不同参数设置的单一聚类算法在相同数据集上得到的分布也可能有所不同，在没有监督信息的情况下，很难判断哪种结构与真实分布最匹配。为解决这个问题，Strehl 等[62]首次提出“聚类集成（Cluster Ensembles）”方法，它将多个聚类结果组合成最终聚类，而无需访问数据的原始特征或算法。与单一聚类算法相比，聚类集成的显著优势是能够明显提高聚类结果的质量和鲁棒性。该方法主要分为两个步骤，生成多个不同的聚类结果以及将它们集成为一个更好的聚类结果。这在一定程度上与集成分类相似，但其难度远高于集成分类任务。一方面，缺乏标签使模型评估相当困难，一般只能使用类簇多样性相关的指标；另一方面，由于基聚类的簇标签是象征性的，且**不同聚类方法的簇标签含义不同，无法和有监督集成一样通过简单的平均或投票完成，需要更复杂的集成策略**。近十年来有大量学者为聚类集成方法的改进和拓展作出大量贡献，现有研究主要可分为基于关联矩阵的方法、基于图划分的方法和基于重标记的方法，其中代表性方法如表 2 所示。

表 2 无监督集成学习代表性方法

类别	代表方法	描述
基于关联矩阵的方法	EAC[63]	将每个基聚类看作一个独立证据，构建基聚类的协关联矩阵，并通过多种层次聚类算法获得集成聚类结果。
	LWCA[64]	考虑类簇的可靠性和集成的局部多样性，将局部自适应性纳入协关联矩阵。
基于图划分的方法	DICLENS[65]	结合簇间相似性和簇内相似性构建无向加权图，基于此创建基于相似性的最小生成树。
	ECPCSMC[66]	将类簇视为图节点，使用 Jaccard 系数计算加权边构建聚类相似度图并定义转移概率矩阵，进行随机游走以探索图中的多尺度结构信息。
基于重标记的方法	SWV[67]	先计算类簇重叠的对象数来对齐标签，通过归一化互信息对基聚类进行加权投票组合。

上述方法都能够有效优化聚类的结果，但仍然存在一定问题。例如，基于关联矩阵的方法和基于图划分的集成聚类一般需要计算类簇或样本间的两两关系，**计算开销较大**；基于重标记的方法虽然便于计算，但在标签对齐的过程中将**不可避免地造成信息损失**。

【小结】根据文献回顾可知，不同的机器学习范式与集成学习结合有不同的侧重点。与研究较为成熟的有监督集成分类方法相比，无监督集成方法、半监督集成方法都尚有很大的研究空间。因此，为适用于不同的应用场景，在改进现有选择性集成方法的基础上，**亟需充分结合各学习范式的特点取长补短，进一步降低计算开销、提升预测性能**。

3 项目应用前景和学术价值

本项目在我国发展新一代人工智能的重大战略背景下，围绕选择性集成学习方法及其应用研究这一主题，拟展开以下四项研究内容：基学习器的评价体系研究；基学习器的组合策略研究；选择性集成方法在不同机器学习研究范式中的拓展；选择性集成学习方法在湖南省财信数据集中的应用，具有较好的应用前景和学术价值，具体说明如下。

3.1 应用前景

随着我国人工智能技术的发展，模型算力需求的逐步提升，机器学习模型的能耗成本与日俱增。考虑到机器学习模型训练与使用的巨大能耗成本，本研究将计算资源投入作为必要考量因素，在构建机器学习算法时，考虑数据本身的性质，力争以尽量少的资源投入达到尽量优的预测效果，进而构建整体高效的集成学习方法，将有助于我国政府与企业在开发机器学习项目、实施数智能化转型时降本增效，对我国发展新一代人工智能，推动经济高质量发展起到积极作用。

3.2 学术价值

针对以上应用前景，本项目将从以下三方面开展理论研究：基学习器的评价体系研究、基学习器的组合策略研究、选择性集成方法在不同机器学习研究范式中的拓展研究，学术价值体现在以下方面。

首先，**完善现有的选择性集成学习方法。**目前选择性集成的选择准则较为单一，大多只关注基学习器的准确度与多样性，并未综合考虑计算资源投入等现实因素，且难以适用于复杂的、动态的、大规模数据下的应用场景。项目将通过构建更加全面的评价指标体系、设计更加高效的组合策略构建整体高效的集成学习器。

其次，**深化 DEA 与集成学习领域的交叉研究。**在现有研究中，鲜有 DEA 与集成学习的交叉研究。为保证集成学习中基学习器选择与组合权重分配的合理性，本项目基于集成学习的背景构建新的基学习器的生产可能集，并结合交叉效率评价、FDH 等方法提出更加合适的基学习器效率评价方法，充分结合 DEA 的理论优势提升集成学习方法的性能。

最后，**在不同机器学习范式下对选择性集成学习方法进行拓展。**与研究较为成熟的有监督集成分类方法相比，无监督集成方法、半监督集成方法、强化集成方法都尚有很大的研究空间。本项目将在改进选择性集成方法的基础上，结合各学习范式的特点取长补短，提出相应的算法框架，进一步优化学习器的性能。

4 现有研究基础、条件、手段

4.1 研究基础

首先，项目申请人及团队成员具有良好基础知识储备，学习 DEA 方法、机器学习理论与方法、评价理论与方法等相关知识已有近两年的时间，**已对本项目密切相关的问题进行**

研究并完成相关论文 2 篇，其中 1 篇论文对非凸假设下大规模样本的效率评价与更新方法进行研究，已投稿至《系统工程理论与实践》在审，1 篇论文结合 DEA 与机器学习提出了一种新的集成分类方法，已投稿至 European Journal of Operational Research 在审，上述**研究成果与本项目研究内容高度匹配**。

其次，项目申请人具备与研究内容相关的充足的技能与经验。对于本项目的理论研究部分，申请人曾获得**第十九届中国研究生数学建模竞赛全国二等奖**、**2020 年 MathorCup 数学建模挑战赛全国一等奖**，且熟练掌握 Matlab、python 等工具以及 Scikit-learn、TensorFlow 等常用机器学习库，**在数学建模与算法设计上具备扎实的基础**；对于本项目应用研究部分，申请人于 2023 年 4 月至 7 月在湖南财信数字科技有限公司业务研发部实习，期间参与**完成了湖南省旅游、医疗等 5 个重点行业非法集资风险智能决策模型的开发与优化工作**，已对研究数据集的样本特征、分布等信息具备充分的了解。

再次，项目申请人参与导师国家自然科学基金面上项目 2 项（环境大数据背景下交互网络系统的环境绩效评价方法及其应用研究；面向共享经济的网络 DEA 合作博弈方法及应用研究），**申请人在参与导师项目的过程中，储备了与大规模样本效率评价、机器学习的相关知识，为本项目的完成打下了坚实的基础**。

最后，项目的研究成员均为在校研究生，有充足的学习时间，具备较好的学习和研究能力，对 DEA 方法、数据分析、机器学习等的具有良好的掌握和应用；项目研究团队的指导老师长期从事该领域的研究，是该领域非常杰出的研究人员；申请人及其他研究人员将在导师的监督和指导下认真完成该项目的研究。

以上都为本项目的顺利实施提供了良好的研究基础。

4.2 研究条件

本项目的执行主要依托于中南大学商学院。学院“管理科学与工程”为一级学科国家重点学科，学院科研教学条件优越，设施齐备，藏书 6 万多册，中外期刊 300 多种；学院拥有 28 个为科研和教学服务的实验室与研究中心，如决策理论与支持平台研究中心、工程管理和项目评价研究中心等，为项目的理论及应用提供了充分的硬件基础。特别地，学院拥有湖南省首个国家基础科学中心“数字经济时代的资源环境管理理论与应用”，本项目所有成员均来自上述实验室。此外，本项目所属研究领域常在全国各地举办学术交流会议，和国内外杰出学者的交流也将有助于本项目的顺利开展，推进项目顺利完成。

4.3 研究手段

本项目的研究主要是以数据分析、数学建模和算法设计为主，数学建模主要涉及 DEA 方法，多目标优化方法等，算法设计主要涉及集成学习框架、分类算法、聚类算法等，数据处理以及算法开发主要依赖于 Matlab、Python 等工具。

5 参考文献

[1] 闫海潮, 胡金旭. 习近平关于人工智能时代意义及实践路径重要论述研究 [J]. 党的文献, 2019, (03): 45-49+71.

[2] Zhou Z. Ensemble methods: Foundations and algorithms [M]. CRC press, 2012.

[3] Sagi O, Rokach L. Ensemble learning: A survey [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1249.

[4] Zhou Z, Wu J, Tang W. Ensembling neural networks: Many could be better than all [J]. Artificial Intelligence, 2002, 137(1-2): 239-263.

[5] 张春霞, 张讲社. 选择性集成学习算法综述 [J]. 计算机学报, 2011, 34(08): 1399-1410.

[6] Wang J, Qian Y, Li F, et al. Generalization performance of pure accuracy and its application in selective ensemble learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(2): 1798-1816.

[7] Lv S, Peng L, Hu H, et al. Effective machine learning model combination based on selective ensemble strategy for time series forecasting [J]. Information Sciences, 2022, 612: 994-1023.

[8] Strubell E, Ganesh A, Mccallum A. Energy and policy considerations for modern deep learning research [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(09): 13693-13696.

[9] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:181004805, 2018.

[10] Strubell E, Verga P, Andor D, et al. Linguistically-informed self-attention for semantic role labeling [C]. Brussels:Association for Computational Linguistics, 2018.

[11] 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.

[12] Research D. Artificial intelligence and machine learning projects are obstructed by data issues [R]. USA: Alegion, 2019.

[13] Banker R D, Charnes A, Cooper W W. Some models for estimating technical and scale inefficiencies in data envelopment analysis [J]. Management Science, 1984, 30(9): 1078-1092.

[14] Giacinto G, Roli F. An approach to the automatic design of multiple classifier systems [J]. Pattern Recognition Letters, 2001, 22(1): 25-33.

[15] Martínez-Muñoz G, Hernández-Lobato D, Suárez A. An analysis of ensemble pruning techniques based on ordered aggregation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 245-259.

[16] Aksela M, Laaksonen J. Using diversity of errors for selecting members of a committee classifier [J]. Pattern Recognition, 2006, 39(4): 608-623.

[17] Wolpert D H, Macready W G. No free lunch theorems for optimization [J]. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 67-82.

[18] Guo H, Zhang S, Wang W. Selective ensemble-based online adaptive deep neural networks for streaming data with concept drift [J]. Neural Networks, 2021, 142: 437-456.

[19] Krawczyk B, Minku L L, Gama J, et al. Ensemble learning for data stream analysis: A survey [J]. Information Fusion, 2017, 37: 132-156.

[20] Ko A H R, Sabourin R, Britto J a S. From dynamic classifier selection to dynamic ensemble selection [J]. Pattern Recognition, 2008, 41(5): 1718-1731.

[21] Charnes A, Cooper W W, Rhodes E. Measuring the efficiency of decision making units [J]. European Journal of Operational Research, 1978, 2(6): 429-444.

[22] Tone K. A slacks-based measure of efficiency in data envelopment analysis [J]. European Journal of Operational Research, 2001, 130(3): 498-509.

[23] Andersen P, Petersen N C. A procedure for ranking efficient units in data envelopment analysis [J]. Management Science, 1993, 39(10): 1261-1264.

[24] Tulkens H. On FDH efficiency analysis: Some methodological issues and applications to retail banking, courts, and urban transit [J]. Journal of Productivity Analysis, 1993, 4(1-2): 183-210.

[25] Zhou X, Luo R, An Q, et al. Water resource environmental carrying capacity-based reward and penalty mechanism: A DEA benchmarking approach [J]. Journal of Cleaner Production, 2019, 229: 1294-1306.

[26] Sueyoshi T, Yuan Y, Goto M. A literature study for DEA applied to energy and environment [J]. Energy Economics, 2017, 62: 104-124.

[27] Zhou H, Yang Y, Chen Y, et al. Data envelopment analysis application in sustainability: The origins, development and future directions [J]. European Journal of Operational Research, 2018, 264(1): 1-16.

[28] Ding L, Lei L, Wang L, et al. A novel cooperative game network DEA model for marine circular economy performance evaluation of china [J]. Journal of Cleaner Production, 2020, 253: 120071.

[29] Fukuyama H, Matousek R. Modelling bank performance: A network DEA approach [J]. European Journal of Operational Research, 2017, 259(2): 721-732.

[30] Ruiz J L, Sirvent I. Performance evaluation through dea benchmarking adjusted to goals [J]. Omega, 2019, 87: 150-157.

[31] Dobos I, Vörösmarty G. Inventory-related costs in green supplier selection problems with data envelopment analysis [J]. International Journal of Production Economics, 2019, 209: 374-380.

[32] An Q, Tao X, Xiong B. Benchmarking with data envelopment analysis: An agency perspective [J]. Omega, 2021, 101: 102235.

[33] Chen L, Wang Y-M. Dea target setting approach within the cross efficiency framework [J]. Omega, 2020, 96: 102072.

[34] Yang G, Fukuyama H, Song Y. Measuring the inefficiency of chinese research universities based on a two-stage network DEA model [J]. Journal of Informetrics, 2018, 12(1): 10-30.

[35] An Q, Yang M, Chu J, et al. Efficiency evaluation of an interactive system by data envelopment analysis approach [J]. Computers & Industrial Engineering, 2017, 103: 17-25.

[36] Esteve M, Aparicio J, Rodriguez-Sala J J, et al. Random forests and the measurement of super-efficiency in the context of free disposal hull [J]. European Journal of Operational Research, 2023, 304(2): 729-744.

[37] Valero-Carreras D, Aparicio J, Guerrero N M. Support vector frontiers: A new approach for estimating production functions through support vector machines [J]. Omega, 2021, 104: 102490.

[38] Zhong K, Wang Y, Pei J, et al. Super efficiency sbm-dea and neural network for performance evaluation [J]. Information Processing & Management, 2021, 58(6): 102728.

[39] Thaker K, Charles V, Pant A, et al. A dea and random forest regression approach to studying bank efficiency and corporate governance [J]. Journal of the Operational Research Society, 2022, 73(6): 1258-1277.

[40] Anouze A L, Bou-Hamad I. Inefficiency source tracking: Evidence from data envelopment analysis and random forests [J]. Annals of Operations Research, 2021, 306(1): 273-293.

[41] Zheng Z, Padmanabhan B. Constructing ensembles from data envelopment analysis [J]. INFORMS Journal on Computing, 2007, 19(4): 486-496.

[42] Zhu D. A hybrid approach for efficient ensembles [J]. Decision Support Systems, 2010, 48(3): 480-487.

[43] Chao X, Kou G, Peng Y, et al. An efficiency curve for evaluating imbalanced classifiers considering intrinsic data characteristics: Experimental analysis [J]. Information Sciences, 2022, 608: 1131-1156.

[44] Bühlmann P. Bagging, boosting and ensemble methods [J]. Handbook of Computational Statistics: Concepts and Methods, 2012: 985-1022.

[45] Wolpert D H. Stacked generalization [J]. Neural Networks, 1992, 5(2): 241-259.

[46] Altman N, Krzywinski M. Ensemble methods: Bagging and random forests [J]. Nature Methods, 2017, 14(10): 933-935.

[47] 曹莹, 苗启广, 刘家辰, 等. Adaboost 算法研究进展与展望 [J]. 自动化学报, 2013, 39(06): 745-758.

[48] Chen T, Guestrin C. XGBoost: A scalable tree boosting system[C]. Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining, 2016.

[49] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree [J]. Advances in Neural Information Processing Systems, 2017, 30.

[50] 史佳琪, 张建华. 基于多模型融合stacking集成学习方式的负荷预测方法 [J]. 中国电机工程学报, 2019, 39(14): 4032-4042.

[51] 屠恩美, 杨杰. 半监督学习理论及其研究进展概述 [J]. 上海交通大学学报, 2018, 52(10): 1280-1291.

[52] 蔡毅, 朱秀芳, 孙章丽, 等. 半监督集成学习综述 [J]. 计算机科学, 2017, 44(S1): 7-13.

[53] Bennett K P, Demiriz A, Maclin R. Exploiting unlabeled data in ensemble methods [C]. New York: Association for Computing Machinery, 2002.

[54] Zhou Z. When semi-supervised learning meets ensemble learning [J]. Frontiers of Electrical and Electronic Engineering in China, 2011, 6: 6-16.

[55] Mallapragada P K, Jin R, Jain A K, et al. Semiboost: Boosting for semi-supervised learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(11): 2000-2014.

[56] 唐焕玲, 鲁明羽. 利用置信度重取样的 semiboost-CR 分类模型 [J]. 计算机科学与探索, 2011, 5(11): 1048.

[57] Li Y, Su L, Chen J, et al. Semi-supervised question classification based on ensemble learning[C]. proceedings of the Advances in Swarm and Computational Intelligence, 2015.

[58] Leistner C, Saffari A, Santner J, et al. Semi-supervised random forests [C]. In proceedings of the IEEE 12th International Conference on Computer Vision, 2009.

[59] Liu X, Song M, Tao D, et al. Random forest construction with robust semisupervised node splitting [J]. IEEE Transactions on Image Processing, 2014, 24(1): 471-483.

[60] Ezugwu A E, Ikotun A M, Oyelade O O, et al. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects [J]. Engineering Applications of Artificial Intelligence, 2022, 110: 104743.

[61] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究 [J]. 软件学报, 2008, (01): 48-61.

[62] Strehl A, Ghosh J. Cluster ensembles---a knowledge reuse framework for combining multiple partitions [J]. Journal of Machine Learning Research, 2002, 3(11): 583-617.

[63] Wang X, Yang C, Zhou J. Clustering aggregation by probability accumulation [J]. Pattern Recognition, 2009, 42(5): 668-675.

[64] Huang D, Wang C-D, Lai J-H. Locally weighted ensemble clustering [J]. IEEE Transactions on Cybernetics, 2017, 48(5): 1460-1473.

[65] Mimaroglu S, Aksehirli E. Diclens: Divisive clustering ensemble with automatic cluster number [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 9(2): 408-420.

[66] Huang D, Wang C-D, Peng H, et al. Enhanced ensemble clustering via fast propagation of cluster-wise similarities [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 51(1): 508-520.

[67] Zhang M. Weighted clustering ensemble: A review [J]. Pattern Recognition, 2022, 124: 108428.

三、研究方案

1 研究目标、研究内容和拟解决的关键问题

1.1 研究目标

本项目面向我国发展新一代人工智能的重大战略需求，围绕选择性集成学习方法及其应用这一主题开展研究，具体可以分为以下几个目标：

（1）**构建全面、客观的选择性集成方法的基学习器评价体系**。针对现有选择性集成方法未考虑计算资源投入的问题，构建更加全面的评价指标体系，使用更加客观的评价方法用于选择基学习器，并将评价的结果作为学习器组合的基础。

（2）**设计高效、合理的选择性集成方法的基学习器组合策略**。针对多源、异构、动态的海量数据，结合基学习器的评价结果，从静态集成和动态集成两个角度设计更加高效、合理的基学习器组合策略，从而实现集成学习方法的整体有效。

（3）**在不同机器学习范式下对选择性集成学习方法进行拓展**。在完善选择性集成理论的基础上，分别针对有监督集成学习、半监督集成学习、无监督集成学习提出相应的机器学习算法框架，充分结合各类机器学习范式的特点，达到优化学习器性能的目的。

（4）基于理论研究成果，**以湖南省财信金融大数据集为依托，开展湖南省中小企业风险评价应用研究**。在检验理论研究正确有效的同时，为促进我国数字科技公司低成本高质量实施机器学习项目、发展新一代人工智能起到积极作用。

上述四个研究子目标中，前三个子目标是基础性目标，承载了本项目的主要学术价值；第四个子目标是应用性目标，是助推实现国家发展新一代人工智能战略需求的主要体现。

1.2 研究内容

为实现上述研究目标，本项目拟展开以下四个部分的研究：基学习器的评价体系研究；基学习器的组合策略研究；选择性集成学习方法在不同机器学习范式中的拓展研究；选择性集成学习方法在湖南省财信数据集中的应用研究。整个项目由理论研究与应用研究两个部分组成，研究内容总体框架如图 2 所示。

1.2.1 学习器的评价体系研究

基于提高预测准确度的同时降低算法能耗成本的现实需求，需要对基学习器的效率进行客观合理的评价，从而为构建整体高效的集成学习器提供依据。因此，本部分研究拟从基学习器的评价环节来改进选择性集成学习方法，即构建全面、客观的基学习评价体系，研究内容从以下两方面展开：

（1）**学习器的评价指标体系**。首先，筛选尽量全面的评价指标。在了解评价指标体系构建方式以及相关理论的基础上，筛选合理的可用指标，形成一套完整的基学习器评价指标体系。**在构建指标体系时，可以分类为通用指标体系和领域指标体系**。通用指标体系以

现有的选择性集成学习相关文献为基础，充分考虑现实需求来制定，具体而言，需要考虑如何衡量基学习器的准确性、多样性、计算资源、能耗等不同方面。领域指标体系需要结合问题背景和领域知识进行增减。如对于不平衡分类问题，基学习器的准确度、召回率可能都无法真正度量出基学习器的预测性能，需要使用 AUC 等指标代替。其次，**将评价指标分为投入类和产出类**。投入类指标包含训练和使用基学习器需要花费的计算资源、训练时间、能源消耗等；产出类指标包含对基学习器性能的多方面测量，如准确度、多样性等；而基学习器的误差、分类错误率等指标，则可以看作非期望产出。

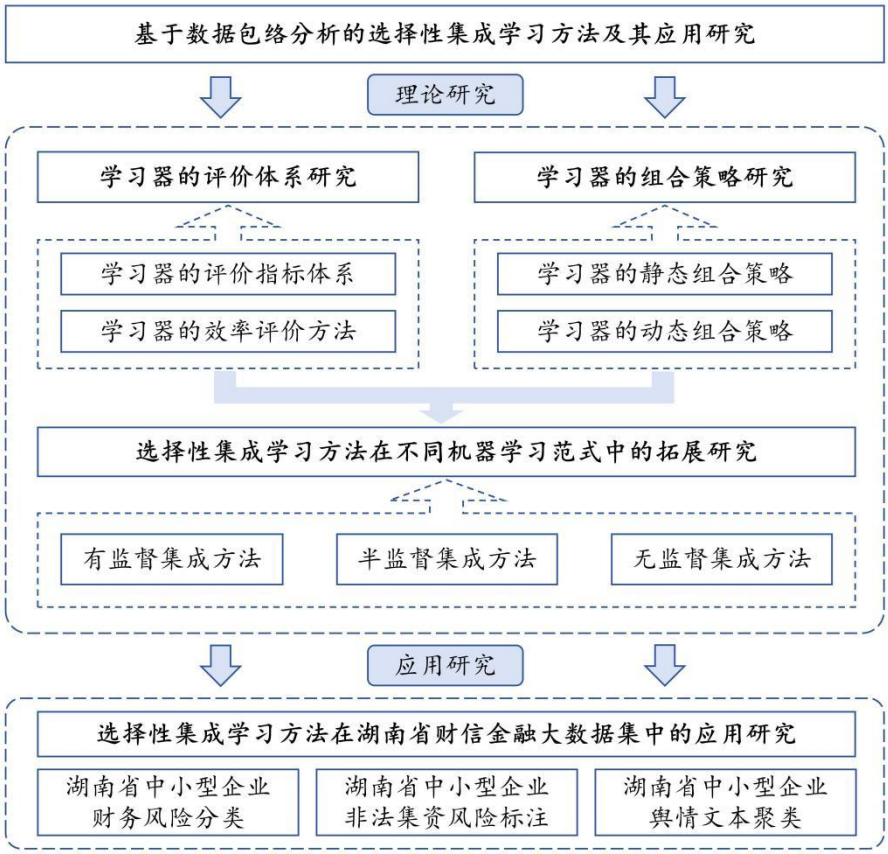

图 2 研究内容总体框架图

（2）**学习器的效率评价方法**。基于构建的基学习器评价指标体系，采用 DEA 对基学习器的效率进行评价，效率评价的结果将作为基学习器组合的依据。一方面，通过对集成学习背景下基学习器的分析发现，传统 DEA 关于生产可能集为凸集的假设很难满足基学习器的计算资源投入与预测性能产出之间的真实情况。例如，通过实验可以观察到基学习器 A 使用 0.6 单位时间取得了 0.6 单位准确度，基学习器 B 使用 0.8 单位时间取得 0.8 单位准确度，而即使在相同的训练环境配置下，二者的线性组合（如使用 0.7 单位时间取得 0.7 单位准确度）却不一定真实存在。基于此，**在对基学习器进行评价时，有必要放松关于生产可能集的凸性假设**。另一方面，传统 DEA 模型都以自评效率为基础，可能会夸大基学习器的多个计算资源投入或预测性能产出的影响，从而高估基学习器的效率。基于此，可以**结合自评与他评的方式，更加准确地反映基学习器效率的相对关系**。基于上述分析，项目拟在

集成学习背景下，提出新的用于评价基学习器效率的方法，并基于评价结果说明模型的管理意义与经济含义，从理论与应用层面比较不同评价方法之间的差异与优劣，从而证明项目提出的效率评价方法的合理性与有效性。

1.2.2 学习器的组合策略研究

通过基学习器的评价体系，能够得到基学习器的效率评价结果。相比于使用性能最优的单一学习器，选择性集成学习方法将选择部分学习器进行组合，从而得到泛化能力强的集成学习器。其中，组合方法起到至关重要的作用。因此，本部分研究拟从基学习器的组合环节来改进选择性集成学习方法，即设计高效、合理的基学习器组合策略，研究内容从以下两方面展开：

（1）**学习器的静态组合策略**。在静态集成背景下，对于所有的待预测样本，都使用相同的集成学习器对其进行预测。因此，**静态组合策略主要研究采用何种方法对基学习器进行组合**。研究分为两个阶段，第一阶段，项目研究如何**从所有生成的基学习器中选择出用于组合的基学习器**，基本思路是根据基学习器的效率评价结果确定，首选应该为有效的基学习器。在特殊情况下，有效的基学习器数量可能非常少，此时可以设定合理的效率阈值，选择效率高于该阈值的高性能学习器。第二阶段，项目研究**所选出的基学习器的权重应该如何分配**，同样地，效率越高的基学习器应该给予更高的权重，即效率越高的基学习器应该对集成学习器预测结果的贡献度越大。此外，项目拟**通过理论和实验两部分内容证明组合方法的合理性与有效性**，理论证明部分将分析基于效率加权组合的集成学习器预测边界，包括预测准确性的上下界、预测结果的偏差和方差等，实验证明部分将通过不同领域、不同样本规模的数据集比较基于效率加权组合的集成学习器与其他组合方法的预测性能。

（2）**学习器的动态组合策略**。在动态集成背景下，对于不同的待预测样本，将构建不同的集成学习器进行预测，从而达到针对每个样本都尽可能的预测准确度高的目的。然而，对所有新的待预测样本都构建新的集成学习器必然导致计算量剧增，且在数据量大、更新速度快的应用场景下，这个问题将更加严重。因此，**动态组合策略主要研究何时对基学习器的组合进行更新**。研究分为两个阶段，第一阶段，项目研究设置监控机制对基学习器的效率进行动态评估。关键在于如何合理设定对基学习器进行更新的判定条件，例如当效率下降至设定的阈值，或基学习器性能下降导致的损失大于对模型进行动态更新的计算资源投入成本时，应对集成模型进行更新。第二阶段，项目研究如何对现有的集成学习器进行更新。项目提出的动态更新算法需要满足以下条件，一是**时效性**，即模型更新的响应时间不能过长、所占用的计算资源不能过多；二是**稳定性**，即更新后的模型不止适用于新的数据，同时也应该能够尽可能保留，甚至优化对于历史数据的预测性能。

1.2.3 选择性集成学习方法在不同机器学习范式中的拓展研究

本部分研究将基于DEA的选择性集成学习理论拓展至不同的机器学习研究范式下，为解决不同的现实问题提供理论基础与方法支撑，研究内容从以下三方面展开：

（1）**有监督集成方法**。以分类预测问题为例，**将基于DEA的选择性集成方法用于分类预测问题，并提出相应的可实施的算法框架**。首先，并非所有评价指标都适用于分类问题，需要根据分类问题选择适用的基分类器评价指标，构建针对分类问题的领域指标体系。其次，考虑是否需要对指标进行降维。如果用于评价的指标个数数量过多，而基学习器的数量较少，则应考虑对指标进行降维。接着，构建合适的DEA模型对基学习器进行效率评价，并基于评价结果对基学习器进行组合。最后，通过公开数据集验证所提出的算法框架相较于其他机器学习分类算法的有效性。

（2）**半监督集成方法**。同样以分类预测为例，通过结合半监督方法改进选择性集成分类器的预测结果。考虑到现实分类问题中，常出现已标记样本稀少、人工标注费时费力、多数样本为未标记数据的特点。虽然在仅使用已标记样本的情况下，集成学习依然能够呈现出相较于传统机器学习算法更出色的性能表现，但大量的未标记数据信息未被有效利用。基于此，本部分研究将**在有监督集成算法框架的基础上，提出新的半监督选择性集成算法框架**。具体而言，将在已构建的基学习器效率评价体系基础上额外考虑数据的标注成本，引入无标记的样本提升学习器的多样性并设计相应的组合策略，通过公开数据集验证研究提出的算法相较于有监督集成算法，能够在降低训练成本的同时提升预测准确度。

（3）**无监督集成方法**。以聚类问题为例，通过选择性集成策略改进学习器的聚类效果。由于在聚类任务中，学习样本没有标签，对于聚类结果的准确性进行度量是比较困难的。虽然多样性便于量化，但在聚类集成的研究中，没有理论能够直接证明多样性能够直接影响聚类的好坏，常用于有监督集成策略中的组合方法无法直接挪用至无监督集成中。对于上述问题，本部分研究将重点研究**如何构建适用于聚类方法的性能指标体系，以及如何使用合适的组合策略将基聚类结果进行融合，进而提出选择性聚类集成算法框架**，并通过公开数据集验证研究提出的算法与已有的聚类算法相比的有效性。

1.2.4 选择性集成学习方法在湖南省财信数据集中的应用研究

本部分拟开展选择性集成学习方法在湖南省财信数据集中的应用研究。湖南省财信数字科技有限公司是湖南省财信集团下提供业务数智化支持服务的子公司。该公司针对财信集团不同的业务搭建数据中台，如财信信托业务数据仓库、湖南省企业风险决策数据仓库等。其中，湖南省企业风险决策数据仓库收集了**湖南省近50万家企业的运营信息、财务信息、法律案件等数据，还会在网络公开平台实时抓取与企业相关的舆情文本、招聘广告等数据**，用于构建企业风险决策模型。由于其监测下的湖南省企业大多为中小型企业，可能出现数据缺失、数据标签不平衡等情况，即存在数据规模大、质量低的问题。普通的机器

学习方法不仅难以取得预期的效果，而且可能导致较高的算力成本。为解决上述问题，在湖南省财信金融大数据的支持下，本项目将通过三个具体的应用：**湖南省中小型企业财务风险分类研究、湖南省中小型企业非法集资风险标注研究、湖南省企业舆情文本聚类研究**，对本文提出的理论结果进行检验，证明本项目提出的方法能够以较少的资源投入取得较好的预测性能，有助于财信数科在实施湖南省中小企业风险智能决策系统的开发过程中降本增效，并为政府或企业实施相关机器学习项目提供参考。

1.3 拟解决的关键问题

本项目拟解决的关键科学问题如下：

（1）**如何构建合理的数据包络分析模型对基学习器的效率进行评价**。评价模型的构建既要考虑评价指标的选择，又要考虑基学习器生产可能集的性质，不同的评价模型可能得到不同的评价结果，需要尽量保证结果的客观和可解释性。因此，如何选择合适的数据包络分析模型是本项目拟解决的第一个关键科学问题。

（2）**如何使用效率评价结果指导基学习器组合权重的分配**。为保证集成学习器的整体高效，需要在对集成学习器进行组合时充分利用效率评价的结果，从而保证预测准确度的同时消耗尽量少的计算资源。因此，如何使用效率评价结果指导基学习器组合权重的分配是本项目拟解决的第二个关键问题。

（3）**如何结合不同机器学习范式的特点改进选择性集成学习方法**。不同的机器学习范式具有不同的特点，需要充分结合其优势，取长补短。对于无监督集成方法，应考虑如何通过集成方法提升聚类结果的稳定性与准确性；对于半监督集成方法，应考虑如何运用无标签数据提升分类器的多样性，减少数据标注成本并达到预期效果。因此，如何结合不同机器学习范式的特点改进选择性集成学习方法是本项目拟解决的第三个关键问题。

2 拟采取的研究方法及可行性分析

2.1 拟采取的研究方法

本项目拟围绕基于 DEA 的选择性集成学习方法及应用研究展开理论研究和应用研究，结合上述研究内容，本项目总体研究技术路线如图 3 所示。

2.1.1 基学习器的评价体系研究方案

为构建全面、客观的基学习器评价体系，本项目将采用内容分析法、生产经济学理论、数学规划等理论与方法进行研究，具体方案如下：

（1）**基学习器的评价指标体系研究方案**。为构建全面、客观的基学习器评价指标，首先，本项目将结合研究的现实背景，广泛搜集有关发展新一代人工智能的政策报告内容，**采用内容分析法归纳出当下发展人工智能、开发机器学习模型应遵循的条件，作为构建学习器评价准则的指导依据**。其次，结合已有关于机器学习的研究中所涉及的性能评价指标

进行全面的总结，并区分投入类指标与产出类指标。基于此，构建通用指标体系和领域指标体系。最后，可通过企业实地调研，深入分析现实中机器学习项目落地实践需要考虑的因素，并参考相关专家的意见对评价指标体系进行完善。

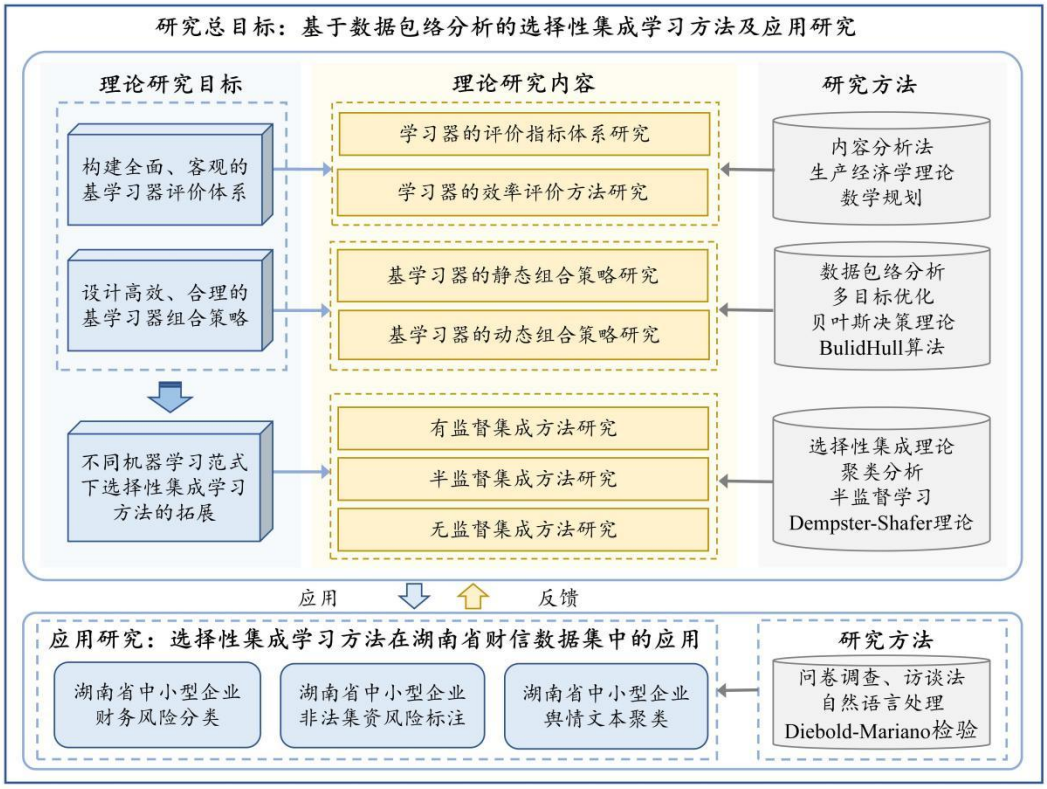

图 3 技术路线图

（2）**基学习器的效率评价方法研究方案**。在构建完成基学习器的评价指标体系的基础上，选择对基学习器进行效率评价的方法。首先，由于在集成学习背景下，传统 DEA 关于生产经济学理论的假设（如生产可能集为凸集、规模收益不变等）难以反映基学习器计算资源投入与预测性能产出之间关系，可基于非凸假设构建基学习器的生产可能集，并写出不同规模收益假设下生产前沿面的数学表达式。其次，采用**线性规划、多目标规划**等方法，**结合数据包络分析中的交叉效率评价、方向距离函数、FDH 模型等效率评价方法，提出集成学习背景下的基学习器效率评价模型**。最后，基于评价模型给出基学习器的效率的定义，并从理论和应用层面比较不同评价模型之间的异同。

2.1.2 基学习器的组合策略研究方案

为设计高效、合理的基学习器组合策略，本项目将采用数据包络分析、多目标优化、贝叶斯决策理论、时间序列分析、BulidHull 算法等理论与方法进行研究，具体方案如下：

（1）**基学习器的静态集成策略研究方案**。静态集成策略将基于基学习的评价结果，对基学习器进行选择与组合，提出基于效率加权的选择性集成算法，并对算法有效性进行分析，如图 4 所示。第一阶段，**采用项目提出的基学习器效率评价模型对所有已生成的基学习器进行评价**，并根据评价结果选择效率值大于α的基学习器。这里α的值可以根据经验确

定，如设定α的值为 0.8，即选取所有效率值高于 0.8 的基学习器，从直觉上来看，所选的基学习器都是表现相对较好的，且数量不会过少。第二阶段，对选出来的基学习器进行组合。**建立多目标优化模型**确定基学习器的权重，**优化目标分别为均方误差最小与多样性最大，约束条件至少应包含效率低的基学习器的权重不高于效率高的基学习器的权重**，可采用**多目标进化算法**加快求解。第三阶段，对所提出的算法的有效性进行检验。一是分析算法复杂度与收敛性，二是分析算法的决策边界，如基于**贝叶斯决策理论**证明算法性能优于单一基学习器与平均组合学习器的条件，三是通过数值实验比较算法与其他集成算法的性能。

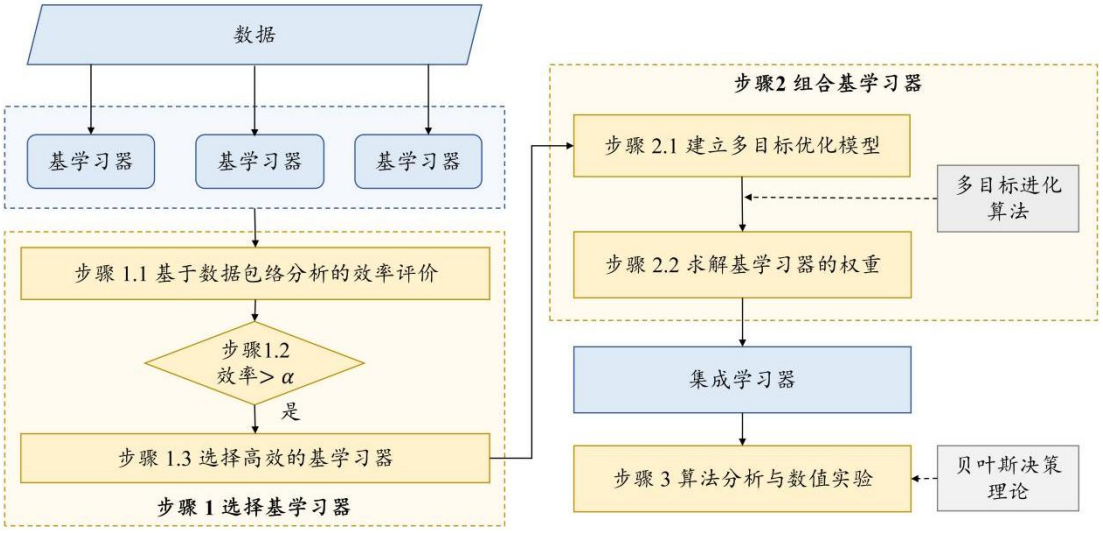

图 4 基学习器的静态集成策略研究方案

（2）**基学习器的动态集成策略研究方案**。动态集成策略对集成学习器预测性能的变化情况进行监控，从而在保证预测性能良好的同时，避免消耗过多的计算资源用于模型更新，如图 5 所示。首先，选用准确性、多样性等集成学习器的预测性能指标，**将历史T个时期内上述指标的时间序列分解为趋势项、周期性和噪音项**，若无监测指标呈下降趋势，即不满足集成学习器的更新条件，则直接使用历史集成学习器对当前数据进行预测。其次，**采用经验回放（Experience Replay）的机制构建基学习器池**。利用当前数据训练新的基学习器并添加至基学习器池中，基学习器池中仍应保留部分历史基学习器，可根据实际情况设定存储条件，如仅保留最新的N个基学习器，或保留历史时期内所有效率大于α的基学习器，从而在一定程度上满足模型预测结果的稳定性。最后，结合静态组合策略从更新后的基学习器池中选择高效的基学习器，并构建新的基于效率加权的集成学习器。需要注意的是，**当基学习器池更新后，历史基学习器的效率将发生改变，需要重新计算池中所有基学习器的效率。**项目拟结合 **BulidHull 算法、FrameStream 算法**等用于解决大规模样本效率计算与动态更新问题的算法，对池中基学习器效率进行快速更新，以保证预测的响应速度。

2.1.3 选择性集成学习理论在不同机器学习范式中的拓展研究方案

为提高不同机器学习范式下学习器的性能，本项目将采用**选择性集成理论、聚类分析、半监督学习方法、Dempster-Shafer 理论**等理论与方法进行研究，具体方案如下。

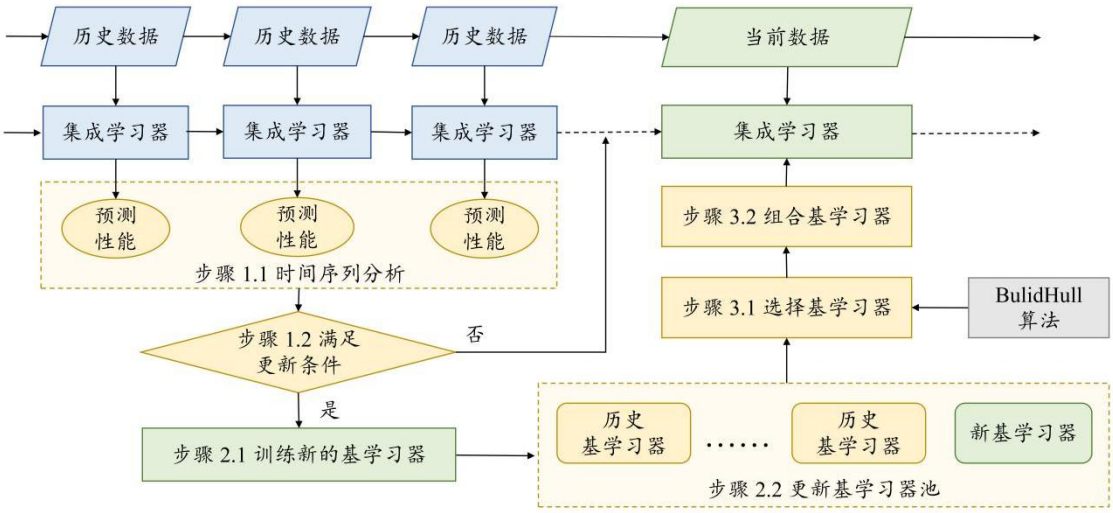

图 5 基学习器的动态集成策略研究方案

（1）**有监督集成方法研究方案。**以分类预测问题为例，本部分研究将基于 DEA 的选择性集成方法用于分类预测问题，提出相应的分类集成算法框架，并通过公开数据集验证算法的有效性。首先，构建基分类器的评价体系。基于已有研究选择基分类器的能耗、准确度和多样性的评价指标，并将其分为投入类与产出类。**投入指标包括模型训练与使用的 CPU 时间、GPU 时间、内存占用量、False Positive Rate（FPR）等；产出指标包括 Recall、Accuracy、F1-score、AUC、Precision、Kohavi-Wolpert Variance、内部一致性等。**其次，基于上述指标对生成的基分类器进行交叉效率评价，并选择高效的基分类器进行组合。**建立多目标优化模型，在考虑效率排序约束**下，最大化集成学习器的 Accuracy 同时最小化集成学习器的内部一致性，采用 **NSGA-II 算法求解基学习器的权重**，从而得到集成分类器。最后，通过公开数据集验证所提出的算法与基准方法（单一学习器和平均集成方法）相比具有更高的分类准确度。

（2）**半监督集成方法研究方案。**以分类预测问题为例，本部分研究将结合半监督学习思想和基于 DEA 的选择性集成方法，提出相应的半监督分类集成算法框架，并通过公开数据集验证算法的有效性。首先，利用有标记样本训练基分类器，并在（1）构建的基分类器的评价体系的基础上选择出部分效率相对较高的基分类器，进入初始基分类器池。其次，使用池中的基分类器对无标记样本进行预测，若预测结果置信度超过阈值，则将该结果视作伪标签，将该样本并入有标记样本集，训练新的基分类器并加入基分类池，同样可**采用 BuildHull 算法快速更新基分类器的效率并保留高效的基分类器**；否则，该样本仍未无标记样本。**重复上述过程至所有无标记样本都具有高置信度类别标签。**最后，通过成本-效益分析验证所提出的算法的有效性。成本指标除算法能耗相关的指标外，应额外包含**数据标注成本（可由有标记样本量占比表示）**，效益指标主要指算法预测结果的准确度相关的指标。以公开数据集为例，设定有标记样本量占比区间为 30%至 70%，其余样本作为无标签样本进行多次实验，证明所提出的方法与基准方法（如有监督分类集成算法、非集成分类算法）

相比，能够在达到相同分类准确度的同时减少需花费的成本。

（3）**无监督集成方法研究方案。**以聚类问题为例，本部分研究将结合聚类分析方法、证据理论和基于 DEA 的选择性集成方法，提出相应的聚类集成算法，并通过公开数据集验证算法的有效性。首先，构建基聚类的评价指标体系。与分类器不同，由于缺少监督信息，很难直接判断哪种聚类结果与真实情况最匹配，即很难直接度量聚类准确性。基于此，评价指标体系主要由能耗和多样性两大类组成，并将其划分为投入和产出指标。**投入指标主要包括 CPU 时间、GPU 时间、内存占用量等；产出指标主要包括调整兰德系数（adjusted rand index，ARI）、归一化互信息（normalized mutual information，NMI）等**，选用交叉效率评价方法对所有聚类结果进行评价并选择部分高效率的基聚类。其次，集成所选出来的基聚类结果得到最终的聚类结果。由于聚类方法不关心样本被划分至哪个类别，而更关注哪些样本被聚为一类，研究拟**通过样本间的共现概率估计每个样本的邻域，再基于邻域样本的类别估计该样本的类别置信度函数**。将每个基聚类视作一个专家，类别置信度函数则为专家对样本划分的评价，然后**基于 Dempster-Shafer 理论对基聚类结果进行融合**，得到最终的聚类结果。最后，拟采用公开数据集验证所提出的方法与基准方法（如基于投票法的聚类集成算法）相比能够有效提高聚类准确度，并基于实验结果分析基聚类的多样性与聚类结果准确度之间的关系。

2.1.4 选择性集成学习方法在湖南省财信数据集中的应用研究方案

为验证上述研究的理论成果的有效性，本项目将采用**问卷调查、访谈法、自然语言处理、Diebold-Mariano 检验**等方法进行应用研究，具体方案如下。

（1）**数据准备阶段。**依托于湖南财信金融大数据的支持，收集湖南省中小企业风险决策相关的指标与文本数据，并对数据进行初步的清洗、分析与整理，分别构造湖南省中小企业财务风险分类数据集、非法集资风险标注数据集与舆情文本数据集。此外，项目组拟通过访谈调查、问卷调查等方法建立并完善评价指标体系，收集企业在开发湖南省企业风险智能决策系统过程中涉及的计算资源成本、能耗成本等相关数据，为下一阶段不同机器学习范式下的选择性集成方法应用做好准备。

（2）**应用研究阶段。**对于不同机器学习范式下的预测问题，分别选用不同的基于 DEA 的选择性集成方法。例如，在湖南省中小企业财务风险分类问题中，风险类别分为高风险、较高风险、中风险和低风险四类，大多数企业具有风险类别标签，可采用有监督集成方法；在湖南省中小企业非法集资风险标注问题中，由于非法集资活动从立案调查到裁判文书公示的过程较为漫长，即非法集资行为标签验证具有延后性，且具有非法集资行为标签的企业占比极少，可采用半监督集成方法，合理利用无标签数据提高分类预测效果；在湖南省中小企业舆情文本聚类问题中，先采用自然语言处理技术对文本抽取关键词并转换为词向量，再采用无监督集成方法完成文本聚类任务。将数据集划分为训练集与验证集，通过

Diebold-Mariano 检验进行实证分析，验证在不同的预测任务中，项目所提出的方法相较于基准方法都能够在提高预测准确度的同时，显著减少投入的计算资源成本。

（3）**检验与反馈阶段。**基于以上研究结果，为决策者开展机器学习项目的计划与实施提供管理建议，同时将应用结果进行反馈以检验理论研究的正确性、有效性和适用性。

2.2 可行性分析

申请人拟从“研究方案的可行性”与“项目团队实施项目的可行性”两方面分别阐述项目实施的可行性。

（1）**研究方案的可行性。**本项目的研究问题基于现实需求提炼而来，面向我国发展新一代人工智能的战略需求，具有重要研究意义。本项目的研究内容分为理论研究和应用研究两部分，理论研究主要围绕选择性集成方法展开，主要分为：基学习器的评价、基学习器的组合和选择性集成学习理论在不同机器学习范式下的拓展三个部分，应用研究以理论研究为基础，对理论研究成果进行检验，研究内容界定明确，思路逻辑清晰。本项目的研究方案提供了解决研究问题、实现研究目标的初步思路，研究路线切实可行，并可在研究过程中逐步调整、完善，符合研究的基本范式。项目拟采取的关键研究方法主要包括 DEA、多目标决策、机器学习等。目前，学术界关于基于 DEA 的效率评价、不同机器学习范式的融合等研究均较为成熟，这些研究为本项目的开展提供了坚实的理论方法依据，且已有研究通过实验表明 DEA 对于指导机器学习模型的训练是有效且合理的。因此，本项目的研究方案具有可行性。

（2）**项目团队实施项目的可行性。**首先，申请人及项目成员在 DEA、机器学习、大规模样本的效率评价等相关领域研究近两年，取得了一定的研究成果，已完成相关论文 2 篇。其次，本项目涉及数学建模与算法设计，申请人曾获得第十九届中国研究生数学建模竞赛全国二等奖，且系统学习过编程语言，熟练掌握 Matlab、python 等工具，以及 Scikit-learn、TensorFlow 等常用机器学习库，故在这两方面均有比较扎实的研究基础。再者，申请人多次参与导师及课题组同学的项目，有一定的经验和能力完成好此项目。此外，团队成员包含两名优秀的富有科研热情和创新精神的博士生和一名刻苦钻研且有建模基础的硕士生，研究团队结构合理，研究领域范围广泛，包括基于 DEA 的聚类方法研究，评价理论与方法研究等，并均在各自研究方向取得了一定的研究成果。最后，项目团队已对本研究某些子问题的相关核心模型构建进行了反复的交流讨论。因此，项目团队为项目顺利开展奠定了坚实的基础，项目实施具有可行性。

综上所述，本项目**研究问题明确、研究思路清晰、研究方法切实可行、研究内容具备较高的研究价值。**申请人及项目组成员具备扎实良好的科研基础，可以胜任本项目的研究工作。基于此，本项目的研究科学可行，并且有望取得高质量的研究成果。

3 本项目的创新之处

本项目结合数据包络分析、集成学习、多目标优化等交叉领域知识，对不同机器学习范式下的选择性集成方法与应用进行研究，具体创新之处如下。

（1）**研究视角创新**。本项目基于我国发展新一代人工智能技术的现实需求，并充分考虑到算法算力提升所带来的能耗问题，力求以尽量少的计算资源投入获得尽量优的预测性能，基于效率评价的视角，提出新的集成学习背景下基学习器的效率评价方法，构建整体高效的选择性集成学习技术，研究视角具有创新性。

（2）**研究方法创新**。本项目将数据包络分析方法与集成学习方法结合，既为选择性集成方法对基学习器的选择与组合提供理论支持，又深化数据包络分析与机器学习领域的交叉研究，同时还结合了经济学的生产理论，多目标优化，贝叶斯决策等理论与方法，能够促进多学科的交叉应用，研究方法具有创新性。

（3）**理论与实践紧密结合**。本项目针对机器学习算法能耗成本日益剧增的实际问题，结合全面、客观的评价体系与高效、合理的组合策略，提出基于数据包络分析的选择性集成理论，将其拓展至不同机器学习范式下并应用于湖南省企业风险智能决策系统中，为相关管理者低成本、高质量地实施机器学习项目提供科学的方法工具。

4 预期研究进展

本项目研究时间从 2024 年 1 月开始，预计用两年的时间完成研究计划，工作进度安排如图 6 所示，其中$Q_i(i=1,2,3,4)$表示当年的第i季度。

研究内容		2024				2025			
		Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4
基学习器的评价体系研究	学习器的评价指标体系	■	■						
	学习器的效率评价方法	■	■						
基学习器的组合策略研究	学习器的静态组合策略			■	■	■			
	学习器的动态组合策略					■	■	■	■
在不同机器学习范式的拓展研究	有监督选择性集成方法	■	■	■	■				
	半监督选择性集成方法			■	■	■	■		
	无监督选择性集成方法					■	■	■	■
在湖南省财信数据集中的应用	数据准备阶段	■	■						
	应用研究阶段			■	■	■	■	■	
	检验与反馈阶段								■
总结	申请专利或软件著作权					■	■	■	■
	撰写学术论文	■	■	■	■	■	■	■	■

注：年度工作计划可能会根据研究中的新问题做出动态弹性调整。

图 6 研究进度计划甘特图

除上述甘特图中所包括的具体研究工作计划外，项目在研期间，申请人及项目成员将通过参加高水平学术会议、阅读相关文献，积极跟进 DEA、集成学习等领域的最新动态，以确保项目的顺利完成。

5 预期成果

本项目的主要工作是结合 DEA 和选择性集成学习方法，针对选择性集成方法中基学习器的评价与组合问题进行研究并提出一套科学的解决方法，预期研究成果包括：

（1）**发表学术论文**：拟在管理科学与工程、运筹学领域的国内外主流学术期刊，如 European Journal of Operational Research，Omega，《系统工程理论与实践》等期刊发表（含录用）论文 1~2 篇。

（2）**申请专利或软件著作权**：拟根据项目提出的基于 DEA 的选择性集成算法优化现有的湖南省企业风险决策预警系统，并申请专利或软件著作权 1~2 项。

（3）提升科研能力：通过参加国际性或国内学术交流会议，增加与国内外优秀学者的交流，并定期开展研究团队内部学术研讨会，在导师的指导下取得科研能力上的进步。

四、研究基础

与本项目有关的研究工作积累、已取得的研究工作成绩及目前承担项目的情况（不能出现相关人员信息）

1 项目负责人

1.1 与本项目有关的工作积累

首先，申请人学习机器学习理论与方法、DEA 相关知识近两年的时间，对机器学习方法、集成学习方法、交叉效率评价、FDH 模型等理论与方法进行了较为深入的研究，并在研一期间对《微观经济学》、《管理研究方法》、《大数据分析》、《最优化理论》等课程进行系统性的学习并取得了满绩的成绩，为本项目研究储备了充足的基础知识。其次，通过大量的英文文献阅读与总结，申请人在文献检索、阅读、信息提取和学术论文的撰写方面具备充足的科研技能。再者，申请人熟练使用 Python、Matlab、Lingo、SPSS 等软件与工具，熟练掌握 scikit-learn、TensorFlow 等主流机器学习库的使用，可为本项目的算法设计与实施提供技术支撑。此外，申请人于 2023 年 4 月至 7 月在湖南财信数字科技有限公司业务研发部实习，期间参与完成了湖南省旅游、医疗等 5 个重点行业非法集资风险智能决策模型的开发与优化工作，对研究数据集的样本特征、分布等信息具备充分的了解。最后，申请人已对项目关键的问题进行初步研究，已完成相关论文 2 篇，具体如下：

（1）论文《非凸技术假设下大规模样本 FDH 效率值的计算与更新方法》，已投稿至期刊《系统工程理论与实践》在审（导师一作，本人二作）。该论文提出了一种基于决策单元之间支配关系与参考关系的快速枚举算法（Fast Enumeration Algorithm，FEA）用于计算大规模样本 FDH 效率值，进而结合支配关系和参考关系的传递性提出了动态快速枚举算法（Dynamic Fast Enumeration Algorithm，DFEA）用于更新 FDH 效率评价结果。FDH 是非凸技术假设下经典的的效率评价方法，**上述研究与本项目涉及的基学习器的非凸生产可能集的构建、基学习器动态组合策略的设计等研究内容密切相关。**

（2）论文 An ensemble learning method for classification: integrating data envelopment analysis with machine learning，已投稿至期刊 European Journal of Operational Research 在审。该论文提出了一种新的用于分类问题的集成学习方法，具体而言，该方法使用范围方向模型（Range directional model，RDM）对基分类器的计算资源消耗和分类预测性能进行评估，并使用交叉效率模型确定分类器的组合权重。结果表明，该方法优于现有的基于投票和基于 BCC 模型的集成方法，且相比于其他集成方法能够有效节省计算资源。**上述研究与本项目涉及的基学习器的效率评价方法、基学习器的权重分配等研究内容高度匹配。**

1.2 已取得的研究工作成绩

通过本科阶段及研一的学习和研究，申请人取得的研究工作成绩（包括已完成的科研工作、获得的成果及奖励）如下：

（1）以导师一作本人二作完成论文《非凸技术假设下大规模样本 FDH 效率值的计算与更新方法》，投稿至《系统工程理论与实践》在审；

（2）第十九届研究生数学建模竞赛全国二等奖（排序为 1）；

（3）MathorCup 数学建模挑战赛全国一等奖（排序为 1）；

（4）全国大学生英语竞赛三等奖；

（5）主持并完成校级大学生自主创新项目一项。

1.3 目前承担项目情况

参与导师国家自然科学基金面上项目两项（项目编号：71871223；72171238）

参与导师国家社科基金重大项目子课题（项目编号：21&ZD103）

参与中南大学研究生自主创新项目一项（项目编号：1053320214632）

2 项目成员 2

2.1 与本项目有关的工作积累

项目成员已在 DEA 方法、机器学习理论与方法、评价理论与方法等领域进行深入研究，具备充足的知识储备与工作的积累。其中，项目成员关于 DEA 与聚类分析方法的交叉研究开展的工作填补了现有文献的空白，并取得了较好的现实应用效果，能够为本项目的开展提供必要的理论与方法支持，具体如下：

（1）论文 DEA-based Internal Validity Index for Clustering 投稿至期刊 Journal of Operational Research Society，已返修。该论文提出了一种基于 DEA 的内部有效性指标（DEA-based Internal Validity Index，DEAI）用于评价聚类结果，通过在 10 个公开数据集上的实验表明，DEAI 识别最佳聚类结果的准确度都优于其余现有的内部有效性指标。**上述研究为本项目构建无监督集成学习方法下的基学习器评价指标体系提供了必要的理论依据。**

（2）论文《基于数据包络分析与多目标软子空间聚类的房地产企业运营风险综合评价研究》投稿至期刊《运筹与管理》，已录用。该论文基于超效率模型和多目标软子空间聚类方法，对湖南省房地产企业运营风险开展综合评价研究，结果表明，该方法相较于传统聚类方法更为准确，且在高维数据上表现更优。**上述研究与本项目涉及的基学习器效率评价方法、选择性集成学习方法在无监督学习范式下的拓展等研究内容提供方法参考。**

2.2 已取得的研究工作成绩

项目成员在 DEA 方法与聚类方法交叉研究领域已发表较高水平论文 1 篇，多篇论文在国内外权威期刊在审和返修，另发表了多篇关于 DEA 方法的研究，主要成果如下：

（1）论文《基于数据包络分析与多目标软子空间聚类的房地产企业运营风险综合评价研究》投稿至期刊《运筹与管理》，已录用；

（2）论文 DEA-based Internal Validity Index for Clustering 投稿至期刊 Journal of Operational Research Society，已返修，该期刊为 SCI 检索，ABS 三星期刊；

（3）在国际期刊 Environment development and sustainability 发表论文 1 篇，该期刊为 SCI 检索，影响因子 3.118，JCR 二区；

（4）在国际期刊 Environmental Science and Pollution Research 发表论文 1 篇，该期刊为 SCI 检索，影响因子 4.306，JCR 二区；

此外，该成员作为负责人主持中南大学研究生科研创新项目（编号：1053320214632）一项，具备充足的研究经验和项目经历，为项目顺利展开提供帮助与支持。

3 项目成员 3

3.1 与本项目有关的工作积累

项目成员对 DEA 方法、机器学习理论与方法已有较长时间的学习经历，主要研究方向为 DEA 与集成学习领域交叉研究，对本项目研究方法与理论都较为熟悉。其次，项目成员通过大量英文文献的阅读以及英文论文的撰写培养了该成员良好的科研能力，为本项目的顺利开展提供了科研基础。最后，项目成员熟练掌握项目所需的工具，包括 Pycharm、Matlab、Gurobi 等，能够有效促进项目的开展。

3.2 已取得的研究工作成绩

（1）通过研一课程完成多篇课程小论文的撰写；

（2）以导师一作本人二作完成论文 An ensemble learning method for classification: integrating data envelopment analysis with machine learning，投稿至期刊 European Journal of Operational Research 在审。

4 项目成员 4

4.1 与本项目有关的工作积累

项目成员已对 DEA 方法、最优化理论、决策方法与理论等领域进行长期的深入学习和研究，并且撰写并发表了多篇英文论文，**具有丰富的项目经历，在理论基础和科研能力等方面均可为本项目的顺利开展提供较大帮助和支持。**

4.2 已取得的研究工作成绩

首先，该项目成员关于 DEA 方法相关的研究领域已发表较高水平论文 4 篇，有多篇论文在国内外权威期刊在审和返修，多篇论文待投，已发表研究成果如下：

（1）在管理运筹学权威期刊 European Journal of Operational Research 发表论文 1 篇，该期刊为 SCI 检索，JCR 一区，ABS 四星级期刊；

（2）在国际权威期刊 Computers & Industrial Engineering 发表论文 2 篇，该期刊为 SCI 检索，JCR 一区，ABS 二星级期刊；

（3）在管理运筹学权威期刊 OR Spectrum 发表论文 1 篇，该期刊为 SCI 检索，JCR 二区，ABS 三星期刊；

其次，该项目成员获得的奖励和荣誉如下：

（1）2017-2018 年度中南大学优秀学生

（2）2020 年湖南省优秀毕业生

（3）2020 年度博士研究生国家奖学金（学院共 5 人获得，排名第 1）

（4）2022 年湖南省优秀硕士学位论文

（5）2022 年中南大学山茶花奖学金

最后，该项目成员主持并完成 2021 年度中南大学研究生科研创新项目（《数字经济背景下网络系统固定成本分摊方法研究》，2021zzts0028）一项。

五、经费预算

支 出 科 目	金 额 （万元）	计 算 根 据 及 理 由
合 计	1.5	
科研业务费	0.9	包括差旅费、调研费、会议费、国际合作交流费、出版/文献/信息传播/知识产权事务费等。 (1) 国际合作交流费指研究人员出国及外国专家来华费用，参加国际会议的会务费共计 4000 元，明细如下： A. 项目研究过程中，参与国内学术交流会议 1-2 人次，会务费每人次 1500 元，共计 3000 元； B. 项目研究过程中，组织开展学术研讨、咨询以及协调项目研究工作等活动而发生的会议费用合计约 1000 元。 (2) 差旅费指市外出差的住宿费、旅费、伙食补助费、杂费等，以及调研和到外地开会的会务费和市内出租车、公交车乘车费，共计 3000 元，明细如下： A. 在项目研究过程中参与国内学术交流会议 2-5 人次，往返车费、住宿费等共计 1000 元； B. 项目组成员前往非同城高校、科研院所进行学术交流探讨 1-3 人次，每人次差旅费平均 500 元，共计 1000 元； C. 项目组成员进行相关数据搜集、调研等 2-4 人次，每人次差旅费平均 300 元，每次一周左右，共计 1000 元。 (3) 出版/文献/信息传播/知识产权事务费指出版费、资料费、印刷费、办公软件购买费、文献检索费、专业通信费、专利申请及其他知识产权事务等费用 1000 元；项目期间预计发表 SCI/SSCI 检索论文 1-2 篇，撰写论文 1-3 篇，语言润色费、版面费等每篇平均 500 元，共计约 2000 元。
实验材料费	0.2	材料费指各种原材料、辅助材料等低值易耗品、办公用品的购置费用，共计 2000 元，明细如下： (1) 项目研究过程中发生的资料查询、打印、复印、检索、翻译等费用合计 800 元； (2) 项目研究过程中购买文具、U 盘等办公用品合计 200 元； (3) 购买中英文书籍费，其中中文书籍 5-10 本，平均每本 50 元，英文书籍 3-5 本，平均每本 150 元，合计约 1000 元。
仪器设备费	0.3	设备费包含专用设备、办公设备、专用软件购买费、办公家具购置费用：项目研究过程中购买打印机一台，约 0.3 万元。
相关经费	0.1	其他，如项目研究过程中产生的设备（如计算机）维护费用等。

注: 预算支出科目按下列顺序填写: 1. 科研业务费 2. 实验材料费 3. 仪器设备费 4.相关经费。