Université Claude Bernard  Lyon 1

# État de l'art (C):

# Sélection de variables *Multi-label (ML) et Multi-label partiel (PML)*

LEFEBVRE Julien

P2105454

MERCIER Loris

P1906860

**Encadrant :** Mr. Khalid BENABDESLEM

**Unité d'Enseignement :** MIF11 – Ouverture à la recherche

**Année :** 2022 - 2023

# Table des matières



# Table des illustrations

# Introduction

La sélection de variables est une tâche primordiale à effectuer en amont de quelconques algorithmes d'apprentissage. Elle consiste à identifier puis sélectionner les variables les plus pertinentes et non redondantes afin d'améliorer la précision du modèle tout en réduisant sa complexité.

Jusqu'alors, nous avions vu son utilité dans le cadre d'une approche supervisée et semi-supervisée mono-label. Néanmoins, la sélection de variables peut être bien plus complexe lorsqu'elle entre dans le paradigme multi-label, c'est-à-dire, un monde où chaque instance est associée à plusieurs labels cibles.

C'est d'ailleurs à partir de ce constat que nous allons vous présenter dans cet état de l'art les techniques de sélection de variables en mode multi-label. Après avoir introduit cette notion et en avoir expliqué ses principales caractéristiques, nous détaillerons les différentes méthodes de sélection de variables supervisées selon la taxonomie filtre, wrapper, embedded. Fort de cette approche, nous nous concentrons ensuite sur l'approche semi-supervisée découlant en partie des algorithmes présentés auparavant. Nous évoquerons enfin un domaine très récent n'étant qu'à l'état embryonnaire des recherches : la sélection de variables en mode multi-label partiel.

# 1 Introduction à l'univers multi-label

Les multiples problèmes faisant appel aux techniques d'apprentissage se classent en différents paradigmes. L'un des plus importants est sans doute l'approche mono-label que nous avions étudiée jusqu'à présent[1]. Dans cette approche, une donnée X en entrée est associée à une valeur Y en sortie, connue ou non, en fonction du mode d'apprentissage (supervisé, semi-supervisé ou non-supervisé). Comme nous l'avons vu, ces méthodes permettent par exemple dans le domaine biologique de détecter des anomalies génotypiques en attribuant à un gène une valeur « dangereux » ou « normal » en fonction de son état. Néanmoins, ce premier paradigme ne permet pas d'assigner plusieurs étiquettes à une même instance initiale.

C'est à partir de ce constat que le paradigme multi-label a commencé à émerger. Avec de nombreuses applications concrètes dans les domaines de la musique [18, 28], de la santé [31], ou de la reconnaissance textuelle [8] pour ne citer qu'eux, cette approche permet d'associer à une donnée X un vecteur Y composé d'au moins 2 étiquettes. Ainsi, il est par exemple possible d'attribuer à une musique donnée, une liste d'instruments présents dans le morceau. Une autre illustration courante de cette approche se retrouve aussi dans la reconnaissance de scène où plusieurs labels doivent s'affecter à une image afin d'en détecter ses différentes composantes.

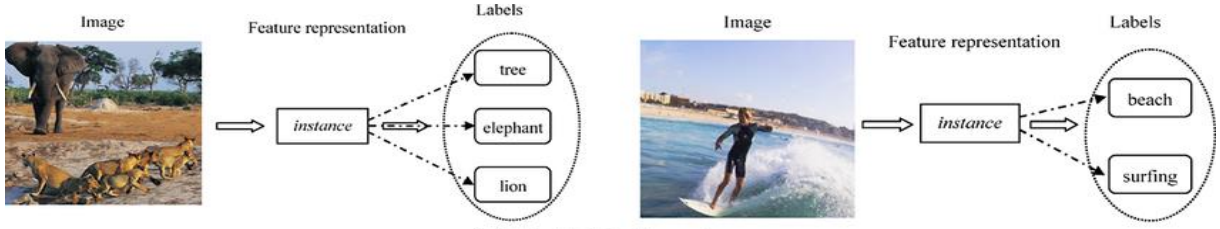


*Figure 1: Exemple de fonctionnement d'une approche d'apprentissage Multi-label [12]*

[1] Voir Etat de l'art (A) : Sélection de variables Supervisée – Mono-Label et Etat de l'art (B) : Sélection de variables Semi-Supervisée – Mono-Label

Tout comme le paradigme mono-label, l'approche multi-label repose sur différents algorithmes de classification dont les premiers travaux ont été effectués par Boutell and al [4] en 2004. Depuis, de nombreux chercheurs ont publié des algorithmes toujours plus efficaces et ceux-ci peuvent se retrouver dans différents états de l'art déjà existant à ce sujet comme celui très complet de Zhang et Zhou parut en 2014 [43]. De manière générale, une taxonomie divisée en trois[2] grandes familles de classifieur est partagée dans les différentes études [21] :

- Les **approches d'apprentissage par adaptation** adaptent et étendent le fonctionnement des algorithmes mono-label afin qu'elles puissent traiter les données multi-label
- Les **approches par transformation** consistent à transformer et diviser le problème multi-label en plusieurs sous-problème mono-label
- Les **méthodes d'ensemble** ou **méthodes ensemblistes** s'appuient et combinent les résultats d'un ensemble de classifieur.

L'explication de ces méthodes dépasse toutefois la portée de notre article.

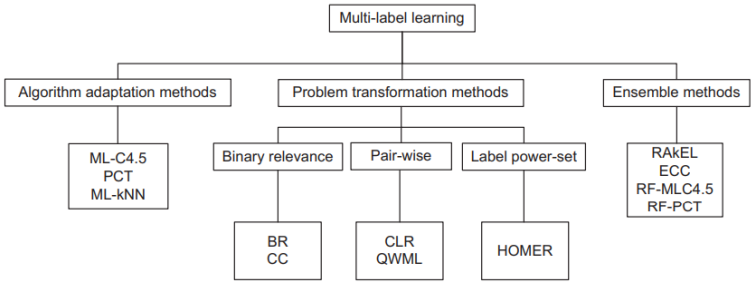


*Figure 2: Catégorisation des algorithmes de classification en approche Multi-label [21]*

A l'instar des modèles mono-label, la sélection de variables consistant à identifier les variables les plus pertinentes pour la tâche de classification est aussi au cœur de l'apprentissage multi-label afin de réduire la complexité des modèles et en augmenter la performance.

Ainsi, nous vous proposons par la suite une introduction aux différentes techniques de sélection de variables existantes dans l'univers multi-label.

[2] Les méthodes ensemblistes étant aussi existantes pour les approches mono-label, certaines études n'évoquent pas cette catégorie dans leur classification « multi-label ».

# 2 La sélection de variables appliquée au multi-label

Tout comme dans l'approche mono-label, de nombreuses techniques de sélection de variables appliquées au multi-label ont fait leur apparition en parallèle du développement des classifieurs. Souvent inspirées et adaptées des méthodes mono-label existantes, la sélection de variables multi-label peut se classer selon différentes taxonomies : type de label, stratégie de recherche, interaction avec le classifieur ou format de données comme l'illustre le schéma ci-contre [13].

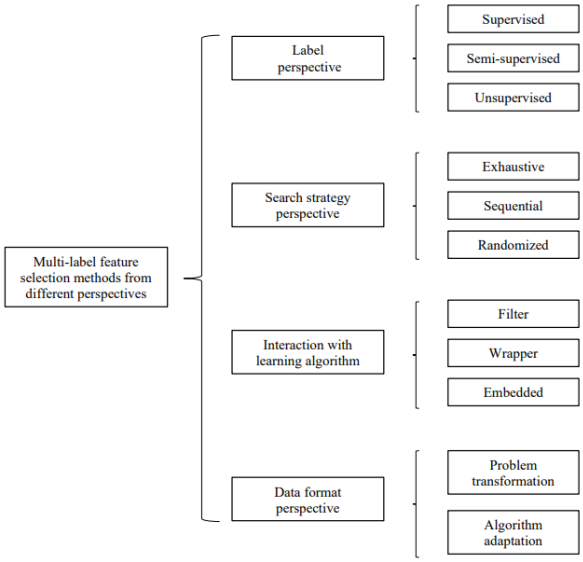


*Figure 3: Catégorisation des approches de sélection de variables dans l'univers Multi-label [8]*

Dans le cadre de notre sujet de recherche, nous adopterons comme précédemment une classification reposant sur le *type de label* et *l'interaction des méthodes de sélection de variables avec le classifieur*. Nous verrons ainsi une liste de méthodes non exhaustives utilisant les approches par filtrage, par symbiose (wrappers) et les méthodes intégrées (embedded) dans le cadre d'une approche supervisée.

Nous traiterons dans une section dédiée l'approche semi-supervisée. (*voir partie 3*)

## 2.1 Les techniques de filtrage

Comme étudiées précédemment, les techniques de filtrage sont une approche de sélection de variables indépendantes du classifieur consistant à classer puis sélectionner les meilleures variables d'entrée en fonction d'une métrique de performance. Les avantages de cette approche résident dans leur simplicité d'implémentation et leur rapidité d'exécution permettant à ces méthodes d'être très largement utilisées dans des ensembles de données à grandes dimensions. De manière générale, de nombreux algorithmes de sélection multi-label s'inspirent ou reposent sur des critères ou méthodes déjà utilisés dans un monde mono-label.

C'est par exemple le cas des algorithmes supervisés *ReliefF-ML*, *PPT-ReliefF* et *RReliefF-ML* proposé en 2015 par Reyes et Al [27] dérivant tous les trois des familles d'algorithmes Relief introduites pour la première fois par Kira & Rendell en 1992 [14]. Ces trois algorithmes sont particulièrement sensibles aux interactions entre les variables et s'appuient, de façon assez rare pour des méthodes de filtrage, sur la notion de « plus proche voisin » afin de capter des dépendances intervariables. Plus concrètement, *ReliefF-ML* est considéré par ses auteurs comme une généralisation pour le multi-label des équations de l'algorithme ReliefF classique. Son extension *RReliefF-ML* est basée sur les principes d'adaptation aux problèmes de régression. Ces deux méthodes utilisent directement les données multi-label tandis que l'algorithme *PPT-ReliefF* utilise une méthode de transformation de problème (méthode Pruned Problem Transformation [25]) afin de se ramener à plusieurs sous-problèmes mono-label.

Autre algorithme provenant du monde mono-label, l'algorithme *MLfR* de Lastra et al (2011) [16] s'appuie sur la méthode « *Fast Corrélation-Based Filter* » introduite en 2002 [37]. Utilisant des mesures de corrélation non-linéaires entre les variables, l'incertitude symétrique noté *SU ;* cette approche multi-label ne fonctionne que pour des données discrètes. L'idée générale derrière cet algorithme est de construire un graphe de relation entre les variables. La méthode commence par créer une matrice de score SU pour toutes les paires de variables, puis elle utilise cette matrice pour générer un arbre couvrant non orienté qui représente les variables sous forme de nœuds et les scores SU sous forme d'arêtes entre ces nœuds. Cet arbre servira à caractériser la pertinence et la redondance des variables d'entrées.

Par ailleurs, la grande famille des algorithmes utilisant la théorie des graphes se voit aussi dotée de plusieurs extensions fonctionnant dans l'univers multi-label. C'est le cas de la méthode *gMLC* conçue par Kong et Yu [15] qui recherche un ensemble optimal de variables à l'intérieur de sous-graphes spécialement conçus pour des objets multi-labellisés. Plus formellement, après construction d'un graphe général, l'algorithme utilise un critère de corrélation pour estimer la dépendance entre les variables des sous-graphes et les multiples labels du graphe. La méthode s'appuie ensuite sur un algorithme de « *séparation et évaluation* » (*branch-and-bound*) pour rechercher efficacement les variables optimales du sous-graphes en élaguant progressivement l'espace de recherche.

Très populaire dans les approches de filtrage supervisé mono-label, les méthodes s'appuyant sur le *gain d'information* connaissent également des équivalents multi-labels. C'est par exemple le cas de l'algorithme *MLInfoGain* [24] proposé par Pereira et al en 2015. A partir d'un ensemble de données multi-label en entrée, l'algorithme va calculer le gain d'information multi-label pour chaque variable à partir de l'entropie définie par Clare et King [7] (*entropie elle-même adaptée de l'algorithme C4.5 présenté lors de notre premier état de l'art*).

## 2.2 Les techniques wrapper (symbioses)

Autre approche de la sélection de variables, la méthode Wrapper consiste à utiliser les algorithmes de classification pour rechercher puis évaluer pas à pas un sous-ensemble de variables jusqu'à obtenir un ensemble dit optimal pour le classifieur utilisé. Ces méthodes ont donc l'avantage de prendre pleinement en compte les interactions entre les variables et le classifieur assurant une optimalité que l'approche par filtrage ne peut assurer. Ces méthodes sont toutefois coûteuses et ne peuvent fonctionner efficacement pour de grands jeux de données.

Différentes familles d'algorithmes peuvent intégrer les approches par wrapper. Parmi elles, nous pouvons citer les *random forest* dont une étude comparative sur la sélection de variables multi-labels est parue par Gharroudi, Elghazel et al en 2014 [10]. Trois algorithmes de sélection sont alors proposés puis étudiés : *BRRF* (**B**inary **R**elevance **R**andom **F**orest), *RFLP* (**R**andom **F**orest **L**abel **P**ower-set) et *RFPCT* (**R**andom **F**orest of **P**redictive **C**lusterin **T**rees). *BRRF* et *RFLP* utilise tout d'abord des algorithmes de transformation pour se ramener à des sous-problèmes mono-label où des *random forest* plus standard pourront être appliqués. *RFPCT* utilise quant à lui le classifieur *PCT* pouvant prédire plusieurs labels à la fois.

D'autres stratégies *wrappers* ont également été mises en place dans le cadre de l'approche multi-label. C'est le cas de l'algorithme proposé en 2017 par Zhang et al [45] basé sur une amélioration des algorithmes d'optimisation par essaims particulaires multi-objectifs (*OEP ou PSO en anglais).* La méthode consiste à adapter le problème discret qu'est la sélection de variables multi-labels à un

problème continu adapté à *PSO*. Pour cela, un encodage probabiliste stratégique est effectué sur les variables d'entrées afin que chaque individu soit une particule dans l'algorithme PSO. Afin d'améliorer les performances, Zhang et al proposent également d'ajouter à PSO des mutations uniformes adaptatives et des stratégies locales d'apprentissage. Les chercheurs mettent aussi en place une méthode d'archivage permettant de sauvegarder les solutions optimales lors des itérations de l'algorithme.

En 2009, Zhang et Al [40] propose une méthode *MLNB* (**M**ulti-**L**abel **N**aive **B**ayes) adaptant les classifieurs traditionnels *Naïves Bayes* à la sélection de variables multi-labels. Pour cela, l'algorithme utilise tout d'abord des techniques d'extraction de variables basées sur l'*analyse en composante principales* (ACP) pour supprimer les données redondantes et/ou non pertinentes. En employant des algorithmes génétiques (GA), la méthode choisit ensuite un sous-ensemble de variables optimal en prenant en compte les corrélations entre les labels de chaque instance. La fonction de fitness utilisée dans les GA traite explicitement ces corrélations afin d'assurer une sélection efficace.

De nombreux autres méthodes par symbiose multi-label s'appuient aussi sur les algorithmes génétiques. C'est le cas d'une technique proposée en 2014 par Yu et Wong [39]. L'algorithme se décompose en deux étapes. Dans un premier temps, il utilise *l'information mutuelle* pour sélectionner localement les variables. Ensuite, il s'appuie sur l'algorithme GA pour sélectionner le sous-ensemble de variables optimal tout en prenant en compte les corrélations entre les différents labels, selon la même logique que la méthode *MLNB*.

## 2.3 Les techniques embedded (intégrées)

Dernière catégorie de cette taxonomie, les méthodes intégrées dîtes « *embedded* » essayent de combiner les atouts des approches par filtrage et par symbiose. Elles intègrent pour cela la partie « entrainement » des données directement au processus de classification. Ces méthodes ont alors l'avantage d'avoir une grande interaction entre les variables initiales et le classifieur utilisés tout en restant assez rapides comparées aux approches *wrapper*. Tout comme les catégories précédentes, c'est à partir de méthodes préexistantes dans un monde mono-label que les techniques multi-labels ont été élaborées.

C'est d'ailleurs le cas de l'algorithme *CMLFS* (**C**orrelated **M**ulti-**L**abel **F**eature **S**election) proposé en 2011 par Gu et al [11]. Cette méthode vise à répondre d'un seul coup aux défis de l'interdépendance, de la corrélation et de la haute dimensionnalité des données. Elle s'appuie pour cela sur un classifieur linéaire tel que *LaRankSVM* [9]. L'idée principale de cet algorithme est de découvrir un sous-ensemble de variables tel que la perte de classement régularisée par corrélation d'étiquettes soit minimisée. Cela se traduit par un programme linéaire à contrainte quadratique pouvant se résoudre par un algorithme de plan de coupe. Cette approche a l'avantage de prendre pleinement en compte la corrélation entre les labels. Néanmoins son coût de calcul conséquent entraine une perte de rapidité importante vis-à-vis de méthodes plus récentes.

Les méthodes utilisant les principes de régularisation existent aussi en multi-label comme l'algorithme *MLSLR* (**M**ultilabel **L**earning via **S**parse **L**ogistic **R**egression) introduit en 2014 par Liu et Al [17]. Cette approche se base sur la régression logistique pour former et créer des modèles multi-label efficaces. Plus concrètement, la sélection de variables proposée ici s'appuie sur la régression

« *Elastic Net* », c'est-à-dire la combinaison des régularisations *LASSO* (norme l1) et *RIDGE* (norme l2), où la norme l1 vise à éliminer les variables non pertinentes tandis que la norme l2 garantit que les variables fortement corrélées possèdent un coefficient de régression similaire.

Autre méthode intégrée, l'algorithme « ***M**ulti-label **E**mbedded **F**eature **S**election* » (*MEFS*) présenté en 2021 par You et Al. [36] utilise tout d'abord le critère de risque de prédiction pour évaluer l'importance des variables. Il emploie par la suite la stratégie d'élimination récursive (*Backward elimination*) pour rechercher un sous-ensemble optimal de variables spécifiques au classifieur utilisé. L'algorithme MEFS a l'avantage de prendre en compte la corrélation entre les labels et est conçu pour s'adapter de manière efficace à plusieurs classifieurs tel que *Rank-SVM* [9], *LEAD* [41], *MLNB* [40], *ML-KNN* [42].

# 3 La sélection de variables semi-supervisées appliquée au multi-label

S'appuyant sur des méthodes déjà présentes en apprentissage supervisé et non supervisé, plusieurs algorithmes de sélection de variables semi-supervisé co-existent dans le cadre de l'approche multi-label.

L'une des approches de sélection de variables semi-supervisée consiste à combiner l'apprentissage Manifold avec la sélection des variables. La régularisation Manifold fondée sur le Laplacien du graphe est largement utilisée dans la sélection des variables multi-labels. Nie et al [23] ont combiné la régression des moindres carrés avec la régularisation de Manifold et ont proposé un cadre de sélection d'éléments semi-supervisés et non supervisés. En outre, la régularisation éparse est largement étudiée dans la sélection des variables [22]. Nie et al [22] ont d'ailleurs proposé une autre méthode de sélection des variables en utilisant cette fois la norme l2,1 dans la fonction objective. De plus, Ma et al [20] ont introduit en 2021 une méthode semi-supervisée de sélection des variables pour les données multimédias en combinant la régularisation éparse et la régularisation. Manifold. Xu et al [35] ont quant à eux proposé un modèle de sélection de variables multi-label semi-supervisée basé sur la cohérence spatiale en utilisant des voisins probabilistes adaptatifs. Ces méthodes ont donné de bons résultats, mais elles ne prennent en compte que la structure locale des données et ignorent leur structure globale, ce qui fait que les variables sélectionnées sont très redondantes [26].

Une autre approche pour la sélection semi-supervisée de variables multi-labels consiste à supposer que les données multi-labels ont un sous-espace partagé. Par exemple, dans les ensembles de données d'images multi-labels, une image étiquetée "jardin" et "arbres" partage le label "arbres" avec une autre image étiquetée "forêt" et "arbres". Chang et al [6] ont proposé une méthode semi-supervisée de sélection des variables en exploitant la corrélation des étiquettes. Wang et al [30] ont proposé une méthode semi-supervisée de sélection de variables multi-labels basée sur l'apprentissage par sous-espace partagé et l'apprentissage Manifold. Bien que l'apprentissage par sous-espace partagé ait montré son efficacité dans les scénarios d'apprentissage multi-label, il présente également des inconvénients. Les dimensions du sous-espace partagé sont difficiles à déterminer. En outre, les dimensions du sous-espace partagé de différentes données sont également différentes, ce qui pose également des difficultés à l'algorithme.

D’autre part, Lv et al [19] ont proposé de fusionner l'apprentissage adaptatif de la structure globale et la régularisation Manifold dans le cadre de la sélection des variables, où la régularisation Manifold basée sur le Laplacien du graphe est utilisée pour capturer la structure locale et les corrélations entre les étiquettes, l'apprentissage adaptatif de la structure globale est utilisé pour extraire la structure globale.

Contrairement aux méthodes ci-dessus, Chang et al [5] ont proposé une méthode semi- supervisée de sélection de variables multi-labels. La méthode de Chang et al. ne fait pas appel à la régularisation Manifold et à l'hypothèse du sous-espace partagé, ce qui permet de l'appliquer à des ensembles de données à grande échelle.

# 4 Une nouvelle approche : L’apprentissage multi-label partiel

## 4.1 Introduction au multi-label partiel (PML)

Dans le cadre des approches mono-label ou multi-labels étudiées jusqu’à présent, l’ensemble des données étiquetées ont toujours été considérées comme fiables. L’objectif de ces approches est alors uniquement de faire correspondre une instance X en entrée à une ou plusieurs cibles Y en sortie sans jamais se soucier de la véracité de la variable Y. Ces approches sont donc performantes lorsque nous sommes certains de pouvoir étiquetées convenablement les données.

Néanmoins, comme nous l’avions vu précédemment[3], il est souvent très coûteux et difficile d’annoter avec précisions les instances de départ notamment dans le cadre multi-labels. En effet, en reprenant l’exemple de la reconnaissance de scène étudié dans la partie 1 (*Introduction à l’univers multi-label)*, il est à première vue facile d’annoter grossièrement les éléments clés de l’image ci-dessous comme avec les étiquettes « building » ou « light ». Il est toutefois beaucoup plus délicat d’annoter avec certitude les éléments de second plan. Peut-on affirmer que nous voyons des nuages ou des personnes au fond de la rue ? La réponse est non, ces éléments sont probables, mais pas « sûr ». C’est ce qu’on appelle alors l’approche Multi-Label Partiel (PML).


*Figure 4: Exemple de fonctionnement d'une approche d'apprentissage Multi-label Partiel [33]*

[3] Voir Etat de l’art (B) : Sélection de variables Semi-Supervisée – Mono-Label

Introduite pour la première fois en 2018 par Xie et Huang [32], cette approche très récente consiste donc à travailler sur un ensemble d'étiquettes cibles Y appelée « labels candidats » dont seule une partie de ces étiquettes est réellement pertinente. La notion de labels candidats est formalisée par les deux chercheurs comme ceci :

a) L'ensemble des labels candidats peut contenir à la fois des labels pertinents et d'autres non-pertinents
b) Le nombre de label pertinent est inconnu mais est au moins de 1.
c) Les étiquettes non-présentes dans les labels-candidats sont considérées comme non-pertinentes.

C'est à partir de cette définition que les premiers classifieurs PML ont émergé à partir de 2018 dont les premiers algorithmes *PML-fc* et *PML-fp* détaillés dans ce même article [32]. Bien que dépassant le cadre de cet état de l'art, l'idée général derrière ces nouveaux algorithmes d'apprentissage est d'attribuer une valeur de confiance à chaque label en fonction d'hypothèses supplémentaires au jeu de données puis de faire varier ces coefficients jusqu'à obtenir des résultats satisfaisants.

Depuis, plusieurs autres classifieurs PML ont été développés comme l'algorithme *PML-LFC* (2020) [38] estimant les valeurs de confiance en utilisant un concept de similarité entre un label et l'ensemble des labels candidats, ou encore le classifieur *PML-MD* (2021) [33] qui estime de manière adaptative la confiance attribuée à chaque label en fonction d'hypothèse interne au classifieur.

Néanmoins, afin d'augmenter les performances de ces différents algorithmes, il est très vite devenu crucial de mettre en place des algorithmes de sélection de variables permettant de réduire le nombre de variables redondantes et/ou non-pertinentes.

## 4.2 Les méthodes de sélection de variables appliquées au multi-label partiel (FS-PML)

Domaine de recherche extrêmement récent, la sélection de variables appliquée aux problèmes PML n'en n'est pour l'heure qu'au stade embryonnaire de son développement. Alors que les premiers classifieurs PML utilisaient des méthodes de sélections multi-label simple, certains algorithmes destinés spécifiquement à l'approche multi-label partiel commence à émerger.

C'est en effet en juillet 2022 que le premier algorithme de sélection de variables PML est apparu. Introduit par Wang, Li et Yu [29], l'algorithme *PMLFS* fait partie de la grande famille des techniques **embedded** (=intégrées) vu précédemment. Dans un premier temps, cette méthode distingue à partir d'une corrélation les étiquettes bruitées des étiquettes considérées comme fiables. La méthode incorpore ensuite une régularisation Manifold afin d'explorer la structure locale des données. Enfin, les chercheurs s'appuient sur une régularisation de norme l1 et l2 pour sélectionner au mieux les variables d'entrée optimales. Basé sur des métriques d'évaluation fréquemment utilisée dans l'univers multi-label, comme Hamming Loss, Macro-F, Micro-F, Average Precision ou le Ranking-Loss, l'algorithme *PMLFS* surpasse globalement les algorithmes de sélection de variables multi-labels classiques. Il connait néanmoins quelques difficultés lorsque les labels candidats sont fortement bruités s'appuyant parfois à tort sur des labels non-pertinents pour réaliser la sélection de variables.

Plus récemment en février 2023, Xu et Al [34] ont proposé une nouvelle approche de sélection de variables appliquée au PML consistant à s'appuyer sur des labels « crédibles » afin de guider le choix des variables initiales. Cette notion de crédibilité est donc le cœur de cet algorithme. Concrètement, les chercheurs s'appuient tout d'abord sur une matrice de scores de confiance pour déterminer les

labels candidats les plus fiables. Ensuite, ils s'appuient sur les labels les plus crédibles pour former un modèle de sélection conjoint, autrement dit, un modèle comprenant un apprentissage simultané de variables spécifiques à chaque label et de variables communes à tous les labels. Enfin, l'algorithme ajoute un critère de corrélation entre les labels afin de faciliter la conception d'un sous-ensemble de variables optimales. Cette méthode faisant intervenir la phase d'apprentissage parallèlement à la sélection de variables est donc également une méthode **embedded**.

De notre connaissance, très peu d'algorithmes autres que ceux mentionnées ci-dessus existent pour la sélection de variables PML. Néanmoins, il est intéressant de noter que certains chercheurs évoquent cette notion de *sélection de variables* pour désigner la réduction de l'espace des labels candidats. C'est par exemple le cas de l'algorithme *CENDA* [2] utilisant le critère d'interdépendance d'Hilbert-Schmidt, de la méthode *DECLIN* [44] se basant une analyse discriminante linéaire inter-labels, ou de la technique *SAUTE* [3] qui évalue la dépendance entre les labels via un critère d'information mutuelle. Ces algorithmes transforment donc des critères classiques de sélection de variables pour sélectionner un sous-ensemble optimal de label candidats.

## Conclusion

Après avoir longuement étudié la sélection de variables dans une approche à simple label, nous avons dans cet état de l'art changé de prisme d'étude pour passer dans une approche multi-label. Ici, chaque instance initiale peut se rattacher à plusieurs labels cibles dont certains, pour l'approche PML, peuvent être bruitées. Les algorithmes de sélection de variables doivent alors s'adapter en conséquence afin d'augmenter la précision et la performance des classifieurs.

De manière générale, les méthodes de sélection de variables multi-label s'appuient sur des algorithmes préexistant dans une approche à simple label. Nous retrouvons ainsi les mêmes trois grandes familles d'algorithmes que sont les filtres, les wrapper (symbioses) et embedded (intégrées). Ces méthodes doivent néanmoins s'adapter à l'univers multi-label et notamment au tout récent domaine de recherche qu'est le multi-label partiel. Dans cette approche, le défi des algorithmes de sélection de variables et des classifieurs est alors de distinguer les labels candidats fiables des labels non-pertinents afin de ne prédire que des étiquettes correctes.

Pour l'heure, le nombre d'algorithme de sélection de variables dans l'approche multi-label partiel reste relativement faible. Cela s'explique notamment par la nouveauté de ce domaine de recherche dont les premiers articles remontent à l'été 2022. Toutefois, l'approche PML apportant de nouvelles perspectives/solutions intéressantes pour l'apprentissage, il parait comme nécessaire et probable que les méthodes de sélection de variables sur le sujet croîtront dans les prochaines années.

# Références


[1] Alalga, A., Benabdeslem, K., & Taleb, N. (2016). Soft-constrained Laplacian score for semi-supervised multi-label feature selection. *Knowledge and Information Systems*, *47*(1), 75-98.

[2] Bao, W. X., Hang, J. Y., & Zhang, M. L. (2021, August). Partial label dimensionality reduction via confidence-based dependence maximization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 46-54).

[3] Bao, W. X., Hang, J. Y., & Zhang, M. L. (2022, August). Submodular Feature Selection for Partial Label Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 26-34).

[4] Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern recognition*, *37*(9), 1757-1771.

[5] Chang, X., Nie, F., Yang, Y., & Huang, H. (2014, June). A convex formulation for semi- supervised multi-label feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 28, No. 1).

[6] Chang, X., Shen, H., Wang, S., Liu, J., & Li, X. (2014). Semi-supervised feature analysis for multimedia annotation by mining label correlation. In *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part II 18* (pp. 74-85). Springer International Publishing.

[7] Clare, A., & King, R. D. (2001). Knowledge discovery in multi-label phenotype data. In Principles of Data Mining and Knowledge Discovery: 5th European Conference, PKDD 2001, Freiburg, Germany, September 3–5, 2001 Proceedings 5 (pp. 42-53). Springer Berlin Heidelberg.

[8] Elghazel, H., Aussem, A., Gharroudi, O., & Saadaoui, W. (2016). Ensemble multi-label text categorization based on rotation forest and latent semantic indexing. *Expert Systems with Applications*, *57*, 1-11.

[9] Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification. *Advances in neural information processing systems*, *14*.

[10] Gharroudi, O., Elghazel, H., & Aussem, A. (2014). A comparison of multi-label feature selection methods using the random forest paradigm. In *Advances in Artificial Intelligence: 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014. Proceedings 27* (pp. 95-106). Springer International Publishing.

[11] Gu, Q., Li, Z., & Han, J. (2011, October). Correlated multi-label feature selection. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1087-1096).

[12] He, J., Gu, H., & Wang, Z. (2012). Bayesian multi-instance multi-label learning using Gaussian process prior. *Machine learning*, *88*(1-2), 273-295.

[13] Kashef, S., Nezamabadi-pour, H., & Nikpour, B. (2018). Multilabel feature selection: A comprehensive review and guiding experiments. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(2), e1240.

[14] Kira, K., & Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *Aaai* (Vol. 2, No. 1992a, pp. 129-134).

[15] Kong, X., & Yu, P. S. (2012). gMLC: a multi-label feature selection framework for graph classification. *Knowledge and information systems*, *31*, 281-305.

[16] Lastra, G., Luaces, O., Quevedo, J. R., & Bahamonde, A. (2011). Graphical feature selection for multilabel classification tasks. In *Advances in Intelligent Data Analysis X: 10th International Symposium, IDA 2011, Porto, Portugal, October 29-31, 2011. Proceedings 10* (pp. 246-257). Springer Berlin Heidelberg.

[17] Liu, H., Zhang, S., & Wu, X. (2014). MLSLR: Multilabel learning via sparse logistic regression. *Information Sciences*, *281*, 310-320.

[18] Lo, H. Y., Wang, J. C., Wang, H. M., & Lin, S. D. (2011). Cost-sensitive multi-label learning for audio tag annotation and retrieval. *IEEE Transactions on Multimedia*, *13*(3), 518-529.

[19] Lv, S., Shi, S., Wang, H., & Li, F. (2021). Semi-supervised multi-label feature selection with adaptive structure learning and manifold learning. *Knowledge-based systems*, *214*, 106757.

[20] Ma, Z., Nie, F., Yang, Y., Uijlings, J. R., Sebe, N., & Hauptmann, A. G. (2012). Discriminating joint feature analysis for multimedia data understanding. *IEEE Transactions on Multimedia*, *14*(6), 1662-1672.

[21] Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern recognition*, *45*(9), 3084-3104.

[22] Nie, F., Huang, H., Cai, X., & Ding, C. (2010). Efficient and robust feature selection via joint ℓ2, 1-norms minimization. *Advances in neural information processing systems*, *23*.

[23] Nie, F., Xu, D., Tsang, I. W. H., & Zhang, C. (2010). Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions onImage Processing*, *19*(7), 1921-1932.

[24] Pereira, R. B., Carvalho, A. P. D., Zadrozny, B., & Merschmann, L. H. D. C. (2015). Information gain feature selection for multi-label classification.

[25] Read, J. (2008). A pruned problem transformation method for multi-label classification. *Proceedings 2008 New Zealand Com-puter Science Research Student Conference* (NZCSRS 2008), *pp. 143–150*

[26] Ren, Y., Zhang, G., Yu, G., & Li, X. (2012). Local and global structure preserving based feature selection. *Neurocomputing*, *89*, 147-157.

[27] Reyes, O., Morell, C., & Ventura, S. (2015). Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context. *Neurocomputing*, *161*, 168-182.

[28] Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. P. (2008, September). Multi- label classification of music into emotions. In *ISMIR* (Vol. 8, pp. 325-330).

[29] Wang, J., Li, P., & Yu, K. (2022, July). Partial Multi-Label Feature Selection. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-9). IEEE.

[30] Wang, X. D., Chen, R. C., Hong, C. Q., Zeng, Z. Q., & Zhou, Z. L. (2017). Semi- supervised multi-label feature selection via label correlation analysis with l1-norm graph embedding. *Image and Vision Computing*, *63*, 10-23.

[31] Wu, J. S., Huang, S. J., & Zhou, Z. H. (2014). Genome-wide protein function prediction through multi-instance multi-label learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(5), 891-902.

[32] Xie, M. K., & Huang, S. J. (2018, April). Partial multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

[33] Xie, M. K., Sun, F., & Huang, S. J. (2021, August). Partial multi-label learning with meta disambiguation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 1904-1912).

[34] Xu, T., Xu, Y., Yang, S., Li, B., & Zhang, W. (2023). Learning Accurate Label-Specific Features From Partially Multilabeled Data. *IEEE Transactions on Neural Networks and Learning Systems*.

[35] Xu, Y., Wang, J., An, S., Wei, J., & Ruan, J. (2018, October). Semi-supervised multi-labelfeature selection by preserving feature-label space consistency. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 783-792).

[36] You, M., Liu, J., Li, G. Z., & Chen, Y. (2012). Embedded feature selection for multi-label classification of music emotions. *International Journal of Computational Intelligence Systems*, *5*(4), 668-678.

[37] Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, *5*, 1205-1224.

[38] Yu, T., Yu, G., Wang, J., & Guo, M. (2020). Partial multi-label learning with label and feature collaboration. In *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part I 25* (pp. 621-637). Springer International Publishing.

[39] Yu, Y., & Wang, Y. (2014). Feature selection for multi-label learning using mutual information and GA. In Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24-26, 2014, Proceedings 9 (pp. 454-463). Springer International Publishing.

[40] Zhang, M. L., Peña, J. M., & Robles, V. (2009). Feature selection for multi-label naive Bayes classification. *Information Sciences*, *179*(19), 3218-3229.

[41] Zhang, M. L., & Zhang, K. (2010, July). Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 999-1008).

[42] Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition*, *40*(7), 2038-2048.

[43] Zhang, M. L., & Zhou, Z. H. (2013). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, *26*(8), 1819-1837.

[44] Zhang, M. L., Wu, J. H., & Bao, W. X. (2022). Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *16*(4), 1-18.

[45] Zhang, Y., Gong, D. W., Sun, X. Y., & Guo, Y. N. (2017). A PSO-based multi-objective multi-label feature selection method in classification. *Scientific reports*, *7*(1), 1-12.