



西安电子科技大学
XIDIAN UNIVERSITY

基于机器学习的医疗不平衡数据分类算法研究

赵梦楠



目录

CONTANTS

01/课题研究意义

02/KM-BRAF算法设计与实现

03/ KMBS-BRAF算法设计与实现

04/发热待查病因预测应用

05/总结与展望



课题研究意义



研究背景

老龄化

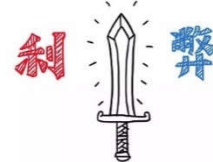
多样化

复合化



成本低

稳定性高



疾病诊断主要方式比较

诊断方式	优点	缺点
人工诊断	安全性高，诊疗经验丰富	培养成本高、稳定性较差、精力有限
机器学习	培养成本低、看诊成本低、稳定性高	缺乏信任、缺少感情、安全性低

不平衡分类发展现状

不平衡数据分类

数据预处理

欠采样

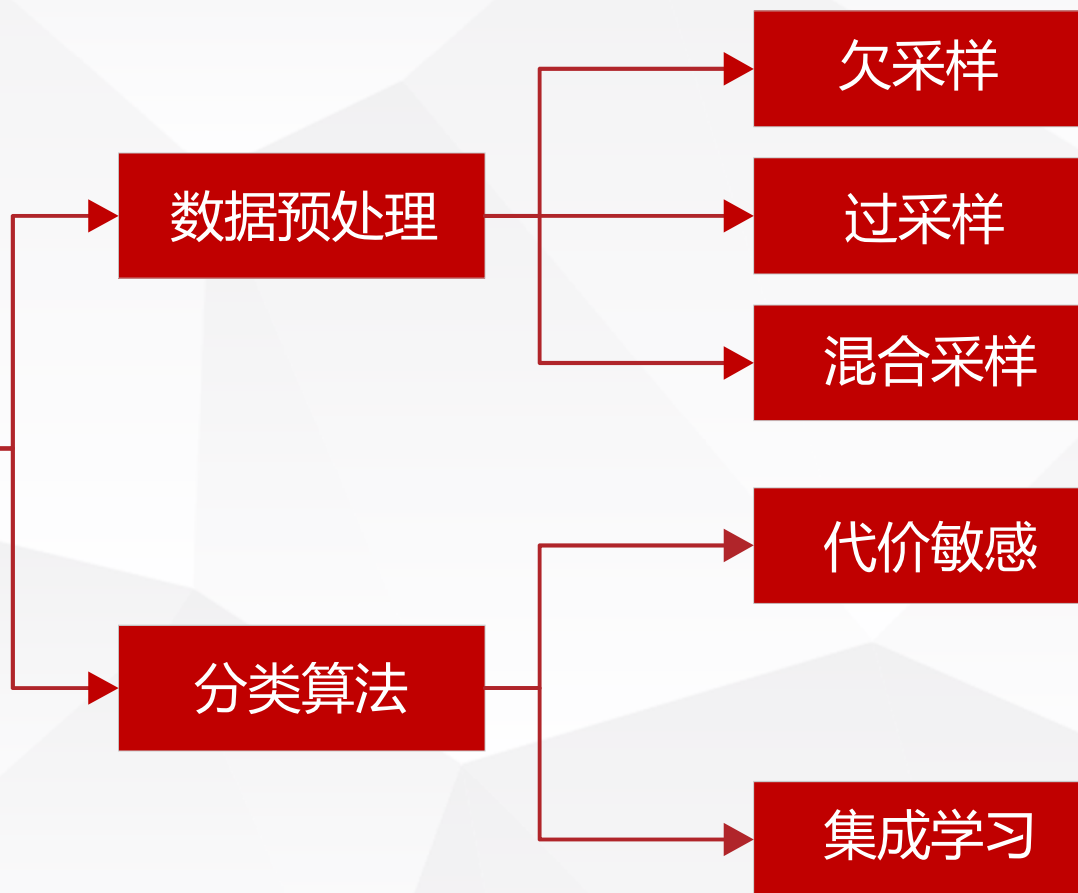
过采样

混合采样

分类算法

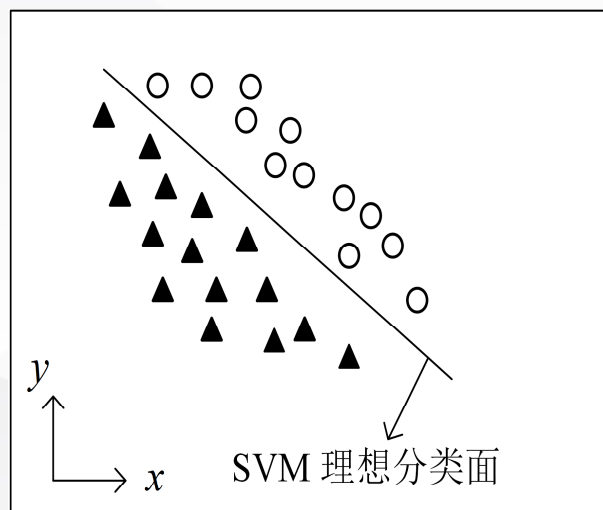
代价敏感

集成学习



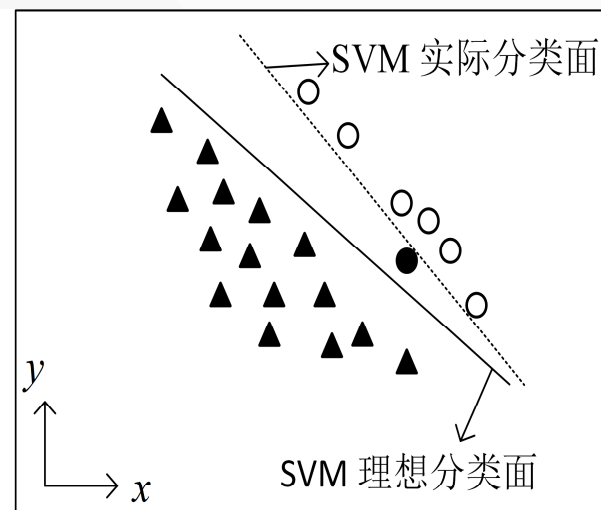
不平衡分类难点

经典分类算法都是基于数据集平衡这一前提的，因为只有在这一条件下，分类算法才可以在学习过程中对数据集中所有类别一视同仁。因此，这些算法在**不平衡数据集**上训练出的模型效果一般不符合预期，因为分类算法为了减少学习过程中的错误分类损失，将不可避免地对本样本数目比较多的类别存在一定程度的偏斜，从而导致**少数类样本**的分类效果较差。



(a) 理想分类平面

- ▲ 正类样本
- 负类样本
- 错分样本
- 理想分类面
- 实际分类面



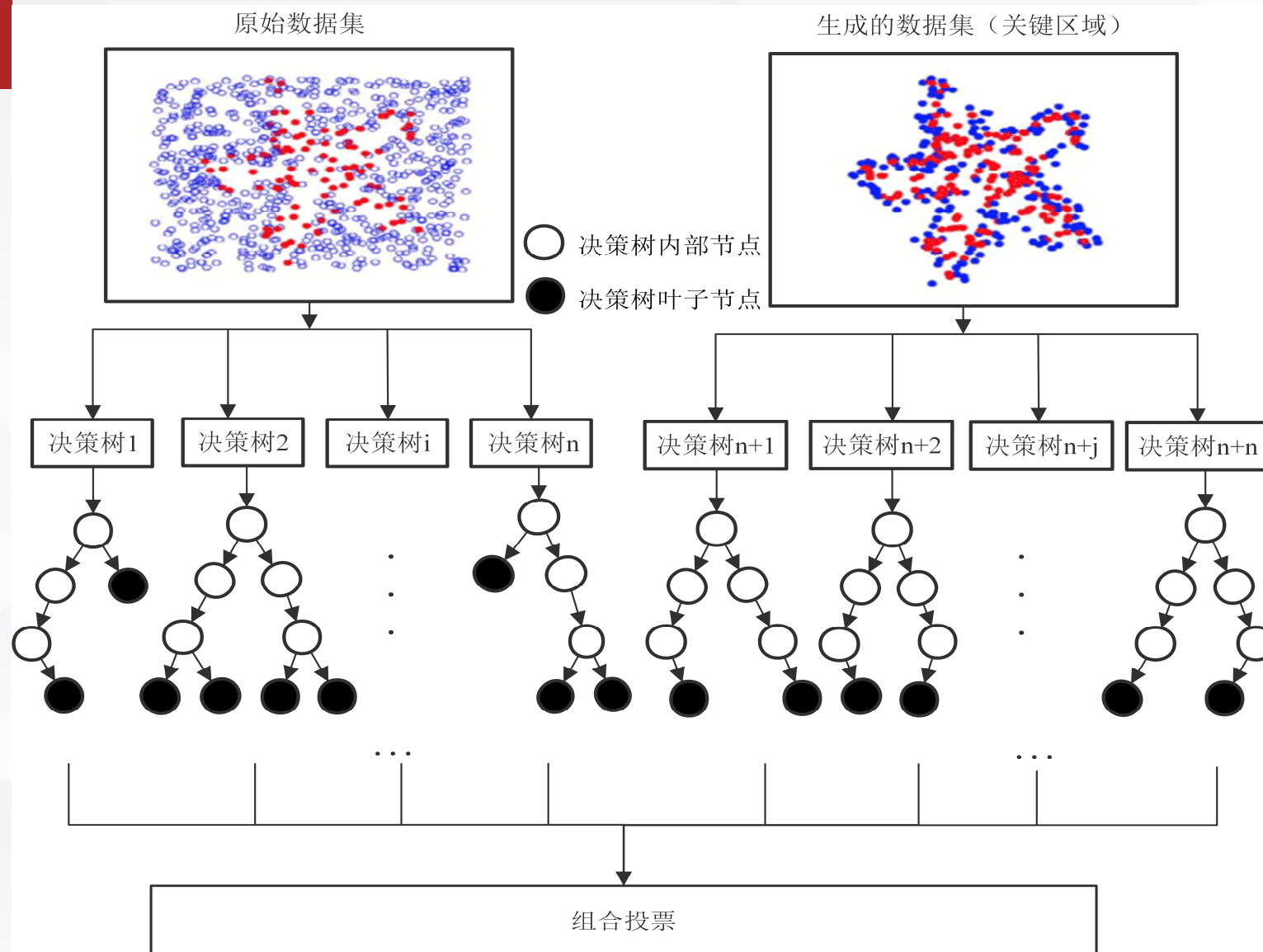
(b) 分类面发生偏移



KM-BRAF算法 设计与实现

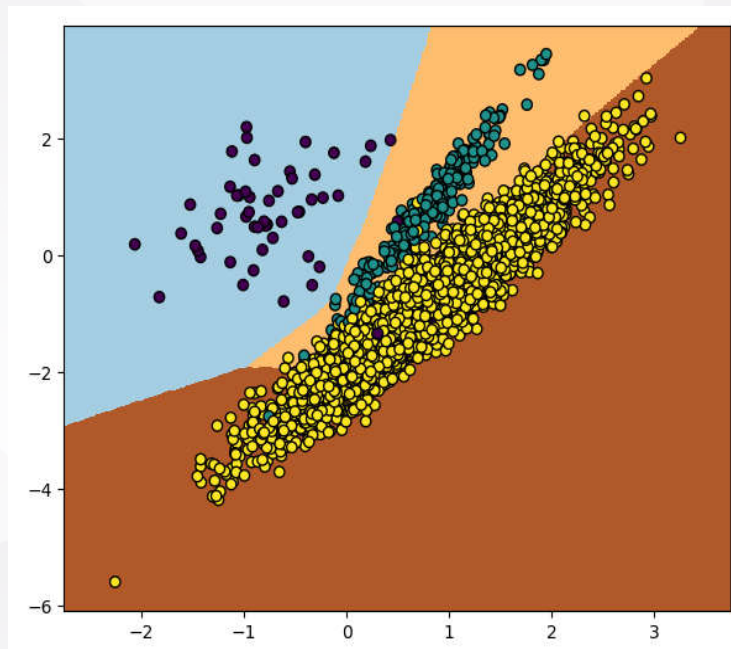


BRAf模型

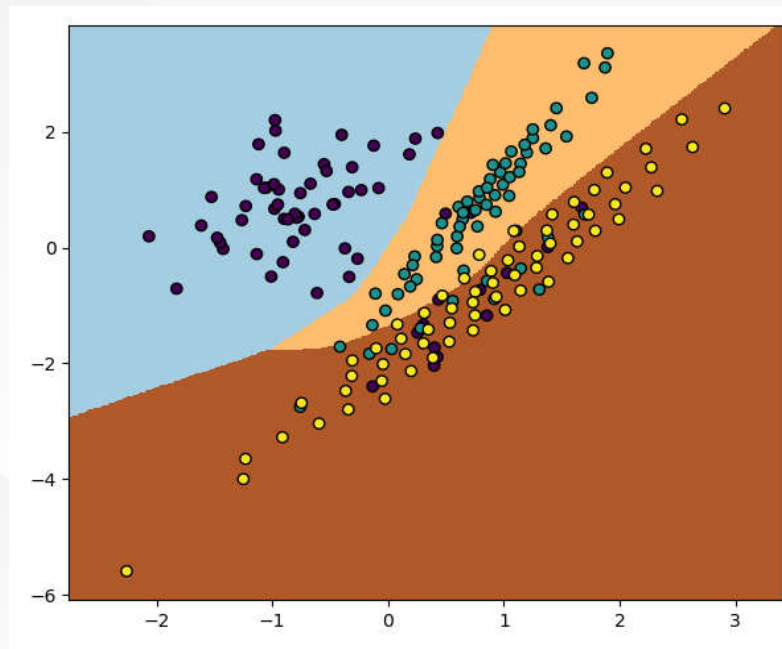


K-Means聚类欠采样方法

K-Means聚类算法对不平衡数据集中的多数类进行聚类，聚类中心或与聚类中心最临近的样本具有较高的代表性和重要性，可以代表该样本所属的整个聚类，将其与少数类样本进行合并组成平衡数据集。



(a) 原始数据分布



(b) K-Means聚类欠采样后数据分布

采样算法对比

K近邻欠采样

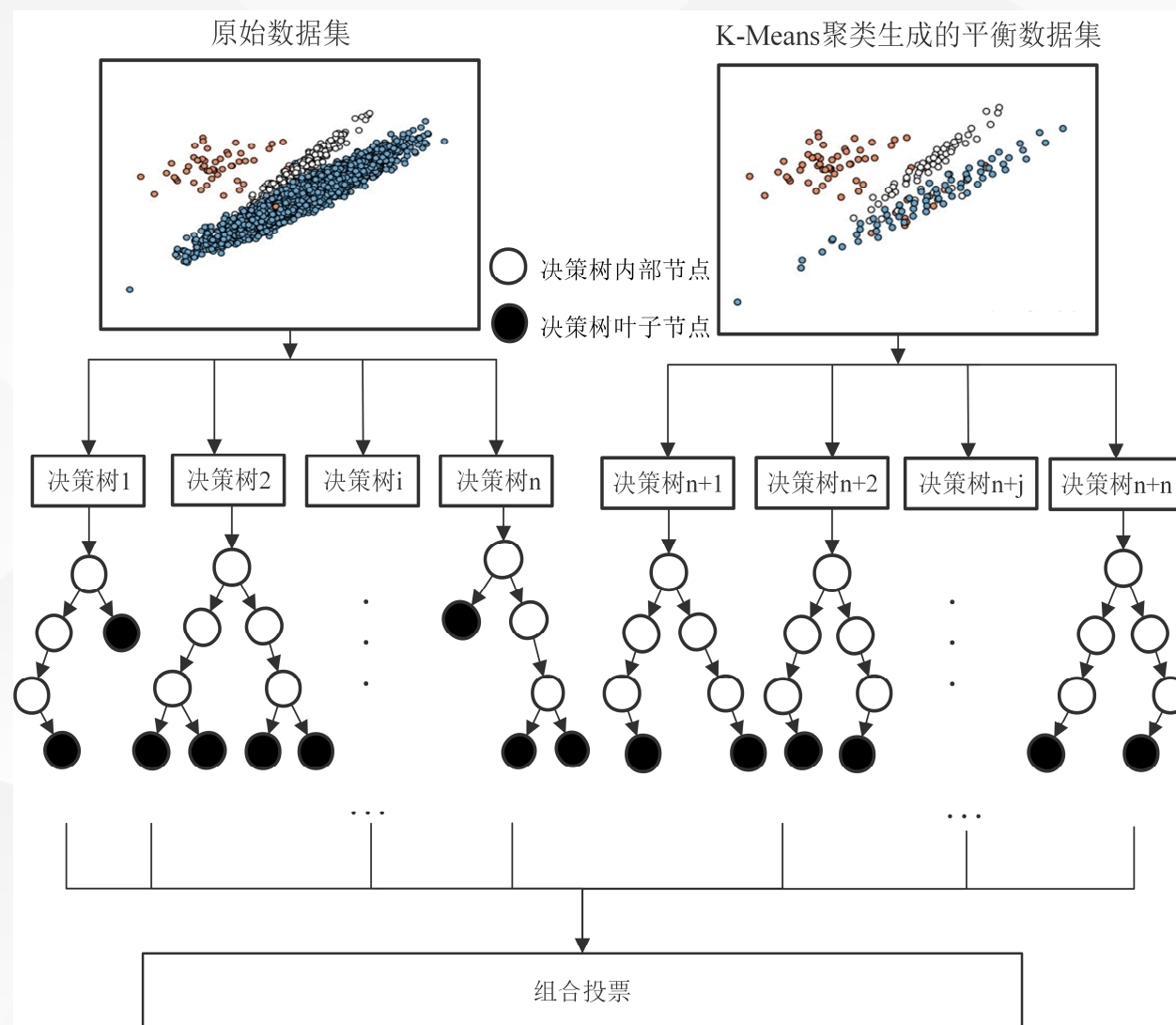
- 1、局部欠采样信息丢失；
- 2、采样后数据仍可能不平衡；
- 3、在多分类问题上难以应用；

K-Means聚类欠采样

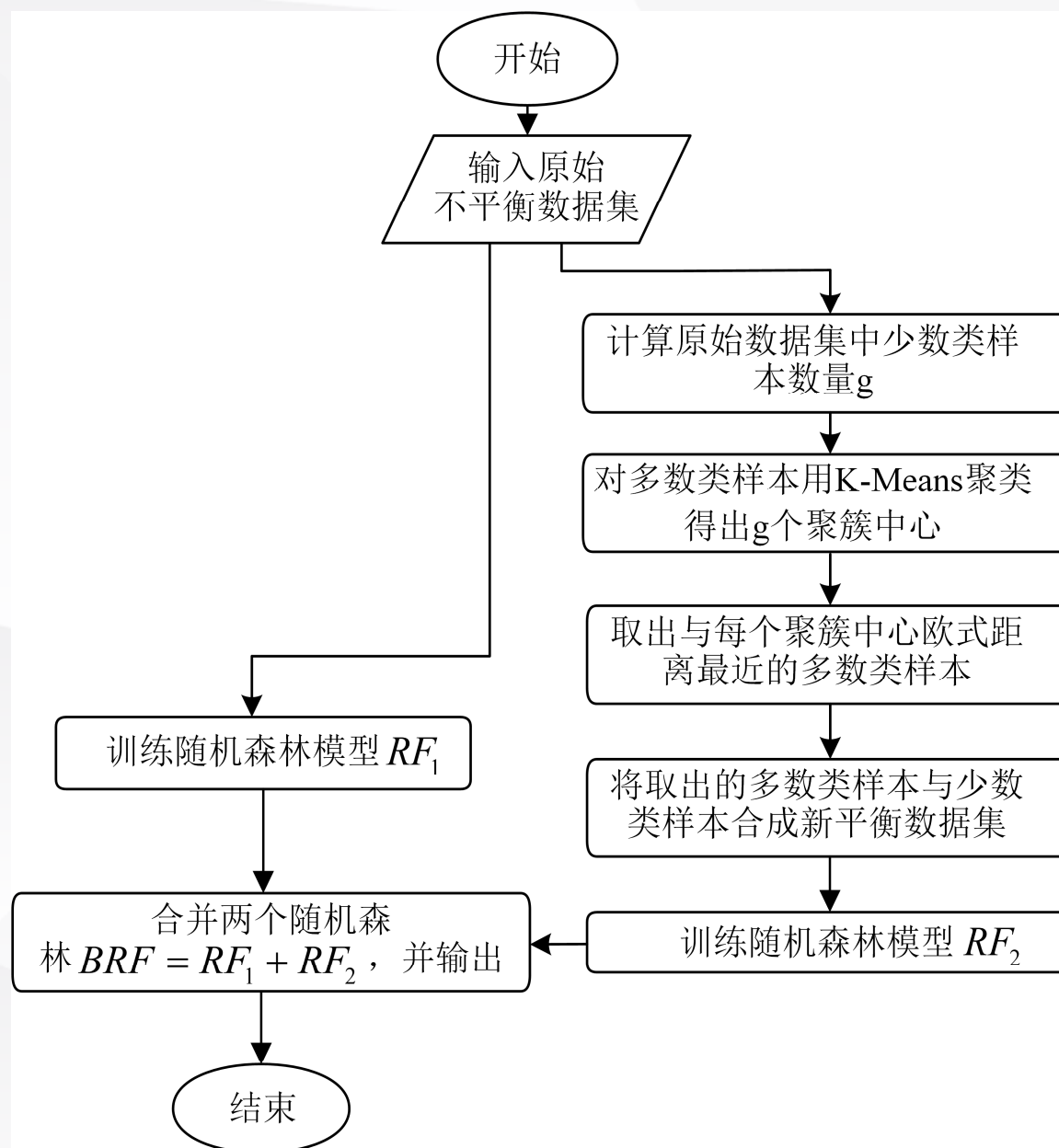
- 1、最大限度保持样本集完整；
- 2、采样后的数据绝对平衡；
- 3、可直接应用于多分类问题；

本文通过结合K-Means聚类欠采样和BRAFF算法的优势，将K-Means聚类欠采样技术融合到了随机森林的构建过程中，提出基于K-Means聚类欠采样的有偏随机森林算法（KM-BRAF），算法同时适用于不平衡二分类和多分类任务。

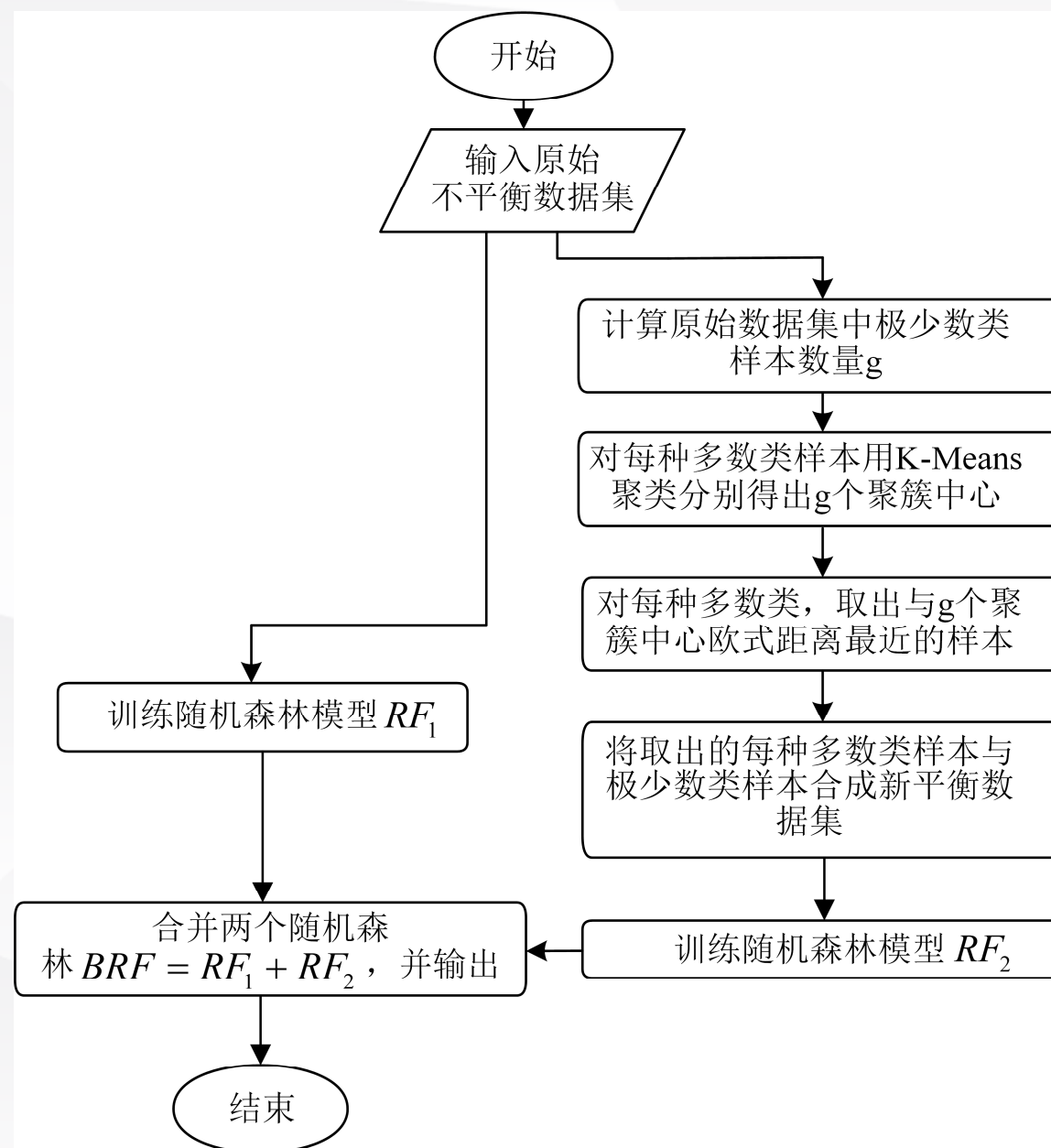
基于K-Means聚类欠采样的有偏随机森林算法 (KM-BRAF)



二分类流程图



多分类流程图



数据集介绍

为证明本文提出的**KM-BRAF算法**在不平衡数据分类任务中有更好的表现，本文使用来自KEEL库公开的13个人工构造或真实的不平衡二分类数据集和6个真实的不平衡多分类数据集进行实验论证。

多分类不平衡数据集基本信息

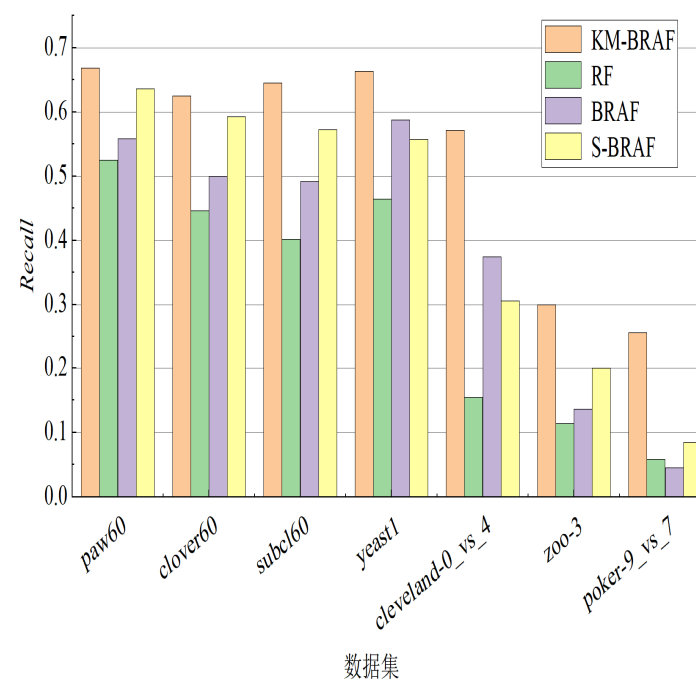
编号	名称	属性个数	样本总量	不平衡率	类别数量	类分布
14	hayes-roth	4	132	1.7	3	51/51/30
15	contraceptive	9	1473	1.89	3	629/511/333
16	new-thyroid	5	215	4.84	3	150/35/30
17	dermatology	34	366	5.55	6	112/72/61/52/49/20
18	lymphography	18	148	40.5	4	81/61/4/2
19	ecoli	7	336	71.5	8	143/77/52/35/20/5/2/2

二分类不平衡数据集基本信息

编号	数据集名称	属性个数	样本总量	不平衡率	类分布
1	paw02a-600-5-50-BI	2	600	5	500/100
2	paw02a-600-5-60-BI	2	600	5	500/100
3	paw02a-600-5-70-BI	2	600	5	500/100
4	04clover5z-600-5-50-BI	2	600	5	500/100
5	04clover5z-600-5-60-BI	2	600	5	500/100
6	04clover5z-600-5-70-BI	2	600	5	500/100
7	03subcl5-600-5-50-BI	2	600	5	500/100
8	03subcl5-600-5-60-BI	2	600	5	500/100
9	03subcl5-600-5-70-BI	2	600	5	500/100
10	yeast1	8	1484	2.46	1055/429
11	cleveland-0_vs_4	13	177	12.62	780/160
12	zoo-3	16	101	19.2	96/5
13	poker-9_vs_7	10	244	29.5	236/8

二分类实验结果分析

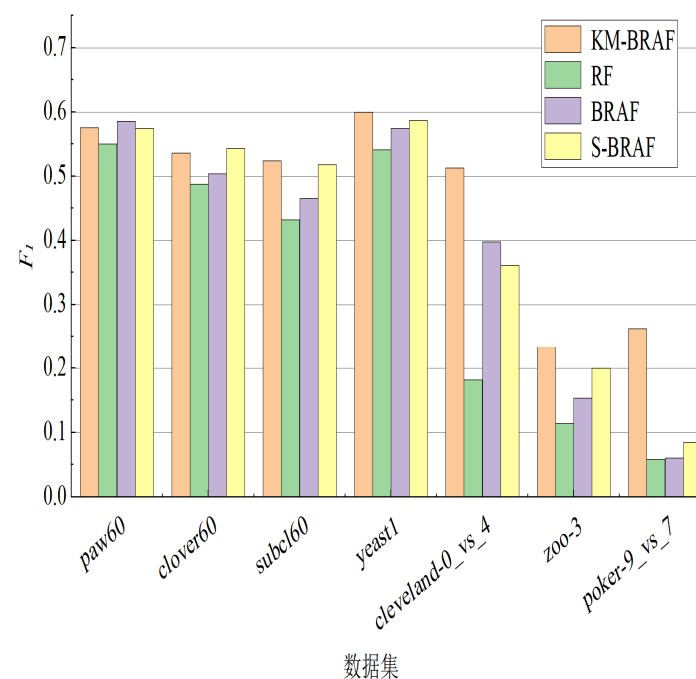
数据集名称	KM-BRAF	RF	BRAF	S-BRAF
paw02a-600-5-50-BI	0.7841*	0.6158	0.7610	0.7487
paw02a-600-5-60-BI	0.6687*	0.5245	0.5582	0.6364
paw02a-600-5-70-BI	0.7355*	0.5543	0.5967	0.6810
04clover5z-600-5-50-BI	0.7157*	0.4047	0.4915	0.5840
04clover5z-600-5-60-BI	0.6248*	0.4462	0.4998	0.5927
04clover5z-600-5-70-BI	0.5872*	0.3209	0.4180	0.5577
03subcl5-600-5-50-BI	0.6443*	0.4174	0.4859	0.5862
03subcl5-600-5-60-BI	0.6448*	0.4017	0.4913	0.5720
03subcl5-600-5-70-BI	0.6115*	0.4239	0.4802	0.4958
yeast1	0.6638*	0.4639	0.5881	0.5574
cleveland-0_vs_4	0.5717*	0.1544	0.3750	0.3055
zoo-3	0.3000*	0.1143	0.1366	0.2000
poker-9_vs_7	0.2566*	0.0587	0.0455	0.0843



不同方法在不同数据集上的 *Recall*

二分类实验结果分析

数据集名称	KM-BRAF	RF	BRAF	S-BRAF
paw02a-600-5-50-BI	0.6766*	0.6158	0.6178	0.6652
paw02a-600-5-60-BI	0.5751	0.5506	0.5861*	0.5744
paw02a-600-5-70-BI	0.6030*	0.5730	0.5823	0.5799
04clover5z-600-5-50-BI	0.6240*	0.4565	0.4849	0.5511
04clover5z-600-5-60-BI	0.5359	0.4876	0.5040	0.5427*
04clover5z-600-5-70-BI	0.5319*	0.3826	0.4196	0.5285
03subcl5-600-5-50-BI	0.5310*	0.4633	0.4902	0.5295
03subcl5-600-5-60-BI	0.5243*	0.4322	0.4651	0.5180
03subcl5-600-5-70-BI	0.5121*	0.4558	0.4862	0.4640
yeast1	0.5997*	0.5411	0.5747	0.5862
cleveland-0_vs_4	0.5128*	0.1818	0.3973	0.3611
zoo-3	0.2333*	0.1143	0.1529	0.2000
poker-9_vs_7	0.2627*	0.0587	0.0607	0.0843

不同方法在不同数据集上的 F_1

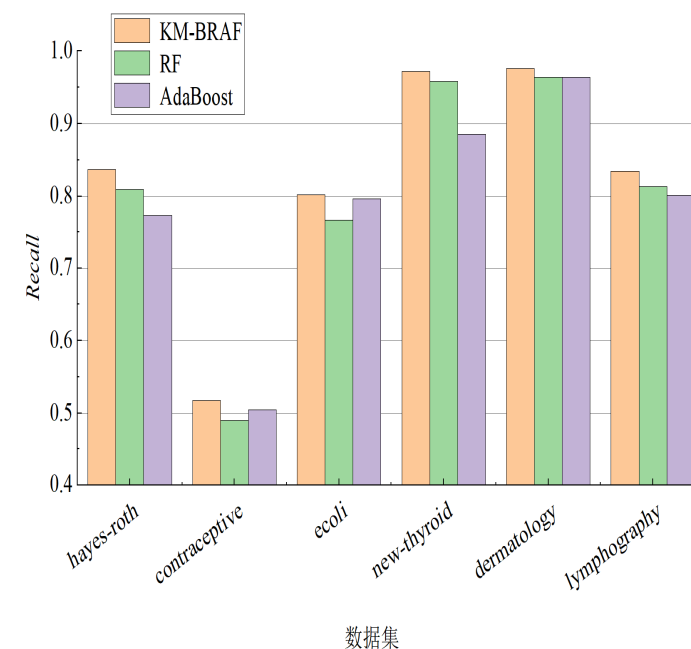
二分类实验结果分析

数据集名称	KM-BRAF	RF	BRAF	S-BRAF
paw02a-600-5-50-BI	0.8824	0.8859	0.8830	0.8680
paw02a-600-5-60-BI	0.8403	0.8597	0.8681	0.8527
paw02a-600-5-70-BI	0.8480	0.8716	0.8625	0.8460
04clover5z-600-5-50-BI	0.8529	0.8474	0.8395	0.8507
04clover5z-600-5-60-BI	0.8254	0.8480	0.8400	0.8391
04clover5z-600-5-70-BI	0.8318	0.8364	0.8359	0.8443
03subcl5-600-5-50-BI	0.8168	0.8503	0.8399	0.8323
03subcl5-600-5-60-BI	0.8144	0.8410	0.8249	0.8187
03subcl5-600-5-70-BI	0.8093	0.8445	0.8366	0.8150
yeast1	0.7468	0.7740	0.7494	0.7769
cleveland-0_vs_4	0.9438	0.9401	0.9573	0.9520
zoo-3	0.9400	0.9614	0.9323	0.9700
poker-9_vs_7	0.9783	0.9691	0.9706	0.9702

不同方法在不同数据集上的 *Accuracy*

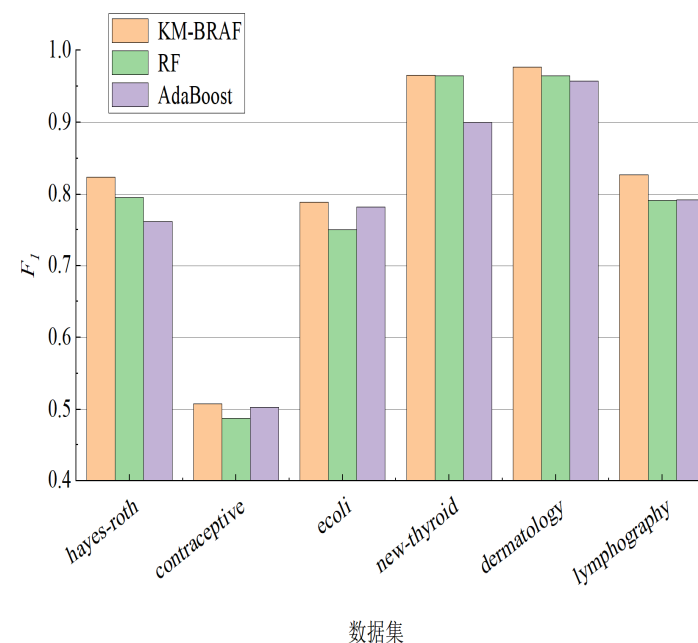
多分类实验结果分析

数据集	KM-BRAF	RF	AdaBoost
hayes-roth	0.8369*	0.8093	0.7720
contraceptive	0.5178*	0.4902	0.5045
ecoli	0.8019*	0.7656	0.7966
new-thyroid	0.9722*	0.9580	0.8856
dermatology	0.9763*	0.9635	0.9635
lymphography	0.8348*	0.8134	0.8016

三种算法在不同数据集上的 *Recall*

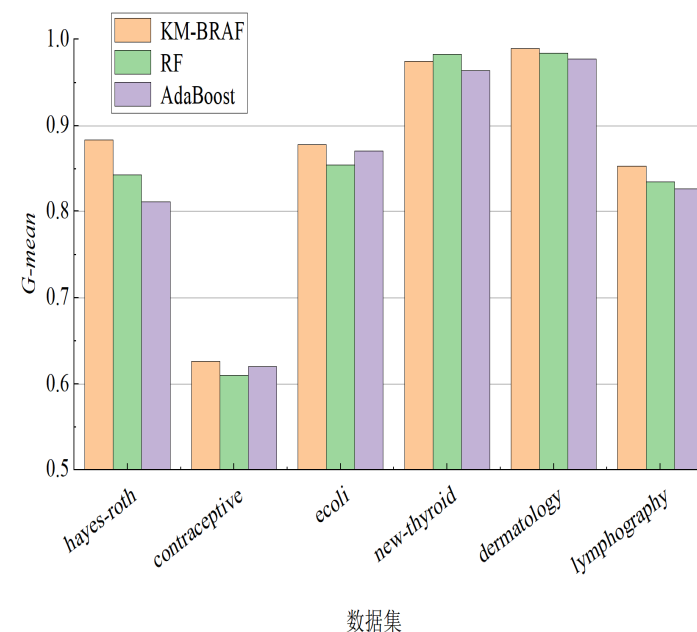
多分类实验结果分析

数据集	KM-BRAF	RF	AdaBoost
hayes-roth	0.8238*	0.7956	0.7609
contraceptive	0.5080*	0.4878	0.5034
ecoli	0.7891*	0.7494	0.7810
new-thyroid	0.9654*	0.9644	0.8996
dermatology	0.9769*	0.9647	0.9575
lymphograph y	0.8272*	0.7921	0.7925

三种算法在不同数据集上的 F_1

多分类实验结果分析

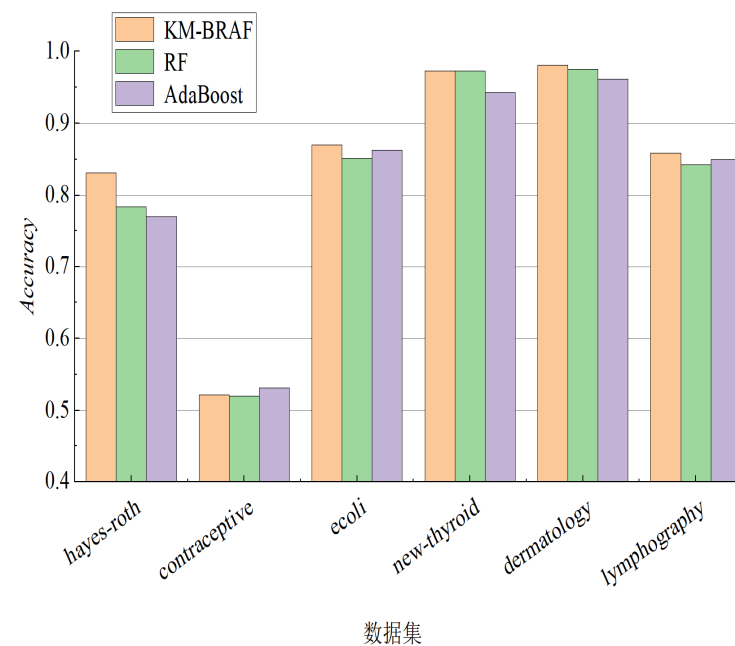
数据集	KM-BRAF	RF	AdaBoost
hayes-roth	0.8839*	0.8435	0.8107
contraceptive	0.6262*	0.6104	0.6203
ecoli	0.8786*	0.8548	0.8706
new-thyroid	0.9747	0.9825*	0.9639
dermatology	0.9891*	0.9839	0.9772
lymphograph y	0.8537*	0.8350	0.8270



三种算法在不同数据集上的G-mean

多分类实验结果分析

数据集	KM-BRAF	RF	AdaBoost
hayes-roth	0.8310*	0.7831	0.7692
contraceptive	0.5217	0.5202	0.5313*
ecoli	0.8697*	0.8512	0.8625
new-thyroid	0.9724*	0.9724	0.9429
dermatology	0.9808*	0.9749	0.9611
lymphography	0.8584*	0.8425	0.8500



三种算法在不同数据集上的Accuracy



KMBS-BRAF

算法设计与实现



KM-BRAF算法局限性

噪声影响

KM-BRAF算法对多数类进行欠采样时利用到了**K-Means聚类算法**，但是K-Means算法对**噪声**孤立点数据和选取的初始聚类中心比较敏感，聚类结果易收敛于局部最小值；

信息缺失

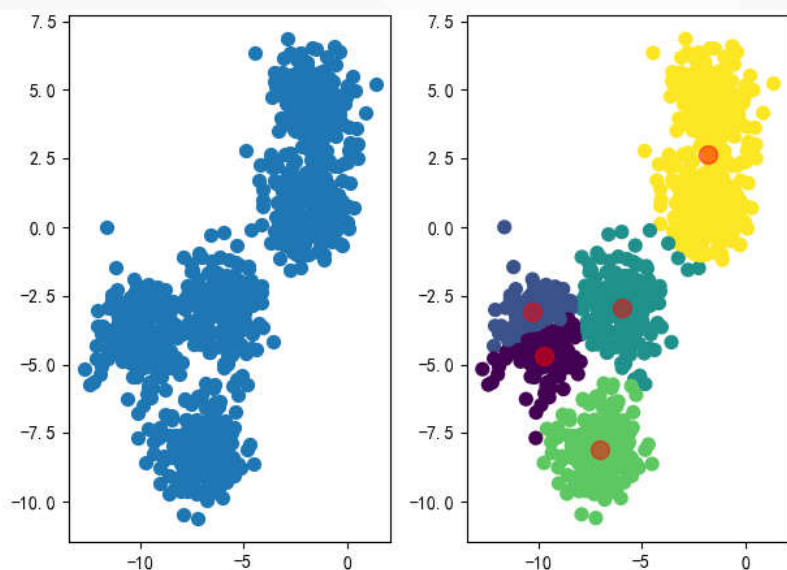
KM-BRAF算法仅利用**K-Means聚类欠采样**技术对多数类样本采样以生成平衡数据集，在少数类样本数量级较小的情况下，欠采样技术会抛弃大量的多数类样本，造成严重的**关键信息缺失**。

基于混合采样的有偏随机森林算法 (KMBS-BRAF) 算法

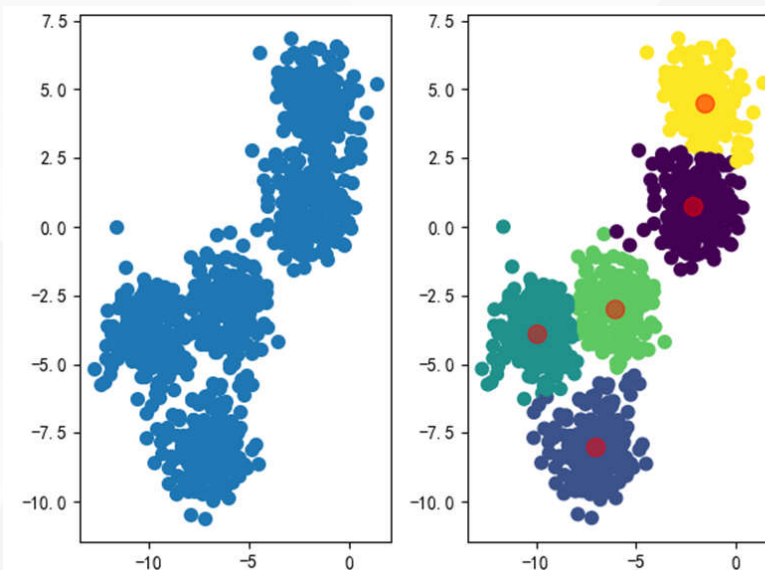
KMBS-BRAF算法中混合采样是指分别用**二分K-Means聚类欠采样**对多数类进行欠采样、用**Borderline-SMOTE**对少数类进行过采样。算法主要优势为：第一，对K-Means聚类欠采样算法进行改进，得到二分K-Means聚类欠采样算法并用于多数类样本欠采样过程，解决了K-Means聚类欠采样技术对**噪声敏感、容易收敛于局部最小值**的问题；第二，对少数类样本进行**Borderline-SMOTE**过采样，解决了小数量级数据集中仅对多数类欠采样引起的**严重信息缺失**问题。最终将二分K-Means聚类欠采样和Borderline-SMOTE过采样组成混合采样算法，替换KM-BRAF中的K-Means聚类欠采样算法用于关键区域平衡数据集的生成，从而使训练出的模型性能得到提升。

二分K-Means聚类欠采样

步骤	
1	将数据集中所有数据看成一个聚簇；
2	当聚簇数量少于 时或者符合其他聚类条件时：
3	对每一个聚簇：
4	计算误差平方和 ；
5	对该聚簇使用K-Means算法将其一分为二；
6	对二分后的所有聚簇总误差平方和 进行计算；
7	选取使总误差平方和 降低最快的一个聚簇继续二分；

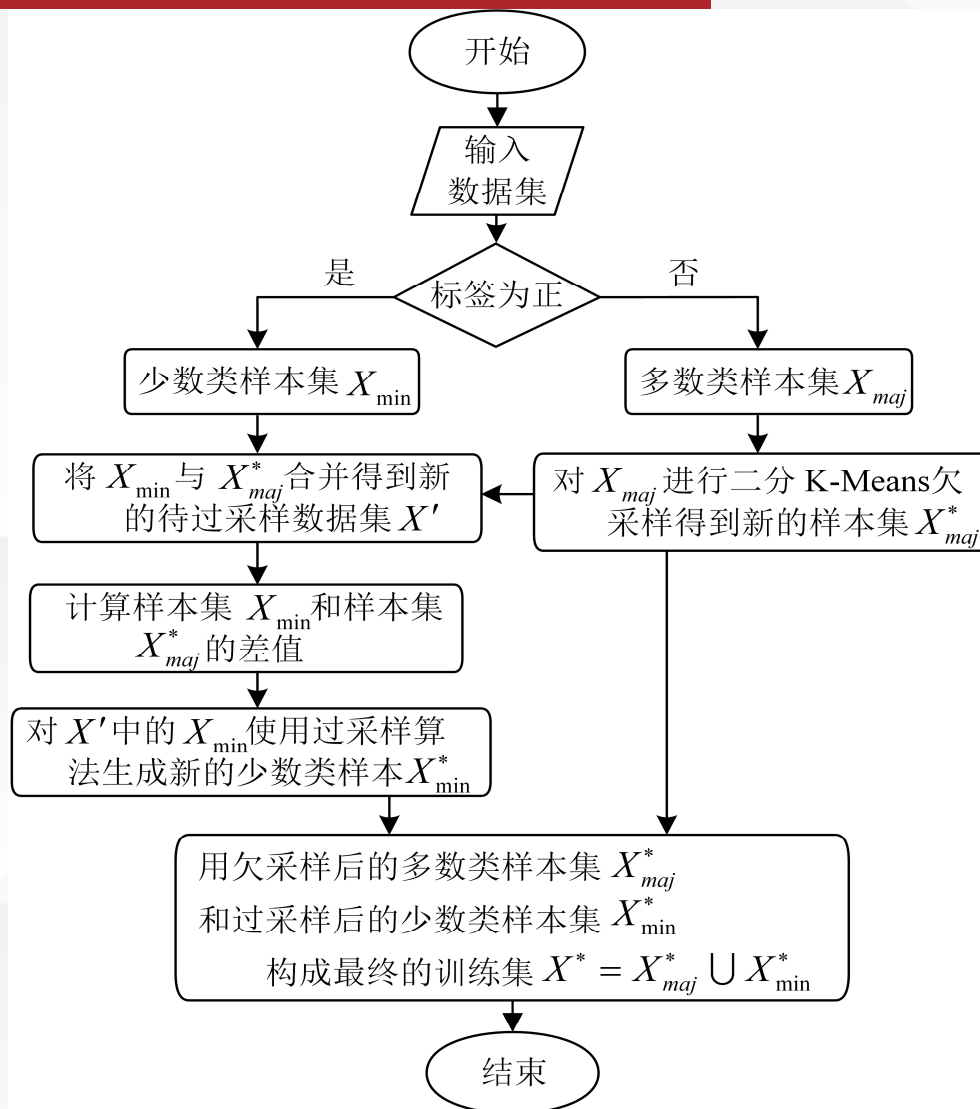


K-Means算法聚类效果

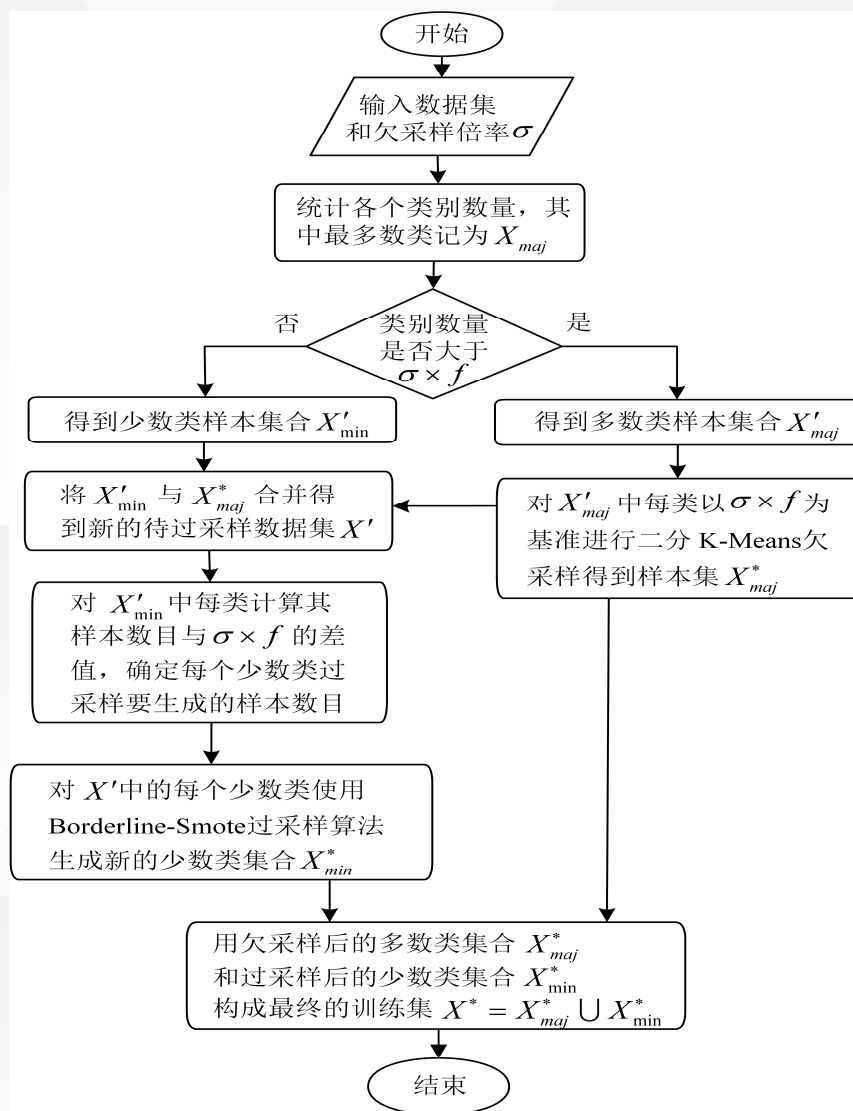


二分K-Means算法聚类效果

二分类不平衡数据集KMBS混合采样算法流程



多分类不平衡数据集KMBS混合采样算法流程



KMBS-BRAF在二分类数据集上各项指标对比

数据集	<i>Recall</i>		F_1		<i>Accuracy</i>	
	KMBS-BRAF	KM-BRAF	KMBS-BRAF	KM-BRAF	KMBS-BRAF	KM-BRAF
paw02a-600-5-50-BI	0.7752	0.7841*	0.6920*	0.6766	0.8848*	0.8824
paw02a-600-5-60-BI	0.7132*	0.6687	0.6155*	0.5751	0.8538*	0.8403
paw02a-600-5-70-BI	0.7397*	0.7355	0.6050*	0.6030	0.8433	0.8480*
04clover5z-600-5-50-BI	0.6882	0.7157*	0.6288*	0.6240	0.8700*	0.8529
04clover5z-600-5-60-BI	0.6726*	0.6248	0.5771*	0.5359	0.8388*	0.8254
04clover5z-600-5-70-BI	0.6317*	0.5872	0.5808*	0.5319	0.8525*	0.8318
03subcl5-600-5-50-BI	0.6493*	0.6443	0.5464*	0.5310	0.8243*	0.8168
03subcl5-600-5-60-BI	0.6606*	0.6448	0.5517*	0.5243	0.8212*	0.8144
03subcl5-600-5-70-BI	0.6207*	0.6115	0.5204*	0.5121	0.8120*	0.8093
yeast1	0.6855*	0.6638	0.6090*	0.5997	0.7475*	0.7468
cleveland-0_vs_4	0.6260*	0.5717	0.5751*	0.5128	0.9462*	0.9438
zoo-3	0.3000*	0.3000*	0.2935*	0.2333	0.9507*	0.9400
poker-9_vs_7	0.3386*	0.2566	0.3515*	0.2627	0.9799*	0.9783

KMBS-BRAF在多分类数据集上各项指标对比

数据集	<i>Recall</i>		F_1		<i>G - mean</i>		<i>Accuracy</i>	
	KMBS-BRAF	KM-BRAF	KMBS-BRAF	KM-BRAF	KMBS-BRAF	KM-BRAF	KMBS-BRAF	KM-BRAF
hayes-roth	0.8528*	0.8369	0.8442*	0.8238	0.8902*	0.8839	0.8358*	0.8310
contraceptive	0.5159*	0.5178	0.5099*	0.5080	0.6245	0.6262*	0.5240*	0.5217
new-thyroid	0.8561*	0.8019	0.8442*	0.7891	0.9059*	0.8786	0.8879*	0.8697
dermatology	0.9748*	0.9722	0.9714*	0.9654	0.9785*	0.9747	0.9747*	0.9724
lymphography	0.9823*	0.9763	0.9797*	0.9769	0.9871	0.9891*	0.9805	0.9808*
ecoli	0.8736*	0.8348	0.8673*	0.8272	0.8691*	0.8537	0.8733*	0.8584



发热待查病因预测应用



发热待查问题介绍

发热待查 (FUO: Fever of Unknown Origin) 是指发热持续3周以上，经过至少一周在门诊或住院的系统全面的检查仍不能确诊的一组疾病引起发热疾病种类繁多，医生难以在有限的时间内给出明确的诊断。



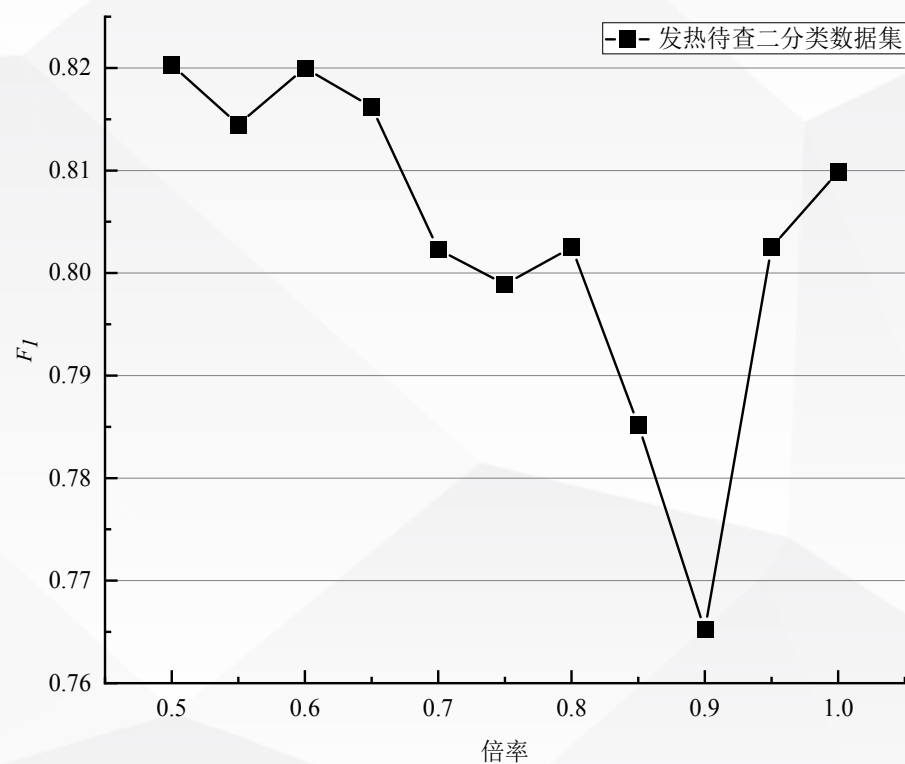
发热待查医疗数据集		类别	诊断	病因	数量（例）	合计（例）
原始数据集中包含的病因类别以及其对应数量		感染性疾病	细菌	肝脓肿	24	419
				感染性心内膜炎	12	
				布氏杆菌病	64	
				军团菌	6	
				脾脓肿	1	
				李斯特菌感染	1	
			病毒	EB病毒感染	77	
				巨细胞病毒感染	14	
				腺病毒感染	1	
				疱疹病毒感染	3	
				传染性单核细胞增多症	38	
				上呼吸道感染	97	
				登革热	1	
			真菌	组织胞浆菌	1	
			寄生虫	黑热病	13	
				阿米巴感染（肝脓肿）	1	
				弓形虫感染	1	
				疟疾	5	
				猪肉绦虫（肠绦虫病）	1	
			不典型病原体	支原体感染	11	
				立克次体感染	45	
				猫抓病	1	

发热待查医疗数据集

类别	诊断	病因	数量（例）	合计（例）
非感染性 疾病	肿瘤	淋巴瘤	33	191
	自身免疫性疾病	Anca相关性血管炎	12	
		大动脉炎	3	
		成人斯蒂尔病	20	
		结缔组织病	7	
		皮肤炎	3	
	自身炎症性疾病	结节性红斑	1	
		全身炎症反应综合征	47	
		嗜血细胞综合征	19	
		嗜酸性粒细胞增多症	2	
		多形红斑	4	
		脂膜炎	4	
	其他	坏死性淋巴结炎	34	
		药物过敏	2	

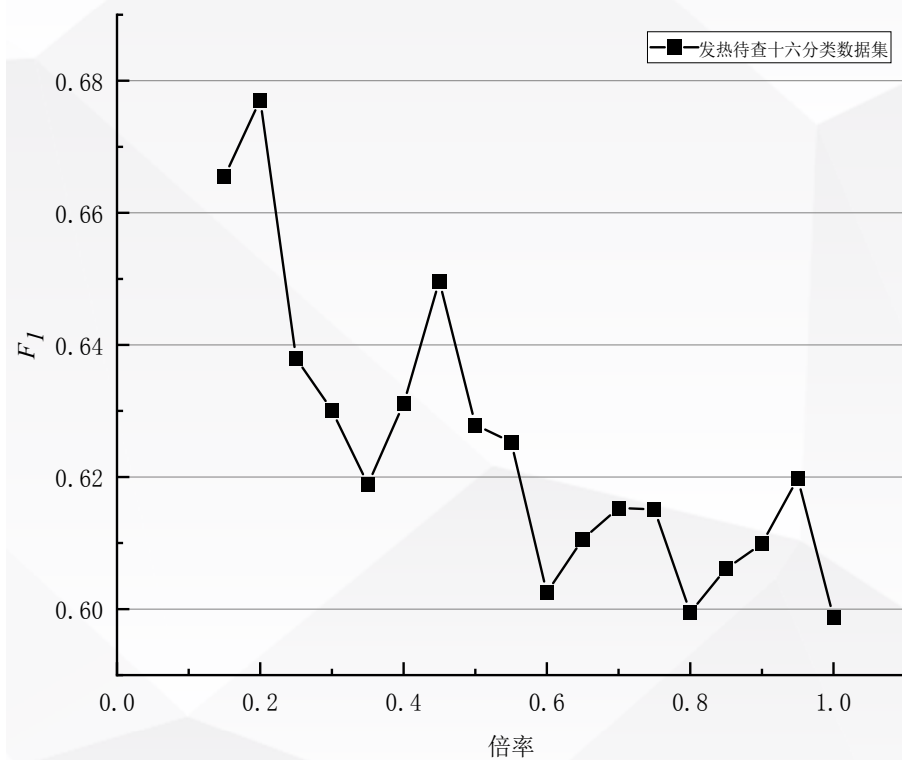
原始数据集中包含的病因类别以及其对应数量

发热待查不平衡二分类实验结果



指标	KMBS-BRAF	RF	BRAF	S-BRAF	AdaBoost
Accuracy	0.8912*	0.8686	0.8839	0.8792	0.8590
F_1	0.8455*	0.6714	0.8210	0.8265	0.7508
Recall	0.8203*	0.7477	0.7973	0.8004	0.7548

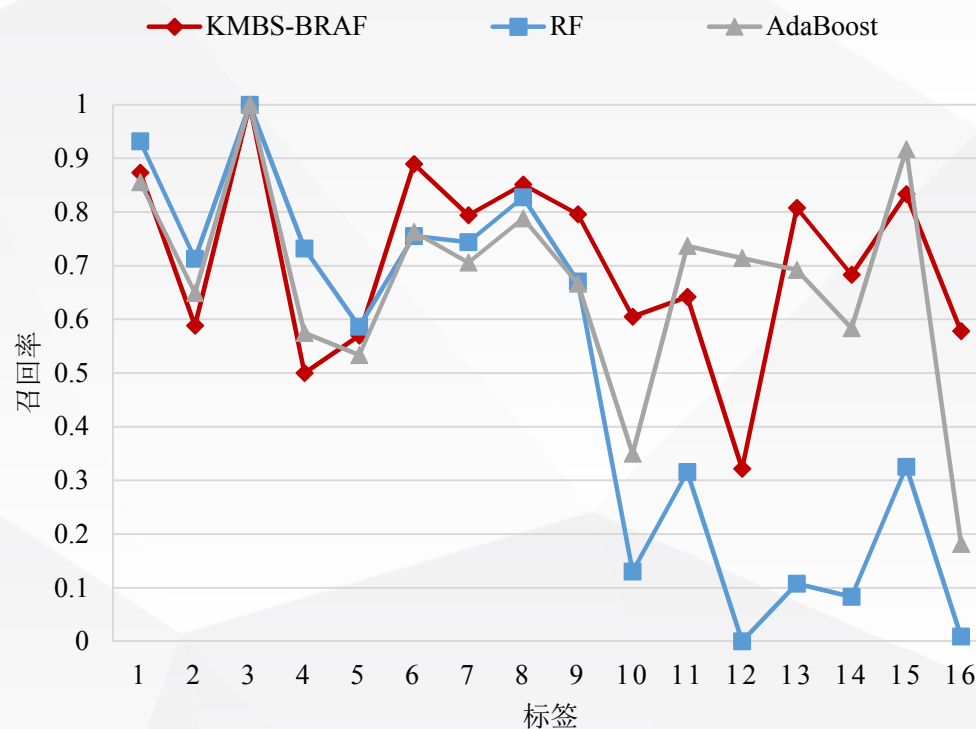
发热待查不平衡十六分类实验结果



指标	KMBS-BRAF	RF	AdaBoost
<i>Accuracy</i>	0.7533*	0.6754	0.7214
F_1	0.6771*	0.5487	0.6385
<i>Recall</i>	0.8308*	0.7191	0.8109

发热待查不平衡十六分类实验结果

标签	KMBS-BRAF	RF	AdaBoost
1	0.8732	0.9320*	0.8557
2	0.5883	0.7130*	0.6494
3	1.0000*	1.0000*	1.0000*
4	0.5000	0.7319*	0.5745
5	0.5698	0.5867*	0.5333
6	0.8895*	0.7553	0.7632
7	0.7941*	0.7441	0.7059
8	0.8515*	0.8273	0.7879
9	0.7958*	0.6708	0.6667
10	0.6050*	0.1300	0.3500
11	0.6421	0.3158	0.7368*
12	0.3214	0.0000	0.7143*
13	0.8077*	0.1077	0.6923
14	0.6833*	0.0833	0.5833
15	0.8333	0.3250	0.9167*
16	0.5780*	0.0091	0.1818
Avg	0.7083*	0.6665	0.6665



发热待查不平衡十六分类召回率对比



总结与展望



**A**

➤ KM-BRAF算法

在对K-Means聚类欠采样的原理和有偏随机森林算法 (BRAF) 的框架的基础上提出KM-BRAF算法。设计实验对所提出的KM-BRAF算法以及对比算法进行验证，即将KM-BRAF算法、RF算法、BRAFA算法和S-BRAF算法在19个公开的不平衡二分类和多分类数据集上实验得出的有关评价指标进行了对比论证。

B

➤ KMBS-BRAF算法

KMBS-BRAF算法在KM-BRAF算法的基础上，对关键区域训练数据样本的采样过程进行了优化，并在对应数据集上进行实验。

C

➤ 发热待查问题应用研究

将KMBS-BRAF算法应用于发热待查的疾病诊断问题，并通过实验证实了模型的性能。



◆ 大数据集计算速度

本文提出的KM-BRAF算法在小样本量不平衡分类任务中具有良好的效果，但在处理大数据集时会增加计算时间的成本。因此，未来的研究可以对该算法进行改进，以提高其适应大数据集的能力，并进一步提高模型的计算速度。

◆ 医疗数据特性

医疗数据质量具有高度异构、模糊、嘈杂和不完整等缺陷。因此利用医疗数据数据集训练好一个机器学习模型，还需要重点对医疗数据集中的数据稀疏性、冗余和缺失值等问题展开研究。

◆ 临时医疗数据处理

疾病总是以不确定的方式随着时间的推移而进展和变化。然而，许多现有的机器学习模型，包括医学领域已经提出的模型，都假设基于静态矢量的输入，无法以自然的方式处理时间因素。因此设计能够处理临时医疗数据的机器学习方法也是开发新解决方案的一个重要方面。