

Apprentissage multi-tâches par transfert de structures de réseaux bayésiens

Sarah Benikhlef,¹ Philippe Leray,¹ Guillaume Raschia,¹
Montassar Ben Messaoud,² Fayrouz Sakly²

¹LS2N, UMR CNRS 6004, University of Nantes, France
{sarah.benikhlef, philippe.leray, guillaume.raschia}@ls2n.fr

²LARODEC, ISG Sousse, Tunisia
montassar.benmessaoud@gmail.com, saklyfayrouz@outlook.fr

Abstract

Nous nous intéressons à l'apprentissage des structures de réseaux bayésiens dans un contexte multi-tâches en tirant parti de l'information communiquée entre tâches similaires. Nous proposons un nouvel algorithme que nous appelons MT-MMHC (Multi-Task Max-Min Hill Climbing) qui combine des méthodes d'apprentissage par transfert à plusieurs approches d'apprentissage de graphes de réseaux bayésiens. Cet algorithme se compose de deux phases principales. La première identifie les tâches les plus similaires et utilise leur similarité pour enrichir l'apprentissage de leurs graphes non orientés correspondants. La seconde oriente les arcs en effectuant une recherche gloutonne contrainte puis optimisée grâce un algorithme de séparation et évaluation (Branch-and-Bound). L'évaluation empirique montre que MT-MMHC parvient à de meilleurs résultats que ceux obtenus par l'apprentissage individuel des structures ainsi que ceux de l'algorithme multi-tâches MT-GS de l'état de l'art, en termes de précision d'apprentissage et de temps de calcul.

Introduction

Apprendre des modèles robustes à partir de petits jeux de données peut parfois s'avérer difficile. L'apprentissage par transfert (TL) (Zhou, Hospedales, and Fenton 2016) permet d'améliorer la fiabilité des modèles construits en exploitant la similarité qui lie certaines tâches (Oyen and Lane 2013). Il a été étudié et employé avec succès dans une variété de domaines d'apprentissage automatique, plus particulièrement les réseaux de neurones (Salaken et al. 2017; Zhou, Hospedales, and Fenton 2016).

L'apprentissage multi-tâches (MT), également appelé apprentissage par transfert parallèle, vise à apprendre simultanément plusieurs tâches en tirant parti de leurs similarités pour améliorer la généralisation de toutes les tâches. Il peut être combiné à d'autres mécanismes d'apprentissage pour augmenter considérablement la performance des modèles construits (Zhang and Yang 2021).

L'apprentissage par transfert et l'apprentissage MT sont étroitement liés. Ils ont tous deux pour but d'exploiter les connaissances entre des problèmes similaires (Niculescu-Mizil and Caruana 2007). La distinction entre ces deux approches réside dans la technique de transfert. Dans l'apprentissage TL, les informations sont transférées de la

tâche source à la tâche cible. L'objectif est ainsi d'améliorer les performances de la tâche cible grâce à l'enrichissement fourni par les tâches auxiliaires. Dans l'apprentissage MT, toutes les tâches jouent le même rôle et cherchent à s'enrichir mutuellement (Thung and Wee 2018).

Les réseaux bayésiens (RB) sont un formalisme de raisonnement probabiliste qui permet de saisir les dépendances conditionnelles et les indépendances entre différentes variables aléatoires. Ils constituent un outil efficace pour représenter la structure de plusieurs applications du monde réel et sont de plus en plus employés dans des tâches aussi diverses que le diagnostic médical (López-Cruz et al. 2014), l'industrie ou encore la robotique. La structure des réseaux bayésiens est généralement fournie a priori par un expert ou déterminée à partir des données. Cependant, dans certains domaines, les données peuvent être rares et occasionnellement recueillies à partir de problèmes différents mais étroitement liés. Dans cette situation, des solutions d'apprentissage par transfert et d'apprentissage multi-tâches de réseaux bayésiens ont été proposées (Luis, Sucar, and Morales 2010; Oyen and Lane 2012; Niculescu-Mizil and Caruana 2007). Ces approches sont principalement des adaptations d'algorithmes d'apprentissage de structures de réseaux bayésiens basés sur un score ou sur la recherche d'indépendances conditionnelles.

Dans cet article, nous proposons une extension à l'apprentissage multi-tâches d'un algorithme efficace de détermination de structures de RB appelé Max-Min Hill-Climbing (MMHC) (Tsamardinos, Brown, and Aliferis 2006). Il s'agit d'une méthode hybride qui combine à la fois la recherche d'indépendances conditionnelles et les algorithmes basés sur un score.

Nos trois principales contributions sont : (1) MT-MMHC, un algorithme hybride d'apprentissage par transfert qui peut apprendre simultanément plusieurs structures de RBs en induisant l'information entre tâches similaires pour améliorer la performance des réseaux construits; (2) une procédure pour générer des benchmarks MT de réseaux bayésiens à partir de n'importe quel modèle de référence tout en contrôlant la similarité entre les tâches; et (3) une validation expérimentale de MT-MMHC en utilisant de tels benchmarks comparé à l'apprentissage individuel des tâches (ST pour single-task), mais aussi à un algorithme multi-tâches de

références MT-GS (Multi Task Greedy Search).

Nous allons commencer par présenter de brefs rappels ainsi que des travaux connexes sur l'apprentissage de structures de réseaux bayésiens. Nous décrirons ensuite notre approche MT-MMHC puis nous détaillerons la procédure de génération de benchmarks MT. Nous exposerons alors l'évaluation empirique de notre proposition. Enfin, nous conclurons sur les résultats obtenus et évoquerons plusieurs perspectives.

Apprentissage multi-tâches de structures de réseaux bayésiens

Un réseau bayésien (RB) $\mathcal{B} = \langle G, \Theta \rangle$ représente la loi de probabilité jointe d'un ensemble de n variables aléatoires $X = \{X_1, X_2, \dots, X_n\}$ (Daly, Shen, and Aitken 2011). Il se caractérise par un graphe orienté acyclique (DAG, pour Directed Acyclic Graph) G et Θ l'ensemble des tables de probabilités conditionnelles (CPTs) également appelées paramètres. Le graphe ou la structure G est déterminée par le couple (V, E) où V est l'ensemble des noeuds du graphe et E l'ensemble d'arcs qui les relie. Chaque noeud correspond à une variable aléatoire et chaque arc représente une dépendance directe entre les deux variables qu'il relie. Ainsi, la partie graphique du réseau bayésien procure un outil visuel de représentation des connaissances qui permet d'appréhender plus facilement l'incertain en quantifiant les dépendances entre les variables.

L'apprentissage d'un seul RB (ST) revient à déterminer la structure du réseau G et à estimer les paramètres Θ du modèle à partir d'un jeux de données D .

Dans le cas de l'apprentissage multi-tâches (MT), nous considérons k tâches correspondant à k jeux de données $\mathcal{D} = \{D_1, D_2, \dots, D_k\}$ à partir desquels nous apprenons k graphes de RBs correspondants $\mathcal{G} = \{G_1, G_2, \dots, G_k\}$ et les paramètres associés. L'objectif est d'apprendre simultanément tous les modèles tout en exploitant l'information induite entre tâches. Dans un contexte similaire, Azzimonti, Corani, and Scutari proposent une approche alternative dans laquelle une structure de RB commune est apprise à partir de plusieurs données connexes.

L'apprentissage MT des paramètres, lorsque \mathcal{G} est connu, a été traité dans la littérature (Daly, Shen, and Aitken 2011; Luis, Sucar, and Morales 2010). Luis, Sucar, and Morales présentent une méthode d'apprentissage par transfert pour déterminer l'ensemble des CPTs à partir de plusieurs sources grâce à des fonctions d'agrégation.

Dans cet article, nous nous intéressons principalement à l'apprentissage MT de structures de réseaux bayésiens. Dans un contexte ST, déterminer la structure d'un réseau bayésien à partir des données revient à trouver un graphe qui représentera et explique le mieux ces données. Pour se faire, il existe trois grandes familles d'approches: (1) les méthodes basées sur un score, (2) la recherche d'indépendances conditionnelles et (3) les méthodes hybrides qui combinent les deux précédentes (Daly, Shen, and Aitken 2011; Leray 2006).

Algorithmes basés sur un score

Pour apprendre la structure d'un réseau bayésien, les approches basées sur un score recherchent dans l'espace des DAGs possibles celui qui maximise une métrique de score donnée (Daly, Shen, and Aitken 2011). De manière générale, ces algorithmes nécessitent: (i) un espace de recherche des états admissibles du problème, chaque état représentant une structure possible du réseau bayésien, (ii) une fonction de score pour évaluer un état et (iii) un mécanisme pour explorer cet espace.

Greedy Search (GS) (Scutari, Vitolo, and Tucker 2019), par exemple, est un algorithme d'apprentissage ST de graphes de RBs qui part d'une certaine structure initiale et explore l'espace des DAGs en sélectionnant à chaque itération la structure voisine au score le plus élevé. La recherche s'arrête lorsque la structure sélectionnée a un meilleur score que tous ses voisins, et peut être répétée plusieurs fois à partir de différents états initiaux pour éviter les maxima locaux. Un voisinage d'une structure G est communément défini comme étant l'ensemble des DAGs obtenus en ajoutant, supprimant ou en inversant un arc du graphe G . Pour l'évaluation des états, plusieurs scores ont été établis dans la littérature tels que AIC, BIC, BDeu (Carvalho 2009) ou encore plus récents tels que qNML (Silander et al. 2018).

Niculescu-Mizil and Caruana étendent cette recherche basée sur le score au contexte de l'apprentissage MT. Ils considèrent une configuration de k structures $\{G_1, G_2, \dots, G_k\}$ et effectuent une recherche gloutonne dans l'espace des DAGs pour déterminer les k graphes qui correspondent le mieux aux k jeux de données respectifs. Le voisinage d'une configuration \mathcal{G} est défini comme étant l'ensemble de toutes les configurations obtenues en appliquant une opération d'ajout, de suppression ou d'inversion au même arc dans chaque graphe de chaque sous-ensemble possible de la configuration \mathcal{G} . Le score à maximiser considéré est la probabilité a posteriori d'une configuration connaissant les données comme défini dans l'équation 1.

$$P(\mathcal{G}|\mathcal{D}) = P(G_1, \dots, G_k | D_1, \dots, D_k) \propto P(\mathcal{G}) \prod_{a=1}^k P(D_a | G_a) \quad (1)$$

Ils définissent deux types de décompositions de la probabilité a priori $P(\mathcal{G})$, (i) la première considère le nombre minimum d'opérations nécessaires pour qu'un arc soit identique dans chaque structure et (ii) la deuxième prend en compte la distance structurale entre chaque paire de graphes tel qu'il est défini dans l'équation 2.

$$P(\mathcal{G}) = Z_{\delta,k} \prod_{1 \leq a \leq k} P(G_a)^{\frac{1}{1+(k-1)\delta}} \prod_{1 \leq a < b \leq k} (1-\delta)^{\frac{d(G_a, G_b)}{k-1}} \quad (2)$$

où $\delta \in [0, 1]$ est un paramètre qui pénalise la différence entre structures, $Z_{\delta,k}$ est un facteur de normalisation et $d(G_a, G_b)$ est le nombre d'arêtes dans la différence symétrique entre G_a and G_b . Ainsi définie, la fonction de

score permet de considérer toutes les données et d'exploiter l'information qui les lie.

Par ailleurs, comme l'espace de recherche peut devenir grand pour des valeurs élevées de k et n (le nombre de variables dans les données), Niculescu-Mizil and Caruana adoptent un algorithme de séparation et évaluation (Branch and Bound) pour optimiser l'exploration du voisinage. À cette fin, ils définissent la notion de configuration partielle $\mathcal{C}_\ell = (G_1, \dots, G_\ell)$ d'ordre $\ell < k$, où seules les premières structures ℓ sont spécifiées et les autres $k - \ell$ ne le sont pas. L'exploration du voisinage passe par à un arbre de recherche de profondeur k pour atteindre la configuration ayant le meilleur score. À chaque niveau $\ell < k$, seul le score de la configuration partielle $\mathcal{C}_\ell = (G_1, \dots, G_\ell)$ est calculé et comparé au meilleur score actuel. Le score d'une configuration partielle \mathcal{C}_ℓ est défini comme borne supérieure aux scores de toutes les configurations complètes dans lesquelles elle est incluse. De cette façon, chaque sous-arbre enraciné dans une configuration partielle dont le score est inférieur au meilleur score actuel pourra être élagué.

Recherche d'indépendances conditionnelles

Cette famille d'approches d'apprentissage de structures, souvent appelée recherche sous contraintes, a pour principale caractéristique de tester les indépendances conditionnelles entre les variables et d'identifier progressivement le graphe qui décrit ces dépendances et indépendances découvertes dans les données (Daly, Shen, and Aitken 2011).

Jia et al. abordent le problème de la recherche sous contrainte avec transfert inductif. Luis, Sucar, and Morales proposent une méthode d'apprentissage de structure sous contrainte dans un contexte multi-tâches qui s'inspire de l'algorithme (ST) PC (Spirtes, Glymour, and Scheines 2000). Ce dernier part d'un graphe non orienté entièrement connecté puis mesure l'association entre les variables pour décider si un arc doit être retiré du graphe ou non. La principale différence réside dans la manière d'évaluer les tests d'indépendances. Le calcul usuel est remplacé par une combinaison linéaire des mesures d'indépendances de la tâche cible et de la tâche auxiliaire la plus similaire. La similarité entre tâches est définie par deux métriques: la similarité globale Sg et la similarité locale Sl .

La mesure de similarité globale Sg_{ab} calcule le nombre de dépendances et d'indépendances communes entre chaque paire possible de variables (X, Y) dans la tâche a et la tâche b (c'est à dire à partir de leurs données respectives D_a et D_b).

$$Sg_{ab} = \sum_{X < Y} \mathbb{1}(I_a(X, Y) - I_b(X, Y)) \quad (3)$$

où $I_a(X, Y)$ et $I_b(X, Y)$ sont respectivement les résultats du test statistique d'indépendance entre les variables X et Y à partir des données D_a et D_b .

La mesure de similarité locale $Sl_{ab}(X, Y|S)$ compare les résultats des tests d'indépendance entre deux variables X et

Y conditionnellement à un ensemble de variables S .

$$Sl_{ab}(X, Y|S) = \begin{cases} 1, & \text{si } I_a(X, Y|S) = I_b(X, Y|S). \\ 0.5, & \text{sinon.} \end{cases} \quad (4)$$

où $I_a(X, Y|S)$ et $I_b(X, Y|S)$ sont respectivement les résultats des tests d'indépendance conditionnelle entre X et Y étant donné S dans les données D_a et D_b . À partir des deux mesures précédentes, Luis, Sucar, and Morales définissent la mesure de similarité combinée $Sc_{ab}(X, Y|S)$ comme suit:

$$Sc_{ab}(X, Y|S) = Sg_{ab} \times Sl_{ab}(X, Y|S) \quad (5)$$

Ils introduisent également la mesure de confiance α_a qui estime la confiance du test d'indépendance entre X et Y conditionnellement à S à partir des données de la tâche a et la définissent ainsi:

$$\alpha_a(X, Y|S) = 1 - \frac{\log N_a}{2N_a} \times T \quad (6)$$

où $T = |X| \times |Y| \times |S|$, avec $|x|$ la cardinalité de x , et où N_a est le nombre d'observations dans les données D_a .

Enfin, la fonction d'indépendance combinée I_a , qui calcule le test d'indépendance entre X et Y conditionnellement à S pour la tâche a en exploitant le transfert d'information, est une combinaison linéaire pondérée des mesures d'indépendances dans a et dans sa tâche la plus similaire b^* (celle qui maximise la similarité combinée Sc):

$$I_c(X, Y|S) = \alpha_a(X, Y|S) \times \text{sgn}(I_a(X, Y|S)) + \alpha_{b^*}(X, Y|S) \times Sc_{ab^*}(X, Y|S) \times \text{sgn}(I_{b^*}(X, Y|S)) \quad (7)$$

où $\text{sgn}(I)$ vaut $+1$ si X et Y sont indépendant conditionnellement à S et -1 sinon.

Méthodes hybrides

Afin de tirer parti des avantages respectifs des algorithmes de recherche d'indépendances conditionnelles et de ceux basés sur un score, de nombreux travaux ont mené à des méthodes hybrides (Daly, Shen, and Aitken 2011). Généralement, ces algorithmes commencent par mettre en oeuvre une stratégie de recherche sur contraintes pour réduire l'espace des DAGs candidats, puis ils exécutent une stratégie basée sur l'utilisation d'un score pour trouver un DAG optimal dans l'espace restreint. Dans ce contexte, nous pouvons citer Max-Mix Hill-Climbing (MMHC) proposé par (Tsamardinos, Brown, and Aliferis 2006). MMHC est une approche hybride d'apprentissage (ST) de structure de réseau bayésien qui identifie d'abord le squelette du graphe avec une recherche sous contraintes (nommée MMPC pour Max-Min Parent Children) puis sélectionne et oriente les arêtes pertinentes en utilisant une procédure de recherche avec score.

L'algorithme MMPC utilise une métrique d'association $A(X, Y|S)$ (telle que l'information mutuelle ou χ^2) pour estimer la dépendance entre X et Y étant donné S et déduit un test d'indépendance conditionnelle $I(X, Y|S)$ à partir de cette métrique. Il identifie progressivement pour chaque

variable X un ensemble de parents et d'enfants candidats $CPC(X)$ (sans distinction entre parent et enfant).

En ce qui concerne l'étape d'orientation des arcs, Tsamardinos, Brown, and Aliferis effectuent une recherche gloutonne pour déterminer le DAG le plus fidèle aux données. La principale différence avec une recherche gloutonne usuelle est que l'espace de recherche des réseaux bayésiens est réduit aux DAGs cohérents avec les $CPCs$ identifiés par MMPC.

MT-MMHC

Les algorithmes d'apprentissage de structure de RBs mentionnés dans les sections précédentes proposent principalement des méthodes d'apprentissage par transfert ST avec des approches de recherche d'indépendances conditionnelles (Luis, Sucar, and Morales 2010) ou de l'apprentissage multi-tâches avec des algorithmes basés sur l'utilisation d'un score (Niculescu-Mizil and Caruana 2007).

Dans notre contribution, nous proposons une approche d'apprentissage hybride dans un contexte multi-tâches. Le but est d'apprendre simultanément k structures de réseaux bayésiens à partir de k problèmes similaires, en tirant parti des avantages de la recherche sous contraintes, des algorithmes basés sur score et aussi des techniques d'apprentissage par transfert et multi-tâches.

Le processus général de MT-MMHC

L'idée principale est d'étendre l'algorithme MMHC au contexte multi-tâches tel qu'il est décrit dans la figure 1. Comme son homologue ST, la procédure commence par une phase de recherche d'indépendances conditionnelles MMPC pour identifier les ensembles de $CPCs$ associés à chaque tâche (voir les cases grises de la figure 1). Pour adapter cette procédure au contexte multi-tâches avec apprentissage par transfert, nous proposons dans la section suivante une mesure d'association combinée pour effectuer les test d'indépendances.

Dans la deuxième phase, pour orienter les arcs, nous appliquons l'algorithme MT Greedy Search (MT-GS) proposé par Niculescu-Mizil and Caruana que nous adaptons à notre contexte en contraignant l'espace de recherche aux $CPCs$ tel qu'il sera décrit dans ce qui suit.

La mesure d'association combinée

La première phase de MT-MMHC, que nous désignons par MT-MMPC, consiste à effectuer k MMPC parallèles avec une nouvelle mesure d'association combinée qui prend en compte la similarité entre les tâches (se référer aux cases grises du schéma de la figure 1). Inspirés par le travail introduit par Luis, Sucar, and Morales, nous proposons la définition suivante:

$$A_{ca}(X, Y|S) = \frac{\alpha_a(X, Y|S)A_a(X, Y|S) + \alpha_{b^*}(X, Y|S)S_{c_{ab^*}}(X, Y|S)A_{b^*}(X, Y|S)}{\alpha_a(X, Y|S) + \alpha_{b^*}(X, Y|S)S_{c_{ab^*}}(X, Y|S)} \quad (8)$$

où $A_a(X, Y|S)$ est la mesure d'association usuelle entre deux variables X et Y étant donné un ensemble de variables S à partir des données D_a et $A_{b^*}(X, Y|S)$ est la mesure d'association entre les deux mêmes variables X et Y étant

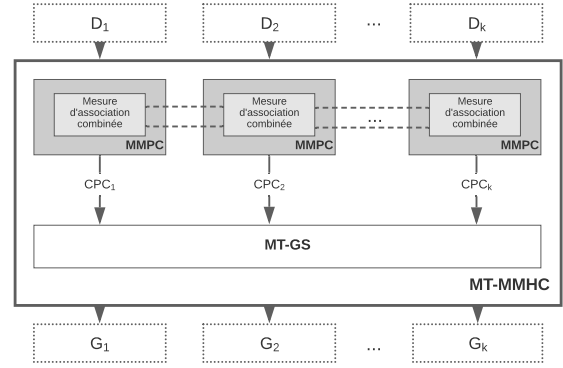


Figure 1: Schéma descriptif du processus global de MT-MMHC

donné l'ensemble de variables S à partir des données D_{b^*} de la tâche la plus similaire b^* déterminée par la mesure de similarité combinée S_c .

Cette mesure d'association combinée nous permet d'enrichir le résultat des tests d'indépendance dans chaque tâche avec ceux de ses tâches les plus similaires. Nous proposons d'utiliser $A_{ca}(X, Y|S)$ au lieu de la fonction d'indépendance combinée I_{ca} définie dans l'équation 7 pour la raison suivante : en tant que combinaison convexe de deux mesures d'association, cette valeur peut également être interprétée comme une mesure d'association. I_{ca} est la combinaison linéaire non convexe de deux signes de tests d'indépendance qui n'est utilisée que pour son signe, alors que l'algorithme MMPC requiert également l'information fournie par la valeur même de l'association entre les variables.

Recherche gloutonne multi-tâches contrainte aux CPCs

MT-MMPC, la première étape de MT-MMHC basée sur une recherche d'indépendances conditionnelles, produit l'ensemble des $CPCs$ pour chaque tâche, $\{CPC_1, CPC_2, \dots, CPC_k\}$. Pour la deuxième phase de MT-MMHC, similairement à MMHC, nous proposons d'appliquer l'algorithme MT-GS. La principale différence réside dans la manière avec laquelle nous contraignons l'espace de recherche de réseaux bayésiens en utilisant les informations apportées par les $CPCs$.

Le voisinage d'une configuration $\mathcal{G} = \{G_1, G_2, \dots, G_k\}$ est obtenu en appliquant pour chaque paire de noeuds dans chaque sous-ensemble possible de \mathcal{G} une des opérations "ajouter", "supprimer", "inverser" ou "ne pas modifier" l'arc qui les relie. Nous adaptons la génération de ce voisinage en ne permettant l'ajout de l'arête $X \rightarrow Y$ dans le graphe G_a que lorsque $X \in CPC_a(Y)$.

En utilisant cette définition de voisinage, nous pouvons effectuer une recherche gloutonne multi-tâches tout en conservant les propriétés du greedy search dans un contexte MT et en respectant les contraintes fournies par le résultat de MT-MMPC. Par la même occasion, nous réduisons l'espace

des DAGs candidats et diminuons ainsi le coût de calcul.

Expérimentations

Dans cette section, nous présentons une évaluation empirique de MT-MMHC sous la forme d'une étude comparative avec les algorithmes d'apprentissage de structures de réseaux bayésiens à tâche unique GS et MMMHC, ainsi qu'avec l'algorithme multi-tâches MT-GS.

Protocole expérimental

Génération de benchmarks multi-tâches Les benchmarks de réseaux bayésiens dans un contexte ST sont disponibles et répandus dans la littérature avec des modèles de référence tels que ASIA (Lauritzen and Spiegelhalter 1988) ou ALARM (Beinlich et al. 1989) (voir Table. 1). Il n'existe cependant aucun benchmark pour évaluer les algorithmes d'apprentissage dans un contexte multi-tâches. Nous proposons ici une procédure simple qui prend en entrée un des modèles de références utilisés dans l'apprentissage ST, et génère k modèles similaires référents pour de l'apprentissage multi-tâches.

Comme première version de cette procédure, nous générons k réseaux bayésiens similaires en appliquant une marche aléatoire entre chaque modèle (c'est à dire en appliquant aléatoirement ℓ opérations usuelles: ajout, suppression ou inversion d'arc) et en recalculant aléatoirement les paramètres (voir figure 2).

Nous pouvons alors générer un jeux de données pour chaque modèle avec la méthode du Forward Sampling, avec potentiellement un nombre différent d'observations que nous appellerons "taille des données" N_a pour chaque tâche a .

Dans le cadre de ce travail, nous avons généré trois séries d'expérimentations, avec des données de petite taille ($N_a \in [500, 1000[$), de taille moyenne ($N_a \in [1000, 5000[$) et de grande taille ($N_a \in [5000, 10000[$) pour $k = 5$ tâches différentes générées avec $\ell = 1$.

Algorithmes Plusieurs algorithmes sont déjà implémentés dans PILGRIM¹, notre bibliothèque C++ dédiée aux modèles graphiques probabilistes. Nous proposons de comparer MT-MMHC, MT-GS et l'exécution indépendante de k algorithmes d'apprentissage à tâche unique kST-GS et kST-MMHC. Dans nos expérimentations, nous utilisons l'information mutuelle comme mesure d'association, avec $\alpha = 5\%$ pour les tests d'indépendance, le score BIC comme approximation de chaque vraisemblance marginale $P(D_a|G_a)$ dans l'équation 1 et $\delta = 1e - 7$ comme pénalité dans l'équation 2.

Métriques d'évaluation Nous mesurons les performances en termes de temps d'exécution et de distance de Hamming (SHD) entre les structures apprises et les structures de référence (celles des réseaux bayésiens desquels nous avons générés les données d'apprentissage). Pour chaque expérimentations, nous présentons la moyenne et l'écart-type de la distance de Hamming (SHD) sur 10

exécutions x 5 tâches, ainsi que la moyenne et l'écart-type du temps d'exécution sur les 10 tours.

Résultats empiriques

MT-MMHC vs kST-GS et kST-MMHC La figure 3 (haut) représente les performances en termes de SHD (plus la valeur est faible, le mieux c'est) pour les trois approches MT-MMHC, kST-GS et kST-MMHC en fonction des trois catégories de tailles de données et pour deux modèles de références ASIA et ALARM.

Les tendances globales sont comme attendues: (1) plus la taille des données est grande, plus les réseaux appris sont précis et (2) kST-MMHC est plus performant que kST-GS sauf pour les petits jeux de données. Il est intéressant de noter que MT-MMHC est capable de construire de meilleurs modèles que les approches d'apprentissage à tâche unique.

La figure 3 (bas) montre le temps d'exécution moyen pour apprendre cinq tâches. MMHC est l'algorithme le plus rapide dans toutes les expérimentations, mais nous pouvons noter que MT-MMHC est bien plus rapide qu'une recherche gloutonne classique lorsqu'il s'agit de réseaux bayésiens de taille moyenne tel que ALARM avec un grand nombre d'observations.

MT-MMHC vs MT-GS La figure 4 affiche les performances en termes de SHD et de temps d'exécution pour MT-MMHC et MT-GS en fonction de la taille des données, pour des benchmarks MT générés à partir du modèle de référence ALARM.

Pour les petits jeux de données, MT-GS présente des résultats légèrement meilleur que MT-MMHC mais avec un temps d'exécution beaucoup plus élevé. Pour des données de taille plus importantes, MT-MMHC est plus performant en termes de précision du modèle appris et s'exécute dans un temps acceptable.

Conclusion

Dans cet article, nous proposons une extension au contexte multi-tâches d'un algorithme efficace de découverte de structures de réseaux bayésiens, MMHC. MT-MMHC est le premier algorithme hybride d'apprentissage multi-tâches de structures de réseaux bayésiens par transfert. Il peut apprendre simultanément plusieurs structures de réseaux bayésiens en induisant de l'information entre les tâches les plus similaires.

Afin de valider notre approche, nous avons également proposé une procédure de génération de benchmarks MT à partir de n'importe quel modèle de référence, en contrôlant la similarité entre les tâches.

Les résultats de nos expérimentations montrent qu'il est plus avantageux d'apprendre des tâches connexes simultanément que de les considérer individuellement. Ils attestent aussi que notre mesure d'association combinée et l'utilisation des CPCs découverts par MT-MMPC pour contraindre la recherche gloutonne ont contribué à construire des modèles plus précis dans un temps d'exécution moindre. Dans ces expériences, MT-MMHC a été capable d'apprendre de meilleurs modèles que kST-GS, kST-MMHC

¹<https://pilgrim.univ-nantes.fr/>

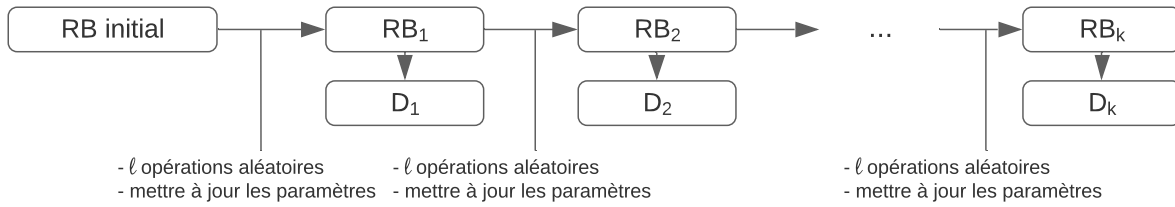


Figure 2: Procédure de génération de benchmarks multi-tâche à partir d'un réseau bayésien de référence

Table 1: Descriptif des réseaux bayésiens utilisés dans nos expérimentations.

Nom	Noeuds	Arcs	Description
Asia	8	8	Utilisé pour diagnostic clinique thoracique compte tenu des symptômes et des facteurs de risque du patient
Alarm	37	46	Diagnostic médical pour le suivi des patients en réanimation

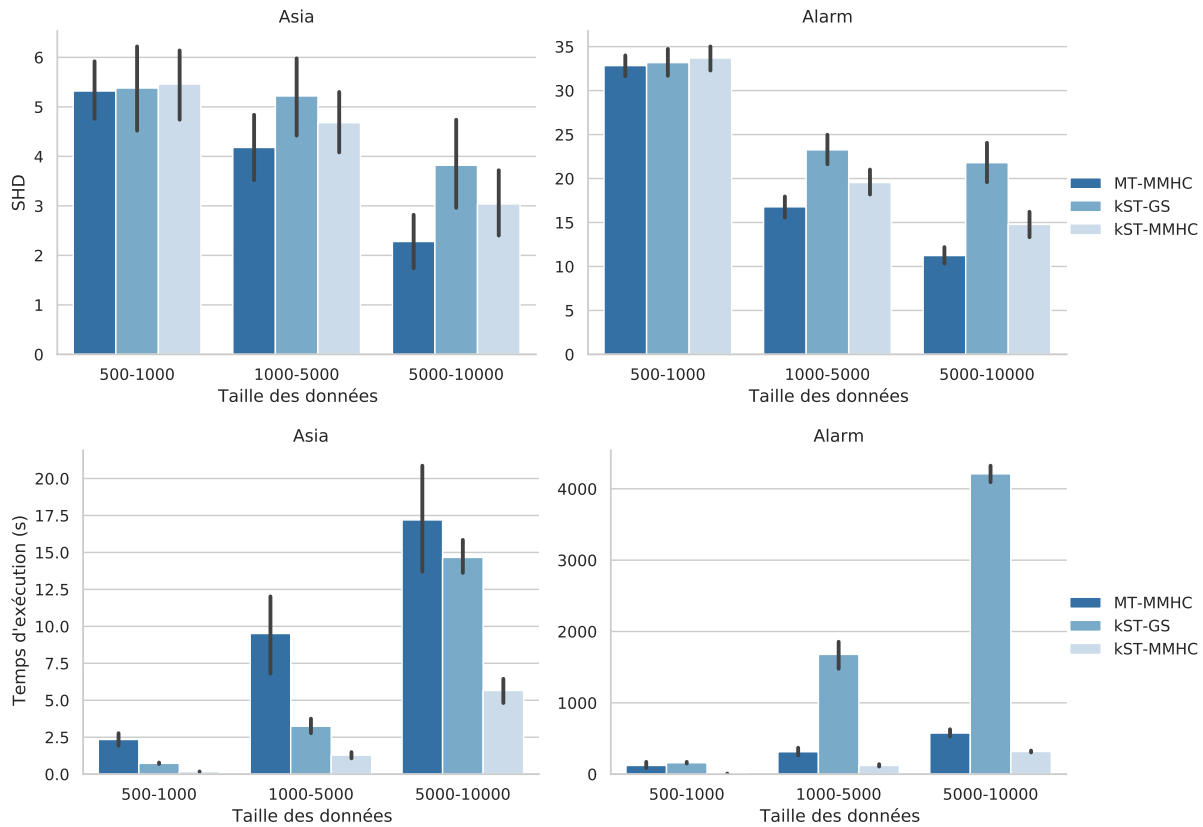


Figure 3: SHD (haut) et temps d'exécution (bas) en fonction de la taille des données pour les algorithmes MT-MMHC, kST-GS and kST-MMHC (MT benchmarks générés respectivement à partir des modèles de références ASIA et ALARM).

et MT-GS pour des benchmarks MT de taille moyenne à grande.

Ce travail est la première étape de notre recherche, avec plusieurs perspectives en vue. À très court terme, nous envisageons de réaliser des expériences plus importantes avec d'autres jeux de données pour consolider l'intérêt de notre proposition. Nous avons également l'intention de travailler sur des procédures de génération de benchmarks

MT supplémentaires, par exemple en utilisant des marches aléatoires plus "longues", ou en créant des tâches qui n'ont pas nécessairement toutes les variables en commun.

Enfin, notre objectif est de combiner ces algorithmes d'apprentissage MT de structures de RBs avec des techniques de confidentialité différentielle (déjà proposé dans le cadre d'apprentissage ST dans (Zhang et al. 2017) par exemple) afin de proposer un cadre général pour de l'apprentissage fédéré de réseaux bayésiens, c'est à dire de

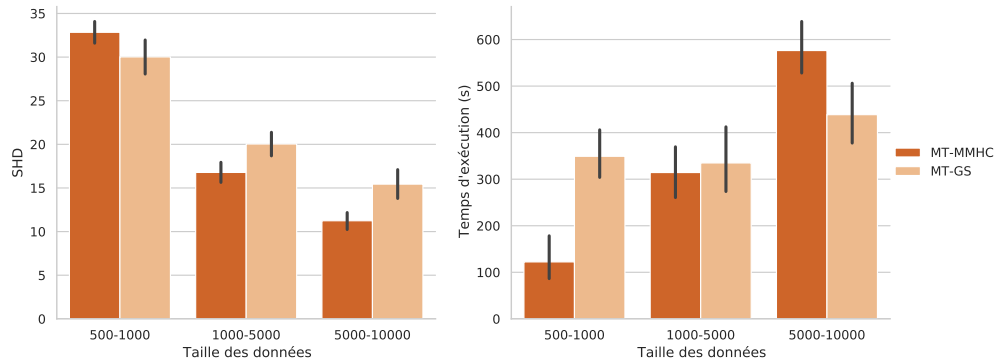


Figure 4: SHD (gauche) et temps d'exécution (droite) en fonction de la taille des données pour les algorithmes MT-MMHC et MT-GS pour des benchmarks MT généré à partir du réseaux bayésien ALARM.

l'apprentissage collaboratif de structure et paramètres de réseaux bayésiens sans partage de données d'apprentissage et en respectant la protection de la vie privée. Notre intérêt est en outre d'utiliser ce cadre pour le développement d'assistants médicaux basés sur les réseaux bayésiens tels que "Medical Companion" (Mouchabac et al. 2021).

Reconnaissance

Ce travail a été soutenu par le programme doctoral ANR Alby4 (Intelligence Artificielle par l'Homme, pour l'Homme) (ANR-20-THIA-0011) et le programme régional Atlanstic2020.

References

Azzimonti, L.; Corani, G.; and Scutari, M. 2020. Structure Learning from Related Data Sets with a Hierarchical Bayesian Score. In Jaeger, M.; and Nielsen, T. D., eds., *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, 5–16. PMLR.

Beinlich, I. A.; Suermondt, H. J.; Chavez, R. M.; and Cooper, G. F. 1989. The ALARM Monitoring System: A Case Study with two Probabilistic Inference Techniques for Belief Networks. In Hunter, J.; Cookson, J.; and Wyatt, J., eds., *AIME 89, Second European Conference on Artificial Intelligence in Medicine*, volume 38 of *Lecture Notes in Medical Informatics*, 247–256. Springer.

Carvalho, A. M. 2009. Scoring functions for learning Bayesian networks. Technical Report 54/2009 Apr 2009, INESC-ID.

Daly, R.; Shen, Q.; and Aitken, S. 2011. Learning Bayesian networks: approaches and issues. *The Knowledge Engineering Review* 26(2): 99–157.

Jia, H.; Wu, Z.; Chen, J.; Chen, B.; and Yao, S. 2018. Causal Discovery with Bayesian Networks Inductive Transfer. In Liu, W.; Giunchiglia, F.; and Yang, B., eds., *Knowledge Science, Engineering and Management - 11th International Conference, KSEM 2018, Changchun, China, August 17-19,*

2018, *Proceedings, Part I*, volume 11061 of *Lecture Notes in Computer Science*, 351–361. Springer.

Lauritzen, S. L.; and Spiegelhalter, D. J. 1988. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society. Series B (Methodological)* 50(2): 157–224.

Leray, P. 2006. Réseaux bayésiens: apprentissage et modélisation de systèmes complexes. *habilitation à diriger les recherches, Université de Rouen*.

López-Cruz, P. L.; Larrañaga, P.; DeFelipe, J.; and Bielza, C. 2014. Bayesian network modeling of the consensus between experts: An application to neuron classification. *International Journal of Approximate Reasoning* 55(1): 3–22.

Luis, R.; Sucar, L. E.; and Morales, E. F. 2010. Inductive transfer for learning Bayesian networks. *Machine learning* 79(1): 227–255.

Mouchabac, S.; Leray, P.; Adrien, V.; Gollier-Briant, F.; and Bonnot, O. 2021. Beyond big data in behavioral psychiatry, the place of Bayesian Network. Example from a pre-clinical trial of an Innovative smartphone application to prevent suicide relapse. *Journal of Medical Internet Research* 16/03/2021:24560: (in press).

Niculescu-Mizil, A.; and Caruana, R. 2007. Inductive Transfer for Bayesian Network Structure Learning. In Meila, M.; and Shen, X., eds., *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, 339–346. San Juan, Puerto Rico: PMLR.

Oyen, D.; and Lane, T. 2012. Leveraging Domain Knowledge in Multitask Bayesian Network Structure Learning. *Proceedings of the AAAI Conference on AI* 26(1).

Oyen, D.; and Lane, T. 2013. Bayesian discovery of multiple Bayesian networks via transfer learning. In *IEEE International Conference on Data Mining*.

Salaken, S. M.; Khosravi, A.; Nguyen, T.; and Nahavandi, S.

2017. Extreme Learning Machine Based Transfer Learning Algorithms. *Neurocomput.* 267(C): 516–524.

Scutari, M.; Vitolo, C.; and Tucker, A. 2019. Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation. *Stat. Comput.* 29(5): 1095–1108.

Silander, T.; Leppä-Aho, J.; Jääsaari, E.; and Roos, T. 2018. Quotient normalized maximum likelihood criterion for learning Bayesian network structures. In *International Conference on Artificial Intelligence and Statistics*, 948–957. PMLR.

Spirites, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. MIT press, 2nd edition.

Thung, K.-H.; and Wee, C.-Y. 2018. A Brief Review on Multi-Task Learning. *Multimedia Tools Appl.* 77(22): 29705–29725.

Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning* 65(1): 31–78.

Zhang, J.; Cormode, G.; Procopiuc, C. M.; Srivastava, D.; and Xiao, X. 2017. PrivBayes: Private Data Release via Bayesian Networks. *ACM Trans. Database Syst.* 42(4).

Zhang, Y.; and Yang, Q. 2021. A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering* 1–20.

Zhou, Y.; Hospedales, T. M.; and Fenton, N. 2016. When and where to transfer for Bayesian network parameter learning. *Expert Systems with Applications* 55: 361–373.