理論談話会 #13
(2023/5/31)

A deep inverse reinforcement learning approach to route choice modeling with context-dependent rewards

Zhan Zhao , Yuebing Liang

Transportation Research Part C 149 (2023) 104079

学部4年　加藤小百合

0. Summary

Summary

敵対的逆強化学習(AIRL：Adversarial inverse reinforcement learning)を適用した新たな経路選択モデリングの方法を開発し、得られたモデルが従来の経路選択モデルよりも精度が高いことを実証分析で検証した

良かった点

- 深層学習を用いることで、従来モデルのデメリット（効用関数が線形関数で制限されていることなど）を改善し、経路選択に影響するより複雑な要素を考慮することができるモデルになっている
- 実証分析として、従来モデルを含めた５つのモデルで結果を予測精度、計算効率など複数の観点から比較・考察しているので説得力がある

課題点

- 深層学習の解釈性に改善の余地がある（人の経路選好が何によって構成されるのか、得られた関数から解釈できると良い

0. 新規性・有用性・信頼度

新規性

- リンクベースで深層学習を用いたモデリング手法の研究はなく、新規性は高い
  （深層学習自体が新しくホットな研究分野とされている）

有用性

- 実証分析の結果からも示されているように、他のモデルと比較しても予測精度が高く、計算効率も高い
- 今まで考慮できなかった経路選択の選好に影響を与える様々な要素を取り入れられているので、かなり有望

信頼度

- 経路選択モデルに関しては、従来のモデルを整理し、モデルの比較実験を行っていて信頼度は高い
- 深層学習の解釈の難しさに関して解決方法が述べられていなかった

目次

章構成

1. 研究の概要・目的

従来の経路選択モデル

動的離散選択モデル（DCM）

- パスベース → 大規模ネットワークではパスの数が非常に多く、パスのサンプリングは困難
- リンクベース →
  - ○パスのサンプリングが不要で、リンクベースで詳細な情報を表せる
    (リンク長、道幅、建物の有無など…)
  - ×既存のモデル(RLモデルなど）では単純な線形効用関数を仮定しているため非線形な人間の複雑な経路選好を明らかにできない

DCMモデルとIRLは
構造的類似性
→相性が良い

DNNs(deep neural networks)：
深層学習の中核アルゴリズム
非線形関係を捉え、高次元の特徴を取り込むことができる

本研究：DNNを用いた深層の逆強化学習のフレームワークを経路選択問題に適用
→非線形な関数も表現可能な新たなモデルを開発

逆強化学習(IRL; inverse reinforcement learning)とは;人間の行動から目的を推定することを目的とした強化学習の一種

目標：観察された人間の行動軌跡から非線形も含めた報酬関数（効用関数に同じ）を復元すること

目次

章構成

2.1 既往研究 [経路選択モデリング]

経路選択モデリングの既往研究

既存の経路選択問題はパスベースとリンクベースに分けられる

- パスベース
  - デメリット：膨大なパスの可能性がある→経路除去などの方策が取られてきた
- リンクベース(RLモデルなど)
  - デメリット：単純な線形効用関数だからトリップやユーザーに関するコンテキスト機能を取り入れられない、複雑な効果の説明ができない

2.2 既往研究　[逆強化学習]

逆強化学習について

- RL (Reinforcement Learning 強化学習)

予め定義された報酬関数を最大化する行動を生み出す決定プロセスを学習することが目的

- IRL (Inverse Reinforcement Learning 逆強化学習)

観察された人間の行動を説明する報酬関数を実証データから抽出することが目的

- 報酬関数：通常、いくつかの特徴によって指定され、学習されたパラメータは人間の嗜好を記述し、これはDCMにおける効用関数と同様
- 逐次的な決定過程をモデル化する点で構造的に経路選択と類似性

2.3 既往研究　[逆強化学習]

逆強化学習の既往研究

- Deep Learningの中核アルゴリズムであるDNNを逆強化学習に適用した研究に関心が高まっている
- 逆強化学習の手法別に、既往研究は以下の通り

①値の反復：配送ルート計画（Liuら、2020）

②GAIL：生成的敵対的フレームワークを組み合わせ

タクシードライバーの戦略学習（Zhang et al., 2020）や合成軌道生成（Choi et al., 2021）に適用

③AIRL:本研究で採用→経路選択問題への適用研究なし

●深層学習は解釈可能性が欠落しているという課題

←解釈可能性は人間の行動の理由説明のために必要

- 近年解釈可能な深層学習の理論と方法が研究されている（後ほど出てくるSHAPなど）

目次

章構成

3.1 定式化

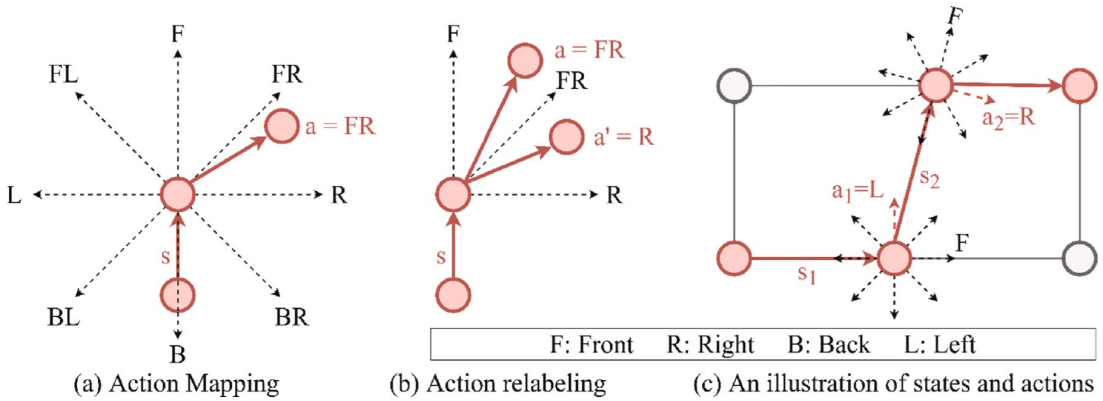
経路選択問題をマルコフ決定過程(MDP)として考える

- State(状態s): 道路ネットワークにおいて旅行者の現在の位置を示すリンク
- Action(行動a)：リンク間の移動選択（下図；８つに分けたもののどれか、R,Fなど）
- Context(コンテキストc) :個人の経路選択に影響を及ぼす要素群ベクトル(トリップ中変化しない)

（例：目的地や目的、個人属性、日時、天気など←非線形で別々に影響を及ぼす可能性が高い）

- Policy(ポリシーπ): Contextに依存し、sとcが与えられた時にどのような行動aを取っていくかを決める関数(π(a|s,c)と表記）（π*は最適な選択行動を返す）
- Reward function(報酬関数R): R(s,a|c)は経路の”preference”のセット(良さ、価値のようなイメージ)（R(s,a|c)はcが与えられた時、sの状態でaの行動を選択した時の効用とほぼ同じと考えられる）

ポリシーより報酬関数の方がより根本的で汎用性が高い



(a) Action Mapping (b) Action relabeling (c) An illustration of states and actions

Point

- これまでの研究にはcontextがあまり含められていなかった（リンクベースなら線形効用として含めることは可能）
- リンクベースモデル(RLモデル)のリンク間の即時効用には含められていなかった目的地の情報がコンテキストに含められている

3.2 経路選択問題における逆強化学習の問題

- **経路選択問題における逆強化学習**とは、軌跡のセット(X=x1,x2,…xN)から報酬関数Rを推定すること

$$x_i = \{(s_1^{(i)}, a_1^{(i)}), (s_2^{(i)}, a_2^{(i)}), \ldots, (s_{T_i}^{(i)}, a_{T_i}^{(i)}) \mid c^{(i)}\},$$

↑i番目の軌跡　↑tステップ目　↑Tステップ目(最後)　T:軌跡の長さ

- 報酬関数Rをパラメータθ関数として定義し、θを学習させることで、Rの関数を形にする
(θの推定方法が実際にはたくさんある)

$$R_\theta(x_i) = \sum_{t=1}^{T_i} \gamma^t R_\theta(s_t^{(i)}, a_t^{(i)} \mid c^{(i)}).$$

観測されたx(軌跡)の起こる確率はRのexpに比例するという法則

$$P_\theta(x) = \frac{1}{Z} \exp\left(R_\theta(x)\right)$$

↑確率　Zは全てのx（軌跡）のRθ(x)の積分和

$$\max_\theta \sum_{i=1}^{N} \log\left(P_\theta(x_i)\right)$$

:→IRL問題は観測された軌跡に基づいて尤度を最大化する問題として組み立てられる

課題:Zを計算すること

従来は値の反復をしたり、RLモデルのように線形と決めることで、パラメータ推定していた
→本研究ではZを直接求めようとするのではなく別のアプローチをとる(敵対学習を利用)

目次

章構成

1. はじめに
2. 既往研究（経路選択モデル/深層学習)
3. 経路選択問題の定式化・経路選択問題における逆強化学習とは
4. AIRLモデルの説明・経路選択問題への応用の仕方
5. 実証分析の方法
6. 実証分析の結果
7. 考察

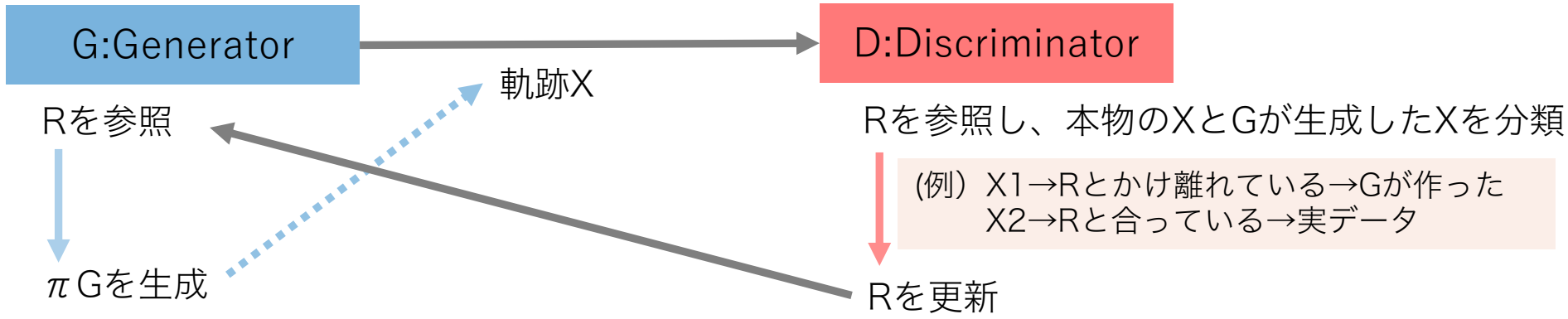
4.1 AIRLモデルについて

AIRL(Adversarial IRL:敵対的逆強化学習)

- 逆強化学習のアルゴリズムの一種で、観測された軌跡データから逆に報酬関数とポリシーを学習する（→最終的には経路選好を明らかにして、経路選択行動を予測する）
- **GとDの２つのモデルが相互に評価し合いながら同時に学習する**

●A discriminator(D)：**inputがGeneratorのoutputなのか、実データなのか分類**する役割を持つ
●A generator(G)： Dにπ^*(最適なポリシー関数)が作ったと思われるoutputを生み出すπGを作る役割を持つ（本物のデータだと分類されるようなデータを作るポリシーを作ることが目標）
Gがポリシーを決定し、DがRを特定する

イメージ図



4.2 AIRLモデル(式の流れ①)

- Dの定義

$$D_{\theta,\phi}(s,a) = \frac{\exp\left(f_{\theta,\phi}(s,a)\right)}{\exp\left(f_{\theta,\phi}(s,a)\right) + \pi_G(a \mid s)}$$

- fは学習するRに関連する関数で以下のように定義される

$$f_{\theta,\phi}(s,a) = g_\theta(s,a) + \gamma h_\phi(s') - h_\phi(s)$$

- gはRの近似関数、($\gamma h\phi(s')$ - $h\phi(s)$は報酬の近似関数における不要な再設定の影響を緩和するための項)

- Dの目標はGのデータと実際のデータのクロスエントロピーを最小にすることで、以下の式

$$\min_{\theta,\phi} -E_D\left[\log\left(D_{\theta,\phi}(s,a)\right)\right] - E_{\pi_G}\left[\log\left(1 - D_{\theta,\phi}(s,a)\right)\right]$$

↑実際の軌跡の期待値　　↑Gが作った軌跡の期待値

4.3 AIRLモデル(式の流れ②)

- 報酬関数Rの定式化にはさまざまな方法があるがAIRLモデルでは以下のように決める

$$R_{\theta,\phi}(s,a) = \log\left(D_{\theta,\phi}(s,a)\right) - \log\left(1 - D_{\theta,\phi}(s,a)\right) = f_{\theta,\phi}(s,a) - \log \pi_G(a \mid s)$$

- Gの目標は最大のRになる軌跡を出力するようなpolicyπ^*を見つけること

$$\max_{\pi_G} E_{\pi_G}\left[R_{\theta,\phi}(s,a)\right]$$

G:Generator ⟷ D:Discriminator

$$\max_{\pi_G} E_{\pi_G}\left[R_{\theta,\phi}(s,a)\right] \qquad \min_{\theta,\phi} -E_D\left[\log\left(D_{\theta,\phi}(s,a)\right)\right] - E_{\pi_G}\left[\log\left(1 - D_{\theta,\phi}(s,a)\right)\right]$$

イタチごっこのように互いに競争し合いポリシー、報酬関数を改善していく

- 経路選択問題に単純応用する場合の課題点：目的地に達するまでに非常に長い時間がかかるか、ループして目的地に達しない←報酬関数がs,aにしか依存していない、目的地の影響の強さを考慮していないから

→Contextベクトルを追加したフレームワークを考えた

4.4 RCM-AIRLモデルについて

RCM-AIRLモデル(Route Choice Modeling AIRL)

AIRLモデルとの相違点

- GとDのモデルに加えてV（特定の目的地への現在の状態の期待されたリターンを計算するため）も追加

GとDの機械学習における計算手法の概要

→次スライドでそれぞれ説明

4.5 RCM-AIRLモデル(G)①

Gの目標：OD（出発地と目的地）のペアが与えられた場合に現実的な人間の軌跡を生成するポリシーπGを学習すること

- Gの入力

①状態特徴量Fs：現在のリンクの特徴（例：リンクの長さ）

②コンテキスト特徴量Fc：目的地に関連する特徴（例：目的地までの最短経路距離）、一般的なコンテキスト（例：旅行者の属性）

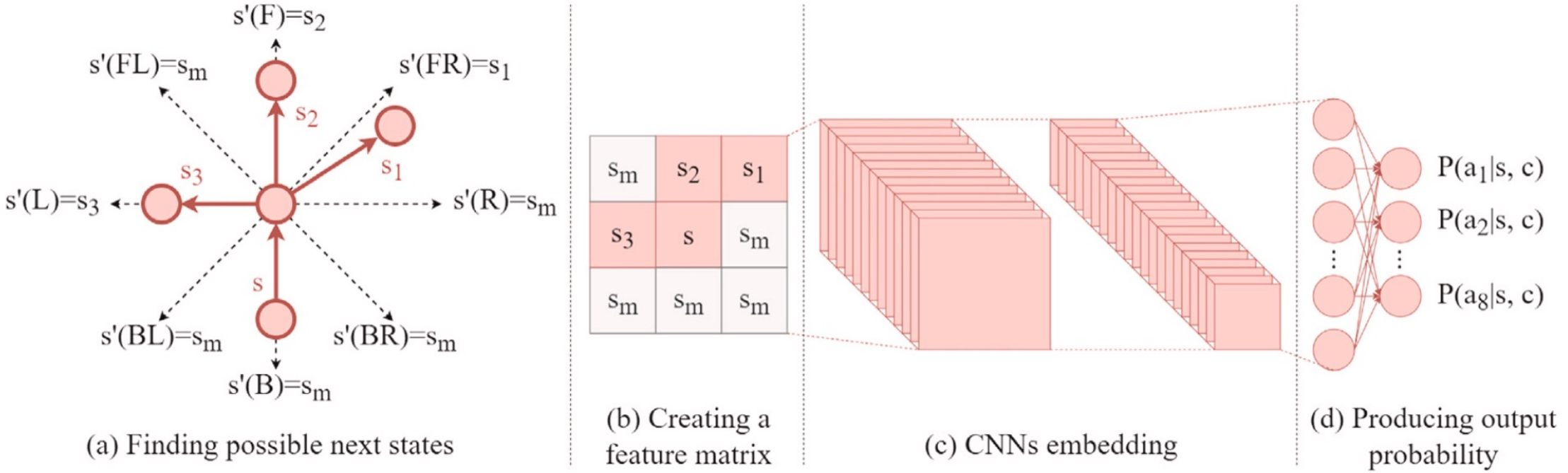
- Gの出力

現在の状態sとコンテキストcが与えられた場合に旅行者が異なるアクションを選択する確率分布

- 隣接する状態間の空間的相関を考慮するために、畳み込みニューラルネットワーク（CNN←DNNの一種）を使用して、現在の状態と潜在的な次の状態のために$F = [Fs ; Fc]$を集約する

4.6 RCM-AIRLモデル(G)②

使用しているネットワーク構造



Step1

可能な次の状態を見つける。状態sが与えられた場合、各アクション（すなわち方向）$a \in A$に対して次の状態$s'(s,a)$を見つける。アクションが有効な状態につながらない場合や、$a \notin A(s)$の場合、$s'(s,a)$はマスク状態としてsmで示される。

Step2

特徴行列の作成。3×3の特徴行列に、状態特徴量Fsと可能な次の状態$S'(s)$のコンテキスト特徴量Fcを集約

Step3

畳み込みニューラルネットワーク（CNN）の埋め込みを学習する。カーネルサイズ2の2層のCNNを使用して、特徴行列から潜在空間ベクトルを学習

Step４

アクションの確率を生成する。CNNから学習した潜在空間ベクトルを入力として、2層の順伝播ニューラルネットワークとsoftmax関数を使用して出力を生成する。

4.7 RCM-AIRLモデル(D)①

Dの役割：実際の人間の軌跡とGeneratorが作った軌跡を見分けること

$$D_{\theta,\phi}(s,a|c) = \frac{\exp\left(f_{\theta,\phi}(s,a|c)\right)}{\exp\left(f_{\theta,\phi}(s,a|c)\right) + \pi_G(a \mid s,c)}$$

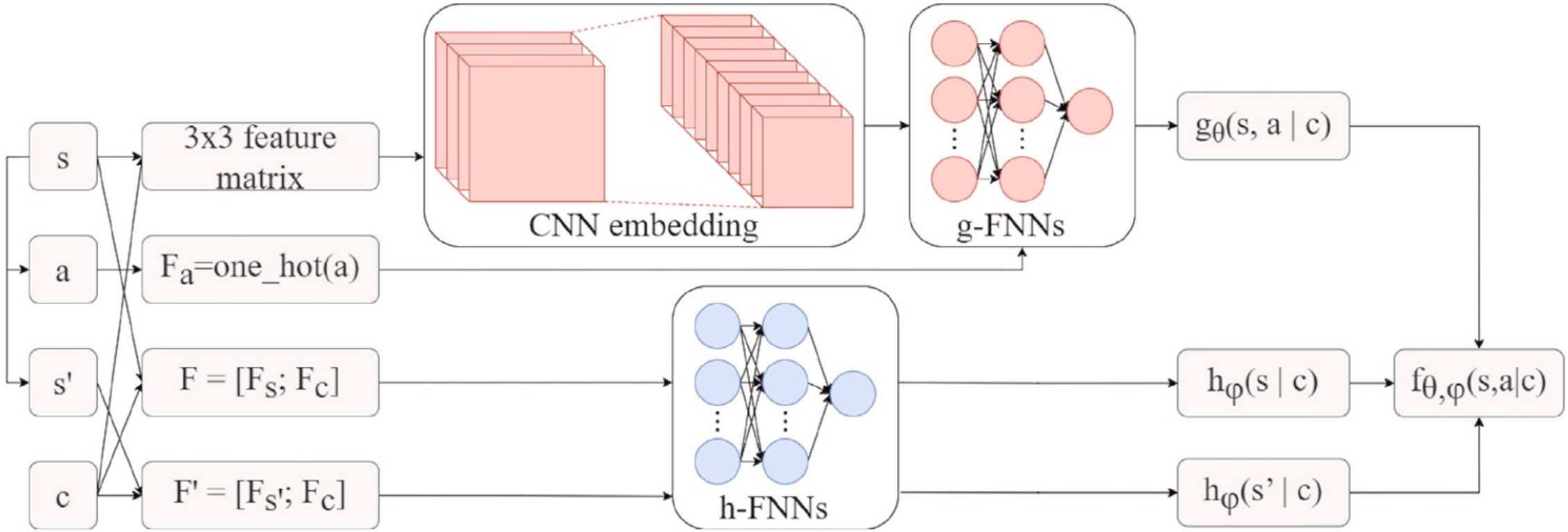
$$f_{\theta,\phi}(s,a|c) = g_\theta(s,a|c) + \gamma h_\phi(s'|c) - h_\phi(s|c).$$

- fを知るにはg(Rの近似関数)とhが必要
- Rは次の状態に依存するので今の状態と、隣接する状態の特徴を畳み込んで入力するためCNNネットワークを使う(CNNにFs,Fcがいる)（Gの時と同じ感じ）
- Rにはactionも影響を及ぼす（人はまっすぐが好きとか、右とか左とかが好きとか）
- Actionの特徴はone-hotベクトルとして表し、CNNに$F = [Fs;Fc]$ がインプットされたレイヤーと繋ぐ
- gはこのCNNレイヤーによって学習される
- hもCNNレイヤーを使って学習できるが計算の都合上、g,hからfを出し、Dを求めたら、Rを以下の式で求める

$$R_{\theta,\phi}(s,a \mid c) = \log\left(D_{\theta,\phi}(s,a \mid c)\right) - \log\left(1 - D_{\theta,\phi}(s,a \mid c)\right)$$

4.8 RCM-AIRLモデル(D)②

使用しているネットワーク構造イメージ



4.8 RCM-AIRLモデル(V)・トレーニングアルゴリズム

Vの役割(Value estimator)：状態sとコンテキストcが与えられた時の状態の価値を計算する

- V(s|c)と表される
- Gが本物に近いサンプルデータを作るために使われる
  - 今の状態sがどのくらい良いかVを参照し判断
  - 次の可能性のある状態s'の状態がどのくらい良いかVを参照し判断
  - 次の状態が一番良くなるaになるようにπを作る

トレーニングアルゴリズム(トレーニングのプロセス)

- 軌跡Xを(s,a)の組に分解する　$\tilde{X} = \{(s_t^{(i)}, a_t^{(i)}, c^{(i)}), \forall t \in \{1, 2, \ldots, T_i\}, i \in \{1, 2, \ldots, N\}\}$
- 実データ$\tilde{X}_e$と生成したサンプルデータ$\hat{X}_e$を用意
- バッチ勾配降下法でDのパラメータを推定
- PPO法でGのパラメータを推定

→それぞれのパラメータを求めることでポリシー関数、報酬関数が求められた！

4.9 2つのAIRL類似モデル

①RCM-BCモデル

- 行動クローニングという教師あり学習手法を利用するアプローチ
- 観測された行動に基づいて、状態と行動を直接対応させるポリシーを訓練する(模倣学習)
- 目的は観測されたデータの尤度を最大化するポリシーπを学習すること
- シンプルで、豊富なデータがある小規模な問題には効果的
- デメリット

  ①連続性を考慮できないので、系列データや時間的な依存関係を適切に捉えられない

  →OD間を結ぶ各(s,a)ペアの時間的な繋がりを考慮できない

  ②観測された入力データが正解になるように直接学習する

  →未知のデータや異なる環境への適応性が制限されやすい

②RCM-GAILモデル

- 敵対学習のフレームワークで、AIRLとかなり近い
- AIRLとの違いは、GAILではデモンストレーションデータから直接ポリシーを学習し、報酬関数の復元を行わないこと

目次

章構成

1. はじめに
2. 既往研究（経路選択モデル/深層学習)
3. 経路選択問題の定式化・経路選択問題における逆強化学習とは
4. AIRLモデルの説明・経路選択問題への応用の仕方
5. 実証分析の方法
6. 実証分析の結果
7. 考察

5.1 実証分析 データ

実証実験で用いたデータ

- 上海の大手タクシー会社の１つから提供されたデータセット
- 2015/4/16-4/21の10609台のタクシーのGPS追跡
- 上海の中でも十分なデータがある範囲に絞った
- 318ノード、712リンク（リンク平均長は199.9m)
- 3500万の記録
- Taxi ID,日付、時間、緯度、経度、空車/満車、
- 目的地が決まったトリップを調べるため、客が乗ったタクシーのデータのみ使う
- 同じルートを繰り返してるトリップ、短すぎるトリップ(15リンク以下）は除外（近距離だと非常に些細な道の違いになってしまうため）
- 結果的に664の目的地リンクをカバーする24470のトリップデータを用いた

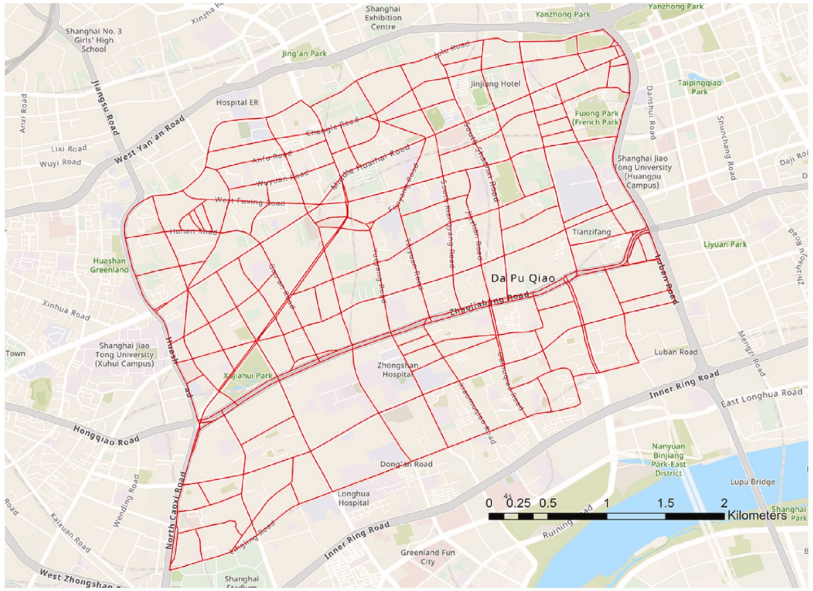


Fig. 5. Selected road network in Shanghai.

5.2 実証分析 データ②

- インプットした特徴
- State features(Fs):リンクの特徴
- Context features(Fc):目的地に関連するものを含んだ特徴
- Action features(Fa):旅行者が動く方向の指示

The description of input features.

Feature	Description
State features F_s	
Link length	The length of a link.
Link level	Whether a link belongs to primary, secondary, tertiary, living street, residential or unclassified link level.
Context features F_c	
Shortest distance	The shortest path distance from the current link to the destination (pre-computed using Dijkstra's algorithm).
Number of links	The number of links along the shortest path from the current link to the destination.
Number of turns	The number of left, right and u-turns along the shortest path from the current link to the destination.
Frequency of link levels	The number of primary, secondary, tertiary, living street, residential and unclassified links along the shortest path from the current link to the destination.
Action features F_a	
Direction	Whether the traveler is moving forward, forward right, right, backward right, backward, backward left, left or forward left.

5.3 実証分析 比較モデル

RCM-AIRLモデル比較する他のモデル

- Path Size Logit(PSL)：パスベースモデルの一つ。それぞれのパスの選択確率の計算で、”各リンクがどのくらい共有されているか”を考慮する項を含めることが特徴。実装では確率の高いパスを５つ候補として用意し、実際に観測された軌跡を候補の中で最も近いパスにマッチさせる
- DNN-PSL：DNNを使用するPSLの拡張版。非線形な関係や柔軟なコンテキストを含めることができるのが特徴
- Recursive Logit：リンクベースモデル。価値関数は線形関数。パスの量は膨大だが、リンクベースなので全てのパスを抽出する必要がない

$$P(s' \mid s) = \frac{\exp(v(s' \mid s) + V(s'))}{\sum_{s'' \in S'(s)} \exp(v(s'' \mid s) + V(s''))}$$

↑瞬間効用(各リンク)　↑次の状態を選択した時の期待効用(各リンク)

- RCM-BCモデル：p.参照
- RCM-GAILモデル：p.参照

従来モデル

AIRL類似モデル

5.4 実証分析 評価指標

モデルの精度の良さを測る指標

- Edit Distance(ED): 2 つの関連するものがどれくらい異なっているかを測る指標。一方を他方に変換する最小の操作回数の計算によって求める。
- BiLingual Evaluation Understudy score (BLEU):スコアは 2 つがどの程度似ているか、近いかを示す
- Jensen–Shannon Distance (JSD):2つのものの確率分布の近さを測ることができる
- Log Probability (LP):与えられたモデルにおいて実データの経路の確率のlogをとったものの平均値

データ成形

- データをランダムに 5 つに分け、 1 つをテスト用、残り 4 つトレーニング用
- 学習用のデータサイズを100,1000,10000の 3 種類で実験
- 実験成功のために重要だとわかったこと：コンテクストに依存する報酬関数の設計

  ←ゴールに関連する情報をきちんと与えることで旅行者が合理的な行動として経路選択すると意味付けられるので、何度もループしたりゴールに到達しないという課題を解消する

目次

章構成

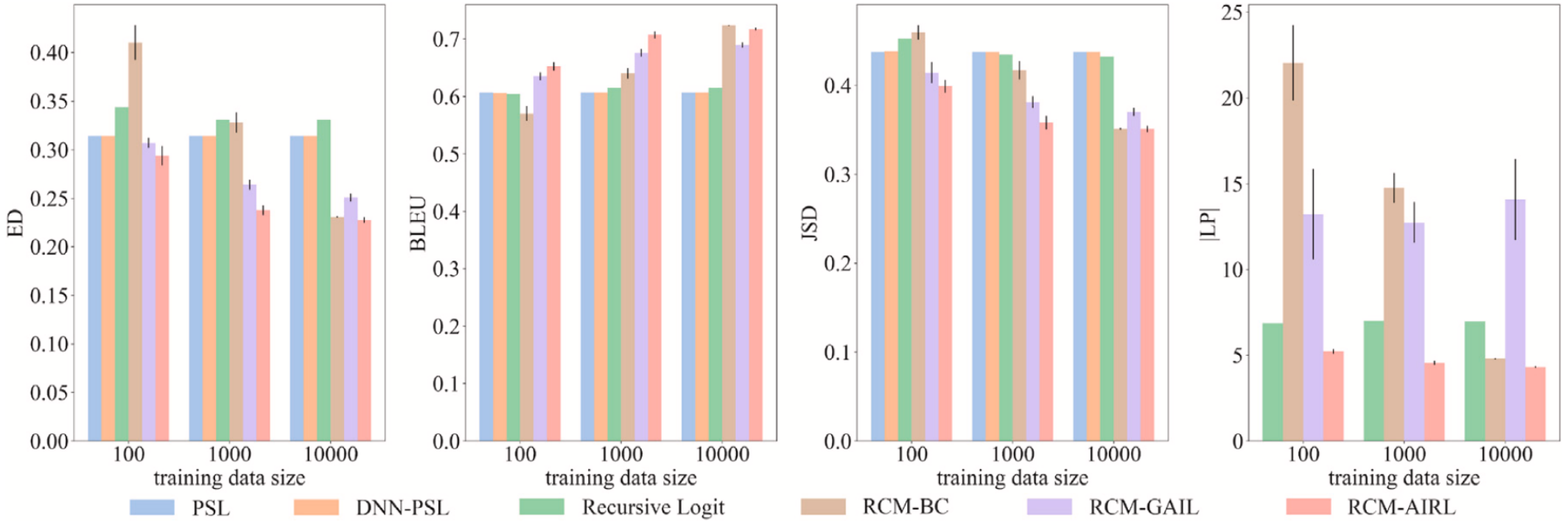
1. はじめに
2. 既往研究（経路選択モデル/深層学習)
3. 経路選択問題の定式化・経路選択問題における逆強化学習とは
4. AIRLモデルの説明・経路選択問題への応用の仕方
5. 実証分析の方法
6. 実証分析の結果
7. 考察

6.1 結果 モデルの予測能力①

- 異なるモデル間の性能差を評価するために、tテストを使用。RCM-BC、RCM-GAIL、およびRCM-AIRLは、既存のベースラインモデルよりも明らかに優れたパフォーマンスを示した（p値が0.001未満）

  →**深層IRL（逆強化学習）/IL（模倣学習）手法がルート選択モデリングにおいて効果的**であることを示唆

- 特にAIRLは全ての評価指標に関して良い結果
- モデルの潜在的な不果実性を考慮するため、３回実験を行なったものの標準偏差をエラーバーとして示したところ、AIRLは割と安定



6.2 結果 モデルの予測能力②

交通流のシミュレーション結果

R^2：回帰モデルの予測の適合度を評価する統計的指標

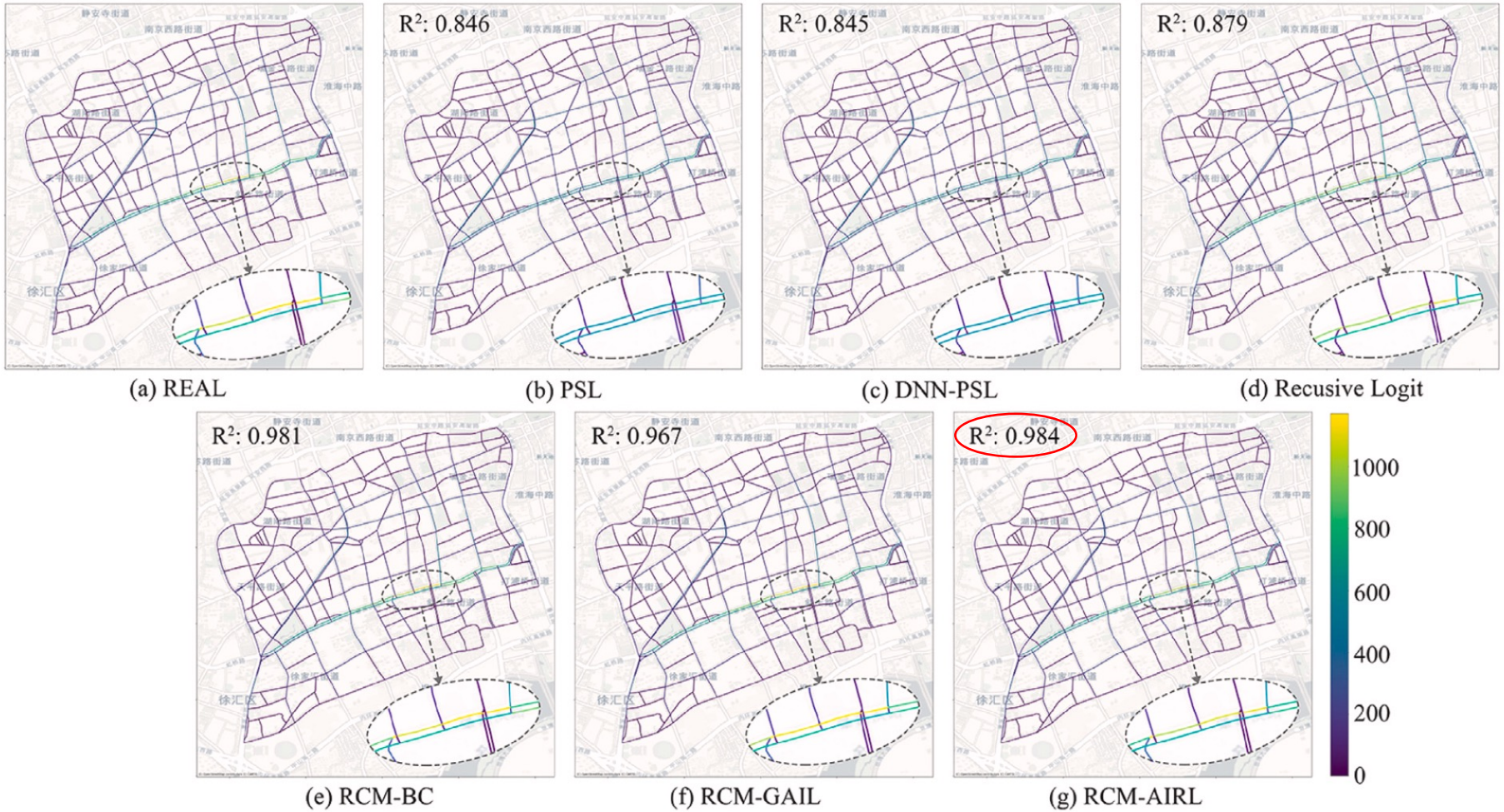
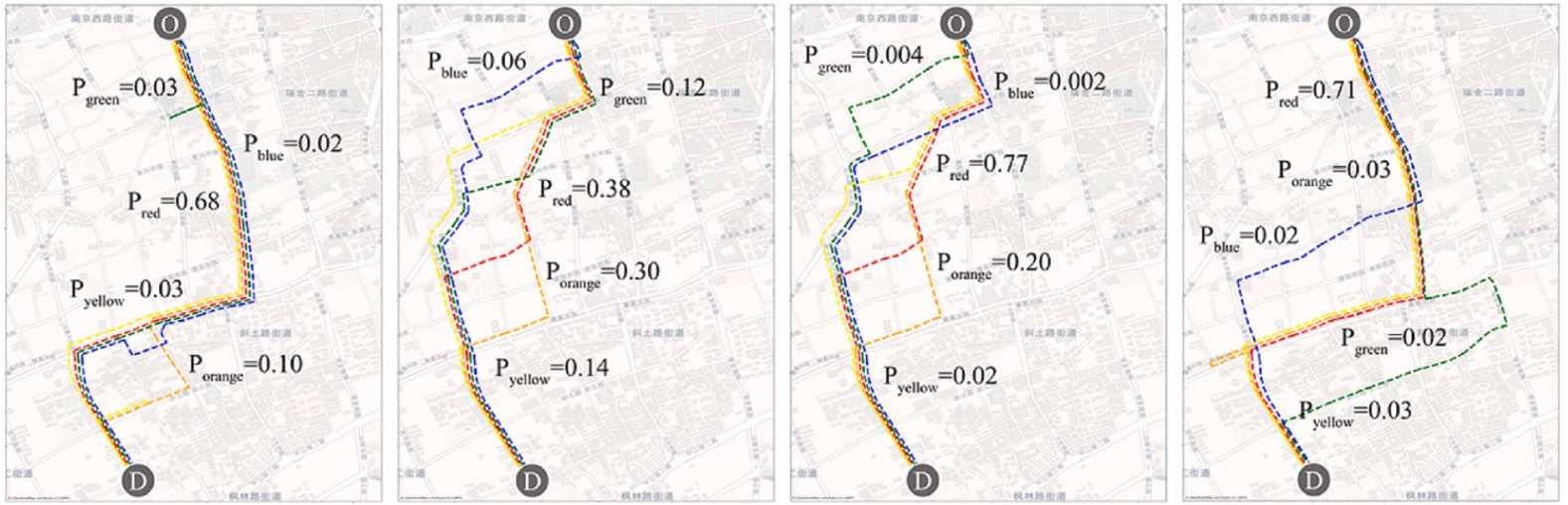


Fig. 7. Predicted link flow distribution using different models.

6.3 結果 モデルの予測能力③

あるODペアに対する実際のルート選択と予測



(a) REAL (b) PSL (c) DNN-PSL (d) Recursive Logit

(e) RCM-BC (f) RCM-GAIL (g) RCM-AIRL

6.4 結果 モデルの計算効率

FE:特徴工学
MO:モデル最適化
PG:パス生成

実装には、モデルから実際のルートを生成するPGが重要

Table 3
Computation time for feature engineering (FE), model optimization (MO), and path generation (PG) of different models. The results are averaged over 5-fold cross validation.

Models		100 training trips			1000 training trips			10 000 training trips		
		FE(s)	MO(s)	PG(s)	FE(s)	MO(s)	PG(s)	FE(s)	MO(s)	PG(s)
Path-based	PSL	28.8	0.3	1370.5	277.8	1.8	1370.5	3166.7	12.6	1370.4
	DNN-PSL	28.8	0.8	1370.4	277.8	5.7	1370.4	3166.7	42.3	1370.4
Link-based	Recursive logit	0.0	22.2	4.0	0.0	44.4	4.0	0.0	238.8	4.1
	RCM-BC	3840.0	124.8	1.7	3840.0	1174.1	1.6	3840.0	11 487.4	1.6
	RCM-GAIL	3840.0	6112.8	1.7	3840.0	9305.3	1.6	3840.0	18 270.9	1.6
	RCM-AIRL	3840.0	5990.4	1.6	3840.0	8962.5	1.5	3840.0	18 307.0	1.5

6.5 結果 未知の目的地に対するモデルの汎用性

- 学習用のデータとは別のテストデータを用いた
- データ数を100,1000,10000の３パターンで比較

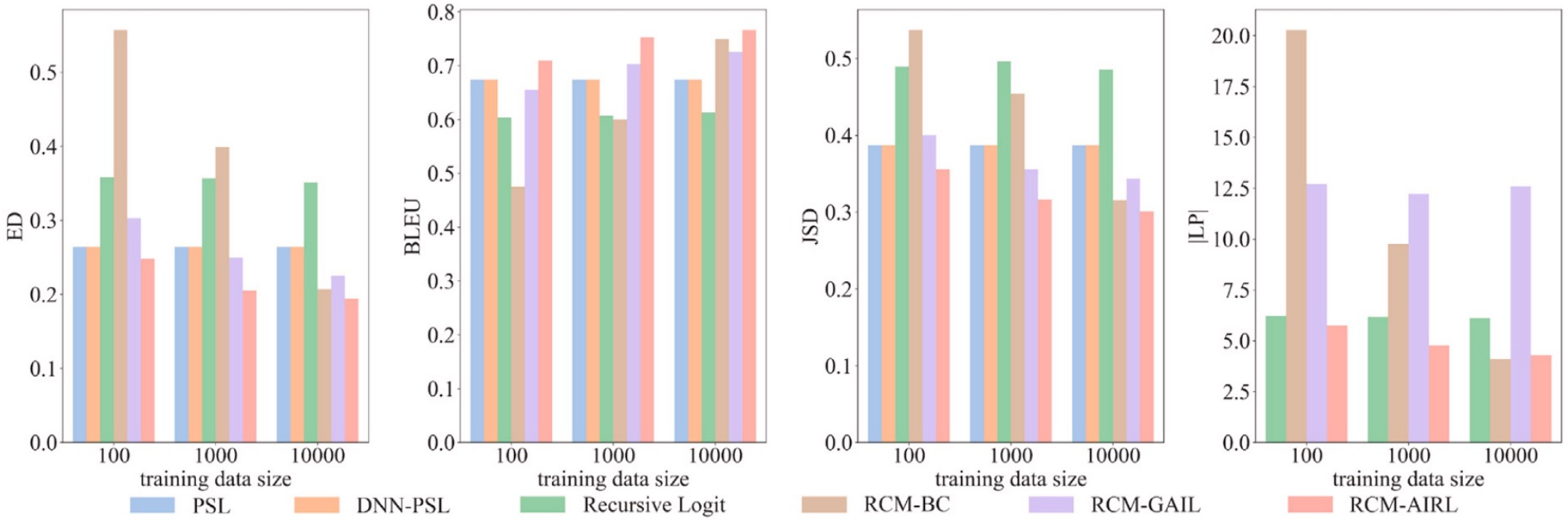
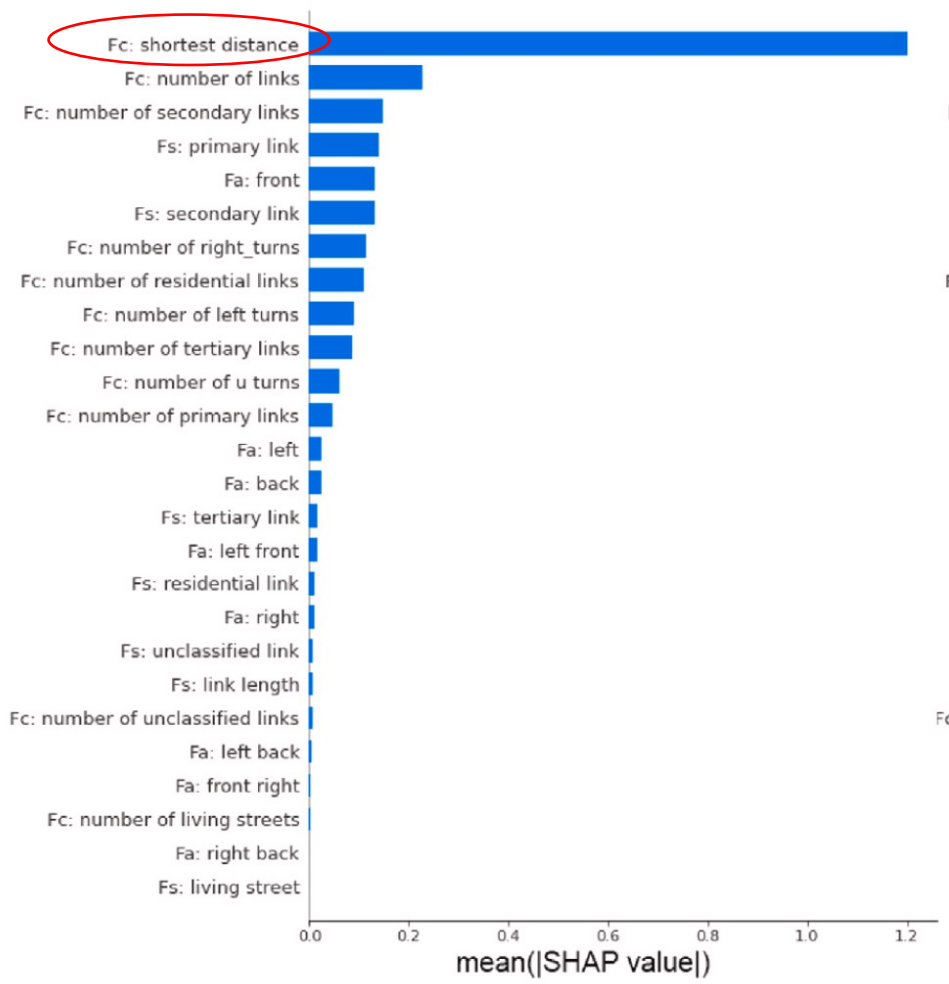


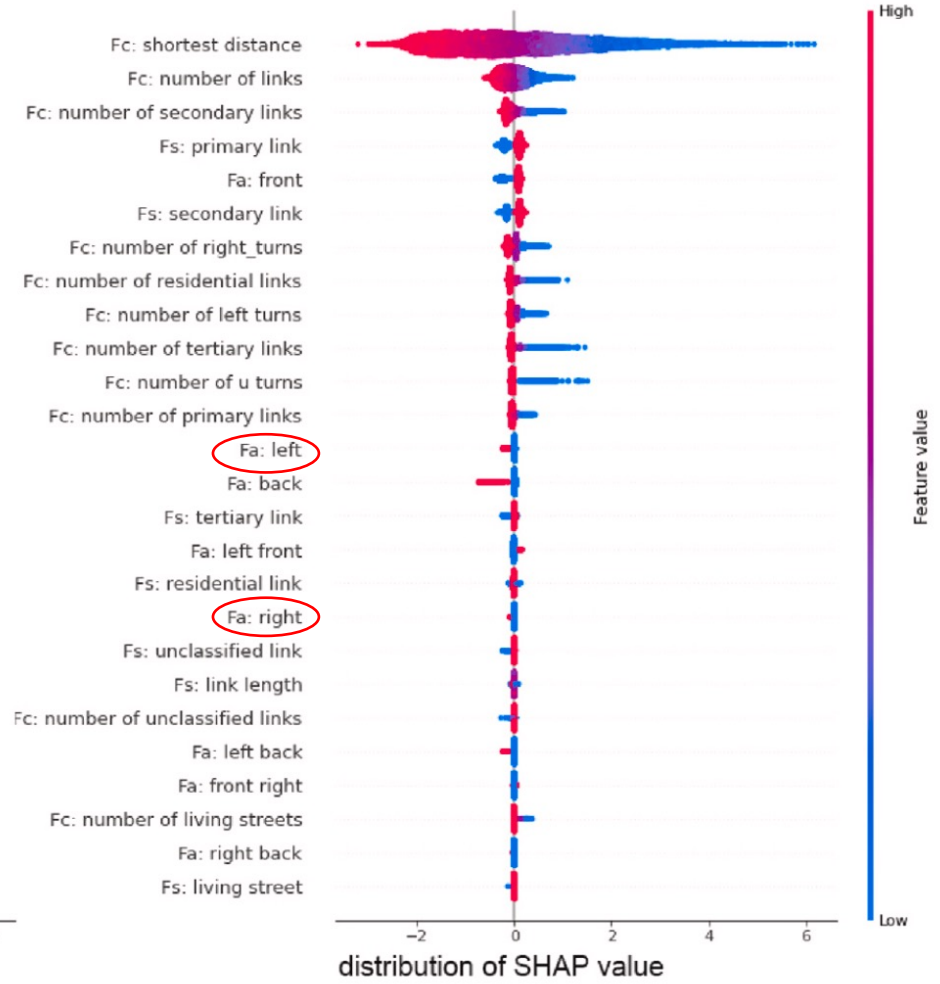
Fig. 9. Route choice prediction performance of different models for unseen destinations.

RCM-AIRLは未知のデータに対して一般化可能で、学習データの必要量が比較的少ない

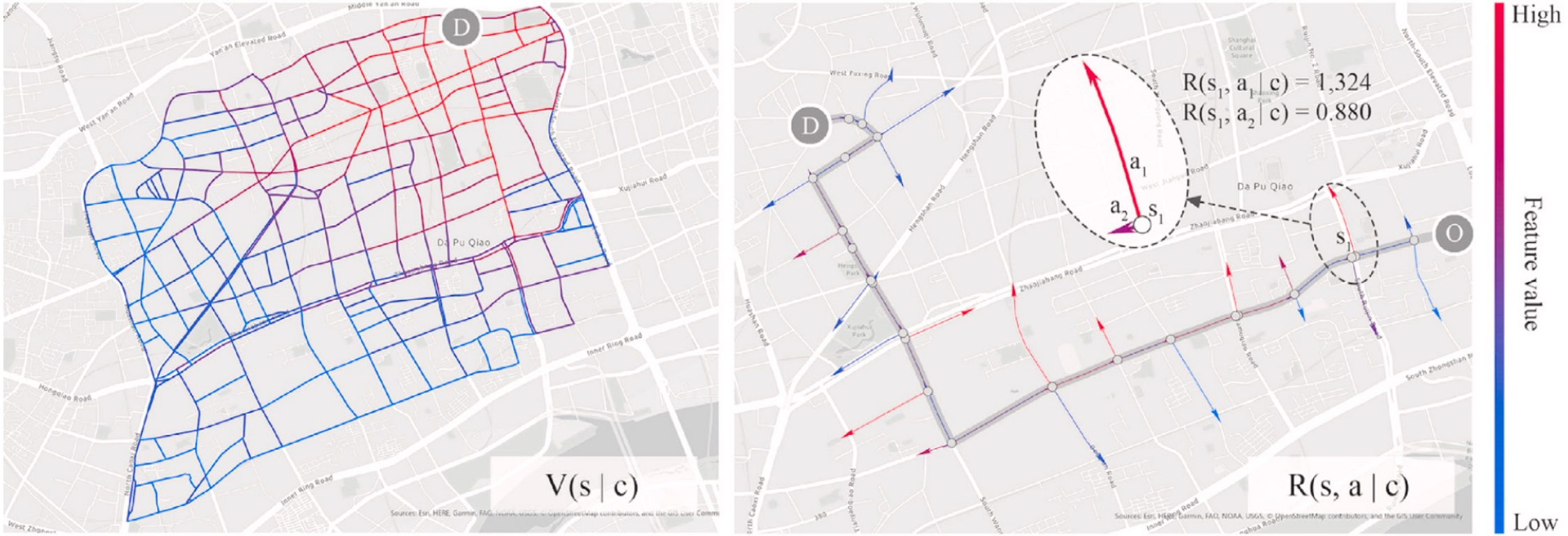
6.6 結果　経路選択行動の”解釈可能性”について

- SHapley Additive exPlanations（SHAP）を使用して、入力特徴量の影響を理解
- SHAPは、機械学習モデルの説明を行うためのゲーム理論的なアプローチを用いる説明可能なAI技術





6.7 結果　経路選択行動の”解釈可能性”について



(a) Estimated value of a selected destination　(b) Estimated reward along a selected route

Fig. 11. Examples of estimated $V(s|c)$ and $R(s,a|c)$ from RCM-AIRL.

目的地に近いほど高い値を示す

- コンテクストcで状態sの時行動aを取った場合の瞬間効用はR(s,a|c)で近似される
- 旅行者が短期的な報酬だけでなく、長期的な報酬/効用も考慮していることを示唆

6.8 結果　潜在的な個人の特徴の取り込み

- 個人の特徴をコンテキストcの一部として扱うことができる
- 本実験ではデータの制約により、実験では明示的な個人の特性を組み込むことはできなかったが、各GPSトレースは車両ID（個人識別子として仮定）と関連付けられているため、DNNから直接潜在的な個人特徴を抽出することが可能

  →経路選択モデルの一部として、各個々の旅行者に対して埋め込みベクトルを学習することができる

  →パーソナライズされたルーティング予測が可能

- DNNから学習された潜在的な個人特徴がモデルの予測性能にどのように貢献するかを探るための実験を行った
- 最も多くのトリップを持つ50人のドライバーを選択、各ドライバーについて、軌跡データのの80%をトレーニングデータ、20%をテストデータに分割

  →予測性能はわずかに向上する

Table 4
Performance comparison of incorporating embedded individual features or not using RCM-AIRL.

With/without individual embedding	ED	BLEU	JSD	LP
Without individual embedding	0.245	0.700	0.391	−4.206
With individual embedding	*0.223*	*0.716*	*0.380*	*−3.869*

7. まとめ　結論と今後の展望

結論

- 本研究では、経路選択モデリングのための深層逆強化学習のフレームワークを提案
- 経路選択問題をマルコフ決定過程として定式化し、実際の人間の経路選択行動を最もよく説明し、予測することができる基礎となる報酬関数（経路選好）を復元することが目標
- 手法として、報酬関数やポリシーが依存するコンテキストをDNNを用いて近似し、敵対的逆強化学習で報酬関数、ポリシーをモデルフリーで効率的に学習

● 優れた点

- リンクベースで、さらにDNN（深層ネットワーク）を用いることで複雑な文脈を取り入れることができる
- 模倣学習と異なり、ポリシーよりも根本的で汎用性が高い報酬関数を推定する

展望

- 深層学習は高次元の特徴を取り込めるので、経路選択に影響を与えうる他の特徴(気象データ、交通フロー、道路の特徴など）を組み込む
- ユーザーの個別の特徴や好みを考慮したパーソナライズドなルート選択モデルを構築する
- DNNはブラックボックスモデルとして知られており、DNNを用いたモデルの解釈性、透明性の研究も重要

所感

- 深層学習分野の知識が全くなく理解に苦労した一方で、深層学習のネットワークで多くの情報を畳み込んで取り入れられることはわかり、深層学習の可能性に圧倒された！
- 実装を考えるには深層学習の理解が追いつかなかった…
- 自分の研究（観光回遊）に当てはめると…

歩行者の経路選択決める関数　　土地

$\pi = f_\theta(x,R)$　　$R = f_\phi(x,R)$

参考

実験の仮定

Table 2
Hyperparameters used for RCM-AIRL.

Hyperparameter	Value
Number of discriminator updates per iteration	1
Number of PPO updates per iteration	10
Batch size for PPO updates	64
Learning rate	3e−4
Discount rate of reward (γ)	0.99
Number of generated samples per iteration	8192
Number of CNN layers	2
Output channel and kernel size in the first CNN layer	20, 3
Output channel and kernel size in the second CNN layer	30, 2