XXX Congresso de Iniciação Científica UNICAMP

Relatório Científico Resumido

Aprendizado de máquina para a predição da Doença de Alzheimer por meio de imagens de tensor de difusão

Aluna: Thais Maria Santos Bezerra – Faculdade de Ciências Médicas (FCM)

Orientadora: Dr.ª Liara Rizzi – Departamento de Neurologia (FCM)

Palavras-Chave: Doença de Alzheimer; Aprendizado de Máquina; DTI

Campinas/SP

2022

1. Introdução e Justificativa

A doença de Alzheimer (DA) é uma patologia crônica e neurodegenerativa, caracterizada pela perda progressiva da memória e de outras capacidades cognitivas, sendo a principal causa de demência associada à idade (1). No Brasil, uma recente meta-análise que incluiu sete estudos brasileiros encontrou uma prevalência combinada de demências de 14,3% em pacientes com mais de 65 anos, embora com significativa heterogeneidade amostral (2). Devido ao envelhecimento populacional, é uma enfermidade com perspectiva de crescimento. Logo, tendo em vista a diminuição das competências cognitivas, a ausência de cura e sua complexidade de evolução, a rápida implementação do manejo adequado é essencial (3).

Um dos principais desafios da prática clínica da DA é a capacidade de realizar diagnósticos precoces de maneira não invasiva e identificar corretamente os indivíduos que estão susceptíveis ao desenvolvimento da patologia. Essa distinção de indivíduos que se encontram em vias de desenvolvimento da DA daqueles que não estão é de extrema importância, pois, quanto mais cedo intervenções forem implementadas, maior será o êxito no prognóstico (4). Nesse contexto, a utilização de técnicas de aprendizado de máquina para diferenciação de indivíduos portadores da DA em relação a grupo controle utilizando DTI mostra-se promissora na literatura (4). Dito isso, há possibilidade dessa capacidade de diferenciação estender-se para a predição de indivíduos com CCL evoluírem para a DA.

Uma compreensão mais profunda dos muitos caminhos complexos que levam à demência permite que cada processo individual seja mais rapidamente identificado e direcionado. Este estudo justifica-se pela importância da identificação indivíduos com CCL com grande potencial de progredirem para a DA, a fim de permitir a implementação de propostas terapêuticas precoces e proporcionar maior participação do paciente nas decisões que envolvem seu tratamento.

2. Preparação e Análise de Dados.

Quatro parâmetros foram gerados a partir da técnica de Imagens por Tensor de Difusão ou DTI (do inglês, Diffusion Tensor. Imaging) no processamento das neuroimagens, a saber: anisotropia fracional (FA), difusividade axial (AD), difusividade radial (RD) e difusividade média (MD). Esses foram os parâmetros utilizados para relacionar o DTI com a capacidade diagnóstica da Doença de Alzheimer via aprendizado de máquina.

Inicialmente, fez-se a carga dos dados que contém os parâmetros numéricos do DTI no formato nativo do software SPSS para o Python 3.0. Em seguida, verificou-se a completude dos

dados de DTI, isto é, a presença e a frequência de valores faltantes. Pacientes sem os valores de DTI foram removidos do estudo.

Alguns sujeitos de pesquisa do banco de dados não possuíam todos os dados de neuroimagem coletados. Haja vista que a maior parte dos algoritmos de aprendizado de máquina não consegue lidar com entradas com dados faltantes, implementou-se algumas abordagens distintas, com base na literatura, para lidar com esse problema: remoção das colunas com valores faltantes ou inserções com diferentes estratégias. As seguintes estratégias de inserção foram testadas: inserção de constante, da mediana dos valores da coluna, maximização de expectativa e vizinhos mais próximos (kNN).

Desse modo, cinco bases de dados foram geradas, cada uma com seu respectivo método de gerenciamento de dados faltantes.

Após a preparação dos dados para inserção nos modelos, considerou-se a elevada dimensionalidade da matriz de entrada, em que o número de variáveis era superior ao número de sujeitos. Considerando o exposto, implementou-se um método de seleção de variáveis, um método de filtro, *Information Gain*, a fim de remover a possibilidade de ruído que a possível existência de atributos não relevantes possa inserir no modelo, comprometendo a sua acurácia.

No fim, um conjunto de 20 técnicas de aprendizado de máquina foram exploradas no estudo: *AdaBoost, Bagging Classifier, Extra Trees, Gradient Boosting, Random Forest, Gaussian Process Classifier, Logistic Regression, Passive Aggressive Classifier, Ridge Classifier, Perceptron, KNeighbors Classifier, Support Vector Classifier, Linear Support Vector Classifier, Bernoulli Naïve Bayes, Gaussian Naïve Bayes, Decision Trees, Extra Trees, Linear Discriminant Analysis* e *Quadratic Discriminant Analysis*.

1.6. Resultados e Conclusão

Na primeira etapa, treinou-se 20 modelos para cada um dos 5 conjuntos de dados produzidos com as 20 técnicas clássicas oriundas do Aprendizado de Máquina, no início, sem considerar a redução de dimensionalidade. Os resultados mais relevantes estão apresentados na **Tabela 1** abaixo. Os valores correspondem à média da acurácia do algoritmo correspondente nos dados separados para teste na validação cruzada com os 5 conjuntos de dados distintos:

Tabela 1. Acurácias obtidas nos casos teste nos modelos que visam distinguir entre indivíduos que irão progredir ou não para DA

Algoritmo	DTI (Nenhum) %	DTI (Constante) %	DTI (EM) %	DTI (kNN) %	DTI (Mediana) %
GradientBoostingClassifier	80,64	**83,97**	**83,97**	82,31	82,31

BaggingClassifier	79,23	82,44	82,44	**83,97**	82,44
AdaBoostClassifier	80,64	80,64	82,31	**85,51**	**83,97**

Nesse estágio, é importante ressaltar que, como temos uma amostra de dados desbalanceada, na qual 83,85% dos dados são de pacientes que não irão progredir para a doença de Alzheimer, consideram-se relevantes os modelos capazes de superar a escolha arbitrária de inserir todos os dados na classe mais frequente, destacados em negrito.

Na segunda etapa, os conjuntos de dados que apresentaram pelo menos um resultado relevante, ou seja, todos os conjuntos com exceção do conjunto de dados sem estratégia de preenchimento, passaram por um filtro de variáveis, na qual se removeu variáveis com base nas relevâncias atribuídas a elas pelo cálculo do ganho de informação (*information gain*), conforme Tabela 2, que apresenta os resultados relevantes:

Tabela 2. Acurácias obtidas nos casos teste nos modelos com variáveis filtradas que visam distinguir entre indivíduos que irão progredir ou não para DA

Algoritmo	DTI reduzido (Constante) %	DTI reduzido (EM) %	DTI reduzido (kNN) %	DTI reduzido (Mediana) %
GradientBoostingClassifier	79,23	80,77	77,31	69,23
KNeighborsClassifier	82,18	**85,38**	**85,38**	**85,38**
BaggingClassifier	83,85	79,10	**83,97**	**86,92**
ExtraTreesClassifier	80,77	**85,38**	**85,51**	80,77
AdaBoostClassifier	**83,97**	80,51	**83,97**	83,85

Como pode ser visto nos resultados destacados da tabela 2, os algoritmos com resultados satisfatórios foram: *Adaboost Classifier*, *Extra Trees*, *Bagging Classifier* e kNN *Classifier*. Considerando a similaridade metodológica com o algoritmo *Extra Trees*, o algoritmo *Random Forest Classifier* também foi considerando para a próxima fase. Com base nisso, construiu-se uma matriz de hiperparâmetros para cada um desses algoritmos e testou-se a performance das diferentes combinações de hiperparâmetros utilizando a ferramenta *GridSearchCV*, com o intuito de determinar qual a melhor configuração de cada modelo testado e selecionado que maximiza a acurácia.

Por fim, obteve-se que, no espectro testado apenas com os dados de DTI, a opção que ofereceu maior acurácia foi o algoritmo *Random Forest*, com 88,72% de acurácia final utilizando o conjunto de dados preenchidos por kNN com dimensões reduzidas. Pelo modelo matemático utilizado na técnica de aprendizado de máquina, esta melhor acurácia indica uma maior compreensão da relação entre os parâmetros de DTI e o diagnóstico da Doença de Alzheimer.

Em seguida, avaliou-se a participação das diferentes áreas do cérebro a partir do recurso *feature importance* embutido no algoritmo, a fim de determinar quais as regiões e parâmetros de DTI considerados mais importantes para classificação, obtendo como mais relevantes a difusividade média (MD) do fascículo fronto-occipital superior e a anisotropia fracional (FA) do Fórnix. Esse resultado também colabora para a interpretação da doença do Alzheimer como uma patologia que afeta múltiplas regiões do cérebro ao invés de apenas uma.

Por fim, utilizando o algoritmo e o conjunto de dados selecionados, incluiu-se os demais tipos de dados, incluindo as variáveis neuropsicológicas, obtendo uma acurácia de 90,26% mesmo após otimização de hiperparâmetros. Comparado ao resultado anterior, pode-se observar que o ganho não foi muito superior ao valor de acurácia obtido anteriormente, sugerindo forte associação dos parâmetros de DTI com a capacidade diagnóstica precoce da Doença de Alzheimer.

Portanto, é possível afirmar que técnicas de aprendizado de máquina, utilizando parâmetros de DTI, que é uma técnica de neuroimagem não-invasiva, são capazes de predizer, com boa acurácia, se o paciente pode ou não evoluir para Doença de Alzheimer antes mesmo do diagnóstico clínico, permitindo ao paciente maior controle sobre a tomada de decisão acerca do curso do seu tratamento antes que a progressão da doença comprometa ainda mais sua autonomia e permita o estudo e a intervenção precoces na doença.

7. Referências

1. Sereniki A, Vital MABF. A doença de Alzheimer: aspectos fisiopatológicos e farmacológicos. Revista de Psiquiatria do Rio Grande do Sul. 2008;30:0-.
2. Farina N, Ibnidris A, Alladi S, Comas-Herrera A, Albanese E, Docrat S, et al. A systematic review and meta-analysis of dementia prevalence in seven developing countries: A STRiDE project. Glob Public Health. 2020;15(12):1878-93.
3. Atri A. The Alzheimer's Disease Clinical Spectrum: Diagnosis and Management. Med Clin North Am. 2019;103(2):263-93.
4. Billeci L, Badolato A, Bachi L, Tonacci A. Machine Learning for the Classification of Alzheimer's Disease and Its Prodromal Stage Using Brain Diffusion Tensor Imaging Data: A Systematic Review. Processes. 2020;8(9):1071.