Aplicação de Algoritmos de Aprendizado de Máquina na Identificação de Recidiva em Diversos Tipos de Câncer

Bolsista Mariangela Dametto (CTI) mdametto@cti.gov.br

Supervisor Rodrigo Bonacin

Co-autores: Sérgio Vechi , Guilherme Ruppert

Resumo

Atualmente, não existe conhecimento estabelecido sobre a correlação entre as características específicas de cada câncer, as características do indivíduo e os tipos de tratamento adotados com a ocorrência ou não de recidiva após tratamento da doença. Portanto, este trabalho tem como objetivo inicial buscar um modelo que classifique pacientes com maior ou menor risco de recidiva, utilizando-se um número mínimo de características. Posteriormente, pretende-se analisar a relação entre as características supracitadas e os tratamentos associados a um risco maior ou menor de ocorrência de recidiva. Foram estudados 233 subtipos de câncer usando algoritmos de classificação supervisionados. Para tanto, foi utilizado um banco de dados público com 905.556 registros disponibilizados pela Fundação ONCOCENTRO de São Paulo. Até o momento, o melhor modelo de classificação obtido apresentou, para um mesmo tipo de câncer, acurácia de 83,08% (+/- 0.2) e média da medida F1 igual a 0.83. Entretanto, novos estudos devem ser conduzidos para obter resultados semelhantes ou superiores para a maioria dos tipos de câncer. Melhorias no pré-processamento dos dados, busca pelos melhores hiperparâmetros do algoritmo e análise mais profunda dos atributos de cada câncer estão em curso visando obter modelos mais eficientes.

Palavras-chave: Bancos de dados, pré-processamento, aprendizagem de máquina, recidiva de câncer.

1. Introdução

De forma geral, a recidiva de câncer está associada a um prognóstico ruim (CHIHARA, 2017; PUGH, 2016; SANTINI, 2016), sendo responsável por um forte estresse emocional nos pacientes, que têm que lidar novamente com o difícil período de tratamento (BUTOW, 2018). Uma grave consequência para o sistema de cuidados à saúde é que este sentimento angustiante acaba favorecendo o abandono das orientações de tratamento fornecidas pelos médicos e profissionais da saúde. Apesar dos avanços nas tecnologias da área da saúde, os tratamentos oncológicos são geralmente muito agressivos para o paciente e a escolha da melhor opção é uma tarefa delicada. Independentemente das características do câncer e de fatores ambientais, saber se algum tipo de tratamento específico está associado a uma maior frequência de recidiva pode auxiliar na tomada de decisão pelos médicos, visando uma terapia mais eficiente para cada caso.

A literatura apresenta trabalhos anteriores que estudaram a relação entre recidiva e alguns tipos específicos de tumor usando aprendizado de máquina (PAN, 2017; ABREU, 2016; JHA, 2018). Pan *et al.*, por exemplo, utilizaram 661 registros eletrônicos de pacientes para construir um modelo para predizer recidiva de leucemia linfoblástica aguda (LLA) em crianças por

meio de características clínicas, sociodemográficas, imunológicas e citogenéticas (PAN, 2016). O modelo mais eficiente foi obtido com o algoritmo *Random Forest* com acurácia 0.827 e AUC 0.902, sendo que também foram testados *Support Vector Machine*, *Logistic Regression* e *Decision Tree*. Além disso, os autores identificaram 14 das 103 características iniciais como sendo o número ótimo de características para realizar a predição de recidiva na LLA. Estas 14 características estavam relacionadas à quantidade de células do sangue produzidas pela medula óssea, idade, hepatomegalia, aumento no baço e também na presença de rearranjo genético BCR-ABL. Este trabalho demonstra que esta abordagem computacional apresenta potencial para auxiliar na tomada de decisão sobre dosagem e quantidade de tratamento quimioterápico. Outro grupo utilizou modelos supervisionados para auxiliar a classificação da doença inflamatória do intestino em crianças (MOSSOTO, 2017). A correta classificação é fundamental para um diagnóstico acurado e utilização do tratamento mais efetivo para cada caso sendo, portanto, muito útil para a prática clínica.

Métodos de aprendizagem de máquina têm sido aplicados à área da saúde com aumento na taxa de sucesso (MOSSOTO, 2017), devido à enorme quantidade de dados que tem sido gerada com os avanços tecnológicos na área médica. O aprendizado de máquina é uma área da inteligência artificial capaz de analisar essa grande quantidade de dados e extrair informações úteis. Neste trabalho, nós analisamos um grande banco de dados público, que foi construído com os esforços de vários hospitais de câncer do país. Além do tipo de tratamento utilizado para cada paciente, foram utilizadas informações demográficas e clínicas, como morfologia e topografia de cada tipo de tumor.

O restante deste artigo está estruturado da seguinte maneira: a seção 2 apresenta os objetivos deste trabalho, a seção 3 detalha a metodologia empregada, já a seção 4 apresenta os resultados obtidos e, por fim, a seção 5 faz as considerações finais e apresenta os próximos passos este trabalho.

2. Objetivos

Como objetivo principal, este trabalho busca encontrar um modelo que classifique acuradamente os casos de recidiva e não-recidiva de acordo com as características dos pacientes existentes no banco de dados da Fundação ONCOCENTRO[1]. Além disso, com o modelo criado, pretende-se predizer, com boa margem estatística, se um novo paciente que possua os mesmos tipos de características mensuradas no banco apresenta risco de recidiva.

Com a obtenção do modelo, pretende-se que sejam identificadas quais características são determinantes para a classificação de um paciente como tendo risco maior ou menor de ocorrência de recidiva.

3. Metodologia

A metodologia empregada neste trabalho é descrita pelos seguintes item: (1) coleta de dados (banco de dados utilizado), (2) construção do *dataset* e pré-processamento dos dados e (3) seleção do modelo e avaliação estatística. Os parágrafos subsequentes detalha cada um desses item.

Coleta de Dados. O banco de dados da Fundação ONCOCENTRO é de acesso público e é alimentado por 77 hospitais do país. O comprometimento dos hospitais é que sejam enviados

1 http://www.fosp.saude.sp.gov.br/

dados a cada 3 meses para a Fundação. Os resultados deste trabalho foram obtidos com os dados da atualização realizada em Dezembro de 2019, que contém 905.556 registros de pacientes e 95 características para cada paciente, sendo 331 tipos diferentes de câncer no total.

Construção do *dataset* e pré-processamento dos dados. Após o pré-processamento, foram mantidos 233 tipos de câncer (com ocorrências maiores que 100 casos), totalizando 901.922 registros e 23 características consideradas relevantes e não-redundantes para determinação da classe. Esta etapa foi a que consumiu mais tempo, pois dados faltantes tiveram que ser tratados, atribuindo-se a eles valores que não havia no banco para o atributo específico. Os dados ambíguos estão sendo analisados minuciosamente por meio de bibliotecas de Python, mas esta etapa ainda não foi concluída. Estes dados ambíguos ainda não processados foram removidos por enquanto das análises atuais.

Seleção do modelo e avaliação estatística. Nesta etapa, 7 algoritmos de classificação foram aplicados para o câncer com maior quantidade de casos do banco de dados, 97080 pacientes com neoplasia maligna da pele de outras partes e de partes não especificadas da face. A escolha destes algoritmos foi determinada pela interpretabilidade do modelo gerado e pelo baixo custo computacional para esta fase inicial de melhor compreensão dos dados da saúde. O melhor classificador encontrado até o momento foi o *Naive Bayes*, sendo que este foi então aplicado para cada um dos outros 232 tipos de câncer. Os outros algoritmos analisados foram *MultilayerPerceptron*, *Bagging*, *RandomCommittee*, *ZeroR*, *Random Forest* e *RepTree*. Todos estes algoritmos fazem parte do software WEKA (*Waikato Environment for Knowledge Analysis*)[2] da Universidade de Waikato, Nova Zelândia. Os classificadores foram treinados e validados por meio do procedimento de *resampling 10-fold cross validation*. A versão utilizada foi a Weka-3-9-3. Como a classe de não-recidiva apresenta um número muito maior de pacientes em todos os tipos de câncer, foi realizado um balanceamento da classe por meio do filtro *ClassBalancer* presente no WEKA. Este método atribui pesos para as instâncias de forma que a soma dos pesos de todas as classes existentes para estas instâncias é mantida.

Os indicadores estatísticos calculados para avaliar a adequação do modelo foram:

a) acurácia: proporção das instâncias corretamente classificadas em relação ao número total de instâncias.

$$\frac{TP+TN}{TP+FP+TN+FN}$$

b) sensibilidade ou *recall*: proporção dos valores preditos como verdadeiros (positivo ou negativo) em relação ao número total de valores verdadeiros reais (positivo ou negativo).

$$\frac{TP}{TP+FN} \text{ e } \frac{TN}{TN+FP}$$

c) especificidade ou precisão: proporção dos valores preditos como verdadeiros (positivo ou negativo) e que são realmente verdadeiros em relação ao número total de preditos (positivo ou negativo).

$$\frac{TP}{TP+FP} \text{ e } \frac{TN}{TN+FN}$$

d) medida F1: é uma média harmônica entre a sensibilidade e a especificidade, e é uma medida importante em situações nas quais não só os casos verdadeiros (positivo ou negativo) são importantes, mas também os falso-negativos ou falso-positivos. Neste trabalho, o

2 https://www.cs.waikato.ac.nz/ml/weka/

conhecimento de falso-negativos é vital, ou seja, um paciente com risco real de recidiva não deve ser classificado como não tendo o risco.

$$\frac{2 * sensibilidade * especificidade}{sensibilidade + especificidade}$$

4. Resultados

O conjunto de dados utilizado para a aplicação de algoritmos de aprendizado de máquina apresenta 901.922 registros de pacientes, 233 tipos de câncer e 23 características dos pacientes. Dentre estas, há características sócio-demográficas, de caracterização do tumor (topografia, morfologia, estadiamento clínico e classificação TNM), de tipo de tratamento recebido (cirurgia, quimioterapia, hormonioterapia, radioterapia) além da classe recidiva. Uma informação que seria bastante relevante para tomada de decisão sobre tratamento a ser escolhido é a quantidade/dose das terapias adotadas, além de resultados de exames bioquímicos, por exemplo, realizados para diagnóstico.

O gráfico da Figura 1 mostra a quantidade de pacientes associados aos diferentes tipos de câncer (curva azul no gráfico). Somente em 22 tipos de câncer do banco de dados há mais de 10.000 pacientes para cada câncer. O ponto de corte, neste primeiro momento, foi selecionar os tipos de câncer com pelo menos 100 pacientes para que pudéssemos verificar a quantidade de dados que seriam suficientes para treinar o modelo de forma significante estatisticamente.

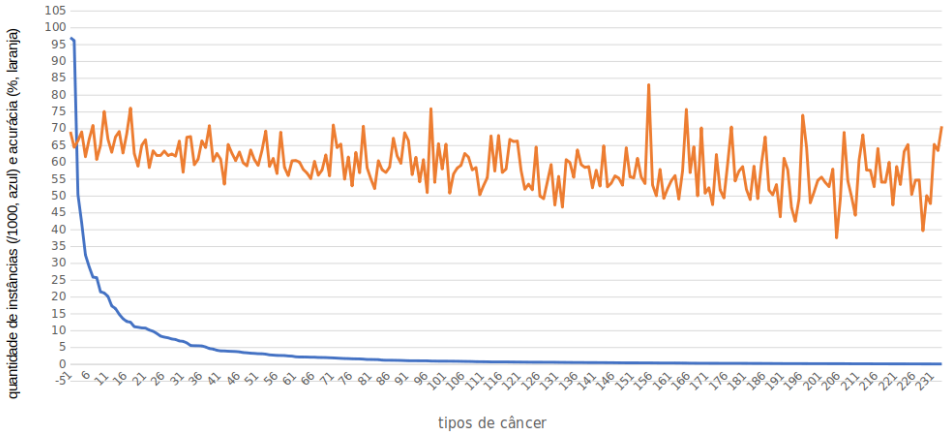

Figura 1 - Medida de acurácia (%, em laranja) e quantidade de instâncias (/1000, em azul) para cada um dos 233 tipos de câncer.

O gráfico da Figura 1 também mostra a quantidade de instâncias corretamente classificadas (acurácia) pelo algoritmo Naive Bayes para cada um dos 233 tipos de câncer. Curiosamente, o tipo de câncer associado ao número 155 do gráfico apresenta somente 433 pacientes e 83,08% de correta classificação pelo algoritmo de aprendizagem. Este câncer é o de pâncreas e apresenta somente 3,7% casos de recidiva e 96,3 % casos de não-recidiva no banco de dados antes da aplicação do filtro de balanceamento da classe. Ou seja, não há inicialmente um número mais equilibrado a ser balanceado entre as classes para que o algoritmo pudesse ter aprendido de forma mais eficiente quando comparado com os outros tipos de câncer.

As medidas de F1 para todos os tipos de câncer estão mostrados no gráfico da Figura 2. É possível observar que a partir do câncer numerado 22, a discrepância entre medidas F1 para recidiva e não-recidiva fica maior, e é a partir deste câncer que há menos do que 10.000 instâncias no banco de dados.

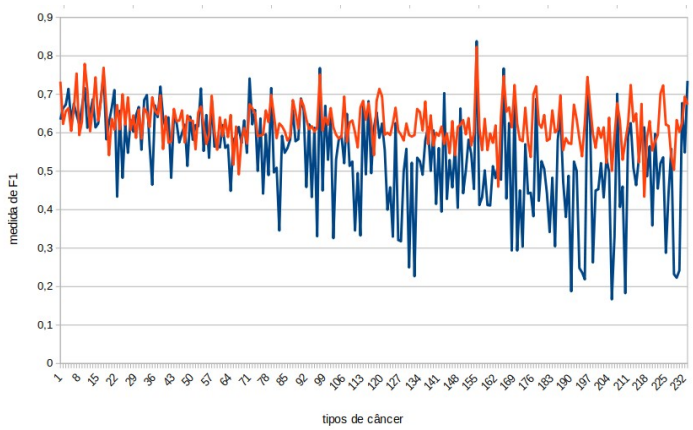

Figura 2 – Medidas F1 para a classe recidiva (azul) e não-recidiva (vermelho)

Curiosamente, assim como no gráfico da Figura 1, há alguns tipos de câncer (97, 155 e 165, por exemplo) que apresentam alto score F1 para as 2 classes, mesmo havendo relativamente poucos registros. Cada um destes tipos apresentam, respectivamente, 977, 433 e 364 casos no banco de dados.

Na tabela 1 abaixo estão mostrados as quantidades de pacientes e as porcentagens de recidiva e não-recidiva para os tipos de câncer que apresentaram acurácia >= 68% e medida de F1 >= 0.68 para cada uma das classes.

Tipo de câncer	Acurácia (%)	F1 – recidiva	F1 – não recidiva	% recidiva	% não recidiva	Quantidade de instâncias
10	75,13	0,72	0,78	2,37	97,63	21217
16	68,56	0,69	0,68	4,18	95,82	12737
17	76,14	0,75	0,77	6	94	12534
38	70,92	0,72	0,7	7,03	92,97	4698
57	68,96	0,68	0,7	10,15	89,85	2641
79	70,75	0,72	0,7	8,34	91,66	1558
90	68,82	0,69	0,7	22,55	77,45	1122
97	75,95	0,77	0,75	7,98	92,02	977
155	83,08	0,84	0,82	3,7	96,3	433
165	75,76	0,77	0,75	5,5	94,5	364
177	70,54	0,69	0,72	12,09	87,91	306
196	74,02	0,74	0,75	11,54	88,46	217

Tabela 1 – Acurácia e medidas F1 para cada classe dos tipos de câncer que apresentaram valores destes indicadores >= 68%. Também é mostrado a porcentagem de cada classe antes do balanceamento e a quantidade de pacientes para cada câncer.

5. Considerações Finais e Trabalhos Futuros

Os resultados apresentados são promissores e apresentam avanços em relação à literatura existente, entretanto são necessários novos estudos para aprimoramento das bases e dos modelos para, assim, torna-los viáveis na prática. Para tanto, estamos analisando as porcentagens de casos de recidiva e não-recidiva para todos os tipos de câncer estudados neste trabalho e verificando a metodologia de aplicação do peso calculado pelo filtro *Classbalancer* do WEKA em cada caso para que se possa encontrar o motivo para estas observações.

Além disso, na próxima etapa do projeto, será investigada a importância de cada característica específica para descrever cada tumor deste trabalho e a correlação entre elas. Será também verificada a qualidade dos valores dos atributos presentes nos dados para cada câncer específico, com a finalidade de esclarecer a razão de se obter indicadores não satisfatórios para grande parte dos tipos de câncer, inclusive dos que apresentam um número muito maior de pacientes, utilizando-se os mesmos atributos.

Referências

ABREU, P.H. et al. *Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review*. ACM Computing Surveys. 49(3): 1-40, 2016.

BUTOW, P. et al. F*ear of Cancer Recurrence: A Practical Guide for Clinicians*. Oncology (Williston Park). 32(1):32-8, 2018.

CHIHARA, D. et al. *The survival outcome of patients with relapsed/refractory peripheral T-cell lymphoma-not otherwise specified and angioimmunoblastic T-cell lymphoma*. Br J Haematol. 176(5):750-758, 2017.

JHA, A. et al. *Deep Convolution Neural Network Model to Predict Relapse in Breast Cancer*. 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, pp. 351-358, 2018.

MOSSOTO, E. et al. *Classification of Paediatric Inflammatory Bowel Disease using Machine Learning*. Sci Rep 7, 2427, 2017.

PAN, L. et al. *Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia*. Sci Rep 7, 7402, 2017.

PUGH, S.A. et al. *Site and Stage of Colorectal Cancer Influence the Likelihood and Distribution of Disease Recurrence and Postrecurrence Survival: Data From the FACS Randomized Controlled Trial*. Ann Surg. 263(6):1143-7, 2016.

SANTINI, D. et al. *Risk of recurrence and conditional survival in complete responders treated with TKIs plus or less locoregional therapies for metastatic renal cell carcinoma*. Oncotarget. 7(22):33381-90, 2016.