Labwork 1 for ML in Medicine

Cao Hieu Vinh

February 28, 2025

1 Introduction

We analyze MIT-BIH Arrhythmia Dataset, one of two collections of ECG heartbeat data sets designed to detect the signals correspond to electrocardiogram (ECG) shapes of heartbeats for the normal case and the cases affected by different arrhythmia and myocardial infarction.

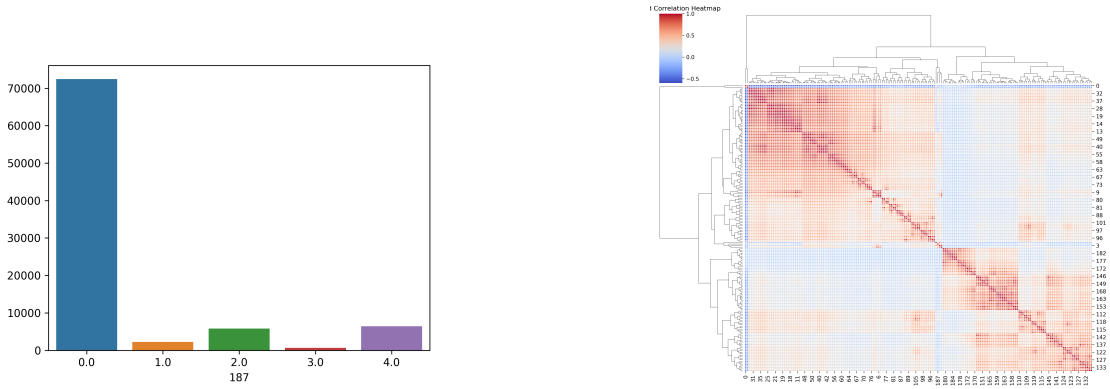
MIT-BIH ground truth dataset involves 5 classes: N (Normal), S (Supraventricular), V (Ventricular), F (Fusion), Q (Unknown). Basically heartbeat will be divided by this 5 classes.

MIT-BIH also has unnamed numerical 187 features each represent an ECG of a heartbeat.

MIT-BIH has 109446 samples,87553 of which has a ground truth, while the rest is meant for testing.

2 MIT-BIH dataset

Though analysis, we found that the MIT-BIH arrhythmia data set is a heavily class imbalance dataset and it also has numerous numerical features that also form a number of pairs of high correlations for each other.



(a) observations count between classes

(b) The correlation between classes

Figure 1: Comparison of class distribution and correlation heatmap

In Figure 1a, the number of normal observations (represented in blue) is significantly higher than the observations of other abnormal classes, highlighting a strong class imbalance in the dataset.

In Figure 1b, the red areas represent highly correlated feature pairs, indicating strong positive correlations. Although these areas are small, the sheer size of the dataset—with its numerous features make even a small portion of high correlation potentially problematic for the model.

3 Preprocess step

3.1 Undersample

We use undersampling technique to handle the class imbalance property in Mit-Bih Dataset. We tried an undersampling value of 0.9, 0.8, 0.7, 0.6 (randomly drop dominant class observation by said

proportions),along with applying PCA, for which the model performance we obtain accuracy of 0.8854, 0.9033 ,0.9497,0.9731 respectively.

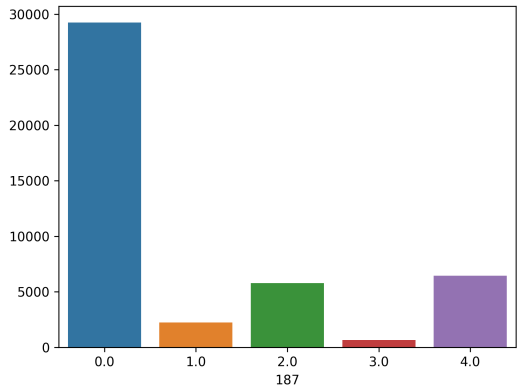


Figure 2: Observations count after undersample

3.2 Principal Component Analysis

Given the large number of features, manually selecting and removing redundant features is not practical. Identifying the most optimal features to drop is difficult. I use PCA as a dimensionality reduction technique to transform the feature space. The PCA output features are almost always uncorrelated, so the model may work fine with it.

4 Modeling

In this lab,i choose to use LightGBM, a reliable and modern general purpose classification ML model using Decision Tree algorithm, to handle the heartbeat.

Parameter tuning of LightGBM:

boosting_type: *gbdt*,
num_leaves: 31,
learning_rate: 0.05,
n_estimators: 200,
max_depth: 5,
min_data_in_leaf: 20,
max_bin: 255,
bagging_fraction: 0.8,
feature_fraction: 0.8.

We also document the output evaluation of model is several other boosting type settings,while we tune learning rate,n estimators and max depth in a for loop each with 3 predefined values,using accuracy as validation metrics.

5 Evaluation

5.1 Model Evaluation

We use basic performance metrics for classification models(Accuracy, Precision, Recall, and F1-score) and ROC curves and AUC scores. Macro-Averaged AUC Score in this model is 0.8960

Class	Precision	Recall	F1-Score	Support
0.0	0.98	0.99	0.99	18118
1.0	0.89	0.63	0.74	556
2.0	0.88	0.93	0.90	1448
3.0	0.78	0.56	0.65	162
4.0	0.98	0.97	0.97	1608
Accuracy	0.97 (Total: 21892)			

Table 1: Classification report showing precision, recall, F1-score, and support for each class.

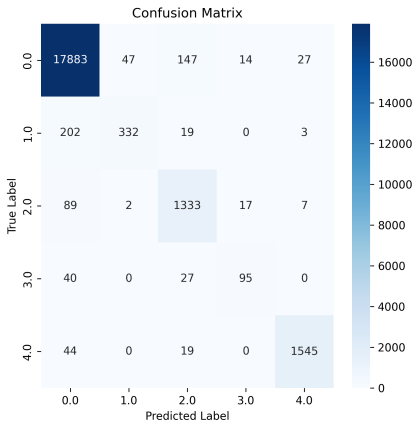


Figure 3: Confusion Matrix

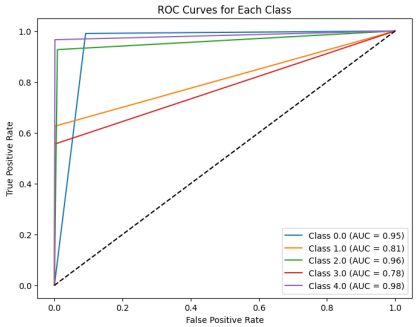


Figure 4: ROC curves

5.2 Other model settings evaluations

Belows is the evaluation of the model when we use rf as boosting type: Macro-Averaged AUC Score of this setting is 0.8601

Class	Precision	Recall	F1-Score	Support
0.0	0.96	0.95	0.96	18118
1.0	0.62	0.56	0.59	556
2.0	0.73	0.70	0.71	1448
3.0	0.27	0.72	0.40	162
4.0	0.87	0.90	0.89	1608
Accuracy	0.92 (Total: 21892)			

Table 2: Classification report showing precision, recall, F1-score, and support for each class in rf setting.

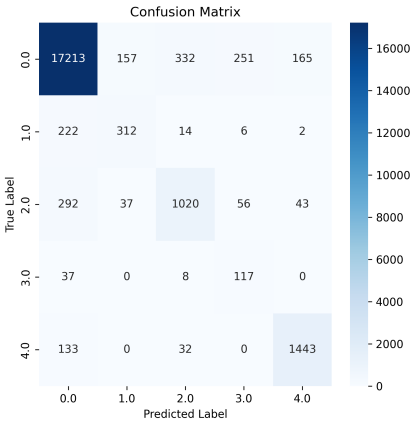


Figure 5: Confusion matrix of model with rf boosting

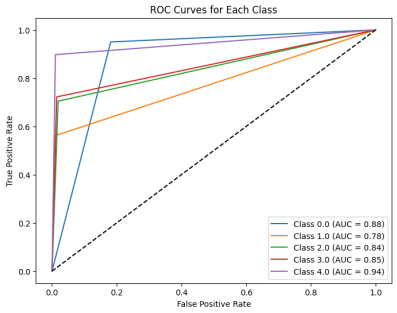


Figure 6: ROC curves of rf boosting type