Analysis of Glass Identification Data

Project Proposal - Surya Suresh (V00998151)

Introduction

Multiclass classification problems are those where a label must be predicted, but there are more than two labels that may be predicted. These are challenging predictive modeling problems because a sufficiently representative number of examples of each class is required for a model to learn the problem. Problems of this type are referred to as imbalanced multiclass classification. The glass identification dataset is a standard dataset for exploring the challenge of imbalanced multiclass classification. The main objective of this project is to fit a regression model for the glass identification dataset.

Dataset Description

The glass identification dataset was obtained from the UCI machine learning repository and credited to Vina Spiehler in 1987. The dataset describes the chemical properties of glass and involves classifying samples of glass using their chemical properties. The dataset consists of 214 instances and 11, all attributes are continuously valued. The main purpose of the classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence, if correctly identified.

Below listed 10 attributes are the input variables that summarize the properties of the glass dataset.

1. Id number – Describes the 1-214 instances
2. RI – Refractive index of the glass
3. Na – Sodium
4. Mg – Magnesium
5. Al – Aluminum
6. Si – Silicon
7. K – Potassium
8. Ca – Calcium
9. Ba – Barium
10. Fe – Iron
11. Type of glass – Class attribute

Among all these variables, Types of glass are the response variable, and the other 10 variables are predictors. Based on their oxide content the predicted variable can be classified into 7 classes

- **Class 1**: Building windows (float processed)
- **Class 2**: Building windows (non-float processed)
- **Class 3**: Vehicle windows (float processed)
- **Class 4**: Vehicle windows (non-float processed)
- **Class 5**: Containers
- **Class 6**: Tableware
- **Class 7**: Headlamps

Here the data is imbalanced, and I observed that the response variable with class 4(non-float processed) has 0 examples. Although there are minority classes, However all classes are equally important in this prediction problem. Here float glass refers to the process used to make glass.