From Foursquare to my Square: Learning Check-in Behavior from Multiple Sources

Eric Malmi
Aalto University and Idiap
Finland / Switzerland
eric.malmi@gmail.com

Trinh Minh Tri Do
Idiap Research Institute
Martigny, Switzerland
do@idiap.ch

Daniel Gatica-Perez
Idiap and EPFL
Switzerland
gatica@idiap.ch

Abstract

Location-based services often use only a single mobility data source, which typically will be scarce for any new user when the system starts out. We propose a transfer learning method to characterize the temporal distribution of places of individuals by using an external, additional, large-scale check-in data set such as Foursquare data. The method is applied to the next place prediction problem, and we show that the incorporation of additional data through the proposed method improves the prediction accuracy when there is a limited amount of prior data.

Introduction

Building location-based context-aware applications and recommender systems requires learning models of user behavior. Despite the availability of a huge amounts of check-in data from various location-based social networks (LBSNs) (Cheng et al. 2011), these models are traditionally learned using only a single check-in data source, since the correspondences between places and between users of different location data sets are typically unknown (Malmi, Do, and Gatica-Perez 2012). The use of external large-scale resources would be especially advantageous for applications which have not yet collected a sufficient amount of data in order to make useful inferences on the behavior of their users. We present an approach to address this problem.

We introduce a transfer learning method which can utilize additional data from any source providing check-in data. The method learns the check-in time distribution of a place by using a generic mixture model on top of the previous check-in times of the place. The generic mixture model can be learned from an additional large-scale data source, which provides prior information of the type of distributions we expect to observe and it is used via the *posterior predictive distribution*. We consider a density estimation problem and do transfer learning in an unsupervised manner which is a relatively new research area (Pan and Yang 2010).

To evaluate our method, we use data from the Nokia Mobile Data Challenge (MDC) and Foursquare (4sq). We apply the method to the prediction of the next place of an MDC user and demonstrate that the method can improve the predictive performance of a personalized model. This is done by incorporating additional information from the 4sq data set when there is limited individual mobility data, a common situation often called *cold start*. To our knowledge, this is the first time transfer learning has been applied to the next place prediction problem.

Methods

Our objective is to model the check-in time distribution of a place with only a few previous check-ins (e.g. a new user of a location-based service), using an additional check-in data set from a different source containing a vast amount of check-ins (e.g. from a LBSN like Foursquare). The method we propose for this transfer learning problem contains three stages: First, we learn place clusters using the additional data set (*source data*). Second, we estimate the time distributions of the original places using the previous check-in times (*target data*) and the clusters via the posterior predictive distribution. The third stage is a next place prediction method which uses this transfer learning approach.

Stage 1. Clustering Places based on Check-in Time

Our goal is to characterize generic place categories based on check-in time. Technically speaking, we want to find place categories so that within a category all places follow the same time distribution. This problem can be seen as a clustering problem where each cluster is represented by a temporal model. Intuitively, one would expect that places have structure from these perspectives. For instance, restaurants are visited mainly at specific times, as are bars and private places. What distinguishes this from a traditional clustering problem is that normally the samples to be clustered are individual data points, but in our case they are places containing a varying number of check-in times. A similar problem has been previously addressed in (Cheng et al. 2011).

Our approach is to learn a mixture model for the data so that each component of the mixture corresponds to a cluster. We introduce latent variables that indicate which component has generated the check-ins of a place. The check-in times are considered discrete hour-weekday pairs: $t = (h, d) \in (\{0, 1, \dots, 23\} \times \{1, 2, \dots, 7\})$. Thus we use two multinomial distributions for each component and model the data with a *mixture of multinomials* which is learned using the Expectation-Maximization algorithm (Rigouste, Cappé, and Yvon 2007). Furthermore, we assume that the hour and weekday are independent, i.e., $p(h,d) = p(h)p(d)$. The resulting density function for a single check-in time is

$$p(t \mid \boldsymbol{\theta}, \boldsymbol{\varphi}) = \sum_{c} \pi_c p(h \mid \boldsymbol{\theta}_c) p(d \mid \boldsymbol{\varphi}_c), \tag{1}$$

where distributions p are multinomial, $\boldsymbol{\theta}_c$ and $\boldsymbol{\varphi}_c$ are the multinomial parameters of the hour and day distribution of mixture component c, respectively, and π_c is the mixing proportion of component c.

Stage 2. Inferring Time Distributions Across Data Sets

We present three models for estimating the time distribution of a place, namely *Multinomial*, *Posterior Predictive Distribution (PPD)*, and *Combined*. The first model does the estimation based on the previous check-in times of the place, i.e. the target data, by calculating check-in counts. In addition, we apply Laplace smoothing to the multinomial distribution so that all times have a non-zero probability. This model does not use the source data set.

In the latter two models, we use a mixture of multinomials learned from the source data. We assume that the underlying place categories are the same in the source and the target data and introduce a latent variable z which indicates the component (= category) that has generated the previous check-in times $\mathbf{t}$. In the second model, the PPD of the next check-in time $\tilde{t}$ takes the form

$$\begin{aligned} p(\tilde{t} \mid \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\varphi}) &= \sum_{c=1}^{C} p(\tilde{t} \mid z=c, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\varphi})\, p(z=c \mid \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\varphi}) \\ &\approx \sum_{c=1}^{C} p(\tilde{t} \mid z=c, \boldsymbol{\theta}, \boldsymbol{\varphi})\, p(z=c \mid \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\varphi}) \\ &\propto \sum_{c=1}^{C} p(\tilde{t} \mid z=c, \boldsymbol{\theta}, \boldsymbol{\varphi})\, p(z=c \mid \boldsymbol{\theta}, \boldsymbol{\varphi})\, p(\mathbf{t} \mid z=c, \boldsymbol{\theta}, \boldsymbol{\varphi}) \\ &\approx \sum_{c=1}^{C} p(\tilde{h} \mid \boldsymbol{\theta}_c) p(\tilde{d} \mid \boldsymbol{\varphi}_c) \pi_c p(\mathbf{h} \mid \boldsymbol{\theta}_c) p(\mathbf{d} \mid \boldsymbol{\varphi}_c). \end{aligned} \tag{2}$$

Note that we have approximated $p(z=c \mid \boldsymbol{\theta}, \boldsymbol{\varphi}) = p(z=c) \approx \pi_c$ which does not strictly hold since the relative proportions of different categories are not necessarily the same in the two data sets. To give an idea of what the PPD does in practice, we show, in Fig. 1, an example of the Multinomial model and the corresponding PPD model learned from the data described in the next section. We can see that the PPD smooths the multinomial distribution, but on the other hand, the effect of Laplace smoothing disappears.

Finally, in the Combined model, we take a linear combination of the Multinomial model $(\boldsymbol{\theta}', \boldsymbol{\varphi}')$ and the PPD model (Eq. 2). This is motivated by the fact that the target and the source data do not follow exactly the same distribution in practice (one of the basic assumptions in several transfer learning techniques) (Pan and Yang 2010) and if the original place has been visited many times in the past, a simple multinomial estimate might be accurate. The resulting model is the following

$$\begin{aligned} &p(\tilde{t} \mid \mathbf{t}, \boldsymbol{\theta}', \boldsymbol{\varphi}', \boldsymbol{\theta}, \boldsymbol{\varphi}) = \\ &\quad \alpha\, p(\tilde{t} \mid \boldsymbol{\theta}', \boldsymbol{\varphi}') + (1-\alpha)\, p(\tilde{t} \mid \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\varphi}), \end{aligned} \tag{3}$$

where α is called *transfer coefficient*.

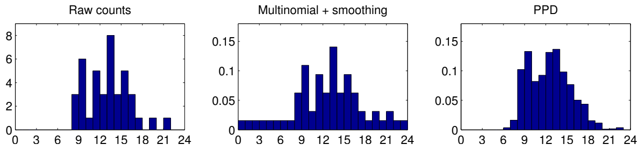

Figure 1: Raw check-in hour counts (left), the corresponding Multinomial model which includes Laplace smoothing (middle) and the corresponding PPD model which uses a mixture model learned from the source data (right).

Stage 3. Next Place Prediction

In next place prediction, the task is to find where a user will go given his mobility history (Laurila et al. 2012). Formally, let $\{(X_1, T_1), ...(X_n, T_n)\}$ be the check-in history of a given user where X_i represents the place of the i^{th} visit and T_i represents the temporal information of the visit. We want to predict the most probable value x of random variable X_{n+1} and, in our problem setting, we assume that we are given the current place $X_n = x'$ and the end time of the current visit $T_n^e = t'$. Thus the task can be formulated as follows

$$x_{pred} = \underset{x}{\operatorname{argmax}} \{p(X_{n+1} = x \,|\, X_n = x', T_n^e = t')\}\,. \tag{4}$$

We shall denote the probability simply by $p(x_{n+1} \mid x_n, t_n^e)$, which is estimated based on the mobility history.

Typically, the amount of data is insufficient to estimate the probability directly. We combine the approaches presented recently by (Etter, Kafsi, and Kazemi 2012) and (Gao, Tang, and Liu 2012), and obtain the following model

$$\begin{aligned} &p(x_{n+1} \mid x_n, t_n^e) \approx \\ &\quad p(x_{n+1} \mid x_n) \sum_{t_{n+1}^s} \Big\{ p(\Delta t_n \mid \Delta x_n) p(t_{n+1}^s \mid x_{n+1}) \Big\}, \end{aligned} \tag{5}$$

where $p(x_{n+1} \mid x_n)$ is a Markov model, $p(\Delta t_n \mid \Delta x_n)$ is a travel time distribution between the end time of the current visit (t_n^e) and the start time (check-in time) of the next visit (t_{n+1}^s). Due to the summation, the time distribution $p(t_{n+1}^s \mid x_{n+1})$ does not depend on the end time of the visit, which allows us to estimate it using the PPD model and a source data set that only contains the check-in times.

Data

We use the two data sets studied previously in (Malmi, Do, and Gatica-Perez 2012). The target data set comes from the *Nokia Mobile Data Challenge* (MDC) (Laurila et al. 2012). The MDC data set contains daily life data from 80 users and about 16 months in Switzerland. The users were given a smartphone which used a variety of sensor data to infer instantaneous locations and visited places of the users. The place detection algorithm is described in (Montoliu and Gatica-Perez 2010).

The source data set comes from *Foursquare* (4sq), which is a highly popular LBSN, and it has been collected through the Twitter API accepting only check-ins from Switzerland. Our aim is to predict the check-in behavior of the MDC users by learning additional temporal characteristics of the MDC places using the 4sq data. While the MDC visits contain both a start and an end time, the 4sq check-ins contain only the check-in time. In this study, we do not consider, e.g., the location but only the check-in times of a place. Table 1 summarizes the key statistics of the two data sets.

Table 1: Key statistics of the MDC and 4sq data sets.

	#places	#users	#visits	First visit	Last visit
MDC	7281	80	51607	30 Sep 2009	4 Feb 2011
4sq	17482	302	40629	19 Dec 2011	21 Jun 2012

Results

We first optimize the parameters of the models, and then evaluate the performance of the transfer learning applied to the time distribution inference and to the next place prediction problem. To calculate the results, the MDC data set is divided into training data (Set B), validation data (Set C), and test data (Set T) as shown in Fig. 2.

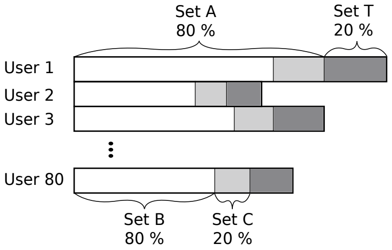

Figure 2: The check-in sequences of the MDC users are divided into different sets used for training (Set B), validation (Set C), and testing (Set T). Set A denotes B ∪ C.

Estimation of Model Parameters

We learn the multinomial mixture model from the 4sq data set to cluster MDC places. Using *k-fold cross-validation*, we find 80 to be the optimal number of mixture components. Fig. 3 shows the three components with the highest mixing coefficients. We observe that different components capture different kind of places. Component 2 contains workplaces as it is mostly visited during weekdays in the morning (when people arrive to work) and in the afternoon (when they return from lunch), whereas component 3 contains nightlife spots and homes where people arrive to in the evening. Note that in our approach, we do not assume that clusters correspond exactly to 4sq venue categories, which are semantic categories rather than data-driven components.

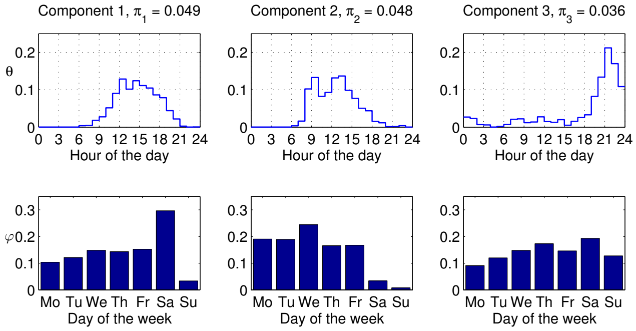

Figure 3: Daily and weekly check-in time distributions of the three components with the highest mixing coefficients in the 4sq mixture model.

The transfer coefficient α from Eq. 3 is learned using Set B as training data and Set C as validation data. The highest prediction accuracy on Set C is obtained when $\alpha = 0.6$. The travel time distribution $p(\Delta t_n \mid \Delta x_n)$ is learned using Set A in whole. Travel time Δt_n is divided into bins of one hour, which is the time resolution of our model, and travel distance Δx_n into bins of five kilometers.

Evaluation of Time Distribution Estimation

To evaluate how accurately we can predict the future check-in times of a place by using only the MDC data or by using the transfer learning approach, we divide Set A as follows: the last 30 check-ins of each place in Set A are used as test data and N_p preceding check-ins are used for training while varying N_p. The higher the log likelihood obtained for the unseen test data, the better the model.

Figure 4 shows log likelihood as a function of N_p. We can see that the PPD model performs clearly better than the Multinomial model when there is little data for a place. However, only about five visits seem to be sufficient in PPD to estimate to which clusters the place belongs to and thus the likelihood converges in relatively few steps. Nevertheless, by combining the two models according to Eq. 3, we observe a better performance than either of the individual methods when N_p is around 35 and elsewhere a better or similar performance compared to the Multinomial model.

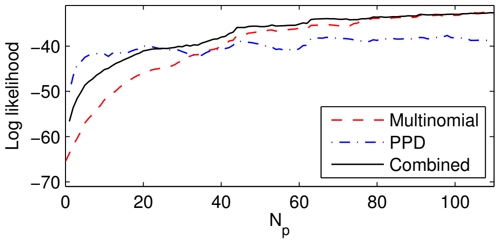

Figure 4: Log likelihood as a function of the number of preceding training visits per place (N_p) for the three methods.

Evaluation of Next Place Prediction Task

Finally, we assess if using transfer learning with data from 4sq helps to predict the mobility patterns of the MDC users. The prediction test accuracies are calculated based on Set T (see Fig. 2) by taking an average of each user's accuracy weighted by the number of test samples the user has. We vary the number of training place transitions per user (N_u) in Set A, starting from the most recent samples, so that we see how the amount of training data affects the performance of transfer learning. Test samples always start right after the last training sample and the test data set (Set T) is thus kept fixed. The only exception is that the users who do not have enough training data are not taken into account.

The next place prediction results are shown in Fig. 5 (top-left). The accuracy naturally increases when N_u increases. However, there are also some drops in the accuracy since the set of users on whom the accuracy is calculated might vary as the users with too few training samples are ignored. When the number of training samples is 450, there are only ten MDC users with enough data. The relative performance of the methods, shown in Fig. 5 (top-right), shows that a 2 percent absolute improvement in prediction accuracy is obtained with transfer learning when N_u is around 20, which corresponds to 9 days of data collection when averaged over all users. One-tailed paired t-test shows that the improvement is statistically significant at the 0.01 level. After about 100 training samples, corresponding to 45 days, transfer learning no longer improves the prediction accuracy, which suggests that the personalized data is better on its own.

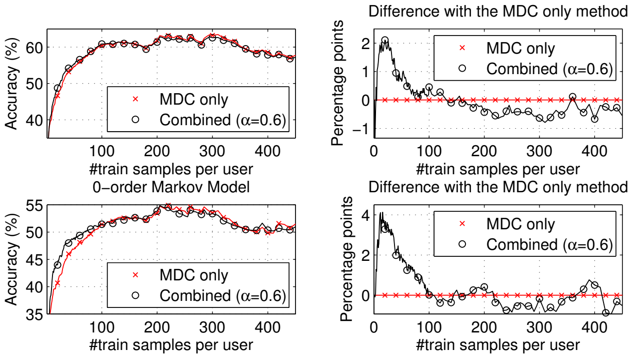

Figure 5: Top-left: next place prediction accuracy for the Multinomial model, which is the baseline method, and the Combined model, which uses transfer learning. Top-right: accuracy differences between the two methods. Bottom: performance when the Markov model part in Eq. 5 has been replaced by a 0-order Markov model.

Transfer learning is only applied to the time distributions but not to the Markov model part of Eq. 5, which mainly dominates the predictions. To suppress the impact of the Markov model and understand the sole effect of transfer learning, we calculate the prediction accuracies using a 0-order Markov model instead of the standard 1st-order Markov model. The results are shown on the bottom of Fig. 5. Now we can see the difference more clearly: a 4 percent absolute improvement is obtained with transfer learning when N_u is around 20.

Conclusions

For a mobility prediction task, we presented a transfer learning method to learn the check-in time distributions of places, using the advantages of additional large-scale check-in data available in LBSNs like Foursquare, in a complementary manner. The additional data set can originate from any check-in data source and it can also be a combination of several data sets. We showed that the proposed method (Combined) outperforms the traditional method (Multinomial) in terms of likelihood. Furthermore, we applied our method to the next place prediction problem, and showed that it improves the predictions when the users have up to 45 days of data. The method thus helps tackling the cold start problem. Furthermore, new LBSNs or games could use the method to improve their predictive analytics and recommender systems.

Acknowledgments

This research was funded by the LS-CONTEXT project (Nokia Research Lausanne) and the HAI project (Swiss National Science Foundation).

References

Cheng, Z.; Caverlee, J.; Kamath, K. Y.; and Lee, K. 2011. Toward traffic-driven location-based web search. In *Proc. ACM Int. Conf. on Information and Knowledge Management*.

Etter, V.; Kafsi, M.; and Kazemi, E. 2012. Been there, done that: What your mobility traces reveal about your behavior. In *Proc. Mobile Data Challenge Workshop*.

Gao, H.; Tang, J.; and Liu, H. 2012. Mobile location prediction in spatio-temporal context. In *Proc. Mobile Data Challenge Workshop*.

Laurila, J.; Gatica-Perez, D.; Aad, I.; Blom, J.; Bornet, O.; Do, T.-M.-T.; Dousse, O.; Eberle, J.; and Miettinen, M. 2012. The mobile data challenge: Big data for mobile computing research. In *Proc. Mobile Data Challenge Workshop*.

Malmi, E.; Do, T.; and Gatica-Perez, D. 2012. Checking in or checked in: comparing large-scale manual and automatic location disclosure patterns. In *Proc. Int. Conf. on Mobile and Ubiquitous Multimedia*.

Montoliu, R., and Gatica-Perez, D. 2010. Discovering human places of interest from multimodal mobile phone data. In *Proc. Int. Conf. on Mobile and Ubiquitous Multimedia*.

Pan, S., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359.

Rigouste, L.; Cappé, O.; and Yvon, F. 2007. Inference and evaluation of the multinomial mixture model for text clustering. *Information processing & management* 43(5):1260–1280.