Análisis y Clasificación de Señales ECG utilizando Random Forest

Macarena Miño

Universidad de Talca

Curso: Machine Learning

Fecha: 30 de junio de 2025

1. Introducción

El objetivo de este proyecto es analizar y clasificar señales de ECG (Electrocardiograma) en dos clases distintas: 'N' (Normal) y 'A' (Anormal). El conjunto de datos utilizado para esta tarea contiene múltiples registros de señales ECG, acompañados de sus respectivas etiquetas de referencia que indican si la señal es normal o anormal. A continuación, se detallan el preprocesamiento de los datos, la extracción de características, el análisis exploratorio de los datos y la clasificación utilizando el algoritmo Random Forest.

2. Preprocesamiento de Datos y Extracción de Características

El conjunto de datos incluye archivos .mat que contienen señales ECG en crudo y archivos .hea con metadatos como la frecuencia de muestreo. El primer paso consistió en cargar los datos de las señales ECG y los metadatos correspondientes, seguido de la extracción de características relevantes para describir las características de las señales.

Las características extraídas fueron:

Media (media_mv): Valor promedio de la señal.

Desviación Estándar (mstd_mv): Variabilidad en la amplitud de la señal.

Asimetría (skewness): Medida de la asimetría de la distribución de la señal.

Curtosis (kurtosis): Medida de la "altitud" de la distribución de la señal.

Media de Intervalos RR (rr_mean_s): Tiempo promedio entre los picos sucesivos de la señal ECG, crucial para evaluar la variabilidad de la frecuencia cardíaca.

Desviación Estándar de Intervalos RR (rr_std_s): Medida de la variabilidad en los intervalos RR.

3. Análisis Exploratorio de Datos

El análisis exploratorio se centró en la distribución de las características extraídas entre las dos clases (Normal 'N' y Anormal 'A'). Se utilizaron gráficos de histograma para visualizar la distribución de características clave como la media de la señal, la desviación estándar, la asimetría, la curtosis y los intervalos RR. El análisis reveló que:

La clase 'N' tiende a mostrar menos variabilidad y menor asimetría.

La clase 'A' presenta una mayor dispersión y curtosis, lo que sugiere que las señales anormales pueden tener irregularidades más pronunciadas en su forma de onda.

4. Limpieza de Datos

El conjunto de datos pasó por un proceso de limpieza donde se eliminaron las filas con valores nulos o infinitos. Después de la limpieza, el conjunto de datos se dividió en dos subgrupos: uno para cada clase ('N' y 'A'). Se seleccionó un subconjunto equilibrado de 738 muestras por clase para entrenar el modelo, asegurando que el modelo no fuera sesgado hacia una clase.

5. Clasificación utilizando Random Forest

Se utilizó un clasificador Random Forest para clasificar las señales ECG según las características extraídas. El clasificador se entrenó con el 80% del conjunto de datos y su rendimiento se evaluó con el 20% restante.

Pasos clave en el proceso de clasificación:

División de Datos: El conjunto de datos se dividió en conjuntos de entrenamiento (80%) y prueba (20%), asegurando una muestra estratificada para mantener la distribución de las clases.

Entrenamiento del Modelo: Se entrenó un modelo Random Forest con 250 árboles utilizando las características: 'media_mv', 'mstd_mv', 'skewness', 'kurtosis', 'rr_mean_s', y 'rr_std_s'.

Evaluación: El rendimiento del modelo se evaluó utilizando matrices de confusión y reportes de clasificación. La matriz de confusión mostró el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, mientras que el reporte de clasificación proporcionó precisión, recall y puntuaciones F1 para cada clase.

6. Resultados

Los resultados de la clasificación mostraron un rendimiento prometedor, con el modelo Random Forest logrando una alta precisión en la clasificación de las señales ECG. La matriz de confusión indicó que el modelo fue efectivo para distinguir entre señales normales y anormales. Además, el análisis de la importancia de las características mostró que la 'media de la señal' y la 'desviación estándar' fueron las características más importantes para la clasificación.

7. Visualizaciones

La importancia de las diversas características se visualizó mediante gráficos de barras, lo que ayudó a comprender qué características fueron más influyentes a la hora de distinguir entre las dos clases. Estas visualizaciones son valiosas para mejorar la interpretabilidad del modelo.

8. Discusión y Mejoras

El rendimiento del modelo podría mejorarse mediante:

Aumento del Tamaño del Conjunto de Datos: Un conjunto de datos más grande proporcionaría más puntos de datos para que el modelo aprenda, mejorando potencialmente la capacidad de generalización.

Optimización de Parámetros de Random Forest: Ajustar el número de árboles en el Random Forest podría ayudar a obtener un clasificador más robusto.

Propuesta de mejora: Características adicionales, como medidas de variabilidad de la frecuencia cardíaca o características en el dominio del tiempo, podrían mejorar el rendimiento del modelo.

9. Conclusión

Este estudio demuestra el potencial de utilizar aprendizaje automático, específicamente Random Forest, para la clasificación de señales ECG. El proceso de preprocesamiento, extracción de características y clasificación proporciona una base sólida para mejoras futuras, incluyendo la expansión del conjunto de datos y la optimización de parámetros. Estos modelos pueden ser valiosos para la interpretación automatizada de ECG, ayudando en el diagnóstico de anomalías cardíacas.