Analyse de séquence : biblio

1 Introduction

Dans de nombreux domaines, on cherche la meilleure séquence d'étiquettes au sens d'une séquence d'observations : bioinformatique (decriptage des séquences d'ADN), reconnaissance de la parole, de l'écrit, extraction d'information textuelle dans des documents. Nous présentons les différents outils rencontrés dans la littérature permettant l'étiquetage de séquences, ainsi que leurs applications sur des signaux réels, et en particulier l'écriture manuscrite.

2 Notations

Dans ce document, nous appelerons x la séquence d'observation, et y les étiquettes.

3 Méthodes à fenêtres glissantes

Les méthodes à fenêtre glissantes utilisent un classifieur "classique" qui se déplace sur la séquence et prend en entrée une fenêtre d'observations centrée pour classer l'élément courant. Pour une fenêtre de largeur $w = 2d + 1$ (d éléments précédents, 1 élément courant et d éléments suivants), il s'agit de déterminer $y_{i,t}$ avec la fenêtre $< x_{i,t-d}, \ldots, x_t, \ldots, x_{i,t+d} >$. Les méthodes à fenêtre permettent ainsi de prendre en compte le contexte au niveau des observations.

L'apprentissage de ces classifieurs est effectué grâce aux algorithmes classiques en convertissant les éléments à classifier en fenêtres à l'aide de leurs voisins.

Si ces méthodes prennent en compte le contexte au niveau des observations, elles ne permettent pas de prendre en compte les corrélations entre les étiquettes. D'où l'introduction des méthodes à fenêtre glissantes récurrentes.

Les méthodes à fenêtre glissantes récurrentes sont basées sur le même principe que les Méthodes à fenêtre glissantes simples, mais les sorties précédentes $y_{i,t-d} \ldots y_{i,t-1}$ sont utilisées par le classifieur en plus de la fenêtre $< x_{i,t-d}, \ldots, x_t, \ldots, x_{i,t+d} >$ pour déterminer $y_{i,t}$. La récurrence permet de prendre en compte le contexte au niveau des étiquettes.

Ces méthodes à fenêtres glissantes récurrentes ont le plus souvent été mise en oeuvre en utilisant des réseaux de neurones, soit en connectant les sorties du réseaux à la couche cachée (recurrent neural network) , soit en bouclant les sorties de la couche cachée sur les entrées de la couche cachée en appliquant un délai (Time delay neural network).

Ce type de réseau a été utilisé dans de nombreux domaines tels que la reconnaissance de codes postaux manuscrits [LeCun 89] ou la reconnaissance de la parole [Pérez-Ortiz 01].

4 Modèles de Markov cachés

Les méthodes markoviennes et en particulier les modèles de Markov cachés sont certainement les approches les plus répandues pour la modélisation de séquences. Le cadre statistique dans lequel les modèles de Markov se placent les rend très robustes aux variabilités des signaux réels. De plus, ce sont des modèles dynamiques capables de segmenter les séquences en faisant intervenir la reconnaissance.

4.1 Formalisation des HMM

Un modèle de Markov caché fournit un cadre probabiliste à la modélisation des séquences. La séquence est modélisée par une succession d'états représentant des entitées constituantes du signal (phonèmes, acide aminé, mots, lettres, graphèmes, etc), et de transitions probabilisées entre ces états. Les états constituent la couche cachée du modèle qui produit une séquence d'observations. On cherche alors à trouver la séquence d'états cachés la plus probable (y) compte tenu de la séquence d'observation (x).

La modélisation probabiliste d'un HMM requiert la définition des éléments suivants : le nombre d'états du modèle, les probabilités de transition entre les états $P(y_t|y_{t-1})$, les probabilités d'émission des symboles $P(x_t|y_t)$, et les distributions des états initiaux.

Après avoir fixé le nombre d'états d'un modèle, une phase d'apprentissage est nécessaire afin de déterminer la matrice des probabilités de transition entre états et les distributions des états initiaux. Cet apprentissage est généralement effectué grâce à l'algorithme de Baum-Welch.

La meilleure séquence d'étiquettes au sens d'une séquence d'observations et d'un modèle est effectué grâce à l'algorithme de Viterbi [Forney 73]. Cet algorithme est issu de la programmation dynamique, repose sur le principe d'optimalité suivant : le meilleur chemin pour aller de $t = 0$ à $t = N$ est composé du meilleur chemin pour aller de $t = 0$ à $t = N - 1$ et du meilleur chemin pour aller de $t = N - 1$ à $t = N$. L'algorithme de Viterbi consiste ainsi à calculer pour toutes les étiquettes et pour tous les instants t la probabilité du meilleur chemin amenant à l'état courant, compte tenu des premières observations.

L'algorithme de Viterbi, s'il permet de déterminer de manière efficace la meilleure séquence d'états pour une séquence d'observations donnée, fait également l'hypothèse d'indépendance des observations, cette hypothèse n'étant pas toujours vérifiée dans les problèmes réels.

4.2 Modèles continus/discrets, et vraisemblance des observations

Les modèles de Markov cachés peuvent être discrets si les observations appartiennent à un alphabet fini de symboles, ou continus si les observations sont continues.

Dans les HMM "classiques", les vraisemblances des observations continues $P(x_t|y_t)$ sont modélisées par des mélanges de gaussiennes [Vinciarelli 04] dont les paramètres sont estimés lors de l'apprentissage du modèle.

Un autre type d'approche, qualifiées d' "hybrides" ou de "neuro markoviennes", consiste à remplacer les mélanges de gaussiennes par des classifieurs de type réseau de neurones. Dans ce cas, les probabilités *a posteriori* fournies en sortie du réseau $P(y_t|x_t)$ sont transformées par la règle de Bayes en vraisemblances normalisées $P(x_t|y_t)/P(x_t)$. Une procédure d'apprentissage itérative du système hybride est généralement mise en

oeuvre, où les sorties désirées du réseau de neurones sont fournies par le HMM. La rétropropagation du gradient est alors appliquée pour la mise à jour des poids du réseau. Ces méthodes ont rencontré un franc succès [Morgan 93, Bengio 95, Gilloux 95, Knerr 98] car elles permettent de bénéficier du pouvoir discriminant des réseaux de neurones allié à la capacité de modélisation des séquences des modèles de Markov cachés.

4.3 Applications des HMM à des signaux réels

Les HMM ont été très largement appliqués à la reconnaissance de signaux réels.

Dans [Rabiner 90, Bahl 83], les auteurs appliquent les HMM à la reconnaissane de la parole. La figure 1 montre un exemple de modèle de Markov pour la reconnaissance de mots.

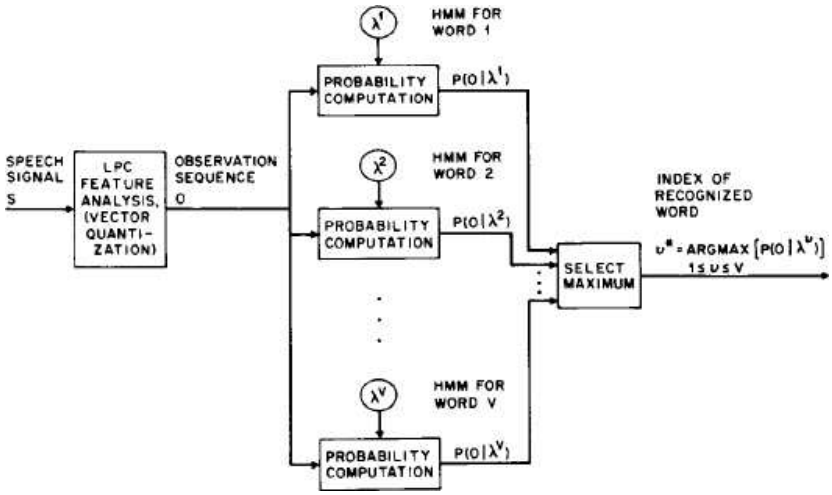

FIG. 1 –

En reconnaissance de l'écriture manuscrite, les HMM ont également été appliqué pour obtenir des modéles statistiques de lettres [El-Yacoubi 99], chiffres [J. Cai 99], mots [Chen 94, Mohammed 96], ou de lignes de texte plus ou moins contraintes : ligne provenant de blocs adresses [Kim 98, El-Yacoubi 02], de dates de chèques [Morita 02, **?**], et textes libres [Vinciarelli 04, Marti 00]. Dans [Marti 00], les modèles de mots sont obtenus par concaténation des modèles de lettres, et les modèles de phrase sont eux mêmes une concaténation des modèles de mots (voir figure 2).

Dans [El-Yacoubi 99, El-Yacoubi 02], les modèles de mots et de phrases sont également obtenus par concaténation des modèles de lettres. Les auteurs rajoutent à gauche et à droite du nom de rue recherché un modèle générique permettant d'absorber les informations non pertinentes telles que le numéro de rue, la nature de la voie, etc. (voir figure 3).

Dans [Morita 02], les auteurs utilisent les HMM pour segmenter les dates provenant de chèques en entités plus petites : jour, mois, année, nom de ville. Un module de reconnaissance spécifique pour chaque type de champs est ensuite appliqué : MLP pour les séquences numériques (jour, année), HMM pour les mots (mois). Des modèles élémentaires pour les noms de ville, les espaces inter-champs, les lettres et les chiffres

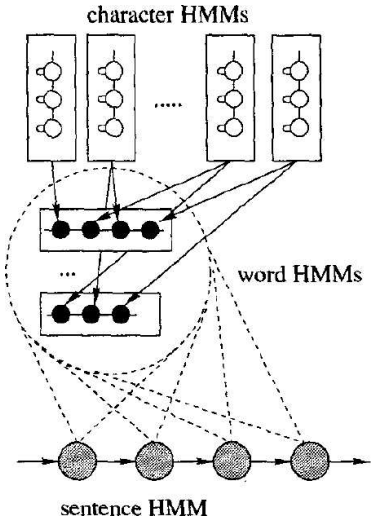

FIG. 2 – Modèles de lettre, mot et phrase [Marti 00]

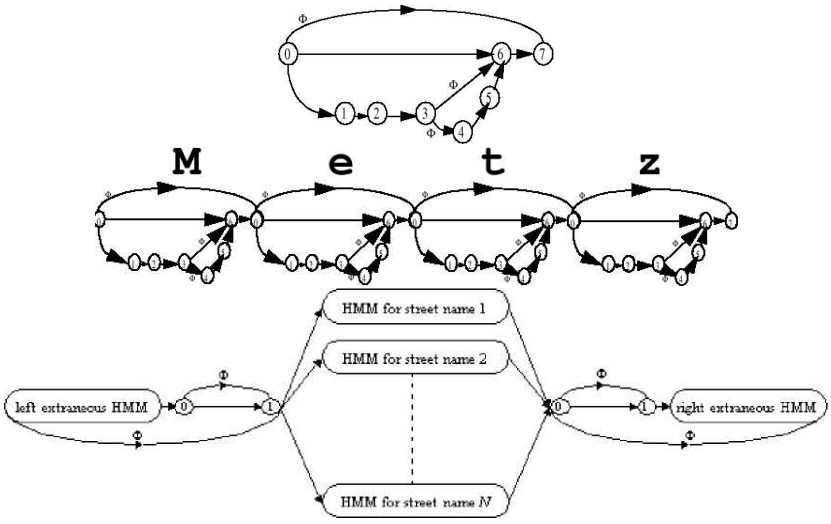

FIG. 3 – Modèlisation d'un caractère, d'un mot et d'une ligne de texte d'adresse postale par concaténation successive [El-Yacoubi 99, El-Yacoubi 02]. Dans le modèle de ligne, deux modèles génériques permettent d'absorber l'information non pertinente (numéro de rue, la nature de la voie, etc.)

sont appris, et les modèles de mot et de date sont obtenus par concaténation des modèles élémentaires. Comme cette étape ne cherche pas à reconnaitre la date mais seulement à la segmenter, la taille du lexique a été diminué en utilisant un seul modèle pour l'ensemble des noms de ville, et le concept de "méta-classe" de chiffre. Nous illustrons le concept de méta-classe de chiffre par un exemple : le premier chiffre de l'année étant obligatoirement un "1" ou un "2", une seule classe est constituée pour ces deux

chiffres. La règle de décision utilisée est basée sur la maximisation de la probabilité *a posteriori* $P(y|x)$.

METTRE LES IMAGES DES MODELES

5 Champs conditionnels aléatoires

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. En effet, les modèles conditionnels considèrent la probabilité conditionnelle $p(x|y)$ 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.

Plusieurs expériences ont montrés la supériorité des CRF sur les HMM et sur les MEMM sur des problèmes réels, en particulier la recherche d'information [Lafferty 01, Sha 03, Pinto 03]. En revanche, les CRF n'ont pas encore à notre connaissance été appliqué à la reconnaissance de l'écriture manuscrite.

5.1 Application

Szummer 2004 [Szummer 04] utilise les CRF sur des documents manuscrits présentés en figure 4. Il s'agit d'étiqueter les traits manuscrits comme "appartenant à un rectangle" ou comme "connecteur". Sans le contexte, il est impossible de dire à quelle classe appartient un trait. Les CRF permettent de prendre une décision globale pour l'ensemble des traits de la figure.

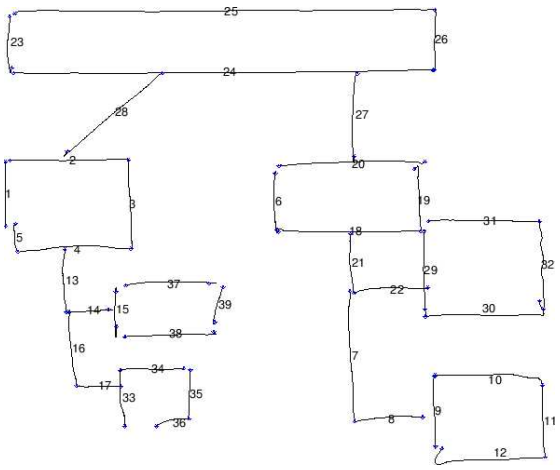

FIG. 4 –

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

[Bahl 83] L. R. Bahl, F. Jelink & R. Mercer. *A maximum likelihood approach to continuous speech recognition*. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 5, pages 179–190, 1983.

[Bengio 95] Y. Bengio, Y. Le Cun, C. Nohl & C. Burges. *LeRec : A NN-HMM hybrid for on-line handwriting recognition*. Neural Computation, vol. 7, no. 6, pages 1289–1303, 1995.

[Chen 94] M.Y. Chen, A. Kundu & J. Zhou. *Off-Line Handwritten Word Recognition Using a Hidden Markov Model Type Stochastic Network*. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 5, pages 481–496, 1994.

[El-Yacoubi 99] A. El-Yacoubi, M. Gilloux, R. Sabourin & C. Y. Suen. *An HMM Based Approach for Off-line Unconstrained Handwritten Word Modeling and Recognition*. IEEE Trans. on PAMI, vol. 21, no. 8, pages 752–760, 1999.

[El-Yacoubi 02] A. El-Yacoubi, M. Gilloux & J.-M. Bertille. *A Statistical Approach for Phrase Location and Recognition within a Text Line : An Application to Street Name Recognition*. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pages 172–188, 2002.

[Forney 73] G.D. Forney. *The Viterbi Algorithm*. Proc. IEEE, vol. 61, pages 268–278, 1973.

[Gilloux 95] M. Gilloux, B. Lemari & M. Leroux. *A hybrid radial basis function network/hidden markov model handwritten word recognition system*. ICDAR'95, pages 394–397, 1995.

[J. Cai 99] Z.Q. Liu J. Cai. *Integration of structural and statistical Information for Unconstrained Handwritten Numeral Recognition*. IEEE Trans. on PAMI, vol. 21, no. 3, pages 263–270, 1999.

[Kim 98] G. Kim & V. Govindaraju. *Handwritten Phrase Recognition as Applied to Street Name Images*. Pattern Recognition, vol. 31, no. 1, pages 41–51, 1998.

[Knerr 98] S. Knerr & E. Augustin. *A neural network-hidden markov model hybrid for cursive word recognition*. ICPR, vol. 2, pages 1518–1520, 1998.

[Lafferty 01] J. Lafferty, A. McCallum & F. Pereira. *Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data*. In Proc. 18th International Conf. on Machine Learning, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[LeCun 89] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard & L. D. Jackel. *Backpropagation Applied to Handwritten Zip Code Recognition*. Neural Computation, vol. 1, no. 4, pages 541–551, 1989.

[Marti 00] U.V. Marti & H. Bunke. *Handwritten sentence recognition*. ICPR, vol. 3, pages 3467–3470, 2000.

[Mohammed 96] M. Mohammed & P. Gader. *Handwritten Word Recognition Using Segmentation-Free Hidden Markov Modeling and Segmentation-Based Dynamic Programming Techniques*. IEEE Trans. Pattern Analysis Machine. Intelligence, vol. 18, no. 5, pages 548–554, 1996.

[Morgan 93] N. Morgan, H. Bourlard, S. Renls, M. Cohen & H. Franco. *Hybrid neural network/hidden Markov model systems for continuous speech recognition*. IJPRAI, vol. 7, no. 4, 1993.

[Morita 02] M. Morita, R. Sabourin, F. Bortolozzi & C.Y. Suen. *Segmentation and Recognition of Handwritten Dates*. IWFHR, pages 105–110, 2002.

[Pinto 03] D. Pinto, A. McCallum, X. Wei & 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, pages 235–242. ACM Press, 2003.

[Pérez-Ortiz 01] Juan Antonio Pérez-Ortiz & Mikel L. Forcada. *Part-of-speech tagging with recurrent neural networks*. In Proceedings of the International Joint Conference on Neural Networks, IJCNN 2001, pages 1588–1592, 2001.

[Rabiner 90] L. R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Apllications in Speech Recognition*. In Readings in Speech Recognition, pages 267–296. Kaufmann, 1990.

[Sha 03] F. Sha & F. Pereira. *Shallow Parsing with Conditional Random Fields*. 2003.

[Szummer 04] M. Szummer & Y. Qi. *Contextual Recognition of Hand-drawn Diagrams with Conditional Random Fields*. IWFHR'9, pages 32–37, 2004.

[Vinciarelli 04] A. Vinciarelli, S. Bengio & H. Bunke. *Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models*. IEEE Trans. on PAMI, vol. 26, no. 6, pages 709–720, 2004.