Apprentissage Dynamique du Nombre d'états d'un Modèle de Markov Caché à Observations Continues : Application au Tri de Formulaires

S. RAMDANE[1,2], B. TACONET[1], A. ZAHOUR[1]

[1]G.E.D., [2]G.R.E.A.H., Université du Havre, Place Robert Schuman, 76610 Le Havre, France

ramdan@iut.univ-lehavre.fr, taconet@iut.univ-lehavre.fr, zahour@iut.univ-lehavre.fr

Résumé – Dans le cadre de la reconnaissance automatique de types de formulaires avec champs manuscrits et sans aucun signe de référence, basée sur une description de la structure physique du formulaire, nous sommes amenés à représenter un formulaire par un modèle de Markov caché pseudo-2D (PHMM). Ce modèle est constitué d'un graphe de super-états. A chaque super-état on associe un modèle de Markov caché secondaire (HMM) dont les observations sont continues. Nous exposons pourquoi la méthode classique des k-moyennes est mal adaptée à notre problème, puis nous détaillons une nouvelle méthode générale qui prend mieux en compte la réalité physique des états, en les situant dans l'espace de représentation des caractéristiques, et en les construisant dynamiquement par agrégation progressive des séquences d'observations. Ce n'est qu'à la fin du processus d'agrégation que le nombre d'états du modèle stochastique initial est connu.

Abstract – Our purpose is to recognise types of forms with handwritten fields. The form is described by a pseudo-2D hidden Markov model (PHMM). The observations are features extracted from the rectangular blocks of its physical structure. This model consists of a graph of super-states. To each super-state, one associates a secondary 1D hidden Markov model (HMM) whose observations are continuous. We expose why the traditional method of the K-means is badly adapted to our problem, then we detail a general method which takes better into account the physical reality of the states, by locating them in the space of representation of the characteristics, and by dynamically building them by progressive aggregation of the sequences of observations. It is only at the end of the process of aggregation that the number of states of the initial stochastic model is known.

1. Le modèle de la structure physique d'un formulaire

Nos travaux visent à trouver une méthode générale et fiable fondée scientifiquement, qui permet de trier automatiquement les formulaires avec champs manuscrits et sans aucun signe de référence. La structure physique est décrite par un modèle de Markov caché pseudo-2D (PHMM), qui incorpore les variations de taille des champs manuscrits et qui prend en compte les phénomènes de fusionnement et de fragmentation des blocs [1,2].

1.1 Architecture générale du modèle PHMM

Un algorithme de rectangulation [3,4] permet d'extraire les blocs rectangulaires englobant les zones d'inscription, et en fournit la liste. L'ensemble de ces rectangles constitue la totalité de l'information retenue pour faire l'identification. Nous nous sommes limités aux paramètres les plus simples de chaque rectangle : hauteur, largeur, coordonnées de son centre.

Puisque le formulaire traité est composé de pavés noirs sur fond blanc, nous observons fréquemment des ensembles de lignes successives identiques ; une super-ligne décrit un tel ensemble. De façon à comprimer la représentation, un formulaire sera décrit par un tableau de super-lignes, composées de super-segments noirs. Nous avons opté pour une architecture planaire à modèle principal vertical ; l'image d'un formulaire doit donc être découpée en bandes horizontales homogènes (dont les lignes sont semblables). Chaque bande horizontale est modélisée par un modèle de Markov secondaire (HMM-1D) de type gauche droite (figure 1). Dans la direction verticale, un modèle de durée explicite a été retenu, ce qui permet de mieux prendre en compte la hauteur des différents super-états [5]. La durée dans un super-état est ainsi assimilée à la hauteur de la bande (c'est-à-dire le nombre de lignes). Dans la direction horizontale, les observations sont relatives aux super-segments noirs et caractérisées par deux composantes : la position et la longueur du super-segment noir. Les phénomènes de fragmentation horizontale, et, paradoxalement, les phénomènes de fragmentation verticale d'un rectangle majeur sont traités par les transitions entre états des HMMs secondaires.

1.2 Les modèles secondaires

Les modèles secondaires du PHMM proposé sont des modèles markoviens continus d'ordre 1, de type gauche-droite. Dans un tel problème, il est recommandé de faire l'apprentissage des HMMs par l'algorithme des k-moyennes (de préférence à l'algorithme de Baum-Welch), pour répartir le mieux possible les observations continues dans les états. Les phénomènes de fragmentation horizontale sont naturellement pris en compte par les transitions entre états des HMMs secondaires. Les phénomènes de fragmentation

verticale, se produisant à l'intérieur d'un même super-état, sont également (et paradoxalement) absorbés par les transitions entre états du modèle secondaire (figure 2).

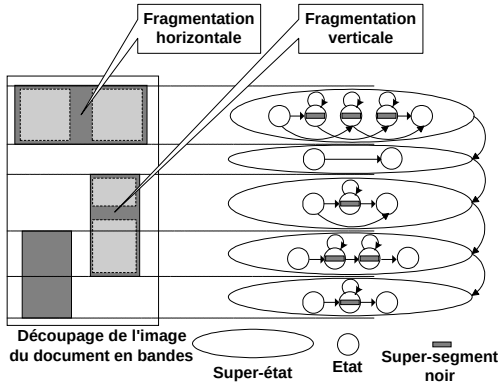

FIG. 1 : architecture générale du modèle markovien PHMM

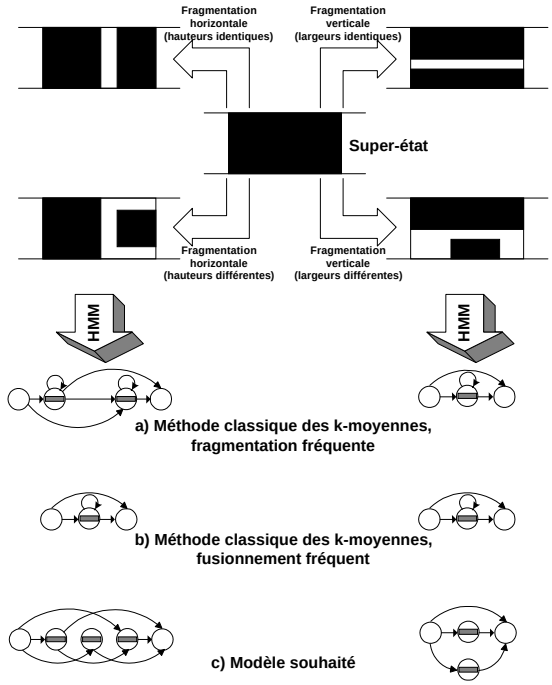

FIG. 2 : fragmentation horizontale ou verticale dans un super-état et HMM associé

Les figures 2-a et 2-b montrent les modèle de Markov obtenus par la méthode classique des k-moyennes : la longueur moyenne de la séquence varie selon que la fragmentation ou le fusionnement est le plus fréquent. Or le nombre d'états se confond avec l'arrondi de la valeur moyenne. On note que la présence de boucles sur un même état est nécessaire pour absorber les fragments.

La figure 2-c reflète mieux la réalité physique : par exemple, dans le cas de la fragmentation horizontale, on distingue trois états, correspondant chacun à une observation précise (un super-segment correspond au bloc de fusionnement, et deux super-segments correspondent aux deux blocs de fragmentation). On remarque qu'il n'y a pas de bouclage d'un état sur lui-même. Ce modèle, en étant au plus près de la réalité physique, évite le phénomène de bouclage, reconnu comme un problème majeur dans la modélisation par HMM (il biaise le calcule des probabilités en surestimant les séquences les plus courtes et en sous-estimant les séquences les plus longues).

Dans la suite, nous exposons une nouvelle méthode d'apprentissage non supervisé des HMMs, destinée à modéliser le mieux possible les bandes de super-états. Cette méthode dépasse largement le cadre du problème posé ; il s'agit d'une méthode générale qui évite la création de boucles dans le modèle de Markov caché. Pour cela, et contrairement à la méthode classique des k-moyennes, le nombre d'états ne doit pas être fixé avant l'agrégation des observations, mais ce nombre doit évoluer avec l'agrégation des observations.

2. Méthode d'apprentissage non supervisé des HMMs

2.1 Limitation de la méthode classique des k-moyennes

L'inconvénient principal de la méthode des k-moyennes est qu'elle ne permet pas d'ajuster le nombre d'états [6,7,8]. Le nombre de classes de regroupement des observations est ainsi fixé, dès le départ, par le nombre d'états choisi initialement. Si ce nombre est trop petit, une longue séquence d'observations sera telle que plusieurs observations seront affectées au même état. S'il s'agit d'un modèle gauche-droite, cela se traduit par un bouclage sur le même état ; s'il s'agit d'un modèle ergodique général, cela se traduit par un bouclage sur le même état ou un bouclage impliquant plusieurs états. Ce type de modèle comportant des boucles biaise la reconnaissance en privilégiant les séquences d'observations les plus courtes. Si le nombre d'états est trop grand, non seulement les temps de calcul sont très élevés, mais la dernière phase de l'algorithme peut ne pas converger en pratique.

2.2 La méthode d'agrégation dynamique (en une seule passe)

Puisque le choix a priori du nombre d'états est en pratique reconnu comme un problème difficile [9,10], nous proposons de modifier la méthode classique des k-moyennes, en tenant compte de deux objectifs : le nombre d'états doit rendre compte de la réalité physique, et le modèle stochastique initial doit être presque optimal. En réalité, le traitement que nous proposons à présent permet à la fois de répartir les observations, tout en faisant évoluer le nombre des états, d'une façon progressive, par intégration des séquences dans l'ordre des longueurs décroissantes et en une seule passe. Les figures 3 à 6 illustrent l'agrégation dynamique de deux séquences.

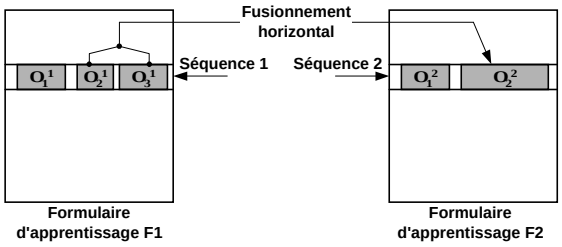

FIG. 3 : les deux premières séquences

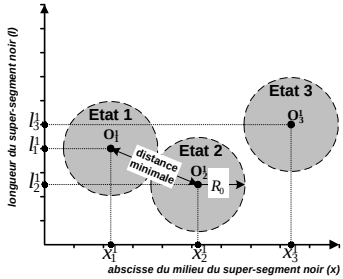

FIG. 4 : représentation des observations et des états de la première séquence

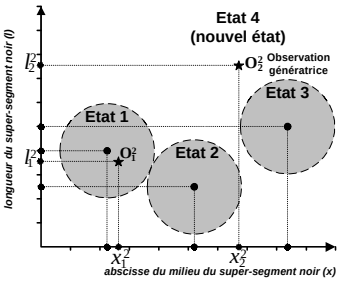

FIG. 5 : localisation des observations de la deuxième séquence

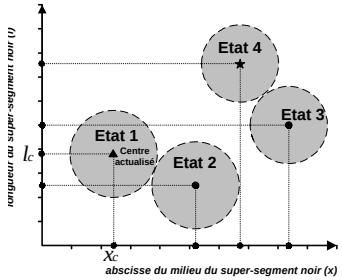

FIG. 6 : agrégation de la deuxième séquence

Puisque l'agrégation est progressive, et puisqu'elle se fait en une seule passe, la construction des états dépend de l'ordre dans lequel sont présentées les séquences. Les courtes séquences contiennent des observations résultant d'un fusionnement, souvent de deux blocs (figure 3). La plus longue séquence fournit statistiquement la plus grande contrainte de séparabilité. La plus petite demi-distance entre toutes les paires de points représentatifs de cette séquence, est donc prise à la fois comme le rayon initial, et le rayon de référence (R_0) des disques (plus généralement, hyper-sphères) qui définissent les lieux géométriques des états (figure 4). Une observation de chaque nouvelle séquence est affectée à un état existant si le point de représentation est situé dans le disque correspondant ; elle crée un nouvel état si le point de représentation est situé en dehors de tous les disques (figure 5). Après cette étape d'affectation, les positions des centres des disques sont recalculées. Le nouveau rayon R_i relatif à un disque (d'un état existant ou d'un nouvel état) est égal à la demi distance au centre de l'état le plus proche, celle-ci ne peux toutefois excéder le rayon de référence R_0 (figure 6). C'est donc une nouvelle représentation des états qui est prise en compte pour la prochaine séquence d'observations. A la fin du processus d'agrégation, on calcule les paramètres du modèle probabiliste de l'observation associée à l'état. Ces paramètres sont estimés statistiquement pour chaque état "i". On calcule le vecteur moyen $\hat{\mu}_i$ et la matrice de covariance $\hat{V}_i$ des observations dans cet état :

$$\hat{\mu}_i = \frac{1}{N_i}\sum_{O_t \in i} O_t \quad , \quad \hat{V}_i = \frac{1}{N_i}\sum_{O_t \in i} (O_t - \hat{\mu}_i)^T (O_t - \hat{\mu}_i)$$

Le modèle probabiliste des observations est un modèle gaussien qui intègre le vecteur moyen $\hat{\mu}_i$ et la matrice de covariance $\hat{V}_i$.

3. Expérimentation et conclusion

3.1 Analyse de l'apprentissage d'un HMM secondaire

La figure 7 montre la composition de chacun des cinq types de bandes affectées à un même super-état, résultant de l'algorithme de programmation dynamique [1,2]. La première observation apparaît dans les quatre premiers types de bandes. On voit que les deux derniers super-segments de la bande A, en fusionnant, ont engendré le dernier super-segment de la bande B. A leur tour, les deux derniers super-segments de la bande B, en fusionnant, ont donné naissance au dernier super-segment de la bande C. L'absence de deuxième super-segment dans la bande D, et l'absence totale de super-segments dans la bande E, s'expliquent par l'existence d'une fragmentation verticale dans chacun des deux cas.

La figure 8 présente le graphe des états du HMM secondaire obtenu par la méthode d'agrégation dynamique. Une transition de bouclage a été ajoutée sur chaque état, avec une très faible probabilité, pour absorber une éventuelle variation non apprise. L'apprentissage supervisé effectué par

la méthode classique des k-moyennes fixe à deux le nombre des états (figure 9). La première observation de chaque séquence est rangée dans le premier état, toutes les autres observations sont rangées dans le deuxième état. Le HMM appris par l'apprentissage dynamique des états reflète fidèlement la réalité physique, alors que le modèle appris par la méthode des k-moyennes traduit la réalité logique. L'apprentissage non supervisé donne un modèle moins généralisé, mais plus précis, alors que l'apprentissage des k-moyennes, en groupant dans un même état des observations physiquement différentes, généralise au détriment de la précision, mesurée par l'inverse de la dispersion (figure 10).

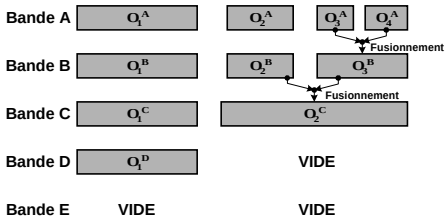

FIG. 7 : les 5 sortes de bandes du super-état

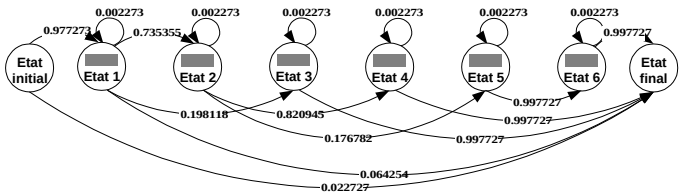

FIG. 8 : HMM secondaire (méthode d'agrégation dynamique)

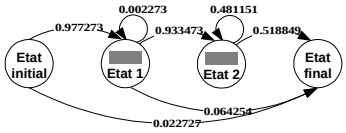

FIG. 9 : HMM secondaire (méthode classique des k-moyennes)

Etat	1	2	3	4	5	6
Dispersion	85.9	8.55	1	18.6	1	1

a) Méthode d'agrégation dynamique

Etat	1	2
Dispersion	85.9	107880

b) Méthode classique des k-moyennes

FIG. 10 : tableau de la dispersion des observations dans les états

3.2 Résultat global de la reconnaissance

La base d'apprentissage est constituée de 50 classes. Chacune des classes comprend 20 formulaires remplis par des scripteurs différents. La reconnaissance a été testée sur une autre base comprenant les 50 mêmes classes. Chaque classe comprend 10 exemplaires remplis par des scripteurs différents. Nous avons obtenu avec la nouvelle méthode (apprentissage non supervisé) un taux de reconnaissance de : 97,6% au lieu de 90,4% par la méthode classique (apprentissage supervisé par les k-moyennes). Nous expliquons l'amélioration du taux de reconnaissance par la méthode de l'apprentissage non supervisé, par le fait que le processus de création des états par agrégation des observations rend compte de la réalité physique (phénomène de fusionnement-fragmentation) avec plus de précision.

4. Bibliographie

[1] S. Ramdane. *Identification automatique de types de formulaires par des méthodes stochastiques markoviennes*. Thèse, Univ. Le Havre, décembre 2002.

[2] S. Ramdane, B. Taconet et A. Zahour. *Classification of Forms with Handwritten Fields by Planar Hidden Markov Models*. Pattern Recognition, vol. 36/4, pp. 1045–1060, April 03.

[3] S. Kebairi, A. Zahour, B. Taconet, L. Boukined. *Segmentation of Composite Documents Into Homogenous Blocks.* Proc. IGS'98, pp. 111-112, 1997.

[4] L. Boukined, B. Taconet. *Recherche de la Structure Physique d'un Document Imprimé par Rectangulation.* Proc. RFIA'91, pp. 1027-1031, 1991.

[5] N. Ben Amara, A. Belaïd. *Printed PAW Recognition Based on Planar Hidden Markov models.* ICPR'96, pp. 220-224, 1996.

[6] D. Fasulo. *An Analysis of Recent Work on Clustering Algorithms.* Technical Report # 01-03-02, Department of Computer Science & Engineering, University of Washington, April 1999.

[7] A. Ben-Dor, Z. Yakhini. *Clustering Gene Expression Patterns.* In Proceedings of RECOMB 99, 1999.

[8] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein. *Cluster Analysis and Display of Genome-Wide Expression Patterns.* Proceedings of the National Academy of Sciences, 95:14863-14868, December 1998.

[9] D. Cosic, S. Loncaric. *New Methods for Cluster Selection in Unsupervised Fuzzy Clustering.* Proceedings of the 41th Conference KoREMA'96, vol. 4, pp. 1-3, Opatija, Croatia, 1996.

[10] C. Fraley, A. E. Raftery. *How Many Clusters? Which Cluster Method? Answers Via Model-based Cluster Analysis. T*echnical Report 329, Department of Statistics, University of Washington, February 1998.