

*Les réseaux bayésiens
et leur application en Data Mining*

Ecole KDD 2000

Paul MUNTEANU

Centre de Recherche du Groupe ESIEA

Plan de l'exposé

- Introduction
- Notions de représentation des connaissances et d'inférence avec des réseaux bayésiens
- **Notions d'apprentissage des réseaux bayésiens**
- Interprétation causale des réseaux bayésiens
- Recommandations bibliographiques
- Conclusion

Introduction

- Les réseaux bayésiens sont le formalisme de représentation des connaissances incertaines le plus utilisé actuellement (voir actes des conférences *UAI*)
- Hybridation réussie entre :
 - la compréhensibilité des résultats des méthodes symboliques
 - les fondements probabilistes rigoureux des méthodes statistiques
 - la structure en réseau de composants simples des approches connexionnistes

Applications des réseaux bayésiens

- Réalisation de systèmes experts par recueil d'expertise

- aide au diagnostic médical, technique, logiciel

- assistants "intelligents" (Office 97)

(selon Bill Gates, l'avantage compétitif à terme de Microsoft serait son expertise dans les réseaux bayésiens [Los Angeles Times, octobre 1996])

- applications militaires de pilotage automatique des systèmes de défense

- etc.

- Plus récemment : *Data Mining* par apprentissage automatique de réseaux bayésiens

Réseaux bayésiens et Data Mining

- Les réseaux bayésiens offrent un cadre unitaire pour traiter un grand nombre de tâches de *Data Mining* :
 - découverte de relations entre les variables plus puissante que la découverte d'associations classique
 - *clustering* (Autoclass, par exemple, cherche un réseau bayésien à structure contrainte)
 - apprentissage supervisé
 - peuvent caractériser n'importe quelle variable à partir de n'importe quelles autres
 - permettent de relâcher les conditions d'indépendance conditionnelle du classificateur naïf de Bayes

Inférence probabiliste

- Le domaine étudié est représenté par un ensemble de variables (attributs) X_1, X_2, \dots, X_n
- On dispose des valeurs d'un sous-ensemble de ces variables (variables observées)
- On veut caractériser les valeurs possibles d'un autre sous-ensemble de variables (variables requises) sous la forme de leurs distributions de probabilités

Représentation des connaissances

- La distribution des probabilités jointes (DPJ) :
$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \equiv P(x_1, x_2, \dots, x_n)$$

pour toutes les combinaisons possibles des valeurs des variables
- La connaissance de la DPJ est suffisante pour mener à bien toute inférence portant sur l'ensemble des variables
- La représentation de la DPJ nécessite un nombre de paramètres exponentiel par rapport au nombre de variables (plus d'un milliard pour 20 variables binaires)

Indépendance conditionnelle

- Soit A , B et C des ensembles de variables et a , b et c des combinaisons de valeurs pour ces ensembles de variables
- Si $P(A = a|C = c) = P(A = a|B = b, C = c) \forall a, b, c$ alors on dit que les variables A sont indépendantes des variables B si on connaît les variables C
- La relation d'indépendance conditionnelle entre les variables A et B est symétrique
- Idée : exploiter les indépendances conditionnelles entre les variables afin de représenter la DPJ d'une manière plus compacte

Exploitation des indépendances conditionnelles

- Règle du ET :

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i | X_{i-1} = x_{i-1}, X_{i-2} = x_{i-2}, \dots, X_1 = x_1)$$

- Soit $Pa_i \subset \{X_{i-1}, X_{i-2}, \dots, X_1\}$ tels que :

$$P(X_i = x_i | X_{i-1} = x_{i-1}, X_{i-2} = x_{i-2}, \dots, X_1 = x_1) = P(X_i = x_i | Pa_i = pa_i)$$

pour toutes les valeurs possibles

- Formulation équivalente :

- la variable X_i est indépendante des variables $\{X_{i-1}, X_{i-2}, \dots, X_1\} \setminus Pa_i$ si on connaît les variables Pa_i

Représentation compacte de la DPJ

- Alors on peut écrire la DPJ sous la forme :

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i | Pa_i = pa_i)$$

– notation plus compacte : $P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | pa_i)$

- Au lieu d'un nombre de valeurs exponentiel par rapport au nombre de variables, on a besoin, pour chaque variable, d'un nombre de valeurs exponentiel par rapport au nombre de ses parents

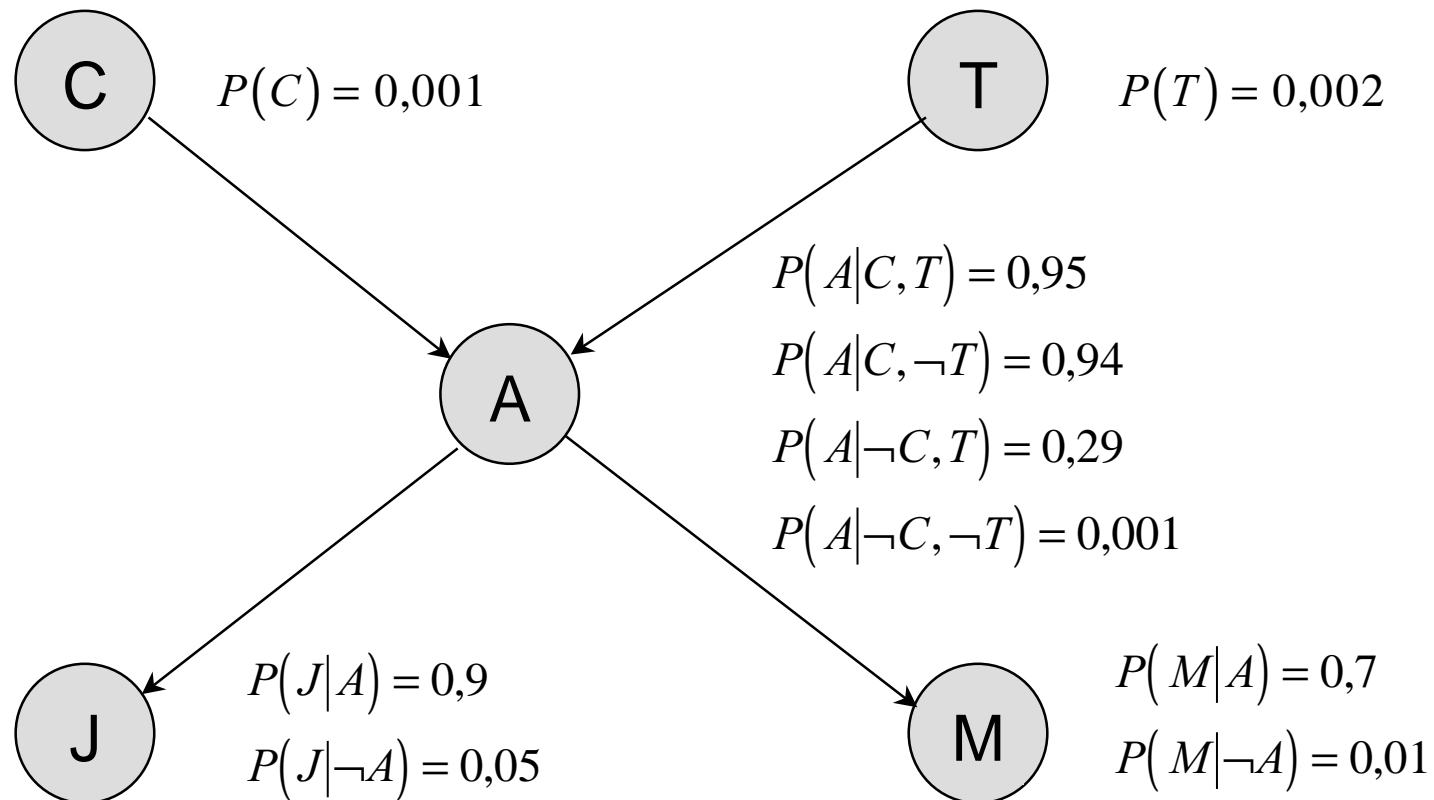
Réseaux bayésiens (RB)

- Graphe orienté sans cycle
 - les nœuds représentent les variables du domaine
 - les arcs représentent les influences directes entre les variables
- Tables de probabilités conditionnelles (TPC) attachées à chaque nœud X_i
 - soit Pa_i les parents du nœud X_i
 - la TPC du nœud X_i contient les probabilités $P(X_i = x_i | Pa_i = pa_i)$ pour toute valeur x_i de X_i et toute combinaison de valeurs pa_i de Pa_i

Exemple - l'histoire

- [Stuart & Russell, *AI - A Modern Approach*]
 - une personne habitant Los Angeles (tremblements de terre fréquents) installe chez elle une alarme contre les cambriolages
 - elle demande à ses amis John et Mary (qui restent chez eux dans la journée) de l'appeler au travail s'ils entendent l'alarme
 - John appelle toujours quand il croît entendre l'alarme mais la confond parfois avec la sonnerie du téléphone
 - Mary aime écouter de la musique forte et, donc, n'entend pas toujours l'alarme

Exemple - le réseau bayésien



Construction de RB

- Algorithme théorique :
 - Choix d'un ordre sur les variables
 - Pour i de 1 à n :
 - création du nœud correspondant à la variable X_i
 - recherche d'un ensemble minimal de variables Pa_i , parmi les variables déjà traitées, tels que la variable X_i est indépendante des variables $\{X_{i-1}, X_{i-2}, \dots, X_1\} \setminus Pa_i$ si on connaît les variables Pa_i
 - création d'arcs orientés entre les variables Pa_i et la variable X_i
 - création de la TPC correspondant à la variable X_i
- La complexité du réseau obtenu est très dépendante de l'ordre des variables
- Dans la pratique, on s'appuie sur les relations de causalité directe entre les variables

Inférence à partir des RB

- La connaissance de la DPJ est suffisante pour mener à bien toute inférence probabiliste portant sur les variables du domaine
- L'inférence dans les réseaux bayésiens est NP-difficile
- Il existe cependant des algorithmes d'inférence suffisamment rapides pour la plupart des applications pratiques
 - inférence exacte : la méthode des arbres de jonction
 - inférence approximative : méthodes type Monte-Carlo

Exemple - types d'inférence possibles

- Diagnostique : $P(C|J)$?
- Inférence causale : $P(J|C)$?
- Inférence inter-causale : $P(C|A,T)$?
- Inférence mixte :
 - $P(A|J,\neg T)$? : combinaison de diagnostic et d'inférence causale
 - $P(C|J,\neg T)$? : combinaison de diagnostic et d'inférence inter-causale

Méthodes d'apprentissage des RB

- Plusieurs catégories de méthodes selon que :
 - la structure du réseau est **connue** ou **non**
 - apprentissage de paramètres
 - apprentissage de structure
 - les valeurs de tous les attributs sont **disponibles** ou non
 - on va s'intéresser uniquement à la première catégorie
 - pour les cas où certaines valeurs des variables sont absentes d'une manière aléatoire (valeurs manquantes) ou systématique (variables cachées) des méthodes d'apprentissage de type échantillonnage de Gibbs ou EM (*Expectation-Maximisation*) ont été développées

Apprentissage des paramètres

- La méthode la plus simple : le comptage

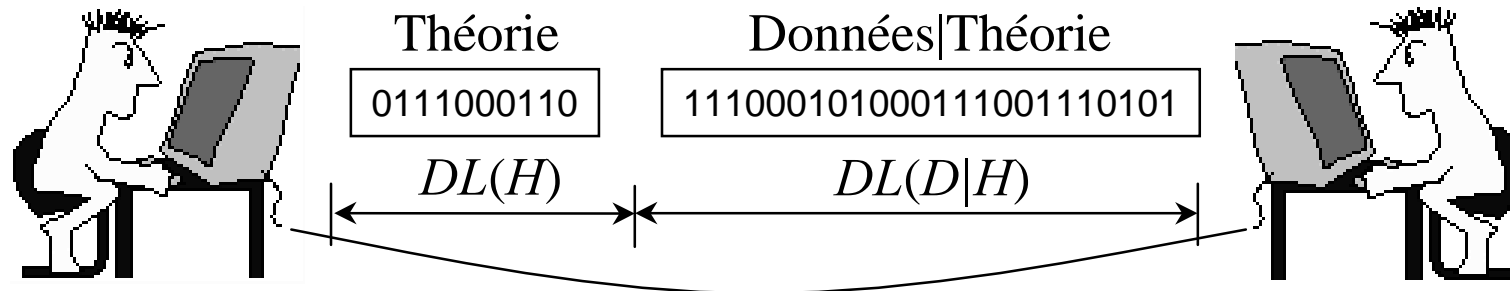
$$P(X = x | U_1 = u_1, U_2 = u_2, \dots, U_n = u_n) = \frac{\|\{e | (X(e) = x) \wedge (U_1(e) = u_1) \wedge (U_2(e) = u_2) \wedge \dots \wedge (U_n(e) = u_n)\}\|}{\|\{e | (U_1(e) = u_1) \wedge (U_2(e) = u_2) \wedge \dots \wedge (U_n(e) = u_n)\}\|} = \frac{N(X = x, U = u)}{N(U = u)}$$

- Risque : certaines combinaisons parents-enfant peuvent être peu ou pas du tout représentées
- Méthodes plus sophistiquées :
 - estimation bayésienne avec prise en compte des probabilités *a-priori*
 - utilisation de modèles locaux (arbres de décision) pour représenter la dépendance parents-enfant

Apprentissage de la structure

- Définir une fonction d'évaluation des différentes structures possibles et utiliser une stratégie de recherche heuristique pour trouver la structure « optimale »
- Deux méthodes génériques en apprentissage pour définir les fonctions d'évaluation :
 - le principe de la description minimale (*MDL = Minimum Description Length*)
 - l'approche bayésienne qui passe par le calcul des probabilités *a posteriori*

Le principe MDL



- L'apprentissage est vu comme un problème de compression de l'information
- La meilleure théorie est celle qui minimise la longueur totale du message transmis : $DL(H) + DL(D|H)$

Codage de la théorie

- Théorie = réseau bayésien R

$$DL(R) = DL(G) + DL(TPC|G)$$

$$\begin{aligned} DL(G) &= \sum_i DL(\text{noeud}_i) = \sum_i \left(DL(\|Pa_i\|) + DL(Pa_i \| \|Pa_i\|) \right) \\ &= \sum_i \left(\log(n) + \log\left(C_n^{\|Pa_i\|}\right) \right) \end{aligned}$$

$$\begin{aligned} DL(TPC|G) &= \sum_i DL(TPC_i|G) = \sum_i \left(\text{val}(Pa_i) \cdot (\text{val}(X_i) - 1) \cdot DL(\text{prob}) \right) \\ &= \sum_i \left(\text{val}(Pa_i) \cdot (\text{val}(X_i) - 1) \cdot \frac{\log(N)}{2} \right) \end{aligned}$$

Codage des données

- Données = somme des descriptions des exemples

$$\begin{aligned} DL(D|R) &= \sum_{j=1}^N DL(e_j|R) = -\sum_{j=1}^N \log P_R(e_j) \\ &= -\sum_{j=1}^N \log \left(\prod_{i=1}^n P_R(x_{ij}|pa_{ij}) \right) = -\sum_{j=1}^N \sum_{i=1}^n \log P_R(x_{ij}|pa_{ij}) \\ &= -\sum_{i=1}^n \sum_{x_i, pa_i} N(X_i = x_i, Pa_i = pa_i) \log P_R(x_i|pa_i) \end{aligned}$$

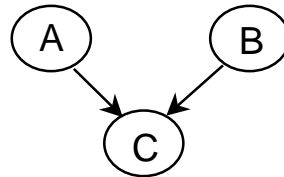
- Les paramètres qui minimisent cette expression sont

$$P_R(x_i|pa_i) = \frac{N(X_i = x_i, Pa_i = pa_i)}{N(Pa_i = pa_i)}$$

- Donc : $DL(D|R) = N \cdot H(X_i|Pa_i)$

Exemple de codage des données

- Soit le réseau



et deux exemples à transmettre : $e_1 = (v, v, v)$ et $e_2 = (v, f, v)$

- $P_R(A=v, B=v, C=v) = P_R(A=v).P_R(B=v).P_R(C=v|A=v, B=v)$
- $P_R(A=v, B=f, C=v) = P_R(A=v).P_R(B=f).P_R(C=v|A=v, B=v)$
- Le nombre de bits nécessaires est donc :

$$DL(D|R) = -(2 \log_2(P_R(A=v)) + \log_2(P_R(B=v)) + \log_2(P_R(B=f)) + \log_2(P_R(C=v|A=v, B=v)) + \log_2(P_R(C=v|A=v, B=f)))$$

Le score MDL

$$\begin{aligned} \text{score}(R, D) &= DL(R) + DL(D|R) \\ &= \sum_i \left(\log(n) + \log\left(\binom{n}{\|Pa_i\|}\right) + \text{val}(Pa_i) \cdot (\text{val}(X_i) - 1) \cdot \frac{\log(N)}{2} + N \cdot H(X_i|Pa_i) \right) \\ &= \sum_i \text{score}_i(R, D) \end{aligned}$$

- Le score se décompose donc en une somme de sous-scores locaux à chaque nœud (propriété importante pour l'optimisation)

L'approche bayésienne

- $P(G|D) = \frac{P(G)P(D|G)}{P(D)}$ avec $P(D)$ indépendant de G
- $scoreB(G, D) = P(G)P(D|G)$: le score "Bayésien"
- $P(G)$ réalise une pénalisation des structures dissemblables à une structure *a priori* ou, en absence d'une telle structure, des structures complexes (rasoir d'Occam)
- $P(D|G)$ doit prendre en compte toutes les valeurs possibles des TPC :
$$P(D|G) = \int P(D|\Theta_G, G)P(\Theta_G|G)d\Theta_G$$

Le score BD

- Si :
 - on accepte certaines hypothèses qui peuvent être résumées sous la forme : "chaque distribution $P(X_i|P\alpha_i)$ peut être apprise indépendamment des autres"
 - les *a priori* sur les paramètres peuvent être représentés sous une certaine forme analytique (loi de Dirichlet)
- Alors :
 - l'intégrale de la vraisemblance peut être calculée analytiquement
 - le score résultant porte le nom de "Bayésien Dirichlet" (BD)

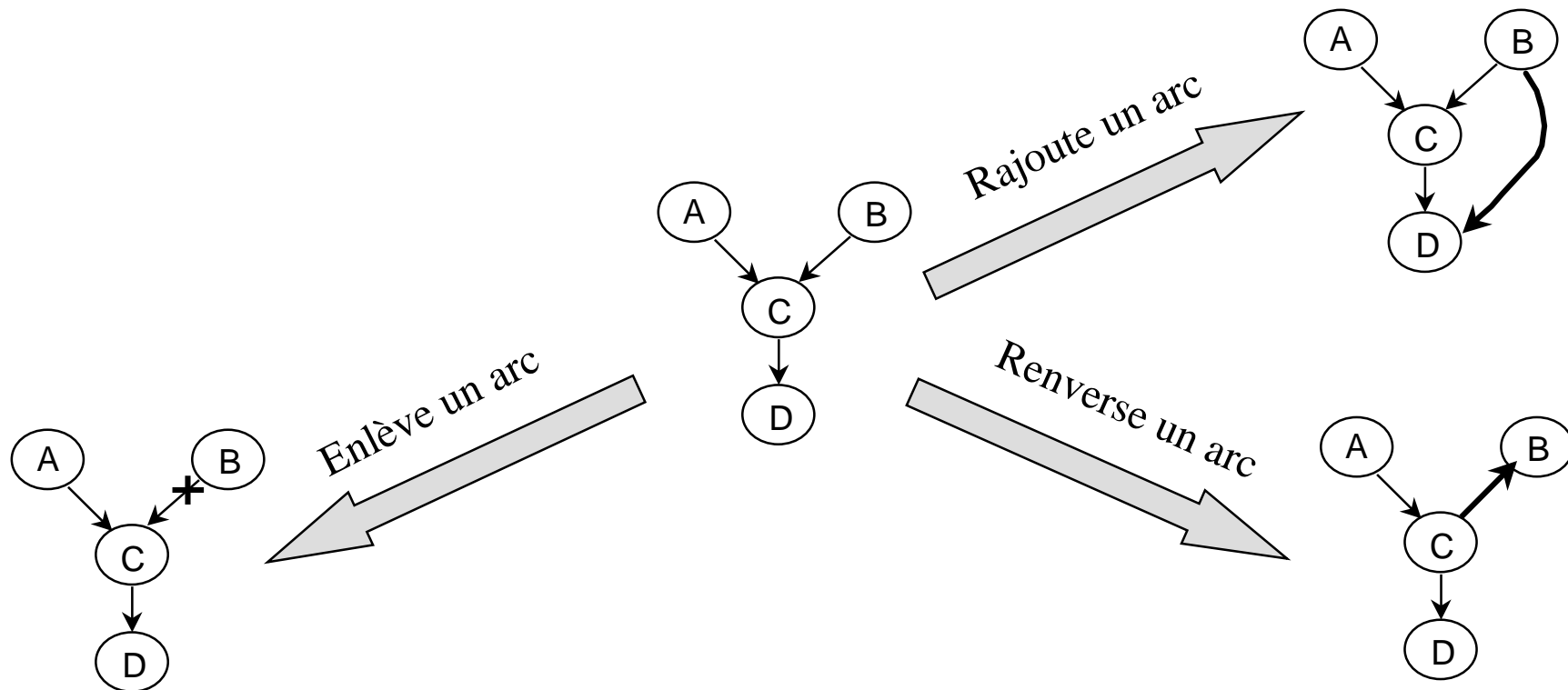
Le score BDe

- Problème avec le score BD :
 - demande l'obtention des *a priori* pour tous les paramètres de toutes les structures évaluées : infaisable !
- Si en plus des hypothèses déjà acceptées :
 - on accepte que la vraisemblance ne puisse pas départager des structures qui représentent les mêmes relations d'indépendance conditionnelle (structures "équivalentes")
- Alors :
 - on aboutit à une formule analytique et à une méthode d'élicitation des *a priori* sur les paramètres qui définissent le score Bde
- Les scores BD et BDe sont décomposables sur les nœuds

Optimisation des scores

- La recherche de la structure optimale d.p.d.v. de ces scores (et de la plupart des autres) est NP-difficile
→ Recherche heuristique
- Le plus souvent : recherche gloutonne, tabou ou recuit simulé
- Ses stratégies de recherche demandent la définition d'un ensemble d'opérateurs de transformation qui permettent l'exploration de l'espace de recherche

Opérateurs de transformation



Détails algorithmiques

- Le point de départ de la recherche :
 - le réseau totalement déconnecté
 - un réseau *a priori* fourni par l'utilisateur
 - un réseau à structure contrainte qui peut être obtenu rapidement (l'arbre de recouvrement maximal)
- La décomposabilité du score :
 - facilite le calcul du score des hypothèses transformées : au plus deux scores locaux à actualiser
 - peut être utilisée pour optimiser les calculs à l'aide d'une mémoire « cache »

Interprétation causale des réseaux bayésiens

- Contestée par certains chercheurs : peut-on vraiment apprendre des causalités à partir des données (sans faire des expériences) ?
- Ses partisans s'appuient sur des hypothèses qui, selon eux, permettent de donner à certains arcs la sémantique de liens de causalité directe entre les variables
- A l'origine d'une famille de méthodes d'apprentissage de structure reposant directement sur des tests d'indépendance conditionnelle :
 - plus efficaces que les méthodes fondées sur un score
 - capables dans certains cas de découvrir la présence de variables cachées
 - très sensibles aux erreurs faites par les tests d'indépendance conditionnelle et à leur paramétrage

Recommandations bibliographiques

- Tutoriaux :
 - D. Heckerman, Bayesian Networks for Data Mining, *Data Mining and Knowledge Discovery*, 1, 1997
 - N. Friedman, M. Goldszmidt, Learning Bayesian Networks from Data, *AAAI 1998 Tutorial*, transparents disponibles sur le Web
- Recueil d'articles sur l'apprentissage :
 - M.I. Jordan (éd.), *Learning in Graphical Models*, Kluwer, 1998
- Ouvrage sur la causalité :
 - P. Spirtes, C. Glymour, R. Scheines, *Causation, Prediction and Search*, Springer-Verlag, 1993
- (Le seul) ouvrage en français :
 - A. Becker, P. Naïm, *Les réseaux bayésiens*, Eyrolles, 1999

Conclusion

- Les réseaux bayésiens ont déjà fait leurs preuves dans des nombreux champs d'application de l'IA
- L'apprentissage des réseaux bayésiens offre des perspectives très prometteuses dans le *Data Mining*
 - déjà des applications finalisées (détection des impayés chez AT&T)
- Domaine scientifique en plein essor, confronté à de nouveaux défis
 - optimisation algorithmique, notamment pour le traitement des valeurs manquantes et des variables mixtes