Improving Activity Mining in a Smart Home using Uncertain and Temporal Databases

Josky Aïzan[1,2], Cina Motamed[2] and Eugene C. Ezin[1]
[1]*Institut de Mathématiques et de Sciences Physiques, Université d'Abomey-Calavi, Benin*
[2]*Laboratoire d'Informatique Signal et Image de la Côte d'Opale, Université du Littoral Côte d'Opale, France*

Keywords: Sequential Pattern Mining, Smart Home, Activity of Daily Living.

Abstract: In the context of smart home, activity mining appears as an interesting and promising solution for learning activity of daily living. This paper is an extension of a previous a research work titled *Activity Mining in a Smart Home from Sequential and Temporal Databases*. It proposes an activity mining method based on uncertain and temporal sequential pattern mining to deal with data uncertainty and events temporal relationships. It allows to track regular activities and to detect changes in an individual's behavioural pattern. Uncertain sequential pattern mining algorithm is firstly applied to the input sequence database to extract typical sequences and secondly a clustering approach based on sequence alignment methods is performed in order to obtain separated typical activities. The results obtained are enough good compared to existing related works.

1 INTRODUCTION

Home assistance is an emerging area that aims to ensure health care, safety and autonomy for the elderly. A smart home capability appears as a promising solution. It is the combination of technologies and services through the networking process for a better quality of life.

Because of the large amount and uncertainty of data that can be generated through daily activities of inhabitant, in this paper, we have extended our previous work (Aïzan et al., 2020) by using sequential pattern mining technique based on probabilistic and temporal databases to discover frequent activities models. Proposed method is consisted of three steps. In the first step, the pre-processing phase converts sensor data into the event sequences. In the second step, the detection of frequent activities is performed by applying uncertain sequential pattern mining algorithm. In the third step, the system estimates temporal constraints between events inside of each set of frequent sequences and tries to cluster temporally similar sequence to obtain more specific sequences. It permits, for example, to differentiate a short and a long activity based on the same sequence.

The paper is organized as follows. In section 2, related works are reviewed. Section 3 gives a theoretical description of the proposed method while section 4 presents experimental results and analysis. A conclusion ends this paper with its future directions.

2 STATE OF ART AND RELATED WORKS

Learning daily activities in a smart home is a research real challenge in the field of Sequential Pattern Mining (SPM).

A frequent sequential pattern mining algorithm is proposed in (Schweizer et al., 2015) to learn consumer behaviour and then reduce energy consumption in smart homes. This algorithm uses a window with a prefixed size over the chronologically ordered events to find all possible frequent patterns and is named Window Sliding with De-Duplication (WSDD). This approach does not consider the time between two events. (Singh and Yassine, 2017) proposed in the same field of energy consumption behaviour analysis, an unsupervised progressive incremental data mining mechanism.

Frequent episode mining is used in (Li et al., 2017) to recognize sequential behaviour patterns. To discover data model for smart home and IoT data analytics, (Suryadevara, 2017) developed a framework. A novel sequential pattern mining algorithm called PBuilder is proposed in (Hassani et al., 2015) that

Aïzan, J., Motamed, C. and Ezin, E.
Improving Activity Mining in a Smart Home using Uncertain and Temporal Databases.
DOI: 10.5220/0009827206370642
In *Proceedings of the 17th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2020)*, pages 637-642
ISBN: 978-989-758-442-8

uses a batch-free approach to mine activities in a smart home. High utility pattern mining is given in (Menaka and Gayathri, 2013) to model activity in a smart home and uses linked sensor data stream approach to save processing time and memory space.

Temporal data mining algorithm is presented in (Moutacalli et al., 2012) to model activities. Their approach uses the mining process of the timestamp of each activity event constituting. A novel method used sequential pattern mining based on the longest common subsequence is proposed in (Raeiszadeh and Tahayori, 2018) to model behaviour in smart home.

All proposed approaches and algorithms in the context of smart home are applied over deterministic database. Therefore they did not consider the uncertainty of real world data. In practice, sensors data are not always reliable and it seems important to consider these imperfections.

Several works are carried out on sequential pattern mining algorithms by integrating uncertainty management. (Muzammal and Raman, 2010) proposed probabilistic models for uncertain sequential pattern mining. (Zhao et al., 2014) have developed algorithm to mine probabilistically frequent sequential patterns in large uncertain databases. (Li et al., 2013) focused on probabilistic frequent spatio-temporal sequential pattern with gap constraints. (Zhang et al., 2017) proposed high utility-probability sequential pattern mining from uncertain databases. (Muzammal et al., 2017) developed a framework for probabilistic trajectory extraction and mining from uncertain data. In (Yang et al., 2002) sequential pattern mining is studied in noisy sequences. Sequential pattern mining in probabilistic databases (Aggarwal, 2009), (Suciu and Dalvi, 2005) is the popular framework for modelling uncertainty.

In our approach, we considered uncertainty of the sensor events and also, the order of the sensors' activation, for the frequent sequential pattern extraction for each activity. Since some activities share some sensor events, in order to solve the conflicts in the classification of activities, we used the sequence alignment-based technique, which basically increases the average accuracy of classification over the previous works.

3 PROPOSED METHOD

In this work, we use uncertain and temporal sequential pattern mining to discover typical activities that frequently occur by an individual in smart home and the sequence alignment-based technique to recognize and predict these activities as they occur. The proposed method has three phases namely preprocessing, uncertain sequential pattern mining and activity modeling. Fig. 1 presents the proposed approach architecture wich differs from that used in our previous work (Aïzan et al., 2020) by the integration of uncertainty management in the seqetial pattern mining phase and the use of sequence alignment in the activity modeling phase. For our experience, we have used the Massachusetts Institute of Technology (MIT) smart home data set (Tapia et al., 2004). This data set needs to be transformed to a realistic probabilistic and temporal sequential database. The pre-processing represents the first stage of the architecture. The second step extracts typical activities by using a sequential pattern mining approach, and then a third stage operates activity modeling based on sequence alignment.

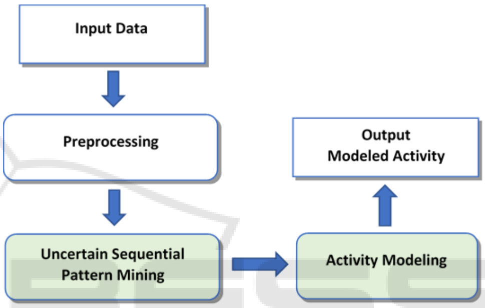

Figure 1: Architecture of proposed approach.

3.1 Pre-processing

This stage is the same as that proposed in our previous work (Aïzan et al., 2020). An activity is a time ordered records of events. Events are generated by sensors and actuators. The decision about activating of an event is linked with the state changes (Boolean) from the sensor or when its value greatly changes numerically. A small change in value is considered as the noise and is therefore ignored. In the real world, each sensor is associated with a confidence c with $0 < c < 1$. The pre-processing phase aims to convert sensor data into probabilistic event sequences. For illustration we show the "Washing dishes" activity from the dataset in Table 1. In the pre-processing phase as shown in Fig. 2, raw sensor data are converted to $(t)eid$ format in which t represents sensor activation or deactivation timestamp, eid represents event id. The event id named eid is in the form XYZ where X represents sensor id, Y represents sensor state which can be 1 for activating or 0 for deactivating. Z represents the number of times the sensor is activated or deactivated during the same activity.

Table 1: Sample of data.

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

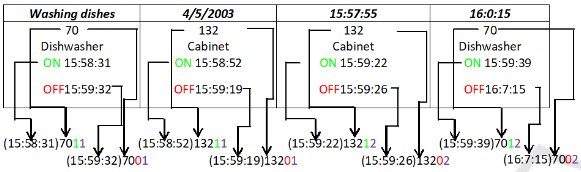

Figure 2: Pre-Processing phase of the sensor data.

3.2 Sequential Pattern Mining

The second stage is performed by a sequential pattern mining to obtain frequent sequences. According to the sequence database, we have a deterministic sequential pattern mining when there is no uncertainty in the data otherwise we must use an uncertain sequential pattern mining.

Deterministic Sequential Pattern Mining. represents the standard sequential pattern mining. Let $\mathcal{S} = \{1, \cdots, p\}$ and $I = \{i_1, i_2, \cdots, i_m\}$ be respectively a set of sources and a set of items. An event e is a collection of items such that $e \subseteq I$. A sequence database $D = \langle s_1, s_2, \cdots, s_p \rangle$ is an ordered list of sequences such that each $s_i \in D$ is of the form (eid_i, e_i, σ_i), where eid_i is a unique event-id, including a timestamp (events are ordered by this timestamp), e_i is an event and σ_i is a source.

A sequence is an ordered list of events $s = \langle e_1, e_2, \cdots, e_n \rangle$ such that $e_k \subseteq I$ $(1 \leq k \leq n)$. A sequence $s_a = \langle A_1, A_2, \cdots, A_n \rangle$ is a subsequence of another sequence $s_b = \langle B_1, B_2, \cdots, B_m \rangle$ denoted $s_a \preceq s_b$, if and only if there exist integers $1 \leq i_1 < i_2 < \cdots < i_n \leq m$ such that $A_1 \subseteq B_{i_1}, A_2 \subseteq B_{i_2}, \cdots, A_n \subseteq B_{i_n}$. Let d_i be the sequence corresponding to a source i. For a sequence s and source i, let $X_i(s, d_i)$ be an indicator variable, whose value is 1 if s is a subsequence of sequence d_i, and 0 otherwise. For any sequence s, its support in D is denoted by:

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

The goal is to find all sequences s such that $Sup(s, D) \geq \theta p$ for some user-defined threshold $0 \leq \theta \leq 1$.

Uncertain Sequential Pattern Mining. In the context of Uncertain sequential pattern mining, there are three different kinds of uncertainty: the source, the event and the time (Muzammal and Raman, 2010). In our work, we will focus on sensors data uncertainty, which is linked with the event uncertainty.

In the event uncertainty, the database $D^p = (D_1^p, D_2^p, \cdots D_m^p)$ is represented by a collection of p-sequence. Each p-sequence is an ordered list of events with its confidence. An example of p-sequence database is presented in Table 2.

Table 2: Sample of p-sequence database.

	p-sequence
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	$(e : 1.0)(e, f : 0.5)(f, g : 0.3)$

The Possible Worlds semantics of D^p is as follows. For each event e_j in a p-sequence D_i^p there are two kinds of worlds; one in which e_j occurs and the other where it does not. For the rest of the algorithm, we use Possible Worlds semantics $PW(D^p)$ of the database D^p. we get $PW(D^p)$ by calculating the $PW(D_i^p)$. each $PW(D_i^p)$ is obtained by considering of the 2^l possibilities with l the length of p-sequence D_i^p.

In this approach, the events at the p-sequences level are considered probabilistically independent, Table 3 presents $PW(D_Y^p)$.

Table 3: Possible worlds of 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)	$(1 - 0.4) \times (0.2) = 0.12$
$(e, h)(e, f)$	$(0.4) \times (0.2) = 0.08$

The same method is used to determine $PW(D_X^p)$ and $PW(D_Z^p)$ and Table 4 presents the possible worlds $PW(D^p)$ of the p-sequence database D^p.

An example of the possible world D^* is shown in Table 5.

A probability of this possible world $Pr(D^*) = 0.294 \times 0.32 \times 0.35 = 0.03$ if we consider that p-sequence stochastically independent. Expected Support is evaluated according (2).

Table 4: Possible worlds of D^p.

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

Table 5: One possible world.

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

$$ES(s,D^p) = \sum_{D^* \in PW(D^p)} Pr[D^*] \times Sup(s,D^*) \tag{2}$$

$Sup(s,D^*)$ is evaluated according the (1) because D^* is deterministic. We have $|PW(D_X^p)| \times |PW(D_Y^p)| \times |PW(D_Z^p)| = 8 \times 4 \times 4 = 128$ and then Equation (2) becomes unexploitable when the database is large. To deal with problem, expected support is evaluated as follows: Let $s = (e)(f)$ be a sequence and database of Table 2. For each source X, Y and Z, the probability that it supports s is calculated. According to $PW(D_X^p$ see Table 4, the probability that source X support s is $(0.054 + 0.294 + 0.084 + 0.126) = 0.558$ and the probability that it does not is $1 - 0.558 = 0.442$. similarly, the probability that Y and Z support s are 0.08 and 0.65. For $i = 0,1,2,3,$ independence of p-sequence is used to compute the probability that exactly i sources support s as shown in Table 6. For example, the probability that s is supported by all three sources is $(0.558 \times 0.08 \times 0.65) = 0.029$. Then $ES(s) = (0 \times 0.142 + \cdots 3 \times 0.029) = 1.228$.

Table 6: Support probability distribution.

No of sources	0	1	2	3
support probability	0.142	0.456	0.372	0.029

3.3 Activity Modeling

At this stage, each input sequence (test sequences) is compared against frequent sequences patterns of all of the activites from uncertain sequential pattern mining stage. A score will be computed for each pair of sequence and activity class. This score determines how similar the triggered sensors of input sequence are with the triggered sensors of discovered frequent sequences patterns for each activity. The class of activity with maximum score will be selected for that sequence.

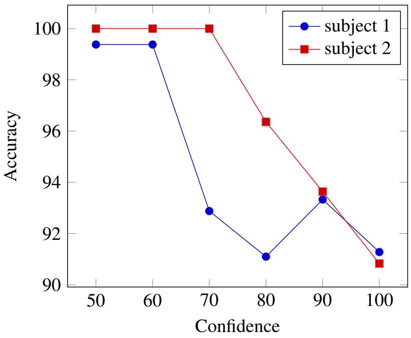

Figure 3: Activities recognition accuray according to sensor confidence.

To define a similarity measure between the two sequences (3), we used Levenshtein distance $e(A,B)$ (Rashidi et al., 2011), which is the number of edits (insertions, deletions, and substitutions) required to transform an event sequence A into another event sequence B.

$$sim(A,B) = 1 - \frac{e(A,B)}{max(|A|,|B|)} \tag{3}$$

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present the results obtained with the proposed method. We used the MIT data smart home testbed better described in subsection 4.1.

4.1 MIT Dataset

MIT dataset is a two weeks collection of human activity in two single-person apartments containing respectively 77 and 84 sensors see Fig. 5 for illustration. The sensors were installed in everyday objects such as containers, drawers, refrigerators, etc. to record opening-closing events (activation deactivation events) as the subject carried out everyday activities. Activities are labeled in to 16 different classes and the number of occurrences of each class by subject is showed in Table 7.

4.2 Results and Analysis

Our implementation in Java, is executed on a machine Intel(R) Core(TM) $i7-7500U$ CPU @2.70 GHz 2.90 GHz running on Windows 10. With a support value fixed to 0.5, our method applied on MIT dataset (see

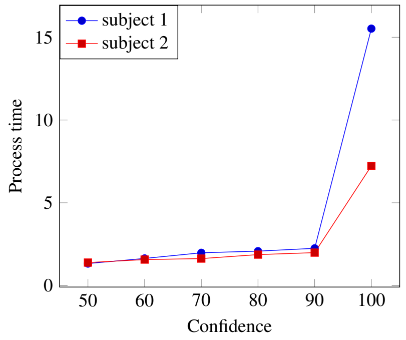

Figure 4: Processing time according to sensor confidence.

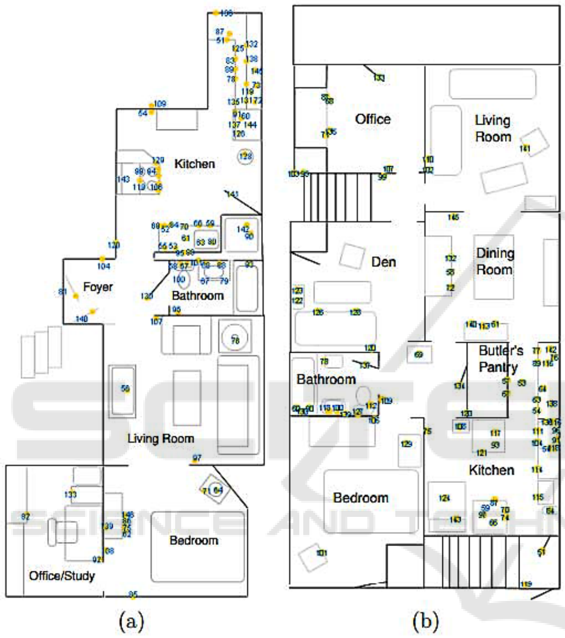

Figure 5: (a) apartment of subject one. (b) apartment of subject two.

Fig. 3) shows that accuracy rate is maximum for confidences between 50% and 70% and decreases for confidence greater than 70%. The maximum values of accuracy are 99.38% and 100% respectively for subject 1 and subject 2.

In summary, the decrease of accuracy rate for confidences greater than 70% is explained by an increase in the part of unreliable frequent pattern resulting from uncertain sequential pattern mining phase by applying these confidences.

Fig. 4 depicted processing time for MIT dataset. We can notice that the processing time increases for confidences greater than 70% and are low for confidences that give maximum values of accuracy rate. This is explained by the fact that when applying these confidences that give maximum values of accuracy rate at the uncertain sequential pattern mining phase, we obtain only useful and reliable frequent pattern. Other confidences give not only useful and reliable frequent pattern but also unreliable frequent pattern and then need more processing time. Table 8 shows a comparison of our results with other methods.

Table 7: Activities labeled.

Number of Examples per Class		
Activity	**Subject 1**	**Subject 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
Watching TV	-	15
Bathing	18	-
Going out to work	12	-
Dressing	24	-
Grooming	37	-
Preparing a beverage	15	-
Doing laundry	19	-
Cleaning	8	-

Table 8: Comparison of results with MIT dataset.

	Approach	Result
Proposed Method	(Uncertain SPM)	Subject 1: 99.38% Subject 2: 100%
(Raeiszadeh and Tahayori, 2018)	(UP-DM+RandomForest)	Subject 1: 97.45% Subject 2: 91.37%
(Tapia et al., 2004)	(Naive Bayes Classifier)	Subject 1: 60.6% Subject 2: 41.09%

5 CONCLUSION

We have used a sequential pattern mining algorithm from probabilistic and temporal databases to bring out typical activities in the smart home. By considering of sensor uncertainty, the recognition focuses on reliable parts of the sensor data. We use temporal relationships between events for a more accurate characterization/classification of frequent activities. The future work is the development of an online adaptive unsupervised learning of activities in the context of uncertain observation and temporal constraints.

REFERENCES

Aggarwal, C. C. (2009). Managing and mining uncertain data.

Aïzan, J., Motamed, C., and Ezin, E. (2020). Activity mining in a smart home from sequential and temporal databases. In *In Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*.

Hassani, M., Beecks, C., ows, D. T., and Seidl, T. (2015). Mining sequential patterns of event streams in a smart home application. In *The LWA 2015 Workshops: KDML, FGWM, IR, and FGD*.

Li, L., Li, X., Lu, Z., Lloret, J., and Song, H. (2017). Sequential behavior pattern discovery with frequent episode mining and wireless sensor network. In *Communications Magazine*. IEEE.

Li, Y., Bailey, J., Kulik, L., and Pei, J. (2013). Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases. In *IEEE International Conference on Data Mining (ICDM)*. IEEE.

Menaka, J. and Gayathri, K. S. (2013). Activity modeling in smart home using high utility pattern mining over data streams. In *The Journal of Computer Science and Network*.

Moutacalli, M. T., Bouzouane, A., and Bouchard, B. (2012). Unsupervised activity recognition using temporal data mining. In *The First International Conference on Smart Systems, Devices and Technologies*.

Muzammal, M., Gohar, M., Rahman, A. U., and Qu, Q. (2017). Trajectory mining using uncertain sensor data. In *The Journal of IEEE Access*. IEEE.

Muzammal, M. and Raman, R. (2010). On probabilistic models for uncertain sequential pattern mining. In *International Conference on Advanced Data Mining and Applications*.

Raeiszadeh, M. and Tahayori, H. (2018). A novel method for detecting and predicting resident's behavior in smart home. In *6th Iranian Joint Congress on Fuzzy and Intelligent Systems*. IEEE.

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

Schweizer, D., Zehnder, M., Wache, H., and Witschel, H. (2015). Using consumer behavior data to reduce energy consumption in smart homes. In *14th International Conference on Machine Learning and Applications*.

Singh, S. and Yassine, A. (2017). Mining energy consumption behavior patterns for house holds in smart grid. In *Transactions on Emerging Topics in Computing*. IEEE.

Suciu, D. and Dalvi, N. N. (2005). Foundations of probabilistic answers to queries. In *the ACM SIGMOD International Conference on Management of Data*.

Suryadevara, N. (2017). Wireless sensor sequence data model for smart home and iot data analytics. In *First International Conferenceon Computational Intelligence and Informatics, Advances in Intelligent Systems and Computing*.

Tapia, E. M., Intille, S. S., and Larson, K. (2004). Activity recognition in the home setting using simple and ubiquitous sensors. In *Pervasive Computing*.

Yang, J., Wang, W., Yu, P. S., and Han, J. (2002). Mining long sequential patterns in a noisy environment. In *ACM SIGMOD international conference on Management of data*.

Zhang, B., Lin, J. C., Fournier-Viger, P., and Li, T. (2017). Mining of high utility-probability sequential patterns from uncertain databases. In *The Journal of PLoS One*.

Zhao, Z., Yan, D., and Ng, W. (2014). Mining probabilistically frequent sequential patterns in large uncertain databases. In *The Journal of IEEE Transactions on Knowledge and Data Engineering*. IEEE.