Reinforcement Learning Framework for Modeling Spatial Sequential Decisions under Uncertainty*

(Extended Abstract)

Truc Viet Le[#], Siyuan Liu[⋆], and Hoong Chuin Lau[#]
[#]School of Information Systems, Singapore Management University
[⋆]Smeal College of Business, Pennsylvania State University
[#]{trucviet.le.2012, hclau}@smu.edu.sg; [⋆]siyuan@psu.edu

ABSTRACT

We consider the problem of trajectory prediction, where a trajectory is an *ordered* sequence of location visits and corresponding timestamps. The problem arises when an agent makes sequential decisions to visit a set of spatial locations of interest. Each location bears a stochastic utility and the agent has a limited budget to spend. Given the agent's observed *partial* trajectory, our goal is to predict the remaining trajectory. We propose a solution framework to the problem considering both the uncertainty of utility and the budget constraint. We use reinforcement learning (RL) to model the underlying decision processes and inverse RL to learn the utility distributions of the locations. We then propose two decision models to make predictions: one is based on long-term optimal planning of RL and another uses myopic heuristics. We finally apply the framework to predict real-world human trajectories and are able to explain the underlying processes of the observed actions.

Keywords

Reinforcement learning; trajectory prediction; stochastic utility; budget constraint; Markov decision process; sequential decisions.

1. INTRODUCTION

How does one decide to visit a set of locations in space? Assuming there are distinct points of interest (POIs), then the act of visiting them has to happen sequentially. We call it *spatial* sequential decision-making. It is reasonable to assume that each location bears a utility (reward) to the decision-maker that would not be fully realized until it is visited. Until then, utilities remain *uncertain* and reflect the agent's prior preferences. When making sequential decisions, a rational agent should also weigh in the long-term costs of visiting each of the locations in order to make an optimal plan, where "costs" here are assumed proportional to physical distances. Hence, answering the question above would require a model of the agent's sequential decisions for selecting locations, whose utilities remain uncertain and costs are dynamic, and weighing in their long-term consequences into the decision-making process.

*This research is supported by the National Research Foundation of Singapore under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority.

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016),*
J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.),
May 9–13, 2016, Singapore.

In many practical situations, the agent typically has a limited amount of resources (e.g., time) to run its plan, which we call *budget*. Such a budget constraint can significantly shape the agent's decision-making process in non-obvious ways. In this paper, we propose a framework based on reinforcement learning [2] to model the agent's spatial sequential decision-making, taking into account the uncertainty of the utilities and the budget constraint. Using the framework, we could discover the underlying processes that drive real-world behaviors such as the conditions for making long-term optimal decisions. Such discoveries give insights into real-world human behaviors and the conditions for rationality that would help bridge the gap between human and machine intelligence [3].

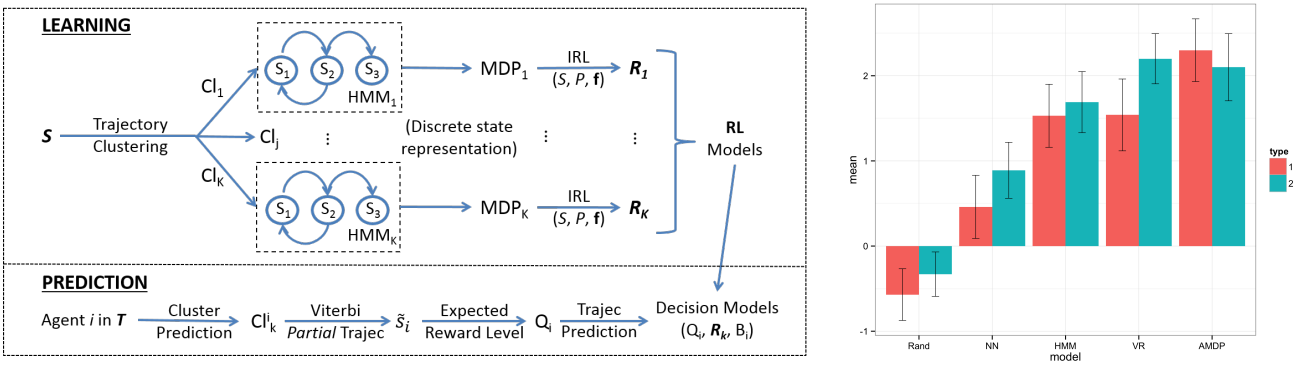
Our motivation comes from the problem of predicting the *next* sequence of location visits (called *trajectory*) of a mobile agent knowing its current trajectory and past observed trajectories of other similar agents. In this paper, we develop on the framework proposed by Le *et al.* [1] for spatial decision modeling. However, we set out to predict an *ordered* sequence of an agent's future locations. Furthermore, our novelty is that we do not rely solely on generative models (e.g., HMMs) to generate sequential actions. Instead, we integrate them into a reinforcement learning framework to model an agent's optimal sequential decision-making. This enables us to explain the underlying processes of the predicted outcomes and the effects of budget constraint on the decision-making.

2. SOLUTION OVERVIEW

We propose an integrated framework to model and predict the next sequence of locations given the observed partial trajectory. The framework consists of two components: learning and prediction. Fig. 1a illustrates the overall proposed framework.

Learning. We first divide agents in the training set $\mathcal{S}$ into K clusters, where each cluster Cl_j ($1 \leq j \leq K$) represents an agent *type*. Using the agents' observed features and the K clusters as class labels, we train a multi-class classifier (e.g., multinomial logistic regression). We also model the environment that the agents interact with as a finite set of states S, where each state $s \in S$ has a distinct vector of features $\mathbf{f}_s$. We use hidden Markov models (HMMs) to transform the observed trajectories into finite sequences of states. Such a representation can then be modeled as a Markov decision process (MDP). The utility of each action (i.e., location visit) can then be derived via the process of inverse reinforcement learning (IRL) [3] using the agents' observed actions (represented in the transition probability matrix P of the MDP). The final outcomes of IRL are the reward matrices $\mathbf{R}$.

Prediction. Given the observed partial trajectory and features of an agent i in the test set $\mathcal{T}$, we first predict i's type Cl_k^i using the trained classifier above. We then use the Viterbi algorithm to find



(a) (b)

Figure 1: (a) The proposed framework to predict the remaining trajectory of test agent $i \in \mathcal{T}$ given observed partial trajectory $\tilde{s}_i$, budget B_i, and trajectories of the agents in the training set $\mathcal{S}$. (b) Similarity measures between the actual and predicted trajectories across the baselines: Random ("Rand"), Nearest Neighbor ("NN"), and HMM; and the proposed models: Value Ratio ("VR") and Adaptive MDP ("AMDP").

the most probable sequence of states $\tilde{s}_i$ for the observed trajectory. We are then able to model i's goal Q_i using the value function and predict the next sequence of visits that can meet this goal within budget B_i. We finally propose two decision models that take into account the uncertainty of the utilities (represented by the matrix $\mathbf{R}_k$ for each type k) and budget B_i. One is based on long-term optimal planning of RL and another uses myopic heuristics.

3. EXPERIMENTS

We collaborated with a large theme park in a major Asian city to conduct experiments and collect data from their visitors from January to April, 2014. The dataset $\mathcal{D}$ contains $3,867$ visitors' trajectories tracked using RFID devices. In the experiments, visitors pay upfront a fixed amount in order to redeem up to 14 attractions (POIs) in the theme park. Visitors can only redeem the attractions during the period from 9 a.m. to 7 p.m. on a chosen day.

We perform 5-fold cross-validation on $\mathcal{D}$. For each fold, the training set $\mathcal{S}$ is used for trajectory clustering and modeling. Our hierarchical clustering results in $K = 2$ clusters. For each agent $i \in T$, let l_i be i's final sequence length. We first predict i's type using its demographic features via a multinomial logistic model. Given i's partial trajectory of length k, we predict i's remaining trajectory while varying $k \in [2, l_i - 1]$. Let s_i^* and $\hat{s}_i$ be i's actual and predicted remaining trajectory, respectively. We use the Levenshtein edit distance to quantify their similarity Each match receives a fixed positive score and each mismatch incurs a negative penalty proportional to the distance between the two locations.

We evaluate two models based on the proposed framework: Adaptive MDP (AMDP), which is based on the long-term optimal planning of MDP, and Value Ratio (VR), which uses a greedy and myopic heuristic to make decisions. The baseline models are the same as in [1]. Our trajectory clustering reveals that the main differences between the two agent types are their temporal behaviors. Agent type 1 tends to arrive earlier and has their peak of visiting activities earlier in the day, and then (their visit frequency) sharply drops off. Whereas, agent type 2 tends to arrive much later and reaches their peak later, and then gradually declines. As a result, we call agent type 1 the "early birds" and agent type 2 the "latecomers".

Fig. 1b summarizes the experimental results, which shows the distributions of the similarity measures (i.e., the means and the 95% confidence bars) across the models. A higher mean similarity implies a more accurate prediction, on average. For type 1, it shows that the AMDP model has the most accurate prediction. The VR and HMM model both have about the same second best average prediction score. The Random baseline model has the least accurate average prediction followed by the NN model. For type 2, the AMDP model performs marginally worse than VR, even though still faring much better than the other baselines. In other words, the VR model makes the most accurate average prediction.

Because type 1 are the early birds and agent type 2 are the latecomers, agent type 1 has a much larger budget than agent type 2. Larger budget means more foresight and better long-term planning, which is what the AMDP model reflects: it embodies the long-term optimal policy of reinforcement learning. This indeed performs better than other short-sighted baselines. On the other hand, a smaller budget, which agent type 2 has, translates into less time for careful planning, which ultimately results in more myopic and suboptimal decisions (i.e., resorting to greedy strategies). This is reflected in the experimental results, where the greedy and myopic Value Ratio (VR) model performs the best for agent type 2.

4. CONCLUSION

In this paper, we address the problem of trajectory prediction using reinforcement learning to model the agent's sequential decisions. By doing so, we have discovered from real-world trajectories how people make decisions: they make more optimal decisions when given enough time to do so. This is perhaps not surprising in retrospect, because it is reasonable that foresighted decisions and careful plans need time to coordinate, while myopic ones do not (as only the immediate rewards are considered). On the other hand, this also validates our framework's ability to model real-world behaviors by finding out what makes reasonable sense in real life.

REFERENCES

[1] T. V. Le, S. Liu, H. C. Lau, and R. Krishnan. Predicting bundles of spatial locations from learning revealed preference data. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1121–1129. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[2] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1998.

[3] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Human behavior modeling with maximum entropy inverse optimal control. In *AAAI Spring Symposium: Human Behavior Modeling*, page 92, 2009.