Introduction

Ce document nous présente différentes techniques d'ingénierie des connaissances pour la construction de réseaux bayésiens pour les diagnostics médicaux lorsque les informations probabilistes disponibles sont incomplètes.

Même si on ne peut pas remplacer des probabilité complètes et précises, il est possible de construire un système de diagnostic utile avec des données imparfaites ou manquantes en introduisant une dépendance conditionnelle non modélisée.

Le document nous présente une solution qui détermine les influences combinées de différentes maladies/facteurs sur l’apparition de certaines autres maladies.

Il s’agit d’un des outils les plus performant au niveau du diagnostic médical et il est assez ressemblant aux processus d'ingénierie des connaissances des systèmes experts.

Mettre en place un réseau bayésien est souvent plus compliqué à mettre en place qu’un système expert car il doit contenir un nombre important d’information quantitative. On va donc devoir utiliser les statistiques partielles pour estimer les tables de probabilité selon les différents facteurs indiqués.

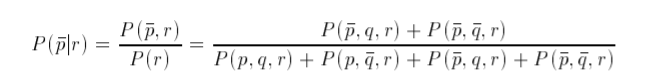
Mais il y à quelques problèmes : beaucoup de données statistiques doivent être estimées et on doit incorporer des connaissances partielles sur le comportement du réseau (additional nodes : états pathologiques cachées) que l’on pourra représenter par des fonctions OR et AND.

Construction

Il s’agit d'une factorisation des distributions de probabilité distinctes. Il faut spécifier une probabilité pour chacune des combinaisons possibles. Si on connaît la distribution de probabilité distinctes on peut obtenir l’inférence en utilisant les probabilité-à-postériori ([probabilité](https://fr.wikipedia.org/wiki/Probabilit%C3%A9) recalculée ou remesurée qu'un [évènement](https://fr.wikipedia.org/wiki/%C3%89v%C3%A9nement_(probabilit%C3%A9s)) ait lieu en prenant en considération une nouvelle information.)

Exemple du document : si on à 3 variable p, q et r on à la probabilité suivante :

Problème : beaucoup trop de variables pour certains critères.

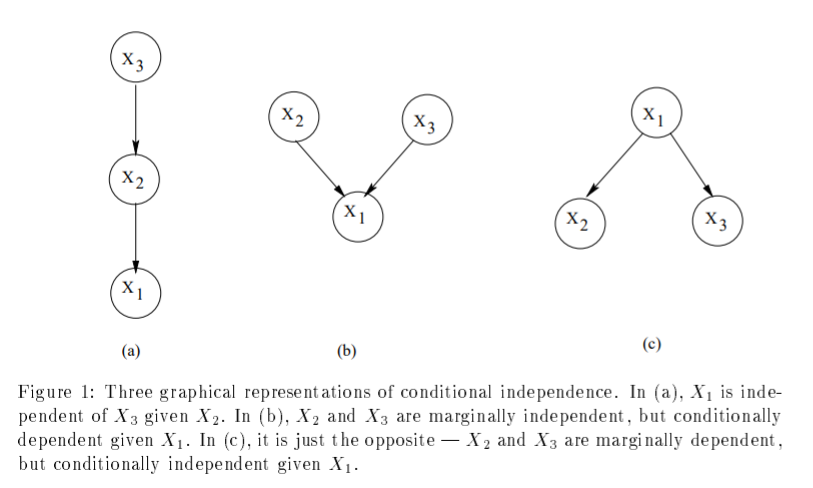


Toutes les probabilités sont des probabilité d’événement élémentaire (événement qui ne contient qu'un seul résultat dans l' espace d'échantillonnage) et peuvent être lues à partir de la représentation de la distribution de probabilité distinctes.

On peut représenter cette distribution à l’aide d’une table de probabilité conjointe multidimensionnelle (tableau avec les différentes probabilités selon les différents critères) dont la clé est la combinaison des valeurs des variables. Ce qui nous amène à un problème : le nombre de données/critères de notre application qui est trop important car la taille de notre table est exponentielle selon notre nombre de données et le nombre de probabilité atomique de l'équation est elle aussi exponentielle au nombre de variable (p q ou r) non mentionné dans la requête. Donc plus nous aurons de données, plus il y aura de variables non utilisées et donc plus elle prendra de temps. Il faut donc mettre en place une forme efficace en utilisant les réseaux bayésiens et les algorithmes de raisonnement associés .

On va donc devoir réduire le nombre de probabilité de notre tableau.

-l'indépendance conditionnelle : À et B sont conditionnellement indépendantes par rapport à C si P(À|C,B) = P(À|C) (si on connait C, connaitre B ne change pas notre connaissance de À). On peut les représenter grâce à des Graphe acyclique dirigée (GAD) et des tables de probabilités . Exemple de GAD :



Il faut donc identifier les variables intéressantes puis identifier lesquelles sont conditionnellement indépendantes. Puis avec différentes conditions indépendantes on peut créer le GAD. Le DAG d'un réseau bayésien est lié au GAD d'un ensemble de règles propositionnelles comme celles utilisées dans les systèmes experts. On doit ensuite préciser les probabilité dans les tables de proba associé au GAD. Nous devrons donc utiliser les documents envoyés par le Docteur Draira pour calculer les probabilités de cette partie.

Construire une table de probabilité conditionnelle complète avec des stats partielles

Le plus gros problèmes dans la construction est la détermination des entrées dans les tables CPT locales pour les nœuds du réseau.

Si n maladie influence une découverte alors la table aura 2 puissance n entrée.

Donc le problème de compilation va apparaître quand on à 2 facteurs ou plus pour l’apparition d’une maladie.

Le plus souvent nous avons la sensibilité du résultats P(F|D) (mesure de la prévalence du résultat F parmi ceux qui ont le critère D) et la spécificité P(F|D) (montre que l’absence de résultats entraîne l’absence de la maladie chez le patient). Si il y à une sensibilité égale, alors la spécificité plus élevée indiquera une probabilité de présence d’une maladie plus élevée. On peut obtenir ces données à partir de statistiques de prévalence. Ces chiffres sont faciles à obtenir car ils sont modulables et car ils représentent la relation objective entre une maladie et un facteur, on pourra donc les obtenir assez facilement à l’aide du docteur.

Partie proba ?? Demander de l’aide à M Faux

Porte ou signification en proba

probabilité de liaison de porte ? (noisy-OR)

affectation de verite particulière ?

probabilité de fuite ? Probabilité de lien

proba de fuite = facteur de fuite ??

Dépendance conditionnelles et excès de confiances