Predicting Mobile Call Behavior via Subspace Methods

Peng Dai, Wanqing Yang, and Shen-Shyang Ho

School of Computer Engineering,
Nanyang Technological University, Singapore
{daipeng,wyang006,ssho}@ntu.edu.sg

Abstract. We investigate behavioral prediction approaches based on subspace methods such as principal component analysis (PCA) and independent component analysis (ICA). Moreover, we propose a personalized sequential prediction approach to predict next day behavior based on features extracted from past behavioral data using subspace methods. The proposed approach is applied to the individual call (voice calls and short messages) behavior prediction task. Experimental results on the Nokia mobility data challenge (MDC) dataset are used to show the feasibility of our proposed prediction approach. Furthermore, we investigate whether prediction accuracy can be improved (i) when specific call type (voice call or short message), instead of the general call behavior prediction, is considered in the prediction task, and (ii) when workday and weekend scenarios are considered separately.

Keywords: Sequential Prediction, Eigenbehavior, Principal Component Analysis, Independent Component Analysis, Behavior Prediction

1 Introduction

To make accurate prediction on individual activities and behavioral patterns are new research directions in data mining, machine learning, and pervasive computing research communities. Based on data collected from mobile devices such as smart phones, one can predict and understand an individual's behavior and provide useful services or information to the individual. The industry takes a serious interest in these research topics with their game-changing potential in the highly competitive mobile device market [6]. Eagle and Pentland [2] introduced the eigenbehavior to represent repeating structures in an individual's behavior using principal components similar to those for eigenface [7]. They further claimed that "dimensionality reduction techniques [...] will play an increasingly important role in behavioral research".

In this paper, the two main contributions are (i) our investigation on whether independent components can be as useful as principal components in their representation of individual behavior and (ii) a sequential prediction approach to predict daily personal behavior modeled at hourly intervals. Our proposed approach assumes that the behavior of interest represented by primary (either

principal or independent) components remain (almost) unchanged in the near future (e.g., the next few days). We demonstrate the feasibility of our proposed sequential prediction approach on the individual mobile call behavior prediction task. For this task, the objective is to predict whether an individual will call (voice call or/and short message) within some hour interval on the next day. Moreover, we investigate (i) whether predicting specific call type (voice call or short message) is a better problem setting than the general call behavior prediction setting; and (ii) whether splitting the training data to workday and weekend data can improve the prediction performance.

2 Dataset and Data Preprocessing

In Section 2.1, we briefly describe the Nokia Mobility Data Challenge (MDC) dataset that is used in this paper. In Section 2.2, we describe how we process the MDC data for the mobile call prediction task.

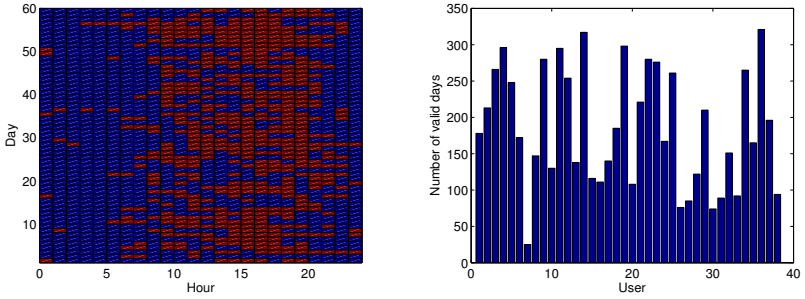
2.1 Nokia MDC Dataset

The MDC dataset consists of smartphone data collected in the Lake Geneva region from October 2009 to March 2011. Data types related to location (GPS, WLAN), motion (accelerometer), proximity (Bluetooth), communication (phone call and SMS logs), multimedia (camera, media player), and application usage (user-downloaded applications in addition to system ones) and audio environment (optional) were collected [6]. A total of 185 participants were involved. 38% of the participants are females and the rest are males. About two thirds of the participants are of age ranging from 22 to 33. Individual data was collected using the Nokia N95 smartphone and a client-server architecture. The open challenge data subset from the MDC dataset consisting of data from 38 participants for 8154 days are used in this paper. We focus on the voice calls, short messages, and the time they occurred.

2.2 Data Preprocessing

The call log data, consisting of call time, call duration, call type (short message or voice call), and etc. Call time and call type are used in our investigation. We, first, categorize about 2 years of daily call information for all participants into valid and invalid days. A valid day is a day where there are some phone activities (either voice calls or short messages). Otherwise, when there is no phone activity, it is a invalid day. Invalid days are ignored in the construction of the call behavior matrix for a participant so that there can be no row of zeros (i.e. no phone activity). Hence, we may not have consecutive days of call behavior vectors in the matrix. This call behavior matrix construction assumes that a person must have daily phone activity. Towards this end, the trivial prediction of no phone activity is not possible for our approach.

The call behavior of a participant is characterized by a $D_i \times 24$ matrix M_i, where i is the unique index for a participant and D_i is the total number of valid day used to construct our matrix for participant i. The call behavior matrix consist of binary values, one and zero, representing the existence or the non-existence of phone activity, respectively. Figure 1(a) shows the first 60 consecutive valid days of the phone activities for participant 2. Figure 1(b) shows the total number of valid days for the thirty-eight participants. Data from participant 7 include only 25 valid days and hence his data are not used in our experiments.



(a) Call behavior matrix (b) Valid days for all participants

Fig. 1. The MDC data

3 Behavioral Representations

In Section 3.1, we introduce eigenbehavior and its implementation using principle component analysis (PCA). In Section 3.2, we introduce independent component analysis (ICA) as an alternative behavioral representation.

3.1 Eigenbehavior and Principal Component Analysis

Eigen representations have become one of the most popular techniques in pattern recognition (e.g. face recognition [7]) because of its strong discriminative ability. Eagle and Pentland [2] proposed using the so-called eigenbehavior to measure the distance between people, which is then used for the construction of a social network. They also apply the eigenbehavior for individual location prediction [1]. Eigenbehavior is based on the application of principal component analysis [5, 2] on task-dependent daily individual behavioral representations.

Given an individual's daily m-dimensional behavior vectors, Γ_1, Γ_2, ..., Γ_i, ..., Γ_D, for a total of D days. Based on the convention used in [1], the average behavior of the individual is

$$\Psi = \frac{1}{D} \sum_{i=1}^{D} \Gamma_i. \tag{1}$$

The behavior deviation for a particular day from the mean behavior is

$$\Phi_i = \Gamma_i - \Psi. \tag{2}$$

Principal components analysis (PCA) is then performed on these vectors generating a set of m orthonormal vectors that can be linearly combined that best describe the distribution of the set of behavior vectors. The vectors and their corresponding scalars computed from PCA are the eigenvectors and eigenvalues of the covariance matrix

$$C = \frac{1}{D}\sum_{i=1}^{D} \Phi_i \Phi_i^T \tag{3}$$

3.2 Representing behavior using Independent Components

The goal of PCA is to find a set of orthogonal components that minimize the error in the reconstructed data. In fact, PCA seeks a transformation of the original data into a new frame of reference with as little error as possible, using fewer factors (i.e., principal components) than the original data. In particular, PCA is a popular approach to perform dimensionality reduction [7].

Here, we investigate whether independent components derived from independent component analysis (ICA) can be used to obtain behavior representation as useful as eigenbehavior for prediction tasks. In contrast to PCA, ICA seeks, not a set of orthogonal components, but a set of independent components. Two components are independent if any knowledge about one implies nothing about the other.

Again, given an individual's daily m-dimensional behavior vectors, Γ_1, Γ_2, $\ldots$, Γ_i, $\ldots$, Γ_D, for a total of D days. Each behavior vector

$$\Gamma_i = \sum_{i=1}^{n} w_i s_i \tag{4}$$

is assumed to be generated by the set of independent components $s_i, i = 1, \ldots, n$ and $w_i, i = 1, \ldots, n$, are the corresponding weights.

Our ICA representation is constructed using the InfoMax algorithm [3]. It is based on maximizing the output entropy (or information flow) of a neural network with non-linear outputs. Assume that $\mathbf{x}$ is the input to the neural network whose outputs are of the form $\phi_i\left(\mathbf{w}_i^T\mathbf{x}\right)$, where the ϕ_i are some non-linear scalar functions, and the $\mathbf{w_i}$ are the weight vectors of the neurons [3]. Then ICA model can be obtained by maximizing the entropy

$$H\left[\phi_1\left(\mathbf{w}_1^T\mathbf{x}\right), \cdots, \phi_n\left(\mathbf{w}_n^T\mathbf{x}\right)\right] \tag{5}$$

of the outputs [4]. The MATLAB implementation of InfoMax algorithm is publicly available in DTU toolbox [8].

4 Behavior Prediction Approaches

In Section 4.1, we introduce the approach proposed by Eagle and Pentland [2] that predicts the later part of the day based on information on the earlier part of the day. In Section 4.2, we describe our proposed sequential prediction approach for the next day(s) based only on data from previous days.

4.1 Single-Day Method

Both PCA (or eigenbehavior) and ICA share the same idea that the daily behavior vector obtained in Section 2.2 can be treated as a combination of several primary daily behavior components generated by either approach. An individual's primary daily behavior components represent a space upon which all of his daily behavior can be projected with different levels of accuracy. Using the primary behavior components, it is possible to predict the future behavior for an individual.

One straightforward way to predict the future behavior for an individual at the later part of a particular day is to reconstruct an entire daily behavior vector using only behavior information from an earlier part for that day [2]. Let

$$\mathbf{A} = [\mathbf{\Phi_1}, \mathbf{\Phi_2}, \ldots, \mathbf{\Phi_i}, \ldots, \mathbf{\Phi_M}] \tag{6}$$

denotes the primary behavior matrix calculated from N days of behavior data such that each column contains one principal/independent component $\mathbf{\Phi_i}$ that is a $24-$dimensional vector corresponding to the ith primary behavior component. Assuming the first p hours behavior for that day, $\mathbf{\Gamma_{1:p}}$, are known. Hence,

$$\mathbf{A_s v} = \mathbf{\Gamma_{1:p}} \tag{7}$$

where $\mathbf{A_s}$ is a $p \times M$ matrix corresponding to the first p row of $\mathbf{A}$. Then one obtains a M-dimensional reconstruction vector

$$\mathbf{v} = \mathbf{A_s}^{-1}\mathbf{\Gamma_{1:p}} \tag{8}$$

where $\mathbf{A_s}^{-1}$ is the pseudo inverse matrix of $\mathbf{A_s}$. To predict the rest of the day, i.e., $p+1$ to 24 hours in $\mathbf{\Gamma}$, one reconstructs the entire behavior vector using

$$\mathbf{\Gamma} = \mathbf{A v}. \tag{9}$$

The above predictive model assumes dependency of behavior within the same day, and the relationship stays relatively stable. We refer to this prediction approach as the **single-day method**.

4.2 Multiple-Day Method

An alternative prediction approach is to model the daily behavior as a whole day event and then predict the next day(s). Assuming a sequence of D days

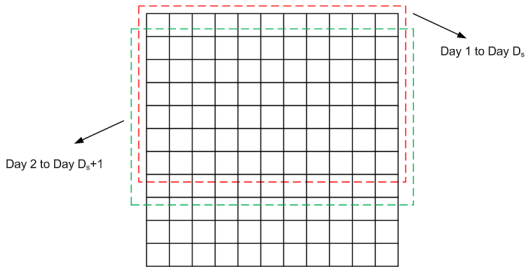


Fig. 2. Multiple-day method for generic future behavior prediction

of 24-dimensional behavior vectors in our prediction scenario, one predicts the behavior for day D_s+1 given behavior information from previous D_s days. Based on the assumption that there cannot be too much changes in a person's behavior within a short time interval, our proposed approach models daily behavior within a fixed temporal window of D_s days (see Figure 2) and predict the next day's (day $D_s + 1$) behavior based on this daily behavior model. We first obtain the primary behavior matrix representing the behavior from day 1 to D_s, denoted as $\mathbf{\Gamma}_1^{D_s}$ corresponding to the red bounding box in Figure 2. According to our assumption, $\mathbf{\Gamma}_2^{D_s+1}$ (denoted as the green bounding box in Figure 2) share the same primary behavior matrix as $\mathbf{\Gamma}_1^{D_s}$. Note that row $D_s + 1$ represents the unknown next day behavior that we want to predict.

Using the D_s days of daily behavior vectors, one obtains the daily behavior model as a set of D_s-dimensional primary components, $\mathbf{\Phi_i'}, i = 1, \ldots, 24$. The first M primary components are chosen to construct the primary behavior matrix

$$\mathbf{C} = [\mathbf{\Phi_1'}, \mathbf{\Phi_2'}, \ldots, \mathbf{\Phi_M'}] \tag{10}$$

where each column of C correspond to a primary vector. Then one obtains the $M \times 24$ reconstruction matrix

$$\mathbf{V} = \mathbf{C_s}^{-1}\mathbf{\Gamma_2^{D_s}} \tag{11}$$

where $\mathbf{C_s}^{-1}$ is the pseudo inverse matrix of the $(D_s - 1) \times M$ matrix $\mathbf{C_s}$ corresponding to the first $D_s - 1$ rows of $\mathbf{C}$, since

$$\mathbf{\Gamma_2^{D_s}} = \mathbf{C_s V} \tag{12}$$

Then, the prediction of day $D_s + 1$ can be obtained as the last row of

$$\mathbf{\Gamma} = \mathbf{CV}. \tag{13}$$

This method makes use of the relationship embedded in the historical data from the previous D_s days. We refer to this prediction approach as the **multiple-day method**. Note that the multiple-day method can be used to predict not only

the behavior for the next day (i.e., $D_s + 1$) but also the next $n(<< D_s)$ day's behavior. The modified prediction scheme for day $D_s + n$ is

$$\mathbf{\Gamma} = \mathbf{C}\left(\mathbf{C'_s}^{-1}\mathbf{\Gamma}^{\mathbf{D_s}}_{\mathbf{n+1}}\right) \tag{14}$$

where $\mathbf{C'_s}$ is a $(D_s - n) \times M$ matrix corresponding to the first $D_s - n$ rows of $\mathbf{C}$.

5 Experimental Results

First, we study the prediction performance of two subspace approaches, PCA and ICA, utilizing different number of daily behavior vectors via the single-day and multiple-day methods. Then, we investigate whether the prediction performance can be improved by (i) considering the call types: short messages and voice calls; and (ii) then further splitting the data into workday and weekend observations. For illustration purposes, we apply only the PCA-based single-day and multiple day methods for this investigation. For all our empirical results, we use 4 primary components for either PCA or ICA. Since the objective of the prediction task is to predict whether an individual will or will not call (i.e., 1 or 0) within some hour interval the next day, a threshold is required to decide on the final prediction. Here, the threshold is set to 0.5.

From Figure 3, we observe that both PCA-based and ICA-based multiple-day methods perform better than the single-day methods. Moreover, their prediction performance are comparable. PCA-based single-day method performs slightly better than ICA-based single-day method. One thing to note is that the single day approach has to use the first p (here, $p = 12$, i.e., using first half of the day to predict the second half) hourly observations to calculate the reconstruction vector, $\mathbf{v}$ in (8). Thus, it is impossible to predict the whole day. On the other hand, the multiple-day method uses the previous days' observations to calculate the reconstruction matrix $\mathbf{V}$ in (11). Therefore, it can predict an individual's behavior for the entire day.

From Figure 3, we see that the number of daily behavior vectors used for optimal prediction performance is seventy for the PCA-based single-day method and the multiple-day methods. However, PCA-based single-day method has comparable prediction performance when the number of days used is between 10 and 80. The performance of ICA-based single-day methods degrades as the number of days used increases.

Using 20 days and 70 days of data to build the behavior matrices, we investigate the distribution of participants at various level of average prediction accuracy shown in Figure 4. Considering using 70 days of data, we observe that the multiple-day method is very competitive due to its use of information from previous 70 days when prediction is made. In particular, 21 out of the 38 participants achieve prediction accuracy of more than 80% for each multiple-day method. Furthermore, one observes that prediction performance for single methods can go as low as 55% for a user while multiple-day methods achieves a minimum of 65% accuracy for the participants. Considering using only 20 days of

data, PCA-based multiple-day method performs the best with 19 out of 38 participants achieve 80% or more prediction accuracy. While PCA-based single day performs relative well with 17 participants achieving accuracy of 80% or more, we observe from the Figure 4 that prediction performance for 6 participants are 65% or below. Compared to the other approaches, the number of participants with poor prediction performance is significantly higher. One notes that when a small number of days of data are used, ICA-based methods have average prediction performance with respect to the number of participants. Again, readers are reminded that multiple-day and single-day methods can be considered to be solutions for two different prediction tasks or problem settings.

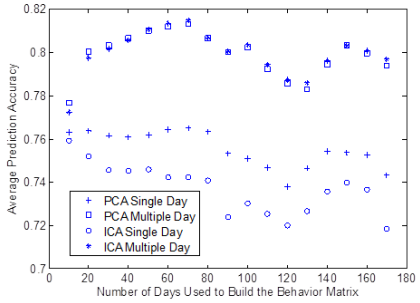


Fig. 3. Effect of different number of daily behavior vectors on prediction performance.

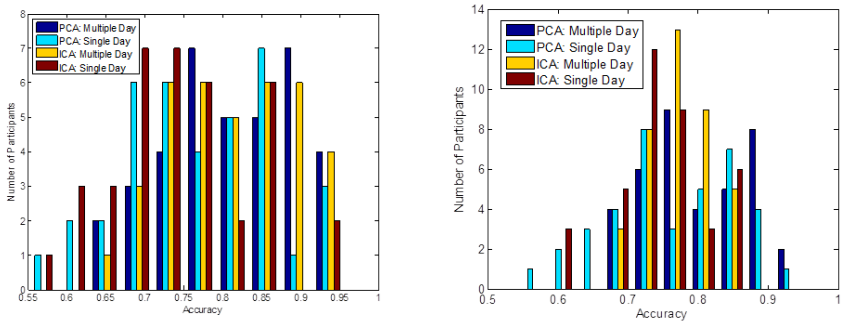


Fig. 4. The number of participants achieving various average prediction accuracy for the different approaches when the number of days to build the behavior matrix are 70 (left) and 20 (right), respectively.

From Figure 5, we observe that prediction performances are improved for both methods when call types: short messages and voice calls, are considered.

Hence, specific (call) behaviors are more predictable. One notes that the number of days of data used has a significant effect on the PCA-based single-day method for short messaging prediction.

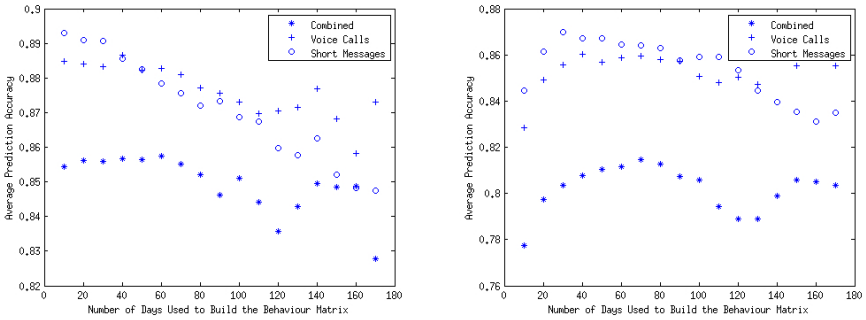


Fig. 5. Prediction accuracy for PCA-based Single-Day Method (left) and PCA-based Multiple-Day Method (right) when call behavior is further categorized into sending short messages and making voice calls.

From Figure 6, we observe that in general the two methods have better prediction performance for workday call behavior than for weekend call behavior. It is particularly significant that short messaging behavior is less predictable during weekends. The significant drop and haphazardness in the prediction performance during weekends as the number of days of observation used increases is most probably due to shortage of data for testing purposes. Hence, the empirical results using more than 60 days of data should be ignored. Towards this end, we conclude that PCA-based multiple-day method predicts well on voice call behavior for both weekend and workday, and occasionally even slightly better than using weekend and workday data together (see bottom right graph in Figure 6).

6 Conclusions

In this paper, we investigate whether independent components can be as useful as principal components in their representation of individual behavior. Moreover, we propose a sequential prediction approach to predict daily personal behavior modeled at hourly intervals. We demonstrate the feasibility of our proposed sequential prediction approach on the individual mobile call behavior prediction task using the Nokia MDC dataset. We observe that formulating a specific (voice call or short message) behavior prediction problem is better than a general (call) behavior prediction problem as one can obtain better prediction accuracy in the former task. Also in general, we observe that workday behavior is more predictable than weekend behavior.

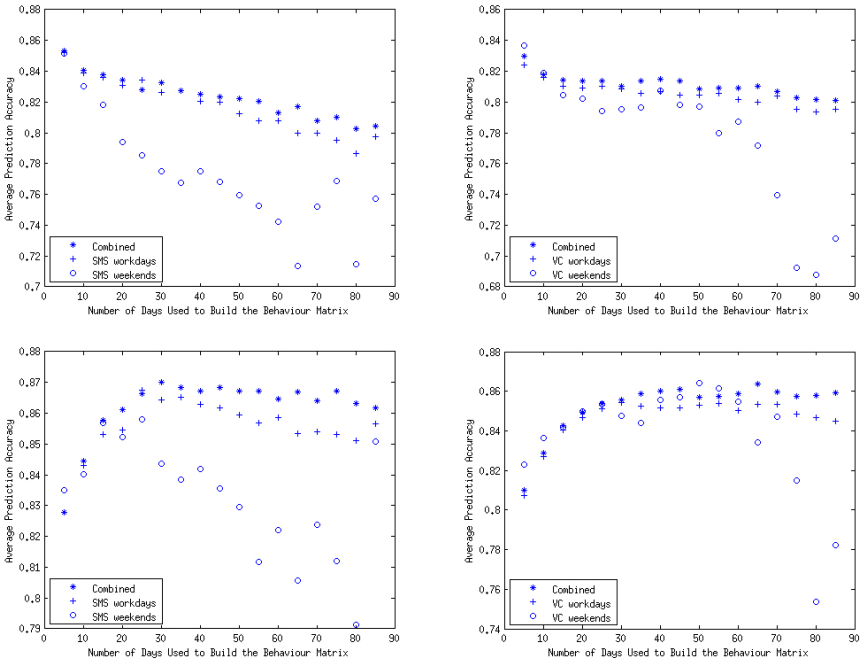


Fig. 6. Prediction accuracy for PCA-based Single-Day Method (top) and PCA-based Multiple-Day Method (bottom) when short message data (left) and voice call data (right) are split based on whether they occurred on weekends or workdays.

References

1. N. Eagle, A. Pentland, and D. Lazer, *Inferring Social Network Structure using Mobile Phone Data*, Proceedings of the National Academy of Sciences, 106(36), pp. 15274-15278, 2009.
2. N. Eagle, and A. S. Pentland, *Eigenbehaviors: Identifying structure in routine*, Behavioral Ecology and Sociobiology, vol. 63, pp. 1057-1066, 2009.
3. A. Hyvarinen, and E. Oja. *Independent component analysis: algorithms and applications*, Neural Netw, 13(4-5), pp. 411-430, 2000.
4. A. Bell and T.J. Sejnowski *An Information-Maximization Approach to Blind Separation and Blind Deconvolution*, Neural Computation, 7, pp. 1129-1159, 1995.
5. I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag New York, Inc., 1997.
6. J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. *The mobile data challenge: Big data for mobile computing research*, Proc. on Mobile Data Challenge by Nokia Workshop in conjunction with Int. Conf. on Pervasive Computing, Newcastle, June 2012.
7. M. Turk, and A. S. Pentland, *Eigenfaces for Recognition*, Journal of Cognitive Neuroscience, 3(1), pp. 71-86, 1991.
8. ICA:DTU Toolbox, *http://http://cogsys.imm.dtu.dk/toolbox/ica/*