文字列カーネルによる旅行時間予測

Travel-Time Prediction using String Kernels

井手剛　　加藤整
Tsuyoshi Idé　　Sei Kato

IBM 東京基礎研究所
IBM Research, Tokyo Research Laboratory

地図上のある 2 点間を結ぶ任意の経路に対する旅行時間予測の新しい手法を提案する。我々の手法の特長は、交通量の多い特定のリンクに注目して時系列モデリングを行う従来手法と異なり、未知リンクを含む経路に対しても確率的な旅行時間の予測を可能にするという点である。新たなアイディアは 2 つある。文字列カーネルを経路の類似度として使うことと、確率的予測のために正規過程回帰を使うことである。本論文は、旅行時間予測の問題をトラジェクトリに対する回帰問題として定式化した初めての試みである。

1. 初めに

最近のセンシングおよび情報技術の進歩により、自動車や人間といった移動体を、広い領域にわたり追跡しその軌跡をデータとして蓄積することが可能になってきた。環境問題に絡んで最近重要性を増している高度交通システム（Intelligent Transportation System; ITS）は、まさにそのようなデータを提供する枠組みである。ちょうど Web ネットワーク上の情報流が Web マイニングという新しい研究分野を作り出したように、交通データも、その多様性と巨大さから新たな研究領域を生み出しつつある。

この論文では、ある所与の始点と終点の間の**旅行時間予測**（または所要時間予測）という問題を考える。これは交通モデリングの最も基本的なタスクのひとつであり、現代的な ITS が現れ始めた 90 年代から関心を持たれてきた。一般に、交通モデリングを行うには 2 つの見方がある。ひとつの見方は「**路傍の視点**」とでも言うべきもので、これはある固定したリンク（隣接する交差点の間の道路）において、移動体の流れを観察するものである。もうひとつが**運転者の視点**で、それぞれの移動体が通過するすべてのリンクを解析対象にする。伝統的には、旅行時間予測の問題は路傍の視点に基づいて行われてきた。状態空間モデルによる時系列予測がその典型的な例である [Nakata 04]。移動体の追跡データが蓄積されはじめたのはつい最近のことなので、これはある意味当然のことである。

従来研究の実用上の問題は、交通量があまり多くないリンクに対しては旅行時間の予測が難しいという点であった。地図上で任意に与えた経路には一般に非幹線道路が含まれており、そこでは、精度の良い時系列モデリングができるほどの交通履歴は普通望めない。交通量情報の取得に必要なインフラ自体が未整備という場合もよくある[*1]。GPS (Global Positioning Systems) などで取得できる広域的なデータを考えた時、少数の特定リンクに対する精緻なモデルより、都市全体を俯瞰的に見た上で、任意の経路に対して旅行時間を概算してくれる方がよい場合も実用的に存在する。例えば、渋滞回避のための迂回路を交通制御センターが各車両に指示するといった状況がそれであり、対象を自動車以外にも広げて想像してみると、例えば店舗内動線解析など、さらに多くの例があることに気づく。

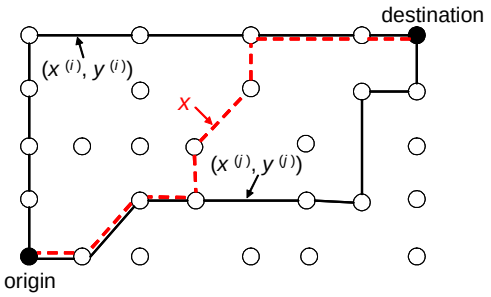

図 1: 問題設定。過去の履歴が、経路と所要時間の組の集合として $\{(x^{(n)}, y^{(n)})|n = 1, ..., N\}$ のように与えられている時、点線で示されているような任意の経路 x に対して所要時間 y を予測する。

従来研究と異なり、本論文では、運転者の視点で旅行時間予測の問題を定式化することを提案する。我々の問題設定を図 1 に示す。クエリとして与えられる経路 x は、一般に裏道なども含み、従来手法では旅行時間予測が難しい。また、可能な経路の個数はリンクの数に対し指数関数的であり、履歴の中に x と同一の経路を見出せるとは限らない。経路 x に対して実数値 y を関係付けるという意味では、これは回帰の問題と見なせるが、入力 x は普通のベクトルデータではないので、標準的な回帰公式は使えない。

やや抽象的に考えると、我々の問題は、移動体の経路、あるいは**トラジェクトリ**と、旅行時間を結びつけるという問題である。トラジェクトリの解析はデータマイニングにおいて最近注目を集めている新しい研究分野であり [Vlachos 05, Lee 07, Kriegel 08]、GPS をはじめとしたセンシング技術の進歩にともない、今後応用面・理論面での研究の進展が期待されている。

本論文では、文字列カーネルと正規過程回帰 (Gaussian process regression) を用いて旅行時間予測の問題を解くことを提案する。まず、トラジェクトリを適当な文字列で表現し、トラジェクトリ同士の類似度を文字列カーネル [Leslie 02] で定義する。そして、ノンパラメトリックなカーネル回帰手法としての正規過程回帰を用いて、旅行時間を予測する。著者の知る限り、本研究はトラジェクトリマイニングの問題として旅行時間予測を扱った最初の仕事である。

連絡先: 井手剛 goodidea@jp.ibm.com

*1　実際、名古屋市においては、幹線道路に限ってみても、VICS (vehicle information and communication system) ビーコンが設置されているのは道路長にしてわずか 22%のみである [Morikawa 07]。

2. 問題設定

本節では改めて旅行時間予測という問題を整理し、扱うデータについて簡単に説明する。

まず定義から始める。

定義 1 (リンク) リンクとは隣接する交差点の間にある道路のことである。

定義 2 (経路) 経路とは、任意の 2 つの連なったリンクがひとつの交差点を共有するようなリンクの系列である。

定義 3 (旅行時間予測問題) 本論文での旅行時間予測タスクとは、訓練データ $\mathcal{D}$ から、任意の経路 x に対する旅行時間 y の確率分布 $p(y|x,\mathcal{D})$ を学習することである。

ここで、訓練データ $\mathcal{D}$ は過去の N 個の履歴情報からなり、

$$\mathcal{D} \equiv \{(x^{(n)}, y^{(n)})|n = 1, 2, ..., N\} \tag{1}$$

と書かれる。ここで $x^{(n)}$ は、第 n 番目の履歴の経路であり、$y^{(n)}$ はその経路に対する所要時間である。$y^{(n)}$ は、たとえば、312 秒、というような実数となるが、$x^{(n)}$ の方は、

$$x^{(n)} = (25020201, 24021102, 222020101, 258020001, ...)$$

のように、地図上のリンク ID の系列として与えられる。

交通データは一般に定常ではないので、$\mathcal{D}$ はある時間区間を指定した上で収集されたと想定する。$\mathcal{D}$ の中のすべての経路と x は、ある同一の始点と終点を通るものとする。先に述べたように、可能な経路の種類は指数個あるので、x が $\mathcal{D}$ に含まれるとは限らないことに注意する。従って、「同じ経路を通った過去 100 台の平均」、のような単純な手法は一般には使えない。データについてのさらに詳しい説明は我々の別論文 [Idé 09] を参照されたい。

3. トラジェクトリに対する回帰モデル

本節では、 ある経路 x を入力とした回帰モデルの説明を行う。ポイントは、入力 x を明示的に何かの特徴ベクトルとして扱うのではなく、入力空間における類似度を介して回帰モデルを構築することである。これはカーネル回帰の枠組みによって実行できる。ここでは非線形関数の記述能力や結果の安定性、さらに確率的出力の得やすさの点から [Bishop 06]、正規過程回帰を採用する。

3.1 正規過程回帰

正規過程回帰における最初の仮定は、$\mathcal{D}$ における第 n 番目の旅行時間 $y^{(n)}$ の分布が、分散 σ^2 の観測ノイズを持つ正規分布により

$$p(y^{(n)}|f_n) = \mathcal{N}(y^{(n)} - \bar{y}|f_n, \sigma^2) \tag{2}$$

と表されるということである。ここで $\mathcal{N}(\cdot|f_n, \sigma^2)$ は平均 f_n、分散 σ^2 の正規分布を表す。後で述べるように、σ はデータから決められる上位パラメター (hyperparameter) であるが、しばらく所与の定数と考えておく。$\bar{y}$ は $\mathcal{D}$ における所要時間の総和平均である。ノンパラメトリック回帰モデルの一般的特徴に従って、このモデルでは、各 $y^{(n)}$ に潜在変数 $f^{(n)}$ が結び付けられており、原理的には $\mathcal{D}$ のいかなる関数関係でも再現することができる。しかしこのままでは過適合し放題であるから、正規過程回帰では、$f^{(1)}, ..., f^{(N)}$ に対して次の事前分布を置く。

$$p(\boldsymbol{f}_N) = \mathcal{N}(\boldsymbol{f}_N|\mathbf{0}, \mathsf{K}) \tag{3}$$

ここで $\boldsymbol{f}_N \equiv (f_1, ..., f_N)^\top \in \mathbb{R}^N$ であり、共分散行列 K の (i,j) 成分は経路 $x^{(i)}$ と $x^{(j)}$ の間のカーネル関数 $k(x^{(i)}, x^{(j)})$ として定義される（引数には入れていないが、$p(\boldsymbol{f}_N)$ は $\{x^{(n)}\}$ に依存していることに注意）。

式 (2) および (3) という二つが正規過程回帰のモデルである。y についての予測分布は、通常のベイズ推論の枠組みに従い、まず、潜在変数 f についての事後分布を求め、それを元に $p(y|f)$ から f を積分消去する。詳細は [Bishop 06] または [Idé 09] を参照されたい。これらの計算はすべて解析的に実行できて、予測分布は以下のように与えられる。

$$p(y|x,\mathcal{D}) = \mathcal{N}(y|m, s^2) \tag{4}$$

$$m = \bar{y} + \boldsymbol{k}^\top \mathsf{C}^{-1} \boldsymbol{y}_N \tag{5}$$

$$s^2 = \sigma^2 + k(x,x) - \boldsymbol{k}^\top \mathsf{C}^{-1} \boldsymbol{k} \tag{6}$$

ただし $\boldsymbol{y}_N$ と $\boldsymbol{k}$ は

$$\boldsymbol{y}_N = (y^{(1)} - \bar{y}, ..., y^{(N)} - \bar{y})^\top \tag{7}$$

$$\boldsymbol{k} = (k(x^{(1)}, x), ..., k(x^{(N)}, x))^\top \tag{8}$$

で定義され、$\mathsf{C} \in \mathbb{R}^{N \times N}$ は

$$\mathsf{C} = \mathsf{K} + \sigma^2 \mathsf{I}_N \tag{9}$$

で定義される。ここで、I_N は N 次元の単位行列である。

3.2 文字列カーネル

二つの経路の類似度を考える際、最も素朴な方法は経路長に注目することである。我々の文脈ではこれは、長い経路には大きな旅行時間を期待する、ということを意味する。これは高速道路などではおそらく悪くないモデリングであると考えられるが、市街地での予測では、交差点での挙動が大きく旅行時間に影響するはずであり、第 0 近似としての意味しか持たないものと想像される

このような見方を一般化して、まず各経路を何らかの文字列で表す。そのアルファベットとしては、先に与えた ID や、リンクの方向を表す東西南北などの文字を想定することができる。次に、その文字列の長さ p の連続する部分文字列を考え、それを用いて、$x^{(i)}$ と $x^{(j)}$ の間のカーネル関数を次のように定義する。

$$k_p(x^{(i)}, x^{(j)}) = \beta \sum_{\boldsymbol{u} \in \Sigma^p} N_{\boldsymbol{u}}(x^{(i)}) N_{\boldsymbol{u}}(x^{(j)}) \tag{10}$$

これは p-spectrum カーネルとして知られているものである [Leslie 02]。上に使った記号の定義は次のとおりである。

- Σ はリンクを表現するのに使われている文字の集合である。
- Σ^p は長さ p の相連なる部分文字列の集合である。
- $N_{\boldsymbol{u}}(x^{(i)})$ は、ある経路（文字の系列）$x^{(i)}$ における部分系列 $\boldsymbol{u}$ の出現回数である。

係数 β は事前分布の分散の大きさを制御する定数で、次節の通り、σ 同様データから決められる上位パラメターとして扱われる。

3.3 上位パラメター σ および β の計算

ベイズ推論の枠組みにおいては、上位パラメター σ と β は周辺化尤度を最大化することで決められる。通常これは勾配法により行われるが、今のモデルでは比較的計算効率のよい固定点方程式を導ける。これを示そう。今、σ の代わりに

$$\gamma \equiv \sigma^2/\beta$$

を使えば、対数周辺化尤度は次のように書かれる。

$$\begin{aligned}\psi(\gamma,\beta) &\equiv \ln \int \mathrm{d}\boldsymbol{f}_N \; p(\boldsymbol{f}_N) \prod_{n=1}^{N} p(y^{(n)}|f_n) \\ &= -\frac{1}{2}\ln\det(\mathsf{C}_1) - \frac{1}{2\beta}\boldsymbol{y}_N^\top \mathsf{C}_1^{-1}\boldsymbol{y}_N - \frac{N}{2}\ln\beta\end{aligned}$$

最後の等式で定数項を省いた。行列 C_1 は次のように定義される。

$$\mathsf{C}_1 \equiv \mathsf{K}_1 + \gamma \mathsf{I}_N$$

ここで、K_1 は $\beta = 1$ の時のカーネル行列である。行列の微分に関する標準的な公式を使えば [Bishop 06]、最適解の条件は次のようになる。

$$0 = \frac{\partial\psi}{\partial\gamma} = \frac{1}{2\beta}\boldsymbol{y}_N^\top \mathsf{C}_1^{-2}\boldsymbol{y}_N - \frac{1}{2}\mathrm{tr}\left(\mathsf{C}_1^{-1}\right) \tag{11}$$

$$0 = \frac{\partial\psi}{\partial\beta} = -\frac{N}{2\beta} + \frac{1}{2\beta^2}\boldsymbol{y}_N^\top \mathsf{C}_1^{-1}\boldsymbol{y}_N \tag{12}$$

それぞれを交代的に繰り返し解くことで解が得られる。後者については、ある γ の値について、解が解析的に求められることに注意する。すなわち、

$$\beta = \frac{1}{N}\boldsymbol{y}_N^\top \mathsf{C}_1^{-1}\boldsymbol{y}_N \tag{13}$$

数値計算的な詳細は別論文 [Idé 09] を参照されたい。

3.4 算法の要約

訓練時。σ と β を次のように求める。

1. 入力: カーネル行列 K、旅行時間のベクトル $\boldsymbol{y}_N$、σ と β についての初期値。
2. 手順: 式 (11) および (12) を交代的に収束するまで解く。
3. 出力: ψ を最大化する σ^2 と β。

予測時。あらかじめ C の Cholesky 因子 L とその逆 L^{-1}、さらに $\boldsymbol{h} \equiv \mathsf{L}^{-1}\boldsymbol{y}_N$ を計算しておく。

1. 入力: 経路 x （と事前計算された L^{-1} および $\boldsymbol{h}$）。
2. 手順:
   - $\boldsymbol{l} \equiv \mathsf{L}^{-1}\boldsymbol{k}$ を計算する。
   - 予測平均 $m = \bar{y} + \boldsymbol{h}^\top\boldsymbol{l}$ を計算する。
   - 予測分散 $s^2 = \sigma^2 + k(x,x) - \boldsymbol{l}^\top\boldsymbol{l}$ を計算する。
3. 出力: 予測平均 m と予測分散 s^2。

4. 実験

本節では実際の地図情報に基づく交通データに基づいて旅行時間予測の実験を行う。

4.1 実験の設定

データ $\mathcal{D}$ を生成するために、我々はエージェントベースのトラフィックシミュレータである IBM Mega Traffic Simulator [*2] を用いた。図 2 にシミュレータのスクリーンショットを示す。画面には京都の市街地図が表示されており、薄く点で表されているのが生成されたエージェント（自動車）である。その地図の左上が経路の始点、右下が終点である。

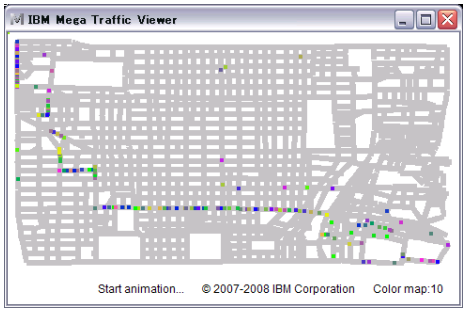

図 2: IBM Mega Traffic Simulator と、その画面上に表示された京都市街地図。

各エージェントはそれぞれに対して事前に定義された経路をたどるが、各リンクの法定速度に加えて、他のエージェントとの相互距離により速度を変化させるようプログラムされているので（その説明はここでは省く）、得られる旅行時間は試行ごとに大きくばらつく。ここではまず、ある N_0 個の経路の候補を、理性的な運転者の行動を模擬する目的で k 最短経路アルゴリズム [Yen 71] を用いて生成し、各エージェントはその中から経路を順に選択するものとした。エージェントの発生間隔は、平均 0.1 秒の指数分布に従うものと仮定し、信号待ちを模擬するために、交差点においては、τ 秒の待ち時間を与える。以下の実験では、$N_0 = 132$ とし、訓練用に $N = 100$ 個をランダムに選択し、残りをテスト用とする[*3]。

4.2 比較対象

上に説明したデータに対し、3 つのカーネル関数を比較する。(1) 文字列集合 Σ としてリンクの ID そのものを採用した p-spectrum カーネル、(2) Σ としてリンクの方角（東西南北）を採用した p-spectrum カーネル、(3) 2 つの経路の囲む面積を非類似度に使ったカーネル、である。それぞれ、ID カーネル、方向カーネル、面積カーネルと呼ぶ。

面積カーネルは、2 つの経路 $x^{(i)}$ と $x^{(j)}$ の囲む面積を $S(x^{(i)}, x^{(j)})$ とした時、

$$k^{\mathrm{area}}(x^{(i)}, x^{(j)}) \equiv \beta\, \mathrm{e}^{-S(x^{(i)},x^{(j)})}$$

のように定義される。$S(x^{(i)}, x^{(j)})$ は地図上の 2 つの経路の ℓ_1 距離の拡張とみなすことができる。この意味で面積カーネルは、動的時間伸縮法など 2 つの経路の間の非類似度を幾何学的な方法で定義する方法の対応物とみなすことができる。

4.3 実験結果

ここではまず、ID および方向カーネルについて、文字列長 p に対する予測精度の依存性を論じ、次いで異なるカーネルを比較する。

いくつかの p の値に対して予測平均 $m(x)$ を計算し、実測値との相関係数 r を求めた。結果を図 3 に示す。この図は ID カーネルに基づくが、方向カーネルでも本質的な傾向は変わらない。予測精度は τ の値によって異なるが、おおむね $p = 2$ で最も精度がよいことがわかる。なお、$\tau = 10$ 秒の時は、交

*2 http://www.ibm.com/jp/press/pressroom/kaiken/20080610a.pdf

*3 データの一部は次の URL で入手できる: http://www.trl.ibm.com/projects/socsim/project_e.htm

差点での待ち時間は総旅行時間の 3 割ほどを平均して占める。これを考えるとこの実験結果は妥当であろう。すなわち、総旅行時間には個々のリンクの影響がまず支配的で、次いで隣接する複数のリンクの関係が効いてくる。つまり右左折や直進など、交差点での振る舞いの相違が結果に効いてくるということである。

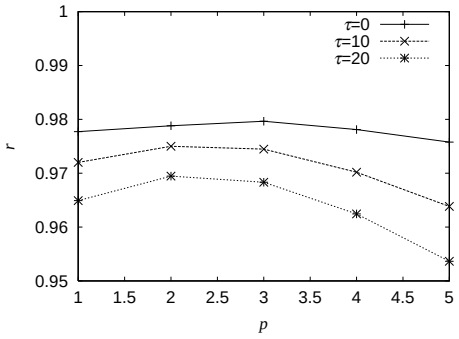

図 3: 文字列長 p の関数として表した r 値。実線が $\tau = 0$、破線が $\tau = 10$、点線が $\tau = 20$ を示す。

図 4 に異なるカーネル同士の比較結果を示す。縦軸が予測値、横軸が実測値であり、点線が完全一致を示す。図から、ID および文字列カーネルについては予測と実測の対応が比較的よいことがわかる。他方、面積カーネルの結果は顕著に悪い。これは、地図上で経路同士の「形」を直接比べるような手法が役に立たないことを意味する。たとえば高速道路と一般道が併走している状況を想像すれば直感的に納得できる結果であろう。トラジェクトリの解析は何らかの意味で「形」の比較に基づいて行われることが多いが、この場合は p-spectrum カーネルが定めるヒルベルト空間において類似性を考えねばならないということであり、興味深い結果と言える。

表 1 に上記の結果を要約してある。表にはテストデータにわたる予測分散の平均の値も示してある。表からわかるとおり、ID カーネルが最もよい r 値を示しており、予測の分散も小さい。

5. まとめ

任意経路に対する旅行時間の予測という問題を設定し、トラジェクトリに対する回帰の問題として定式化した。提案した手法は未知リンクを含む経路についても旅行時間を予測できるという、従来技術にはなかった特徴を持つ。京都市街地図上でのシミュレーションデータを用いて、文字列カーネルに基づく正規回帰という我々の手法が十分な予測能力を持つことを検証した。

今後の課題としては、まず、実データでの検証とシミュレーションの精緻化が挙げられる。我々のシミュレータは、自由流相と渋滞相との間のメタ安定状態を再現する能力を持つという意味で現実の交通流の特徴を再現できるが、今回の実験の設定

表 1: 異なるカーネル同士の r と平均 s^2 値の比較 ($\tau = 10$)。

	ID	direction	area
r	**0.980**	0.933	0.059
$\sqrt{\bar{s^2}}$	**4.5**	10.0	10.3

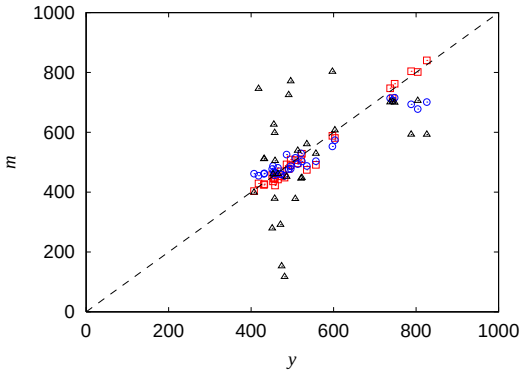

図 4: 予測旅行時間 (m) と実測値 (y) との比較。ID カーネルを □、方向カーネルを ◯、面積カーネルを △ で表す。

がどの程度現実的かは必ずしも明らかでない。また、理論面でも、データのスケーラビリティへの対応や、カーネル最適化手法の検討など興味深い課題がいくつか残されている。

参考文献

[Bishop 06] Bishop, C. M.: *Pattern Recognition and Machine Learning*, Springer-Verlag (2006)

[Idé 09] Idé, T. and Kato, S.: Travel-Time Prediction using Gaussian Process Regression: A Trajectory-Based Approach, in *Proc. SIAM Intl. Conf. Data Mining* (2009)

[Kriegel 08] Kriegel, H.-P., Renz, M., Schubert, M., and Zuefle, A.: Statistical Density Prediction in Traffic Networks., in *Proc. SIAM Intl. Conf. Data Mining*, pp. 692–703 (2008)

[Lee 07] Lee, J., Han, J., and Whang, K.-Y.: Trajectory Clustering: A Partition-and-Group Framework, in *Proc. 2007 ACM SIGMOD Intl. Conf. Management of Data*, pp. 593–604 (2007)

[Leslie 02] Leslie, C., Eskin, E., and Noble, W. S.: The spectrum kernel: A string kernel for SVM protein classification, in *Proc. the Pacific Symposium on Biocomputing*, pp. 564–575 (2002)

[Morikawa 07] Morikawa, T., Yamamoto, T., Miwa, T., and Wang, L.: Development and Performance Evaluation of Dynamic Route Guidance System PRONAVI, *Journal of the Japan Society of Traffic Engineers*, Vol. 42, No. 3, pp. 65–75 (2007)

[Nakata 04] Nakata, T. and Takeuchi, J.: Mining traffic data from probe-car system for travel time prediction, in *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, pp. 817–822 (2004)

[Vlachos 05] Vlachos, M.: Elastic Translation Invariant Matching of Trajectories, *Machine Learning Journal*, Vol. 58, No. 2-3, pp. 301–334 (2005)

[Yen 71] Yen, J. Y.: Finding the K Shortest Loopless Paths in a Network, *Management Science*, Vol. 17, No. 11, pp. 712–716 (1971)