Team V Project Mid Report

David Young-Jae Kim[a], Seoung-Joon Kim[a], Hao Wu[a], Aman Satya[a]

[a] *University of Colorado, Boulder*
Computer Science Department

Keyword: Convolution Neural Network, Imbalanced Dataset, ROC-AUC curve, Precision and Recall, Multi-class Classification.

Abstract: Class imbalance is a common problem that has been aroused in Machine Learning, Deep Learning. In terms of image processing oversampling has been proven to best method[1]. However, there is not much study addressing issues when a certain class is more important than others especially using the Precision and Recall curve which is known to be better than Receiver Operating Characteristic curve(ROC AUC)[2][3] in this case. In our study, we use Traffic Sign dataset attain from German Traffic Sign Benchmarks to investigate which methods best fits solving data imbalance problem when certain classes are more important than others. The evaluation metric we will mainly use is area under the receiver operating characteristic curve (ROC AUC) and Precision Recall curve.

1. Introduction

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. It has been studied that class imbalance can lead to very misleading model training and eventually effecting generalization of the model. While there were systematic study on this topic [1], there are still issues that need to be addressed such as what is the best method when certain classes are more important than other classes. There are lots of classical machine learning methods that deals with data imbalanced. The most common approach is manipulating the data itself to make the imbalanced data balanced. The most common method used is oversampling and under sampling. We can also approach this in the classifier level. Learning algorithms are modified by giving different weights to different class classification (cost sensitive learning)[4], adjusting prior class probabilities, or introducing new loss functions for neural networks. There has been research that the best method to use in these cases is oversampling for deep learning[1]. We will tackle this belief and prove if it is true with our dataset and also test it's robustness when certain classes are more important than others.
Furthermore, this paper used ROC-AUC area to evaluate the best model. However, there are many research that when certain class is more important than the other ROC-AUC could be misleading. In this case, precision recall curve is believed to be better[3]. We will evaluate our model using this method and investigate if oversampling is still the best method when there is a data imbalanced.

2. Project Description

2.1. Methods

2.1.1. Data Level

Oversampling Over-sampling refers to various methods that aim to increase the number of instances from the underrepresented class in the data set. Duplicate the observations of the minority class to obtain a balanced dataset which is a random naive over-sampling. Although it is known to be a very effective method it is also prone to overfitting, in order to overcome this the new algorithm SMOTE(Synthetic Minority Over-Sampling Technique) has been developed[5]. While SMOTE still oversamples the minority class, it does not rely on reusing previously existing observations. Instead, SMOTE creates new (synthetic) observations based on the observations in your data.

Undersampling This method drop observations of the majority class to obtain a balanced dataset. While intuitively this seems ineffective since it discards datasets, there is some evidence that in some situations undersampling can be the better option compared to oversampling. There are two popular undersampling methods: Cluster Centroid based Majority Under-sampling Technique (CCMUT) and Extended Cluster Centroid based Majority Under-sampling Technique (E-CCMUT)[6].

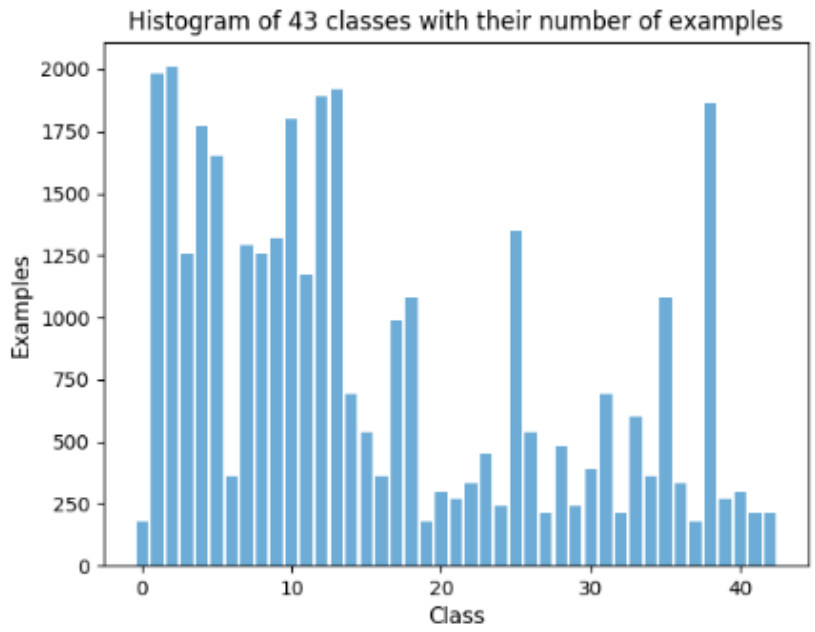
2.1.2. Model Level

Cost sensitive learning Cost-Sensitive Learning is a type of learning in data mining that takes the miss classification costs (and possibly other types of cost) into consideration[7]. Different cost function for example: MSE, Hinge Loss, MAE, etc. These cost functions can have different emphasis, so we can choose different cost/objective functions to see how it works differently.

Hybrid of methods/models This is an approach that combines multiple techniques from one or both above categories. We can train different models based on different methods and cost functions, then combine these models[8].

2.2. Data

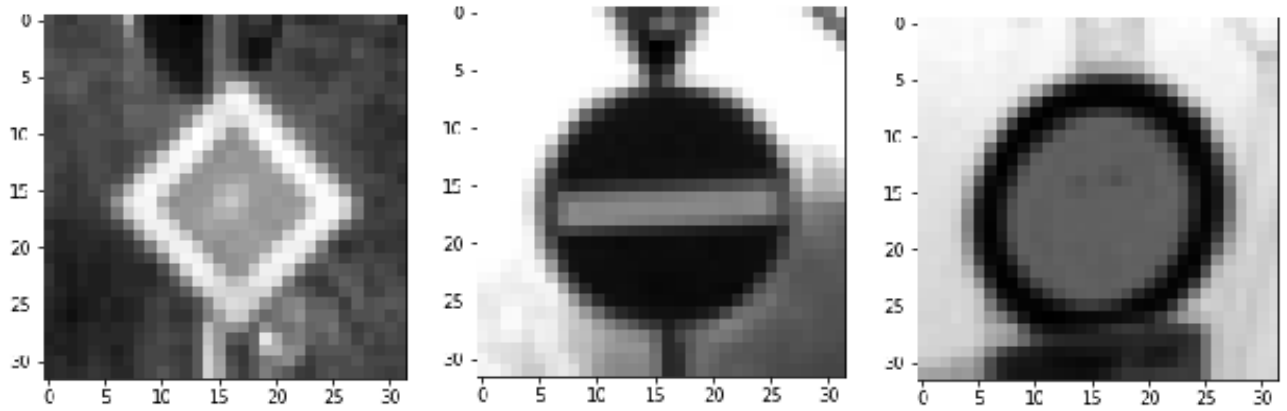
Traffic Sign dataset: Size of the data set 86989 training data, 4410 Validation data, 12630 Test data.
The images are 3 * 32 * 32 pixels composed by RGB layers.



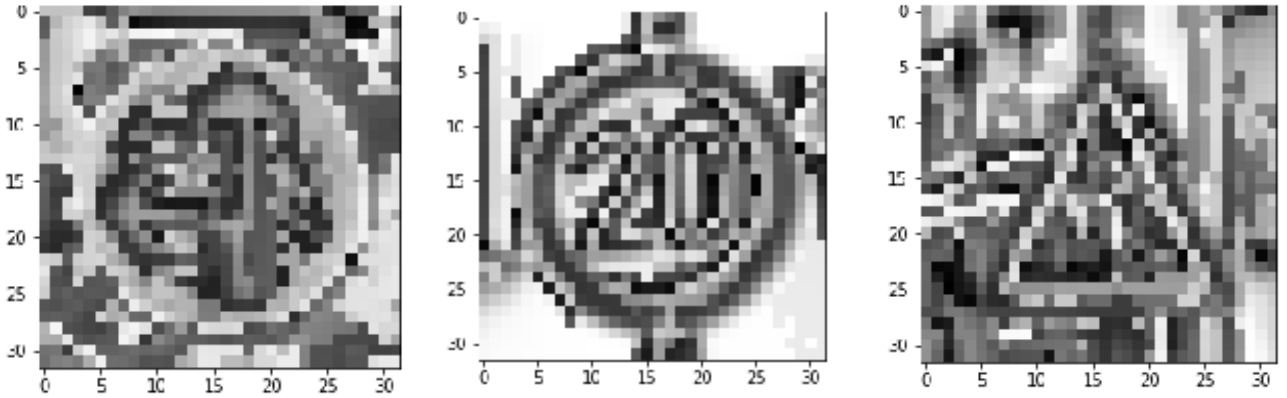
This dataset has 43 labels:
0: Speed limit (20km/h) 1: Speed limit (30km/h) 2: Speed limit (50km/h) 3: Speed limit (60km/h) 4: Speed limit (70km/h) 5: Speed limit (80km/h) 6: End of speed limit (80km/h) 7: Speed limit (100km/h) 8: Speed limit (120km/h) 9: No passing 10: No passing for vehicles over 3.5 metric tons 11: Right-of-way at the next intersection 12: Priority road 13: Yield 14: Stop 15: No vehicles 16: Vehicles over 3.5 metric tons prohibited 17: No entry 18: General caution 19: Dangerous curve to the left 20: Dangerous curve to the right 21: Double curve 22: Bumpy road 23: Slippery road 24: Road narrows on the right 25: Road work 26: Traffic signals 27: Pedestrians 28: Children crossing 29: Bicycles crossing 30: Beware of ice/snow 31: Wild animals crossing 32: End of all speed and passing limits 33: Turn right ahead 34: Turn left ahead 35: Ahead only 36: Go straight or right 37: Go straight or left 38: Keep right 39: Keep left 40: Roundabout mandatory 41: End of no passing 42: End of no passing by vehicles over 3.5 metric tons
Official site of this dataset(link): German Traffic Sign Benchmarks.

3. Data Preprocessing

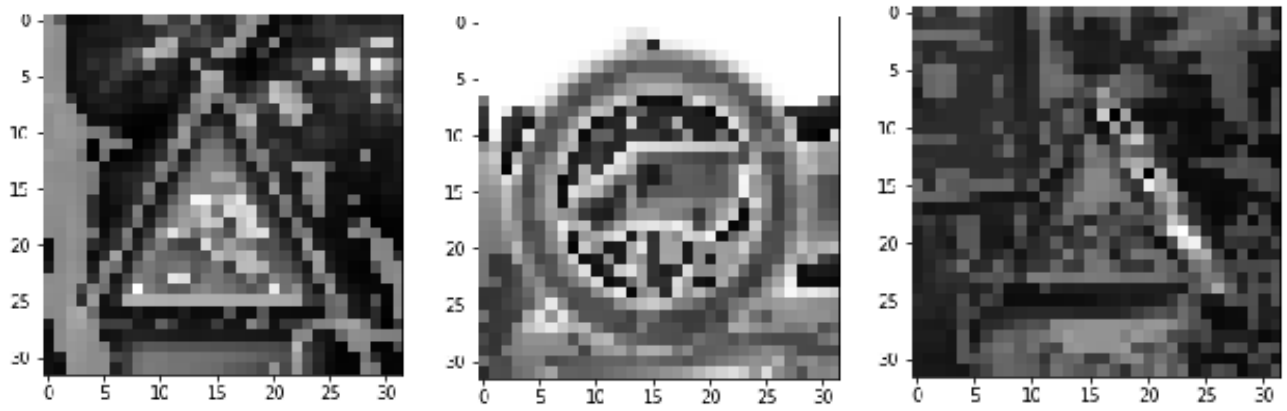
For the preprocessing, we resized the images to 36x36 with and without applying grey scale. With the that being done we applied 3 types of oversampling (SMOTE[9], Borderline SMOTE[10], Kmean SMOTE[11]) and 4 types of undersampling (ClusterCentroids[12], NearMiss, RandomUnderSampler, InstanceHardnessThreshold).



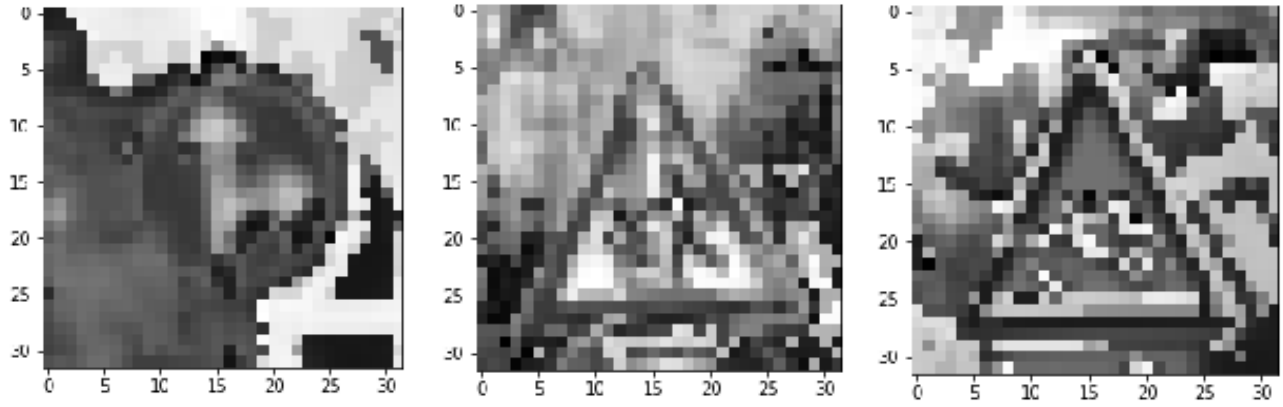
Above are images of the original images resized to 36x36 and grey scaled.



Above are images of the SMOTE sampled images



Above are images of the Kmeans SMOTE sampled images



Above are images of the Borderline SMOTE sampled images.

4. Results

For the baseline result, we ran a two layer simple CNN architecture with the SMOTE dataset. The data was divided into train, validation, test. We optimized our parameters according to the precision and recall value in the validation set. The result below is from the test data.

	precision	recall	f1-score	support
accuracy			0.84	12630
macro avg	0.84	0.79	0.80	12630
weighted avg	0.87	0.84	0.84	12630

5. Further Research

The first thing we must do further is to construct our own CNN architecture or implement some of the state of the art CNN architectures. After that we must test all of our manipulated dataset(3 oversample data and 4 undersample data) and implement the threshold method by making our own loss function, we will compare which one performed the best in method wise and CNN architecture wise according to the average precision recall value. Also, we discovered that this dataset is fairly easy to distinguish thus, it seems that using the Receiver operating characteristic(ROC) curve or the precision recall curve does not make much difference. We would like to further investigate if this is true and generalize when it is better to use the precision recall curve rather than the ROC curve.

References

[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] *S. S. Japkowicz Nathalie, The class imbalance problem: A systematic study., Intelligent Data Analysis 6 (2002) p429. 21p.* doi:10.3233/IDA-2002-6504.

[5] *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.

[6] *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.

[7] *V. S. S. Charles X. Ling, Cost-sensitive learning and the class imbalance problem, Encyclopedia of Machine Learning 1.* doi:10.1.1.164.4418&rep=rep1&type=pdf.

[8] *Q. Wang, A hybrid sampling svm approach to imbalanced data classification, Hindawi Publishing Corporation Volume 2014.* doi:10.1155/2014/972786.

[9] N. V. Chawla, Smote: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research Volume 16. doi:https://doi.org/10.1613/jair.953.

[10] H. Han, Borderline-smote: A new over-sampling method in imbalanced data sets learning, International Conference on Intelligent Computingdoi:https://doi.org/10.1007/11538059_91.

[11] F. Last, G. Douzas, F. Baão, Oversampling for imbalanced learning based on k-means and smote, ArXiv abs/1711.00837.

[12] S.-J. Yen, Cluster-based under-sampling approaches for imbalanced data distributions, Expert Systems with Applicationsdoi:https://doi.org/10.1016/j.eswa.2008.06.108.