A Comparative Analysis of the Impact of Data Imbalances on the Learning Curve

Jia Jie Feng
J.J.Feng-1@student.tudelft.nl
Supervisors: Dr. Tom Viering and Mr. Taylan Turan

1.Introduction

Learning Curves are visual representations that show the progression of the learning model's performance as the train size increases.

Real-life data typically involves imbalance datasets. **Imbalance datasets** are datasets where one class consist of relatively more data points than other class in the dataset.

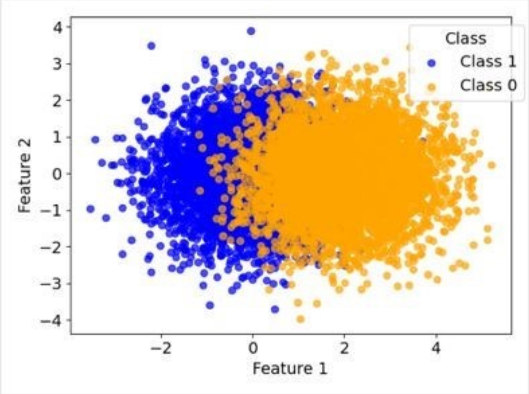
Definitions for the experiment

- **Imbalanced Ratio** is defined as [0.1,0.2,0.3,0.4,0.5], they represent the percentage of the dataset that consist of the minority class. Ratio 0.5 indicates a balanced dataset.
- **Positive Class** is defined as Class 1, representing the minority class
- **Negative Class** is defined as Class 0, representing the majority class

Research Question: How do Data Imbalances affect the Learning Curve using Nearest Mean Model?

2.Methodology

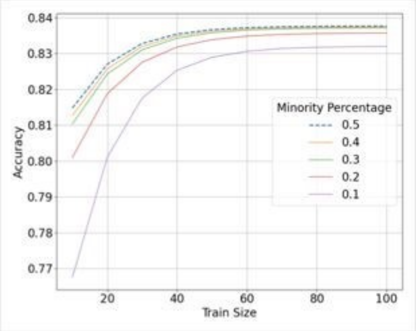
The dataset used for the experiment is sampled from a **Multivariate Gaussian Distribution**. The distribution is defined as: Class 0 with mean [2,0] and Class 1 with mean [0,0]. Both covariance is an identity matrix.



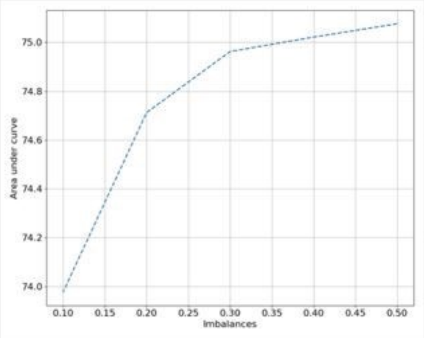
The experiment uses 100 training samples based on the imbalance ratio and 10000 balanced test samples. The nearest mean model is trained on subsets of [10,20,30,40,50,60,70,80,100] training size. The entire experiment is repeated 1000 repetitions to ensure reliability of the results by averaging the accuracy.

3. Results

Performance

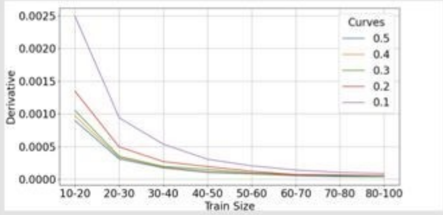


The accuracy learning curves of all the imbalance ratios tested. The **balanced ratio performed the best** throughout the training of the nearest mean model.



The area under the accuracy curve also indicates that balanced ratio performed the best throughout the training. The **training performance decreases as the imbalance increases**.

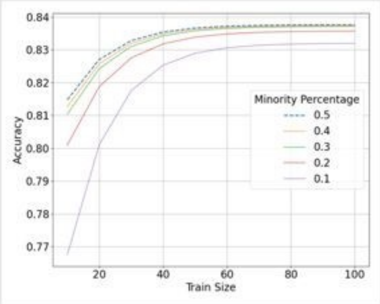
Plateau Analysis



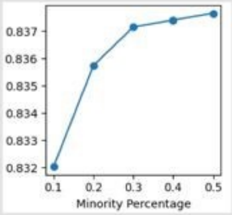
Plateau analysis determines the point at the learning curve where increased of training size will not significantly improve the accuracy of the model. A threshold of 0.0001 was used to determine the plateau for this analysis. Curve ratios of [0.5,0.4,0.3] plateau at train size 50. Curve ratios 0.2 and 0.1 plateau at train size 60 and 80 respectively. **Increased in imbalances result in the increase of the plateau point.**

Shape of the Curve Analysis

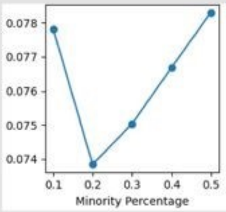
Curves were fitted onto the original curve using the logistic function and the estimated parameters "a", "b", and "c" were extracted. Below illustrate the results of the fitted curves.



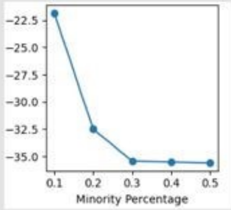
The fitted curves missed a few data points on the original curves but the relationship among the fitted curves mirror the original curves. Additionally, the square error were minimal, indicating relatively reliable curves.



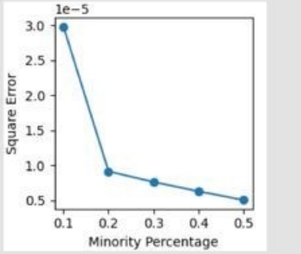
This figure represents the "a" parameter compared with other curves. Parameter "a" represents the maximum achievable accuracy. **As imbalances increase, the maximum achievable accuracy decreases**. Differences are negligible.



This figure represents the "b" parameter compared with other curves. Parameter "b" represents the growth or steepness of the curve. **No discernible relationship can be established among the curves**. Differences are negligible.

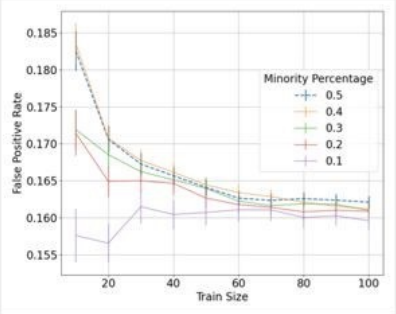


This figure represents the "c" parameter compared with other curves. Parameter "c" represents horizontal shift of the logistic function. **As imbalances increase, the curve shifts more to the right.** Differences are significant.

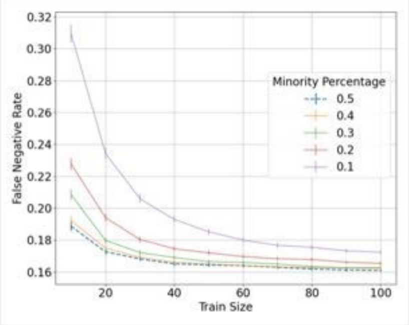


This figure represents the squared error of the fitted curve when compared with the original curve. As imbalances increase, the fitting of the curve represent less of the original curve. The errors are non-significant.

False Positive



The figure illustrates the false positive rate of the model during training. It behaved as expected. **As imbalances increase, the false positive rate decrease**. Since there are more majority data points for the model to learn from.



This figure illustrates the false negative rate of the model during training. It behaved as expected. **As the imbalances increase, the false negative rate increases**. Since there are less minority data points for the model to learn from.

4. Limitation

1. Train size limited at 100 due to simplicity of the data.
2. Imbalance ratio intervals are fixed at 0.1 and subset size of the training sample fixed at intervals of 10, due to the train size limit of 100 and effort to maintain the consistency of class representation throughout the training.
3. Only one model was used.
4. Only one data set was used.

5. Conclusion

1. The overall performance of the model decreases as imbalances increases in the dataset.
2. The plateau point at which the curves no longer significantly increase in accuracy using the threshold 0.0001 increases as imbalances in the data increases.
3. The amplitude and growth rate of the curves showed no significant differences among the curve. The curves shift to the right as imbalances increase. The differences are significant.
4. False positive rate exhibits increasing noise as the imbalances increase. However, it follows the expectation that with increase in imbalance, the false positive rate decreases.
5. False negative rate behaved as expected. The false negative rate increases as the imbalances increases, indicating a monotonic relationship.
6. The false negative are more impactful on the curve than false positives due to the relatively large difference between the false negative rates between the curves.

Future Works

1. Experiment with diverse set of models
2. Experiment with diverse set of data
3. Experiment with higher imbalances in the dataset