Temporal Mining Mixture Model for Residential Occupancy Prediction

Huijuan Shao[1,2], Yaowei Li[3], Fei Li[2], Erin Griffiths [4], Kamin Whitehouse[4], Naren Ramakrishnan[1,2]
[1]Discovery Analytics Center, Virginia Tech, Blacksburg, VA 24061
[2]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061
[3] Game source, McLean, VA 22101
[4]Department of Computer Science, University of Virginia, VA 22904

ABSTRACT

Conserving energy and optimizing its use has been a long standing challenge. Apart from the monetary benefits associated with successfully tackling these problems, saving energy has a significant positive environmental impact. One potentially fruitful approach would be to automatically adjust the HVAC of residential buildings based on occupancy. In this study, we mine people's individual energy activity profiles to predict the occupancy of residential buildings. A novel hybrid method is proposed that uses episode mining for target event detection and a mixture of episode-generating HMM (EGH), combined with standard kNN approaches. We conclude by demonstrating how this hybrid approach consistently yields the best results.

KEYWORDS

occupancy prediction, motif mining, hidden Markov model

1 INTRODUCTION

Modeling the activity of daily life (ADL) has become a fertile research topic, satisfying the demand for a comfortable home life at a lower cost. Since heating and cooling our living spaces consumes $\sim$53% of the total electrical used by an average household, automating the operation of HVAC devices to minimize energy usage is clearly important. One of the crucial components required to achieve this goal is the ability to accurately model and predict the occupancy of a home. Supervised learning approaches based on the analysis of indoor temperature [5], GPS data from smart phones [6], historical electricity consumption data [3] and sensor data generated by tracking indoor activities [1, 13] have all been shown to be effective ways to approach this prediction problem. Prediction of occupancy using sensor data has been broadly studied generally by capturing daily activities like room occupancy within the house, usage of electrical devices, and usage of water systems to model occupancy [2, 3, 9] and using these results to automate the control of the HVAC system.

Although supervised learning techniques—kNN [13], neural network [9] and Markov model [3]—have all shown to be effective, the fine-grained details of household activities represented as a time series have not yet been fully utilized. Daily activities such as waking up, cooking, washing, and commuting to and from work/school have different patterns based on the day of week. For instance, the schedule on a working day is significantly different from that on a weekend or a holiday. Thus, this scenario leads itself to episode mining analysis such as the approach adopted in this study to predict household occupancy. Applying this strategy of episode mining for occupancy prediction has three advantages. First, episode mining, a temporal mining approach, mines the data according to the time distribution for each type of activity. Second, it can be used to build an activity scenario and connect different episodes with a probabilistic hidden Markov model (HMM). Unlike earlier models, the time and order of each kind of activity can be fully utilized using this approach. Third, the algorithm predictions are based on a scenario-based probabilistic model episode generative HMM (EGH). The prediction accuracy is consequently a marked improvement over that achieved by existing models.

The contributions of this study can be summarized in terms of the way it addresses the three questions below.

(1) How can we mine for meaningful scenarios? Episode mining can mine many frequent episodes, but not all these episodes will be useful for occupancy prediction. By narrowing down the episodes included in the analysis according to their start state, end time, event dwelling time and the gap between two activities, we can interpret these episodes more accurately thus pinpoint the episodes more likely to convey useful information .
(2) How can we predict occupancy more accurately? Our dataset comprises detailed information regarding the various activities of a household tracked as a time series on a daily basis. Thus our episodes contain richly detailed information based on the occupancy/unoccupancy status of the household. Since we are mining episodes from this data, the accuracy of occupancy prediction improves significantly.
(3) Can the new approach proposed here help reduce electricity usage in the home? The prediction gives at least 15 minutes advance notice of the time a person leaves or returns home. By connecting this prediction result to an automatic HVAC control system, the HVAC can be turned on or off ahead of the occupancy change. Since this means that the HVAC does not heat or cool the home when there is no one home, this can substantially reduce the household's electricity usage.

UrbComp'17, August 14, 2017, Halifax, Nova Scotia, Canada
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

Accurately predicting whether a home is occupied is a difficult task. People in the same household will have different daily schedules: some go out to work while others stay at home for a period of time. A great deal of research has been done to track the activities of people thus infer a home's occupancy, with researchers utilizing data from sensors, smart phones, the calendar, and weather information. Most of the approaches that model and predict occupancy are primarily based on the use of sensor data to detect conditions such as room occupancy, use of electrical appliances, and water usage. Supervised learning approaches such as kNN, neural networks, rule-based models, and Markov chain models have all been used to model and predict building occupancy [1–3, 9, 13]. Applying the kNN supervised learning algorithm and monitoring sensor data for a portion of the day, Scott et al. successfully predicted an entire day's occupancy [13]. A neural network approach using a binary time series based on occupancy/unoccupancy along with exogenous input network (NARX) has also been proposed [9]. Mahmoud et al. tackled the problem by presenting a non-linear autoregressive model with an exogenous input (NARX) network. Several Markov chain models, including the blended Markov chain, closest distance Markov chain, and moving-window Markov chain, have been suggested [3]. A mixture of multi-lag Markov chains has also been used to predict the occupancy of single-person offices [10], with the authors comparing the performance of their model to those achieved by the Input Output Hidden Markov Model, First Order Markov Chain and the NARX neural network.

A recent survey [5] compared the major occupancy prediction algorithms utilizing the LDCC dataset [4], which was collected by GPS and other sensors. They concluded that time-based presence probability [7] performs slightly better than the preheat kNN approach [13]. Nevertheless, since the preheat kNN approach [13] is more widely applicable, as can be used for both GPS and sensor datasets, we chose to use it as a baseline method for comparison in the present study.

3 PROBLEM FORMULATION

Given M time series, each time series $X^{(m)} = X_1^{(m)}, ..., X_t^{(m)}, ..., X_T^{(m)}$ represents a sequence describing the room occupancy of person m inside a home over K days, where $X_t \in s$ denotes that X belongs to a finite room set s at the sequence number of t, and $m \in \{1, ..., M\}$. Let Z denote that the home is unoccupied and $Z \in s$. We can then predict whether person m will stay at home for the rest of a day from time T, i.e. during $T+1$, $T+2$, ..., ΔT,

$$\hat{Y}^{(m)} = \hat{Y}_{T+1}^{(m)}, ..., \hat{Y}_{T+\Delta T}^{(m)} \tag{1}$$

where $Y_{T+\Delta t}^{(m)} = Z$ if person m does not stay at home at time $T + \Delta t$; otherwise, $Y_{T+\Delta t}^{(m)} \neq Z$. If any person m stays at home $Y_{T+\Delta t}^{(m)} \neq Z$, then this house is occupied $Y_{T+\Delta t} \neq Z$.

4 TEMPORAL MINING MIXTURE MODEL

We use a three-pronged approach to tackle the problem of mining and predicting unoccupancy, as shown in Figure 1. Given a time series representing the indoor activities of a person over a period of time, first, we use an episode mining algorithm to discover frequent episodes from previous days' data, enabling us to connect each episode with an EGH and build a mixture EGH model. Based on this mixture model, we can now predict when each person will leave and return to the house. If all the people leave, then the house is unoccupied.

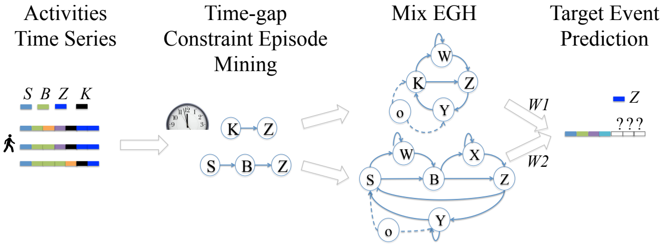


Figure 1: Occupancy Prediction Framework.

Episode An episode is a collection of ordered events. Here an episode refers to a set of ordered events which are highly relevant to the occupancy status inside a building. For instance, let us represent 'S' as sleep, 'K' as kitchen, and 'Z' as going out. If an episode $S \rightarrow K \rightarrow Z$ is found, this can be interpreted as a person getting up, going to the kitchen for breakfast, and then leaving the house. An episode α is composed of a series of ordered events $\alpha = \langle X_1, .., X_t, ...X_T \rangle$, where X_t denotes that X occurs at a sequence time of t. The event X_t may be a point event or extend over a longer time, in which case it is referred to as a dwelling event with a start time $X.start$ and end time $X.end$. In this paper, X denotes a dwelling event and indicates the room where a person is located inside the building, i.e., this building is *occupied*. Since Z denotes that a room is unoccupied, $Z.start$ is the point at which a person or all the people inside the building leave, and $Z.end$ is the point at which one or more of the people come back.

4.1 Time-gap Constraint Episode Mining

Episode Mining Episode mining [11] uses a non-overlap mining approach to identify frequent episodes. It has been applied to energy disaggregation to help conserve energy in buildings [14] in sustainability research. However, in contrast with previous research, the events in the current application of occupancy prediction take into account the dwell time of an event. This forces us to extend the above two episode mining algorithms [8, 12] and enforce extra constraints. In addition to adopting an appropriate alignment for the first element in the episode mining process, time constraints must be added and gap duration constraints applied between two consecutive events inside an episode. Figure 2 shows a time-gap constraint episode mining example. Let us assume we have a frequent episode $S \rightarrow B \rightarrow K \rightarrow Z$. We must then add the time constraints to each event $\{S, B, K, Z\}$: the dwelling duration for S is 3 to 9 hours, for B it is 2 to 20 minutes, for K it is 2 to 60 minutes, and for Z it is 3-9 hours. In addition, we must specify the gap duration between any two consecutive events. The gap duration of SB is calculated as $\Delta SB = B.start - S.end$. For example, we set the maximum gap time between SB, BK, and KZ as 10 minutes, 10 minutes and 100 minutes, respectively; the minimum gap time is 0. We now have a stream composed of the sequence of dwelling

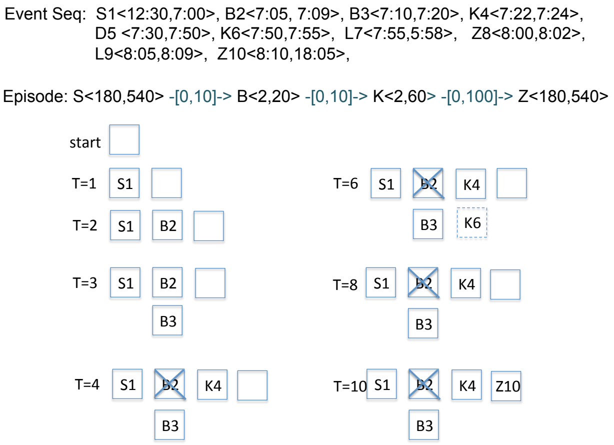

Event Seq: S1<12:30,7:00>, B2<7:05, 7:09>, B3<7:10,7:20>, K4<7:22,7:24>, D5 <7:30,7:50>, K6<7:50,7:55>, L7<7:55,5:58>, Z8<8:00,8:02>, L9<8:05,8:09>, Z10<8:10,18:05>,
Episode: S<180,540> -[0,10]-> B<2,20> -[0,10]-> K<2,60> -[0,100]-> Z<180,540>
start
T=1
T=2
T=3
T=4
T=6
T=8
T=10
S1
B2
B3
K4
K6
Z10

Figure 2: Time-gap constraint episode mining example.

events "Event seq," as shown in Figure 2. The time-gap constraint episode mining process applied to discover a frequent episode uses the following method. Let the $node$ structure denote each element in any episode, depicted as a square box in Figure 2. Let $waits$ refer to a structure that pairs with an episode and has the same length as this episode. This creates an initial $waits$ structure related to the episode of $S \to B \to K \to Z$. A $node$ structure related to S is created that waits for the first element of the episode $S\langle 180, 540\rangle$. When $T = 1$, the duration of S_1 is checked. Since S_1 is in the range of $3 - 9$ hours, S_1 passes and is put into the node structure $node$ related to S. Next a new $node$ structure is created to wait for $B\langle 2, 20\rangle$. When $T = 2$ and $T = 3$, both B_2 and B_3 are qualified in terms of the time constraints and the gap constraints; this means that the gap between S and B ΔSB should be between 0 and 10 minutes. These two nodes B_2 and B_3 are then input into the $waits$ structure. At the same time, a new $node$ structure is created for $K\langle 2, 60\rangle$. When $T = 4$, the gap between $\langle B_3, K_4\rangle$ is satisfied with a distance condition between B and K of 0-10 minutes. However, the gap between $\langle B_2, K_4\rangle$ is longer than the constraint gap, so B_2 is canceled out. Now a new Z waits for the symbol $Z\langle 180, 540\rangle$. When $T = 6$, the gap from B_2 to K_6 is too large, so K_6 is not added into the $node$ K structure in $waits$. When $T = 8$, the duration of Z_8 does not satisfy the condition of between 3-9 hours, so Z_8 is not added, but when $T = 10$, the duration of Z_{10} does meet the requirement of between 3-9 hours and its distance to $K4$ also meets the requirement of $\Delta KZ \in [0, 100]$ minutes. Thus, Z_{10} is added into the $node$ Z structure in $waits$. The mining of this episode is now complete, and we have mined the instance $S_1 B_3 K_4 Z_{10}$.

4.2 Mixture EGH

Episode Generating HMM The episode generative HMM (EGH) model is a type of HMM that connects frequent episodes, and the more frequently an episode occurs inside a sequence, the greater the likelihood of the state sequence including that episode [8]. The unique feature of EGH is that the transition matrix and emission matrix are solely decided by a noise parameter η. The noise parameter η of a frequent episode α is calculated as $\eta = \frac{T - N f_\alpha}{T}$, where T is the training data stream length, α is the frequent episode, N is the length of frequent episode α, and f_α is the frequency over time T.

In the mixture EGH model shown in Figure 3, the transition matrix for EGH is given as an example. Let us assume we have an N-node frequent episode $S \to B \to Z$, where $N = 3$. We define 2N hidden states: N episode states and N noise states, where the noise states are $\{W, X, Y\}$. An episode state transfers to another episode state with a probability of $1 - \eta$; an episode state transfers to a noise state with a probability of η; and a noise state transfers to another noise state with a probability of $1 - \eta$. To calculate the emission matrix, first let M denote the total number of symbols in the event stream. For any hidden states in the episode, M has a delta function emission. Whenever it is visited (right alignment of the first element in the episode, left alignment for the left elements in the episode), it will generate the same observation symbol. For any noise hidden states, it emits any of the symbols from the M observation symbols with a uniform distribution at probability $\frac{1}{M}$.

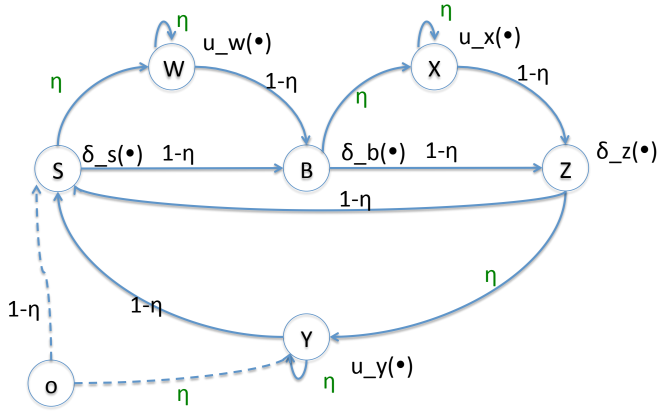


Figure 3: States Transition of Episode Generating HMM (EGH).

Theorem 1 from [8] is crucial here as it proves that the more frequently an episode occurs within a sequence, the greater the likelihood that the state sequence will include this episode. The proof for this theorem is explained in detail in [8].

THEOREM 1. *[8] Let $D_Z = X_1, ..., X_K$ represent the given sequence data, where ε is the symbol set, and the size of these symbols is M. Given two frequent N-node episodes α and β with frequencies f_α and f_β, respectively, their corresponding EGH will be Λ_α and Λ_β. The most likely state sequences for episodes α and β are q^*_α and q^*_β. The noise parameters for these two EGH are η_α and η_β. Assuming that both of these noise parameters are less than $\frac{M}{M+1}$, we have (1) if $f_\alpha > f_\beta$, then $P(D_Z, q^*_\alpha|\Lambda) > P(D_Z, q^*_\beta|\Lambda)$ (2) if $P(D_Z, q^*_\alpha|\Lambda) > P(D_Z, q^*_\beta|\Lambda)$, $f_\alpha > f_\beta$*

Mixture Model The mixture EGH model is fully discussed in previous work [8], which shows how the model gives different weights to each EGH when predicting a target event. We can assume we know whether an episode occurs on a certain day. Let $D_Z =$

$\{X_1, ..., X_K\}$ denote the data set for K days. $F = \{\alpha_1, ...\alpha_J\}$ denotes the frequent episodes in the dataset D_Z. An EGH Λ_{α_j} is associated with frequent episode α_j. Λ_Z denotes a mixture EGH model. The likelihood of D_Z under the mixture model is written as Equation (2).

$$\begin{aligned} Pr(\Lambda|Z) &= \prod_{i=1}^{K} P[X_i|\Lambda_Z] & (2)\\ &= \prod_{i=1}^{K}(\sum_{j=1}^{J}\theta_j P[X_i|\Lambda_{\alpha_j}]) & (3)\end{aligned}$$

where θ_j is the mixture coefficient of Λ_{α_j} and is subject to $\sum_{j=0}^{J}\theta_j = 1$

The parts inside Equation (2) are additive; the coefficients θ are computed by an EM algorithm. During the initialization part of the EM algorithm, the episode frequency over the time series T over K days is calculated and specific frequent episodes ending with target event 'Z' selected. Optionally, we could add special constraints on episodes, starting with certain types of event 'S'. In the expectation step, one key part is the likelihood value of each episode α_j in time series X_i. The likelihood value is computed as Equation (4), after which Bayes rule is applied to compute the new coefficient θ_{new}.

$$Pr(X_i|\Lambda_{\alpha_j}) = (\frac{\eta_{\alpha_j}}{M})^{|X_i|}(\frac{1-\eta_{\alpha_j}}{\eta_{\alpha_j}/M})^{|\alpha_j|f_{\alpha_j}(X_i)} \qquad (4)$$

In the maximization step, we update the objective value based on Equation (2) until it converges, i.e., until the difference between two consecutive objective values is smaller than a given threshold.

4.3 Predict When the Target Event Occurs

Although target event prediction has been studied in [8], researchers only sought to predict whether a target event would occur, rather than when it would happen. Our occupancy prediction algorithm enriches the previous event prediction algorithm by breaking it up into three sub-problems: whether the target event un-occupancy Z will appear; when the target event Z will start; and when the target event Z will end.

Since the solution to the first sub-problem is similar to that addressed in previous work, this sub-section focuses on the last two sub-problems. After obtaining the results for the first sub-problem, we can assume we already know that the target event Z will indeed happen, and so the next step is to predict when the person will leave or come back. The departure time corresponds to the start time of the dwelling event, $Z.start$, and the return time refers to the end time of the dwelling event, $Z.end$.

After running the episode mining algorithm and mixing the EGH model, we have obtained all the frequent episodes $F = \langle\alpha_1, ..., \alpha_J\rangle$, the corresponding EGH $\Lambda_{\alpha_j}, j = 1...J$ with noise parameter η_j, and the mixture models Λ_Z with coefficients θ_j. We can use the coefficient of these mixture models to predict the departure and return times for target event Z. Each day is separated into three phases: 1) before a person gets up; 2) after the person has got up but before he/she goes out; and 3) after the person goes out but before he/she comes back.

(1) Usually before a person gets up, there is only one frequent episode, namely 'SZ'. The start time and end time of Z depends on 'S'. Therefore $Z.start$ and $Z.end$ are calculated based on the probability density function of the departure and return times for previous days.
(2) After the person gets up, if he/she has engaged in a number of activities at home, several frequent episodes will be mined before the person leaves home. If there are several frequent episodes ending with Z, the leave time and return time of each episode is checked to determine whether they fall within the range of probability density function (PDF) values observed previously. If the answer is yes, the mean value of these episodes is recorded. Since each episode generates an EGH, the mixture EGH model computes a weight for each EGH as a coefficient. The departure time $Z.start$ and return time $Z.back$ are the weighted mean departure and return times of these frequent episodes.
(3) After a person has left home, we already know when the departure time $Z.start$. If the person has come back nothing needs to be predicted, but if the person has not come back, the return time $Z.end$ is the weighted historical return time of the mined frequent episodes, viz. the probability density function of the return time is based on the time-constraint for the departure.

5 EXPERIMENTAL RESULTS

We conduct experiments on three datasets, where each dataset is obtained by monitoring the 24-hour activities of two adults in a house via RFID. All these activities occur in twelve rooms; the house's basement, bathroom, bedroom, dining room, hallway, kitchen, living room, mudroom, nursery, outside-front, outside-back, and upstairs. The dataset comprises a set of events in the form of timestamped room occupancy data points. For instance, an event can correspond to person 1 being in the kitchen at 7:00 am. The three datasets are summarized in Table 1. Here, *unoccupancy* of a person is defined as one of the following conditions: the person has left the *outside-front* or *outside-back* of the house for more than 30 minutes; the person has remained in the living room or dining room for more than 9 hours with no other activities; or the gap between any two events is more than 30 minutes. Since our research goal is to automate switching the HVAC system on or off at least 30 minutes before a change in occupancy occurs, the first and third constraints are in place. We are only interested in events where the *unocupancy* period is for an extended duration (> 30 minutes). The second constraint comes from our observation that if a person appears to remain in the same room for more than 9 hours without moving to another room, this usually means that that person has gone out but left the RFID equipment at home. We also delete events with a duration of less

Table 1: Datasets summary.

Dataset	Number of entries	Period(day)	Start date
study 10	6596	12	02/10/2014
study 11	1696	10	01/29/2014
study 14	3453	13	12/09/2013

than 2 minutes since these correspond to the individual walking back and forth across rooms and generally do not contribute to meaningful episodes. We conduct four types of experiments to compare the four approaches; kNN, mixture EGH, PDF, and support vector regression (SVR). For each dataset, we use $2/3$ of the data for training and the remaining $1/3$ of the data as the test set. Following the approach in [13], we organize each day's data into 96 15-minute periods. For the test data, we assume that we only know some of the 15-minute periods. Our target is thus to predict the occupancy for the reminder of the day, or for 30 minutes ahead.

5.1 Occupancy Prediction for Individuals

The individual occupancy prediction results for the datasets Study 10 and Study 11 are summarized in Table 2. In Study 10, the mixture EGH performs better than kNN for the occupancy prediction for person 2 on 02/17/2014. The mixture EGH also outperforms kNN for both subjects on 02/20/2014. However, for person 1 on 02/19/2014, kNN works slightly better. When checking the original data for the test date, we discover that the activities on this date were very similar to the historical activities in the training data, which leads us to the conclude that when the test data is highly similar to the historical data, the kNN approach may perform a little better. In the Study 11, mixture EGH achieves higher precision, recall and f-measure scores on 02/05/2014, but not on 02/04/2014. We again analyze the original data to find the reason for kNN's better performance, and find a deviation from the normal pattern, since both individuals went to sleep later than usual that day (after 12:00am). Also, before they went to sleep, person 1 stayed in the kitchen for around two hours. The frequent episode KZ, which represents 'kitchen-unoccupied', usually occurs in the morning instead of around midnight. However, the mixture EGH model still assumes that the KZ pattern occurred during the morning, so the prediction results are not accurate in this case. Since kNN ignores this fine granular activity pattern in the household and only considers the occupancy status over the previous five most similar days, its performance is better. Generally speaking, the mixture EGH helps predict when a person would leave home and his/her sleep period, delivering a performance comparable to that achieved using the kNN approach. For all these experiments, the SVR approach performs the worst because of the limitations of this approach. SVR uses several of the most recent data points for the occupancy state as the training vector for the prediction of the next occupancy state. Here we set eight past data points as the predictor, but other features such as the time of day and day of week can not be utilized fully.

We also conduct experiments to predict the participants' *rest-of-day* occupancy at different times. Figure 4 illustrates a person's occupancy prediction result from Study 10. There are three sub-figures, with each sub-figure describing the precision, recall, and f-measure for person 1 on 02/20/2014. The blue line represents the mixture EGH model, the green line represents the PDF model, and the red line denotes the kNN model. The x-axis is the number of known 15-minute periods for the test day. For instance, at $x = 20$, we already know $20 * 15$ minutes' data and need to predict whether the home will be occupied during the remaining 76 time periods. The y-axis denotes the precision, recall and f-measure values in the three sub-figures from the top down. The first sub-figure shows that mixture EGH has the highest precision, recall and f-measure on test day 02/20/2014 for occupancy prediction. The other two baseline approaches are comparable, except that kNN performs better than PDF when the person returned home after slot 72. Looking at the original data, we find that person 1 actually came home later than usual in the data used for the training dataset.

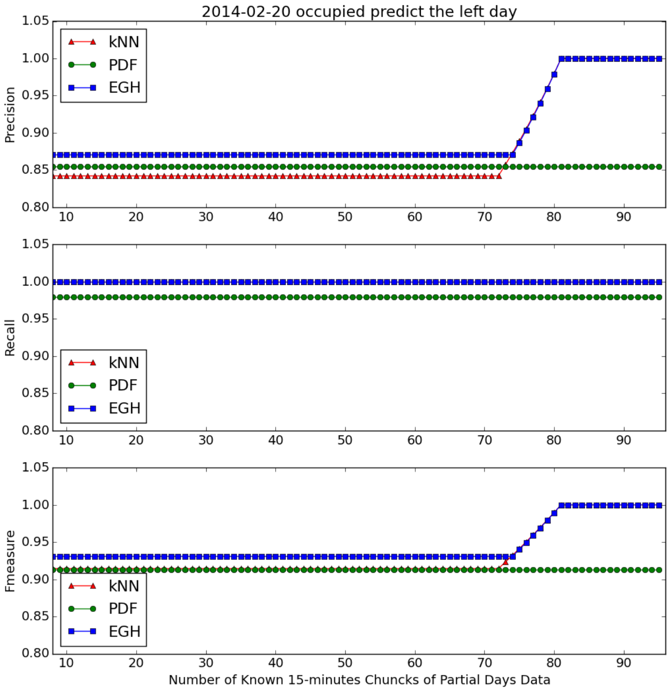


Figure 4: Occupancy prediction precision, recall and f-measure comparison of three approaches of person 1 on 02/20/2014 on Study 10.

5.2 Occupancy Prediction of Residential Buildings

Based on the individual prediction results, we go on to deduce when the house will be occupied using logic OR operations on the prediction results for the two study participants. The whole house occupancy prediction results are listed in Table 2 and marked in bold. In Study 10, the precision, recall, and fmeasure values of the whole house are 0.92, 0.92 and 0.91, respectively, higher than the values from either the kNN approach of 0.91, 0.90 and 0.91, or the SVR approach's 0.79, 0.74 and 0.74. Similarly, the mixture EGH model outperforms kNN in Study 11. However, it is important to note that in Study 11 on 02/04/2015, EGH does not perform as well as kNN on individuals but performs a little bit better than kNN, and much better than SVR, for the occupancy prediction for the whole house. This is likely because the activities of the two people inside the home were not synchronized. The mixture EGH model is able to predict the occupancy for each person and capture each person's activities more accurately.

Table 2: Precision Recall F-measure Comparison of Individual and Whole House Occupancy Prediction in Study 14.

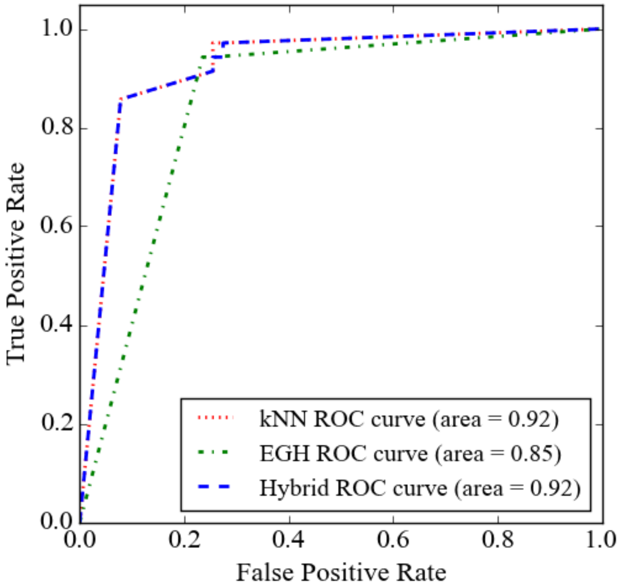
Dataset	Date	Person	EGH			kNN			SVM		
			precision	recall	fmeasure	precision	recall	fmeasure	precision	recall	fmeasure
study10	02/17/2014	person2	**1.00**	**1.00**	**1.00**	0.99	0.98	0.98	0.71	0.76	0.71
	02/19/2014	person1	0.98	0.99	0.98	**0.99**	**0.99**	**0.99**	0.71	0.76	0.70
	02/20/2014	person2	**0.93**	**0.92**	**0.92**	0.92	0.91	0.90	0.72	0.77	0.72
	02/20/2014	person1	**0.95**	**0.94**	**0.94**	0.94	0.93	0.93	0.71	0.77	0.72
	02/20/2014	**wholehouse**	**0.92**	**0.92**	**0.91**	0.91	0.89	0.91	0.79	0.74	0.74
study11	02/04/2014	person2	0.93	0.93	0.92	**0.95**	**0.95**	**0.95**	0.71	0.77	0.72
	02/04/2014	person1	0.93	0.93	0.92	**0.95**	**0.95**	**0.95**	0.70	0.77	0.71
	02/05/2014	person2	**0.85**	**0.92**	**0.86**	0.87	0.87	0.84	0.71	0.76	0.71
	02/05/2014	person1	**0.84**	**0.90**	**0.84**	0.79	0.90	0.80	0.70	0.77	0.71
	02/04/2014	**wholehouse**	**0.918**	**0.924**	**0.913**	0.916	0.921	0.911	0.77	0.69	0.71
	02/05/2014	**wholehouse**	**0.90**	**0.84**	**0.84**	0.88	0.81	0.81	0.74	0.70	0.70

5.3 Limitations of Mixture EGH Model

Although the temporal mixture EGH model performs well on the datasets Study 10 and Study 11, the same cannot be said for the dataset Study 14. Table 3 shows that in Study 14, the mixture EGH model performs better for the individual and whole-house occupancy predictions on 12/18/2013 but not on either 12/19/2013 or 12/20/2013. Checking the activities of the study participants on these two days, we find that both went out again after coming back and staying home for a while. Since the episodes of going out again after returning home from work do not occur frequently, the mixture EGH is unable to detect this pattern. Thus, the occupancy prediction probability for these events is completely absent. However, kNN performs well for such events because it leverages all the historical data and so if an abnormal event occurs just once, this prediction approach incorporates it when calculating the average value. To relax the limitations for abnormal events, we therefore propose a hybrid model for prediction: when deploying this occupancy prediction, for example for a prediction that is 30 minutes ahead, just 15 minutes before the prediction, if a person goes out again after coming back, the deployed system switches to the kNN approach rather than the mixture EGH model for the prediction. In such cases, this hybrid model should always yield the best prediction results.

5.4 House Occupancy Prediction 30 Minutes Ahead with Hybrid Approach

To adequately preheat the house prior to the occupants' return, we need to evaluate how much time to allow in advance to automatically turn on/off the HVAC. Here, the advance notice time estimate given in [13] is utilized and a prediction time of 30 minutes ahead of house occupancy applied. We compare the receiver operating characteristic (ROC curve) of three approaches: the mixture EGH model, kNN, and a hybrid approach consisting of the mixture EGH model plus kNN. In this hybrid approach, we set the mixture EGH results as the baseline, but replace the values of the mixture model with the values obtained from the kNN model in the following two situations: 1) after a person returns home; and 2) when the prediction probability of kNN is greater than 0.8. Figures 5, 6, and 7 illustrate the ROC curve for the whole house occupancy prediction on 02/20/2014 for the dataset Study 10, 02/04/2014 for the dataset Study 11 and 12/20/2013 for the dataset Study 14, respectively. The red and green lines represent the kNN and mixture EGH models; the blue line denotes the hybrid approach. The ROC curves show that the hybrid approach consistently produces the largest areas for the three cases, namely 0.96, 0.92 and 0.92, respectively, which indicate that the hybrid approach always performs best.



(a)

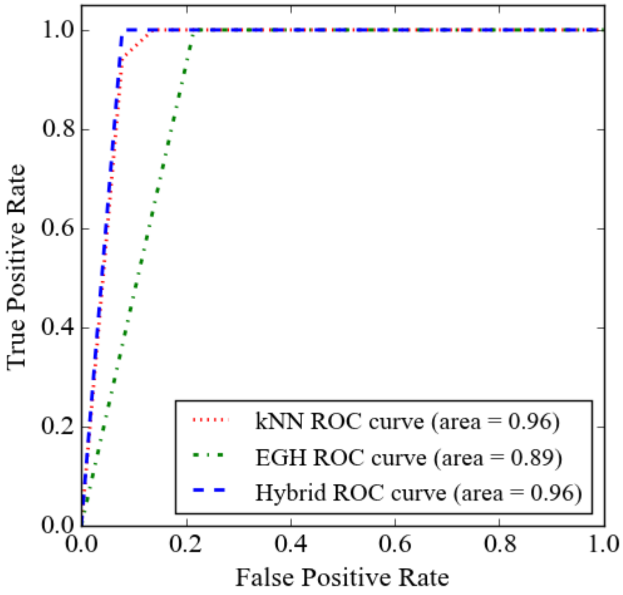
Figure 5: ROC curve of house occupancy prediction in Study 10 (02/20/2014).

6 CONCLUSION

Residential occupancy prediction is a hot research topic supporting efforts to control HVAC systems more efficiently, with a consequent reduction in energy consumption. Increasing the accuracy of occupancy prediction allows these saving to be made without sacrificing the comfort of those living inside the home. In order to achieve the

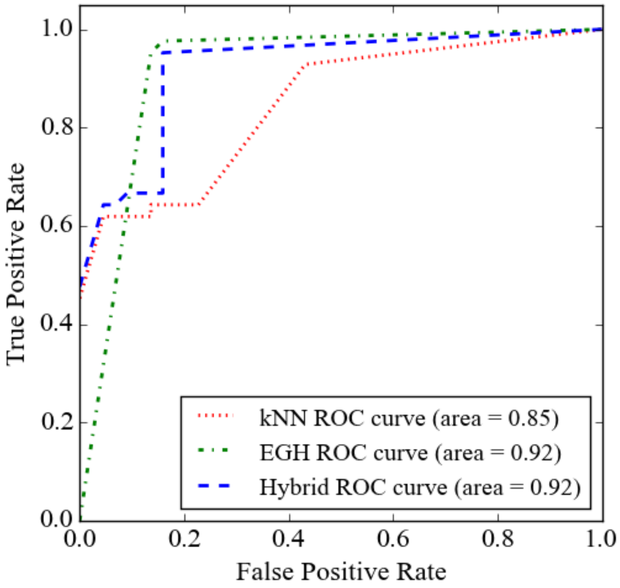
Table 3: Precision Recall F-measure of Individual and Whole House Occupancy Prediction in Study 14.

Dataset	Date	Person	EGH			kNN			SVM		
			precision	recall	fmeasure	precision	recall	fmeasure	precision	recall	fmeasure
study14	12/18/2013	person2	**0.91**	**0.91**	**0.89**	0.87	0.87	0.84	0.73	0.77	0.71
	12/18/2013	person1	**0.92**	**0.92**	**0.91**	0.90	0.90	0.89	0.73	0.76	0.71
	12/19/2014	person2	0.86	0.86	0.85	**0.90**	**0.90**	**0.88**	0.73	0.76	0.71
	12/19/2014	person1	0.85	0.84	0.84	**0.86**	**0.86**	**0.85**	0.73	0.76	0.71
	12/20/2014	person2	0.92	0.94	0.92	**0.98**	**0.97**	**0.97**	0.75	0.79	0.75
	12/20/2014	person1	0.90	0.91	0.90	**0.95**	**0.95**	**0.95**	0.75	0.79	0.75
	12/18/2013	**wholehouse**	**0.91**	**0.91**	**0.90**	0.88	0.88	0.86	0.75	0.72	0.70
	12/19/2013	**wholehouse**	0.841	0.845	0.838	**0.848**	**0.853**	**0.842**	0.79	0.74	0.74
	12/20/2013	**wholehouse**	0.92	0.90	0.90	**0.94**	**0.93**	**0.93**	0.74	0.72	0.70



((b)

Figure 6: ROC curve of house occupancy prediction in Study 11 (02/04/2014).



(c)

Figure 7: ROC curve of house occupancy prediction in (c) Study 14 (12/20/2013).

best prediction results, we propose integrating the mixture EGH model and kNN to create a new hybrid approach.

Our work differs from previous research based on the main contributions listed below:

(1) We formulate the problem as one of temporal mining, where the activities inside the building are abstracted as episodes, and each episode is connected via an episode generative HMM model.
(2) We mine the activity patterns according to the times and the gaps between them: both the duration of each type of activity, and the gap between two consecutive events are limited within an appropriate range. This range is extracted from historical data according to the day of the week and holidays.
(3) Our hydrid prediction solution performs best for workday occupancy prediction: in the case of normal activities, a mixture EGH model is applied and in the case of abnormal events, kNN is utilized, as this is generally considered a benchmark in occupancy prediction problems.

In future work, we plan to continue working on improving holiday occupancy prediction. This is a more intractable problem because the occupancy patterns for these days are completely different. For example, on certain days, a person may never leave the house. Therefore the occupancy prediction for holidays probably depends more on the date than on the indoor activities performed. We also plan to apply this new temporal mining approach to GPS datasets [6] to test the effectiveness of occupancy prediction with different kinds of data.

7 ACKNOWLEDGMENTS

This research has been supported in part by US National Science Foundation grant DGE-1545362 and the Institute for Critical Technology and Applied Science, Virginia Tech. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

[1] Abdullah Alrazgan, Ajay Nagarajan, Alexander Brodsky, and Nathan E Egge. 2011. Learning occupancy prediction models with decision-guidance query language. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. IEEE, 1–10.

[2] Alex Beltran and Alberto E Cerpa. 2014. Optimal HVAC building control with occupancy prediction. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM, 168–171.

[3] Varick L Erickson and Alberto E Cerpa. 2010. Occupancy based demand response HVAC control strategy. In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*. ACM, 7–12.

[4] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin* (2010).

[5] Wilhelm Kleiminger, Friedemann Mattern, and Silvia Santini. 2014. Predicting household occupancy for smart heating control: A comparative performance analysis of state-of-the-art approaches. *Energy and Buildings* 85 (2014), 493–505.

[6] Christian Koehler, Brian D Ziebart, Jennifer Mankoff, and Anind K Dey. 2013. TherML: occupancy prediction for thermostat control. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 103–112.

[7] John Krumm and A Brush. 2011. Learning time-based presence probabilities. *Pervasive Computing* (2011), 79–96.

[8] Srivatsan Laxman, Vikram Tankasali, and Ryen W White. 2008. Stream prediction using a generative model based on frequent episodes in event sequences. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 453–461.

[9] Sawsan Mahmoud, Ahmad Lotfi, and Caroline Langensiepen. 2013. Behavioural pattern identification and prediction in intelligent environments. *Applied Soft Computing* 13, 4 (2013), 1813–1822.

[10] Carlo Manna, Damien Fay, Kenneth N Brown, and Nic Wilson. 2013. Learning occupancy in single person offices with mixtures of multi-lag Markov chains. In *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*. IEEE, 151–158.

[11] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. 1997. Discovery of frequent episodes in event sequences. *Data mining and knowledge discovery* 1, 3 (1997), 259–289.

[12] Debprakash Patnaik, PS Sastry, and KP Unnikrishnan. 2008. Inferring neuronal network connectivity from spike data: A temporal data mining approach. *Scientific Programming* 16, 1 (2008), 49–77.

[13] James Scott, AJ Bernheim Brush, John Krumm, Brian Meyers, Michael Hazas, Stephen Hodges, and Nicolas Villar. 2011. PreHeat: controlling home heating using occupancy prediction. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 281–290.

[14] Huijuan Shao, Manish Marwah, and Naren Ramakrishnan. 2013. A Temporal Motif Mining Approach to Unsupervised Energy Disaggregation: Applications to Residential and Commercial Buildings. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*. AAAI, Bellevue, U.S.A., 1327–1333.