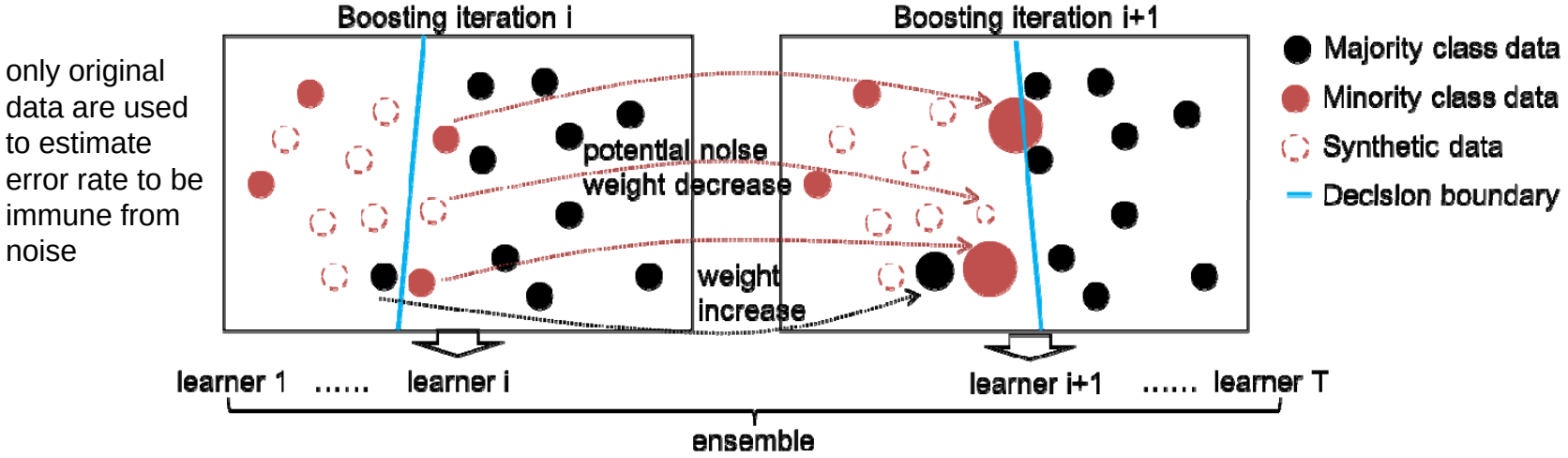
Transfer synthetic over-sampling for class-imbalance learning with limited minority class data

Xu-Ying LIU , Sheng-Tao WANG, Min-Ling ZHANG

Frontiers of Computer Science, DOI: 10.1007/s11704-018-7182-1

Problems & Ideas

- Synthetic over-sampling could introduce potential noise to class-imbalanced tasks with limited minority class data
  - SMOTE: the synthetic samples are not i.i.d. samples of minority class
  - Other sophisticated synthetic sampling methods: denoise or generate samples more consistent with ground-truth data distribution. But the assumptions about true noise or ground-truth data distribution may not hold.
- Ideas: Transfer valid knowledge from the synthetic data to the minority class



Idea of TrasoBoost Method

Main Contributions

Table 9 Summary of comparison results.

	win/tie/lose [a]	avg. rank [b]
CART	35/1/0 •	15.1
OVERCART	18/17/1•	8.2
SMOTECART	17/18/1•	7.6
ADASYN	19/16/1•	9.5
AdaBoost	18/17/1•	8.6
OverAda	31/5/0 •	10.9
SMOTEAda	26/10/0•	10.0
ADASYNAda	24/12/0•	11.9
RWOAda	23/13/0•	8.7
STomekAda	22/14/0•	9.1
SMOTEBoost	23/13/0•	10.5
RUSBoost	23/13/0•	9.0
UnderBagging	17/16/3•	7.2
SMOTEBagging	17/18/1•	8.4
AdaC1	21/15/0•	10.6
EE	8/25/3	5.0
TrasoBoost	—	**2.6**

[a)] Total win/tie/loss counts of F-measure, AUC and G-mean results. •/◦ indicates whether TrasoBoost is statistically superior/inferior to the comparing algorithm (sign test at 0.05 significance level).

[b)] Mean of average ranks of F-measure, AUC and G-mean results.