Correlating Sensors and Activities in an Intelligent Environment: A Logistic Regression Approach

Fahd Al-Bin-Ali[1, 2], Prasad Boddupalli[1, 2], Nigel Davies[1, 2], and Adrian Friday[2]

[1] Computer Science Department, University of Arizona, Tucson, AZ 85721, USA
{Albinali, Bprasad}@cs.arizona.edu
[2] Computing Department, Lancaster University, Lancaster, LA1 4YR, UK
{Nigel, Adrian}@comp.lancs.ac.uk

Abstract. An important problem in intelligent environments is how the system can identify and model users' activities. This paper describes a new technique for identifying correlations between sensors and activities in an intelligent environment. Intelligent systems can then use these correlations to recognize the activities in a space. The proposed approach is motivated by the need for distinguishing the critical set of sensors that identifies a specific activity from others that do not. We compare several correlation techniques and show that logistic regression is a suitable solution. Finally, we describe our approach and report preliminary results.

1 Introduction

In his classic paper "The Computer for the 21st Century" [14] Weiser envisions a world of intelligent environments that are highly aware of their inhabitants. In this vision, physical spaces are enhanced with computing capabilities to act more intelligently: they observe, interact with and react to humans in meaningful ways. They understand human reasoning, analyze behaviors and infer intentions. Furthermore, intelligent environments actively collaborate with their inhabitants to assist them in making their surroundings more pleasant. Intelligent environments even take decisions and execute actions on their own. They become integral participants in the daily human activity.

A critical element that Weiser anticipated, yet has not been achieved, is the invisibility of pervasive systems. The ability of such systems to disappear into the background of everyday life is dependant on their ability to correctly interpret the state of the environment and to act accordingly: intelligent systems that incorrectly interpret the state of the world or the intentions of users are likely to take inappropriate actions that are not naturally anticipated by users [6]. Such incorrect actions could become very disruptive and intrusive to users, they distract the inhabitants of intelligent spaces from their ongoing activity and therefore, they make them more aware of the

E. Aarts et al. (Eds.): EUSAI 2003, LNCS 2875, pp. 318–333, 2003.

system. This paper begins to address the challenge of designing less intrusive intelligent environments that can engage in richer and more meaningful interactions with users. We believe that such systems must have a deep understanding of user context and, specifically, should have an understanding of activities that a user is engaged in. Our approach is thus inspired by concepts from activity theory [9] and requires support for three basic system functions:

- *Sensing context:* By observing and monitoring users' context, intelligent systems can collect information about the intelligent space and its inhabitants.
- *Analyzing context:* By analyzing users' context, intelligent systems can estimate and interpret users' activities.
- *Gracefully reacting to the inhabitants:* By understanding users' activities, intelligent systems can react unobtrusively to their inhabitants and therefore can potentially become more invisible.

In this paper, we focus on one aspect of our system design, i.e. how to identify sensors that correlate with activities in an intelligent space. First, we motivate our use of an activity-centric approach and justify the need for precisely identifying correlations between sensors and activities. Second, we identify a number of desirable properties for activity-aware intelligent systems. We then analyze different techniques for identifying the correlations between sensors and activities and show that statistical logistic regression has the desired properties. Third, we describe in detail our regression technique. Finally, we report preliminary results and state our conclusions.

2 Why an Activity-Centric Approach?

Intelligent environments are inherently social and collaborative spaces. Understanding the "behavioral-level" interaction in such environments require modeling the context in which the inhabitants of the space interact [6]. Early research [3],[12] in intelligent environments focused on establishing simple relationships between tangible context and appropriate actions, for example, switching on and off devices based on user proximity. Intangible context such as activities, human moods and human intentions and complex relationships between sensor data and actions have not received significant attention to date. However, to be invisible, intelligent systems must understand both tangible and intangible aspects of context and the complex relationships between sensors and actions.

We believe that the best method for capturing these complex relationships is using the notion of 'activities'. Many earlier projects acknowledge a need for such a capability. For example, MIT utilizes an activity based approach in their second generation iRoom [11]. EasyLiving [3] from Microsoft acknowledges the need for tracking activity in an intelligent environment. Responsive Offices [8] from Xerox PARC identifies activity as an essential ingredient for determining appropriate reactive behaviors. Moreover, numerous studies in psychology [9] advocate that individual and group behavior should be interpreted in relation to the activities people participate in. Indeed, recent work on groupware [2] has employed many of these concepts (in par-

ticular concepts from activity theory) for modeling collaborative tasks. Such systems interpret behaviors by considering the activity as the fundamental unit of analysis.
Figure 1 shows a high-level view of our activity analysis system. Initially, sensors in the intelligent space are correlated with activities that interest the inhabitants. The system uses empirical data (collected from the space) to derive causal correlations between activities and sensors. The correlations are then used to create a *correlation matrix* that captures all the correlations between activities and sensors in an intelligent space. Subsequently, the intelligent system can use the matrix to interpret the activities in the intelligent space, for example, a probabilistic reasoner can use the matrix for building a Bayesian network to analyze the activities. This might involve assessing the uncertainties in the reasoner's inferences or establishing a dialogue with the inhabitants of the space to disambiguate activities in situations of high uncertainty as proposed by [4],[5].

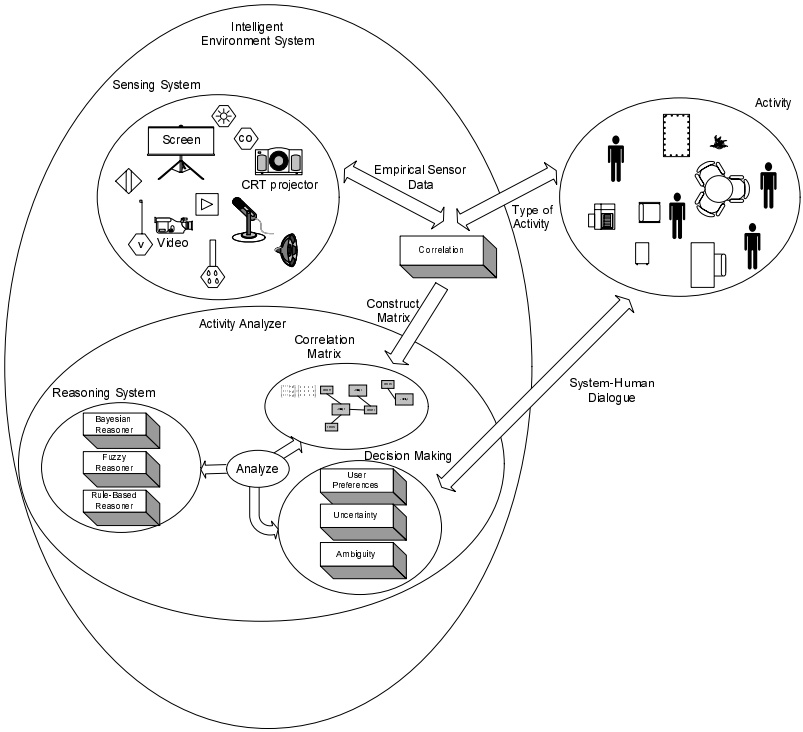

Fig. 1. Activity-aware intelligent space

It is important to emphasize that in a ubiquitous environment that is saturated with sensors, it is extremely important to distinguish the *critical set* of sensors that correlate with a specific activity from others that do not. For example, imagine constructing a Bayesian network for the activities in a ubiquitous space without knowing the dependencies between sensor readings and the activities. Including uncorrelated sensors in the Bayesian network will result in inaccuracies that can potentially misguide

the reasoner. Similarly, excluding correlated sensors from the Bayesian network could result in ignoring some important aspects of the activities that can also misguide the reasoner. This paper focuses on how to determine the *critical set* of sensors that correlate to activities and proposes a new technique for accomplishing that. We begin our discussion by examining some of the desirable properties for activity-aware intelligent systems.

3 Desirable Properties for Activity-Aware Intelligent Systems

Few intelligent environments exist, and those that do are confined within research labs. Therefore, to identify the desirable properties for activity-aware intelligent systems, we examined recent work on intelligent environments [3],[12],[13], studies from psychology on individual and group behavior [9], work on natural and multimodal human-computer interaction (HCI) [6,7] and connectionist and statistical modeling techniques [5],[10],[12]. These efforts led us to the following desirable properties:

3.1 Transparency and Comprehensibility

Intelligent systems must support transparent activity modeling. Transparent modeling enables intelligent environments to reason in ways that are comprehensible to their inhabitants. Such a property is critical in order that it is possible to formulate precisely how systems reached particular decisions. Subsequently, this information could be relayed to the inhabitants of an intelligent space to support a dialogue with the system as proposed by [4,5] to fix any incorrect actions.

3.2 Adaptability

Intelligent systems must be adaptable to endure the highly dynamic nature of ubiquitous environments. Such adaptability must apply to both physical reconfiguration of spaces (e.g. changes in the availability of sensors) and to changes in activity patterns within these spaces. Different systems will require different forms of adaptability including offline adaptability in which sensor data is logged for later analysis and online adaptability in which sensor data is examined and adaptation is performed while the system is in use.

3.3 Accuracy

Clearly, achieving high accuracy in terms of identifying the activity in an intelligent environment from a given set of sensor data is crucial. However, it should be noted that the exact requirements in terms of accuracy are actually a property of the entire

system and are influenced by the significance of the actions that will be triggered: users will perceive the activity analysis process as accurate and indeed as invisible when the system's reactions are correct. However, this does not necessarily mean that the system has identified the users' activities correctly. For example, imagine a user having a nap while watching TV. An intelligent system might detect a reduction in the overall mobility in the space and therefore infers that no one is in the room; resulting in switching off the TV and the lights. Clearly, the analysis process misdiagnosed the activity, but the outcome is still considered correct by the user.

3.4 Knowledge Portability

It is important that knowledge about users and their activity patterns can be moved between intelligent environments, reflecting user mobility inherent in the real world. This will require a clear separation between the models that represent the system's knowledge about activities and the system-specific assumptions and mechanisms. In practice, achieving portability is likely to be extremely complex, raising many technical challenges (e.g. determining the equivalence between sensors in different environments) and non-technical challenges in areas such as legal and social ethics (e.g. can models about activity patterns be exchanged between private places and public places without violating the privacy of people?).

So far, we have described 4 desirable properties for activity-aware intelligent systems. It should be clear that the above properties are not exhaustive, but we have deliberately chosen them because of their importance in the context of intelligent environments. It should also be noted that many of the properties discussed above are greatly exacerbated when multiple people are participating in an activity.

4 Techniques for Correlating Activities and Sensors

Several techniques can be conceived for correlating activities and sensors including: expert correlation, statistical correlation and connectionist correlation. We briefly describe these approaches and we analyze their merits and demerits.

4.1 Expert Correlation

The easiest way to correlate activities and sensors in an intelligent space is to use the opinion of an expert who is familiar with the space. For example, in a smart classroom, a teacher can identify different activities that students participate in such as pop quiz, discussion, on-board problem solving exercise etc. Subsequently, a rough mapping could be made between these activities and the available sensors in the classroom. Gaia [13] uses this approach for activity analysis where the inhabitants of the intelligent space identify the correlations and construct a belief network that models

their activities. This network is then used by a Bayesian reasoner to identify the activities.

Expert correlation suffers from several limitations. Firstly, it does not scale well: as more activities and sensors are introduced, it becomes harder for human experts to assess the correlations. Secondly, people might have different views about the degrees of correlation between sensors and activities. Therefore, relying on the subjective assessment of a particular individual might lead to inaccuracies. Thirdly, adapting the correlations to the dynamic nature of a ubiquitous space requires a human expert: an undesirable proposition especially when intelligent spaces host rapidly changing activities. Hence, we believe that expert correlation is of limited use in ubiquitous environments that are heavily saturated with sensors.

4.2 Connectionist and Statistical Correlation

Alternatively, connectionist or statistical techniques can be used to identify correlations between sensors and activities. Connectionist correlation relies on neural network analysis that identifies patterns between different inputs. This approach has been used in the neural house project [12] where a neural network observes the lifestyle of the inhabitants of a house and programs itself accordingly. Similarly, statistical techniques such as regression can identify potential causal relationships between different variables. In ubiquitous environments, these two techniques can certainly handle large amounts of data that human experts find cumbersome. Several research studies [5],[12] have affirmed that neural techniques are more accurate than regression techniques owing to their ability to capture non-linear correlations automatically. However, they suffer from the incomprehensibility of the decision making process, i.e. it is very hard to reconstruct the rationale of a neural network of why a particular correlation between a sensor and an activity is strong. In contrast, statistical techniques are based upon “well understood models of behavior” and therefore, it is usually easier to reconstruct the rationale behind their decision making process [5]. Moreover, adapting neural networks to the continuous changes in an intelligent space might often require retraining the whole network which can be an expensive process especially in cases that require on-line adaptation.

4.3 Analysis

In light of the above discussion, we can see that expert correlation is not a viable solution due to its vulnerability to inaccuracies and its inability to deal with the abundance of sensor information in ubiquitous environments. In contrast, regression and neural networks can deal with the richness of sensor information in such environments. However, regression provides a more comprehensible framework than neural based techniques thereby making it more suitable for supporting transparent modeling where users can establish a dialogue with the system. Moreover, reapplying regression to adapt to the dynamic nature of a ubiquitous space is likely to incur less over-

head than retraining a neural network. Finally, it should be noted that although regression is less accurate than neural techniques, its outcome is still comparable [5]. However, it would be unfair to give the impression that neural analysis is unusable, while regression is completely without problems. The major conceptual limitation in regression is that it can never identify the underlying causal mechanism. For example, one would find a strong positive correlation between the number of users attached to a particular access point in a conference hall and the presentation activities taking place. Do we conclude that a presentation activity causes an increase in the number of users attached to an access point? Even though that might be the case in this simple example, in many other cases, the causal explanations might not be obvious. Moreover, as the number of variables increase, more empirical observations are required to avoid having significant correlations while in fact one or more variables are capitalizing on chance. Finally, even though the rationale behind correlations is potentially easier to reconstruct using regression, it is unclear how easy it is to relay that information to regular inhabitants of an intelligent space that have no prior knowledge of statistics. Undoubtedly, friendly means should be developed to enable such system-user dialogues. We acknowledge these problems and recognize the need for exploring them further.

5 Multinomial Logistic Regression

Multinomial Logistic Regression (MLR) [10] is a statistical technique that investigates and models relationships between a dependent variable and one or more independent variables. It is typically used when a dependent variable has the following properties:

- *Categorical*: The dependent has a limited set of values (e.g. for an activity {presentation=0, break=1, lunch=2}) that could be ordinal (e.g. {strongly agree, agree, disagree}) or non-ordinal.
- *Mutually Exclusive*: Any instance of a dependent cannot be classified as belonging to more than one category. For example, considering an activity as a dependant, an instance of an activity cannot be a presentation and a break at the same time.
- *Polychotomous:* The dependent can have 2 or more categories. A special case of MLR is the binomial logistic regression that deals with the dependent when it is a dichotomy.

MLR can deal with independents of any type (e.g. continuous, discrete, dichotomous, polychotomous etc.). Generally speaking, MLR has less stringent requirements than conventional regression techniques including:

1. It does not assume linearity of relationship between the independent variables and the dependent.
2. It does not require normally distributed variables.
3. It does not assume homoscedasticity (i.e. the variance around the regression fit is the same).

Details of logistic regression techniques can be found in [10], below we explain only those aspects critical to our discussion. In particular, we explain how to assess the adequacy of a logistic regression.

5.1 Logistic R^2

The logistic R^2 measures the strength of the association between the dependent variable and the independents. It should be noted that the logistic R^2 is different from the R^2 in conventional regression. The latter measures the goodness of fit relying on the variance around the regression fit. However, the variance of categorical dependent variables depends on the frequency distribution of that variable and therefore logistic R^2 just reflects the strength of the association.

5.2 Classification Percentage

The classification percentage reflects how good a logistic regression formula is in estimating the correct categories of a dependant. In a perfect model, the estimated values are the exact actual values making the overall classification percentage 100%. It should be noted that the classification process relies on a probability cutoff where higher cutoffs mean more sensitivity in the classification process.

5.3 Model Chi-Square Test

It is very important to determine the effect of each independent in the logistic formula. For example, the formula might show better correlation without some independents or with some additional independents. Model Chi-Square is a technique that measures the improvement in a fit that an independent variable makes compared to the null model (i.e. model without independents). This technique uses the null hypothesis to test for individual significance. The null hypothesis says that an independent variable coefficient has no effect on the dependent variable. Therefore, rejecting the hypothesis means that the independent should not be deleted from the formula because it has a significant contribution. While accepting the null hypothesis means that the independent variable is insignificant and therefore should be deleted. Generally speaking, when the probability of the Model Chi-Square is less than 0.05, the null hypothesis is rejected.

So far, we have described some important concepts for our following discussion. Next, we describe how to use logistic regression for identifying the correlations in an intelligent environment.

6 Correlation in Intelligent Environments

In the context of intelligent environments, we are using MLR for identifying the *critical set* of sensors that highly correlate with activities in an intelligent space. Sensor data is collected for some period of time while users are required to record their activities. The system records this information along with statistical data from all sensors. The collected data is then analyzed by a logistic regression engine to identify the sensors that are showing high correlation with the activities. The output of the regression engine takes the form of a correlation vector.

Definition 1. In an activity-aware environment with the following properties:

- A is a set of n+1 activities defined by $\{a_1, a_2, \cdots, a_n\} \cup \{a_\varphi\}$ where a_φ denotes the unrecognized activity, and
- S is a set of k sensors in the intelligent space defined by $\{s_1, s_2, \cdots, s_k\}$,

a Correlation Vector (CV) identifies the critical set of sensors that highly correlate with 1 or more activities (and is thus influential in identifying the activities). A CV has the following form:

$$CV(A') =< c_1, c_2, ..., c_k > \quad where\ A' \subset A\ and \quad c_i \in \{0,1\}\ and\ 1 \le i \le k \tag{1}$$

where c_i reflects the correlation between the activities belonging to A' and sensor s_i such that: $c_i = 0$ indicates no or insignificant correlation and $c_i = 1$ indicates significant correlation. For example, in a space with three sensors $\{ s_1 = temperature\ sensor,\ s_2 = projector\ \ sensor,\ s_3 = people\ count\ sensor \}$, a correlation vector for a presentation activity might look as follows:

$$CV(A' = \{presentation\}) =< 0,1,1 > \tag{2}$$

This indicates that a presentation activity is highly correlated with the projector sensor and the people count sensor but not with the temperature sensor.

6.1 Sensor Selection and the Correlation Matrix

Our current regression engine relies on the Chi-Square test and the classification percentage to determine the CV. We can configure the engine to select the highly correlated sensors in one of two ways:

1. *Forward Selection:* In this procedure, the best sensor is found. Next, the sensor that adds the most to the logistic fit is included. This process continues until specific cutoff thresholds (in the Chi-Square and in the classification percentage) are reached or none of the sensors add a significant value to the strength of the association.

2. *Backward Selection:* In this procedure, all the sensors are initially included in the logistic model. Subsequently, sensors are deleted from the model based on their level of significance. Again, this process continues until all the sensors left are at a specific significance level.

It should be noted that higher cutoff values reduce the number of sensors that correlate with particular activities. This potentially simplifies the reasoner's logic, for example, a rule-based reasoner that relies on a small set of sensors is likely to generate simpler rules than reasoners that account for a big set of sensors.

Furthermore, sensor selection is directly influenced by the number of categories of the dependent. When the logistic engine is given some empirical data, it tries to account for all the categories of the dependent using one single logistic formula. For example, imagine a space with the configuration shown in Figure 2 where the links between sensors and activities indicate the presence of a correlation.

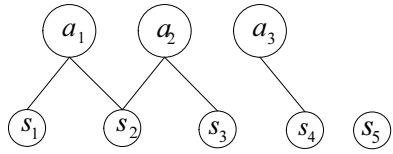

Fig. 2. Example intelligent environment

Analyzing empirical readings from the above space, the logistic regression engine produces the following CV:

$$CV(A' = \{a_1, a_2, a_3\}) =< 1,1,1,1,0 > \tag{3}$$

Obviously, the CV fails to reflect the exact dependencies shown in Figure 2. This can potentially result in inaccuracies when disambiguating activities. For example, suppose an intelligent system wants to disambiguate two particular activities $\{a_1, a_2\}$, relying on the above CV includes s_4 which is uncorrelated to the two activities. Clearly, this can potentially misguide the reasoner. To resolve this issue, we use binomial logistic regression to identify the CV of sensors for each activity with respect to a_ϕ (i.e. the no activity state). We give the resulting CV a special name: the *Reference Correlation Vector* (RCV). In the example shown in Figure 2, the RCVs are:

$$\begin{aligned} RCV(a_1) &= CV(A' = \{a_1, a_\phi\}) =< 1,1,0,0,0 > \\ RCV(a_2) &= CV(A' = \{a_2, a_\varphi\}) =< 0,1,1,1,0 > \\ RCV(a_3) &= CV(A' = \{a_3, a_\phi\}) =< 0,0,0,1,0 > \end{aligned} \tag{4}$$

Notice that the RCVs precisely reflect the dependencies shown in Figure 2. More importantly, the disjunction of RCVs is the CV for the union of their activities. For

example, the disjunction of the above 3 equations is the correlation vector for all the activities shown in Figure 2:

$$\begin{aligned} & RCV(a_1) \vee RCV(a_2) \vee RCV(a_3) \\ & =< 1,1,0,0,0 > \vee < 0,1,1,0,0 > \vee < 0,0,0,1,0 > \\ & =< 1,1,1,1,0 > \\ & = CV(A' = \{a_1, a_2, a_3\}) \end{aligned} \tag{5}$$

Furthermore, combining RCVs of all activities is simply a matrix that represents all the *critical correlations* in an intelligent environment. We call this a *correlation matrix.* The following equation shows the general form of this matrix:

$$\begin{bmatrix} RCV(a_1) \\ RCV(a_2) \\ \vdots \\ RCV(a_n) \end{bmatrix} = \begin{bmatrix} c_1^1 & c_1^2 & \cdots & c_1^k \\ c_2^1 & c_2^2 & \cdots & c_2^k \\ \vdots & \vdots & \vdots & \vdots \\ c_n^1 & c_n^2 & \cdots & c_n^k \end{bmatrix} \tag{6}$$

where $n = (|A| - 1))$ *and* $k = |S|$ *and* $c_j^i \in \{0,1\}$ *and* $1 \le i \le k$ *and* $1 \le j \le n$

The above matrix can also be represented as a simple correlation graph. Figure 3 shows an example of a graph that correlates 4 sensors with 3 activities.

Finally, we note that a correlation matrix does not reflect the exact degree of correlation between a particular activity and its sensors. However, the degree of correlation can be roughly estimated using the reduction in the logistic R^2 of the model as a result of omitting the term of a particular sensor from the regression formula. We further elaborate on this particular issue in the results section.

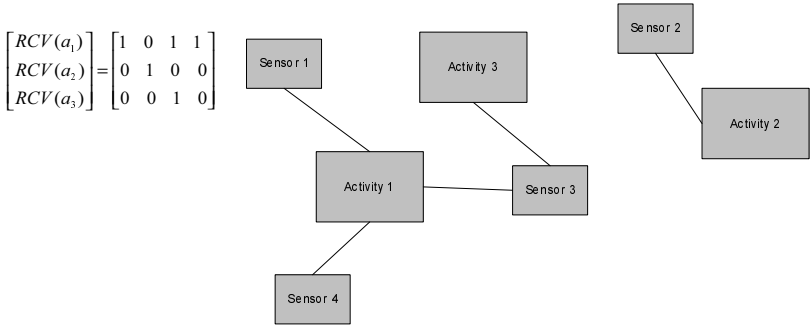

Fig. 3. Correlation matrix and correlation graph

6.2 Using the Correlation Matrix

Referring back to our discussion about an activity centric approach, we highlighted the need to identify the activities in an intelligent space. We explained that the abundance of sensors in ubicomp environments complicates activity analysis: including

uncorrelated sensors or excluding correlated sensors from the decision making process can potentially mislead any reasoner. The correlation matrix (described above) serves as a filter that reflects the strong dependencies between activities and sensors in an intelligent space. Reasoners that use the correlation matrix will deal with a reduced set of sensors that are highly correlated with the activities they are trying to recognize. Clearly, this simplifies the task of a reasoner.

In situations of high uncertainty, intelligent systems fail to identify the activities with reasonable confidence. The logistic engine can be used with higher cutoff values to determine the sensors that show the highest correlation and therefore could be considered more reliable. These sensors can then be used to identify the activities. Moreover, when sensors are removed from the space, their values are replaced with zeros in the matrix. Depending on the accuracy of the classification process and the weight of the removed sensors, the system might decide to include one or more correlated sensors to compensate for the removed sensors. Similarly, when sensors are added to the space, the system gathers empirical data from the new sensors. Subsequently, activities that are frequently misclassified can be reexamined with the new sensors included for potentially improving the classification process.

7 Preliminary Results

In this section, we describe preliminary results of our approach. We use publicly available traces recorded over three days at the ACM SIGCOMM'01 conference (held at U.C. San Diego in August 2001) to demonstrate that logistic regression is effective in correlating sensors with activities. A detailed description of these traces can be found in [1]. The traces record data samples from wireless access points serving the conference. Note that due to the lack of availability of information on the no activity state (a_ϕ), we are unable to calculate the RCV and therefore we present measurements based on analysis of the CV.

Two important pieces of information can be identified: the number of mobile nodes attached to a particular access point and the load on each access point. These two quantities will serve as sensors for our experiment. In addition, two different activities can be identified including: sessions and breaks. Intuitively, we would expect a corre lation between these sensors and the activities. For example, during breaks the load over an access point is likely to drop and therefore the load sensor will show a negative correlation with breaks. Our experiments used 100 sample readings from one day to identify the logistic regression formulae. We also performed 3 experiments to classify 100 activities using the regression formula with the throughput sensor only, the number of nodes sensor only and both sensors.

7.1 Throughput and Activities

First, we characterize the strength of the correlation between the throughput at the access point and the activities in the conference hall. Figure 4 shows the proportion of the correctly classified activities using the regression formula for different cutoffs. From the figure, we see that the throughput sensor can indeed classify all the activities correctly when the cutoff is very low (i.e. we accept classifications with a broad error margin). However, its accuracy decreases rapidly as the cutoff is increased (i.e. demanding less deviation from the categories). With a 0.5 cutoff the regression formula classifies 83% of our test cases correctly.

We also found that the logistic R^2 for the regression formula is equal to 0.55. This reflects a moderate association between the throughput and the activities.

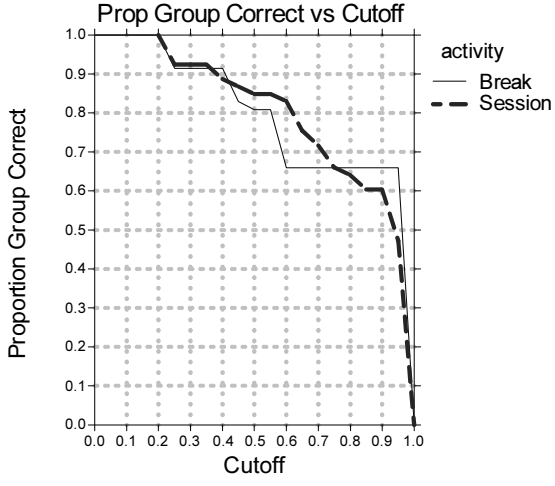

Fig. 4. Proportion of correct classification Vs. cutoff (using throughput)

7.2 Number of Nodes and Activities

Our second experiment characterizes the strength of the correlation between the number of nodes attached to the access point and the activities in the conference hall. Figure 5 shows the proportion of the correctly classified activities using the regression formula for different cutoffs. From the figure, we see that the number of nodes sensor can also classify all the activities correctly for low cutoffs. However, the sensor is more robust to higher cutoffs than the throughput sensor. In other words, its accuracy decreases more slowly than that in the throughput case as the cutoff increases. With a 0.5 cutoff the regression formula classifies 92% of our test cases correctly. We also found that the logistic R^2 for the regression formula is equal to 0.966. This reflects a strong association between the number of nodes and the activities.

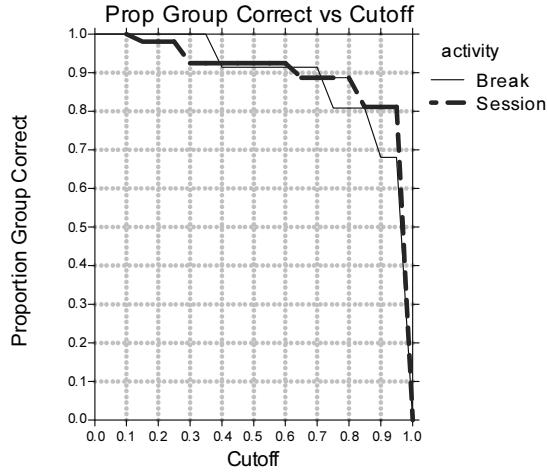

Fig. 5. Proportion of correct classification Vs. cutoff (using number of users)

7.3 Number of Nodes, Throughput, and Activities

Finally, our third experiment characterizes the strength of the correlation between both the number of nodes and their throughput, and the activities in the conference hall. Figure 6 shows the proportion of the correctly classified activities using the regression formula for different cutoffs. From the figure, we see that the regression formula can still classify all the activities correctly with low cutoffs. However, the classification percentage does not seem to improve from the one that uses the number of nodes only. With a 0.5 cutoff the regression formula classifies 92% of our test cases correctly.

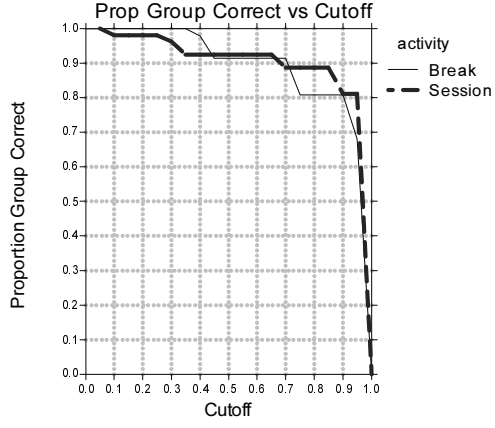

Fig. 6. Proportion of correct classification Vs. cutoff (using throughput and number of users)

We also found that the logistic R^2 for the regression formula is equal to 0.80. This means that the strength of the association between the number of nodes and the throughput, and the activities is significant.

7.4 Sensor Selection and the Correlation Matrix

When the logistic regression engine performed forward and backward selection on the (throughput and number of nodes) regression, it omitted the throughput in both cases. First, the engine measured the reduction in R^2 when omitting the throughput term. This reduced the strength of the association by 0.00172. Second, the engine measured the reduction in R^2 when omitting the number of nodes term. This resulted in a reduction of 0.24. Clearly, including the throughput sensor does not improve the strength of the association between the activities and the independents significantly. Moreover, including the throughput has not improved the classification percentage beyond 92%. Therefore, the regression engine omitted the throughput from the correlation matrix:

$$\begin{bmatrix} CV(break) \\ CV(session) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad where \quad s_1 = numberOfUsers \qquad (7)$$

$$s_2 = throughput$$

Finally, we note that the reduction in R^2 can be used as a rough estimate for the weights of sensors.

8 Discussion and Future Work

In this paper, we have highlighted the importance of correlating sensors with activities in an intelligent space. Our approach uses logistic regression. We described some desirable properties for activity-aware environments including: transparency and comprehensibility, adaptability, accuracy and knowledge portability. In light of these properties, we analyzed several techniques for correlating activities with sensors including: expert correlation, regression correlation and connectionist correlation. We concluded that regression provides a more comprehensible framework for correlating activities than the other approaches. We then described in detail our logistic regression approach. Finally, we reported preliminary results.

Our plan for future work is to assess our approach in the intelligent environment in our research lab. We are currently developing software components for hardware and software sensors to use them for collecting empirical data. We are also working on building a probabilistic reasoning system that will use our correlation matrix to identify activities in the intelligent space. In addition, we are developing techniques for exporting and importing contextual knowledge across intelligent environments to allow spaces to identify unfamiliar activities using imported knowledge from other

spaces. Finally, we intend to deploy all these components in our research lab and to make our system accessible to a user community that can report on the impact of our system on user perceptions of activity analysis.

References

1. Balachandran, A., Voelker, G., Bahl, P., and Rangan, P.: Characterizing User Behavior and Network Performance in a Public Wireless LAN. In: Proc. ACM SIGMETRICS (2002)
2. Barthelmess, P., and Anderson, K.: View of Software Development Environments Based on Activity Theory. Computer Supported Cooperative Work. Vol. 11 Issue 1–2 (2002)
3. Brumitt, B., Meyers B., Krumm, J., Kern, A., and Shafer, S.: EasyLiving: Technologies for Intelligent Environments. In: HUC (2000)
4. Dey, A., Mankoff, J., Abowd, G. and Carter, S.: Distributed mediation of ambiguous context in aware environments. In: Proc. of UIST (2002) 121–130
5. Dix, A.: Human Issues in the Use of Pattern Recognition Techniques. Workshop on Neural Networks and Pattern Recognition in Human Computer Interaction. King's Manor, York (1991)
6. Dix, A.: Managing the Ecology of Interaction. In: Proc. of Tamodia, First International Workshop on Task Models and User Interface Design. Bucharest, Romania (2002)
7. Dix, A., Finaly, J., Abowd, G., and Beale, R.: Human-Computer Interaction. Prentice Hall (1998)
8. Elrod, S., Hall, G., Costanza, R., Dixon, M., and Des Rivieres, J.: Responsive Office Environments. Communications ACM Vol. 36, 7 (1993) 84–85
9. Engeström, Y.: Learning by Expanding: An Activity-Theoretical Approach to Developmental Research. Helsinki: Orienta-Konsultit (1987)
10. Hosmer, D., and Lemeshow, S.: Applied Logistic Regression. John Wiley & Sons, 2nd Edition (2000)
11. Kulkarni, A.: A Reactive Behavioral System for the Intelligent Room. Master's Thesis in Computer Science and Engineering at the Massachusetts Institute of Technology. Cambridge, MA (2002)
12. Mozer, M. C.: The Neural Network House: An Environment that Adapts to its Inhabitants. In: Proc. of AAAI Spring Symposium. Menlo, Park, CA. (1998) 110-114
13. Roman, M., Ziebart, B., and Campbell, R.: Dynamic Application Composition: Customizing the Behavior of an Active Space. In: PerCom 2003, Dallas-Fort Worth, Texas (2003)
14. Weiser M.: The Computer for the Twenty-First Century. Scientific American (1991)