Avis de Soutenance

Madame Anna Karen GÁRATE ESCAMILLA

Présentera ses travaux en soutenance

Thèse soutenue le ***mardi 26 mai 2020*** à 14h00
Lieu : 13 Rue Thierry Mieg, 90000 Belfort
Salle : I102

Titre des travaux : Big Data et machine learning pour améliorer le suivi et la télésurveillance médicale

Ecole doctorale : SPIM - Sciences Physiques pour l'Ingénieur et Microtechniques

Section CNU : 27

Unité de recherche : Laboratoire de Nanomédecine, Imagerie, Thérapeuthique

Directeur de thèse : Amir HAJJAM EL HASSANI

Codirecteur de thèse : ☐ HDR ☑ NON HDR

Soutenance : ☑ Publique ☐ A huis clos

Membres du jury :

Nom	Qualité	Etablissement	Rôle
M. Amir HAJJAM EL HASSANI	Maître de conférences	Université Bourgogne - Franche-Comté	Directeur de thèse
M. Vincent HILAIRE	Professeur des Universités	Université Bourgogne - Franche-Comté	Président
M. Dan ISTRATE	ECC	UTC	Examinateur
M. Emmanuel ANDRES	Professeur des universités – praticien hospitalier	CHRU Strasbourg - Université de Strasbourg	Examinateur
Mme Parisa GHODOUS	Professeur des Universités	Université Lyon 1	Rapporteure
M. Germain FORESTIER	Professeur des Universités	Université de Haute Alsace	Rapporteur

Mots-clés : Big Data,Apprentissage,esanté,Ingénieurie de la connaissance,

Résumé de la thèse (en français) :

The growth in medical data collection presents a new opportunity for physicians to improve patient diagnosis. In recent years, practitioners have increased the use of computer technologies to provide better decision-making support. One way to address this challenge is to build models with empirical data from large-scale health experiments, such as those that have been established in recent decades for some e-health systems, mobile applications and telemedicine. To this end, Machine learning is an analytical tool that identified patterns and relationships by learning from experience. This thesis used individual measurements from some datasets, including a heart disease dataset, a data set of pregnant women in the first trimester with hypothyroidism, and synthetic medical data (with more than 5,000 patients and 1,000,000 features). Specifically, I pursued the following objectives: (i) to optimize execution time and scalability of the large-scale health datasets predictions and to deploy the optimized setup (chapter 1); (ii) to predict whether a patient has heart disease by using classification techniques and to describe the features most correlated with the heart condition (chapter 2); and (iii) to use unsupervised learning models to enrich medical ontologies and generate new knowledge based on the search of similar characteristics and symptoms (chapter 3). After establishing the required databases, I performed different machine learning models that related the correlation of the features (i.e. hearth disease related to cholesterol, heart rate and chest pain). Finally, I created the clusters for pregnant women in the first trimester with hypothyroidism (i.e. the cluster using thyroid pathology, risk, anthropometric, gynecological factors). My results reveled that: (i) the logistic regression presented an optimal execution time and scalability without being affected by complex computational operations; (ii) cache and persist methods are powerful in reducing the consumption time; (iii) PCA outperformed the results after using chi-square, otherwise, when PCA is used directly from the raw data, the performance is poor; (iv) cholesterol, maximum heart rate, chest pain and heart vessels are the anatomically and physiologically relevant features of heart disease; and (v) cluster analysis proved to be a practical approach for the heterogeneity of the hypothyroidism risk factors in women in the first trimester of pregnancy in clinical studies, identifying three clusters: women over 30 years lacking of signs and symptoms of thyroid hypofunction, women under 30 years without signs and symptoms of thyroid hypofunction, and women under 30 years with some risk factors and signs or symptoms suggesting thyroid hypofunction. My results underline that dimensionality reduction techniques and machine learning improved the detection of abnormal situations in the context of medical remote monitoring and show that the models are not affected by complex operations.

Abstract (in English):

L'évolution de la quantité de données médicales collectée offre aux médecins une nouvelle opportunité d'améliorer le diagnostic des patients. Ces dernières années, les praticiens ont accru l'utilisation des technologies informatiques afin de fournir une meilleure aide à la décision. Une façon de relever ce défi est la construction des modèles avec des données empiriques issues d'expériences de santé à grande échelle, comme celles qui ont été établies au cours des dernières décennies pour certains systèmes de santé en ligne, pour des applications mobiles ou encore pour la télémédecine. L'apprentissage est un outil analytique qui permet d'identifier des modèles et des relations en tirant les leçons de l'expérience. Cette thèse a utilisé des mesures individuelles provenant de certaines ensembles de données, notamment un ensemble de données sur les maladies cardiaques, bases de données sur les femmes enceintes au premier trimestre souffrant d'hypothyroïdie et des données médicales synthétiques (avec plus de 5 000 patients et 1 000 000 de caractéristiques). Plus précisément, les objectifs de cette thèse sont : (i) optimiser le temps d'exécution et l'extensibilité des prévisions des bases de données sanitaires à grande échelle et déployer une configuration optimisée, (ii) prédire si un patient souffre d'une maladie cardiaque en utilisant des techniques de classification et décrire les caractéristiques les plus corrélées avec l'état cardiaque et (iii) utiliser des modèles d'apprentissage non supervisés pour enrichir les ontologies médicales et générer de nouvelles connaissances

basées sur la recherche de caractéristiques et de symptômes similaires. Après avoir établi les connaissances nécessaires, j'ai réalisé différents modèles d'apprentissage qui mettaient en relation les caractéristiques (c'est-à-dire les maladies du cœur liées au cholestérol, au rythme cardiaque et aux douleurs thoraciques). Par ailleurs, j'ai créé les clusters pour les femmes enceintes au premier trimestre souffrant d'hypothyroïdie (c'est-à-dire le cluster utilisant la pathologie thyroïdienne, le risque, les facteurs anthropométriques, gynécologiques). Mes résultats ont révélé cela : (i) la régression logique présentait un temps d'exécution et une évolutivité optimaux sans être affectés par des opérations de calcul complexes, (ii) les méthodes de mise en cache et de persistance sont puissantes pour réduire le temps de consommation, (iii) l'ACP a donné de meilleurs résultats après utilisation du chi carré, sinon, lorsque l'ACP est utilisée directement à partir des données brutes, les performances sont médiocres, (iv) le cholestérol, la fréquence cardiaque maximale, les douleurs thoraciques et les vaisseaux cardiaques sont les caractéristiques anatomiques et physiologiques pertinentes des maladies cardiaques et (v) l'analyse par groupes s'est révélée être une approche pratique pour l'hétérogénéité des facteurs de risque d'hypothyroïdie chez les femmes au cours du premier trimestre de la grossesse dans les études cliniques. Pour ce dernier point, les trois groupes identifiés sont : les femmes de plus de 30 ans ne présentant pas de signes ni de symptômes d'hypofonctionnement de la thyroïde, les femmes de moins de 30 ans ne présentant pas de signes et symptômes d'hypofonctionnement de la thyroïde et les femmes de moins de 30 ans présentant certains facteurs de risque et des signes ou symptômes suggérant un hypofonctionnement de la thyroïde. Mes résultats soulignent que les techniques de réduction de la dimensionnalité et l'apprentissage ont amélioré la détection de situations anormales dans le cadre d'une télésurveillance médicale et montrent que les modèles ne sont pas affectés par des opérations complexes.