第 62 卷　第 2 期
2026 年 3 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 62, No. 2
Mar., 2026

DOI:10.13232/j.cnki.jnju.2026.02.012

用于多实例嵌入学习的层次化关键实例选择方法

潘　臻[1,2],张雨轩[3],张佳慧[4],闵　帆[5],杨　梅[5*]

(1.成都师范学院计算机科学学院,成都,611130;2.四川省教育数字化发展与评价重点实验室,成都,610000;
3.西南交通大学信息科学与技术学院,成都,611756;4.兰卡斯特大学计算机与通信学院,兰卡斯特,LA1 4YW,英国;
5.西南石油大学计算机科学与软件工程学院,成都,610500)

摘　要:在多实例学习(Multi-Instance Learning,MIL)中,数据对象以层次结构的形式被组织为由多个实例组成的包.传统的MIL嵌入方法通过选择具有代表性的实例来将每个包嵌入为向量以简化MIL问题,然而大多数现有方法忽略了包的层次结构,导致生成的关键实例集(Key Instance Set,KIS)中包含大量离群实例.此外,这些方法没有利用KIS去除包中的离群点,影响了包的嵌入效果.为此,提出一种层次化关键实例选择的多实例嵌入学习算法(Hierarchical Key Instance Selection for Multi-Instance Embedding Learning,HKMIL),其包括三个关键技术:首先,层次化实例选择技术(Hierarchical Instance Selection,HIS)结合子空间与相似度更新机制,用于识别和优化KIS,同时根据实例密度生成新的包;其次,Fisher向量嵌入技术(Fisher Vector Embedding,FVE)利用高斯混合模型从新包中提取关键统计信息,将其转化为向量;最后,集成分类技术(Ensemble Classification Technique,ECT)动态加权融合KIS更新前后的信息,以提升包级别标签预测的准确性.在六个典型的MIL任务上的实验结果表明,HKMIL优于九种当前最先进的算法,取得了更优异的分类性能.

关键词:多实例学习,关键实例,实例选择,嵌入方法,集成学习

中图分类号:TP181　　**文献标志码:**A

Hierarchical key instance selection for multi-instance embedding learning

Pan Zhen[1,2], Zhang Yuxuan[3], Zhang Jiahui[4], Min Fan[5], Yang Mei[5*]

(1. School of Computer Science, Chengdu Normal University, Chengdu, 611130, China;
2. Sichuan Provincial Key Laboratory of Education Digitalization Development and Evaluation, Chengdu, 610000, China;
3. School of Information Science and Technology, Southwest Jiaotong University, Chengdu, 611756, China;
4. School of Computing and Communications, Lancaster University, Lancaster, LA1 4YW, UK;
5. School of Computer and Software Engineering, Southwest Petroleum University, Chengdu, 610500, China)

Abstract: In MIL (Multi-Instance Learning), data objects are hierarchically organized as bags containing multiple instances. The well-known MIL embedding approach embeds each bag as a vector by selecting representative instances. However, most existing methods ignore the hierarchical structure of bags, leading to the generated KIS (Key Instance Set) that contains substantial outlier instances (the instances where bag labeling cannot be triggered). Additionally, KIS is not utilized to exclude outliers in bags, which will impact embedding quality. To address these issues, we propose HKMIL (Hierarchical Key Instance Selection for Multi-Instance Embedding Learning) algorithm with three technologies. First, HIS (Hierarchical Instance Selection) uses subspace- and affinity-based updates to identify and refine KIS, generating new bags while

基金项目:成都师范学院科研项目(YJRC202449),南充市政府高校科研合作项目(23XNSYSX0084,23XNSYSX0062),浙江省海洋大数据挖掘与应用重点实验室开放课题(OBDMA202102)

收稿日期:2026－01－26

*通信联系人,E-mail:yangmei@swpu.edu.cn

considering instance density. Second, FVE (Fisher Vector Embedding) technique uses Gaussian mixture models to extract key statistical information from the new bags, converting them into vectors to simplify the MIL problem. Third, ECT (Ensemble Classification Technique) dynamically weights the information before and after KIS updates for improved bag label predictions. Experiments on six MIL tasks show that HKMIL outperforms nine state-of-the-art algorithms, achieving superior classification performance.

Keywords: multi-instance learning, key instance, instance selection, embedding, ensemble learning

多实例学习(Multi-Instance Learning, MIL)是一种旨在处理复杂数据结构的学习范式. 在MIL中,每个数据样本被表示为一个包含多个实例的包,且监督信息仅在包级提供,而实例级标签是未知的,或者获取代价高昂. 根据MIL基准假设[1],当一个包中至少包含一个正实例时,该包被标记为正,否则为负. 这种学习范式与现实世界中的许多应用高度契合,例如图像分类[2-3]、医学诊断[4]和网络推荐[5].

根据算法的实现原理,现有的MIL方法可分为三类[6]. 基于实例的方法[7]对每个实例分别进行分类,然后将单个实例的预测结果整合以估计包的标签,然而,由于实例监督信息的缺失,其预测结果不可靠,导致误差累积. 基于包的方法[8]将整个包视为一个整体,重点建模不同包之间的相似性或者邻域关系,虽然这种方式能够捕获包的层次结构,但由于没有考虑潜在的实例标签,容易遗漏关键信息,影响整体性能. 基于嵌入的方法[9]将包映射到新的特征空间中,将MIL问题转化为标准的单实例分类任务,通过这种嵌入,能够有效地刻画包与实例之间的关系.

近年来,基于关键实例选择(Key Instance Selection, KIS)的嵌入学习已成为MIL研究的热点之一,其中关键实例代表性可通过聚类分析[10]、相似度计算[11]或判别式优化[12]等方式来评估,但这些方法通常仅依赖单一度量指标来生成关键实例集(Key Instance Set, KIS),忽略了包的层次结构,而且没有利用KIS来去除包中的离群实例,在嵌入过程中引入噪声信息,导致分类性能下降且可扩展性低.

为了应对上述问题,本文提出一种面向多实例嵌入学习的分层关键实例选择算法(Hierarchical Key Instance Selection for Multi-Instance Embedding Learning, HKMIL),其总体框架如图1所示. 具体地,在给定的MIL数据集中,天然存在实例级与包级这两个层次结构. 由于实例数量远多于包的数量,直接生成KIS难度较高,因此,首先提出一种三阶段的分层实例选择技术(Hierarchical Instance Selection, HIS)用于初始KIS的生成与更新. 对于初始化阶段,HIS在实例空间中采样多个随机子空间,利用判别策略快速生成初始正负关键实例集,以确定可行域. 对于更新阶段,HIS采用基于相似度的更新策略,进一步清除初始KIS中的离群实例. 对于生成阶段,则利用更新后的KIS计算包内实例的相似度与密度,从而筛选关键实例并生成新包. 其次,提出Fisher向量嵌入技术(Fisher Vector Embedding, FVE),其利用高斯混合模型从新包中提取关键统计信息,并将其转换为固定长度的向量,从而将MIL分类任务简化为传统的单实例监督分类问题.

最后,设计集成分类技术(Ensemble Classification Technique, ECT),充分利用更新前后的KIS信息,并采用动态加权策略融合不同阶段的特征,实现更稳定、更强健的包级分类性能. HKMIL在23个数据集上进行了实验,实验结果证明其在突变性预测与医学图像分类任务中表现尤为出色,显著优于九种最新的MIL算法.

本文的主要贡献如下.

(1)提出一种分层实例选择技术,充分利用包的层次结构,有效去除离群实例,并通过快速定位可行域生成新包. 在此基础上,FVE准确提取统计信息,将包嵌入为向量,简化MIL分类问题.

(2)设计一种集成分类技术,充分利用KIS更新前后的关键信息,防止重要知识遗忘,并通过动态加权策略构建更强的集成分类器.

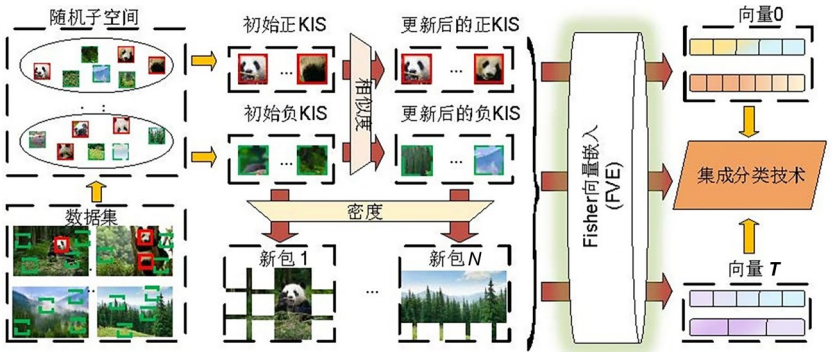

图1　HKMIL的整体框架

Fig. 1　The overview of HKMIL

1　相关工作

1.1　多实例学习　1997年Dietterich[1]首次将MIL引入药物活性检测领域.在这一框架下,每个分子被视为一个包,而同一分子的同分异构体被视为实例.如果一个包中至少包含一个能够产生药效的实例,则该包被标记为正,否则为负.通过这种方式,可以预测新分子是否具有药物活性.基于这一思想,2006年MILES[13]将实例空间视作原型集,并通过高斯相似度来评估包与原型之间的关系,从而在新的特征空间中对包进行特征化表示.2013年MILFM[14]通过学习正负相关概念的分布来评估实例的重要性与代表性.2018年ABMIL[15]采用注意力机制来识别包内的关键实例,并结合门控机制来提升模型性能.2020年StableMIL[16]专注于识别因果实例以确定最具代表性的正样本.同年,SMDP[17]提出一个基于密度峰值聚类的代表性发掘方法,能够最大化实例选择的多样性.2024年MINTL[5]提出一种基于优化理论的类别边界信息学习方法,能更高效地确定分类超平面.INS[18]在MIL框架下首次设计了实例级弱监督对比学习算法,以更好地学习实例特征表示.ELIMIPL[19]利用共轭标签信息来有效提高消歧性能.CAMIL[20]引入邻域约束注意力,在包内建模实例间依赖关系,并将上下文约束作为先验知识融入MIL模型.MIPLMA[21]提出一种基于裕度调整的MIL算法,可用于注意力分数和预测概率的裕度自适应调整.

2025年CDL[22]提出一种可插拔的消歧策略,大幅提升了MIL算法的准确度.PSMIPL[23]提出一种用于MIL的倾向性评分框架,有效利用了标签集中的弱监督信息.MSFF[24]提出一种融合体素块内部、体素块之间以及高置信度体素块的MIL多尺度特征融合框架,可以有效提升辅助诊断效果.GDF-MIL[25]提出一种基于原型压缩的图MIL方法,能快速捕获包中的语义与拓扑结构信息.上述大多数方法属于基于嵌入MIL算法,按照实现原理可分为四类,即实例方法、统计方法、核函数方法以及基于包的方法.本文主要聚焦于基于实例的嵌入方法,后续将对此进行详细介绍.

1.2　基于实例的MIL嵌入方法　在实际的MIL应用中,实例数量通常远远大于包的数量.由于实例标签不可见或缺失,MIL成为一种典型的弱监督学习问题,因此,基于实例的MIL嵌入方法的关键挑战在于有效利用未标注实例中的潜在信息,以更好地训练分类模型.2009年MILD[26]通过计算实例之间的相似度来评估实例对包标签的贡献,从而识别正包中的真实正样本.2018年MILDM[27]引入实例判别准则,并定义了实例可区分性度量.2019年ISK[28]基于数据相关的隔离集核,设计了一种基于数学期望的稀疏特征映射方法.2022年IMIL[29]提出一种基于因果干预的期望最大化框架,以提升实例级预测的可靠性.2023年FCBE-miFV[7]采用模糊聚类计算每个实例的选择概率,从而实现鲁棒的包嵌入.DEMIPL[30]提出消歧注意力机制,学习包内实例的注意力权重.2024年CaMIL[31]引入可学习的因果建模机制与跨注意力,有效消除了包内的伪相关性.2026年ProtoMIL[32]提出一种基于原型引

导和注意力增强的多实例嵌入学习,其可以很好地处理混淆实例的分布问题,并可以作为一个即插即用部件来提升已有模型的性能.

这些算法在实例选择与包转换方面提供了宝贵的启发,然而,它们存在一个共性问题,即生成KIS的过程通常依赖单一度量指标,忽视了包的层次结构,这种忽略会在嵌入过程中引入离群实例,降低了分类性能. 对此,本文采用分层策略并同时考虑相似度与密度信息,以更全面地挖掘实例特征. 此外,HKMIL的差异不在于引入新的监督信号,而在于将关键实例集从一次性选择转换为可迭代优化的中间变量,并进一步把KIS用于离群实例剔除与新包构建,同时通过集成策略显式利用更新前后的互补信息,从而在不增加实例标注成本的前提下提升嵌入质量与分类稳定性.

2 算法

首先对基本符号进行定义并说明MIL范式,然后依次介绍所设计的三项核心技术,即用于关键实例挖掘的分层实例选择技术、用于获取包表示的Fisher向量嵌入技术以及用于获取预测标签的集成分类技术.

2.1 符号定义与MIL范式说明 令 $T=\{B_i\}_{i=1}^{N}$ 表示给定的数据集,其中,$B_i=\{x_{ij}\}_{j=1}^{n_i}$ 表示 T 中的第 i 个包,x_{ij} 表示 B_i 中的第 j 个实例,N 和 n_i 分别表示 T 和 B_i 的基数. 令 $\boldsymbol{X}=\bigcup_i B_i\subseteq R^d$ 表示包含所有 x_{ij} 的实例空间,其中,d 表示 x_{ij} 的维度. MIL中每个包对应一个监督信息 $y_i\in\{-1,+1\}$,所有的 y_i 构成对应于 T 的标签向量 $Y=[y_1,y_2,\cdots,y_{n_i}]$. 相应地,尽管 x_{ij} 没有监督信息,但其对应于一个潜在的语义标签 $y_{ij}^*\in\{-1,+1\}$. 基于标准MIL假设[1],包的标签可以由实例标签导出:

$$y_i=\begin{cases}+1, & \text{if } \exists j\in 1,2,\cdots,n_i, y_{ij}^*=+1\\ -1, & \text{otherwise}\end{cases} \tag{1}$$

在基于关键实例的MIL嵌入方法中,主要任务包括生成关键实例集 $\boldsymbol{K}$ 并设计基于 $\boldsymbol{K}$ 的嵌入函数 $F:B_i\mapsto V_i$,从而将 B_i 映射为向量表示 V_i,并通过类似SVM的分类器输出预测标签 $\hat{y}_i$.

2.2 分层实例选择技术 现有的关键实例选择方法[12,25]存在以下缺点:(1)解空间过大,使得KIS的生成过程耗时;(2)KIS未被更新或只保留更新后的KIS,造成信息损失;(3)学得的KIS没有用于包内的关键实例筛选,导致嵌入过程中仍存在离群实例. 为此,本文提出一种三阶段的KIS优化策略HIS,其充分利用包的层次结构,有效去除离群实例并生成新包,为后续嵌入与分类做准备.

2.2.1 阶段一:KIS初始化 KIS初始化的目标是通过随机子空间策略[31]快速识别实例空间中的关键节点,同时利用判别关系生成初始的正、负关键实例集[11-12]. 对此,首先生成 Φ 个随机子空间:

$$S=\{S_\phi\}_{\phi=1}^{\Phi}=\left\{\{c_k^\phi\}_{k=1}^{K}\right\}_{\phi=1}^{\Phi} \tag{2}$$

其中,c_k^ϕ 表示子空间 $\boldsymbol{X}^*$ 的聚类中心. 具体地,子空间 $\boldsymbol{X}^*\in R^{n^*\times d^*}$ 是从原始实例空间 $\boldsymbol{X}\in R^{n\times d}$ 中随机采样而来,且对于每个采样的实例,只保留 d^* 个随机特征维度[33]. 一旦 S_ϕ 确定,$\boldsymbol{X}^*$ 便可自然地被划分为 K 个簇 C_k^ϕ. 基于簇中实例所对应的包的标签,可以计算每个簇的正、负实例的比例:

$$\begin{aligned}\delta_{\phi k}^{+}&=\frac{\left|\{x_{ij}^*\,|\,x_{ij}^*\in C_k^\phi, y_i=+1\}\right|}{\left|C_\phi^k\right|}\\ \delta_{\phi k}^{-}&=\frac{\left|\{x_{ij}^*\,|\,x_{ij}^*\in C_k^\phi, y_i=-1\}\right|}{\left|C_\phi^k\right|}\end{aligned} \tag{3}$$

据此,将实例在当前子空间中的相关性权重定义为:

$$s_{ij}^\phi=\begin{cases}\delta_k^{+\phi}, & x_{ij}^*\in C_k^\phi, y_i=1\\ \delta_k^{-\phi}, & x_{ij}^*\in C_k^\phi, y_i=-1\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

该权重体现了实例的判别性. 因此,对于 Φ 个子空间上,实例的平均权重计算为:

$$s_{ij}=\frac{1}{\Phi}\sum_{\phi=1}^{\Phi}s_{ij}^\phi \tag{5}$$

最终,从每个包中选择权重最大的实例分别构建初始正、负KIS:

$$\boldsymbol{K}_0^{+}=\left\{x_{ij^*}\,\middle|\,x_{ij}\in B_i, y_i=1, j^*=\operatorname{argmax}_j s_{ij}\right\} \tag{6}$$

$$\boldsymbol{K}_0^{-}=\left\{x_{ij^*}\,\middle|\,x_{ij}\in B_i, y_i=-1, j^*=\operatorname{argmax}_j s_{ij}\right\} \tag{7}$$

2.2.2 阶段二:KIS更新 虽然初始化阶段可以快速生成KIS,但由于随机采样存在一定程度的

信息丢失，因此需对KIS进行更新．首先，定义实例与KIS之间的相似度为：

$$A(x^*,\boldsymbol{K})=\frac{1}{|\boldsymbol{K}|}\sum_{x_{ij}^*\in\boldsymbol{K}}e^{-|x^*-x_{ij}^*|^2} \tag{8}$$

该相似度越高，表示输入的实例与KIS的信息一致性越强．进一步，设计了内部竞争与外部评估这两个KIS更新策略，具体如下．

对于内部竞争策略，其主要用于KIS的提纯，即去除KIS中的低相似度实例：

$$\begin{aligned}&x_{it\tau}^+,\tau=\operatorname{argmin}_{x_{ij}^+\in\boldsymbol{K}_t^+}A(x_{ij}^+,\boldsymbol{K}_t^+)\\&x_{it\tau}^-,\tau=\operatorname{argmin}_{x_{ij}^-\in\boldsymbol{K}_t^-}A(x_{ij}^-,\boldsymbol{K}_t^-)\end{aligned} \tag{9}$$

其中，x_{it}^+和x_{it}^-分别表示正、负KIS中需要被去除的实例，t表示更新轮次．外部更新则通过遍历$\boldsymbol{X}^*$中所有的实例x^*，具体的更新机制如下：

$$\begin{aligned}x_{it\tau}^+=x^*,\quad&\text{if } A(x^*,\boldsymbol{K}_t^+)>A(x^*,\boldsymbol{K}_t^-)\&\\&A(x^*,\boldsymbol{K}_t^+)>A(x_{it\tau}^+,\boldsymbol{K}_t^+)\\x_{it\tau}^-=x^*,\quad&\text{if } A(x^*,\boldsymbol{K}_t^-)>A(x^*,\boldsymbol{K}_t^+)\&\\&A(x^*,\boldsymbol{K}_t^-)>A(x_{it\tau}^+,\boldsymbol{K}_t^-)\end{aligned}$$

2.2.3　阶段三：新包生成　在大多数现有方法中，KIS并未用于清除包内的离群实例，导致嵌入向量中仍包含噪声，因此，本文利用KIS生成新包．直观地，$\boldsymbol{K}_t^+$和$\boldsymbol{K}_t^-$分别是包含最多正例和负例的集合．根据MIL基本假设[1]，正实例的数量明显少于负实例的数量，自然导致生成负KIS的准确率高于生成正KIS．此外，由于测试包的标签未知，无法简单地利用测试包与正负KIS的相关性来去除异常值．为了解决这个问题，利用实例密度来更全面地评估包中的实例：

$$\rho_{ij}=\sum_{k\neq j}e^{-\left(d_{jk}/d_c\right)^2} \tag{10}$$

其中，d_c为距离阈值，d_{ij}表示实例x_{ij}^*与x_{ik}^*之间的欧式距离．进一步，结合实例与KIS的相似度，实例的关键性评估值计算为：

$$\delta_{ij}=\zeta_{ij}\times\rho_{ij} \tag{11}$$

在此基础上，为了去除包中的离群实例并获取新包，首先计算包的平均相似度：

$$A_i=\frac{1}{n_i}\sum_{j=1}^{n_i}A(x_{ij}^*,K_t^-) \tag{12}$$

其归一化值计算为：

$$A_i\leftarrow\frac{A_i-\min_iA_i}{\max_iA_i-\min_iA_i} \tag{13}$$

若A_i大于阈值ξ，则删除包中δ_{ij}最小的50%实例，以生成新包B_{it}^*，用于后续嵌入．

2.3　Fisher向量嵌入技术　Fisher向量嵌入技术源自Fisher核[7,10]，其主要利用高斯混合模型（Gaussian Mixture Model，GMM）将实例空间$\boldsymbol{X}_t^*=\bigcup_iB_{it}^*$划分为$K$个成分$\boldsymbol{G}_k\subseteq\boldsymbol{X}_t^*$．对于所有的成分，可以使用参数集合$\lambda=\left\{\alpha_{tk},\mu_{tk},\Sigma_{tk}\right\}_{k=1}^K$来统一描述，其中$\alpha_{tk}$，$\mu_{tk}$以及$\Sigma_{tk}$分别表示用于描述$\boldsymbol{G}_k$的权重、均值以及协方差矩阵．基于此，每个包的Fisher向量嵌入计算为：

$$\boldsymbol{V}_{it}=\left[f_{\alpha_{t1}}^{B_{it}^*},f_{\mu_{t1}}^{B_{it}^*},f_{\Sigma_{t1}}^{B_{it}^*},\cdots,f_{\alpha_{tK}}^{B_{it}^*},f_{\mu_{tK}}^{B_{it}^*},f_{\Sigma_{tK}}^{B_{it}^*}\right]\in R^{K(2d^*+1)} \tag{14}$$

其中$f_{\alpha_{tk}}^{B_{it}^*},f_{\mu_{tk}}^{B_{it}^*},f_{\Sigma_{tk}}^{B_{it}^*}$分别表示GMM中的统计向量．为了提升模型性能，参照Wei et al[10]的工作对嵌入向量进行幂归一化与L_2归一化：

$$\boldsymbol{V}_{it}\leftarrow\frac{\boldsymbol{V}_{it}}{|\boldsymbol{V}_{it}|_2},v_{itl}\leftarrow sign(v_{itl})\sqrt{|v_{itl}|} \tag{15}$$

其中，v_{itl}是$\boldsymbol{V}_{it}$的第l个特征值．

2.4　集成分类技术　集成分类技术构建了一个加权集成模型，其由多个基分类器及其对应权重组成：

$$\boldsymbol{M}=(M_t,\omega_t)_{t=0}^T \tag{16}$$

其中，M_t为单实例分类器（如SVM）；ω_t表示分类器权重，其通过在验证集上预测的包分类准确率获得；T为KIS的更新次数．最后，利用每次更新生成的特征空间训练分类器，再通过加权投票获得最终预测结果：

$$\hat{y}_i=sign\left(\sum_{t=0}^T\omega_t\hat{y}_{it}\right) \tag{17}$$

其中，$\hat{y}_{it}=M_t(\boldsymbol{V}_{it})$.

3　实验

使用来自六个不同应用场景的23个MIL数据集来验证提出的HKMIL算法的有效性．针对每个数据集，将HKMIL与九种当前最先进的MIL方法进行对比．此外，还全面分析了五类实验，依次为性能对比、消融实验、参数分析、统计显著性检验以及效率分析．

3.1　实验设置　使用的MIL数据集涵盖六类典型任务：药物活性预测[1]、致突变性预测[34−35]、医学图像分析[10]、图像分类[11]、文本分类[36]以及视频

异常检测[37]. 此外,在基于嵌入的MIL算法的四个子类中依次选择具有代表性的九个算法用于性能对比,分别为:(1)基于统计的算法Simple-MI[38],没有超参数配置;(2)基于核函数的算法MSK[11],其权重比设为(0.5,0.5),距离函数采用最小值策略;(3)基于包的算法ELDB[39],包选择模式设为"a",代表性包比例设为0.9;(4)基于嵌入的算法,包括FCBE-miFV[7],MILFM[14],MILDM[27],miVLAD[10],StableMIL[16]以及DP-MIL[9].

参数设置如下:FCBE-miFV的子空间比例设为0.05,子空间数量为15;MILFM的聚类中心数量为40;MILDM的判别性实例数为包的数量;miVLAD的密码本大小为1;StableMIL的实例阈值为0.25;DPMIL的候选实例数量级从$[1,1.25,\cdots,2.5]$中选取.

对于提出的HKMIL算法,其主要控制参数包括参与离群实例去除的包比例$\{0.25,0.5,0.75,1.0\}$(即对于给定的数据集,只有给定比例且随机选择的包会被执行离群实例去除操作)和随机子空间数量$\{5,10,\cdots,25\}$,其他参数通过经验值进行设置,每个子空间的聚类数量设为1,高斯混合模型的分量设为1,包内删除的离群实例比例为0.5. 此外,所有实验均采用10次10折交叉验证($10\times10CV$),以平均准确率与相应的标准差作为最终的评价指标.

3.2 性能对比 将HKMIL与九种最新算法在六类数据集上进行了性能对比,结果以平均准确率和标准差表示,如表1所示,表中黑体字表示最佳结果. 综合来看,HKMIL在23个数据集中取得优异表现,其中在16个数据集上排名第一,三个数据集上排名第二. 进一步,HKMIL在Messidor与Ucsb breast医学数据集上的性能优势最显著. 在大多数文本数据集上(超过一半),HKMIL得到最优或接近最优的结果,这种性能提升主要得益于HKMIL的层次化结构,在实例选择过程中逐步更新关键实例集,并利用其指导新包生成和模型训练,而现有的多数方法没有充分利用这些阶段性信息. 在图像数据集Tiger上,HKMIL的结果略逊于ELDB,这表明其仍有优化空间. 该差距的原因在于ELDB使用了判别性策略与强化更新机制,可以更高效地提取图像特征,而传统MIL方法普遍缺乏此类能力. 总之,HKMIL在六个领域的数据集上展现了很高的性能优势与可迁移性,尤其在药物活性预测与致突变性预测上.

表1 HKMIL与前沿算法在六个类型数据集上的性能对比

Table 1 Performance of HKMIL with state-of-the-art algorithms on six types of MIL datasets

Dataset	$n\times d$	Simple-MI	MSK	ELDB	FCBE-miFV	MILFM	MILDM	miVLAD	StableMIL	DPMIL	HKMIL
Musk1	476×166	0.800±0.016	0.884±0.013	0.880±0.018	0.879±0.021	0.856±0.008	0.802±0.033	0.849±0.011	0.840±0.017	0.828±0.018	**0.911±0.012**
Musk2	6598×166	0.750±0.014	0.818±0.026	**0.861±0.018**	0.835±0.014	0.776±0.020	0.804±0.029	0.766±0.012	0.820±0.014	0.836±0.063	0.846±0.015
Mutagenesis1	10486×7	0.852±0.020	0.842±0.012	0.844±0.017	0.765±0.017	0.861±0.007	0.828±0.013	0.817±0.036	0.852±0.025	0.797±0.010	**0.869±0.014**
Mutagenesis2	2132×7	0.765±0.030	0.745±0.020	0.713±0.028	0.800±0.010	0.798±0.032	0.755±0.026	0.830±0.010	0.740±0.014	**0.857±0.011**	0.750±0.032
Messidor	12352×687	0.618±0.008	0.627±0.009	0.569±0.015	0.685±0.005	0.622±0.005	0.640±0.002	0.676±0.003	0.623±0.005	0.655±0.008	**0.737±0.015**
Ucsb breast	2002×708	0.837±0.043	0.532±0.027	0.630±0.076	0.824±0.027	0.556±0.023	0.560±0.022	0.800±0.018	0.552±0.010	0.640±0.010	**0.876±0.027**
Elephant	1391×230	0.825±0.005	0.781±0.008	0.754±0.020	0.829±0.006	0.815±0.012	0.765±0.016	0.847±0.010	0.632±0.026	0.849±0.032	**0.866±0.012**
Fox	1320×230	0.619±0.009	0.534±0.010	0.588±0.027	0.590±0.012	0.608±0.026	0.542±0.035	0.633±0.018	0.597±0.043	0.623±0.020	**0.642±0.019**

续表

Dataset	$n\times d$	Simple-MI	MSK	ELDB	FCBE-miFV	MILFM	MILDM	miVLAD	StableMIL	DPMIL	HKMIL
Tiger	1220×230	0.811±0.012	0.767±0.014	0.674±0.027	0.785±0.013	0.763±0.013	0.690±0.014	**0.849±0.007**	0.657±0.021	0.788±0.020	0.796±0.005
News.aa	5443×200	0.836±0.008	**0.854±0.008**	0.846±0.018	0.826±0.010	0.526±0.079	0.546±0.071	0.840±0.023	0.526±0.040	0.845±0.021	0.842±0.012
News.co	5175×200	0.574±0.036	0.734±0.009	0.631±0.031	0.731±0.011	0.496±0.022	0.522±0.043	0.692±0.016	0.474±0.040	0.721±0.021	**0.752±0.008**
News.csi	4827×200	0.754±0.008	0.776±0.005	0.781±0.020	0.799±0.014	0.576±0.032	0.566±0.066	0.800±0.016	0.502±0.055	0.662±0.005	**0.816±0.010**
News.csm	4473×200	0.778±0.008	0.832±0.004	0.764±0.039	0.721±0.004	0.528±0.068	0.434±0.034	0.780±0.011	0.510±0.050	**0.846±0.023**	0.752±0.008
News.cw	3,110×200	0.710±0.032	0.840±0.012	0.796±0.014	0.858±0.014	0.578±0.028	0.568±0.043	0.822±0.010	0.542±0.045	0.840±0.017	**0.858±0.012**
News.mf	5306×200	0.588±0.010	0.724±0.017	0.644±0.024	0.730±0.019	0.512±0.023	0.468±0.027	0.722±0.019	0.526±0.067	0.721±0.012	**0.748±0.008**
News.ra	3458×200	0.754±0.005	0.768±0.008	0.715±0.023	0.787±0.016	0.524±0.016	0.518±0.067	**0.816±0.010**	0.520±0.050	0.714±0.011	0.796±0.010
News.rm	4730×200	0.772±0.026	0.822±0.008	0.811±0.018	0.832±0.013	0.548±0.033	0.570±0.051	0.828±0.008	0.540±0.019	0.814±0.019	**0.868±0.012**
News.rsb	3358×200	0.746±0.010	0.832±0.009	0.792±0.033	0.832±0.010	0.546±0.034	0.482±0.034	0.816±0.010	0.542±0.031	0.765±0.011	**0.842±0.008**
News.rsh	1982×200	0.808±0.010	0.884±0.010	0.772±0.032	0.844±0.015	0.504±0.005	0.474±0.059	0.896±0.010	0.510±0.048	0.868±0.006	**0.896±0.015**
News.se	3192×200	0.920±0.000	**0.938±0.004**	0.882±0.013	0.930±0.009	0.530±0.000	0.556±0.020	0.924±0.005	0.510±0.036	0.851±0.004	0.932±0.008
News.ss	3655×200	0.822±0.004	0.814±0.005	0.788±0.018	0.869±0.013	0.542±0.017	0.502±0.028	0.856±0.012	0.500±0.014	0.831±0.013	**0.876±0.022**
News.trm	4606×200	0.616±0.010	0.742±0.015	0.664±0.023	0.757±0.009	0.526±0.010	0.472±0.039	0.774±0.008	0.510±0.036	0.732±0.045	**0.776±0.008**
Shanghai-Tech	7616×2048	0.612±0.025	0.400±0.038	0.433±0.070	0.849±0.065	0.854±0.007	0.735±0.011	0.851±0.070	0.734±0.050	0.832±0.036	**0.864±0.006**
Average		0.748	0.769	0.737	0.796	0.613	0.580	0.804	0.583	0.780	**0.818**
Mean rank		5.9	4.4	6.1	3.9	7.5	8.6	3.5	8.7	4.53	**2.17**

3.3　消融实验　HKMIL的核心步骤包括KIS的初始化与更新,前者用于识别信息丰富的实例以剔除噪声实例,后者通过迭代更新进一步提升KIS的质量.在此基础上,FVE模块将每个包嵌入固定长度向量,使传统机器学习分类器可以直接应用.由于嵌入是该框架不可或缺的核心环节,因此FVE模块在消融实验中始终保留.此外,当移除更新模块时,HKMIL会退化为仅包含单一分类器的模型.因此,重点分析了KIS更新机制与集成分类器对性能的提升,具体如表2所示,表中黑体字表示最优性能.由表可见,包含更新与集成机制的完整HKMIL模型在所有代表性数据集上均取得最佳性能(除Musk2).例如,完整模型和两种消融版本相比,在Mutagenesis2数据集上,准确率分别提升了4.5%和3.5%,在Musk2上略有0.2%的下降,原因在于该数据集中包大小的差异极大(最大与最小包的实例数比例超过1000),这种极端差异干扰了HKMIL的核心机制.总体上,KIS更新与集成模块的结合能在大多数情况下显著提升模型的分类性能.

3.4　参数分析　HKMIL包含两个关键参数,即包去除比例和随机子空间的数量,其中,前者决定后续子空间的规模,后者影响实例评估的精度与运行时间.为了分析这两个参数对算法的影响,

表 2 HKMIL 在六个代表数据集上的消融实验

Table 2 Ablation study of HKMIL on six representative datasets

Number	Update	Ensemble	Musk1	Musk2	Mutagenesis1	Mutagenesis2	News0. se	News. ss
1	×	×	0.887±0.013	**0.866±0.005**	0.854±0.025	0.705±0.029	0.926±0.016	0.871±0.007
2	√	×	0.907±0.015	0.846±0.008	0.861±0.018	0.715±0.034	**0.932±0.004**	0.872±0.008
3	√	√	**0.911±0.012**	0.846±0.015	**0.869±0.014**	**0.750±0.032**	**0.932±0.008**	**0.876±0.022**

在来自三个领域的六个代表数据集上进行参数敏感性分析,结果如图 2 所示,其中,在不同场景下的参数均独立优化. 实验结果表明,对所有包删除所有离群实例是不现实的,这会增加错误剔除关键实例的风险并显著延长处理时间. 较高的去除比例仅适用于如 Musk2 这类数据集,而对于 Mutagenesis2 与 News. ss 等数据集,需要较低的比例. 在子空间数量方面,性能随子空间数量增加而总体提升,但在 10~20 个子空间时已达到最佳平衡点. 因此,HKMIL 推荐的参数设置为包去除比例为{0.25, 0.5, 0.75},随机子空间数量为{10, 15, 20}.

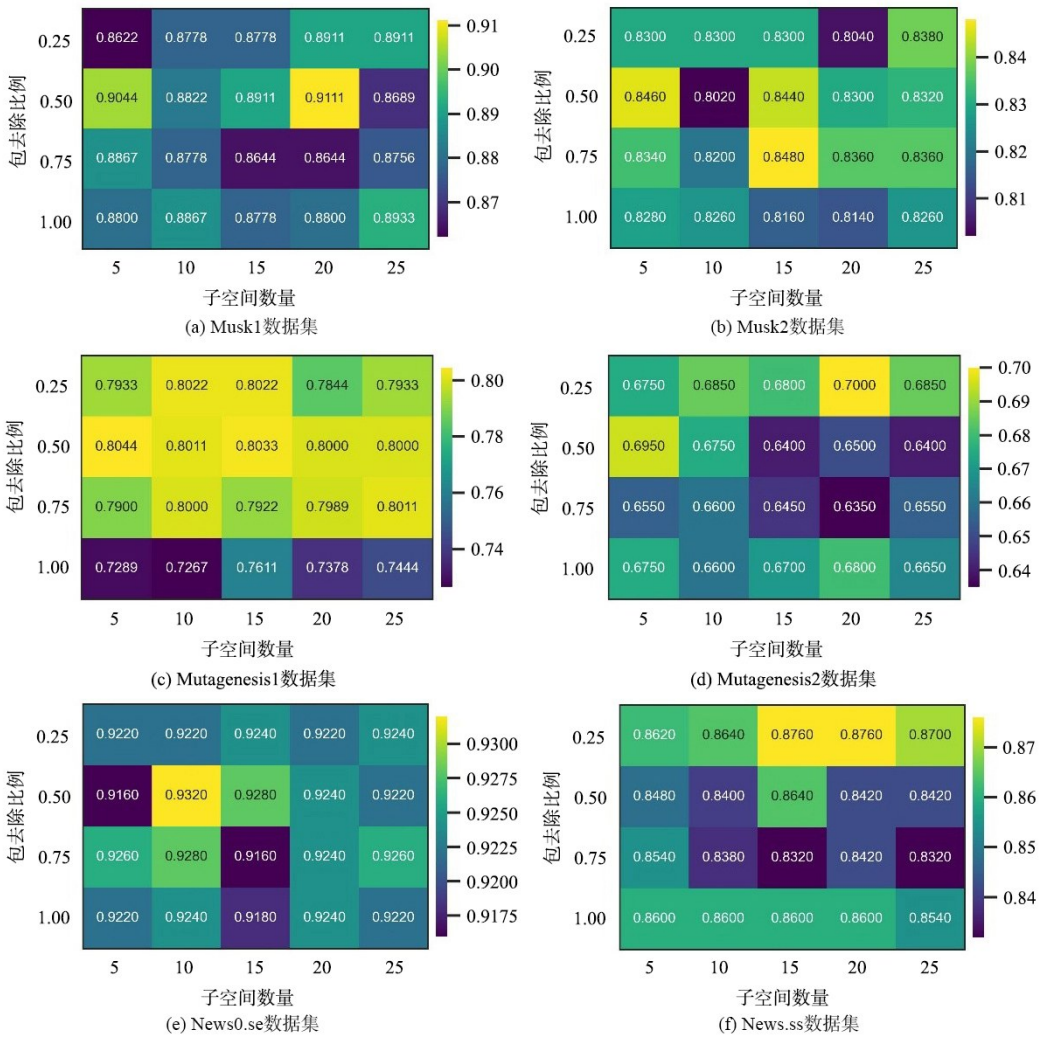

图 2 HKMIL 在六个代表数据集上的参数分析

Fig.2 Parameter analysis of HKMIL on six representative datasets

3.5　统计显著性对比　为了进一步验证HKMIL与对比算法之间的性能差异,采用显著性水平设置为5%的Friedman检验[40−41]. 具体地,检验结果的统计量为144.9215,$p=0.0000$,表明各算法之间存在显著差异,需要进行事后分析. 对此采用Nemenyi检验[42]进行多算法间的显著性比较,结果如图3所示. 由图可见,HKMIL的平均排序位于最优区间,与FCBE-miFV和miVLAD属于同一显著性区间,且显著优于其他所有算法. 说明HKMIL在整体性能上有统计的显著优势.

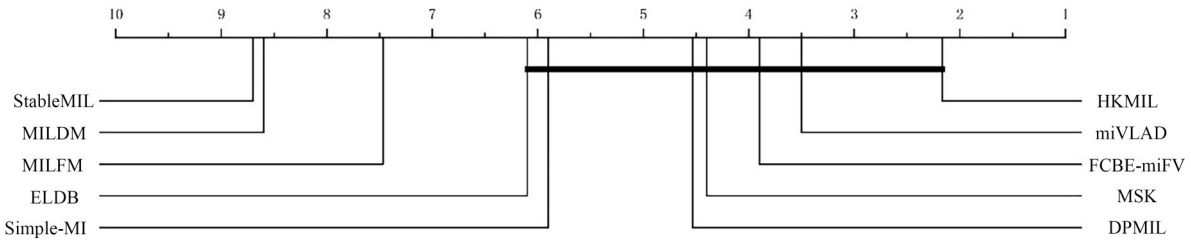

图3　HKMIL与前沿算法的统计显著性对比

Fig. 3　Statistically significant of HKMIL with state-of-the-art algorithms

3.6　时间效率对比　最后,比较HKMIL与其它对比算法的时间复杂度与运行效率. 结果为一次10折交叉验证的平均CPU运行时间. 具体地,HKMIL的时间开销主要由三部分组成. 首先是分层实例选择,其包括实例打分、KIS更新及关键实例筛选. 若使用k-means聚类,随机子空间的聚类复杂度为$O(CRdn)$,其中,C为聚类数,R为子空间数;更新过程与密度计算的复杂度均为$O(dnN)$. 其次是特征映射,其复杂度为$O(KdN)$. 最后是集成模型构建,其基于多个特征空间训练多个分类器,复杂度为$O(MdN^2)$,其中,M为分类器数量. 综上,HKMIL的总体复杂度为$O(dN^2)$.

为了验证以上分析进行了运行时间分析,结果如表3所示. 由表可见,HKMIL的运行效率优于MILFM,MILDM与Stable-MIL等算法. 尽管HKMIL的计算成本略高,但由于其在每个阶段均对实例进行多次评估,因而获得了显著的分类性能提升. 总体上,HKMIL在多数任务中以略高的时间开销换取了更高的精度与稳定性,表现出较优的综合性能.

4　结论

本研究提出一种面向多实例嵌入学习的分层关键实例选择算法HKMIL,以应对MIL中的关键挑战. 具体地,设计了一种三阶段的分层实例选择技术,通过同时利用实例级与包级的层次信息,实现了对关键实例的高效筛选. 在26个MIL数据集上的实验结果表明,HKMIL在多种任务(尤其是致突变性预测与医学图像分类)中表现出显著的性能优势,优于九种最先进的对比算法.

表3　不同算法直接的时间复杂度与平均排名

Table 3　Time complexity and mean rank of different algorithms

Dataset	(d,n,N)	Simple-MI	MSK	ELDB	FCBE-miFV	MILFM	MILDM	miVLAD	Stable-MIL	DPMIL	HKMIL
Musk1	(166,476,92)	0.111	4.975	3.697	5.777	5.072	3.937	1.023	7.405	0.330	9.253
Mutagenesis1	(7,2132,42)	0.089	11.342	994.706	9.590	1231.717	1520.635	1.543	2147.020	0.623	26.911
Ucsb breast	(708,2002,58)	1.242	5.069	56.912	6.791	84.982	61.841	2.539	176.339	1.648	16.858
Elephant	(230,1320,200)	0.187	13.293	20.213	12.139	30.486	25.110	1.483	17.021	0.353	23.724
News0.aa	(200,3048,100)	0.154	6.170	272.508	12.540	348.883	312.990	1.761	64.516	1.158	12.291
Time Complexity		$O(dN)$	$O(dN^2)$	$O(dN^2)$	$O(dN^2)$	$O(dN^2)$	$O(dN^2)$	$O(dN)$	$O(dN)$	$O(dN)$	$O(dN^2)$
Mean rank		1.0	4.8	6.6	5.4	8.8	8.0	3.0	8.4	2.0	7.0

尽管如此,HKMIL 仍存在局限性:(1)更新与密度评估会带来额外计算开销,在实例规模较大或包大小差异极端的场景下效率下降;(2)当前采用固定比例的离群实例剔除与替换式更新,可能误删关键实例或导致阶段性信息遗忘;(3)方法对上游实例特征质量与分布稳定性较敏感,存在跨域漂移时性能波动的风险. 未来将引入记忆/原型机制以减少遗忘并提高鲁棒性,设计更高效的近似评估与自适应阈值策略以降低开销. 此外,面向工程部署,还将探索 HKMIL 的增量式更新与监控机制. 例如,在数据流持续到达时对 KIS 与 GMM 进行轻量更新,并通过漂移检测触发重训练,同时结合关键实例可视化与人工复核形成数据闭环,以提升在真实业务中的长期稳定性.

参考文献

[1] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 1997, 89(1/2):31—71.

[2] Tang W, Yang Y F, Wang Z F, et al. Multi-instance partial-label learning with margin adjustment// Proceedings of the 38th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2024: 26331—26354.

[3] 朱越,姜远,周志华. 一种基于多示例多标记学习的新标记学习方法. 中国科学:信息科学, 2018, 48(12):1670—1680.

[4] Li C T, Huang P, Qin J, et al. Knowledge-driven multiple instance learning with hierarchical cluster-incorporated aware filtering for larynx pathological grading. IEEE Journal of Biomedical and Health Informatics, 2025:1—13.

[5] Xiao Y S, Liu B, Hao Z F. Multi-Instance nonparallel tube learning. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(2):2563—2577.

[6] Zhang Y L, Zhou Z H. Multi-instance learning with key instance shift// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Menlo Park, CA, USA: AAAI Press, 2017:3441—3447.

[7] Waqas M, Tahir M A, Khan S A. Robust bag classification approach for multi-instance learning via subspace fuzzy clustering. Expert Systems with Applications, 2023, 214:119113.

[8] Pal S, Valkanas A, Regol F, et al. Bag graph: Multiple instance learning using Bayesian graph neural networks// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA, USA: AAAI Press, 2022:7922—7930.

[9] Yang M, Chen T L, Wu W Z, et al. Dual-perspective multi-instance embedding learning with adaptive density distribution mining. Pattern Recognition, 2025, 158:111063.

[10] Wei X S, Wu J X, Zhou Z H. Scalable algorithms for multi-instance learning. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(4): 975—987.

[11] Yang M, Zhang Y X, Zhou Z C, et al. Multi-embedding space set-kernel and its application to multi-instance learning. Neurocomputing, 2022, 512: 339—351.

[12] Zhang Y X, Zhou Z C, He X X, et al. Data-Driven knowledge fusion for deep Multi-Instance learning. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(5):8292—8306.

[13] Chen Y X, Bi J B, Wang J Z. MILES: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12):1931—1947.

[14] Hong R C, Wang M, Gao Y, et al. Image annotation by multiple-instance learning with discriminative feature mapping and selection. IEEE Transactions on Cybernetics, 2014, 44(5):669—680.

[15] Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning// Proceedings of International Conference on Machine Learning. New York, NY, USA: PMLR, 2018: 2127—2136.

[16] Zhang W J, Liu L, Li J Y. Robust multi-instance learning with stable instances// Proceedings of the 24th European Conference on Artificial Intelligence. Amsterdam, Netherlands: IOS Press, 2020: 1682—1689.

[17] 杨梅,张雨轩,闵帆. 密度峰值聚类的半监督多示例学习. 山西大学学报(自然科学版), 2020, 43(4): 803—816.

[18] Qu L H, Ma Y F, Luo X Y, et al. Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(10): 9732－9744.

[19] Tang W, Zhang W J, Zhang M L. Exploiting conjugate label information for multi-instance partial-label learning//Proceedings of International Joint Conference on Artificial Intelligence. Jeju Island, Korea (South): IJCAI, 2024: 4973－4981.

[20] Fourkioti O, De Vries M, Bakal C. CAMIL: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images//Proceedings of International Conference on Learning Represen - tations. Washington DC, USA: ICLR, 2024: 36205－36220.

[21] Tang W, Yang Y F, Wang Z F, et al. Multi-instance partial - label learning with margin adjustment// Proceedings of the 38th International Conference on Neural Information Processing Systems. Red Hook, NJ, USA: Curran Associates Inc., 2024: 26331－26354.

[22] Tang W, Yang Y F, Zhang W J, et al. Calibratable disambiguation loss for multi - instance partial - label learning. https://arxiv.org/abs/2512.17788, 2025－12－19.

[23] Luo H, Zhang Y X, Zhou Z, et al. Propensity scoring for multi-instance partial-label learning//Proceedings of APWeb-WAIM Joint International Conference on Web and Big Data. Shenyang, China: Springer, 2026: 1－14.

[24] 安曾,志富帅,丹潘,等. 基于多示例学习与多尺度特征融合的阿尔茨海默病分类诊断模型. 生物医学工程学杂志, 2025, 42(1): 132.

[25] Zhang Y X, Zhou Z C, Liu W S, et al. Rethinking multi-instance learning through graph-driven fusion: A dual - path approach to adaptive representation// Proceedings of AAAI Conference on Artificial Intelligence. Menlo Park, NY, USA: AAAI, 2026: 28510－28518.

[26] Li W J, Yeung D Y. MILD: Multiple - instance learning via disambiguation. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(1): 76－89.

[27] Wu J, Pan S R, Zhu X Q, et al. Multi - instance learning with discriminative bag mapping. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1065－1080.

[28] Xu B C, Ting K M, Zhou Z H. Isolation set-kernel and its application to multi - instance learning// Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery, 2019: 941－949.

[29] Lin T C, Xu H T, Yang C Q, et al. Interventional multi - instance learning with deconfounded instance - level prediction//Proceedings of AAAI Conference on Artificial Intelligence. Menlo Park, NY, USA: AAAI, 2022: 1601－1609.

[30] Tang W, Zhang W J, Zhang M L. Disambiguated attention embedding for multi - instance partial - label learning//Proceedings of Advances in Neural Information Processing Systems. Red Hook, NJ, USA: Curran Associates, Inc., 2023: 56756－56771.

[31] Chen K T, Sun S L, Zhao J. Camil: Causal multiple instance learning for whole slide image classification// Proceedings of AAAI Conference on Artificial Intelligence. Menlo Park, NY, USA: AAAI, 1120－1128.

[32] Yang M, Chen T L, Wu W Z, et al. Dual-perspective multi - instance embedding learning with adaptive density distribution mining. Pattern Recognition, 2025, 158: 111063.

[33] Wichitaksorn N, Kang Y Y, Zhang F Q. Random feature selection using random subspace logistic regression. Expert Systems with Applications, 2023, 217: 119535.

[34] Reutemann P, Pfahringer B, Frank E. A toolbox for learning from relational data with propositional and multi - instance learners//Advances in Artificial Intelligence. Berlin, Germeny: Springer, 2004: 1017－1023.

[35] Decencière E, Zhang X W, Cazuguel G, et al. Feedback on a publicly distributed image database: The messidor database. Image Analysis & Stereology, 2014, 33: 231－234.

[36] Zhou Z H, Sun Y Y, Li Y F. Multi-instance learning by treating instances as non - I. I. D. samples//

Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery, 2009: 1249－1256.

[37] Liu W, Luo W X, Lian D Z, et al. Future frame prediction for anomaly detection: A new baseline// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 6536－6545.

[38] Amores J. Multiple instance classification: Review, taxonomy and comparative study. Artificial Intelligence, 2013, 201: 81－105.

[39] Yang M, Zhang Y X, Wang X Z, et al. Multi-instance ensemble learning with discriminative bags. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(9): 5456－5467.

[40] Qian K, Min X Y, Cheng Y S, et al. Weight matrix sharing for multi-label learning. Pattern Recognition, 2023, 136: 109156.

[41] Qian K, Tang J Y, Zhao Q M, et al. Multi-label learning for fault diagnosis of pumping units with one positive label. Applied Soft Computing, 2025, 174: 113014.

[42] Demšar J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 2006, 7: 1－30.

(责任编辑 杨可盛)