O USO DE ÁRVORES DE DECISÃO PARA EXPLICAR OS RESULTADOS DE ALGORITMOS DE INTELIGÊNCIA ARTIFICIAL OPACOS

Julia Machado Lechi

Prof. Dr. Luciano Antonio Digiampietri

Escola de Artes, Ciências e Humanidades / Universidade de São Paulo

julia.lechi@usp.br; digiampietri@usp.br

Objetivos

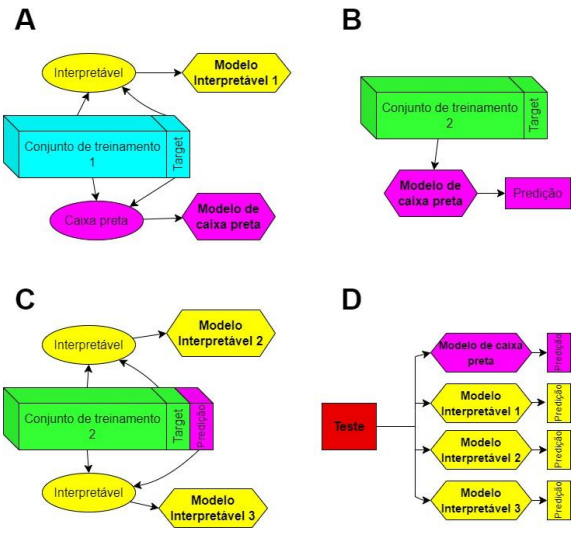
Na sociedade atual, conhecida como sociedade da informação, os algoritmos de inteligência artificial estão ganhando espaço na tomada de decisões. Projetados para analisar grandes volumes de dados de empresas e aplicativos, oferecem percepções e recomendações em diversas áreas. Por exemplo, na área financeira, existem diversas empresas que utilizam algoritmos para auxiliar no processo de concessão ou não de financiamentos. Também existem empresas que usam modelos de IA em seus setores de recursos humanos. Além disso, vários serviços públicos e de segurança usam diferentes modelos de IA, incluindo algoritmos de reconhecimento facial, como parte de seus procedimentos. [1, 2, 3, 4]. O objetivo do presente trabalho é analisar o desempenho de Árvores de Decisão que são algoritmos de inteligência artificial considerados inerentemente explicáveis ou interpretáveis tanto na tarefa de explicar os resultados de algoritmos opacos como também em tentar construir modelos explicáveis com base nos resultados de modelos de caixa preta.

Métodos e Procedimentos

O desenvolvimento deste projeto foi composto por: Revisão bibliográfica sobre trabalhos correlatos; Seleção de oito conjuntos de dados públicos; Seleção de métricas para avaliar a classificação, como métricas mais ligadas ao desempenho e à fidelidade de modelos; Seleção dos algoritmos usados como opacos; Pré-processamento e implementação das ferramentas utilizadas da biblioteca Python Scikit Learn, no ambiente Colab da Google.

Para os oito conjuntos, os dados categóricos foram convertidos em dados numéricos. Após o pré-processamento, houve a separação aleatória em três subconjuntos: 40% para o primeiro conjunto de treinamento, 40% para o segundo conjunto de treinamento e 20% para o conjunto de teste. A necessidade de dois conjuntos de treinamento distintos se justifica porque em diversas situações reais, o desenvolvedor não têm acesso aos dados que produziram o modelo caixa preta.

Figura 1 - Representação do método utilizado



O algoritmo Árvore de Decisão foi usado em dois contextos: como modelo explicável com altura máxima três e como modelo caixa preta (ou opaco) com valores padrão para os parâmetros. Os demais algoritmos usados como caixa preta são: Random Forest; SVM (com kernel linear e polinomial); Logistic Regression; Multilayer Perceptron; Gaussian Naive Bayes; and KNN.

Resultados

O primeiro conjunto de treinamento teve o objetivo de analisar os desempenhos entre o modelo interpretável e os modelos opacos, comparando-os, usando as métricas de precisão e medida F1 macro. O Random Forest foi o que apresentou a melhor precisão em sete dos oito conjuntos de dados, enquanto o modelo interpretável teve, em média, 93,2% do desempenho do melhor modelo para cada conjunto de dados. Considerando a medida F1 macro, o melhor desempenho foi do algoritmo Random Forest novamente, obtendo melhor resultado em seis dos oito conjuntos de dados, enquanto para o modelo interpretável, teve em média, correspondente a 90,5% do desempenho do melhor modelo para cada conjunto de dados. Assim demonstrando que, de fato, os modelos caixa-pretas são melhores em comparação aos explicáveis, entretanto, há pouca diferença entre os dois.

Sobre a fidelidade do modelo interpretável, foi avaliada a habilidade de imitar os resultados produzidos por modelos opacos, tendo em média, a fidelidade em 88,9%, e considerando cada modelo opaco, a maior fidelidade do modelo interpretável foi em relação ao SVM usando kernel linear (94,6%). O modelo de árvore de decisão tratado como caixa preta teve o pior valor para o F1-score e acurácia, mostrando sua ineficácia, no contexto deste projeto, para outro uso além de um modelo interpretável.

Conforme a Figura 1C, modelos interpretáveis foram construídos a partir da saída de modelos de caixa preta (chamado de "modelo interpretável 3"), cuja fidelidade foi apresentada anteriormente. Além da fidelidade, levantou-se a hipótese de que esses modelos interpretáveis construídos das saídas de modelos de caixa preta podem superar os modelos interpretáveis treinados diretamente com a variável alvo do conjunto de treinamento. A análise comparativa entre os modelos interpretáveis 2 e 3, ambos construídos no conjunto de treinamento 2, mostrou que, os modelos interpretáveis obtidos das saídas dos modelos de caixa preta tiveram resultados superiores de acurácia e medida F1 em relação aos modelos baseados na variável alvo. Os resultados sugerem ser promissora a construção de modelos explicáveis a partir das saídas de modelos de caixa preta, indicando um caminho interessante para investigações.

Conclusões

Com base nos resultados, verificou-se a esperada diferença de desempenho entre os modelos caixa preta e o modelo interpretável. Isto, para muitos problemas, serve de justificativa para o uso de modelos caixa preta. Uma capacidade satisfatória por parte do modelo interpretável para imitar o comportamento do modelo caixa preta também foi observada. A fidelidade, em média, foi maior do que a precisão do modelo para resolver o problema em questão.

Referências

[1] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine Bias, p. 11. **Auerbach Publications, New York, NY (2022)**, https://doi.org/10.1201/9781003278290

[2] Coelho, J., Burg, T.: **Uso de inteligência artificial pelo poder público (2020) Transparência Brasil.**

[3] Francisco, P.A., Hurel, L.M., Rielli, M.M.: **Regulação do reconhecimento facial no setor publico (2020)**, https://igarape.org.br/regulacao-do-reconhecimento-facialno-setor-publico/

[4] Ramos, S. **Retratos da violência - cinco meses de monitoramento, análises e descobertas. Rede de Observatórios da Segurança**/CESeC 1(1), 1–72 (11 2019),https://cesecseguranca.com.br/textodownload/retratos-da-violencia-cincomeses-de-monitoramento-analises-e-descobertas/