Inferência Estatística para Classificação de Sinais Cardíacos

Mikaela Baldasso[1]

Marcio Valk [2]

Resumo: Doenças cardiovasculares são responsáveis por milhões de mortes anualmente, segundo a Organização Mundial da Saúde e, dado isso, várias são as iniciativas, em todo o mundo, que visam estimular o desenvolvimento de novas técnicas que permitam diagnosticar e prevenir essas enfermidades. Diferentes técnicas de diagnósticos são utilizadas para detectar e prevenir esses desfechos, em que busca-se, principalmente, utilizar métodos não invasivos, baratos e que resultem em respostas rápidas e confiáveis, como por exemplo, aqueles baseados em Eletrocardiogramas e Fonocardiogramas. A partir disso, nosso objetivo nesse trabalho é utilizar a estatística para fazer inferência sobre classificação, ou seja, mensurar a confiabilidade de uma técnica de diagnóstico, em particular testar o método baseado em U-estatística para classificação e agrupamento de dados.

Palavras-chave: *Doenças Cardíacas, Classificação, Inferência.*

2 Introdução

As doenças cardiovasculares (DCV) continuam sendo a principal causa de morbidade e mortalidade no mundo todo, de acordo com Liu et al. (2016). Estima-se que 17,5 milhões de pessoas morreram de DCV em 2012, representando 31% de todas as mortes globais (OMS 2015). Um dos primeiros passos na avaliação do sistema cardiovascular é o exame físico: a auscultação dos sons do coração é parte essencial do exame e pode fornecer importantes pistas iniciais na avaliação da doença, servindo de guia para um exame diagnóstico posterior.

A análise automatizada do som cardíaco nas aplicações clínicas geralmente consiste em três passos; Pré-processamento, segmentação e classificação. Nas últimas décadas, métodos para segmentação automatizada e classificação de sons cardíacos foram amplamente estudados. Muitos métodos demonstraram potencial para detectar com precisão patologias em aplicações clínicas. Infelizmente, as comparações entre técnicas foram dificultadas pela falta de bases de dados de alta qualidade, rigorosamente validadas e padronizadas de sons cardíacos obtidos a partir de uma variedade de condições saudáveis e patológicas. Em muitos casos, ambos os dados experimentais

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *mikaelabaldasso@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *marcio.valk@ufrgs.br*

e clínicos são coletados a custos consideráveis, mas apenas analisados uma vez por seus colecionadores e, em seguida, arquivados indefinidamente por variados motivos, como mencionado em Liu et al. (2016).

Algoritmos baseados em aprendizado supervisionado são amplamente utilizados na classificação de dados, como *Support Vector Machine* (SVM, citeScholkopf2001, Scholkopf2002) ou *support vector data description* (SVDD, Tax e Duin (2004)). Outras abordagens concentram-se na estimativa de densidade paramétrica. Essas metodologias também podem ser aplicadas na detecção de novidades ou *outliers*, que são dois tópicos importantes em estatística e aprendizado de máquina, devido a sua relevância prática em cenários do mundo real. A detecção de novidade é a tarefa de classificar os dados que diferem em alguns aspectos dos dados usados durante o treinamento Pimentel et al. (2014). A detecção de anomalias, também chamada de análise outlier, é a tarefa de identificar dados que se desviam de algum comportamento esperado Chandola et al. (2009).

Com base nessa estrutura, Cybis et al. (2018) propõe um teste para avaliar a significância estatística no problema da classificação de um elemento. A abordagem baseada na U-estatística é apresentada e uma extensão de uma U-estatística de teste crucial é proposta. Para a utilização dessa técnica no contexto de séries temporais é necessário transformar os dados de alguma maneira. Para isso utilizamos o periodograma, que é uma estimativa da densidade espectral do sinal, ou seja, é uma medida que descreve como a força do sistema se comporta conforme a variação da frequência, que pode ser aplicado em análise e processamento dos eletrocardiogramas. Em termos gerais, uma maneira de estimar essa densidade espectral é encontrar a transformada de Fourier de tempo discreto das amostras do processo e apropriadamente e calcular a distância euclideana entre esses resultados.

3 Sinais cardíacos e seus padrões

Nosso objeto de estudo são sinais cardíacos provenientes de diferentes fontes; podem ser eletrocardiogramas (ECG's) ou fonocardiogramas (PCG's). Para esse trabalho, escolhemos alguns sinais do banco de dados *MIT-BIH Arrhythmia DataBase*, Goldberger et al. (2000), que foi disponibilizado, como material de teste padrão para avaliação de detectores de arritmia, em 1980. O conjunto contém 48 trechos de meia hora de registros obtidos de 47 indivíduos e as gravações foram digitalizadas com resolução de 11 bits em uma faixa de 10 mV. Na figura 3, apresentamos alguns sinais desse banco de dados. Existem 3 grupos de sinais: os sinais normais, sem qualquer tipo de anomalia; os sinais com algum tipo de arritmia considerada comum; e os sinais com arritmias não tão comuns. Dois sinais de cada grupo são apresentados no gráfico juntamente com

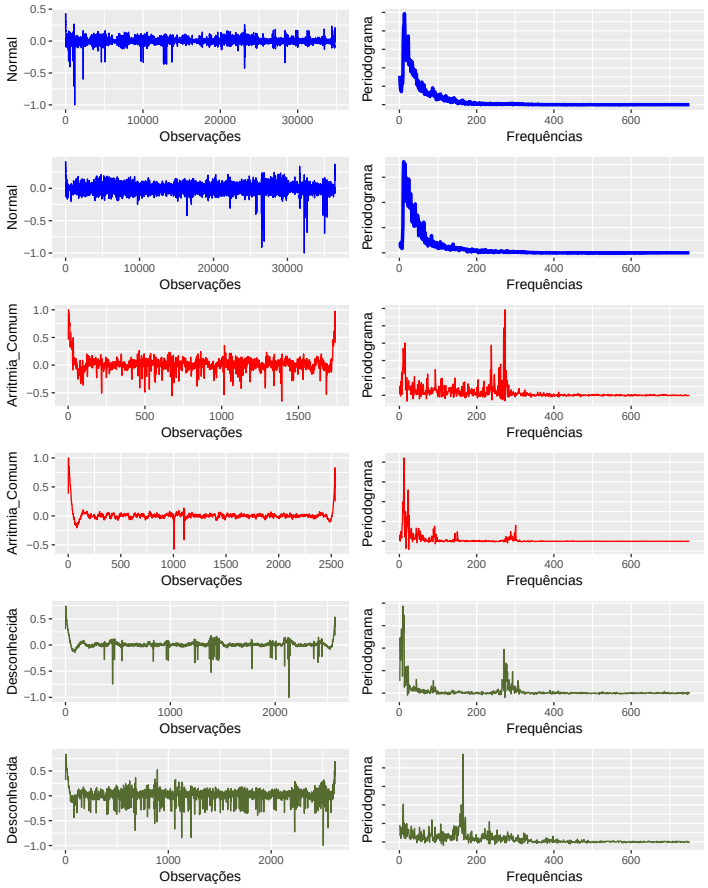

Figura 1: Sinais cardíacos com arritmia e sem arritmia com seus respectivos periodogramas que são transformações dos dados usadas na busca por padrões.

os seus respectivos periodogramas. Podemos observar que essa transformação dos dados captura padrões importantes. Isso se repete na maioria dos sinais observados. No grupo de sinais normais podemos observar um único pico no periodograma. Nos sinais com arritmias comuns podemos observar dois picos e nas arritmias não comuns o que prevalece é a não necessária existências de picos. É claro que há exceções a essa análise eurística e por isso a necessidade de um método estatístico para ajudar a decidir sabendo-se a probabilidade de errar.

Neste trabalho, temos por objetivo mensurar a confiabilidade do método baseado em u-estatísticas e, a partir disso, avaliar ECG's como séries temporais, em que a técnica de classificação e agrupamento pode ser aplicada.

O método de *clustering* é um conjunto de técnicas computacionais cujo propósito consiste em separar objetos em grupos distintos de acordo com as características que eles apresentam. De forma geral, a técnica consiste em colocar elementos similares em um mesmo grupo de acordo com algum critério já estipulado.

Uhclust - Método baseado em U-estatísticas

Dada uma amostra $X = (X_1, ..., X_n)$ de n vetores L-dimensionais dividida em dois grupos G_1 e G_2 de tamanhos n_1 e n_2 respectivamente onde $n = n_1 + n_2$. Sejam $X_1^{(g)}, ..., X_{n_g}^{(g)}$ as observações do g-ésimo grupo, independentes e com distribuição F_g. Define a distância funcional $\theta(F_1, F_2)$ por

$$\theta(F_1, F_2) = \int\int \phi(F_1, F_2) dF_1(x_1) dF_2(x_2)$$

onde $x_1, x_2 \in \mathbb{R}^L$.

Da teoria das U-estatísticas segue que um estimador não-viesado deste funcional para um mesmo grupo é uma estatística generalizada, com kernel $\phi(.,.)$ dada por

$$U_{n_g}^{(g)} = \binom{n_g}{2}^{-1} \sum_{1 \le i < j \le n_g} \phi(X_i^{(g)}, X_j^{(g)}).$$

Analogamente, o estimador para dois grupos diferentes é dado por

$$U_{n_1,n_2}^{(1,2)} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_i^{(1)}, X_j^{(2)}).$$

Note que a U-estatística pode ser decomposta por

$$\begin{aligned} U_n &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n_g} \phi(X_i, X_j) \\ &= \sum_{g=1}^{2} \frac{n_g}{n} U_{n_g}^{(g)} + \frac{n_1 n_2}{n(n-1)} (2U_{n_1 n_2}^{(1,2)} - U_{n_1}^{(1)} - U_{n_2}^{(2)}) \\ &= W_n + B_n. \end{aligned}$$

Assim, o teste, proposto por Cybis et al. (2018), consiste em verificar se G_1 e G_2 constituem grupos separados ou se derivam da mesma distribuição. Basicamente, quando os grupos derivam da mesma distribuição temos $F_1 = F_2$ e portanto $\mathbb{E}(B_n) = 0$, e quando os grupos diferem temos $\mathbb{E}(B_n) > 0$.

Para evitar maiores complicações computacionais, o problema resume-se em minimizar a função

$$f(G_1, G_2) = -\frac{B_n}{\sqrt{(Var(B_n))}},$$

que também caracteriza o menor p-valor que a configuração pode assumir. De certa forma, se

esse p-valor for menor que um certo nível de significância α então há uma certa "confiança" na conclusão a respeito da separabilidade dos grupos.

3.1 Extensão da estatística de teste para grupos de tamanho 1

Valk e Cybis (2018) propõe explorar o método de *clustering* apresentado em Cybis et al. (2018) para construir um algoritmo de detecção de outliers. Contudo, o método de *clustering* hierárquico não deve ser restrito a *clusters* com tamanhos $g_i \geq 2$. Essa restrição de tamanho de grupo é uma consequência da definição da B_n de um argumento de decomposibilidade de um subgrupo, resultando em somas ponderadas de distâncias entre e dentro de *clusters*.

Para construir um algoritmo de *clustering* que considere grupos de tamanho 1, é proposta uma extensão das estatísticas de teste B_n. Define-se

$$B_n = \begin{cases} \frac{n-1}{n(n-1)}(U^{(1,2)}_{1,n-1} - U^{(2)}_{n-1}) & \text{if } n_1 = 1, \\ \frac{n_1 n_2}{n(n-1)}(2U^{(1,2)}_{n_1 n_2} - U^{(1)}_{n_1} - U^{(2)}_{n_2}) & \text{if } 2 \leq n_1 \leq n-2, \\ \frac{n-1}{n(n-1)}(U^{(1,2)}_{1,n-1} - U^{(1)}_{n-1}) & \text{if } n_1 = n-1, \end{cases} \tag{1}$$

Primeiro notamos que a decomposição apresentada na expressão ainda é válida para o B_n estendida com um grupo de tamanho $2 \leq n_1 \leq n-2$, bem como a decomposição de Hoeffding e a teoria sobre convergência.

O método de Valk e Cybis (2018) está implementado no pacote *uhclust* o qual foi utilizado para as simulações. Cabe ressaltar que nenhuma abordagem a séries temporais foi proposta ainda utilizando esse método.

3.2 Simulações de Monte Carlo

Para verificar o desempenho do método de agrupamento *uclust* quando utilizado em um contexto de séries temporais, propomos um estudo de simulação em que os cenários são controlados. Nesse estudo, sabemos quem são os verdadeiros *clusters* e então podemos verificar a qualidade do método em encontrá-los e também a capacidade de detectar diferença entre os mesmos, quando ela existe.

Assim, utilizamos os processos autorregressivos de ordem 1 (AR(1)) para gerar os grupos. O processo é definido por $Y_t = \phi y_{t-1} + \varepsilon_t$, em que o parâmetro ϕ deve satisfazer $|\phi| < 1$ e ε_t é um ruído branco gaussiano.

Na Tabela 1, as $n_1 = 10$ séries temporais que compõem o grupo 1 (G1) são geradas com

$\phi = 0.3$ (conforme coluna do ϕ_1) e as $n_2 = 7$ séries que compõem o grupo 2 (G2) são geradas a partir de diferentes valores para ϕ (conforme a coluna do ϕ_2). Os resultados mostram a proporção de rejeição em 100 replicações de cada cenário, além de uma medida de "qualidade de cluster"(ARI) proposta por Rand (1971). Dessa forma, a partir do cálculo da ARI, comparamos a qualidade do nosso método com o método clássico de agrupamento hierárquico *hclust* "complete linkage", do pacote *stats* do R.

Sob a hipótese de homogeneidade de grupos, ou seja, que todos os componentes tenham mesma distribuição, que nesse contexto pode ser traduzido para mesmo processo gerador, espera-se que o método não encontre mais do que α% de rejeição, onde α é o nível de significância. Neste estudo, usamos $\alpha = 5\%$ e podemos observar que quando os parâmetros ϕ_1 e ϕ_2 são iguais a 0.3, a proporção de rejeição é muito próxima a 5%, o que indica que o método está bem "calibrado", não rejeitando mais do que α. A medida em que ϕ_1 se diferencia de ϕ_2, a proporção de rejeição aumenta, indicando que o método detecta dois grupos.

Além disso, é importante ressaltar que, quando $n_1 = 10$ e $n_2 = 7$, o ARI do método *uhclust* é melhor que o tradicional *hclust*. No entanto, em um segundo cenário em que $n_1 = 10$ e $n_2 = 1$, o ARI do método tradicional *hclust* é mais satisfatório, como mostra a Tabela 2.

$n_1 = 10$ e $n_2 = 7$				
ϕ_1	ϕ_2	Proporção de Rejeição	ARI *hclust*	ARI *uhclust*
0.30	−0.20	1.00	0.99	1.00
0.30	−0.10	1.00	0.77	0.99
0.30	0.00	0.97	0.46	0.88
0.30	0.10	0.24	0.11	0.42
0.30	0.20	0.05	0.02	0.06
0.30	0.30	0.04		
0.30	0.40	0.08	0.02	0.06
0.30	0.50	0.53	0.17	0.61
0.30	0.70	1.00	0.99	1

Tabela 1: Proporção de rejeição do *uhclust* e ARI do *uhclust* e *hclust*

$n_1 = 10$ e $n_2 = 1$				
ϕ_1	ϕ_2	Proporção de Rejeição	ARI *hclust*	ARI *uhclust*
0.30	−0.20	0.18	0.77	0.35
0.30	−0.10	0.1	0.54	0.20
0.30	0.00	0.08	0.26	0.05
0.30	0.10	0.07	0.17	0.04
0.30	0.20	0.03	0.02	0.01
0.30	0.30	0.06		
0.30	0.40	0.03	0.05	0.01
0.30	0.50	0.06	0.14	0.04
0.30	0.70	0.41	0.87	0.61

Tabela 2: Proporção de rejeição do *uhclust* e ARI do *uhclust* e *hclust*

4 Resultados

Durante a realização do presente trabalho, exploramos vários bancos de dados de diferentes fontes e características, e neles aplicamos diversas transformações na busca por padrões. Simulações de Monte Carlo foram realizadas em um contexto controlado e sugerem que o método *uhclust* pode ser usado para caracterizar sinais com dinâmicas diferentes desde que a métrica correta seja utilizada. Os próximos passos serão na direção da aplicação aos dados reais apresentados nesse trabalho.

Referências

Chandola, V., Banerjee, A., e Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15.

Cybis, G. B., Valk, M., e Lopes, S. R. (2018). Clustering and classification problems in genetics through u-statistics. *Journal of Statistical Computation and Simulation*, pages 1–21.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C.-K., Stanley, H., PhysioBank, PhysioToolkit, e PhysioNet (2000). Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):215.

Liu, C., Springer, D., Li, Q., Moody, B., Juan, R. A., Chorro, F. J., Castells, F., Roig, J. M., Silva, I., Johnson, A. E. W., Syed, Z., Schmidt, S. E., Papadaniil, C. D., Hadjileontiadis, L., Naseri, H., Moukadem, A., Dieterlen, A., Brandt, C., Tang, H., Samieinasab, M., Samieinasab, M. R., Sameni, R., Mark, R. G., e Clifford, G. D. (2016). An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12):2181.

Pimentel, M. A., Clifton, D. A., Clifton, L., e Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99(Supplement C):215 – 249.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Tax, D. M. e Duin, R. P. (2004). Support vector data description. *Machine learning*, 54(1):45–66.

Valk, M. e Cybis, G. B. (2018). U-statistical inference for hierarchical clustering. *arXiv preprint arXiv:1805.12179*.