Detecção de alterações respiratórias na fibrose cística com o uso de algoritmos de aprendizado de máquinas

Noemi P. Pinto[1], Jorge L. M. Amaral[2], Pedro L. Melo[3]

[1]Programa de Pós-Graduação em Engenharia Eletrônica – Universidade do Estado do Rio de Janeiro (UERJ)
20.550-013 – Rio de Janeiro – RJ – Brasil

[2]Departamento de Eletrônica e Telecomunicações – UERJ

[3]Instituto de Biologia – UERJ

{noemidpp}@gmail.com, jamaral@uerj.br, plopes@uerj.br

Abstract. *Advances in the treatment of cystic fibrosis have allowed patients to reach adulthood. As an alternative, the Forced Oscillations Technique (FOT) is being used in the respiratory system analysis and must prove its efficiency. Thus, this work proposes the use of machine learning algorithms to aid the investigation and diagnosis of respiratory changes in cystic fibrosis through the data provided by FOT. During the experiments, the used models presented an AUC value varying from 0.87 to 0.89, showing that the use of machine learning algorithms increased accuracy in diagnosis of respiratory changes in patients who suffer from cystic fibrosis.*

Resumo. *Avanços no tratamento da fibrose cística têm permitido que pacientes chegassem até a fase adulta. Como uma alternativa, a Técnica de Oscilações Forçadas (FOT) vem sendo usada na análise do sistema respiratório e precisa comprovar sua eficácia. Sendo assim, este trabalho propõe o uso de algoritmos de aprendizado de máquinas (AM) para auxiliar a investigação e diagnóstico de alterações respiratórias na fibrose cística através dos dados fornecidos pela FOT. Durante os experimentos realizados, os modelos usados apresentaram valores de AUC variando de 0,87 a 0,89, mostrando que o uso de algoritmos de AM aumentou a acurácia no diagnóstico de alterações respiratórias da fibrose cística.*

1. Introdução

A fibrose cística (FC) é uma doença genética que inicialmente era diagnosticada em recém-nascidos (ANDERSEN, 1938), sendo essas crianças levadas a óbito ainda no primeiro ano de vida. Porém, os últimos avanços apresentados no tratamento e diagnóstico da doença, fizeram com que esses pacientes chegassem à idade adulta (DALCIN et al., 2008). A espirometria é um dos principais métodos usados atualmente para o diagnóstico da FC, porém não caracteriza em detalhes o sistema respiratório e não permite identificar os processos das doenças. Esse fato tem sido uma grande motivação para aprimorar a identificação da FC (LIMA et al., 2010).

Dentre os métodos pesquisados, a Técnica de Oscilações Forçadas (FOT – *Forced Oscillation Technique*) vem sendo estudada para avaliar de forma mais detalhada as propriedades mecânicas do sistema respiratório, sendo uma técnica de fácil execução para o profissional (DUBOIS et al., 1956). Os parâmetros obtidos pela FOT, associados aos métodos de aprendizado de máquinas (AM), trouxeram importantes avanços no estudo de doenças respiratórias, mostrando que é possível realizar essa associação (AMARAL et al., 2013; AMARAL et al., 2015; AMARAL et al., 2017). Embora apresente elevado potencial, essa aplicação ainda não foi amplamente

investigada. Neste sentido, este artigo se insere na linha de pesquisa dos trabalhos citados, propondo o uso de algoritmos de AM e dos dados fornecidos pela FOT, para aprimorar a detecção de alterações respiratórias em portadores de FC.

2. Algoritmos de Aprendizado de Máquinas

Com o intuito de auxiliar a equipe médica no diagnóstico de alterações respiratórias na FC por meio da FOT, foi usada a técnica de AM. Os algoritmos descritos a seguir foram selecionados devido ao bom desempenho apresentado em trabalhos anteriores (AMARAL et al., 2013; AMARAL et al., 2015; AMARAL et al., 2017) e nos primeiros testes realizados neste estudo.

O *K*-NN é considerado um dos mais simples algoritmos de AM, com o aprendizado feito por instâncias e o conjunto de treinamento armazenado durante a aprendizagem (FACELI et al., 2011). As florestas aleatórias (*Random Forests*) são comitês de árvores de decisão e se baseiam em dois conceitos principais: seleção aleatória dos atributos de entrada e o *bagging* (*Bootstrap Aggregation*) (LIAW et al., 2002). Já o *Adaptative Boosting* (Adaboost) tem como objetivo criar um classificador forte (alta acurácia), através da combinação de classificadores fracos (baixa acurácia) (MARGINEANTU et al., 1997). No caso do algoritmo *Support Vector Machines* (SVM), o aprendizado é baseado na teoria de aprendizagem estatística, tendo como ideia principal a criação de um hiperplano como uma fronteira de classificação. Essa formulação pode ser generalizada pela aplicação de funções para o mapeamento dos dados em uma dimensão maior (FACELI et al., 2011).

3. Experimentos

O conjunto de dados usado foi obtido através de exames realizados por um sistema de Técnica de Oscilações Forçadas (FOT), desenvolvido no Laboratório de Instrumentação Biomédica da UERJ (LIMA et al., 2015). Os exames foram realizados em 23 indivíduos do grupo controle e 27 portadores de FC, que formam um grupo de teste. Em cada exame foram feitas três medidas, o que totalizou um conjunto de dados de 150 instâncias. Dos parâmetros fornecidos pela FOT, oito foram usados nesse trabalho: Resistência no Intercepto (R_o), Resistência Média (R_m), Reatância Média (X_m), Complacência Dinâmica (C_{din}), Inclinação da Curva de Resistência (S), Impedância em 4Hz (Z_{4Hz}), Frequência de Ressonância (F_r) e Elastância Dinâmica (E_{din}). A área sob a curva ROC (AUC) foi a medida de desempenho usada, por ser uma ferramenta normalmente aplicada em diagnósticos médicos (METZ, 1978), fornecendo informações sobre a eficácia de algoritmos de AM (HUANG et al, 2005).

Ao todo, foram realizados cinco experimentos com duas possíveis saídas nos classificadores: 0 (não portador de FC) ou 1 (portador de FC). Primeiramente, o desempenho de cada parâmetro da FOT foi avaliado separadamente na classificação da FC. No segundo experimento, os oito atributos fornecidos pela FOT foram submetidos aos quatro algoritmos de AM, implementados por meio da *toolbox Pattern Recognition* (prtools) do *software* Matlab. O algoritmo *K*-NN foi configurado com o valor de *K* igual a 1 (1NN). O número de classificadores fracos usado no Adaboost (ADAB) foi determinado igual a 200. Já no classificador *Random Forest* (RF), há dois parâmetros a serem definidos: o número de árvores construídas, configurado igual a 50, e o tamanho do subconjunto de atributos usado na construção dessas árvores, configurado igual a 1. Os parâmetros desses três modelos foram definidos com base em seus desempenhos ao serem apresentados a diferentes configurações. Já no caso do *Support Vector Machine*, a função Kernel de base radial foi usada para realizar o mapeamento do conjunto de treinamento (RSVM). Esse modelo possui dois parâmetros: desvio padrão da base radial e o parâmetro de regularização, cuja busca foi realizada por uma validação cruzada interna durante o treinamento. A técnica de validação cruzada por *k*-pastas foi usada para evitar o *overfitting* durante o treinamento dos modelos. Com esse método, os dados

são divididos em *k* pastas, sendo geralmente uma pasta usada para teste e *k*-1 pastas usadas para o aprendizado (HASTIE, 2008). O valor de *k* escolhido foi igual a 10.

No terceiro experimento, foram selecionados cinco dos oito atributos originais da FOT, com o intuito de escolher as características que melhor descrevessem o problema e melhorassem o desempenho dos algoritmos. Essa seleção foi realizada pelo método *Wrapper*, realizando a busca por um conjunto de atributos que maximize o valor da AUC média. No quarto e quinto experimentos, foi aplicado o produto cruzado nos oito parâmetros originais da FOT e nos cinco parâmetros selecionados no terceiro experimento, respectivamente. Em ambos os testes, o objetivo foi investigar se através dos produtos cruzados seriam observadas melhoras no desempenho dos algoritmos.

4. Resultados

Durante o primeiro experimento, todos os parâmetros tiveram suas medidas de AUC calculadas (DELONG et al., 1988). De acordo com a Figura 1, a reatância X_m, melhor parâmetro da FOT (MPF), e a frequência F_r apresentaram melhor desempenho individual. O desempenho de todos os atributos analisados separadamente se enquadra na faixa de acurácia moderada (0,70 a 0,90). No caso do segundo experimento, com os oito parâmetros da FOT, o algoritmo 1NN apresentou melhor desempenho (AUC=0,87).

No terceiro experimento, foram selecionados cinco parâmetros da FOT: R_o, R_m, X_m, C_{din} e Z_{4Hz}. Essa seleção também está de acordo com as variáveis escolhidas por um especialista. O algoritmo 1NN apresentou o mesmo desempenho do experimento anterior, com AUC igual a 0,87. Os demais resultados podem ser vistos na Figura 1.

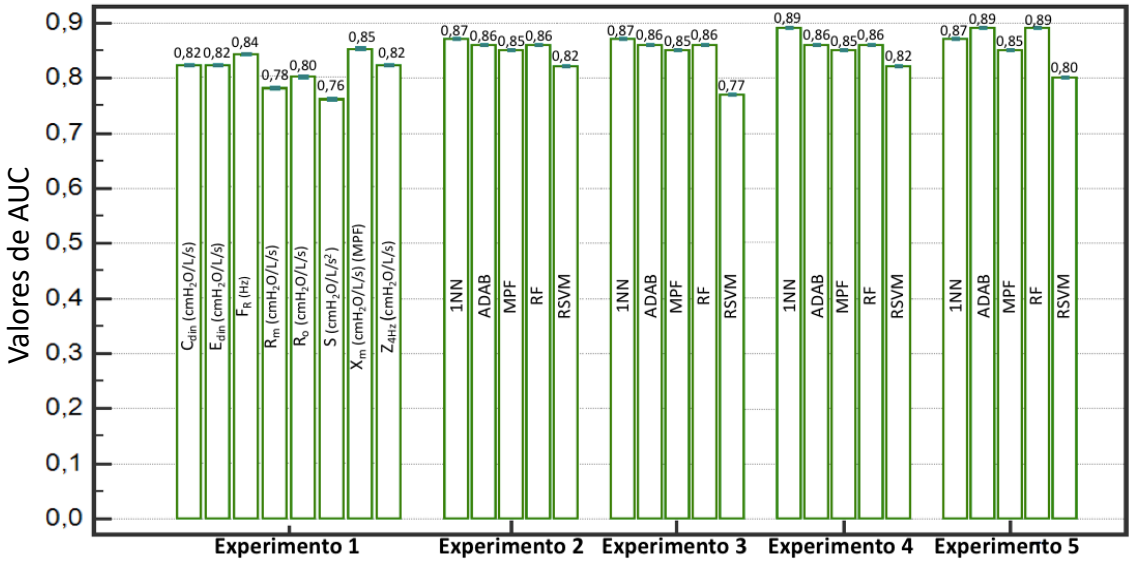

Figura 1. Valores de AUC obtidos em cada um dos cinco experimentos realizados

Já no quarto experimento, o produto cruzado dos oito parâmetros da FOT foi usado como entrada nos classificadores. No algoritmo 1NN houve um aumento no valor da AUC de 0,87 para 0,89, com relação aos experimentos anteriores. Os demais valores podem ser observados na Figura 1. No quinto e último experimento, o produto cruzado dos atributos selecionados no terceiro experimento, foram usados como entrada nos classificadores. Os algoritmos RF e ADAB apresentaram melhor desempenho, com valores de AUC iguais a 0,89. Os demais resultados estão na Figura 1.

5. Conclusão

Este trabalho mostrou que a aplicação dos dados fornecidos pela FOT em algoritmos de AM, são eficientes na identificação de portadores de FC. Em todos os

experimentos, pelo menos um desses classificadores apresentou valor de AUC maior do que o método atual, que consiste na classificação feita pelo MPF, neste caso, a reatância X_m (AUC=0,85). Dessa forma, houve maior acurácia na detecção de alterações respiratórias na FC.

Este trabalho foi desenvolvido em ambiente Matlab por se tratar de um protótipo. Entretanto, como melhoria futura, será feita a implementação dos algoritmos no *software* Python, devido à facilidade em usar um ambiente aberto e gratuito. Outra melhoria será o teste em outros algoritmos, com o intuito de melhorar o desempenho no diagnóstico da FC.

6. Referências

Amaral, J., Lopes, A., Faria, A. e Melo, P. (2015) "*Machine learning algorithms and forced oscillation measurements to categorise the airway obstruction severity in chronic obstructive pulmonary disease*", Em: *Computer Methods and Programs in Biomedicine*, Elsevier, Volume 118, páginas 186-197.

Amaral, J., Lopes, A., Jansen, J., Faria, A. e Melo, P. (2013) "*An improved method of early diagnosis of smoking-induced respiratory changes using machine learning algorithms*", Em: *Computer Methods and Programs in Biomedicine*, Elsevier, Volume 112, páginas 441-454.

Amaral, J., Lopes, A., Veiga, J., Faria, A. e Melo, P. (2017) "*High-accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements*", Em: *Computer Methods and Programs in Biomedicine*, Elsevier, Volume 144, páginas 113–125.

Andersen, D. (1938) "Cystic fibrosis of the pancreas and its relation to celiac disease clinical and pathologic study", Em: *American Journal of Diseases of Children*, Volume 56 (2), páginas 344-399.

Dalcin, P. e Silva, F. (2008) "Fibrose cística no adulto: aspectos diagnósticos e terapêuticos", Em: Jornal Brasileiro de Pneumologia, páginas 107-117.

Dubois, A., Brody, A., Lewis D. e Burgess, B. (1956) "*Oscillation mechanics of lungs and chest in man*", Em: *Journal of applied physiology*, Volume 8, páginas 587-594.

Faceli, K., Lorena, A., Gama, J. e Carvalho, A. (2011) "Inteligência Artificial: uma Abordagem de Aprendizado de Máquina", Editora LTC.

Hastie, T., Tibshirani, R. e Friedman, J. (2008) "*The Elements of Statistical Learning: Data Mining, Inference and Prediction*", Em: Springer.

Huang, J. e Ling C. (2005) "*Using AUC and Accuracy in Evaluating Learning Algorithms*", Em: *IEEE Transaction Knowledge and Data Engineering*, Volume 17, Número 3, páginas 299–310.

Liaw, A. e Wiener, M. (2002) "Classification and Regression by Random Forest", Em: R. News, Volume 2/3, páginas 18–22.

Lima, A., Faria, A. e Lopes, A. (2010) "Técnica de oscilações forcadas na avaliação funcional de pacientes com fibrose cística com idade superior a 18 anos", Em: Pulmão RJ.

Margineantu, D. e Dietterich, T. (1997) "*Pruning Adaptive Boosting, Machine Learning*", Em: *Proceedings of the Fourteenth International Conference*, páginas 211-218, 1997;

Metz, C. (1978) "*Basic Principles of ROC Analysis*", Em: *Seminars in Nuclear Medicine*, Volume 8, Número 4.