The Pennsylvania State University

The Graduate School

Department of Electrical Engineering

A MIXED MARKOV MODEL APPROACH TO PREDICT FUTURE POINTS OF INTEREST IN INDOOR SPACE

A Thesis in

Electrical Engineering

by

Ying Sun

© 2013 Ying Sun

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2013

The thesis of Ying Sun was reviewed and approved* by the following:

David J. Miller
Professor of Electrical Engineering
Thesis Advisor

Tim Kane
Professor of Electrical Engineering

Conrad Tucker
Assistant Professor of Industrial and Manufacturing Engineering

Kultegin Aydin
Professor of Electrical Engineering
Head of the Department of Electrical Engineering

*Signatures are on file in the Graduate School

ABSTRACT

With the advances in GPS and other location acquisition technologies, an increasing amount of trajectory data is being captured and recorded. This trajectory data has attracted the attention of researchers from many domains due to the potential benefits of discovering the underlying behavior patterns and predicting future trajectories of the users. In this thesis, we focus on indoor trajectory data in an attempt to detect heterogeneous trajectory patterns across users within the raw x-y coordinated data records in an indoor environment and further predict individuals' future trajectories. Due to the heterogeneity in human behaviors, individuals do not exhibit the same pattern in their trajectories. Consequently, one single trajectory model is not capable of capturing these behavior patterns. To tackle this problem, we propose a Mixed Markov Model (MMM) approach that models the latent trajectory patterns from all input trajectories without the need of user identification. This study may show high impact in domains, such as health care, transportation systems, public security, etc., where privacy issue and heterogeneity in behavior patterns are concerned. A case study using real indoor trajectory data of workers in an engineering design space is presented for validating our prediction model.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Foremost, I would like to express my deepest appreciation to my advisor Dr. Miller for the support of this study. His patient guidance and profound knowledge helped me all the way in doing the research and writing this thesis. Without his guidance and persistent help this thesis would not be possible to finish.

I also would like to thank my committee members, Dr. Tucker and Dr. Kane for their words greatly encouraged me to keep on doing this research. Special thanks to Dr. Tucker. This work is a continuing of a summer project working with Dr. Tucker, who gave me the opportunity to start this fascinating research work and gave me inspiration and motivation in the start of this work.

Finally, I would like to thank my grandparents, who bring me up to pursue a Master's degree. This thesis is a gift to them for their loving care and persistent support of my study, life and spirit.

Chapter 1

Introduction

The past few years have witnessed the fast development of wireless communication techniques (GPRS, Bluetooth, RF). Some technologies are synchronized with Global Positioning System (GPS) which can provide accurate time-tagged x-y coordinates of moving objects. These location acquisition technologies have been embedded in a huge number of mobile devices to capture all kinds of trajectories, leading to an era of ubiquitous position tracking. The availability of tremendous spatial-temporal data urges researchers to develop data mining algorithms to extract interesting patterns of people’s behaviors. The valuable knowledge discovered in the spatial data has been applied to many location-based services to benefit people’s lives everywhere. For example, people can search for the nearest restaurants or hotels with high rates from other people’s recommendation when they travel to a new place; decision makers from the areas of business, marketing and government are also interested in the data with spatial-temporal data: marketing managers want to know customers’ behavior patterns so that they can optimize their marketing strategies to offer the right services for the targeted customer; police can make use of the data to detect the patterns of criminals who commit crimes to do a better job in guarding public security. Thus, the ability to collect and store spatial-temporal data enables people from various areas to discover useful patterns and knowledge for their different purposes.

In the meantime, a lot of indoor tracking systems have also been set up to monitor people's trajectories, among which RFID-based systems have been widely used for indoor trajectory tracking. RFID (RF Identification) is a means of storing and retrieving data through electromagnetic transmission to an RF compatible integrated circuit and is now being seen as a radical means of enhancing data handling processes (Chiesa et al., 2002). Users wear tags to be tracked. There is communication between the receiver and tags for sampling users' locations periodically. However, the trajectory is recorded real-time or may even be delayed. That is, it may be too late to take measures when there are abnormal signs detected, such as abnormal behavior in a nursing home or changes of customers' purchasing behaviors in shopping malls, etc. Besides, it costs a lot of human resource for manual monitoring. Thus, there is an urgent need to develop spatio-temporal data mining techniques to automatically extract interesting patterns and knowledge in user trajectories to facilitate quick responses and proper measures taken instantly, according to the system prediction.

Several proposals for predicting people’s future trajectories are based on people’s trajectory histories. However, in an indoor space there are a lot of customer flows. It will be a waste to build database to record individual’s trajectories. Besides, people may also have privacy concerns. Thus, we aim to predict the trajectories of people in an indoor environment to realize an intelligent indoor tracking system while simultaneously maintaining people's privacy. For example, the employees in nursing homes can prevent potential wandering or elopement behavior of the elders or patients; The manager or owner of groceries are able to adjust selling strategies according to customers' purchasing behavior patterns. The basic idea is: there exist some latent patterns among the trajectory data generated by all users, so that the trajectories can be grouped into several clusters with their unique trajectory features, instead of studying each individual’s trajectory separately. The methodology follows two steps: first, we transfer the x-y coordinated locations into states. Here, a state means a region or place of interest. So, the trajectories can be presented as a sequence of states; second, we model these trajectories by an MMM and train the parameters of the MMM by applying the EM algorithm to learn the latent patterns of the trajectories and the state transition probabilities of each pattern; finally, we can estimate the current state of the coming trajectory and make prediction for its future states based on the state transition matrix trained for each cluster. In order to make prediction more precise, we also make use of the specific time and location information. We assume that (a) people's behavior varies temporally. For example, one may come and stay in his office from 9am to12am every weekday. So, the state transition probability will be different at different times of the day; (b) the state transition probability will also depend on the location. That means, people "on the way" to (with a trajectory toward) region B from region A will definitely have higher transition probability of going to B (A->B) than going to X (A->X, where X denotes other regions except B).

The potential contributions of this thesis include: 1) the technique of this paper is free from domain knowledge, so that our method can be applied in other scenarios for trajectory prediction. For example, car renting, terrorism alert, shopping at stores, etc.; 2) we count different patterns of user behaviors by applying Mixed Markov Model; 3) we take the specific time and location information into consideration to make better predictions.

The rest of the thesis is organized as follows. In chapter 2, we conduct a literature review on general methods of trajectory pattern analysis, ideas of trajectory prediction in previous papers and related methods that will be applied in the methodology of this paper. Chapter 3 proposes a methodology of Mixed Markov model based trajectory pattern learning and prediction. Chapter 4 presents results of the experiment. Chapter 5 evaluates the results and gives a discussion. Chapter 6 concludes the thesis.

Chapter 2

Literature Review

In this chapter, we review previous studies related to trajectory pattern mining. First, we summarize some general methods that have been widely used in trajectory pattern analysis. Second, we take a look at ideas that have been proposed by researchers that address the problem of trajectory prediction. Finally, we introduce some basic concepts of MMM, which was employed in our prediction model.

2.1 Trajectory Pattern Analysis

Owing to the tremendous increase in the amount of spatial-temporal data, data mining techniques, such as clustering, pattern detection, etc., applied to the analysis of this kind of data have been widely studied in different scenarios (e.g., epidemic surveillance (Pfeiffer et al., 2007), evacuation tracking (Andrienko et al., 2009), etc.). For example, Andrienko et al. (2009) applied density-based clustering OPTICS to two types of spatiotemporal data: trajectories of moving objects (evacuation traces) and events (landings and interdictions of moving entities). Several spatio-temporal data mining methods (including clustering, association rules and sequential pattern detection) were investigated to help address RFID trajectory data challenges.

2.1.1 Spatial-temporal data clustering

Data Mining Clustering aims to discover groups of objects that share similar attributes/patterns in a high dimensional space. However, clustering methods often consider objects as points with multidimensional properties, which may not be appropriate for structurally complex data sets (e.g., trajectories of objects and other spatio-temporal data types). Specifically, clustering in general assumes that coordinates of data points are not ordered according to some independent variable, such as time. In order to overcome these limitations and accommodate to the clustering of spatio-temporal data, a generic notion of clustering can be applied by defining a specific distance function which measures the degree of dissimilarity between data items. I.e., spatial-temporal clustering here refers to a process of grouping objects based on their spatial and temporal similarities (Kisilevich et al., 2010). In the following, we present two clustering methods for trajectory data related to our methodology: model-based clustering and clustering for extracting places of interest.

Model-based clustering

The objective of model-based clustering is to derive a global model capable of describing the whole trajectories in the dataset (Kisilevich et al., 2010). In Gaffney et al.'s work, the authors proposed a generative probabilistic approach for clustering continuous trajectories based on mixture models (Gaffney et al., 1999). In their method, each cluster is modeled in a regression manner as a smooth function of time with additive noise. The trajectories which are likely to be generated from the same regression model component are grouped into one cluster. Generally used model types (referred to a priori) include Gaussian, multinominal, Hidden Markov Models (HMM), etc., among which, Markov Chains and HMMs (Smyth, 1997; Bengio, 1999; Alon et al., 2003) have been recognized as the dominant models for time sequences. Compared to similarity-based methods, model-based methods offer better interpretability since the resulting model for each cluster directly characterizes that cluster (Zhong et al., 2003).

Places of Interest

Enriching trajectories with semantic meaning offers an insightful perspective in trajectory data analysis. Especially for our purpose of next-location prediction, detecting of places of interest or important places in the trajectory data, i.e., frequently visited places, places that people stay for a long time or pass regularly during some time of the day, is critical in the first step. Alvares et al. proposed a model of enriching trajectories with semantic geographical information by detecting the stops and moves (Alvares et al., 2007). In their definition, stops refer to interesting spatial locations, such as traffic lights in transportation system or parks in people's traveling routes. By constructing the table of stops, their application was able to query the important places during a given time. Also, the trajectory was conferred with semantics by being transferred into the "stop"s and "move"s. Their algorithm for finding "stop"s and "move"s verifies for each point of a trajectory T if it intersects the geometry of a candidate stop. In the following work of Alvares et. al, they presented an alternative solution to find interesting places which are not expected by the user (Palma et al., 2008). The proposed solution called CB-SMoT (Clustering-Based Stops and Moves of Trajectories) is a spatio-temporal clustering method based on speed, dealing with single trajectories. This approach mainly deals with the cases where the low speed is the main indicator of an interesting event, such as traffic jams.

2.1.2 Association rule mining

Association rule mining (ARM) seeks to discover association among transactions encoded within a database (Agrawal et al., 1993). The associations detected in spatio-temporal data can offer people an insightful view of mobility characteristics of moving objects and can be further used to predict their future movements. For example, detecting the movement patterns of customers over time in a shopping mall may help the manager to optimize the utilization of space and inform the store owners to pre-prepare their service, which is predicted to be in need in the future few hours. Attempts to explore these spatio-temporal relationships have been made by researchers from spatio-temporal domain by mapping the data to transactions. Mennis and Liu applied ARM to reveal the association among processes of socioeconomic change and urban growth in Denver, Colorado, U.S.A (Mennis and Liu, 2005). In the technical report of Verhein et al., they provide a comprehensive definition of spatio-temporal association rules (STARs) that describe how an object moves between regions over time (Verhein et al., 2006). Further, they are devoted to finding regions with useful temporal characteristics (thoroughfares, sinks, sources and stationary regions) and to predicting how objects move through the regions based on spatio-temporal rules that are discovered. Their approach presents the spatio-temporal patterns over a time window so that it enables the observation of seeing the changing nature of patterns over time.

2.1.3 Sequential pattern detection

Patterns that are mined from trajectories are called trajectory patterns and characterize interesting behaviors of a single object or group of moving objects (Fosca et al., 2008). These patterns provide useful information for inferring the past trajectories or future movements of objects. Given a spatio-temporal series, Cao et al. study the problem of discovering sequential patterns, which refer to routes frequently followed by the object (Cao et al., 2005). The idea is: they transfer the original series of spatial locations into a list of frequently visited spatial regions by summarizing the line segments. A substring tree structure was employed to find patterns.

Instead of focusing on the trajectory segments, Giannotti et al, on the other hand, developed an extension of the sequential pattern mining paradigm to find frequent movement patterns by taking both space (i.e., the regions of space visited during movements) and time (i.e., the duration of movements) into account (Giannotti et al., 2007). A T-pattern is introduced to describe a sequence of consecutive points with annotation of temporal transitions. Their methodology follows two steps: first, the regions of interest need to be detected. Then the input trajectories are transformed from sequences of points to that of regions of interest; second, temporal annotation is counted for mining the underlying patterns by applying the method proposed in (Giannotti et al., 2006).

2.2 Trajectory Prediction

With the pervasiveness of location-based services, there is an unprecedented availability of huge amounts of spatio-temporal data, which attracts researchers' interest to tackle research problems on human mobility. Among these problems, prediction of people's next location has been an essential issue in many studies. Attempts from different perspectives have been made to solve this problem. Some of them employ a model based on frequent trajectory patterns (Yavas et al., 2005; Monreale et al., 2009). In Yavas et al.'s work, the authors proposed an algorithm to predict the next intercell movement of a mobile user in a mobile computing system (Yavas et al,, 2005). The algorithm follows three phases: first, they extract user mobility patterns from the history of mobile users; second, association rules are mined from these patterns; finally, a mobile user's next location is predicted by using these rules. The notion of support is used for the selection of rules. In Monreale et al.'s work, they also use the movement patterns, which is extracted by employing the Trajectory Pattern algorithm proposed by Giannotti et al. in 2006 (Giannotti et al., 2006; Monreale et al., 2009). Then, a T-pattern Tree is built to capture the global model of the underlying mobility data, which is similar to the use of association rules as predictive rules. Their idea for prediction is to compute the best matching score of all the candidate paths of the T-patterns Tree with children of the best node to be selected as next possible locations. The algorithm to find the best matching pattern computes the matching score by introducing the concepts of punctual score and path score. A similar idea was also proposed by Jeung et al. earlier (Jeung et al., 2008). A Hybrid Prediction Algorithm (HPA) presented by them provides predictions for both near and distant time by combining the predefined motion functions with movement behavior patterns.

Ying et al. present an innovative approach for predicting future locations from the perspective of semantic trajectory (Ying et al., 2011). By taking both geographic and semantic features of users' trajectories into account, they propose the location prediction framework called SemanPredict. Furthermore, instead of detecting frequent patterns among the whole dataset, they give a cluster-based prediction strategy based on the frequent behaviors of similar users in the same cluster for the purpose of predicting a mobile user's next location.

In addition, the extended Kalman Filter(EKF) was also proposed for state estimation and trajectory prediction. In Prévost et al.'s work, the authors presented an EKF approach for state estimation and trajectory prediction of a moving object detected by an unmanned aerial vehicle (Prévost et al., 2007). The filter is used to estimate its true setpoints, model states, model outputs and position in space, in spite of model uncertainty and measurement uncertainty. They assume the object is a non-linear controlled system with Gaussian noise and the dynamic characteristics of the system have been already known.

2.3 Mixed Markov Model

Mixture Markov Model (MMM) is defined as a finite mixture of Markov chains, allowing for individual differences in transition probabilities. For example, it considers the heterogeneity of customer behaviors in a market. Each chain represents a group of sequences sharing similar characteristics. Each sequence is generated by one of the chains with some probability, denoted by a latent variable, X. $X_k = 1$ when the kth model generates the sequence; $X_k=0$, otherwise. MMM has been employed for years in the analysis of customer behavior. Poulsen applied MMM to model the brand choice behavior of customers (Poulsen, 1990). Not limited to marketing, MMM has also been employed in other applications. For instance, by learning a mixture of first-order Markov models, Cadez et al. discover navigation patterns of users on a Web Site (Cadez et al., 2000). Asahara et al. applied MMM to spatio-temporal data, i.e., trajectories, to detect the underlying behavior patterns of customers in a commercial shopping complex for the purpose of evaluation of levels of customer satisfaction (Asahara et al, 2011). An EM algorithm is usually applied to optimize the parameters of MMM by an iterative procedure.

Chapter 3

Methodology

Trajectory mining techniques have been extensively developed and employed to study people's trajectories. In order to predict next locations in people's trajectories, it is essential to understand the underlying patterns in them. A great challenge in the data set we had access to was that an individual's identity is anonymous in the trajectories (on any given day, an individual might pick up any of the RFID tags, with this information unrecorded). So, we are not able to speculate their future trajectories based on their past trajectory histories. But people may share the same pattern during a particular time of the day. For example, two colleagues go to have lunch from their office every noon during the workday. Likewise, the trajectories of an individual person may carry several patterns according to the person's schedule. Thus, we focus on the trajectory patterns among all the (a training set of) input trajectories. MMM is employed in this thesis to detect the different patterns in these trajectories using a model-based clustering method. We chose it because it considers the trajectory patterns as latent variables. This model-based clustering method can directly characterize parameters of each cluster in terms of its state transition matrix, as well as the cluster's prior probability (essentially, its frequency of use). Our methodology follows three steps. First, we cluster all the two-dimensional spatial measurements into "states". And, we represent the original x, y coordinated trajectories as a sequences of clusters (states). Second, the Mixed Markov Model is employed to build the model for state transitions of each unique pattern. The EM algorithm is used for training the model parameters in an iterative way. Finally, prediction is made according to state transition probabilities of the mixing components in the operational phase, applied to (test) trajectories that are temporally "revealed", with the predictions for a given trajectory at each time point based on the trajectory's causal subsequence, relative to the current time point. Besides our basic model, we also exploit time and location information as conditioning context for the state transitions in the second step. Also, a model order selection method, Bayesian Information Criterion (BIC), helps to determine the number of latent trajectory patterns.

3.1 Track Representation: Sequence of States

In this first step of the methodology, the input trajectories, given as a sequence of x-y coordinated points with time stamps, are converted into a sequence of states. Here, state means point of interest. This step endows semantic meaning to the trajectories and helps in the comprehension of the trajectory data, serving as an essentially important preprocessing step for the following pattern extraction and further prediction of future locations of the users.

Alvares et al. proposed a general model for enriching trajectory with semantic meaning (Alvares et al., 2007). They presented an algorithm to extract "stop"s and "move"s for the trajectories, where stops denote important places where people stay for a relatively long time (e.g., hotels, parks, etc.) and moves are sub-trajectories between two consecutive stops, by integrating trajectory data with the background geographic information.

Borrowing the concept of "stop"s and "move"s, we propose a method to extract the points of interest, i.e., states, without the need of knowing the background geographic information. We assume people involved in activities in a room should exhibit dense trajectory points in this local area; on the other hand, people should have sparse points in the sub-trajectory when they move from room to room and the shape is line-like. So, the trajectories can be segmented into several parts. The segments are primarily composed of two main categories: points of interest (stops) and sub-routes (moves). In indoor space, for example, sub-routes can be paths from a meeting room to an office, and points of interest can be an office and a meeting room. A density-based DBSCAN clustering method (or, in fact, any clustering method) can in principle separate the stops from moves for trajectories. Specifically, our method for extracting states from trajectory data follows three steps:

(1) We partition the trajectory into segments, "stop"(in-room) and "move" (room-to-room) by applying a clustering method to extract the dense-regions as candidate stops;

(2) We cluster the candidate "stop" segments to detect points of interest for all trajectories;

(3) The input trajectories are thus converted to sequences of points of interest.

There are two potential forms for representation. Either, the track is represented as a sequence of states; alternatively, by a sequence of unique cluster transitions, plus the run-length (i.e., how long the user stays in a cluster). The first representation of a discrete sequence of states is given as

$X^{(n)} = (X1^{(n)}, X_2^{(n),} ..., \ X_{Tn}^{(n)})$, n=1, 2,…, N

where N is the number of measurements (points) in a trajectory; $X_i^{(n)} \in \{1, 2, \ldots, M\}$, which is the set of state indices. Actually, the two presentations are equal and can be converted to each other since the sampling rate is almost uniform. We use the first representation in this thesis. *The above presents a general method to extract the points of interest for transforming coordinated location points into sequences of states. However, in our indoor space case, the sampling rate is about two minutes, which allows users to move from one location to another between two of our data measurement points. Thus, the "moves" are not well-captured in the experiments of our indoor trajectory data.*

3.2 Mixed Markov Model

Mixed Markov chain is proposed to deal with the heterogeneity in human behaviors in order to overcome the assumption of homogeneity in a traditional Markov chain, which is not able to capture different user behaviors in the data. The Mixed Markov chain is more appropriate to describe real-world cases where human behaviors or reactions vary in response to even the same offering or input. Thus, MMM has been employed in many domains, including marketing, psychology, biology, etc. for discovering unique user clusters, within which users share some similarities, and with clear differences in behaviors exhibited across clusters. In [Poulsen, 1990], the author applied Mixed Markov chain to analyze brand choice behaviors in the market. In our case, for indoor trajectory data, we denote each chain representing a group of trajectories that carry similar characteristics in state transitions. Each trajectory is generated by one of the chains in probability. Z denotes the latent variable which represents one unique cluster, namely, $Z_k = 1$ means the kth cluster generates the trajectory; $Z_k = 0$, otherwise.

Let X = {X(n), n=1, 2,…, N} be the training dataset for all input trajectories. We model the trajectory data by a mixture of first order Markov Models, i.e.,

$$\mathrm{P[x]} = \sum_{l=1}^{K} \alpha_l \rho[x/\theta_l] \tag{3.1}$$

where $P[x/\theta_l]$ is the mixture density (joint likelihood), given by

$$\prod_{i=1}^{T_n} \rho[x_i/x_{i-1}; l] \tag{3.2}$$

Thus, P[x] can also be written as

$$\sum_{n=1}^{N} \log(\sum_{l=1}^{K} \alpha_l (\pi_{x0/l} \prod_{i=1}^{T_n} \rho[x_i/x_{i-1}; l])) \tag{3.3}$$

where K is the number of mixture components; α is mixing coefficient; π is the probability of component l starting with state x0; ρ is the transition probability conditioned on component l; and $0 \leq \alpha_l \leq 1, \sum_{l=1}^{K} \alpha_l = 1$.

The training objective is to maximize the training set log-likelihood

$$\log P[X] = \sum_{n=1}^{N} \log(\sum_{l=1}^{K} \alpha_l (\pi_{x0/l} \prod_{i=1}^{T} \rho[x_i/x_{i-1}; l])) \tag{3.4}$$

An Expectation-Maximization(EM) algorithm is used to estimate the parameters following the steps below:

Step 1: Initialization

$$\alpha_l^{(0)} = \frac{1}{K}, \forall l \tag{3.5}$$

$$\pi_{x_0/l}^{(0)} = \frac{1}{M}, \forall l \tag{3.6}$$

$$P[x_i = j/x_{i-1} = k; l]^{(0)} = \frac{N(j,k,l)}{N(k,l)} \tag{3.7}$$

where N is the number of times a joint event occurs.

Step 2: E-step

$$P[L = l/x^{(n)}]^{(t+1)} = \frac{\alpha_l^{(t)} \rho[x^{[n]}/\theta_l^{(t)}]}{\sum_{l=1}^{K} \alpha_l^{(t)} \rho[x^{[n]}/\theta_l^{(t)}]}, \text{ l=1,2,...K; n=1,2,...,N.} \tag{3.8}$$

Step 3: M-step

$$\alpha_l^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} P[L = l/x^{(n)}]^{(t+1)}, \text{ l=1,2,...,K} \tag{3.9}$$

$$\pi_{k/l}^{t+1} = \frac{\sum_{n=1:x_0^{(n)}=k}^{N} P[L = l/x^{(n)}]^{(t+1)}}{\sum_{n=1}^{N} P[L = l/x^{(n)}]^{(t+1)}}, \text{ k=1,2,...,M; l=1,2,...,K} \tag{3.10}$$

$$P[x_m = j/x_{m-1} = k; l]^{(t+1)} = \frac{\sum_{n=1}^{N} \sum_{i=1:x_i^{(n)}=j, x_{i-1}^{(n)}=k}^{T_n} P[L = l/x^{(n)}]^{(t+1)}}{\sum_{n=1}^{N} \sum_{i=1:x_{i-1}^{(n)}=k}^{T_n} P[L = l/x^{(n)}]^{(t+1)}} \tag{3.11}$$

Step 4: go to step 2 unless converged.

This EM algorithm monotonically ascends in the training set log-likelihood objective, and is thus finds a locally optimal solution. However, as the log-likelihood objective is a non-convex function of the model parameters, only local optimality is ensured – the global maximum likelihood solution is not ensured to be found by this method.

After apply the EM algorithm, the trajectories are clustered into groups with common latent variables, which thus indicate common patterns across the training trajectories.

However, we lose a lot of information regarding state transition when we represent the original x-y coordinated points into states in the first step. For example, if an individual is on the way to B from A, it is very likely that he will transfer from state A to state B. Thus, the location information that we reduced in the first step (by clustering the spatial measurements) may still be able to help us predict state transitions. We consider this information a "hint". On the other hand, we may also get a transition hint from temporal information. It is reasonable to assume people exhibit different state transition probabilities during different times of the day. For example, at 12pm in the noon, it is more likely for a person to transfer from office to dining room; while at 5 pm, he should have high probability to go home from his office. Thus, we also evaluated embedding the spatial and temporal information in the MMM (as conditioning context) in order to improve the prediction accuracy of state transitions.

Time-embedded MMM

In order to make use of the information of time stamp for each location point, we embed temporal factor into the MMM. The idea is pretty straightforward: we make the state transition probability condition on time. Specifically, the method follows two steps. First, we quantized the time stamps into a discrete set of times (e.g., within one hour). Second, a dimension of temporal factor was added to the state transition probability matrix, which is expressed as

$$P[S_t = k / S_{t-1} = j; l, T_{n-1}]$$

where T_{n-1} denotes the time of the $(t-1)_{th}$ state.

The E-M step for training the parameters can be modified in the same way.

Location-embedded MMM

We quantified the trend for a user to go to one state from another in order to embed location hint for their state transitions. Conditioned on the previous two location points, the state transition probability is given as

$$P[S_t = k / S_{t-1} = l, x_{t-1}, x_{t-2}] = \frac{e^{\gamma_{kl} + \alpha_1(\|C_k - x_{t-1}\|^2) + \alpha_2(\|C_k - x_{t-2}\|^2 - \|C_k - x_{t-1}\|^2)}}{\sum_k e^{\gamma_{kl} + \alpha_1(\|C_k - x_{t-1}\|^2) + \alpha_2(\|C_k - x_{t-2}\|^2 - \|C_k - x_{t-1}\|^2)}} \quad (3.12)$$

where C_k is the center of state k; $\|C_k - x_{t-1}\|^2$ indicates the absolute distance between current location and center of state k; $(\|C_k - x_{t-2}\|^2 - \|C_k - x_{t-1}\|^2)$ is dynamic distance to state k considering previous two locations, which suggests the direction of going to state k.

Another simpler way to consider the location hint is to quantize the 2D measurement space in the same way as counting the time hint, and condition on which spatial cell the trajectory is in at times t-1 and t, in predicting the state at time t+1. Similarly, the transition probability conditioned on pattern l, spatial cells at time t-1 and t, is given as

$$P[S_t = k / S_{t-1} = j; l, C_{t-1,} C_t]$$

where C_{t-1} and C_t denote indices of spatial cells at time t-1 and t.

3.3 Prediction of Next Point of Interest

With the parameters of MMM trained, we can make prediction for users' next point of interest. Given the current state k, the predicted probability of going to state j can be computed as

$$P[x_i = j / x_{i-1} = k] = \sum_l P[L = l / X^{(n)}] \bullet P[x_i = j / x_{i-1} = k; l] \quad (3.13)$$

where $P[L=l/X^{(n)}]$ is the probability of the coming trajectory $X^{(n)}$ being cluster l, which can be computed using equation (3.8) in the E-step of the EM algorithm.

3.4 Model Order Selection: BIC

In order to determine the number of latent components (clusters) in the MMM, we employ BIC for model order selection. BIC for choosing the number of components is given as

$$BIC(K) = N_{Free}(K) \cdot \frac{1}{2} \log_2(N) - \log_2 P[x \mid \theta^{(k)}] \tag{3.14}$$

where $N_{Free}(K)$ is the number of free parameters in the model.

In our case,

$$N_{Free}(K) = K \cdot ((K-1) + (M-1) + M(M-1)) \tag{3.15}$$

where K-1 is for α, M-1 is for π, M(M-1) is for state transition probability matrix.

$$K^* = \arg_K \min BIC(K) \tag{3.16}$$

- X in BIC is the set of all training trajectories.

Chapter 4

Experiments and Results

The data source we employ to do the experiments is provided by Buzby Networks, LLC, collected by the BuzNet Real-Time Locating System (RTLS) which monitors people's real-time indoor positions in Penn State Learning Factory (http://www.buzbynetworks.com/). A total number of twelve battery-power wireless tags are used for tracking people's locations over time. The received data contains information of all 12 tags from January to April, 2012, which shows people's locations sampled approximately every two minutes every day for those three months. The table to store such spatio-temporal data is comprised of three columns, Time_Stamp (time stamp), Message (location) and Destination_EUI (tag number): Time_Stamp indicates the time point of the corresponding location in the same row, e.g., 2012-01-21 20:39:50.000, Message describes the 3D (x,y,z) location of people in that moment, e.g., 18.70,15.20,Floor 1, Destination_EUI records the number of the tag which reports the spatio-time information. There are 12 different number, each identifying one of the 12 tags. The original data format we received is .bak file. We extended and restored the data in SQL Server 2008. The original data schema is clear and succinct in recording the time and location information of people's trajectories. A user picked up one tag from the tag container when his/her trajectory was starting to be monitored and he/she put it back in the container before leaving the building of the learning factory. Thus, the trajectories are round-trip, starting from the location of tag contain (18.6, 11.8) and ending in the same place. Also, as we know, most of the time, the tags lie in the tag container. So, we filtered these points in the tag container. And we also select the data points with time stamps from 8am to 10pm, which is the open time of Learning Factory. We pre-processed a total number of 1438 trajectories from the data of three months. There are totally 36356 measurements (points). We divide the data into two parts for training and testing. The training data set is 1091 tracks for two months (1/20/12~3/20/12); and, the testing data set is 347 tracks for one month (3/21/12~/4/20/12).

The experiment is done following the steps in the methodology: (1) we cluster the point-based locations into states; and transfer trajectories made of the x-y coordinated points into sequences of states. (2) We build the MMM model and employ EM algorithm to train and BIC to select the models. (3) Next point of interest is made by computing how it belongs to each of the patterns, and then the transition probability is based on the transition probability of all patterns. The results of experiments of extended MMM conditioned on time and location is also presented.

State definition

In this indoor setting, we simplified the extraction of places of interest by employing k-means to extract regions of interest in the indoor space. As Fig. 4-1 shows, we get five clusters. Each color denotes a cluster: red, shop room; pink, part #1 of work room; green, part #2 of work room; yellow, lab; blue, lobby. Choosing these five clusters is visually plausible since the clusters correspond to different rooms (most of them). Besides, Bayesian Information Criterion (BIC) could be used in practice for choosing the number of clusters, which can help make the local decisions about whether the current centroids should be split.

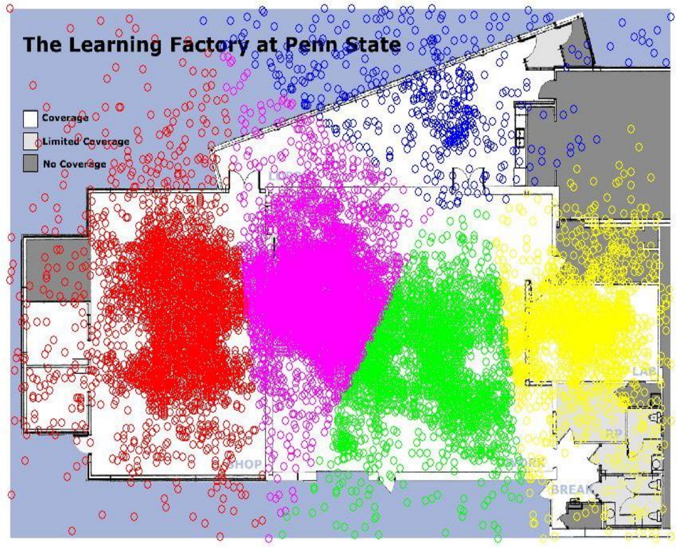


Figure 4-1 Five Regions of Interest for Learning Factory.

MMM model Building

Next, we build the MMM on the training data for detecting the underlying trajectory patterns. MMM considers the heterogeneity in human behaviors as latent patterns. The objective is to maximize the log likelihood. As Fig. 4-2 shows, the log-likelihood is strictly increasing versus iteration times by EM algorithm.

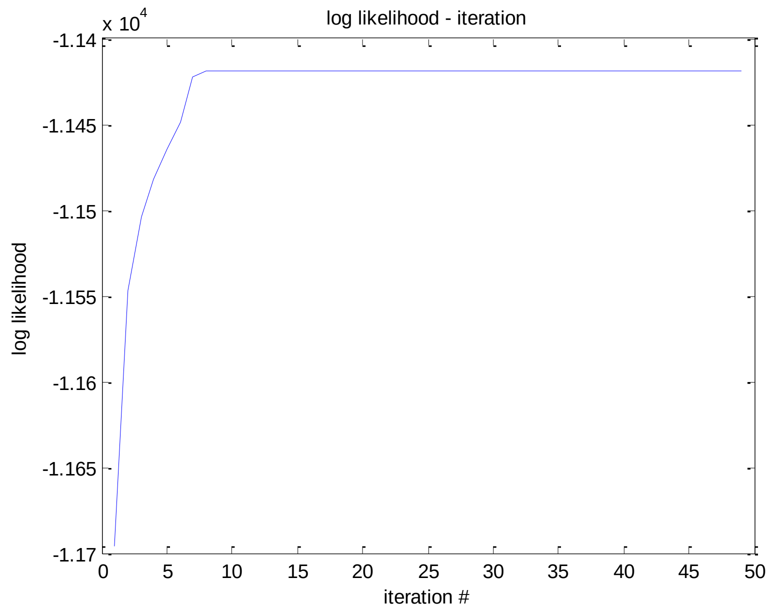


Figure 4-2 Log-likelihood Vs. Iteration.

The total prediction accuracy we get is 0.9155. And the prediction accuracy for each step is presented in Fig. 4-3.

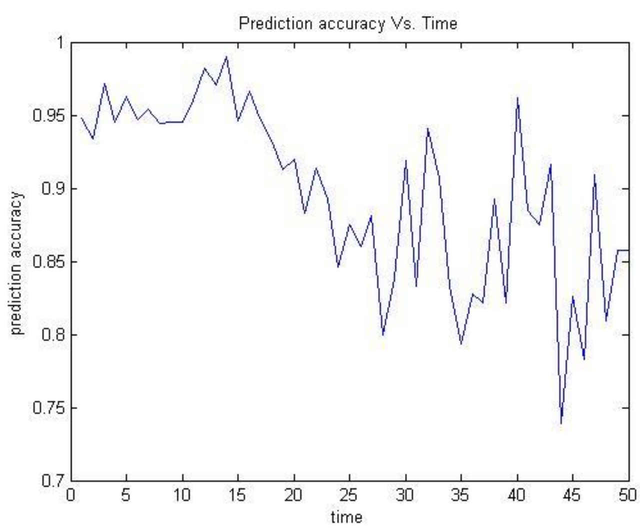


Figure 4-3 Prediction Accuracy for Each Step.

As we see, the prediction accuracy is high in the first few steps. This is because users usually stay in the initial state for a while. Then, the curve drops since states start to transition and the ongoing state varies -- it is hard to tell which one the user is going to without enough information about the coming trajectory. Finally, the curve goes up and oscillates without dropping too much. This suggests the prediction accuracy increases as more information of the coming trajectory is available. (Also, the ending state is known – it is the same as the starting state).

Model Order Selection: BIC

The number of components is determined by BIC. As Fig. 4-4 shows, the minimum point is when # of components equals 4. It suggests there are four unique trajectory patterns across all the users.

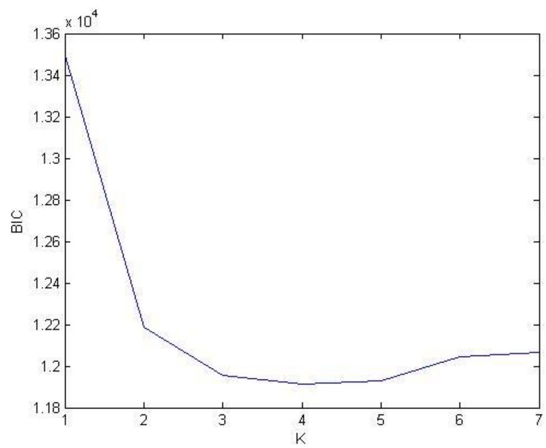


Figure 4-4 BIC for Determining # of Latent Variable K

In order to observe the four components, we present the clustering of all training trajectories into four patterns detected as show in Fig. 4-5.

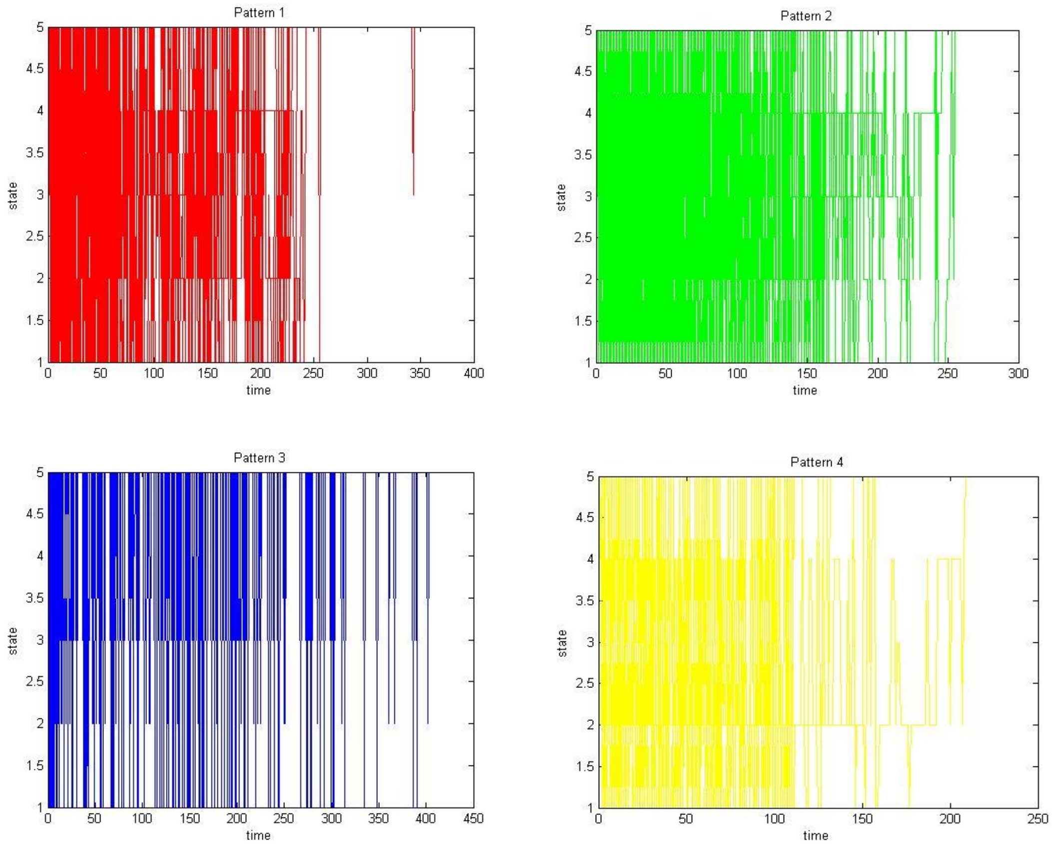


Figure 4-5 Clusters of Training Trajectories for Each Pattern

The transition probabilities for each component are presented as follows.

Probj_k_l(:,:,1) =

0.7024	0.1347	0.1371	0.0732	0.0572
0.0605	0.3905	0.0947	0.0941	0.0486
0.0325	0.0364	0.5103	0.0321	0.0129
0.0432	0.1220	0.0804	0.7300	0.0137
0.1613	0.3164	0.1775	0.0706	0.8675

Probj_k_l(:,:,2) =

0.7034	0.1510	0.0622	0.0456	0.2681
0.0779	0.4580	0.0969	0.1044	0.2132
0.0200	0.0237	0.6718	0.0259	0.0427
0.0393	0.1327	0.0801	0.7562	0.0637
0.1594	0.2346	0.0890	0.0679	0.4123

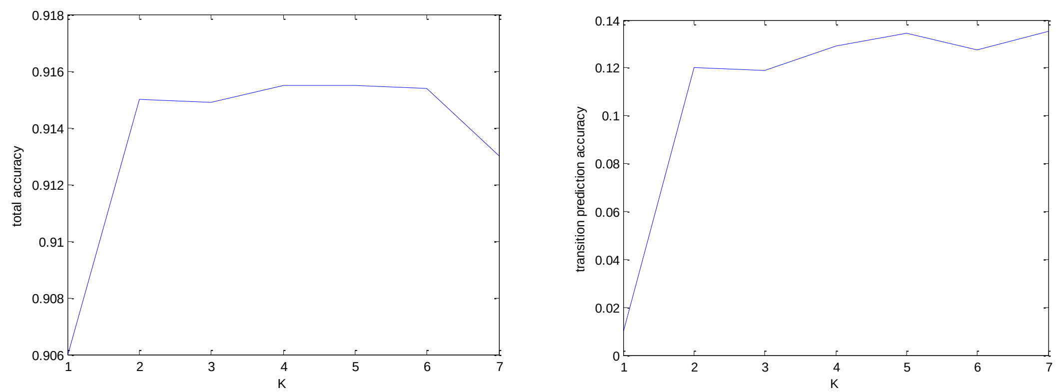
Probj_k_l(:,:,3) =

0.0278	0.0204	0.0749	0.0000	0.0077
0.0000	0.0123	0.0341	0.1443	0.0052
0.0000	0	0	0	0.0007
0	0	0	0	0.0004
0.9722	0.9673	0.8909	0.8557	0.9859

Probj_k_l(:,:,4) =

0.8416	0.0628	0.1115	0.0284	0.2261
0.0680	0.8110	0.3643	0.3952	0.3080
0.0041	0.0077	0.3353	0.0326	0.0041
0.0143	0.0541	0.1515	0.5180	0.0079
0.0719	0.0644	0.0373	0.0258	0.4539

Also, it is interesting to know how the prediction accuracy varies depending on the number of components. Figure 4-6 (a) shows the relation between total accuracy and number of components. The largest accuracy achieves at K=4 or 5, which matches the result of model order selection. Figure 4-6 (b) shows the relation between transition prediction accuracy and number of components. The transition accuracy is still smoothly increasing after K=4. Thus, BIC is more likely to select the optimum number of components to maximize the total accuracy.



(a) (b)

Figure 4-6 (a) Total Prediction Accuracy Vs. # of components (b) Transition Prediction Accuracy Vs. # of components

Sample predicted tracks

Several predicted trajectories are shown in Fig. 4-7 (a)-(d) to know how the prediction algorithm works. The results suggest that (1) prediction is good when the state is stable (in Fig. 4-7 (a)); (2) the prediction accuracy increases when more track history is available (in Fig. 4-7 (b)); (3) sharp changes are hard to predict; while general changes can be predicted (in Fig. 4-7 (c)); (4) bad prediction when the trajectory changes a lot in early times (in Fig. 4-7 (d)).

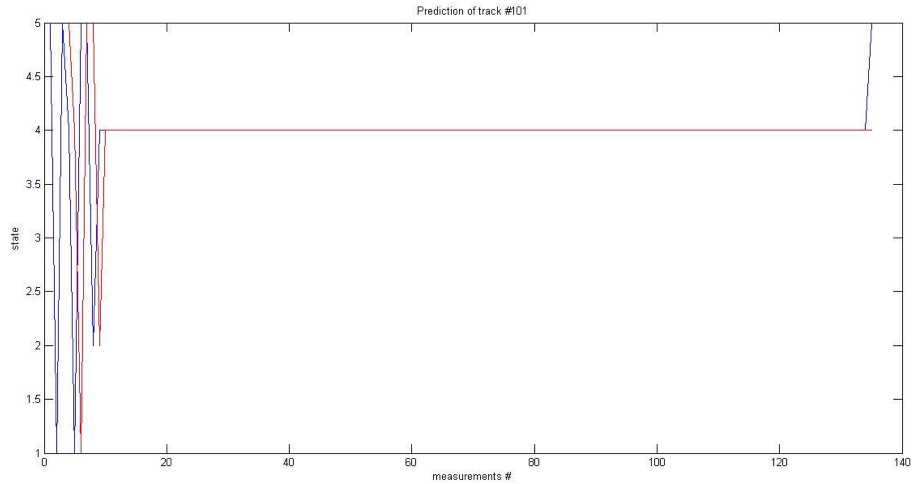


Figure 4-7 (a) Prediction of Test Track #101

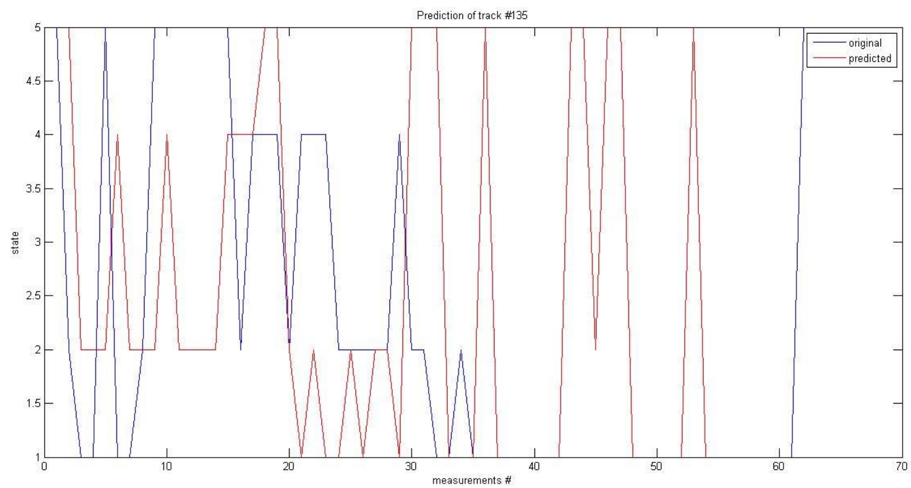


Figure 4-7 (b) Prediction of Test Track #135

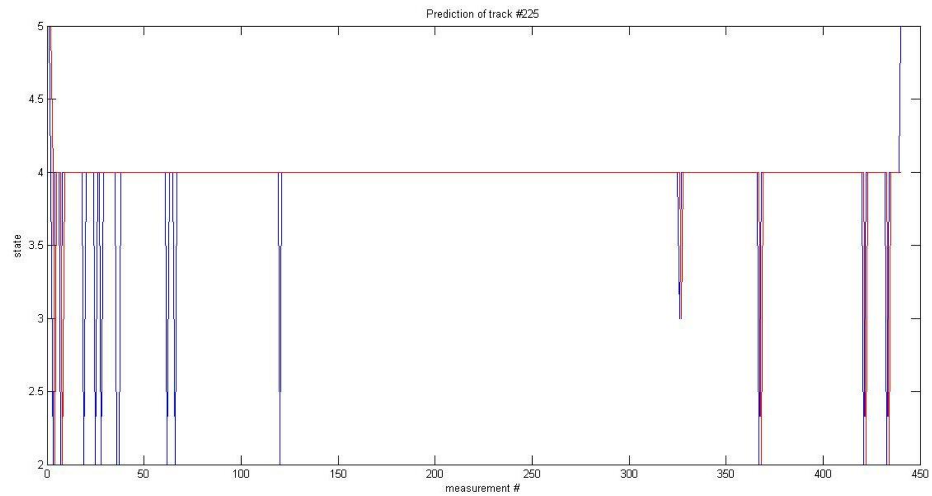


Figure 4-7 (c) Prediction of Test Track #225

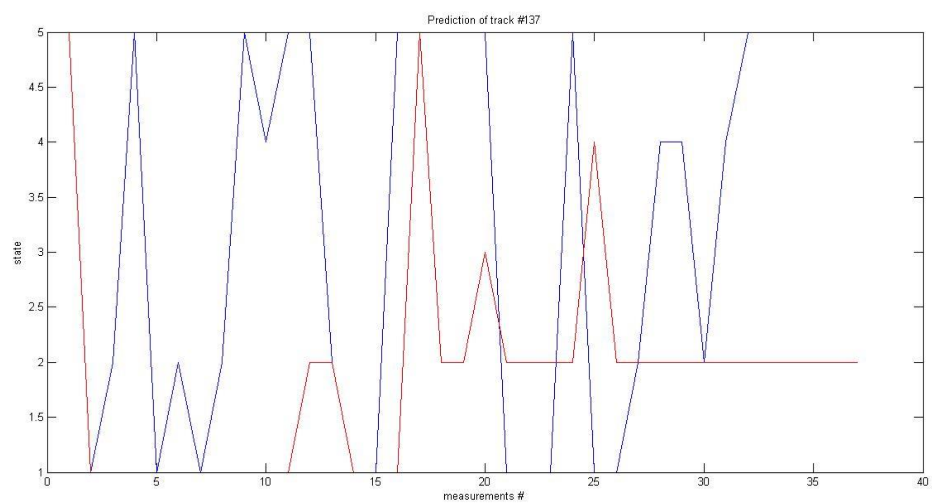


Figure 4-7 (d) Prediction of Test Track #137

In general, our algorithm learned from the training trajectories that when a user enters into a state, he/she will stay in the stable state there for a time. So, the prediction is pretty good when the state is stable; when it comes to the state transition, the algorithm needs to learn more information about the history of the trajectory to detect its pattern before it can make correct predictions. Otherwise, the prediction does not work well in the initial stage when there are a lot state transitions.

The transition accuracy of basic MMM is 0.1293. Thus, we embed time and location information in order to improve the prediction in state transitions.

Time-embedded Prediction

Time is embedded by quantifying the time stamp on a hourly basis in order to catch the potential transition hint based on time. The transition accuracy of time-embedded MMM is 0.1234 and the total accuracy is 0.9013. From the result, we cannot see effective improvement in both transition prediction accuracy and total prediction accuracy. Two reasons may cause this. First, the periodicity is more likely to be observed in the trajectory history of individual users, while it is not obvious among all the input trajectories across the users since different users have different schedules. Second, the contextual information regarding how data was collected may explain the reason. The data is collected from employees in Penn State Learning Factory, who randomly picked up one tag for recording their trajectories, which suggest, the trajectories may not be complete. Also, for most of the time, the tags are not worn, which means the effective trajectories data is a small portion.

Location-embedded Prediction

We implement the "location hint" embedded prediction by defining a 2-dim spatial quantification. Specifically, we quantified the 1490x1000 gridded layout into 620 cells, with an average of 50 points in a cell. The prediction accuracy of state transition has risen to be 0.3306, with the predictable rate (namely, the fraction of transitions with plausible hints that are predicted by the location embedded MMM) of the whole test data to be 0.3687. The total prediction accuracy is 0.89. And the false state transition prediction rate (i.e., how often a state transition is predicted that does not occur) is 0.6513. The false state transition prediction rate is a little bit high, suggesting the location-embedded prediction is a little bit overreacted. It is reasonable since the workers in Learning Factory always walk around during their work, especially in the “work” room, which takes a large portion of the data points. Thus, it generates false signs for state transitions.

Comparison of the performance of three models

We compute the performance measure for the three models, denominated by the number of predictable state transitions. We get 0.35, 0.33 and 0.89, respectively for MMM, time-embedded MMM and location-embedded MMM. The chart below shows the comparison of the performance of the three models, in terms of performance measure, transition prediction accuracy and total prediction accuracy.

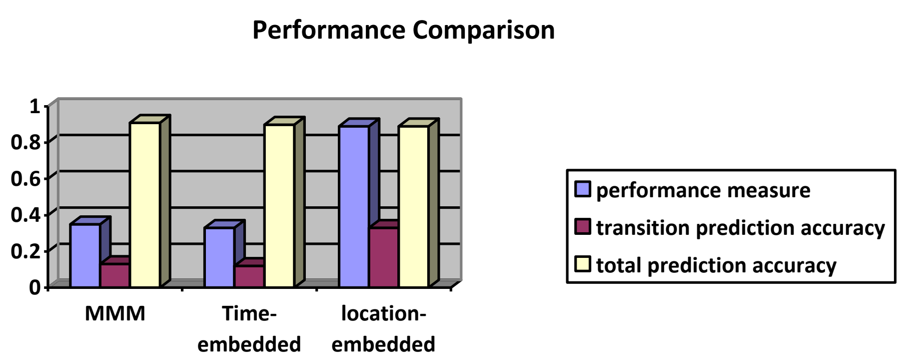


Figure 4-8 Comparison of the Performance of Three Models

The performance measure of Fig. 4-8 shows that the location-based MMM presents remarkable increase in prediction accuracy of state transitions. It suggests that the “location hint” is recognized as a strong sign for state transitions.

Chapter 5

Discussion

In this thesis, we focused on prediction of users' next point of interest in the trajectory data of anonymous user identification. Several attempts have been made by researchers before. The methodologies can be generalized into two kinds: using the user's trajectory history (Yavas et al., 2005) and learning from the patterns in all the input trajectories (cluster into patterns in nature) (Monreale et al., 2009). Our methodology belongs to the latter since the user identity is not known. By employing the MMM, we detected four trajectory patterns within the data. Using the parameters trained by EM, we get a total prediction accuracy of 0.9133 depending on 5 states (points of interest) in Penn State Learning factories. The step prediction (Fig. 4-3) shows that the prediction accuracy of first few steps is high on account of the states being relatively stable; then the accuracy drops as states transitions happen and users' states vary; however, as Fig. 4-3 tells the accuracy goes up again and oscillates, which suggests, when more trajectory information is available, our prediction algorithm detects some state changes as well as predicting the stable ones. It is consistent with the predictions of sample trajectories: the stable trajectory maintains high prediction rates (Fig. 4-7(a)); the algorithm can detect general changes (Fig. 4-7(b)(c)) but does not work well when there are a lot of changes in the first few states(Fig. 4-7(b)(d)). In Monreale et al.'s work, their prediction rate is no less than 20% (Monreale et al., 2009). And our results shows higher prediction rate, up to 90%. It suggests the prediction based on pattern detection is being validated. Our experimental results outperform that of Monreale et al.'s. This may be caused by the data set we employed; or the patterns in this data set are simple and small in number; or trajectory patterns are not fully detected by applying the Trajectory Pattern algorithm to insufficient trajectory data in their work. But the performance of our prediction algorithm should also count.

In general, the MMM for prediction has three unique characteristics compared with other pattern-detection based location prediction: (1) The patterns are considered as latent variables in MMM and the model is learned with no need of prior pattern input. (2) It predicts not only state transitions, but also determines whether it will stay in the current state, further, suggesting how long it will stay in the state. (3) MMM model is simple in computing and easy to interpret. Whilst the results for prediction is good, but the prediction rate is low as 10% when it comes to state transitions. In order to overcome this problem, we make use of "time hint" and "location hint". However, the MMM embedded with "time hint" does not rise up to our expectation. It may be caused by several reasons: (1) there is no obvious periodicity in the data among trajectory data across the users and due the incompletion of user trajectories; (2) the initialization of EM for training MMM counts; (3) in terms of applying BIC to choose the number of patterns, the parameters maybe inter-dependent when the time dimension is added. When "location hint" is embedded, the transition prediction is significantly improved to be 0.33. Thus, location information is proved to be plausible hints for state transition in this indoor context. However, we should also be careful about the trade-offs of the location-embedded MMM. While it can suggest signs of state transition according to two previous points, it also causes "overreact", which false predicts a state transition.

Also, the prediction results depend on places of interest. It is reasonable that a more accurate prediction can be achieved when we reduce the number of clusters. Since this algorithm can be applied to other domains, the definition of clusters can be determined by the users of this methodology so that they can select the points of interest according to their specific needs, making it as a user input for domain applications.

The MMM we proposed in this thesis solves the issue of prediction of points of interest. Due to performance that showed in results, it is an effective methodology to solve the problem. It fits to the literature of this problem for the three unique characteristics discussed above. The significance of solving this problem also lies in that it can be applied to many domains where user preference is diverse and privacy (anonymous user identification) is highly valued. Potential contribution can be: (1) In shopping malls. The manager can know the patterns from customer flow, and predict the future customer flows, so that he can inform the store owners to prepare their service well in advance. (2) In health care center. The stuff can deal with the wandering case more efficiently by knowing what the next area would be for the patient so that tragedies may be avoided by fast actions. (3) Serial criminal or terrorism. It can help the police predict and identify where will happen the next criminal and do a better job in guaranteeing public security.

The future work may be: (1) to further improve the accuracy of state transitions by embedding other information which can be a hint to the state transitions. (2) to discover the correlation between the trajectories of individuals, and explore the underlying social networks within the anonymous trajectory data set. One interesting research question can also be to un-anonymize user identification.

Chapter 6

Conclusion

We introduced a new methodology to predict the next point of interest of a moving object. The idea of prediction is based on all the input trajectories with anonymous user identifications. By applying MMM, we modeled the latent patterns with EM for parameter training. The results of our experiment demonstrated that this methodology gives significant prediction accuracy and allow users to define the points of interest according to their own needs. Besides, the MMM is easy to compute and interpret. And the model can determine the patterns without prior input patterns. It can also tell whether it will stay in the current state as well as the probability of transit to another state. Besides, the location-based MMM exhibits significant performance in the prediction of state transitions.

References

[1] C. S. Poulsen, “Mixed Markov and latent Markov modelling applied to brand choice behaviour,” International Journal of Research in Marketing, vol. 7, no. 1, pp. 5–19, 1990.

[2] Y. Bengio, “Markovian Models for Sequential Data,” Neural Computing Surveys, vol. 2, no. 1049, pp. 129–162, 1999.

[3] P. Smyth, “Clustering sequences with hidden Markov models,” Advances in neural information processing systems, vol. 9, pp. 648–654, 1997.

[4] J. Alon, S. Sclaroff, G. Kollios, and V. Pavlovic, Discovering clusters in motion time-series data, vol. 1. IEEE Comput. Soc, 2003, p. I–375–I–381.

[5] A. Monreale, F. Pinelli, and R. Trasarti, “WhereNext : a Location Predictor on Trajectory Pattern Mining,” Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, vol. 115, no. 3, pp. 637–645, 2009.

[6] S. Gaffney and P. Smyth, “Trajectory clustering with mixtures of regression models,” Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 99, vol. 10, no. 99, pp. 63–72, 1999.

[7] G. Yavas, D. Katsaros, O. Ulusoy, and Y. Manolopoulos, “A data mining approach for location prediction in mobile environments,” Data & Knowledge Engineering, vol. 54, no. 2, pp. 121–146, 2005.

[8] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, “Trajectory pattern mining,” Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining KDD 07, vol. San Jose,, p. 330, 2007.

[9] H. Cao, N. Mamoulis, and D. W. Cheung, “Mining Frequent Spatio-Temporal Sequential Patterns,” Fifth IEEE International Conference on Data Mining ICDM05, no. ii, pp. 82–89, 2005.

[10] L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. De Macedo, B. Moelans, and A. Vaisman, "A model for enriching trajectories with semantic geographical information," Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems GIS 07, no. i, p. 1, 2007.

[11] H. J. H. Jeung, Q. L. Q. Liu, H. T. S. H. T. Shen, and X. Z. X. Zhou, A Hybrid Prediction Model for Moving Objects, vol. 00. Ieee, 2008, pp. 70–79.

[12] S. Zhong and J. Ghosh, "A Unified Framework for Model-based Clustering," Journal of Machine Learning Research, vol. 4, no. 6, pp. 1001–1037, 2003.

[13] D. U. Pfeiffer, P. Q. Minh, V. Martin, M. Epprecht, and M. J. Otte, "An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data.," Veterinary journal London England 1997, vol. 174, no. 2, pp. 302–309, 2007.

[14] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," Proceedings of the 1993 ACM SIGMOD international conference on Management of data SIGMOD 93, vol. 22, no. May, pp. 207–216, 1993.

[15] J. Mennis and J. W. Liu, "Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change," Transactions in GIS, vol. 9, no. 1, pp. 5–17, 2005.

[16] S. Kisilevich, F. Mansmann, M. Nanni, and S. Rinzivillo, "Spatio-Temporal Clustering: a Survey," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. Springer, 2010, pp. 1–22.

[17] M. Chiesa, R. Genz, F. Heubler, K. Mingo, C. Noessel, N. Sopieva, D. Slocombe and J. Tester, RFID (March 2002), http://people.interaction-ivrea.it/c.noessel/RFID/research.html

[18] A. Palma, V. Bogorny, B. Kuijpers, L. Alvares, "A clustering-based approach for discovering interesting places in trajectories", Proceedings of the 2008 ACM Symposium on Applied Computing (SAC), Fortaleza, Ceara, Brazil, pp. 863-868, 2008.

[19] F. Verhein and S. Chawla, “Mining spatio-temporal association rules, sources, sinks, stationary regions and thoroughfares in object mobility databases,” in Database Systems for Advanced Applications, 2006, no. 574, pp. 187–201.

[20] G. L. Andrienko, N. V. Andrienko, “Interactive cluster analysis of diverse types of spatiotemporal data”, SIGKDD Explorations 11(2): 19-28 (2009).

[21] G. Fosca, P. Dino, “Mobility, Data Mining and Privacy: Geographic Knowledge Discovery”. Springer, 2008.

[22] F. Giannotti, M. Nanni, D. Pedreschi, “Efficient Mining of Temporally Annotated Sequences”, SDM 2006.

[23] J. J. Ying, W. Lee, T. Weng, and V. S. Tseng, "Semantic trajectory mining for location prediction", In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 34-43. ACM, 2011.

[24] I. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. Smith, "Visualization of navigation patterns on a web site using model-based clustering", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.

[25] A. Asahara, K. Seto, A. Shimura, and S. Yamakawa, “Creating a virtual-entertainment field in a real commercial shopping complex: customer-trajectory analysis of experimental event “HIKARINO Magical Quest””, Proceedings of the Second International Workshop on Regional Innovative Studies: (IWRIS 2010), p. 91-94, 2011.

[26] C. G. Prevost, A. Desbiens, E. Gagnon, "Extended Kalman Filter for State Estimation and Trajectory Prediction of a Moving Object Detected by an Unmanned Aerial Vehicle", American Control Conference, 2007. ACC '07 , vol., no., pp.1805,1810, 9-13 July 2007.