O Impacto da Métrica de Distância na Classificação de Arritmias com o Algoritmo KNN

Renan Catini Amaral[1]
Reginaldo José da Silva[2]
Angela Leite Moreno[3]

As arritmias cardíacas representam um risco significativo para a saúde, sendo fundamental um diagnóstico preciso para um acompanhamento clínico adequado. Entretanto, a análise de eletrocardiogramas (ECG) baseada na interpretação visual é uma tarefa complexa, que pode ser subjetiva e gerar divergência entre profissionais [1]. Assim, ao se utilizar algoritmos de Aprendizado de Máquina (*Machine Learning*-ML) e padrões do ECG, consegue-se prever o diagnóstico com maior precisão e confiabilidade.

É nesse cenário que este trabalho está inserido, separando os diagnósticos em normais e anormais, buscando analisar qual configuração do modelo é mais sensível para identificar corretamente cada tipo de batimento. O conjunto de dados utilizado para tal tarefa foi o tratado por [2], obtido diretamente do "*MIT-BIH Arrhythmia Database*" [3]. Este conjunto contém registros de ECG com cerca de 30 minutos de duração. No pré-processamento, cada exame foi separado batimento por batimento, visando treinar o modelo para reconhecer os diferentes tipos de batimentos. Cada batimento foi rotulado seguindo o padrão AAMI, sendo eles: batimentos normais (N), supraventriculares (S), ventriculares (V), fusão (F) e não classificáveis (Q). Ao se analisar a quantidade de dados em cada classe, verificou-se que ele é altamente desbalanceado, com cerca de 93% de casos classificados como normais. Portanto, tais rótulos foram separados em dois grupos: os normais (N), compostos somente por batimentos rotulados como N, e os anormais (AN), composto por todos os outros rótulos, visando reduzir o desbalanceamento e possibilitar a formulação de um problema binário, para uma primeira investigação.

Para realizar a classificação, o algoritmo de aprendizado de máquina escolhido foi o K-*Nearest Neighbors* (KNN). A base de dados foi dividida em 70% para treinamento e 30% para teste. A busca por hiperparâmetros foi realizada no conjunto de treinamento utilizando a técnica 5-*fold* comparando-se o desempenho do modelo sob diferentes métricas de distância (*Minkowski*), controladas pelo hiperparâmetro p. O resultado indicou que, dentre os valores testados para o número de vizinhos ($\texttt{n\_neighbors} \in \{5, 7, 9, 11, 15\}$, o valor ideal foi 5 e que o uso de pesos ponderados pela distância ($\texttt{weights} = \texttt{distance}$) apresentou os melhores resultados consistentemente. Para os casos $p = 1$ e $p = 2$, esses valores foram encontrados por meio de uma busca em grade (*Grid Search*). No entanto, para os testes com $p = 4$ e $p = 8$, a busca completa se mostrou inviável devido ao alto

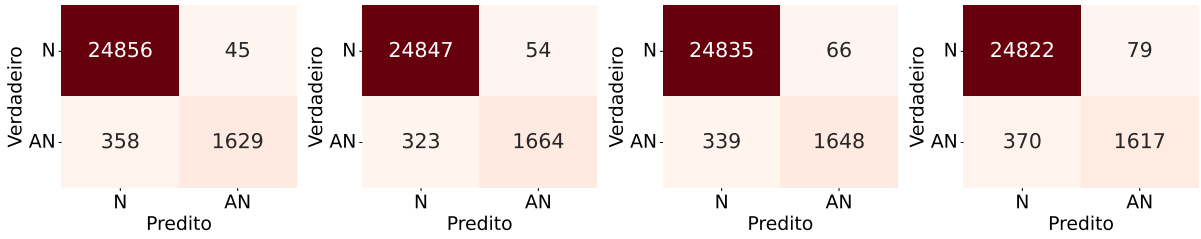
[1]Departamento de Ciência da Computação, Universidade Federal de Alfenas, renan.amaral@sou.unifal-mg.edu.br
[2]Faculdade de Engenharia de Ilha Solteira, Universidade Estadual Paulista, reginaldo.silva@unesp.br
[3]Departamento de Matemática, Universidade Federal de Alfenas, angela.moreno@unifal-mg.edu.br

custo computacional. Nesses casos, o valor de `n_neighbors` foi fixado em 5 (com base no resultado anterior) e a otimização focou no hiperparâmetro `weights`. Todo o processo de otimização foi focado em maximizar a Sensibilidade.

Os resultados dos testes, reunidos na Tabela 1, apontam a distância Euclidiana ($p = 2$) como a de melhor desempenho. Esta configuração se destacou por alcançar a maior Sensibilidade, com 83,74%. A prioridade foi dada a esta métrica, pois em um diagnóstico clínico, a importância de um Falso Negativo (não detectar uma arritmia real) é muito maior do que o de um Falso Positivo. A superioridade do modelo com $p = 2$ é confirmada em sua matriz de confusão (Figura 1), que mostra o menor número de Falsos Negativos entre os testes: 323, ou seja, 323 batimentos anormais foram classificados como normais, ao custo de somente 9 batimentos normais classificados como anormais em relação ao melhor modelo apresentado. Portanto, esta configuração provou ser a mais eficaz para a tarefa. Os próximos passos da pesquisa são verificar se isso se mantém para o problema multiclasse e também utilizar outros classificadores para tratar o problema.

Métrica	Accuracy	Recall	ROC	F1	Precision	Specificity
1	98,50	81,98	0,9090	0,8899	97,31	99,82
2	98,60	83,74	0,9176	0,8982	96,86	99,78
4	98,49	82,94	0,9134	0,8906	96,15	99,73
8	98,33	81,38	0,9053	0,8781	95,34	99,68

Tabela 1: Resultados do treinamento do modelo KNN.



(a) $p = 1$. (b) $p = 2$. (c) $p = 4$. (d) $p = 8$.

Figura 1: Matrizes de confusão do modelo KNN para cada métrica. Fonte: Os autores.

Referências

[1] A. A. Ahmed, W. Ali, T. A. Abdullah, and S. J. Malebary, "Classifying cardiac arrhythmia from ecg signal using 1d cnn deep learning model," *Mathematics*, vol. 11, no. 3, p. 562, 2023.

[2] R. J. Silva *et al.*, "Classificação de arritmias no tempo e tempo-frequência: uma abordagem baseada em subproblemas," in *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, vol. 11, pp. 010360–1–7, 2025.

[3] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.