Probabilistic Learning From Incomplete Data for Recognition of Activities of Daily Living in Smart Homes

Shuai Zhang, Sally I. McClean, *Member, IEEE*, and Bryan W. Scotney

***Abstract*—Learning behavioral patterns for activities of daily living in a smart home environment can be challenged by the limited number of training data that may be available. This may be due to the infrequent repetition of routine activities (e.g., once daily), the expense of using observers to label activities, and the intrusion that would be caused by the presence of observers over long time periods. It is important, therefore, to make as much use of any labeled data that are collected, however, incomplete these data may be. In this paper, we propose an algorithm for learning behavioral patterns for multi-inhabitants living in a single smart home environment, by making full use of all limited labeled activities, including incomplete data resulting from unreliable low-level sensors in this environment. Through maximum-likelihood estimation, using Expectation–Maximization, we build a model that captures both environmental uncertainties from sensor readings and user uncertainties, including variations in how individuals carry out activities. Our algorithm outperforms models that cannot handle data incompleteness, with increasing performance gains as incompleteness increases. The approach also enables the impact of particular sensors to be assessed and can thus inform sensor maintenance and deployment.**

***Index Terms*—Activity recognition, activities of daily living (*ADLs*), Expectation–Maximization (EM) algorithm, incomplete data, probabilistic learning.**

I. Introduction

ELDERLY people often experience difficulties in performing activities of daily living (*ADL*s). Along with their poor health conditions, it is difficult for them to remain in their own home and live independently. With the increasing number of elderly people in the population, effort has been made in smart home research to provide a living home environment with support for assisted living and health monitoring through intelligent analysis of interactions between the home and its inhabitants [1]. Monitoring inhabitants' performance on key *ADL*s, such as eating and medication consumption, provides valuable information about their well-being.

Recognition of *ADL*s from sensor data in the smart home is a key element of a support system. At the core of the smart home concept is sensor technology. Sensors capture the status of the home, the movement of inhabitants, and their interactions with the home. Inhabitants' activities are thus reflected by low-level sensor activation data that provide fundamental information for learning and activity recognition. Models of inhabitants' activity patterns can then be used to recognize and interpret future performed activities, and thus to monitor and assist the completion of an activity if necessary. However, deriving models in a smart home environment is a complex problem, as the activities being studied are not necessarily repeated regularly every day. Therefore, the volume of training data collected during a reasonable period of time, e.g., three months, may be limited, and in many cases, we do not have complete control over the training data available for learning. At the same time, the collection of training data may require labeling by caregivers to follow the inhabitants and record their activities. Such approaches are very expensive or may be intimidating for the inhabitants. Using a larger volume of training data to improve the learning performance is, therefore, not necessarily an available option.

Another challenge is that information used for training recorded from low-level sensors is not always reliable [2]. Common situations that may occur include instances when sensors are in low-battery status or alternatively a sensor may malfunction and not produce its triggered event even if it has been activated correctly. Under these circumstances, the observed sensor data do not reveal complete information describing the events that may have happened while an inhabitant carries out an activity. Unreliable sensors cause the observed information to be incomplete.

The issue of learning on limited amounts of unreliable training data caused by environmental uncertainties from the low-level sensor information is considered to be both a research and a practical challenge. Therefore, making full use of the information from collected data is crucial in deciding the performance of the activity profile model. User uncertainty is another important issue for long-term deployment of activity recognition systems in smart home environments. User uncertainties include various decisions made by a user in different contexts, or by different users within the same environment [3]. There may be multiple colocated persons: in some smart environments, for example, one patient (or an elderly couple) with the caregiver(s). The ability to distinguish an individual from another promotes resource saving, service improvement and personalized intervention delivery.

The purpose of this research is to investigate whether we can achieve good activity recognition performance for multiple

Manuscript received February 11, 2011; revised December 31, 2011; accepted February 9, 2012. Date of publication March 9, 2012; date of current version May 4, 2012. This work was supported by the Engineering and Physical Sciences Research Council for the MATCH project under Grant GR/S29874/01.

The authors are with the School of Computing and Information Engineering, University of Ulster, Coleraine, BT52 1SA, U.K. (e-mail: s.zhang@ulster.ac.uk; si.mcclean@ulster.ac.uk; bw.scotney@ulster.ac.uk).

Digital Object Identifier 10.1109/TITB.2012.2188534

inhabitants, in situations where the number of training data is limited, and also data may be incomplete due to sensor unreliability in the environment. An algorithm that is able to learn from incomplete data as well as complete data is thus required.

In the literature, work has been carried out on sensor analysis for different application purposes. The collected low-level sensor input data can be seen as a sequence of events of possibly varying data types. Frequent episodes can be discovered from sequences [4] to identify regularly occurring device interactions sequence based on techniques from text compression to minimize entropy of inhabitant mobility and daily activity [5]. Uncertainty in low-level information is often addressed using Bayesian networks, as for example, in the PROACT project [6], or by Hierarchical Markov Model in [7] and [8]. Bayesian networks must be trained using labeled data that can be expensive to obtain. In [9], activities were inferred by a model learned in a smart home via Dempster–Shafer theory using the fusion of contextual information from uncertain sensor data with default manufacturer statistics for sensor unreliability. Fuzzy logic was used to preprocess the sensor activation sequences to fill in logical gaps in [10] by removing inconsistencies and converting the data to a more suitable format for use by the analysis algorithms. This approach tackles the data incompleteness problem in the data proprocessing stage; a data model is then based on complete information rather than addressing the issue of incomplete sensor readings. In contrast, our algorithm learns directly from incomplete data, and inhabitants' behavioral patterns are characterized using the learned probability distribution over various activities. The model is used to infer the activities and the inhabitants who have carried them out. The comparison of inhabitants' profiles learned in different time periods can assist long-term health monitoring and reveal indications of the development of functional deterioration.

II. Probabilistic Learning Framework

A. Data Model

Activities in smart homes are represented by four attributes, namely, *Person*, *ADL*, *Episode,* and *Time*. An *ADL* may be carried out in a way described by an *Episode* in the context represented by *Person* and *Time*. Each attribute has a list of categorical domain values.

Person: For the situation of multi-inhabitants at home, individuals with different activity profiles need to be represented differently.

ADL: Self-care activities include basic but fundamental tasks of preparing a meal, eating, getting in or out of bed or a chair, using the toilet, etc. [11]. The *ADL*s to be learned in a smart home vary and depend on applications.

Time: A context variable indicating the time of day at which an activity is carried out, which takes discretized categorical values at appropriate levels according to the requirement of applications.

Episode: An episode is a sequence of low-level sensor events activated during the process of performing an activity. Episode values reflect the typical routine actions for completing an activity. Different episodes for the same activity represent different ways of carrying out that activity.

For each of four attributes, we have a set of domain values $\{c_1^{(j)}, \ldots, c_{k_j}^{(j)}\}$, where k_j is the number of domain values and j denotes attribute ($j = Person, ADL, Episode, Time$). The domain values at the finest level of detail for each attribute form the base schema S. In schema structures for incomplete data, attribute values of coarser levels of granularity arise according to the type of incompleteness.

A logical datacube of cells v_{paet} is formed by the schema together with corresponding cardinalities n_{paet}, representing the number of occurrences of this activity represented by the cell v_{paet}. Cell i of datacube D represented as v_{paet} is the Cartesian product of the four attributes in the form $\{c_{i_p}^{(Person)} \times c_{i_a}^{(ADL)} \times c_{i_e}^{(Episode)} \times c_{i_t}^{(Time)}\}$. For example, $c_{i_p}^{(Person)}$ is the value of attribute *Person* occurring in cell i of the datacube.

In the case of incomplete data, we have datacubes D_w of schema S_w ($w = 1,\ldots,W$) where each schema represents different types of incompleteness. Correspondingly, we label n_{paet}^w as the cardinality for cell v_{paet}^w of schema S_w.

Extensible notations have been defined previously. For simplicity of presentation in the rest of the paper, these general notations are modified. When writing sum or product over a range of attribute values, we substitute $k_p = P$, $k_a = A$, $k_e = E$, $k_t = T$, and for example, a sum $\sum_{\{c_s^{(p)}\}_{s=1}^{k_p}}$ over attribute values for *Person* as $\sum_{p=1}^{P}$ and a product $\prod_{\{c_s^{(p)}\}_{s=1}^{k_p}}$ as $\prod_{p=1}^{P}$.

Mapping between episodes indexed by e^w and e in schemas S_w and S (base schema), respectively, is represented as $q_{e^w e}^w$ ($e^w = 1,\ldots,E^w$; $e = 1,\ldots,E$), taking value 1 if e^w and e relate to the same underlying base episode, and 0 otherwise.

We consider an example of the activity of "making a drink" in the kitchen, more specially "making a black coffee," "making coffee with milk," "making a tea with milk and sugar," "getting a juice," etc. When the *Sugar* sensor is not working, if the observed episode has sensor activation sequence *Coffee*, *Kettle*, represented as "CK," we are not able to know from this observation if the inhabitant put sugar in the coffee or not, nor when he or she put it in. Thus, the observed episode "CK" can correspond to a number of possible base episodes, namely "SCK", "CSK", "CKS", or "CK".

B. Learning

A straightforward method to deal with incompleteness is to take only the complete data into the learning process [12]. However, as discussed previously, it is important to learn from incomplete data given the limited volume of collected training data. Given a mixture of data $D_1,\ldots, D_W$ of schemas $S_1,\ldots, S_W$, ($w = 1,\ldots,W$), our target is to find the most likely model that can fit the data. We propose a probabilistic activity learning algorithm using maximum-likelihood estimation (MLE) [13], a mixture of complete and various incomplete smart home data.

Although the computational cost of deriving a model using the full joint distribution may be potentially high due to the large number of domain values for each of the attributes, the learning in our application is performed offline, and therefore,

the computation time is not a crucial issue. In practice, as each inhabitant has habits of performing activities with relatively low variation, the derived probability distribution is sparse, thus reducing the volume of calculations required.

An activity profile model of probability distributions [14] $\{\pi_{paet}\}$ is built in the form of probability distributions over the Cartesian product of the four attributes *Person*, *ADL*, *Episode,* and *Time* described at the base schema of the finest level of information granularity, where π_{paet} is the probability value associated with the cell v_{paet} over the base schema *S*.

The likelihood function given the mixture data $D_1,\ldots,D_W$ of *W* different data schemas $S_1,\ldots,S_W$, following a multinomial distribution is thus given as

$$L \propto \prod_{w=1}^{W}\left(\prod_{p=1}^{P}\prod_{a=1}^{A}\prod_{t=1}^{T}\prod_{e^w=1}^{E^w}\left(\sum_{e=1}^{E} q^w_{e^w e}\pi_{paet}\right)^{n^w_{pae^w t}}\right)$$

for $p=1,\ldots,P; a=1,\ldots,A; e=1,\ldots,E; t=1,\ldots,T$.

For any schema S_W, the probability distribution can be represented using the distribution π_{paet} of the base schema and the relationship between the two schemas. The probability value $\sum_{e=1}^{E} q^w_{e^w e}\pi_{paet}$ of cell $v^w_{pae^w t}$ in schema S_w is derived by the sum of probabilities of all possible base cells to which $v^w_{pae^w t}$ maps, achieved by the mappings $q^w_{e^w e}$ between episodes indexed by e^w and *e* in schemas S_w and *S*, respectively.

Our solution is formulated by maximizing the probability model likelihood using the Expectation–Maximization (EM) algorithm [15]. The computation of joint probability π_{paet} for cell v_{paet} in base schema *S* is carried out iteratively between the expectation (E) step and the maximization (M) step. We update the partition of incomplete data in the E-step, followed by calculating the data descriptors of probability distribution in the M-step. The stopping criterion for the iteration is that the relative improvement of the log-likelihood is below a specified threshold.

In the E-step, given the current model description, incomplete data of schema S_w are partitioned according to the base schema *S*, given the mappings between schemas:

$$n^w_{paet} = \pi^{(n-1)}_{paet} \times \sum_{e^w=1}^{E^w} \frac{n^w_{pae^w t} q^w_{e^w e}}{\sum_{e'=1}^{E} q^w_{e^w e'}\pi^{(n-1)}_{pae't}}$$

where n^w_{paet} is the contribution of datacube D_w of schema S_w to the cell v_{paet} of the base schema *S*.

In the M-step, the algorithm re-estimates the probability distribution to describe the mixture data given the updated cardinality values in the previous E-step:

$$\pi^{(n)}_{paet} = \frac{1}{N}\sum_{w=1}^{W} n^w_{paet}$$

where *N* is the total number of observations, $N = \sum_{w=1}^{W}\left(\sum_{p=1}^{P}\sum_{a=1}^{A}\sum_{e^w=1}^{E^w}\sum_{t=1}^{T} n^w_{pae^w t}\right)$.

The EM algorithm is a widely used general class of iterative procedures for learning in the presence of missing information, which provides an intuitive approach to learning directly from complete and incomplete data [16]. Its use here for a practical application of learning in smart home environment is novel. If complete knowledge at the base schema were available, the probabilities would be estimated by simply dividing the number of cases that take each respective value by the total cardinality of observations. In the algorithm, where the number of cases that take a particular value is unknown, this number is replaced by apportioning the corresponding data from the local aggregates that might take that value. We have undertaken the aggregation of uncertain and incomplete information in databases in previous work [17], and this paper extends this concept to new application areas.

C. Evaluation Criteria

The performance criteria are the model classification accuracies of activities of different levels of detail and the distance measure of the model as a whole to the true pattern. In an uncertain environment, some activity recognition may not be sufficiently confident in difficult scenarios. In such cases, *ADL*s can be defined at different levels of detail, and coarser level activity prediction, denoted by *HA*, can still be provided. Given the observed episode e^o, time t^o, predictions are made for *ADL* class along with the *Person* who has carried out the *ADL*, represented by the joint class of (*Person*, *ADL*) (*PA* for abbreviation), and on the higher level task (*Person*, *HA*) (*PHA*). Prediction on the class (*Person*, *Episode*) (*PE*) is also made to indicate the specific way in which the activity has been carried out, which can be used potentially for activity intervention. In the following prediction formulas, in order to show explicitly how the computation is performed, we use Pr as a generic logical representation of a probability and π_{paet} to represent values used in the operational execution of the calculation.

Starting from the prediction task of *PE* at the lowest level of detail, the classification calculation is formulated as follows:

$$\begin{aligned}\Pr(p=i,e=j|e^w=e^o,t=t^o) &= \frac{\sum_{u=1}^{A}\Pr(p=i,a=u,e=j,t=t^o)}{\Pr(e^w=e^o,t=t^o)}\\ &= \frac{\sum_{a=1}^{A}\pi_{iajt^o}}{\sum_{p=1}^{P}\sum_{a=1}^{A}\sum_{e=1}^{E}(\pi_{paet^o}\times q^w_{e^o e})}\end{aligned}$$

where $q^w_{e^o e}$ is the mapping from the observed episode e^o of schema S_w to the episode e in base schema *S*.

The prediction can then be assigned to the joint (*Person*, *Episode*) (*PE*) class c^{PE} with the highest probability:

$$c^{PE} = \arg\max_{i,j}\Pr(p=i,e=j\,|e^w=e^o,t=t^o)\ .$$

The (*Person*, *Activity*) *PA* prediction task is formulated as:

$$\begin{aligned}&\Pr(p=i,a=j|e^w=e^o,t=t^o)\\ &=\sum_{k=1}^{E}\Pr(p=i,a=j,e=k|e^w=e^o,t=t^o)\\ &=\sum_{k=1}^{E}\left(\frac{\pi_{ijkt^o}}{\sum_{p=1}^{P}\sum_{a=1}^{A}\sum_{e=1}^{E}(\pi_{paet^o}\times q^w_{e^o e})}\right).\end{aligned}$$

Crisp prediction can then be assigned to the *PA* class where

$$c_k^{PA} = \arg\max_{(i,j)} \Pr(p = i, a = j | e^w = e^o, t = t^o).$$

Prediction on the highest level task of *PHA* is achieved through *HA*'s relationship with the corresponding *ADL*s. If $\Omega(HA)$ denotes the set of *ADL*s to which *HA* corresponds, the probability distribution over *PHA* classes is then given by

$$\Pr(p = i, ha = k \,| e^w = e^o, t = t^o) = \sum_{j \in \Omega(ha=k)} \Pr(p = i, a = j | e^w = e^o, t = t^o)$$

where $j \in \Omega(ha = k)$ represents that the activity with index j is a subclass of *HA* value $ha = k$.

The learned model takes the general form of a probability distribution on the base schema of the four attributes *Person*, *ADL*, *Episode,* and *Time*. As another measure of the model performance, the Euclidean distance is calculated between the learned model as a whole and the true underlying distribution.

III. Data Collection

A. Smart Kitchen Environment

For data collection in our experiment, a 17-m^2 smart kitchen laboratory was used, located at the University of Ulster. ADL such as "making a drink" were monitored within this environment, for multiple users, through a suite of contact switch sensors embedded in specific objects relevant to our target activities. These include the kettle, sugar jar, coffee jar, tea pot, fridge, cups, cupboards, and door sensors. Actions carried out in the kitchen environment are detected from the sensor activations. Each contact switch sensor contains two magnetic parts, where one is attached to a base and the other on a target object. When the object is being used, the two magnetic parts are separated, triggering sensor activation that sends a signal to a base receiver.

In our experiments, the task of "making a drink" refers to nine possible activities: $\{ADL_j\} = \{$"making black tea," "making tea with milk," "making tea with sugar," "making tea with both milk and sugar," "making black coffee," "making coffee with milk," "making coffee with sugar," "making coffee with both milk and sugar," and "making a cold drink"$\}$ ($j = 1,\ldots,9$). The higher level description of this task includes HA_1 = "making a cup of tea," HA_2 = "making a cup of coffee," and HA_3 = "making a cold drink." Two participants were involved in the experiments, and a model of their activity patterns was learned for the task of "making a drink." To obtain labeled data, a computer interface was designed to collect values of labels of *Person* and *ADL* for each individual activity. Before each activity, participants were asked to select who they were and what activity they were about to perform from drop-down lists on the interface. The participants click the "start" and "end" buttons on the interface before and after they have performed the activity. Using the start and end times for each activity, the corresponding episode value is extracted from the sensor events. The class labels are thus collected that correspond to the sequence of sensor activations. The value for attribute *Time* has categorical values of "Morning," "Afternoon," or "Evening." The discretization can be application oriented. As for our activity of making a drink, three categories of a day are sufficient to characterize various habits of making a drink in different sections of a day. Finally, by matching "Episode" and "Time" with their activity class labels, we derive data tuples in the corresponding schema, as illustrated in Table I. "Status" is used to indicate the information reliability, as each sensor can malfunction or be in low-battery status as reported by routine sensor broadcast.

TABLE I
SAMPLE OF DATA FROM THE ACTIVITY TABLE

ACTIVITY ID	*PERSON*	*ADL*	*EPISODE*	*TIME*	STATUS
…	…	…	…	…	…
15	David	Tea	TSFK	Afternoon	-
35	David	Coffee	CK	Afternoon	*S* Sensor unreliable
44	Emma	Coffee	CSKF	Morning	-
…	…	…	…	…	…

In total, 47 labeled records of activities have been collected in this environment over a period of four weeks; within this sample, seven of the nine *ADL* types referred to above are present.

B. Data Simulation

Given that we have collected a limited number of real data in the smart kitchen environment, parameter estimates from the model built using these real data were then used to create synthetic data in order to augment the volume of data used for evaluation and to provide the opportunity for systematic and controlled performance evaluation of various aspects. Additionally, we assume that it is equal likely that any sensor(s) are malfunction or in low-battery status. The algorithm applies in the same manner to scenarios in which sensor unreliability may be different for different sensors and such scenarios could be readily simulated by adjusting the generation of the synthetic data. Using synthetic data, we can evaluate the algorithm in dealing with incompleteness caused by a wide range of realistic situations of sensor(s) unreliability, which is not possible by using only real data collected in a short period.

Synthetic data are generated using MATLAB [18]. Random numbers between 0 and 1 fall into particular intervals, according to cumulative probability distribution, which correspond to the activity indicated by a cell v_{paet}. The categorized random data were then aggregated into datacube *D* of schema *S*, where cardinality n_{paet} is the number of random values that fall into the corresponding probability interval corresponding to v_{paet}. Incomplete data corresponding to other schema structures were generated from the simulated complete data of the base schema by combining the appropriate intervals in the base schema using the mappings between the two schemas for complete and incomplete data.

The synthetic datasets for training and evaluation purposes were generated separately, based on the same distribution representing the behavioral model. Models are learned from the simulated data where labels of activities are known. They are

evaluated by classification on the generated test data, where the class labels are used later to assess the performance.

IV. Evaluation Results and Performance Analysis

Different degrees and types of incomplete data caused by missing sensor reading(s) have diverse effects on the performance in terms of activity recognition. The quantity of incomplete data depends on the reliability of the sensors and also on the replacement strategy once they malfunction, or require a change of battery. A systematic evaluation framework has been established to investigate, in these different scenarios, how our algorithm can be used to make full use of all complete and incomplete information to retain good performance, in situations where additional training data to improve learning performance are not available. Given a certain amount of training data, learning is carried out in situations where there are different percentages of incomplete data, ranging from 10% to 70%, indicating the proportion of the data points that have missing sensor readings. We have chosen to develop a general and principled method of handling incomplete data that minimizes imputation assumptions. Our results are compared with "Bench" models in order to show the contribution made by incomplete data through appropriate learning. By "Bench" models, we mean models that are learned from only the complete data and make no use of incomplete data. In the literature, other methods have been developed for the imputation of values where data are missing. However, in most cases, these have been designed for numerical data and thus they are not suitable for use here. Also, most methods in the literature perform the imputation by using nonmissing data to predict the values of missing data, e.g., k-means. However, in our case, the data are partially missing, but with information still available at a higher level of granularity. Data imputation that replaces the possible value before the analysis may not make use of the existing knowledge available through the partial value mappings to finer levels of granularity or may lead to overcorrection and generate bias in the data. MLE implemented by EM makes fewer assumptions about the data and provides an intuitive and principled approach to be used in our problem.

Model performance is evaluated on the criterion of classification accuracy on tasks of *PHA* (*Person*, *Higher-level ADL*), *PA* (*Person*, *ADL*), and *PE* (*Person*, *Episode*) and the distance measure of the learned probability distribution to the true distribution. All of the evaluations are carried out using ten repetitions to obtain the average model performance and their standard mean error.

Fig. 1 shows the performance of models learned from different volumes of complete data. The size of the training data for the following experiments is then chosen to be $N = 100$ because it achieves good performance without using large volumes of expensive labeled data.

Assuming that it is equally likely for any sensor(s) to be unreliable during the period of data collection, in the first set of experiments, we investigate the case of one unreliable sensor during the period of data collection. Experiments are then generalized to settings with more than one unreliable sensor and again with varying degrees of unreliability or sensor malfunction.

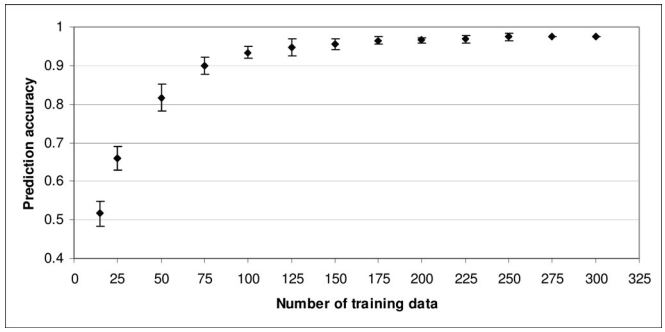


Fig. 1. *PA* prediction accuracies by models learned from different volumes of training data.

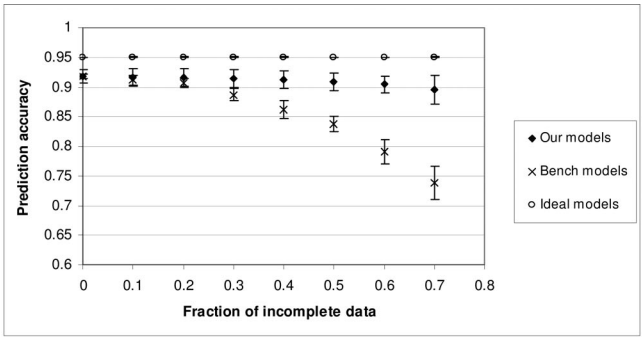
A. Experiments With a Single Unreliable Sensor

Experiments are first carried out with intermittent failure of the "Kettle" sensor during the training data generation. All evaluations are carried out on the same set of test data comprised of base and other schema structures. Prediction performance by models trained on different proportions of incomplete data is shown in Fig. 2(a)–(c) for tasks of *PHA*, *PA,* and *PE* respectively. The Euclidean distance measure for our learned models is shown in Fig. 3.

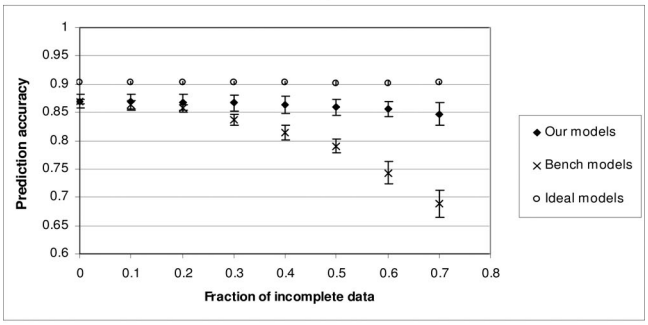
Results in Fig. 2 show that with increasing proportion of incomplete data, the prediction performance decreases for all three prediction tasks, as we would expect. Compared with the performance of the "Bench" models, our algorithm has substantially increased prediction accuracies, and the advantage of our models is greatest when the proportions of incomplete data are largest. Our model can tolerate a fairly large percentage of incomplete data in the training dataset without affecting the prediction performance greatly. Prediction performance is better on tasks described at coarser levels of detail. The tradeoff between the level of prediction class granularity and the prediction performance may be dependent on the requirements of real applications.

As a comparison, we investigate the performance with the scenario of learning when the training data are sufficient to build a model ($N = 1000$) according to the results in Fig. 1. Given a certain amount of training data $N = 1000$, in similar settings, learning using our algorithm is carried out in situations where there are different proportions of incomplete data, ranging from 10% to 70%. We call these derived models "Ideal" models, as shown in Fig. 2. As expected, we see that in the "Ideal" model, there are sufficient complete training data to achieve consistently high prediction accuracy, even in the situation where the percentage of incomplete data is high.

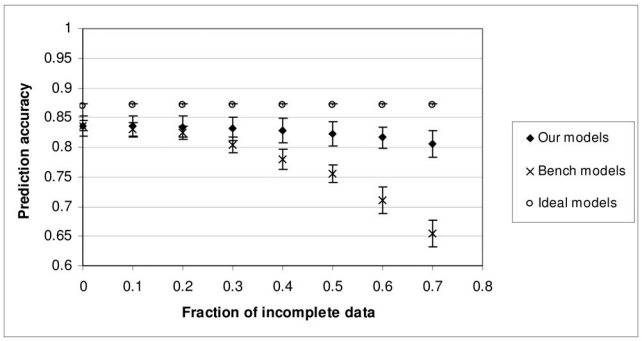
Of course, in this paper, our focus is on situations in which only limited amounts of complete data are available, but we may consider the results achieved by the "Ideal" models as upper limits of achievable prediction accuracy against which we can assess our approach. Using the "Ideal" model results for comparison with our current approach, we see that the improvement



(a)



(b)



(c)

Fig. 2. Prediction accuracies by models learned from incomplete data due to an unreliable "Kettle" sensor in comparison with performance of "Bench" models and "Ideal" models on hierarchical tasks *PHA*, *PA,* and *PE*. (a) *PHA* prediction task. (b) *PA* prediction task. (c) *PE* prediction.

achieved by making use of the incomplete data compared with the complete data only is a significant proportion of the maximum improvement achievable, particularly in the case when the proportion of incomplete data is high. Hence, we show that our approach is a promising way to tackle the tradeoff of performance and data collection cost.

For the distance measures shown in Fig. 3, similarly, distance measures for "Bench" models deteriorate much more rapidly with increased incompleteness in the training data, compared to our learned models.

In a similar way to the experiment with the intermittent "Kettle" sensor, experiments have been carried out simulating various scenarios with a single unreliable sensor during training data collection. In each scenario, the single unreliable sensor is "Tea," "Coffee," "Fridge," or "Sugar," respectively. Performance is again evaluated on prediction accuracy for *PHA*, *PA*, and *PE* and also the model's Euclidean distance measure from the true distribution.

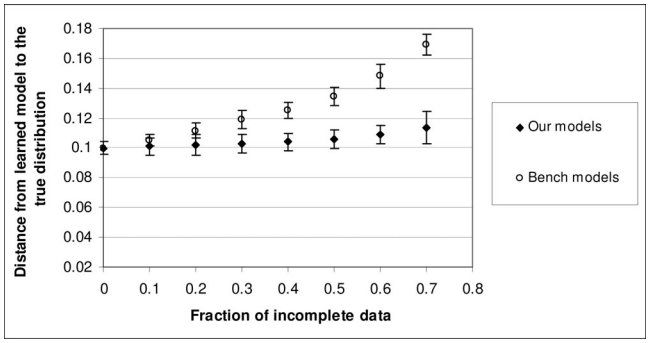


Fig. 3. Comparison of distances to the true model from our learned models and "Bench" models.

In each set of experiments, compared with the performance of the "Bench" model, the models derived using our algorithm have improved prediction accuracies, and the performance improvement is the highest when the proportions of incomplete data are largest. All the results follow a pattern similar to the experiment with the unreliable "Kettle" sensor. With an increase in incomplete training data, the learned probability distributions become less accurate, but are still much better than the corresponding "Bench" models.

To summarize, our learning algorithm can effectively handle incompleteness in the training data due to a single unreliable sensor and make full use of information to achieve good performance. The algorithm outperforms the "Bench" models, with performance gains more obvious as the extent of the incomplete data increases. Given the upper limit performance of "Ideal" models, prediction improvement made using our algorithm, compared with performance on the complete data only, is a significant proportion of the maximum improvement achievable, particularly in the case when the proportion of incomplete data is high. Detailed results and evaluation analysis for other sensor are shown in [19].

B. Experiments With Multiple Unreliable Sensors

A further sensor may also malfunction or be in a low-battery condition at various times during the period of data collection. As demonstrated previously, in each of the periods, data are generated with their own schema structure depending on the corresponding configurations of sensors. We now investigate the scenario of two unreliable sensors in the environment. Learning and evaluation are carried out for different combinations of two unreliable sensors, and similar trends have been consistently observed in each case. Therefore, as an illustrative example, we show the algorithm performance in the scenario of "Sugar" and "Fridge" as the two unreliable sensors, evaluated with various degrees of incomplete data. We calculate the proportion of the incomplete data based on the total number of cases for which at least one sensor is unreliable. The prediction performance is presented in Fig. 4 for individual classification tasks. Also, the distance of the learned models from the true model increases progressively as the fraction of incomplete data increases, indicating the deteriorating overall model performance. Performance can be compared with "Bench" models plotted in

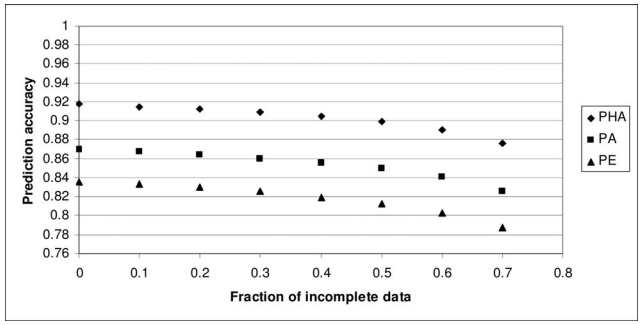


Fig. 4. Prediction comparisons by models learned from incomplete data due to unreliable "Sugar" and "Fridge" sensors.

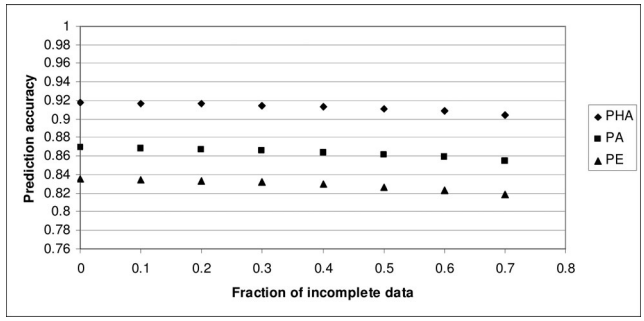


Fig. 5. Prediction accuracies by models learned from incomplete data due to unreliable "Tea," "Coffee," and "Fridge" sensors.

TABLE II
NUMBER OF ITERATIONS REQUIRED FOR CONVERGENCE BY MODELS LEARNED FROM INCOMPLETE DATA DUE TO UNRELIABLE SENSORS

Scenario 1: unreliable 'sugar' and 'fridge' sensors		**Scenario 2**: unreliable 'tea', 'coffee' and 'Fridge' sensors	
Fraction of incomplete data	Average Iterations	Fraction of incomplete data	Average Iterations
0.1	2.3	0.1	2.3
0.2	3.9	0.2	3.9
0.3	18	0.3	6.2
0.4	24.2	0.4	9.6
0.5	44.4	0.5	13.5
0.6	45.4	0.6	21.2
0.7	53	0.7	30.2

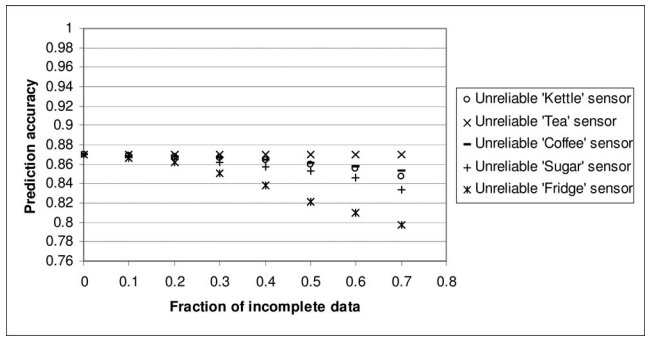


Fig. 6. *PA* prediction performance comparison among models learned from incomplete data due to an unreliable sensor.

Figs. 2(a)–(c) and 3 for hierarchical prediction accuracies and distance measure, respectively. These comparisons support the trend that the advantage of our models is greatest when the proportions of incomplete data are largest, for the scenario of learning with two unreliable sensors.

The average numbers of iterations required for models to achieve convergence are shown in Table II Scenario 1. The larger the amount of incomplete data which correspond to any schema other than the base schema, i.e., the coarser the information available, the longer it takes for the EM algorithm to achieve convergence for the distribution at the base schema level. However, the induction process is typically carried out off-line, so the increased training time is not a crucial issue.

We then extended the analysis to the learning scenario carried out in the situation of three unreliable sensors, and performance is evaluated using the same criteria. As a demonstration we show the learning of the model from incomplete data due to unreliable "Tea," "Coffee," and "Fridge" sensors, with different amounts of incomplete data. The prediction performance is shown in Fig. 5 and the numbers of iterations to convergence are shown in Table II Scenario 2. Also, the model distance measure to the true distribution reveals the overall performance trend of the models as the fraction of incomplete data increases. Learning and evaluation for other combinations of three unreliable sensors have also been carried out, and the results follow a similar pattern.

To summarize, we have investigated our algorithm in handling incomplete data through systematic sets of experiments. All results have provided promising feedback. They show that our algorithm can learn effectively from incomplete information to achieve a good stable performance in different situations of multiple unreliable sensors in the smart home environment. The overall performance characteristics are further analyzed in the next section.

V. ANALYSIS OF IMPACT OF SENSORS

Comparing across performance of models derived in various situations, it gives rise to the opportunity of the investigation of the importance of each of the individual sensors associated with monitoring and identifying *ADL*. In Fig. 6, we present the prediction performance for task *PA* for models each trained with a different single unreliable sensor. For each value of the proportion of incomplete data, performance comparison between models derived from different individual unreliable sensors indicates the relative impact of each sensor in identifying daily activities.

For further investigation, we consider the extreme case of model performance when one sensor is completely removed. The results are displayed in Fig. 7 for the three prediction tasks of *PHA*, *PA*, and *PE*.

In the results shown in Fig. 6, the "Fridge" sensor is seen to have the greatest impact on identification of activities in our experiment; the next greatest impact is the "Sugar" sensor, with the "Tea" sensor having the least influence. Also, the results show a greater effect on the more detailed prediction tasks for a given missing sensor in Fig. 7. To analyze this effect, we have defined the detail level of the base schema for complete data as

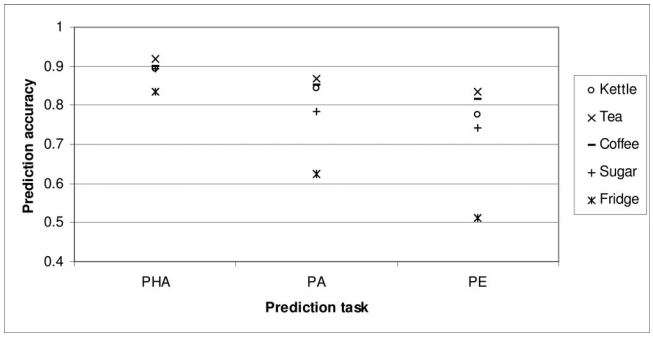


Fig. 7. Prediction accuracies on tasks of *PHA*, *PA*, and *PE* by models learned from incomplete data generated when removing an individual sensor.

the finest level of information. The schema structures for different types of incomplete data due to various unreliable sensor(s) have a coarser information level than the base schema, the level varying according to the type of missing sensor. The schema for incomplete data in the case of the unreliable "Fridge" sensor is much coarser than the schema for other single malfunctioning sensors, and thus the unreliable "Fridge" sensor has the greatest affect on the model performance.

We have thus established a method that can help to evaluate the minimal configuration of sensors to recognize activities in a smart home. We can identify sensors that have a high impact on activity identification when missing, so that these sensors can be provided with regular maintenance. Less informative sensors can be identified and removed to reduce costs. Such information can, therefore, guide the system design by providing information on redundancy of noncritical sensors.

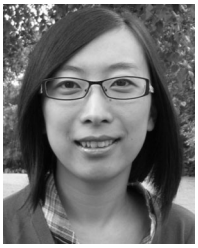
VI. Conclusion and Future Work

In this paper, we have proposed a learning algorithm to address the challenge of learning multi-inhabitant activity profiles from a limited number of data gathered from routine activities carried out in the smart home, where unreliable incomplete information is caused by unreliable low-level sensors. Activity recognition in such an environment is carried out based on the learned model on three levels of detail. The algorithm is evaluated systematically in scenarios of different types and amount of incompleteness. All our experiments yield consistent results that demonstrate that our algorithm is capable of learning from incomplete information, and can tolerate a reasonably high proportion of incomplete data while still achieving good prediction performance. Incomplete data are still informative and can be used to improve learning performance. Results have shown that performance improvements are made by incorporating these incomplete data, with the improvements becoming greater with an increase in the proportion of incompleteness compared to the prediction performance of "Bench" models derived from complete data only. Our algorithm provides a flexible learning framework across a range of schema structures; the approach also provides an opportunity to assess the impact of particular sensors in activity recognition. Such assessment can inform the effective selection and maintenance of sensor devices for monitoring activities in a smart home environment.

Part of this work is further used in decision making for a personalized intervention process to support independent living [12], [19]. Other future work includes exploring the deductive capability of the learning framework in a real-time application, detection of change, scalability evaluation, and the incorporation of other information such as duration data in the learning. In this study, we have used only labeled data. We will also investigate the generalization of our learning algorithm to semisupervised learning from both labeled and unlabeled data, seen as incomplete information, in order to reduce the cost of collecting expensive labeled training data.

References

[1] R. Harper, *Inside the Smart Home*. London: Springer, 2003, pp. 17–39.
[2] A. Ranganathan, J. Al-Muhtadi, and R. H. Campbell, "Reasoning about uncertain contexts in pervasive computing environments," *IEEE Pervasive Comput*, vol. 3, no. 2, pp. 62–70, Apr./Jun. 2004.
[3] H. Hagras, "Embedding computational intelligence in pervasive spaces," *IEEE Pervasive Comput*., vol. 6, no. 3, pp. 85–89, Jul./Sep. 2007.
[4] E. O. Heierman (III) and D. J. Cook, "Improving home automation by discovering regularly occurring device usage patterns," in *Proc. 3rd IEEE Int. Conf. Data Mining*, 2003, pp. 537–540.
[5] K. Gopalratnam and D. J. Cook, "Online sequential prediction via incremental parsing: The active LeZi algorithm," *IEEE Intell. Syst*., vol. 22, no. 1, pp. 52–58, Jan./Feb. 2007.
[6] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel, "Inferring Activities from Interactions with objects," *IEEE Pervasive Comput*., vol. 3, no. 4, pp. 50–57, Oct./Dec. 2004.
[7] T. S. Barger, D. E. Brown, and M. Alwan, "Health status monitoring through analysis of behavioral patterns," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 35, no. 1, pp. 22–27, Jan. 2005.
[8] D. J. Cook, "Health monitoring and assistance to support aging in place," *J. Universal Comput. Sci*., vol. 12, no. 1, pp. 15–29, 2006.
[9] X. Hong, C. Nugent, M. Mulvenna, S. McClean, B. Scotney, and S. Devlin, "Evidential fusion of sensor data for activity recognition in smart homes," *Pervasive Mobile Comput*., vol. 5, no. 3, pp. 236–252, Jun. 2009.
[10] T. Martin, B. Majeed, B.-S. Lee, and N. Clarke, "A third-generation telecare system using fuzzy ambient intelligence," *Studies Comput. Intell*., vol. 72, pp. 155–175, 2007.
[11] M. G. Meghan, "Activities of daily living evaluation," in *Encyclopedia of Nursing & Allied Health*, K. Krapp, Ed. Farmington Hills, MI: Gale Group, Inc., 2006.
[12] S. Zhang, S. McClean, B. Scotney, X. Hong, C. D. Nugent, and M. D. Mulvenna, "An intervention mechanism for assistive living in smart homes," *J. Ambient Intell. Smart Environments*, vol. 2, no. 3, pp. 233–252, 2010.
[13] L. Le Cam, Lucien, "Maximum likelihood: An introduction," *ISI Rev*., vol. 58, no. 2, pp. 153–171, 1990.
[14] M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*, 3rd ed. New York: Wiley, 2000, pp. 134–136.
[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood for incomplete data via the EM algorithm," *J. Amer. Statist. Soc., Series B*, vol. 39, pp. 1–38, 1977.
[16] S. McClean, B. Scotney, P. Morrow, and K. Greer, "Integrating semantically heterogeneous aggregate views of distributed databases," *Distrib. Parallel Databases*, vol. 24, no. 2, pp. 73–94, 2008.
[17] S. McClean, B. Scotney, and M. Shapcott, "Aggregation of imprecise and uncertain information in databases," *IEEE Trans. Knowledge Data Eng*., vol. 13, no. 6, pp. 902–912, Nov./Dec. 2001.
[18] MATLAB, The language of Technical Computing (Service Pack 3), in Version 7.1.0.246 (R14), The MathWorks, Inc., Natick, MA, Aug. 2005.
[19] S. Zhang, "Learning from semantically heterogeneous aggregate data in a distributed environment," Ph.D. dissertation, Faculty Comput. Eng., Univ. Ulster, 2009, pp. 98–115.



Shuai Zhang received the B.Sc. degree in computer science from Heilongjiang University, Harbin, China, the M.Phil. degree in visual arts data mining from the University of Bradford, Bradford, U.K., and the Ph.D. degree in intelligent data analysis from the University of Ulster, Coleraine, U.K.

She is currently a Research Associate in the School of Computing and Information Engineering, University of Ulster. Her research interests include intelligent data analysis on semantically heterogeneous aggregate data in a distributed environment, learning inhabitants' behavioral patterns from unreliable low-level sensor data in smart environment to support assisted living, modeling for connected health applications, and health technology assessment.

Sally I. McClean (M'98) received the B.Sc. degree in mathematics from Oxford University, Oxford, U.K., the M.Sc. degree in mathematical statistics and operational Research from Cardiff University, Cardiff, U.K., and the Ph.D. degree from the University of Ulster, Coleraine, U.K.

She is currently a Professor of mathematics in the School of Computing and Information Engineering, University of Ulster. Her research interests include mathematical modeling, applied probability, multivariate statistical analysis and applications of mathematical and statistical methods to computer science, particularly database technology.

Dr. McClean is a Fellow of the Royal Statistical Society, a member of the IEEE Computer Society, a member of the Operational Research Society, and an Associate Fellow of the Institute of Mathematics and its Applications.

Bryan W. Scotney received the B.Sc. degree in mathematics from the University of Durham, Durham, U.K., and the Ph.D. degree in mathematics from the University of Reading, Reading, U.K.

He is currently a Professor of informatics and Director of the Computer Science Research Institute, University of Ulster, Coleraine, U.K. He joined the University of Ulster in 1984 as a Lecturer in mathematics. He has more than 160 publications, spanning a range of research interests in mathematical computation, especially in digital image processing and computer vision, pattern recognition and classification, statistical databases, reasoning under uncertainty, and applications to healthcare informatics, official statistics, biomedical and vision sciences, and telecommunications network management.

Dr. Scotney is the President of the Irish Pattern Recognition and Classification Society, and a member of Council of the International Federation of Classification Societies.