"Application of Ensemble methods for solving imbalanced data problems"

Author : Henock Makumbu Mboko

Supervisor : Prof. Kafunda Katalay

University of Kinshasa_RDC

Department of Mathematics and Computer Sciences, Faculty of Sciences

Abstract

Our work is about the study of performance and eligibility of decision trees for solving imbalanced data problems.

First, we have investigated classical decision trees on imbalanced data, we noticed this yield a poor performance on minority class, this stands for decision trees is sensitive to imbalance class problem. To improve this, we have been able to combine multiple decision trees. This combination lead us to think about random forest (RF), including Balanced random forest (BRF) and weighted random forest (WRF). Our results show that Balanced random forest outperform Random forest, decision tree and weighted random forest.

Keywords: Machine Learning, imbalanced classes, Decision trees, Ensemble methods.

Optimall
Optimization and Machine Learning Laboratory

Data sources and subject context

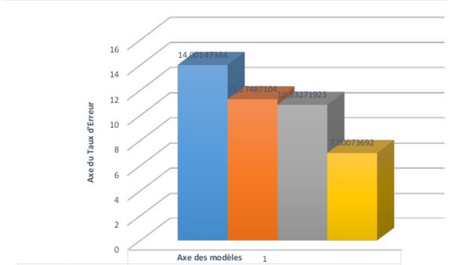
	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	30	3	3	1	0	1787	0	0	2	19	10	79	1	-1	0	0	-1
1	33	10	3	2	0	4789	1	1	2	11	5	220	1	339	4	2	-1
2	35	4	1	3	0	1350	1	0	2	16	4	185	1	330	1	2	-1
3	30	4	3	3	0	1476	1	1	0	3	6	199	4	-1	0	0	-1
4	59	6	3	2	0	0	1	0	0	5	5	226	1	-1	0	0	-1
5	35	4	1	3	0	747	0	0	2	23	2	141	2	176	3	2	-1
6	36	7	3	3	0	307	1	0	2	14	5	341	1	330	2	1	-1
7	39	9	3	2	0	147	1	0	2	6	5	151	2	-1	0	0	-1
8	41	12	3	3	0	221	1	0	0	14	5	57	2	-1	0	0	-1
9	43	10	3	1	0	-88	1	1	2	17	4	313	1	147	2	2	-1

In this experiment, we use data from a bank which contains the transactional information of customers contacted in advance by the bank in order to invite them to make the deposit operations and are observable over a given period by the bank.

These data are so imbalanced meaning that customers who responded positively and negatively constitute the minority and majority classes respectively.

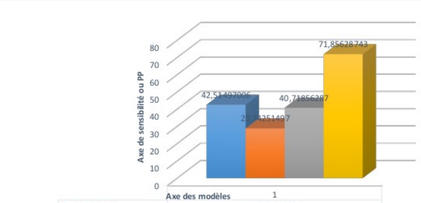
those who responded positively constitute the minority class and those who did not respond negatively form the majority class.

Estimation of the error rate in relation to the models studies



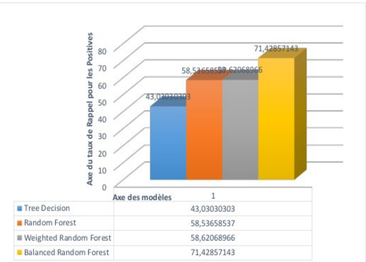
Tree Decision	14,00147384
Random Forest	11,27487104
Weighted Random Forest	10,83271923
Balanced Random Forest	7,00073692

Estimation of Sensitivity with respect to the models studied



Tree Decision	42,51497006
Random Forest	28,74251497
Weighted Random Forest	40,71856287
Balanced Random Forest	71,85628743

Estimation of Recall for the Positives in relation to models studied



Tree Decision	43,03030303
Random Forest	58,53658537
Weighted Random Forest	58,62068966
Balanced Random Forest	71,42857143

Tabular Presentation of the results

Data	Distribution (D/N/ & P.D/		Models	Error Rate	Sensitivity	Specificity	RP	RN
Data set	4523	4000 ND	TD	14,0014738	42,5149701	97,1428571	43,030303	91,9463087
		521 PD	RF	11,274871	28,742515	97,1428571	58,5365854	90,6666667
Training set	3164	2810 D.N	WRF	10,8327192	40,7185629	95,9663866	58,6206897	92,0225624
		354 D.P						
Test Set	1357	1190 D.N	BRF	7,00073692	71,8562874	95,6912029	71,4285714	95,7771788
		167 D.P						

Interpretation of results

Our dataset highlighted that the minority class is the positive class and this constitute our interest class on which we aim to improve the performance. Our measures of performance are error rate, specificity, sensitivity and recall for positive/negative.

The above table and figures highlighted that BFR outperform RF, WRF and decision tree due to its lower error rate, higher values for both, sensitivity and specificity.

Although the BFR improves the minority class performance, but we are still have a problem, because BFR will hurt the majority class performance for the benefit of minority class.

WRF seems to outperform slightly BFR with respect to the performance of negative class whereas BRF outperform all other methods with respect to positive class performance.

As future work, we aim to improve BFR in order to trade-off the majority and minority classes performance.