Extraction non supervisée de motifs temporels, multidimensionnels et hétérogènes Application à la télésurveillance médicale à domicile

Florence Duchêne*, Catherine Garbay* et Vincent Rialle* **

*Laboratoire TIMC-IMAG, Faculté de médecine de Grenoble
Florence.Duchene@sophia.inria.fr, Catherine.Garbay@imag.fr
**Département d'Informatique Médicale (SIIM), Hôpital Michallon, Grenoble
Vincent.Rialle@imag.fr

Résumé. Une méthode générique pour l'extraction non supervisée de motifs dans des séquences temporelles multidimensionnelles et hétérogènes est proposée, puis expérimentée pour l'identification des comportements récurrents d'une personne à domicile. L'objectif est de concevoir un système d'apprentissage des habitudes de vie, à partir des données de capteurs, pour la détection d'évolutions critiques à long terme.

1 Introduction

Dans l'objectif de détecter les évolutions critiques à long terme de personnes à domicile, on souhaite mettre en place un système d'apprentissage d'un profil comportemental dans la vie quotidienne. Toute modification des activités habituelles pouvant correspondre à une dégradation de l'état de santé, un écart par rapport à ce profil est considéré inquiétant. Il s'agit d'extraire des *motifs* "haut niveau" de séquences temporelles "bas niveau" collectées de capteurs installés au domicile. Un *motif* est le représentant d'une classe de sous-séquences récurrentes, et correspond à un comportement type de la personne. Les caractéristiques de ce problène sont les suivantes :

1. **Méthode** – L'extraction de motifs est *non supervisée* pour s'adapter aux spécificités individuelles de comportement et au manque de connaissances *a priori*.
2. **Séquences temporelles** – Les séquences analysées sont multidimensionnelles, hétérogènes (données qualitatives ou quantitatives), et *mixtes* : elles contiennent à la fois des sous-séquences représentatives de *motifs* et des "*non motifs*".
3. **Motifs** – On recherche des *motifs multidimensionnels* afin d'éviter une sursimplification du système observé, et la non détection de certaines évolutions critiques. Par ailleurs, les instances d'un motif ont les caractéristiques suivantes :
   - **Variabilité dans les valeurs**, due à celle des comportements humains.
   - **Présence d'interruptions** dans la réalisation d'une activité (toilettes, etc.).
   - **Déformations et translation dans le temps**, car une même activité se répète à des instants et sur des durées variables.

Dans ce contexte non supervisé, et concernant de larges ensembles de données temporelles, l'extraction de motifs se rapporte à un problème de *fouille de données temporelles* [Antunes et Oliveira 2001, Roddick et Spiliopoulou 2002]. Pour prendre en compte l'écart entre le bas niveau des données des capteurs et les objectifs d'apprentis-

sage à long terme, *plusieurs niveaux d'analyse* sont considérés, pour aboutir à des informations efficaces pour la fouille de données et adaptées au sens de la décision. En particulier, une étape d'abstraction [Höppner 2002] fournit une représentation symbolique interprétable des séquences issues des capteurs. La constitution des séquences analysées – succession de *motifs* et *non-motifs* – et la quantité exponentielle des sous-séquences possibles impose également une étape de *fouille de caractères* [Kudenko et Hirsh 1998, Lesh et al. 2000], pour l'identification des sous-séquences les plus susceptibles de correspondre aux instances de motifs.

L'originalité de ce travail par rapport aux recherches déjà effectuées sur l'extraction de motifs temporels est la prise en compte de séquences de données multidimensionnelles et hétérogènes, pour l'identification de motifs multidimensionnels. La section 2 décrit la méthode proposée. Les résultats expérimentaux sont discutés en section 3, et nos conclusions présentées en section 4.

2 Méthode pour l'extraction de motifs

La méthode proposée est illustrée sur la figure 1. La première étape consiste en l'**abstraction** des données issues des capteurs. 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).

2.1 Abstraction

La représentation des séquences issues des capteurs est une *abstraction*, pour leur donner un sens en fonction des objectifs d'analyse. Il s'agit de résumer les situations stationnaires observées pour mettre en évidence les tendances de variation au niveau de détail pertinent pour l'analyse, en trois étapes :

1. **Prétraitement** – On supprime la variabilité très haute fréquence.
2. **Discrétisation** – On agit sur l'*axe des valeurs* en discrétisant les paramètres quantitatifs. L'idée est de disposer de séquences homogènes, discrètes, plus faciles à analyser, et sur lesquelles les méthodes de fouille de données sont applicables.
3. **Agrégation** – On intervient ensuite sur l'*axe temporel* en agrégeant dans le temps, selon des intervalles de longueur variable, les vecteurs discrets successifs ne présentant pas de forte variation au regard du niveau de décision. Les intervalles de stationnarité sont déterminés en fonction d'un seuil maximum de distance entre les vecteurs discrets, le symbole associé étant alors le vecteur discret moyen.

Une séquence est ainsi représentée par une succession de symboles estampillés. Chaque symbole est significatif de la continuité d'une même action, à l'échelle de décision.

2.2 Fouille de caractères

Une étape de fouille de caractères est nécessaire pour identifier les sous-séquences récurrentes – les *caractères* – et réduire ainsi l'espace de recherche des instances effectives de motifs. D'après [Lesh et al. 2000], les caractères doivent être (1) fréquents, (2) distinctifs d'au moins une classe, et (3) non redondants.

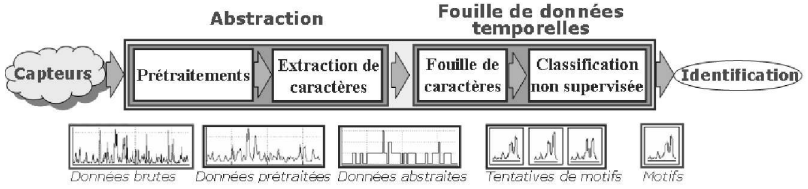

FIG. 1 – *Principe de la méthode proposée pour l'extraction de motifs temporels.* Les signaux illustrent le type des données et informations disponibles après chaque étape.

Fréquence
Une méthode efficace d'identification de sous-séquences récurrentes en contexte bruité est basée sur les *projections aléatoires* [Buhler et Tompa 2002, Chiu et al. 2003] de l'ensemble des sous-séquences dites *de base* (longueur fixe de symboles), pour leur comparaison deux à deux sur une partie seulement de leurs symboles. Une *matrice de collisions* enregistre pour chaque couple de sous-séquences le nombre de fois où leurs projections ont été identiques. Les valeurs élevées de cette matrice montrent ainsi une forte présomption de similarité. L'application de cette méthode est étendue aux séquences multidimensionnelles, et à la prise en compte de déformations temporelles possibles entre sous-séquences récurrentes.
Signification
Les valeurs de la matrice de collisions sont examinées selon deux critères : (1) un *seuil minimum de collisions* entre sous-séquences discrètes, puis (2) un *seuil maximum de distance* entre les sous-séquences *réelles* correspondantes. Les sous-séquences récurrentes significatives pour l'extraction de motifs sont identifiées en trois étapes :
- Identification d'un couple de sous-séquences *de base* satisfaisant ces deux critères ;
- Extension de ces sous-séquences pour déterminer les sous-séquences similaires "complètes", plus longues en terme du nombre de symboles qui les représente.
- Vérification d'un seuil minimum de durée de chaque sous-séquence.

Non redondance
Ce critère impose une démarche de synthèse des sous-séquences récurrentes identifiées d'après la matrice de collisions, afin de regrouper les sous-séquences représentatives d'une même instance de motif. Leurs représentants forment alors un ensemble de *tentatives de motifs* disjointes. On utilise une méthode de classification divisive.

2.3 Classification

La classification des *tentatives de motifs* est enfin réalisée sur la base d'une mesure de distance. On applique une méthode agglomérative non supervisée : la classification ascendante hiérarchique. Le représentant de chaque classe définit un *motif*.

2.4 Mesures de similarité

Une mesure approchée de similarité, appliquée sur des séquences discrètes, est nécessaire à l'abstraction pour l'agrégation des vecteurs successifs correspondant à un

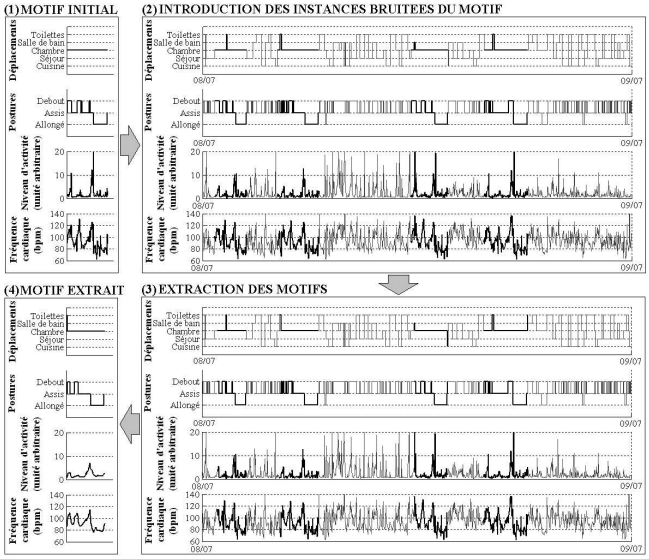

FIG. 2 – *Illustration de l'extraction des instances d'un motif insérées initialement dans une séquence de non-motifs issue d'un processus de simulation.*
Les figures proposées montrent les instances de motifs sur les données prétraitées.

état stationnaire du système. Par ailleurs, la comparaison des séquences réelles est nécessaire à la fouille de données. Les fonctions de similarité non métriques s'avèrent particulièrement efficaces entre séquences bruitées. La mesure de distance proposée s'appuie ainsi sur la plus longue sous-séquence commune, définie selon deux seuils de similarité sur les valeurs et dans le temps [Vlachos et al. 2002]. Cette notion de similarité est étendue à la comparaison de séquences multidimensionnelles hétérogènes.

3 Résultats et Discussion

3.1 Contexte expérimental

L'approche proposée est expérimentée dans le contexte de la télésurveillance médicale à domicile. Les *tentatives de motifs* représentent les comportements récurrents d'une personne dans sa vie quotidienne ; les *motifs* sont les classes de ces comportements, représentatives d'activités "types". Face au manque d'un ensemble complet et représentatif d'enregistrements réels, l'expérimentation est réalisée sur des données simulées. On considère des séquences hétérogènes à quatres dimensions :

- **Déplacement** – Pièces successivement occupées par la personne dans l'habitat.
- **Posture** – Postures successives : allongé, assis ou debout.
- **Niveau d'activité** – Norme de l'accélération sur l'axe antérieur-postérieur, calculée en moyenne toutes les minutes, et représentative des mouvements effectués.
- **Fréquence cardiaque** – Valeur moyenne calculée toutes les minutes.

L'utilisation de données simulées permet de générer plusieurs types de situations :

- **Modifications "normales"** dans la réalisation répétée d'une même activité : variabilité dans les valeurs, interruptions et déformations temporelles.
- **Modifications inquiétantes** : ruptures dans les principes de variation habituels.

Des indices de **sensibilité** et **spécificité** permettent d'évaluer les performances, *i.e.* :

- L'identification de *tentatives de motifs* correspondant effectivement aux instances d'un *motif*, sans inclure de points issus d'intervalles de *non-motifs*.
- Le regroupement de toutes et d'uniquement les instances d'un *motif* dans une seule classe.

3.2 Résultats obtenus

On constate de bonnes performances moyennes de l'identification et de la classification des motifs (voir tableau 1). Ces résultats sont encourageants, d'autant plus qu'ils montrent qu'une classification parfaite est possible dans 20% des cas. Un exemple est présenté sur la figure 2. La variabilité observée peut s'expliquer par la prise en compte de motifs qui ne sont peut-être pas tous réellement significatifs, car sélectionnés aléatoirement dans des séquences simulées.

	Identification		**Classification**	
Indices	S_e	S_p	S_e	S_p
Moyenne	**0.71**	**0.92**	**0.66**	**0.79**
Écart-type	0.18	0.07	0.34	0.26
Indices parfaits	–	–	35%	60%
Classification parfaite			**20%**	

TAB. 1 – *Indices moyens de sensibilité (S_e) et de spécificité (S_p) de l'extraction de motifs, dans une configuration par défaut du système et contexte relativement bruité.*

Les **tests de sensibilité** montrent une bonne résistance aux modifications normales de comportement. Les **tests de spécificité** permettent de vérifier la diminution progressive, avec la dégradation des instances de motifs, du taux de reconnaissance comme "normales" des instances anormalement modifiées (voir tableau 2). Enfin, l'application de la méthode sur des séquences simulées dans des conditions habituelles de vie d'une personne permet d'extraire des motifs interprétables *a posteriori* en terme de la réalisation d'activités quotidiennes.

Modifications	**Normales**	**Inquiétantes**			
Taux de modification	*0%*	*10%*	*20%*	*30%*	*40%*
Taux de reconnaissance	**77.5%**	**25%**	**10%**	**10%**	**5%**

TAB. 2 – *Taux de reconnaissance comme "normales" d'instances bruitées puis anormalement modifiées d'un motif, selon des taux croissants de dégradation.*

4 Conclusion

Une approche générique pour l'extraction de motifs temporels, multidimensionnels et hétérogènes a été proposée puis expérimentée dans le cadre de la télésurveillance médicale à domicile. Les résultats ont mis en évidence les potentialités de la méthode. Sa mise en pratique est cependant complexe étant donné le nombre et les difficultés de réglage des paramètres. Le manque de données réelles ne permet par ailleurs pas une validation complète des résultats. La généricité de la méthode doit cependant permettre son application à différents niveaux d'analyse et contextes de surveillance.

Références

[Antunes et Oliveira 2001] Antunes C. et Oliveira A. (2001). Temporal data mining : an overview. Dans *Proceedings of the Workshop on Temporal Data Mining at the 7^{th} International Conference on Knowledge Discovery and Data Mining (KDD'01), San Francisco, CA*, pages 1–15.

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

[Chiu et al. 2003] Chiu B., Keogh E. et Lonardi S. (2003). Probabilistic discovery of time series motifs. Dans *Proceedings of the 9^{th} ACM International Conference on Knowledge Discovery and Data Mining (KDD'03), Washington DC., USA*, pages 493–498.

[Höppner 2002] Höppner F. (2002). Time series abstraction methods – a survey. Dans *Proceedings of the GI Jahrestagung Informatik, Workshop on Knowledge Discovery in Databases, Dortmund, Germany*, pages 777–786.

[Kudenko et Hirsh 1998] Kudenko D. et Hirsh H. (1998). Feature generation for sequence categorization. Dans Press A., éditeur, *Proceedings of the 15^{th} Nat'l Conf. Artificial Intelligence (AAAI'98), Menlo Park, California*, pages 733–739.

[Lesh et al. 2000] Lesh N., Zaki M. et Ogihara M. (2000). Scalable feature mining for sequential data. *IEEE Intelligent Systems*, 15(2) :48–56.

[Roddick et Spiliopoulou 2002] Roddick J. et Spiliopoulou M. (2002). A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering*, 14(4) :750–767.

[Vlachos et al. 2002] Vlachos M., Kollios G. et Gunopulos G. (2002). Discovering similar multidimensional trajectories. Dans *Proceedings of the 18^{th} International Conference on Data Engineering (ICDE'02), San Jose, CA*, pages 673–684.

Summary

A generic unsupervised learning process for mining heterogeneous multivariate time-series and identifying temporal patterns is proposed, and experimented in the context of identifying recurrent behaviors of a person at home. We aim at learning daily living habits from sensors data, in order to detect worrying changes over the long term.