SEQTREE, un outil de fouille de données séquentielles par visualisation

Christine Largeron

Université Jean Monnet de Saint-Etienne
EURISE
23, rue du docteur Paul Michelon
42023 Saint-Etienne Cedex 2
Christine.Largeron@univ-st-etienne.fr

Résumé. Dans cet article, nous présentons un outil de visualisation de séquences modélisées par des arbres de suffixes probabilistes (Prediction suffix trees - PST). Ce type d'arbre permet de représenter une chaîne de Markov d'ordre variable. Dans différentes applications, il s'est avéré plus efficace qu'une chaîne de Markov d'ordre fixe avec un coût calculatoire moindre. Pour ces raisons, il nous a paru intéressant d'exploiter le caractère arborescent de ce mode de représentation non seulement d'un point de vue algorithmique mais aussi d'un point de vue visuel.

1 Introduction

Avec l'émergence de la fouille de données visuelle [Card et al., 1999, Spence, 2001, Keim, 2002, Davidson and Soukup, 2002, Poulet, 2004], les techniques de visualisation permettent de mieux appréhender les données et d'impliquer davantage l'utilisateur dans le processus d'extraction de connaissances. C'est dans cette perspective, que nous avons développé un logiciel de représentation et de comparaison de séquences par visualisation. Par séquence, nous entendons une suite de valeurs observées dans le temps. Il peut s'agir par exemple en climatologie du temps observé quotidiennement dans une région pendant une période donnée, en bioinformatique de séquences d'ADN. Si on suppose que le phénomène étudié présente une dépendance temporelle ; ce qui signifie que la valeur observée à un instant dépend des valeurs observées antérieurement ou du moins de certaines d'entre elles, on peut avoir recours à un modèle de Markov d'ordre variable [Rissanen, 1983]. Un modèle de Markov d'ordre variable peut être représenté par un arbre. Cet arbre, construit à partir de séquences d'apprentissage, peut être utilisé ensuite pour classer de nouvelles séquences ou pour prédire le caractère suivant dans une séquence. Cependant, à notre connaissance, les possibilités offertes par ce mode de représentation arborescent n'ont pas été exploitées dans une perspective de fouille de données visuelle. C'est la raison pour laquelle, nous avons conçu un outil de visualisation et de comparaison de séquences reposant sur ce modèle. Cet outil sera décrit dans la troisième section ; la suivante étant consacrée au modèle de Markov d'ordre variable et à sa représentation sous forme de PST.

2 Chaînes de Markov d'ordre variable et Arbres de suffixes probabilistes

2.1 Définition

Par rapport à une chaîne de Markov d'ordre fixe L, une chaîne d'ordre variable (Variable Memory Markov Model) [Rissanen, 1983] exploite l'idée que dans certaines séquences naturelles la longueur de la mémoire dépend du contexte et n'est pas fixe. Ceci conduit en fait à conserver un historique de longueur maximale L dans certains cas mais à le limiter lorsque la prise en compte d'un événement supplémentaire ne modifie pas significativement la distribution des probabilités conditionnelles. Un modèle de Markov d'ordre variable peut être représenté par un arbre de suffixes probabilistes (Prediction Suffix Tree - PST) tel qu'il a été défini dans Ron [Ron et al., 1996]. Ainsi, par exemple, la séquence $s = (aabaabaabaab)$ définie l'alphabet $\Sigma = \{a,b\}$ peut être représentée avec une longueur de mémoire L égale à 2 par le PST S décrit dans la figure 1. Dans ce PST, chaque noeud interne est l'extrémité initiale de $|\Sigma|$ arcs correspondant chacun à un symbole distinct et unique de Σ. De plus, chaque noeud de l'arbre est étiqueté par un couple (k,φ^k) où k est le mot correspondant au chemin parcouru depuis ce noeud jusqu'à la racine ε de l'arbre, et φ^k est un vecteur de dimension $|\Sigma|$ dont chaque composante φ^{kj} est égale à la probabilité conditionnelle d'observer le caractère j de Σ après le mot k.

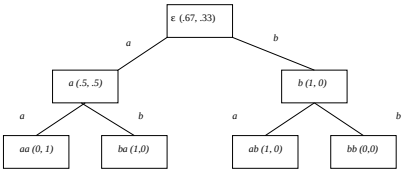

FIG. 1 – *PST S associé à la séquence* $s = (aabaabaabaab)$

Lorsque l'arbre est complet (*ie.* tous ses noeuds internes ont exactement $|\Sigma|$ fils) et que le nombre de caractères du mot associé à une feuille de l'arbre est L, le PST représente une chaîne de Markov d'ordre fixe L. Par élagage, on obtient un arbre qui correspond à une chaîne de Markov d'ordre variable dont le coût de stockage est réduit par rapport au modèle d'ordre fixe. Parmi les algorithmes d'apprentissage de modèles de Markov d'ordre variable proposés [Willems, 1998, Seldin et al., 2001], nous avons retenu celui développé par Ron et *al.* [Ron et al., 1996] et repris par Bejerano et Yona [Bejerano and Yona, 2001]. Il permet d'apprendre un modèle d'ordre variable L en un temps polynomial en L. En combinant l'élagage de l'arbre et un lissage des probabilités, il permet de construire un PST S à partir d'une ou plusieurs séquences d'apprentissage. Dans un contexte de classement supervisé, cet arbre S peut ensuite être utilisé pour calculer la probabilité qu'une nouvelle séquence ait été générée par le modèle d'ordre variable décrit par S.

3 Représentation et comparaison visuelle de séquences

Compte tenu des avantages et des performances obtenues dans différentes applications par le modèle d'ordre variable comparativement à d'autres modèles [Bejerano and Yona, 2001, Largeron-Leténo, 2003], il nous a paru intéressant d'exploiter dans une perspective de fouille de données visuelle son mode de représentation arborescent. Ceci nous a conduit à développer un logiciel graphique de visualisation et de prévision de séquences disponible en ligne[1]. Ses différentes fonctionnalités sont décrites dans les sections suivantes et illustrées à partir d'un exemple simple de prévision climatique. Dans cet exemple, on se propose à partir du temps observé quotidiennement dans une région de caractériser son climat. On distingue quatre types de climat possibles : océanique, continental, méditerranéen et montagnard. Pour identifier le climat d'une zone géographique, on relève chaque jour le temps qu'il fait dans cette zone pendant une période de temps suffisamment longue. Chaque jour, le temps peut être décrit suivant quatre modalités : beau temps, pluie, vent, nuageux notées respectivement B, P, V et N. On suppose que non seulement la fréquence d'apparition de chacune de ces modalités mais aussi leur succession permettent de différencier les différents climats. On peut donc employer un modèle de Markov d'ordre variable pour modéliser chaque climat. Pour construire ce PST pour chaque type de climat, on utilise en phase d'apprentissage supervisé, une séquence d'observations effectuées dans une zone déjà identifiée par un expert comme relevant de ce climat. A partir des modèles représentatifs des différents climats, on peut ensuite, en phase de classement, prédire le climat d'une nouvelle zone en disposant uniquement de la séquence des relevés de temps effectués quotidiennement et en calculant la probabilité pour que cette séquence ait été générée par chacun des modèles.

3.1 Représentation d'une séquence par PST

La première fonctionnalité offerte par le logiciel est la représentation visuelle sous forme d'arbre du modèle de Markov d'ordre variable sous-jacent à une séquence. Cette première fonctionnalité paraît d'autant plus importante que le modèle construit à l'aide d'un logiciel d'apprentissage de PST se présente sous un formalisme logique qui répond bien aux contraintes d'implémentation mais qui s'avère difficilement compréhensible par un utilisateur non averti.
Ainsi, dans l'exemple climatique cité précédemment, on obtient un PST S_m à partir d'une séquence s_m de temps relevés pendant 86 jours dans une région caractérisée comme ayant un climat montagneux par un expert et définie par : $s_m = (bnpvbvbvbpv$ $nbvbvbvbvbnvbvbvnbvbvbpbvbvbvbvbvbvbvnbvbvbvbvbvbvbvbvbvbvbnbvbvbvbvbnvbvbbv)$.
A partir du fichier[2] contenant le modèle S_m appris à partir de la séquence s_m, l'application fournit la représentation graphique arborescente correspondant à la figure 2, exploitable visuellement grâce à différentes capacités offertes par le logiciel telles que le zoom vers des points d'intérêt, l'élagage des éléments les moins intéressants, la

1. `http://eurise.univ-st-etienne.fr/~largeron/RNTI_visualisation/index.htm`
2. Fichier montagnePST.TXT sur le site

personnalisation de la représentation graphique selon les préférences de l'utilisateur.

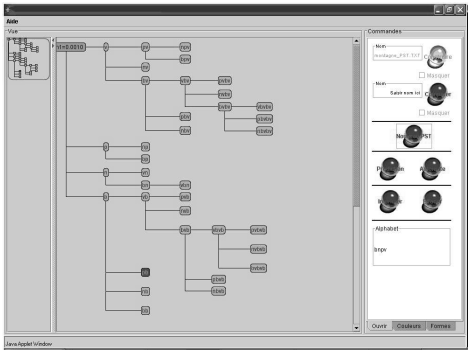

FIG. 2 – *Visualisation graphique arborescente du PST* S_m

3.2 Représentation simultanée de plusieurs séquences

Pour permettre une analyse comparative de deux modèles, l'application fournit simultanément leur représentation graphique arborescente. Des couleurs différentes indiquent les noeuds qui appartiennent au premier modèle, au second et ceux qui sont communs aux deux. L'utilisateur peut ensuite conserver uniquement les noeuds communs aux deux modèles, mettant ainsi en évidence leur différence ou au contraire leur similarité. Ainsi, dans l'exemple climatique, la figure 3 fait apparaître en plus du modèle S_m précédent, le modèle S_o[3] caractérisant le climat océanique et défini à partir de la séquence s_o suivante : s_o = (*bnpvbnvpbnnvpbbbnvpbnvpbnvpppbbbnvpbnvpbnvpbnvpbnvpbbbnvpbnvpbnvpbnvpbnvpbbbnvpbnvppbnvpppbnvpbnvp*). On peut noter que les deux séquences n'ont pas nécessairement la même longueur.

3.3 Classement d'une nouvelle séquence

La dernière fonctionnalité offerte par l'interface graphique est le classement d'une nouvelle séquence. Le logiciel permet l'affichage sur le modèle des noeuds intervenant dans le calcul de la probabilité et enfin du résultat final. Ainsi, en reprenant l'exemple climatique, la figure 4 illustre le déroulement de la phase de classement pour une séquence s' = (*bnvbvbvbpvnbvnbvvbvbnvnvbvnvbv*) relevée dans une région dont on souhaite identifier le climat. Le logiciel affiche la probabilité pour que cette séquence s' ait été générée par le modèle S_m à savoir $1{,}53 \times 10^{-10}$.

3. Fichier oceanPST.TXT sur le site

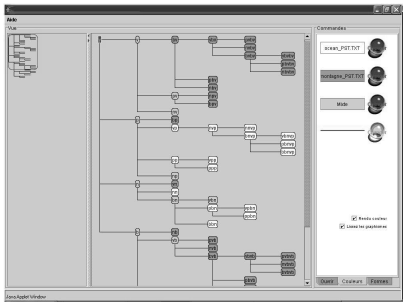

FIG. 3 – *Visualisation graphique simultanée de S_m et de S_o*

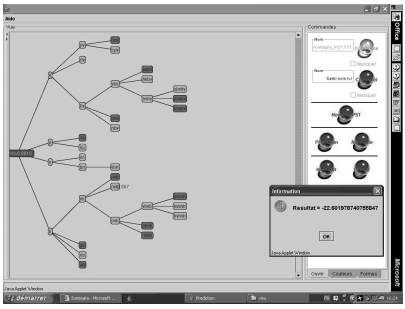

FIG. 4 – *Prévision d'une séquence à partir de S_m*

4 Conclusion

Les arbres de suffixes probabilistes permettent de représenter des modèles de Markov d'ordre variable. En ce sens, ils fournissent une généralisation des modèles de Markov d'ordre fixe, capables grâce à leur taille de mémoire variable, de capturer des dépendances à long terme présentes dans des séquences. Pour exploiter cette représentation arborescente nous avons développé un outil graphique qui permet de visualiser une séquence sous forme de PST, puis de la comparer à une autre en distinguant les noeuds communs aux deux modèles. Cet outil peut aussi être employé pour classer une nouvelle séquence. Dans ce contexte de classement supervisé, il apporte également une information complémentaire par rapport au modèle de Markov d'ordre variable en mettant en évidence les sous-séquences qui n'ont pas été observées dans la nouvelle séquence bien qu'elles soient caractéristiques du modèle. Ainsi, cet outil permet de mieux appréhender la structure des séquences et d'améliorer le processus de fouille de données par leur visualisation.

Références

[Bejerano and Yona, 2001] Bejerano, G. and Yona, G. (2001). Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics*, 17(1):23–41.

[Card et al., 1999] Card, K., Mackinlay, J., and Schneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

[Davidson and Soukup, 2002] Davidson, I. and Soukup, T. (2002). *Visual data mining*. Wiley.

[Keim, 2002] Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Trans. on visualizations and computer graphics*, 7(1):100–107.

[Largeron-Leténo, 2003] Largeron-Leténo, C. (2003). Prediction sufffix tree for supervised classification of sequences. *Pattern recognition letters*, 24:3153–3164.

[Poulet, 2004] Poulet, F. (2004). Towards visual data mining. In *Proceedings of the 6th International Conference on Enterprise Information Systems ICEIS*, pages 349–356, Portugal Porto.

[Rissanen, 1983] Rissanen, J. (1983). A universal data compression system. *IEEE Trans Infor Theory*, 29(5):656–664.

[Ron et al., 1996] Ron, D., Singer, Y., and Tishby, N. (1996). The power of amnesia : learning probabilistic automata with variable memory length. *Machine learning*, 25:117–149.

[Seldin et al., 2001] Seldin, Y., Bejerano, G., and Tishby, N. (2001). Unsupervised sequence segmentation by mixture of switching variable memory sources. In *Proceedings of the Eighteenth International Conference of Machine Learning*, pages 513–520. ICML.

[Spence, 2001] Spence, B. (2001). *Information visualization*. Addison-Wesley, ACM Press.

[Willems, 1998] Willems, F. (1998). The context tree weighting method: extensions. *IEEE Trans. Infor. Theory*, pages 792–798.

Summary

This paper presents a visual tool for sequences analysis. Sequences are represented by prediction suffix trees (PST). PST can be used to efficiently describe variable order chain. It performs better than the Markov chain of order L and at a lower cost. For this reason, it is interesting to exploit the tree representation not only from a computational point of view but also from a visual one. The tool, developped in this aim, provides a graphic representation of a PST constructed from sequences. It can also be used to compare two models. In supervised classification context, it brings more information than the model by underlying sub-sequences observed in the new sequence and which are not in the model of its predicted class. By the way of visualisation, this tool upgrades the data mining process and the comprehension of the information encoded in sequences.