Réseaux bayésiens jumelés et noyau de Fisher pondéré pour la classification de documents XML

Ait Ali Yahia Yassine et Amrouche Karima

Ecole Nationale Supérieure d'Informatique
BP 68M Oued Smar
Oued Smart 16270 Alger
ALGERIE

y_ait_ali_yahia@esi.dz, k_amrouche@esi.dz

RÉSUMÉ. Dans le cadre de cet article, nous nous intéressons à la classification supervisée de documents structurés de type XML. Nous présentons tout d'abord un modèle génératif arborescent jumelé, basé sur le formalisme des réseaux bayésiens, afin de modéliser les documents structurés qui permet de prendre en compte simultanément l'information de contenu et l'information de structure. Ensuite nous appliquons une variante du noyau de Fisher, basée sur la pondération des composantes du vecteur, à ce modèle pour obtenir un modèle discriminant. Enfin, nous testons les deux modèles avec et sans pondération sur un corpus de documents XML en utilisant les méthodes CBS et SVM.

ABSTRACT. In this paper, we are presenting a learning model for XML document classification based on Bayesian networks. Then, we are proposing a model which simplifies the arborescent representation of the XML document that we have, named coupled model and we will see that this approach improves the response time and keeps the same performances of the classification. Then, we will study an extension of this generative model to the discriminating model thanks to the formalism of the Fisher's kernel. At last, we have applied a ponderation of the structure components of the Fisher's vector. We finish by presenting the obtained results on the XML collection by using the CBS and SVM methods.

.**MOTS-CLÉS :** Documents XML, réseaux bayésiens, noyau de Fisher, classification, modèles discriminants.

KEYWORDS: XML documents, Bayesian networks, Fisher's kernel, classification, discriminating models.

1. Introduction

Le développement du web et le nombre croissant de documents électroniques disponibles ont permis l'émergence de formats semi-structurés permettant la représentation et le stockage de documents textuels ou multimédias. Différents formats comme le HTML, le XHTML ou le XML sont aujourd'hui très populaires. Ces formats prennent en compte la structure logique des documents. Nous étudions ici le problème de la classification de documents structurés textuels de type XML. La classification supervisée de documents est une problématique générique de la recherche d'information. Elle est utile pour différentes tâches telles que le filtrage d'email ou de Spam, l'indexation de documents, l'organisation de corpus, etc.

Nous présentons, dans cet article, pour la classification de documents structurés, un modèle génératif, issu du formalisme des réseaux bayésiens, qui permet la prise en compte simultanée de l'information de structure et de l'information de contenu des documents structurés inspiré de [2][3]. Ensuite on lui applique la technique de « jumelage » des nœuds qui simplifie la structure arborescente du document XML. Cette technique a déjà fait l'objet d'un travail précédent [1] où l'on montre qu'elle réduit le temps d'exécution de 30% par rapport aux réseaux classiques tout en préservant les mêmes performances pour la classification des documents XML en termes d'efficacité. On a appelé le modèle résultant : modèle génératif arborescent avec jumelage. C'est ce dernier qu'on a ensuite implémenter. Puis, à partir de ce modèle génératif simplifié, nous sommes passés à un modèle discriminant à l'aide du noyau de Fisher auquel on a appliqué une pondération pour les composantes de structure des vecteurs de Fisher. Nous présentons finalement les résultats obtenus sur un échantillon de la collection de documents XML, issue de la campagne d'évaluation INEX. Dans le travail présenté dans [2] [3], les auteurs utilisent un réseau bayésien classique avec le noyau de Fisher pour la classification. Dans ce travail, on applique les réseaux bayésiens jumelés. De plus, dans le noyau de Fisher on utilise une pondération des composantes du vecteur en tenant compte des fréquences des nœuds dans les classes et ceci, comme on va le voir dans les résultats expérimentaux, améliore sensiblement les résultats de la classification. Par ailleurs on montre que, dans le modèle discriminant, la méthode CBS (Classifier-Based Search) donne de meilleurs résultats que la méthode SVM (Support Vecteur Machine).

2. Modèle génératif avec jumelage pour les documents structurés.

Dans [2], les auteurs ont développé un modèle génératif pour la classification des documents structurés permettant de prendre en compte simultanément l'information

dégagée par la structure du document et de son contenu textuel. Issu du formalisme des réseaux Bayésiens, ce modèle repose sur le principe suivant : l'auteur va tout d'abord décrire à priori la structure (plan) de son document puis « remplir » chacune de ces entités structurelles. Autrement dit, le texte apparaissant dans un nœud du document ne dépend que de l'entité structurelle qui le contient.

2.1. Document structuré

Nous représentons un document structuré comme un graphe orienté sans cycle (DAG pour Directed Acyclic Graph), ce qui correspond à la représentation usuelle utilisée dans les langages à base de balises (HTML et XML). Chaque nœud du graphe représente une entité structurelle (paragraphe, titre, section...) du document et chaque arc représente une relation hiérarchique entre deux entités (par exemple, un paragraphe est inclus dans une section). Ainsi un document *d* peut être vu comme un ensemble de nœuds, où chaque nœud n_i^d est composé de :

– Une *étiquette* (ou label) : un label peut être par exemple paragraphe, section, introduction etc. L'ensemble des étiquettes dépend des documents du corpus que nous traitons.

– Un *contenu* : qui est le texte associé au label du nœud n_i^d , un nœud peut ne pas avoir de contenu (auquel cas, nous considérons le contenu comme vide).

2.2. Modèle génératif pour les documents structurés

Les hypothèses du modèle peuvent être résumées par : « la structure d'un document ne dépend pas du texte contenu dans ce document tandis que le texte du document dépend uniquement de l'unité structurelle qui le contient ». Plusieurs réseaux Bayésiens peuvent être associés à un document en fonction des dépendances à prendre en compte, dans notre cas nous ne nous intéressons qu'aux dépendances père-fils. Ce modèle est appelé modèle parent.

Soit $(n_1^d \ldots\ldots n_{|d|}^d)$ l'ensemble des nœuds d'un document d. On va considérer que le document structuré est modélisée par un réseau bayésien de $|d|$ variables aléatoires $n_1^d \ldots\ldots n_{|d|}^d$ qui correspondent aux nœuds du réseau (DAG) représentant le document. Les arcs du réseau seront modélisés par la fonction $pa(n_i^d)$ qui revoie le parent de la variable n_i^d dans le réseau.

Posons $n_i^d = (s_i^d, t_i^d)$, où s_i^d représente l'étiquette du $i^{éme}$ nœud du document d et t_i^d représente le contenu du $i^{éme}$ nœud du document d. Ainsi un document peut être vu comme la réalisation d'un couple de deux variables aléatoires : $d = (s^d, t^d)$, où : $s^d = (s_1^d, \ldots, s_{|d|}^d)$ et $t^d = (t_1^d, \ldots, t_{|d|}^d)$.

Un modèle ainsi construit nous permet de calculer la probabilité qu'un document ait été généré par une classe : $P(d/\Theta^c)$. Nous allons présenter la décomposition de cette

probabilité pour le cas général c'est-à-dire $\Theta^c = \Theta$**.** La probabilité qu'un document ait été généré par un modèle de paramètres Θ est :

$P(d|\Theta) = P(s^d, t^d|\Theta)$.

Sous les hypothèses posées ci-dessus, cette équation peut être réécrite comme suit :

$$P(d|\Theta) = P(s^d, t^d|\Theta) = P(t^d|s^d, \Theta) \times P(s^d|\Theta) \quad (1)$$

$P(s^d|\Theta)$ est appelée probabilité structurelle et $P(t^d|s^d, \Theta)$ est appelée probabilité textuelle du document ou probabilité du contenu. En considérant les dépendances statistiques entre les labels (hypothèse sur la structure) et les dépendances entre le contenu (hypothèse sur le contenu), nous obtenons le modèle final suivant :

$$P(d|\Theta) = \left(\prod_{i=1}^{|d|} P(t_i^d|s_i^d, \Theta^t_{s_i^d})\right) \times \left(\prod_{i=1}^{|d|} P(s_i^d|pa(s_i^d), \Theta^s)\right). \quad (2)$$

Avec $\Theta = \Theta^s \cup \Theta^t$ et $\Theta^t = \bigcup_{l \in \Lambda} \Theta_l^t$. (3)

Où θ^s et θ^t représentent respectivement l'ensemble des paramètres de structure et l'ensemble des paramètres de contenu.

2.3. Apprentissage des paramètres

2.3.1. Apprentissage des paramètres de structure

Posons $\theta^s_{n,m}$ l'estimation de la probabilité $P(s_i^d = n|pa(s_i^d) = m, \Theta^s)$**.**

Notre but est donc d'estimer la probabilité $\theta^s_{n,m}$ pour chaque valeur $(n, m) \in \Lambda^2$, pour cela nous allons utiliser l'approche du *maximum de vraisemblance* qui consiste à estimer la probabilité d'un événement par la fréquence d'apparition de l'événement dans la base de données. Cette approche nous donne alors :

$$\forall\, n, m \in \Lambda^2, \theta^s_{n,m} = \frac{\sum_{d \in D^c{}_{TRAIN}} NS^d_{n,m} + 1}{\sum_{d \in D^c{}_{TRAIN}} \sum_{n' \in \Lambda} NS^d_{n',m} + |\Lambda|} \quad (4)$$

$NS^d_{n,m}$ est le nombre de fois qu'un nœud de label n possède un père de label m dans l'ensemble des documents de la base d'apprentissage D_{TRAIN}. Λ l'ensemble des étiquettes possibles (i.e. $s_i^d \in \Lambda$).

2.3.2. Apprentissage des paramètres de contenu

2.3.2.1. Modèle génératif local de type Naïve Bayes pour le contenu

L'information de contenu d'un nœud n_i^d est une information textuelle de la forme $t_i^d = \{w^d_{i,k}\}, k \in [1, \dots, |t_i^d|]$, Où $w^d_{i,k}$ est le $k^{ème}$ *mot* du nœud $\boldsymbol{i}$ du document d. L'aspect naïf de ce modèle réside dans le fait qu'on suppose l'indépendance entre les différents mots d'un contenu d'un nœud, cette indépendance nous permet de décomposer la probabilité de contenu.

$$P(t^d|s^d, \Theta^t) = \prod_{i=1}^{|d|} P\left(t_i^d \middle| s_i^d, \Theta^t_{s_i^d}\right) = \prod_{i=1}^{|d|} \left(\prod_{k=1}^{|t_i^d|} P(w^d_{i,k}|s_i^d, \Theta^t_{s_i^d})\right) \quad (5)$$

2.3.2.2. Apprentissage du modèle Naïve Bayes pour le modèle structuré

Comme précédemment, nous allons utiliser la méthode du *maximum de vraisemblance.* Notons $\theta_{w,l}$ l'estimation de la probabilité $P(w_{i,k}^d = w | s_i^d = l, \Theta_l^t)$.

L'estimation de cette probabilité est donnée par la formule suivante :

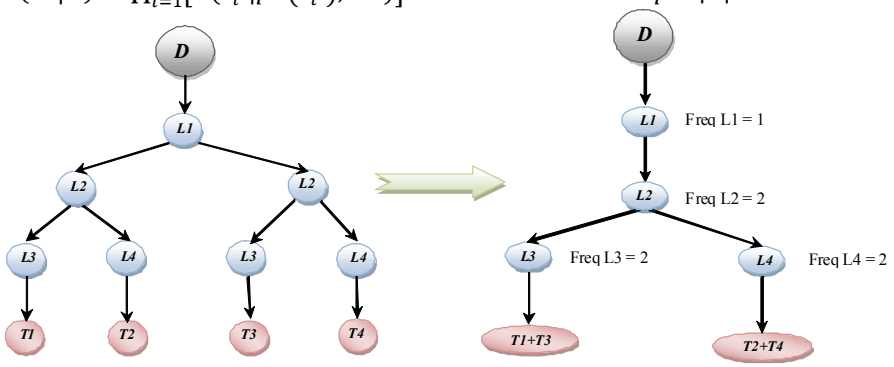
$$\theta_{w,l} = \frac{\sum_{d \in D_{TRAIN}} NW_{w,l}^d + 1}{\left(\sum_{d \in D_{TRAIN}} \sum_{w' \in V} NW_{w',l}^d\right) + |V|} . \quad (6)$$

$NW_{w,l}^d$ est le nombre de fois que le mot w apparaît dans le document d dans un nœud de label l. V l'ensemble de mots possibles (i.e. $w \in V$), appelé vocabulaire.

2.4. Jumelage

Dans le modèle arborescent de base présenté dans la Figure 1. a), les probabilités $P(l_i | pa(l_i), \Theta^s)$ sont estimées pour tous les nœuds de l'arbre. Si un document *d* contient *k* fois une relation « *nœud-Père(nœud)* », les probabilités de ces relations sont calculées séparément les unes des autres bien qu'elles ont la même probabilité, on calculera donc *k* fois la même probabilité. On a remédié à ce problème en proposant une nouvelle représentation du document donnée par la Figure 1. b), que nous avons appelé *modèle jumelé*. Chaque relation « *nœud-Père(nœud)* » est représentée une seule fois en lui associant un poids qui est la fréquence d'apparition de cette relation dans l'arbre initial. Avec cette nouvelle représentation du document structuré, si un document *d* contient *k* fois un nœud « *Nœud* » de père « *pére-Nœud* », la probabilité est calculée une seule fois ensuite elle est élevée à la puissance *k*. Ainsi, le calcul de la probabilité structurelle se fait selon la formule suivante :

$$P(s^d | \Theta) = \prod_{i=1}^{q} \left[P(s_i^d | pa(s_i^d), \Theta^s)\right]^{freq(s_i^d)} \quad \text{avec } q < |d| \quad (7)$$



a) Modèle de base b) Modèle jumelé

Figure 1. *Jumelage des nœuds de la structure arborescente.*

3. Passage au modèle discriminant par le noyau de Fisher

3.1. Le noyau de Fisher :

La méthode du noyau de Fisher a été développée au départ pour la classification de séquences biologiques par Jaakkola ([4], [5]). Son but est de transformer un modèle génératif en un modèle discriminant afin d'accroître ses performances pour la tâche de classification. L'idée de ce modèle consiste à créer à l'aide d'un modèle génératif une fonction noyau qui pourra ensuite être utilisée dans différentes machines discriminantes (CBS dans notre travail). Soit un modèle génératif$P(d/\Theta)$, Jaakkola propose, dans [4], de calculer le score de Fisher du document *d* comme suit :

$U_d = \nabla_\Theta \log P(d/\Theta)$.

Où l'opérateur ∇_Θ représente le gradient par rapport à Θ. U_d est alors un vecteur dont la dimension est égale au cardinal de Θ. En ce sens, U_d est une représentation vectorielle du document *d* par rapport à un modèle génératif de paramètres Θ.

3.2. Application du noyau de Fisher à notre modèle génératif :

En appliquant le score de Fisher sur le modèle *« parent »*, le vecteur de Fisher s'écrit:

$$\left(\frac{NS^d_{l_1,l_1}}{\theta^s_{l_1,l_1}}, \dots, \frac{NS^d_{l_{|\Lambda|},l_1}}{\theta^s_{l_{|\Lambda|},l_1}}, \frac{NS^d_{l_1,l_2}}{\theta^s_{l_1,l_2}}, \dots, \frac{NS^d_{l_{|\Lambda|},l_{|\Lambda|}}}{\theta^s_{l_{|\Lambda|},l_{|\Lambda|}}}, \frac{NW^d_{w_1,l_1}}{\theta^t_{w_1,l_1}}, \dots, \frac{NW^d_{w_{|V|},l_1}}{\theta^t_{w_{|V|},l_1}}, \frac{NW^d_{w_1,\lambda l_{|\Lambda|}}}{\theta^t_{w_1,l_{|\Lambda|}}}, \dots, \frac{NW^d_{w_{|V|},l_{|\Lambda|}}}{\theta^t_{w_{|V|},l_{|\Lambda|}}}\right)$$

Sous-vecteur correspondant au gradient sur le modèle de structure

Sous-vecteur correspondant au gradient pour les nœuds de label l_1

Sous-vecteur correspondant au gradient pour les nœuds de label $l_{|\Lambda|}$

Figure 2. *Vecteur de Fisher final obtenu pour un document d.*

Le vecteur obtenu est une composition d'un vecteur représentant l'information de structure (partie gauche) et d'un ensemble de vecteurs représentant les contenus pour les différentes étiquettes possibles des nœuds.

– $NW^d_{w,l}$ est le nombre de fois que le mot w apparaît dans le document d d'entraînement pour le nœud de label l.

– $NS^d_{n,m}$ est le nombre de fois qu'un nœud de label n possède un père de label m dans le document d.

– $\Theta^{s}_{n,m}$ représente la probabilité structurelle donnée par la formule (4)

– $\Theta^{t}_{w,l}$ représente la probabilité du contenu donnée par la formule (6)

3.3. Pondération des composantes du vecteur de Fisher

Comme les composantes du vecteur de Fisher (voir Figure 2.) sont les différentes combinaisons possibles *nœud-parent(nœud)* et *nœud-terme*, l'idée est de pondérer chaque relation comme suit :

Chaque composante structurelle correspondant à une relation *nœud-parent(nœud)* sera pondérée par le nombre d'apparition du nœud parent *« parent(nœud) »* dans la collection. Chaque composante de contenu correspondant à une relation *nœud-terme* sera pondérée par le nombre d'apparition du nœud *« nœud »* qui contient le terme dans la collection. Soit k_{l_i} le nombre d'apparition du nœud de label l_i dans la collection. Une telle pondération nous donne un vecteur de la forme suivante (Figure 3.) :

$$\left(\boldsymbol{k_{l_1}} \frac{NS^{d}_{l_1,l_1}}{\theta^{s}_{l_1,l_1}}, \dots, \boldsymbol{k_{l_1}} \frac{NS^{d}_{l_{|\Lambda|},l_1}}{\theta^{s}_{l_{|\Lambda|},l_1}}, \dots, \boldsymbol{k_{l_{|\Lambda|}}} \frac{NS^{d}_{l_1,l_{|\Lambda|}}}{\theta^{s}_{l_1,l_{|\Lambda|}}}, \dots, \boldsymbol{k_{l_{|\Lambda|}}} \frac{NS^{d}_{l_{|\Lambda|},l_{|\Lambda|}}}{\theta^{s}_{l_{|\Lambda|},l_{|\Lambda|}}}, \right.$$

$$\left. \boldsymbol{k_{l_1}} \frac{NW^{d}_{w_1,l_1}}{\theta^{t}_{w_1,l_1}}, \dots, \boldsymbol{k_{l_1}} \frac{NW^{d}_{w_{|V|},l_1}}{\theta^{t}_{w_{|V|},l_1}}, \dots, \boldsymbol{k_{l_{|\Lambda|}}} \frac{NW^{d}_{w_1,\lambda l_{|\Lambda|}}}{\theta^{t}_{w_1,l_{|\Lambda|}}}, \dots, \boldsymbol{k_{l_{|\Lambda|}}} \frac{NW^{d}_{w_{|V|},l_{|\Lambda|}}}{\theta^{t}_{w_{|V|},l_{|\Lambda|}}} \right)$$

Figure 3. *Vecteur de Fisher final obtenu pour un document d avec pondération.*

4. Tests et résultats

Dans ce qui suit, nous allons présenter une série d'expériences pour tester notre modèle. Ces expériences ont été réalisées sur différentes collections de documents XML prises de la campagne d'évaluation INEX. Afin de mettre en évidence l'apport de la pondération, nous avons implémenté, pour la classification, les méthodes SVM (Support Vecteur Machine) et CBS (Classifier-Based Search) [6] suite au passage au modèle discriminant par le noyau de Fisher. On a construit un échantillon extrait d'INEX, pour effectuer les différents tests. Le système calcule la F-mesure micro moyenne et macro-moyenne ainsi que le rappel macro et micro moyenne en prenant en compte le type de méthode (CBS et SVM), les résultats obtenus sont résumés dans le tableau 1.

On voit à partir du tableau 1 de résultats que la pondération des composantes des vecteurs de Fisher nous permet une augmentation de 4% de performances. De plus, les résultats montrent que la méthode CBS permet d'obtenir de meilleures performances que la méthode SVM.

	Avec Pondération		*Sans pondération*	
	CBS	***SVM***	***CBS***	***SVM***
F-Mesure micro moyenne	0,9181	0,903903	0,863344	0 ,87546
F-Mesure macro moyenne	0,9141	0,899777	0,877536	0,865585
Rappel micro moyenne	0,9189	0,905405	0,884918	0,870094
Rappel macro moyenne	0,9047	0,888888	0,86766	0,853332

Tableau 1. *Résultats obtenus pour les différents tests avec et sans pondération.*

5. Conclusion

Notre principal objectif est de comparer entre les deux modèles discriminants, obtenus à l'aide du noyau de Fisher à partir du modèle génératif, avec et sans pondération des composantes du score de Fisher. On a montré que la pondération nous permet d'obtenir de meilleures performances. Un autre objectif été d'exploiter l'apport du jumelage de nœuds car ce dernier réduit le temps de réponse (environ 30%) tout en préservant les performances en termes d'efficacité.

6. Bibliographie

[1] Amrouche K. and Ait Ali Yahia Y., Nodes coupling in a Bayesian Network for the automatic classification of XML documents, *International Conference on Machine and Web Intelligence*, *ICMWI'2010*, Algiers.

[2] Denoyer, L., Gallinari, P., Bayesian Network Model for semi-structured document classification, IP&M *Bayesian Network and Information Retrieval*, 1-25, 2004

[3] Denoyer, L., Gallinari, P., Un modèle de mixture de modèles génératifs pour les documents structurés multimédias, DN-8/2004, *fouille de textes*, pages 35 à 54.

[4] Jaakkola T., Haussler D., Exploiting generative models in discriminative classifiers, *Advanced in Neural information processing systems,* 11, 1998.

[5] Jaakkola T., Diekhans M. and Haussler D.,Using the Fisher kernel method to detect remote protein homologies, *Intelligent Systems for Molecular Biology Conference* (*ISMB'99*), Heidelberg, Germany, AAAI.

[6] Schühmacher J.P.C., Classifier-Based search in large document collections, Master's thesis 2011.