Assessing modern classification models and resampling techniques for imbalanced data

Maxime Jan maxime.jan@epfl.ch
Alexander Rusnak amr@alexrusnak.com
Arnaud Robert arnaudemmanuel@epfl.ch

Abstract

Extending the work of another paper [1], we assess the performance of Multilayer Perecptron and Gradient Boosted Classifier on Civil War Onset prediction, which is based on very unbalanced data. We observe that ensemble based methods appear to be the best at handling such data. Furthermore we investigate whether data augmentation and resampling through the SMOTE algorithm can help mitigate the problems induced by class unbalance, and conclude that it does have such potential. Finally, as a creative extension, we explore the predictive power of the features of the Civil War Dataset on a measure of urban unrest from the Urban Social Disorder dataset [2], and find out that these features are reasonable predictors of such events.

1 Introduction

In the paper we are extending [1], the authors are comparing the predictive performance of Random Forest and Logistic on unbalanced data from the Civil War Dataset (CWD) and conclude that the Random Forest is the superior classifier for this type of problem. Each sample in the data represents a specific country, in a given year. The task is to predict based on a selection of features, whether or not a civil war has broken out in that country for this year. We develop this research through three additional questions.

The first is to study if other models are able to make better predictions on this unbalanced data. We will try another ensemble method - Gradient Boosting Classifier - and a Deep Learning one - Multilayer Perceptron. We intend to compare their performance to the paper's classifiers.

Secondly, we are interested in balancing the data before training a classifier. We will explore this solution using the Synthetic Minority Oversampling Technique (SMOTE), which will augment and resample the data. We will then measure how this method improves the Civil War Onsets predictions.

Finally we bring a creative extension to this project using the Urban Social Disorder (USD) dataset [2]. We attempt to reuse the features of the CWD to predict if a country had any urban social unrest during a particular year.

2 Datasets

2.1 Civil War Dataset

The Civil War data were measured annually for each recognized country in the world from 1945 to 2000 [1]. Each entry records various variables (e.g. GDP, illiteracy) as well as a binary one determining whether a Civil War started during the given year in this country. This variable is heavily unbalanced as only 116 of the 7141 entries indicate a civil war Onset.

2.2 Urban Social Disorder

Collected by the *Peace Research Institute Oslo* (PRIO) [2] the USD dataset compiles urban social disorder events occurring in capitals and other major cities through the 1960-2014 period. Each of the 9018 elements of the set constitutes an event, described by 30 features. Among them we are interested in the beginning year of an event and its *PTYPE* (Problem type). This last feature is an ordinal ranking of societal conflict magnitude, based on the degree of institutionalization of political activity. For example, a value of 10 indicates that the event is a *General Warfare* while *50* represents an *Organized Violent Riot*.

3 Methods

3.1 Using new models

We introduce two new classifiers, namely the Gradient Boosting Classifier and the Multilayer Perceptron.

3.1.1 Gradient Boosted Classifier

Similarly to Random Forests, Boosted Decision Trees are an ensemble method. The main difference resides in the fact Trees of the latter are trained sequentially, using the negative gradient of the preceding Tree, thus reducing bias along training. This contrasts with Random Forests for which Trees are trained independently.

3.1.2 Multilayer Perceptron

The MLP is a feedforward neural network used in supervised learning. It consists of fully connected successive layers of varying sizes between the input and the output layers. Nodes in the layers use non linear activation functions such as rectified linear units (relu) or

sgimoid. This architecture allows the model to transform its inputs into more complex representations in order to better separate the data. The model is iteratively trained with the samples and its parameters are updated by computing the negative gradient on the loss function relative to its weights, which is the so-called gradient descent algorithm.

Since neural networks generally rely on geometric relationships between features, it is very important to have a standardized scale. This substantially simplifies the loss topology of the model and allows for faster convergence, and much better results. Thus for this model type, we scaled all features using scikit-learn's standard scaler.

3.2 Augmenting data with SMOTE

As the poor performances of classifiers are partially due to the severe unbalance of the CWD (roughly a 1:100 ratio), we want to test these models again after having balanced the data. To do so, we use one of the most commonly used oversampling method: the SMOTE algorithm.

The SMOTE creates new instances of the minority class (in our case, label *warstds* when set to 1) and makes use of the K-nearest neighbors procedure to do so. The algorithm loops through all the "real" minority class instances. At each iteration, it finds its K nearest neighbors, and then creates "synthetic" minority class instances in the interval between the "real" instance and its "real" neighbors, such that the new instances are convex combinations of the original instance and of one of its neighbors. The procedure is illustrated in Fig 1.

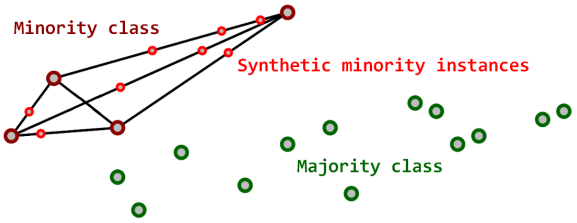

Figure 1: Minority class instance generation through SMOTE [3]

We are interested in testing whether training on balanced data is enough to increase the predictive power of the Random Forest and of the logistic regression in both normal and L2-penalized versions.[1] For this, we split our data between a training and a testing set, and then proceed to train each model twice: once on the raw training data and once on a SMOTE-resampled version. Both versions of each trained models are then tested on the test set, which remained unbalanced. Testing is done by computing the F1 score.

[1]For logistic regression, features used are those of the model in Hegre and Sambanis (2006) [4]

3.3 Predicting urban social unrest

Our third extension is to assess whether or not the features from the CWD can be used to predict the presence of urban social unrest in a given country and year [2].

First of all, we must define what we mean by social unrest and craft a corresponding feature. For a given year, we consider a country is experiencing social unrest if a violent event involving multiple groups takes place. Such events include not only warfare and armed battles, but also pro/anti government terrorism and violent riots. To obtain this feature, we first filter the USD dataset by retaining only events whose *PTYPE* value is different from 60, 61, 62 and 70, thus eliminating all events of a peaceful nature. All these events are then given a *social_unrest* feature equal to 1. We also eliminate all objects from both datasets that have year values that do not overlap with the other, namely we only preserve data from 1960 to 2000. We then proceed to merge the two datasets on the year (through features *year* and *BYEAR*, for CWD and USD, respectively) and the country (through features *cowcode* and *GWNO*). The merge is performed such that elements from the CWD that do not match on the filtered USD are preserved, and assigned a *social_unrest* feature equal to 0.

We then proceed to train four models to predict this new feature, namely the Random Forest, the penalized logistic regression (using the same features as in Hegre and Sambanis, 2020), and our two new models: the MLP and the Gradient Boosting Classifier.

4 Results

4.1 Using new models

Performance results of the new classifiers are displayed on 2. Just like in the original paper, we compute the AUC scores for each methods. The prediction used to draw the curves were computed using cross validation over 10 folds.

4.1.1 Gradient Boosted Classifier

The Gradient Boosted Classifier obtains a AUC score equal to 0.91 which is only slightly more than the score obtained by the Random Forest. We conduct a Kolmogorov-Smirnoff test over each model's distribution of the AUCs scores computed over 10 fold cross validation. The null hypothesis is that the mean of these two distribution are the same. The resulting p-values is equal to 0.99 ($p > 0.05$), and thus we cannot reject the null hypothesis. Similarly we conduct the same test for GBR and the penalized logistic regression, but this time the null hypothesis is rejected as the p-value is equal to 0.012 ($p < 0.05$).

4.1.2 Multilayer Perceptron

Our second model, the MLP, achieves a AUC score of 0.79 which is below the performance of the Random Forest and comparable to that of the logistic regression based model.

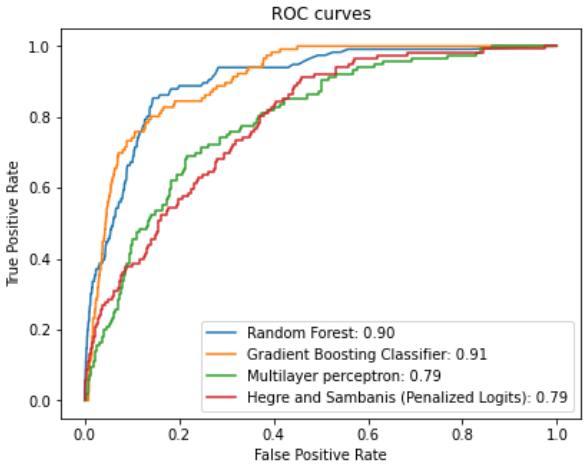

Figure 2: Model performance on predicting civil war onset

4.2 Augmenting data with SMOTE

Results of the SMOTE procedure can be observed on Fig 3. The procedure detailed in the Methods section is reproduced 40 times to obtain confidence intervals on the F1 scores. It appears that resampling the data through SMOTE has a clear and significant effect on the F1 score of all original models. The logistic regression based models perform particularly better even though the score remains rather low, at a value of 0.07 for both models. This suggests that the low predicting power of logistic regression in class imbalanced settings can be somewhat mitigated by training them on artificially balanced data. The Random Forest, which already performed significantly better at predicting civil war onset compared to logistic regression, also obtains a higher F1-score, raising from 0.09 to 0.2.

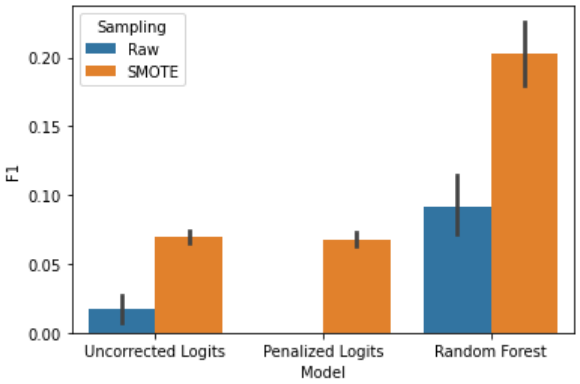

Figure 3: SMOTE effect on F1 score

4.3 Predicting urban social unrest

As described in the Methods section, we compute once again the AUC scores but using our crafted social unrest feature as the target label. The corresponding curves and scores are displayed in Fig 4. One can see that all the models deliver an acceptable performance, with AUC scores superior to 0.7. Interestingly, the two ensembles methods are not the best performers anymore. Indeed, the logistic regression based model now has the highest curve and reaches a AUC score equal to 0.81.

It is important to note that the class imbalance is lesser when considering this new target label. Indeed, after merging the datasets we have that 1285 of the county-years have experienced social unrest, while the remaining 4677 did not. This can partly explain the increased performance of the logistic regression in this new context.

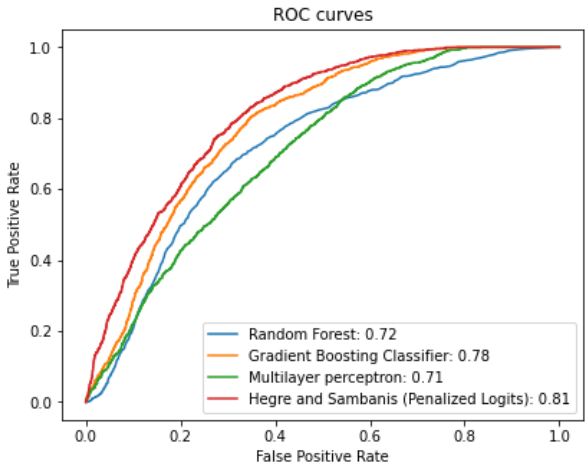

Figure 4: Model performance on predicting social unrest

5 Conclusions

The results displayed by the two new models further confirm the original paper results according to which ensemble methods perform best on classifying imbalanced data. The Multilayer Perceptron does not achieve outstanding results in itself.

We also found out that resampling data through SMOTE during training was a possible method to mitigate class imbalance induced problems, not only for logistic regression based models but also for ensemble based methods such as Random Forest.

Finally, we tried to predict the presence of violent social unrest events in different county-years using the original data from the CWD and found out that these were able to do so acceptably. Interestingly, the best performing method for predicting social unrest, namely a logistic regression based one, was not the same as the best performing method for predicting civil war onsets. This can be explained by the fact that the class imbalance was more mitigated for social unrest prediction, which suggests indeed that logistic regression conserves a strong predictive power when class imbalance is not too important.

References

[1] Muchlinski, D., Siroky, D., He, J., Kocher, M. 2016. *Political Analysis*, 24(1), 87-103.

[2] PRIO visited in December 2020. *https://www.prio.org/Data/Armed-Conflict/Urban-Social-Disorder/*

[3] Rikunert visited in December 2020. *https://rikunert.com/SMOTE_explained*

[4] Hegre, H., and N. Sambanis. 2006. *Journal of Conflict Resolution*, 50(4), 508-135.