ESTUDOS SOBRE APLICAÇÕES DE APRENDIZADO DE MÁQUINA NA DETECÇÃO DE DOENÇAS RARAS

VINICIUS ALEXANDRE DE O. ZEVAREX[1], CARLOS HENRIQUE DA S. SANTOS[2]

1 Curso Técnico em Informática Integrado ao Ensino Médio em Andamento, Bolsista PIBIFSP, IFSP, Câmpus Itapetininga e
2 Professor do curso Técnico em Informática Integrado ao Ensino Médio, IFSP, Câmpus Itapetininga.
vinicius.zevarex2002@gmail.com, carlos.santos@ifsp.edu.br

Área de conhecimento (Tabela CNPq): 1.03.04.03-7 Software Básico
Apresentado no
10° Congresso de Inovação, Ciência e Tecnologia do IFSP ou no 4º Congresso de Pós-Graduação do IFSP
27 e 28 de novembro de 2019- Sorocaba-SP, Brasil

RESUMO: As doenças são tidas raras quando afetam até 65 pessoas em cada 100.000 indivíduos, geralmente, sendo que usualmente o diagnóstico tardio dessas doenças é tardio e pode prejudicar no seu tratamento, deixando assim muitas sequelas ou não sendo eficaz. O emprego da computação para auxiliar tanto no levantamento de informações dessas doenças quanto no descobrimento de doenças raras têm agilizado e facilitado esse trabalho. Nesse contexto, este trabalho é uma pesquisa exploratória que visa compreender, por meio de um referencial teórico sistêmico, estudos sobre imunodeficiência primárias e como possivelmente diferentes técnicas de aprendizado de máquina tem contribuído com essa área. Para isso, duas diferentes bibliotecas de bioinformática para análise gênica têm sido estudada para a realização de testes posteriores, incluindo bibliotecas baseadas no descobrimento de doenças e análise genéticas chamadas Selene e o Expecto. Assim, espera-se que este trabalho contribua com uma sistematização de referencial teórico para auxiliar em futuros trabalhos de aprendizado de máquina e bioinformática associado a doenças raras ser desenvolvido no grupo.

PALAVRAS-CHAVE: Doenças Raras; Aprendizado de Máquina; Bioinformática.

STUDIES ABOUT MACHINE LEARNING APPLICATIONS IN RARE DISEASE DETECTION

ABSTRACT: Rare diseases are described when they affect up to 65 people in every 100,000 individuals, being the diagnosis usually delayed of these diseases can often hinder their treatment, leaving many sequelae or not effective. The use of computing to assist both in gathering information on these diseases and discovering rare diseases has streamlined and facilitated this work. In this context, this work is an exploratory research that aims to understand, through a systemic theoretical review, studies on primary immunodeficiency and how different machine learning techniques would contribute to this area. Therefore, two different bioinformatics libraries for gene analysis have been studied for further testing with including libraries based on disease discovery and genetic analysis called Selene and the ExPecto. Thus, it is expected that this work contributes to a systematization of theoretical framework to assist in future work on machine learning and bioinformatics associated with rare diseases to be developed in the research group.

KEYWORDS:Rare diseases; Machine Learning; Bioinformatics.

INTRODUÇÃO

As Imunodeficiência primárias raras são enfermidades de origem genética e que afetam os níveis de anticorpos, interferindo nos mecanismos de defesa do corpo e aumentando a

susceptibilidade de doenças. Há pouco mais 300 variações já levantadas na literatura sobre essas doenças raras, que segundo Pires et al. (2013), a Imunodeficiência Comum Variável (ICV) é tido como o tipo de imunodeficiência primária mais comum com incidência estimada de 1:10.000 a 1:50.000. Há uma grande dificuldade em diagnosticá-la com antecedência, muitas vezes a identificando quando ela já está em um estágio avançado ou crônico.

A Imunodeficiência comum variável é uma enfermidade que afeta os níveis de anticorpos, interferindo nos mecanismos de defesa do corpo e aumentando a susceptibilidade de doenças. As células B desses pacientes são incapazes se diferenciar em células plasmáticas, resultando em hipogamaglobulinemia. A característica laboratorial dessa doença é redução do nível de Imunoglobulina G(IgG) no sangue, sendo menor que 500 mg/dL em adultos, acompanhado da redução dos níveis de IgA e, em aproximadamente 50% dos casos, o IgM é reduzido juntamente (TORRES, 2007).

Os portadores dessa doença possuem susceptibilidade a adquirir doenças auto-imunes, neoplasias com grande risco de desenvolver câncer gástrico e distúrbios digestivos, sendo a manifestação mais comum a diarréia recorrente na forma de esteatorréia em 20%, seguida de enteropatia grave em 10% dos casos (ERRANTE, 2008). Segundo Pires et al. (2013), no tratamento , utiliza-se imunoglobulina endovenosa para diminuir o número de infecções. Essa aplicação deve ser realizada de três a quatro semanas de maneira contínua com dose inicial de 400-600 mg/kg, isso garante que o IgG permaneça acima de 500 mg/dL.

Para auxiliar o diagnóstico de algumas variações dessas imunodeficiências primárias, inclusive a ICV, tem-se utilizado recursos computacionais por meio de técnicas de bioinformática e com recente potencialização na exploração de técnicas de aprendizado de máquinas.

Assim, esse projeto de pesquisa é de natureza exploratória e baseou-se em estudos de conceitos da Imunodeficiência Comum Variável, diferentes técnicas de aprendizado de máquina, como o Naive Bayes, Redes Neurais Artificiais e Máquina de Vetores-Suporte, e na realização de um levantamento bibliográfico sobre sobre aprendizado de máquina aplicado no descobrimento de doenças raras, com o estudos de duas Bibliotecas, o Selene e o Expecto. Levando em consideração que esta pesquisa ainda não chegou ao fim, este relatório apresenta resultados parciais.

MATERIAL E MÉTODOS

Inicialmente foi realizado um levantamento bibliográfico referente as imunodeficiências primárias, com especial atenção à ICV por ser a de maior incidência, buscando entender qual a origem, o diagnóstico e tratamento dessa enfermidade. Posteriormente, foi realizado o estudos de três técnicas de aprendizado de máquina, classificador Naive Bayes, Redes Neurais Artificiais e Máquinas de Vetores-Suporte. Por fim, tem-se estudado o funcionamento de duas bibliotecas o Selene (Chen, et al., 2019) e a Expecto (Zhou et al., 2018) e que agora entra em etapa de execução de testes com a biblioteca Selene utilizando base dados gênicas gratuitas e abertas disponibilizadas na Internet para compreender seu funcionamento.

RESULTADOS E DISCUSSÃO

Os estudos de conceitos associados à aprendizado de máquina a serem utilizados em aplicações de análises de dados associados à doenças raras de imunodeficiência possibilitaram compreender diferentes classificadores baseados Naive Bayes, Redes Neurais Artificiais e Máquinas de vetores de suporte, incluindo dois *frameworks* de análise genética para o descobrimento de doenças Selene e Expecto.

Assim, verificou-se que as análises de genomas por algoritmos de aprendizado de máquina requerem treinamento em sequências de dados para que sejam capazes de prever o impacto e doenças causadas por mutações. Entretanto, esses tipos de análises requerem conhecimentos específicos sobre *machine learning* e o desenvolvimento de novos códigos para atenderem a bioinformática (CHEN, 2019) como, por exemplo, os disponibilizados pelos *frameworks* o Selene e o Expecto.

O Selene é um *framework* aplicado de Redes Neurais Artificiais de aprendizado profundo que fornece um apoio aos cientistas biomédicos um apoio abrangente no treinamento, avaliação e aplicação de modelos. Esse *framework* possui duas estruturas de auxílio ao pesquisador, na primeira ele pode criar ou modificar modelos para aplicações específicas e, no segundo, ele pode utilizar o modelos novos ou já existentes para a previsão e visualização dos impactos de uma mutação.

Em Chen (2019), o autor menciona quatro casos de usos do Selene. No primeiro caso de uso, um pesquisador de câncer deseja utilizar o modelo DeepSea para modelar o elementos do fator de transcrição GATA1, focando em um recurso genômico específico do tecido que o DeepSEA não prevê. Com o Selene o pesquisador pode treinar esse modelo sem escrever linha de código alguma e obtém uma avaliação de desempenho. No segundo caso, ele utiliza o novo modelo para aplicar uma mutagênese *in silico* e analisar as figuras geradas pelo *framework*. No terceiro caso, um pesquisador pode querer aprimorar um modelo já existente, o Selene compara o desempenho de seu novo modelo com o modelo original e publica-o. Por fim, no quarto caso, um geneticista humano estudando a doença de Alzheimer quer aplicar o modelo desenvolvido no estudo de caso anterior, para avaliar sua capacidade de priorizar as variantes associadas à doença.

O ExPecto é um *framework* baseado na técnica *deep-learning* prevê com precisão *ab initio,* a partir de uma sequência de DNA, os efeitos transcricionais específicos de tecido de mutações, inclusive, aqueles que são raros ou que não foram observados (ZHOU, 2018). Ele é utilizado para modelagem e pode prever os níveis de expressão gênica de sequências para mais de 200 tecidos e tipos de células, sendo seu grande diferencial a previsão para variantes comuns e até mesmo raras, posto que em seu treinamento não utiliza-se informação de variantes (ZHOU, 2018). Em Zhou(2018), é descrito um estudo da utilização ExPecto priorizando a análise de supostas variantes causais associadas a traços e doenças humanas, avaliando experimentalmente supostas variantes causais para a doença de Crohn, colite ulcerativa, doença de Behçet e infecção pelo vírus da hepatite B (VHB).

CONCLUSÕES

A associação das áreas de aprendizado de máquina e na análise genética surge com o objetivo de facilitar o trabalho de detecção de variantes causais de doenças, inclusive as tidas como raras, posto que esses tipos de atividade requerem muito tempo investido na obtenção conhecimentos específicos e desenvolvimento de algoritmos. O principal resultado dessa mesclagem são as bibliotecas que visam otimizar esse esse trabalho, como as que foram objetos de estudo neste projeto, Selene e o ExPecto, que utilizam a construção de modelos de Redes Neurais Artificiais de Aprendizado profundo para a análise de expressão gênica. Agora essa equipe concentra-se em realizar testes com a biblioteca Selene, analisando suas funcionalidades e possibilidades de aplicação, com a utilização de dados genéticos gratuitos disponibilizados na Internet.

AGRADECIMENTOS

Os autores agradecem o IFSP pelo apoio financeiro via o programa PIBIFSP.

REFERÊNCIAS

CHEN, Kathleen M. et al. Selene: a PyTorch-based deep learning library for sequence data. **Nature Methods**, [s.l.], v. 16, n. 4, p.315-318, 28 mar. 2019. Springer Nature. http://dx.doi.org/10.1038/s41592-019-0360-8.

ERRANTE, Paolo R.; CONDINO-NETO, Antonio. Imunodeficiência comum variável: revisão da literatura. **Rev bras alerg imunopatol**, v. 31, n. 1, p. 10-18, 2008.

Ministério da Saúde. **Doenças raras: o que são, causas, tratamento, diagnóstico e prevenção.** Disponível em: <http://www.saude.gov.br/saude-de-a-z/doencas-raras>. Acesso em: 16 maio 2019.

TORRES, J. et al. Diarreia num doente com imunodeficiência comum variável: a propósito de um caso clínico. **Jornal Português de Gastrenterologia**, v. 14, n. 4, p. 199-203, 2007.

ZHOU, Jian et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. **Nature Genetics**, [s.l.], v. 50, n. 8, p.1171-1179, 16 jul. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41588-018-0160-6.