Seleção de features para classificação de ECG: análise de novo método baseado em diversidade em grafos de visibilidade

Paulo Coelho[1], **Samir Saliba**[1], **Luís Ramos** [1], **Renato Vimieiro** [1]

[1]Instituto de Ciências Exatas - Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil.

{paulohdscoelho,samirsaliba,luisfeliperamos,rvimieiro}@dcc.ufmg.br

Abstract. *Here, we propose an innovative approach for feature selection in electrocardiogram classification, employing visibility graphs and a diversity metric. The methodology is evaluated through a classification pipeline, comparing the effectiveness of feature selection with random choices. Preliminary results are shown.*

Resumo. *É proposta uma abordagem inovadora para a seleção de características em classificação de eletrocardiogramas, empregando grafos de visibilidade e uma métrica de diversidade. A metodologia é avaliada por meio de um pipeline de classificação, comparando a eficácia da seleção de características com escolhas aleatórias. Resultados preliminares são apresentados.*

1. Introdução

Doenças cardíacas são a principal causa de mortes globalmente, tornando essencial um diagnóstico preciso e precoce que aumente a eficácia do tratamento e potencialmente salve vidas. Contudo, a escassez de cardiologistas em regiões remotas e desfavorecidas, particularmente no contexto brasileiro, pode resultar em atrasos no diagnóstico e tratamento dessas doenças. O Eletrocardiograma (ECG) de 12 derivações é o principal exame para avaliar a saúde cardiovascular, fornecendo uma representação gráfica da atividade cardíaca. No entanto, nem todas as 12 derivações são igualmente informativas para o diagnóstico. Algumas podem ser irrelevantes, prejudicando a classificação automática de doenças cardíacas. Portanto, o estudo de técnicas de seleção de features torna-se crucial para agilizar o ajuste de modelos e melhorar a precisão da classificação de doenças.

Este trabalho contribui para a discussão da classificação automática de doenças cardíacas ao propor uma abordagem matemática baseada na representação de sinais de ECG como grafos e aplicação de uma métrica de diversidade para selecionar as features mais relevantes para treinar modelos de classificação. A abordagem proposta tem como objetivo identificar as derivações mais relevantes para se prever determinada anomalia, evitando características redundantes que possam prejudicar a precisão dos modelos, e, assim, reduzir o tempo de ajuste dos classificadores.

2. Trabalhos relacionados

A representação de Séries Temporais como grafos visa construir um grafo em que as arestas estão ativas em pontos específicos no tempo, incorporando informações adicionais da série temporal [Holme and Saramäki 2012]. O Grafo de Visibilidade (GV), proposto

por [Lacasa et al. 2008], é um método que mapeia séries temporais em um grafo $G(V, E)$, representando observações como vértices dispostos em barras verticais e arestas entre vértices refletindo a visibilidade mútua das barras, proporcionando uma representação gráfica e intuitiva das relações de visibilidade na série temporal.

No contexto da detecção de arritmias cardíacas, [Oliveira et al. 2022] propõem o mapeamento de ECG em GV e utiliza Graph Neural Networks (GNN) para classificação. Apesar da inovação, a falta de validação de baseline e a utilização de apenas uma derivação para classificação sem explicação clara da seleção de features são limitações. Em contraste, [Ribeiro et al. 2020] implementa uma ResNet para classificar ECGs em 6 anomalias, superando em alguns casos a performance de cardiologistas. Contudo, a complexidade temporal elevada para treinar o modelo e a falta de explicabilidade para a escolha da ResNet são observações pertinentes.

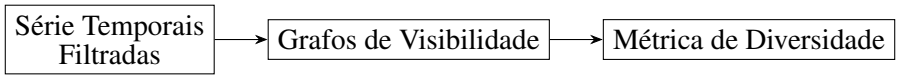
Introduzindo a Métrica de diversidade em grafos (MDG), [Carpi et al. 2019] propõem uma abordagem matemática para analisar a diversidade em sistemas. Aplicamos essa métrica para reduzir as camadas da rede preservando sua estrutura e informações essenciais. Esse processo recursivo calcula a distância entre as camadas, removendo as menos relevantes em termos de diversidade, resultando em um ranking ordenado das camadas menos contributivas, permitindo uma exclusão seletiva.

3. Metodologia

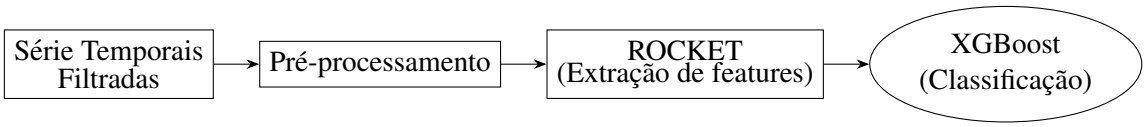
Para a representação e classificação de sinais de ECG propomos a sua modelagem como GV, onde as derivações do ECG atuam como camadas. A diversidade entre os grafos é calculada pela ordenação ascendente das derivações com base na Métrica de diversidade proposta por [Carpi et al. 2019], seguida da aplicação do classificador de séries temporais ROCKET, modificado com o uso de XGBoost como classificador [Dempster et al. 2020, Chen et al. 2015]. Experimentos foram conduzidos, explorando a remoção seletiva das derivações a partir do ranking de diversidade, comparando-a com a eliminação aleatória, proporcionando insights sobre o papel de cada derivação na representação e classificação eficaz de ECG.
Foi utilizada a base "CODE-15%,"com 345.779 exames de 233.770 pacientes, sendo uma amostra aleatorizada de 15% da base "CODE,"coletada e rotulada pelo Centro de Telessaúde do Hospital das Clínicas da UFMG [Ribeiro et al. 2021]. Cada exame é rotulado entre seis classes de anomalias cardíacas: bloqueio atrioventricular de primeiro grau (**1dAVb**), bloqueio de ramo direito (**RBBB**), bloqueio de ramo esquerdo (**LBBB**), bradicardia sinusal (**SB**), fibrilação atrial (**AF**) e taquicardia sinusal (**ST**).

Para avaliar a eficácia do método de seleção de features proposto, foi implementado o pipeline de classificação dos ECGs da figura 1 comparando ao fim o desempenho na tarefa de classificação utilizando as três melhores features selecionadas a partir do ranking de diversidade contra 34 seleções aleatórias de três features. Esse teste investiga se a seleção de features pelo método apresenta resultados superiores à uma escolha aleatória. Os dados foram divididos em subconjuntos de treinamento e teste, na proporção 70% e 30% respectivamente. Essa divisão proporciona uma avaliação razoável do desempenho do método proposto. Mantendo variáveis de semente de aleatoriedade constantes, o LabelEncoder do scikit-learn foi utilizado para codificar os rótulos [1].

[1]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html



(a) Criação do ranking de métrica de diversidade



(b) Avaliação dos testes na classificação dos ECGs

Figura 1. Pipeline proposto para classificação de ECGs com VG e MDG

4. Resultados preliminares

Os valores F1-score para as 6 classes consideradas no estudo estão na figura 2, incluindo testes com todas as derivações, seleção pelo método de diversidade e 9 testes de seleções randomizadas de 3 derivações. Ao analisar os resultados, observamos que, nos testes com a seleção de features obtida através do ranking de diversidade, o desempenho superou a maioria das seleções aleatórias - embora não tenha sido a melhor opção geral.
É relevante destacar, no entanto, que os testes na 2ª e 3ª posições, que obtiveram resultado pior em relação à classificação incluindo todas as derivações, utilizaram as 3 melhores features ranqueadas, sendo V1, V3, V5 e V1, V3 V6, respectivamente. A derivação V1, presente em todos os testes no top 10, não foi identificada como uma das mais importantes pelo método proposto.

Tipo de seleção	Modelo	1davb	af	lbbb	rbbb	sb	st
Todas as Derivações		0.71	0.79	0.93	0.91	**0.9**	0.94
Seleção do modelo	"V3, V5, V6"	0.65	0.77	0.9	0.83	0.89	0.94
Seleção aleatória	"V1, V3, V5"	0.71	0.78	0.92	0.9	0.91	0.95
	"V1, V3, V6"	0.69	0.77	0.92	0.9	0.9	0.94
	"V1, V3, V4"	0.68	0.77	0.92	0.89	0.9	0.94
	"V1, V4, V5"	0.69	0.77	0.92	0.89	0.9	0.94
	"V1, V5, V4"	0.68	0.76	0.92	0.89	0.9	0.94
	"V1, V4, V6"	0.69	0.76	0.92	0.89	0.9	0.93
	"V1, V5, V6"	0.68	0.76	0.91	0.88	0.9	0.93
	"DI, V1, V5"	0.67	0.77	0.91	0.88	0.89	0.93
	"V1, V2, V6"	0.67	0.76	0.92	0.88	0.89	0.92

Figura 2. Top 10 F1-score para classificação com Rocket, selecionando 3 LEADS para análise. O treino foi realizado sem validação cruzada. Por se tratar de um resultado preliminar, não apresentamos aqui análises como desvio-padrão. Dada a limitação de páginas, deixamos as discussões para a apresentação do pôster

Destacamos que, nos problemas de classificação de ECGs, cada classe de doença está associada a uma das derivações, evidenciando uma correlação com as classes. A escolha da derivação V1, por exemplo, resultou em melhores resultados de F1 para a classe RBBB, relacionada ao bloqueio de ramo direito, indicando que ela pode refletir efetivamente a região direita do coração. Essa observação abre portas para futuras investigações, como abordagens binárias para cada classe, visando identificar qual derivação melhor codifica informações para uma determinada doença, o que permitiria uma compreensão mais profunda de suas relações intrínsecas e sua capacidade de discriminação das diferentes classes de doenças.
Os testes realizados são limitados nessa perspectiva, uma vez que o método de seleção

busca as três derivações mais relevantes para um problema de seis classes, enquanto cada classe pode ser melhor representada por uma derivação diferente.

5. Conclusão

Neste trabalho, apresentamos uma abordagem inovadora para a seleção de features em classificação de ECGs utilizando grafos de visibilidade e métrica de diversidade. Para avaliar essa seleção, implementamos um pipeline de classificação de ECGs, buscando determinar se o desempenho na classificação com o conjunto de features escolhido pelo método proposto supera o de um subconjunto de features escolhido aleatoriamente. Os resultados indicam que, embora o método de classificação proposto tenha alcançado resultados preliminares de F1-score aceitáveis, a seleção de features não encontrou um subconjunto ótimo para a tarefa de classificação, não superando a seleção aleatória.

Contudo, este estudo destaca aspectos que serão aprimorados com a continuidade da pesquisa, como a consideração de uma abordagem de classificação binária para cada classe, buscando evidenciar uma correlação entre ela e uma determinada derivação. Outra possibilidade sugerida seria utilizar o método proposto para identificar qual derivação codifica melhor a informação para cada classe, validando diretamente com especialistas da área, eliminando a etapa de classificação da questão.

Referências

Carpi, L. C., Schieber, T. A., Pardalos, P. M., Marfany, G., Masoller, C., Díaz-Guilera, A., and Ravetti, M. G. (2019). Assessing diversity in multiplex networks. *Scientific reports*, 9(1):1–12.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Dempster, A., Petitjean, F., and Webb, G. I. (2020). Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495.

Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97–125. Temporal Networks.

Lacasa, L., Luque, B., Ballesteros, F., Luque, J., and Nuno, J. C. (2008). From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13):4972–4975.

Oliveira, R., Freitas, V., Moreira, G., and Luz, E. (2022). Explorando redes neurais de grafos para classificação de arritmias. In *Proceedings of the 22nd Brazilian Symposium on Computing Applied to Health*, pages 178–189, Porto Alegre, RS, Brasil. SBC.

Ribeiro, A. H., Paixao, G. M., Lima, E. M., Horta Ribeiro, M., Pinto Filho, M. M., Gomes, P. R., Oliveira, D. M., Meira Jr, W., Schon, T. B., and Ribeiro, A. L. P. (2021). CODE-15%: a large scale annotated dataset of 12-lead ECGs.

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P., Andersson, C. R., Macfarlane, P. W., Meira Jr, W., et al. (2020). Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1–9.