Location-Based Activity Recognition with Hierarchical Dirichlet Process

Negar Ghourchian
McGill University, Montreal, Canada
negar@cs.mcgill.ca

Abstract

We consider the problem of analyzing people's mobility and movement patterns from their location history, gathered by mobile devices. Human mobility traces can be extremely complex and unpredictable, by nature, which makes it hard to construct accurate models of mobility behavior. In this work, we present a novel high-level strategy for mobility data analysis based on Hierarchical Dirichlet process, which is a powerful probabilistic model for clustering grouped data. We evaluate our unsupervised approach on two real-world datasets.

1 Introduction

The growing popularity of mobile computing devices with integrated location sensing technologies has promoted a huge interest in analyzing data collected from such devices. These modern intelligent devices have the ability to accurately track moving objects, store a huge amount of mobility data and transfer this information. Given this massive volume of information, wide research efforts have been put towards different directions in trajectory analysis. Most of the classical studies focus on defining a spatial-temporal model that synthesize or summarize the movement patterns into a statistical model [Zheng, 2015]. For instance, many studies have used Markov models to represent the mobility behavior of an individual and predict their next location. However, these approaches are limited while facing geometrical complexity of long-term human mobility traces, specially when the number of clusters are not known *a priori*. Therefore, increasing number of studies have started to consider higher-level analysis of long and complex trajectories, where semantic information and personal interests are used to infer meaningful properties.

We are interested in clustering mobility traces of human subjects to infer social ties from their physical proximity. Our main strategy is to discover and characterize the places of interest (POIs) frequently visited by each user, and subsequently build a similarity measure between users based on the proximity of their POIs. We assume that different social groups exhibit distinct profiles in terms of the places where they hang out, so users can be clustered based on the distribution of their POIs. We propose to employ a Hierarchical Dirichlet Process (HDP), which automatically adapts the number of identified clusters and allows sharing of components between clusters. However, the HDP also faces some challenges in practice: the number of clusters grows logarithmically with the size of the data, which can yield to overfitting. We propose a regularization step on top of the HDP model, where we use a distance measure to prune very similar clusters and to compute the proximity between pairs of individuals, by comparing their *a posteriori* distribution of POIs.

2 Hierarchical Location Clustering

Inspired by the topic modeling approach used for text documents, we model the location traces (i.e., "documents") using a Dirichlet process mixture model, in which each mobility behaviour profile (i.e., "topic") is a distribution across the location points(i.e., "words"). There are two practical challenges that motivate the usage of topic modelling framework for human mobility analysis. First is that users belonging to different mobility behaviour profiles, may share a large set of common location points. Second, in many real-world scenarios there is no *prior* knowledge on the number of mobility profiles. We aim to model the data points in each observation group with a Dirichlet process (DP) mixture model. While different groups share the same set of mixture components, each mixture has a mixing proportion specific to the group.

2.1 Inferring Places of Interest

Suppose there are J groups of location trajectories, each consisting of n_j exchangeable location points, $L_j = (l_{j1}, l_{j2}, \cdots, l_{jn_j})$, from total N possible locations. These J observation groups represent the daily mobility of U individual users, where each user has J_u entries in the data, such that $\sum_{u=1}^{U} J_u = J$. While all individuals travel across the same set of locations, each daily trace has its own characteristics due to a specific combination of visit frequencies. This naturally fits into the HDP setup, where the clusters (mobility profiles) are probability distributions over all possible locations. A given mobility trace can be related to several clusters and is modelled as a sample from a mixture of corresponding clusters. The actual mixing proportions are defined by the location counts and "interest" in a particular place is estimated by the number of timestamps that the user was at that place.

The HDP defines a conditional distribution over cluster assignments $P(\boldsymbol{c}|\boldsymbol{L})$, where $\boldsymbol{L} = \{L_1, \cdots, L_J\}$ are the loca-

tion traces and $\boldsymbol{c} = \{c_1, \cdots, c_r\}$ are r assigned clusters. Using Bayes rule, $P(\boldsymbol{c}|\boldsymbol{L}) = \frac{P(\boldsymbol{L}|\boldsymbol{c})P(\boldsymbol{c})}{P(\boldsymbol{L})}$, where $P(\boldsymbol{L})$ is computed from the location traces by counting the occurrences of each location and $P(\boldsymbol{c})$ from the clustering of the entire traces. The distribution $P(\boldsymbol{L}|\boldsymbol{c})$ is estimated from the visited location frequencies of each trace, as:

$$P(L_j|\boldsymbol{c}) = \sum_{i=1}^{N} P(L_j, l_i|c) = \sum_{i=1}^{N} P(L_j|l_i, c)P(l_i|c) \quad (1)$$

In $P(\boldsymbol{c}|\boldsymbol{L})$, each particular cluster represents a *mobility profile* among user mobility patterns.

2.2 Regularization

Due to the nonparametric nature of HDP models, the number of clusters is a random variable whose mean grows at a logarithmic rate with respect to the number of data points. The drawback is the emergence of too many similar clusters that represent the same behavior profile. Moreover, the resulting posterior probability distribution from the HDP will rarely provide an explicit correspondence between clusters and POIs. We address these problems by introducing a regularization step on top of the HDP, in order to *a*) measure the distance between derived mixture models, in order to prune the clusters generated by HDP that are too similar; and *b*) compute the proximity between pair of individuals by comparing their posterior distribution of POIs. Depending on the application scenario, many distance measures have been used to compute a distance between two probability distributions. In practice, we found that ℓ_2-norm and KL-Divergence measures are the most effective approaches for our application. Given the distribution $P(L|c)$, where $P(l_k|c_i)$ denotes cluster i with corresponding probability distributions over location point $k = \{1, \cdots, N\}$, we define $L2(i,j)$ and $D(i,j)$ scores between each pair of $P(l|c_i), P(l|c_j) \in P(l|c)$ as:

$$L2(i,j) = \sum_{k=1}^{N} \|P(l_k|c_i - P(l_k|c_j)\|^2, \quad (2)$$

$$\begin{aligned} D(i,j) = &\sum_{k=1}^{N} P(l_k|c_i) \log \frac{P(l_k|c_i)}{P(l_k|c_j)} + \\ &+ \sum_{k=1}^{N} P(l_k|c_j) \log \frac{P(l_k|c_j)}{P(l_k|c_i)}. \end{aligned} \quad (3)$$

$L2$ is a standard ℓ_2-norm score that efficiently discover correlation between clusters and $D(i,j)$ is a variation of KL-divergence, developed for comparison of the DP mixtures.

3 Experimental Results

Proposed approaches are evaluated on two datasets from MIT Reality Mining projects [Olguín *et al.*, 2009; Eagle and Pentland, 2006].

Badge Dataset documents the work performance of 23 employees at an IT facility over one month. In total 1900 hours of indoor location data were collected from employees while they were asked to perform their daily tasks, from a set of $\{configuration, pricing, coordinator\}$ job titles. Each trajectory correspond to one specific completed task, which includes a lot of walking around the workspace and interaction with others, and can take from minutes to hours. Employees from different job titles share the same workspace and they regularly interact with other employees from different task titles. In this realistic scenario, we employed the $L2$-HDP and KL-HDP algorithms to predicting the task titles from location trajectories, and these algorithms obtained 75% and 82% accuracy rates, respectively.

Reality Mining Dataset includes mobility data of 106 subjects' (students and staff at MIT) outdoor location, collected using their mobile phones, over nine months. The location of each subject at each timestamp was estimated from the cell towers present in their vicinity, which permits localization within 100-200 meters. The dataset also contains a self-report survey in which subjects were asked about their level of physical proximity with others. In this project our goal was to infer the level of pairwise proximity between users by discovering the distribution of their POIs. We validated the results obtained by our method against the labels provided by users in the survey responses. As an additional baseline model, we computed the most frequented places estimated from the probability distribution of visited locations, and applied k-means clustering to discover social ties among users. Our approaches, $L2$-HDP and KL-HDP have obtained 87% and 92% accuracy rate, respectively, and outperformed the k-means baseline with 82% accuracy rate.

4 Conclusion and Future Work

We proposed a method for analyzing location data obtained from mobile devices, in order to glean information about the social behaviour and interactions of the users. We demonstrated that this approach can be successful using two different real datasets. While a lot of the work on HDPs assumes that the strength of the prior will be sufficient to control the model size, our regularization approaches provide a tighter control over this aspect. We are currently working on providing a theoretical explanation of the effect of the KL divergence use on the model from a Bayesian perspective. We anticipate that the proposed methodology would have a positive impact on other cases in which the possible model complexity is in fact bounded, but we still want it to grow as more data become available.

References

[Eagle and Pentland, 2006] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[Olguín *et al.*, 2009] Daniel Olguín Olguín, Benjamin N Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1):43–55, 2009.

[Zheng, 2015] Yu Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29, 2015.