Machine Learning in Medicine
Labwork 1

Nguyen Tuan Khai

February 2025

1 Introduction

This is the report on the process of modelling ECG Heartbeat Categorization Dataset. For this project, I employed methods such as k-Mean and SMOTE to handle class imbalance, as well as SVM and LSTM for classification purpose.

2 Dataset

In this project, we utilised the data derived from the MIT-BIH Arrhythmia Dataset and the PTB Diagnostic ECG Database. These are two renowned collections of heartbeat signals used for cardiovascular disorders classification. For each sample, there are a total of 187 features, representing the beat value (ranging from 0 to 1) at a certain time line. Since the sampling frequency is 125Hz [1], each observation capture the activity of the heart over a 1,496-second interval.

There exist 5 different categories among over these 115,000 samples, denoted as followed [2]:

- N: normal beat (0)
- A (or S): atrial premature beat (1)
- V: ventricular premature beat (2)
- F: ventricular fusion beat (3)
- VT (or Q): ventricular tachycardia (4)

It is worthwhile to note that this dataset suffer from tremendous class imbalance. As shown in **??**, the training set is disproportionately populated with normal beat N, while fusion beat F only makes up 654 samples.

3 Experiments

To cope with class imbalance, the method used was a combination between oversampling and undersampling. That is for majority classes (N and S), the k-Mean clustering algorithm is applied to reduce the number of samples. Since V and F have roughly 6000 observations, the goal was to balance N and S to a similar amount. Additionally, Synthetic Minority Over-sampling Technique (SMOTE) is applied for fusion beat class.

3.1 Undersampling

With the help of k-Mean algorithm, dominant classes can be properly downsampled [3] [5]. From given data of class 0, I extracted N_num cluster centers, where N_num also equal to the sample size of V and F (i.e 6000). These cluster centers are considered new synthetic data points and are directly used to train the model. To verify that all produced data points belong to the same class, I implemented silhouette score metric, which measures how well the clusters are separated. A score of 1 indicates perfectly separated clusters, while values near 0 shows overlapping clusters. The result for 6000 centroids was 0.15, meaning that these clusters are highly similar, proving the validity of the method.

3.2 Oversampling

The SMOTE algorithm create synthetic data out of existing ones. By locating the k-nearest neighbor of these samples, new data points can be generated [4].

3.3 Modelling

3.4 Results

4 Conclusion

References

[1] Shayan Fazeli,"ECG Heartbeat Categorization Dataset" , 2018

[2] George B. Moody and Roger G. Mark,"The Impact of the MIT-BIH Arrhythmia Database History, Lessons Learned, and Its Influence on Current and Future Databases",May/June 2001

[3] Sergey Feldman,"How could one build models for skewed classes?", Quora, 2018

[4] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer,"SMOTE: Synthetic Minority Over-sampling Technique", 2002

[5] Chih-Ming Huang et al ,"A K-means Clustering Based Under-Sampling Method for Imbalanced Dataset Classification "