A INTELIGÊNCIA ARTIFICIAL E A FASE DOS PORQUÊS

O recente desempenho notável da Inteligência Artificial (IA), suportado em técnicas de Aprendizagem Profunda, criou expectativas sobre o seu potencial transformador positivo na sociedade. Contudo, as questões éticas e morais ressurgiram também com grande intensidade. A IA interpretável emerge como uma resposta parcial a estas preocupações e à necessária melhoria contínua.

JAIME S. CARDOSO[1,2],
LUÍS F. TEIXEIRA[1,2]
[1] INESC TEC;
[2] FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO
jaime.cardoso@inesctec.pt
luis.f.teixeira@inesctec.pt

A chamada "idade dos porquês" é um período clássico e habitual no desenvolvimento das crianças. A criança, ávida de conhecer o mundo que a rodeia, começa a questionar o adulto sobre tudo o que quer compreender; o adulto, com paciência e respeito, ajuda-a a esclarecer as suas dúvidas, contribuindo assim para o seu processo de aprendizagem.

Será excitante quando a Inteligência Artificial (IA) desempenhar esse papel do adulto e nós, o da criança que quer aprender. Quando a inteligência artificial estiver suficientemente desenvolvida poderá, explicando as suas decisões, contribuir para o nosso próprio crescimento intelectual.

Até lá, ainda existe um longo caminho a percorrer. A IA ainda erra. Por isso, muito do trabalho atual de interpretar a decisão automática tem como objetivo compreender o erro para melhorar o algoritmo de decisão e aumentar a nossa confiança na máquina. É interessante verificar que, em alguns domínios 'fechados', apesar de decidir globalmente bem, muito bem, a máquina comete erros 'infantis' e é facilmente manipulada. Este comportamento estatisticamente positivo, mas com casos individuais aberrantes, levanta dúvidas sobre os conceitos que o algoritmo integrou; são dúvidas que importa dissipar e ultrapassar.

Muito do trabalho atual em IA versa os ditos algoritmos de Aprendizagem Profunda (Deep Learning). A Aprendizagem Profunda é uma área específica da aprendizagem automática, onde os algoritmos de aprendizagem geram modelos a partir dos padrões encontrados nos exemplos que são processados.

Uma das diferenças mais evidentes da Aprendizagem Profunda é que, para além de serem aprendidos modelos de decisão, também são aprendidos modelos de representação dos dados. Ou seja, é aprendido um modelo que transforma os dados de entrada, por exemplo uma imagem, numa representação abstrata de conceitos representativos dessa imagem. O desempenho alcançado por estes algoritmos é notável, sendo o estado da arte em vários domínios, por exemplo em análise de imagem médica, capaz de desafiar os especialistas nas suas próprias áreas. Há, no entanto, um obstáculo à interpretação do processo de decisão desses modelos - a sua opacidade. A sua elevada complexidade e elevada abstração tornam a decisão automática de difícil interpretação pelos humanos, sejam estes especialistas ou leigos em medicina ou em IA.

Numa tentativa de ultrapassar essa dificuldade, a IA interpretável tenta justificar uma decisão com base em informação complementar. Por exemplo, podem ser evidenciadas as regiões da imagem mais relevantes para a tomada de decisão. Os algoritmos de 'interpretação' fornecem um mapa de relevância (representado por uma imagem) onde são identificadas as zonas que condicionaram a decisão. O cálculo destes mapas pode ser feito de formas diversas, mas sempre tentando imputar a responsabilidade da decisão às diferentes entradas do modelo. Por exemplo, se o modelo previu cancro com probabilidade de 80% baseado num mamograma, os modelos de interpretação vão tentar responsabilizar as diferentes regiões da imagem por essa decisão. Que regiões contribuíram de forma mais significativa para a previsão de cancro? Existem várias partes interessadas na resposta a esta questão: os especialistas em IA que desenvolvem e treinam os modelos e os utilizadores finais do modelo, por exemplo os radiologistas. Nem sempre a mesma interpretação da decisão é igualmente útil para todos os consumidores da interpretação e por isso é necessário adequar as técnicas da IA interpretável a quem vai tirar partido desta informação. Por exemplo, a informação visual poderá não ser suficiente e outras formas de explicação são úteis para a melhor compreensão do processo de decisão, tais como um texto descritivo ou um conjunto de exemplos similares. Atualmente, estas interpretações, apesar de ainda bastante básicas, já são úteis para diagnóstico e melhoria do próprio algoritmo de IA. Se a interpretação realçar na imagem zonas fora do pulmão como relevantes para um diagnóstico de pneumonia baseado num raio-X, provavelmente o algoritmo de IA, mesmo tendo decidido bem, terá "raciocinado" mal. Se a análise de um algoritmo de AI de apoio ao recrutamento de recursos humanos revela que este está a favorecer os homens em relação às mulheres, há um viés que importa remover.

O Workshop em "iMIMIC - Interpretability of Machine Intelligence in Medical Image Computing", que organizámos no dia 4 de outubro como parte da conferência International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), foi revelador da dinâmica da área, do potencial e das virtudes destas abordagens. Contudo, foi também revelador das limitações e do muito que ainda falta fazer. Sendo as explicações / interpretações elas próprias geradas por um algoritmo de IA automático, também este tem limitações e erros. As explicações também podem ser manipuladas. Por exemplo, um algoritmo pode usar a origem do cliente como característica que condiciona a concessão de crédito e uma explicação que não a usa para explicar a decisão. Noutra direção, é importante generalizar as explicações para casos para além da classificação. Como explicar que a previsão para o valor de venda da casa é de €437,52K e não outro valor qualquer? Qual é a explicação adequada neste caso? Ainda noutra direção, como explicar uma decisão suportada simultaneamente em múltiplas fontes de informação (áudio, texto, vídeo)?

A área da interpretabilidade da IA está a dar os primeiros passos. Ainda nos falta um longo caminho pela frente, a ser trilhado com otimismo, passo a passo. Este progresso conjunto, ora focado na melhoria da decisão, ora focado na explicação da decisão, está a permitir um crescimento mútuo das soluções para ambas as tarefas, em que todos saímos a ganhar. Não é utópico querer aprender com a máquina.