Project Proposal: Supervised Clustering

Eliot Seo Shekhtman; Alan Caldera

October 2021

1 Introduction

This is an Algorithm Development project, revolving around the utility of clustering in supervised machine learning. Specifically, we seek to find an algorithm that would allow us to find optimal clusters that satisfy constraints on their contents in a supervised learning setting. We then seek to evaluate how this method performs in multiple settings, including imbalanced data settings when a major class is vastly overrepresented in a dataset, versus current state of the art methods. In general, our clusters will be bounded by the class labels of the dataset such that clusters designated for one class will have the majority of their probability mass within their class. The results of this project will hopefully extend the field of supervised clustering into a new area or improve upon current methods.

2 Significance

There is high utility for this general class of algorithms. In general, this could lend some interpretability to learning methods, as clusters not predicated entirely by the classes but delineating them nevertheless could show different subgroups relevant to the classes. This could be used by itself possibly as a method for classifying data, or could be paired with a downstream supervised learning model to hopefully improve accuracy and granularity of predictions.

A particularly relevant subfield of machine learning for which this method could be beneficial is imbalanced data. Imbalanced data is an issue in multi-class classification where there's an uneven distribution of data representation among the classes. This can cause many learning algorithms to prioritize outputting certain labels solely based on their prevalence in the training data, resulting in trivial and uninformative models which converge to a local loss minimum rather than accurately learning the true distribution of data. With an upstream clustering algorithm which can partition the dominating classes of a dataset, each cluster could be treated as its own class, allowing classification to progress on a more granular dataset which would both remove the issue of one class having the majority of the datapoints, and possibly create clusters which reveal additional information about the major class which could improve classification as a whole.

3 Related Methods

Imbalanced data, being a common issue in the real world, has several solutions already present and used in practice. The first widely used solution is in balanced errors, where models that optimize an objective function might assign a larger loss towards the misclassification of less common classes, promoting a more effective learning of their characteristics. Two other widely used solutions revolve around the concepts of undersampling and oversampling, whereby data scientists may seek to either undersample by randomly choosing a subset of the major class for classification to reduce its dominance, or oversample to generate new minor class samples using algorithms such as SMOTE.

4 Datasets

For preliminary analysis, datasets will be generated with specific parameters to simulate the features of data we wish to analyze. We will also use real-world datasets with certain augmentations to enhance this concept, such as https://www.kaggle.com/jsphyg/weather-dataset-rattle-package (a dataset for predicting next-day weather in Australia, where 76% of observations report no rain: this will allow us to predict in an imbalanced class scenario).