Análise de Características a Partir do Classificador MLP Para Auxílio no Diagnóstico da COVID-19

Rhyan Ximenes de Brito[1], **Adonias Caetano de Oliveira**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Av. Tabelião Luiz Nogueira de Lima S/N – Tianguá – CE – Brazil

{rxbrito,adonia.ifce}@gmail.com

Abstract. *Artificial Intelligence (AI) has gained prominence in the most varied areas of science. From this perspective, it is known that COVID-19 is a disease that can cause everything from common colds to more serious diseases such as the Severe Acute Respiratory Syndrome (SARS). This work aims to perform a characteristic analysis based on the Multilayer Perceptron (MLP) neural network using feature selection techniques for the pre-diagnosis of COVID-19 disease. The methodology was based on features selection techniques and the use of a public database. The results were expressive considering that the feature selection chose attributes according to the metric of the MLP neural network, emphasizing those of greater relevance for accuracy.*

Resumo. *A Inteligência Artificial (IA) tem ganhado destaque nas mais variadas áreas da ciência. Nessa perspectiva sabe-se que a COVID-19 é uma doença que pode causar desde resfriados comuns a doenças mais graves como a Sindrome Aguda Respiratória Severa (SARS). Este trabalho objetiva fazer uma análise de características baseada na rede neural MultiLayer Perceptron (MLP) utilizando técnicas de feature selection para o pré-diagnóstico da doença COVID-19. A metodologia foi baseada em técnicas de features selection e na utilização de uma base de dados pública. Os resultados foram expressivos tendo em vista que o feature selection escolheu atributos de acordo com a métrica da rede neural MLP, enfatizando os de maior relevância para a acurácia.*

1. Introdução

É cada vez mais frequente o uso de técnicas com Inteligência Artificial (IA) na busca por respostas nas mais diversas áreas do conhecimento. Este é o caso da área de saúde que tem sido beneficiada com várias técnicas, como recursos que auxiliam na compreensão ou mesmo no pré-diagnóstico para encaminhamento de exames complementares a posteriori ou forma de tratamento mais adequada.

Atualmente a pandemia global de COVID-19 está relacionada a uma doença respiratória aguda causada pelo novo coronavírus (SARS-CoV-2), altamente contagioso e de evolução ainda pouco conhecida. Considerando-se a atual definição de caso baseada no diagnóstico de pneumonia, milhões de infecção por COVID-19 foram confirmados em todo o mundo com taxa de mortalidade associada oscilado em torno de 2% [Araujo-Filho et al. 2020].

Nessa perspctiva este trabalho teve como objetivo desenvolver uma solução utilizando técnicas de IA como a rede neural *MultiLayer Perceptron* (MLP) e seleção de

atributos como *Sequential Forward Selection* (SFS) que auxialiasse à triagem de sintomas de COVID-19, levando em consideração uma série de fatores químicos e biológicos do paciente com base nos critérios estabelecidos pela Organização Mundial de Saúde (OMS). Assim o trabalho pretende-se mensurar o quão eficiente pode ser a utilização de ferramentas computacionais inteligentes como apoio no processo de triagem de pacientes em situação de risco.

O presente artigo está organizado em mais cinco seções. A seção 2 apresenta os trabalhos relacionados. A fundamentação teórica acerca da COVID-19 e das técnicas de IA utilizadas no trabalho são abordadas pela seção 3. Na seção 4 é apresentada a metodologia utilizada no trabalho, enquanto que os resultados são analisados na seção 5. Por fim, as principais conclusões e trabalhos futuros são descritos na seção 6.

2. Trabalhos Relacionados

Esta seção resume alguns trabalhos científicos que aplicam técnicas de Inteligência Artificial em problemas de saúde.

[Fonseca 2019] propos um modelo de RNA para auxiliar no diagnóstico de transtorno bipolar, da depressão maior e da esquizofrenia, utilizando biomarcadores e características simples da população amostrada. O método de análise para o primeiro artigo é o treinamento de RNA aplicada à um banco de dados de distribuição livre da *Stanley Neuropathology Consortium*, consistindo de biomarcadores inflamatórios e características da população com diagnósticos de esquizofrenia, transtorno bipolar e um grupo controle (semtranstornos); assim como outro banco de dados, com variáveis bioquímicas, características da população e respostas de questionários com diagnósticos de depressão maior, transtorno bipolar e um grupo controle (sem transtornos). O programa de treinamento da RNA utilizado foi o OpenNN de distribuição livre. Como resultado tem-se RNAs treinadas com mais de 80% de acurácia nas classificações dos diagnósticos.

[de Brito et al. 2019a] implementaram uma rede *MultiLayer Perceptron*, objetivando usá-la como auxílio na identificação de pessoas com ou sem problemas cardíacos, com ênfase no treinamento e teste para classificação desses indivíduos. A metodologia foi implementada com base em dez treinamentos utilizando dados balanceados e normalizados com 270 amostras e 14 atributos. Os resultados foram analisados estatísticamente por meio de percentuais de acertos e erros da rede implementada, obtendo-se uma medida da qualidade atingida.

[Rocha et al. 2020] propôs um modelo de previsão de curto prazo baseado em Análise dos Componentes Principais (PCA) e Redes Neurais Artificiais (RNAs), capaz de estimar o número de casos e de óbitos causados pelo SARS-CoV-2. O modelo adotado foi baseado em dados, onde toda inferência foi feita a partir do conhecimento descoberto. O estudo apresentou resultados para o estado do Pará e Brasil, com séries temporais como base para analisar o impacto da capacidade de atendimento nos leitos de Unidade de Terapia Intensiva (UTI), servindo de suporte à tomada de decisões por parte dos órgãos de vigilância em saúde.

[Artoni et al.] através de um banco de dados contendo amostras de aplicações do teste AQ-10, em adultos, construíram uma aplicação com algoritmos de aprendizado de máquina usando técnicas de classificação para demonstrar possíveis soluções e alternativas para realizar ou auxiliar no diagnóstico do Transtorno do Espectro Autista (TEA).

[Frota et al. 2020] propuseram a utilização de um banco de dados público sobre TEA e as características mais relevantes apresentadas por um paciente. Assim, analisaram com os algoritmos Árvore de Decisão (AD), *Suport Vector Machine* (SVM), *MultiLayer Perceptron* (MLP) e *K–Nearest Neighbors* (KNN) na construção de um modelo capaz de simplificar uma estratégia de decisão para ajudar no diagnóstico desse tipo de distúrbio.

[de Brito et al. 2019b] realizaram um estudo com a implementação e análise das redes neurais *MultiLayer Percepron* (MLP) e *Radial Basis Function* (RBF), cujo objetivo era comparar resultados baseados no treinamento, teste e classificação de crianças com ou sem autismo. Para isso foi realizada uma validação cruzada dividida em 10 partes (*k-Fold*) a partir de 292 amostras de um banco de dados público. Os resultados foram analisados considerando as características e os comportamentos diferentes das redes implementadas, obtendo-se uma medida da qualidade atingida.

[Frota et al. 2019] utilizando o algoritmo de Árvore de Decisão (AD) construíram um modelo capaz de simplificar um conjunto complexo de decisões e produzir uma estratégia a partir de uma base de dados pública e internacional sobre o Transtorno do Espectro Autista, permitindo desenvolver um complemento no seu diagnóstico e possıvel tratamento.

3. Fundamentação Teórica

3.1. COVID-19

A COVID-19 foi detectada em Wuhan, China, em dezembro de 2019. Com o crescimento no número de casos, óbitos e países afetados, a OMS declarou que o evento constituía-se uma Emergência de Saúde Pública de Importância Internacional (ESPII), em 30 de janeiro de 2020. No Brasil, a epidemia foi declarada Emergência em Saúde Pública de Importância Nacional (ESPIN) em 3 de fevereiro de 2020 [Garcia and Duarte 2020].

O período de incubação do SARS-CoV-2 parece ser de quatro a sete dias. Há quadros clínicos variados relacionados à COVID-19, desde infecção assintomática até insuficiência respiratória grave. Os principais sintomas relatados são febre, mialgia, fadiga, tosse seca e dispneia. Sintomas incomuns, mas que também foram relatados, incluem escarro purulento, cefaleia, hemoptise e diarreia [Quintão et al. 2020].

Muitos fatores podem afetar a rapidez com que práticas eficazes de controle de doenças são implementadas, como campanhas de informação, práticas locais de saúde, comportamento social e sistema de crenças. A transmissão de pessoa para pessoa ocorre principalmente pelo contato direto ou por gotículas espalhadas pela tosse ou espirro de um indivíduo infectado. Sendo assim, o combate à disseminação da COVID-19 preconiza lavar as mãos frequentemente, evitar abraços, beijos e apertos de mãos e adotar medidas de afastamento social, como quarentena [Lima et al. 2020].

3.2. *MultiLayer Perceptron* (MLP)

A rede neural MLP consiste de um conjunto de unidades sensoriais (neurônios de fonte) que compõe a camada de entrada, uma ou mais camadas ocultas de neurônios computacionais e uma camada de saída de neurônios computacionais. O sinal de entrada se propaga para frente, camada por camada através da rede [Gonçalves et al. 2010].

Para [Bonifácio 2010] uma rede MLP consiste de uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. É do tipo *feedforward*, ou seja, nenhuma saída de um neurônio de uma camada k será sinal de entrada para um neurônio de uma camada menor ou igual a k, e é completamente conectada, tal que cada neurônio fornece sua saída para cada unidade da camada seguinte.

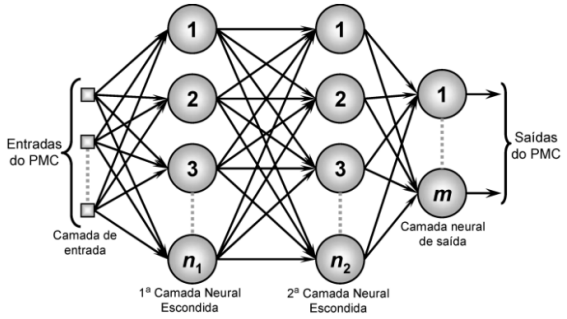

Figura 1. Arquitetura da Rede Neural MLP

3.3. *Sequential Feature Selection*

Para [Torres 2010] os algoritmos de seleção sequencial avaliam de forma independente todas as características de um conjunto para depois selecionar as melhores usando um critério personalizado.

O método *sequential feature selection* possui duas variantes: (i) seleção sequencial direta, na qual os recursos são adicionados sequencialmente a um conjunto candidato vazio até que a adição de outros recursos não diminua o critério; e o (ii) seleção sequencial para trás, onde os recursos são removidos sequencialmente de um conjunto completo de candidatos [MathWorks 2020].

De acordo com [Torres 2010], seja $X_n^2 \{x_1^2, x_2^2, ..., x_n^2\}$ um subconjunto de n características do conjunto $X_m^1 \{x_1^1, x_2^1, ..., x_m^1\}$, a melhor característica x_i^2 do subconjunto X_n^2 é definido por

$$x_i^2 = argmin_{x_j^2 \in X_n^2}(J(X_n^2 \backslash \left\{x_j^2\right\})) \tag{1}$$

A pior característica do conjunto X_n^2 é definida por

$$x_i^2 = argmax_{x_j^2 \in X_n^2}(J(X_n^2 \backslash \left\{x_j^2\right\})) \tag{2}$$

A melhor característica x_i^2 em relação ao subconjunto X_n^2 pode ser definida como

$$x_i^2 = argmax_{x_j^2 \in Y \backslash X_n^2}(J(X_n^2 \cup \left\{x_j^2\right\})) \tag{3}$$

A pior característica x_n^2 em relação ao subconjunto X_n^2 pode ser definida por

$$x_i^2 = argmin_{x_j^2 \in Y \backslash x_n^2}(J(X_n^2 \cup \left\{x_j^2\right\})) \tag{4}$$

3.3.1. *Sequential Forward Selection* (SFS)

A ideia principal por traz do algoritmo SFS é a adição de características sequencialmente a partir de um subconjunto inicial. A cada adição é realizada uma chamada a uma função critério que realiza a avaliação através da Equação 3. Pode-se dizer que a função critério mensura o conjunto de características [Torres 2010].

Assim o algoritmo precisamente parte de um conjunto vazio e vai adicionando as melhores características em relação ao subconjunto obtido na etapa anterior. Deve-se salientar que o SFS é usado quando se deseja selecionar poucas características [Torres 2010].

4. Metodologia

A metodologia desse trabalho foi desenvolvida em etapas: (1) pré-processamento dos dados; (2) treinamento e; (3) teste com a base de dados.

4.1. Base de Dados Utilizada

A base foi implementada a partir de um banco de dados público obtido no site *kaggle* através do link: https://bit.ly/2YjEqfa. A base de dados é composta 86 atributos e 395 amostras divididas entre 2 classes (0 - negativo, 1 positivo) para o teste de COVID-19. Onde 206 amostras são de pacientes que testaram negativo para COVID-19 e 189 são de pacientes que testaram positivo. Ressalta-se que utilizou-se validação cruzada *k-fold* com k=10, normalização (*zscore*) e balanceamento dos dados.

4.2. Treinamento e Teste

Para a escolha da configuraçãao da rede MLP embasou-se no desempenho entre várias configurações testadas. Assim utilizou-se para o treino e teste com a rede neural MLP a seguinte configuração, uma camada de entrada composta por 85 neurônios, uma camada oculta com 43 neurônios, uma de saída com 2 neurônios e para o treino utilizou-se de 6.000 *epochs*.

5. Resultados e Discussões

Nesta seção são feitas discussões sobre os resultados adquiridos com base na rede neural MLP e no *feature selection Sequential Forward Selection* (SFS).

A Tabela 1 mostra os resultados adquiridos com a base de dados normalizada (*zscore*) percebendo-se que o *fold* 3 obteve uma taxa de acerto de 96,91% e 3,09% de erro. Já o *fold* 8 teve o pior desempenho com 78,80% de acerto e 21,20% de erro. Ressalta-se que a taxa média de acerto foi de 87,97% e a de erro 12,03%.

Como observado na Tabela 2 a rede neural MLP juntamente como o *feature selection* SFS selecionaram os atributos, paciente internado em enfermaria, hemoglobina, leucócitos, teste rápido de influenza A, quantil da idade do paciente e saturação de oxigênio, como os que contribuiram para o melhor resultado com 89,17% de acerto. O algoritmo testou várias combinações de atributos nos quais observou que os acima mencionados obtiveram a melhor taxa de acerto.

Logo abaixo na Tabela 2 tem-se o resultado da MLP com base na utilização do *feature selection* SFS.

Tabela 1. Resultados dos Testes (MLP com Dados Nornalizados)

K-fold	Taxa de Acerto Geral (%)	Taxa de Erro Geral (%)	Situação
1º	91,90%	8,10%	—
2º	89,90%	19,10%	—
3º	96,91%	3,09%	Melhor Caso
4º	93,80%	6,20%	—
5º	82,90%	17,10%	—
6º	92,90%	7,10%	—
7º	80,91%	19,09%	—
8º	78,80%	21,20%	Pior Caso
9º	80,40%	19,60%	—
10º	91,30%	8,70%	—
Taxa Média	87,97%	12,03%	Médio Caso

Tabela 2. Resultados dos Testes (MLP normalizada com SFS)

Passo Realizado	Atributo Adicionado	Descrição do Atributo
1	2	Paciente Internado em Enfermaria
2	6	Hemoglobina
3	12	Leucócitos
4	37	Influenza A, teste rápido
5	1	Quantil da Idade do paciente
6	80	Saturação de oxigênio

6. Conclusão e Trabalhos Futuros

Como pode ser comprovado, o artigo relatou a utilização da rede neural artificial MLP juntamente com a técnica de seleção de atributos SFS, tendo como finalidade selecionar os melhores e mais importantes atributos de forma a contribuirem para as melhores acurácias.

Foi constatado que a rede MLP com normalização *zscore* obteve 87,97% de acerto e 89,17% de acerto quando aplicada juntamente com o SFS, de forma que se atribui a esse resultado aos atributos selecionados: paciente internado em enfermaria, hemoglobina, leucócitos, influenza A, teste rápido, quantil da idade do paciente e saturação de oxigênio.

Sugere-se como trabalho futuro a implementação de outras bases de dados relacionadas a COVID-19 com o classificador *Naive Bayes* em seguida fazer uma comparação e análise dos resultados encontrados.

Referências

Araujo-Filho, J. d. A. B., Sawamura, M. V. Y., Costa, A. N., Cerri, G. G., and Nomura, C. H. (2020). Pneumonia por covid-19: qual o papel da imagem no diagnóstico? *Jornal Brasileiro de Pneumologia*, 46(2).

Artoni, A. A., Prece, B., Scaranti, G., Junior, S. B., and de Barbosa, C. R. Aplicação de aprendizado de máquina para auxílio no diagnóstico do transtorno do espectro autista em adultos.

Bonifácio, F. N. (2010). Comparação entre as redes neurais artificiais mlp, rbf e lvq na classificação de dados. *Paraná: Universidade Estadual do Oeste do Paraná*.

de Brito, R. X., de Sousa Ximenes, J. N., da Silva, P. H. A., and de Sousa, R. N. (2019a). Sistema de análise de dados através de uma rede neural artificial mlp na prediç ao de doença cardıaca. *ANAIS ELETRÔNICOS CAIS TECH 2019*, page 102.

de Brito, R. X., Fernandes, C. A. R., and Amora, M. A. B. (2019b). Análise de desempenho com redes neurais artificiais, arquiteturas mlp e rbf para um problema de classificação de crianças com autismo. *iSys-Revista Brasileira de Sistemas de Informação*, 13(1):60–76.

Fonseca, M. B. (2019). *ClassificaÇão do Transtorno Bipolar, Esquizofrenia e Depressão Utilizando Redes Neurais Artificiais*. PhD thesis, Unviversidade Católica de Pelotas.

Frota, M., Hericles, S., Aguiar, G., Renoir, P., Nunes, R., Vilela, M., Gomes, D., and Paula Jr, I. C. (2020). Análise de características a partir de algoritmos de aprendizagem de máquina para auxílio ao diagnóstico do transtorno do espectro autista. *Revista de Sistemas e Computação-RSC*, 10(1).

Frota, M., Vilela, M., Hericles, S., Aguiar, G., Renoir, P., Nunes, R., Gomes, D., and Cavalcante, I. (2019). Aplicação de Árvore de decisão para auxılio ao diagnóstico do transtorno do espectro autista. In *Anais da VII Escola Regional de Computação Aplicada à Saúde*, pages 205–210, Porto Alegre, RS, Brasil. SBC.

Garcia, L. P. and Duarte, E. (2020). Intervenções não farmacológicas para o enfrentamento à epidemia da covid-19 no brasil.

Gonçalves, R. M., Coelho, L. d. S., Krueger, C. P., and Heck, B. (2010). Modelagem preditiva de linha de costa utilizando redes neurais artificiais. *Boletim de Ciências Geodésicas*, 16(3):420–444.

Lima, D. L. F., Dias, A. A., Rabelo, R. S., Cruz, I. D. d., Costa, S. C., Nigri, F. M. N., and Neri, J. R. (2020). Covid-19 no estado do ceará, brasil: comportamentos e crenças na chegada da pandemia. *Ciência & Saúde Coletiva*, 25:1575–1586.

MathWorks (2020). Sequential feature selection. [Online; acessado em: 07-Junho-2020].

Quintão, V. C., Simões, C. M., Navarro, L. H., Barros, G., Salgado-Filho, M. F., Guimarães, G. M. N., and Carmona, M. (2020). O anestesiologista e a covid-19. *Rev Bras Anestesiol*.

Rocha, J. E. C. d., Souza Júnior, G. N. d., Brito, S. R. d., Folador, A. R. C., RAMOS, R. T. J., Braga, M. d. B., Botelho, M. d. N., et al. (2020). Redes neurais artificiais na previsão de contágio e óbitos por covid-19: um estudo no estado do pará, brasil.

Torres, L. C. B. (2010). Seleção de características em expressões gênicas utilizando estratégias de busca e regressão logística. *Proceedings Seminário Interno da disciplina de Técnicas Clássicas de Reconhecimento de Padrões*, page 156.