A Combination Method of Resampling and Random Forest for Imbalanced Data Classification

作者：LIU Zheng、QIU Han、ZHU Junhu

來源：IEEE Access

15 August 2022

報告者：林承緯

目錄

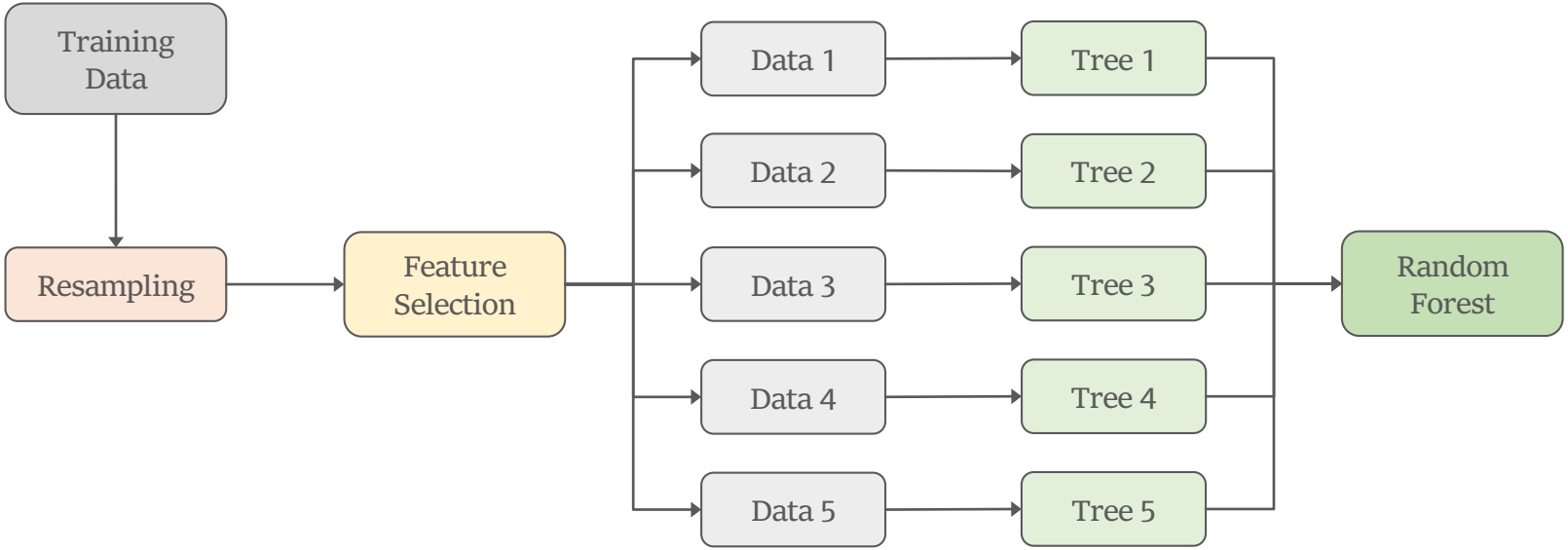
CONTENTS

摘要

提出了一種結合取樣和隨機森林的方法來進行不平衡資料分類，在不同的子樹中使用不同的子資料集進行取樣，通過減少單次取樣中使用的特徵數量，降低特徵維度對取樣效果的影響，以提高基本分類器的多樣性。

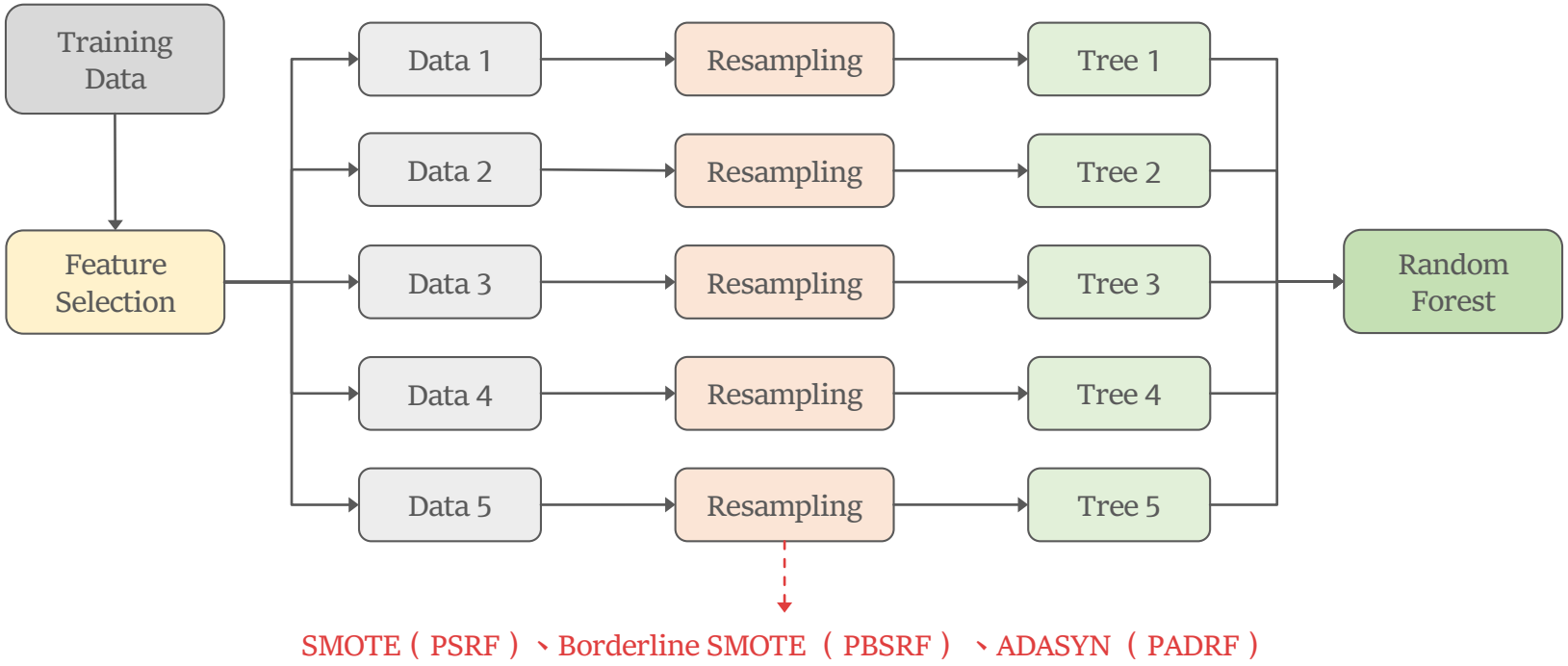
文獻探討

Resampling & Random Forest



研究方法

研究架構圖


Training Data
Feature Selection
Data 1
Data 2
Data 3
Data 4
Data 5
Resampling
Resampling
Resampling
Resampling
Resampling
Tree 1
Tree 2
Tree 3
Tree 4
Tree 5
Random Forest
SMOTE（PSRF）、Borderline SMOTE（PBSRF）、ADASYN（PADRF）

研究方法

資料集

資料集名稱	資料量	特徵數量	IR
vehicle0	846	18	3.3
newthyroid2	215	5	5.1
segment0	2308	19	6.0
glass6	214	9	6.4
yesat3	1484	8	8.1
ecoli3	336	7	8.6
page-blocks0	5472	10	8.8
vowel0	988	13	10.0
ecoli4	336	7	15.8
yeast5	1484	8	32.7

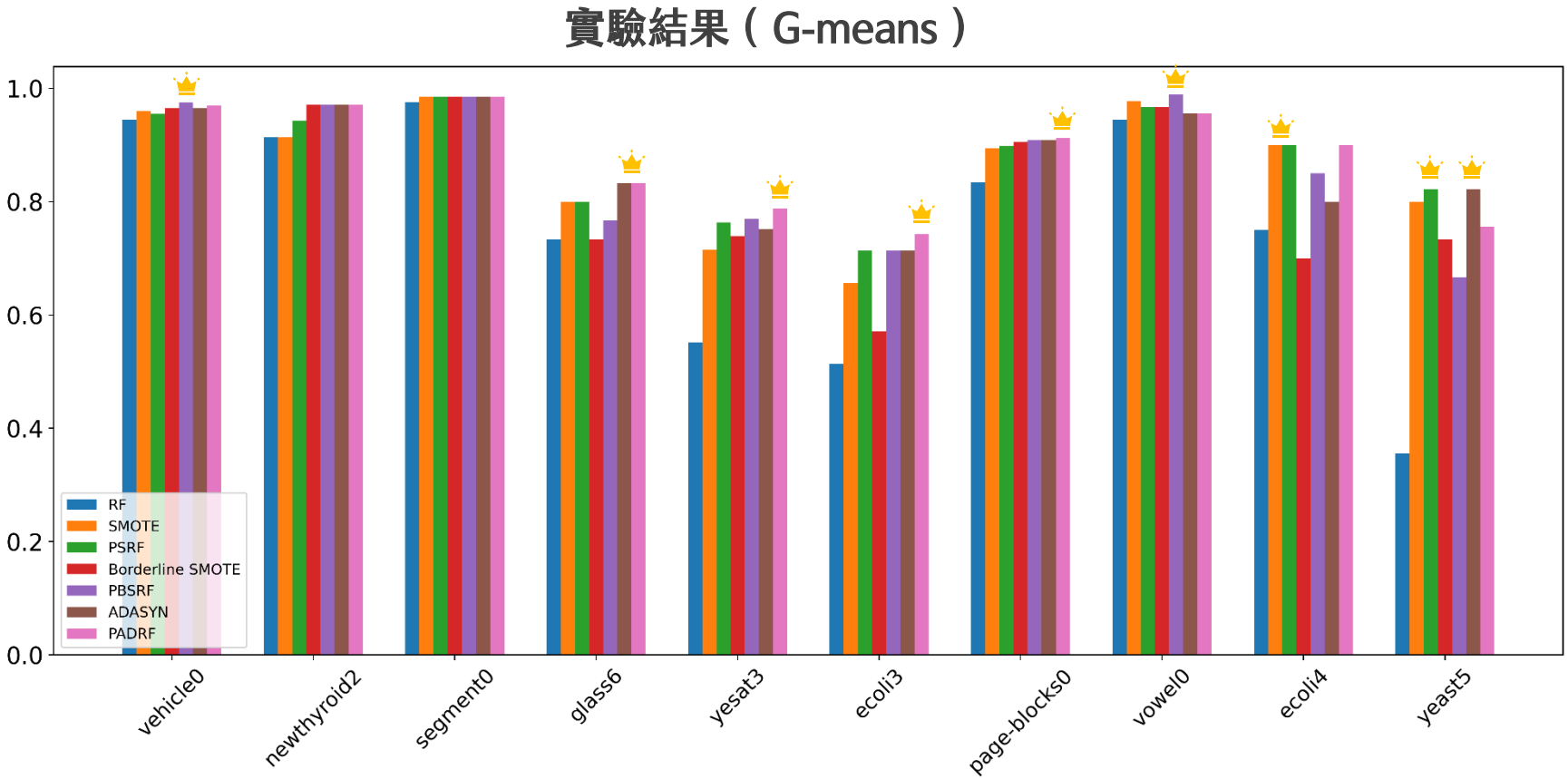
資料來源：KEEL

研究方法

實驗結果（G-means）

資料集名稱	RF	SMOTE	PSRF	Borderline SMOTE	PBSRF	ADASYN	PADRF
vehicle0	0.9450	0.9600	0.9550	0.9650	0.9750	0.9650	0.9700
newthyroid2	0.9143	0.9143	0.9429	0.9714	0.9714	0.9714	0.9714
segment0	0.9758	0.9848	0.9848	0.9848	0.9848	0.9848	0.9848
glass6	0.7333	0.8000	0.8000	0.7333	0.7667	0.8333	0.8333
yesat3	0.5515	0.7152	0.7636	0.7394	0.7697	0.7515	0.7879
ecoli3	0.5143	0.6571	0.7143	0.5714	0.7143	0.7143	0.7429
page-blocks0	0.8339	0.8946	0.8982	0.9054	0.9089	0.9089	0.9125
vowel0	0.9444	0.9778	0.9667	0.9666	0.9889	0.9556	0.9556
ecoli4	0.7500	0.9000	0.9000	0.7000	0.8500	0.8000	0.9000
yeast5	0.3556	0.8000	0.8222	0.7333	0.6667	0.8222	0.7556

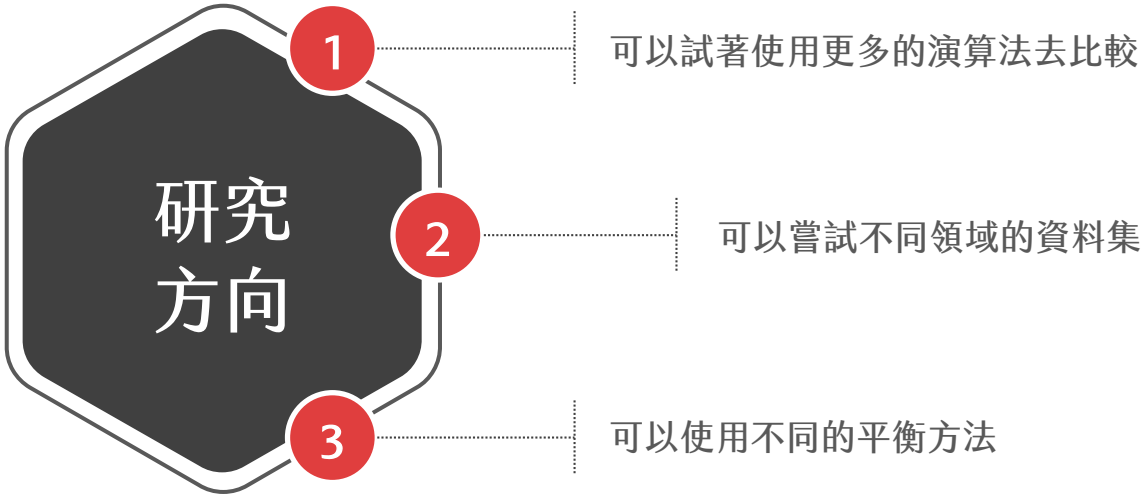
研究方法


實驗結果（G-means）
1.0
0.8
0.6
0.4
0.2
0.0
RF
SMOTE
PSRF
Borderline SMOTE
PBSRF
ADASYN
PADRF
vehicle0
newthyroid2
segment0
glass6
yesat3
ecoli3
page-blocks0
vowel0
ecoli4
yeast5

結論

- 不同子樹使用不同的特徵進行取樣，增加了基本分類器的多樣性
- 每次取樣中使用的特徵數量減少，減輕高維特徵對取樣效果的影響

後續研究方向



Thank You