中山大學
SUN YAT-SEN UNIVERSITY

集成学习方法在财务舞弊中的应用
——以CUSBoost算法为例

莫书琪

2021/4/29

引入

- **研究背景：**财务舞弊的危害性与财务舞弊预测的困难性
- **研究目的：**针对财务舞弊预测这一情境，寻找合适的机器学习方法
- **主要挑战：**传统机器学习方法在非平衡数据学习任务下表现不佳

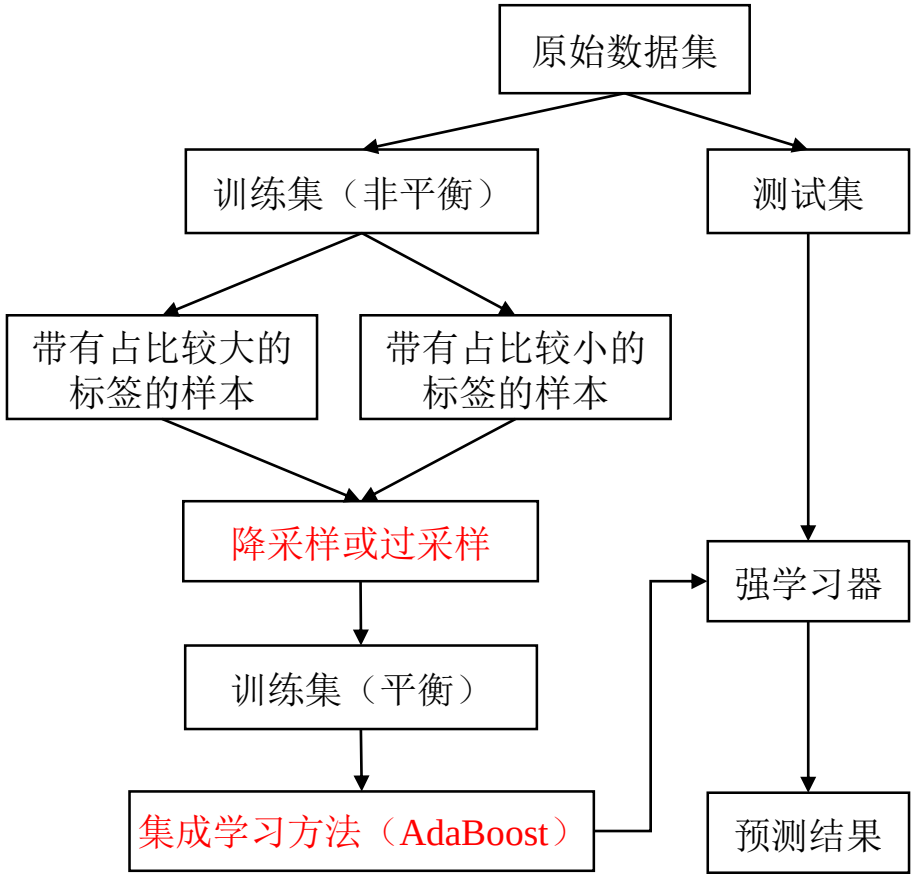
相关工作

- JAR期刊上一篇使用RUSBoost模型预测财务舞弊的工作(Bao et al.2020)

The Out-of-Sample Performance Evaluation Metrics for the Test Period 2003–08

			Performance Metrics Averaged over the Test Period 2003–2008			
			Metric 1	Metric 2		
Input Variables		Method	AUC	NDCG@k	*Sensitivity*	*Precision*
14 financial ratios	1)	logit	0.672 (0.167)	0.028 (0.479)	3.99%	2.63%
28 raw financial data items	2)	SVM-FK	0.626 (0.012)	0.020 (0.171)	2.53%	1.92%
	3)	Logit	0.690 (0.211)	0.006 (0.041)	0.73%	0.85%
	4)	RUSBoost	0.725	0.049	4.88%	4.48%

模型介绍



- **RUSBoost**: 随机降采样(Random Under-Sampling)+集成学习方法(Boosting)
- **CUSBoost**: 基于聚类的随机降采样(Cluster-based Random Under-Sampling)+集成学习方法(Boosting)
- **优点**：考虑了更全面的信息

实验设置

- **为了更公平地进行性能对比，本文在实验设置上主要参考了Bao et al.(2020)的工作**
- **数据集：**CFRM数据集，包含了美国证券交易委员会历年公布的会计审计监管文件
- **输入变量：**28个财务报表科目的财务数据
- **输出变量：**是否舞弊（0代表没有舞弊，1代表舞弊）

Bao et al.(2020)的开源数据集: https://github.com/JarFraud/FraudDetection

实验结果

- 研究问题一：CUSBoost模型的性能如何？

模型	超参数		评价指标	
	决策树最大深度	迭代次数	AUC	NDCG@k
AdaBoost	2	20	0.52	0.01
	2	30	0.52	0.01
	2	40	0.52	0.01
	12	20	0.64	0.04
	12	30	0.64	0.04
	12	40	0.64	0.04
RUSBoost	2	20	0.23	0.00
	2	30	0.72	0.03
	2	40	0.64	0.05
	12	20	0.72	0.01
	12	30	0.73	0.06
	12	40	0.74	**0.08**
CUSBoost	2	20	**0.71**	**0.07**
	2	30	0.67	**0.07**
	2	40	**0.73**	**0.06**
	12	20	**0.75**	**0.08**
	12	30	**0.76**	**0.15**
	12	40	**0.76**	0.06

实验结果

- **研究问题二：CUSBoost模型的可解释性如何？**
- **轮廓系数(silhouette coefficient)：**一个评估聚类结果好坏的评价指标，越接近1表明效果越好
- **实验方法：**在CUSBoost完成聚类后计算轮廓系数
- **实验结果：**轮廓系数均值大于0.75，表明财务舞弊数据集具有良好的聚类特性，可以借助基于聚类的随机降采样方法提高性能

总结

- 强调了财务舞弊预测中的非平衡数据学习问题
- 讨论了当前最优的财务舞弊预测模型RUSBoost的缺陷，并使用CUSBoost模型进行财务舞弊预测
- 通过实验证明了CUSBoost模型的现实意义下的有效性和理论意义下的结构可解释性
- **Q&A**

本文的数据集、代码与实验记录文档： https://github.com/shuqi-mo/FraudDetectionByCUSBoost