Unsupervised Segmentation of Human Habits in Smart Home Logs Through Process Discovery*

Lucia Esposito, Silvestro Veneruso, Francesco Leotta, Flavia Monti, Jerin George Mathew, and Massimo Mecella

Sapienza Università di Roma
{surname}@diag.uniroma1.it

Abstract. Smart homes represent examples of cyber-physical environments realizing the paradigm known as *ambient intelligence*. An information system supporting ambient intelligence takes as input raw sensor measurements and analyzes them to eventually make decisions following final user preferences and needs. Unfortunately, algorithms in this research area are mostly supervised, thus requiring a manual labeling of training instances usually involving final users in annoying and imprecise training sessions. In this paper, we propose an unsupervised approach allowing, given a sensor log, to automatically segment *human habits* on a temporal basis, by applying a bottom-up discretization strategy to the timestamp attribute of the sensor log.

Keywords: Ambient intelligence · habit mining · unsupervised log segmentation · process mining.

1 Introduction

Smart spaces connect computing devices and other smart devices to everyday settings and tasks, realizing the paradigm known as *ambient intelligence*. The aim is (i) to understand what is happening in the environment, i.e., *context extraction*, and (ii) eventually use this information to trigger automated actions, following final user preferences and needs, i.e., *decision making*.

In particular, this process is supported by a set of models [12] representing the contextual situation of the monitored environment, the *activities* performed by the inhabitant(s) (e.g., cleaning the house) and his/her/their needs and *habits* (e.g., what the user does every morning between 08:00 and 10:00). Such models, in the vast majority of cases, are specifically trained for a specific home and/or inhabitant(s). Unfortunately, the vast majority of the state-of-the-art algorithms are supervised, thus requiring a manual labeling of training instances usually involving final user(s) in annoying and imprecise training sessions.

In [10] authors argued that human activities and habits can be modelled by using business process modelling approaches. With respect to this, the term activity can be confusing for the Business Process Management (BPM) [4] research community, which often use it to denote tasks in business processes. In

this paper, we use instead the terms *activity* and *habit* to indicate human daily processes/routines, with habits that can be represented in terms of either composing activities (i.e., hierarchically) or atomic actions.

In this paper, we propose a methodology that allows to automatically segment *human habits* by applying a classical bottom-up discretization strategy. Such class of algorithms find the best division of a continuous attribute by iteratively merging contiguous sub-ranges (also called bins) following a quality evaluation heuristic based on measures computed over the process models automatically mined over the intermediate bins. Obtained bins allow to automatically segment logs based on the time of day. Human habits models can be used, in conjunction with condition mining techniques, to eventually anticipate user actions.

The paper is organized as follows. Section 2 introduces background concepts and related works. Section 3 describes the proposed approach. Finally, Section 4 concludes the paper with discussion and future works.

2 Background and Related Works

As already discussed in Section 1, the vast majority of approaches in ambient intelligence are inherently supervised, thus requiring a segmented and labeled dataset at least at training time, whereas windowing techniques can be used to roughly segment the sensor stream at runtime [6]. In any case, there are some works which can be defined as fully unsupervised.

For instance, in [2] authors propose the APUBS algorithm to automatically extract *Event-Condition-Action* (ECA) rules by considering the typology of the sensors involved in the measurements and the time relations between their activations. An ECA rule has the form "ON event IF condition THEN action", thus automatizing the execution of an action, as soon as a specific event is detected and if and only if certain contextual conditions are met.

Our approach differs from APUBS as we are not directly discovering enactment rules. Instead, we discover the process models underlying each human habit. These models can be then employed to derive enactment rules, which are strongly related to human habit.

In [3] authors proposed instead an approach based on the minimum description length (MDL) principle to automatically extract activity patterns. The algorithm takes as input a dataset consisting of a sequence of sensor events witnessing human interactions with the environment.

Differently from the latter approach, we focus on habits instead of activities. Additionally, in [3] patterns are extracted with the sole goal of recognizing them at runtime, without providing neither a visual analysis tool nor a structured description of human routines.

2.1 Process Mining and Ambient Intelligence

Process mining [1] is a fairly recent research discipline which combines techniques derived from Machine Learning and Data Mining with approaches used

in Business Process Management (BPM) [4], such as process modeling and process analysis. Its main goal is to extract meaningful information from event logs.

Process discovery, in particular, is a process mining technique used to discover the process model describing the behavior recorded in the event log. Thus, it takes in input an event log and automatically generates the correspondent process model. In this paper, we propose to apply process discovery to smart homes [5,10] for log-segmentation purposes. In particular, we employ the Inductive Miner [9], which produces as output a process model represented using the Petri net formalism, i.e., a directed graph composed of arcs and nodes.

The structure of the Petri nets mined through process discovery can be analyzed providing several different quality measures. Throughout this paper, we are interested in particular in structuredness and simplicity. *Structuredness* [7] is a measure obtained by iteratively disassembling the observed model in small sub-models, assigning a score to each of them and combining these scores. In particular, the score will be lower for patterns perceived as simple (e.g., *sequences*, *while* and *choice* patterns) and higher for the complex ones. *Simplicity* is a metric that depends only on the size and the structure of the model, without considering its behavior. With a lower value of simplicity, we expect that the related Petri net has a complex structure, i.e., the number of arcs is much bigger than the number of total nodes in the net; in general this leads to models that are not easy to read.

However, in order to apply process mining techniques to raw sensor measurements coming from smart environments, the sensor log must be properly converted into an event log.

Authors of VPM - Visual Process Maps system [11], uses TRACLUS [8] for this conversion task. Originally designed for describing trajectories of hurricanes, TRACLUS is used to segment the log into subtrajectories with a homogenous velocity. Each of these trajectory is then classified into three categories with labels `MOVEMENT,AREA,STAY` by considering information features such as duration, velocity and heterogeneity.

The classification allows to replace sequences of measurements with human actions consisting of a category and a location, the latter inferred by the position of the corresponding sensors in the environment (e.g., `<STAY Kitchen_table>`).

In this paper, we apply the same technique used in VPM to turn sensor measurements into human actions. Then, we provide a principled method to segment the dataset on human habits. In this way, for each habit, an event log is provided that can be used to mine the human process underlying the habit.

3 Proposed Approach

As discussed in Section 2, our methodology applies the same approach proposed by the VPM system [11] to turn a raw unsegmented sensor log $\mathcal{S}$ into an unsegmented action log $\mathcal{A}$.

The resulting action log is a sequence of tuples $\langle day, start\_ts, end\_ts, action \rangle$ containing information about the day and timestamp of the related action, i.e.,

Algorithm 1 Pseudocode of our proposed discretization algorithm based on Chi-Merge.

Input: finite set *intervals* of chronologically ordered time ranges; an integer $minN$ denoting the minimum number of intervals to be returned; a float $minScore$ denoting the minimum acceptable score for a merged interval;
Output: discretized intervals

```
1: procedure DISCRETIZATION(A, intervals, minN, minScore)
2:     while len(intervals) > minN do
3:         max ← 0
4:         index ← null
5:         for i ∈ [0, len(intervals) − 2] do
6:             pair ← concat(intervals[i], intervals[i + 1])
7:             pn ← inductiveMiner(eventLog(A, pair))
8:             score ← 100 × pn.simplicity − pn.struct
9:             if score > maxScore then
10:                max ← score
11:                index ← i
12:            end if
13:        end for
14:        if maxScore < minScore then
15:            return intervals
16:        end if
17:        intervals = merge(intervals, index, index + 1)
18:    end while
19:    return intervals
20: end procedure
```

in our case, a movement action labeled as `MOVEMENT,AREA` or `STAY` followed by the identifier of the location within the monitored smart home.

Once we have an action log, we can then proceed to segmentation. Our approach is based on Chi-merge [13]. The original Chi-merge algorithm starts by dividing the entire range of an attribute at the finest level of granularity possible. Then, it iteratively merges adjacent bins showing the highest value of the χ^2 statistical measure until merging bins do not improve the homogeneity of bins.

We start our segmentation by dividing the entire range of the time of the day attribute (i.e., 00:00 - 24:00) in bins of constant width. E.g., if 15 minutes is chosen as minimum bin width, the time of the day attribute will be divided into $24 * (60/15) = 96$ bins. Each bin is associated to the correspondent event sub-log (e.g., all the actions in a specific day happening from 00:00 to 00:15).

Once this initial segmentation is provided, Algorithm 1 is executed. The algorithm takes as input an ordered array of time intervals (96 in the previous example), a minimum number `minN` of bins that must be returned from the algorithm, and a minimum score `minScore` required to stop the discretization.

The algorithm finds the best possible subdivision of the day in a set of habits. It ensures (see row 2) that no less than `minN` bins are produced. At each iteration

(rows from 3 to 17), the algorithm iterates over all of the intervals, and for each pair of adjacent bins (see row 6) applies the inductive miner (see row 7) to the event log over the concatenation of those bins.

For each couple of adjacent bins, we obtain a Petri net *pn* from the execution of the inductive miner. For each of these Petri nets, we compute a *score*. This score is obtained (see row 8) starting from the *simplicity* and *structuredness* measures introduced in Section 2. Simplicity in particular is multiplied by a factor of 100, which has been empirically chosen to make the two quality measures uniform. We keep track of the maximum score computed and of the corresponding couple of adjacent bins (rows from 9 to 12).

After all the adjacent couples have been considered, if the maximum score computed is above the `minScore` threshold, the `intervals` array is modified accordingly by merging the adjacent bins corresponding to the maximum score. Otherwise the algorithm terminates as additional merging are not convenient.

It is worthwhile to note how the algorithm always terminates. Merging stops either if it is not convenient to keep merging bins or if a minimum number of bins is reached (note that iteration by iteration the number of bins always decreases by one).

After the algorithm terminates, the `intervals` variable contains a set of bins each corresponding to a habit, as defined in Section 1. Here the rationale is that bins will be merged only if the resulting Petri net is simpler and less structured, meaning that the process model of the underlying habit is easy to read. The principle driving the segmentation is then very similar to the *Occam razor*.

4 Conclusions and Future Work

In this paper, we have introduced a way to automatically segment a sensor log into habits by defining a new heuristic for the seminal Chi-merge discretization method based on the structuredness and simplicity measures of Petri nets automatically discovered at each merging step by applying inductive miner. The final segmentation can then be used to obtain Petri nets describing human habits. The proposed approach is, to the best of our knowledge, the first to provide automatic segmentation of a full smart space log. The proposed approach has been validated with a state-of-the-art dataset provided within the CASAS project (see `http://casas.wsu.edu/datasets/`).

At the current stage, the proposed solution still suffers of some limitations. In first place, the proposed algorithm requires as input an action log $\mathcal{A}$, which can be hard to obtain. In this paper, in order to obtain $\mathcal{A}$ we applied the method suggested in [11], which only supports Presence InfraRed (PIR) sensors. By only using this kind of sensors, we limit the number of different actions that can be recognized, which in turn limits the level of details of the Petri nets that are used to extract quality measures driving the segmentation process. It can be argued that a sensor log richer in terms of sensor could lead to better results.

In principle, once the habit identification is performed, any kind of process discovery algorithm can be employed, as Petri nets are only required when qual-

ity measures are computed. A possible application of our system is to define Petri nets that can be used for automation. Future work includes the definition of a condition mining algorithm able to turn this methodology into a fully functioning system able to provide automation in addition to provide only Petri nets for analysis.

Finally, in this paper, we have only discussed temporal segmentation targeted at defining *habits*. In the next future we would like to implement a segmentation based on *activities*. Unsupervised modelling of activities would allow a finer grain control over human routines.

References

1. van der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer, 2 edn. (2016)
2. Aztiria, A., Augusto, J.C., Basagoiti, R., Izaguirre, A., Cook, D.J.: Discovering frequent user–environment interactions in intelligent environments. Personal and Ubiquitous Computing **16**(1), 91–103 (2012)
3. Cook, D.J., Krishnan, N.C., Rashidi, P.: Activity Discovery and Activity Recognition: A New Partnership. IEEE Transactions on Cybernetics **43**(3), 820–828 (2013)
4. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A., et al.: Fundamentals of business process management, vol. 1. Springer (2013)
5. Janiesch, C., Koschmider, A., Mecella, M., Weber, B., Burattin, A., Di Ciccio, C., Fortino, G., Gal, A., Kannengiesser, U., Leotta, F., et al.: The internet of things meets business process management: A manifesto. IEEE Systems, Man, and Cybernetics Magazine **6**(4), 34–44 (2020)
6. Krishnan, N.C., Cook, D.J.: Activity recognition on streaming sensor data. Pervasive and mobile computing **10**, 138–154 (2014)
7. Lassen, K.B., van der Aalst, W.M.: Complexity metrics for workflow nets. Information and Software Technology **51**(3), 610–626 (2009)
8. Lee, J.G., Han, J., Whang, K.Y.: Trajectory clustering: a partition-and-group framework. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. pp. 593–604 (2007)
9. Leemans, S.J., Fahland, D., Van Der Aalst, W.M.: Process and deviation exploration with inductive visual miner. BPM (Demos) **1295**(8) (2014)
10. Leotta, F., Mecella, M., Mendling, J.: Applying process mining to smart spaces: Perspectives and research challenges. In: Persson, A., Stirna, J. (eds.) Advanced Information Systems Engineering Workshops. pp. 298–304. Springer International Publishing, Cham (2015)
11. Leotta, F., Mecella, M., Sora, D.: Visual process maps: A visualization tool for discovering habits in smart homes. Journal of Ambient Intelligence and Humanized Computing pp. 1–29 (2019)
12. Leotta, F., Mecella, M., Sora, D., Catarci, T.: Surveying human habit modeling and mining techniques in smart spaces. Future Internet **11**(1), 23 (2019)
13. Liu, H., Hussain, F., Tan, C.L., Dash, M.: Discretization: An enabling technique. Data mining and knowledge discovery **6**(4), 393–423 (2002)