Instituto de Matemática e Estatística - USP

MAC0459/MAC5865 - Data and Engineering Science

General Test - 2021 - QUESTÃO 3:

A CONFIANÇA EM ALGORITMOS DE MACHINE LEARNING: Sobre o artigo "Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making "por Suresh et. al.(2020)

Edilson Pereira dos Santos - N° USP:

RESUMO: O artigo ao qual este ensaio se propõe tratar, apresenta estudo de medição de confiança de usuários em algoritmos de Machine Learning. Para tal, baseou-se em experimento onde os participantes da pesquisa foram apresentados à dois tipos de questionários estruturados a partir de imagens de multidões e animais e suas respectivas recomendações realizadas por algoritmos de ML, tendo que, para a primeiro questionário, informar a quantidade de seres humanos presentes nas imagens e para a segundo questionário, informar a semelhança entre os animais das fotos. Foram usados questionários em formato de webapp com participantes encontrados em redes sociais como Facebook e Instagram. Os modelos de ML utilizados durante o estudo basearam-se em ResNet50 para previsão de semelhança entre animais em formato binário e MCNN para avaliação de imagens de multidões. O artigo conclui que independente do nível de conhecimento do usuário em técnicas de ML, existe uma tendência humana a superestimar as previsões e recomendações realizadas por estes algoritmos.

Palavras-Chave: Confiança, Machine Learning, subjetivismo.

ARGUMENTO PRINCIPAL

Existe uma certa fascinação entre o público em geral com relação aos algoritmos de Machine Learning. Em parte por conta das grandes vantagens funcionais que estas técnicas apresentam para a vida das pessoas, em outra medida, por conta do 'aroma' de novidade que estes algoritmos sugerem.

Todavia, nenhum sistema tecnológico é perfeito e livre de erros e como esperado, mesmo os algoritmos com maior precisão não podem ser considerados integralmente confiáveis.

O artigo em análise de modo metódico tenta medir a confiança dos usuários em relação aos algoritmos de Machine Learning, tarefa árdua, pois a medição da confiança de um ser humano não é tão simples como parece em um primeiro instante e embora exista grande carga de trabalhos nesta área, principalmente, por parte da psicologia, existem subjetivismos naturais inerentes ao processo de medição conforme observam *C. Castelfranchi e R. Falcone* [1], em continuidade, o rigor ao se conduzir um experimento desta magnitude exige cuidados fundamentais para que não haja desvios de interação por parte dos participantes.

MÉTODOS E CONCLUSÕES

O estudo avaliou três áreas do conhecimento (raciocínio lógico, matemática e machine learning), a fim de, inicialmente, medir o grau de conhecimento dos participantes selecionados. Além disso, foi utilizado um conjunto de perguntas baseadas em tarefas para medir a confiança com recomendações reais de ML e material explicativo sobre os sistemas de ML e recomendações para cada tarefa/pergunta. A primeira fase foi composta por 22 questões de avaliação de conhecimento divididas em 3 categorias: Lógica (3 questões), Matemática (8 questões) e ML (11 questões). Os autores esclarecem que as questões foram elaboradas pela equipe que conduziu o experimento devido à falta de material disponível neste sentido.

Para a área de lógica, os temas abordados foram: Correlação e causalidade, lógica da árvore de decisão e estabelecer relacionamentos entre eventos. As questões de matemática cobriram as seguintes áreas de conceito: ajuste de curva, variância de uma amostra, probabilidade, funções lineares, média e mediana, modelos gráficos, falsos positivos e negativos, e pontos em um hiperplano. Os assuntos abordados para a área de Machine learning foram: definição de ML , problema gerais de ML, aprendizagem supervisionada e não supervisionada, treinamento versus dados de teste, SVM e funções doe kernel,Gradiente descendente, clusterização por k-means, transferência de aprendizagem, regularização, otimização e modelagem de tópicos.

Na segunda fase do estudo, os participantes responderam às questões com ou sem recomendações de ML. Os webapps utilizados para os questionários, fundamentalmente, atuaram de modo a induzir os participantes a apontar imagens de animais de aparência semelhante e comparar o número de pessoas em fotos de multidões. Vídeos explicativos foram utilizados como apoio para algumas das questões dos questionários, o critério de atribuição não foi revelado pelos autores. Ademais, estes vídeos tiveram duração entre 15 e 45 cada, sendo 2 vídeos de dados (1 para cada webapp), 2 vídeos de ML (1 por webapp), 8 vídeos de desempenho (1 por questão) e 32 vídeos de previsão (1 para cada questão).

A análise dos resultados foi baseada em ANOVA unilateral e no teste de Tukey-Kramer (todos os pares, Tukey HSD) entre os tratamentos. Usamos um nível

alfa de 0,05 para todos os testes estatísticos. Todos os valores de p são significativos para o número de comparações de pares independentes usando a correção de Bonferroni. Os autores sintetizam sua conclusões conforme a seguir:

As pessoas geralmente seguem as recomendações de ML e se estas recomendações são acompanhadas por vídeos explicativos, esta tendência sobe; As pessoas seguem recomendações incorretas de ML para tarefas que fariam de maneira correta se não tivessem recomendações, incluindo especialistas. Embora os participantes tenham seguido recomendações incorretas em geral, eles foram seguidos com, significativamente, menos frequência do que as recomendações corretas, e recomendações incorretas e/ou anormais foram seguidas significativamente menos do que recomendações incorretas normais. Nenhum modelo, dados, desempenho ou informações de previsão melhoraram a precisão das pessoas, mesmo para recomendações anormais incorretas em que as características do vídeo de previsão ou da imagem sugeriram que a recomendação deveria ser menos confiável.

ANÁLISE CRÍTICA

As métricas utilizadas durante o experimento ao qual o artigo analisado se debruça apresentam forte embasamento, além disso, o artigo em si como um todo faz uso do método científico com todo o rigor que lhe cabe. Onde foi possível, procurou-se aleatorizar questões, participantes e os diversos aspectos dos métodos. Prosseguindo, o arcabouço estatístico utilizado para as análises pertinentes também seguiu o rigor acadêmico exigido para um artigo desta importância. Suas conclusões são assertivas e convincentes, entretanto, conduzir o experimento de forma remota pode ter contribuído negativamente para os resultados.

A este respeito, não garantir que os participantes da pesquisa estivessem em lugar tranquilo e adequado para a resolução dos questionários de modo a que sua atenção estivesse concentrada na tarefa pode ter gerado certo grau de descomprometimento com o estudo. Em continuidade, *Daniel J. Levitin*[2] esclarece em sua obra *'A Mente organizada'* sobre a alternância do modo de foco e devaneio do cérebro humano, traz também informações sobre o funcionamento do 'filtro de atenção cerebral' e como estes três processos podem interferir profundamente no modo como fazemos tarefas.

Além disso, é esclarecido como um ser humano tende a ser menos comprometido com uma atividade quando não é supervisionado. Neste contexto, não adotar medidas que pudessem induzir os participantes a estados de concentração sem dúvida alguma pode ter comprometido o estudo como um todo. Mas de um cuidado para este fim poderia ter sido tomado e um simples exercício caberia bem a este fim: Indução de estados de concentração baseados em respiração consciente como ilustrado por *Renan Vivas Zanotto*[3], induz o cérebro humano a uma maior capacidade de se ater à tarefas através do aumento da quantidade proporcional de dopamina, endorfina, ocitocina e serotonina na corrente sanguínea.

Contudo, entende-se que o experimento adotado pelo estudo foi realizado em meio a grande pandemia de covid-19 no ano de 2020 e que os cuidados de conduzir a pesquisa remotamente se deram por estes motivos, principalmente. Em conclusão, para um estudo futuro deve ser pensado com mais cuidado, fatores como a infraestrutura e ambiente e a preparação dos participantes para a pesquisa, considerando aspectos fisiológicos da faculdade de atenção do cérebro humano para obter o máximo de foco dos envolvidos.

Referências Bibliográficas

[1] C. Castelfranchi e R. Falcone - **A confiança é muito mais do que uma probabilidade subjetiva**: componentes mentais e fontes de confiança. 33ª conferência internacional do Havaí de ciência de sistemas (HICSS2000). 2020. volume 6.

[2]LEVITIN, D. J. **A mente organizada**: Como pensar com clareza na era da sobrecarga de informação (Edição em português). 1ª ed. 2014. 464 p.

[3] - ZANOTTO, R. V. - **Significado da prática da reaprendizagem respiratória para participantes de um programa orientados**. 2020. 50 p.