Projeto do TCC

Lourenço Henrique Moinheiro Martins Sborz Bogo

May 3, 2022

1 Tema e Motivação

Meu TCC será sobre uso de aprendizado não supervisionado para clusterização de tumores de boca e pescoço a partir de dados radiômicos (dados retirados de imagens tomográficas). Ao separar esses tumores em clusters, pretendemos conseguir identificar padrões em pacientes com tumores dentro de um mesmo grupo de forma menos invasiva (principalmente considerando que a maioria dos pacientes realizam tomografias durante o tratamento). Com a identificação desses padrões, torna-se possivel também, relacionar esses grupos com dados de análise clínica usando modelos de aprenzidado supervisionado (como Máquinas de Vetores de Suporte).

Esse trabalho é interessante, pois diminuiria os custos e a dificuldade de identificar características de tumores, além de permitir que tratamentos sejam escolhidos levando em conta essas características identificadas, possivelmente aumentando a taxa de sobrevivência dos pacientes.

2 Objetivos

O objetivo principal é conseguir agrupar os tumores em subtipos usando apenas dados radiômicos (dados retirados de imagens do tumor). Se esse objetivo for atingido rapidamente, a ideia seguinte seria tentar relacionar os grupos criados a partir dos dados radiômicos com dados de análise clínica usando aprendizado supervisionado. Por exemplo, descobrir se é possível, com eficiência, prever a expressão de alguns genes no tumor usando apenas dados radiômicos. Essa segunda ideia é especialmente útil para compreender os tratamentos aplicáveis a cada caso.

3 Metodologia

Usaremos o banco de dados do TCGA (N = 113) como fonte para conseguir clínicos, de transcriptômica e de imagem. Depois disso, aplicaremos algoritmos de redução de dimensionalidade (por exemplo, o TSNE) para diminuir o número de features do conjunto de dados e, em seguida, clusterizar os dados em grupos. Posteriormente, iremos interpretar os resultados e verificar se a clusterização obtida oferece informações relevantes, como por exemplo, informações sobre a sobrevida dos pacientes e, talvez, relacionar os resultados com dados de expressão gênica.

Exemplificando, suponhamos que os tumores foram agrupados em 3 categorias a partir dos dados radiômicos: A, B e C. Pode ser que pacientes com tumores do tipo A tenham uma sobrevida maior do que pacientes com tumores do tipo C.

4 Cronograma inicial

Os primeiros meses (até maio) serão utilizados para familiarizar-me com o tema e com os jargões da área e, também, para decidir as ferramentas (linguagens, bibliotecas e algoritmos) que serão utilizadas na implementação do modelo. Os meses seguintes (até agosto) serão usados para implementar modelos de clusterização e de previsão de dados de expressão gênica. O tempo restante será utilizado na interpretação dos dados e decidir se os resultados foram promissores ou não para incentivar o uso de dados radiômicos na análise de tumores na prática médica.

- Até 16/04: Delimitação do tema.
- Até 30/04: Produção do site e do projeto do TCC.
- Até 31/05: Leitura de artigos para entendimento do tema e de documentações para delimitação das ferramentas.
- Até 31/08: Implementação e teste dos modelos.
- Até 01/12: Interpretação dos resultados e talvez prever expressão gênica usando os dados radiômicos. Escrita da Monografia.

5 Orientador(es)

Meu orientador nesse projeto de TCC é o André Fujita. Além dele, também estou sendo auxiliado pelo doutorando dele Vinícius Jardim Carvalho,

pelo professor da FMUSP Gilberto de Castro Junior e pelo seu aluno Mateus Cunha.