Pervasive and Mobile Computing 9 (2013) 784–797

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Where to go from here? Mobility prediction from instantaneous information

Vincent Etter *, Mohamed Kafsi, Ehsan Kazemi, Matthias Grossglauser, Patrick Thiran

School of Computer and Communication Sciences, EPFL, CH-1015 Lausanne, Switzerland

A R T I C L E I N F O

Article history:
Available online 19 July 2013

Keywords:
Machine learning
Mobility prediction
Probabilistic Graphical Models
Dynamical Bayesian Network
Artificial Neural Networks
Gradient Boosted Decision Trees

A B S T R A C T

We present the work that allowed us to win the Next-Place Prediction task of the Nokia Mobile Data Challenge. Using data collected from the smartphones of 80 users, we explore the characteristics of their mobility traces. We then develop three families of predictors, including tailored models and generic algorithms, to predict, based on instantaneous information only, the next place a user will visit. These predictors are enhanced with aging techniques that allow them to adapt quickly to the users' changes of habit. Finally, we devise various strategies to blend predictors together and take advantage of their diversity, leading to relative improvements of up to 4%.

1. Introduction

Mobility is a central aspect of our life; the locations we visit reflect our tastes and lifestyle and shape our social relationships. The ability to predict the places a user will visit is therefore beneficial to numerous applications, ranging from forecasting the dynamics of crowds to improving the relevance of location-based recommendations.

This paper reports our winning contribution to the Nokia Mobile Data Challenge (NMDC). Our task was [1] "to predict the next destination of a user given the current context, by building user-specific models that learn from their mobility history, and then applying these models to the current context to predict where the users go next". A context is described by the data collected from the mobile phone of the user (date, location of the user, cell tower id, WLAN, phone calls, etc.).

First, we examined carefully the data and their characteristics and implemented techniques to overcome some of the roots of unpredictability. For instance, we noticed that some users change their home location during the observation period so we developed aging techniques that allow us to detect and adapt to these changes. Then, we developed several mobility predictors, based on graphical models, neural networks, and decision trees. These predictors exhibited close average prediction accuracies, yet we observed, for each user, a high performance variability between predictors.

In order to take advantage of this variability, we finally combined these predictors using different blending strategies. Blending is an ensemble method that combines different predictors in order to obtain a predictor which is more accurate than any of the individual ones. We submitted five sets of predictions to the challenge, based on these blending techniques. Each submission obtained a higher prediction accuracy than all the submissions from other participants, allowing our team to win the first place of the challenge.

The paper is organized as follows: we describe the NMDC and its dataset in Section 2. In Section 3, we introduce a framework where we define the notations, the learning process and the prediction performance measure. Then, we present,

* Corresponding author. Tel.: +41 216931260.
E-mail addresses: vincent.etter@epfl.ch (V. Etter), mohamed.kafsi@epfl.ch (M. Kafsi), ehsan.kazemi@epfl.ch (E. Kazemi), matthias.grossglauser@epfl.ch (M. Grossglauser), patrick.thiran@epfl.ch (P. Thiran).

1574-1192/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.pmcj.2013.07.006

in Section 4, various models serving as basis for our predictors, and we show their individual performance. In Section 5, we describe different blending strategies and establish their performance gain over individual predictors. Finally in Section 6, we briefly discuss other works related to predicting the mobility of users, before concluding in Section 7.

2. Nokia mobile data challenge: next-place prediction

The NMDC is [1] "a large-scale research initiative aimed at generating innovations around smartphone-based research, as well as community-based evaluation of related mobile data analysis methodologies". It was organized by Nokia and took place from January 2012 to April 2012. It featured an open track, in which participants were able to propose their own problem to study, and three dedicated tracks, each defining a specific problem for teams to solve: semantic place prediction, next-place prediction and demographic attributes' prediction.

At the heart of this challenge is the dataset gathered during the Lausanne Data Collection Campaign (LDCC) [2]. This dataset consists of a rich set of features (locations, phone calls, text messages, application usage, etc.) recorded from the smartphones of 170 participants, over periods of time ranging from a few weeks to almost two years. These data were collected in a privacy-preserving manner, allowing for meaningful statistics to be gathered while the anonymity of participants was protected.

Each task had its own subset of the LDCC data. The Next-Place Prediction task, the focus in this paper, was assigned a subset of 80 users. For each user, the last 50 days of data were kept as a test set for the evaluation of each team's submissions, and the rest was used as training data.

For privacy reasons, all identifiers (phone numbers, WLAN SSIDs, contact names, etc.) were encrypted, but more importantly, physical locations were not released. Instead, for each user, the organizers of the NMDC first identified *places* – corresponding to discs with a 100 m radius – by using both GPS and WLAN data. Then, they represented each place by a unique identifier. Consequently, a sequence of geographic coordinates is represented as a sequence of place identifiers.

These visits represent the basic unit for the prediction task. They are defined by their starting and ending times, and the corresponding place. In addition, several types of data are available: accelerometer, application usage, GSM, WLAN, media plays, etc. The complete list can be found in the dataset description [1]. Given a visit and the data characterizing it, our task was to predict the next place visited by the user.

At the end of the challenge, each participating team was allowed to submit five different sets of predictions, corresponding to visits from the undisclosed part of the LDCC data. Then, the organizers of NMDC evaluated the prediction accuracy of each participating team's submissions.

We present below two major constraints (imposed by the rules of the NMDC) that restrict the range of methods we can use and make our task more challenging.

User specificity. To prevent cross-referencing people and places between users, all sensitive data are user-specific: the identifiers are encrypted using different keys, and places are defined and numbered for each user independently. Moreover, the rules of the challenge explicitly forbade all participants to try and reverse this process, or make some links between users. We were therefore not allowed to build joint models over the user population, *i.e.*, to learn from one user to make a prediction about another. For this reason, we built user-specific predictors and consider each user independently.

Memoryless predictors. As explained above, the input for the Next-Place Prediction task is the *current* visit, along with all additional data recorded from the user's phone during that time. However, we do not have access to the *history* of the user, *i.e.*, the sequence of previous visits. If we did, we could develop higher-order predictors that not only take into account the current place but also the sequence of places visited just before. Indeed, such information is very useful: if a user is currently at a transportation hub, *e.g.* a bus station, knowing whether he was home or at work just before greatly helps in predicting his next move. Because this information is not available to us in this challenge, we limit ourselves to *memoryless* predictors, *i.e.*, methods that take into account only the current context, without any knowledge of the past.

3. Place prediction framework

In this section, we explore some characteristics of the dataset and define the framework within which we develop our predictors.

3.1. Dataset characteristics

We show in Fig. 1 an intuitive representation of the mobility traces of three users selected from the dataset. The figure depicts a user's behavior over a year as a matrix, where each column is a day of the year and each line an interval of 1 h. We map each place to a color and leave blank intervals of time during which we have no information about the user's location.

User 143, whose mobility is represented in Fig. 1(a), has a very regular behavior, which seems to support the results (such as those presented by Song et al. [3]) claiming that human mobility is very predictable. However, similarly to User 1