SCC222: AI Concepts

Glass Identification for Criminological Investigation

Introduction

Identifying glasses by their concentration of chemical properties and refractive indices has proved to be a crucial factor that forensic specialists have used to identify or link a crime to an individual. There have recently been rapid developments in the field of forensics lately through the technological advancements in machine learning which aid in criminological investigations. This piqued my curiosity about the accuracy in which certain types of machine learning algorithms, namely KNN, Decision Trees, Naïve Bayes and SVMs can solve the glass dataset problem. For this problem our main aim is to develop each of these algorithms from scratch and identify which algorithm performs best in terms of accuracy (percentage of predictions made which yield to the true glass type) and the speed at which they generate their results.

Data and Preparation

The glass dataset used includes 9 features, RI (refractive index), Na, Mg, Al, Si, K, Ca, Ba and Fe weight percentages of their corresponding oxides and 6 classes, building windows both float and non-float processed, vehicle windows float processed, containers, tableware and headlamps. The samples must be validated through scientific laboratory experiments to ensure that the class assigned for them is correct. Data is then prepared in the following order: 1. All feature values within the dataset are normalized as to avoid dominance by large-scale features. 2. Dataset is divided into k equal stratified folds for more inclusive training and testing and less 'luck' bias. Additionally, a correlation matrix is used to identify strong correlations between features amongst themselves and in comparison, to labels, which helps in getting rid of redundant or weak predictors which introduce unnecessary complexity.

Methodology

When investigating this problem, I faced a few challenges such as deciding on which methods to use for fine tuning and finding a relevant train test split that could be used for all algorithms. For fine tuning I would always start by running an initial correlation matrix on the 9 features then running a grid solution approach which removes strongly correlated features (>|75|) and gives an accuracy for each combination of features used. I would then change the feature set to the combination which yielded the highest accuracy and repeat the process until the accuracy starts decreasing. When experimenting with the number of k folds, a 5-fold cross-validation with 80% training balanced between accuracy and overfitting prevention across all models. Furthermore, when training the algorithms for decision trees I used the ID3 algorithm as it is more suitable for categorical data than CART. For Naïve Bayes I used PDEs as the features were continuous. I used One vs Rest for SVMs because the dataset included more than two classes. For KNN I used the Manhattan distance as it led to the highest accuracy gain.

Results

After fine tuning I discovered that KNN was the most accurate while algorithms like Decision Trees and Naïve Bayes had issues in handling the dataset's feature dependencies. In summary, KNN might have generated the best results due to its ability to accurately find local patterns within the dataset. SVM's lower accuracy suggests that the model had difficulties in separating classes based on their features due to the overlapping feature spaces which may not align with maximum-margin hyperplanes. Maybe an increase in the dimensions of the SVMs could have better separated the classes. The decision tree was expected to not perform well as it overfits on classes which don't appear to often in the dataset. Finally, the Naïve Bayes algorithm assumes feature independence which is not the case for the glass dataset. In terms of time both the KNN and the Naïve Bayes algorithms had the least run time due to their simplistic models.

Conclusion

Based on the results, KNN is clearly the superior model for the glass dataset due to its low computational overhead and adaptability to feature-scaling, which makes it appropriate for forensic applications where precision and accuracy are prioritised. However, KNN depends largely on the choice of K and the distance metric which require careful tuning as the dataset changes.