Machine Learning in Medicine: Pratice 1 (Heartbeat type Classification)

Vinh Thanh Nguyen

I. Introduction

Machine Learning nowadays provide many method that help many field: Finance & Stock, Math, and so on. Especically, Machine Learning play the critical role in medical field, the most common task that Machine Learning does is **Diagnosis decease**. The input will be symptoms, health state, measurement of body (weight, blood pressure, height, etc.). This implementation of Machine Learning save a significant amount of times for doctors and prevent any human faults.

II. Abstract

Heartbeat type test is one the most important stage to measure and health assessment. The result of the test in the most of times will be the class of your heartbeat. In this experiment, the outputs are classes from 1-5. Input is the ECG measurement statistics, from this doctor can easily evaluate your health condition.

III. Dataset Overview

This dataset has been used in exploring heartbeat classification. The signals correspond to electrocardiogram (ECG) shapes of heartbeats for the normal case and the cases affected by different arrhythmias and myocardial infarction. These signals are preprocessed and segmented, with each segment corresponding to a heartbeat.

Content

Arrhythmia Dataset

- **Number of Samples**: 109,446
- **Number of Categories**: 5
- **Sampling Frequency**: 125 Hz
- **Data Source**: PhysioNet's MIT-BIH Arrhythmia Dataset
- **Classes**: `['N': 0, 'S': 1, 'V': 2, 'F': 3, 'Q': 4]`

The PTB Diagnostic ECG Database

- **Number of Samples**: 14,552
- **Number of Categories**: 2
- **Sampling Frequency**: 125 Hz
- **Data Source**: PhysioNet's PTB Diagnostic Database

IV. Visualization

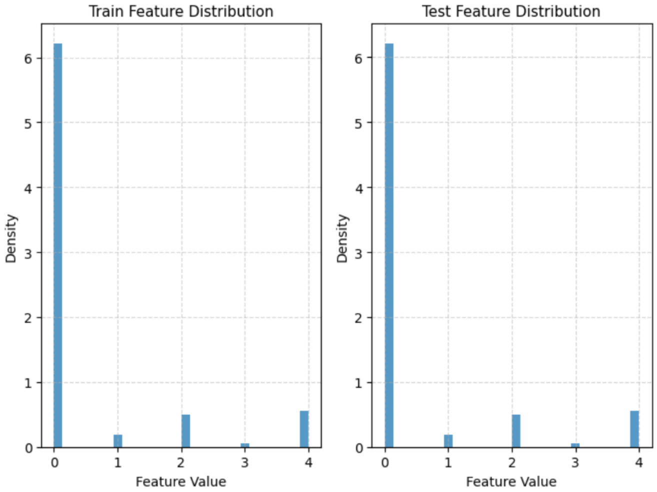

Fig. 1. Class distribution

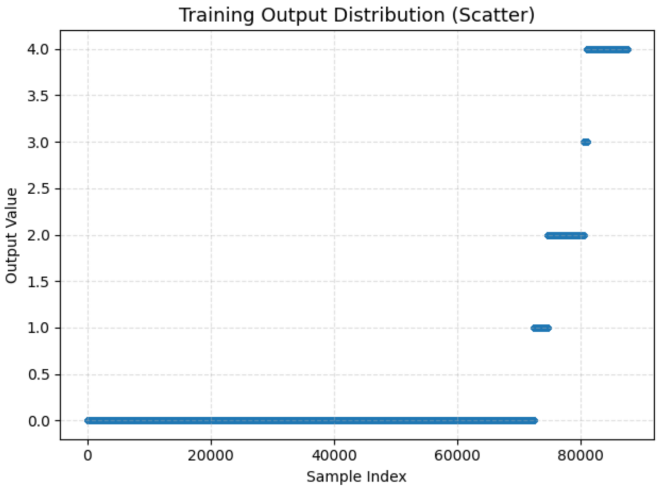

Fig. 2. Output Distribution

The classes outputs distribution are not balance, **Class 0** has dominate appearance, which can cause imbalance data, a common cause lead to poor prediction model performance.

But the advantage of this dataset is every classes have been well clustered which ideal for almost every type of cluster-based models like KNN, SVM, .etc

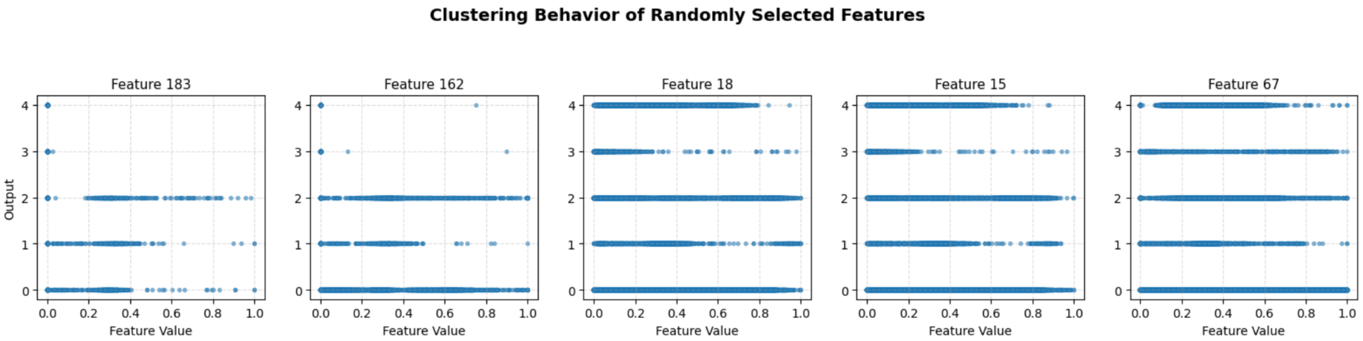

Fig. 3. Random 5 columns sample distribution

V. Experiment: SVM and KNN

In this experiment, we consider two scenarios: *imbalanced-data usage* and *balanced-data usage*,in order to analyze the performance differences and highlight the importance of the data preprocessing stage. In particular, this experiment demonstrates the limitations of directly applying machine learning models to raw, imbalanced data and emphasizes the necessity of preprocessing techniques such as resampling for improving minority-class recognition.

A. SVM: Imbalanced Data

TABLE I
Classification Performance of SVC on Imbalanced Data (cache_size = 500)

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	0.9672	0.9982	0.9825	18,118
1	0.9632	0.5647	0.7120	556
2	0.9667	0.8626	0.9117	1,448
3	0.7500	0.4815	0.5865	162
4	0.9959	0.9111	0.9516	1,608
Accuracy			**0.9680**	21,892
Macro Avg	0.9286	0.7636	0.8289	21,892
Weighted Avg	0.9676	0.9680	0.9657	21,892

B. KNN: Imbalanced Data

TABLE II
Classification Performance of kNN on Imbalanced Data

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	0.9777	0.9946	0.9861	18,118
1	0.8970	0.6421	0.7484	556
2	0.9395	0.9012	0.9200	1,448
3	0.7630	0.6358	0.6936	162
4	0.9941	0.9509	0.9720	1,608
Accuracy			**0.9736**	21,892
Macro Avg	0.9143	0.8249	0.8640	21,892
Weighted Avg	0.9727	0.9736	0.9725	21,892

VI. Summary

We can see that the precision and other metrics result are quite good as expected, class 1 , class 2 and 4 has higher precision , recall , F1 than other class. This happened because the number of samples belonging to those classes is much more than the others, so the model learns and biases for those classes. In an intuitive way, the more data of the class that feed into the model, the better prediction the model perform. Next time we will test with balanced data.

References

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[2] ECG Heartbeat Categorization Dataset, Kaggle, Available: https://www.kaggle.com/datasets/shayanfazeli/heartbeat. Accessed: 2026-01-XX. Dataset composed of segmented/preprocessed ECG signals for heartbeat classification derived from the MIT-BIH Arrhythmia Dataset and PTB Diagnostic ECG Database. :contentReference[oaicite:0]index=0

[3] Scikit-Learn Documentation, "Support Vector Machines," Available: https://scikit-learn.org/stable/modules/svm.html. Accessed: 2026-01-XX. SVM methods are supervised learning models for classification, regression, and outlier detection, with kernel functions defining decision boundaries. :contentReference[oaicite:1]index=1