

H5 – Testing two resampling methods or classifiers for solving class imbalance problems

Introduction

Class imbalance is a serious and real problem in Machine Learning. For solving these kind of problems, two solutions are proposed, resampling methods and classifiers. In this homework, two different Resampling methods are used with a CostSensitiveClassifier.

Development & Results

Resample with CostSensitiveClassifier and LMT

The first step was using the Resample Method in Weka with the original dataset of Training.csv

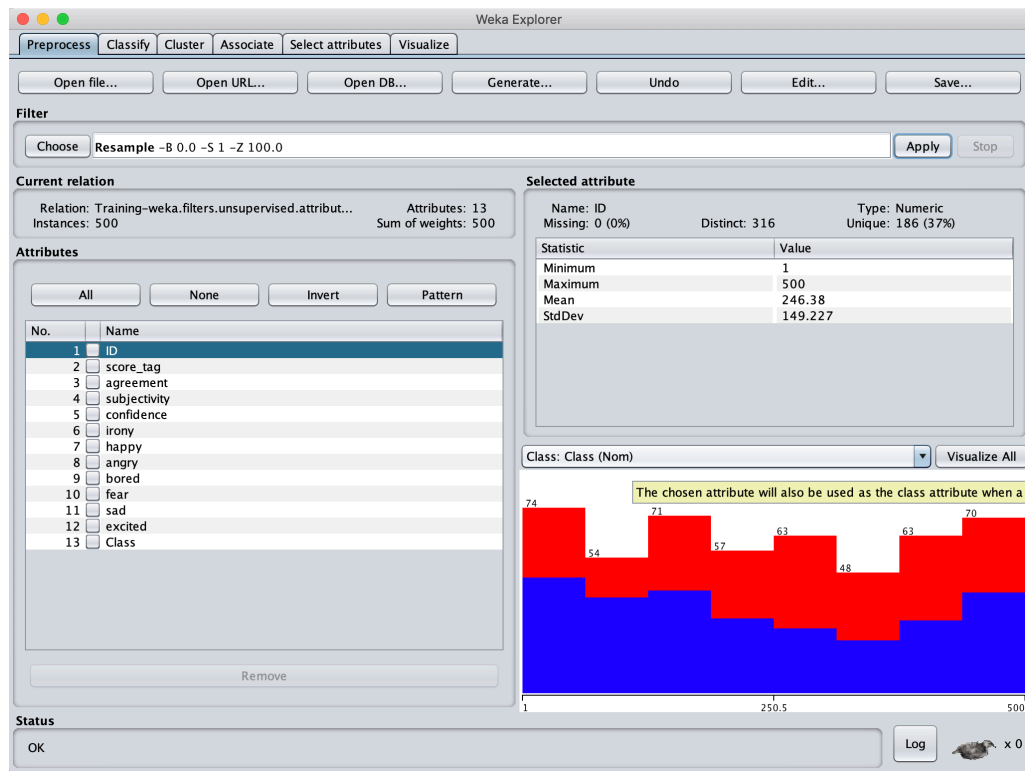


Figure 1. Resampling Training.csv in WEKA

Then, a CostSensitiveClassifier was applied with an LMT.

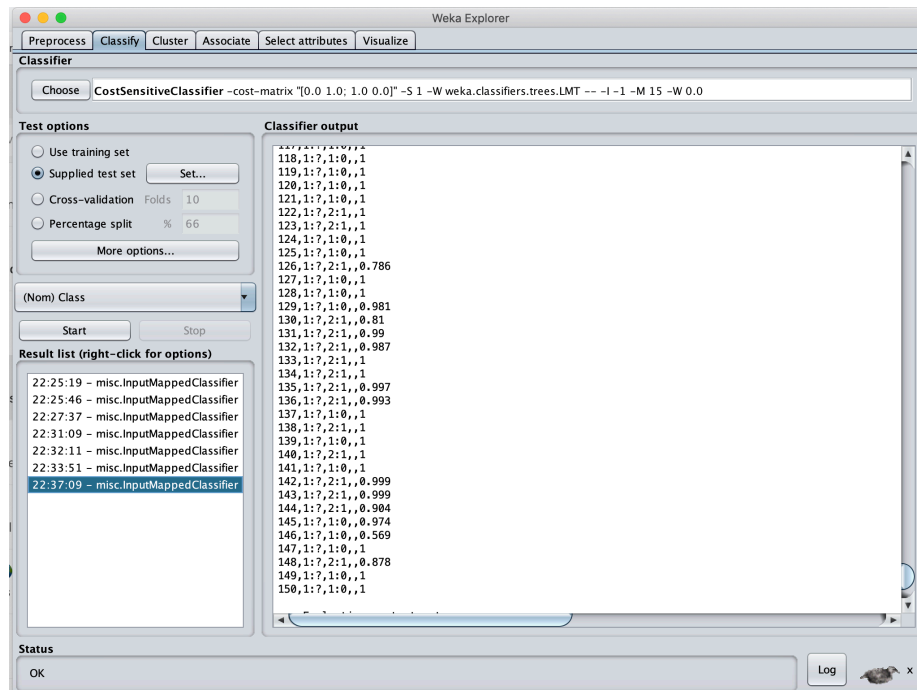


Figure 2. Results of the LMT in WEKA

After uploading the WEKA's Output to Kaggle, it can be appreciated that sadly it did not scored better than by highest score.

Public Leaderboard Private Leaderboard							
This leaderboard is calculated with all of the test data.							
				Raw Data	Refresh		
#	Team Name	Notebook	Team Members	Score	Entries	Last	
1	Saúl Ghenno Hernández			0.72500	72	4h	
2	Hector Duran Herrera			0.67777	36	4d	
3	Begoña Montes Gómez			0.67500	23	3h	
4	Diego C.			0.67222	17	2h	
5	JJoseCortesSarmiento			0.66944	48	3h	
6	Nicolas Albo			0.66666	35	now	
Your Best Entry ↑							
Your submission scored 0.51944, which is not an improvement of your best score. Keep trying!							
7	Alma del Carmen Ayaquica Ont...			0.66666	36	1h	

Figure 3. Score in Kaggle

SpreadSubsample with CostSensitiveClassifier and LMT

In this case, the first step was using the SpreadSubsample Method in Weka with the original dataset of Training.csv

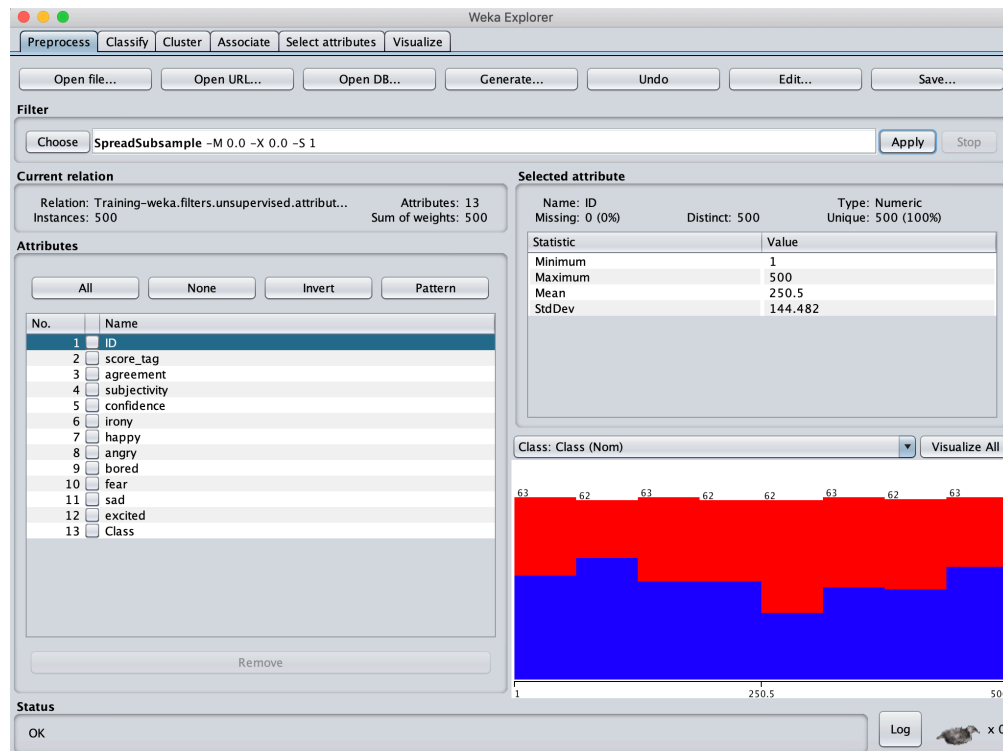


Figure 4. SpreadSubsample method applied in Training.csv in WEKA

Then, a CostSensitiveClassifier was applied with an LMT.

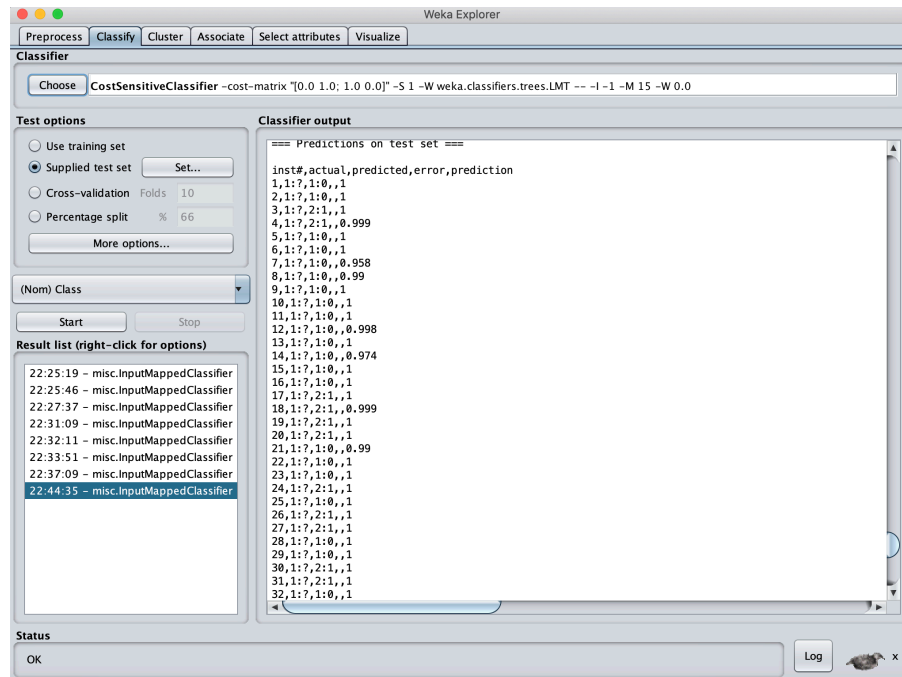


Figure 5. Results of the LMT in WEKA

After uploading the WEKA's Output to Kaggle, it can be appreciated that sadly it did not scored better than by highest score.

Overview	Data	Notebooks	Discussion	Leaderboard	Rules	Team	My Submissions	Submit Predictions
Public Leaderboard Private Leaderboard								
This leaderboard is calculated with all of the test data.								
Raw Data Refresh								
#	Team Name	Notebook	Team Members	Score	Entries	Last		
1	Saúl Ghenno Hernández			0.72500	72	4h		
2	Hector Duran Herrera			0.67777	36	4d		
3	Begoña Montes Gómez			0.67500	23	3h		
4	Diego C.			0.67222	17	2h		
5	JJoseCortesSarmiento			0.66944	48	3h		
6	Nicolas Albo			0.66666	36	now		
Your Best Entry ↑								
Your submission scored 0.55833, which is not an improvement of your best score. Keep trying!								

Figure 6. Score in Kaggle

Resample with CostSensitiveClassifier and Hoeffding Tree

In this last example, the first step was using the Resample Method in Weka with the original dataset of Training.csv

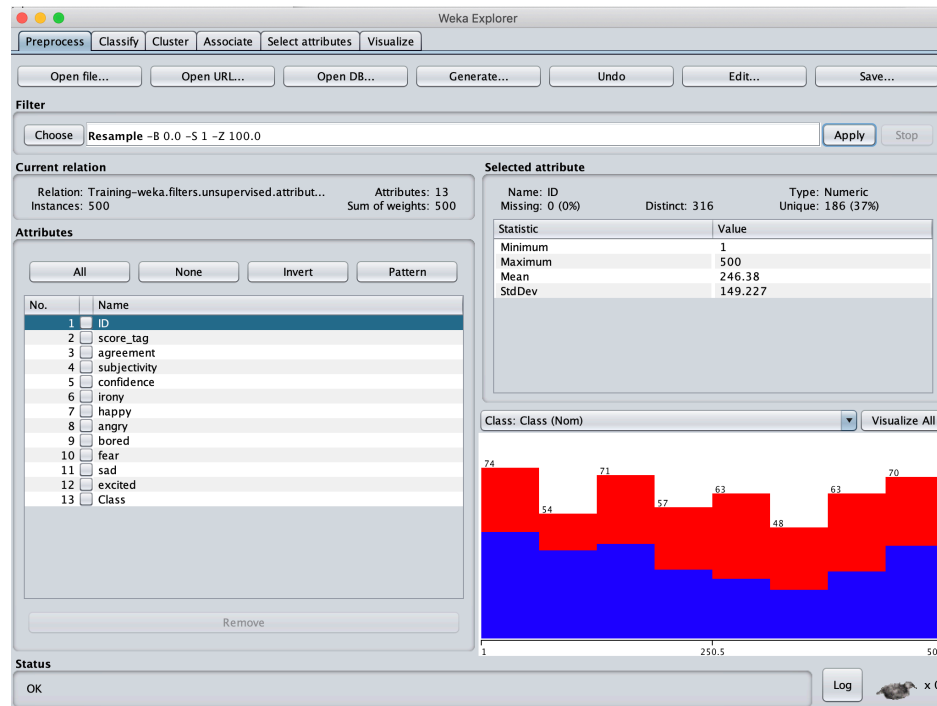


Figure 7. Resampling Training.csv in WEKA

Then, a CostSensitiveClassifier was applied with a Hoeffding Tree.

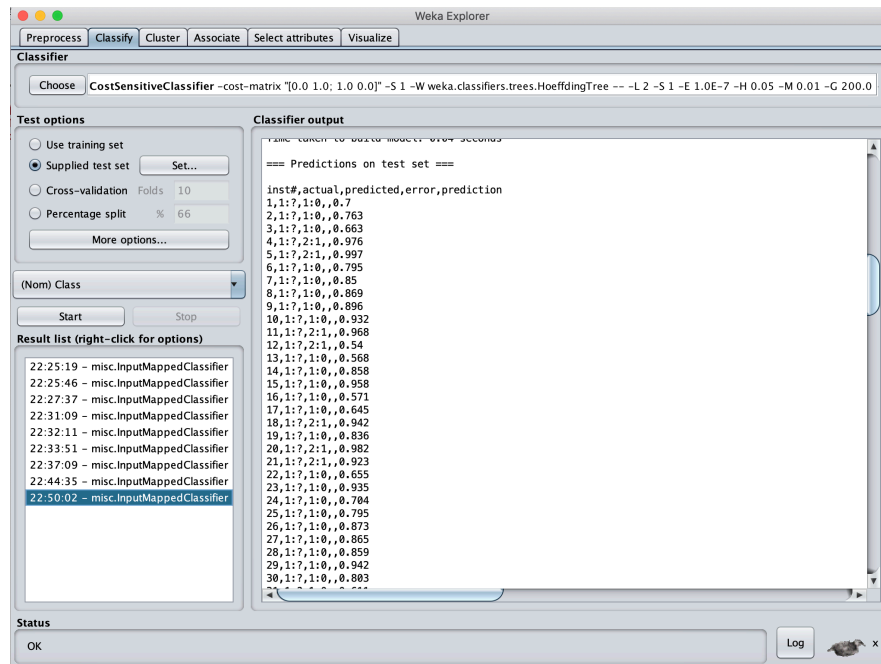


Figure 8. Results of the Hoeffding Tree in WEKA

Public Leaderboard						
Private Leaderboard						
This leaderboard is calculated with all of the test data.						
				Raw Data	Refresh	
#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Saúl Ghenno Hernández			0.72500	72	4h
2	Hector Duran Herrera			0.67777	36	4d
3	Begoña Montes Gómez			0.67500	23	3h
4	Diego C.			0.67222	17	2h
5	JJoseCortesSarmiento			0.66944	48	3h
6	Nicolas Albo			0.66666	37	now
Your Best Entry						
Your submission scored 0.57499, which is not an improvement of your best score. Keep trying!						

Figure 9. Score in Kaggle

Conclusion

Resampling methods or classifiers are a proper solution for class imbalance problems. But, in this case, these methods did not help me to improve my high score in Kaggle. Maybe I did not use an appropriate resampling method that could have helped me to rank higher in Kaggle.