Université Claude Bernard  Lyon 1

Etat de l'art :

Feature Selection Partial Multi-Label Learning

CAHIER DES CHARGES

LEFEBVRE Julien

P2105454

MERCIER Loris

P1906860

Encadrant : Mr. Khalid BENABDESLEM

Unité d'Enseignement : MIF11 – Ouverture à la recherche

Année : 2022 - 2023

1 Présentation du contexte global

Dans un monde de plus en plus interconnecté, les datas prennent aujourd'hui une place centrale dans le monde informatique et plus généralement dans le monde de l'entreprise. Avec de nombreuses applications dans les domaines de la santé, de l'industrie ou de la finance, les différentes facettes de l'analyse et du traitement de données offrent aux entreprises ou laboratoire de recherche de nouvelles perspectives de développement.

Branche de l'intelligence artificielle, l'apprentissage de données (dit *machine learning* en anglais) est un regroupement de méthode permettant à nos ordinateurs d'engranger de la connaissance à partir de données afin d'analyser et prédire le comportement de nouvelles données. Il utilise pour cela différentes approches que l'on peut regrouper en quatre grandes branches : **l'apprentissage supervisé**, **non-supervisé**, **semi-supervisé** ou par **renforcement**. La principale différence entre ces concepts réside sur la présence ou non de données « étiquetées », c'est-à-dire un ensemble de données marqués où l'on connaît donc déjà le résultat attendu.

[1]

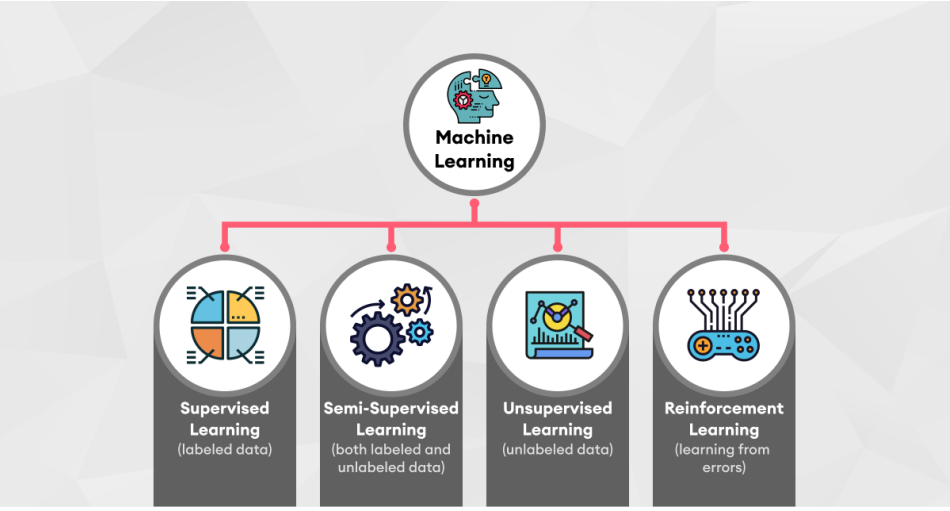


Figure 1: Schéma des différentes branches du Machine Learning

A noter qu'une donnée peut posséder plusieurs étiquettes. On parle alors de Multi-Label Learning. Une donnée X en entrée correspond alors à plusieurs étiquettes Y en sortie. Pour aller plus loin, ces étiquettes peuvent aussi être incomplètes ou incertaines, on est alors dans cadre du **Partial Multi-Label Learning (PML),** une approche très récente qui n'en est pour l'heure qu'au stade de recherche.

C'est d'ailleurs dans le cadre de cette nouvelle approche que nous réaliserons notre projet de recherche. Encadré par Monsieur Khalid BENABDESLEM, maître de conférences au Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) dans l'équipe Data Mining and Machine Learning, notre sujet d'ouverture à la recherche porte sur la sélection de variable (= choix d'un sous ensemble de variables parmi les entrées X) dans l'approche Partial Multi-Label Learning.

[1] Crédit image : https://blog.superannotate.com/supervised-learning-and-other-machine-learning-tasks/

2 Objectifs du projet

Le principal objectif de ce projet de recherche consiste à :

« *Expliquer les méthodes de sélection de variable dans le cadre de l'approche semi-supervisé Partial Multi-Label Learning (PML).*[2] »

Concrètement, cette étude vise à réaliser un état de l'art sur le domaine PML par le prisme de la sélection de variable. Il sera alors important d'éclaircir par quelles méthodes les algorithmes de prédiction PML nettoient et combinent les variables initiales afin d'optimiser leur performance.

En parallèle de cet état de l'art, un objectif complémentaire orienté programmation est attendu : l'obtention de statistiques de performances détaillées des algorithmes de Partial Multi-Label Learning.

3 Travail à réaliser

En lien avec notre encadrant de projet, l'objectif principal de notre recherche se traduit par la réalisation progressive de trois états de l'art. D'abord une vision globale de la **sélection de variable** et de **l'approche multi-label**, puis **l'application de ces deux concepts dans l'approche PML**. En complément, nous devrons réaliser une étude comparative des algorithmes PML à l'aide de tests statistiques.

Tache 1 : Etat de l'art au sujet des sélection de variables

Tâche primordiale dans les processus de machine Learning, la sélection de variable consiste à supprimer les données redondantes et/ou non pertinentes pouvant amener à réduire les performances de nos modèles de prédiction.

Le premier état de l'art présentera le rôle de la sélection de variable dans l'apprentissage de données puis les différentes stratégies de mise en œuvre. Il devra détailler plusieurs métriques de sélection regroupées, entre autres, parmi les méthodes de filtres, de wrapper ou encore l'approche embedded.

Tache 2 : Etat de l'art sur l'approche multi-label

Le multi-label learning est une approche de l'apprentissage de données où une donnée initiale X peut être associée à plusieurs labels Y en sortie. Formellement, cela se traduit par un algorithme mappant des entrées X en vecteur binaire (*0,1*) Y.

Le travail est ici de comprendre comment les algorithmes de prédiction analysent les données multi-étiquetés et comment ceux-ci s'en servent pour apprendre. Ces recherches feront office d'un deuxième état de l'art et mentionneront une liste non exhaustive de méthodes utilisées.

[2] Objectif défini en lien avec notre encadrant de recherche.

Tache 3 : Etat de l'art sur l'approche Partial Multi-Label Learning (PML)

Issu de l'approche multi-label du mode semi-supervisé, l'approche PML repose sur des données multi-labélisés partielles. Autrement dit, l'algorithme traite des données associées à un ensemble de labels candidats, dont plusieurs sont pertinents et quelques-uns sont bruités et/ou incertains.

A travers l'ensemble des connaissances acquises lors des deux premiers états de l'art, le travail consiste à expliquer les mécanismes de sélection de variables dans l'approche PML. Dans le cadre d'un nouvel état de l'art, il sera essentiel de mentionner les travaux de recherche en cours ainsi que les enjeux de cette nouvelle approche.

Tache 4 : Etude comparative des approches PML

A partir du recensement d'algorithmes effectué en parallèle des états de l'art, plusieurs tâches sont attendues pour réaliser l'étude comparative :

- Recherche d'une base de données PML sur lequel exécuter nos algorithmes
- Etablissement d'un protocole expérimental de test
- Réalisation des tests
- Etude et présentation des résultats

Nous tenterons ensuite de proposer une innovation aux algorithmes PML en réalisant notre propre algorithme tirant profit des forces et faiblesses constatées lors de notre phase d'étude.

4 Contraintes

Les états de l'art devront être réalisés séparément et pouvoir être compris de manière unitaire. Ils seront réalisés de manière pédagogique et contiendront une recherche bibliographique approfondies.

L'étude comparative s'effectuera sous le langage Python. Elle portera uniquement sur l'approche PML à partir d'une base de tests commun à toutes les méthodes. L'étude sera hébergée sur la forge de l'université Lyon 1 et sera ouverte au partage.

Concernant l'organisation, plusieurs rendus intermédiaires seront attendus au cours des prochains mois. (*Date à préciser par le responsable d'UE*). Il est donc essentiel de rédiger proprement les états de l'art au fur et à mesure des recherches afin de pouvoir livrer les travaux dès que demandés.

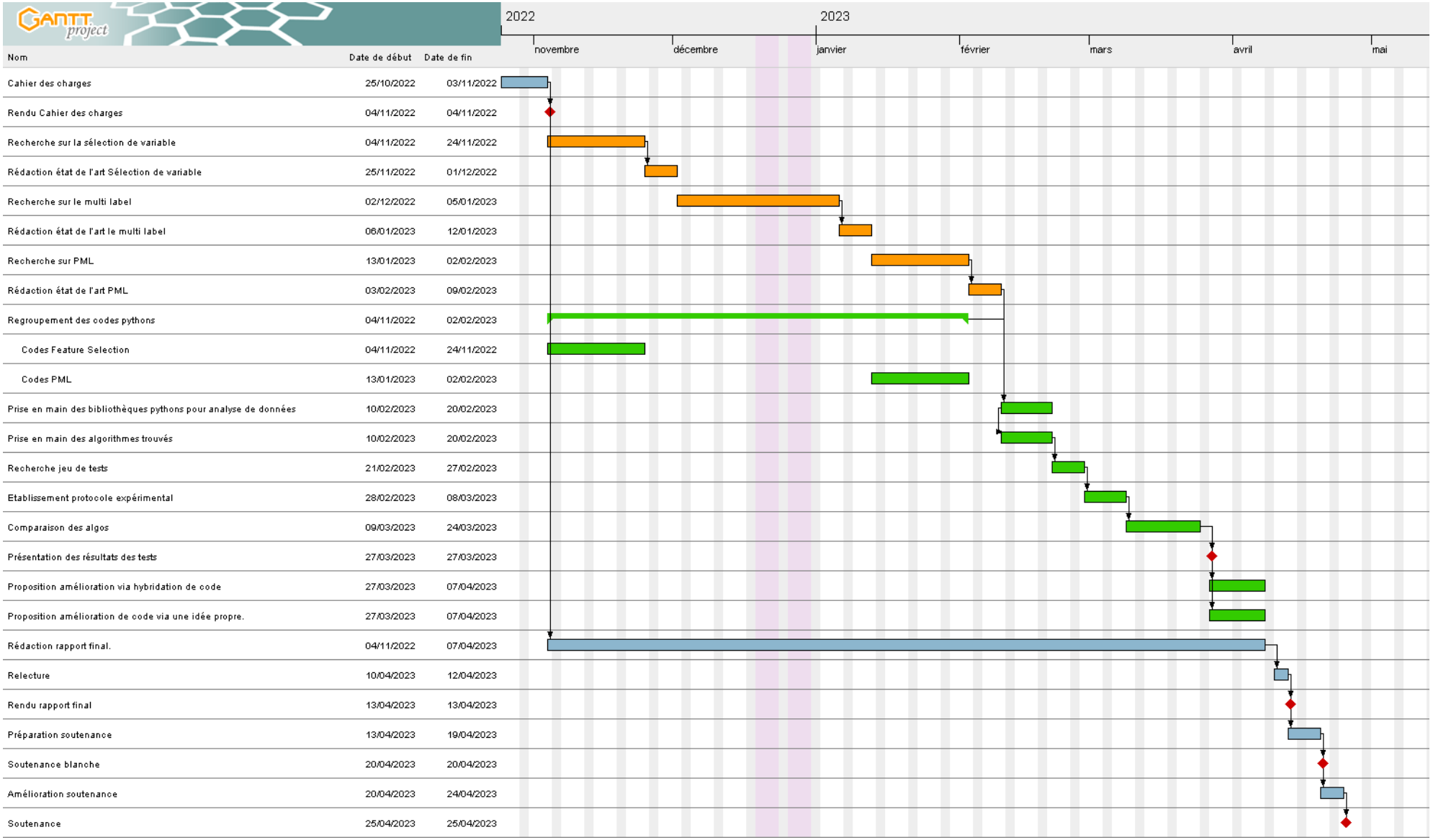
Un rapport de projet sera à produire d'ici fin avril 2023 accompagné par la suite d'une vidéo de vulgarisation[3].

5 Calendrier prévisionnel (Diagramme de Gantt)

(Voir annexe ci-après)

[3] A l'heure où nous écrivons ce cahier des charges, nous n'avons reçu aucune information sur l'organisation et les critères de cette vidéo. Nous ne détaillerons donc pas ce travail dans le calendrier prévisionnel.

Annexe 1 : Diagramme de Gantt



Nom	Date de début	Date de fin
Cahier des charges	25/10/2022	03/11/2022
Rendu Cahier des charges	04/11/2022	04/11/2022
Recherche sur la sélection de variable	04/11/2022	24/11/2022
Rédaction état de l'art Sélection de variable	25/11/2022	01/12/2022
Recherche sur le multi label	02/12/2022	05/01/2023
Rédaction état de l'art le multi label	06/01/2023	12/01/2023
Recherche sur PML	13/01/2023	02/02/2023
Rédaction état de l'art PML	03/02/2023	09/02/2023
Regroupement des codes pythons	04/11/2022	02/02/2023
Codes Feature Selection	04/11/2022	24/11/2022
Codes PML	13/01/2023	02/02/2023
Prise en main des bibliothèques pythons pour analyse de données	10/02/2023	20/02/2023
Prise en main des algorithmes trouvés	10/02/2023	20/02/2023
Recherche jeu de tests	21/02/2023	27/02/2023
Etablissement protocole expérimental	28/02/2023	08/03/2023
Comparaison des algos	09/03/2023	24/03/2023
Présentation des résultats des tests	27/03/2023	27/03/2023
Proposition amélioration via hybridation de code	27/03/2023	07/04/2023
Proposition amélioration de code via une idée propre.	27/03/2023	07/04/2023
Rédaction rapport final.	04/11/2022	07/04/2023
Relecture	10/04/2023	12/04/2023
Rendu rapport final	13/04/2023	13/04/2023
Préparation soutenance	13/04/2023	19/04/2023
Soutenance blanche	20/04/2023	20/04/2023
Amélioration soutenance	20/04/2023	24/04/2023
Soutenance	25/04/2023	25/04/2023