Indexation de structures de documents par réseaux bayésiens

Mohamed Ali Mahjoub[*,**] **- Khlifia Jayech**[**]

[*] *Institut Préparatoire aux Etudes d'Ingénieurs de Monastir*
Rue Ibn Eljazzar Monastir 5019
Tunisie
Medali.mahjoub@ipeim.rnu.tn

[**] *Unité de recherche en Signaux, image et documents, Ecole nationale d'ingénieurs de Sousse Eniso Tunisie*
jayech_k@yahoo.fr

RÉSUMÉ : *Notre objectif est d'étudier l'apport des réseaux naïfs augmentés dans les problèmes de classification d'images. Les images utilisées dans notre étude représentent la structure d'un type de documents qui contiennent des blocs de textes et de graphiques. Nous avons proposé trois variantes des réseaux bayésiens. En premier lieu les réseaux bayésiens naïfs RN qui malgré leur structure simple ont donnés un très bons résultats. En second lieu, les réseaux bayésiens naïfs augmentés par un arbre TAN. En effet, l'hypothèse d'indépendance entre les attributs est généralement fausse. A cet effet, l existe différentes techniques pour assouplir cette hypothèse. En troisième lieu, les réseaux bayésiens naïfs augmentés par une forêt FAN qui bien qu'ils soient assez connus dans les problèmes de classification, n'ont pas été explorés à notre connaissance en imagerie. Les résultats obtenus ont montré une nette amélioration du FAN par rapport aux réseaux RN et TAN*

MOTS-CLÉS : *Réseaux bayésiens – structure de documents – Réseau naïf – TAN - FAN*

Abstract : Our objective is to study the contribution of naive increased Bayesian networks in problems of image classification. The images used in this study represent the structure of a document containing text blocks and graphics. We proposed three variants of Bayesian networks. First naive Bayesian networks RN who, despite their simple structure and strong assumption on independence have given very good results. Secondly, the naive Bayesian networks augmented by a tree TAN. Indeed, the assumption of independence among attributes is in general false. Thus, there are different techniques to relax this assumption. Thirdly, the naive Bayesian networks augmented by a forest called FAN who they are rather well known classification problems have not been investigated to our knowledge in image classification. The results showed a marked improvement over the FAN network RN and TAN.

KEYWORDS: *Bayesian network – document structure – RN – TAN - FAN*

1. Introduction

Le domaine d'indexation par le contenu est depuis des années un domaine de recherche très actif. En particulier le développement d'approches pour l'indexation de structures de documents est un axe qui prend de plus en plus d'intérêt vu l'apport que peut apporter surtout pour tout ce qui outils d'aide à la navigation dans les grandes bases documentaires tels que les documents d'archives ou les documents anciens.

Alors qu'un objet est décrit par un ensemble de pixels au niveau physique, il est, après annotation manuelle, segmentation ou reconnaissance de forme, identifié et décrit par un symbole ou une étiquette au niveau logique. Une image symbolique est donc représentée par un ensemble de symboles ou icônes représentant les objets d'intérêt identifiés dans l'image (Hsu 1999, Huang 2007). Les images utilisées représentent la structure d'un type de documents qui contiennent des blocs de textes et de graphiques.

Dans notre travail, nous avons opté une approche spatiale basée sur une méthodologie assez récente qui s'intéresse plus particulièrement à la modélisation des images par des réseaux bayésiens. Des variantes des réseaux bayésiens appelées réseaux naïfs ou encore réseaux naïfs augmentés par un arbre ou part une forêt (Freidman 1996, François 2006) sont au centre d'intérêt de notre étude.

On commence par présenter la structure des images utilisées ainsi que la base d'image d'images traitées. En effet, puisque notre étude porte essentiellement sur l'exploration des modèles des réseaux bayésiens, nous allons considérer des images pré-segmentées à régions identifiées.

Par la suite, il est question d'une présentation des réseaux bayésiens et leurs variantes Rn, TAN et FAN. L'accent sera particulièrement sur les algorithmes associés aux phases d'apprentissage et d'inférence.

La dernière partie sera consacrée à l'évaluation des différents algorithmes.

2. Structure d'images de documents

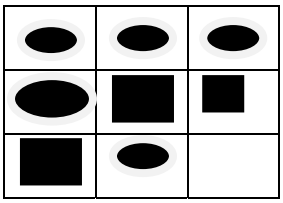
Plusieurs approches ont été proposées pour décrire les relations spatiales entre objets dans une image. Mais les approches, les plus puissantes pour la correspondance et la recherche par le contenu spatial des images, sont la projection symbolique (2D String), les arbres et les graphes relationnels attribués (Attribut Relation Graph). Dans le cadre de ce travail, Nous avons opté pour les réseaux bayésiens comme étant des outils de raisonnement utilisant des graphes acycliques orientés pour la représentation des relations causales et des probabilités conditionnelles pour exprimer l'incertitude sur ces relations. Mais le problème qui se pose comment adapter ces réseaux face au problème d'indexation.

La structure graphique d'une image peut être chiffrée pour qu'elle soit transformée en une représentation arithmétique descriptive.

En prenant comme exemple de chiffrement :

0 : consacré pour le vide. 1 : pour remplacer la forme carré de taille petite.
2 : pour remplacer la forme carré de taille grande.
3 : pour remplacer la forme cercle de taille petite.
4 : pour remplacer la forme cercle de taille grande.

On peut transposer l'image si dessous sous forme de suite de 0,1, 2,3 et 4 (figure 1):



a

1	**1**	**1**
2	**4**	**3**
4	**1**	**0**

b

Figure 1. *Image symbolique*
a) image structuée en 9 zones (A1,…,A9) b) image codée (1 1 1 2 4 3 4 1 0)

Ainsi considérant une base d'apprentissage, on peut la structurer en cinq classes (tableau 1) modélisant par exemple des entités de type graphique ou texte. En effet, notre objectif à long terme est le développement d'outils d'indexation et de navigation dans les bases de documents anciens présentant des structures particulières.

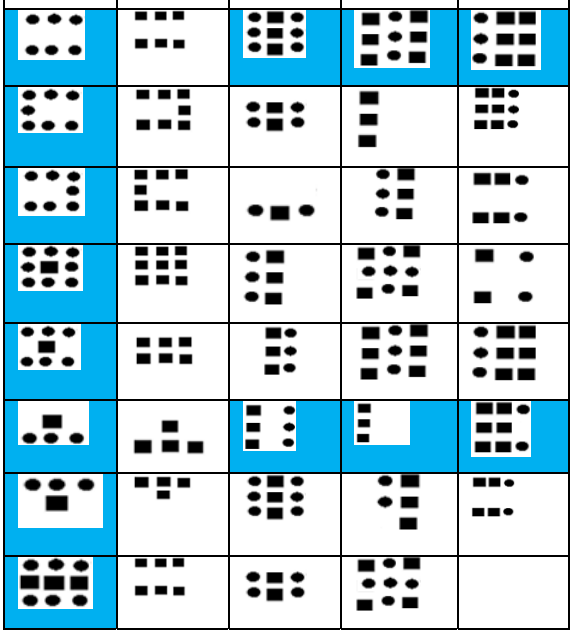
Classe1	Classe2	Classe3	Classe4	Classe5
				

Tableau 1. *Structuration d'images en 5 classes*

Dans le cadre des expérimentations réalisées et dont les résultats sont présentées dans la section 7 nous avons considéré une base d'apprentissage comportant environ 3000 images réparties uniformément sur les 5 classes et une base de test d'environ 1800 images.

3. Rappel sur les réseaux bayésiens

3.1. *Principe*

Les réseaux bayésiens ont été développés au début des années 1980 pour résoudre certains problèmes de prédiction et d'abduction, courants en intelligence artificielle (IA). Ils s'appuient sur un théorème : Le théorème de Bayes. C'est un résultat de base en théorie des probabilités, issu des travaux du révérend Thomas Bayes (1702-1761), présenté en 1763. Voici ces résultats :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A)$ est la *probabilité a priori* de *A*. Elle est aussi appelée la *probabilité marginale* de *A*.
- $P(A|B)$ est appelée la *probabilité a posteriori* de *A* sachant *B*.
- $P(B|A)$, pour un *B* connu, est appelée la *fonction de vraisemblance* de *A*.
- De même, le terme $P(B)$ est la *probabilité marginale* ou *a priori* de *B*.

Un réseau bayésien est définie par :

- Un graphe oriente sans circuit (DAG) G= (V, E), où V est l'ensemble des nœuds de G, et E l'ensemble des arcs de G ;
- Un *espace probabilisé* fini (Ω, Z, p) ;
- Un ensemble de *variables aléatoires* associées aux nœuds du graphe et définies sur (Ω, Z, p), tel que :

$$p(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} p(V_i | C(V_i))$$

Où $C(V_i)$ est l'ensemble des causes (parents) de V_i dans le graphe G.
Un réseau bayésien est donc un graphe causal auquel on a associé une représentation probabiliste sous-jacente. Cette représentation permet de rendre quantitatifs les raisonnements sur les causalités que l'on peut faire à l'intérieur du graphe (Naim 2007).
L'utilisation essentielle des réseaux bayésiens est de calculer des probabilités conditionnelles d'événements reliés les uns aux autres par des relations de cause à effet. Cette utilisation s'appelle *inférence* qu'on va le détailler dans le paragraphe suivante.

Une difficulté essentielle des réseaux bayésiens se situe précisément dans l'opération de transposition du graphe causal à une représentation probabiliste.

4. Réseaux bayésiens naïfs

Une variante des réseaux bayésiens est appelée réseaux bayésiens naïfs. Ces réseaux ont une structure simple et unique qui comprend deux niveaux. Le premier niveau contient un seul nœud parent et le second plusieurs enfants avec la forte hypothèse *naïve* d'indépendance conditionnelle des enfants (**X**) conditionnellement au parent.

Les réseaux bayésiens naïfs sont largement utilisés pour résoudre des problèmes de classification (Smail 2004, Cerquide 2003). En effet, la classification est assurée en considérant le nœud parent comme une variable *non observée* précisant à quelle classe appartient chaque objet et les nœuds enfants comme étant des variables *observées* correspondant aux différents attributs spécifiant cet objet. La figure 3 rappelle le principe de fonctionnement et de classification ra réseau bayésien naïf.

Par conséquent, en présence d'un ensemble d'apprentissage, la seule investigation à faire est de calculer les probabilités conditionnelles en appliquant la règle de décision de Bayes comme suit :

$$\begin{aligned} d(X) &= argmax_{classe} P(classe|X) \\ &= argmax_{classe} P(X|classe) x P(classe) \\ &= argmax_{classe} \prod_{i}^{N} P(X_i|classe) x P(classe) \end{aligned}$$

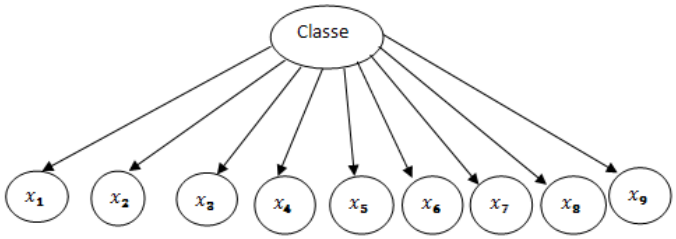


Figure 3. *Réseau naïf*

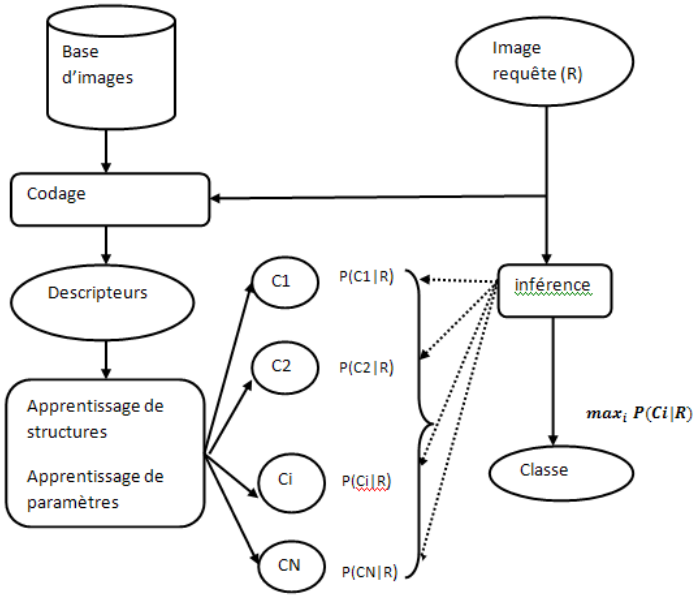


Figure 4. *Architecture du système*

4.1. *Apprentissage du modèle naïf*

Soit un ensemble d'apprentissage, le réseau bayésien naïf correspond à cet ensemble est représenté par la figure 3. Notons que les neuf nœuds enfants correspondent aux neuf attributs caractérisant les neuf zones de l'image et qu'ils sont considérés comme indépendants dans le contexte du nœud parent (Classe) qui correspond au classe. D'autre part, l'apprentissage des paramètres peut être réalisé selon le principe de l'algorithme suivant (Ben Amor 2006) :

Entrée :

- I : image symbolique chiffrée sous forme d'une suite de 0, 1, 2, 3 et 4.
- L'ensemble d'apprentissage.

Sortie : - déterminer la classe de l'image

$I=(I_{1I}, I_{2I}, \ldots, I_{nI})$

Algorithme :

Apprentissage :

- Déterminer les fréquences d'apparition de chaque classe à partir de l'ensemble d'apprentissage. Ces fréquences sont calculées comme suit:

$$p(C_i) = \frac{N(C_i)}{N} \quad \text{et} \quad \sum_i p(C_i) = 1$$

Sachant que : $-\ N(C_i)$ est le nombre d'images appartenant à la classe C_i.

N est le nombre total d'images d'apprentissage.

- Déterminer les fréquences d'apparition de chaque terme sachant la classe C_i en utilisant la formule suivante :

$$p(x_j = 1|C_i) = \frac{N(C_i, x_j)}{N(C_i)}$$

Sachant que : - $N(C_i, x_j)$ est le nombre d'images de la classe C_i qui contiennent le terme x_j.

- Inférence :
  - Déterminer la classe de l'image I= $(I_{11}, \dots, I_{ml})$ selon la formule suivante :

$$p(C_i|I) = p(C_i|I_{11}, \dots, I_{ml}) = \frac{p(I_{11}, \dots, I_{ml}|C_i) * p(C_i)}{p(I_{11}, \dots, I_{ml})}$$

D'après l'hypothèse d'indépendance on a : $p(I_{11}, \dots, I_{ml}|C_i) = \prod_j p(I_{jI}|C_i)$. Si la distribution des I_i est égale on doit seulement comparer le numérateur $p(I_{11}, \dots, I_{ml}|C_i) * p(C_i)$.

Le calcul des probabilités conditionnelles et *a priori* se basant sur les fréquences peut s'avérer entaché d'erreur si la valeur d'un attribut n'apparaît pas avec toutes les classes dans l'ensemble d'apprentissage. En effet, ceci peut entraîner des probabilités conditionnelles nulles qui vont réduire à zéro les probabilités de certaines classes.
La technique standard pour éviter ce problème s'appelle *estimateur de Laplace* et elle consiste à ajouter 1 à tous les numérateurs et de compenser ces ajouts dans les dominateurs (tableau 3). En utilisant cet estimateur, le calcul des probabilités conditionnelles et *a priori* est donné dans le tableau 2.

P(Classe)				
C1	C2	C3	C4	C5
0,300	0,200	0,200	0,150	0,150

Tableau 2. *Probabilités à priori des classes*

	P(A3|Classe)				
	C1	C2	C3	C4	C5
0	0,200	0,200	0,102	0,200	0,329
1	0,003	0,005	0,673	0,581	0,465
2	0,003	0,005	0,215	0,206	0,194
3	0,538	0,522	0,005	0,006	0,006
4	0,256	0,268	0,005	0,006	0,006

Tableau 3. *Probabilités non nulles*

4.2. *Inférence dans le modèle naïf*

Une fois le réseau quantifié, il peut être utilisé pour classer de nouveaux images étant données leurs valeurs d'attributs en utilisant la règle de Bayes exprimée par :

$$p(C_i|I) = p(C_i|I_{1I}, \ldots, I_{mI}) = \frac{p(I_{1I}, \ldots, I_{mI}|C_i) * p(C_i)}{p(I_{1I}, \ldots, I_{mI})} = \frac{\prod_j p(I_{jI}|C_i) * p(C_i)}{p(I_{1I}, \ldots, I_{mI})}$$

Notons que nous n'avons pas besoin de calculer explicitement le dénominateur $p(I_{1I}, \ldots, I_{mI})$ puisqu'il est déterminé par la condition de normalisation. Donc il est suffisant de calculer pour chaque classe C_i son degré de vraisemblance exprimé par :

$$\prod_j p(I_{jI}|C_i) * p(C_i)$$

Afin de classer n'importe quelle nouvelle image caractérisée par ses valeurs d'attributs : $I_{1I}, \ldots, I_{mI}$. Les classes choisies seront celles dont la probabilité est la plus grande.

5. Réseaux bayésiens naïfs augmentés par un arbre

L'hypothèse d'indépendance entre les attributs utilisée dans le classifier de Bayes naïf est généralement fausse (hypothèse naïve). Il existe différentes techniques pour assouplir cette hypothèse. Elles consistent à identifier les dépendances conditionnelles entre les attributs. Nous obtenons alors une sous-structure optimale sur les observations en adaptant la méthode de recherche de l'arbre de recouvrement de poids maximal (Maximal Weight Spanning Tree ou MWST) (François 2006,2008).

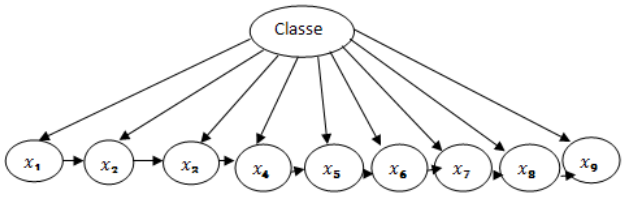


Figure 4. *Réseau augmenté par un arbre*

Cette méthode s'applique à la recherche de structure d'un réseau bayésien en fixant un poids à chaque arête potentielle A–B de l'arbre. Ce poids peut être par exemple l'information mutuelle entre les variables A et B. Une fois cette matrice de poids définie, il suffit d'utiliser un des algorithmes standards de résolution du problème de l'arbre de poids maximal comme l'algorithme de Kruskal ou celui de Prim. L'arbre non dirigé retourné par cet algorithme doit ensuite être dirigé en choisissant une racine puis en parcourant l'arbre par une recherche en profondeur. La racine peut

être choisie soit aléatoirement, soit à l'aide de connaissance a priori (Li 2003, Abbid 2007).

5.1. *Algorithme d'apprentissage de la structure*

Entrée :

- L'ensemble d'apprentissage.
- RB Naïf

Sortie :

- RB Naïf augmenté par un arbre

Algorithme :

1- Calculer l'information mutuelle[1] $I(O_i, O_j|C)$ pour chaque couple d'attributs (O_i, O_j) avec i≠j ;

$$I(O_i, O_j|C) = \sum_{\substack{x \in O_i \\ y \in O_j \\ z \in C}} P(x,y,z) \log \frac{P(x,y|z)}{P(x|z)P(y|z)}$$

2- Construire un graphe complet et non orienté entre les attributs ;

3- Pondérer $I(O_i, O_j|C)$ sur l'arc (O_i, O_j)

4- Transformer le graphe en un arbre dont la somme des arcs est maximum en utilisant l'algorithme MWST.

5- Choisir un nœud comme racine de l'arbre et le transformer en un arbre orienté ;

Ajouter le nœud C et un arc de C vers chaque O_i .

Comme pour le réseau naïf, l'apprentissage des paramètres dans le cas du TAN est basé sur un calcul de fréquences comme illustré dans l'algorithme suivant :

Entrée :

- L'ensemble d'apprentissage.
- Structure des réseaux bayésiens naïf augmenté par un arbre.

Sortie : les paramètres des réseaux bayésiens naïf augmenté par un arbre.

Algorithme :

- Déterminer les fréquences d'apparition de chaque classe à partir de l'ensemble d'apprentissage. Ces fréquences sont calculées comme suit:

$$p(C_i) = \frac{N(C_i)}{N} \qquad et \qquad \sum_i p(C_i) = 1$$

[1] Dans la théorie des probabilités et la théorie de l'information, l'information mutuelle de deux variables aléatoires est une quantité mesurant la dépendance statistique de ces variables.

Sachant que :

- $N(C_i)$ est le nombre d'images appartenant à la classe C_i.
- N est le nombre total d'images d'apprentissage.
- Déterminer les fréquences d'apparition de chaque terme sachant les parents P_i, C_i en utilisant la formule suivante :

$$\begin{cases} p(x_j|C_i,P_i) = \dfrac{N(C_i,P_i,x_j)}{N(C_i,P_i)} & si\ P_i\ existe \\ p(x_j|C_i,P_i) = \dfrac{N(C_i,x_j)}{N(C_i)} & si\ P_i\ n'existe\ pas \end{cases}$$

Sachant que :

- $N(C_i,P_i,x_j)$ est le nombre d'images qui contient le terme x_j et qui a comme parents P_i, C_i.
- $N(C_i,x_j)$ est le nombre d'images de la classe C_i qui contient le terme x_j.

5.2. *Expérimentation*

Nous avons utilisé Matlab, et plus précisément la Bayes Net Toolbox de Murphy et Structure Learning Package décrit dans pour faire l'apprentissage de structure. En effet, en appliquant l'algorithme ci-dessus la matrice de scores[2] obtenus est la suivante :

score_mat =

```
 0
 5.7      0
 5.4   4.8816       0
 5.8   4.8171   5.6694       0
 5.71  4.906    5.5713   5.3432       0
 5.82  4.8425   5.6753   5.0993   3.9640       0
 5.7   4.9203   5.6000   5.3607   3.7945   5.0068       0
 5.8   4.7902   5.6621   5.1859   3.9522   4.8646   5.7847       0
 5.7   4.911    5.551    5.349    3.8201   5.041    5.504    5.50         0
```

D'où l'obtention de la structure suivante (figure 5) :

[2] La matrice de scores est une matrice symétrique. La moitié de la matrice suffit pour l'affichage.

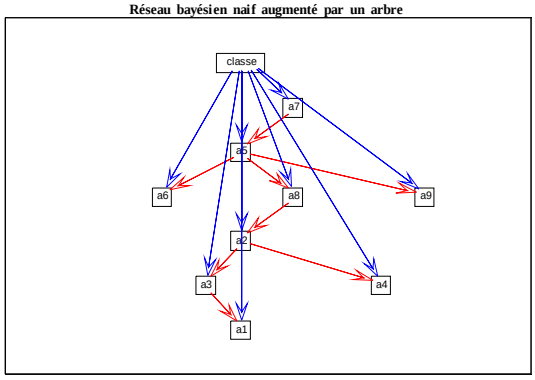


Figure 5. R*éseau augmenté par un arbre TAN*

Une fois le réseau quantifié, il peut être utilisé pour classer de nouveaux images étant données leurs valeurs d'attributs en utilisant la règle suivante:

$$P(C|I) = P(C)\prod_{j} P(I_j|I_{jp}, C)$$

Sachant que I_{jp} est le parent de I_j et :

$P(I_j|I_{jp}, C) = P(I_j|I_{jp}, C)$ si $I_{jp} \neq \emptyset$ sinon
$P(I_j|I_{jp}, C) = P(I_j|C)$ si $I_{jp} = \emptyset$,

6. Réseaux bayésiens naïfs augmentés par une forêt

D'après nos expérimentations, on observe que le réseau bayésien naïf a donné un bon résultat que TAN. Deux facteurs peuvent être à l'origine :

- Les directions de liens dans un TAN sont cruciales. Dans l'étape 5 du l'algorithme TAN, un attribut est choisi aléatoirement comme racine de l'arbre et les directions de tous les liens sont mises après. On remarque que la sélection de l'attribut racine détermine en réalité la structure du TAN résultant, puisqu'un TAN est un graphique dirigé. Ainsi la sélection de l'attribut racine est importante pour construire un TAN.
- Des liens non nécessaires peuvent exister dans un TAN. Dans l'étape 4 du TAN, un arbre de recouvrement de poids maximal est construit. Ainsi, le nombre des liens est fixé à n-1. Parfois, il pourrait être sur adapter avec les données, puisque quelques liens ne peuvent pas être nécessaires d'exister dans le TAN.

En se basant sur les observations précédentes, nous modifions l'algorithme TAN également comme suit :

1- Nous choisissons l'attribut A_{racine} qui a l'information mutuelle maximale avec la classe, définie par l'équation ci-dessous, comme racine.
$A_{racine} = argmax_{A_i} I_p(A_i; C)$
Où $i = 1, \ldots, n$. Il est naturel d'utiliser cette stratégie, c'est à dire l'attribut qui a la plus grande influence sur la classification devrait être la racine de l'arbre.
2- Nous filtrons les liens qui ont des informations mutuelles conditionnelles moins qu'un seuil. À notre compréhension, ces liens ont un risque élevé pour sur adapter les données et ensuite l'évaluation de probabilité. Plus précisément, nous utilisons l'information mutuelle conditionnelle moyenne définie dans l'équation ci-dessous comme un seuil. Tous les liens avec les informations mutuelles conditionnelles moins que $Iavg$ sont enlevés.

$$Iavg = \frac{\sum_i \sum_{j, j \neq i} I_p(A_i; A_j | C)}{n(n-1)}$$

où n est le nombre des attributs

Puisque la structure du modèle résultant n'est pas un arbre strict, nous appelons notre algorithme naïve bayes augmenté par une forêt FAN (Forest Augmented Naive bayes).

6.1. *Algorithme d'apprentissage de la structure*

Entrée :

- L'ensemble d'apprentissage.
- TAN

Sortie :

- RB Naïf augmenté par une foret (FAN)

Algorithme :

1- Calculer l'information mutuelle conditionnelle $I(O_i, O_j | C)$ pour chaque couple d'attributs (O_i, O_j) avec i≠j

$$I(O_i, O_j | C) = \sum_{\substack{x \in O_i \\ y \in O_j \\ z \in C}} P(x, y, z) \log \frac{P(x, y | z)}{P(x|z) P(y|z)}$$

et calculer l'information mutuel conditionnel moyenne $Iavg$ définie dan l'équation ci-dessous :

$$Iavg = \frac{\sum_i \sum_{j, j \neq i} I_p(A_i; A_j | C)}{n(n-1)}$$

2- Construire un graphe complet et non orienté entre les attributs ;
3- Pondérer $I(O_i, O_j | C)$ sur l'arc (O_i, O_j)
4- Transformer le graphe en un arbre dont la somme des arcs est maximum en utilisant l'algorithme MWST.

5- Calculer l'information mutuelle $I_p(A_i; C), i = 1,2,\ldots,n$ entre chaque attribut et la classe, et choisir l'attribut A_{racine} qui a la plus grande information mutuel avec la classe.
$A_{racine} = argmax_{A_i} I_p(A_i; C)$

6- Transformer l'arbre non orienté résultante en un arbre orienté en mettant A_{racine} comme racine.
7- Supprimer les liens dirigés qui ont le poids des informations mutuelles conditionnelles au-dessous des informations mutuelles conditionnelles $Iavg$
8- Ajouter le nœud classe C et un arc de C vers chaque O_i .

En outre, on rappelle que l'algorithme d'apprentissage de paramètres est une extension de celui du TAN.

6.2. Expérimentation

score_mat =
0
5.7 0
5.42 4.8 0
5.83 4.81 5.66 0
5.71 4.9 5.57 5.3 0
5.82 4.84 5.67 5.09 3.9 0
5.74 4.92 5.60 5.36 3.79 5.0 0
5.8 4.71 5.6 5.122 3.9 4.864 5.78 0
5.73 4.91 5.5 5.35 3.81 5.004 5.504 5.5 0

L'apprentissage de structure s'effectue en appliquant l'algorithme FAN comme suit :

- Une fois que nous calculons la matrice de score,
- Nous choisissons la racine en calculant (tableau 4) :

$$A_{racine} = \text{argmax}_{A_i} I_p(A_i; C)$$

$I_p(A_i; C)$	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
C	5.2624	4.4460	5.1211	4.8822	3.6111	4.5738	5.2684	4.1191	4.8951

Tableau 4. *Choix de la racine*

- Nous calculans maintenant $Iavg$:

$$Iavg = \frac{\sum_i \sum_{j,j\neq i} I_p(A_i; A_j|C)}{n(n-1)} = 5.1733$$

- On supprimer les liens qui ont $I <$ $Iavg$

D'où l'obtention de la structure suivante (figure 6) :

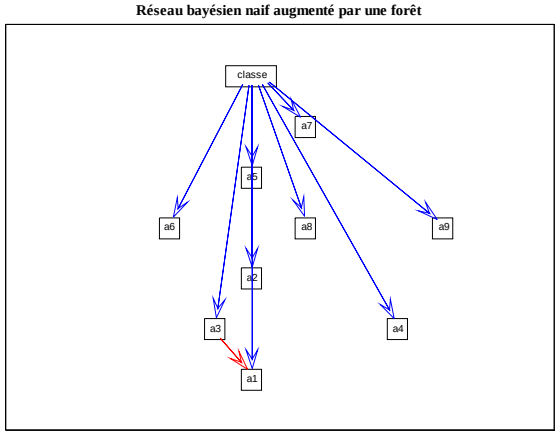


Figure 6. *Réseau augmenté par une forêt FAN*

Sachant qu'on confirme comme dans la section 4.1 et d'après le tableau 4 que uniquement quelques valeurs de probabilité sont significatives ce qui reflète encore une fois la disposition des liens/arcs du graphe de l'image.

7. Résultats

Dans chacune des expérimentations effectuées, l'évaluation de l'efficacité de classification est basée sur le pourcentage de classification correcte PCC des instances appartenant à l'ensemble test défini par:

$$PCC = \frac{\text{nombre des instances correctement classées}}{\text{nombre total des instances classées}}$$

D'autre critères seront également utilisés, à savoir :

- P(X=C1 | classe=C1) et P(X=C2 | classe=C2) représente le taux de bonne classification.
- P(X=C1 | classe=C2) et P(X=C2 | classe=C1) représente le taux de mauvaise classification.

Les résultats des expérimentations sont résumés dans le tableau ci-dessous (tableau 5). Les réseaux bayésiens naïfs, TAN et FAN utilisent la même base d'apprentissage et sont évalués exactement sur les mêmes données test. Le tableau ci-dessous montre que Les réseaux bayésiens naïfs ainsi que TAN et FAN sont complètement en accord avec l'ensemble d'apprentissage qui est donc cohérent. En d'autres

termes, la majorité des instantes d'apprentissage caractérisées par les mêmes valeurs d'attributs appartiennent à la même classe.
D'autre part, considérant la base de test, et d'après tableau 5 ainsi que la figure 7, on remarque que le RN a donné un bon résultat par rapport au TAN. Cependant, le TAN bien qu'il soit un réseau augmenté par un arbre n'est plus satisfaisant comparé au RN. En contre partie le FAN a donné une légère amélioration par rapport au RN et au TAN.

		classe1	classe2	classe3	classe4	classe5	moyenne
R N	Ensemble d'apprentissage	100%	100%	100%	100%	100%	100,00%
	Ensemble test	95%	98%	92%	89%	84%	91,60%
TAN	Ensemble d'apprentissage	100%	100%	100%	100%	100%	100,00%
	Ensemble test	100%	80%	74%	89%	82%	85,00%
FAN	Ensemble d'apprentissage	100%	100%	100%	100%	100%	100,00%
	Ensemble test	97%	98%	92%	93%	86%	93%

Tableau 5. *Résultats de classification*

6. Conclusion

Dans notre travail de recherche, nous avons essayé d'explorer une variante des réseaux bayésiens qui sont les réseaux naïfs RN, les réseaux naïfs augmentés par un arbre et les réseaux naïfs augmentés par une forêt, et ce en vue de les exploiter dans une application de classification d'images structurées. L'objectif principal était de développer des méthodes d'indexation de structures de documents par des approches bayésiennes de calssification.
Pour les expérimentations, nous avons utilisé Matlab, et plus précisément la Bayes Net Toolbox pour faire l'apprentissage de structure. Les résultats obtenus ont montré une nette amélioration du FAN par rapport aux réseaux RN et TAN.

D'après la section précédente, on peut conclure que l'étude de dépendances entre les attributs a une grande influence sur les résultats de classification donc on doit bien

choisir les liens c'est-à-dire choisir les liens qui présentent une forte dépendance entre les attributs donc influent positivement sur les résultats de classification. En effet, un choix de liens bien étudié scientifiquement implique une bonne discrimination - donc un pouvoir discriminant augmenté- entre les classes, ce qui induit certainement une nette amélioration dans le taux de bonne classification.

Bibliographie

Abbid Sharif, Ahmet bakan Tree augmented naïve baysian classifier with feature selection for FRMI data, 2007

Ben Amor, S.Benferhat, Z.Elouedi. « Réseaux bayésiens naïfs et arbres de décision dans les systèmes de détection d'intrusions » *RSTI-TSI*. Volume 25 - n°2/2006, pages 167 à 196.

Chikering DM Learning bayesian network is NP-complete, In Learning from data: artificial intelligence and statistics V, pages 121-130. *Springer-Verlag*, New York 1996

Cerquides J., R.L.Mantaras. Tractable bayesian learning of tree augmented naïve bayes classifiers, January 2003

Cooper JF. Computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, vol.42, p.393-405, 1990.

Nicolas Loménie, Nicole Viencent, rémy Mullot, Les relations spatiales : de la moélisation à la mise en oeuvre, *Revue des nouvelles technologies de l'information*, cépadues éditions 2008

Friedman N., N.Goldszmidt. Building classifiers using bayesian networks. *Proceedings of the American association for artificial intelligence conference*, 1996.

Francois O. De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes. Thèse de doctorat, Institut National des Sciences Appliquées de Rouen, 2006.

Francois O., P.Leray. Learning the tree augmented naïve bayes classifier from incomplete datasets, LITIS Lab., INSA de Rouen, 2008.

Hsu F., Y.Lee, S.Lin. 2D C-Tree Spatial Representation for Iconic Image. *Journal of Visual Languages and Comp*, pages 147-164, 1999.

Huang G., W.Zhang, et L.Wenyin. A Discriminative Representation for Symbolic Image Similarity Evaluation. Workshop on Graphics Recognition, Brazil, 2007.

Leray P. Réseaux Bayésiens : apprentissage et modélisation de systèmes complexes, novembre 2006.

Li X.. Augmented naive bayesian classifiers for mixed-mode data, December, 2003.

Naim P., P.H.Wuillemin, P.Leray, O.Pourret, A.Becker. *Réseaux bayésiens*, Eyrolles, Paris, 2007.

Patrakis EGM. Design and evaluation of spatial similarity approaches for image retrieval. *In Image and Vision Computing* 20, 59-76, 2002.

Smail L.. Algorithmique pour les réseaux Bayésiens et leurs extensions. Thèse de doctorat, 2004