社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

# ネットワーク上の経路に対する回帰問題について

井手　　剛†

† IBM 東京基礎研究所　〒 242-8502 神奈川県大和市下鶴間 1623–14
E-mail: †goodidea@jp.ibm.com

**あらまし**　空間を時系列的に移動する物体の軌跡、すなわちトラジェクトリに対するデータマイニングの問題は、実用的にも理論的にも興味深い研究テーマである。本稿では、ネットワーク上のトラジェクトリ（経路）のコスト予測問題を、経路に対する回帰問題として定式化する。この回帰問題はカーネル回帰の枠組みで扱うことが可能であるが、本稿では、その双対問題として別の定式化を導き、両者の関係を論ずる。

**キーワード**　トラジェクトリ回帰, グラフ回帰, ネットワーク, 旅行時間, グラフラプラシアン

# On regression problem for paths on networks

Tsuyoshi IDÉ†

† IBM Research – Tokyo　1623–14 Shimotsuruma, Yamato-shi, 242-8502 Kanagawa, Japan
E-mail: †goodidea@jp.ibm.com

**Key words**

## 1. はじめに

最近のセンシング技術の発展により、空間を移動する物体の軌跡を記録することが可能になってきた。最近は多くの携帯電話に搭載されている GPS（Global Positioning System）は、その最も身近な例である。移動体の軌跡、もしくはトラジェクトリの解析技術の開発は、新しい計測技術が可能にした新しい研究テーマである。伝統的な機械学習では、独立同一分布のベクトルの集合がデータとして主に想定されてきた。しかし、トラジェクトリは空間において連続的な曲線として表されるため、従来の学習手法の多くは何らかの修正を必要とする。

トラジェクトリからの知識発見は、データマイニングにおける最近の主要なテーマの一つとなっている。Gaffney と Smyth [4] によるトラジェクトリ・クラスタリングの研究は、トラジェクトリからの知識発見に言及した最初期の仕事の一つである。最近では、クラスタリング [5,12,16] のほか、分類 [14]、外れ値検出 [1,13]、密度推定 [10]、変化点検出 [21]、トラジェクトリ予測 [8,15] といった多彩なテーマが研究されている。

本稿では、トラジェクトリ回帰というタスクを考える。これは、過去データに基づいて、トラジェクトリがひとつ与えられた時のその「コスト」を返す関数を学習するものである。今のところ、トラジェクトリマイニングの文脈で回帰の問題を解いた研究は少なく、おそらく [7] は最初期の仕事の一つであると思われる。トラジェクトリといっても、制約がない空間での自由なトラジェクトリを数理的にきれいに定式化するのは簡単ではないので、本稿では、ネットワーク上に制約されたトラジェクトリを考える。この場合の典型的な応用は、地図上の経路に対する旅行時間予測である。とりわけ市街地で GPS で位置の追跡を行った場合、欠損値・異常値の問題は不可避なので、各リンク（隣接する交差点の間の道）個々のコストではなくて、経路の総所要時間のみをデータとして要求するトラジェクトリ回帰の定式化は、実用上好ましい特徴を持っている。

先行研究 [7] では、ネットワーク上のトラジェクトリを文字列で表し、文字列カーネルと共にカーネル回帰の枠組みで扱うという手法が提案されている。しかしそこでは、カーネル関数の最適な選び方が明らかではないこと、また、文字列カーネルで十分な予測精度を得るためには始点と終点を固定したようなトラジェクトリデータを想定せざるをえなかったこと、などの限界があった。

本稿ではこれらの限界を解決するために、カーネル回帰の枠組みの双対な定式化、すなわち、トラジェクトリの特徴ベクトルを明示的に構成するアプローチを考える。ネットワーク上のトラジェクトリは有向グラフとして表現できるので、このアプローチは、Tsuda や Saigo らにより精力的に研究されてきたグラフ回帰 [2,18,19] の特別な場合と考えることができる。ただし、彼らは主に化学構造を念頭において、グラフ回帰の方法論を追求してきたため、本稿で興味を持つ移動体の軌跡に対する回帰の問題が同等の方法で扱えるという保証は必ずしもない。本稿では、主に旅行時間予測という応用を念頭において、どのような定式化が可能かを考え、また、先行研究 [7] との理論的

関係について論じる。

## 2. トラジェクトリ回帰問題

トラジェクトリ回帰の問題では、通常の回帰問題と同様、訓練データとして $N$ 個のトラジェクトリとコストの組が与えられると仮定する

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

ここで $x^{(n)}$ は第 $n$ 番目のトラジェクトリで、$y^{(n)}$ はそれに対応するコストである。我々のゴールは、任意に与えたトラジェクトリのコストを予測することである。

問題を数理的に健全にするために、我々はトラジェクトリが、ネットワーク上に制約されていると仮定する。したがって、あるトラジェクトリは、リンク ID の系列として表現される。この際、あるネットワーク上のトラジェクトリのバラエティはほとんど無限にあるので、クエリとして与えられたトラジェクトリと同一、または非常に似ているトラジェクトリが訓練データの中にあるとは限らないことに注意する必要がある。したがって、一般にはいわゆる $k$ 最近傍回帰のような方法はうまく働かない。しかも一般には、リンクの交通量には著しい偏りがあり、訓練データの中に通過履歴がほとんどないリンクさえ大量にあるはずである。

## 3. 回帰モデル

本節では、基本的に旅行時間予測を念頭において、トラジェクトリ回帰問題の解法を提案する。

### 3.1 指標ベクトルによるコストの表現

旅行時間予測の場合、各リンクで消費される時間はリンクの長さと、その上での制限速度に依存する。通常、リンク長と制限速度は電子地図によって与えられているから、渋滞も何もない場合の所要コストはおおよそ計算できる。そこで、あらかじめ訓練データから基本コストを差し引いておく。すなわち

$$y^{(n)} \leftarrow y^{(n)} - \sum_{e \in x} l_e f_e^0$$

という変換をしておく。ここで $l_e$ はリンク（もしくは edge）$e$ の長さであり、$f_e^0$ は同じく単位長さあたりのコストである。これによって $y^{(n)}$ はベースラインのコストからのずれという意味を持つ。我々は例えば都市部において、渋滞等によりベースラインのコストからのずれが著しい場合に興味を持つ。そのずれをどのようにパラメトライズするかが問題である。ネットワーク上のトラジェクトリ回帰の場合、経路 $x$ に対するコストとして、

$$y = \sum_{e \in x} l_e f_e + \sum_{e \in x} \sum_{e' \in N(e)} f_{e,e'} + \sum_{e \in x} \sum_{e' \in N(e)} \sum_{e'' \in N(e')} f_{e,e',e''} + \cdots \tag{2}$$

のような表現を自然に想定できる。ここで、訓練データから決めるべき係数が $\{f_e, f_{e,e'}, f_{e,e',e''}, \cdots\}$ であり、$N(e)$ はリンク $e$ に接続されているリンクを表す。具体的に旅行時間予測について考えれば第 1 項は各リンク個別の寄与であり（混んでいる道は時間がかかる）、第 2 項は交差点での挙動についてのコスト（右折には時間がかかる）、第 3 項以降は通過リンクの履歴による何らかの効果を表す。上式は経路 $x$ を、長さ 1、2、3、、、の部分経路に分解した上で、それぞれについて係数を考えていることに対応する。今仮に、全部の係数をまとめた長いベクトルを $\boldsymbol{f}$ とし、経路 $x$ を部分構造に分解した際、それぞれの部分構造を含む場合に非ゼロ、含まない場合に 0 をとる指標ベクトル（indicator vector） $\boldsymbol{q}$ と表せば、上式は単に

$$y = \boldsymbol{f}^\top \boldsymbol{q}$$

のように書ける。

### 3.2 目的関数

回帰問題に対する標準的な処方に従って、我々は係数ベクトル $\boldsymbol{f}$ を、観測されたコスト $y^{(n)}$ とその予測値 $\boldsymbol{f}^\top \boldsymbol{q}^{(n)}$ との差がなるべく小さくなるように選びたい。我々は次の目的関数を最小化することを考える。

$$\Psi(\boldsymbol{f}|\lambda) = \sum_{n=1}^{N} \left( y^{(n)} - \boldsymbol{f}^\top \boldsymbol{q}^{(n)} \right)^2 + \lambda \sum_{e=1}^{M} \sum_{e'=1}^{M} S_{e,e'} \left| f_e - f_{e'} \right|^2 \tag{3}$$

ここで、第 1 項は通常の 2 乗誤差であり、第 2 項が、直感的には「渋滞しているリンクの隣のリンクも渋滞しているはずだ」という知識を表す正規化項である。$M$ はネットワークにおけるリンクの総数である。$S_{e,e'}$ はリンク同士の類似度を表す。先に述べたとおり、ネットワーク上でのトラジェクトリ回帰問題では、トラジェクトリの多様性が非常に高く、訓練データの中のトラジェクトリによりネットワークの全てが被覆されるということは期待できない。したがって、単に損失関数を考えただけでは多くの係数が目的関数に現れず、不定のままとなる。しかし、もし近隣のリンクに通過履歴があれば、それをもとにして、今着目しているリンクの状況についても何がしかの推測ができると想像することは自然である。そのような効果を表すのが第 2 項である。

リンク同士の類似度については、たとえば次のような決め方ができる。

$$S_{e,e'} \equiv \begin{cases} 1 & e = e' \\ \omega^{d(e,e')} & d(e,e') \leqq d_0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

ここで $\omega$ は 1 未満の定数であり、指数 $d(e,e')$ はリンク $e$ と $e'$ の間の何らかの距離尺度である。もっとも素朴には、$e$ から $e'$ に到達するためのホップ数を使うことができる。カットオフ値 $d_0$ の決め方や $\omega$ の選び方には任意性があるが、我々の実験によれば、これらへの結果の依存性は弱いことがわかっている。

なお、指標ベクトルを特徴ベクトルとした線形のモデルを想定する点は Tsuda らのグラフ回帰と同様であるが、化学構造を対象にしたグラフ回帰と異なり、各部分構造、少なくとも 1 次と 2 次の項に関しては損失関数に直感的な意味が付けられる

点が面白いところである。また、正則化項に関しても、単にスパース化のための便法というよりは、「ここが渋滞しているのなら回りも渋滞しているはずだ」という直感的な信念に対応している点が面白い。

### 3.3 目的関数の行列表現

さて、上に導入した目的関数を行列表示することを考えよう。これにより式が非常に見通しがよくなる上、後段の理論解析が非常に容易になる。今、簡単のため、式 (2) において2次以上の項を無視して、主要項と思われる第1項のみを残す。すると指標ベクトルとしては、

$$q_e^{(n)} = \begin{cases} l_e, \text{ for } e \in \boldsymbol{x}^{(n)} \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

のようなものになる。この時、さらに、

$$\mathsf{Q} \equiv [\boldsymbol{q}^{(1)}, ..., \boldsymbol{q}^{(N)}] \in \mathbb{R}^{M \times N} \tag{6}$$

と定義すれば、簡単な計算から、我々の目的関数が次のように書けることがわかる。

$$\Psi(\boldsymbol{f}|\lambda) = \left|\left|\boldsymbol{y}_N - \mathsf{Q}^\top \boldsymbol{f}\right|\right|^2 + \lambda \boldsymbol{f}^\top \mathsf{L} \boldsymbol{f} \tag{7}$$

ただし、$\mathsf{L}$ は類似度行列 $\mathsf{S}$ から導かれるグラフラプラシアンであり、$\mathsf{L}_{i,j} \equiv \delta_{i,j} \sum_{k=1}^{M} S_{i,k} - S_{i,j}$, と定義される。ラプラシアンの存在以外は、この目的関数はリッジ回帰のそれと同様であり、次の連立方程式を解くことで容易に解が求まる。

$$\left[\mathsf{Q}\mathsf{Q}^\top + \lambda \mathsf{L}\right] \boldsymbol{f} = \mathsf{Q}\boldsymbol{y}_N \tag{8}$$

幸い、$\mathsf{Q}\mathsf{Q}^\top + \lambda\mathsf{L}$ は高度に疎であることが期待されるので、共役勾配法 [6] などの反復法を用いることで、非常に効率よく解が求められる。$\lambda$ は交差検証法で決める。

## 4. カーネル回帰との関係

我々はトラジェクトリ回帰の問題がリッジ回帰として解けることを示した。ここで、正規過程回帰を用いた解法 [7] との関係を調べることは興味深い。まず、明らかに次の性質が成り立つ。

［Proposition 1］ 式 (7) を最小化する最適化問題は、コスト $y$ のノイズ分布を、$\boldsymbol{q}^\top \boldsymbol{f}$ を平均とする正規分布とした時、次の事前分布による MAP (maximum a posteriori) 推定と等価である。

$$p(\boldsymbol{f}) \equiv \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \mathsf{L}^{-1}) \tag{9}$$

事前分布式 (9) はいわゆる improper な分布であるが、半教師あり学習タスクで空間的な連続性を表すためにしばしば使われるものである（たとえば [9]）。

さて、この事前分布の下でのベイズ的な予測分布を求めてみよう。$y$ のノイズ分散を $\sigma^2$ とし、訓練データの観測コストを $N$ 個並べたベクトルを $\boldsymbol{y}_N$ とすると、通常の正規過程回帰の定式化に従って [17]、予測分布が

$$p(y|x, \boldsymbol{y}_N) = \mathcal{N}\left(y \left| \boldsymbol{k}_q^\top \mathsf{C}_q^{-1} \boldsymbol{y}_N, \sigma^2 + k_q - \boldsymbol{k}_q^\top \mathsf{C}_q^{-1} \boldsymbol{k}_q \right.\right) \tag{10}$$

となることを示せる。ただし、

$$\boldsymbol{k}_q \equiv \mathsf{Q}^\top \mathsf{L}^{-1} \boldsymbol{q}, \quad k_q \equiv \boldsymbol{q}^\top \mathsf{L}^{-1} \boldsymbol{q}$$
$$\mathsf{C}_q \equiv \sigma^2 \mathsf{I}_N + \mathsf{K}_q, \quad \mathsf{K}_q \equiv \mathsf{Q}^\top \mathsf{L}^{-1} \mathsf{Q}$$

などである。これらの式を通常の正規過程回帰の表現と比較すれば、次の事実が直ちに導ける

［Proposition 2］ 前節の定式化は、$(n, n')$ 成分が次で与えられるカーネルを持つ正規過程回帰の予測平均を与える。

$$\mathsf{K}_{n,n'} = \boldsymbol{q}^{(n)\top} \mathsf{L}^{-1} \boldsymbol{q}^{(n')} \tag{11}$$

この結果は、もしリンク同士の類似度が自明に求められるならば（これは通常妥当な想定である）、[7] で使われた文字列カーネルは最適な選択ではないということを示している。すなわち、何か恣意的なカーネルを選んでしまうと、ネットワークのトポロジーと矛盾した類似度行列を与える可能性がある。これは実験精度を向上させる上で非常に重要な示唆を与える。

その他、上記のようなカーネルの表現に基づけば、トラジェクトリ同士の commute time [3, 11, 20] という概念を導くことができて興味深いのだが、詳細は別論文に譲る。

## 5. ま と め

本稿では、ネットワーク上のトラジェクトリ回帰という問題に対する新しい定式化を提案した。カーネル回帰の枠組みとの対応を考えることで、最適なカーネルが何かという問題に対して実用上有益な知見が得られた。

本稿で触れなかったさらにいくつかの理論的話題について、はまた稿を改めて論じたい。実トラジェクトリデータに基づく実験結果についてもまた別論文で議論したい。なお、実験結果について言えば、恣意的に文字列カーネルを選択した場合よりも、本稿の方法の方が圧倒的な高精度を与える。

## 謝　　辞

本研究の一部は、総務省の地球温暖化対策 ICT イノベーション推進事業 (PREDICT) の助成により行われました.

## 文　　献

[1] Y. Bu, L. Chen, A. W.-C. Fu, and D. Liu. Efficient anomaly monitoring over moving object trajectory streams. In *Proc. of the 15th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD 09)*, pages 159–168, New York, NY, USA, 2009. ACM.

[2] S. Chiappa, H. Saigo, and K. Tsuda. A bayesian approach to graph regression with relevant subgraph selection. In *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM 2009)*, pages 295–304, 2009.

[3] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007.

[4] S. Gaffney and P. Smyth. Trajectory clustering with mixtures of regression models. In *Proc. the fifth ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD 99)*, pages 63–72, New York, NY, USA, 1999. ACM.

[5] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *Proc. of the 13th ACM SIGKDD Intl. Conf. on Knowledge discovery and Data Mining (KDD*

*07)*, pages 330–339, New York, NY, USA, 2007. ACM.

[6] G. H. Golub and C. F. V. Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, 1996.

[7] T. Idé and S. Kato. Travel-time prediction using Gaussian process regression: A trajectory-based approach. In *Proc. SIAM Intl. Conf. Data Mining*, pages 1183–1194, 2009.

[8] N. Jetchev and M. Toussaint. Trajectory prediction: learning to map situations to robot trajectories. In *Proc. of the 26th Intl. Conf. on Machine Learning (ICML 09 )*, pages 449–456, New York, NY, USA, 2009. ACM.

[9] A. Kapoor, Y. A. Qi, H. Ahn, and R. Picard. Hyperparameter and kernel learning for graph based semi-supervised classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 627–634. MIT Press, Cambridge, MA, 2006.

[10] H.-P. Kriegel, M. Renz, M. Schubert, and A. Zuefle. Statistical density prediction in traffic networks. In *Proc. SIAM Intl. Conf. Data Mining*, pages 692–703, 2008.

[11] J. Kunegis and A. Lommatzsch. Learning spectral graph transformations for link prediction. In *Proc. International Confernence on Machine Learning*, pages 561–568, 2009.

[12] J. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: A partition-and-group framework. In *Proc. 2007 ACM SIGMOD Intl. Conf. Management of Data*, pages 593–604, 2007.

[13] J.-G. Lee, J. Han, and X. Li. Trajectory outlier detection: A partition-and-detect framework. In *Proc. of the 2008 IEEE 24th Intl. Conf. on Data Engineering (ICDE 08)*, pages 140–149, Washington, DC, USA, 2008. IEEE Computer Society.

[14] J.-G. Lee, J. Han, X. Li, and H. Gonzalez. Traclass: trajectory classification using hierarchical region-based and trajectory-based clustering. *Proc. VLDB Endow.*, 1(1):1081–1094, 2008.

[15] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proc. of the 15th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD 09)*, pages 637–646, New York, NY, USA, 2009. ACM.

[16] N. Pelekis, L. Kopanakis, E. Kotsifakos, E. Frentzos, and Y. Theodoridis. Clustering trajectories of moving objects in an uncertain world. In *Proc. of the 2009 Ninth IEEE Intl. Conf. on Data Mining (ICDM 09)*, pages 417–427, Washington, DC, USA, 2009. IEEE Computer Society.

[17] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[18] H. Saigo, N. Krämer, and K. Tsuda. Partial least squares regression for graph mining. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

[19] H. Saigo, S. Nowozin, T. Kadowaki, T. Kudo, and K. Tsuda. gboost: a mathematical programming approach to graph classification and regression. *Machine Learning*.

[20] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens. Graph nodes clustering based on the commute-time kernel. In *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007). Lecture notes in Computer Science, LNCS*, volume 4426, pages 1037–1045, 2007.

[21] H. Yoon and C. Shahabi. Robust time-referenced segmentation of moving object trajectories. In *Proc. of the 2008 Eighth IEEE Intl. Conf. on Data Mining (ICDM 08)*, pages 1121–1126, Washington, DC, USA, 2008. IEEE Computer Society.