Modélisation et reconnaissance d'activités quotidiennes au sein d'une maison intelligente : application à la surveillance des personnes âgées

Josky Aïzan

Modélisation et reconnaissance d'activités quotidiennes au sein d'une maison intelligente : application à la surveillance des personnes âgées

THÈSE

présentée et soutenue publiquement le 21 Octobre 2020

pour l'obtention du

Doctorat delivré conjointement par l'Université d'Abomey-Calavi et l'Université du Littoral Côte d'Opale

(spécialité : Sciences et Technologies de l'Information et de la communication. Traitement du signal et des images)

par

Josky AÏZAN

Composition du jury

Président :	**Leonard Todjihoundé** Professeur titulaire à l'Université d'Abomey-Calavi, Bénin
Rapporteurs :	**Pierre Gouton** Professeur titulaire à l'Université de Bourgogne Franche Comte, France **Antonio Pinti** Maître de conférences HDR à l'Université Polytechnique Hauts-de-France, France **Marc Kokou Assogba** Maître de conférences HDR à l'Université d'Abomey-Calavi, Bénin
Examinateur :	**Jean Marie Dembele** Maître de conférences HDR à l'Université Gaston Berger, Sénégal
Membre :	**Michel Dossou** Maître de conférences HDR à l'Université d'Abomey-Calavi, Bénin
Directeurs de thèse :	**Cina Motamed** Professeur titulaire à l'Université du Littoral Côte d'Opale, France **Eugène C. Ezin** Professeur titulaire à l'Université d'Abomey-Calavi, Bénin

Unité de Recherche d'Informatique et Sciences Appliquées
Laboratoire d'Informatique, Signal et Image de la Côte d'Opale

Remerciements

Ce travail est le résultat d'une convention de thèse en cotutelle entre l'Université d'Abomey-Calavi et l'Université du Littoral Côte d'Opale. Il n'aurait jamais existé sans le soutien moral, intellectuel et financier de personnes auxquelles je voudrais exprimer ma profonde gratitude.

Je voudrais tout d'abord adresser mes sincères remerciements à mes directeurs de thèse le Professeur Cina MOTAMED et le Professeur Eugène C. EZIN. Vous m'avez fait vivre une très belle expérience de recherche par votre encadrement, votre soutien et vos encouragements.

Je tiens à témoigner ma reconnaissance aux rapporteurs de ma thèse, dont les remarques et suggestions permettront d'améliorer la qualité de ce travail.

Je remercie également le personnel et le corps enseignant de l'IMSP (l'Institut de Mathématiques et de Sciences Physiques) et particulièrement son Directeur le Professeur Léonard TODJIHOUNDE pour son accompagnement et son soutien.

Mes remerciements vont à l'endroit de toute l'équipe du Laboratoire d'Informatique Signal et Image de la Côte d'opale (LISIC) particulièrement son Directeur et Madame Gaëlle Compiègne. Vous m'avez accueilli dans vos locaux et offert de meilleurs conditions de travail lors de mes différents séjours en France.

A mes collègues de la Direction Générale du Trésor et de la Comptabilité Publique (DGTCP) et particulièrement à son Directeur Général Monsieur Oumara KARIMOU ASSOUMA j'adresse mes sincères remerciements. Vous m'avez fait confiance et offert des facilités administratives lors de mes différents voyages dans le cadre de mes recherches doctorales.

Je m'en voudrais de ne pas remercier le Professeur Joël TOSSA. Vous avez su me tendre votre main au moment où j'en avais le plus besoin. votre sagesse a été hautement appréciée.

A mon collègue et ami Monsieur Raoul ALABI du ministère de l'économie et des finances je dis merci. Tu as cru en moi et tu m'a beaucoup aidé à commencer cette belle aventure.

Merci enfin à mes parents, mes amis et tous ceux qui de près ou de loin ont contribué à la réalisation de ce travail.

A toi mon épouse et à vous mes enfants
Trouvez en cette thèse le fruit de vos sacrifices.

Table des matières

Table des figures

Liste des tableaux

Glossaire

AAL Ambient Assisted Living. 1, 7, 8, 103, 104

AD Activity Discovery. 2, 17, 19, 20, 22

ADL Activity of Daily Living. 1, 2, 7, 8, 16, 17, 21, 80

AP Activity Prediction. 2, 17

AR Activity Recognition. 2, 17, 19, 22

BADL Basic Activities of Daily Living. 16

BSN Body Sensor Network. 10

CASAS Center of Advanced Studies in Adaptive System. 80, 82, 83

DBN Dynamic Bayesian Network. 18

DD Detection of Deviation. 2, 17

EFA Extended Finite Automata. 20, 21

HaH Health at Home. 1, 2, 8, 9

HCS Home Care Systems. 8, 9

HMM Hidden Markov Model. 18

HSMM Hidden Semi-Markov Model. 18

IADL Instrumental Activities of Daily Living. 16

MIT Massachusetts Institute of Technology. 71, 72, 80, 82, 83

OWL Web Ontology Language. 22

PIR Passive Infrared. 13

RFID Radio-frequency identification. 12, 88

SED systèmes à événement discrets. 89

Introduction générale

1 Contexte et problématique

Les décennies récentes ont été marquées dans le monde, comme dans la plupart des pays d'Afrique, par un accroissement continu du taux de la population vieille (personnes âgées de 65 ans et plus). Les perspectives prévoient un accroissement spectaculaire de ce taux qui passera du simple (9.5%) en 2020 au double à l'horizon 2075 [1].

Cette augmentation du taux de la population vieille entraine une augmentation du ratio de dépendance de la population (ratio des personnes dépendantes par rapport à la population active) qui passera de 101.2% en 2020 à 105.8% à l'horizon 2075 [1]. Dans ces études sociétales, une personne est considérée comme dépendante sur le reste de la population active si son âge est inférieur à 24 ans ou supérieur à 65 ans. La population active étant considérée comme celle dont l'âge est compris 25 et 64 ans.

Les enjeux humains et économiques de cette évolution démographique sont importants pour les années à venir et les établissements de santé et de bien-être actuels ne suffiront pas pour traiter cette proportion de population vieille. Tous les pays doivent de ce fait relever des défis majeurs pour préparer leurs systèmes sociaux et de santé à tirer le meilleur parti de cette mutation démographique. Par conséquent, des solutions alternatives doivent être trouvées et rapidement développées afin d'apporter de l'aide et de l'indépendance aux personnes âgées.

Pour faire face au futur déficit des systèmes de santé, il est important de développer des systèmes d'aide à la vie ambiante : Ambient Assisted Living (AAL), qui permettent le maintien à domicile des personnes âgées. Les AAL dédiés à la surveillance de la santé, également appelés systèmes de santé à domicile : Health at Home (HaH), consistent à maintenir les personnes âgées à la maison aussi longtemps que possible, grâce à une

surveillance automatique de leur vie quotidienne. Dès que, une action dangereuse ou un comportement anormale est détecté, le personnel médical responsable ou la famille est informée.

La surveillance des activités de la vie quotidienne : Activity of Daily Living (ADL), est l'une des principales branches du domaine des HaH. Il vise à fournir au personnel médical des informations très utiles et précises concernant le patient surveillé. Une ADL est par définition, une activité qui est quotidiennement effectuée par une personne (par exemple préparer le repas, faire le ménage, avoir des loisirs, etc.) et dont la surveillance est utile au médecin [2]. Les études actuelles sur les ADL portent principalement sur quatre problématiques à savoir : la découverte d'activités : Activity Discovery (AD), la reconnaissance d'activités : Activity Recognition (AR), la prédiction d'activités : Activity Prediction (AP) et la détection d'anomalies : Detection of Deviation (DD). Dans cette thèse, seuls les AD et les AR sont traités. L'objectif de la AD est de générer un ou plusieurs modèles d'activités plus ou moins formels en étudiant les habitudes des patients au cours d'une période d'apprentissage. L'objectif de la AR est de détecter qu'une activité est effectivement effectuée par une personne lors de sa réalisation. À ce stade, l'acceptation du système proposé par la personne surveillée est un problème majeur. Le système doit être considéré comme non intrusif par le patient tout en donnant des renseignements pertinents au personnel médical.

L'objectif de cette thèse est de proposer à la fois une méthode de découverte d'activité et une méthode de reconnaissance d'activité qui intègrent les questions d'acceptation, de contrainte temporelle entre évènements et de gestion de l'incertitude en utilisant les méthodes de fouille de séquences fréquentes.

2 Contributions de cette thèse

La contribution de cette thèse est résumée par la figure 1

Dans un premier temps, lors de la phase de découverte d'activités, des modèles d'activités sont générés en tenant compte des activités effectuées quotidiennement par l'habitant d'une maison intelligente. La découverte d'activités s'effectue en utilisant une base de

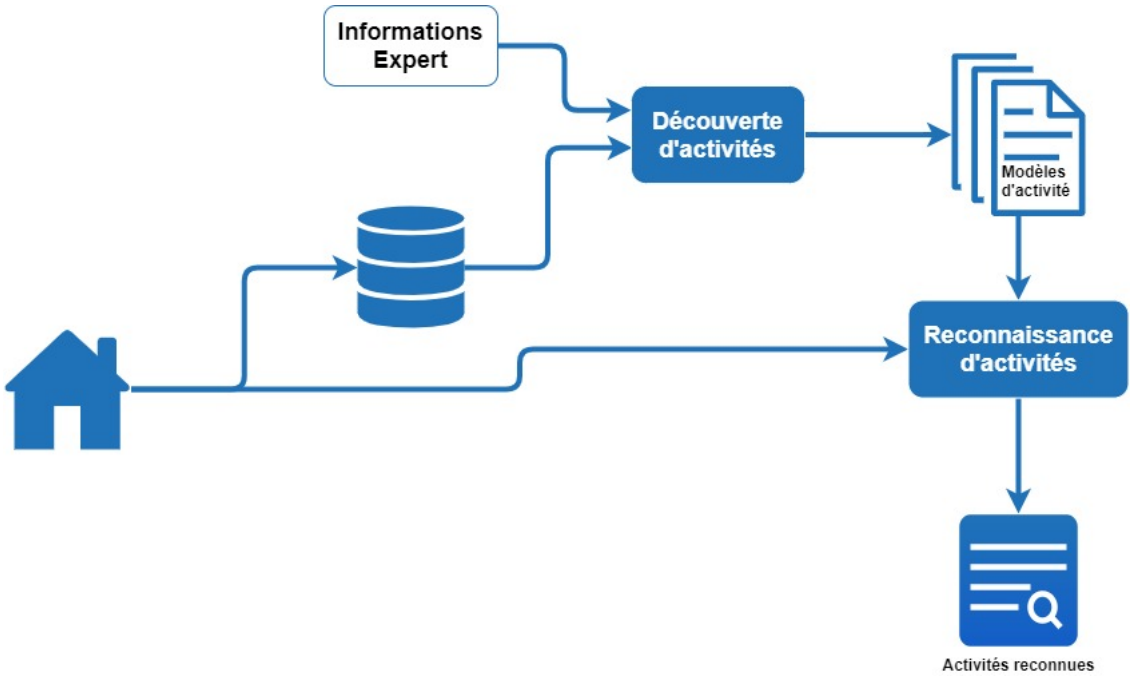

FIGURE 1 – Aperçu des contributions.

données obtenue par enregistrement des événements émis dans la maison intelligente au cours d'une période d'apprentissage. Une expertise supplémentaire est nécessaire afin de produire des modèles efficaces. Dans la majorité des études existantes [3, 4], cette expertise consiste en l'étiquetage des activités exécutées pendant la période d'apprentissage (apprentissage supervisé). Au cours de cette phase de découverte d'activités, nous proposons d'utiliser dans notre étude, contrairement aux approches existantes, une fouille de séquences fréquentes avec gestion des contraintes temporelles pour modéliser les activités. Cette approche nous permet d'intégrer l'ordre des évènements, leur fréquence et une contrainte temporelle inter-événement dans le processus de modélisation des activités [5]. Une extension de cette approche a été également proposée en intégrant la gestion de l'incertitude (fouille incertaine de séquences fréquentes) pour prendre en compte le caractère incertain des données issues des capteurs pour plus de précision dans la reconnaissance [6]. En outre, pour faire face à la volonté et à la vie privée des habitants, nous décidons d'utiliser uniquement des capteurs binaires.

Dans la deuxième phase, un protocole de reconnaissance d'activités a été développé à l'aide des modèles précédemment découverts afin de reconnaître les activités effectuées

par un habitant au cours de sa vie quotidienne. Contrairement aux méthodes de reconnaissance d'activités déjà existantes, nous proposons dans le cadre de notre étude, une méthode basée sur un modèle de classification (forêt aléatoire) qui réalise un mappage entre les séquences découvertes et les activités et utilise ce mappage pour la reconnaissance d'activités [5]. Le mappage se réalise par extraction des caractéristiques temporelles discriminatives de chaque séquence fréquente. Différentes activités peuvent avoir en commun des séquences fréquences et créer un conflit lors de la phase de reconnaissance. Pour résoudre ce conflit, nous proposons une méthode de reconnaissance basée sur l'alignement de séquences et la mesure de score [6]. Cette méthode de reconnaissance se fait en deux phases. La première phase consiste à s'assurer que l'activité à reconnaitre respecte bien la contrainte temporelle inter-événement définie et la seconde phase se charge dévaluer le degré de similitude de l'activité aux modèles découverte lors de la phase de découverte d'activités.

3 Organisation du manuscrit

Le manuscrit est structuré en six chapitres regroupés en trois parties. La première partie est composée de deux chapitres. Le premier chapitre de cette partie dénommé revue de littérature, fait un état de l'art des techniques de surveillance des activités de la vie quotidienne. Les concepts de maison intelligente, de découverte d'activités et de reconnaissance d'activité sont donnés et plusieurs technologies, capteurs et méthodes existants sont présentés dans ce chapitre. Le second chapitre de cette première partie, énonce le problème et présente un résumé des hypothèses et des considérations qui se trouvent dans cette thèse.

La deuxième partie consacrée à la découverte d'activités est constituée de deux chapitres. Le premier chapitre de cette partie présente une approche de découverte d'activités basée sur une fouille de séquences fréquentes. Le deuxième chapitre utilise une extension de la fouille de séquences fréquentes pour la découverte d'activités. Cette extension intègre la gestion de l'incertitude et la durée entre deux évènements pour assurer une meilleur modélisation.

La troisième et dernière partie est consacrée à la reconnaissance d'activités réalisées par l'habitant d'une maison intelligente. Elle est constituée de deux chapitres. Le premier chapitre présente une méthode de reconnaissance basée sur un modèle de classification nommé *Random Forest*. Le deuxième chapitre de cette partie présente une méthode de reconnaissance basée sur l'alignement de séquences et la mesure de score.

Pour terminer ce manuscrit, un résumé des travaux réalisés dans le cadre de cette thèse est présenté et des perspectives pour les travaux futurs sont donnés.

Première partie

Activités quotidiennes : enjeux et objectifs

Chapitre 1

Revue de littérature

Sommaire

1.1 Introduction

En raison de l'augmentation de la population âgée, des alternatives devraient être trouvées par les structures sanitaires pour des soins efficaces aux personnes en situation de dépendance. L'une des solutions les plus répandues est celle qui consiste à maintenir à domicile les personnes dont les pathologies ne sont pas trop graves. À cet effet, il est nécessaire de surveiller ces personnes à l'aide de données extraites d'un ensemble de capteurs adéquat. Pour atteindre un tel objectif, plusieurs approches et modèles basées

sur des études scientifiques et plusieurs considérations sur les capteurs qui peuvent être installés ont été développées.

Ce chapitre vise à examiner l'état de l'art relatif à ce sujet. Dans un premier temps, certaines définitions et termes génériques sont donnés et les concepts de AAL et de maison intelligente sont développés. Ensuite, une description technologique des capteurs existants est donnée de même que leurs utilisations. Enfin, une définition complète des ADL est présentée et les principaux sujets développés autour de ces ADL sont énumérés. Les deux sujets développés dans cette thèse sont présentés en détail et une étude de l'existant réalisée.

1.2 Terminologies et définitions

Le XXe siècle a été marqué par une grande évolution technologique, notamment dans les domaines de l'information et des réseaux électroniques. Des capteurs compacts à faible consommation d'énergie et donnant des informations spécifiques ont vu le jour. cette évolution technologique a donné naissance à la notion de réseau d'ubiquité.

Définition 1. *(Réseau d'ubiquité). D'après [7], le réseau d'ubiquité, aussi connu sous le nom de réseau pervasif, est la distribution d'infrastructures de communication et de technologies sans fil dans un environnement afin de permettre une connectivité continue. Cette capacité est un élément essentiel de l'informatique pervasif.*

Les termes sont interchangeables, avec de légères variations, soit "ubiquité" ou "pervasif", ce qui signifie essentiellement la même chose.

L'équipement d'un environnement adéquat peut nous permettre de résoudre plusieurs problématiques. En effet, l'environnement pervasif pourrait améliorer la consommation d'énergie d'un bâtiment [8, 9] ou améliorer la sécurité à domicile [10]. En outre, les environnements pervasifs peuvent être utilisés pour gérer la vie de la population vieillissante ou handicapée. Le système d'aide à la vie ambiante a donc été envisagé pour surveiller la vie de la personne à tout instant. A ce stade, la surveillance peut se faire n'importe où : à la maison, au travail, dans les supermarchés etc.

Définition 2. *(Aide à la vie ambiante AAL). D'après [11], l'aide à la vie ambiante peut être définie comme "l'utilisation des technologies de l'information et de la communication dans le milieu de vie et de travail quotidien d'une personne pour lui permettre de rester active plus longtemps, de rester socialement connectée et de vivre de façon autonome la vieillesse".*

En considérant le cas de la surveillance de la santé, le principal environnement de vie des personnes vieilles et handicapées est leur propre domicile. Dans la présente thèse, l'objectif est donc de surveiller la santé des personnes à domicile. Dans ce cas précis, les termes systèmes de santé à domicile HaH ou systèmes de soins à domicile : Home Care Systems (HCS) sont utilisés et définis comme suit :

Définition 3. *(Systèmes de soins à domicile). D'après [12], l'objectif des systèmes de soins à domicile est de permettre aux personnes assistées de vivre plus longtemps dans leur environnement préféré (la maison), tout en conservant leur indépendance, même lorsqu'elles ont des handicaps ou des maladies.*

Il est important de noter que les défis que posent HaH (ou HCS) et leurs solutions dépendent fortement des technologies. Il est à la fois porté et limité par les technologies existantes et leur coût [13]. Pour développer le HCS, il est nécessaire d'équiper le domicile des personnes âgées ou handicapées : la notion de maison intelligente est donc apparue [14].

Définition 4. *(Maison intelligente). [15] Une maison intelligente est une résidence équipée d'un réseau de communication, reliant les capteurs, les appareils électroménagers et les appareils, qui peuvent être surveillés, accessibles ou contrôlés à distance [16] et fournissant des services qui répondent aux besoins de ses habitants [17, 18]. En général, le terme "maison intelligente" désigne toute forme de résidence, par exemple une maison autonome, un appartement ou un logement dans un ensemble de logements sociaux.*

Dans cette définition, les capteurs sont des dispositifs utilisés pour détecter l'emplacement des personnes et des objets, ou pour recueillir des données sur les états (température, consommation d'énergie, fenêtres ouvertes etc). Les appareils peuvent être électroniques

(téléphones, téléviseurs, ordinateurs etc) ou électriques (grille-pain, bouilloires, ampoules etc).

Le réseau, reliant et coordonnant ces différentes composants technologiques (capteurs, appareils, appareils électroménagers) est au cœur du concept de la maison intelligente [16, 19]. C'est l'existence de ce réseau qui permet de distinguer une maison intelligente d'une maison simplement équipée de composantes technologiques autonomes et très avancées [20].

La figure 1.1 est une représentation de maison intelligente dans sa définition basique : une maison équipée de capteurs et de déclencheurs connectés via un réseau de communication.

FIGURE 1.1 – Maison intelligente : une maison équipée de plusieurs capteurs et déclencheurs.

Les définitions des termes de base, allant des environnements pervasifs aux maisons intelligentes sont données. La description des technologies utilisées dans les différentes maisons intelligentes est faite dans la prochaine section.

1.3 Technologies de maison intelligente

Les objectifs réalisables, et les moyens de les atteindre, sont fortement liés aux technologies utilisées dans les maisons intelligentes. Les capteurs sont répartis selon le niveau sémantique d'information qu'ils offrent. Nous pouvons définir une échelle à trois niveaux :

- niveau sémantique élevé,
- niveau sémantique moyen,
- faible niveau sémantique.

1.3.1 Niveau sémantique élevé

Certains capteurs donnent des informations complexes et complètes sur la personne à surveiller. Dans cette catégorie, nous pouvons citer les capteurs de signes vitaux portables et les caméras.

(i) Les capteurs de signes vitaux, utilisés dans [21, 22], sont des capteurs portables, également appelés capteurs corporels, qui envoient périodiquement des données sur les signes vitaux de l'utilisateur à un serveur distant. Le tableau 1.1 montre quelques exemples de capteurs corporels typiques. Ces types de capteurs forment ensemble un réseau de capteurs portables : Body Sensor Network (BSN) [23, 24]. Ces capteurs ont des configurations et des infrastructures qui les rendent facile à implanter ou à porter sur le corps humain (Figure 1.2). Certains capteurs peuvent être implantés dans les vêtements de l'utilisateur ; ceux-ci sont connus sous le nom de capteurs textiles portables. Ces capteurs à faible puissance, peuvent communiquer sans fil et surtout surveiller la santé et l'activité de l'utilisateur cible.

(ii) Les caméras sont des capteurs donnant des informations via le traitement d'image et la reconnaissance des mouvements. En effet, comme le précise [25], la reconnaissance du mouvement est à la base de l'estimation de la pose humaine, de la direction du regard (également appelée centre d'attention) et des tâches d'analyse du comportement humain. Le mouvement peut être considéré comme une série de poses dans le temps. Le corps humain est un système articulé de segments rigides reliés par des

joints (comme les modèles utilisés dans [26, 27] supposent). Le mouvement humain est souvent considéré comme une évolution continue de la configuration spatiale de ces segments ou postures corporelles (comme indiqué dans [28] et exploité dans [26, 27]). D'autre part, le regard peut être vu soit comme une ligne dans l'espace 3D ou un cône, soit comme une direction et un angle dans le plan horizontal.

TABLE 1.1 – Exemples de quelques capteurs portables typiques et leur utilisation [46].

Capteur	Signal mesuré	Fonction
ECG	Onde électrocardiogramme	Fréquence cardiaque
PPG	Onde photoplethysmogramme	Pouls de volume sanguin
BP	Pression sanguine en mm Hg	Pression sanguine
EEG	Onde électroencéphalogramme	Anomalie
EMG	Onde électromyographique	Activité musculaire
Accéléromètre	Accélération dans l'espace 3D	Reconnaissance d'activité
Capteur de mouvement	Signal de mouvement	Mouvement de l'utilisateur
Capteur d'activité	Mouvement à 3 axes	Reconnaissance d'activité
Capteur inertiel	Signal de mouvement	Détection de position
Capteur BG	Taux de sucre dans le sang	Détection de diabète
Gyroscopes	Angle de rotation	Orientation corporelle
Thermomètre	Température corporelle en F	Détection de la fièvre
Antenne RF	Onde RF	Détection de position
Détecteur de chute	Signal de mouvement	Détection de chute

1.3.2 Niveau sémantique moyen

Contrairement aux capteurs à niveau sémantique élevé, d'autres capteurs donnent des informations moins complexes mais d'un niveau sémantique non moins important. C'est le cas des capteurs portables ne donnant pas des informations sur les signes vitaux et des microphones.

(i) Les capteurs portables qui ne donnent pas d'information sur les signes vitaux sont des capteurs partiellement portés par l'habitant et qui fournissent des informations binaires au système de santé à domicile. C'est le cas des capteurs basés sur les technologies d'identification par radio-fréquence : Radio-frequency identification (RFID). Le RFID est une technologie utilisée pour identifier les personnes qui portent des

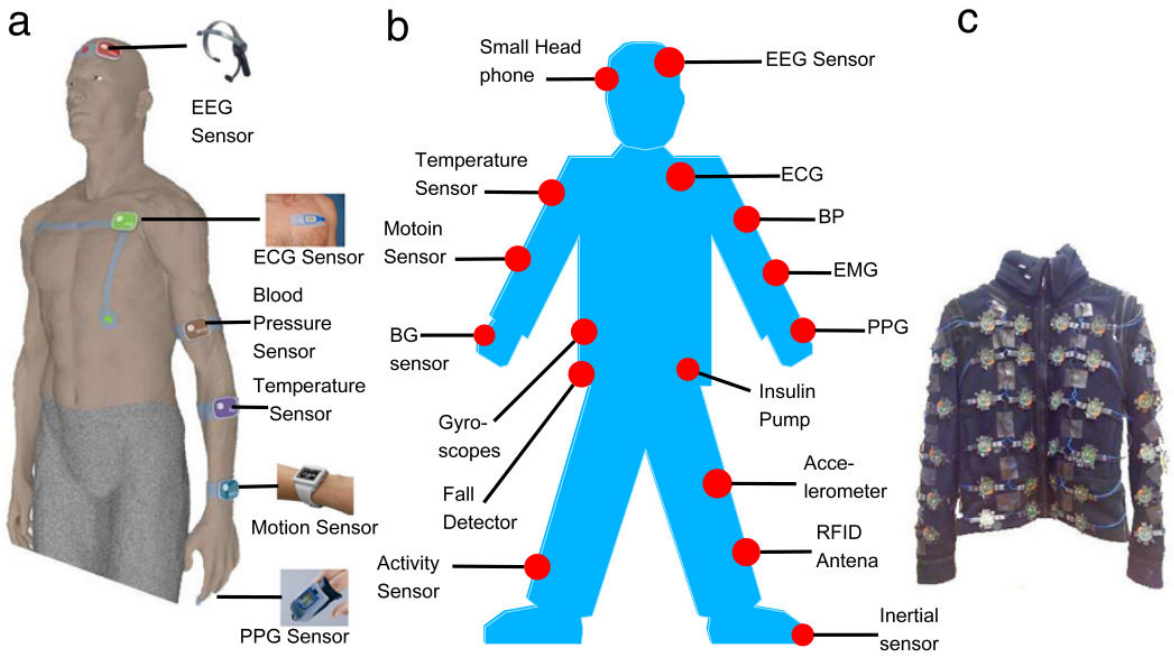

FIGURE 1.2 – (a) Une démonstration de capteurs portables sur un corps humain. (b) Une architecture conceptuelle BSN du système AAL proposé. (c) Un exemple de capteurs portable textile [46].

badges d'identification (ou des étiquettes) ou un lecteur. Si la personne porte une étiquette [29], cette technologie implique des lecteurs qui lisent l'étiquette pour identifier la personne proche du lecteur. Autrement, si la personne porte le lecteur [30], il identifie les objets marqués ou les zones étiquetées proche de la personne.

(ii) Les microphones sont également utilisés comme capteurs donnant des informations complémentaires à d'autres capteurs sous forme de caméras [31] ou RFID [32].

1.3.3 Faible niveau sémantique

Enfin, certaines technologies de capteurs fournissent un faible niveau d'information sémantique. Ces capteurs donnent des valeurs binaires extraites de la détection environnementale. D'après [33], ces capteurs binaires environnementaux ne sont pas apposés sur les personnes effectuant l'activité mais sont placés dans l'environnement qui les entoure. Ces capteurs sont utiles pour des lectures passives sans obligation aux utilisateurs à se

conformer aux règles liées au port ou au transport de capteurs de manière prescrite. Les capteurs binaires environnementaux les plus fréquents sont : les capteurs infrarouges passifs, les capteurs de contact magnétiques, les capteurs de vibration, les capteurs de pression, les capteurs d'écoulement, et les capteurs de température, de lumière, et d'humidité.

Capteurs infrarouges passifs

Les capteurs infrarouges passifs : Passive Infrared (PIR) ou capteurs de mouvement, détectent les rayonnements infrarouges émis par les objets dans leur champ de vision à travers plusieurs points. Si la différence dans le rayonnement détecté entre les point multiples d'un capteur PIR est supérieure à un seuil prédéfini (comme cela se produirait lorsqu'un corps chaud se déplace dans ou hors de la portée du capteur), il génère un message. Un capteur PIR détecte le mouvement de tout objet qui génère de la chaleur, même si l'origine est inorganique.

Capteurs de contact magnétiques

Les capteurs de contact magnétiques, ou capteurs de porte magnétique, sont formés de deux composants : un interrupteur à lames et un aimant. Lorsque la porte est fermée, le composant d'aimant tire l'interrupteur métallique dans le deuxième composant fermé de sorte que le circuit électrique est fermé, changeant ainsi l'état du capteur. Le capteur peut signaler ce changement d'état comme un événement de capteur. Lorsque l'aimant est déplacé en ouvrant la porte, le ressort enclenche l'interrupteur en position ouverte. Cela coupe le courant et ferme le relais, provoquant une nouvelle fois un changement d'état du capteur qui peut être signalé comme un événement de capteur. Cette fonctionnalité est utile pour détecter si les portes, fenêtres, tiroirs ou armoires sont ouverts ou fermés.

Capteurs de vibration

Les capteurs de vibration sont souvent attachés aux éléments ou placés sur les surfaces afin de détecter les interactions avec l'objet correspondant. Certains capteurs sont conçus pour être sensibles à la fois aux vibrations (accélération dynamique) et à l'inclinaison

(accélération statique). Bien qu'ils puissent être utiles pour générer des événements lorsque l'objet auquel ils sont attachés est manipulé, ils peuvent également générer des événements quand ils sont accidentellement heurtés ou lorsque les surfaces voisines tremblent.

Capteurs de pression

Les capteurs de pression sont utilisés pour la surveillance des activités dans un environnement particulier. Les capteurs tactiles sont sensibles au toucher, à la force ou à la pression. Ces capteurs de pression détectent et mesurent les interactions entre un individu et une surface de contact. La force combinée peut être comparée à une valeur seuil pour noter qu'il y a un objet en contact avec le capteur. Des capteurs de pression peuvent être placés sur ou sous des chaises, des tapis de porte, des planchers et des lits pour surveiller l'emplacement et la répartition du poids d'un individu dans l'espace.

Capteurs d'écoulement

Les capteurs d'écoulement fournissent des lectures indiquant la quantité d'électricité ou d'eau qui a été consommée par un bâtiment particulier pendant une unité de temps. Le compteur calcule la quantité d'électricité ou d'eau actuellement consommée et peut signaler des valeurs instantanées, des valeurs accumulées ou des changements suffisants dans la consommation. La quantité d'électricité ou d'eau actuellement consommée peut être comparée à une valeur seuil pour noter qu'un objet est utilisé.

Capteurs de température, lumière et humidité

Les capteurs de température, les capteurs de lumière et les capteurs d'humidité peuvent être placés dans des environnements pour mesurer la température ambiante, l'éclairage et l'humidité. Ces types de capteurs sont souvent regroupés en un seul paquet. Ces capteurs sont calibrés pour déclarer périodiquement leur état actuel (niveau de luminosité, niveau d'humidité et niveau de température) ou ils signalent leur niveau actuelle lorsqu'il y a un changement suffisamment important de la valeur par rapport au moment précédent.

Une autre approche consiste à classer les technologies de capteurs selon leurs interactions avec l'homme. Dans [34], les capteurs sont regroupés en deux familles : les capteurs

effectuant une détection externe et les capteurs portables. Une autre classification, telle que proposée dans [35], contrôle les événements en tenant compte du sentiment humain vis à vis des capteurs en termes de vie privée. Les capteurs sont donc considérés comme intrusifs ou non intrusifs. La figure 1.3 illustre la différence entre ces deux nouvelles classifications.

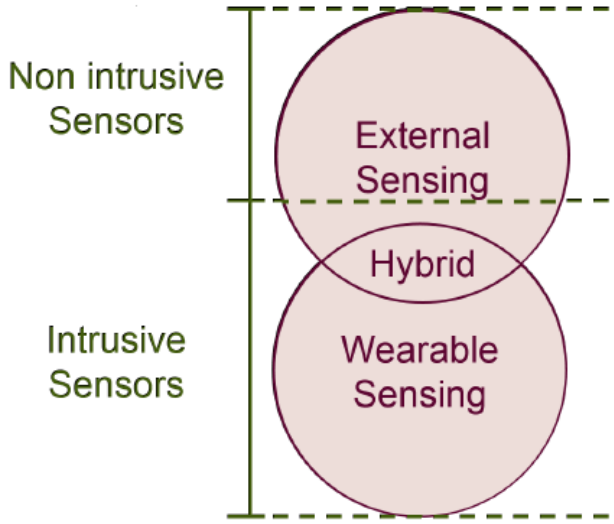

FIGURE 1.3 – Taxonomie des capteurs.

La question de l'acceptation est essentielle pour permettre la surveillance à domicile. Même si les capteurs générant un niveau élevé d'informations sémantiques, comme les caméras, ont une haute performance dans le domaine de la surveillance de la vie quotidienne, ils sont souvent considérés comme trop intrusifs et soulèvent des problèmes d'acceptation par les personnes surveillées [36]. Cette considération concerne également les microphones.

En outre, l'efficacité des capteurs portables dépend fortement de la capacité et de la volonté des patients de les porter tous les jours, et parfois pendant la nuit. Comme dans le cas des caméras, cette technologie de capteur soulève également quelques problèmes d'acceptation et, en outre, n'est parfois pas compatible avec la pathologie des patients à surveiller (par exemple la perte de mémoire).

Pour résumer, deux questions concernant les paramètres humains doivent être prises en considération lors de l'équipement d'une maison intelligente : l'intrusion des capteurs et la capacité des patients à vivre avec. Le tableau 1.2 fait un récapitulatif des capteurs

existants, leur emplacement, leurs niveaux sémantique et leur compatibilité avec les deux problèmes présentés.

TABLE 1.2 – Récapitulatif des capteurs existants.

Type de capteurs	**Position par rapport à l'homme**	**Niveau sémantique**	**Intrusion**	**Capacité du patient**
Capteurs de signes vitaux portables	Portable	Haut	Très intrusif	Parfois incompatible
Caméras	Externe	Haut	Très intrusif	Compatible
Autres capteurs portables	Hybride	Moyen	un peu intrusif	Parfois incompatible
Microphones	Externe	Moyen	Intrusif	Compatible
Capteurs de mouvement	Externe	Bas	Non intrusif	Compatible
Capteurs de portes magnétiques	Externe	Bas	Non intrusif	Compatible
Capteurs de vibrations	Externe	Bas	Non intrusif	Compatible
Capteurs de pression	Externe	Bas	Non intrusif	Compatible
Capteurs d'écoulement	Externe	Bas	Non intrusif	Compatible
Capteurs de température, de lumière et d'humidité	Externe	Bas	Non intrusif	Compatible

Dans cette thèse, les capteurs non intrusifs compatibles avec n'importe quelle pathologie sont préférés pour permettre à notre méthode d'être applicable à la majorité des cas. Par conséquent, dans les approches proposées, seuls les capteurs environnementaux binaires sont utilisés, même s'ils ne fournissent qu'un très faible niveau d'information sémantique.

1.4 Activités de la vie quotidienne : définition et principaux sujets

L'une des façons possibles de prendre soin de la santé des personnes à domicile est de surveiller leurs activités quotidiennes ADL. ADL est défini d'après [37] comme suit :

Définition 5. *(Activité de la vie quotidienne ADL). Tâches effectuées par des personnes dans une journée typique qui permettent la vie autonome. Les activités de base de la vie quotidienne : Basic Activities of Daily Living (BADL) comprennent l'alimentation, l'habillage, l'hygiène et la mobilité. Les activités instrumentales de la vie quotidienne : Instrumental Activities of Daily Living (IADL) comprennent des compétences plus avancées telles que la gestion des finances personnelles, l'utilisation des transports, l'utilisation du téléphone, la cuisine, les tâches ménagères, la lessive et les courses.*

La capacité d'effectuer des activités de la vie quotidienne peut être entravée par une maladie ou un accident entraînant une déficience physique ou mentale. Les professionnels de la réadaptation en soins de santé jouent un rôle important à ce niveau. Ils enseignent des méthodes permettant de maintenir ou de réapprendre des compétences afin d'atteindre le plus haut degré d'indépendance possible.

En outre, les auteurs précisent dans [33] que les activités à surveiller dans une maison intelligente comprennent des activités physiques (ou de base) ainsi que des activités instrumentées (introduites dans [2]). Par conséquent, l'utilisation du thème générique ADL est préférable.

De plus, comme on le considère dans [25, 33, 38, 39] les activités menées par une personne peuvent être décomposées en plusieurs actions. Par exemple, la "cuisson" peut être considérée comme l'ensemble des actions "préparation des pâtes", "préparation d'un plat prêt-à-cuire", "commander un repas sur le net", etc. De plus, les actions peuvent être décrites comme une succession de mouvements élémentaires. Cette décomposition hiérarchique des activités dans les actions et les mouvements est représentée en figure 1.4.

En lisant les articles existants traitant de l'ADL, il apparaît quatre sujets principaux basés sur les études de l'ADL : la découverte d'activités AD [40, 34], la reconnaissance

d'activités AR [41, 34], la prédiction d'activités AP [42, 43] et la détection d'anomalie comportemental DD [44, 21]. Dans ce travail, nous nous concentrons sur l'AD et l'AR.

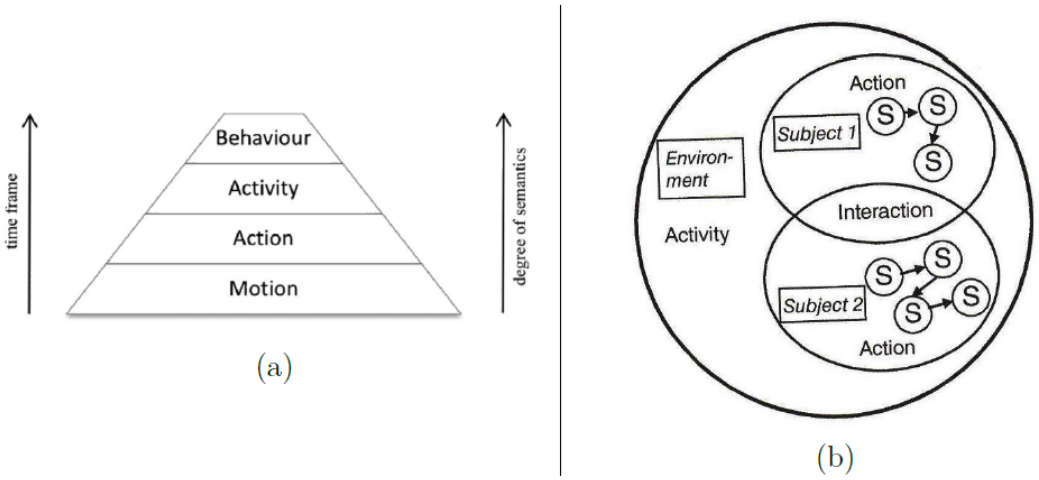

FIGURE 1.4 – (a)Classification des tâches d'analyse du comportement humain [25] (b) Degré de sémantique du comportement humain [33].

1.5 Découverte d'activités

Tout d'abord, les ADL doivent être modélisés. Lorsque le modèle est généré par l'apprentissage, nous utilisons le terme de découverte d'activité.

Définition 6. *(Découverte d'activité AD). [45] La découverte d'activités est un algorithme d'apprentissage supervisé ou non permettant de découvrir les activités dans les données brutes de séquence d'événements de capteur.*

Dans la littérature, une grande variété de méthodes utilisant différentes entrées et sorties peuvent être trouvées. Un bref examen des principales méthodes est développé dans cette section. En cohérence avec le tri effectué dans la section 1.3, ces méthodes sont regroupées en tenant compte du niveau sémantique des capteurs utilisés. Le niveau sémantique des modèles générés est également mis en évidence pour améliorer la compréhension des avantages et des inconvénients liés à chaque méthode.

1.5.1 Entrées de niveau sémantique haut

D'après [46], les auteurs modélisent le comportement humain à l'aide de connaissances d'experts et des capteurs de signes vitaux. En outre, les auteurs sautent le processus de découverte d'ADL en choisissant de ne pas utiliser l'apprentissage des données, mais plutôt des modèles ontologiques comme visible dans le tableau 1.3. Par conséquent, les modèles de comportement humain de cette méthode ont un niveau sémantique très élevé car ils correspondent à des situations très spécifiques. Avec ces modèles, il est possible de détecter directement les situations dangereuses et de réagir rapidement en cas d'urgence. Cependant, ces modèles sont entièrement construits en utilisant des connaissances d'experts, et donc soumis à des erreurs humaines.

D'après [47], les auteurs utilisent des caméras pour détecter l'emplacement d'un l'habitant. Puis, en utilisant des connaissances d'experts, les auteurs relient différentes successions d'emplacement aux activités afin de générer des modèles semi-Markov cachés : Hidden Semi-Markov Model (HSMM) [48]. Un modèle de Markov caché : Hidden Markov Model (HMM) est un modèle stochastique d'un processus avec une partie sous-jacente considérée comme non observable. En outre, un HSMM est un HMM dans lequel une connaissance de durée est ajoutée. La figure 1.5 montre la structure graphique du réseau bayésien dynamique : Dynamic Bayesian Network (DBN) pour HSMM avec une distribution générique de durée de l'état. À chaque tranche de temps, un ensemble de variables $V_t = \{x_t, m_t, y_t\}$ est maintenu où x_t est l'état actuel, m_t est la variable durée de l'état actuel, et y_t est l'observation actuelle. La durée m_t est une variable de comptage propre, qui non seulement spécifie combien de temps l'état actuel durera, mais agit également comme un contexte influençant la façon dont la prochaine tranche de temps $t+1$ sera générée à partir de la tranche de temps actuelle t.

En effectuant des activités connues et recherchées, des auteurs utilisent une distribution de Coxian pour modéliser efficacement l'information sur la durée. Cette information ajoutée au HSMM donne lieu à une nouvelle forme de modèle stochastique : le modèle de semi-Markov caché coxian (CxHSMM).

La méthode de découverte présentée utilise le HSMM donné par un expert comme bases

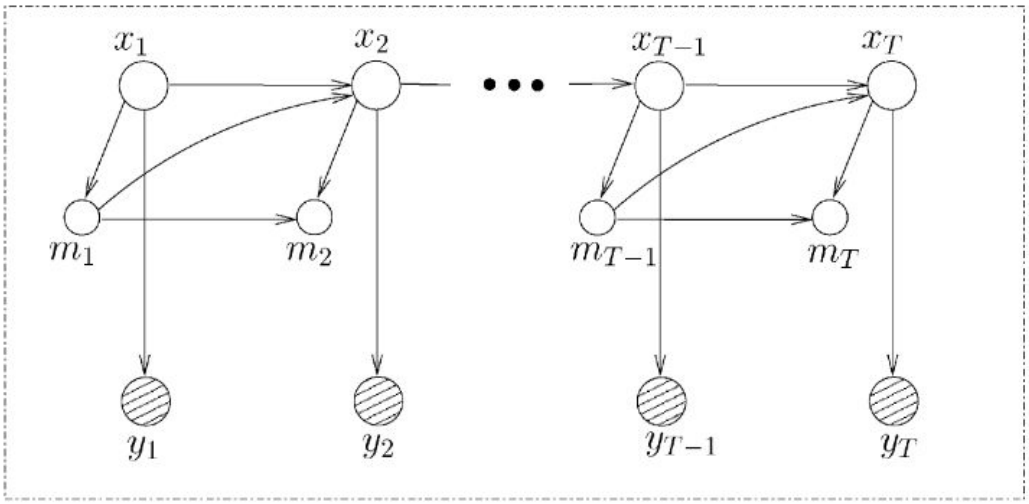

FIGURE 1.5 – représentation DBN d'un HSMM standard. Les nœuds hachurés représentent l'observation [47].

pour l'apprentissage de la durée. Selon les auteurs, l'utilisation des HMM est appropriée et efficace pour apprendre des données séquentielles simples. Il est à noter que, dans ce travail, les informations des caméras sont rapidement transformées en de simples informations de localisation.

1.5.2 Entrées de niveau sémantique moyen

D'après [34], l'AD, appelée étape de l'apprentissage, nécessite initialement un ensemble de données de séries chronologiques d'attributs mesurés sur des personnes effectuant chaque activité. Les séries chronologiques sont divisées en fenêtres temporelle pour appliquer l'extraction des fonctionnalités et filtrer ainsi les informations pertinentes dans les signaux bruts. Par la suite, des méthodes d'apprentissage sont utilisées pour générer un modèle de reconnaissance d'activité à partir du jeu de données des fonctionnalités extraites. De même, les données sont recueillies au cours de la fenêtre temporelle, qui est utilisée pour extraire des fonctionnalités. Cet ensemble de fonctionnalités est évalué dans le modèle d'apprentissage préalablement formé, générant une étiquette d'activité prévue (voir figure 1.6).

Une acquisition de données génériques est également une architecture identifiée pour

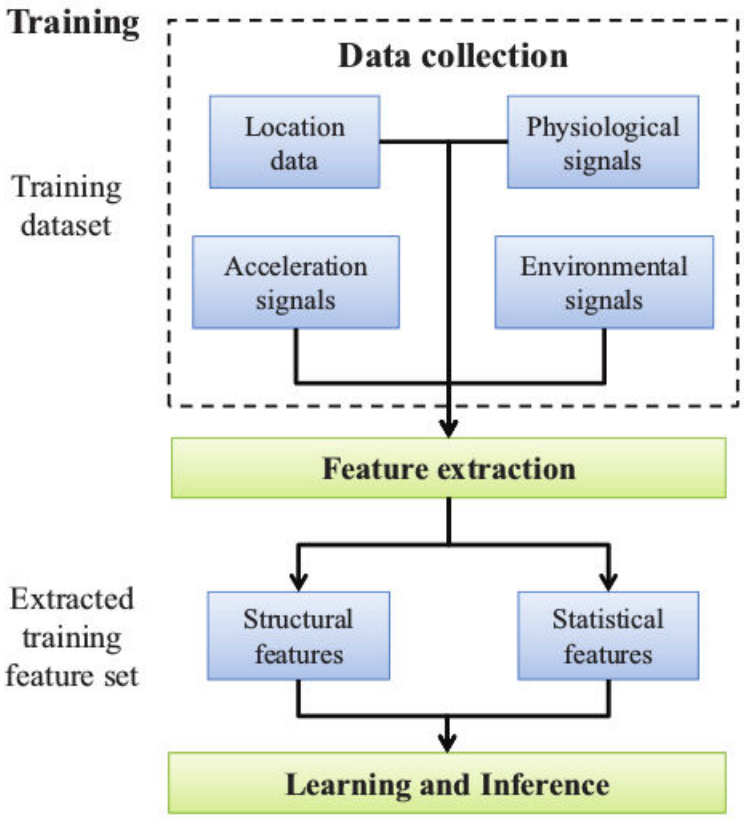

FIGURE 1.6 – Flux de données généraux des systèmes d'apprentissage basés sur des capteurs portables [34].

les systèmes AD et AR, comme le montre la figure 1.7. Dans la première étape, des capteurs portables sont fixés au corps de la personne pour mesurer les attributs d'intérêt tels que le mouvement, l'emplacement, la température, ECG, entre autres. Ces capteurs doivent communiquer avec un dispositif d'intégration, qui peut être un téléphone cellulaire, un PDA, un ordinateur portable ou un système intégré personnalisé.

Les modèles ainsi obtenus peuvent être probabilistes ou non. Cependant, la nécessité de scinder les données enregistrées au cours de la période d'apprentissage à "personne exécutant chaque activité" conduit à l'enregistrement des étiquettes de l'activité effectuée

1.5.3 Entrées de faible niveau de sémantique

Dans [49], les auteurs proposent une méthode de modélisation des habitudes de l'habitant à partir d'un journal d'événements de capteurs binaires. Ces modèles sont extraits

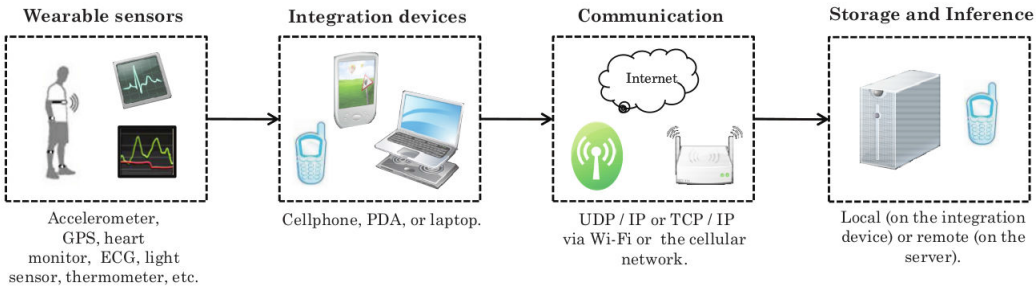

FIGURE 1.7 – Architecture générique d'acquisition de données pour la découverte et la reconnaissance d'activités humaines [34].

par des techniques d'extraction de séquences et modélisés par des automates finis étendus : Extended Finite Automata (EFA). Les habitudes apprises sont ensuite étiquetées par un expert. Cette méthode d'AD, basée sur la méthode d'extraction de modèles [50], est une méthode de découverte de boite noire. Cependant, comme cette méthode d'extraction de modèle distingue chaque modèle récurrent, chaque inversion aléatoire et mineure d'événement crée un nouveau modèle. Le travail d'expert est donc fastidieux s'il s'agit de traiter des données provenant d'une grande maison intelligente. En sortie, une carte globale des activités représentées par un EFA est donnée. Ce modèle de sortie a un niveau sémantique moyen puisque la principale partie intéressante (c'est-à-dire l'étiquetage) se fait a posteriori par l'expert.

Les auteurs présentent dans [33], plusieurs méthodes d'apprentissages automatiques à l'aide de capteurs binaires et d'individus exécutant chaque activité comme entrées. Le classificateur naïf bayésien, modèle Gaussien mixte, modèle de Markov caché, l'arbre de décision, la machine support vectorielle (SVM) et le champ aléatoire conditionnelle (CRF) sont des modèles probabilistes qui peuvent être générés avec ces entrées. Les modèles produits ont des informations sémantiques élevées puisqu'ils sont formés directement avec des données adaptées et étiquetées.

1.6 Reconnaissance d'activités

Les méthodes de reconnaissance d'activité sont des approches basées sur des modèles pour surveiller les personnes. Les modèles utilisés peuvent être donnés par un expert ou obtenus par l'apprentissage (c.-à-d. en appliquant une méthode de AD).

Définition 7. *(Reconnaissance d'activité AR). D'après [33], le domaine de la reconnaissance d'activité AR concerne la question de savoir comment étiqueter les activités à partir d'une perception de l'environnement basée sur les capteurs. Le problème de l'AR est de cartographier une séquence de sorties de capteurs sur une valeur à partir d'un ensemble d'étiquettes d'activités prédéfinies*

Comme pour le AD, une grande variété de méthodes utilisant différentes entrées et sorties peuvent être trouvées dans la littérature. Dans cette section, les méthodes existantes seront classées selon le type de modèle utilisé pour modéliser les activités.

1.6.1 AR utilisant un modèle descriptif lié aux données des capteurs de signes vitaux

Les auteurs adoptent dans [46], le modèle de contexte basé sur l'ontologie [51, 52] pour reconnaître l'activité effectuée ou une situation dangereuse. La figure 1.8 montre le modèle de contexte fondé sur l'ontologie proposé basé sur le langage en ontologie web : Web Ontology Language (OWL). Chaque entité contextuelle a des attributs pour décrire certaines propriétés de base de l'entité. Les entités contextuelles font partie de l'entité mère, comme les caractéristiques, les maladies, les préférences, les ontologies sociales et de santé, font partie de l'ontologie des personnes. Chacune de ces entités a encore des enfants pour les décrire. La relation entre les différentes entités est également démontrée.

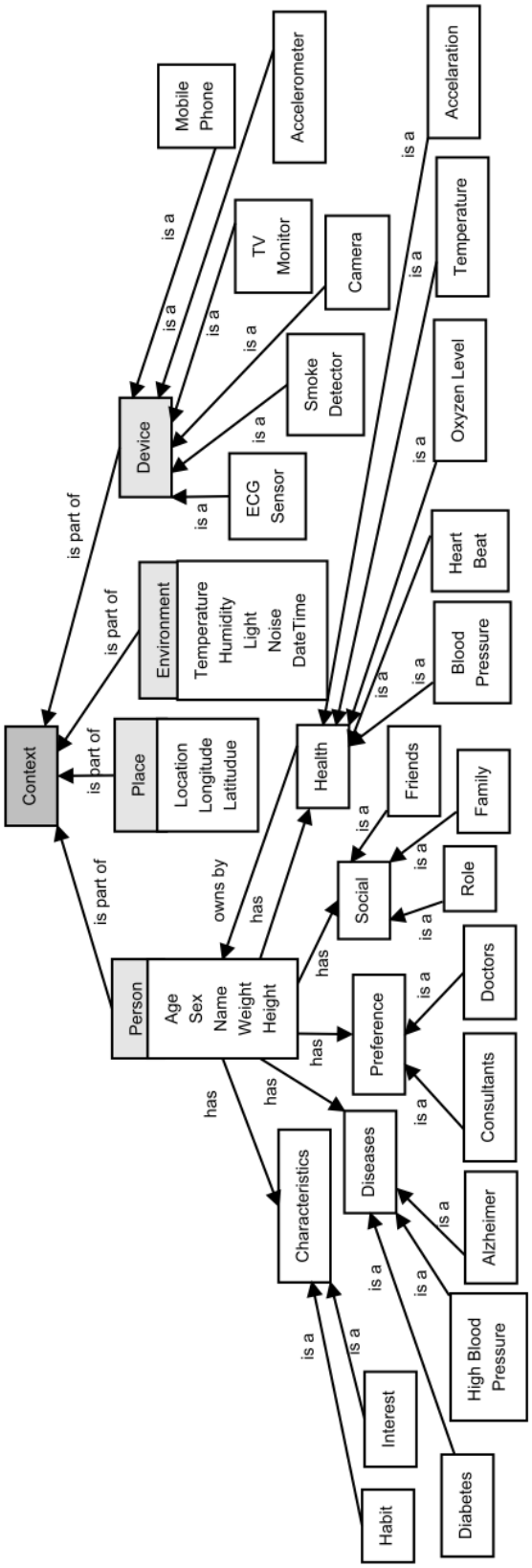

FIGURE 1.8 – Modèle de contexte utilisant OWL. [46]

L'espace contextuelle est décrit en quatre entités principales :

- l'ontologie des personnes est utilisée pour identifier l'utilisateur du système AAL et son profil, ses maladies, ses conditions de santé, ses médecins, ses interactions sociales, etc.
- la place ontologie décrit la position actuelle de l'utilisateur.
- l'ontologie de l'environnement est utilisée pour identifier les conditions des environnements environnants. L'environnement a un certain impact sur la prise de décisions pour les mesures d'assistance.
- l'ontologie de l'appareil contient les détails des capteurs du corps et des dispositifs du système.

Pour détecter une activité ou une situation anormale, une règle de service peut être définie. Par exemple, le tableau 1.3 montre les règles de service de détection d'une crise cardiaque possible en utilisant le modèle d'ontologie.

TABLE 1.3 – Règles de service pour la détection d'une crise cardiaque [46].

Ontology	Instance	Raw data	Context attr.	Values
Person	User X	Profile	Age	≤ 65
Person	User X	Profile	Weight	≤ 80
Person	Disease	Profile	Cardiac patient	Have cardiac issue
Device	ECG sensor	ECG wave	Heart rate	Abnormal
Device	BP sensor	BP readings	Blood pressure	Normal or high
Device	PPG sensor	Sensor readings	O_2 consumption	Low
Device	Audio sensor	Sound wave	Breathing	Irregular
Device	Camera, radar, accelerometer	Video, images, 3D acceleration, motion path	Motion	Tripping or falling or flailing of arms or any rapid motion

En utilisant ces modèles très détaillés, les auteurs peuvent générer des informations très précises sur l'état de santé du patient. Malheureusement, ces modèles experts sont génériques et ne sont pas adaptés à chaque humain vivant chaque pathologie différemment. À titre d'exemple, le cas d'une crise cardiaque pour une personne âgée de moins de 65 ans

n'est pas détecté avec le tableau de description présenté 1.3. Générer un modèle adapté pour chaque humain hors normes devrait être un processus fastidieux et coûteux.

1.6.2 AR à l'aide de modèles basés sur la vision

D'après [25], les auteurs font une excellente revue sur les techniques de vision appliquées à l'analyse du comportement humain pour AAL. Selon les auteurs, on peut voir qu'au niveau de l'estimation du mouvement, de la pose et du regard, plusieurs méthodes atteignent des taux de réussite robustes et élevés.

Les acteurs dans [53] sont en mesure de suivre l'activité du lavage des mains pour aider les personnes âgées atteintes de démence. Plusieurs ordres dans le processus peuvent être correctes, mais pas toutes. Leur système est en mesure d'alerter l'utilisateur si une étape nécessaire est manquante ou l'ordre des actions est incorrect. La vision est utilisée comme seul capteur dans le système développé à deux fins : le suivi de l'emplacement de la main et le suivi des emplacements d'objets spécifiques à l'étape.

En ce qui concerne ce type de reconnaissance d'activité, les travaux réalisés dans [54] se distinguent par la reconnaissance d'activité basée sur l'utilisation d'objets. Les activités sont définies comme des combinaisons d'actions et d'objets. La reconnaissance et le suivi de l'utilisation des objets permet d'inférer les activités humaines. Les modèles d'objets sont acquis automatiquement à partir de la vidéo, tandis que l'identification d'objet est basée sur des étiquettes RFID. Lors de la phase d'apprentissage, l'utilisateur porte un bracelet RFID qui lit les étiquettes RFID attachées aux objets environnants dans un environnement familial. En supposant que l'objet déplacé est toujours l'objet utilisé et qu'un seul objet est déplacé à la fois, le système apprend la relation entre l'image segmentée et l'étiquette RFID active à l'aide d'un réseau bayésien dynamique. Au fur et à mesure que les bras et les mains se déplacent avec les objets, le filtrage de la peau est appliqué à l'avance. Lors de la phase de test, le système fonctionne sans les données RFID car les objets sont reconnus par la détection des fonctionnalités dans la zone segmentée. Ces points clés sont appariés en fonction de la probabilité maximale des points précédemment formés.

D'après [55], la reconnaissance de l'activité est abordée différemment. La silhouette individuelle est obtenue à différentes positions d'un salon. Regroupée en prototypes de

10 à 20, chaque silhouette stocke son centre, sa largeur et sa hauteur et est étiquetée manuellement avec un emplacement. Une méthode d'inférence floue est utilisée pour estimer l'emplacement physique le plus probable des silhouettes d'essai. L'estimation de localisation et les coordonnées précédemment assignées permettent la mesure de la vitesse moyenne, qui est utilisée en plus de l'emplacement afin de reconnaître les activités humaines. Un arbre de décision d'action hiérarchique : Hierarchical Action Decision Tree (HADT) est utilisé pour classer les actions humaines à l'aide de plusieurs niveaux. Au premier niveau, les actions humaines sont classées en fonction de l'emplacement et de la vitesse. Avec les K-moyens, des modèles de fonctionnalités de regroupement sont obtenus ; et les activités de la vie quotidienne, comme la marche ou la visite de la salle de bain, sont reconnues.

Toutes ces méthodes présentées sont efficaces dans leur domaine d'application. Néanmoins, selon [25], à des niveaux plus élevés, en particulier au comportement, il reste encore beaucoup à faire pour obtenir des produits prêts à l'emploi. Pourtant, d'énormes progrès ont été réalisés au cours des dix dernières années. Mais le défi de concevoir et de développer des systèmes stables et généraux persiste, car la plupart des systèmes ne résolvent que des problèmes spécifiques dans des environnements très particuliers.

1.6.3 AR à l'aide de modèles liés à l'information binaire

Les auteurs décrivent dans [56] toutes les activités d'habitant par un seul HMM. Ensuite, les auteurs reconnaissent les activités en appliquant l'algorithme Viterbi [48]. En fait, comme expliqué dans [33], la déduction de la séquence d'étiquettes qui explique le mieux une nouvelle séquence d'observations peut être effectuée efficacement avec cet algorithme. Cette technique de programmation dynamique est couramment utilisée pour les calculs HMM. Si toutes les activités sont modélisés dans le même HMM, l'algorithme de Viterbi génère la séquence d'étiquettes d'activité la plus probable à partir d'une séquence d'observation du capteur. Malheureusement, la complexité du modèle augmente considérablement avec le nombre d'activités et de capteurs. En outre, le modèle utilisé n'a pas de niveaux sémantiques intermédiaires entre les activités et les capteurs et la précision de la reconnaissance n'est pas garantie.

Selon [57], après avoir converti l'information vidéo en événements binaires traduisant la posture humaine, les auteurs présentent un système qui reconnaît un ensemble d'activités modélisés par des HMM. En outre, ils classent les activités par une probabilité qui permet de reconnaître l'activité comme étant celle qui est représentée par le modèle le plus probable.

Cependant, les méthodes précédentes ne peuvent comparer que les modèles liés aux mêmes capteurs et ensembles d'événements. Dans la pratique, les activités sont liées à différents capteurs parce qu'ils sont effectués dans différentes zones d'origine et sont réalisés en utilisant divers équipements dans différents espaces.

1.7 Discussion

Ce chapitre a présenté l'état de l'art des principaux sujets abordés dans cette thèse à savoir : la découverte d'activités et la reconnaissance d'activités.

La découverte d'activités est le premier module dans la réalisation d'un système de surveillance des activités de la vie quotidienne. Elle consiste à générer un ou plusieurs modèles d'activités en étudiant les habitudes du sujet surveillé au cours d'une période d'apprentissage. Les algorithmes de découverte d'activités dépendent fortement du niveau sémantique de leurs entrées (haut, moyen et faible).

La majorité des méthodes existantes modélisent les ADL par des modèles probabilistes. En plus de leur robustesse naturelle à de petites variations, ces modèles ont l'avantage d'être cohérents avec le non-déterminisme humain. Par conséquent, dans cette thèse, une modélisation des ADL humains par modèle probabiliste est préférable. Cependant, la génération des modèles telle que expliquée par les chercheurs, utilise des individus effectuant des activités pour apprendre les probabilités. Il est donc nécessaire d'enregistrer pendant la période d'apprentissage, un journal des activités effectuées. Malheureusement, cette information est en pratique très difficile à obtenir.

D'après [3], le patient surveillé indique quelle activité il effectue. L'efficacité de cette approche est confrontée à la capacité et à la volonté de la personne de déclarer son activité. D'après [4], les experts sont chargés de l'enrichissement de la base de données

en étudiant les journaux de capteurs pendant la phase d'apprentissage. Dans les deux cas, l'étape d'étiquetage est réalisée. C'est la raison pour laquelle, dans les méthodes proposées dans cette thèse, la connaissance des activités réellement exécutées pendant la phase d'apprentissage est nécessaire.

La reconnaissance d'activités est le module qui suit celui de la découverte d'activité dans un système de surveillance des activités de la vie quotidienne. Elle consiste à cartographier une séquence de sorties de capteurs sur une valeur à partir d'un ensemble d'étiquettes d'activités prédéfinies.

Les algorithme de reconnaissances d'activités sont classées selon le type de modèle utilisé pour modéliser les activités (modèles descriptifs, modèles basés sur la vision, modèles basés sur l'information binaire).

Dans la majorité des méthodes présentées pour la reconnaissance d'activités, les modèles probabilistes sont utilisés pour estimer quelle activité est la plus susceptible d'être effectuée. Pour toutes les œuvres existantes de AR, la méthode de reconnaissance est fortement liée aux modèles utilisés pour représenter l'activité. Par conséquent, si un nouveau type de modèles est utilisé lors de la découverte de l'activité, une nouvelle méthode, idéalement basée sur celles existantes, devrait être développée. La grande utilisation de modèles probabilistes conforte la décision de modéliser les activités par des modèles probabilistes dans cette thèse.

1.8 Conclusion

Ce chapitre, après la définition des terminologies utilisée dans le contexte de l'étude, a présenté une revue de littérature des différentes méthodes de découverte et de reconnaissance d'activités dans un contexte de maison intelligence. La mise en évidence des avantages et inconvénients des méthodes existantes nous a permis de justifier le choix des méthodes retenues dans le cadre de cette étude.

Chapitre 2

Énoncé du problème

Sommaire

2.1 Introduction

Une revue de littérature des différents travaux existants dans le domaine étant présentée, l'énoncé du problème de cette thèse est développé dans ce chapitre. Les hypothèses retenues et l'approche proposée pour la découverte et la reconnaissance d'activités sont présentées.

2.2 Objectif général

La revue de littérature présentée au chapitre 1 montre que la surveillance des activités de la vie quotidienne est une solution prometteuse pour gérer le taux croissant de la population dépendante. Plusieurs méthodes plus ou moins intrusives existent et sont applicable sous certaines conditions.

Bien que ces méthodes existantes sont différentes, elles ont des points communs. L'un des point commun est la nécessité de découvrir et de reconnaître les ADL des habitants. Ces deux missions sont les opérations de base à réaliser afin de gérer les activités humaines de la vie quotidienne.

Par conséquent, l'objectif de cette thèse est de développer une approche globale pour découvrir et reconnaître les activités de la vie quotidienne. Cette approche intègre les questions d'acceptation, de contrainte temporelle entre évènements, de gestion de l'incertitude et utilise les méthodes de fouille de séquences fréquentes. Pour y parvenir, trois sources de données sont utilisées :

- le personnel médical donne une liste des ADL à surveiller correspondant aux pathologies des habitants ;
- les capteurs de la maison intelligente donnent des informations sur la vie de l'habitant. Ces données de capteurs peuvent être enregistrées au cours d'une période d'apprentissage ou interprétées directement ;
- un expert peut, au besoin, compléter l'information de base (plan de l'appartement, emplacement des capteurs,...) ou ceux donnés par les deux sources précédemment présentées (le personnel médical ou les capteurs).

En outre, pour être applicable à grande échelle et sur la majorité de la population, quatre points sont considérés, dans cette thèse, comme essentiels :

- la nature humaine doit être considérée dans le choix des modèles ;
- la vie privée du patient est une priorité et son sentiment quant à l'intrusion des capteurs utilisés doit être pris en compte ;
- la compatibilité avec la pathologie du patient doit être considérée

- les méthodes développées doivent être applicables à grande échelle et les informations données par le personnel médical, les capteurs et l'expert doivent être viables et faciles à obtenir.

Ces différents points nous amènent à retenir quatre hypothèses décrites dans la section 2.3

2.3 Hypothèse considérées

Afin d'atteindre les objectifs suivant les quatre points essentiels présentés ci-dessus, des décisions pratiques et éthiques doivent être prises et assumées. Dans cette thèse, quatre hypothèses principales sont considérées.

(i) **Hypothèse 1 : Les activités sont représentées par des modèles probabilistes** Le premier point essentiel à considérer lors du choix de notre modèle est la nature humaine. Cela signifie que l'humain ne peut pas être considéré comme une machine répétant strictement les mêmes mouvements. En effet, le comportement humain est, par nature, non déterministe et peut même être irrationnel. Par conséquent, les modèles déterministes et les méthodes d'identification classiques, plus adaptés aux comportements cartésiens et répétitifs, ne peuvent pas être utilisés dans le problème actuel.

Un humain peut au cours de sa vie, varier sa façon d'effectuer une activité en inversant deux mouvements au cours de l'activité "cuisine", par exemple. En effet, comme une variation moyenne, un humain peut choisir, un jour, de prendre le paquet de pâtes après avoir bouilli l'eau et l'inverse un autre jour. Comme une petite variation, un humain peut ouvrir et fermer inutilement un placard pour préparer le thé parce qu'il a oublié où le thé est stocké.

Pour être compatibles avec ce non-déterminisme humain, les modèles choisis doivent être robustes aux variations. Par conséquent, dans cette thèse comme dans beaucoup de travaux dans la littérature, les ADL sont modélisés en utilisant des modèles probabilistes.

(ii) **Hypothèse 2 : Seuls les capteurs binaires et environnementaux sont utilisés** Le deuxième point essentiel est que la vie privée du patient est une priorité et son sentiment quant à l'intrusion des capteurs utilisés doit être pris en compte. Par conséquent, comme introduit précédemment, les caméras sont rejetées. En effet, ils peuvent être considérés comme trop intrusifs et peuvent être rejetés par les patients [36].

En outre, l'efficacité des capteurs portables dépend fortement de la capacité et de la volonté des patients de les porter tous les jours, et parfois pendant la nuit. Cette technologie de capteur soulève également des problèmes d'acceptation pour la personne surveillée.

De plus, les capteurs portables ne sont pas parfois compatibles aux pathologies des patients à surveiller (par exemple la perte de mémoire). Cette propriété est en contradiction avec notre troisième point essentiel.

L'élimination des capteurs trop intrusifs et portables conduit à l'utiliser seulement les capteurs binaires, tels que des détecteurs de mouvement ou des barrières de porte. Malheureusement, en utilisant uniquement des capteurs binaires, une difficulté due au très faible niveau d'information sémantique détectée existe.

(iii) **Hypothèse 3 : La maison intelligente considérée est occupée par un seul habitant** Selon l'hypothèse 2, seuls les capteurs binaires sont utilisés à la fois pour la AD et la AR. Dans le cas où plusieurs habitants vivent dans le même appartement, il n'est pas possible de distinguer quel habitant génère des événements observés à travers des capteurs binaires. C'est donc nécessaire de faire l'hypothèse qu'un seul habitant vit dans la maison intelligente.

Cette hypothèse est assez restrictive, mais permet de proposer une solution complète pour la AD et la AR qui est basée sur l'utilisation de capteurs binaires uniquement. Une façon d'assouplir cette hypothèse est de considérer que chaque habitant porte un capteur qui permet de l'identifier (par exemple un capteur RFID) et donc de savoir qui a généré quel événement.

Comme nous l'avons vu précédemment, l'utilisation de ce type de capteurs est incompatible avec toutes les pathologies. Cependant, dans le cas de plusieurs habitants

atteints de pathologie compatible, l'utilisation de capteurs portables peut être un compromis acceptable entre l'applicabilité et la vie privée.

La limitation à un seul habitant est supposée dans cette thèse, mais la méthode présentée peut s'appliquer à chaque personne individuellement si une maison intelligente de plusieurs habitants est équipée d'un capteur RFID qui permet d'attribuer chaque événement de capteur à une personne unique.

(iv) **Hypothèse 4 : La connaissance de l'activité réellement effectuée est requise** Comme il est précisé précédemment, plusieurs études utilisent les connaissances des activités exécutées au cours de la phase d'apprentissage pour effectuer une découverte efficace. En effet, ces informations permettent une décomposition facile des données observées pendant la période d'apprentissage et permet d'étiqueter directement les données.

Par conséquent, dans la thèse présentée, la connaissance de l'activité réellement effectuée pendant la période d'apprentissage est nécessaire vu le niveau sémantique bas des entrées.

2.4 Architecture proposée pour découvrir et reconnaitre les activités

L'objectif de cette thèse est de proposer de nouvelles méthodes de découverte et de reconnaissance d'activité suivant les quatre hypothèses présentées ci-dessus. La figure 1 représente l'architecture proposée pour atteindre ces deux objectifs.

2.4.1 Découverte d'activités

Dans l'architecture proposée, la découverte d'activité est effectuée pour générer en sortie les modèles d'activité. Le principe de la découverte d'activités est de générer des modèles à l'aide de données observées au cours d'une période d'apprentissage. Ainsi, une base de données obtenue en enregistrant les événements générés au cours de cette observation est fournie en entrée.

Partant de ces données en entrée (intervention de l'expert et base de données d'événements) la découverte d'activités génère un ensemble de modèles probabilistes conformément à la première hypothèse, chacun représentant une activité.

Le modèle probabiliste utilisé, sa définition et plus généralement la méthode proposée pour la découverte d'activités sont données à la partie II.

2.4.2 Reconnaissance d'activités

La reconnaissance d'activités est une procédure qui permet de reconnaître l'activité effectivement effectuée par l'habitant au cours de sa vie.

La reconnaissance d'activités est effectuée en utilisant une observation directe des événements générés par la maison intelligente. Il ne s'agit pas seulement du dernier événement observé, mais du journal des derniers événements observés.

La deuxième entrée est un ensemble de modèles représentant les activités à surveiller. Comme la reconnaissance d'activités présentée répond à un cadre global contenant la découverte d'activité et la reconnaissance d'activité, les modèles d'activités sont ceux obtenus en appliquant notre méthode de découverte d'activité.

Partant des données en entrée, le protocole reconnaissance d'activités proposé permet de reconnaitre une activité exécutée.

2.5 Discussion

Dans ce chapitre, nous avons précisé le périmètre de notre études autour de quatre hypothèses définies. Ces hypothèses correspondent à quatre objectifs à atteindre à savoir : La prise en compte du non déterminisme du comportement humain, la préservation de la vie privé du sujet surveillé, La prise en compte de la pathologie du sujet surveillé et la fiabilité de la méthode développée.

L'utilisation dans cette thèse des modèles probabilistes pour représenter les activités permet d'intégrer dans la modélisation la nature non déterministe du comportement humain. La limitation aux capteurs binaires permet de tenir compte de la vie privé du sujet surveillé et son sentiment quant à l'intrusion des capteurs. L'utilisation seulement des

capteurs binaires implique la surveillance d'une seule personne à la fois bien que la méthode proposée dans cette thèse soit applicable à la surveillance de plusieurs personnes si les pathologies des personnes surveillées ne sont pas incompatible aux capteurs portables. Pour plus d'efficacité et de fiabilité de la méthode proposée, une connaissance de l'activité est nécessaire à la phase d'apprentissage.

2.6 Conclusion

Dans ce chapitre, nous avons présenté l'objectif général de notre thèse. Ensuite, les différentes hypothèses d'étude ont été énumérées et expliquées pour enfin aboutir à l'énoncé du problème qui peut se résumer comme suit : Développer une architecture globale pour découvrir et reconnaitre les activités de la vie quotidienne d'une personne vivant seul dans une maison intelligente. Cette maison doit être équipée uniquement de capteurs binaires, une connaissance préalable des activités devrait être nécessaire et les activités peuvent être représentées par des modèles probabilistes.

Deuxième partie

Découverte d'activités : Fouille de séquences fréquentes

Chapitre 3

Fouille déterministe de séquences fréquentes

Sommaire

3.1 Introduction

Dans ce chapitre, après la définition des termes utilisés dans le domaine de fouille de séquences fréquentes, nous présenterons l'état de l'art des différents algorithmes de fouille de séquences fréquentes. Enfin, l'algorithme CM-SPADE sera présenté.

3.2 Terminologies et définitions

Cette section présente les définitions des différentes terminologies utilisées dans le domaine de la fouille de séquences fréquentes.

Définition 8. *(Séquences). Considérons un ensemble d'éléments (symboles)* $I = \{i_1, i_2 \cdots i_m\}$, X *un ensemble d'éléments tel que* $X \subseteq I$. *Une séquence est une liste ordonnée* $s = \{I_1, I_2 \cdots I_n\}$ *tel que* $I_k \subseteq I$ *avec* $1 \leq k \leq n$. *Une séquence est dite de taille* k *si elle contient* k *éléments.*

Définition 9. *(Base de séquences). Une base de séquence SDB est une liste de séquences* $SDB = \langle s_1, s_2 \cdots s_p \rangle$ *ayant des identifiants de séquences (SID)* $1, 2 \cdots p$.

Définition 10. *(Sous-séquence). Une séquence* $s_a = \langle A_1, A_2 \cdots A_n \rangle$ *est contenue dans une autre séquence* $s_b = \langle B_1, B_2 \cdots B_m \rangle$ *(noté* $s_a \sqsubseteq s_b$*) si et seulement si il existe des entiers* $1 \leq i_1 < i_2 < \cdots < i_n \leq m$ *tel que* $A_1 \subseteq B_{i1}, A_2 \subseteq B_{i2}, \cdots A_n \subseteq B_{in}$. *Si une séquence* s_a *est contenue dans une séquence* s_b, *on dit que* s_a *est une sous-séquence de* s_b. *Considérons la base de séquences du tableau 3.1 et* S *la séquence ayant pour SID* 1. *Le premier événement* $\{a, b\}$ *de cette séquence se produit au temps* $t = 0$ *et son dernier événement* $\{e\}$ *se produit au temps* $t = 4$. *La séquence* $\langle \{b\}, \{f, g\} \rangle$ *est une sous-séquence de la séquence* S *alors que la séquence* $\langle \{b\}, \{f\}, \{g\} \rangle$ *n'est pas une sous-séquence de* S.

Définition 11. *(Support). Le support ou support absolu d'une séquence* s_a *dans une base de séquences SDB est définie comme le nombre de séquences de la base de séquences contenant* s_a *et est noté* $sup(s_a)$. *Le support relatif de* s_a *noté* $relSup(s_a)$ *est obtenu en divisant le support absolu par le nombre de séquences de la base de séquences. Soit* d_i *une séquence correspondant à la source* i. *Pour une séquence* s *et une source* i, *soit* $X_i(s, d_i)$ *une variable indicatrice , dont la valeur est* 1 *si* s *est une sous-séquence de la séquence* d_i, *et* 0 *dans le cas contraire. Pour toute séquence* s, *son support dans la base* D *est calculé comme suit :*

$$Sup(s, D) = \sum_{i=1}^{p} X_i(s, D) \tag{3.1}$$

Définition 12. *(Fouille de séquences fréquentes). Une fouille de séquences fréquentes est une tâche qui consiste à énumérer toutes les sous-séquences fréquentes d'une base de sé-*

quences. Une séquence s est dite fréquente par rapport à une base de séquences, si et seulement si $sup(s) \geq minsup$ avec minsup un seuil fourni par l'utilisateur [62].

3.3 État de l'art de la fouille déterministe de séquences fréquentes

La fouille de séquences fréquentes est un problème d'énumération (il énumère toutes les séquences qui ont un support supérieur ou égal au seuil indiqué).

Une approche naïve de résolution d'un problème de fouille de séquences fréquentes consiste à calculer le support de toutes les sous séquences possibles de la base des séquences et de renvoyer celles dont le support est supérieur ou égal au seuil indiqué. Cette approche est inefficace parce que le nombre de sous séquences peut être très grand ($2^q - 1$ pour une séquence de q éléments) de ce fait elle est inappropriée à la plupart des problèmes de fouille de séquences fréquentes. Pour éviter d'explorer toutes les sous séquences possibles, des algorithmes ont été mise au point pour la fouille de séquences fréquentes les uns plus efficients que les autres à savoir : le GSP [62], le Spade [117], le Spam [58], le CM-Spam [59], le CM-Spade [59] et le PrefixSpan [60].

Le GSP (Generalized Sequential Pattern) Inspiré du premier algorithme de fouille de séquences fréquentes appelé AprioriAll [50] dont il est une version améliorée, l'algorithme GSP utilise la représentation horizontale de la base des séquences voir tableau 3.1 et le parcours en largeur *(breadth-first search)* pour découvrir les sous séquences fréquentes. L'algorithme GSP utilise une exploration par niveau afin de découvrir les séquences fréquentes. Pour se faire, il parcourt la base de séquences, pour déterminer et garder en mémoire les séquences fréquentes de taille 1. Ainsi, de façon récursive, l'algorithme GSP explore les séquences fréquentes de taille plus grande. Au cours de cette exploration, l'algorithme utilise les séquences fréquentes de taille k pour générer les potentiels séquences fréquentes de taille $k+1$ puis, parcourt la base de séquences pour déterminer et garder en mémoire les séquences fréquentes de taille $k+1$. l'algorithme GSP répète ce processus

pour générer les séquences fréquentes de taille 2, 3 etc. jusqu'à ce qu'il n'y ait plus de séquences à générer. Les limites de l'algorithme GSP se présentent comme suit :

- le parcours répété de la base de séquences pour la détermination des séquences fréquentes peut être couteux pour les bases de grande taille
- la génération de séquences qui n'existe pas dans la base de séquences
- le maintien en mémoire de séquences à cause du parcours en largeur *(breadth-first search)* qui peut induire une importante consommation de la mémoire

TABLE 3.1 – Base de séquences.

SID	**Séquences**
1	$\langle(0)\{a,b\},(1)\{c\},(2)\{f,g\},(3)\{g\},(4)\{e\}\rangle$
2	$\langle(0)\{a,d\},(1)\{c\},(2)\{b\},(3)\{a,b,e,f\}\rangle$
3	$\langle(0)\{a\},(1)\{b\},(2)\{f,g\},(3)\{e\}\rangle$
4	$\langle(0)\{b\},(1)\{f,g\}(2)\rangle$

Le Spade Le Spade est un algorithme qui est inspiré de l'algorithme Eclat [63] et qui utilise le parcours en profondeur *(depth-first search)*. C'est une alternative à l'algorithme GSP parce qu'il résout les limites de l'utilisation de cet algorithme. Le parcours en profondeur permet de résoudre le problème de maintien en mémoire des sous-séquences que pose le parcours en largeur de l'algorithme GSP. Il utilise la représentation verticale de la base des séquences. La représentation verticale indique les sous-ensembles d'éléments (IDList) dans lesquels un élément donné i apparait dans la base des séquences [64, 58, 59]. La figure 3.1 montre la représentation verticale de la base de données horizontale présentée dans le tableau 3.1. Dans cet exemple, l'IDList de l'élément g indique que g apparait dans les sous-ensembles 3 et 4 de la séquence 1, le sous-ensemble 3 de la séquence 3 et le sous-ensemble 2 de la séquence 4. La représentation verticale permet de résoudre le problème de parcours répété de la base des séquences que pose l'algorithme GSP car la base des séquences est parcourue une seule fois et ceci lors de la construction des IDList. La représentation verticale a deux avantages fondamentaux à savoir :

- le IDList d'une sous-séquence permet de renseigner directement sur le nombre de séquence de la base contenant cette sous-sequence ;
- le IDList d'une sous-séquence S_a obtenu par extension d'une sous séquences S_b et un élément i peut être obtenu par jointure des IDList de Sb et i.

a	
SID	Itemsets
1	1
2	1,4
3	1
4	

b	
SID	Itemsets
1	1
2	3,4
3	2
4	1

c	
SID	Itemsets
1	2
2	2
3	
4	

d	
SID	Itemsets
1	
2	1
3	
4	

e	
SID	Itemsets
1	5
2	4
3	4
4	

f	
SID	Itemsets
1	3
2	4
3	3
4	2

g	
SID	Itemsets
1	3,4
2	
3	3
4	2

FIGURE 3.1 – Base de données verticale.

Le Spam L'algorithme Spam est une optimisation de l'algorithme Spade qui représente les IDList comme des vecteurs binaires. Il est adapté aux bases contenant de longues séquences où la jointure des IDList est coûteuse. L'utilisation des vecteurs binaires réduit considérablement l'utilisation de la mémoire dans le processus de fouille de séquences fréquentes. La représentation en vecteur binaire d'un IDList se fait de la façon suivante. Considérons une base de m séquences, où $Y(i)$ est le nombre d'ensemble de la séquence i. Soit une sous-séquence s_a et $IDList(s_a)$ son IDList. La représentation en vecteur binaire de $IDList(s_a)$ noté $BList(s_a)$ est un vecteur binaire contenant $\sum_{i=1}^{m} Y(i)$ bits, où le bit j représente l'ensemble p de la séquence t de la base de séquences. Le bit j est mis à 1 si $(s_t, p) \in IDList(s_a)$ et 0 dans le cas contraire. Le tableau 3.2 montre la représentation binaire des IDList de la figure 3.1.

Le CM-Spam et le CM-Spade Les algorithmes CM-Spam et CM-Spade sont des améliorations respectives des algorithmes Spam et Spade motivées par le fait que les

Algorithm 1 The pseudocode of CM-Spade

```
1: procedure CM-SPADE(D,minsup)
2:     for all d ∈ D do
3:         create V(D)
4:         identify F1 the list of frequent items
5:     end for
6:     ENUMERATE(F1)
7: end procedure

8: procedure ENUMERATE(an equivalence class F)
9:     for all pattern A_i ∈ F do
10:        Output A_i
11:        T_i ← φ
12:        for all pattern A_j ∈ F, with j ≥ i do
13:            R ← MergePatterns(A_i, A_j)
14:            if Prune(R) = FALSE then
15:                if sup(R) ≥ minsup then
16:                    T_i ← T_i ∪ {R}
17:                end if
18:            end if
19:        end for
20:        ENUMERATE(T_i)
21:    end for
22: end procedure
```

TABLE 3.2 – Représentation en vecteur binaire de la base de données verticale.

Elément x	**IDList de x en vecteur binaire**
a	100001001100000
b	100000011010010
c	010000100000000
d	000001000000000
e	000010001000100
f	001000001001001
g	001100000001001

algorithmes Spam et Spade génèrent beaucoup de sous-séquences et les opérations de jointure de liste pour créer les différentes IDList sont coûteuses. Pour réduire le nombre d'opérations de jointure, les algorithmes CM-Spam et CM-Spade utilisent le concept de *Cooccurence pruning* qui consiste à parcourir initialement la base de séquences pour créer une structure appelée *Co-occurence Map (CMAP)* qui stocke toutes les sous-séquences fréquentes de taille 2. Ainsi, pour toute sous-séquence s découverte lors du parcours du domaine de recherche des sous-séquences, si les deux derniers éléments de la sous-séquence s ne forment pas une sous-séquence fréquente de taille 2, la sous séquence s est ignorée sans construction de son IDList et donc sans réaliser l'opération de jointure de IDList.

Le PrefixSpan Le PrefixSpan est un algorithme basé sur le parcours en profondeur *(depth-first search)* et une représentation verticale de la base de séquences qui utilise l'approche de *pattern-growth*. Il est inspiré de l'algorithme FPGrowth [61] et répond aux limites des précédentes qui génèrent des sous-séquences n'appartenant pas à la base des séquences. L'algorithme utilise le parcours en profondeur (depth-first search). Il commence par parcourir la base de séquences initiale afin de déterminer toutes les sous séquences fréquentes de taille 1. Ensuite, l'algorithme utilise ces sous-séquences fréquentes de taille 1 pour réaliser le parcours en profondeur. Durant ce parcours, pour toute sous-séquence s de taille k, une base de données projetée de la sous-séquence s est créée. La base de données projetée est parcourue pour rechercher les éléments à ajouter à la sous-séquence s de taille k pour construire la sous-séquence de taille $k+1$. Ce processus est répété récursivement en utilisant le parcours en profondeur afin de déterminer toutes les séquences fréquentes.

L'avantage de cet algorithme est qu'il permet d'explorer uniquement les sous-séquences qui apparaissent dans la base de séquences contrairement aux autres algorithmes. L'inconvenant de cet algorithme est qu'il peut être coûteux de parcourir la base de séquences de façon répétitive et de créer des projections, en terme de temps d'exécution. Aussi, en termes d'utilisation de la mémoire, la création de projection de base de séquences peut être couteuse car dans la plupart des cas, elle requiert la copie de l'ensemble de la base à chaque projection de base.

3.4 Discussion

Nous avons présenté un état de l'art des algorithmes de fouille de séquences fréquentes dans un contexte où les données issues des capteurs sont déterministes. La fouille de séquences fréquentes est une technique utilisé à la phase de découverte d'activités et est la toute première étape du processus de reconnaissance d'activités de la vie quotidienne.

Le tableau 3.3 fait un résumé des avantages et inconvénients de chaque algorithme de fouille de séquences fréquente. A partir de ce tableau, nous pouvons conclure que l'algorithme CM-Spade est l'algorithme de fouille de séquences fréquentes le plus efficace [59] car alliant le mieux le temps d'exécution et la gestion de l'espace mémoire.

L'efficacité de l'algorithme CM-Spade justifie le choix de cet algorithme dans nos travaux à la phase de découverte d'activités. La section 3.5 présente en détail l'algorithme CM-Spade.

3.5 Algorithme de fouille de séquences fréquentes CM-Spade

Parmis les algorithmes de fouille de séquences fréquentes présentés dans l'état de l'art, l'algorithme CM-Spade est le plus rapide en terme de temps d'exécution [59]. L'algorithme 1 présente son pseudocode. Il prend en entrée la base de séquences D et le seuil *minsup*. CM-Spade construit dans un premier temps la base de données verticale $V(D)$ et identifie l'ensemble des séquences fréquentes $F1$ de taille 1. Ainsi, la procédure ENU-

MERATE est appelée en prenant en entrée $F1$ comme paramètre. Chaque élément A_i de $F1$ est une séquence fréquente. L'ensemble T_i, représentant l'ensemble de toutes les séquences fréquentes des extensions de A_i est initialisé à l'ensemble vide. Pour chaque séquence $A_j \in F$ tel que $j \geq i$, la séquence A_i est fusionnée à la séquence A_j pour former une séquence de taille plus grande. Pour chacune de ces séquences r, le support est calculé en réalisant une opération de jointure entre les *IdLists* de A_i et A_j. La fonction *Prune* dans [59] utilise l'approche *co-occurrence pruning*. Si la cardinalité *l'IdList* résultat n'est pas inférieure à *minsup*, on retient que r est une séquence fréquente et on l'ajoute alors à T_i. Après que toutes les séquences A_j soient comparées aux A_i, l'ensemble T_i contient la totalité de l'ensemble des séquences fréquentes préfixées de A_i. la procédure ENUMERATE est alors appelée avec T_i pour découvrir les séquences fréquentes de taille plus grande préfixées de A_i. A la fin de toutes les boucles, toutes les séquences fréquentes sont renvoyées.

3.5.1 Expérimentations

Nous présenterons ici les expérimentations réalisées selon l'approche qui consiste à utiliser l'algorithme de fouille déterministe de séquences fréquentes (CM-Spade) lors de la phase de découverte d'activités. Cette présentations des expérimentations sera suivi d'une analyse des résultats obtenus.

Dans le cadre de ces expérimentations, nous avons utilisé la base de données Massachusetts Institute of Technology (MIT). Pour transformer les données réels issues des capteurs en une base de séquences, une phase de pré-traitement est nécessaire. Les expérimentations ont été conduites en utilisant une machine équipée d'un processeur Intel(R) Core(TM)$i7-7500U$ CPU @2.70GHz 2.90GHz, d'une mémoire RAM de 8GB et tournant sous Windows 10. Pour appliquer l'algorithme de fouille de séquences fréquentes sur la base de séquence construite à l'issu de la phase de pré-traitement, un seuil de 0.5 a été choisi. Ce seuil indique que pour qu'une séquence d'évènements modélise une activité, il faut qu'elle apparaisse dans 50% au moins des séquences d'évènement de l'activité. nous avons choisi ce seuil car c'est la limite acceptable pour qu'une séquence soit dite fréquente dans une base de séquences.

La base de données MIT

La base de données MIT est une collection d'activités humaines sur une période de deux semaine dans deux appartements individuels contenant respectivement 77 et 84 capteurs (voir figure 3.2 pour illustration). Le premier sujet est une femme professionnelle de 30 ans qui habite dans l'appartement présenté à la figure 3.2(a) et qui passe son temps libre à la maison. Le second sujet quant à lui est une femme de 80 ans qui passe la plus part de son temps à la maison et vie dans l'appartement présenté à la figure 3.2(b). les capteurs sont installés sur les objets quotidiens tels que les tiroirs, réfrigérateurs, etc. afin d'enregistrer les évènements ouverture/fermeture (évènement d'activation/désactivation) au moment où les sujets effectuent les activités quotidiennes. Les activités sont étiquetées en 16 différentes classes et le nombre d'occurrences de chaque classe par sujet est indiqué dans le tableau 3.4.

FIGURE 3.2 – (a) Appartement du premier sujet. (b) Appartement du second sujet.

La phase de pré-traitement

Nous verrons ici comment transformer les données issues des capteurs en une base de séquences exploitable par l'algorithme de fouille de séquences fréquentes en y intégrant une contrainte temporelle entre les évènements.

Une activité est une suite d'évènements ordonnée dans le temps. Les évènements sont générés par les capteurs. La décision d'activation d'un événement est liée aux changements d'état (Booléen) du capteur ou lorsque sa valeur numérique change considérablement. Une variation non significative de valeur est considéré comme un bruit et est donc ignoré. Pour illustrer le pré-traitement, nous avons utilisé l'activité *"Washing dishes"* de la base d'exemple présenté dans le tableau 3.5. Dans la phase de pré-traitement comme le montre la figure 3.3, les données issues des capteurs sont converties sous le format $(t)eid$ dans lequel t représente l'horodatage d'activation/désactivation du capteur, eid représente l'identifiant de l'évènement. l'identifiant de l'évènement eid est sous la forme XYZ où X représente l'identifiant du capteur, Y représente l'état du capteur qui peut être 1 si le capteur est activé ou 0 si le capteur est désactivé. Z représente le nombre de fois où le capteur est activé ou désactivé au sein de la même activité.

En considérant la figure 3.3 et l'ordre croissant de l'horodatage d'activation/désactivation du capteur, on obtient la séquence suivant $7011, 13211, 13201, 13212, 13202, 7001, 7012, 7002$ après la phase de pré-traitement.

L'intégration de la contrainte temporelle entre les évènements à cette phase de pré-traitement se fait en s'assurant que le temps entre deux évènements consécutifs σ respecte la contrainte $\tau \leq \sigma \leq T$ avec τ le temps minimum inter-événement et T le temps maximum inter-événement.

En supposant $\tau = 5s$ et $T = 60s$, la séquence construite précédemment sans la prise en compte de la contrainte temporelle devient $7011, 13211, 13201, 13202, 7001, 7012$

3.5.2 Analyse des résultats

L'approche utilisée nous a permis de découvrir 30 séquences fréquentes sur un total de 278 séquences, de longueur variant de 1 à 11 évènements pour le sujet 1 et 39 séquences

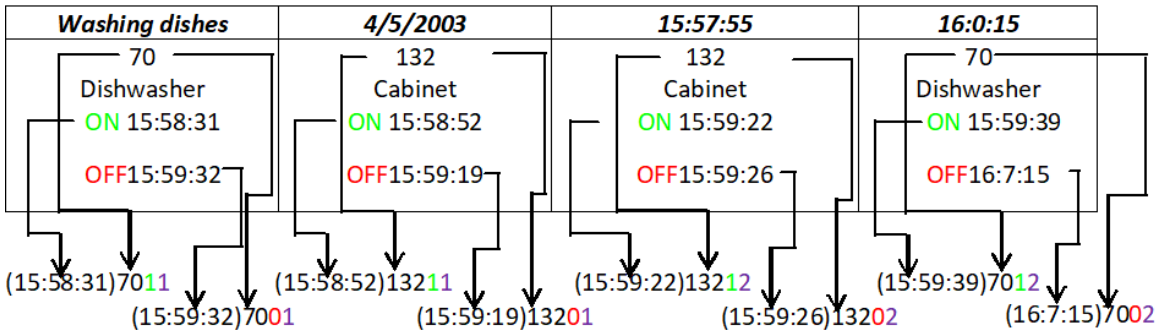

FIGURE 3.3 – Phase de pré-traitement des données issues des capteurs.

fréquentes sur un total de 176, de longueur variant de 1 à 6 évènements pour le sujet 2 après application de l'algorithme de fouille de séquences fréquentes sur la base de séquences. Ce résultat nous permet de dire que l'algorithme de fouille de séquences fréquentes renvoie les activités typiques qui modélisent l'ensemble des activités de la base de séquences. La fréquence d'une séquence est déterminée en évaluant le nombre de séquences de la base de séquences qui contient cette séquence.

3.6 Conclusion

Dans ce chapitre, Nous avons présenté la fouille déterministe de séquences fréquentes. Il commence par la définition des termes utilisés dans le domaine. Ensuite, un état de l'art des algorithmes les plus populaires de fouille déterministe de séquences fréquentes est présenté. Enfin, une discussion reprend à travers un tableau les avantages et inconvénients des algorithmes de l'état de l'art. A la suite de cette discussion, l'algorithme CM-spade est présenté en détail. Cet algorithme est utilisé en expérimentation sur la base de donnée MIT avec une phase de pré-traitement qui intègre la contrainte temporelle inter-événement et les résultats analysés.

TABLE 3.3 – Tableau comparatif des algorithmes de fouille déterministe de séquences fréquentes.

Algorithmes	**Avantages**	**Inconvénients**
GSP	– mise en œuvre facile – s'adapte linéairement au nombre de séquences	– parcours multiple de la base des séquences ce qui peut être couteux pour les grandes bases – inefficace pour de longues séquences car induit une grande consommation mémoire
Spade	– parcours unique de la base de séquences ce qui minimise les coûts d'entrée/sortie – deux fois plus rapide que le GSP	– pas adapté aux longues séquences
Spam	– adapté aux grandes bases – gère efficacement les longues séquences	– gère inefficacement la mémoire
PrefixSpan	– explore uniquement les sous séquences qui apparaissent dans la base de séquences	– couteux en terme de temps d'exécution et d'espace mémoire
CM-Spam	– deux à huit fois plus rapide que le Spam – adapté aux grandes bases – gère efficacement les longues séquences	– gère inefficacement la mémoire
CM-Spade	– parcours unique de la base de séquences ce qui minimise les coûts d'entrée/sortie – gère efficacement la mémoire – deux à huit fois plus rapide que le Spade	– pas adapté aux longues séquences

TABLE 3.4 – Activité étiquetée.

Nombre d'occurrence par classe		
Activité	**Sujet 1**	**Sujet 2**
Preparing dinner	8	14
Preparing lunch	17	20
Listening to music	-	18
Taking medication	-	14
Toileting medication	85	40
Preparing breakfast	14	18
Washing dishes	7	21
Preparing a snack	14	16
Waching TV	-	15
Bathing	18	-
Going out to work	12	-
Dressing	24	-
Grooming	37	-
Preparing a beverage	15	-
Doing laundry	19	-
Cleaning	8	-

TABLE 3.5 – Exemple de données.

Going out to work	***4/1/2003***	***12 :11 :26***	***12 :15 :12***
81	139	140	
Closet	Jewelry box	Door	
12 :12 :29	12 :13 :27	12 :13 :45	
12 :13 :0	12 :13 :35	12 :13 :48	
Toileting	***4/4/2003***	***12 :30 :17***	***12 :31 :10***
100	67		
Toilet Flush	Cabinet		
12 :30 :30	12 :30 :51		
14 :2 :12	12 :30 :54		
Washing dishes	4/5/2003	15 :57 :55	16 :0 :15
70	132	132	70
Dishwasher	Cabinet	Cabinet	Dishwasher
15 :58 :31	15 :58 :52	15 :59 :22	15 :59 :39
15 :59 :32	15 :59 :19	15 :59 :26	16 :7 :15

Chapitre 4

Fouille incertaine de séquences fréquentes

Sommaire

4.1 Introduction

Dans le chapitre précédent, nous avons présenté les algorithmes de fouille de séquences fréquentes dans leur forme classique c'est-à-dire en tenant compte du fait que les données à explorer sont déterministes. Cependant, les données issues d'un large éventail de sources de données sont incertaines. Dans le présent chapitre nous nous intéresserons à la fouille de séquences fréquentes qui intègre la gestion de l'incertitude des données. A la suite d'une

définition des termes utilisés dans le contexte de fouille incertaine de séquences fréquentes, nous présenterons un état de l'art des différent travaux dans ce domaine.

4.2 Terminologies et définitions

Définition 13. *(p-sequence). Une p-sequence est une liste ordonnée d'évènements avec leur degré de confiance respectif.*

Définition 14. *(Base de données probabiliste). Une base de données probabiliste est une collection de p-séquences $D_1^p, D_2^p \cdots D_m^p$ où D_i^p est le p-séquence de la source $i \in S$.*

Définition 15. *(L'univers des possibilités). Encore appelé possible worlds semantics, l'univers des possibilités d'une p-séquence est l'ensemble de combinaisons de tous les évènements qui la compose. En effet, pour chaque évènement e_j de la p-séquence D_i^p, il existe deux sortes d'univers ; un dans lequel l'évènement se produit et l'autre dans lequel il ne se produit pas. L'ensemble de ces deux univers donne l'univers des possibilités. Le tableau 4.3 présente l'univers des possibilités de D_Y^p.*

4.3 État de l'art de fouille incertaine de séquences fréquentes

La fouille de séquences fréquentes utilisant une base de données probabiliste est le cadre le plus rependu de modélisation de l'incertitude. Plusieurs problèmes de fouille de données et de classement ont été étudiés dans ce cadre, à savoir le top-k [65, 66] et l'extraction d'ensembles d'éléments fréquents : Frequent Itemset Mining (FIM) [67, 68, 69, 70].

L'incertitude dans la fouille de séquences fréquentes peut résider à trois niveaux à savoir :

- l'incertitude au niveau source. Dans ce cas, les évènements sont déterministes tandis que les source associées à ces évènements sont incertaines et suivent une distribution de probabilité. Le tableau 4.1 présente une base de données avec incertitude au niveau source.

- l'incertitude au niveau événement. Dans ces scénarios, la source des données est déterministe, mais les événements sont incertains. Ces scénarios pourraient être modélisés comme une séquence d'événements produite par une source, où chaque événement a une certaine probabilité de se produire réellement. Le tableau 4.2 présente un base de données avec incertitude au niveau évènement.

- l'incertitude au niveau horodatage. Dans ce cas, les sources de données sont déterministes de même que les évènements produits par ces sources. Par contre, le moment d'apparition de ces évènements est incertain et suit une distribution de probabilité.

TABLE 4.1 – Base de données avec incertitude au niveau source.

eid	**event**	**W**
e_1	(a, d)	$(X : 0.6)(Y : 0.4)$
e_2	(a)	$(Z : 1.0)$
e_3	(a, b)	$(X : 0.3)(Y : 0.2)(Z : 0.5)$
e_4	(b, c)	$(X : 0.7)(Z : 0.3)$

TABLE 4.2 – Base de données avec incertitude au niveau évènement.

	p-séquence
D_X^p	$(e, h : 0.6)(e, f : 0.3)(f, g : 0.7)$
D_Y^p	$(e, h : 0.4)(e, f : 0.2)$
D_Z^p	$(e : 1.0)(e, f : 0.5)(f, g : 0.3)$

De plus, d'après [71], deux mesures de fréquence, à savoir le *expected support* et le *probabilistic frequentness*, utilisées pour le FIM dans les bases de données probabilistes [68, 70], ont été adaptées à la fouille de séquences fréquentes.

Dans le cadre de notre travail, du fait que nous voulons intégrer la gestion de l'incertitude au niveau des données issues des capteurs, nous utiliserons l'incertitude au niveau évènement comme modèle et le *expected support* pour mesurer la fréquence d'une séquence.

4.4 Évaluation du expected support

Soit $PW(D^p)$ l'univers des possibilités de la base de données D^p. Nous obtenons $PW(D^p)$ en calculant les $PW(D_i^p)$. Chaque $PW(D_i^p)$ est obtenu en considérant les 2^l possibilités avec l la taille de la p-séquence D_i^p.

Dans cette approche, les évènements au niveau des p-séquences sont considérés comme probablement indépendants. Le tableau 4.3 présente $PW(D_Y^p)$.

TABLE 4.3 – Univers des possibilités de D_Y^p.

$\langle\rangle$	$(1-0.4)\times(1-0.2)=0.48$
(e,h)	$(0.4)\times(1-0.2)=0.32$
(e,f)	$(1-0.4)\times(0.2)=0.12$
$(e,h)(e,f)$	$(0.4)\times(0.2)=0.08$

La même méthode est utilisée pour déterminer $PW(D_X^p)$ et $PW(D_Z^p)$ et le tableau 4.4 présente l'univers des possibilités $PW(D^p)$ de la base de p-séquence D^p.

TABLE 4.4 – Univers des possibilités de D^p.

$PW(D_X^p)$	$\{\langle\rangle=0.084\};\{(e,h)=0.126\}$; $\{(e,f)=0.036\};\{(f,g)=0.196\}$; $\{(e,h)(e,f)=0.054\}$; $\{(e,h)(f,g)=0.294\}$; $\{(e,f)(f,g)=0.084\}$; $\{(e,h)(e,f)(f,g)=0.126\}$
$PW(D_Y^p)$	$\{\langle\rangle=0.48\};\{(e,h)=0.32\}$; $\{(e,f)=0.12\};\{(e,h)(e,f)=0.08\}$
$PW(D_Z^p)$	$\{(e)=0.35\};\{(e)(e,f)=0.35\}$; $\{(e)(f,g)=0.15\};\{(e)(e,f)(f,g)=0.15\}$

Un exemple de l'univers des possibilités D^* est présenté dans le tableau 4.5.

Une probabilité de cet univers des possibilités est $Pr(D^*)=0.294\times0.32\times0.35=0.03$ si nous considérons que les p-séquences sont indépendantes du point de vue probabilité, le *Expected Support* est évalué selon l'équation (4.1).

TABLE 4.5 – Un univers des possibilités.

D_X^*	$\{(e,h)(f,g)\}$	0.294
D_Y^*	$\{(e,h)\}$	0.32
D_Z^*	$\{(e)(e,f)\}$	0.35

$$ES(s, D^p) = \sum_{D^* \in PW(D^p)} Pr[D^*] \times Sup(s, D^*) \ . \tag{4.1}$$

$Sup(s, D^*)$ est évalué selon l'équation (3.1) parce que D^* est déterministe. Nous avons $|PW(D_X^p)| \times |PW(D_Y^p)| \times |PW(D_Z^p)| = 8 \times 4 \times 4 = 128$ et donc l'équation (4.1) devient inexploitable lorsque la base de données est grande. Pour résoudre ce problème, le *expected support* est évalué comme suit : Soit $s = (e)(f)$ une séquence et la base de données du tableau 4.2. Pour chaque source X, Y et Z, la probabilité qu'elle contienne s est calculée. Selon le $PW(D_X^p)$ (voir tableau 4.4), la probabilité pour que la source X contienne s est $(0.054 + 0.294 + 0.084 + 0.126) = 0.558$ et la probabilité qu'elle ne contienne pas s est $1 - 0.558 = 0.442$. De façon similaire, les probabilités pour que Y et Z contiennent s sont respectivement 0.08 et 0.65. Pour $i = 0, 1, 2, 3$, l'indépendance des p-séquences est utilisée pour calculer la probabilité pour que la source i contienne s comme le montre le tableau 4.6. Par exemple, la probabilité pour que s soit contenu dans les trois sources est $(0.558 \times 0.08 \times 0.65) = 0.029$. Ainsi $ES(s) = (0 \times 0.142 + \cdots 3 \times 0.029) = 1.228$.

TABLE 4.6 – Distribution de probabilité du support.

No de sources	0	1	2	3
probabilité support	0.142	0.456	0.372	0.029

4.5 Expérimentations et Analyse des résultats

Cette section présente les expérimentations réalisées en suivant l'approche qui consiste à utiliser la fouille incertaine de séquences fréquentes de la phase de découverte d'activités

pour intégrer la gestion de l'incertitude au niveau des données issues des capteurs. Une analyse des différents résultats issus de cette expérimentation sera également présentée.

4.5.1 Expérimentations

Dans le cadre de nos expérimentations, nous avons utilisé la base de données MIT présentée à la sous-section 3.5.1. Pour transformer les données réels issues des capteurs en une base de séquences, une phase de pré-traitement est nécessaire. Les expérimentations ont été conduites en utilisant une machine équipée d'un processeur Intel(R) Core(TM)$i7-7500U$ CPU @2.70GHz 2.90GHz, d'une mémoire RAM de 8GB et tournant sous Windows 10. Pour appliquer l'algorithme de fouille incertaine de séquences fréquentes sur la base de séquence construite à l'issu de la phase de pré-traitement, un seuil de 0.5 est utilisé. Ce seuil indique qu'une séquence peut être considérer comme modèle d'une activité si elle apparaît dans au moins 50% des séquences de cette activité. Cette valeur de seuil est choisie car elle est la limite acceptable pour qu'une séquence soit dite fréquente dans une base de séquences. Pour les besoins l'expérience, nous avons utilisé différentes valeurs de taux de confiance des capteurs (voir figure 4.1).

La phase de pré-traitement

La phase de pré-traitement est presque identique à celle présentée à la sous-section 3.5.1. La différence est que ici pour chaque évènement de la séquence, il faut préciser le taux de confiance du capteur l'ayant généré.

4.5.2 Analyses des résultats

Les expérimentations réalisées sur la base de données choisie montrent selon la figure 4.1 une augmentation du nombre de séquences typiques d'activités pour les taux de confiance croissants. Les valeurs maximale du nombre de séquences typiques sont obtenues pour les taux de confiance supérieurs à 90%. Cette augmentation du nombre de séquences typiques d'activités pour les taux de confiance croissants s'explique par une diminution du taux d'erreurs évalué comme suit : $TauxErreurs = 1 - TauxConfiance$. Cette di-

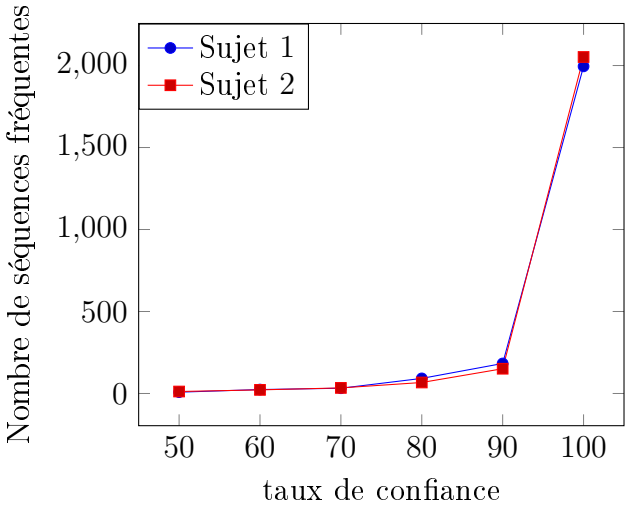

FIGURE 4.1 – Nombre de séquences fréquentes selon le taux de confiance (Base MIT).

minution du taux d'erreurs dans l'évaluation des séquences typiques d'activités induit une augmentation d'informations utiles nécessaires à la modélisation des activités et par conséquent une augmentation de séquences typiques d'activités. Ce résultat nous permet de dire que l'algorithme de fouille incertaine de séquences fréquentes renvoie la part utile des activités typiques qui modélisent l'ensemble des activités de la base de séquences.

4.6 Conclusion

Ce chapitre expose les techniques de fouille de séquences fréquentes intégrant une gestion de l'incertitude. Après une revue des différentes terminologies utilisées dans ce domaine, un état de l'art des différents travaux réalisés a été fait avant de présenter une méthode d'évaluation du *expected support* qui nous permet de mesurer la fréquence d'une séquence fréquente. Les expérimentations réalisées nous ont permis déduire que l'utilisation de la fouille incertaine de séquences fréquentes avec prise en compte de la contrainte temporelle entre événements à la phase de découverte d'activités nous permet d'obtenir des modèles d'activités assez fiables.

Troisième partie

Reconnaissance d'activités

Chapitre 5

Reconnaissance d'activité par utilisation du modèle de forêt aléatoire

Sommaire

5.1 Introduction

Dans ce chapitre, nous allons présenter la forêt aléatoire (Random Forest) en tant que classificateur et son mode de fonctionnement. La différence entre un arbre de décision et une forêt aléatoire sera également abordée dans ce chapitre. A la fin de ce chapitre, mention sera faite sur les avantages et inconvénients des forêts aléatoires.

5.2 Terminologies et définitions

Définition 16. *(Apprentissage automatique ou machine learning). L'apprentissage automatique est un sous-ensemble de l'intelligence artificielle qui permet aux systèmes d'apprendre et de s'améliorer automatiquement à partir de l'expérience sans être explicitement programmés. Il développe des algorithmes d'apprentissage automatique qui construisent un modèle mathématique basé sur des données d'échantillons, connues sous le nom de données d'apprentissage, afin de faire des prédictions ou de prendre des décisions.*

Les algorithmes d'apprentissage automatique sont souvent classés comme étant supervisés, semi supervisé ou non supervisés.

Définition 17. *(Apprentissage supervisé). L'apprentissage supervisé est une tâche de l'apprentissage automatique qui construit un modèle mathématique d'un ensemble de données annotées [72].*

Les données sont connues sous le nom de données d'apprentissage et se composent d'un ensemble d'exemples d'apprentissage. Chaque exemple d'apprentissage a une ou plusieurs entrées et la sortie souhaitée, également connu sous le nom de signal de supervision. Dans le modèle mathématique, chaque exemple d'apprentissage est représenté par un tableau ou un vecteur, parfois appelé vecteur caractéristique, et les données d'apprentissage sont représentées par une matrice. Grâce à l'optimisation itérative d'une fonction objective, les algorithmes d'apprentissage supervisés apprennent une fonction qui peut être utilisée pour prédire la sortie associée aux nouvelles entrées [73]. Une fonction optimale permettra à l'algorithme de déterminer correctement la sortie pour les entrées qui ne faisaient pas partie des données d'apprentissage. Un algorithme qui améliore la précision de ses sorties ou prédictions au fil du temps est dit avoir appris à effectuer cette tâche [75]. Les Types d'apprentissage supervisé sont : l'apprentissage active, la classification et la régression [74].

Définition 18. *(Apprentissage non supervisé). Les algorithmes d'apprentissage non supervisé prennent un ensemble de données ne contenant que des entrées, et trouvent une structure dans les données, comme le regroupement ou la mise en cluster de points de données. Les algorithmes apprennent donc à partir des données de test qui n'ont pas été*

étiquetées, classées ou catégorisées. Au lieu de répondre à un retour d'information, les algorithmes d'apprentissage non supervisés identifient les points communs des données et réagissent en fonction de la présence ou de l'absence de ces points communs dans chaque nouvelle donnée. Le regroupement consiste à l'affectation d'un ensemble d'observations à des sous-ensembles appelés cluster de sorte que les observations au sein d'une même cluster sont similaires selon un ou plusieurs critères prédéfinis, tandis que les observations tirées de différentes clusters sont dissemblables.

Définition 19. *(Apprentissage semi supervisé). L'apprentissage semi-supervisé se situe entre l'apprentissage non supervisé (sans aucune donnée d'apprentissage étiquetée) et l'apprentissage supervisé (avec des données d'apprentissage entièrement étiquetées). Certains des exemples d'apprentissage ne comportent pas d'étiquettes de formation, mais de nombreux chercheurs en apprentissage automatique ont constaté que les données non étiquetées, lorsqu'elles sont utilisées conjointement avec une petite quantité de données étiquetées, peuvent produire une amélioration considérable de la précision de l'apprentissage.*

Définition 20. *(Apprentissage actif). L'apprentissage actif est un cas particulier de l'apprentissage automatique dans lequel un algorithme d'apprentissage peut interroger de façon interactive un utilisateur (ou une autre source d'information) pour étiqueter de nouveaux points de données avec les résultats souhaités [76, 77, 78].*

La source d'information est également appelée enseignant ou oracle. Il existe des situations dans lesquelles les données non étiquetées sont abondantes et où l'étiquetage manuel est coûteux. Dans un tel scénario, les algorithmes d'apprentissage peuvent activement interroger l'utilisateur/enseignant pour obtenir des étiquettes.

Définition 21. *(Classification). En apprentissage automatique et en statistique, la classification est un type d'apprentissage supervisée dans laquelle un programme informatique apprend à partir des données d'entrée et utilise ensuite cet apprentissage pour classer de nouvelles observations [74]. Il existe quatre types de classification à savoir : La classification binaire qui se réfère à la prédiction d'une classe sur deux (l'univers des classes est de taille deux) et la classification multi-classes implique la prédiction d'une classe sur plus de deux classes (l'univers des classes est de taille supérieure à deux). La classification*

multi-étiquettes consiste à prévoir une ou plusieurs classes pour chaque exemple et la classification déséquilibrée se réfère à des tâches de classification où la distribution d'exemples à travers les classes n'est pas égale.

Définition 22. *(Régression). La régression est un algorithme d'apprentissage automatique basé sur l'apprentissage supervisé.*

La régression modélise une valeur de prédiction cible basée sur des variables indépendantes. Elle est principalement utilisée pour déterminer la relation entre les variables et les prévisions [79]. Les différents modèles de régression diffèrent selon le type de relation entre les variables dépendantes et indépendantes qu'ils prennent en compte et le nombre de variables indépendantes utilisées.

Définition 23. *(Méthode d'ensemble). La méthodes d'ensemble est une technique d'apprentissage automatique qui combine plusieurs modèles de base afin de produire un modèle de prédiction optimal [80].*

Dans les modèles d'apprentissage, les principales causes d'erreur sont le le biais (Correspond à l'incapacité du système de représentation à mettre en évidence la relation pertinente entre les attributs d'entrée et l'attribut cible) et la variance (Sensibilité aux petites fluctuations des données d'apprentissage). Les méthodes d'ensemble aident à minimiser ces causes d'erreur. Ces méthodes sont conçues pour améliorer la stabilité et la précision des algorithmes d'apprentissage automatique.

Définition 24. *(Arbre de décision). Un arbre de décision (decision tree) est une structure très utilisée en classification de formes. Son fonctionnement repose sur des heuristiques construites selon des techniques d'apprentissage supervisé [81]. C'est un classificateur interprétable représenté sous forme d'arbre (Voir figure 5.1) tel que :*

- *les nœuds de l'arbre testent les attributs ;*
- *chaque nœud réalise un test portant sur la valeur d'un attribut dont le résultat indique la branche à suivre dans l'arbre ;*
- *il y a une branche pour chaque valeur possible de l'attribut testé ;*
- *les feuilles spécifient les catégories (deux ou plus).*

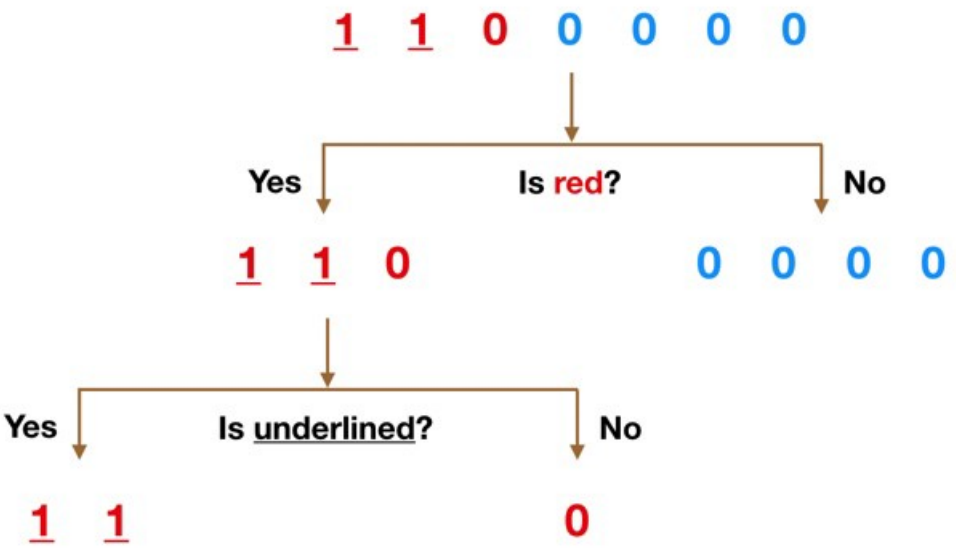

FIGURE 5.1 – Exemple d'arbre de décision.

Définition 25. *(Bagging). Le bagging est un acronyme qui provient de bootstrap aggregating. C'est un algorithme qui combine la technique de bootstrap (tirage avec remise) et celle de aggregating. C'est un principe d'apprentissage d'ensemble de classifieurs qui peut être utilisé avec tout type de classifieur élémentaire, mais dont l'efficacité est démontrée principalement dans le cadre de combinaison d'arbres de décision [82]. L'idée de base est d'entraîner un algorithme d'apprentissage élémentaire sur plusieurs bases d'apprentissage obtenues par tirage avec remise (Voir figure 5.2).*

Définition 26. *(Forêt aléatoire). La forêt aléatoire est un algorithme d'apprentissage supervisé. Il s'agit d'un ensemble d'arbres de décision dépendants de variables caractéristiques aléatoires, combiné pour obtenir une meilleure prédiction.*

La technique des forêts aléatoires modifie la méthode du Bagging appliquée ici aux arbres en ajoutant un critère de dé-corrélation entre ces arbres. L'idée de cette méthode est de réduire la corrélation sans trop augmenter la variance. Le principe consiste à choisir de façon aléatoire un sous-ensemble de variables caractéristiques qui sera considéré à chaque niveau de choix du meilleur nœud de l'arbre.

Formellement, soit $\{\hat{h}(.,\Theta_l) \quad 1 \leq l \leq q\}$ une collection de prédicteurs par arbre, avec

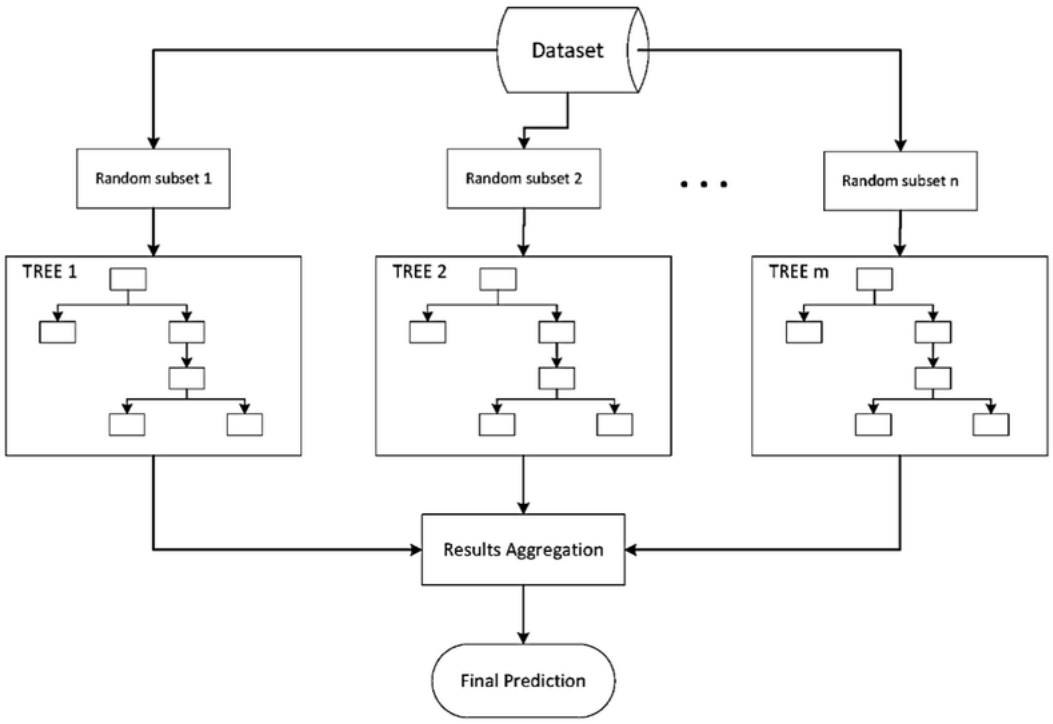

FIGURE 5.2 – Illustration du bagging.

$(\Theta_l)1 \leq l \leq q$ variables aléatoires indépendantes de L_n. Le prédicteur des forêts aléatoires $\hat{h}_{RF}$ est obtenu en agrégeant cette collection d'arbres aléatoires de la façon suivante :

- $\hat{h}_{RF}(X) = \frac{1}{q}\sum_{l=1}^{q} \hat{h}(X, \Theta_l)$ (moyenne des prédictions individuelles des arbres) en régression,
- $\hat{h}_{RF}(X) = \underset{1\leq c\leq L}{argmax} \sum_{l=1}^{q} 1_{\hat{h}(X,\Theta_l)=c}$ (vote majoritaire parmi les prédictions individuelles des arbres) en classification.

Cette définition est illustrée par le schéma de la figure 5.3

5.3 Algorithme forêt aléatoire

Considérons un ensemble d'apprentissage $\Omega = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, p le nombre d'attributs des exemples de X, c le nombre d'éléments distincts de Y. Considérons également Ω_b un *bootstrap* contenant n instances obtenus par tirage avec remise de Ω. Soit $\{M_1, \cdots, M_B\}$ un ensemble de B arbres de décision. Chaque arbre M_b est construit à

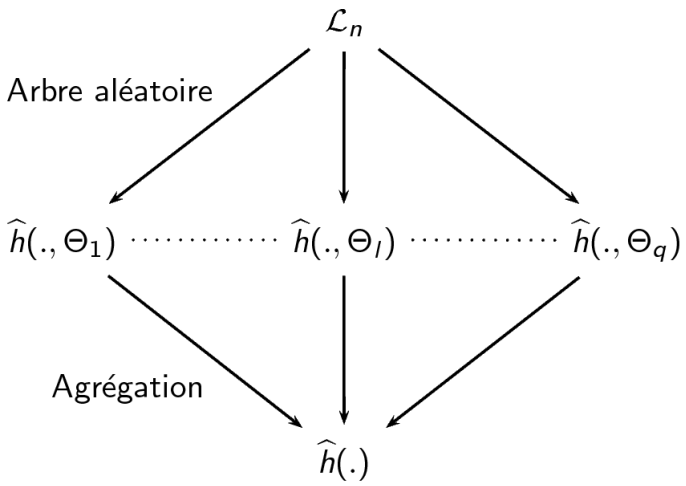

FIGURE 5.3 – Schéma général des forêts aléatoires.

partir de Ω_b. Pour chaque nœud de l'arbre, l'attribut de partitionnement est choisi en considérant un nombre d'attributs m ($m < p$) choisis aléatoirement (parmi les p attributs). Pour classifier une nouvelle instance x, le classificateur des forêts aléatoires effectue un vote de majorité uniformément pondéré des classificateurs de cet ensemble pour l'instance x. L'algorithme 2 illustre ce principe.

5.4 Avantages et inconvénients des forêts aléatoires

L'un des plus grands avantages de la forêt aléatoire est sa polyvalence. Il peut être utilisé à la fois pour des tâches de régression et de classification, et il est également facile de voir l'importance relative qu'il accorde aux caractéristiques en entrée.

La forêt aléatoire est également un algorithme très pratique car les hyperparamètres par défaut qu'elle utilise produisent souvent un bon résultat de prédiction. Comprendre les hyperparamètres est assez simple, et il n'y en a pas beaucoup non plus.

L'un des plus gros problèmes de l'apprentissage automatique est le sur-apprentissage, mais la plupart du temps le classificateur de forêt aléatoire permet de l'éviter.

La forêt aléatoire est un excellent algorithme pour l'apprentissage simple à mettre en œuvre.

L'algorithme est également un excellent choix pour quiconque a besoin de développer

Algorithm 2 The pseudocode of Random Forest

procedure LEARNING(Ω,B)
    $MODELES \leftarrow \phi$
    for $b = 1 \rightarrow B$ **do**
        Générer un échantillon bootstrap Ω_b de taille n à partir de Ω
        for all $Node \in M_b$ **do**
            Sélectionner aléatoirement m attributs parmi les p attributs
            Choisir l'attribut de partitionnement parmi les m est partitionner le nœud
        end for
        $MODELES \leftarrow MODELES \cup \{M_b\}$
    end for
end procedure

procedure PREDICTION(x)
    LEARNING(Ω,B)
    $Y \leftarrow \phi$
    for all $M_b \in MODELES$ **do**
        Soit $\hat{y}_b(x)$ une application de M_b sur x
        $Y \leftarrow Y \cup \{\hat{y}_b(x)\}$
    end for
    for $k = 1 \rightarrow c$ **do**
        $I_k \leftarrow 0$
        for all $\hat{y}_b \in Y$ **do**
            $Test \leftarrow \hat{y}_b = y_k$
            $I_k \leftarrow I_k + Test$
        end for
        $I[y_k] \leftarrow I_k$
    end for
    $\hat{y}_{rf} \leftarrow argmax(I)$
end procedure

rapidement un modèle. En plus, il fournit un assez bon indicateur d'importance accordé aux caractéristiques. Les forêts aléatoires sont également très performantes.

Dans l'ensemble, la forêt aléatoire est un outil (principalement) rapide, simple et flexible, mais non sans certaines limitations.

La principale limitation de la forêt aléatoire est qu'un grand nombre d'arbres peuvent rendre l'algorithme trop lent et inefficace pour les prédictions en temps réel. En général, ces algorithmes sont rapides à former, mais assez lents à créer des prédictions une fois formés. Une prédiction plus précise nécessite plus d'arbres, ce qui se traduit par un modèle plus lent. Dans la plupart des applications du monde réel, l'algorithme de forêt aléatoire est assez rapide, mais il peut certainement y avoir des situations où les performances d'exécution sont importantes et d'autres approches seraient préférées.

La forêt aléatoire est un outil de modélisation prédictive et non un outil descriptif, pour une description des relations dans les données, d'autres approches seraient indiquées.

5.5 Expérimentations et Analyse des résultats

Cette section présente les expérimentations réalisées en suivant l'approche qui consiste à utiliser la fouille déterministe de séquences fréquentes de la phase de découverte d'activités couplée avec le modèle de forêt aléatoire à la phase de reconnaissance d'activité. Une analyse des différents résultats issus de cette expérimentation sera également présentée.

5.5.1 Expérimentations

Dans le cadre de nos expérimentations, nous avons utilisé la base de données MIT présentée à la sous-section 3.5.1. Pour transformer les données réels issues des capteurs en une base de séquences, une phase de pré-traitement est nécessaire. Les expérimentations ont été conduites en utilisant une machine équipée d'un processeur Intel(R) Core(TM)$i7-7500U$ CPU @2.70GHz 2.90GHz, d'une mémoire RAM de 8GB et tournant sous Windows 10. Pour appliquer l'algorithme de fouille incertaine de séquences fréquentes sur la base de séquence construite à l'issu de la phase de pré-traitement, un seuil de 0.5 a été choisi. Ce seuil indique qu'une séquence peut être considérer comme modèle d'une activité si elle

apparaît dans 50% ou plus des séquences de cette activité. Pour les besoins l'expérience, nous avons choisi 80% comme taux de confiance des capteurs. 70% des données sont utilisées pour l'apprentissage (découverte d'activités) et 30% pour le test (reconnaissance d'activités).

Extraction de caractéristiques

Nous présenterons ici la phase d'extraction de caractéristiques appliquée aux séquences issues de a phase de fouille de séquences fréquentes.

En plus des informations sur les évènements générés par les capteurs, cette phase intègre les information temporelles nécessaire à une bonne reconnaissance. Les caractéristiques utilisées se présentent comme suit :

- **Heure de début de l'activité :** L'heure de début des activités est une des caractéristiques distinctives de la reconnaissance d'activité. Selon l'heure de début des activités, on distingue quatre périodes illustrées par la figure 5.4. Ces périodes sont classé comme le montre le tableau 5.1.
- **Durée de l'activité :** Selon leur durée, les activités peuvent être rangées dans quatre classes comme illustré dans la figure 5.5. ces quatre classes are étiquetées comme le montre le tableau 5.2.
- **Densité des évènements :** Le nombre des évènements issus des capteurs pour une activité donnée dépend de la durée et la mobilité. On utilise la densité d'évènement pour mettre en évidence cette caractéristique. Pour calculer la densité d'évènement, le nombre d'évènement de l'activité est divisé par sa durée tel que le montre l'équation (5.1).
- **Activité précédente :** L'activité exécutée précédemment peut fournir des indices de reconnaissance de l'activité courante.

$$Densite = \frac{Nombre\_evenements}{Duree\_activite} \tag{5.1}$$

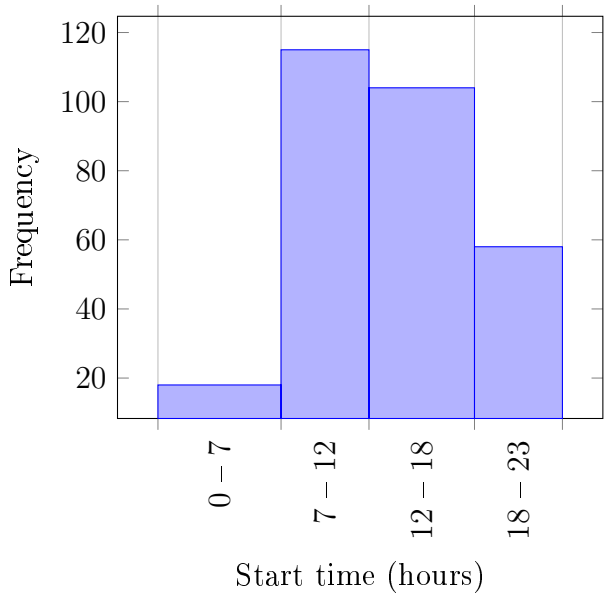

FIGURE 5.4 – Fréquence des activité selon l'heure de début pour le sujet 1.

TABLE 5.1 – Classification des activités selon leur heure de début.

Intervalle d'heure de début	**Classe**
$0 \leq heure < 7$	Nuit
$7 \leq heure \leq 12$	Matin
$12 < heure \leq 18$	Après-midi
$18 < heure < 0$	Soir

TABLE 5.2 – Classification des activités selon leur durée.

Intervalle de temps (minutes)	**Classe**
$duree \leq 5$	Ultra-Court
$5 < duree \leq 15$	Court
$15 < duree \leq 60$	Moyen
$duree > 60$	Long

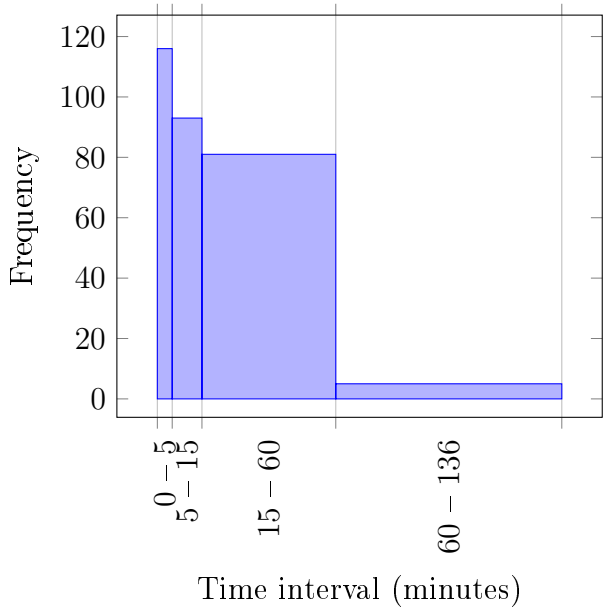

FIGURE 5.5 – Fréquence des activités selon leur durée pour le sujet 1.

5.5.2 Analyses des résultats

Le modèle de forêt aléatoire est utilisé à la phase de reconnaissance d'activités couplée avec l'utilisation d'une fouille déterministe de séquences fréquentes à la phase de découverte d'activités. Pour utiliser le modèle de forêt aléatoire, après la phase de découverte d'activités, une phase d'extraction de caractéristiques intégrant les informations temporelle est introduite. Les expérimentations réalisées nous ont permis d'aboutir à des résultats satisfaisants à savoir un taux de reconnaissance de 98.2% pour le premier sujet et de 95.45% pour le second sujet. Ces taux de reconnaissance sont supérieurs à ceux obtenus avec l'approche proposé par Raeiszadeh dans [89] (voir tableau 5.3).

TABLE 5.3 – Comparaison des résultats.

	Approche	**Résultat**
Méthode Proposée	(SPM+RandomForest)	Sujet 1 : 98.2% Sujet 2 : 95.45%
[89]	(UP-DM+RandomForest)	Sujet 1 : 97.45% Sujet 2 : 91.37%
[3]	(Naive Bayes Classifier)	Sujet 1 : 60.6% Sujet 2 : 41.09%

5.6 Conclusion

Dans ce chapitre nous avons expérimenter l'approche qui consiste à utiliser l'algorithme de fouille de séquences fréquentes à la phase de découverte d'activités et le modèle de forêt aléatoire à la phase de reconnaissance d'activités. Entre la phase de découverte d'activités et la phase de reconnaissance d'activités, une phase d'extractions de caractéristiques a été introduite et intègre les informations temporelles nécessaire à une bonne reconnaissance des activités. Les résultats issus des expérimentations nous montre l'efficacité de notre approche qui améliore de taux de reconnaissance d'activités.

Chapitre 6

Reconnaissance d'activité en utilisant une approche basée sur l'alignement de séquence

Sommaire

6.1 Introduction

Dans ce chapitre, nous aborderons les concepts liés à l'alignement et à la similitude de séquence puis nous présenterons les relations qui existent entre ces deux concepts et qui permettent la reconnaissance d'activités.

6.2 Alignement de séquence

La première étape de comparaison de deux séquences est généralement l'alignement.

Définition 27. *(Alignement de séquence). L'alignement de séquence est un concept de la bioinformatique qui consiste à disposer une séquence test au dessus d'une séquence référence afin de mettre en évidence les points de similitude.*

De façon formelle, soient S_1 et S_2 deux séquences. Un alignement A fait correspondre les séquences S_1 et S_2 respectivement aux séquences S_1' et S_2' qui peuvent contenir des espaces où :

- $|S_1'| = |S_2'|$, et
- La suppression des espaces de S_1' et S_2' (sans changement de l'ordre des caractère résultants) permet d'obtenir S_1 et S_2 respectivement.

Par exemple, considérons $S_1 = \{A, T, A, T, T, G, C, T, A, C, G, T, A, T, A, T, C, A, T\}$ et $S_2 = \{A, T, A, T, A, T, G, C, T, A, C, G, T, A, T, C, A, T\}$. La figure 6.1 montre quelques exemples d'alignement de séquences.

6.2.1 Évaluation des alignements de séquences

Pour être en mesure de comparer les alignement possibles de séquences, nous allons évaluer leur score. Idéalement, on pourrait créer un système d'évaluation qui accorde plus de scores aux alignements qui ordonnent les positions homologues. Un système d'évaluation

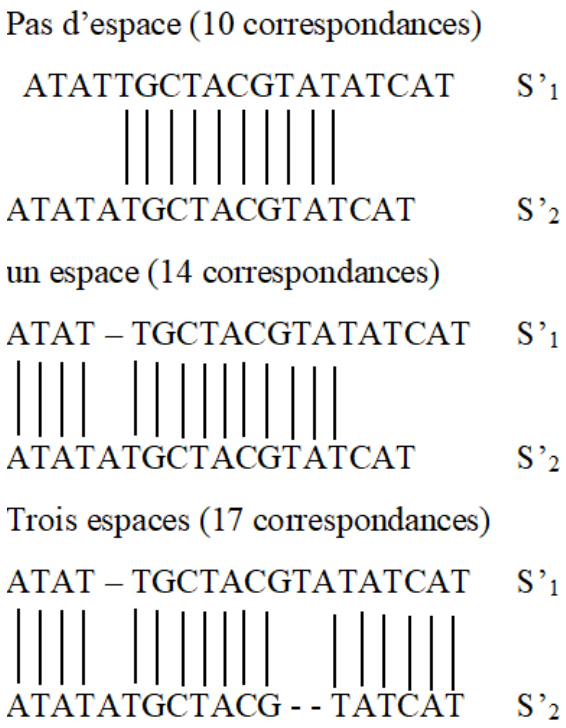

FIGURE 6.1 – Illustration d'alignement.

naïf pourrait consister à compter le nombre de positions correspondantes. Habituellement, les systèmes d'évaluation tiennent compte du nombre et de la taille des espaces. Les alignements sont pénalisés par rapport au nombre et à la taille des espaces qu'ils contiennent. Ainsi, les critères à prendre en compte lors de la création d'un système d'évaluation sont habituellement :

- nombre d'éléments qui correspondent ;
- nombre d'éléments qui ne correspondent pas ;
- nombre d'espaces ;
- taille des espaces.

Différents schémas d'évaluation peuvent découler de ces critère. Par exemple :

- schéma 1 : correspond $+1$ correspond pas 0 création espace -1 extension espace -1
- schéma 2 : correspond $+1$ correspond pas 0 création espace -1 extension espace 0

Le score d'un alignement diffère selon le schéma d'évaluation choisi. Il n'est donc pas

nécessaire de comparer les scores issus de schémas d'évaluation différents. Dès que le système d'évaluation est choisi, l'algorithme d'alignement essaie de créer l'alignement qui obtient le score maximum sous ce schéma.

Définition 28. *(Score d'alignement). Si x et y sont des éléments respectives des séquences S_1 et S_2 ou des espaces, alors $\sigma(x, y)$ désigne le score de l'alignement de x et y. σ est appelé la fonction d'évaluation.*

En se référant à la définition formelle de l'alignement à la section 6.2, le score de l'alignement A est : $\sum_{i=1}^{l} \sigma(S'_1[i], S'_2[i])$, avec $l = |S'_1| = |S'_2|$

6.2.2 Types d'alignement de séquences

On distingue deux type d'alignement de séquence : l'alignement global et l'alignement local.

Définition 29. *(Alignement global). L'alignement global est celui qui fait correspondre les éléments des séquences en tenant compte l'entièreté de leur longueur.*

Définition 30. *(Alignement local). L'alignement local est celui qui fait correspondre les éléments des séquences en s'intéressant aux régions des séquences qui ont le plus de similitude.*

6.2.3 Méthode d'alignement de séquences

Les méthodes d'alignement de séquences peuvent être regroupées en deux grande catégories à savoir : les méthodes d'alignement par paire et les méthodes d'alignement de séquences multiple.

Définition 31. *(Alignement par pair). Les méthodes d'alignement de séquence par paire sont utilisées pour trouver le meilleur alignement (local ou global) de deux séquences.*

Les alignements par paire ne peuvent être utilisés qu'entre deux séquences à la fois, mais ils sont efficaces du point de vue rapidité d'exécution et sont souvent utilisés pour des méthodes qui ne nécessitent pas une haute précision.

Définition 32. *(Alignement multiple). L'alignement de séquences multiples est une extension de l'alignement par paires pour incorporer plus de deux séquences à la fois.*

Plusieurs méthodes d'alignement tentent d'aligner toutes les séquences d'un ensemble donné. Des alignements multiples sont souvent utilisés pour identifier des régions de séquence à travers un groupe de séquences supposées être liées. Les alignements de séquences multiples sont difficiles à produire par calcul et la plupart des formulations du problème conduisent à des problèmes d'optimisation combinatoire NP-complets [83, 84]. Néanmoins, l'utilité de ces alignements en bioinformatique a conduit au développement d'une variété de méthodes appropriées pour aligner trois séquences ou plus.

6.3 Similitude de séquences et approches de mesure

Définition 33. *(Similitude de séquences). La similitude de séquence est un concept issu de la bioinformatique et de l'informatique. Elle désigne un nombre qui montre à quel point deux séquences sont similaires. La similitude de séquence est parfois, mais pas toujours, définie par la distance de séquence : plus la distance est petite, plus les séquences sont similaires.*

6.3.1 Approches basée sur la distance d'édition

Dans cette approche, on calcul le nombre minimum d'opérations nécessaire pour transformer une séquence en une autre. Plus est le nombre d'opérations, moins est la similitude entre les deux séquences. Un point à noter, dans ce cas, chaque élément de la séquence a la même importance.

Définition 34. *(Distance de Hamming). La distance de Hamming est égale au nombre de positions auxquelles les éléments correspondants sont différents. En d'autres termes, elle mesure le nombre minimum de substitutions nécessaires pour transformer une séquence en une autre. Elle s'applique aux séquences de même longueur.*

La distance de Hamming est largement utilisée en télécommunication pour estimer les erreurs en comptant le nombre de bit inversé dans les mots binaires de longueur fixe, c'est

pourquoi on l'appelle aussi la distance du signal.

Définition 35. *(Distance de Levenshtein). La distance de Levenshtein communément appelée distance d'édition, est le nombre de modifications nécessaire pour transformer une séquence en une autre.*

Les transformations autorisées sont l'insertion (ajout d'un nouveau élément), la suppression (la suppression d'un élément) et la substitution (le remplacement d'un élément par un autre) [86]. En effectuant ces trois opérations, l'algorithme essaie de modifier la première séquence pour correspondre à la seconde.

Définition 36. *(Distance de Jaro-Winkler). La distance de Jaro-Winkler mesure la similarité entre deux séquences. Il s'agit d'une variante proposée en 1999 par William E. Winkler [87], découlant de la distance de Jaro [88] qui est principalement utilisée dans la détection de doublons.*

Le résultat est normalisé de façon à avoir une mesure entre 0 et 1, donc zéro représente l'absence de similarité et 1, l'égalité des chaines comparées.

Cette mesure est particulièrement adaptée au traitement de chaînes courtes comme des noms ou des mots de passe.

Distance de Jaro La distance de Jaro entre les séquences s_1 et s_2 est définie par : $d_j = \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m})$ où :

- $|s_i|$ est la longueur de la séquence s_i ;
- m est le nombre d'éléments correspondants ;
- t est le nombre de transpositions.

Deux éléments identiques de s_1 et de s_2 sont considérés comme correspondants si leur éloignement (i.e. la différence entre leurs positions dans leurs séquences respectives) ne dépasse pas : $\lfloor \frac{max(|s_1|,|s_2|)}{2} \rfloor - 1$

Le nombre de transpositions est obtenu en comparant le i-ème élément correspondant de s_1 avec le i-ème élément correspondant de s_2. Le nombre de fois où ces éléments sont différents, divisé par deux, donne le nombre de transpositions.

Distance de Jaro-Winkler La méthode introduite par Winkler utilise un coefficient de préfixe p qui favorise les séquences commençant par un préfixe de longueur ℓ (avec $\ell \leq 4$). En considérant deux séquences s_1 et s_2, leur distance de Jaro-Winkler d_w est :

$d_w = d_j + (\ell p(1 - d_j))$ où :

- d_j est la distance de Jaro entre s_1 et s_2
- ℓ est la longueur du préfixe commun (maximum 4 éléments)
- p est un coefficient qui permet de favoriser les séquences avec un préfixe commun. Winkler propose pour valeur $p = 0.1$

Exemple Soit deux séquences s_1 DIXON et s_2 DICKSONX. Nous allons dresser leur table de correspondance. Ici, l'éloignement maximal vaut $\frac{8}{2} - 1 = 3$. Dans les cases jaunes de la table ci-dessous, on inscrira donc 1 lorsque les éléments sont identiques (il y a correspondance) et 0 sinon figure 6.2

	D	I	X	O	N
D	1	0	0	0	0
I	0	1	0	0	0
C	0	0	0	0	0
K	0	0	0	0	0
S	0	0	0	0	0
O	0	0	0	1	0
N	0	0	0	0	1
X	0	0	0	0	0

FIGURE 6.2 – Illustration de correspondance.

- $m = 4$ (les deux X ne correspondent pas, car ils sont éloignés de plus de 3 caractères)
- $|s_1| = 5$
- $|s_2| = 8$
- $t = 0$

La distance de Jaro :

$d_j = \frac{1}{3}\left(\frac{4}{5} + \frac{4}{8} + \frac{4-0}{4}\right) = 0.767$

La distance de Jaro-Winkler avec $\ell = 2$

$d_w = 0.767 + (2 \times 0.1 \times (1 - 0.767)) = 0.813$

Les méthodes basées sur la distance de modification ont l'avantage d'être simple, et adaptée aux séquence de taille moyenne. Par contre, elles sont inefficaces pour de longues séquences et ne tient pas compte de la sémantique.

6.3.2 Approches basée sur les jetons

Dans cette approche, l'entrée attendue est un ensemble de jetons, plutôt que des séquences complètes. L'idée est de trouver les jetons similaires dans les deux ensembles. Plus est le nombre de jetons communs, plus est la similitude entre les ensembles. Une séquence peut être transformée en ensembles de jetons par fractionnement à l'aide d'un délimiteur. Nous pouvons de ce fait transformer une phrase en jetons de mots ou en caractères n-grammes (une sous-séquence de n éléments construite à partir d'une séquence donnée). Les jetons de longueur différente ont la même importance.

Définition 37. *(Jaccard Index). Relevant du domaine de similitude d'ensemble, les formules consistent à trouver le nombre de jetons communs et à le diviser par le nombre total de jetons uniques. Il est exprimé en termes mathématiques par,*

$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$

où, le numérateur est l'intersection (jetons communs) et le dénominateur est l'union (jetons uniques). Comme l'entrée requise est un ensemble de jetons au lieu de séquences complètes, il incombe de diviser efficacement et intelligemment les séquences en jetons, selon le cas d'utilisation.

Définition 38. *(Sorensen-Dice). La logique ici consiste à trouver les jetons communs et à les diviser par le nombre total de jetons présents en combinant les deux ensembles. La formule se présente comme suit :*

$DSC = \frac{2|A \cap B|}{|A| + |B|}$

où, le numérateur est deux fois l'intersection de deux ensembles (séquences). L'idée est que si un jeton est présent dans les deux chaînes, son nombre total est évidemment le double de l'intersection (ce qui supprime les doublons). Le dénominateur est une simple addition de tous les jetons des deux séquences. On remarque que c'est bien différent du dénominateur du jaccard, qui était l'union des deux séquences. Comme dans le cas de l'intersection, l'union supprime également les doublons et cela est évité dans l'algorithme de Dice. Pour cette raison, Dice surestimera toujours la similitude entre deux séquences.

L'approche basée sur les jetons a l'avantage d'être efficace en terme de temps de calcul, adaptée aux longues séquences et de tenir compte de la sémantique. l'approche basée sur les jetons n'est pas adaptée aux courtes séquences.

6.3.3 Approches basée sur les séquences

Dans l'approche basée sur séquences, la similitude est un facteur de sous-séquence communes entre les deux séquences. Les algorithmes essaient de trouver la sous-séquence la plus longue qui est présente dans les deux séquences, plus ces séquences sont trouvées, plus le score de similitude est élevé. La combinaison d'éléments de même longueur a la même importance.

Définition 39. *(Ratcliff-Obershelp). Ratcliff-Obershelp est un algorithme qui permet de déterminer la similitude de deux séquence, développé en 1983 par John W. Ratcliff et John A. Obershelp. Selon l'algorithme, la similitude en deux séquences est déterminée en calculant le double du nombre d'éléments trouvés en commun K_m divisé par le nombre total de d'éléments dans les deux séquences. nombre d'éléments trouvés en commun K_m est égale à la longueur de la plus longue sous-séquence plus récursivement le nombre d'éléments trouvés en commun dans les régions situées de part et d'autre de la plus longue sous-séquence.*

$D_{ro} = \frac{2K_m}{|s_1|+|s_2|}$

Les méthodes basées sur les séquences sont simple, et adaptée aux séquence de taille moyenne. Par contre, elles sont inefficaces pour de longues séquences, couteuse en terme de temps d'exécution puis ne tient pas compte de la sémantique.

6.4 Relation entre alignement et similarité de séquence

Un alignement pourrait être converti en un score de similitude à l'aide d'une fonction d'évaluation. La fonction d'évaluation la plus simple pourrait être « ajouter 1 pour chaque élément correspondant dans l'alignement, 0 pour chaque décalage ou espace (espace)». Mais il existe de nombreuses autres fonctions d'évaluations, et il n'y a donc pas une similitude unique correspondant à un alignement.

Habituellement, lorsqu'on parle d'alignement, on pense à l'alignement optimal (le meilleur) par rapport à une fonction d'évaluation donnée.

La similitude pourrait être définie par un alignement. En particulier, comme le score du meilleur alignement (score le plus élevé) entre les deux séquences.

Une façon de mesurer la similitude entre deux séquences est de rechercher l'alignement optimal entre ces deux séquence.

6.5 Discussion

Dans ce chapitre, nous avons présenté l'état de l'art des approche de mesure de similitude de séquences. Il existe plusieurs approches de comparaison de deux séquences regroupées en trois catégories. Chaque catégorie a ses avantages et inconvénients et est adaptée à une gamme de cas d'utilisation. Le tableau 6.1 permet de mieux comprendre la différence entre les catégories d'approche.

A travers ce tableau, nous pouvons conclure que l'approche basée sur la distance d'édition a l'avantage d'être simple et adaptée à la mesure de similitude entre séquences issues des activités de la vie quotidienne. Il convient également de dire que l'approche basée sur la distance d'édition la plus connue est la distance de Levenshtein. La distance de Levenshtein est souvent confondue à l'approche basée sur la distance et est communément appelée *distance d'édition*. Ainsi, dans le cadre de notre étude, nous avons choisi d'utiliser à la phase de connaissance d'activités de la vie quotidienne, la distance de Levenshtein pour mesurer la similitude de séquences.

TABLE 6.1 – Tableau comparatif des approches de similitude de séquences.

Catégorie	**Avantages**	**Inconvénients**
Approche basée sur la distance d'édition	– Simplicité – Adaptée aux courtes séquences	– ne tient pas compte de la signification sémantique – inefficace pour de longues séquences
Approche basée sur les jetons	– Applicable aux longue séquences – peu tenir compte de la signification sémantique	– n'est pas adaptée aux courtes séquences
Approche basée sur les séquences	– Simplicité – Adaptée aux courtes séquences	– ne tient pas compte de la signification sémantique – inefficace pour de longues séquences – coûteux en termes de temps d'exécution

6.6 Expérimentations et analyse des résultats

Cette section présente les expérimentations réalisées en suivant l'approche qui consiste à utiliser la fouille incertaine de séquences fréquentes avec gestion de la contrainte temporelle à la phase de découverte d'activités. la phase de reconnaissance se fait en deux étapes : une première étape qui consiste à s'assurer que l'activité à reconnaitre respecte bien la contrainte temporelle inter-événement définie et une deuxième étape qui consiste à utiliser la technique d'alignement de séquences pour évaluer la similitude entre l'activité à reconnaitre et un des modèles obtenus à la phase de découverte d'activités. Une analyse des différents résultats issus de cette expérimentation sera également présentée.

6.6.1 Expérimentations

Dans le cadre de nos expérimentations, nous avons utilisé la base de données MIT et la base de données Center of Advanced Studies in Adaptive System (CASAS). Pour

transformer les données réels issues des capteurs en une base de séquences, une phase de pré-traitement est nécessaire (voir sous-section 4.5.1). Les expérimentations ont été conduites en utilisant une machine équipée d'un processeur Intel(R) Core(TM)$i7-7500U$ CPU @2.70GHz 2.90GHz, d'une mémoire RAM de 8GB et tournant sous Windows 10. Pour appliquer l'algorithme de fouille incertaine de séquences fréquentes sur la base de séquence construite à l'issu de la phase de pré-traitement, un seuil de 0.5 a été choisi car c'est la limite acceptable pour qu'une séquence soit dite fréquente dans une base de séquences. Lors de la phase de reconnaissance d'activité, nous avons utilisé la méthode d'alignement de séquences basée sur la distance de Levenshtein. 70% des données sont utilisées pour l'apprentissage (découverte d'activités) et 30% pour le test (reconnaissance d'activités). l'algorithme 3 présente le pseudo-code de la reconnaissance en utilisant l'alignement de séquences. La procédure prend en paramètre ATD et FSP_DB. Le paramètre ATD représente la partie test (30%) des données brute issues des capteurs. Ces données seront converties en séquences en utilisant la procédure de pré-traitement des données décrite à la sous-section 3.5.1. Le paramètre FSP_DB représente la base de séquences fréquentes issue de la phase de découverte d'activités.

La base de données CASAS

La base de données CASAS est une collection de données dans deux scénarios réalistes différents pour détecter les ADL normales et entrelacées. La figure 6.3 montre la disposition de la maison intelligente du projet CASAS qui comprend trois chambres à coucher, une salle de bain, une cuisine et un salon/salle à manger. Les activités des résidents de la maison intelligente sont enregistrées à l'aide de capteurs de mouvement, de porte, de température et d'objets. Le tableau 6.2 présente les données brutes des capteurs pour une activité de restauration.

Les ADL normales correspondent aux cas où les habitants d'une maison intelligente se concentrent sur une seule activité à la fois. Ces activités sont effectuées séparément, sans entrelacement ni interruption. Dans l'ensemble de données CASAS, 24 utilisateurs ont effectué cinq activités normales (voir tableau 6.3). Les données ont été enregistrées pour chacune des cinq activités effectuées par les 24 utilisateurs. Ainsi, un nombre total

Algorithm 3 Pseudo-code de l'agorithme de reconnaissance par alignement de séquences

```
1: procedure RECONNAISSANCE(ATD, FSP_DB)
2:     for all d ∈ ATD do
3:         convert d to σ (sequences)
4:     end for
5:     for all sequences σ_j do
6:         MaxScore ← 0
7:         for all a_i ∈ Activities do
8:             Score_a_i ← MATCHSCORE(σ_j, FSP_DB_a_i)
9:             if Score_a_i > MaxScore then
10:                MaxScore ← Score_a_i
11:                ActivityClass_σ_j ← a_i
12:            end if
13:        end for
14:    end for
15: end procedure

16: procedure MATCHSCORE(σ, FSP_DB)
17:    Score ← 0
18:    for all pattern P_i ∈ FSP_DB do
19:        Score ← Score + Levenshtein_Distance(σ, P_i)
20:    end for
21:    return Score
22: end procedure
```

de 120 ensembles de données a été recueilli ;

Dans les ADL entrelacées, les activités sont souvent effectuées non seulement de manière isolée (c'est-à-dire équidistante), mais aussi de manière complexe (c'est-à-dire entrelacée et simultanée). L'ensemble de données CASAS contient 24 participants qui ont effectué huit activités entrelacées (voir tableau 6.4) dans la maison intelligente. Les données ont été enregistrées pour chacune des huit activités effectuées par les 24 utilisateurs, ce qui nous donne 192 ensembles de données recueilli.

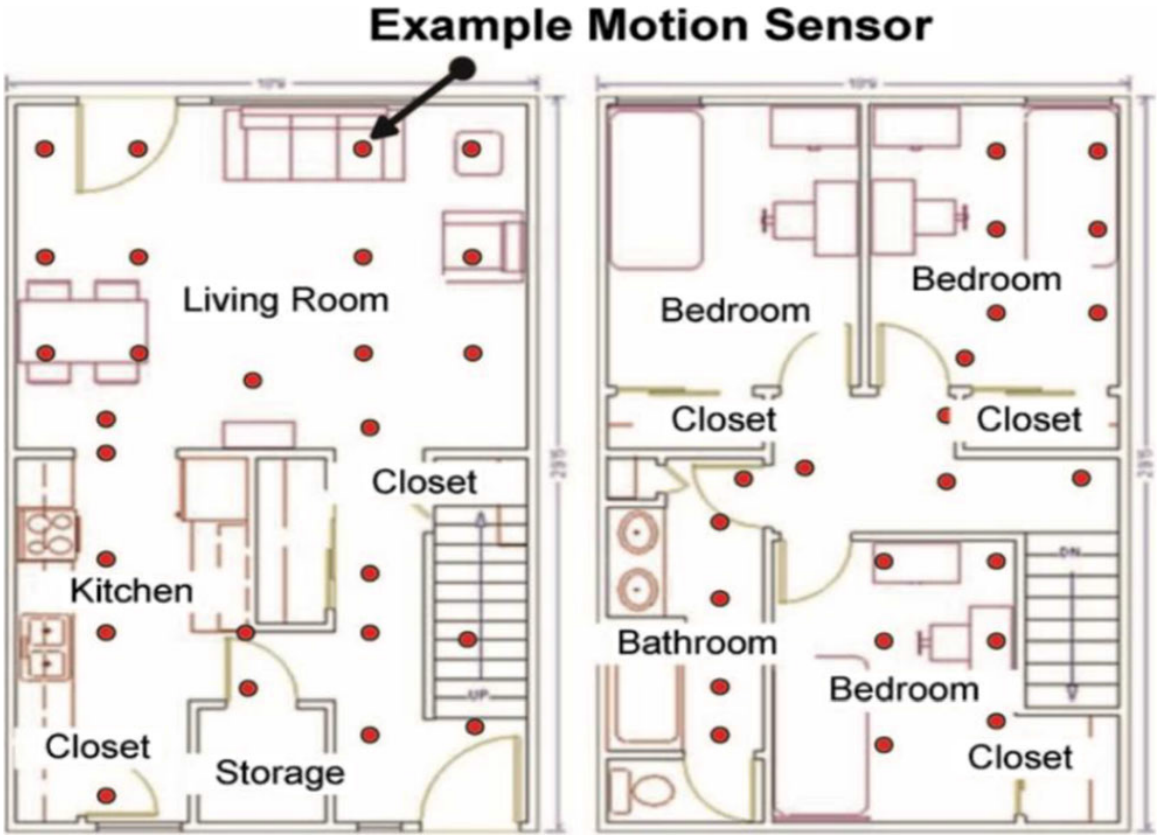

FIGURE 6.3 – Aménagement de l'appartement utilisé dans le CASAS pour la collecte de données.

6.6.2 Analyses des résultats

Les expérimentations réalisées sur les bases de données choisies montrent une amélioration du taux de reconnaissance d'activités par rapport aux études existantes (voir tableau 6.5). Par ailleurs, on observe une forte corrélation entre : le taux de confiance

TABLE 6.2 – Une partie des données brutes issues capteurs Base CASAS

2011-11-04	12 :43 :27.416392	M08	ON
2011-11-04	12 :43 :27.8481	M07	ON
2011-11-04	12 :43 :28.487061	M09	ON
2011-11-04	12 :43 :30.28561	M08	OFF
2011-11-04	12 :43 :31.491254	M07	OFF
2011-11-04	12 :43 :31.491254	M08	ON
2011-11-04	12 :43 :32.18904	M07	ON
2011-11-04	12 :43 :34.23456	M08	OFF

TABLE 6.3 – Activités normales.

Id	**Activity**
1	Phone Calling
2	HandWashing
3	Cooking
4	Eating
5	Cleaning

TABLE 6.4 – Activités entrelacées.

Id	**Activity**
1	Fill Medication Dispenser
2	Watch DVD
3	Water Plants
4	Converse on Phone
5	Write Birthday Card
6	Prepare Meal
7	Sweep and Dust
8	Select an Outfit

des capteurs et le taux de reconnaissance d'activités d'une part, le taux de confiance des capteurs et le temps d'exécution d'autres part.

TABLE 6.5 – Comparaison des résultats.

	Approche	**Dataset**	**Résultat**	
Méthode Proposée	Supervised (Uncertain SPM)	CASAS	ADL Normale : 100%	
			ADL Entrelacée : 94.69%	
		MIT	Sujet 1 : 93.89%	
			Sujet 2 : 90.83%	
[94]	Supervised	CASAS	ADL Normale : 95%	
	(DMVP+RandomForest)		ADL Entrelacée : 94%	
		MIT	Sujet 1 : 93.64%	
			Sujet 2 : 84.73%	
[90]	Supervised	CASAS	ADL Entrelacée	Naive Bayes : 66.08%
				HMM : 71%
[91]	Hybrid Unsupervised	CASAS	ADL Normale : 73.8%	
	(Clustering + HMM)		ADL Entrelacée : 77.3%	
[92]	Hybrid Unsupervised	CASAS	ADL Normale : 78%	
	(K-pattern clustering +NN)			
[3]	Supervised	MIT	Sujet 1 : 60.6%	
	(Naive Bayes Classifier)		Sujet 2 : 41.09%	
[93]	Supervised	MIT	Sujet 1	QP : 89.23%
	(Algorithms of ANNs)			LM : 92.81%
				BBP : 87.61%

Relation entre le taux de confiance et le taux de reconnaissance

Nos expérimentations sur la base de données MIT montrent que le taux de reconnaissance d'activités est faible pour les taux de confiance compris inférieurs à 60% et accroit pour les taux de confiance supérieurs à 60% pour atteindre sa valeur maximale lorsque le taux de confiance est de 100%. Les valeurs maximales des taux de reconnaissance d'activités sont de 93.89% et 90.83% respectivement pour les sujets 1 et 2 (voir figure 6.4).

En utilisant la base de données CASAS (voir figure 6.5), nos expérimentation montrent que le taux de reconnaissance d'activités est faible pour les taux de confiance compris entre 50% et 60% et accroit pour les taux de confiance supérieurs à 60% avant d'atteindre sa valeur maximale à 100% de taux de confiance. Les valeurs maximales des taux de reconnaissance d'activités sont de 100% et 94.69% respectivement pour les activités normales

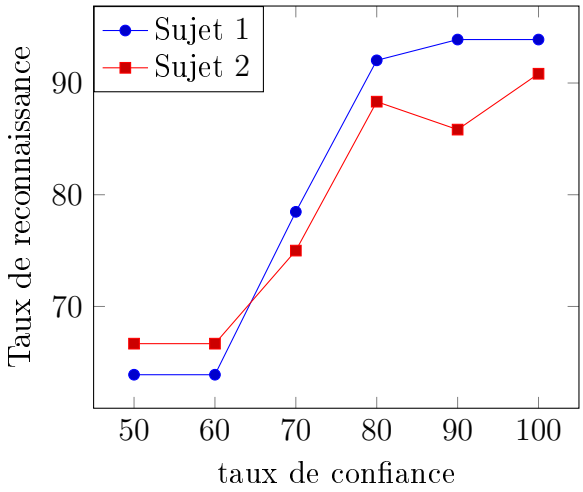

FIGURE 6.4 – Taux de reconnaissance d'activités selon le taux de confiance (Base MIT).

et entrelacées.

En résumé, nous pouvons expliquer cet accroissement du taux de reconnaissance d'activités pour les taux de confiance supérieurs à 60% par une augmentation du nombre de séquences fréquentes fiables issues de fouille incertaine de séquences fréquentes utilisant ces taux de confiance.

Relation entre le taux de confiance et le temps d'exécution

La relation entre le taux de confiance et le temps d'exécution nous montre que le temps d'exécution est bas pour les taux de confiance inférieurs à 60% puis accroit pour les taux de confiance supérieurs à 60% et atteint sa valeur maximale lorsque le taux de confiance est à 100%. Cette relation s'explique par le fait que, en utilisant les taux de confiance élevé, la fouille incertaine de séquences fréquentes renvoie plus de séquences fréquentes ce qui nécessite plus de temps de traitement.

6.7 Conclusion

Ce chapitre nous a permis de de présenter la notion d'alignement de séquences et les différents méthodes d'alignement de séquence. Dans ce chapitre nous avons également mis en exergue la notion de similitude de séquence qui est un outil de classification de

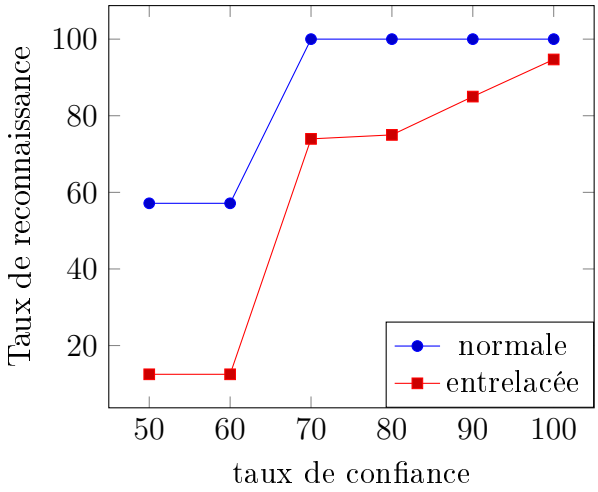

FIGURE 6.5 – Taux de reconnaissance d'activités selon le taux de confiance (Base CASAS).

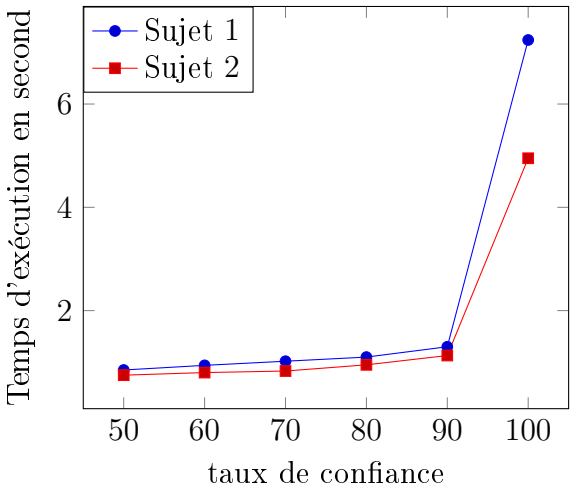

FIGURE 6.6 – Temps d'exécution selon le taux de confiance (Base MIT).

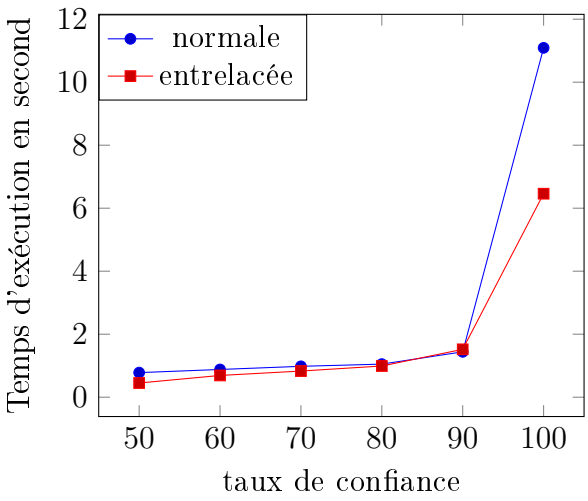

FIGURE 6.7 – Temps d'exécution selon le taux de confiance (Base CASAS).

séquences. Les différentes approches d'évaluation de la similitude de séquences on été également présentées avec leurs avantages et leurs inconvénients. nous avons utilisé l'approche basée sur l'algorithme de fouille incertaine de séquences fréquentes avec gestion de la contrainte temporelle à la phase de découverte d'activités et la méthode d'alignement de séquences à la phase de reconnaissance d'activités. Les résultats issus des expérimentations nous montrent l'importance de considérer l'incertitude des capteurs et les contraintes temporelles entre événements dans le processus de modélisation et de reconnaissance des activités. La gestion de l'incertitude des capteurs couplée avec les contraintes temporelles nous a permis d'une part d'améliorer le taux de reconnaissance d'activités avec en contre partie une augmentation du temps d'exécution.

Conclusion et perspectives

L'objectif de cette thèse était de proposer une contribution dans le domaine du maintien à domicile et du suivi des activités de la vie quotidienne en limitant le niveau sémantique du problème. En effet, les trois principaux limites du domaine sont non négligeable : le coût de l'instrumentation, le caractère non-déterminisme de l'humain et la préservation de la vie privée. Ces trois points nous amènent à émettre quatre hypothèses :

Hypothèse 1 Les activités sont représentées par des modèles probabilistes ;

Hypothèse 2 Seuls les capteurs binaires et environnementaux sont utilisés ;

Hypothèse 3 La maison intelligente considérée n'a qu'un seul habitant ;

Hypothèse 4 La connaissance de l'activité effectivement exécutée est nécessaire.

Ces hypothèses nous amènent à utiliser les paradigmes, la théorie et les outils du domaine des systèmes à événements discrets. En considérant un habitant générant des événements à travers des capteurs installés dans une maison intelligente et en utilisant une connaissance experte initiale, un nouveau cadre a été proposé pour découvrir et reconnaître les activités effectuées par l'habitant. Afin de développer ce cadre, nous avons utilisé des données adaptées à notre hypothèse et nous avons développé deux contributions principales, détaillées et présentées dans cette thèse.

Méthode de découverte d'activités avec gestion d'incertitude

Une approche de découverte des activités humaines de la vie quotidienne a été proposée et illustrée. Cette méthode utilise les algorithmes de fouille incertaine de séquences fréquentes et intègre de ce fait la gestion de l'incertitude des données issues de capteurs. L'utilisation des ces algorithmes nécessite une phase des pré-traitement des données issues des capteurs afin de construire la base de séquences. Cette phase de pré-traitement prend

en compte la notion de contrainte temporelle entre événements d'une séquence. Une procédure de modélisation des activités a été développée sur la base des connaissances d'une base de données de journal d'événements d'apprentissage et grâce à une intervention de l'expert. Le principal avantage de cette méthode est lié à l'hypothèse 4 de cette thèse qui permet l'étiquetage des activités pendant la phase d'apprentissage pour générer les modèles. Cette méthode de découverte d'activités proposée est expérimentée sur des bases de données et donne d'excellents résultats.

Méthode de reconnaissance d'activités utilisant les modèles générés précédemment

Enfin, des approches de reconnaissance d'activités ont été présentées. l'une des approches utilise le concept de forêt aléatoire avec en amont une extraction de caractéristiques du domaine temporel. La seconde approche de reconnaissance d'activités utilise le concept d'alignement de séquence pour évaluer la similitude entre deux séquences avec des distances définies et expliquées dans cette thèse. Enfin, les approches de reconnaissance ont été appliquées à plusieurs activités de test provenant des données expérimentales des laboratoire et la qualité des résultats obtenus est bonne.

Perspectives

Un cadre global pour la découverte et la reconnaissance des activités est proposé dans cette thèse et plusieurs améliorations peuvent être envisagées.

Assouplir l'hypothèse concernant la technologie des capteurs utilisée

Dans cette thèse, outre le rejet des capteurs trop intrusifs (comme les caméras), nous rejetons dans l'hypothèse 2 l'utilisation de capteurs portables parfois incompatibles avec certaines pathologies. Cette hypothèse conduit à ne considérer que les maisons intelligentes à un seul habitant. Une perspective est d'assouplir cette hypothèse, en autorisant, lorsque cela est possible, l'utilisation de capteurs portables (capteurs RFID) qui ne sont pas trop intrusifs tenant compte de la vie privée des habitants. En autorisant l'utilisation de capteurs binaires et portables, l'hypothèse d'un seul habitant peut facilement être supprimée puisque l'étiquetage automatique des données d'entrée avec le nom de la personne qui porte les capteurs est possible. En outre, l'utilisation de capteurs binaires fixées sur

certains objets peut conduire à une meilleure granularité dans les connaissances des experts. La méthode présentée, développée pour traiter les capteurs binaires, devrait être directement applicable sans changement si l'hypothèse 2 est assouplie en autorisant le port de capteurs binaires.

Assouplir l'hypothèse concernant la découverte d'activités

Dans l'hypothèse 4, nous exigeons l'étiquetage des activités à la phase d'apprentissage. Cet étiquetage a l'avantage d'offrir plus d'efficacité dans la découverte d'activités. Cependant, les tâches d'étiquetage peuvent être difficiles, sujettes à des erreurs et donc peu fiable. Une perspective est d'assouplir cette hypothèse, en autorisant, lorsque cela est nécessaire, de se passer de cette tâche d'étiquetage des activités. Pour se faire, afin de compenser la perte d'information que cela pourrait entrainer, il faut produire une connaissance supplémentaire d'un expert.

Utilisation de modèles de découverts et de reconnaissance d'activités pour traiter des problèmes de détection d'anomalies et de prédiction d'activité

Dans la continuité des travaux présentés, on peut envisager de traiter les deux autres principaux objectifs liés à l'activité de surveillance de la vie quotidienne : la détection d'anomalies et la prédiction d'activités. En effet, comme nous l'avons fait pour la découverte et la reconnaissance de d'activité, ces deux autres objectifs pourraient être reformulés pour être compatibles avec les paradigmes du système d'événements discrets. Par exemple, l'utilisation de méthodes de diagnostic bien connues du domaine industriel peut être étendue à la détection d'anomalies humaine si nous considérons les anomalies humaines comme des défauts à détecter. Puisque, dans cette thèse, les comportements humains sont des modèles utilisant les paradigmes systèmes à événement discrets (SED), ces extensions peuvent facilement être envisagées.

Annexe A

Liste des publications

A.1 Conférences internationales avec comité de lecture

1. Josky AIZAN, Cina MOTAMED, Eugène C. EZIN. *Improving Activity Mining in a Smart Home using Uncertain and Temporal Databases.* 17ième conférence internationale sur l'informatique dans le contrôle, l'automatisation et la robotique (ICINCO), 2020

2. Josky AIZAN, Cina MOTAMED, Eugène C. EZIN. *Activity Mining in a Smart Home from Sequential and Temporal Databases.* 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM), pp 542-547, 2020. DOI : 10.5220/0009061105420547.

3. Josky AIZAN, Cina MOTAMED, Eugène C. EZIN. *Learning Trajectory Patterns By Sequential Pattern Mining From probabilistic Databases.* 3rd International Conference on Data Mining and Knowledge Management, pp 73-83, 2018. DOI : 10.5121/csit.2018.815

4. Josky AIZAN, Eugène C. EZIN, Cina MOTAMED. *A Face Recognition Approach Based on Nearest Neighbor Interpolation and Local Binary Pattern.* 12th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS), Naples, pp. 76-81, 2016. DOI : 10.1109/SITIS.2016.21.

A.2 Forums, colloques et séminaires

1. Josky AIZAN. *Surveillance des personnes âgées dans un contexte de maison intelligente*. Séminaire, Université d'Abomey-Calavi, Institut de Mathématiques et de Sciences Physiques, Bénin, 13/01/2020.
2. Josky AIZAN, Cina MOTAMED. *Recent Advances and Research in Smart Homes for Elderly Healthcare, two examples of monitoring systems*. RNI Forum Innovation IX and Summer school, Naples, Italie, 17/07/2019.
3. Josky AIZAN, Cina MOTAMED. *Techlonogies de surveillance intelligente des personne du troisième âge*. colloque Silver Economie, vulnérabilités et territoires, Université d'Artois, pôle d'Arras, France, 19/03/2019.

Bibliographie

[1] United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospects 2019 : Volume II : Demographic Profiles, 2019.

[2] M. P. Lawton and E. M. Brody. Assessment of older people : self-maintaining and instrumental activities of daily living. The gerontologist, 9(3 Part 1) :179–186, 1969.

[3] E. M. Tapia, S. S. Intille and K. Larson. Activity recognition in the home using simple and ubiquitous sensors. Springer, 2004.

[4] S. Gaglio, G. L. Re and M. Morana. Human activity recognition process using 3-d posture data. IEEE Transactions on Human-Machine Systems, 45(5) :586–597, 2015.

[5] J. Aïzan, C. Motamed and E. Ezin. Activity Mining in a Smart Home from Sequential and Temporal Databases. 9th International Conference on Pattern Recognition Applications and Methods, pages 542–547, 2020.

[6] J. Aïzan, C. Motamed and E. Ezin. Improving Activity Mining in a Smart Home using Uncertain and Temporal Databases. 17th International Conference on Informatics in Control, Automation and Robotics, 2020.

[7] M. Rouse and I. Wigmore. Defintion : ubiquitous networking 2003, 2017

[8] R. C. Shah and J. M. Rabaey. Energy aware routing for low energy ad hoc sensor networks. In Wireless Communications and Networking 2003 Conference, WCNC2002. 2002 IEEE, volume 1, pages 350–355. IEEE, 2002.

[9] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei and T. Weng. Occupancy driven energy management for smart building automation. In Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building, pages 1–6. ACM, 2010.

[10] M. Flöck Activity monitoring and automatic alarm generation in AAL-enabled homes. Logos Verlag Berlin GmbH, 2010.

[11] D. Monekosso, F. Florez-Revuelta and P. Remagnino. Ambient assisted living [guest editors' introduction]. IEEE Intelligent Systems, 30(4) :2–6, 2015.

[12] T. Kleinberger, M. Becker, E. Ras, A. Holzinger and P. Muller. Ambient Intelligence in Assisted Living : Enable Elderly People to Handle Future Interfaces, pages 103–112. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[13] J. Barlow and T. Venables. Smart home, dumb suppliers ? the future of smart homes markets. In Inside the Smart Home, pages 247–262. Springer, 2003.

[14] R. J. Robles and T. H. Kim. Applications, systems and methods in smart home technology : A. Int. Journal of Advanced Science And Technology, 15, 2010.

[15] N. Balta-Ozkan, B. Boteler and O. Amerighi. European smart home market development : Public views on technical and economic aspects across the united kingdom, germany and italy. Energy Research and Social Science, 3 :65–77, 2014.

[16] N. King. Smart home–a definition. Intertek Research and Testing Center, pages 1–6, 2003.

[17] M. Chan, D. Esteve, C. Escriba and E. Campo. A review of smart homes—present state and future challenges. Computer methods and programs in biomedicine, 91(1) :55–81, 2008.

[18] A. S. Taylor, R. Harper, L. Swan, S. Izadi, A. Sellen and M. Perry. Homes that make us smart. Personal and Ubiquitous Computing, 11(5) :383–393, 2007.

[19] L. Jiang, D.-Y. Liu and B. Yang. Smart home research. In Machine Learning and Cybernetics. Proceedings of 2004 International Conference on, volume 2, pages 659–663. IEEE, 2004.

[20] F. Scott. Teaching homes to be green : smart homes and the environment. Green Alliance, 2007.

[21] A. R. M. Forkan, I. Khalil, Z. Tari, S. Foufou and A. Bouras. A context-aware approach for long-term behavioural change detection and abnormality prediction in ambient assisted living. Pattern Recognition, 48(3) :628–641, 2015.

[22] I. Sadek, J. Biswas, V. F. S. Fook and M. Mokhtari. Automatic heart rate detection from fbg sensors using sensor fusion and enhanced empirical mode decomposition. In Signal Processing and Information Technology (ISSPIT), IEEE International Symposium on, pages 349–353. IEEE, 2015.

[23] C. Otto, A. Milenkovic, C. Sanders and E. Jovanov. System architecture of a wireless body area sensor network for ubiquitous health monitoring. Journal of mobile multimedia, 1(4) :307–326, 2006.

[24] F. Bellifemine, G. Fortino, R. Giannantonio, R. Gravina, A. Guerrieri and M. Sgroi. Spine : a domain-specific framework for rapid prototyping of wbsn applications. Software : Practice and Experience, 41(3) :237–265, 2011.

[25] A. A. Chaaraoui, P. Climent-Perez and F. Florez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. Expert Systems with Applications, 39(12) :10873–10888, 2012.

[26] M. Andriluka, S. Roth and B. Schiele. Pictorial structures revisited : People detection and articulated pose estimation. In Computer Vision and Pattern Recognition. CVPR 2009. IEEE Conference on, pages 1014–1021. IEEE, 2009.

[27] B. Sapp, A. Toshev and B. Taskar. Cascaded models for articulated pose estimation. In European conference on computer vision, pages 406–420. Springer, 2010.

[28] W. Li, Z. Zhang and Z. Liu. Expandable data-driven graphical modeling of human actions based on salient postures. IEEE transactions on Circuits and Systems for Video Technology, 18(11) :1499–1510, 2008.

[29] S. Hussain, S. Schaffner and D. Moseychuck. Applications of wireless sensor networks and rfid in a smart home environment. In Communication Networks

and Services Research Conference. CNSR'09. Seventh Annual, pages 153–157. IEEE, 2009.

[30] M. Darianian and M. P. Michael. Smart home mobile rfid-based internet-ofthings systems and services. In Advanced Computer Theory and Engineering. ICACTE'08. International Conference on, pages 116–120. IEEE, 2008.

[31] O. Brdiczka, J. L. Crowley and P. Reignier. Learning situation models in a smart home. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(1) :56–63, 2009.

[32] S. Park and H. Kautz. Hierarchical recognition of activities of daily living using multi-scale, multi-perspective vision and rfid, 2008.

[33] D. J. Cook and N. C. Krishnan. Activity Learning : Discovering, Recognizing, and Predicting Human Behavior from Sensor Data. John Wiley and Sons, 2015.

[34] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. Communications Surveys and Tutorials, IEEE, 15(3) :1192–1209, 2013.

[35] B. Chikhaoui, S. Wang and H. Pigot. A frequent pattern mining approach for adls recognition in smart environments. In Advanced Information Networking and Applications (AINA), IEEE International Conference on, pages 248–255. IEEE, 2011.

[36] S. Himmel, M. Ziefle and K. Arning. From Living Space to Urban Quarter : Acceptance of ICT Monitoring Solutions in an Ageing Society, pages 49–58. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[37] Farlex. activities of daily living definition, 2018.

[38] M. A. R. Ahad, J. Tan, H. Kim and S. Ishikawa. Human activity recognition : Various paradigms. In Control, Automation and Systems. ICCAS 2008. International Conference on, pages 1896–1901. IEEE, 2008.

[39] A. Fleury, M. Vacher and N. Noury. Svm-based multimodal classification of activities of daily living in health smart homes : sensors, algorithms, and first

experimental results. Information Technology in Biomedicine, IEEE Transactions on, 14(2) :274–283, 2010.

[40] E. Kim, S. Helal and D. Cook. Human activity recognition and pattern discovery. Pervasive Computing, IEEE, 9(1) :48–53, 2010.

[41] T. V. Duong, H. H. Bui, D. Q. Phung and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 838–845. IEEE, 2005.

[42] S. Mahmoud, A. Lotfi and C. Langensiepen. Behavioural pattern identification and prediction in intelligent environments. Applied Soft Computing, 13(4) :1813–1822, 2013.

[43] J. Krumm and E. Horvitz. Predestination : Inferring destinations from partial trajectories. In International Conference on Ubiquitous Computing, pages 243–260. Springer, 2006.

[44] V. Chandola, A. Banerjee and V. Kumar. Anomaly detection for discrete sequences : A survey. IEEE Transactions on Knowledge and Data Engineering, 24(5) :823–839, 2012.

[45] D. J. Cook, N. C. Krishnan and P. Rashidi. Activity discovery and activity recognition : A new partnership. IEEE transactions on cybernetics, 43(3) :820–828, 2013.

[46] A. Forkan, I. Khalil and Z. Tari. Cocamaal : A cloud-oriented context-aware middleware in ambient assisted living. Future Generation Computer Systems, 35 :114–127, 2014.

[47] T. Duong, D. Phung, H. Bui and S. Venkatesh. Efficient duration and hierarchical modeling for human activity recognition. Artificial intelligence, 173(7-8) :830–856, 2009.

[48] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2) :257–286, 1989.

[49] J. Saives, C. Pianon and G. Faraut. Activity discovery and detection of behavioral deviations of an inhabitant from binary sensors. Automation Science and Engineering, IEEE Transactions on, 12(4) :1211–1224, 2015.

[50] R. Agrawal and R. Srikant. Mining sequential patterns. The International Conference on Data Engineering, pages 3-14, 1995.

[51] J. Mocholi, P. Sala, C. Fernandez-Llatas and J. Naranjo. Ontology for modeling interaction in ambient assisted living environments. In XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010, pages 655–658. Springer, 2010.

[52] A. Devaraju and S. Hoh. Ontology-based context modeling for user-centered context-aware services platform. In Information Technology. ITSim 2008. International Symposium on, volume 2, pages 1–7. IEEE, 2008.

[53] A. Mihailidis, B. Carmichael and J. Boger. The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home. IEEE Transactions on information technology in biomedicine, 8(3) :238–247, 2004.

[54] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In Computer Vision. ICCV 2005. Tenth IEEE International Conference on, volume 1, pages 90–97. IEEE, 2005.

[55] Z. Zhou, X. Chen, Y.-C. Chung, Z. He, T. X. Han and J. M. Keller. Activity analysis, summarization, and visualization for indoor human activity monitoring. IEEE Transactions on Circuits and Systems for Video Technology, 18(11) :1489–1498, 2008.

[56] T. Van Kasteren, A. Noulas, G. Englebienne and B. Krose. Accurate activity recognition in a home setting. In Proceedings of the 10th international conference on Ubiquitous computing, pages 1–9. ACM, 2008.

[57] V. Kellokumpu, M. Pietikainen and J. Heikkila. Human activity recognition using sequences of postures. In MVA, pages 570–573, 2005.

[58] J. Ayres, J. Flannick, J. Gehrke and T. Yiu. Sequential pattern mining using a bitmap representation. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 429-435, 2002.

[59] P. Fournier-Viger, A. Gomariz, M. Campos and R. Thomas. Fast Vertical Mining of Sequential Patterns Using Co-occurrence Information. The Paci c-Asia Conference on Knowledge Discovery and Data Mining, pages 40-52, 2014.

[60] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal and M. C. Hsu. Mining sequential patterns by pattern-growth : The pre xspan approach. IEEE Transactions on knowledge and data engineering, vol. 16(11), pages 1424-1440, 2004.

[61] J. Han, J. Pei, Y. Ying and R. Mao. Mining frequent patterns without candidate generation : a frequent-pattern tree approach. Data Mining and Knowledge Discovery, vol. 8(1), pages 53-87, 2004.

[62] R. Srikant and R. Agrawal. Mining sequential patterns : Generalizations and performance improvements. The International Conference on Extending Database Technology, pages 1-17, 1996.

[63] M. J. Zaki. Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, vol. 12(3), pages 372-390, 2000.

[64] M. J. Zaki. SPADE : An efficient algorithm for mining frequent sequences. Machine learning, vol. 42(1-2), pages 31-60, 2001.

[65] G. Cormode, F. Li and K. Yi. Semantics of ranking queries for probabilistic data and expected ranks. In : ICDE, pages. 305–316, IEEE 2009

[66] Q. Zhang, F. Li and K. Yi. Finding frequent items in probabilistic data. In : Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD, pages 819–832, 2008

[67] C. C. Aggarwal, Y. Li and J. Wang. Frequent pattern mining with uncertain data. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 29–38, 2009

[68] T. Bernecker, H. P. Kriegel, M. Renz, F. Verhein and A. Zufle. Probabilistic frequent itemset mining in uncertain databases. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 119–128, 2009

[69] C. K. Chui and B. Kao. A decremental approach for mining frequent itemsets from uncertain data. In : PAKDD, pages 64–75, 2008

[70] C. K. Chui, B. Kao and E. Hung. Mining frequent itemsets from uncertain data. In PAKDD. LNCS, vol. 4426, pages 47–58, Springer 2007

[71] M. Muzammal and R. Raman. On probabilistic models for uncertain sequential pattern mining. In ADMA (1). LNCS, vol. 6440, pages 60–72, Springer 2010

[72] J. Stuart Russell and Peter Norvig. Artificial Intelligence. A Modern Approach (Third ed.). Prentice Hall, 2010

[73] Mehryar Mohri , Afshin Rostamizadeh and Ameet Talwalkar. Foundations of Machine Learning. The MIT Press, 2012

[74] Ethem Alpaydin. Introduction to Machine Learning. MIT Press, page 9, 2010

[75] T. Mitchell. Machine Learning. McGraw Hill, page 2, 1997

[76] Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648. University of Wisconsin–Madison, 2010

[77] Neil Rubens, Mehdi Elahi, Masashi Sugiyama and Dain Kaplan. Active Learning in Recommender Systems. In Ricci, Francesco ; Rokach, Lior ; Shapira, Bracha (eds.). Recommender Systems Handbook (2 ed.). Springer US, 2016

[78] Shubhomoy Das, Weng-Keen Wong, Thomas Dietterich, Alan Fern and Andrew Emmott. Incorporating Expert Feedback into Active Anomaly Discovery. In IEEE 16th International Conference on Data Mining. IEEE, pages. 853–858, 2016

[79] David A. Freedman. Statistical Models : Theory and Practice. Cambridge University Press, 2009

[80] L. Rokach. Ensemble-based classifiers. Artificial Intelligence Review, vol 33 (1–2), pages 1–39, 2010

[81] Shai Shalev-Shwartz and Shai Ben-David (2014). 18. Decision Trees. Understanding Machine Learning. Cambridge University Press.

[82] L. Breiman. Random forests. Machine learning, vol 45(1), pages 5–32, 2001

[83] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. J Comput Biol. vol 1 (4), pages 337–48, 1994

[84] Isaac Elias. Settling the intractability of multiple alignment. J Comput Biol. vol 13 (7), pages 1323–1339, 2006

[85] A. Monge and C. Elkan. The field matching problem. Algorithms and applications.In Proceedings of The Second International Conference on Knowledge Discoveryand Data Mining (KDD), 1996

[86] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, vol 10 (8), pages 707-710, 1966

[87] W. E. Winkler. The state of record linkage and current research problems. Statistics of Income Division, Internal Revenue Service Publication, 1999

[88] M. A. Jaro. Advances in record linking methodology as applied to the 1985 census of Tampa Florida. Journal of the American Statistical Society, vol. 84 (406), pages 414-420, 1989

[89] M. Raeiszadeh and H. Tahayori. A Novel Method for Detecting and Predicting Resident's Behavior in Smart Home. 6th Iranian Joint Congress on Fuzzy and Intelligent Systems IEEE, 2018

[90] G. Singla and D. J. Cook. Interleaved activity recognition for smart home residents. 5th IEEE Int. Conf. Intelliget environ., pages 145–152, 2009

[91] P. Rashidi, D. J. Cook, L. B. Holder and M. Schmitter-Edgecombe. Discovering activities to recognize and track in a smart environment. IEEE Trans Knowl Data Eng, vol. 23(4) pages 527–539, 2011

[92] S. T. M. Bourobou and Y. Yoo. User activity recognition in smart homes using pattern clustering applied to temporal ANN algorithm. Sensors (Switzerland), vol. 15(5) pages 11953–11971, 2015

[93] H. D. Mehr, H. Polat and A. Cetin. Resident activity recognition in smart homes by using artificial neural networks. 4th Int. Istanbul Smart Grid Congr. Fair, 2016

[94] M. Raeiszadeh, H. Tahayori and A. Visconti. Discovering varying patterns of Normal and interleaved ADLs in smart homes. Appl Intell, vol. 49, pages 4175, 2019

Résumé

Les systèmes d'aide à la vie ambiante permettant le maintien à domicile des personnes âgées sont en pleine expansion de nos jours. Les nouvelles approches consistent à mettre en place un système automatisé de surveillance d'activités au sein d'une maison intelligente équipée de capteurs portables tels que les GPS, les bracelets électroniques ou les puces RFID. Ces capteurs malheureusement ont la contrainte d'être portés constamment. L'utilisation des capteurs binaires est une alternative de plus en plus proposée. Dans cette thèse nous avons proposé la modélisation et la reconnaissance d'activités quotidiennes au sein d'une maison intelligente équipée de capteurs binaires. La première phase de l'architecture proposée concerne la modélisation d'activités. Les algorithmes de fouilles de séquences fréquentes déterministes et incertaines ont été utilisés. Ces algorithmes contiennent une phase de pré-traitement qui intègre la contrainte temporelle entre évènements. Les performances de ces algorithmes ont été évaluées sur la base de données MIT qui contient une collection d'activités humaines issues de deux appartements instrumentés respectivement de 77 et 84 capteurs. Ces expérimentations nous montrent que le nombre et la qualité des modèles issus de la phase de modélisation sont fortement liés au taux de confiance des capteurs. La seconde phase de l'architecture concerne la reconnaissance d'activités. Au cours de cette phase, deux approches sont proposées. La première approche consiste à coupler la méthode de forêt aléatoire avec l'algorithme de fouille déterministe de séquences fréquentes. Cette approche intègre une caractérisation temporelle des modèles d'activités découverts. Une expérimentation est effectuée sur la base de données MIT et les résultats en terme de reconnaissance d'activités sont de 98% pour le sujet 1 et 95% pour le sujet 2. Ces résultats sont comparés à ceux de la littérature pour rendre compte de la performance de l'approche proposée. La seconde approche utilise la méthode de reconnaissance par alignement de séquences basée sur la distance de Levenshtein couplée à la fouille incertaine de séquences fréquentes. A ce niveau, l'algorithme de fouille incer-

taine de séquences fréquentes, intègre à la fois la gestion des contraintes temporelles entre évènements et la gestion de l'incertitude des données issus des capteurs. Les performances de cette méthode ont été évaluées sur les bases de données MIT et CASAS. La base de données CASAS contient une collection de données issues de deux scénarios réalistes pour détecter les activités de la vie quotidiennes normales et entrelacées. Les résultats obtenus des expérimentations sur ses deux bases de données montrent que le taux de reconnaissances est une fonction croissante du taux de confiance des capteurs. Ces résultats sont de 100% et 94% respectivement pour les activités normales et entrelacées de la base CASAS puis 93% et 90% respectivement pour les activités des sujets 1 et 2 de la base MIT. Comparés avec ceux de la littérature, ces résultats mettent en évidence l'efficacité de notre méthode.

Mots-clés: fouille de séquence fréquentes, maison intelligente, activités de la vie quotidienne.

Abstract

The ADL systems for keeping seniors at home are expanding today. The new approaches involve setting up an automated activity monitoring system in a smart home equipped with wearable sensors such as Global Positioning System (GPS), electronics bracelets or RFID chips. These sensors unfortunately have the constraint to be worn constantly. The use of binary sensors is an increasingly common alternative. In this thesis we proposed modeling and recognition of daily activities within a smart home equipped with binary sensors. The first phase of the proposed architecture concerns activity modelling. Deterministic and uncertain sequential pattern mining algorithms were used. These algorithms contain a pre-processing phase that integrates the temporal constraint between events. The performance of these algorithms was evaluated on the MIT database, which contains a collection of human activities from two instruments of 77 and 84 sensors respectively. These experiments show that the number and quality of models from the modeling phase are strongly linked to the confidence rate of the sensors. The second phase of architecture involves the recognition of activities. During this phase, two approaches are proposed. The first approach is to pair the random forest method with the deterministic sequential pattern mining algorithm. This approach incorporates a temporal characterization of the activity models discovered. An experiment is carried out on the MIT database and the results in terms of activity recognition are 98% for the subject 1 and 95% for the subject 2. These results are compared with those in the literature to reflect the performance of the proposed approach. The second approach uses the sequence alignment recognition method based on the Levenshtein distance coupled with the uncertain sequential pattern mining. At this level, the uncertain sequential pattern mining algorithm integrates both the management of time constraints between events and the management of the uncertainty of data from the sensors. The performance of this method was evaluated on the MIT and CASAS databases. The CASAS database contains a collection of data from two realistic scenarios to detect normal and intertwined daily activities. The results of the experiments on its two databases show that the recognition rate is an increasing function of the confidence rate of the sensors. These results are 100% and 94%

respectively for the normal and interweave activities of the CASAS base and 93% and 90% respectively for the activities of subjects 1 and 2 of the MIT base. Compared with those in the literature, these results highlight the effectiveness of our method.

Keywords: uncertain sequential pattern mining, smart home, activity of daily living.