COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Judy Fox (University Virginia) January 30, 2024

Title: Expéditions : Recherche collaborative : Épidémiologie computationnelle omniprésente à l'échelle mondiale

Judy Fox CIC Database Profile

NSF Award #: 2151597

YouTube Recording with Slides

Winter 2024 CIC Webinar Information

Transcript Editor: Lauren Close

Transcript

Slide 1

Merci, Lauren, pour votre aimable présentation, et merci aux hôtes de 2024 COVID Information Commons et à nos collègues conférenciers, étudiants et membres du personnel. Je vais aujourd'hui vous présenter notre projet Expeditions, financé par la NSF. Il s'agit d'une recherche collaborative pour une épidémiologie computationnelle omniprésente à l'échelle mondiale. Je m'appelle Judy Fox. Je travaille à l'école de science des données et d'informatique de l'université de Virginie.

Slide 2

Il s'agit d'un projet multi-institutionnel dirigé par le Dr Madhav Marathe, qui en est le coordinateur principal. Nous avons des collègues à l'université de Virginie et au Biocomplexity Institute. Il s'agit d'une collaboration multi-institutionnelle et nous avons des collègues et des chercheurs formidables. C'est une expérience très stimulante pour moi.

Slide 3

Je voulais parler de l'avenir. C'est la partie la plus passionnante de ma présentation d'aujourd'hui. Imaginez COVID en 2025. Que serons-nous ? Nous voulons passer de l'intervention à la prévention, car les maladies infectieuses sont un problème de société. D'ici à 2050, on prévoit plus de 10 millions de décès par an et un impact économique de plus de 100 000 milliards de dollars. Il y a quelques années, nous sommes sortis d'une pandémie qui a fait plus de 1,1 million

de morts rien qu'aux États-Unis et qui compte plus de 100 millions de cas. Cela représente près d'un tiers des familles infectées. Une grande partie de cette situation aurait pu être évitée grâce à une politique gouvernementale mieux informée. Cependant, COVID-19 est un problème de données complexe. Tout d'abord, nous recevons des données non stationnaires. Il est très difficile d'apprendre et de prédire les tendances avec un manque de données, des données bruyantes. Rendre les prévisions d'infection compréhensibles ou explicables peut également aider à la prise de décision. Elles peuvent aider à identifier les zones géographiques et temporelles importantes afin que nous puissions signaler aux gouvernements une allocation plus efficace des ressources pour prévenir la propagation de la maladie.

Slide 4

Je souhaite axer le reste de mon exposé sur les recherches menées par mon groupe. Nous voulions interpréter les infections COVID-19 au niveau des comtés aux États-Unis. Nous avons appliqué le modèle Transformer AI, qui est un type de modèle d'apprentissage profond utilisé par les grands modèles de langage.

Slide 5

L'un des domaines sur lesquels nous nous concentrons est la question suivante : pourquoi avons-nous besoin de prédictions ? Les prédictions utilisant des données en temps réel ont été mises en évidence en 2009 par le Dr Harvey V. Fineberg et le Dr Mary Elizabeth Wilson. Ils soulignent l'importance de l'utilisation des données les plus récentes pour étudier la lutte contre les maladies et tenter d'observer et de prédire. Les interventions seront mises en œuvre au moment du pic, mais elles permettront d'aplanir la courbe avant l'heure.

Slide 6

Nous utilisons un modèle d'apprentissage profond, le Temporal Fusion Transformer (TFT), qui permet de faire des prévisions en temps réel. Dans nos expériences, nous utilisons les 13 derniers jours pour prédire les 15 jours à venir. Les données collectées proviennent de différents modèles de prédiction. Nous les classons en ensembles de données co-variantes statiques et dynamiques, comme les cas et les décès. Nous disposons également de données connues telles que les fêtes de fin d'année.

Slide 7

Avec un tel modèle, notre objectif est d'essayer de comprendre comment utiliser l'IA interprétable pour obtenir des connaissances et des informations sur le lieu et le moment où l'infection se produira. Quels sont les pays les plus exposés ? Quelles sont les communautés vulnérables ? Et nous essayons de les aider. Le parcours de cette étude est tout cela, avec de nombreux obstacles que nous devons surmonter.

Slide 8

Le problème vient de la question générale de la précision. Comment notre modèle peut-il prédire de manière à ce que nous soyons sûrs que notre prédiction est exacte ? De plus, lorsque nous avons cette prédiction, comment expliquer aux décideurs politiques quels sont les facteurs importants de l'augmentation actuelle du nombre de cas ? Pour cela, nous devons avoir une compréhension plus approfondie des données elles-mêmes à un niveau beaucoup plus fin, comme les caractéristiques au niveau du comté. Nous voulons également prendre une décision en temps réel afin de garantir sa pertinence.

Slide 9

Il s'agit d'une étude très riche, car la disparité en matière de santé n'est pas seulement liée aux personnes, mais aussi à l'impact socio-économique. Nous avons recueilli plus de deux ans et demi de données pour 3 142 comtés américains. Nous catégorisons et réduisons les caractéristiques de l'ensemble des décès de vingt à six caractéristiques. Deux d'entre elles sont statiques : la disparité en matière de santé et le groupe d'âge de la population. Il existe des caractéristiques observables, notamment la vaccination, la propagation de la maladie, les cas transmissibles et la mobilité. Nous incorporons des événements connus et inconnus, de sorte que nous disposons d'un modèle d'IA articulée multimodale très complexe.

Slide 10

Permettez-moi de passer à la vitesse supérieure et de vous présenter quelques-uns de nos résultats en matière de prédiction. Nous avons comparé le modèle TFT avec le LSTM, un modèle de base de séquence à séquence. Nous pouvons montrer que dans le graphique de gauche, le modèle TFT est le plus performant. Il donne le message d'erreur, la précision est plus élevée et l'erreur est plus faible.

Slide 11

Qu'est-ce qui sous-tend la compréhension de l'IA interprétable ? Cela vient du mécanisme de tension qui utilise une architecture codeur-décodeur et le mécanisme de tension est à la base du grand modèle de langage actuel, y compris ChatGPT. Grâce à ce mécanisme de multi-attention, nous pouvons saisir le contexte de la maladie au fil du temps, ce qui permet d'affiner l'espace dans lequel nous examinons la caractéristique - la cause et l'effet. Nous pouvons ainsi mettre l'accent sur l'importance des schémas spatiaux et temporels dans les zones sensibles. Sur la droite, vous pouvez voir l'architecture d'un modèle TFT. Nous saisissons les caractéristiques passées et essayons de prédire les événements futurs. Dans ce cas, ce sont les cas d'infection qui permettent d'intégrer les caractéristiques, en particulier les caractéristiques statiques, en utilisant un modèle de séquences pour capturer les modèles temporels fiables. Nous propageons tous ces

schémas à l'auto-attention pour tenter de masquer l'attention multi-têtes interprétable afin de pouvoir nous concentrer sur les schémas et les zones importants.

Slide 12

Ce modèle de prévision est en mesure de nous fournir le modèle cyclique qui rend compte des cas COVID. Il tient également compte des événements particuliers, tels que les vacances et les week-ends. Nous pouvons clairement les identifier dans ce graphique. Sur le côté droit, nous pouvons même regarder en arrière pour voir quelle période a le plus d'impact sur la prédiction future dans le laps de temps de 0 à 13 jours.

Slide 13

Voici le tableau des tendances que nous pouvons prédire. En choisissant les 100 comtés les plus peuplés, vous pouvez voir que nous avons une prédiction par rapport à la vérité de terrain. En outre, nous pouvons comparer les comtés moins peuplés - vous pouvez voir que le résultat correspond beaucoup mieux et qu'il y a moins de pics et de valeurs différentes.

Slide 14

Qu'en est-il des informations relatives à la localisation ? Ce graphique montre à gauche l'auto-attention du modèle d'IA et saisit l'intensité au niveau du comté. Cela n'est pas possible si nos données se situent au niveau de l'État. À droite, on trouve une représentation des données des cas cumulés provenant du CDC pour plus de 3 000 comtés américains. Si vous regardez ces deux résultats, vous pouvez voir la corrélation entre ces deux ensembles de données et les résultats.

Slide 15

Nous avons mesuré la corrélation et conclu que nous pouvons interpréter le modèle d'IA en saisissant les poids d'auto-attention au niveau du comté. Il existe une forte corrélation entre le comportement du modèle et la prédiction des cas par rapport à la vérité de terrain.

Slide 16

R2 fournit les informations dont les décideurs politiques ont besoin. Nous pensons qu'une petite réduction de la transmission dans les points chauds peut conduire à une forte réduction des infections, en particulier au stade précoce. Il est essentiel d'établir des prévisions en temps réel et de concentrer notre attention sur les régions les plus importantes en termes d'infections quotidiennes. Nous disposons d'une méthode fine pour détecter ces infections au niveau du comté, ce qui réduirait considérablement le risque. De nombreux travaux futurs peuvent être réalisés à partir de ce résultat existant. Nous pouvons explorer de nombreux impacts et disparités sociaux et économiques dans le cadre de travaux futurs.

Slide 17

À l'UVA, nous avons lancé le programme AI for science et attiré plus de 3 000 étudiants de premier cycle. Nous avons sélectionné une douzaine d'étudiants impliqués dans notre projet.

Slide 18

Nous souhaitons étudier plus avant la sensibilité des groupes d'âge de la population en utilisant l'étude de l'indice de Morris pour savoir, dans chaque groupe de population, qui est le plus vulnérable aux infections par COVID.

Slide 19

Nous avons choisi un modèle d'apprentissage profond de séries temporelles parce que le mécanisme de tension mentionné ci-dessus peut donner un aperçu et comprendre comment le modèle est prédit.

Slide 20

Nous avons tiré plusieurs leçons de ce voyage. Les maladies infectieuses sont un problème socio-économique mondial qui a un impact sur la santé publique et l'économie à grande échelle. La première leçon que nous avons tirée de la pandémie de COVID-19 est que les tests sont essentiels pour comprendre l'évolution de la pandémie. La deuxième est que de nombreuses infections sont très différentes d'une région à l'autre. La situation est très dynamique. La meilleure façon d'aborder cette question dans la politique est de l'adapter au niveau local. Nous pouvons nous améliorer à bien des égards, notamment en construisant les outils que nous avons étudiés pour prédire avec précision l'infection par le COVID et les maladies infectieuses à venir. Cela aidera les décideurs politiques à intervenir sur une base scientifique. L'intervention est l'avenir. Nous voulons être préparés et prêts pour les crises futures, telles que la pandémie.

Slide 21

Avant de conclure, je tiens à dire à tout le monde que nous devons instaurer la confiance. S'il y a des leçons à tirer de la dernière pandémie, nous voulons mieux expliquer au public, aux experts, aux décideurs politiques ce qui se passe et comment nous pouvons tirer parti de la politique pour avoir un impact. Nous voulons disposer d'une méthode interprétable pour modéliser et évaluer nos méthodes de manière quantifiable. Nous voulons également expliquer le comportement de notre modèle et les prédictions basées sur l'IA aux non-experts, y compris au public et aux étudiants. Collectivement, nous espérons construire un avenir qui nous permettra d'être prêts pour les événements futurs.

Slide 22

Sur ce, je tiens à vous remercier tous d'avoir participé à l'atelier d'aujourd'hui. Je présente ici quelques-uns de nos travaux. N'hésitez pas à me contacter. J'attends avec impatience les discussions qui auront lieu à la fin de cet atelier. Je vous remercie.