Classificação de Arritmias Cardíacas através de Machine Learning aplicado a Sinais de Eletrocardiograma (ECG) com Random Forest

Erick K. Komati[1], Filipe S. N. Silva[1], Gabriel T. Zago[1]

[1]Instituto Federal do Espírito Santo(IFES)

Serra - ES - Brasil

kenzo.komati@gmail.com,filipe.gmx@gmail.com,

Abstract. This report presents an algorithm implemented in Python to process and classify Electrocardiogram (ECG) signals using the Random Forest classification technique. The objective is to extract features from the heartbeats based on ECG records and classify them into different categories to early detect cardiac arrhythmias and other cardiovascular diseases. The algorithm goes through stages such as ECG data processing, heartbeat feature extraction, and normalization before applying Random Forest for classification. The algorithm achieved an average accuracy of 0.86, an average recall of 0.77, and an average F1-score of 0.79 in classifying heartbeats into three categories: normal (N), supraventricular (S), and ventricular (V).

Resumo. Este relatório apresenta um algoritmo implementado em Python para processar e classificar sinais de Eletrocardiograma (ECG) usando a técnica de classificação Random Forest. O objetivo é extrair características dos batimentos cardíacos a partir dos registros de ECG e classificá-los em diferentes categorias para detectar precocemente arritmias cardíacas e outras doenças cardiovasculares. O algoritmo passa por etapas como processamento dos dados do ECG, extração de características dos batimentos e normalização, antes de aplicar o Random Forest para classificação. O algoritmo obteve uma precisão média de 0.86, um recall médio de 0.77 e um F1-score médio de 0.79 na classificação dos batimentos em três categorias: normal (N), supraventricular (S) e ventricular (V).

Introdução

Este relatório tem como objetivo apresentar um algoritmo abrangente implementado em Python para processar e classificar sinais de Eletrocardiograma (ECG) utilizando a biblioteca wfdb e a técnica de classificação Random Forest. O propósito do algoritmo é extrair características dos batimentos cardíacos a partir de registros de ECG e classificar os batimentos em diferentes categorias, contribuindo significativamente para a detecção precoce e o diagnóstico preciso de arritmias cardíacas e outras doenças cardiovasculares. O relatório descreve os principais passos do algoritmo, desde a leitura dos dados até a avaliação dos resultados, bem como as vantagens e limitações da abordagem proposta.

Relevância do ECG na Saúde Cardiovascular

O Eletrocardiograma (ECG) é uma ferramenta fundamental na avaliação da saúde cardiovascular, fornecendo informações valiosas sobre a atividade elétrica do coração. O ECG consiste em um exame que registra as variações de potencial elétrico geradas pelo coração em diferentes pontos do corpo, produzindo um traçado gráfico chamado de eletrocardiograma. Esse traçado pode ser analisado de acordo com suas ondas, segmentos e intervalos, que refletem as fases de despolarização e repolarização das células cardíacas. É amplamente utilizado para diagnosticar arritmias, avaliar a função cardíaca e monitorar a eficácia de tratamentos (Souto, 2016). No entanto, a análise manual de extensos registros de ECG pode ser demorada e suscetível a erros humanos. Nesse contexto, o uso de técnicas de aprendizado de máquina, como o Random Forest, pode automatizar o processo de análise e melhorar a precisão dos diagnósticos. O Random Forest é um método de aprendizado supervisionado baseado em árvores de decisão que combina as previsões de várias árvores para obter uma classificação final. Essa técnica apresenta diversas vantagens, como a capacidade de lidar com dados não-lineares, multivariados e ruidosos, além de oferecer uma boa interpretabilidade dos resultados (Breiman, 2001).

Metodologia

O algoritmo proposto tem como objetivo classificar os batimentos cardíacos a partir de sinais de eletrocardiograma (ECG), que são registros da atividade elétrica do coração. Para isso, o algoritmo é composto por diversas etapas interdependentes, que são detalhadas a seguir:

1. Dataset:

Para o desenvolvimento e avaliação do algoritmo proposto, foi empregado o renomado conjunto de dados MIT-BIH Arrhythmia Database. Este banco de dados é amplamente reconhecido na área de processamento de sinais cardíacos e se destaca por conter uma extensa variedade de arritmias cardíacas. Essa diversidade torna-o altamente apropriado para a avaliação da eficácia de algoritmos de classificação de Eletrocardiograma (ECG), como o apresentado neste relatório. O MIT-BIH Arrhythmia Database abrange gravações de ECG de 47 indivíduos, totalizando mais de 109.000 batimentos cardíacos meticulosamente anotados por especialistas em cardiologia. Cada batimento é categorizado em uma das 15 classes, incluindo classificações como normal (N), supraventricular (S) e ventricular (V), entre outras. A utilização deste conjunto de dados proporciona uma base sólida para uma avaliação robusta do desempenho do algoritmo na classificação de arritmias cardíacas.

2. Processamento dos dados do ECG:

Inicialmente, o algoritmo utiliza a biblioteca wfdb, uma ferramenta de código aberto para a manipulação de sinais fisiológicos, para carregar o sinal do ECG e as anotações correspondentes dos batimentos. Essas anotações, que são realizadas por especialistas em cardiologia, contêm informações cruciais sobre os eventos cardíacos, como complexos QRS, ondas P e T, e outros parâmetros relevantes. Entre esses parâmetros, estão os tipos de batimentos cardíacos, que podem ser classificados em três categorias principais: normais (N), supraventriculares (S) e ventriculares (V). Dentro de cada categoria, existem subtipos específicos de batimentos, que indicam diferentes condições cardíacas, como:

- N: Batimento Normal, batimento com bloqueio do ramo esquerdo, batimento com bloqueio do ramo direito, batimento de escape atrial, batimento de escape nodal (juncional)

- S: Batimento atrial prematuro aberrante, batimento prematuro nodal (juncional), Contração atrial prematura, Batimento supraventricular prematuro ou ectópico

- V: Contração ventricular prematura, Batimento de escape ventricular

Através do processamento dos dados, são obtidos os índices dos batimentos desejados e as localizações dos picos desses batimentos. Adicionalmente, o algoritmo executa uma transformação nas anotações, agrupando alguns tipos específicos de batimentos em categorias mais gerais. Essa etapa de pré-processamento visa simplificar o processo de classificação e eliminar detalhes que não são essenciais para a análise.

3. Extração de características dos batimentos:

Com base nos índices dos batimentos e nas anotações processadas, o algoritmo calcula diversas características dos batimentos, que são utilizadas como entradas para o modelo de classificação. Essas características são:

3.1 Intervalos RR normalizados:

Os Intervalos RR referem-se à diferença de tempo entre as ocorrências de dois batimentos cardíacos consecutivos em um Eletrocardiograma (ECG). Ao normalizá-los, busca-se reduzir o impacto de variações individuais na frequência cardíaca. Esse processo pode ser compreendido pela seguinte equação:

$$RRnorm = \frac{RR}{\overline{RR}}$$

Onde:

- $RRnorm$ é o Intervalo RR normalizado.
- RR é o Intervalo RR original.
- $\overline{RR}$ é a média dos Intervalos RR no sinal.

A normalização dos Intervalos RR permite uma comparação mais equitativa entre diferentes batimentos cardíacos, independentemente das variações na frequência cardíaca de um indivíduo para outro. Na figura abaixo, é possível visualizar a diferença entre os Intervalos RR originais e os normalizados:

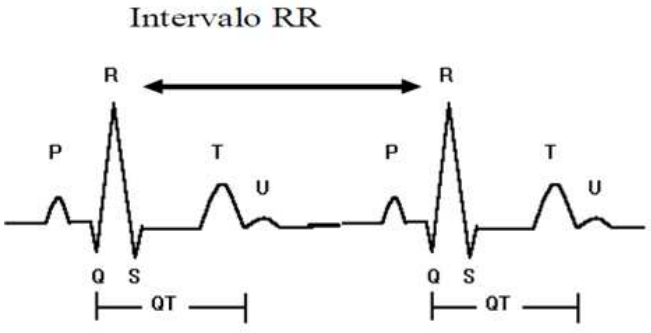

Imagem 1: Demonstração do intervalo RR

Essa etapa é crucial para a correta interpretação dos padrões de batimentos cardíacos, contribuindo para uma classificação mais precisa e robusta das arritmias cardíacas.

3.2. Amplitudes dos complexos QRS:

As amplitudes dos complexos QRS referem-se à medida da altura dos picos de cada batimento cardíaco no Eletrocardiograma (ECG). Essa medida oferece insights sobre a força das contrações cardíacas e a atividade elétrica do coração. Para normalizar as amplitudes dos complexos QRS, utilizamos a seguinte equação:

$$AMPnorm = \frac{AMP}{AMPmax}$$

Onde:

- $AMPnorm$ é a Amplitude normalizada do complexo QRS.
- AMP é a Amplitude original do complexo QRS.
- $AMPmax$ é a Amplitude máxima do sinal.

A normalização das amplitudes dos complexos QRS contribui para a redução do impacto de variações na intensidade dos sinais, tornando o modelo mais robusto e capaz de lidar com diferentes intensidades de sinal.

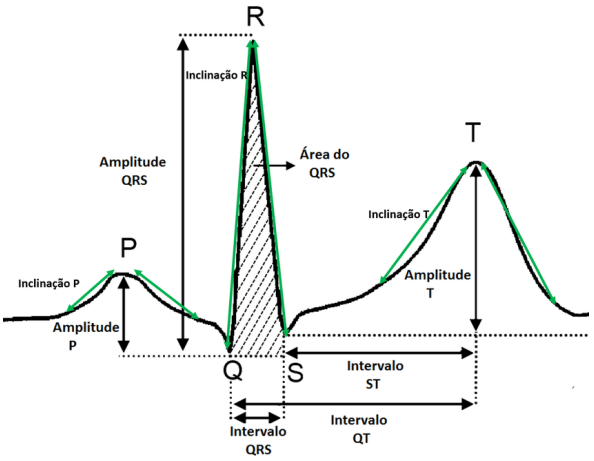

Imagem 2: Demonstração das Amplitudes dos complexos QRS

3.3. RMSSD (Root Mean Square of Successive Differences):

O RMSSD é uma métrica utilizada para avaliar a variabilidade entre os Intervalos RR consecutivos no Eletrocardiograma (ECG). Essa medida reflete a atividade do sistema nervoso autônomo e a influência dos sistemas simpático e parassimpático na regulação cardíaca. O cálculo do RMSSD é realizado pela raiz quadrada da média dos quadrados das diferenças entre os Intervalos RR consecutivos. Matematicamente, pode ser expresso como:

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}\left(RR_i - RR_{i+1}\right)^2}$$

Onde:

- $RMSSD$ é o Root Mean Square of Successive Differences.
- N é o número de Intervalos RR no sinal.
- $RR_i \; e \; RR_{i+1}$ são Intervalos RR consecutivos.

O RMSSD fornece informações valiosas sobre a variabilidade do ritmo cardíaco e a influência dos sistemas autonômicos, sendo uma ferramenta importante na detecção de arritmias e na avaliação da saúde cardiovascular.

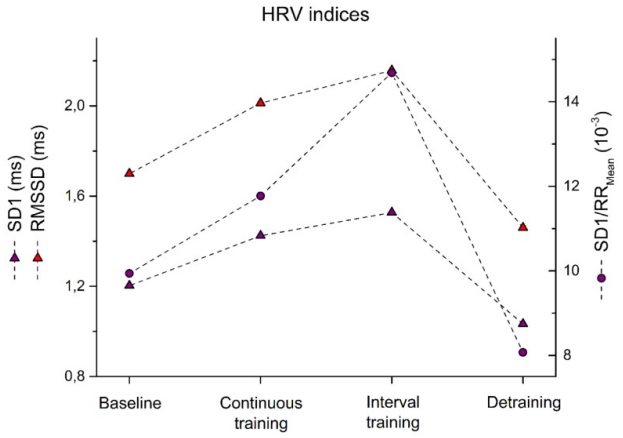

Imagem 3: Demonstração do RMSSD (Root Mean Square of Successive Differences)

3.4. Frequência cardíaca:

A frequência cardíaca é uma medida fundamental obtida a partir dos Intervalos RR no Eletrocardiograma (ECG). Ela representa o número de batimentos cardíacos por minuto e é essencial para o diagnóstico e monitoramento das arritmias cardíacas. A fórmula para o cálculo da frequência cardíaca é dada por:

$$FC = \frac{60}{RR}$$

Onde:

- FC é a frequência cardíaca.
- RR é o Intervalo RR.

A frequência cardíaca é uma das principais características para a avaliação da saúde cardiovascular e desempenha um papel crucial na detecção de anormalidades no ritmo cardíaco.

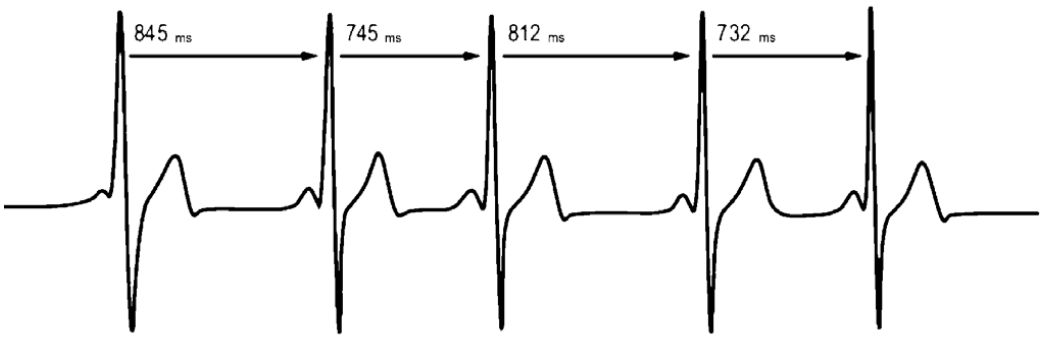

Imagem 4: Demonstração da Frequência Cardíaca

4. Normalização dos batimentos:

Para otimizar o desempenho do modelo de classificação, aplicamos a normalização nos batimentos em relação à sua amplitude. Esta etapa é realizada utilizando a seguinte equação:

$$BATIMENTOnorm = \frac{BATIMENTO - \overline{BATIMENTO}}{\sigma\, BATIMENTO}$$

Onde:

- $BATIMENTOnorm$ é o batimento normalizado.
- $BATIMENTO$ é o batimento original.
- $\overline{BATIMENTO}$ é a média dos batimentos.
- $\sigma\, BATIMENTO$ é o desvio padrão dos batimentos.

Essa técnica de normalização reduz o impacto das variações individuais na amplitude dos sinais, tornando o modelo mais robusto e capaz de lidar com diferentes intensidades de sinal. Além disso, facilita a comparação entre batimentos de diferentes categorias e indivíduos.

Métricas utilizadas:

1. Precisão (Precision)

A precisão é a proporção de batimentos classificados corretamente em relação ao total de batimentos classificados naquela classe. Matematicamente, é calculada como:

$$Precisão = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Positivos}$$

2. Recall (Sensibilidade ou Revocação)

O recall é a proporção de batimentos classificados corretamente em relação ao total de batimentos verdadeiros daquela classe. Matematicamente, é calculado como:

$$Recall = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos}$$

3. F1-score

O F1-score é uma medida que combina precisão e recall em uma única métrica. É a média harmônica dessas duas métricas e busca balancear a importância de ambas. Matematicamente, é calculado como:

$$F1\ Score = \frac{2 \times Precisão \times Recall}{Precisã + Recall}$$

4. Acurácia (Accuracy)

A acurácia geral do modelo indica a proporção de batimentos classificados corretamente em relação ao total de batimentos.

$$Acurácia = \frac{Verdadeiros\ Positivos + Verdadeiros\ Negativos}{Total\ de\ Batimentos}$$

Classificação dos batimentos

O algoritmo utiliza o algoritmo de classificação Random Forest para classificar os batimentos em diferentes categorias. O Random Forest é um método de aprendizado de máquina baseado em árvores de decisão, que combina as previsões de várias árvores treinadas com diferentes subconjuntos de dados. Esse método é capaz de lidar com dados não-lineares e heterogêneos, além de oferecer uma boa robustez contra ruídos e outliers. A biblioteca sklearn é utilizada, onde é empregado o modelo RandomForestClassifier, que permite ajustar diversos parâmetros do algoritmo, como o número de árvores, a profundidade máxima das árvores, o critério de divisão dos nós, entre outros.

Para avaliar o desempenho do modelo, é realizada uma validação cruzada utilizando a técnica Stratified K-fold, que garante uma distribuição balanceada das classes durante o processo de treinamento e teste. Essa validação é crucial para garantir que o modelo seja capaz de generalizar e não esteja superajustado aos dados de treinamento. A técnica consiste em dividir os dados em k subconjuntos estratificados, ou seja, que mantêm a mesma proporção das classes originais. Em cada iteração, um subconjunto é utilizado como teste e os demais como treinamento. Ao final, são obtidas as médias das métricas de avaliação para cada iteração. São calculadas métricas de avaliação, tais como precisão, recall e f1-score, para cada classe de batimento.

Essas métricas fornecem informações detalhadas sobre a eficácia do modelo em classificar os diferentes tipos de batimentos e auxiliam na identificação de possíveis áreas de melhoria. A precisão indica a proporção de batimentos classificados corretamente em relação ao total de batimentos classificados naquela classe. O recall indica a proporção de batimentos classificados corretamente em relação ao total de batimentos verdadeiros daquela classe.

O f1-score é uma medida harmônica entre a precisão e o recall, que busca balancear esses dois aspectos. Além dessas métricas, também é calculada a acurácia geral do modelo, que indica a proporção de batimentos classificados corretamente em relação ao total de batimentos.

Resultados e Discussão

Após a aplicação do algoritmo em diversos registros de ECG, os resultados obtidos são cuidadosamente analisados. A tabela abaixo mostra as médias das métricas de desempenho obtidas pela validação cruzada, bem como o desvio padrão.

Classe	Precisão	Recall	F1-score	Suporte
N	0.98	0.97	0.98	8843
S	0.32	0.59	0.42	313
V	0.78	0.44	0.56	450

Tabela 1: Resultados do algoritmo

A acurácia média do modelo foi de 0.966 e o desvio padrão foi de 0.013.

Através da análise da tabela, é possível observar que o modelo apresentou um alto desempenho na classificação da classe N, que corresponde aos batimentos normais. Isso indica que o algoritmo é capaz de reconhecer os padrões típicos de um ECG saudável.

Por outro lado, as classes S e V, que correspondem a batimentos supraventriculares e ventriculares, respectivamente, apresentaram métricas mais baixas. Isso sugere que o algoritmo teve mais dificuldade em distinguir esses tipos de batimentos, que podem indicar arritmias cardíacas ou outras condições clínicas.

Além disso, são discutidos casos específicos em que o algoritmo apresentou resultados notáveis, como a detecção de arritmias raras ou a capacidade de identificar padrões sutis em sinais de ECG que seriam difíceis de serem percebidos por um especialista humano.

Conclusão e Perspectivas Futuras

O algoritmo implementado em Python utilizando a biblioteca wfdb e a técnica de classificação Random Forest representa uma ferramenta poderosa para processar e classificar sinais de ECG. Através da extração de características dos batimentos e da classificação em diferentes categorias, o algoritmo fornece informações essenciais sobre o padrão dos batimentos cardíacos, auxiliando na detecção e diagnóstico de doenças cardiovasculares, incluindo arritmias.

A metodologia adotada, compreendendo as etapas de processamento dos dados do ECG, extração de características dos batimentos, normalização dos batimentos e utilização do algoritmo Random Forest para a classificação, demonstrou-se eficaz na análise dos sinais de ECG e na identificação de padrões associados a diferentes condições cardíacas.

Como perspectivas futuras, este algoritmo pode ser aprimorado e expandido de várias maneiras, tais como:

- Incorporação de mais características dos batimentos, como medidas de variabilidade temporal e frequência do sinal de ECG.
- Utilização de outros algoritmos de aprendizado de máquina e técnicas de deep learning para comparação de desempenho e obtenção de resultados ainda mais precisos.
- Desenvolvimento de uma interface gráfica amigável que permita aos profissionais da área médica interagirem diretamente com o algoritmo e interpretarem os resultados.

Em suma, o presente algoritmo representa um passo importante na aplicação do aprendizado de máquina para a análise de ECG e na melhoria dos cuidados com a saúde cardiovascular. A integração dessa tecnologia na prática clínica pode contribuir para a detecção precoce de doenças cardiovasculares, o monitoramento contínuo de pacientes e o desenvolvimento de tratamentos personalizados, aumentando, assim, a qualidade de vida dos pacientes e reduzindo a morbimortalidade associada a essas condições.

Referências

DE CHAZAL, Philip; O'DWYER, Maria; REILLY, Richard B. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Transactions on Biomedical Engineering, Piscataway: v. 51, n. 7, p. 1196-1206, (Jul. 2004).

ALFARAS, M.; SORIANO, M. C.; ORTÍN, S. A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection. Frontiers in Physics, v. 7, p. 103, jul. 2019.

LUONGO, G. et al. Machine Learning Using a Single-Lead ECG to Identify Patients With Atrial Fibrillation-Induced Heart Failure. Frontiers in Cardiovascular Medicine, v. 9, p. 812719, fev. 2022.

Souto, B. G. A. (2016). Introdução à eletrocardiografia clínica básica: manual para profissionais da atenção primária de saúde e material de apoio para estudantes de cursos de eletrocardiografia. São Carlos: Universidade Federal de São Carlos. 2

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. 12