Apprentissage sur des données multivariées pour la modélisation de comportement

Florence DUCHÊNE[1]

[1]INRIA – Projet ORION – 2004 route des Lucioles BP 93 – 06902 Sophia-Antipolis Cedex – Florence.Duchene@sophia.inria.fr

1. Introduction

Un système de surveillance nécessite souvent une démarche d'apprentissage pour la constitution d'une base de connaissances *a priori* utiles à la décision, i.e. à la reconnaissance de situations indésirables. Si ces situations sont difficiles à décrire ou observer, l'apprentissage vise à modéliser le comportement dit « normal » du système pour la détection des situations critiques par comparaison à ce modèle. On propose dans ce contexte une méthode générique d'extraction non supervisée de motifs temporels, représentatifs de comportements récurrents, à partir de données multidimensionnelles et hétérogènes issues de capteurs variés. Une application concerne l'apprentissage des habitudes de vie d'une personne à domicile.

2. Méthode

La méthode proposée est illustrée sur la figure 1. La première étape consiste en l'**abstraction** des données issues des capteurs, pour leur donner un sens au niveau de décision. Il s'agit de résumer les situations stationnaires par des symboles estampillés (vecteurs discrets), significatifs de la continuité d'une action. La fouille de données comprend ensuite la **fouille de caractères** pour l'identification des *tentatives de motifs* (sous-séquences récurrentes), puis leur **classification** en *motifs* (classes de ces sous-séquences).

La **fouille de caractères** est basée sur les *projections aléatoires* [1,2] de l'ensemble des sous-séquences d'une longueur fixée, pour leur comparaison deux à deux sur une partie seulement de leurs symboles. Une *matrice de collisions* enregistre le nombre de projections identiques. Les valeurs de collisions sont examinées selon deux seuils : (1) *minimum de collisions* entre sous-séquences discrètes, puis (2) *maximum de distance* entre les sous-séquences réelles correspondantes. Une *synthèse* des récurrences observées permet d'identifier les *tentatives de motifs*, selon des critères de signification et non redondance.

Une distance non métrique sur des séquences multidimensionnelles hétérogènes est proposée. Elle s'appuie sur la plus longue sous-séquence commune, définie selon deux seuils de similarité sur les valeurs et dans le temps [3].

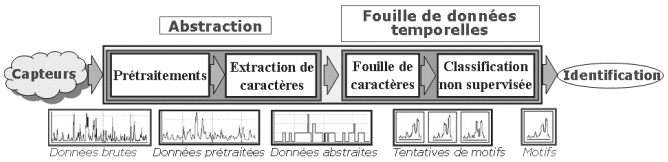

FIG. 1 : Principe de la méthode d'apprentissage proposée

3. Résultats et discussion

Les bonnes performances moyennes de l'extraction de motifs sur des données simulées sont encourageantes. Une classification parfaite est par ailleurs possible dans 20% des cas. Un exemple obtenu à partir de données simulées pour une personne à domicile est présenté sur la figure 2. L'application actuelle de la méthode sur des données issues de l'interprétation de vidéos est illustrée sur la figure 3.

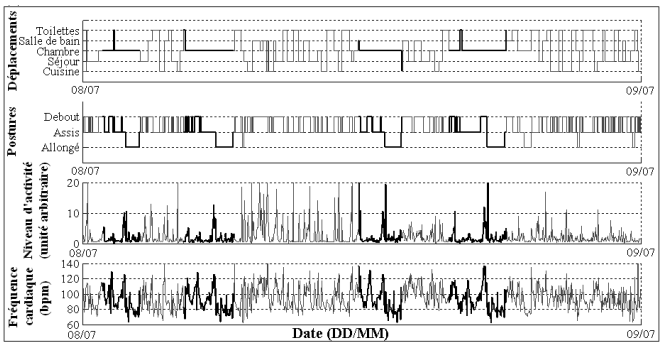

FIG. 2 : Illustration de l'extraction des instances de motif

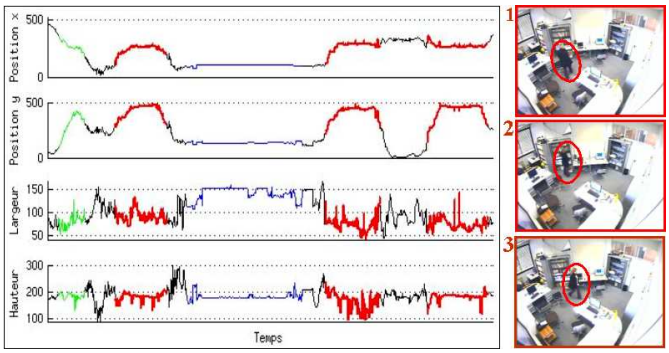

FIG. 3 : Extraction de motifs sur des données vidéos

Références

[1] J. Buhler et M. Tompa, *Finding motifs using random projections*, Journal of Computational Biology, 9(2), pp. 225-242, 2002.

[2] B. Chiu, E. Keogh et S. Lonardi, *Probabilistic Discovery of Time Series Motifs*, Proc.s of the 9th ACM International Conference on Knowledge Discovery and Data Mining, Washington DC, USA, pp.493-498, 2003.

[3] M. Vlachos, G. Kollios et G. Gunopulos, *Discovering Similar Multidimensional Trajectories*, Proc. of the 18th International Conference on Data Engineering, San Jose, CA, pp. 673-684, 2002.