Previsão de estágio da infecção por COVID-19

MAC0425 Inteligencia artificial - EP04

Carolina Senra Marques - 10737101
Felipe Castro de Noronha - 10737032

Resumo

Neste trabalho exploramos o uso de redes neurais para a previsão de resultados de exames acerca da COVID-19. Atualmente, três principais exames são usados para detectar o virus no corpo de um paciente, são eles: PCR, que detecta a presença do material genético do virus no organismo; igM, que detecta se o organismo teve contato com o virus recentemente; igG, que detecta se o organismo teve contato com o virus previamente[1]. Utilizando 372 exames como *features* treinamos 3 redes neurais que prevem o resultado de cada um dos exames acima.

Introdução

A seguir, falaremos da motivação e objetivos deste trabalho, assim como a estrutura deste relatório e dos arquivos.

Motivação

As redes neurais neurais estão se tornando cada vez mais populares, sendo utilizadas nos mais variados campos, desde visão computacional[2] até a astronomia[3]. Por que não tentar utilizar essa tecnologia para ajudar na questão de saúde publica que vivemos atualmente?

Alem disso, esta tarefa se mostra uma excelente oportunidade para colocarmos em pratica os conhecimentos obtidos na disciplina MAC0425 - Inteligencia artificial.

Objetivos

Esperamos criar uma rede de acurácia minimamente satisfatória e que possa ser utilizada para ajudar na previsão e diagnósticos epidemiológicos.

Estrutura deste relatório

O restante do relatório é dividido em 4 partes, são elas:

- Metodologia, onde explicamos como foi realizado o pré-processamento dos dados, a arquitetura da rede neural e a descrição dos experimentos realizados para decisão do modelo.
- Resultados, onde apresentamos a qualidade obtida pela nossa rede.
- Discussão, onde expomos os possíveis usos da rede.
- Bibliografia, fontes das citações aqui presentes.

Estrutura dos diretórios

O diretório principal deste trabalho contem 3 subdiretórios, são eles:

- `NN/`: Contem programas relacionados exclusivamente ao treinamento, uso e avaliação da rede neural. Além disso, a rede gerada fica armazenada em `/trained_nn/`.
- `PRE/`: Contem as planilhas brutas de dados, os scripts responsáveis por realizar a preparação inicial ficam em `/scripts/`.
- `REL/`: Contem o relatório.

Metodologia

Aqui descrevemos o pré-processamento dos dados, a arquitetura da rede neural e fazemos a descrição dos experimentos.

Pré-processamento

~max 2 paginas

Para o treinamento de uma boa rede neural é necessário um bom numero de dados consistentes. Por isso, utilizamos os *datasets* disponibilizados pelo Grupo Fleury[4] (mais de 120000 pacientes) e pelo Hospital Sírio-Libanês[5] (mais 2700 pacientes). Ambos *datasets* se resumem a dois arquivos, planilhas que contem informações dos pacientes e dos resultados de vários exames que os pacientes realizam.

Primeiramente corrigimos os caracteres que estavam errados por causa da conversão de formatação, usando o *script* `fix_utf.py` e um pouco da ajuda de nosso editor de texto. Com o *script* `get_unique.py` pegamos todas as entradas únicas das colunas *DE_ANALITO* da tabela de exames do Fleury e HSL e que estão presentes em ambas tabelas, a lista completa destes exames utilizados pode ser vista em `used_analitos.csv`. Em seguida tivemos que refinar esta lista de exames usados, retirando da lista exames que tivesses resultados de difícil mapeamento numérico e também todos os exames como frequência de resultado menores que 1000.

O *script* `generate.py` é responsável por criar o *dataset* final a ser utilizado. Nele é criado um dicionario que aceita como chave os *ids* relacionados a cada paciente, assim, cada entrada retorna uma lista que corresponde aos exames realizados por um dado paciente. Para lidar com o fato de que um paciente pode realizar o mesmo exame diversas vezes em um dado período de tempo usamos o conceito de *batch*, em particular, processamos todos os exames realizados em um certo dia, atualizando as respectivas colunas de um dado paciente, finalmente, apos o processamento de um dia, todos os pacientes que tiveram algum exame modificado tem sua lista de exames escrita no arquivo `data.csv`.

No *dataset* final `data.csv`, elementos com valores igual a zero significam resultados de exames que não estavam disponíveis na base de dados. Alem disso, logo antes dos dados serem usados pelas redes, normalizamos os dados, colocando-os no intervalo [0,1], e fazemos *over-sampling* para garantir que as classes fiquem balanceadas, isto é, termos o mesmo numero de pacientes com resultados positivos e negativos para um dos 3 exames principais. Essa fase final de adequação é feita pelo modulo `data.py`.

Arquitetura da rede neural

Para definir a melhor arquitetura para a rede neural optamos por testar tres modelos diferentes onde cada modelo tem tres camadas ocultas e foi chamada de NNX, X = {1,2,3}. As camadas de NN1 tem tamanho 512,256,128, as de NN2 1024,512,256 e as de NN3 256,128,64.

Optamos por usar 0.001 como taxa de aprendizado, Adam como otimizador, BCELoss() como loss function e 5 como numero de folds.

A partir dos dados obtidos com os testes (contidos em `NN/trained_nn/accuracy_table.png`) foi possível concluir que a melhor rede é a NN2.

Descrição dos experimentos

Os experimentos foram feitos a partir no arquivo `experiment.py`. Nele sao testadas tres reder neurais usando Stratified K-Fold Cross Validation com k = 5 no lugar de k = 10 para diminuir um pouco o longo tempo de execução. Os resultados de acurácia de cada rede para cada exame estão em `NN/trained_nn/accuracy_table.png`.

Para prever os resultados de uma pessoa, basta rodar `use.py`. Ele utiliza `min_max.csv` que contém o menor e o maior valor para cada exame utilizado para que se possa normalizar os dados fornecidos da pessoa.

Resultados

Discussão

Bibliografia

- [1] "Testes para Covid-19: perguntas e respostas". ANVISA. Disponível em: http://portal.anvisa.gov.br/documents/219201/4340788/Perguntas+e+respostas+-+testes+para+Covid-19.pdf. Acesso em: 07 de Julho de 2020.
- [2] LEE, Yuchun. "Handwritten Digit Recognition Using K Nearest-Neighbor, Radial-Basis Function, and Backpropagation Neural Networks". Disponível em: https://www.mitpressjournals.org/doi/abs/10.1162/neco.1991.3.3.440. Acesso em: 05 de Julho de 2020.
- [3] NUNES, Jorge; LLACER, Jorge. "Astronomical image segmentation by self-organizing neural networks and wavelets". Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S089360800300011X. Acesso em: 05 de Julho de 2020.
- [4] "Dados COVID Grupo Fleury". FLEURY, Grupo. Disponível em: https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/99. Acesso em: 01 de Julho de 2020.
- [5] "Dados COVID Hospital Sírio-Libanês". SIRIO-LIBANES, Hospital. Disponível em: https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/97. Acesso em: 01 de Julho de 2020.