

竹内 純一 (TAKEUCHI Jun'ichi)

## 研究概要

機械学習の基礎と実応用の両面に取り組んでいる。機械学習 (マシン・ラーニング) とは、機械 (コンピュータ) に人間のような学習能力を持たせることを目指す技術である。大きく分けると、与えられたデータからいかに多くの情報を取り出すかという情報論的側面と、いかに速く行うかという計算論的側面があるが、特に前者に着目した研究 (情報論的学習理論) を行っている。

理論については、MDL 原理 (Minimum Description Length Principle; 記述長最小原理) と情報幾何が柱である。MDL 原理は、1978 年に Rissanen が提唱した「学習とはデータ (記号列) をコンパクトな法則に圧縮する過程である」とする考え方であり、情報理論におけるユニバーサル符号と密接な関連がある。特に、情報源 (確率分布) の範囲のある集合 (統計モデル) に限定したときの符号長の限界を確率的コンプレキシティ (Stochastic Complexity; SC) と呼ぶ。例えば、情報源を一次の定常 Markov 連鎖全体の集合  $M$  に限定すると、記号列について  $M$  に関する SC が定義される。これはモデル  $M$  を用いた学習の性能限界を示す指標であると同時に、モデル選択のための情報量基準として用いることができる。したがって、SC の性質を調べることにより、モデルを用いた機械学習問題の情報量的な性質を知ることができる。一方、モデルの性質を、情報幾何の立場から調べることもできる。情報幾何は確率分布を点とする空間の微分幾何的な性質を調べる方法論である。これら二つの観点を基準として、様々なモデルの性質を調べている。応用については、ネットワークセキュリティにおける異常検知を中心に、SC の考え方を基盤としながら多様な実応用に取り組んでいる。

### 1. Markov 連鎖モデルの SC

ベイズ混合の方法により、SC の値を定数オーダーまで評価した。モデルが i.i.d. (独立同分布) の指数型分布族の場合には、Jeffres 事前分布を用いた Bayes 混合 (Jeffreys 混合) が SC が定数オーダーまで達成することが知られていた。Markov 連鎖モデルは i.i.d. ではないが、指数型分布族であることが知られていた。これに着目し、Markov モデルの場合も Jeffreys 混合が定数オーダーまで SC を達成することを示した。

### 2. 木情報源モデルの幾何学

Markov 連鎖モデルの状態集合について木構造を仮定することにより、その部分空間を定義することができる。このように定義した情報源モデルを木情報源モデルと呼ぶ。木構造の深さが一様な場合は、通常の Markov 連鎖モデルに対応する。前述のように、Markov 連鎖モデルは指数型であるが、さらに一般には、FSM (Finite State Machine) モデルが指数型であることが知られている。FSM モデルは、有限オートマトンによって定義される情報源モデルである。一般の木情報源モデルは FSM ではないが、木構造がある条件を満たせば FSM となる。例えば木の深さが一様な場合は FSM となる。FSM となる木情報源モデルを FSMX モデルと呼ぶ。FSMX でない木情報源モデルが指数型かどうかは不明であったが、指数型でないことを証明した。すなわち、木情報源モデルについては、指数型であることと FSMX であることが同値である。これは情報幾何の観点からは、FSM でない木情報源モデルがゼロでない埋め込み指数曲率をもつことを意味する。また、SC

の観点では、指数型でないモデルでは、Jeffreys 混合は SC を達成出来ないため、FSMX でない木情報源モデルでは、Jeffreys 混合は SC を達成しないことを意味する。

### 3. Markov 連鎖モデルの体積要素

SC の定数項は、モデルの体積の対数に等しいことが以前から知られている。ここの体積は、Fisher 計量から自然に誘導される体積要素の積分である。したがって、SC の解析的評価のためには、Fisher 情報行列の行列式を求めれば良い。Markov 連鎖モデルの場合、Markov カーネルパラメータを用いた場合に行列式の計算は容易であるため、前項の研究ではこれを用いている。ところが、情報幾何の観点ではより重要な期待値パラメータ（混合接続に関するアフィンパラメータ）の Fisher 情報行列については、その形が複雑であるため行列式の表式が知られていなかった。この問題に対して、Markov 連鎖モデルの拡大モデルの Fisher 情報行列を利用することで、この形を求めることに初めて成功した。

### 4. 統計多様体の拡大モデルの研究

前項の研究を行う過程で、Markov 連鎖モデルの拡大モデルとその Fisher 情報行列を定義した。ところが、この Fisher 情報行列は退化しており、奇妙な構造となっていることが判明した。これは、i.i.d. の場合とはまったく様相が異なる。すなわち、i.i.d. (多項 Bernoulli モデル) の場合、拡大モデルは非正規化 (denormalization) という操作で定義され、Fisher 情報行列は対角行列となる。この場合も、これを利用して元のモデルの体積要素を求めることができる。しかし、Markov 連鎖モデルの場合は拡大モデルの体積はゼロであるため、求め方が全く異なる。さらに、i.i.d. の場合、期待値パラメータに関する体積要素の表式が、元のモデルと拡大モデルとで同じ形になるが、この命題の幾何学的解釈を与えた。

### 5. 学習応用

ネットワークセキュリティの分野で、ポットネット検知という問題をターゲットに研究を進めている。ポットとは、マルウェア（悪意のあるプログラム）の一種であり、知らないうちに多くの端末に潜っているとされる。これは普段は何もアクションを起こさず、コマンダーからの命令を受けてスキャン（脆弱性を探す調査行動）や DoS 攻撃（サービス不能攻撃）を行う。ポットネットとは、このようなマルウェアが構成するネットワークのことである。ポットネットは普段は何もしないため検出が困難とされ、この検出はセキュリティ上重要な課題である。本研究では、インターネット上にあるホストからダークネット（実ホストが割り当てられていない IP アドレス群）に送信されるパケット数等に基づく手法を検討している。まず、これら時系列に現れる異常を検知する手法を検討した。特に、パケット数等の時系列に長期記憶性があり、ウェーブレットや長期記憶モデルが有効であることを実験的に確認した。さらに、複数のホスト間の行動の相関を考慮し、ポットネット出現時に生じる変化を定量化して検出する手法を開発している。そのために、多数の確率変数間の疎なネットワークを学習する「スパース構造学習」の手法（モデル選択の一種）の適用し、実証実験を行っている。スパース学習は最近の機械学習分野で盛んに研究されているテーマである。本研究は、独立行政法人情報通信研究機構インシデント対策グループとの共同研究である。また、ITS (Intelligent Transport Systems; 高度交通システム) における旅行時間予測についても、長期記憶性を考慮することが重要であることを確認した。これは NEC

との共同研究である．

#### 「マス・フォア・インダストリ」にかかわる H20, 21 年度の研究実績概要

基礎研究として，有限アルファベットの Markov 連鎖モデルを対象に，その情報幾何学的構造と確率的コンプレキシティ(SC) について考察を進めた．i.i.d.(独立同分布) の場合と同様，Markov 連鎖モデルでも，Fisher 計量から誘導されるモデルの体積が SC の定数項に現れることが分かっており，その値は Markov カーネルパラメータによる積分表示で得られていた．今期間には，従来は組み合わせ論を用いて間接的に得られていた期待値パラメータと自然パラメータによる積分表示を，幾何学的技法のみで得ることに成功した．その過程で，Markov 連鎖モデルの拡大モデルを導入し，i.i.d. の場合と構造が全く異なることを明らかにした．また，応用としては，主にネットワークセキュリティにおけるインシデント検知と ITS(Intelligent Transport Systems; 高度交通システム) における旅行時間予測に取り組み，これらの対象ではウェブレットや長期記憶モデルが有効であることを示した．

#### 研究業績

- 1.D. Inoue, K. Yoshioka, M. Eto, M. Yamagata, E. Nishino, J. Takeuchi, K. Ohkouchi, K. Nakao, “An Incident Analysis System nicter and Its Analysis Engines Based on Data Mining Techniques,” *Proc. of the 15th International Conference on Neural Information Processing*, pp. 268-269, November 2008.
2. 柿原雄介, 竹内純一, 藤田貴司, 姚恩建, 岡大雅, “ウェブレット変換を用いた旅行時間時系列の予測,” 電子情報通信学会技術研究報告, IT2009-67, pp. 151-156, March 2009.
3. 西野瑛介, 竹内純一, 吉岡克成, 井上大介, 衛藤将史, 中尾康二, “独立成分分析を用いたインシデント予測手法の検討,” 電子情報通信学会技術研究報告, IT2008-92, pp. 315-320, March 2009.
4. 北川潤也, 竹内純一, 吉岡克成, 井上大介, 衛藤将史, 中尾康二, “ネットワークトラフィックデータ間の相関に基づくインシデント検知の検討,” 電子情報通信学会技術研究報告, IT2008-93, pp. 321-328, March 2009.
- 5.J. Takeuchi, “Fisher Information Determinant and Stochastic complexity for Markov Models,” *Proc. of 2009 IEEE International Symposium on Information Theory*, pp. 1894-1898, 2009.
6. 村上慎太郎, 濱崎浩輝, 川喜田雅則, 竹内純一, 吉岡克成, 井上大介, 衛藤将史, 中尾康二, “確率的依存関係に基づくボットネット検知の検討,” 電子情報通信学会技術研究報告 IA2009-1, pp. 1-6, June 2009.
- 7.J. Takeuchi, “On volume element of Markov chain models,” 第 7 回シャノン理論ワークショップ予稿集, pp. 29-34, September 2009.
8. 北川潤也, 川喜田雅則, 竹内純一, 吉岡克成, 井上大介, 衛藤将史, 中尾康二, “NMF を用いたボットネット検出の検討,” 2010 年暗号と情報セキュリティシンポジウム予稿集, January 2010.
9. 川喜田雅則, 竹内純一, “ラベル無しデータを用いた回帰の改良,” 第 12 回データマイニングと統計数理研究会, March 2010(予定).

#### 講演

1. “マルコフモデルの幾何学について,” 大阪市立大学数学研究所 情報幾何学研究集会 2009, 大阪, January, 2009.

2. “マルコフ連鎖の幾何と確率的コンプレキシティ,” 研究集会「log P の情報学」, 福岡, June 2009.
3. “統計多様体の体積要素について,” 情報とダイナミクス III 研究集会, 福岡, October 2009.
4. “最尤符号とベイズ符号,” 研究集会「ベイズ統計への情報理論的アプローチとその周辺」, 東京, December 2009.

#### 学位

博士(工学)(東京大学)

#### 受賞歴

- ・ SITA97 奨励賞, 情報理論とその応用学会, 1998.
- ・ 先端技術大賞フジサンケイビジネスアイ賞 (山西健司氏, 丸山祐子氏と共同受賞), 2005.

#### 研究集会の主催

1. 研究集会「log P の情報学」オーガナイザー (田中利幸氏と共同), June 2009.
2. 第 12 回情報論的学習理論ワークショップ 実行委員長, October 2009.

#### その他特記事項

- ・ 電子情報通信学会和文論文誌 A(基礎・境界), 編集委員, 2005-2009.
- ・ 電子情報通信学会情報論的学習理論時限研究専門委員会 委員長, 2009-2010.
- ・ 情報理論とその応用学会 理事, 2008-2010 .
- ・ 情報理論とその応用学会 評議委員, 2010-2012 .
- ・ 小西貞則, 竹内純一 (著), 若山正人 (編), 統計的モデリング/情報理論と学習理論, 講談社, September 2008.