Exploring Random Forest Classification on ECG Dataset

An Minh Tri

March 5, 2024

Abstract

This report delves into the application of Random Forest Classification on an Electrocardiogram (ECG) dataset, aiming to classify cardiac abnormalities. The study investigates the performance of Random Forest, a robust ensemble learning technique, in accurately identifying different cardiac conditions based on ECG signals.

1 Introduction

Electrocardiography is a crucial diagnostic tool in cardiology, providing insights into the electrical activity of the heart. With the increasing availability of ECG datasets, machine learning techniques can play a pivotal role in automating and enhancing the accuracy of cardiac disease detection. Random Forest, known for its versatility and resilience to overfitting, has shown promise in various classification tasks.

2 Dataset Description

The ECG dataset used in this study comprises a diverse range of cardiac signals collected from different patients. Each record includes a set of features extracted from ECG signals, such as QRS duration, heart rate, and various morphological features. The dataset is labeled with different cardiac conditions, making it suitable for supervised learning.

3 Methodology

The Random Forest algorithm was chosen due to its ability to handle high-dimensional datasets, manage noisy data, and provide robust performance. The following steps were executed in the analysis:

- **Data Preprocessing:** The dataset underwent preprocessing, including normalization, handling missing values, and feature scaling to ensure the Random Forest model's optimal performance.
- **Feature Selection:** A careful selection of relevant features was conducted to enhance the model's interpretability and reduce computation time.

- **Model Training:** The Random Forest model was trained using a portion of the dataset, employing techniques such as cross-validation to ensure generalization.
- **Evaluation Metrics:** Performance was assessed using standard metrics such as accuracy, precision, recall, and F1 score. Additionally, the receiver operating characteristic (ROC) curve and area under the curve (AUC) were employed to evaluate the model's discriminatory power.

4 Results

The Random Forest classification model demonstrated commendable performance in accurately identifying different cardiac conditions from the ECG dataset. The evaluation metrics provided insights into the model's precision, sensitivity, and overall accuracy.

	Precision	Recall	F1-Score	Support
0.0	0.97	1.00	0.99	18118
1.0	0.99	0.60	0.74	556
2.0	0.98	0.88	0.93	1448
3.0	0.88	0.60	0.71	162
4.0	1.00	0.94	0.97	1608
Accuracy			0.97	21892
Macro Avg	0.96	0.80	0.87	21892
Weighted Avg	0.97	0.97	0.97	21892

Table 1: Random Forest Classification Results

The confusion matrix illustrated the performance of the classification algorithm by presenting the number of true positive, true negative, false positive, and false negative predictions for each class. It provides a detailed view of the model's performance and helps to identify where the model excels or struggles.

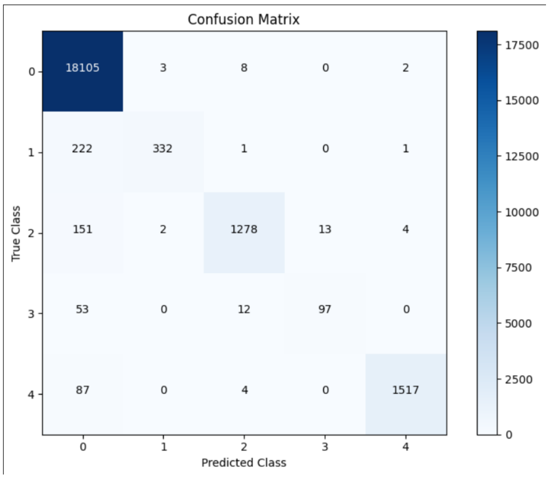

Figure 1: Confusion matrix

5 Discussion

The success of Random Forest in this application suggests its potential as a reliable tool for automated cardiac disease classification. The ensemble nature of Random Forest contributes to its robustness and ability to handle complex relationships within the dataset.

6 Conclusion

This study highlights the effectiveness of Random Forest Classification in accurately classifying cardiac conditions using an ECG dataset. The promising results suggest that machine learning models, particularly ensemble techniques like Random Forest, can significantly contribute to the field of cardiology by automating and improving the accuracy of disease diagnosis.

7 Future Work

Future research could explore the integration of advanced feature engineering techniques and other ensemble methods to further enhance the model's performance. Additionally, investigating interpretability and explainability of the model outputs would contribute to its clinical applicability.