Axe 5 - Modélisation, modèles probabilistes et méthodes de recherche

Avner Bar-Hen (PRCM Cnam), Yvon Pesqueux (PRCM)et Michel Béra (PRCM CNAM)

Chercheur associé: Pr. Michael Spence ; **Didier Le Ruyet** (PR1 - CNAM), **Pr. Robert Cario**, **Pr. Martine Herzog-Evans**, **Pr. Jean-Philippe Denis**

Cet axe de recherche, dirigé par les Professeurs Bar-Hen, Pesqueux et Béra, s'intéresse à l'usage (ou au dévoiement) des méthodologies de recherche appliquées à des objets sécuritaires. Cet axe regroupe aussi bien le développement de nouvelles normes actuaires, le questionnement des modèles à large donnée (LLMs ; Avner Bar-Hen) ou de l'incidence des méthodes de gestion sur l'ordre public ou les organisations sociétales (Y Pesqueux).

Par exemple, Michel Béra examine l'importance de cette quantification et les exigences considérables en matière de calcul dans les grands ensembles de données (variables : ~100, individus : ~Ms) , et propose des mesures plus rapides à calculer qui pourraient servir d'indicateurs de la contribution des enregistrements au risque global extrême de réidentification (Ottawa conference paper)

Les données personnelles concernant de larges segments de la population (qu'il s'agisse d'êtres humains ou de personnes morales) prolifèrent dans les administrations publiques et privées. Elles sont reconnues comme un bien précieux à des fins politiques et commerciales.

Malheureusement, elles sont également précieuses à des fins malveillantes et font souvent l'objet de tentatives de compromission. Les législations sont très exigeantes : les dépositaires de bases de données personnelles/professionnelles peuvent faire l'objet de sanctions sévères, ainsi que d'une atteinte à leur réputation, en cas de vol de données. Les conséquences sont encore plus graves si l'identité de personnes physiques ou morales est révélée. Les dépositaires sont tenus de protéger l'identité des personnes concernées par les données, qu'elles soient ou non destinées à être partagées avec les autorités ou les chercheurs.

L'anonymisation et la « pseudonymisation » sont deux catégories de mesures destinées à protéger l'identité des personnes concernées. Alors que l'anonymisation est considérée comme irréversible et la pseudonymisation comme réversible, la prolifération d'énormes quantités de données générées et collectées par des appareils et des outils logiciels repousse la frontière entre les données anonymes et les données pseudonymes vers ces dernières. Ce qui est anonyme aujourd'hui ne le sera peut-être plus demain.

Selon ce point de vue, tous les ensembles de données comportent un risque de réidentification. La méthode QaR (AFNOR, 2020) propose une mesure du risque de réidentification d'un jeu de données et une technique statistique, basée sur la théorie des valeurs extrêmes, pour l'estimer. Ce risque a une grande valeur, comme le montrent les montants des dédommagements attribués dans les jugements de « class action suits » liés aux vols de données. Il permet d'évaluer l'efficacité du contrôle de la divulgation que les dépositaires appliquent aux données ; il pourrait être communiqué aux autorités réglementaires afin de démontrer le niveau d'attention accordé par les dépositaires à la protection de la vie privée des personnes concernées (Obligations SOLVENCY II pour les assureurs, obligations du RGPD article 35 pour les détenteurs de données); il peut être utilisé pour calculer une prime d'assurance contre la divulgation non autorisée ou le montant dont les dépositaires ont besoin dans leur bilan pour couvrir les dommages financiers potentiels dus à une telle divulgation. YES !!! quand on ne met pas de radar ou de porte blindée dans une maison, l'assurance qui est obligatoire coûte plus cher… les anglo-saxons ont très bien compris : plus que de faire de la techno, on frappe au portefeuille (primes d'assurance)

L'évaluation fiable de la confiance d'un réseau neuronal profond et la prédiction de ses défaillances sont d'une importance primordiale pour le déploiement pratique de ces modèles. Dans cet article, nous proposons un nouveau critère cible pour la confiance des modèles, correspondant à la probabilité de classe réelle (TCP).

Puisque la vraie classe est par essence inconnue au moment du test, nous proposons d'apprendre le critère TCP sur l'ensemble d'apprentissage, en introduisant un schéma d'apprentissage spécifique adapté à ce contexte. Des expériences approfondies sont menées pour valider la pertinence de l'approche proposée. Nous étudions différentes architectures de réseaux, des ensembles de données à petite et grande échelle pour la classification d'images et la segmentation sémantique. Nous montrons que notre approche surpasse systématiquement plusieurs méthodes solides, du MCP à l'incertitude bayésienne, ainsi que des approches récentes spécifiquement conçues pour la prédiction des défaillances.

Cet axe de recherche interroge également, autour de travaux d'Yvon Pesqueux, les dimensions ontologiques, épistémologiques et méthodologiques de la recherche appliquée en sécurité, défense, renseignement, criminologie.

Au-delà du débat relatif à la distinction entre l'approche qualitative et l'approche quantitative, cet axe de recherche aborde les questionnements suivants : L'ethnographie comme outil privilégié pour la production d'un savoir local socialement utile sur la base des éléments suivants: discussion de l'ouvrage « Les ficelles du métier » de H. S. Becker, l'importance de la tâche descriptive, l'enquête, le storytelling, l'étude de cas et la narration ; l'étude de cas, l'étude de cas longitudinale, la situation comme étude de cas. C'est un lieu de discussion des méthodes qualitatives "classiques", des techniques de la méthodologie qualitative (les techniques d'"observation directe, les techniques d'entretien, les méthodes de collecte des données dites « actives », la triangulation dans le cadre des méthodes de collecte, la validité interne et externe d'une recherche qualitative, l'observation participante), la théorie enracinée. L'axe méthodologie de recherche appliquée décrypte également les étapes d'une recherche qualitative (la pré-analyse, la phase d'analyse des données ou de codification, la catégorisation, la mise en relation et la représentation des résultats, la phase de la vérification des données) ; la sociologie de la traduction et la théorie de l'acteur réseau ; les Visual Studies ; la Méthode des Incidents Critiques (MIC) ; les méthodes de recherche art-based ; le journal de bord ; recherche-action et recherche

Questions de recherche associées

ØDeep Learning et Big Data : usages et limites pour la détection et la prévention des crimes

ØModélisation prédictive distribuée des engagements criminels (bascules, passages à l'acte, convergences comportementales)

ØApplication des modèles de données extrêmes à l'étude des phénomènes criminels

ØSingletons et « outliers » dans la prédiction des phénomènes de loups solitaires
ØDonnées aberrantes et données tronquées dans l'étude des p hénomènes criminels

https://esd.cnam.fr/axes-de-recherche/axe-5-modelisation-modeles-probabilistes-et-methodes-de-recherche-1242042.