Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

数据集不平衡评估方法

硕士研究生 马西洋

2025年07月27日

问题回溯

- 总结反思
  - 时间过长，语速过快
  - 缺少互动
  - 基础知识较少

- 相关内容
  - 2018.11.25 胡雅娴：《机器学习中的数据不平衡问题》

内容提要

- 预期收获
- 内容引入
- 内涵解析与研究目标
- 研究背景与研究意义
- 研究历史与现状
- 知识基础
- 算法原理
  - MFII
- 特点总结与未来展望
- 参考文献

预期收获

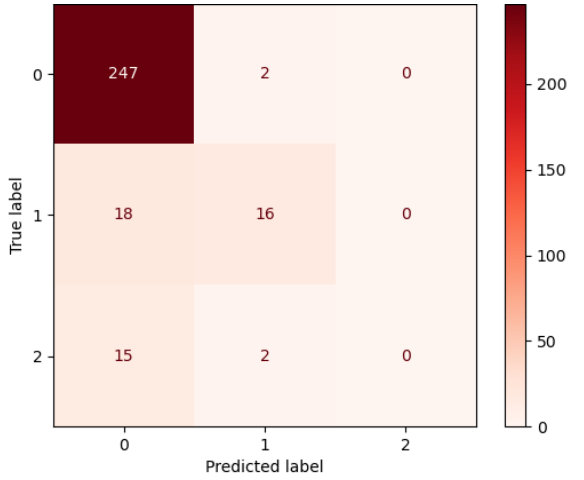
- 预期收获
  - 1. 了解数据集不平衡评估的基本概念
  - 2. 掌握数据集不平衡评估的基本原理
  - 3. 理解数据集不平衡评估的发展历程
  - 4. 了解数据集不平衡评估的未来发展

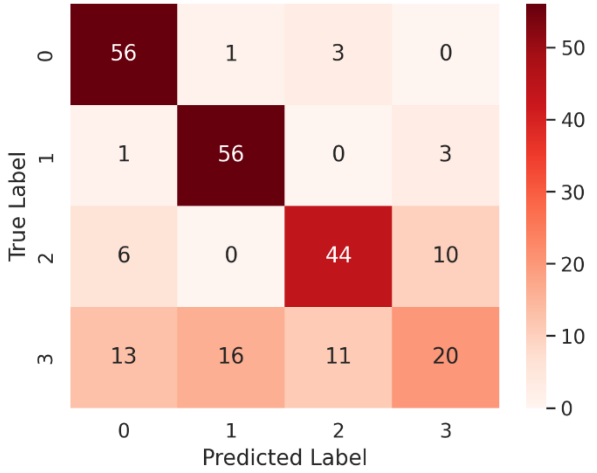
内容引入

- 在常见的多分类任务中，我们常以“平均准确率”衡量模型表现

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}$$

  - 模型关注集中“多数类”
  - “少数类”的准确率严重低下
- 传统定义：
  - Imbalance Ratio(IR) = 最大类样本数 / 最小类样本数
- 事实上，在真实任务中，即使两个类别样本数量相同，分类性能也可能出现天差地别
  - “样本数量比例”无法解释分类难度差异
- 什么才是真正意义上的“数据集不平衡”？





内涵解析与研究目标

- 内涵解析
  - 数据集不平衡不仅仅指类别样本数量的分布不均，更包含类别识别难度的差异
    - 数量不平衡：不同类别之间的样本数量差异显著
    - 结构不平衡：类别在特征空间中的分布形态差异，如边界复杂性、重叠区域、噪声比例等
- 研究目标
  - 以多类数据集为研究对象
  - 支持样本采样策略选择、损失函数设计与类别加权等典型下游任务
  - 解决现有方法存在面对结构性复杂数据时仅考虑样本比例、忽视样本结构的局限性

研究背景与研究意义

- 研究背景
  - 数据不平衡普遍存在于医学图像、文本分类等实际场景中
  - 传统的不平衡评估方式（如 Imbalance Ratio）仅关注样本数量，无法有效揭示不同类别的判别边界复杂度、结构嵌套或局部密度稀疏等导致的学习难度
  - 现有研究多聚焦于如何处理不平衡问题（如采样、损失函数等），但如何评估不平衡程度这一前置任务缺乏系统研究
- 研究意义
  - 构建能够全面反映数据集不平衡程度的评价指标
  - 准确刻画数据集中各类样本的分类难度
  - 为下游模型设计与训练策略提供依据
  - 弥补现有不平衡评估方法的局限，推动从“比例导向”向“结构感知”的评估转变

研究历史

2021

Barella在真实的不平衡数据集上评估了经典的数据复杂度度量，指出类别不平衡本身并不一定会导致性能下降，类别重叠和复杂边界等问题往往是罪魁祸首。并根据不平衡的情境动态调整了复杂度指标

2023

Seoane Santos等人构建类重叠复杂度度量的系统性分类体系，分为四大类代表性“重叠类型”：特征层面重叠、样本邻域重叠、结构形态重叠、多分辨率重叠，整合与分类现有复杂度指标并分析其对类别重叠的敏感性及是否考虑类别不平衡

2024

Aguilar 等人提出了一种全新的评估方法 Imbalanced Multi-class Classification Performance (IMCP) 曲线，专门用于评估多类且类别分布不平衡的数据集下的分类性能。该方法通过将每个样本的预测概率分布与真实概率分布计算Hellinger距离，并引入分布敏感的加权策略生成IMCP曲线，进一步提出其面积作为整体性能指标。

2023

Gøttcke等人提出了一种基于Tomek Link（托梅克链接）的复杂度评估指标，专注于捕捉“边界重叠”。通过计算少数类样本中有多少被卷入了与多数类样本形成的 Tomek Link（即互为最近邻且类别不同的样本对），来评估类间边界的混淆程度。TLCM的数值反映了边界重叠导致的分类难度，具有良好的可解释性与计算效率

2025

Cortes等人提出了一种解决类别不平衡问题的新方法，核心是设计了一种不对称的边界损失函数，对多数类和少数类分别设置不同的“置信边界”要求，让模型在训练时对少数类更加宽容、对多数类更严格。从理论上证明了这种损失在分类准确率下是一致的，比传统的代价敏感方法或加权交叉熵方法更可靠。还提出了一个统一的优化算法，在多个不平衡数据集上取得了优于现有主流方法

A new data complexity measure for multi-class imbalanced classification tasks

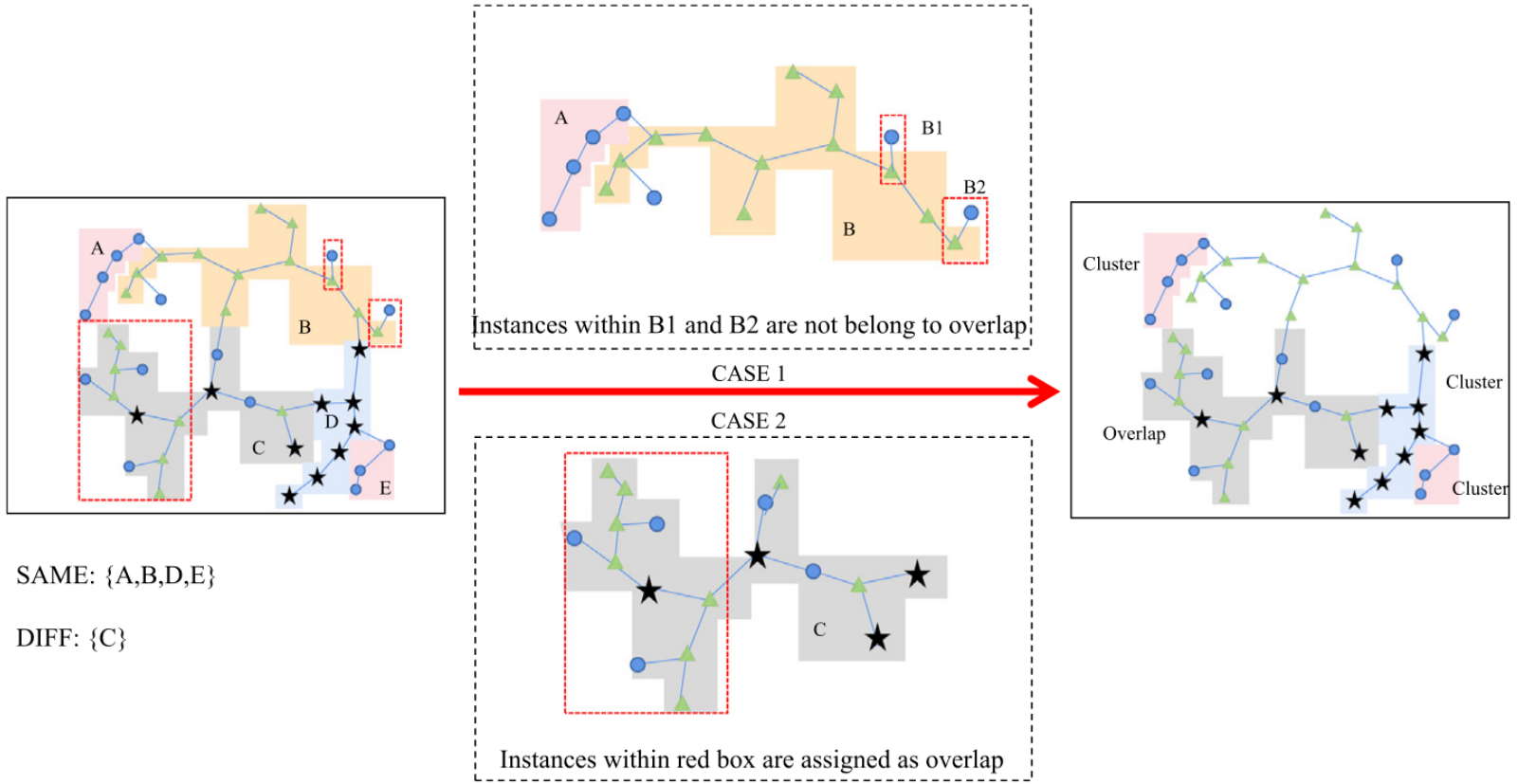
算法原理 TIPO

T	目标	评估多类分类任务中数据集的不平衡程度
I	输入	多类分类数据集
P	处理	1.构建样本之间的结构图 2.样本划分与类型判定 3.计算数据集不平衡指数
O	输出	数据集不平衡指数

P	问题	现有方法未能完全考虑不同数据特征之间的相互作用，或类别不平衡和这些特征之间的联系
C	条件	可获得每类样本的标签及其特征空间分布
D	难点	如何同时建模特征间交互与类别不平衡的结构关系
L	水平	2024中科院1区（Pattern Recognition）

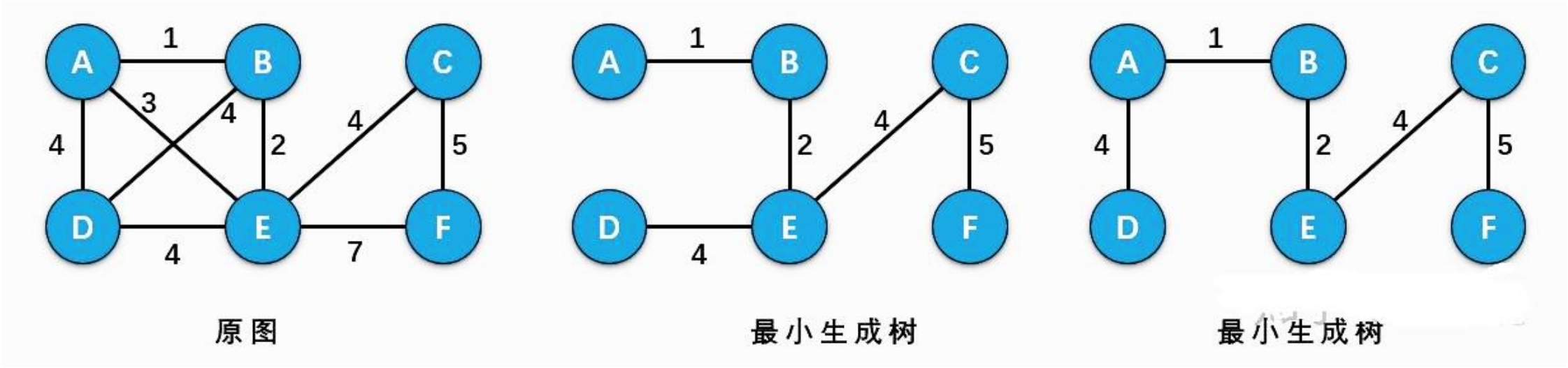
算法原理

- 算法原理图
  - 最小生成树构建：将样本视为节点，以样本间距离为边权，构建最小生成树
  - 样本划分与类型判定：将样本划分为簇或重叠，用于衡量类别混杂程度
  - 计算不平衡指数MFII



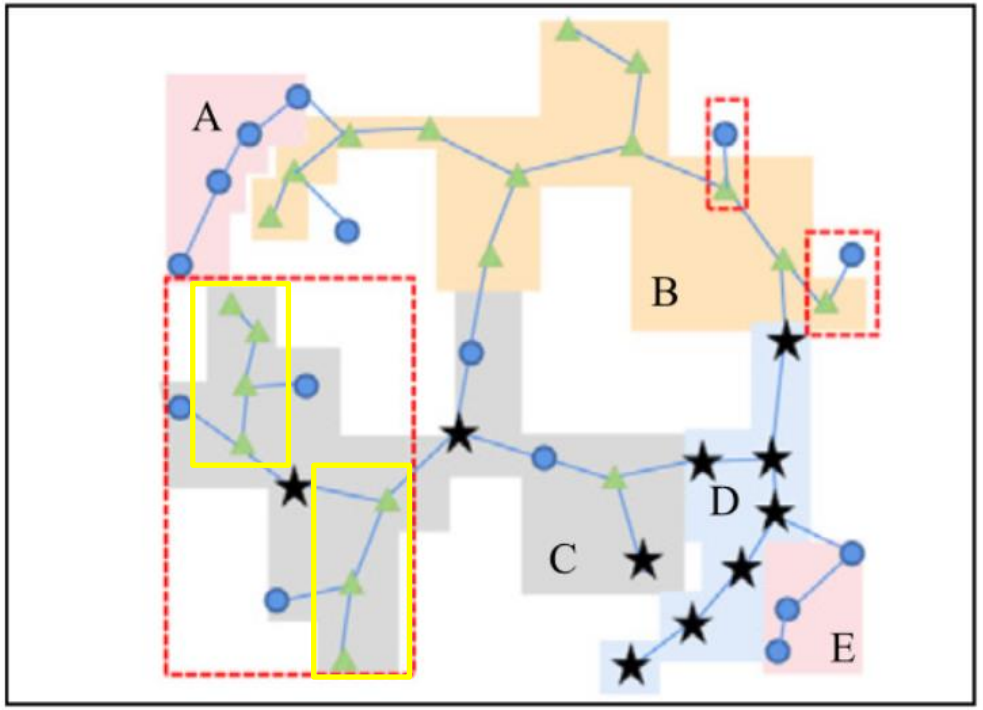
算法原理　最小生成树构建

- 最小生成树（Minimum Spanning Tree）
  - 生成树：一个连通图中选取部分边，形成一个包含所有顶点且无环的子图
  - 最小生成树：边权之和最小的生成树
- MST的作用
  - 提供最低冗余连接，仅保留最关键结构路径



算法原理　样本划分与类型判定

- 样本划分与类型判定
  - SAME：在最小生成树中，结构上连通且类别一致的子图集合
  - DIFF：结构上连通但类别混合的子图集合
  - Cluster：SAME中满足以下两个条件集合
    - 样本规模$|comp_i|$不超过类别i的样本总数的$1/k$（即$|comp_i| < n_i/k$）
    - 该子图中大多数样本与其相连样本类别一致（仅允许一个样本与其他类样本相连）
  - Overlap：
    - SAME中满足$|comp_i| < n_i/k$，但多个样本与其他类样本相连

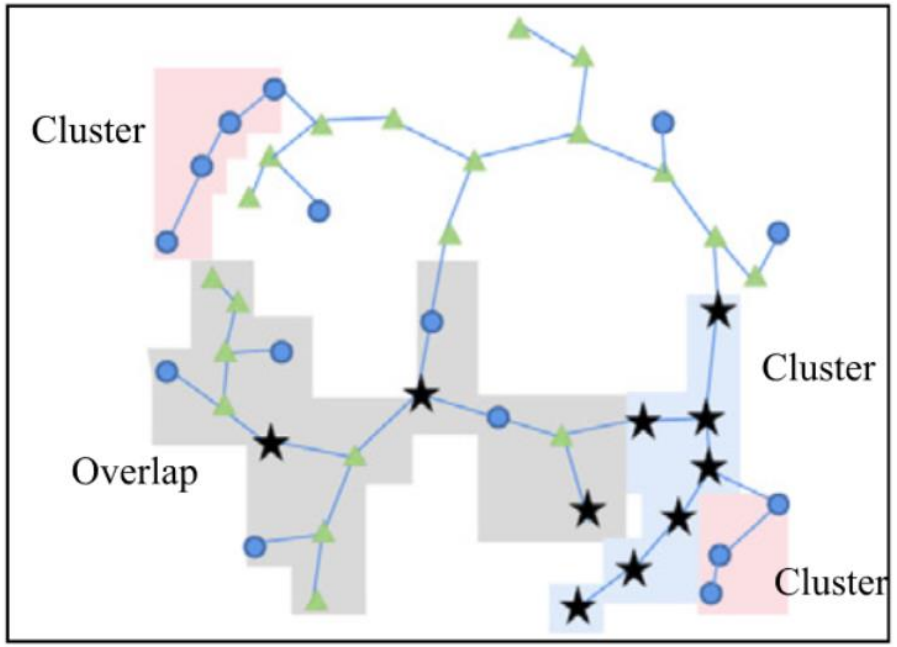


SAME: {A,B,D,E}

DIFF: {C}

算法原理　样本划分与类型判定

- 样本划分与类型判定
  - Overlap：
    - DIFF中满足$|comp_i| > k$，且其不平衡度$IR(comp_i)$小于数据集的平均不平衡度$IR(D)$
  - 分类难度源于两大核心问题：
    - 类内结构松散
    - 类间边界重叠
  - Cluster的划分刻画类的内部结构稳定性，Overlap的划分刻画类间干扰和边界模糊性；两者共同揭示了分类难度背后的结构性复杂性来源



算法原理　不平衡指数计算

- 多类不平衡指数（ MFII ）

$$MFII(\mathrm{D}) = \frac{1}{2(k-1)} \sum_{i=1}^{k} w_i^1 Factor_i^1 + w_i^2 Factor_i^2$$

- $Factor_i^1$：第i类的类内因素，衡量该类样本内部结构破碎程度
- $Factor_i^2$：第i类的类间因素，衡量该类与其他类重叠样本的影响

$$Factor_i^1 = \frac{1}{n_i \sum_{j=1}^{k} \frac{1}{n_j}} \frac{n_i^{cluster}}{n_i}$$

- $1/n_i \sum_{j=1}^{k} \frac{1}{n_j}$：合理增加少数类的贡献，同时防止极端放大
- $n_i^{cluster}/n_i$：反映重叠样本在类内的相对占比，即样本内部结构破碎程度

算法原理　不平衡指数计算

- 多类不平衡指数（ MFII ）

$$w_i^1 = \sum_{j=1,j\neq i}^{k} \frac{n_j}{n}\left(1 - \frac{|SAME_j| - n_j^{cluster}}{n_j}\right) = \sum_{j=1,j\neq i}^{k} \frac{n_j - (|SAME_j| - n_j^{cluster})}{n}$$

– w_i^1反应了当前类i所面临的类间结构干扰强度

$$Factor_i^2 = \frac{n_i^{overlap}}{n_i} + \frac{1}{n_i \sum_{j=1}^{k} \frac{1}{n_j}}\left(\frac{|DIFF_i| - n_i^{overlap}}{n_i}\right)$$

– $n_i^{overlap}/n_i$表示该类中被判定为重叠样本的比例

– $(|DIFF_i| - n_i^{overlap})/n_i$表示该类中被判定为重叠样本的比例

算法原理　不平衡指数计算

- 多类不平衡指数（ MFII ）

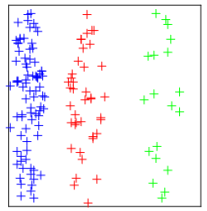
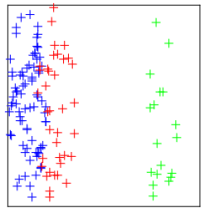
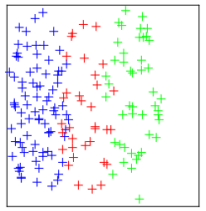
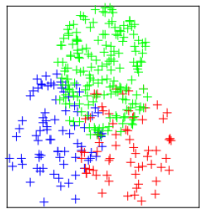
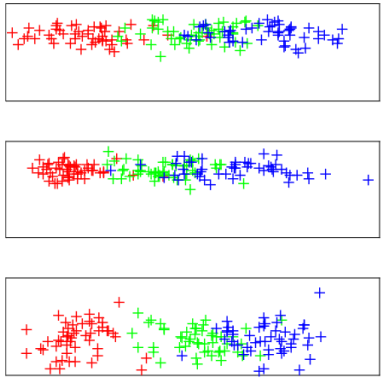
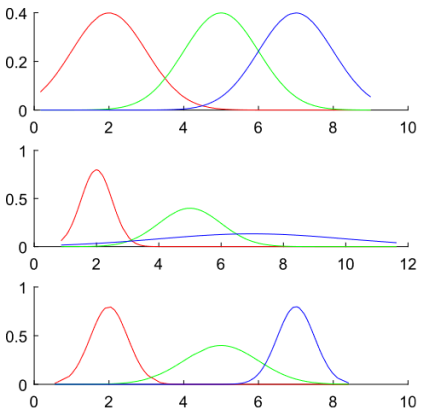
$$w_i^2 = \sum_{j=1, j\neq i}^{k} \frac{n_j^{overlap}}{n}$$

– w_i^2反应了其他类别中的模糊样本（overlap）的影响；

组成项	名称	描述	作用
$Factor_i^1$	子簇因子	类内样本聚集情况，子簇越多说明结构越分散	衡量类内结构复杂度
$Factor_i^2$	重叠因子	包括类别间重叠样本和DIFF中的噪声影响	衡量类别间模糊性
w_i^1	子簇权重	考虑其他类中子簇数量和类别大小	表示外部类结构对当前类的干扰
w_i^2	重叠权重	来自其他类中重叠样本的影响	表示外部类模糊度对当前类的影响

实验设计

- 数据资源
  - 合成数据：
    - 少数类样本数固定为 50
    - 三分类最大不平衡比 225:1
    - 四分类最大不平衡比 125:1

类别数量	分布类型	重叠度	不平衡比例
3类	均匀分布	0,10,20,50,80,100	(1,i,i),i=1,10,50,100
	高斯分布		(1,i,i^2),i=1,5,10,15
4类	均匀分布	0,10,20,50,80,100	(1,i, i^2, i^3),i=1,5
	高斯分布		(1,i, i^2, i^3),i=1,5

实验设计

- 数据资源
  - 真实数据： 包括特征维度、样本数量、类别数及不平衡比
    - 覆盖了从轻度到极度不平衡（IR从3到439.6），以及特征维度从低维（如Iris1）到高维（如movementlibars_12），类别数从3到26

Dataset	#Att.	#Ins.	#Class	IR	Dataset	#Att.	#Ins.	#Class	IR
svmguide4_2	10	390	6	3	ecoli0147v56	6	332	6	12.28
Iris1	4	95	3	3.3	vehicle1	18	365	4	13.3
wine_3	13	178	3	3.8	yeast	8	1484	10	23.15
svmguide2	20	391	3	4.17	Isolet	617	7794	26	25
segment_6	19	2310	7	5	yeast5	8	1484	10	32.73
balance	4	625	3	5.88	wine_red_8v67	11	855	3	46.5
vowell	10	540	11	6	winewhite_39v5	11	1482	3	58.28
led7digit_1v789	7	500	10	6.96	winered_134v5	11	1200	4	68.1
texture_12	40	5500	11	8	winequalityred	11	1599	6	68.1
glass	9	214	6	8.44	ecoli	7	336	8	71.5
penbased_8	16	1100	10	8.47	pageblocks	10	548	5	164
movementlibars_12	90	360	15	11	winequalitywhite	11	4898	7	439.6

实验设计

- 分类器
  - KNN，SVM，CART，RANDOR FOREST（RF）和ADABOOST
- 对比方法
  - 分为两大类：基于特征的方法与基于邻域的方法

	Measures	Definition
Feature-based	$F1$	$F1 = \frac{1}{1+\max\limits_{j=1,2,\ldots,m} \frac{(\mu_{1j}-\mu_{2j})^2}{\sigma_{1j}^2+\sigma_{2j}^2}}$
	$F3$	$F3 = \min_i^m \frac{n_{in}(f_i)}{n}$
Neighbor-based	$N1$	$N1 = \frac{1}{n}\sum_{i=1}^{n} I((x_i, x_{i1}) \in MST \wedge y_i \neq y_{i1})$
	$N2$	$N2 = \frac{1}{n}\sum_{i=1}^{n} \frac{d(x_i, NN_1(x_i) \in y_i)}{d(x_i, NN_1(x_i) \notin y_i)}$
	$N3$	$N3 = \frac{1}{n}\sum_{i=1}^{n} I(NN_1(x_i) \notin y(i))$
	LSC	$LSC = \frac{1}{n}\sum_{i=1}^{n} \frac{\lvert\{x_j \lvert d(x_i,x_j) < d(x_i, NN_1(x_i)) \cap NN_1(x_i) \notin y_i\}\rvert}{n}$
	$T1$	$T1 = \frac{n_{sphere}}{n}$

实验设计

- 分类性能评价指标

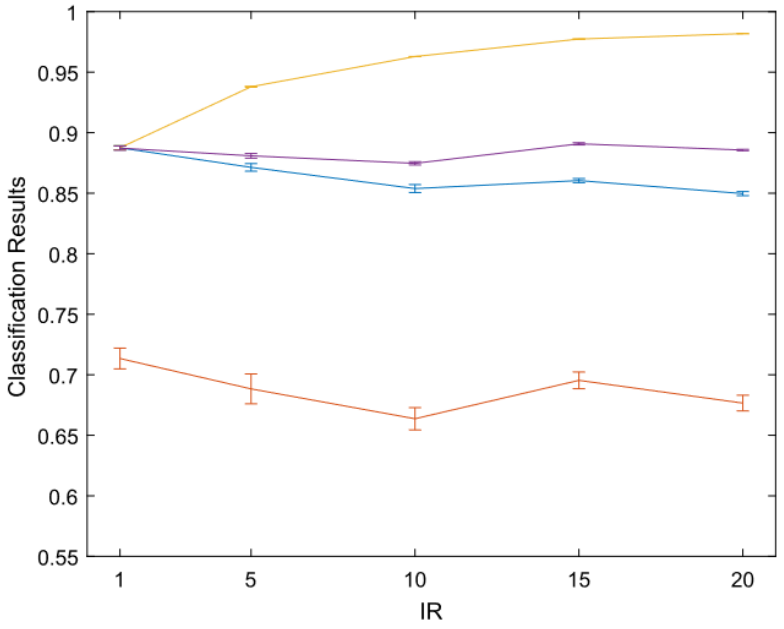
指标名称	全称	含义与说明
IAM	Imbalance-Aware Measure	综合考虑少数类与多数类的识别能力，适用于不平衡分类
CBA	Class-Balanced Accuracy	所有类别赋同等权重，避免多数类主导评估结果
MicroF1	Micro-averaged F1 Score	全局样本级评估，偏向多数类
G-mean	Geometric Mean	所有类别召回率的几何均值，衡量类别均衡性
Pave	Performance Average	各类性能均值，反映总体模型表现

- 数据复杂度指标

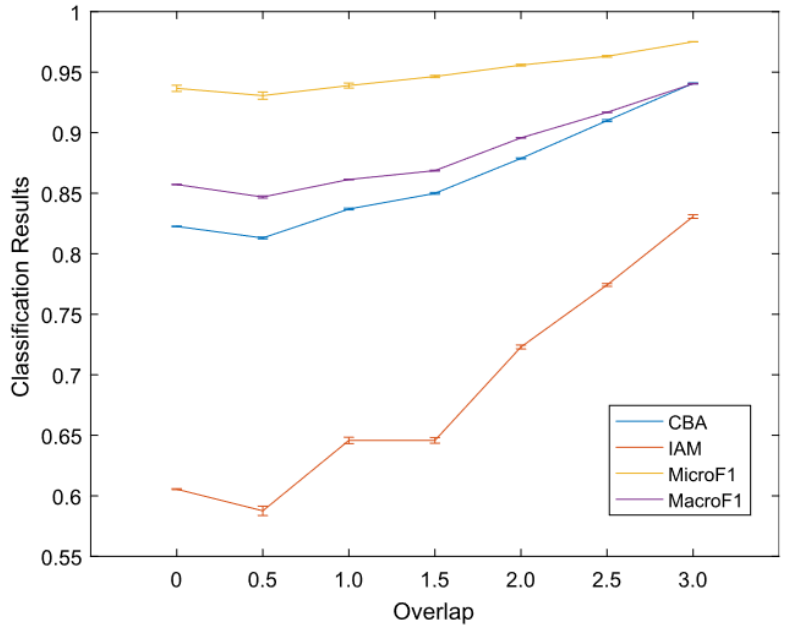
指标名称	含义	数学定义 & 解释
VoR(Value of Resolution)	分辨率能力	衡量数据复杂度值对于不同分类任务结果的分辨能力，VoR越小，说明复杂度值能更好地区分不同分类难度的任务
DoC(Degree of Consistency)	稳定性指标	衡量复杂度指标与实际分类结果之间的趋势一致性，DoC越大，说明复杂度度量值能更稳定地反映任务难度变化

实验分析　分类性能受什么影响?

- 实验结果
  - 在IR增大时，IAM和CBA受到显著影响，表现下降，表明其对类别不平衡更敏感
  - Overlap值越大，代表类间距离越大，越容易区分，几乎所有指标都显示出性能提升趋势



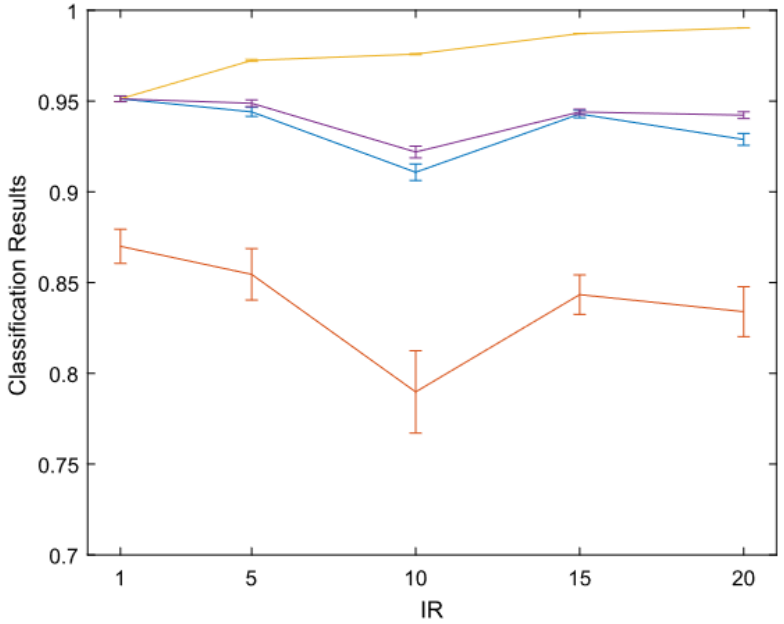
(a) under specific IR



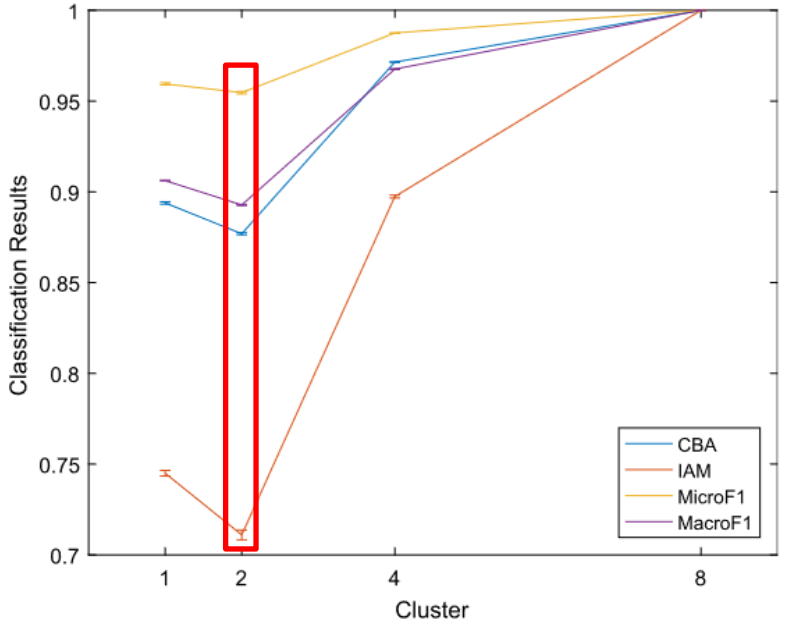
(b) under specific overlap

实验分析　分类性能受什么影响?

- 实验结果
  - 随着簇数量的增加，所有类的簇同步增加。随着簇数量增加，重叠样本数量减小，使每个群集与其他群集之间的区别更为明显
  - 必须同时考虑多个数据特征因素（比如IR、overlap和cluster），才能全面反映分类难度



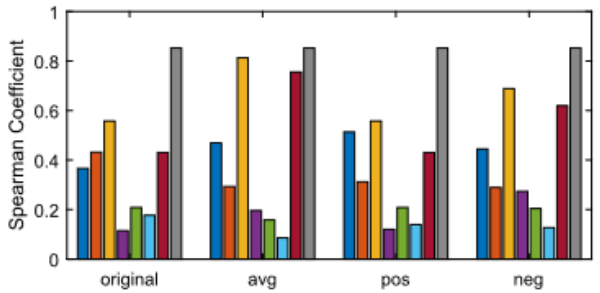
(a) under specific IR



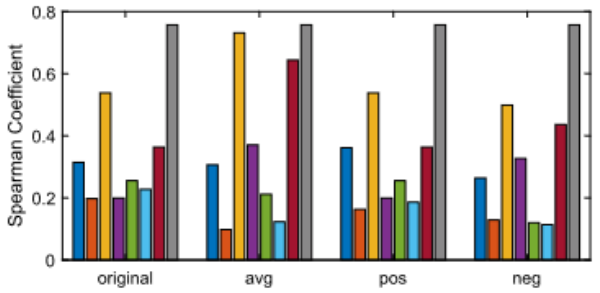
(b) under specific cluster

实验分析　真实数据集实验结果

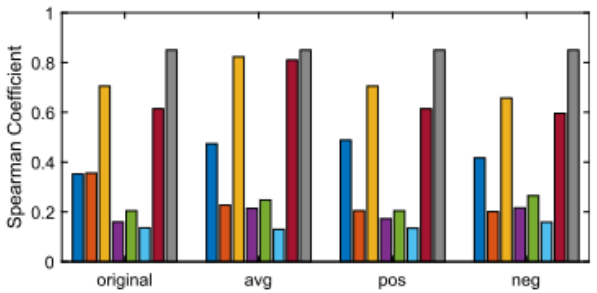
- 复杂度度量与不同分类算法的结果之间的相关系数
  - original、avg、pos、neg表示不同的度量策略或加权方式
  - 除了AdaBoost，MFII与实际分类表现的匹配程度最好



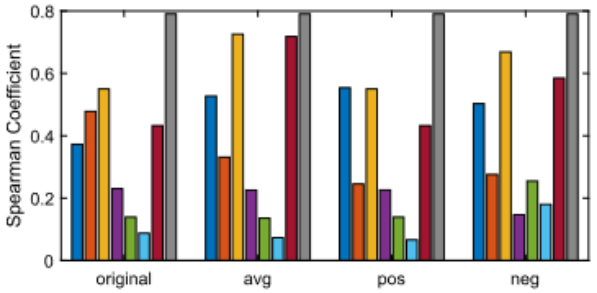
(a) KNN



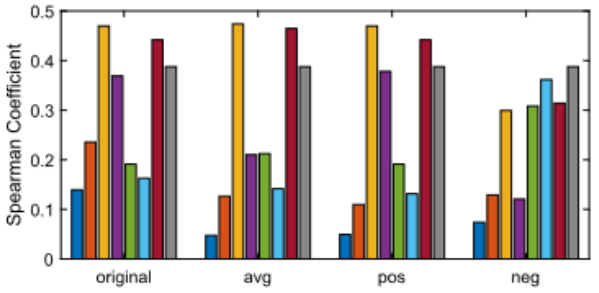
(b) SVM



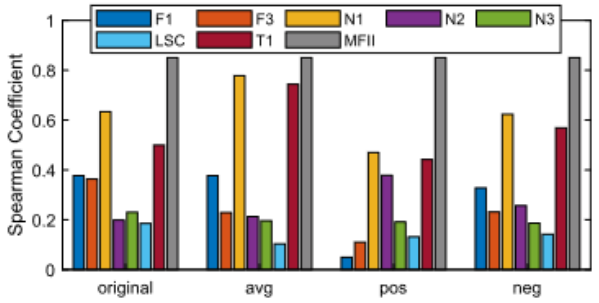
(c) DT



(d) RF



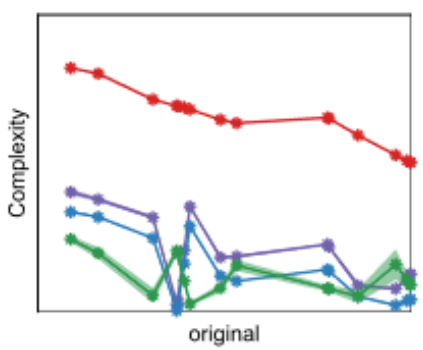
(e) AdaBoost



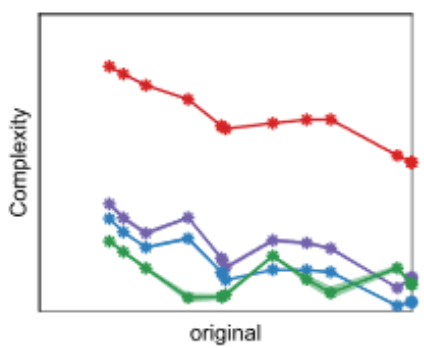
(f) Avg

实验分析　真实数据集实验结果

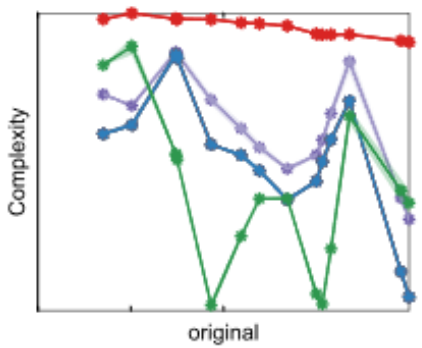
- 复杂度度量与分类性能的对应变化趋势
  - 不论是哪个分类器，MFII的变化幅度小，线条稳定，说明它的波动小、鲁棒性强。



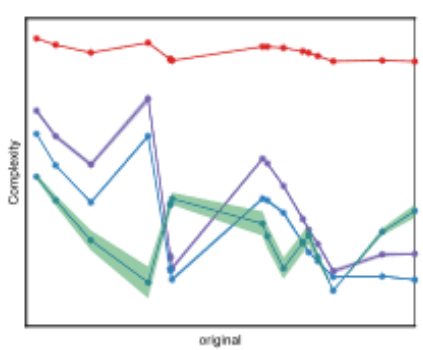
(a) kNN-org



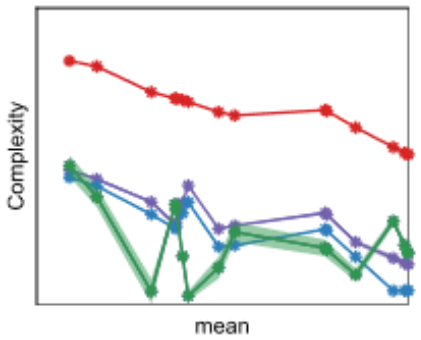
(b) SVM-org



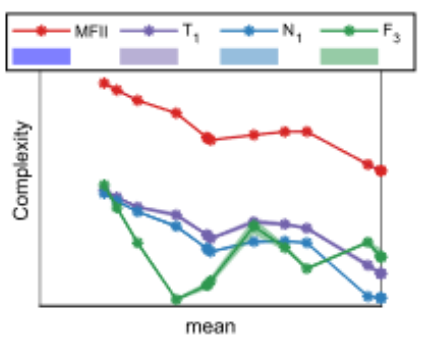
(c) DT-org



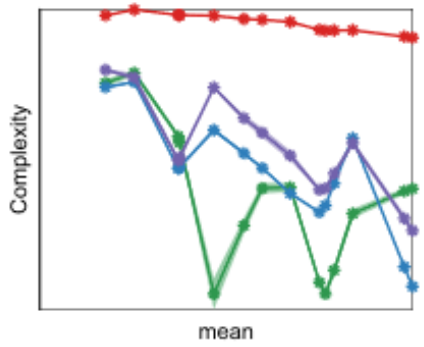
(d) AdaBoost-org



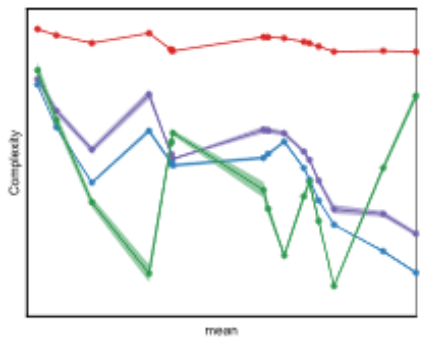
(e) KNN-avg



(f) SVM-avg



(g) DT-avg



(h) AdaBoost-avg

实验分析　真实数据集实验结果

- 不同数据复杂度度量的最佳性能
  - MFII保持了很好的性能
  - 将多个复杂度度量加权组合的策略也没有带来显著的性能提升，虽然这些指标从不同角度来刻画复杂度（比如边界模糊程度、样本密度等），但它们大多集中在“类别间重叠程度”这一单一维度，而忽略了其他类型的复杂性来源

	Spearman	Pearson	DoC	VoR	Mean.R
F1	0.3769	0.2092	0.0900	0.0023	8.00
F3	0.3633	0.3087	0.1575	0.0157	6.75
N1	0.7777	0.8242	0.1248	0.0030	4.00
N2	0.3782	0.1430	0.0849	0.0187	9.50
N3	0.2298	0.2372	**0.1734**	0.0187	7.50
LSC	0.1419	0.2637	0.0927	0.0187	9.25
T1	0.7436	0.7901	0.1717	0.0019	3.00
DC_{avg}	0.4464	0.4739	0.1405	0.0034	5.75
DC_{coe}	0.4869	0.5686	0.0936	0.0015	4.50
DC_{rnd}	0.4065	0.4894	0.0701	0.0014	6.00
MFII	**0.8499**	**0.8838**	0.1541	**0.0000**	**1.75**

特点总结与未来展望

算法总结

- 算法优势：
  - 更强的分辨能力、更高的稳定性、
  - 在合成、小规模、真实大规模数据集上均优于现有方法
- 局限性：
  - 依赖于样本在特征空间中的表达，效果取决于向量表示质量
  - 更复杂的任务（高维、多类别）中性能有轻微下降
  - 仅考虑了两种数据特性（cluster和overlap），尚未全面建模所有影响因素
- 未来展望
  - 多视角特征建模：挖掘更多影响分类难度的特征因素
  - 适配非向量型数据：在图神经网络、序列模型等深度学习架构中研究表示学习与复杂度之间的联动
  - 与样本难度（Instance Hardness）结合

参考文献

[1] L. Yu, Z. Wang, Y. Wang, J. Li, and D. Wang. A new data complexity measure for multi-class imbalanced classification tasks[J]. Information Sciences, 2024, 668: 93–112.

[2] Lorena A, Garcia L P F, Lehmann J, Souto M C P, Gama J. How complex is your classification problem? A survey on measuring classification complexity [J]. ACM Computing Surveys, 2019, 52(5): 1-34.

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！