Titre de la thèse

Analyse de données patient par apprentissage machine ; application au contexte d'infections microbiales persistantes.

Thesis supervisor and supervisory team:

Encadrement	
Nom	Shawky
Prénom	Marc
Ecole doctorale	Université de Technologie de Compiègne (UTC)
Laboratoire	COSTECH (Connaissance Organisation et Systèmes TECHniques), EA2223, UTC,
Adresse	Centre de Recherches, Rue Personne de Roberval, 60200 Compiègne, France
Email	shawky@utc.fr
Phone	+33 3 4423 4423

Co-encadrement	
Nom	Jollivet-Courtois
Prénom	Pascal
ED	UTC
Laboratoire	COSTECH (Connaissance Organisation et Systèmes TECHniques), EA2223, UTC,
Professional address	Centre de Recherches, Rue Personne de Roberval, 60200 Compiègne, France
Email	Pascal.jollivet-courtois@utc.fr
Phone	03 44 23 44 34

Co-encadrement	
Nom	Dr. Zedan
Prénom	Ahed

Profil candidat

Ingénieur informatique ou Data Scientist. Connaissances en apprentissage machine appréciées mais pas indispensables.

Processus de candidature selon programme ISCD de Sorbonne Université.

RESUME

Ce projet de doctorat vise à appliquer l'apprentissage automatique pour analyser 300 à 500 éléments de données provenant de patients suspectés d'être atteints d'infections à long terme, y compris à transmission vectorielle. Les données sont constituées de signes cliniques, analyses sanguines biochimiques, tests sérologiques, rapports textuels d'IRM et de tomodensitométrie, et d'autres données épidémiologiques. L'objectif est de développer des algorithmes de classification capables de mettre en évidence les effets des infections à long terme par un seul vecteur ou plusieurs vecteurs.

La première phase du projet relèvera le défi d'obtenir des informations fiables et complètes à partir d'ensembles de données provenant de formulaires en ligne remplis par les patients sous la supervision d'un médecin ou d'assistants médicaux. Les algorithmes d'apprentissage automatique seront adaptés pour traiter les données cliniques subjectives, nuancés selon leur amplitude, leur fréquence et leur évolution dans le temps.

Le projet visait initialement à fournir aux cliniciens un résultat de diagnostic binaire, mais le travail a été réorienté pour fournir d'autres résultats d'analyse, tels que la « distance » ou les « similitudes » entre les profils des patients, qui peuvent aider à choisir la meilleure approche de traitement.

Le système imitera également le comportement du médecin pendant le processus de diagnostic, comme demander potentiellement des examens supplémentaires plutôt que de fournir un diagnostic simple.

Dans la phase suivante, les données de traitement seront ajoutées à la base de données des patients, qui impliquent le traitement de données capturées à différentes dates. L'ensemble de données d'apprentissage sera constamment mis à jour avec les données des nouveaux patients confirmés. Les méthodes de classement utilisées doivent rester simple, en commençant par K plus proches voisins ou les classifieurs/régresseurs SVM.

Le projet explorera également l'apprentissage par renforcement pour identifier les caractéristiques et paramètres de la pathologie considérée. Nous évaluerons son utilisation pour réduire la durée du transfert d'expertise pendant la phase d'apprentissage supervisé du système. Le projet évaluera également la manière dont les patients perçoivent ce processus de diagnostic assisté.

Les travaux futurs porteront sur la capacité d'un système d'apprentissage automatique à justifier ses décisions, un problème connu sous le nom d'« explicabilité », particulièrement délicat dans les implémentations d'apprentissage profond.

This PhD project aims to apply machine learning to analyze 300 to 500 data elements from patients suspected of long-term infections, including vector-borne. The data consists of

clinical signs, biochemical blood analyses, serological tests, MR and CT scan text reports, and epidemiological data. The goal is to develop classification algorithms that can highlight the effects of long-term infections with a single vector or multiple vectors.

The first phase of the project will address the challenge of obtaining reliable, complete datasets from standard online forms filled out by patients under the supervision of medical assistants. The machine learning algorithms will be adapted to process subjective clinical signs, which are nuanced according to amplitude, frequency, and evolution in time.

The project initially aimed to provide clinicians with a binary result, but this did not meet their expectations. The work was reoriented to provide other analysis outcomes, such as the "distance" or "similarities" between patients' profiles, which can help choose the best treatment approach. The system will also mimic the physician's behavior during the diagnosis process, potentially asking for further exams rather than providing a straightforward diagnosis.

In the following phase, treatment data will be added to the patient database, which will involve processing data captured at different timestamps. The learning dataset will be constantly updated with data from new confirmed patients. The classification methods used should remain simple, starting with K neighbor or SVM classifiers/regressors.

The project will also explore reinforcement learning to identify the most relevant features and parameters for the considered pathology. We will assess its use in reducing the duration of expertise transfer during the supervised learning phase of the machine learning system. The project will also evaluate how patients perceive this assisted diagnosis process.

Future work will address the ability of a machine learning system to justify its decisions, an issue known as "explainability", particularly tricky in deep learning implementations.

DIRECTION DE THESE / MAIN SUPERVISOR

Marc Shawky, Pr, COSTECH Laboratory, Université de Technologie de Compiègne

DESCRIPTION OF THE PROJECT

Machine learning (ML) as a scientific field combining statistics, data processing and knowledge analytics has brought definite advances in medical image analysis with astounding diagnosis help to clinical physicians. The area of application of machine learning extended progressively to the analysis of other patient data.

In this project, we present a PhD subject to analyze from 300 to 500 data elements of patients, suspected of long-term infections, including vector-borne. We focus on the type of microbial infections where treatment decision does not only depend on serological tests, but for a large extent, upon reactivation revealed by clinical signs.

The data comprise clinical signs as expressed by the patient through validated questionnaires, biochemical blood analyses, serological tests, MR and CT scan text reports, epidemiological data, etc.

The classification algorithms to develop should highlight the effects of long-term infections with a single vector, or with associations of multiple vectors. The first phase of the work will tackle the difficulty of having reliable, complete data sets even through standard online forms filled by the patients under the supervision of medical assistants.

In order to improve the efficacy of machine learning algorithms for this type of analysis, we need to adapt them to process subjective clinical signs; a type of data which is nuanced according to three characteristics: amplitude, frequency and evolution in time.

Our database will be fed with data coming from centres of infectious diseases at the national level, including Polyclinique Saint Côme, Compiègne, CHIV of Villeneuve St. Georges and the CHU Raymond-Poincaré, Garches. We're currently signing an agreement with ILADS network of institution/physicians to have access to data from other clinics in Boston, Dublin, Montreal, etc. This will lead us to tackle anonymous patient data sharing between different research institutes without the need for the data to travel between countries. This smart remote data processing on site will also preserve the access rights of the original data (3).

In the first version of this undergoing work (1), we aimed to provide the clinicians with a positive or negative binary result. Progressively, we noticed that this does not fulfill their expectations. On one hand, we reoriented the work to obtain other analysis outcomes, like the "distance" or "similarities" between patients' profiles, which is particularly useful in order to choose the best treatment approach comparatively with patients already treated. In the same time, it is a challenge for the selection of the "revolving" learning and validation datasets. We need to consider also state of the art techniques for "continuous learning" in ML approaches (6). This function has to be designed and developed during this PhD.

On another hand, we aim to let the ML system mimic the physician behavior during the diagnosis process, yielding for example output decisions asking for further exams, rather than just providing a straightforward diagnosis.

Hence, and according to the need of our clinicians, in the following phase, the treatment data has to be added to the patient database. This implies to involve also the follow-up data, i.e. to get the patients fill updated versions of the questionnaires. This means that we will process data "captured" at different time stamps, yielding another challenge to the coherency of learning and validation datasets. Furthermore, the learning dataset will be constantly updated with data of new confirmed patients, which represents an interesting issue according to the state of art of the continuous automatic learning. However, the classification methods that will be used should remain simple. We suggest to start with K neighbor or SVM classifiers/regressors (4), (5), (6).

We will also address reinforcement learning approach, in order to identify the most relevant features for the considered pathology, together with the most impacting parameters. We will also assess its use on reducing the duration of expertise transfer that the clinician has to devote during the supervised learning phase of our ML system (2), (7).

In cooperation with colleagues from our Human Sciences lab., we aim also to evaluate how this assisted diagnosis process by ML will be perceived by patients. Currently, we observed that patients welcome entering their complaints through standard questionnaires, when they realize that they are recognized as "legitimate" symptoms (8).

As prospective trends, we wish also tackle the ability of an ML system to justify its decisions, issue usually known as "explainability", particularly tricky in deep learning implementations.

Finally, this database and the developed functions will represent a key tool to help national but also international clinicians in providing a reliable and inexpensive diagnosis for persistent microbial infections, especially when they have limited experience with the suspected disease.

Keywords

Artificial Intelligence for medical diagnosis, machine learning based diagnosis, supervised and non-supervised learning, long-term infection diagnosis.

References

1. BOLLE-REDDAT, Chloé, GUERIN, Mickaël, PADIOLLEAU-LEFEVRE, Séverine, OCTAVE Stéphane, BIHAN-AVALLE, Bérangère, MAFFUCCI, Irene, ZEDAN, Ahed, SHAWKY, Marc, Experience of patient data analysis with long term infections using Machine Learning, *ILADS 2023,* October 2023.
2. CHUMACHENKO, Dmytro, PILETSKIY, Pavlo, SUKHORUKOVA, Marya, et al. Predictive model of lyme disease epidemic process using machine learning approach. *Applied Sciences, 2022*, vol. 12, no 9, p. 4282.
3. GUÉRIN, Mickaël, SHAWKY, Marc, ZEDAN, Ahed, et al. Lyme borreliosis diagnosis: State of the art of improvements and innovations. *BMC microbiology*, 2023, vol. 23, no 1, p. 204.
4. KOHLI, Pahulpreet Singh et ARORA, Shriya. Application of machine learning in disease prediction. *In : 2018 4th International conference on computing communication and automation (ICCCA). IEEE*, 2018. p. 1-4.
5. PEIFFER-SMADJA, Nathan, RAWSON, Timothy Miles, AHMAD, Raheelah, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, 2020, vol. 26, no 5, p. 584-595.
6. PIANYKH, Oleg S., LANGS, Georg, DEWEY, Marc, et al. Continuous learning AI in radiology: implementation principles and early applications. Radiology, 2020, vol. 297, no 1, p. 6-14.
7. UDDIN, Shahadat, KHAN, Arif, HOSSAIN, Md Ekramul, et al. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making,* 2019, vol. 19, no 1, p. 1-16.
8. JOLLIVET, P., DUARTE, A.-B. Symbolic and Connectionist Artificial Intelligence: A Technical and Scientific Controversy? Socio-Economics in a Transitioning World: Breaking Lines and Alternative Paradigms for a New World Order. *Society for the Advancement of Socio-Economics (SASE)*, Rio de Janeiro, Brésil (ACL)