Champs aléatoires conditionnels : synthèse de [3, 6, 5, 4, 8, 7]

1 Introduction

Dans de très nombreux domaines, on cherche la meilleure séquence d'étiquettes au sens d'une séquence d'observations (bioinformatique, reconnaissance de la parole, de l'écrit, extraction d'information textuelle dans des documents). Des outils très répandus pour cette opération d'étiquetage et de segmentation sont les HMM (Hidden Markov Models ou Modèles de Markov Cachés), qui sont des automates probabilistes à états finis. Les HMM sont une forme de modèles génératifs qui définissent une probabilité jointe $P(X, Y)$, avec X : la séquence d'observation et Y : la séquence d'étiquette. Pour définir cette probabilité jointe sur les observations et les étiquettes, un modèle génératif doit énumerer toutes les séquences d'observations possibles, ce qui dans la pratique n'est pas envisageable à cause de l'explosion combinatoire (Rabiner [2] donne un exemple de cette explosion combinatoire en considérant un modèle à 5 états et 100 observations : il faudrait $2 * 100 * 5^{100} \simeq 10^{72}$ calculs pour énumerer tous les chemins possibles !). L'hypothèse d'indépendance des observations permet de contourner cette explosion combinatoire et d'obtenir des complexités plus raisonnables. Cette hypothèse d'indépendance des observations est la limitation majeure des HMM puisqu'elle n'est que très rarement vérifiée dans les problèmes réels. Nous devons donc trouver une représentation des données qui ne fait pas d'hypothèses sur l'indépendance des observations. C'est le cas des modèles conditionnels que nous présentons dans ce document.

2 Les modèles conditionnels

Les modèles conditionnels considèrent la probabilité conditionnelle $P(Y|X)$ plutôt que la probabilité jointe $P(X, Y)$. On donne donc les probabilités des séquences d'étiquettes possibles pour une séquence d'observation donnée, et non les probabilités des séquences d'étiquettes *et* des séquences d'observation. Contrairement aux modèles génératifs, on ne cherche donc pas à modéliser les observations. De plus, l'hypothèse d'indépendance des observations n'est plus faite, les probabilités de transitions entre étiquettes peuvent ainsi dépendre

des observations passées et futures, et pas seulement de l'observation courante, ce qui correspond davantage à la réalité des séquences réelles. Deux exemples de modèles conditionnels sont les MEMM (Maximum Entropy Markov Model) et les champs aléatoires conditionnels (CAC ou CRF : Conditionnal Random Field).

3 Les champs aléatoires conditionnels

Les champs aléatoires conditionnels (ou Conditional Random Field : CRF) se situent dans un cadre probabiliste et sont basés sur une approche conditionnelle pour étiqueter et segmenter les séquences de données. Le principal avantage des CRF sur les HMM est que leur nature conditionnelle permet de relaxer les hypothèses faites sur l'indépendance des observations. Les CRF évitent le problème du "label bias" rencontré avec les MEMM (Une bonne illustration du problème du label bias est présenté dans [3]). Plusieurs expériences ont montrés la supériorité des CRF sur les HMM et sur les MEMM sur des problèmes réels [3, 5, 4].

Définition

Soit X : variable aléatoire à étiqueter (observations), et Y : variable aléatoire représentant la séquence d'étiquettes, les Y_i appartiennentt à un alphabet fini.

Les distributions des variables aléatoires X et Y sont liées, mais dans une approche discriminante, nous construisons un modèle conditionnel $p(Y|X)$ qui ne nécessite pas la modélisation de $p(X)$.

Soit $G = (V, E)$ un graphe tel que $Y = \{Y_v\}_{v \in V}$. V définit l'ensemble des noeuds, E l'ensemble des arcs. Y est donc indéxé par les noeuds de G.

(X, Y) est un champs conditionnel aléatoire si, conditionné à X, les variables aléatoires Y_v vérifient la propriété de Markov vis à vis du graphe :

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$$

avec $w \sim v$ signifie que w et v sont voisins dans G. Cette propriété est donc satisfaite si l'état du système (le noeud dans lequel on se trouve) ne dépend que des états (noeuds) voisins, ainsi que des probabilités de transitions entre les états.

Si le graphe est un arbre (la chaine en fait donc partie), la distribution jointe des séquences d'étiquettes Y sachant X $p_\theta(y|x)$ est de la forme :

$$\exp\left(\sum_{e\in E,k}\lambda_k f_k(e,y|_e,x)+\sum_{v\in V,k}\mu_k g_k(v,y|_v,x)\right)$$

avec $y|_e$ et $y|_v$: ensembles des étiquettes de Y associés aux noeuds de e et v ; f_k et g_k sont des caractéristiques (fixées) ; λ_k et mu_k sont des paramètres à estimer lors de l'apprentissage.

Wallach [8] utilise une autre notation un peu plus "parlante" pour $p_\theta(y|x)$:

$$\exp\left(\sum_{j}\lambda_j t_j(y_{i-1},y_i,x,i)+\sum_{k}\mu_k s_k(y_i,x,i)\right)$$

Cette notation montre bien les deux types de caractéristiques :

- Les t_j (ou f_k) sont des caractéristiques sur les *transitions* du graphe, définies pour chaque paire d'état (y,y') et fonctions de la séquence d'observation complète et des étiquettes en positions i et $i-1$. Il s'agit de caractéristiques sur les cliques du graphe.
- Les s_k (ou g_k) sont des caractéristiques locales sur les *états* (noeuds du graphe), fonctions de l'étiquette en position i et de la séquence d'observations. Lafferty [3] donne un exemple de caractéristique g_k : "une caractéristique booléenne g_k peut être vraie si le mot X_i est majuscule et le tag Y_i est *nom propre*".

Szummer [7] distingue les deux types de caractéristiques : les "site feature" et les "interaction feature", que l'on peut traduire par caractéristique locale et caractéristique globale.

Une fois les paramètres appris, l'analyse de séquence se fait par l'algorithme de Viterbi dans le cas d'un graphe ayant une structure d'arbre (1D).

4 apprentissage

La phase d'apprentissage consiste à estimer les valeurs des paramètres $(\lambda_1,\lambda_2,\ldots;\mu_1,\mu_2,\ldots)$ à partir d'une base d'apprentissage $D=\{(x^{(i)},y_{(i)})\}_{i=1}^N$ avec les distributions empiriques $\tilde{p}(x,y)$.

Le but est de maximiser la log-vraisemblance $O(\theta)$:

$$O(\theta)=\sum_{i=1}^{N}\log p_\theta(y^{(i)}|x^{(i)})$$

$$O(\theta)\simeq\sum_{x,y}\tilde{p}(x,y)\log p_\theta(y|x)$$

La procédure de maximisation est basée sur l'algorithme itératif IIS (Improved Iterative Scaling) de Della Pietra [1]. Les mises à jour des paramètres $\delta\lambda_k$ (tel que $\lambda_k \leftarrow \lambda_k + \delta\lambda_k$) sont les solutions de l'équation :

$$\tilde{E}[f_k] = \sum_{x,y} \tilde{p}(x,y) \sum_{i=1}^{n+1} f_k(e_i, y|_{e_i}, x)$$

$$= \sum_{x,y} \tilde{p}(x) p(y|x) \sum_{i=1}^{n+1} f_k(e_i, y|_{e_i}, x) e^{\delta\lambda_k T(x,y)}$$

avec

$$T(x,y) = \sum_{i,k} f_k(e_i, y|_{e_i}, x) + \sum_{i,k} g_k(v_i, y|_{v_i}, x)$$

La mise à jour des paramètres μ_k se fait de la même manière.

Le calcul de la somme exponentielle du terme de droite pose cependant problème car $T(x,y)$ est une fonction de (x,y), et la programmation dynamique somme sur les séquences avec un T variable. Pour résoudre ce problème, Lafferty propose deux algorithmes : l'algorithme S qui utilise une caractéristique "slack" ; le second, algorithme T, conserve le chemin des totaux T partiels. Dans les deux cas, nous considèrons la matrice des probabilités conditionnelles ainsi que les variables *forward* et *backward* $\alpha_i(x)$ et $\beta_i(x)$, définies de la manière suivante :

matrice des probabilités conditionnelles :

Pour une structure de chaine, la probabilité conditionnelle d'une séquence d'étiquette peut s'exprimer sous la forme d'une matrice qui sera utile pour calculer l'inférence du modèle. Pour chaque position i dans la séquence d'observation, on définit la matrice $M_i(x) = [M_i(y', y|x)]$ par :

$$M_i(y', y|x = \exp(\Lambda_i(y', y|x))$$

$$\Lambda_i(y', y|x) = \sum_k \lambda_k f_k(e_i, Y|e_i = (y', y), x) +$$

$$\sum_k \mu_k g_k(v_i, Y|v_i = y, x)$$

Contrairement aux modèles génératifs les modèles conditionnels ne nécessitent pas d'énumerer toutes les observations possibles x, ces matrices peuvent donc être calculées pour une séquence d'observation (test ou apprentissage) et un vecteur de paramètre θ. Nous introduisons un facteur de normalisation $Z_\theta(x)$:

$Z_\theta = (M_1(x)M_2(x)\dots M_{n+1}(x))_{start,stop}$

Avec cette notation, la probabilité conditionnelle d'une s'équence d'étiquette y s'écrit :

$$p_\theta(y|x) = \frac{\prod_{i=1}^{N+1} M_i(y_{i-1}, y_i|x)}{\left(\prod_{i=1}^{N+1} M_i(x)\right)_{start,stop}}$$

avec $y_0 = start$ et $y_{n+1} = stop$

forward :

– initialisation : $\alpha_0(y|x) = 1siy = start; 0sinon$

– récurrence : $\alpha_i x = \alpha_{i-1}(x) M_i(x)$

backward :

– initialisation : $\beta_{n+1}(y|x) = 1siy = stop; 0sinon$

– récurrence : $\beta_i x^T = M_{i+1}(x)\beta_{i+1}(x)$

4.1 algorithme S

Nous définissons une caractéristique "slack" (en fr. : desseré, distendu) $s(x, y)$:

$$s(x, y) = S - \sum_i \sum_k f_k(e_i, y|_{e_i}, x) + \sum_i \sum_k g_k(v_i, y|_{v_i}, x)$$

avec S : constante choisie pour que $s(x^{(i)}, y) > 0$ pour tout y et tous les vecteurs d'observations $x^{(i)}$ de la base d'apprentissage.

Les mises à jour des paramêtres sont :

4.2 Algorithme T

Les mises à jour des poids sont $\delta\lambda_k = \log \beta_k$ et $\delta\mu_k = log\gamma_k$; où β_k et γ_k sont les seules racines positives des équations polynomiales suivantes :

$$\sum_{i=0}^{T_{max}} a_{k,t}\beta_k^t = \tilde{E}f_k$$

,

$$\sum_{i=0}^{T_{max}} b_{k,t}\gamma_k^t = \tilde{E}g_k$$

qui se résolvent grâce à la méthode de Newton.

5 Etude de quelques applications

Szummer 2004 [7] utilise les CRF sur des documents manuscrits présentés en figure 1. Il s'agit d'étiqueter les traits manuscrits comme "appartenant à un rectangle" ou comme "connecteur". Sorti du contexte, il est impossible de

dire à quelle classe appartient un trait. Les CRF permettent donc de prendre une décision globale pour l'ensemble des traits de la figure, en prenant en compte le contexte.

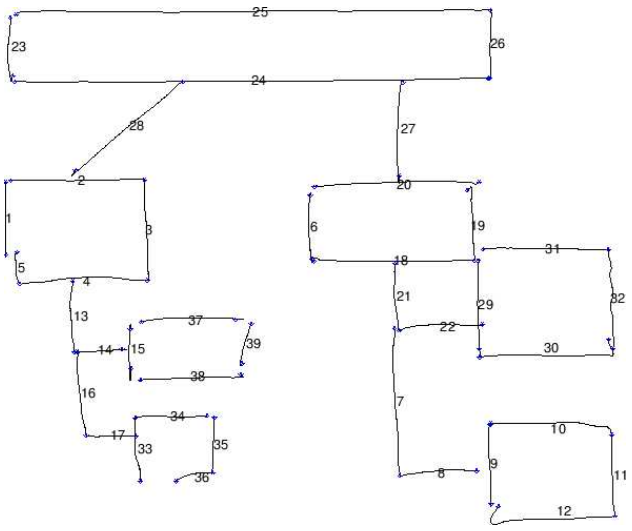

FIG. 1 –

Une fois les traits segmentés (certains sont liés) en "fragments", les auteurs construisent un CRF où chaque fragment est représenté par un noeud du graphe. Les noeuds ont une variable étiquette associée : $+/-1$.

Les caractéristiques (ou "fonctions potentielles") évaluent la compatibilité des étiquettes en fonction du segment courant et des étiquettes des noeuds voisins. Dans ce système, les caractéristiques sont binaires et réelles.

- Caractéristiques locales : longueur et orientation du fragment ; histogramme des distances et anges relatifs avec les fragments voisins.
- Caractéristiques globales : caractéristiques sur des paires de fragments (distances et angles relatifs) ; recherche de formes simples (coins, jonctions), test si présence de deux coins autour d'un fragment ; mesures d'alignements pour voir si deux fragments peuvent être parallèles et alignés dans une même forme.

Références

[1] PIETRA, S. Della, V. Della PIETRA and J. LAFFERTY, "Inducing features of random fields", *IEEE Trans. on PAMI*, vol. 19, 1990, pp. 380–393.

[2] Rabiner, L. R. "A tutorial on hidden markov models and selected aplli-cations in speech recognition". In *Readings in Speech Recognition*. Kaufmann, 1990, pp. 267–296.

[3] Lafferty, J., A. McCallum and F. Pereira. "Conditional random fields : Probabilistic models for segmenting and labeling sequence data". In *Proc. 18th International Conf. on Machine Learning* (2001), Morgan Kaufmann, San Francisco, CA, pp. 282–289.

[4] Pinto, D., A. McCallum, X. Wei and W. B. Croft. "Table extraction using conditional random fields". In *SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (2003), ACM Press, pp. 235–242.

[5] Sha, F. and F. Pereira. "Shallow parsing with conditional random fields".

[6] Gregory, M. L. and Y. Altun. "Using conditional random fields to predict pitch accents in conversational speech.". In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)* (2004).

[7] Szummer, M. and Y. Qi, "Contextual recognition of hand-drawn diagrams with conditional random fields", *IWFHR'9*, 2004, pp. 32–37.

[8] Wallach, H., "Conditionnal random fields : an introduction", *Technical Report*, 2004.