Demographic Prediction of Mobile User from Phone Usage

Shahram Mohrehkesh Shuiwang Ji Tamer Nadeem Michele C. Weigle
Department of Computer Science
Old Dominion University
{smohrehk, sji, nadeem, mweigle}@odu.edu

ABSTRACT
In this paper, we describe how we use the mobile phone usage of users to predict their demographic attributes. Using call log, visited GSM cells information, visited Bluetooth devices, visited Wireless LAN devices, accelerometer data, and so on, we predict the gender, age, marital status, job and number of people in household of users. The accuracy of developed classifiers for these classification problems ranges from 45-87% depending upon the particular classification problem.

Categories and Subject Descriptors
C.1.3 [Other Architecture Styles]- Cellular architecture (e.g., mobile). **I.2.6 [Learning]**- Parameter

General Terms
Algorithms, Design, Experimentation.

Keywords
Profile, prediction, classification, regression, svm, ensemble, age, gender, job.

1. INTRODUCTION
Mobile devices have created a new paradigm of communication between people. Now, the concept of ubiquitous and pervasive computing is not a dream anymore. People use their cell phones at different times and locations to communicate with other people and run various applications. This wireless usage trend has attracted the attention of researchers in all fields of science. For example, the reality mining project investigates the social behavior of users based on their mobile data usage [1]. One of the other aspects that can be investigated from tempo-spatial mobile data sets is user profile, i.e. gender, job, etc.

Even though there are some limited studies such as [2] that investigate the mobile usage for different genders and ages of users, there is not much study in this domain. Particularly, when it comes to prediction of user profiles based on mobile usage, there is not any public literature because usually this data is only available to mobile operators. In this paper, prediction of mobile users has been conducted based on a dataset collected by Nokia [3]. The dataset was published for participants to compete for three predefined challenge questions and some open tasks. A more detailed description of dataset and challenge participation can be found in [3].

In the following section, we first describe the dataset briefly and then features that were extracted from dataset. In section 4, we describe the methods and parameters that we used for our prediction classification and regression model. We explain the accuracy result for our model in section 5. Finally we conclude the paper with some suggestions to improve the accuracy.

2. Problem and Dataset

2.1 Problem
Our goal was to predict the profile of users in each five classification problems with respect to their phone usage. Various aspects of phone usage were collected for each user. Examples of collected information include details of calls, short text messages, and used applications

Table 1 shows the class labels and their definition for users. The goal is to identify these classes based on features obtained from phone usage. The performance of classification prediction is measured in terms of accuracy. However, the challenge asks us to calculate the performance of the *age* and *number of people in household* labels in terms of Root Mean Square Error (RMSE). Therefore, it is a regression problem indeed. However, we will call all of the problems as classification without loss of generality.

Table 1. Class labels and descriptions

Class	Labels and Descriptions
Gender	1 : female 2 : male
Age	1 : 16-21 2 : 22-27 3 : 28-32 4 : 33-38 5 : 39-44 6 : 45+
Marital Status	1 : Single or Divorce 2 : In a relationship 3 : Married or living together with my partner
Job	1 : Training 2 : PhD student 3 : Employee without executive function 4 : Employee exercising executive functions
Number of People in the Household	1 : 1 2 : 2 3 : 3 4 : 4 5 : 4+

2.2 Overview of solution
In the following, we describe our solution steps. More details for each step will be explained in later sections. The first step is to convert the raw data from CSV files to MAT format which is the data format file in MATLAB. Then, we had a brain storm meeting to decide which features can be extracted from the converted dataset. We will explain the extracted features in next section. After feature extraction, we perform the feature selection to choose the most efficient features. The next step is to develop

classifier. In this step, we try to find the most appropriate type of classification method for our dataset, called *SetA*. The performance of classifiers was evaluated by dividing the data into training and test samples. In this stage various parameters are changed to find the best combination of parameter to achieve the highest accuracy. Finally after release of the *SetB* dataset, the developed classifiers are applied to them and predicted results are submitted for challenge. The *SetB* contains the same features in *SetA* for 34 users while the SetA contains data of 80 users. Clearly, converting of CSV data files to MAT files and feature extraction should be performed on *SetB* to be able to use the classifier that were developed in model development based on *SetA*. Because the ground truth for *SetB* is not available, we do not include our prediction result in the paper.

3. Extracted features

We extract the corresponding features from different aspects of the dataset (first column of Table 2). This features list provides 1100 features (columns) per user that are used in our classification model and prediction.

The dataset contains data from 80 mobile users for a duration of approximately 10 months. This number of users could be an appropriate number of samples for a two class problem. However, it is not completely practical for a four or five class problem, especially where there are only four samples for some of labels, e.g. user with age group of 45+. One method to increase the number of samples in this case is to split the data into sub-samples. For example, features can be extracted per month. So each sample will represent a user-month. After predicting the labels, aggregation of labels should be performed to determine the label for each user. A problem that arises in this case is how to aggregate the predicted labels. The most common method is to use voting which works based on the majority of labels. However, we will show later that results using this method cannot provide high accuracy.

Table 2. Extracted features from each dataset

Dataset Aspect	Features
Accelerometer	Mean of *avdelt*[1] values Variance of *avdelt* values Number of accelerometer records Mean of *avdelt* values in each hour of day Mean of *avdelt* values on weekdays and weekends
Application	Number of applications run in each hour of day Duration of usage for top used application Number of applications run on weekdays and weekends
Bluetooth	Number of Bluetooth devices visited in each hour of day Number/Duration of times a device is visited for top 20 most visited device Duration of times a device is visited for top 20 most visited device Number of Bluetooth visited on weekdays and weekdays
Calendar	Number of calendar entries of each entry type (appointment or event) Number of calendar entries of each class type (public or private) Number of calendar entries in each hour of day Number of calendar entries on weekdays and weekdays
Call Log[2]	Mean and variance of call duration Number of all short messages Number of all calls Number of sent short messages Number of received short messages Number of incoming /outgoing calls Mean of call duration for incoming /outgoing calls Number of incoming /outgoing calls from/to most top 10 contacts Number of sent/ received short messages call to/from most top 10 contacts Number of sent/received messages in each hour of day Number/duration of incoming/outgoing calls in each hour of day Duration of incoming/outgoing calls in weekdays and weekends Number of sent/received short messages on weekdays and weekends
GSM	Number of visiting times of GSM cells for top 50 most visited cells Duration of visiting of GSM cells for top 50 most visited cells Number of visiting times of GSM cells for top 50 most visited cells on weekdays vs. weekends Duration of visiting times of GSM cells for top 50 most visited cells in each hour of day
Media	Number of media items Mean and variance of media size Number of media items created in each hour of day Number of media items created on weekdays and weekends
Media Play	Number of times media play plays songs Number of times media is played in each hour of day Number of times media is played on weekdays and weekends
System	Active duration of device profile (general, silent, etc.) Active duration of device ringing type (normal, beep, ascending, etc.) Mean and variance of duration of inactivity of phone
Wireless LAN	Number of WiFi devices visited in each hour of day Number/Duration of times a device is visited for top 20 most visited devices Number of WiFi devices visited on weekdays and weekdays

[1] *Avdelt* value is the average of all recorded accelerometer values for all axes sampled every 15 sec. More details in [3].

[2] Call log contains both call and short text message information.

4. Prediction model

4.1 Feature selection

Because there are many extracted features, and it is not clear which of the extracted features are relevant for our labels, we select only some features. We use the RELIEFF [4] method for feature selection. We evaluated the effect of different values of K, which specifies the K nearest neighbor for selection. RELIEFF weights the features from -1 to 1, with large positive weights assigned to important attributes. For most of our classification problems, around 300-500 features will result in the best performance of classification that selects the features with positive weight. One interesting observation in feature selection was that features related to the duration of visiting of a Bluetooth device, GSM Cell or Wireless LAN device had high weight in most of the classification problems. The *Call Log* related features also have high weights. Extracted features from *Media*, *Media Play*, and *Calendar* aspects have the lowest weighted features. The main reason is that their datasets have very few samples or even no samples for many users. In addition, it appears that many users use the default value settings for calendar, such as type or class of calendar entry. So, not much can be concluded related to users from their calendar entries.

4.2 Model

We apply multiple different classification models on the features that survive our feature selection module. We first employ the support vector machine (SVM) classifier [5], since it uses the hinge loss function, and thus provides robust classification. Linear kernels are used in our experiments, since the number of features is comparable to the number of samples. The LIBSVM [6] toolbox is used to solve the SVM optimization in our experiments. It is also worthwhile to mention that we used the Linear Discriminant Analysis (LDA); however, SVM outperforms LDA. So we do not include result of LDA.

To boost the performance of our classification module, we employ ensemble methods to build multiple classification models from a single set of data. This class of method combines some weak classifiers to achieve a better prediction via majority voting. Many different forms of ensemble learning models have been developed, and we employ the random forests [7], which combine multiple tree models with the bagging scheme. We use the ensemble toolbox from MATLAB. We use the *bag* method with *tree* learners. This method works for the multi-class problem as well as the regression problem.

5. Evaluation of model

To evaluate the performance of classifiers, the accuracy is measured for *SetA*. We use 70 percent of samples for training and rest for test. This divides the 80 users into 56 training samples and 24 test samples. It is worthwhile to mention that for some of users, the value of labels was not known. In these cases, the sample is removed. So, the final number of available samples can be smaller than 80. For the situation where the number of sample is increased through splitting data to user-month samples, we use the same portion, 70%, for dividing the samples between training and testing. Dataset contains 10 months of data per user. So, in total approximately 800 samples could be extracted for training and test. In the following, we describe the accuracy of our classifiers.

Table 3 shows the best classification accuracy that can be achieved for the SVM method with 56 training and 24 test samples.

Table 3. SVM classification accuracy

Class	Accuracy	Random
Gender	83%	50%
Age	41%	16.6%
Marital Status	45%	33.3%
Job	50%	25%
Number of People in the Household	65%	20%

As can be observed, the accuracies are better than random guess. The lowest accuracy occurs for the *Age* label. The main issue is that the classifier cannot classify the samples for label 1 (youngest age group 16-21) and label 6 (oldest group, age 45+). Table 4 shows the confusion matrix for the six class *age* problem. Each row shows for how many of labels in that class, we have correct classification. Also it indicates the other classes that the main class has misclassification with them. One reason of misclassification in *age* label is that there are only a few number of samples in the training set. It is likely that extracting more features can improve the *Age* classifier. Likewise, for the *number of people in the household* label, the classifier performs poorly in determining the people in the first and last classes (Table 5); in fact, people who live alone or live in a household of more than four are not recognized by this classifier in most scenarios. In the case of *Marital Status*, the classifier could not successfully recognize the correct user with label 2, *in a relationship*.

Table 4. Confusion matrix for the age classification problem

	1	2	3	4	5	6
1	0	1	0	0	0	0
2	0	4	2	2	0	0
3	0	2	7	1	0	0
4	0	0	1	2	0	0
5	0	0	0	1	0	0
6	0	0	1	0	0	0

Table 5. Confusion matrix for the number of people in household classification problem

	1	2	3	4	5
1	1	1	0	1	0
2	0	12	1	0	0
3	1	3	1	0	0
4	0	0	0	0	0
5	1	1	0	0	0

We try to improve this case using the ensemble methods. The best results that can be achieved with the ensemble method are shown in Table 6. As can be viewed, accuracy is improved slightly for all of classification problems in comparison to SVM. We also show the error in terms of RMSE.

Table 6. Bagging tree ensemble results

Class	accuracy	RMSE
Gender	87%	NA
Age	45%	0.94
Marital Status	47%	NA
Job	52%	NA
Number of People in the Household	66%	1.14

As mentioned before, to increase the number of samples, we split the user data to user-month. The accuracy results with this method are shown in Table 7. Accuracy is increased when calculated for split samples for the *Age* and *number of people in the household* classification problems. However, there is worse accuracy for other labels in comparison with the non-split method (Table 6). The other interesting observation is that for the *Gender* label and *number of people in household* label, we achieve better accuracy by aggregating the results of split sample through majority voting.

Table 7. Bagging tree ensemble results for split samples

Class	Accuracy for split data	Accuracy for aggregated based on voting
Gender	76%	78%
Age	50%	12%
Marital Status	47%	4%
Job	47%	25%
Number of People in the Household	57%	73%

6. Conclusion and future work

In this paper, we developed classification and regression models for the prediction of a mobile user profile based on their mobile usage. Our model shows that information about call log and surrounding wireless device or antenna are the most significant features in this type of classification problem. We can predict the gender, age, marital status, job and number of people in the household of the user with 87, 45, 47, 52 and 73 percent accuracy, respectively.

We argue that location information can improve accuracy. In the provided dataset, there is a trace of user movement. However, it is not helpful unless the logical location (such as home, office, etc.) can be specified. Even though it is possible to develop a classifier to predict the logical location, we did not develop such a model because we did not have the ground truth data to evaluate the classifier's accuracy.

Another improvement in the accuracy of some classifiers such as *number of people in household* can be achieved by extracting features related to finding the relationship among the users that are in vicinity of each other through WiFi or Bluetooth logs. However, the anonymization method used in this dataset does not allow extracting this sort of information; the same MAC address is anonymized differently for two different users.

7. REFERENCES

[1] Eagle, N., Pentland, A. and Lazer, D. Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences, 106, 36, September 8, 2009, 15274-15278.

[2] Falaki, H., Mahajan, R., Kandula, S., Lymberopoulos, D., Govindan, R. and Estrin, D. Diversity in smartphone usage. In Proceedings of the Proceedings of the 8th international conference on Mobile systems, applications, and services ,San Francisco, California, USA, 2010.

[3] Juha K. Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle and Miettinen, M. The Mobile Data Challenge: Big Data for Mobile Computing Research. Newcastle, UK, June 2012.

[4] Kononenko, I., Simec, E. and Sikonja, M. R.-. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. Applied Intelligence, 7, 1 1997, 39-55.

[5] Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2, 2 1998, 121-167.

[6] Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol., 2, 3 2011, 1-27.

[7] Breiman, L. Random Forests. Mach. Learn., 45, 1 2001, 5-32.