University of Jean Monnet

Advanced Machine Learning

Assignment - 1

Imbalanced Learning

Author:
Allwyn Joseph

Supervisor:
Mr. Jordan Frery

November 30, 2018

1 Abstract

This report aims at summarizing the leanings from experimentation of machine learning algorithms over an imbalanced dataset. The report first skims over the mundane machine learning tasks such as reading, cleaning and exploring. This is followed by a brief run through of the basic techniques to counter the imbalanced nature of the dataset. Lastly, an overview of the techniques along with the machine learning algorithm used in combination with will be expounded upon, followed by results achieved on implementing the same.

2 The Dataset

The dataset received has a total of 100,000 rows and 54 columns signaling fraudulent transactions with label 1 and non-fraudulent transactions with a label 0. Only about 0.2 % of the data are classed as fraudulent, hence considered as imbalanced. The first 21 columns of the dataset are categorical features while the remaining are numerical features. Before moving forward with re-sampling techniques and applying machine learning algorithms the dataset is first cleaned. To do so, the dataset was scrutinized for missing values, high and zero variance attributes - instances of which were deleted. Highly correlated attributes - correlation greater than 0.98 - were deleted keeping only the relevant ones. An additional step of normalizing the numerical features followed by one-hot encoding the categorical features was carried out to be able to run algorithms like SVM and logistic regression on the dataset.

3 The Approach and Algorithms

To begin with, weighted versions of Logistic Regression, Random Forest, AdaBoost and XgBoost were used to get a benchmark on each of their performances. Using the weighted version assigned the equal weights to the sum of both the majority and minority classes. Unfortunately they performed badly on the test set, as assigning high weights to the minority class could have translated to the classifier over-fitting on the noise, resulting in low predictive power. I wasn't expecting the results to change much with simple use of

re-sampling techniques and algorithms so I opted out directly for ensemble methods in combination with re-sampling techniques or Hybrid methods.

3.1 Easy Ensemble [1]

This is an unsupervised bagging method in combination with random under-sampling. Here, the train dataset is split into subsets (25 in my case, as the classifiers' prediction didn't improve further than this point), each subset carrying all of the minority class examples and randomly picked majority class examples. This in-turn leads to balanced subsets with equal number of positive (1) and negative (0) examples. A classifier (I tested Logistic Regression, Random Forest, AdaBoost and XgBoost) is then trained on each of these subsets and tested on a test set. The results of the classifiers trained on each subset is then averaged to produce the final output.

3.2 Balance Cascade [1]

This is a supervised boosting strategy in combination with radom under-sampling techniques. Here, the train dataset is split into subsets, each subset carrying all of the minority class examples and randomly picked majority class examples. The idea is at each iteration on each subset a classifier is applied (I tested Logistic Regression, Random Forest, AdaBoost and XgBoost) the predictions of which will be confronted by the real labels using another classifer (AdaBoost in my case). The well classified true negatives will be removed and iterations continue as the weights learnt are able to better classify harder examples (minority class). This is repeated over the created subsets.

4 Results and Conclusion

The above hybrid ensemble imbalance learning techniques were chosen and tuned with the assumption that 'False negatives' (label '1' positive, label '0' negative) instances are too costly, in other words predicting a fraudulent case to be non-fraudulent would be a costly mistake.

Table 1 summaries the results of the trained classifier (XgBoost) on the test set (created during the train-test split of the labeled data). While Easy Ensemble and Balance Cascade techniques had high recall values on the

Method	Classifier	Class	Precision	Recall	F1-score
Easy Ensemble	XgBoost	0	1.00	0.86	0.93
		1	0.01	0.93	0.03
		Avg	**1.00**	**0.86**	**0.92**
Balance Cascade	XgBoost	0	1.00	0.90	0.95
		1	0.02	0.93	0.04
		Avg	**1.00**	**0.90**	**0.94**

Table 1: Comparison of Ensemble methods for the imbalanced data problem

minority class (1), they suffered in terms of precision, i.e about 10% of the test data was classified fraudulent, when in reality they were not fraudulent. This translated to poor f1-scores for the minority class. So while the classifier succeeded in achieving has a low false-negative rate (which was the goal), it suffered when it came to false-positive rate. So future work could involve finding the right balance between the false-negative and false-positive rates to augment the precision and consequently the f1-score for the minority class.

References

[1] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39:539 – 550, 05 2009.