Um Método Para a Previsão de Resposta a Medicamentos em Pacientes com Epilepsia

Helen Cristina F. da Silva, Mariana Saragiotto S. Alves e Tiago Carvalho
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - *Campus* Campinas
13069-901, Campinas - SP, Brasil
Email: helensilva14@gmail.com, mariana.alves@ifsp.edu.br, tiagojc@gmail.com

***Resumo*—Grande parte dos pacientes com epilepsia não respondem ao tratamento medicamentoso disponível no mercado. A descoberta dessa refratariedade ao tratamento, atualmente, se dá por meio de um processo empírico, em que o paciente é submetido a diferentes drogas por um longo período de tempo até ser constatado que ele não responde a nenhum tratamento e deve ser encaminhado para cirurgia. Entretanto, há indícios que tal refratariedade possa estar ligada a características genéticas específicas do indivíduo. Este trabalho propõe uma nova abordagem, baseada em aprendizado de máquina, para o problema de detecção de refratariedade medicamentosa em pacientes com epilepsia. Usando uma combinação do algoritmo XGBoost e de dados genéticos dos pacientes, nossa abordagem teve um aumento de 319.05% na acurácia de predição de pacientes refratários em relação ao estado da arte.**

***Abstract*—Most patients with epilepsy do not respond to the drug treatment available on the market. This lack of response to treatments, nowadays, is just discovered through an empirical process in which the patient is subjected to different drugs for a long period of time until it is found he does not respond to any treatment and he should undergo surgery. However, there are signs that this refractoriness situation can be related with patient genetic features. This work proposes a new approach, based on machine learning, for refractoriness detection problem. Using a combination of XGBoost algorithm and patients genetic data, our approach improved the accuracy of state-of-the-art methods in 319.05% for refractoriness prediction.**

I. Introdução

Um medicamento receitado para uma pessoa com uma determinada doença nem sempre será eficaz da mesma maneira para outro paciente que tenha a mesma doença. Enquanto um medicamento tem ação benéfica para um grupo de pacientes, outros podem manifestar reações indesejáveis a partir do seu uso, ou ainda, não responder ao uso do medicamento [1].

A farmacogenética é a área de estudo das variações genéticas individuais que levam à obtenção de diferentes respostas à ação dos medicamentos. Segundo Pirazzoli e Recchia [1], esta ciência estuda os genes que codificam proteínas envolvidas no mecanismo de ação, transporte e/ou metabolização das drogas no organismo humano em níveis celular, tecidual, individual ou populacional.

Uma série de estudos da farmacogenética voltam-se para as doenças neurológicas e para a eficácia dos tratamentos realizados [2]. De acordo com a *World Health Organization* [3], aproximadamente cinquenta milhões de pessoas ao redor do mundo sofrem de epilepsia e grande parte deles possuem crises recorrentes que precisam ser tratadas. Mais de 30% desses pacientes não respondem ao tratamento medicamentoso disponível para a doença e são classificados como refratários. Aqueles que respondem ao tratamento são classificados como responsivos.

A Epilepsia do Lobo Temporal Mesial (ELTM) é uma das principais formas de epilepsia, destacando-se por conta de seu difícil tratamento. De acordo com Toledo [4], a ELTM acontece por conta de alterações no funcionamento de neurônios localizados nas estruturas mais profundas do cérebro, como o hipocampo e a amígdala, onde são controladas funções importantes para o ser humano, tais como a memória, atenção e a ansiedade. As crises epilépticas são causadas por descargas elétricas anormais em um grande grupo de neurônios. Elas podem ou não resultar em uma convulsão, interferir na memória e em outras funções cerebrais, e até mesmo colocar o paciente em risco de acidentes e/ou morte.

Segundo Toledo [4], no que se refere à resposta aos medicamentos antiepilépticos, é ainda mais notável a quantidade de pacientes com ELTM que são refratários, podendo chegar a 40% do total de pacientes com ELTM. Muitas vezes, a cirurgia para remoção da área cerebral que provoca as crises é a melhor opção. Porém, o consenso atual da área procura medidas para controlar as crises epilépticas a partir de diferentes terapias medicamentosas antes de se considerar a cirurgia.

Contudo, a detecção da resistência aos medicamentos decorre de um processo empírico, no qual o paciente é submetido a diferentes drogas por um longo período de tempo e continua sofrendo com as crises que não são controladas [4].

Acredita-se que a predisposição genética possa ser um dos principais causadores de tal refratariedade medicamentosa. Atualmente, há estudos na farmacogenética [5], que buscam entender quais genes estão envolvidos na absorção, no metabolismo e no transporte das drogas antiepilépticas no organismo dos pacientes. Para lidar com os dados de pacientes que devem ser coletados e analisados nestes estudos, além da busca para

redução de custos, muitos pesquisadores estão recorrendo ao uso de técnicas de aprendizado de máquina, tal como no estudo de Silva-Alves *et. al* [6].

Atualmente, as aplicações de *Machine Learning* (ML) na genômica dividem-se em diversas categorias. Segundo Sennaar [7], a principal é o sequenciamento genômico, principalmente no contexto da medicina personalizada, na qual os pesquisadores estão usando aprendizado de máquina na identificação de padrões dentro de extensos conjuntos de dados genéticos. Esses padrões são então traduzidos para modelos computacionais que podem ajudar a prever a probabilidade de um indivíduo desenvolver certas doenças ou colaborar para o estudo de respostas a medicamentos.

Nesse contexto, o estudo de Silva-Alves *et. al* [6] procurou desenvolver uma metodologia capaz de predizer o tipo de resposta medicamentosa de um paciente de ELTM, refratário ou responsivo, com base na análise do material clínico e genético dos participantes. Utilizando uma combinação de ML, dados genéticos e clínicos dos pacientes, os autores obtiveram uma área sob a curva (AUC) de 81%.

Neste trabalho, investigamos a utilização de métodos de ML mais robustos e propícios a lidar com estes tipos de dados. Utilizando o algoritmo XGBoost, nosso método é capaz de detectar pacientes refratários com uma AUC de 88% usando apenas dados genéticos.

De forma geral, as principais contribuições deste trabalho são: (1) a investigação da eficácia de diferentes tipos de algoritmos de ML aplicados ao problema de detecção de refratariedade medicamentosa; (2) o aumento de 8.64% na AUC na detecção de resposta medicamentosa usando somente os dados genéticos, com a expressiva melhora de 319.05% na acurácia de predição de pacientes refratários em relação ao estado da arte; (3) a colaboração para a redução do tempo de constatação da refratariedade; (4) o fomento a estudos de previsão de resposta a medicamentos.

O restante deste trabalho está dividido da seguinte forma: na Seção II apresentamos em detalhes a metodologia proposta neste trabalho. Na Seção III apresentamos os principais experimentos realizados para a validação do método proposto. Por fim, na Seção IV apresentamos nossas principais conclusões e perspectivas para investigações futuras.

II. METODOLOGIA

De forma geral, a metodologia proposta para a resolução do problema deste trabalho consiste de duas etapas: (1) pré-processamento dos dados dos pacientes; (2) treinamento de diferentes classificadores para reconhecimento de padrões entre o tipo de resposta medicamentosa (refratária ou responsiva) e os dados dos pacientes. A Figura 1 apresenta a visão geral da metodologia proposta.

A. *Pré-processamento dos Dados*

Para serem utilizados pelos métodos de ML, os dados dos pacientes precisam ser preparados, seguindo um conjunto específicos de etapas de pré-processamento. Os principais passos são descritos nas subseções seguintes.

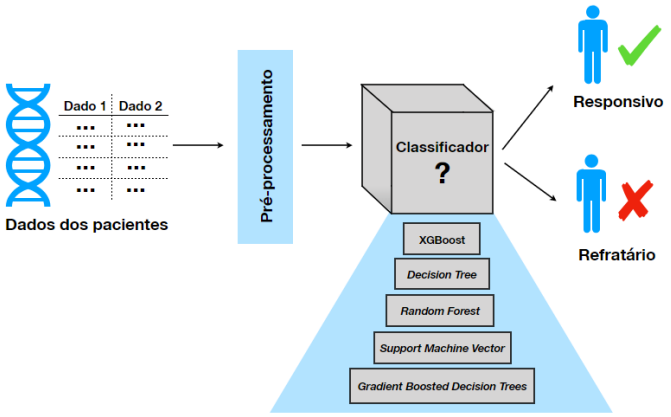

Figura 1. Visão geral da metodologia proposta.

1) *Conversão das variáveis categóricas para variáveis numéricas*: quando sequenciados em laboratório, cada gene passa a ser representado como um par de bases nitrogenadas (Adenina - A, Citosina - C, Guanina - G, Timina - T), também conhecidos como alelos. Entretanto, para serem utilizados em algoritmos de ML, é preciso representá-los numericamente. Assim, a etapa de conversão atribui um valor numérico a cada alelo possível. A Tabela I apresenta a conversão de alelos para valores numéricos.

Tabela I
VALORES ATRIBUÍDOS AOS ALELOS

CC	TT	CG	AA	AC	GT	CT	AG	GG	AT	TC
2	3	4	5	6	7	8	9	10	11	12

2) *Tratamento de valores ausentes:* no que se refere aos dados genéticos, são vários os motivos para a não obtenção de um determinado valor e o mais recorrente é a baixa qualidade da amostra, que pode ter sido comprometida na extração ou no armazenamento. A solução para esse problema foi a utilização do conceito da frequência alélica a fim de atribuir o alelo mais comum de cada gene aos valores faltantes, isto é, aquele que é o mais frequente (relaciona-se com o conceito estatístico conhecido como moda).
3) *Conversão das variáveis numéricas para variáveis dummy*: valores numéricos podem implicitamente induzir o algoritmo a criar uma relação de ordenação entre os valores. Para evitar este tipo de problema, nossa abordagem transforma cada valor numérico em variáveis *dummy* (fictícias), que são mais expressivas para os modelos, a partir da técnica *One-Hot Encoding*, que codifica cada variável numérica para um vetor binário de D posições (D = quantidade de classes/categorias do conjunto), na qual uma delas tem o valor 1 para indicar a presença da sua classe correspondente (é a chamada posição quente) enquanto outros são 0, indicando a ausência das demais classes.

B. Construção dos Modelos de ML

As características pré-processadas na etapa anterior servem de entrada aos modelos de ML nesta etapa. Com o propósito de comparação, utilizamos o protocolo de validação cruzada *Leave-One-Out* (LOO), como adotado por Silva-Alves *et. al* [6], para a avaliação de diferentes métodos para a classificação das amostras. Os classificadores usados são brevemente descritos nos itens seguintes.

1) *Decision Tree [8]* (DT): a árvore de decisão é um dos modelos de classificação mais básicos na literatura. Deve ser percorrida de cima para baixo, seguindo pelos arcos dos nós, nos quais as características das amostras satisfazem os testes realizados em cada nó até alcançar um nó-folha, que contém a classificação da amostra;
2) *Random Forest [8]* (RF): este método utiliza a estratégia de *bagging* para combinar o resultado de várias árvores de decisão, por meio de um mecanismo de votação, para determinar a classificação de uma amostra;
3) *Support Machine Vector [8]* (SVM): é um dos métodos mais utilizados em problemas de classificação na literatura de ML. Busca por um hiperplano de separação entre as classes do problema de forma a maximizar as margens que separam os pontos mais próximos a cada uma das classes;
4) *Gradient Boosting [8]* (GB): assim como o algoritmo RF, utiliza a abordagem de *boosting* para gerar uma sequência de árvores simples, onde cada árvore sucessiva é construída para os resíduos de previsão da árvore anterior. Desta forma, cada nova árvore tende a aprender com os erros que foram obtidos na etapa anterior;
5) *XGBoost [9]*: é uma evolução do algoritmo GB, a qual foi especialmente projetada para uma execução de alta velocidade computacional e alto desempenho, além de oferecer regularização, flexibilidade e processamento paralelo.

III. EXPERIMENTOS E RESULTADOS

Esta seção apresenta os principais experimentos e resultados obtidos ao longo do trabalho.

A. Ambiente de Desenvolvimento

O ambiente de desenvolvimento utilizado foi um computador Windows 10 Intel(R) Core(TM) i7-4600U CPU 2.10GHz 16GB RAM. Foi utilizada a linguagem de programação Python (3.6.1) e as bibliotecas Numpy (1.12.1), Pandas (0.20.1), Matplotlib (2.0.2), Scikit-learn (0.18.1) e XGBoost (0.71).

B. Bases de Dados

A base de dados clínicos e genéticos de pacientes usada nos experimentos é a mesma do estudo de Silva-Alves *et. al* [6]. Para os dados genéticos, foram selecionados 11 genes (*ABCB1, ABCC2, CYP1A1, CYP1A2, CYP1B1, CYP2C9, CYP2C19, CYP2D6, CYP2E1, CYP3A4, CYP3A5*) que estão envolvidos na absorção, no metabolismo e no transporte de medicamentos antiepilépticos. Nesses genes, foram genotipados 119 marcadores moleculares do tipo polimorfismo de base única (SNP), os quais são utilizados como características de entrada para os algoritmos de ML. Em relação ao número de amostras, a base de dados conta com 241 pacientes com ELTM que já estavam sendo acompanhados na Universidade de Campinas (UNICAMP) há pelo menos dois anos. Deste grupo, foi constatado que 162 pacientes eram refratários ao tratamento, enquanto os outros 79 respondiam bem aos medicamentos. Desse modo, temos um conjunto de dados desbalanceado, no qual, a quantidade de pacientes refratários (instâncias negativas) superam a de responsivos (instâncias positivas), o que pode afetar negativamente o desempenho dos classificadores.

C. Métricas de Desempenho

Como métricas para medir o desempenho dos modelos, nos baseamos nas seguintes medidas:

1) *Accuracy:* taxa de identificação de pacientes refratários e responsivos dentre todos os pacientes do conjunto.
2) *Precision:* quantidade de pacientes responsivos reais (*True Positive*) dividida pelo total de pacientes classificados como responsivos;
3) *Recall:* quantidade de pacientes responsivos reais (*True Positive*) dividida pelo total de pacientes responsivos reais no conjunto de teste;
4) *F1-Score*: média ponderada entre os valores de *Precision* e *Recall*, a qual é significativa em cenários de dados desbalanceados;
5) *Sensitivity:* acurácia de predição de responsividade;
6) *Specificity:* acurácia de predição de refratariedade.

D. Pré-processamento dos Dados

A partir da base de dados descrita na Seção III-B, aplicamos os passos de pré-processamento descritos na Seção II-A, resultando em um vetor de características com 241 linhas e 1309 colunas (cada característica é agora representada por 11 dimensões).

E. Comparação entre Diferentes Classificadores

Todos os classificadores descritos na Seção II-B foram treinados utilizando os vetores de características resultantes do pré-processamento e avaliados utilizando o protocolo LOO, confome adotado no trabalho de Silva-Alves *et. al* [6].

O classificador DT utilizou a métrica *gini* como critério de construção da árvore com a estratégia de encontrar a melhor divisão de nós. Todos os valores do vetor de características foram considerados, tendo os rótulos das amostras com peso 1. A árvore não teve profundidade máxima definida e utilizou o mínimo de 2 amostras necessárias para a obtenção de um novo nó. Os classificadores RF e GB fazem uso desses mesmos parâmetros.

O classificador RF utilizou 10 árvores de decisão para compor a "floresta", ao passo que, o classificador GB usou a regressão logística como função de perda a ser otimizada com 100 árvores de decisão como estimadores, tendo o hiperparâmetro *learning_rate* com valor padrão igual a 0.1 usado para desacelerar a aprendizagem do modelo. O XGBoost

utilizou 100 instâncias do algoritmo *Gradient Boosting* como estimador, nos quais as árvores possuem profundidade máxima de 3 níveis.

O classificador SVM utilizou o kernel RBF e teve seus demais parâmetros definidos via *grid search* nos intervalos: η ϵ [1e-3, 1e-4] e α ϵ [0.01, 1, 10, 100], onde η e α denotam a configuração de $gamma$ e C, respectivamente.

A Tabela II apresenta os resultados obtidos e a Figura 2 apresenta a curva ROC para cada um dos classificadores avaliados, somente com os dados genéticos dos pacientes.

Tabela II
RESULTADOS OBTIDOS SOMENTE COM OS DADOS GENÉTICOS

Classificador	*Accuracy*	*Precision*	*Recall*	*F1-score*	AUC
Decision Tree	0.75	0.76	0.76	0.76	0.70
Random Forest	0.73	0.73	0.74	0.71	0.75
SVM	0.79	0.79	0.80	0.79	0.83
Gradient Boosting	0.79	0.79	0.80	0.79	0.86
XGBoost	0.81	0.81	0.81	0.81	0.88

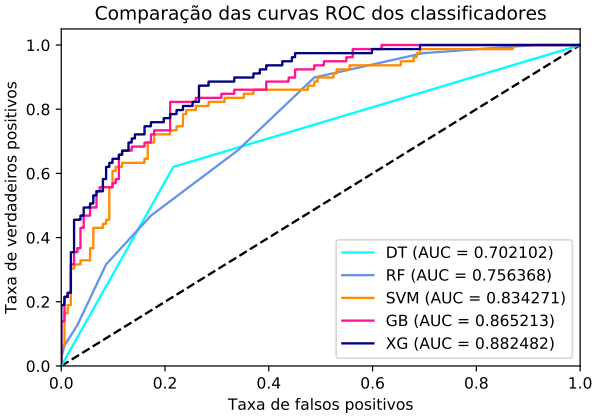

Figura 2. Curva ROC dos classificadores somente com os dados genéticos

Tanto pela comparação gráfica (via curvas ROC) quanto numérica (via tabela) o algoritmo XGBoost apresentou o melhor desempenho, seguido pelo *Gradient Boosting*, SVM, *Random Forest* e por último o *Decision Tree*.

No geral, é possível notar que classificadores baseados em árvores são particularmente satisfatórios para aplicação em dados genéticos, tendo em vista que são capazes de avaliar as diferentes combinações entre as características fornecidas.

F. Comparação com o Estado da Arte

O algoritmo XGBoost foi o melhor classificador ao detectar pacientes refratários com uma AUC de 88% usando somente os dados genéticos. Portanto houve um aumento de 8.64% com relação a AUC de 81% obtida no estudo de Silva-Alves *et. al* [6] usando a combinação dos dados clínicos e genéticos. Neste mesmo contexto, ao obter 88% de *specificity* em relação ao valor de 21% de Silva-Alves *et. al* [6], tivemos um aumento de 319.05%, porém tivemos uma redução de 29% no valor da *sensitivity*, visto que obtivemos 67% em comparação ao valor de 95% de Silva-Alves *et. al* [6]. Contudo, vale ressaltar que, dado que a detecção de refratariedade é o objetivo deste trabalho, uma maior *specificity* é mais representativa em contraste à detecção de *sensitivity*.

Apesar do *XGBoost* emprestar técnicas do RF, tais como amostragem em coluna [9], ele se mostra mais eficiente, tendo em vista que ele reduz tanto a variância quanto o viés estatístico, ao passo que o RF reduz somente a variância, além de facilitar muito o uso de diferentes funções de perda.

IV. CONCLUSÕES E TRABALHOS FUTUROS

Os experimentos iniciais realizados somente com os dados genéticos dos pacientes demonstraram que os modelos propostos alcançaram resultados satisfatórios, quando comparados aos resultados do estudo de Silva-Alves *et. al* [6], considerando que ultrapassamos os resultados obtidos com a combinação dos dados clínicos e genéticos. Além disso, tivemos um aumento de 319.05% na acurácia de predição de pacientes refratários, um resultado significativo para este cenário, visto que temos mais pacientes refratários do que responsivos.

Para os próximos passos, pretende-se: (1) realizar o pré-processamento dos dados clínicos e aplicá-los nos classificadores implementados até o momento, (2) elaborar procedimentos de refinamento de parâmetros com validação cruzada de $5 folds$ e 10 *folds* nos modelos que obtiveram os melhores resultados e (3) implementar um algoritmo para meta-classificação, a partir das probabilidades de predição dos modelos com os melhores resultados, tanto na base de dados genéticos quanto na base de dados clínicos.

REFERÊNCIAS

[1] A. Pirazzoli and G. Recchia, "Pharmacogenetics and pharmacogenomics: are they still promising?" *Pharmacological Research*, vol. 49, no. 4, pp. 357–361, 2004.

[2] A. Chan, M. Pirmohamed, and M. Comabella, "Pharmacogenomics in neurology: Current state and future steps," *Annals of Neurology*, vol. 70, no. 5, pp. 684–697.

[3] World Health Organization. (2018) Epilepsy. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs999/en/

[4] K. Toledo. (2017) Método identifica pacientes com epilepsia que podem se beneficiar com cirurgia. São Paulo. [Online]. Available: http://agencia.fapesp.br/metodo_identifica_pacientes_com_epilepsia_que_podem_se_beneficiar_com_cirurgia_/24697

[5] S. Balestrini and S. M. Sisodiya, "Pharmacogenomics in epilepsy," *Neuroscience Letters*, vol. 667, pp. 27 – 39, 2018, epilepsy: Advances in Genetics and Pathophysiology.

[6] M. S. Silva-Alves, R. Secolin, B. S. Carvalho, C. L. Yasuda, E. Bilevicius, M. K. M. Alvim, R. O. Santos, C. V. Maurer-Morelli, F. Cendes, and I. Lopes-Cendes, "A prediction algorithm for drug response in patients with mesial temporal lobe epilepsy based on clinical and genetic information," *PLOS ONE*, vol. 12, no. 1, pp. 1–15, January. 2017.

[7] K. Sennaar. (2018) Machine learning in genomics: Current efforts and future applications. [Online]. Available: https://www.techemergence.com/machine-learning-in-genomics-applications/

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[9] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings Of The 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining*. [s.l.]: ACM Press, mar. 2016, pp. 785–794.