A Study of Class Imbalance in Image Classification using Precision-Recall as Evaluation

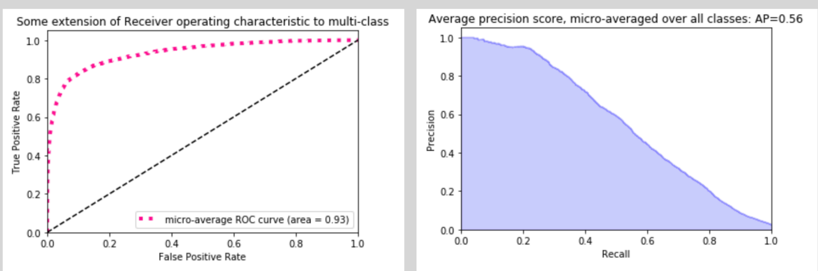
David Young-Jae Kim, SeungJoon Kim, Hao Wu, Aman Satya
University of Colorado, Boulder

Abstract

A common problem in real life applications of classifiers is that some classes have a significantly higher number of examples than other classes, which is referred to as **class imbalance**. In terms of image processing **oversampling** has been proven to best method[1]. However, there is not much study addressing issues using the **Precision and Recall curve** as evaluation which is known to be better than **Receiver Operating Characteristic curve(ROC AUC)**[2][3] in data imbalanced cases. In our study, we use Traffic Sign dataset attain from German Traffic Sign Benchmarks to investigate which methods best fits solving data imbalance problem. <u>Our results show that oversampling did outperform other methods, however, is not enough to justify that oversampling is the general better method.</u>

Introduction

There are lots of classical machine learning methods that deals with data imbalanced. The most common method used is oversampling and undersampling. There has been research that the best method to use is oversampling for deep learning[1]. We investigate this belief and prove if it is true with our dataset. Furthermore, this paper used ROC-AUC area to evaluate the best model. However, there are many research that when the data is imbalanced toward certain classes ROC-AUC could be misleading. In this case, precision recall curve is believed to be better[3].

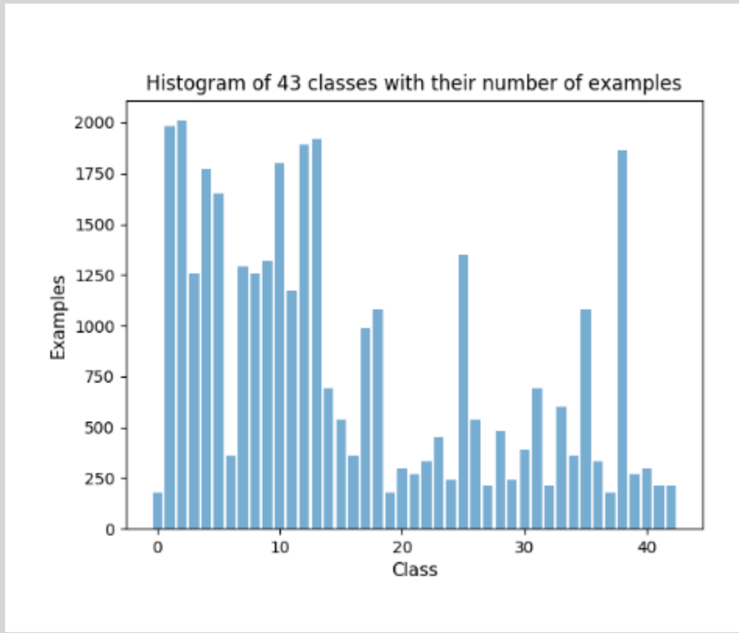


<Figure 1>

Above is a result we got using a simple model and the base data set. The results in the ROC curve show that the model is relatively good, however the PR curve is showing something different.
We will evaluate our model and see if oversampling is still the best method when Evaluated using the Precision-Recall curve

Material & Method

Data set



The data set we used for this study Traffic Sign dataset: Size of the data set 34799 training data, 4410 Validation data, 12630 Test data. The images are 3 * 32 * 32 pixels composed by RGB layers.

The method we attempted with this data is as follows

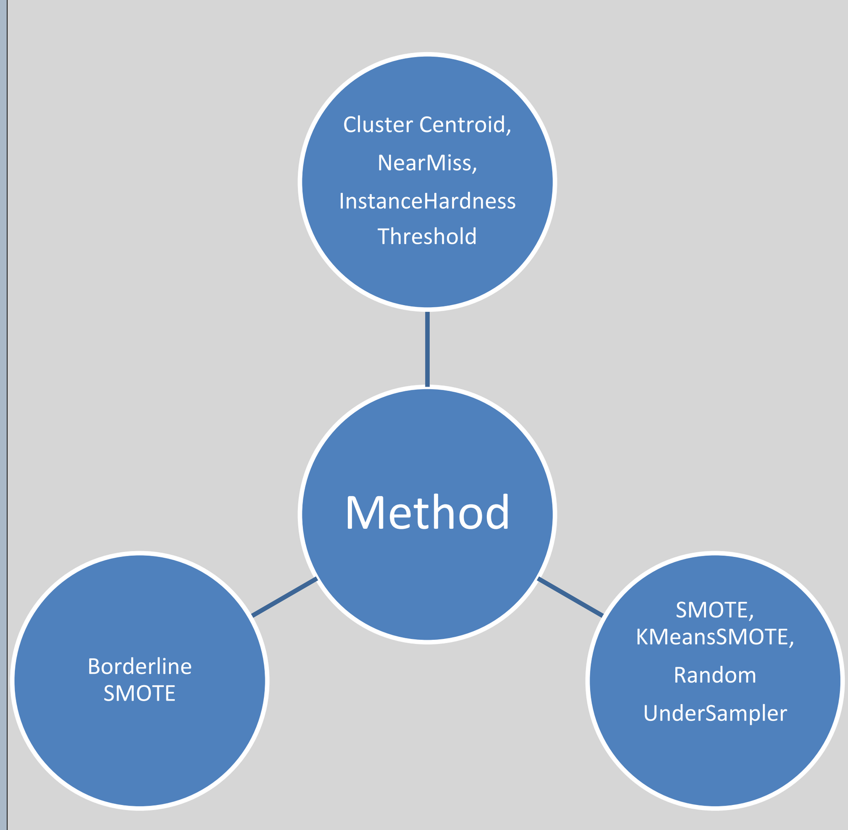
Oversampling
SMOTE
BorderlineSMOTE
KmeansSMOTE

Size of the data set 86430 training data, 4410 Validation data, 12630 Test data.

Undersampling
RandomUnder Sampler
NearMiss
InstanceHardness Threshold
Cluster Centroids

Size of the data set 7740 training data, 4410 Validation data, 12630 Test data.

Below is a grouping according to our interpretation



For the Convolutional Neural Network Architecture we have constructed three architecture. One that is very simple, one that is a little more complex and the final one that is most complex.

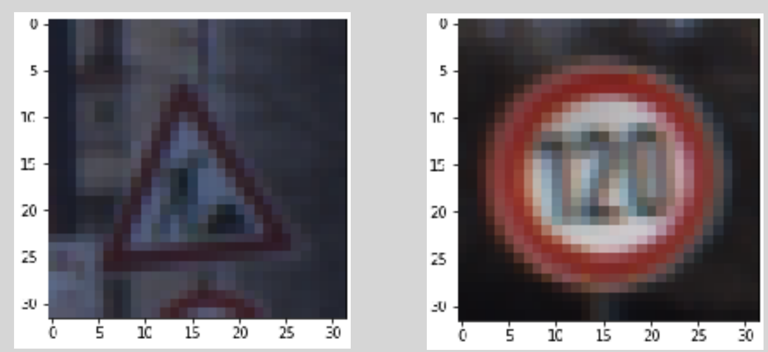
Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 30, 30, 32)	896
max_pooling2d_1 (MaxPooling2	(None, 15, 15, 32)	0
conv2d_2 (Conv2D)	(None, 13, 13, 32)	9248
max_pooling2d_2 (MaxPooling2	(None, 6, 6, 32)	0
conv2d_3 (Conv2D)	(None, 4, 4, 64)	18496
max_pooling2d_3 (MaxPooling2	(None, 2, 2, 64)	0
flatten_1 (Flatten)	(None, 256)	0
dense_1 (Dense)	(None, 500)	128500
activation_1 (Activation)	(None, 500)	0
dense_2 (Dense)	(None, 43)	21543
activation_2 (Activation)	(None, 43)	0

Total params: 178,683
Trainable params: 178,683
Non-trainable params: 0

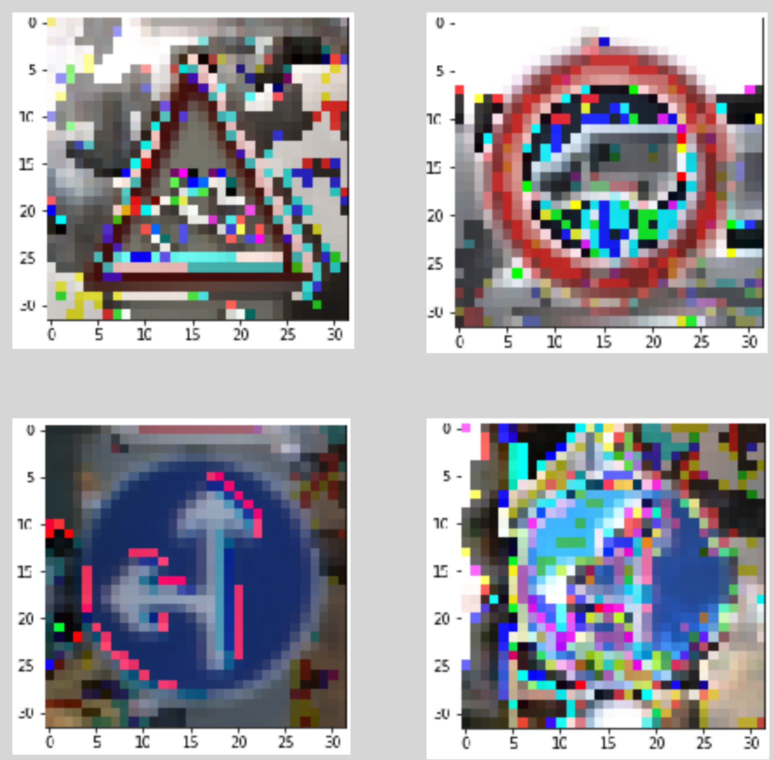
Above is an example of one of the model.

Results and Discussion

The following images are samples of the original image after normalized and rescaled Figure 2 and oversampled images Figure 3



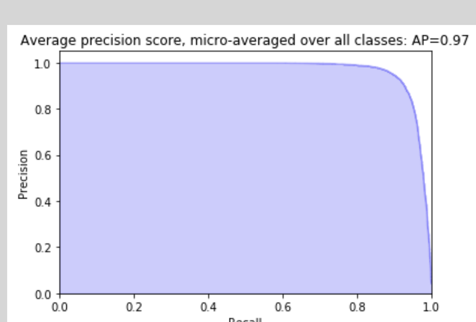
<Figure2>



<Figure3>

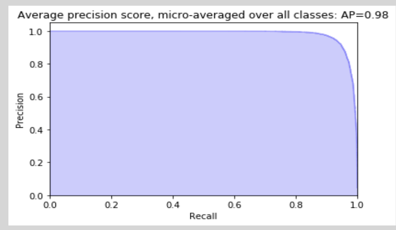
The following is a results using the original data

test_accuracy: 91.83%

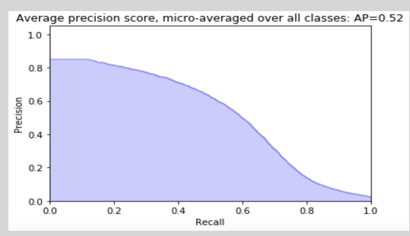


The following is a results using Borderline SMOTE and NearMiss method respectively

test_accuracy: 93.52%



test_accuracy: 55.62%



As we can see above the undersampling method made the results worse. The oversampling method did improve but not enough to justify the conclusion that it is a better method.

Conclusions

The conclusion that we have arrived is that oversampling did come out to be the better method even when we use the Precision Recall as the evaluation metric in imbalanced image data. However, not enough to justify to say that oversampling is the rule of thumb method. <u>We believe this is because that latent structure of the original data is already distinguishable</u>, which is why undersampling didn't help because it merely decreased data and oversampling created data that was redundant. <u>We believe our next step should be to study the structure of the data and its relation to methods.</u>

Reference

[1] M. et al., A systematic study of the class imbalance problem in convolutional neural networks, Neural Networks 106 (1-12) (2018) 249–259.
doi: 10. 1016/ S0031-8914(53) 80099-6

[2] M. G. Jesse Davis, The relationship between precision-recall and roc curves, Proceedings of the 23rd international conference on Machine learning 24 (2006) 233–240.
doi: 10. 1145/ 1143844. 1143874

[3] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, PLoS One
doi: 10. 1371/ journal. pone. 0118432
.

[4] R. B. F. H. Enislay Ramentol, Yail Caballero, A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory, Springer-Verlag London Limited 1.
doi: 10. 1007/ s10115-011-0465-6

[5] Y.-S. L. Show-Jane Yen *, Cluster-based under-sampling approaches for imbalanced data distributions, Expert Systems with Applications 36.
doi: 10. 1016/ j. eswa. 2008. 06. 108