Ministério da Educação
Universidade Tecnológica Federal do Paraná
Pró-Reitoria de Pesquisa e Pós-Graduação

Relatório Final de Atividades

Detecção de padrões em sinais por técnicas de aprendizado de máquinas

vinculado ao projeto

Controle de Sistemas Lineares Variantes no Tempo

Gabriel Lechenco Vargas Pereira
Bolsista PIBIC/CNPq
Engenharia de Computação
Data de ingresso no programa: 08/2019
Prof. Dr. Cristiano Marcos Agulhari

Área do Conhecimento: 3.04.05.03-3 — Controle de processos eletrônicos, retroalimentação.

CÂMPUS CORNÉLIO PROCÓPIO, 2019

GABRIEL LECHENCO VARGAS PEREIRA
CRISTIANO MARCOS AGULHARI

DETECÇÃO DE PADRÕES EM SINAIS POR TÉCNICAS DE APRENDIZADO DE MÁQUINAS

Relatório de Pesquisa do Programa de Iniciação Científica da Universidade Tecnológica Federal do Paraná.

CORNÉLIO PROCÓPIO, 2019

SUMÁRIO

INTRODUÇÃO

A arritmia cardíaca é uma condição que causa alterações na geração de impulsos elétricos no coração, podendo ocasionar em perturbações no ritmo cardíaco, podendo ser fatal como em casos de pacientes com taquicardia ou bradicardia. Esses episódios são chamados de morte súbita cardíaca, ocorrendo de forma instantânea e inesperada pela falência do órgão cardíaco. São cerca de 20 milhões de pacientes diagnosticadas com algum tipo de arritmia, causando mais de 320 mil mortes por ano no Brasil [1].

Devido sua gravidade e grande incidência na população brasileira (cerca de 10%) seu diagnóstico deve ser identificado o mais rápido possível e de forma precisa. A forma mais tradicional de se identificar um artefato de arritmia é de responsabilidade de um cardiologista, que irá analisar o eletrocardiograma de cada paciente manualmente. Essa técnica pode, por vários fatores internos ou externos, estar sujeita à falhas no diagnóstico, além de ser muito custosa por depender de recursos humanos, não podendo monitorar uma grande parte da população em tempo real.

Uma forma de automatizar esta tarefa é pela utilização de técnicas de *Machine Learning*, que realizam um treinamento sobre uma quantidade relevantes de dados a fim de reconhecer padrões presentes nestes dados, podendo os separar por classes, e reconhecer quais classes novas entradas melhor se encaixam. O algoritmo gerado por este processo é também conhecido como classificador de padrões, sendo diversas as técnicas para este fim, como por exemplos as redes neurais artificiais, o algoritmo dos k-vizinhos mais próximos (KNN) e as máquinas de vetores de suporte (SVM).

Será utilizado neste trabalho este último, por sua simplicidade e por ter obtido bons resultados com sinais de eletrocardiogramas na detecção de arritmia [2]. Entretanto, serão utilizados como conjunto de características os coeficientes de energia provenientes de transformadas Wavelet.

METODOLOGIA

Devido as dificuldades apresentadas anteriormente, este trabalho propõem o desenvolvimento de um classificador para identificar trechos de eletrocardiograma que contém artefatos de arritmia. Foram utilizadas técnicas de aprendizado supervisionado junto com máquinas de vetores de suporte (SVM) para se chegar ao objetivo proposto. Os detalhes referentes à implementação do classificador e análises dos dados examinados serão descritos nesta seção.

Base de dados. Duas foram as bases de dados utilizadas durante o projeto, sendo ambas recolhidas e disponibilizadas abertamente pelo *Boston's Beth Israel Hospital* (BIH) em conjunto com o Instituto Tecnológico de Massachussets (MIT). A primeira base selecionada contém sinais de eletrocardiograma coletados de pacientes diagnosticados com arritmia, os sinais e os outros dados dos pacientes estão alocados no *MIT-BIH Arrhythmia Database* (mitdb) [3]. Também foram utilizados os sinais de pacientes saudáveis para uma comparação mais generalista das classes, esses sinais foram extraídos do *MIT-BIH Normal Sinus Rhythm Database* (nrsdb) [4].

Esses conjuntos de dados trazem não só os sinais de eletrocardiograma, mas também a localização de cada pico R junto com uma anotação se este está contido em um batimento normal ou se apresenta alguma anomalia. São diversas as classes dessas anomalias, identificando desde artefatos específicos de arritmia ou até se houve algum problema na leitura devido a algum movimento do paciente. Informações como essas são de suma importância quando se trata de um classificador de aprendizado supervisionado, facilitando a definição das classes e a remoção de

Tabela 1. Número de instâncias extraídas para cada tipo de janelamento.

Database	Classe	Nº de instâncias
nrsdb	saudável	36655
mitdb	patológico	36655
mitdb	saudável	45705

ruídos durante o procedimento de treino.

Para o trabalho apresentado, as classes informadas nos dados foram aglomeradas em apenas duas: sinais saudáveis, se tratando de trechos que contém apenas batimentos rotulados como normais; e sinais patológicos, os quais contém pelo menos um batimento rotulado como alguma forma de arritmia. Trechos que apresentam outros tipos de informação, como ruídos ou problemas de leitura não foram considerados.

Janelamento dos dados. Após identificar e se adquirir os dados, foi identificado algumas características que divergiam entre as duas base de dados, valores como a taxa de amostragem encontravam-se incompatíveis, sendo de $F_s = 360Hz$ para o mitdb e de $F_s = 128Hz$ para o nrsdb. Para evitar que essa informação influenciasse no classificador, optou-se por realizar uma reamostragem do sinais, padronizando a taxa de amostragem para $F_s = 128Hz$.

Ao adotar essas padronizações, evita-se que o classificador seja influenciado por estes dados que são paralelos e irrelevantes ao problema, o que poderia ser visto como uma boa classificação entre sinais saudáveis e patológicos poderia na verdade estar julgando informações externas ao problema, invalidando o estudo. Para padronizar também os trechos de eletrocardiograma, optou-se por dividir e salvar em diferentes arquivos intervalos de $8s$, ou 1024 pontos, dos sinais, sendo cada um desses arquivos uma instância de dados com os rótulos atribuídos a cada batimento daquele trecho junto com o sinal em si. Este janelamento dos dados com múltiplos batimentos se justifica pois a percepção da presença da arritmia se deve principalmente pela irregularidade no espaçamento entre os batimentos, logo, ao se observar diversos batimentos tem-se um melhor contexto sobre a questão.

Foram realizados a princípio 3 janelamentos: o janelamento de trechos saudáveis, apenas com notações de batimentos normais, extraídos dos sinais de nrsdb; o janelamento dos trechos patológicos, com pelo menos uma anomalia encontrada nos sinais de mitdb; e por último escolheu-se também realizar o janelamento de trechos saudáveis extraídos também da base de dados mitdb. Essa escolha se deve para se ter uma maior generalidade dos dados, garantindo que o classificador não tentará separar os sinais pelos dois *dataset*, mas sim pelas duas classes que foram propostas. Dessa forma, foram gerados 119015 instâncias de dados distribuídos entre as classes, como observado pela tabela 1.

Extração de características Entretanto, apenas realizar o janelamento dos sinais não é o suficiente. Os trechos recortados são formados por 1024 pontos, os quais podem por si só não expressar muita informação, muitos desses pontos são leituras entre batimentos, por exemplo, e acabam não tendo muita relevância sobre o problema. Dessa forma, deve-se pensar em outras características mais expressivas, tentando assim otimizar o número de características utilizadas pelo classificador sem comprometer a qualidade do mesmo.

Portanto, cada trecho de sinal retirado das bases de dados passou por um pré processamento, decompondo eles por meio de transformadas wavelet e reunindo a energia relativa de cada um dos nós folhas da árvore de decomposição. Utilizando essa técnica, que mostrou resultados promissores em [5] e em [6], podemos analisar o trecho não apenas no domínio do

Tabela 2. Conjuntos de características testados

Teste	Níveis de decomposição Wavelet	Numero de Caracteristicas
I	3 níveis	8
II	4 níveis	16
III	4 níveis + expansão da primeira folha em mais 2 níveis	20

tempo, já que a Transformada Wavelet pode ser interpretada como uma série de filtros passa alta e passa baixa, agrupando em cada nó uma combinação diferente entre o domínio do tempo e o da frequência. Dessa forma, utilizando a função 3 da família *Daubechies* foram analisados algumas combinações referentes ao número de níveis de decomposição e como estes influenciaram nos resultados finais da classificação, a tabela 2 especifica os conjuntos de características testados.

O terceiro teste surgiu de uma hipótese, já que a arritmia se caracteriza por perturbações nas baixas frequências do sinal, a expansão da primeira folha pode trazer uma melhor separação dessas frequências, podendo beneficiar a performance do classificador.

Reduzir significantemente a quantidade de características a serem analisadas acarretará em uma diminuição do esforço computacional para realizar o treinamento, melhorando a performance do classificador. Além disso, os valores extraídos dos dados, são mais expressivos para a solução do problema em comparação com o trecho apenas no domínio do tempo, incrementado também a eficácia do algoritmo.

Máquina de Vetor de Suporte Para a construção do classificador optou-se por utilizar Máquina de Vetor de Suporte (SVM), devido a sua simplicidade e por ser amplamente utilizada na literatura em estudos com sinais de eletrocardiograma, como visto em [5] e em [6]. Sua classificação ocorre de forma binária ao encontrar um hiperplano $\hat{w}$ que separe da melhor forma possível as duas classes analisadas. A definição do problema de classificação pode ser descrito pela equação 1.

$$f(t) = sign(k(\hat{w}, x) + w_0) \tag{1}$$

Onde x é o vetor de característica que será classificada, k é o *kernel* utilizado durante o treinamento e $w_0/||\hat{w}||$ determina a distância do hiperplano até a origem. A classificação é definida por qual lado do hiperplano o ponto se encontra, isso pode ser facilmente determinado pelo sinal do valor obtido.

Para este trabalho, os dados recolhidos foram separados aleatoriamente entre treino e teste, em uma proporção de 80% para treino e 20% para teste. Devido ao grande número de instâncias, optou-se por manter os dados balanceados, utilizando então a mesma quantidade de instâncias rotuladas como saudáveis e patológicas. O treinamento foi realizado utilizando os kernels o linear e gaussiano, dois dos mais comuns kernels implementados em SVM's. O primeiro se trata do kernel original em que o algoritmo de máquinas de vetores de suporte foi proposto, tendo bons resultados para dados linearmente separáveis, já o kernel gaussiano é mais adaptativo, conseguindo bons resultados mesmo em dados não lineares. Dessa forma, ao realizar o treinamento com múltiplos kernels espera-se não só apresentar um bom classificador, mas também identificar a natureza dos dados e do problema observado.

Tabela 3. Resultados classificadores SVM

Teste	Kernel	Acurácia	Precisão	Recall	F1score
I	linear	-	-	-	-
I	gaussiano	0.830 ±0.003	0.883 ±0.007	0.799 ±0.006	0.839 ±0.003
II	linear	0.856 ±0.004	0.887 ±0.005	0.835 ±0.006	0.860 ±0.004
II	gaussiano	0.904 ±0.003	0.917 ±0.004	0.893 ±0.006	0.905 ±0.003
III	linear	0.851 ±0.005	0.879 ±0.005	0.832 ±0.007	0.855 ±0.004
III	gaussiano	0.942 ±0.004	0.935 ±0.004	0.948 ±0.006	0.941 ±0.004

RESULTADOS E DISCUSSÃO

Para cada um dos conjuntos de características listados na tabela 2 foram realizados duas baterias de testes, uma utilizando o kernel linear e a outra com o kernel gaussiano. As baterias foram realizadas selecionando os dados para treinamento e testes de forma randômica, armazenando ao final métricas comuns na análise de classificadores como a acurácia, a precisão, o recall e a f1score, esse procedimento foi repetido 30 vezes para cada bateria. A média dessas métricas para cada bateria podem ser observadas na tabela 3, na qual para o primeiro teste, o kernel linear acabou não encontrando uma solução ótima para a classificação dos dados

Observando os resultados obtidos (tabela 3) é possível notar um incremento das métricas a medida que a quantidade de características foi acrescentada, indicando os coeficientes de energia wavelet como melhor identificador para arritmia quando o sinal é decomposto um maior número de vezes. Essa afirmação porém, não pode ser vista como um axioma, evidentemente esta melhoria nos resultados irá acontecer até certo ponto, e para decompor para o próximo nível da árvore pode não valer a pena computacionalmente. Com isso em mente, foi proposto a decomposição apenas da primeira folha por dois níveis adicionais (Teste III), o que apresentou uma melhora significativa nos resultados para o kernel gaussiano, indo de 90,4% de acurácia para as energias apenas do quarto nível (Teste II) para 94,2% ao se utilizar a expansão.

Ao comparar os dois kernels utilizados, nota-se que os dados extraídos dos eletrocardiogramas conseguiram bons resultados para ambos, o que demonstra que os dados apresentam algumas características linearmente separáveis. Entretanto, o kernel gaussiano acabou sendo melhor para todos os casos, devido a sua natureza adaptativa. Nas figuras 1 e 2 podemos observar a curva ROC de ambos quando testados com o conjunto de características III.

As duas curvas mostram claramente o melhor desempenho do kernel gaussiano quando comparadas, se aproximando mais de um classificador ideal.

CONCLUSÕES

Arritmias cardíacas estão presentes em cerca de um décimo da população brasileira [1], o desenvolvimento de dispositivos de monitoramento que conseguem distinguir batimentos saudáveis de patológicos podem oferecer um diagnóstico eficaz, antes mesmo de haver complicações sérias para o paciente. Posto isso, a utilização de energias wavelet e máquinas de vetores de suporte para classificação se mostrou promissora com os estudos apresentados neste trabalho.

Os resultados apresentados na tabela 3 corroboram também com a hipótese apresentada, a qual a expansão apenas da primeira folha em mais níveis poderia melhorar os resultados, já que a arritmia costuma alterar o sinal em frequências mais baixas. Dessa forma, o algoritmo foi capaz de alcançar 94,2% de acurácia e uma curva ROC (figura 2) satisfatória ao utilizar o kernel gaussiano.

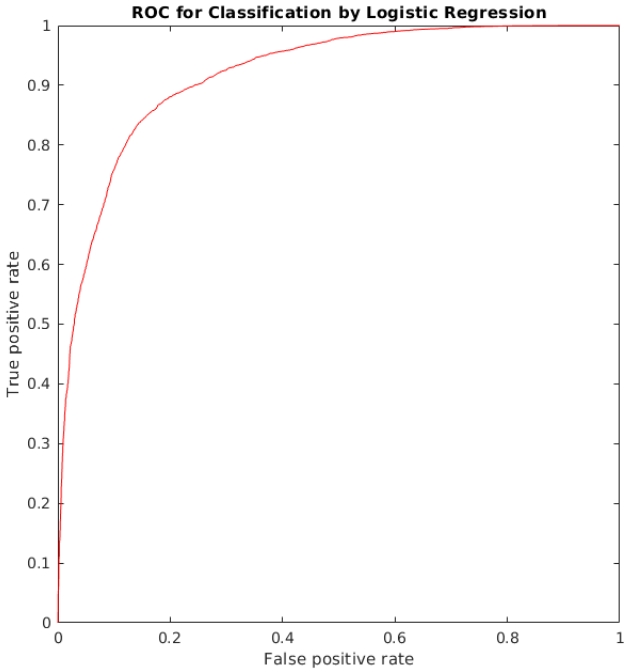

Figura 1. Curva ROC para o kernel linear

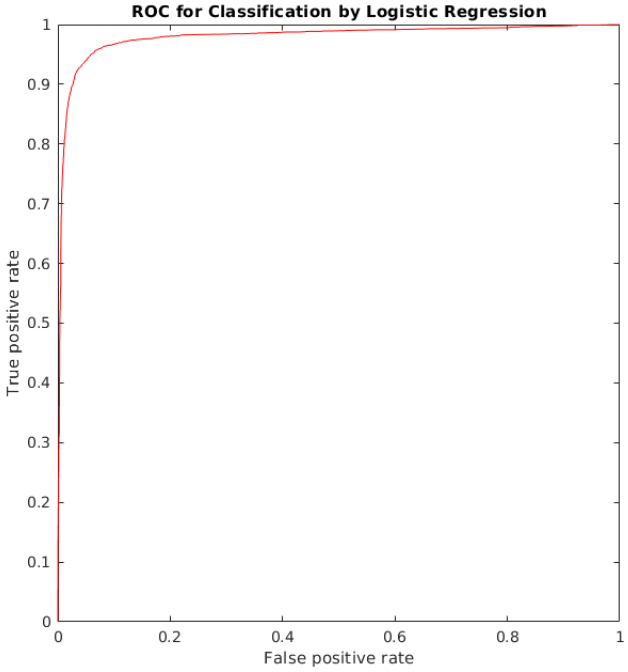

Figura 2. Curva ROC para o kernel gaussiano

Futuramente, espera-se otimizar e melhorar esses resultados com a combinação de outras técnicas como *cross-validation* e a otimização de hiper-parâmetros durante o processo de treinamento. Pode-se explorar também a combinação de novas características, a fim de melhorar ainda mais a confiabilidade do classificador junto com a adição de novas classes para oferecer um diagnóstico mais completo.

AGRADECIMENTOS

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo fomento da pesquisa, e ao meu professor orientador por ter me indicado à bolsa e me instigado a realizar este trabalho.

REFERÊNCIAS

[1] SOBRAC. *Dados sobre Morte Súbita – Coração na Batida Certa*. pt-BR. [S.l.: s.n.]. Disponível em: <http://www.sobrac.org/campanha/arritmias-cardiacas-mortes-subita/>. Acesso em: 29 nov. 2019.

[2] MORA, L. A.; AMAYA, J. E. Proposal of asymmetric multi-classifier of arrhythmias. In: 2012 XXXVIII Conferencia Latinoamericana En Informatica (CLEI). [S.l.: s.n.], out. 2012. p. 1–7. DOI: 10.1109/CLEI.2012.6427169.

[3] MOODY, G. B.; MARK, R. G. *MIT-BIH Arrhythmia Database*. [S.l.]: physionet.org, 1992. type: dataset. DOI: 10.13026/c2f305. Disponível em: <https://physionet.org/content/mitdb/>. Acesso em: 1 out. 2019.

[4] MIT-BIH Normal Sinus Rhythm Database. [S.l.]: physionet.org. Disponível em: <https://www.physionet.org/content/nsrdb/1.0.0/>.

[5] QIBIN ZHAO; LIQING ZHANG. ECG Feature Extraction and Classification Using Wavelet Transform and Support Vector Machines. In: 2005 International Conference on Neural Networks and Brain. [S.l.: s.n.], out. 2005. v. 2, p. 1089–1092. DOI: 10.1109/ICNNB.2005.1614807.

[6] FAZILUDEEN, S.; SABIQ, P. V. ECG beat classification using wavelets and SVM. In: 2013 IEEE Conference on Information Communication Technologies. [S.l.: s.n.], abr. 2013. p. 815–818. DOI: 10.1109/CICT.2013.6558206.