Exploring Subclass Imbalance Effect Models Mispredictions In Balanced vs Unbalanced Hidden-Subclasses

Ella Shalom 208288423

Abstract

Analysis of model faults and mispredictions is one of the main challenges researchers face today. One reason for those mispredictions may arise from poor identification of important subsets of a population. The identification of these subsets plays an important role in many applications, for example model validation and testing, or evaluation of model fairness. In model fairness, as well as in medical field, the costs of different types of mistakes are not equal. All of those raise the need to address this problem, which is also referred as Hidden Stratification. My solution consist of three parts: subclass detection in an unsupervised manner, analysis of model misprediction in each subsets, and further exploring the Hidden Stratification by artificially changing the relation between the set and the subset. My main results are that using the feature space is helpful and Gaussian Mixture gets very good results. I found strong indication for Hidden Stratification in all the datasets that I experimented on.

1 Problem Description:

Explaining faults and mispredictions is crucial for improving models, create better ones and understanding how to use them correctly. Even if we think we understand the reason for the misprediction, it may result from a deeper reason that is not easily understood by us. One possible reason for it is subclass imbalance, also known as Hidden Stratification. This is a phenomenon when a class contains small subsets of samples that have hidden unique characteristics that cause failures in these cases, for seemingly unknown reasons.

Awareness of this problem is especially crucial in safety-critical applications such as medicine. For example, overall performance of a cancer detection model may be high, but the model may still consistently miss a rare but aggressive cancer subtype. In this case, the poor performance on subset, might not be observed because of lack identification of it, and the good performance of the set overall[1].

Hidden Stratification is also crucial in models fairness. For example, when calculating the chances of criminal defendant of committing another offense in a period of two years. Out of dataset consists of different people, the subset of grown african-american men prediction was different from the one for the overall set. People belonging to this data subset tend to be wrongly assigned to high risk of recidivism than the dataset overall[2]. This relates to Fairness as we discussed in class.

There are three main methods to detect Hidden Stratification[1]:(1) exhaustive prospective human labeling of the data, called schema completion,(2) retrospective human analysis of model predictions, called error auditing, and (3) automated algorithmic measurement methods to detect hidden strata. Each of these methods is applied to the test dataset, allowing for analysis and reporting of subclass performance. Schema completion involves providing a more complete set of subclasses to the test set, allowing for consensus on subclass definitions. However, it is time and money consuming and limited by the author's understanding of the data. Error auditing involves examining model outputs for unexpected regularities, but is dependent on the ability of the auditor to visually recognize differences and is still time-consuming. Algorithmic measurement involves designing a method to search for subclasses automatically, such as unsupervised clustering, and is the most efficient method, and the one I will explore here.

2 Related work

The issue of Hidden Stratification has been extensively researched in the medical field due to its significant implications for potential failures of deep learning in clinical practice. In [1], an assessment of the utility of several techniques for measuring Hidden Stratification effects on multiple real-world medical imaging datasets was presented. The researchers found evidence that Hidden Stratification can occur in unidentified imaging subsets with low prevalence, low label quality, subtle distinguishing features, or spurious correlates. They also found that it can result in relative performance differences of over 20% on clinically important tasks.

My project was inspired by this research and sought to explore different subclass detection methods with the aim of uncovering Hidden Stratification in datasets. While some existing works has focused on measuring the effects of Hidden Stratification, such as [1] and [2], my solution takes a different approach by exploring various unsupervised clustering algorithms that can detect subclass structure in the data. In [2] they develop an algorithms that allow to efficiently estimate the divergence in classifier behavior for all subgroups. They define a concept to measure difference in statistics on the subgroup compared to the entire dataset, and call it Divergence.

Additionally, I was also inspired by the work on robustness and fairness in machine learning, such as the GDRO method presented in [3] and the GEORGE algorithm developed in [4]. These methods aim to optimize worst-case performance over a known set of subgroups, but often encounter difficulties when subgroup labels are unavailable. To address this, the GEORGE algorithm was developed, which is a two-step procedure that first estimates subclass labels and then exploits these estimates to train a robust classifier. The robust classifier aims to maximize robust accuracy, which is defined as the worst-case expected accuracy over all subclasses.

While my project did not directly explore these methods, they provided important context for my research where I tried to develop a method which will not require from researchers to train a new model but to use the data available to them from their already trained model.

3 Solution Overview

3.1 Step 1: Choosing the clustering space

When a researcher tries to explore if there exist a problem of Hidden Stratification there are two possible spaces available to him. The raw data after preprocessing and the feature space from the model which was created. Raw data clustering can still be useful when the data is relatively low-dimensional and the patterns are easily observable. Therefor, I expected clustering in the raw data space to be good comparison to my solution. The advantage of clustering in feature space is that it can be more effective at identifying patterns and similarities between high-dimensional data points. Additionally, the transformed feature space can often reveal important underlying structure in the data that may not be immediately obvious in the raw data. I wanted to explore both options available to a researcher, so I trained a model at the beginning, stimulating a condition where the model is already available.

3.2 Step 2: Finding Organic Hidden Stratification

I wanted to try to find subclasses for each of the classes, and explore the model prediction on each subclass. The purpose of this is step is to find organic sub-classes which exist in the data, without any manipulation to it as I did later on. When the divergence equals to zero, it means that there is no difference in the evaluation metric between the class and the subclass. Meaning, this clustering does not contain valuable information which help explain model misprediction. When there are large deviations between the subclass divergence scores, it indicates that we found meaningful sub-classes which shows that the dataset is suffering from Hidden Stratification.

As my goal is to find Hidden Stratification, I believed that trying different clustering methods could be beneficial for researchers, as each method may perform better in different scenarios. To achieve this goal, I decided to inspect several unsupervised clustering algorithms: K-means: groups data points into k clusters based on their similarity. DBSCAN: groups data points based on their density and separates outliers. Gaussian Mixture Model: probabilistically assigns data points to clusters based on their probability density. BIRCH: hierarchical clustering algorithm that builds a tree structure from the data points. OPTICS: density-based clustering algorithm that extracts clusters using a reachability graph and hierarchical clustering.

The clustering is performed on the whole dateset, and then I randomly select 20% of the data as test set. In my research proposal I presented two ways to select the test est. But later on I understood that choosing uniformly at random from the dataset is similar to choosing randomly from each cluster proportional to the cluster size, when the size of the data is sufficient.

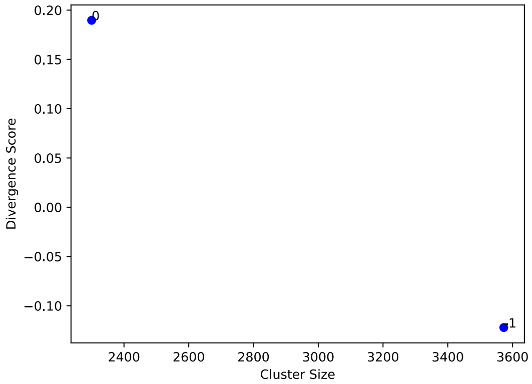

Figure 1: divergence score with respect to cluster size on Bank dataset with DBSCAN and feature space

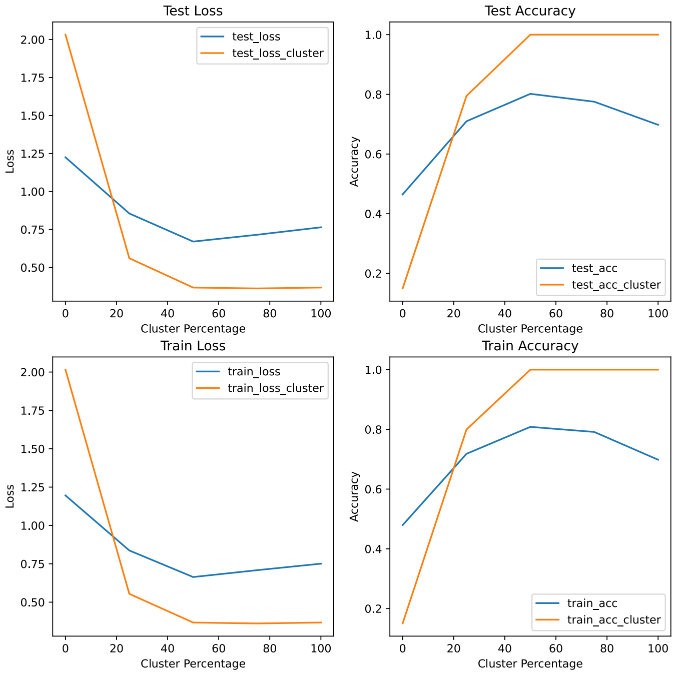

Figure 2: Imbalance creation with respect to overall test and train accuracy and chosen cluster test and train accuracy. Presented on Bank dataset with DBSCAN and feature space

3.3 Step 3: Exploring Artificial Imbalance Creation

I wanted to further investigate the affect of subclass imbalance on the class performance. To do that we can take two approaches. Undersampling a well represented subclass or oversampling an underrepresented subclass. The problem of synthetic data creation is much harder, so i decided on the former. I chose the subclass with the greatest divergence score, which is the subclass the model is most successful on. Choosing this subclass has higher probability that it is not a small subclass.

I decreased the size of the chosen subclass in the train set while maintaining the number of samples in the train set to neutralize the impact of train set size on the predictions. I varied the proportion of the chosen subclass used for training, starting from 100% and descending to 0%. For each trained model I tested it on a test set. I analyzed the influence of this technique on the accuracy for both the entire class and each subclass. My theory was that if the performance of the whole class gradually declines as the subclass imbalance worsens, and if the underrepresented subclass experiences a decrease in performance as well, it indicates that the imbalance in the subclass is the reason for the decline in performance.

4 Experimental Evaluation

4.1 Datasets

Adult dataset: Demographic and income data from the 1994 U.S. Census used to predict income thresholds. German Credit Data: Credit history and demographic data from a German bank used to predict loan defaults. Bank Marketing dataset: Portuguese bank marketing campaign data used to predict term deposit subscriptions. Heart dataset: Patient demographic and medical data used in medical research to predict heart disease likelihood.

4.2 Evaluation

For all datasets I preformed the process described above. Using a different combination of input space and unsupervised clustering algorithms. This combination creates a specific experiment, and for each experiment I perform, I produce three graphs to represent the results.

4.2.1 Divergence score vs. subclass size

Examining the size of a subclass can be intriguing, as a smaller subclass with distinctive features can pose a more significant difficulty for the model. However, it's important to note that the subclass size doesn't always directly correlate with the divergence score, even if there are indications of Hidden Stratification. For instance, in a five-class clustering, the third smallest subclass may receive the lowest evaluation metric score.

To quantify the quality of clustering and the existence of Hidden Stratification, I computed the maximum difference in subclass divergence from zero, and the standard deviation of the divergence scores of the subclasses. I found. As seen In Figure 1, there is one big cluster which outperforms the whole class and a smaller cluster with smaller accuracy than the class, which might indicate Hidden Stratification.

4.2.2 TSNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique that reduces high-dimensional data into a lower-dimensional space while retaining similarities between the data points. Typically, this is done in two dimensions for easy visualization. I did it in a two-step process to optimize the results and speed up computation. Firstly, Principal Component Analysis (PCA) to reduce the dimensions to 25, followed by t-SNE to further reduce it to 2 dimensions. This approach provides better results while being more computationally efficient. In Figure 3, there are two observable clusters.

4.2.3 Imbalance plot

After conducting the imbalance experiment outlined in section 3.3, I created a set of four graphs. The initial two graphs illustrate the loss incurred by the train and test sets. The remaining two graphs, plot the accuracy of the entire class and chosen clusters against the cluster percentage. These graphs provide insights into the effectiveness of the clustering algorithm used in the experiment. By analyzing the loss incurred by the train and test sets, we can determine if the algorithm is overfitting or underfitting the data. In Figure 2 the accuracy of both test and train improves as the chosen cluster percentage is getting bigger.

It is generally expected that an underrepresented class would perform worse, but the overall performance of the entire class may still be better because the other clusters have better representation.

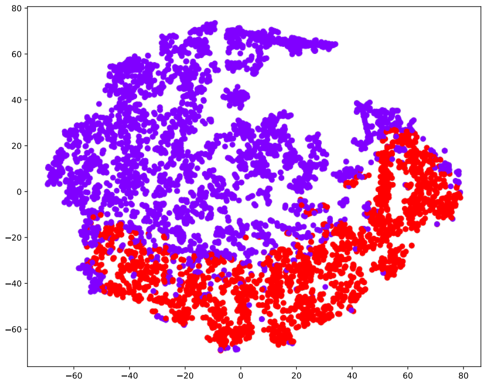

Figure 3: Result of TSNE plot on Bank dataset with DBSCAN and feature space

4.3 Baseline

I needed baseline to compare all my experiments. I chose a simple version of my solution: input clustering space with K-meams clustering algorithm. Presented in Table 1 is a comparison between the results of the datatsets using the input sapce as clustering space or using the extracted features from the model. for each experiment there are two results- The standard deviation of the divergence scores of the subclasses and the maximum difference in subclass divergence from zero. The right column (Input space) will serve as baseline for the experiments.

Dataset		Clustering Space	
Dataset Name	Subclass	Feature	Input
Adult	Class 0	**(0.12, 0.23)**	(0.10, 0.18)
	Class 1	**(0.39, 0.67)**	(0.19, 0.33)
German Credit	Class 0	**(0.55, 0.82)**	(0.07, 0.12)
	Class 1	**(0.24, 0.52)**	(0.05, 0.06)
Bank Marketing	Class 0	**(0.49, 0.81)**	(0.11, 0.16)
	Class 1	**(0.41, 0.70)**	(0.13, 0.19)
Heart	Class 0	**(0.29, 0.38)**	(0.22, 0.37)
	Class 1	(0.13, 0.24)	**(0.14, 0.27)**

Table 1: Input clustering space with K-meams clustering algorithm. In each cell (std, max sdivergence).

4.4 Inputs Space Vs. Feature Space

In 1 presented the result of all datasets with baseline parameters except for feature space instead of input space. It seems that feature space can improve the baseline results in almost all cases. It may be even more effective for sophisticated models trained by researchers with enough time and resources that I did not have.

4.5 Exploring Different Evaluation Metrics of Hidden Stratification

In order to determine the presence and measure the extent of Hidden Stratification, I employed accuracy metric across the entire class and computed accuracy for the chosen subclass. Other metrics that rely on false positive and false negative values may not provide significant insights when applied to a single cluster, as it is reasonable to assume that the subclasses will consist of only one class.

I also computed the divergence of each subclass [2]. Subclass divergence refers to the difference in statistics between different subgroups of a population and the overall population. Positive divergence means better prediction for the subclass in comparison to the class, and negative means the opposite. It measures how much more or less likely a particular subgroup is to receive certain outcomes compared to the overall population.

4.6 Clustering Methods

Based on the results shown in Table 2, GMMs appear to perform well, although other methods may be more effective in certain situations. Therefore, it can be inferred that each clustering approach has the potential to be successful. Additionally, experimenting with various parameters could prove advantageous in tailoring the clustering method to specific scenarios.

4.7 Results Summery

In my research, I discovered evidence of Hidden Stratification in all of the datasets I analyzed. This is a crucial finding that can help to explain model misspredictions and improve model Fairness. Additionally, I found that Hidden Stratification doesn't always occur in the smallest subclasses, as demonstrated in Figure **??**.

Different clustering methods can be good for different scenarios. The TSNE plots can be useful tool for researchers in examining the subclasses, for example the Birch clustering method which received the best result on class 1 of the bank marketing dataset as presented in **??**. The effect of imbalance procedure I explored resulted in an overall trend which seems to support the quality of the clustering, as presented in **??**. However, it's important to note that each training process have a random start, and many other factors can influence the results. The feature space is a excellent clustering space for identifying Hidden Stratification, but it can also be challenging to address. Nonetheless, techniques like feature space augmentation can be attempted. Overall, if using this clustering space helps to identify the problem, which can be complicated at times, it is worthwhile to pursue.

5 Conclusion

In conclusion, this project provided valuable insights into the phenomenon of Hidden Stratification in datasets and the impact of subclass imbalance on clustering performance. Through my analysis, I learned several key lessons that will be valuable for future research.

Dataset		Clustering Method				
Dataset Name	Subclass	K-means	DBSCAN	OPTICS	BIRCH	GMMs
Adult	Class 0	(0.12, 0.23)	(0.06, 0.07)	(0.06, 0.07)	(0.09, 0.15)	(**0.14**, **0.28**)
	Class 1	(0.39, **0.67**)	(0.31, 0.46)	**(0.43, 0.67)**	(0.40, **0.67**)	**(0.43, 0.67)**
German Credit	Class 0	(**0.55**, **0.82**)	(0.00, 0.00)	(0.00, 0.00)	(**0.55**, **0.82**)	(0.54, **0.82**)
	Class 1	(0.24, 0.52)	(0.00, 0.00)	(0.00, 0.00)	(0.19, 0.40)	(**0.37**, **0.82**)
Bank Marketing	Class 0	(**0.49**, **0.81**)	(0.16, 0.19)	(0.00, 0.00)	(0.42, **0.81**)	(0.41, **0.81**)
	Class 1	(0.41, **0.70**)	(0.41, **0.70**)	(0.00, 0.00)	(**0.45**, **0.70**)	(0.37, **0.70**)
Heart	Class 0	(0.29, 0.38)	(0.00, 0.00)	(0.00, 0.00)	(0.30, **0.41**)	(**0.31**, 0.37)
	Class 1	(0.13, 0.24)	(0.00, 0.00)	(0.00, 0.00)	(**0.24**, **0.51**)	(**0.24**, **0.51**)

Table 2: Clustering methods of all datasets. In each cell (std, max divergence)

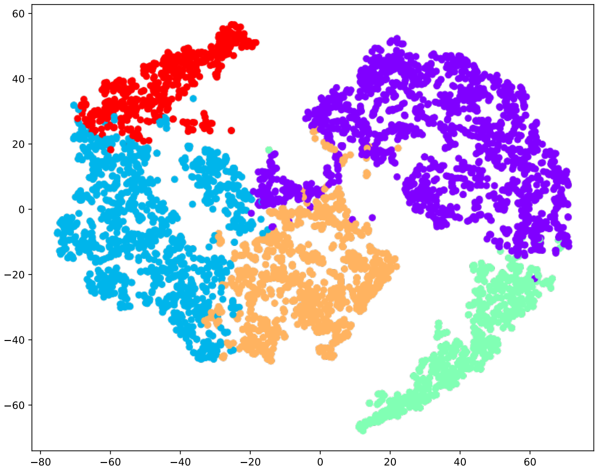

Figure 4: TSNE plot

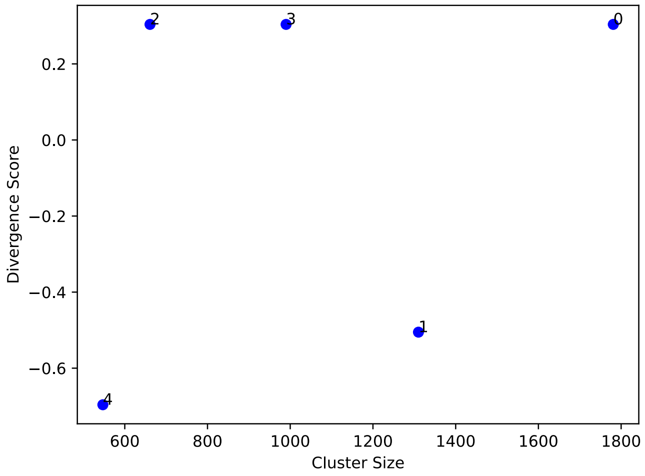

Figure 5: divergence plot

First, I discovered that Hidden Stratification is a real phenomenon that can have a significant impact on clustering results. By carefully examining different clustering spaces and techniques researchers can gain a more accurate understanding of the structure of their data.

Second, I learned that there is no one-size-fits-all approach to clustering analysis. Different datasets may require different clustering techniques, but using the feature space should be considered in finding Hidden Stratification. As I only used data which is already available to researchers, I believe it will not be hard to implement in real research.

On A personal level, this project was challenging for me, because it required combination of real theoretical Knowledge and a lot of effort to program it alone. It was a learning experience, in which I understood new things in its process. Through this project I learned how much a researcher needs to have deep understanding very different concepts, and how much time and effort it takes.

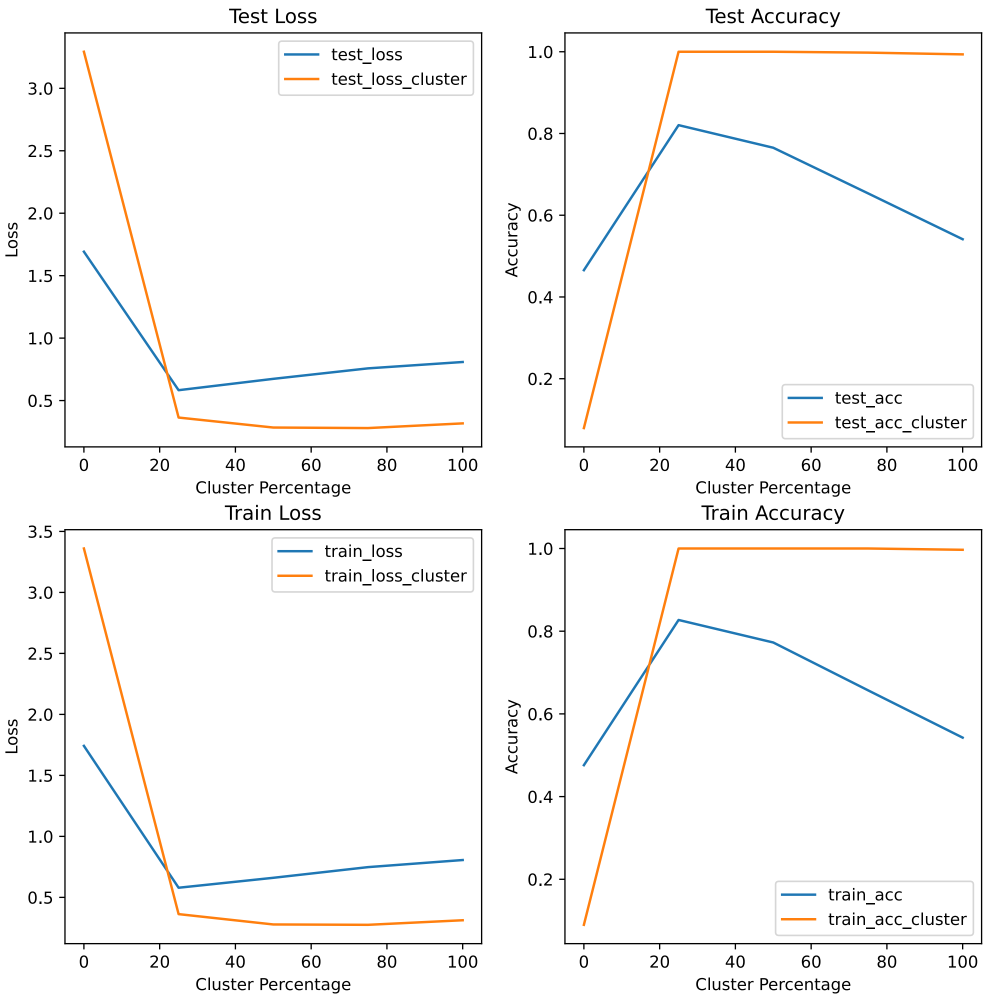

Figure 6: Experiment details: Bank dataset, Feature space, Class #1, Birch clustering method

References

[1] et al Oakden-Rayner, Luke. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proceedings of the ACM conference on health, inference, and learning*, 2020.

[2] Luca de Alfaro Pastor, Eliana and Elena Baralis. Looking for trouble: Analyzing classifier behavior via pattern divergence. *Proceedings of the 2021 International Conference on Management of Data*, 2021.

[3] Tatsunori B. Hashimoto Shiori Sagawa, Pang Wei Koh and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *In International Conference on Learning Representations (ICLR)*, 2020.

[4] et al. Sohoni, Nimit. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems 33*, 2020.