Modeling Activity Patterns

Nipun Dave, Aditya Gudipati, Shantanu Kurhekar

CS 229, Fall 2010, Stanford University

***Abstract* - This paper describes a simple model based on supervised learning that can learn an individual's activity pattern with the data acquired through their cell phones. This model can thereafter be used to predict an individual's location after a transition and the time of the transition. We also present two supervised learning models that can accurately predict if a strong friendship exists between two individuals or not.**

***Index Terms* – Activity profiling, Activity modeling, location prediction, relationship prediction.**

I. Introduction

Over the past few years, cell phones have become a necessity that people always carry on person. Cell phones have a wide variety of sensors that can gather information about the individual's social context, including location with respect to time, other individuals in proximity and interactions like calls and text messages exchanged. This data can be used to learn and predict an individual's future activity, specifically, in terms of location and time. It can also be used to infer strengths of social relationships between individual users. These predictions enable better targeted-advertising, personalization of smart-phone applications and more efficient allocation of infrastructure, among other purposes.

The Reality Mining group at MIT, in collaboration with Nokia, carried out a project to study complex interactions within social groups in 2004-05 [1]. The original Reality Mining experiment was one of the largest mobile phone projects attempted which leveraged the increasingly widespread use of mobile phones to provide insight into dynamics of individual as well as group behavior. More specifically, the research at MIT addressed questions related to evolution of social networks over time, predictability of user activity, flow of information and inferring topology of a social network from proximity data.

Nathan Eagle and Alex Pentland employed PCA to identify the individual's principal behaviors in terms of their location as a function of time, called as eigenbehaviors. These locations were described with 5 different labels, namely, Home, Elsewhere, Work, No Signal and Off [2]. The individual's behavior during the first half on the day is projected on these eigenbehaviors to obtain their weights, which is then used to predict their behavior for the rest of the day. The eigenbehaviors for the individuals are also used to cluster them and thereby infer their community affiliations. K. Laasonen et al present an adaptive framework, implementable on a phone, for identifying locations and routes important to an individual from cellular network data [3]. This framework uses adaptive weighting of locations based on the number of times the individual goes there and also the time since the last visit. Thereafter, the important route is determined by sorting the locations in terms of their weights in a descending order. A more detailed description of the related work can be found in [2].

Our project presents a model to predict an individual's future location at a much finer level. This model predicts the transitions, rather than the actual locations at each instant of time. We also present two models that determine if two individuals are close friends. We use the data made public by MIT Reality Mining project for building and evaluating our model [1]. This dataset contains logs gathered for 94 individuals at MIT over the course of the academic year of September 2004 - June 2005. These individuals were Sloan Business school students, Media Lab incoming students, Media Lab senior students or MIT Staff. Overall, this data represents over 350,000 hours of continuous data on human behavior. More specifically, it contains information about passive behavior such as location (from cell tower ids), other proximate individuals (from Bluetooth device discovery scans), and active behavior such as phone activity, including voice calls and text messages, active applications (such as the calendar or games), and the phone's charging status. It also had self-reported relational data for each individual, where they reported their proximity to, and friendship with, others.

Section 2 gives a description of the different models employed predicting an individual's location after a transition and the time of transition. Section 3 presents the models for capturing the factors in a relationship and thereafter predicting if a particular relationship is close friendship or not. An evaluation of these models is given in Section 4. Section 5 presents our conclusions.

II. Location Prediction

Most individuals have a well-defined routine in terms of their daily activities and social context of these activities, which includes time, location and their companions. This regularity is definitely not a trait exclusive to the subjects considered for the study. Typically, a weekday comprises leaving home in the morning, traveling to work, breaking for lunch, and returning home in the evening. These daily routines are also coupled with routines across other temporal scales, such as going out with friends on weekends, or spending time with family during the holidays.

By capturing knowledge about an individual's past activities and the context into an appropriate model, their future activity can be predicted, given the current context. There are two equivalent ways of modeling this system - based on the individual's transitions or based on their actual location versus time. We focus on the former wherein we predict the individual's transitions. Specifically, this includes predicting

when the transition will occur and where the individual would be after the transition.

In the data, the location is determined in terms of the cell tower IDs. Each handoff from one cell to another is logged with the transition time and the tower ID of the new cell. The cells are then grouped together into areas and are assigned an area ID. We focus on predicting the location at the level of these areas, rather than individual cells. We are not assuming any restrictions on the memory or computational complexity presently.

A. Feature Selection

The future location of an individual typically depends on the following factors:

1. **Current location:** An individual can move to only one of the neighboring areas from the current location.
2. **Day of the week:** Depending on the day of the week, an individual might or might not visit a particular area.
3. **Time of the day:** The time of the day determines if the individual is more likely to be at home, work or traveling.

While there are many more parameters that have an impact on an individual's future location, we consider only these three for the sake of simplicity.

The individuals visited 992 different areas over a span of 9 months. Taking into account the fact that, most of the transitions from a particular area are into the same area or one of the adjacent areas, the number of nonzero transition probabilities become significantly smaller. We determined the set of areas adjacent to an area for all the 992 areas through the transitions logged across all users. Thereafter, the transition probabilities from one area were learned only for a transition to the same area or one of these neighboring areas.

B. Supervised Learning Models

For the listed set of parameters, we tested a supervised learning algorithm, Naïve Bayes, to obtain results for accuracy of prediction. These parameters are certainly not independent of each other. We evaluated this model over different sets of independency assumptions to determine the one which gave the best performance. The baseline for performance was established with the Random Guessing model.

1) Random Guessing

The probability of being in a location is learned using the data logs for each individual. This distribution was learned for each individual. The future location for the individual was thereafter guessed by sampling from this distribution. This very rudimentary model does not incorporate any additional knowledge we have about the system. Hence, it was used to establish a baseline for measuring the performance of other more complex prediction models.

2) Naïve Bayes

A Naïve Bayes classifier assumes that the value of a particular parameter is independent of the value of any other parameters, even if they could be dependent on each other or upon the existence of other parameters. The primary advantage of this classifier is that it simplifies the classification problem by making strong assumptions and requires a small amount of training data to estimate the classification parameters.

The current location parameter models the probability of transitioning from the current area ID to a particular location. Current location parameter ensures that the predicted future area is adjacent to the current area.

The day of the week parameter models the probability of transition to a particular location on the current day of the week. For eg., this parameter would incorporate the fact that people are less likely to go to their offices on weekends.

The entire day was broken into 24 slots of one hour each. The time of the day parameter models the probability of transition to a particular location at the current time slot. For eg., this parameter would incorporate the fact that people are more likely to stay in their homes at night.

C. Transition Prediction

As mentioned previously, we predict when the user makes the next transition and where the individual would be after the transition.

1) Time of Transition Prediction

The duration between successive transitions can vary from less than a minute to more than a day. Typically, the former case is observed when the individual is traveling through a particular area and the latter case is seen when the individual stays at a particular location for a long time. We want to model the order of magnitude of time spent before the next transition, rather than the exact time spent. Hence, we considered different discretizations of the duration between successive transitions and compared accuracy of the models obtained from each method. These boundaries were empirically chosen. One such discretization is given below:

1. Between 0 and 5 minutes
2. Between 5 and 15 minutes
3. Between 15 and 30 minutes
4. Between 30 and 60 minutes
5. Between 1 and 6 hours
6. More than 6 hours

The training data was used to train a model based on Naive Bayes to predict the next transition time and this model was applied on the test set to find the training error over all individuals. The parameters used for training are their previous time of transition and their current location. The model had an accuracy of 91.55%. This may be an indicator that the time slots chosen are inappropriate as it may to be fairly easy to classify transition times in terms of these time slots. So, we changed the discretization boundaries and compared the results. The new discretization considered is given below. This is considered to be more meaningful for the applications we envision.

1. Between 0 and 1 minute
2. Between 1 and 5 minutes
3. Between 5 and 10 minutes
4. Between 10 and 15 minutes
5. Between 15 and 20 minutes
6. More than 20 minutes

The accuracy with this model was 78.70%, much lower than that for the previous case. The fact that the accuracy goes down can be countered by training the model on more data.

It is clear from this experiment that the accuracy of the model depends on the discretization chosen and hence, the degree of accuracy desired by the application can be a crucial parameter in deciding the discretization boundaries. A better method of choosing the distribution of duration between successive transitions can be a part of the future work for this project. The experiment was repeated by adding the previous location too as a parameter. In this case, the accuracy improved by only 0.04% which is not significant. This also confirms our initial assumption that the Markov property applies to the parameters used for modeling the predictor.

2) Location Prediction

Given the fact that a cell-to-cell transition occurred recently and given the information about the current location, current time and the day, we predict the future location of the user after the next transition. As mentioned before, the random guess was used as the baseline for the performance of our models. The Naïve Bayes model was implemented with different sets of parameters and different independence assumptions. These are given in Table 1 along with the accuracy.

III. Social Relationship Prediction

Broadly, the ways of communication between two individuals includes talking face-to-face, talking and texting over the phone or messaging through various online media. The strength of a relationship between them has a significant impact on these interactions. The stronger a relationship, the more frequent and longer are these interactions. We focus on learning a model that captures the differences in these communications with a close friend versus that with others. Thereafter, this model can be used to estimate the strength of relationships, primarily if two individuals are close friends or not.

In the dataset, information about proximity between individuals was captured through Bluetooth. The devices visible to a phone were logged periodically (every 5 minutes). The interactions over the phone in the form of voice calls and text are also logged. However, the interactions over online media are not available. While the available data does not sufficiently span the different kinds of relationships and their manifestation in the communication, we believe that the patterns learned are sufficiently representative of the group of individuals for which the data was gathered.

Triangle-closure is a dominant mechanism for edge-formation in social networks [4]. This has been shown to be true for formation of connections in sites like LinkedIn, Flickr etc. However, these connections might not always correspond to a close relationship. To ensure a simple model, we assume that the labels of the relationships are independent of others. We implemented two algorithms, Naïve Bayes and SVM for labeling a particular relationship. We considered the 7 features given below in our models:

1. Mean Call Duration
2. Standard Deviation of Call Duration
3. Number of Calls
4. Mean Duration of Proximity
5. Standard Deviation of Proximity
6. Number of times proximate to each other
7. Listed as a contact in the phone book.

These parameters were extracted from the Bluetooth logs of each individual. We also observed that the data contained some asymmetry. In some cases the durations for which two people were together as recorded by their phones were significantly different. We believe this can be attributed to temporal staggering of the logs and occasional low reception power at one of the phones resulting in missing the transmitter's presence.

A. Logistic Classifier

With only two possible labels for a relationship, we model the conditional distribution of the label given the values of parameters as a Binomial distribution. Hence, the hypothesis function $h_\theta(x)$ is the logistic function $1/(1+e^{-\theta^T x})$. The parameter θ is of size $n \times 1$ where n is the number of features chosen. It was learned using batch gradient ascent rule with learning rate $\alpha = 0.01$.

B. SVM

We used the SVM Toolbox completely implemented in Matlab [5]. We evaluated the performance of SVM with 2 different kernel functions: Gaussian Radial basis function (RBF) and the linear kernel. Since we have a small number of features, we expect the SVM with Gaussian kernel $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, which map the data to a higher dimensional space to be able to label the relationships more accurately. There are two parameters which impact the performance of an SVM with a Gaussian RBF, namely, C and γ. C is L1 regularization parameter. As mentioned in [6], we performed a grid search over the values of C and γ to determine the values resulting in the best performance. The performance was measured as the fraction of the true positives identified by the SVM. This is based on the fact that any application utilizing such a system would be expected to identify most of the close friends even if there are few false positives.

IV. Results

A. Future Transition Prediction

We have considered the accuracy to be the fraction of predictions which are correct. The accuracy averaged over all 94 individuals is used as the metric for evaluating the performance of the models. The models are learned using 70% of the log data for each individual and then tested over the remaining 30%. The predictions of time until transition and the location after transition are done independently of each other. We are not imposing a time limit on the validity of the prediction of the next location. However, doing this could potentially improve the performance of the models even further.

Future Location Prediction Model	**Accuracy**
Random Prediction	30%
NB Prediction given Current Location	77%
NB Prediction given Current Location & Day of Week	68%

NB Prediction given Current Location & Time of Day	72%
NB Prediction given Current Location, Day of Week & Current Time	66%
NB Prediction given (Current Location, Day of Week)	67%
NB Prediction given (Current Location, Day of Week) & Time of Day	65%

Table 1. Prediction of Future Location after Transition

The accuracy we obtained for models with different independence assumptions are given in Table 1. We obtained a 47% improvement in performance on an average when we had knowledge about the individual's current location. However, relative to this model, we observed a slight deterioration in performance when we included our knowledge about additional temporal parameters such as the current time of the day and the day of the week. The sparsity of the data also contributes to the increase in error to some extent. This decrease in accuracy is also because of the fact that we train over the first 70% of the data set (~5 months) and then test over the remaining 30% (~3 months). However, the weekly routines of individuals might change over a span of 3 months and hence predictions based on the day of the week might be erroneous. To capture these changes in weekly routines, the training and testing data should be done differently. For example, the training could be done over 3 weeks and tested in the fourth week.

Some individuals have a fairly consistent weekly routine whereas some do not. The ones with steady routines stayed in the same location for a longer duration. For these individuals, we saw that the Naive Bayes model performed much better.

Future Transition Time Prediction Model	**Accuracy**
NB Prediction given Current Location & Prev. Transition Time	78.70%
NB Prediction given Current & Previous Location	78.74%

Table 2. Prediction of Future Transition Time

B. Relationship Prediction

The data had overall 8556 labeled relationships. Out of these, only 108 were labeled as close friends (considered as a positive). We observed that this labeling was not symmetric. If one individual called another as a close friend, the reverse was not always true. To ensure that the our model does not get too biased by either label, our training set had 54 data points of close friends and 54 data points of not close friends. The models were tested over 54 data points of positive labels and 500 data points with negative labels. The accuracy obtained with both models is given in Table 3.

Relationship Prediction Model	**False Pos.**	**False Neg.**	**True Pos.**	**True Neg.**	**Accuracy**
Logistic Classifier	394	8	46	107	28%
SVM (Gaussian Kernel)	128	20	34	373	73%

Table 3. Prediction of Future Transition Time

Logistic Classifier had many more false positives than SVM while SVM correctly labeled most true negatives. However, Logistic classifier identified the positive scenarios much more accurately than did SVM. Based on the weights assigned to the different parameters, it was seen that the higher the average duration of call and of proximity and the lower their standard deviations, the higher was the likelihood of the relationship being a close friendship. SVM with a linear kernel was not able to separate the points, which implies that the data is not linearly separable. Hence, we did not perform any further evaluation with a linear kernel.

On increasing the number of data points with negative labels, both methods accurately labeled a smaller number of positives and a larger number of negatives. The training set was chosen to ensure that the positive labels were also classified accurately. The performance of the algorithms, in terms of the ability to accurately identify the data points with positive labels, can be improved further with more data points having positive labels.

V. Conclusion

In conclusion, we note that the daily routines of different individuals can be learned using simple supervised learning algorithms and these models can then be used to predict their locations across time. These predictions can be further improved by incorporating additional factors such as the current mode of transportation of the individual, other individuals in their proximity, information about their schedule, etc. It is also possible to predict the location at a finer level, i.e. individual cells, by first predicting the area and then predicting the most likely cell within the area. An alternative technique is to incorporate metadata about each location into the model and learn their transitions. This metadata should capture the user's activity in that location. For example, it could include the fact that the location is a theater or an office or a train station.

We also showed that it is possible to use information about the communication and proximity between two individuals, which can be acquired using a phone, to predict the strength of the relationship between them. In this case, it is seen that the models are quite accurate after training with a training set having a small number of labeled relationships. The SVM performs well in terms of identifying the negative cases whereas the Logistic Classifier performs better in identifying strong friendships. These predictions can be enhanced further by exploiting the fact that many friendships are formed by the triangle-closure mechanism.

VI. References

[1] N. Eagle, A. Pentland, and D. Lazer (2009), "Inferring Social Network Structure using Mobile Phone Data"**,** *Proceedings of the National Academy of Sciences (PNAS)*,106(36), pp. 15274-15278

[2] N. Eagle and A. Pentland (2009), "Eigenbehaviors: Identifying Structure in Routine"**,** *Behavioral Ecology and Sociobiology* 63:7, 1057-1066.

[3] K. Laasonen, M. Raento, H. Toivonen, "Adaptive On-Device Location Recognition", *Proceedings for Pervasive*, Springer Verlag, pp 287-304, 2004.

[4] J. Leskovec, L. Backstrom, R. Kumar and A. Tomkins, “Microscopic Evolution of Social Networks”, *KDD’08*, Las Vegas, USA.
[5] S. Canu and Y. Grandvalet and V. Guigue and A. Rakotomamonjy, “SVM and Kernel Methods Matlab Toolbox”, *Perception Systèmes et Information, INSA de Rouen, Rouen, France*, 2005.
[6] C. Hsu, C. Chang, and C. Lin. “A Practical Guide to Support Vector Classification” *Technical Report Department of Computer Science and Information Engineering, National Taiwan University*, 2003.