Um problema ao usar machine learning na neurociência

João Avelar Lobato

Diretor de Pesquisa da Cooper & Sacks
j.lobato@cooperandsacks.org

Abstrato

Este artigo examina a relevância de altos níveis de acurácia (accuracy) em estudos de neurociência que utilizam a machine learning para prever psicopatologias ou resultados de intervenções. Em uma simples experiência, mostramos como os algoritmos podem “dar sentido” dados aleatórios, retornando resultados tecnicamente corretos mas falhos. Neste sentido, é essencial questionar altos índices de precisão em amostras pequenas, a fim de usar machine learning de uma forma que possa influenciar com segurança a escolha de intervenções e práticas clínicas.

Introdução

Vários artigos relatam mais de 70% de acurácia na identificação de psicopatologias através da aplicação de machine learning em dados como variação da freqüência cardíaca (Byun et al., 2019), atividade eletrodérmica (Kim et al., 2018), imagens cerebrais (Gao et al., 2018, de uma revisão da RM), imagens de conectividade cerebral (Liu et al., 2022), análise da fala (Lin et al., 2022), e também dados derivados de questionários (Haque et al., 2021).[1] Embora não critiquemos nenhum estudo específico, perguntamos quão válida é uma taxa de acurácia de 70% ou mesmo 80%+ em uma amostra pequena, o que é comum em neurociência.

Um estudo recente mostrou, por exemplo, que os 1038 estudos de neuroimagem mais citados nas últimas três décadas tinham um tamanho de amostra de apenas 12. Enquanto os estudos em 2017 e 2018 tinham em média um tamanho de amostra maior, mas ainda assim 23 e 24 (Szucs e Ioannidis, 2020). Em uma revisão dos estudos que utilizaram machine learning para distinguir indivíduos autistas e não autistas, Vabalas et al. (2019) indicaram que cerca da metade dos estudos tinham uma amostragem abaixo de 80. Os autores também disseram que amostragem pequenas estão associadas a **maior** acurácia.

Métodos

[1] Consulte também Aleem et al. (2022) para uma revisão de dados e modelos utilizados na detecção de depressão.

Para explorar esta questão, treinei sete algoritmos sem ajustes em um pequeno conjunto de dados que pode ser encontrado em https://osf.io/yhk2p/[2] . O conjunto de dados consiste em dados de 26 indivíduos, um dos quais foi excluído por não ter um “label” (rótulo), de pais de crianças com autismo que tinham completado um treinamento na web. O conjunto de dados tem quatro características: renda familiar, grau mais avançado dos pais, o funcionamento social da criança e as notas de base do uso de intervenções comportamentais em casa por parte dos pais. O rótulo é um campo binário, com 0 indicando nenhuma melhoria do comportamento da criança e 1 melhoria. O conjunto de dados foi projetado para avaliar "os efeitos de um treinamento interativo na web para ensinar aos pais procedimentos analíticos de comportamento para reduzir comportamentos desafiadores em crianças com desordem do espectro do autismo" (Turgeon & Lanovaz, 2020, tradução DeepL).

O conjunto de dados é altamente enviesado (skewed), com 17 (ou 68%) das 25 rótulos pertencentes a classe 1, o que significa que um algoritmo que simplesmente produza 1 obteria uma taxa de precisão de 68%, como mencionado pelos autores do artigo. No artigo (versão pré-impressa), a taxa de acurácia de um random forest foi de 77%.

Executei sete modelos (regressão logística, análise linear discriminante, classificador K neighbors, gaussian naive bayes, SVC, random forest, e XGBoost) no conjunto de dados. Apenas um deles teve uma acurácia abaixo de 70% (análise linear discriminante). A XGBoost teve uma precisão de 88%. O código comentado está disponível em https://tinyurl.com/4xzbvzws.

Será que isso significa que os dados são bons indicadores de quais famílias serão beneficiadas pelo treinamento? Afinal, quão relevante é uma acurácia em um conjunto de dados tão pequeno e enviesado? A previsão pode realmente nos dizer quem se beneficiará mais com o treinamento em autismo?

Para responder a isto, executei os mesmos algoritmos acima, mas desta vez usando um **conjunto aleatório de rótulos** no teste, onde cerca da metade deles pertencia a um 0. Fiz isto 39 vezes, criando 39 conjuntos de rótulos aleatórios. Em 26 (67%) dos casos, pelo menos um algoritmo teve uma acurácia de mais de 70%.

Talvez isso não seja uma surpresa, pois alguns dos rótulos, mesmo que aleatórios, poderiam se correlacionar com uma das características, dado o pequeno tamanho da amostra. No entanto, era preocupante que em tantos casos pelo menos um algoritmo tivesse uma acurácia acima de 70% quando usando rótulos **aleatórios.**

Em seguida, usei rótulos aleatórios **tanto** para **treinamento quanto para teste**. Ou seja, os algoritmos foram treinados em dados puramente aleatórios. Eu criei 40 grupos de rótulos

[2] Os autores que o criaram advertiram que amostras maiores são mais adequados, mas não exploraram as implicações

aleatórios. Em 18 (45%) deles, pelo menos um algoritmo teve uma acurácia superior a 70%. Quando os rótulos de treinamento e teste foram estratificados, em 23 (58%) das vezes pelo menos um algoritmo tinha uma acurácia superior a 70%. Não houve ajuste de parâmetros em nenhum ponto, o que poderia ter aumentado ainda mais o desempenho dos algoritmos.

Provavelmente foi novamente devido ao pequeno tamanho da amostra, pois as variáveis poderiam simplesmente coincidir com os rótulos aleatórios. O ponto-chave, no entanto, é que **os algoritmos de machine learning são tão poderosos que podem fazer sentido de dados espúrios.** Esta não é uma afirmação nova, mas pode levar a resultados enganosos. Estudos podem tratar características como preditores **confiáveis** do resultado de uma intervenção ou classificar patologias quando na realidade não o são.

Em mais um passo, criei **40 conjuntos de dados aleatórios** com seis características aleatórias que variam de 0 a 9 e rótulos binárias [0, 1], com um tamanho de amostra de 30. Apesar de serem alimentados com dados puramente aleatórios, pelo menos um modelo teve uma acurácia de 70% ou mais em 26 (65%) dos casos, ou 17 (43%) quando os rótulos não foram estratificados. Isso significa que você poderia criar questões ridículas, como o número de letras no nome das pessoas, o comprimento de seu antebraço, ou pedir-lhes para escolher um número entre 0 e 9, e usar essas características para prever com alta acurácia se elas estão deprimidas ou não, por exemplo. Você poderia então dizer corretamente (mas enganosamente) que um algoritmo teve 70% ou até 80% de acurácia e relacionar características espúrias com a depressão.

Mesmo quando com um tamanho de amostra de 80, os modelos tiveram uma acurácia acima de 70% em 10% a 15% dos casos. Ao criar rótulos irregulares, com 60% delas representando uma classe, o que é comum em experimentos reais, e diminuindo o nível de acurácia para 65%, os algoritmos fizeram sentido de 17 (43%) dos conjuntos de dados. Aumentando o número de variáveis para 64 e alterando a sua amplitude não alterou significativamente a proporção de acurácia acima de 70%.

Conclusão

A afirmação de que um alto grau de acurácia em um pequeno tamanho de amostra não é necessariamente significativo é antiga. No entanto, ela tem sido ignorada em um grande número de estudos. É preocupante que características sem sentido possam influenciar a escolha de intervenções e práticas clínicas. O simples experimento acima mostrou que altas taxas de acurácia podem ser encontradas mesmo nos casos mais extremos de algoritmos sendo treinados em dados aleatórios. Obviamente, isto está ligado à natureza estocástica dos dados gerados, mas isto não parece ser levado em conta em muitos estudos de neuroscience.

Aleem et al. (2022) discutiram um estudo que teve 97,54% de acurácia na detecção de pessoas com depressão usando uma amostra de apenas 66 (58% dos quais estavam deprimidos, o que significa que os rótulos não estavam equilibrados). Qualquer profissional

estaria preocupado com uma acurácia tão alta em uma amostra pequena, particularmente porque 58% dos rótulos representavam um grupo.

Embora machine learning seja uma ferramenta que pode (e já está) apoiando a pesquisa neurocientífica, protocolos precisam ser desenvolvidos para entender melhor seus potenciais e limitações.

Código

O código pode ser encontrado em:
https://github.com/j-lobato/machine-learning-and-neuroscience

O código comentado está disponível em https://tinyurl.com/4xzbvzws [3]

Referências

Aleem, S., Huda, N. U., Amin, R., Khalid, S., Alshamrani, S. S., & Alshehri, A. M. (2022). Algoritmos de Aprendizagem de Máquinas para Depressão: Diagnóstico, Percepções e Direções de Pesquisa. Eletrônica, 11(7), 1111. https://doi.org/10.3390/electronics11071111

Byun S, Kim AY, Jang EH, Kim S, Choi KW, Yu HY, Jeon HJ (2019). Detecção de grande desordem depressiva de características de variabilidade linear e não-linear da freqüência cardíaca durante o protocolo de tarefa mental. Computadores em biologia e medicina. Recuperado em 17 de março de 2023, de https://pubmed.ncbi.nlm.nih.gov/31404718/

Gao, S., Calhoun, V. D., & Sui, J. (2018, novembro). Aprendizagem da máquina em grande depressão: Da classificação à previsão dos resultados do tratamento. Neurociência &amp do SNC; terapêutica. Recuperado em 17 de março de 2023, a partir de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6324186/

Haque, U. M., Kabir, E., & Khanam, R. (2021). Detecção de depressão infantil usando métodos de aprendizagem de máquinas. PLOS ONE, 16(12), e0261131. https://doi.org/10.1371/journal.pone.0261131

Lin, Y., Liyanage, B. N., Sun, Y., Lu, T., Zhu, Z., Liao, Y., Wang, Q., Shi, C., & Yue, W. (2022). Um modelo baseado no aprendizado profundo para detectar a depressão na população idosa. Frontiers in Psychiatry, 13. https://doi.org/10.3389/fpsyt.2022.1016676

[3] Devido à natureza estocástica dos modelos e ao fato de que o código irá gerar novos dados, os resultados podem ser ligeiramente diferentes

Liu, Z., Wong, N. M., Shao, R., Lee, S., Huang, C., Liu, H. A., Lin, C. P., & Lee, T. M. (2022). Classificação dos principais distúrbios depressivos usando a aprendizagem mecânica sobre a estrutura cerebral e a conectividade funcional. Journal of Affective Disorders Reports, 10, 100428. https://doi.org/10.1016/j.jadr.2022.100428

Kim, A. Y., Jang, E. H., Kim, S., Choi, K. W., Jeon, H. J., Yu, H. Y., & Byun, S. (2018, 19 de novembro). Detecção automática de distúrbio depressivo maior usando atividade eletrodérmica. Notícias da Natureza. Obtido em 17 de março de 2023, em https://www.nature.com/articles/s41598-018-35147-3

Szucs, D.; Ioannidis, J. (2020, 15 de julho). Evolução do tamanho da amostra na pesquisa de neuroimagens: Uma avaliação de estudos altamente citados (1990-2012) e das práticas mais recentes (2017-2018) em revistas de alto impacto. NeuroImagem. Recuperado em 13 de março de 2023, a partir de https://www.sciencedirect.com/science/article/pii/S1053811920306509

Turgeon, S., & Lanovaz, M. J. (2020). Tutorial: Aplicando o aprendizado de máquinas na pesquisa comportamental. Perspectives on Behavior Science, 43(4), 697-723. https://doi.org/10.1007/s40614-020-00270-y