

Classificação de Fibrilação Atrial a partir de um registro curto de ECG de derivação única

A.B.S.Telles¹, L.B.G.Trindade², P.H.A.Gomes¹, T.M.S.Barbosa² e W.P.Santos²

¹Departamento de Engenharia Eletrônica, ² Departamento de Engenharia Biomédica (DEBM)
Universidade Federal de Pernambuco, Recife, Brasil

Resumo— O presente estudo visa analisar o dataset disponível no *physionet* chamado “AF Classification from a Short Single Lead ECG Recording: The PhysioNet/Computing in Cardiology Challenge 2017”[1]. Por meio da análise do conteúdo presente no dataset buscou-se fazer uma filtragem do sinal para que fossem recolhidos seus atributos utilizando a ferramenta Colaboratory do Google e então, usou-se o pacote de software Weka para obter resultados referentes aos testes com o algoritmo Random Forest. E assim, saber qual dos testes identifica melhor se o registro mostra ritmo sinusal normal, fibrilação atrial (FA), um ritmo alternativo, ou se possui muito ruído para ser classificado.

Palavras-chave— ECG, Fibrilação Atrial, Mineração de Dados, Algoritmo.

I. INTRODUÇÃO

O ECG é uma reprodução gráfica do comportamento elétrico do coração durante seu funcionamento, este comportamento elétrico é a diferença de potencial ao qual é obtido pelo posicionamento de dois eletrodos em diferentes partes do corpo.

Além disso, através deste estudo é possível classificar se o registro mostra ritmo sinusal normal, fibrilação atrial, ritmos alternados ou ritmos ruidosos como mostrado respectivamente obtidos através do eletrocardiograma. A fibrilação atrial definida como “taquiarritmia” é a arritmia mais comum na população geral e está associada a uma quantidade significativa de morte. Devido a isso, faz-se necessária uma análise de sinais de forma minuciosa que auxilie no diagnóstico desta doença.

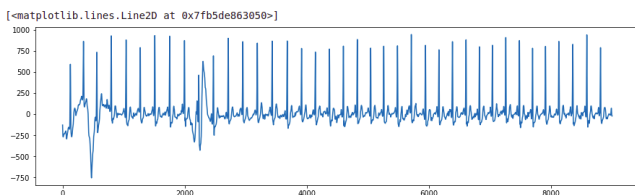


Figura 1. Plot normal

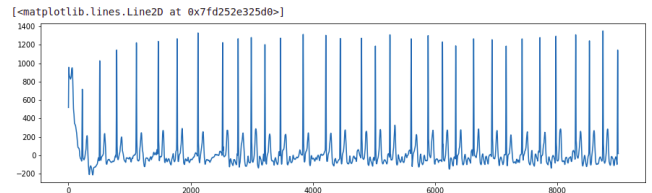


Figura 2. Plot fibrilação atrial

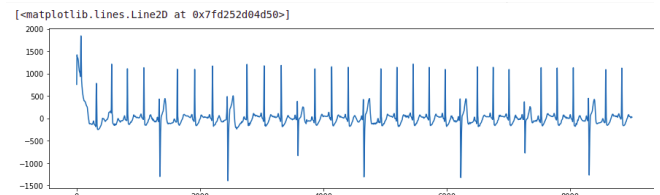


Figura 3. Plot ritmo alternativo

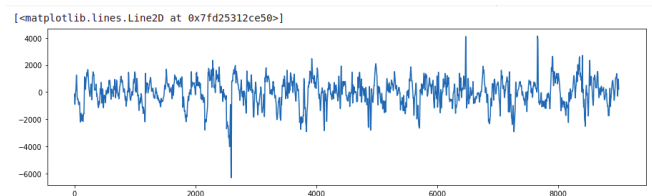


Figura 4. Plot com muito ruído

II. ANÁLISE DA BASE DE DADOS

A base de dados utilizada no projeto foi disponibilizada para que se resolva um desafio, desafio este, que visa incentivar o desenvolvimento de algoritmos para classificar, a partir de um único registro curto de derivação de ECG (entre 30 s e 60 s de duração), se o registro mostra ritmo sinusal normal, fibrilação atrial (FA), uma alternativa ritmo, ou é muito barulhento para ser classificado.

Esta base é composta por 91.506 instâncias, onde 53.832 são de ritmo sinusal normal, 8.087 de fibrilação atrial, 27.392 de ritmos alternativos e 2.195 possuem muito ruído para serem classificados.

III. METODOLOGIA

Esta análise caracteriza-se por ser uma pesquisa exploratória de acordo com seu objetivo, com abordagem além de quantitativa uma vez que métricas são assumidas para avaliar os modelos obtidos, também qualitativa, visto que houve o desenvolvimento de código em python se utilizando do Colaboratory para manipular a base de dados. O método de pesquisa aplicado foi o experimental e o material utilizado foi proveniente da plataforma physionet.

Inicialmente foi feito um janelamento dos sinais a fim de facilitar o trabalho com os mesmo, já que o trabalho estende-se a uma base de dados de algumas centenas de sinais com tamanhos diferente, tratar todos estes por inteiro sem nenhum tipo de tratamento prévio seria inviável. Desta forma foram utilizados alguns parâmetros para o janelamento que facilitariam a leitura e tratamento dos dados de ECG. Foi definido que a frequência utilizada seria de $F_s = 300\text{Hz}$, frequência esta que foi determinada pela natureza dos dados armazenados da pesquisa, além disto também foi utilizada uma duração de 3s para o janelamento, permitindo uma análise parcial que contenha uma quantidade significativa de dados.

Após a obtenção das janelas, foi aplicado um filtro passa faixa individualmente a cada uma delas. Foi utilizado um Filtro Butterworth de ordem 7 e foram testadas algumas faixas de frequência para obter a melhor resposta do mesmo. Assim, como pode ser visto no estudo publicado por Tereshchenko LG e Josephson ME. [2], as frequências abaixo de 100Hz carregam grande parte da informação do ECG, possibilitando assim que a mesma seja utilizada como a frequência de corte superior mantendo as informações necessárias do sinal. Já para a frequência de corte inferior foram utilizados inicialmente valores mais baixos de 0.15Hz e 0.5Hz , já que sinais abaixo destes valores poderiam ser considerados ruídos causados por movimentações do corpo. Realizando alguns testes foi escolhida uma frequência de corte inferior de 5Hz e pode-se observar que o resultado acabou por ser inferior ao obtido anteriormente com uma frequência de 0.5Hz provando assim que a faixa de frequência inicial realmente apresenta a melhor faixa de sinal a ser filtrada.

Para o modelo de classificação funcionar bem foi preciso gerar valor das janelas filtradas, desta forma, foi utilizado a função `dada` ('atributos') pelas monitoras. A função recebe como entrada as janelas e retorna os atributos de cada janela empilhada em um data frame.

O *software* usado para testar o algoritmo na base de dados foi o Weka versão 3.8.4. O algoritmo rodado foi o Random Forest com 10, 50 100 árvores/iterações.

Por fim, as métricas avaliadoras usadas foram: Acurácia, Kappa, MCC, Precisão, Sensibilidade, Área da curva ROC, taxa de falso positivo.

Todos os algoritmos foram rodados com os parâmetros:

- Experiment type
Number of folds: 10
Classification

IV. RESULTADOS

1. Janelamento

A partir do janelamento descrito anteriormente, foram obtidas janelas de 3s como a que pode ser observada na figura 5.

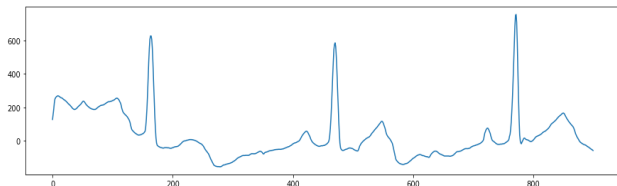


Figura 5: Janela de um sinal normal

Estas janelas foram utilizadas para poder realizar uma análise otimizada do sinal mantendo a qualidade dos dados. Para realizar tal análise foi criada uma função ('gera janelas') que realiza a quebra do sinal em loop descartando janelas que não venham a possuir dados suficientes a fim de manter a qualidade das análises dos sinais.

2. Filtragem

Após realizado o janelo de tais sinais será aplicado um filtro Butterworth de 7º ordem. Este filtro tem como objetivo remover ruídos do sinal que possam ser causados por movimentos do corpo durante as medições, assim como pela atividade elétrica dos músculos.

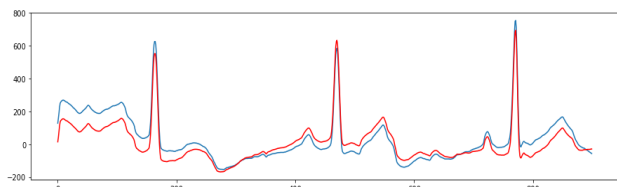


Figura 6: Sinal de entrada do filtro (azul) e sinal filtrado 0.5Hz-100Hz (vermelho).

Na Figura 6 acima é possível observar o funcionamento do filtro através da sobreposição do sinal de entrada (azul) e de saída (vermelho). Para este caso, foi utilizada uma frequência de corte inferior de 0.5Hz, evitando assim os ruídos causados por movimentos e ainda assim mantendo a qualidade no sinal. Já para a Figura 3, pode-se observar que as curvas de entrada (azul) e saída (vermelho) estão mais afastadas, isto ocorreu pois para este filtro foi utilizado uma frequência de corte inferior de 5Hz a fim de tentar reduzir ainda mais os ruídos do sinal.

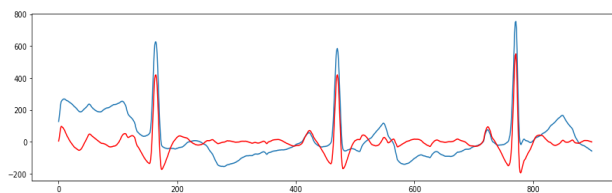


Figura 7: Sinal de entrada do filtro (azul) e sinal filtrado 5 Hz-100 Hz (vermelho).

Observando os resultados finais que serão abordados mais à frente, pode-se ver que ocorreu uma perda da qualidade do sinal ao utilizar uma frequência inferior, fazendo assim com que o percentual de classificação correto caísse.

3. Atributos

Depois da filtragem gerou-se o valor das janelas com a coleta de atributos, estes que estão listados na figura 8.

MVAL	Mean Value	WFL	Waveform Length
VAR	Variance	ZCS	Zero Crossings
STD	Standard Deviation	SSC	Slope Sign Changes
RMS	Root Mean Square	MNF	Mean Frequency
AAC	Average Amplitude Change	MDF	Median Frequency
DASDV	Difference Absolute Deviation	MNP	Mean Power
IAV	Integrated Absolute Value	PKF	Peak Frequency
LOGD	Logarithm Detector	PSR	Power Spectrum Ratio
SSI	Simple Square Integral	TTP	Total Power
MAV	Mean Absolute Value	SM1	1st Spectral Moments
MLOGK	Mean Logarithm Kernel	SM2	2nd Spectral Moments
KURT	Kurtosis	SM3	3rd Spectral Moments
MAX	Maximum Amplitude	VCF	Variance of Central Frequency
TM3	Third Moment		
TM4	Fourth Moment		
TMS	Fifth Moment		

Figura 8: Listagem de atributos

3. Weka

A “Random Forest” é um algoritmo que consiste na junção de várias árvores de decisão (daí o termo *Forest* ou Floresta) de forma aleatória, sendo os parâmetros mais comuns de serem alterados a profundidade máxima das árvores e a quantidade delas. Nesse caso foi alterada apenas a quantidade de árvores, variando em 10, 50, e 100.

A. Usando a base de dados com os sinais filtrados no intervalo de 0.5Hz-100Hz

1.1 Random Forest com 10 árvores

Tabela 1

Acurácia	Kappa	AUROC	Precisão	Sensibilidade	MCC	FPR
64.35	0.254	0.697	0.625	0.644	0.284	0.418

1.2 Random Forest com 50 árvores

Tabela 2

Acurácia	Kappa	AUROC	Precisão	Sensibilidade	MCC	FP R
67.26	0.297	0.754	0.689	0.673	0.350	0.415

1.3 Random Forest com 100 árvores

Tabela 3 Melhor resultado da Random Forest

Acurácia	Kappa	AUROC	Precisão	Sensibilidade	MCC	FP R
67.64	0.302	0.766	0.699	0.676	0.360	0.415

Neste caso, a floresta com 100 árvores foi a que obteve melhores resultados em todas as métricas avaliadas, como visto na tabela 3.

B. Usando a base de dados com os sinais filtrados no intervalo de 5 Hz-100 Hz

2.1 Random Forest com 10 árvores

Tabela 4

Acurácia	Kappa	AUROC	Precisão	Sensibilidade	MCC	FP R
60.02	0.155	0.623	0.556	0.60	0.179	0.462

2.2 Random Forest com 50 árvores

Tabela 5

Acurácia	Kappa	AUROC	Precisão	Sensibilidade	MCC	FP R
62.89	0.189	0.664	0.614	0.629	0.237	0.466

2.3 Random Forest com 100 árvores

Tabela 6 Melhor resultado com filtragem no intervalo 5Hz-100Hz

Acurácia	Kappa	AUROC	Precisão	Sensibilidade	MCC	FP R
63.20	0.191	0.673	0.623	0.632	0.244	0.468

Repetiu-se o processo já feito anteriormente, porém com um intervalo de filtragem diferente e, novamente, a floresta com 100 árvores foi a que melhor se destacou, porém pode-se observar que o resultado teve uma queda no desempenho. O que faz sentido porque da forma que os dados foram filtrados dessa vez, apesar deles apresentarem menos ruído a qualidade do filtro diminuiu. E com isso, é interessante observar que as FP R.(taxa de falso positivo), tiveram um aumento mostrando que claramente houver perda na qualidade do sinal.

V. CONCLUSÃO

Assim como citado anteriormente, foram realizados diversos testes para obter os melhores resultados dos dados. Tais testes foram de extrema importância para o grupo aprender com o problema e se motivar a pesquisar e buscar novas informações a fim de conseguir a melhor resposta possível. Com isso, apesar de todos os parâmetros avaliadores terem sido levados em questão na decisão final do modelo de algoritmo, o melhor modelo obtido foi o algoritmo Random Tree com 100 árvores e Usando a base de dados com os sinais filtrados no intervalo de 0.5Hz-100Hz.

VI. AGRADECIMENTOS

Este trabalho recebeu apoio dos professores Wellington Pinheiro dos Santos, Maira Araujo de Santana e Clarisse Lins de Lima, que contribuíram com a escolha da base de dados, mas também dando suporte e esclarecendo dúvidas durante toda execução do projeto.

VII. REFERÊNCIAS

1. AF Classification from a Short Single Lead ECG Recording: The PhysioNet/Computing in Cardiology Challenge 2017
Disponível em:
<<https://physionet.org/content/challenge-2017/1.0.0/>>
Acesso em: 30 de abr. 2022.
2. Tereshchenko LG, Josephson ME. Frequency content and characteristics of ventricular conduction. J Electrocardiol. Disponível em:
<<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4624499/>>
Acesso em: 30 de abr. 2022.
3. BROWNLEE, Jason. A gentle introduction to k-fold Cross-Validation. Machine Learning Mastery. Disponível em:
<<https://machinelearningmastery.com/k-fold-cross-validation/>>.
Acesso em: 01 de mai. 2022.