自适应子空间插补的不完整数据证据集成分类

张　震[1,2], 春美洁[1], 田鸿朋[2], 李友好[3], 黄伟涛[3], 张俊杰[3]

(1.郑州大学 河南先进技术研究院,河南 郑州 450001;2.郑州大学 电气与信息工程学院,河南 郑州 450001;3.河南汇融油气技术有限公司,河南 郑州 450001)

摘　要:针对基于插补的分类方法在处理缺失数据时的估计值偏差会影响分类性能的问题,提出一种基于自适应子空间插补的不完整数据证据集成分类方法,利用自适应子空间插补和双重证据集成来提升模型对不完整数据集的分类能力。首先,使用谱聚类将特征空间动态划分为多个子空间,在每个子空间内独立进行基于近邻的缺失值插补;其次,设计了一种双重重要性评估机制,计算插补前后训练集数据分布的差异性来评估全局重要性,并通过评估分类模型对测试集样本在训练集中近邻的分类能力来评估其分类结果的局部重要性;最后,在证据理论的基础上融合局部重要性和全局重要性,利用不同子空间信息的互补性提升分类性能。在标准数据集上的对比实验表明:所提方法在 *ARI* 和 *AP* 指标中相比次优方法的提升幅度分别高达 6.23 百分点和 0.82 百分点,验证了所提方法的有效性和先进性。

关键词:不完整数据; 分类; 全局重要性; 局部重要性; 证据理论

中图分类号:TP181　　**文献标志码**:A　　**doi**:10.13705/j.issn.1671-6833.2026.04.006

数据缺失的原因多且难以有效避免,例如受访者出于保护隐私的目的拒绝提供某些信息、设备在某一时刻出现故障、调查时的失误导致信息遗漏。通常情况下这种存在缺失值的数据称之为不完整数据[1]。

近几年来,不完整数据分类问题备受关注,研究成果丰硕,其中对缺失数据处理方面主要有 3 种策略,最简单的策略是直接丢弃不完整的数据,仅使用完整的样本构建分类模型;另一种策略是模式基础法[2],估计输入数据的分布并将其用于模式分类,这种方法依赖分布假设的准确性,若数据偏离假设,分类性能可能显著下降;目前广泛应用的不完整数据分类策略是基于插补的方法[3],这种方法通过构建缺失值估计模型,实现对不完整样本的修复与再利用,其核心优势体现在能保留更多样本信息,从而最大化利用数据以及实现技术多样性。D-S 证据理论[4]是由 Dempster 在 1967 年最先提出的,已广泛应用于包括不完整数据分类任务[5-7]在内的一系列领域。利用 D-S 证据理论,研究者们不仅可以将样本划分给单个类别,也可以将样本划分给 2^{Ω} 中的任何一个子集,拓展了硬划分和模糊划分的现有概念。这种灵活的划分方法能够使人们更加深入了解数据,并提高对数据中异常值处理的鲁棒性。

虽然上述方法在不同场景下展现出一定的效果,但由于传统的插补方法难以有效捕捉高维数据中局部特征间的复杂关联,插补结果失真,会影响分类性能。传统单一评估方法因依赖单维度分析,对数据缺失敏感,在处理不完整数据时存在鲁棒性差、易产生偏差等固有缺陷。在多源决策中,传统的加权平均、多数投票等确定性融合策略无法有效量化和传递不确定性。

针对上述问题,本文提出了一种自适应子空间插补的不完整数据证据集成分类方法。首先,为了克服传统全局插补方法对特征局部相关性建模不足的缺陷,设计了一种自适应子空间插补方法,通过聚类方法将数据集划分为多个子空间,然后在每个子空间内独立进行插补;其次,采用了双重重要性评估机制,既增强了模型对整体数据分布的适应力,又强化了对局部样本分类的准确性,解决传统模型存在的误分类问题;最后,在证据推理的融合决策下[4],进一步整合了多模型优势,显著提升了分类结果的

收稿日期:2025-09-30;**修订日期**:2025-10-30

基金项目:河南省重点研发专项(231111211600);河南省国际科技合作重点项目(231111520300)

作者简介:张震(1966—),男,河南郑州人,郑州大学教授,博士,博士生导师,主要从事计算机视觉、数据挖掘、智能信息处理研究,E-mail:zhangzhen66@126.com。

鲁棒性和准确性。

1 相关工作

1.1 不完整数据插补方法

不完整数据插补是指当数据集中存在缺失值时,通过一定的方法估计并填充这些缺失值的过程。常用的不完整数据插补方法有3种。第一种是基于统计的传统插补方法,计算简单但无法捕捉复杂的数据关系。其中,均值插补用特征的均值填充缺失值,如Razavi-Far等[8]提出EM算法,通过融合模块完成插补值的融合。第二种是基于机器学习的插补方法,利用模型预测缺失值,典型方法有KNN[9]、随机森林插补,这类方法能够捕捉复杂特征关系。第三种是基于深度学习的插补方法及一些前沿方法,如生成对抗性插补网络GAIN[10]、DiffImpute[11]、HOEC[12]这类方法能自动学习复杂的数据表示。

1.2 D-S证据理论

假设Ω是一个识别框架,或称为假设空间,在证据推理中,识别框架从$\Omega=\{\omega_1,\omega_2,\ldots,\omega_c\}$扩展到幂集$2^{\Omega}$,其包含了$\Omega$的所有子集。一个证据的基本信任分配(basic belief assignment,BBA)指从幂集2^{Ω}到[0,1]上的一个映射函数$m(\cdot)$:$2^{\Omega}\rightarrow[0,1]$,并满足以下条件:

$$\begin{cases}\sum\limits_{G\in 2^{\Omega}} m(G)=1;\\ m(\varphi)=0。\end{cases}\tag{1}$$

如果$m(G)>0$,则称G为焦元;如果$m(G)=\max(m(\cdot))$,则称G为主焦元。

设识别框架上有两个独立证据L和H,他们的mass函数分别为m_1,m_2时的Dempster合成规则为

$$m_1\oplus m_2=\frac{1}{Z}\sum_{L\cap H=G} m_1(L)m_2(H)。\tag{2}$$

式中:Z为归一化常数即矛盾因子,计算公式为

$$Z=\sum_{L\cap H\neq\varphi} m_1(L)m_2(H)=1-\sum_{L\cap H=\varphi} m_1(L)m_2(H)。\tag{3}$$

2 本文方法

2.1 自适应子空间插补

假设数据集的训练集和测试集样本都存在缺失值,训练集$Y=\{y_1,y_2,\ldots,y_m\}$,共包含了C个类别,测试集$X=\{x_1,x_2,\ldots,x_n\}$。本文针对训练集和测试集均存在缺失值的情况,提出了一种基于特征相关性聚类的自适应缺失值填补方法。假设一个数据集有p个样本,6个属性,分别为A、B、C、D、E、F,具体实现过程如下。

首先,利用数据集中已知属性使用皮尔逊相关系数[13]ρ_{XY}计算这6个属性间的相关性来构建相关性矩阵$\boldsymbol{R}$:

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{E(X-EX)E(Y-EY)}{\sqrt{D(X)}\sqrt{D(Y)}}。\tag{4}$$

式中:E为数学期望;D为方差;$\sqrt{D}$为标准差;$\mathrm{Cov}(X,Y)$称为两个随机变量间的协方差,用来衡量两个变量的总体误差;ρ_{XY}为两个变量之间的协方差和标准差的商值,也称为变量X与Y的相关系数,X与Y在本文中是两个独立的属性。

其次,使用谱聚类算法[14]将这6个属性中具有强相关性的属性划分到同一子空间。本文前期构建的相关性矩阵可直接作为谱聚类的输入。谱聚类的优化目标与本文意图高度一致,能最大化子空间内特征连接的强度,同时最小化子空间之间的连接。谱聚类具有更高的计算效率,尤其适用于中等规模的特征划分问题。谱聚类要求输入矩阵为非负的相似度矩阵$\boldsymbol{S}$,而相关性矩阵$\boldsymbol{R}$中包含负值,因此需要进行映射转换。采用线性缩放将相关性值域[-1,1]、[-1,1]分别映射到相似度值域[0,1]、[0,1]。这是最直接且常用的方法,具体公式如下:

$$S_{uv}=\frac{R_{uv}+1}{2}。\tag{5}$$

式中:当$R_{uv}=1$时,$S_{uv}=1$,完全相似;当$R_{uv}=-1$时,$S_{uv}=0$,完全不相似;当$R_{uv}=0$时,$S_{uv}=1/2$,中等程度相似。

其次,使用轮廓系数[15]辅助选择最优的簇数。假如属性A、C、F之间相关性高,其次是B、E,单独剩一个D属性,则将这6个属性划分到3个子空间中。

最后,在每个特征子空间内,基于样本间的相似性关系,采用KNN进行缺失值插补。这种方法既保留了全局特征关联性,又能在局部子空间内实现更精准的样本匹配,从而显著提高了缺失值填补的准确性,具体流程图如图1所示。此外,还对Diabets数据集[16]的3个重要特征进行插补前后对比分析,从图2中可以看出,插补后的数据既未因过度拟合原始分布而丢失潜在规律,也未因填补导致特征分布失真,而是保留了数据固有特性,为后续的重要性评估和证据融合提供了更可靠的特征基础。

2.2 双重重要性评估

完成训练集、测试集的自适应子空间插补后,得到T个子集$\{Y_1,Y_2,\ldots,Y_T\}$。每个训练子集训练1个随机森林分类模型[17],可用于对测试集的待测样本进行分类:

$$P_i^t=\Gamma^t(x_i\mid Y^t)。\tag{6}$$

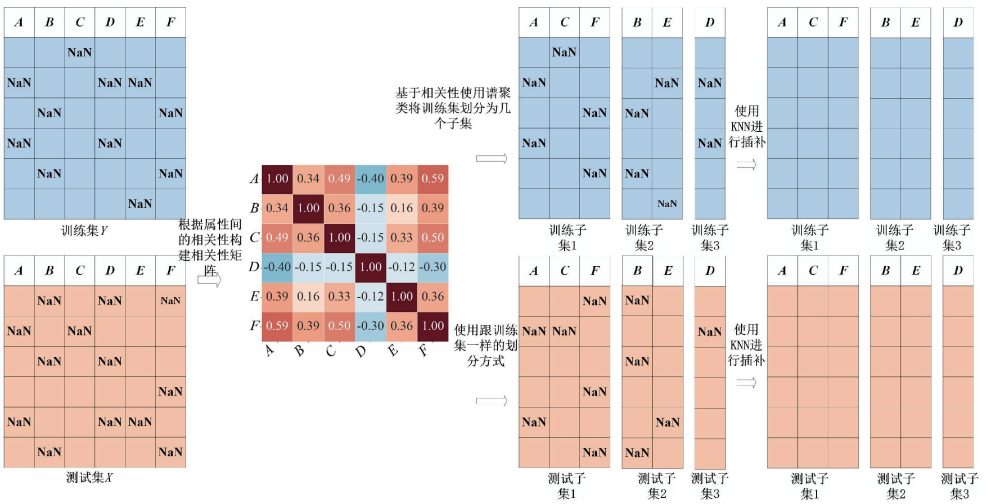
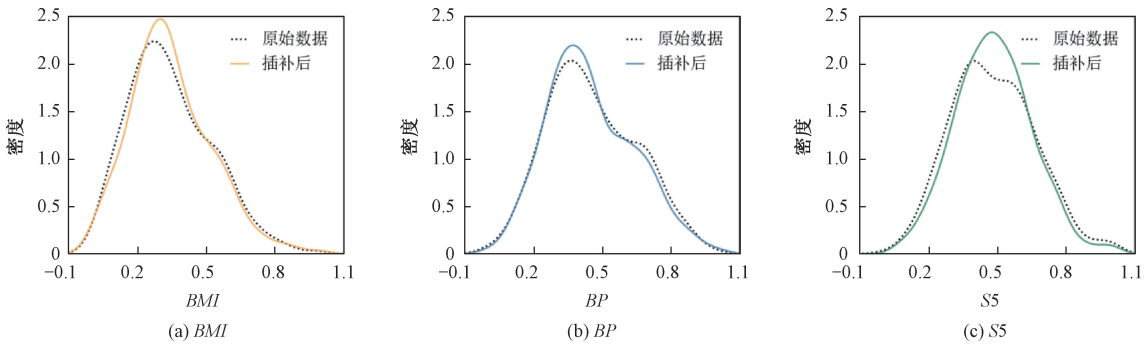


图 1　自适应子空间插补流程图

Figure 1　Flowchart of adaptive subspace imputation



(a) *BMI*　(b) *BP*　(c) *S5*

图 2　插补前后特征分布对比图

Figure 2　Comparison chart of feature distribution before and after imputation

式中:Γ^t 为随机森林模型;Y^t 为第 t 个训练子集;P_i^t 为 x_i 的分类结果。

在分类任务中,通常假定训练数据和测试数据是独立同分布的。然而对于不完整数据而言,将其划分为多个子集后,所得到的训练集和测试集的数据分布难免会存在一定偏差。为了评估这种分布差异对分类结果重要性的影响,本文采用 JS 散度[18]来衡量插补前后训练集与测试集之间的分布差异。因此,每个训练模型所得到的分类结果的全局重要性因子 γ_{it} 可定义为

$$\gamma_{it}=\frac{T}{\sum_{t=1}^{T}\mathrm{JS}(fre_{t,i}^{\mathrm{train}},fre_{t,i}^{\mathrm{test}})+\varepsilon}。\tag{7}$$

式中:$fre_{t,i}^{\mathrm{train}}$ 为训练集中第 t 个子子集中第 i 个特征的概率分布;$fre_{t,i}^{\mathrm{test}}$ 为测试集中第 t 个子子集中第 i 个特征的概率分布;ε 为误差项。

如果测试集样本的分布与训练集高度相似,则模型在训练阶段学到的决策边界能够较好地泛化到测试数据,从而提高预测的全局重要性。

在不完整数据分类中,局部重要性用于评估模型在特定数据区域中的预测稳定性和可信度。其核心思想是通过比较测试集样本的预测结果与其在训练集特征空间邻域内真实标签分布的差异,动态衡量模型局部预测的可靠性。作为重要性评分,分数越高,预测越可靠。由于原始特征尺度差异显著,需要根据数据分布选择合适的归一化方法,以保证近邻搜索时各特征贡献均衡,并通过高斯核加权使邻域标签分布更准确反映局部结构。

因此,每个训练模型所得到的分类结果的局部重要性因子 β_{it} 可定义为

$$\beta_i^t=\mathrm{JS}_i^t\Big/\sum_{t=1}^{T}\mathrm{JS}_i^t;\tag{8}$$

$$\mathrm{JS}_i^t=\sum_{j=1}^{K}\mu_{ij}\mathrm{JS}_{ij}^t。\tag{9}$$

式中:μ_{ij} 为每个近邻的权重;i 为样本的索引;JS_{ij}^t 为逆散度加权。

2.3　证据动态集成

得到样本 x_i 的全局重要性因子和局部重要性因子后,假设样本 x_i 有 T 个结果,在证据推理框架下每个分类结果都可以作为一个证据源;此外,每个

证据源都有一个全局重要性和局部重要性。为了整体评估 T 个分类结果的重要性,本文定义了一种复合重要性因子 λ_{it}[19]:

$$\lambda_{it} = (\gamma_{it} + \beta_{it}) \Big/ \sum_{t=1}^{T} (\gamma_{it} + \beta_{it}) \tag{10}$$

在获得复合重要性因子后,使用经典的证据折扣法根据 λ_{it} 来对每个分类结果进行折扣处理,定义如下:

$$\begin{cases} m_i^t(\omega_c) = \lambda_{it} P_i^t(\omega_c); \\ m_i^t(\Omega) = 1 - \sum_{\omega_c \in \Omega} \lambda_{it} P_i^t(\omega_c)。 \end{cases} \tag{11}$$

式中:$m_i^t(\cdot)$ 为基本信任值,类似于概率框架下的分类概率,表示待测数据属于某一类别的置信度;Ω 为全集,也称为完全未知类。

基于重要性的折扣通过修改证据能够有效降低证据间冲突,被折扣后的第 i 类分类结果 $P_i(\omega_c)$[19] 定义如下:

$$P_i(\omega_c) = \frac{\sum_{m_u \cap m_v} m_i^1(\omega_u) m_i^2(\omega_v)}{1 - \sum_{m_u \cap m_v = \varnothing} m_i^1(\omega_u) m_i^2(\omega_v)}。 \tag{12}$$

证据理论中,pignistic 概率[20]是将基本信任分配转化为经典概率分布的方法,这 t 个分类结果可能将样本目标分给不同的类别,经过 D-S 融合后,pignistic 概率计算调整为

$$BetP_i^t(\omega_c) = \sum_{\omega_c \in \Omega} \frac{m_i^t(\Omega)}{|\Omega|} + m_i^t(\omega_c)。 \tag{13}$$

式中:$|\Omega|$为 Ω 中元素的数量;$BetP(\omega_c)$ 为测试集样本可以分配到概率最大的类别。

3 实验及结果分析

3.1 数据集

本文采用 UCI 数据库[16]中 10 个具有代表性的数据集进行实验验证,这 10 个数据集最初没有不完整的数据。将原始的完整数据先划分为训练集和测试集,然后按照训练集和测试集各随机缺失 20%数据的方式进行处理。采用 10 次独立重复实验之后的平均值来消除随机因素对实验结果的影响。表 1 详细列出了实验所用的不完整数据集基本信息。

3.2 评价指标和对比方法

在实验中,本文用 *ARI*[21] 和 *AP*[22] 作为评价指标评估不同的不完整数据分类方法的性能。*ARI* 衡量预测标签与真实标签之间的相似性,校正了随机分配的影响;*AP* 值从排序角度评估模型性能,综合考虑不同召回率下的精准率,适用于需要衡量预测多个含置信度排序质量的场景。

表 1 不完整数据集基本信息

Table 1 Basic information of incomplete datasets

数据集	样本数量	属性个数	ω_1	ω_2
Tic	957	9	625	322
Stalog	270	13	150	120
Ionosphere	351	34	225	126
Diabets	520	16	320	200
Ecoli	335	7	142	77
Acoustic	400	50	100	100
Raisin	900	7	450	450
Credit	689	15	383	306
Biodeg	1 055	41	699	356
HTRU_2	17 898	8	16 259	1 639

注:其中,Ecoli 的 $\omega_3, \omega_4, \cdots, \omega_8$ 分别为 52,35,20,5,2,2;Acoustic 的 ω_3, ω_4 均为 100。

采用若干典型的不完整数据插补方法进行对比实验。其中,MEAN[23] 和 EM[8] 是最基本的数据插补方法;MICE[24] 使用完整特征建立回归模型,通过多轮迭代逐步优化插补值;KNN[9] 对含缺失值的样本找到 K 个最相似的完整样本,用这些邻居的均值或者众数填充缺失值;LLA[25]、GAIN[10]、DiffImpute[11]、TGANMTL[26] 通过组合多个生成模型组件来实现数据的插补。

3.3 本文方法对比相关不完整数据分类方法

表 2 和表 3 分别为各方法在不同数据集分类的 *ARI* 与 *AP*。实验随机选取每个数据集的 80%的数据作为训练集,剩余 20%作为测试集,并对测试集施加随机缺失。

由表 2 和表 3 可知,本文方法在绝大多数情况下均显著优于对比方法,本文方法比次优方法的 *ARI* 和 *AP* 分别相比次优方法提升了 6.23 百分点和 0.82 百分点 。其优势主要源于基于自适应子空间的插补更逼近真实完整数据,以及双权重融合与证据集成有效降低了分类错误。本文方法在 Credit 数据集上的效果提升不佳,可能是由于 Credit 数据集存在明显的类别重叠与模糊决策边界,导致基于局部邻域的一致性评估面临困难。样本在边界区域预测不确定性高,JS 散度难以区分进一步降低了重要性评分的区分度。

为评估所提方法在不同缺失率下的分类性能,在 Biodeg 数据集上与其他方法进行对比,结果如图 3 所示,可以看出,在 10%~60%的缺失率范围内,所提方法始终优于其他方法,验证了其处理不完整数据分类任务的有效性与鲁棒性。

表 2　不同方法在不同数据集分类的 *ARI*

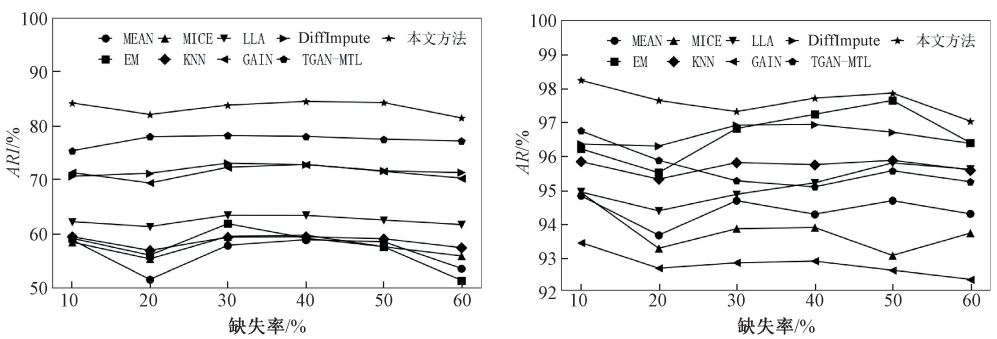
Table 2　*ARI* of different datasets classified by various methods　%

数据集	MEAN	EM	MICE	KNN	LLA	GAIN	DiffImpute	TGAN-MTL	本文方法
Tic	34.05	32.84	33.47	29.57	41.25	47.35	42.67	52.48	66.32
Stalog	96.03	95.34	96.21	93.67	97.24	98.14	98.54	99.07	99.58
Ionosphere	72.95	72.89	73.41	73.93	78.47	82.33	87.09	92.32	98.85
Diabets	72.69	69.47	70.35	74.47	80.44	81.35	85.42	88.63	93.11
Ecoli	29.35	31.65	32.43	33.13	37.56	35.33	40.10	45.18	57.19
Acoustic	45.79	51.37	52.36	47.07	54.71	41.95	60.35	68.52	81.47
Raisin	73.67	76.92	77.35	73.72	79.98	65.27	81.24	90.76	93.83
Credit	38.55	37.41	38.32	39.68	40.31	33.54	39.87	39.75	39.94
Biodeg	51.55	56.05	55.37	56.94	61.35	69.43	71.24	77.98	82.12
HTRU_2	83.95	84.07	84.31	83.24	87.24	85.67	81.53	89.67	94.29
平均值	59.86	60.80	61.36	60.54	65.86	64.04	68.81	74.44	80.67

表 3　不同方法在不同数据集分类的 *AP*

Table 3　*AP* of different datasets classified by various methods　%

数据集	MEAN	EM	MICE	KNN	LLA	GAIN	DiffImpute	TGAN-MTL	本文方法
Tic	81.19	80.73	79.24	83.05	81.47	84.54	85.89	84.62	85.39
Stalog	80.47	84.84	80.36	84.27	84.24	86.91	83.74	86.07	87.96
Ionosphere	96.43	93.20	93.26	98.34	97.31	95.74	94.48	98.28	99.21
Diabets	93.10	94.09	91.27	96.30	94.26	96.32	95.27	97.40	98.24
Ecoli	76.88	75.22	79.31	80.25	78.63	79.85	85.69	82.47	82.43
Acoustic	78.37	80.34	81.45	79.97	81.13	85.77	81.64	83.64	83.81
Raisin	91.37	92.52	86.97	92.03	90.54	91.25	93.87	93.55	94.08
Credit	88.73	88.57	87.43	89.55	89.78	78.24	84.33	91.26	92.73
Biodeg	93.63	95.48	92.24	95.29	95.35	92.65	96.27	95.85	97.62
HTRU_2	90.06	91.45	88.33	92.32	91.90	77.54	92.30	93.51	93.36
平均值	87.02	87.64	85.99	89.14	88.46	86.88	89.35	90.67	91.48



(a) *ARI*　(b) *AP*

图 3　不同方法对不同缺失率的 Biodeg 数据集分类结果

Figure 3　Classification results for the Biodeg dataset with various missing rates by different methods

3.4　算法参数对分类性能的影响

在对比方法中,KNN、LLA 和本文方法均涉及参数 K。为分析 K 对算法性能的影响,分别选取 Stalog、Biodeg、Ionosphere 这 3 个数据集进行测试。如图 4 所示,X 轴表示近邻个数,Y 轴表示 KNN、LLA 和本文方法在不同近邻个数的 *ARI*。可以看出,随着 K 的增大,本文方法的结果都高于其他方法,且受 K 的影响较小,证实了本文方法对 K 的选择对于最终结果没有太大的影响。

3.5　消融实验

为验证各个模块的增益,在 Acoustic 数据集上进行了消融实验,结果如表 4 所示。

由表 4 可知,各模块单独使用时均能显著提升模型性能,自适应子空间插补通过优化缺失数据重建质量,使 *ARI* 提升 7.21 百分点;双重重要性评估凭借局部-全局协同分析机制,带来最大的单模块增

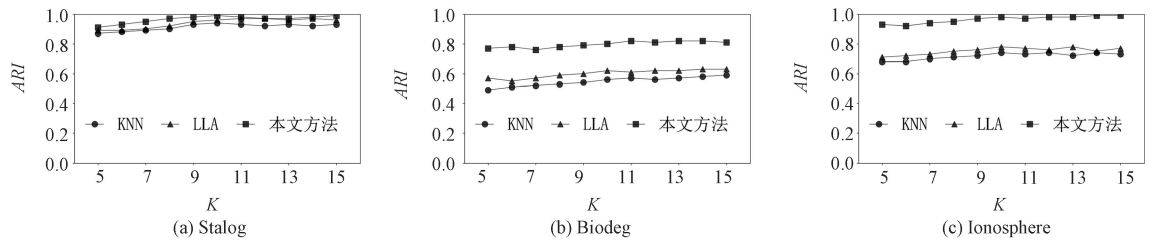


图 4　选择不同 *K* 时不同方法的 *ARI*

Figure 4　*ARI* of different methods with different *K*

表 4　消融实验

Table 4　Ablation experiments

自适应子空间插补	双重重要性评估	证据动态集成	*ARI*/%	*AP*/%
			47.97	79.07
√			55.18	80.21
	√		58.46	80.73
	√	√	67.24	82.46
√		√	69.31	82.95
√	√		70.29	83.04
√	√	√	81.47	83.81

益;证据动态集成通过多源信息融合,实现 *AP* 指标稳定提升。

3.6　计算复杂度

本文所提方法的计算开销主要集中在 3 个模块:在自适应子空间插补模块中,子空间 KNN 插补更占主导;在双重重要性评估模块中,局部模态的邻域搜索是该模块的主导项;在证据动态集成模块中,证据融合为该模块的主导项。整体计算复杂度为 $O(k\cdot N\_train^2+N\_test\cdot N\_train\cdot d+E\cdot N\_test\cdot C^2)$。其中,各项的权重取决于特征数、样本量、子空间维度等参数的实际取值。

表 5 为不同方法的执行时间。由表 5 可知,MEAN 只需要计算每个特征的均值,执行时间短但精度低;EM 需要多次迭代直至收敛;MICE 需要构建多个回归模型;执行时间随特征数平方级增长。KNN 的全局距离计算带来较大开销;相比之下,LLA 需要花费很长时间训练模型估计缺失值;GAIN、DiffImpute、TGAN-MTL 的计算时间相对较长,因为它们需要大量时间来训练网络模型;本文方法需要计算相当长的距离来搜索 KNN,所以执行时间很长。尽管本文方法在计算开销上有所增加,但显著提升了分类性能,且其运行时间仍处于可接受范围。

表 5　不同方法的执行时间

Table 5　Execution time of different methods

数据集	执行时间/s								
	MEAN	EM	MICE	KNN	LLA	GAIN	DiffImpute	TGAN-MTL	本文方法
Tic	0.723 1	1.473 1	1.326 3	0.972 3	2.416 5	5.384 6	4.982 7	7.385 1	6.572 4
Stalog	0.796 5	1.546 3	1.282 1	0.634 7	3.540 7	9.880 5	3.653 1	8.215 6	7.390 1
Ionosphere	0.869 3	8.319 9	7.795 8	1.099 9	4.376 3	10.804 2	5.830 2	11.237 4	10.944 9
Diabets	0.386 9	1.268 6	1.336 3	0.553 7	7.275 9	4.114 0	7.048 3	8.735 1	6.840 2
Ecoli	0.482 4	1.314 7	1.938 5	0.864 3	6.227 9	10.051 1	6.912 8	13.574 2	7.908 5
Acoustic	1.435 8	177.633 3	195.292 8	2.412 6	48.578 3	12.406 7	8.603 2	12.027 1	15.619 0
Raisin	0.701 8	1.135 6	1.017 7	0.735 3	2.894 4	9.519 2	10.317 9	15.329 4	11.899 4
Credit	0.441 8	1.283 9	1.370 6	0.794 4	5.176 8	11.067 1	11.651 7	14.035 7	13.047 1
Biodeg	0.822 7	177.482 5	198.969 0	3.623 4	13.362 9	6.747 5	7.822 5	10.728 4	23.201 0
HTRU_2	58.211 6	68.642 7	68.356 8	94.111 4	77.915 8	20.763 9	20.318 6	25.285 1	48.347 3

4　结论

本文提出了一种基于自适应子空间插补与双重证据集成的不完整数据分类方法,通过自适应子空间插补和局部-全局重要性评估,在插补阶段有效保留少数类局部特征,并在证据融合时引入动态加权机制以避免多数类主导,从而显著提升了模型在不完整和不平衡数据中的分类鲁棒性与准确性。在多个标准数据集上的对比实验表明,该方法相比传统算法具有明显优势。未来研究将进一步融合多种先进插补策略,发展更强大的集成学习框架,以提升模型在复杂实际场景中的分类精度。

参考文献:

[1]　吴晗, 王士同. 不完整数据分类与缺失信息重要性识别特权 LSSVM[J]. 智能系统学报, 2023, 18(4): 743-

753.
WU H, WANG S T. Privileged LSSVM for classification and simultaneous importance identification of missing information on incomplete data[J]. CAAI Transactions on Intelligent Systems, 2023, 18(4): 743-753.

[2] RADFORD A, METZ L, CHINTALA S, et al. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. (2016-01-07)[2025-09-19]. https://arxiv.org/abs/1511.06434.

[3] 徐鸿艳, 孙云山, 秦琦琳, 等. 缺失数据插补方法性能比较分析[J]. 软件工程, 2021, 24(11): 11-14.
XU H Y, SUN Y S, QIN Q L, et al. Comparative analysis of the performance of interpolation methods for missing data[J]. Software Engineering, 2021, 24(11): 11-14.

[4] SHAFER G. A Mathematical Theory of Evidence[M]. Princeton: Princeton University Press, 1976.

[5] TIAN H P, ZHANG Z W, DING W P. Incomplete data transfer calibration classification[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 9(5): 3244-3256.

[6] TIAN H P, WANG X L, TAN Y G. Incomplete data evidential classification with inconsistent distribution[J]. Information Sciences, 2024, 676: 120824.

[7] 段中兴, 毕瀚元, 张作伟. 基于D-S证据理论的不完整数据混合分类算法[J]. 信息与控制, 2020, 49(4): 455-463, 471.
DUAN Z X, BI H Y, ZHANG Z W. A D-S evidence reasoning based hybrid classification algorithm for incomplete data[J]. Information and Control, 2020, 49(4): 455-463, 471.

[8] RAZAVI-FAR R, CHENG B Y, SAIF M, et al. Similarity-learning information-fusion schemes for missing data imputation[J]. Knowledge-Based Systems, 2020, 187: 104805.

[9] 詹兆康, 胡旭光, 赵浩然, 等. 基于多变量时空融合网络的风机数据缺失值插补研究[J]. 自动化学报, 2024, 50(6): 1171-1184.
ZHAN Z K, HU X G, ZHAO H R, et al. Study of missing value imputation in wind turbine data based on multivariate spatiotemporal integration network[J]. Acta Automatica Sinica, 2024, 50(6): 1171-1184.

[10] YOON J, JORDON J, VAN DER SCHAAR M. GAIN: missing data imputation using generative adversarial nets[EB/OL]. (2018-06-07)[2025-09-19]. https://arxiv.org/abs/1806.02920.

[11] WEN Y Z, WANG Y W, YI K, et al. DiffImpute: tabular data imputation with denoising diffusion probabilistic model[C]//2024 IEEE International Conference on Multimedia and Expo (ICME). Psicataway: IEEE, 2024: 1-6.

[12] ZHANG Z, TIAN H P. Hybrid imputation-based optimal evidential classification for missing data[J]. Applied Intelligence, 2024, 55(1): 1-18.

[13] 陶洋, 祝小钧, 杨柳. 基于皮尔逊相关系数和信息熵的多传感器数据融合[J]. 小型微型计算机系统, 2023, 44(5): 1075-1080.
TAO Y, ZHU X J, YANG L. Multi-sensor data fusion based on Pearson coefficient and information entropy[J]. Journal of Chinese Computer Systems, 2023, 44(5): 1075-1080.

[14] 原虹, 张鸿雁. 基于高效谱聚类算法的文本特征分割研究[J]. 长江信息通信, 2025, 38(5): 171-173.
YUAN H, ZHANG H Y. Research on text feature segmentation based on the efficient spectral clustering algorithm[J]. Changjiang Information & Communications, 2025, 38(5): 171-173.

[15] 晏玲, 赵海良. 基于差分隐私 k-means++的一种隐私预算分配方法[J]. 信息安全研究, 2025, 11(8): 710-717.
YAN L, ZHAO H L. A privacy budget allocation method based on differential privacy k-means++[J]. Journal of Information Security Research, 2025, 11(8): 710-717.

[16] KELLY M, LONGJOHN R, NOTTINGHAM K. The UCI Machine Learning Repository[DB/OL]. [2025-09-19]. https://archive.ics.uci.edu.

[17] 付海涛, 张智勇, 王增辉, 等. 改进SHO算法优化随机森林模型[J]. 吉林大学学报(理学版), 2025, 63(3): 861-866.
FU H T, ZHANG Z Y, WANG Z H, et al. Improve SHO algorithm to optimize random forest model[J]. Journal of Jilin University (Science Edition), 2025, 63(3): 861-866.

[18] 李松, 刘晓楠, 刘娟. 基于JS散度的不确定数据密度峰值聚类算法[J]. 吉林大学学报(工学版), 2024, 54(7): 2038-2048.
LI S, LIU X N, LIU J. Peak clustering algorithm for uncertain data density based on JS divergence[J]. Journal of Jilin University (Engineering and Technology Edition), 2024, 54(7): 2038-2048.

[19] 田鸿朋, 张震, 张思源, 等. 复合可靠性分析下的不平衡数据证据分类[J]. 郑州大学学报(工学版), 2023, 44(4): 22-28.
TIAN H P, ZHANG Z, ZHANG S Y, et al. Imbalanced data evidential classification with composite reliability[J]. Journal of Zhengzhou University (Engineering Science), 2023, 44(4): 22-28.

[20] SMETS P. Decision making in the TBM: the necessity of the pignistic transformation[J]. International Journal of Approximate Reasoning, 2005, 38(2): 133-147.

[21] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193-218.

[22] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.

[23] 岳勇, 田考聪. 数据缺失及其填补方法综述[J]. 预防医学情报杂志, 2005, 21(6): 683-685.
YUE Y, TIAN K C. Summary of data missing and its filling methods[J]. Journal of Preventive Medicine Information, 2005, 21(6): 683-685.

[24] VAN BUUREN S, GROOTHUIS-OUDSHOORN K. Mice: multivariate imputation by chained equations in R [J]. Journal of Statistical Software, 2011, 45(3): 1-67

[25] 马宗方, 马祥双, 宋琳, 等. 异常信息的智能分类算法研究[J]. 计算机测量与控制, 2021, 29(10): 164-169.
MA Z F, MA X S, SONG L, et al. Incomplete data belief classification algorithm based on adaptive KNN imputation[J]. Computer Measurement & Control, 2021, 29(10): 164-169.

[26] LAI X M, ZHANG Z, CHEN H, et al. Tracking-removed neural network with graph information for classification of incomplete data [J]. Applied Intelligence, 2025, 55(3):1-20.

Ensemble Classification of Incomplete Data Evidence on Adaptive Subspace Imputation

ZHANG Zhen[1,2], CHUN Meijie[1], TIAN Hongpeng[2], LI Youhao[3], HUANG Weitao[3], ZHANG Junjie[3]

(1. Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou 450001, China; 2. School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China; 3. Henan Huirong Oil and Gas Technology Co., Ltd., Zhengzhou 450001, China)

Abstract: In response to the issue of biased estimates affecting classification performance in imputation-based classification methods when dealing with missing data, an incomplete data evidence ensemble classification method based on adaptive subspace imputation was proposed. The proposed method utilized adaptive subspace imputation and dual evidence integration to enhance the model′s classification ability on incomplete datasets. Firstly, spectral clustering was used to dynamically partition the feature space into multiple subspaces, where missing value imputation based on neighbors was performed independently within each subspace. Secondly, a dual importance evaluation mechanism was designed, which calculated the difference in data distribution before and after imputation in the training set to assess global importance, and evaluated the local importance of classification results by assessing the classification capacity of the classification model on the test set samples′ neighbors in the training set. Finally, based on evidence theory, local and global importance were fused to enhance classification performance by leveraging the complementarity of information from different subspaces. Comparative experiments on standard datasets showed that the proposed method achieved improvements of up to 6.23 percentage and 0.82 percentage, respectively, in the *ARI* and *AP* metrics compared to suboptimal methods, validating the effectiveness and advancement of the proposed method.

Keywords: incomplete data; classify; global importance; local importance; evidential reasoning