「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」
平成 25 年度採択研究代表者

H28 年度
実績報告書

山西　健司

東京大学大学院情報理工学系研究科
教授

複雑データからのディープナレッジの発見と価値化

§1. 研究実施体制

(1)「山西」グループ
① 研究代表者:山西　健司(東京大学　大学院情報理工学系研究科、教授)
② 研究項目
・ディープナレッジのモデル論、推定論の構築

(2)「増田」グループ
① 主たる共同研究者:増田　直紀　(ブリストル大学 Department of Engineering Mathematics, Senior Lecturer)
② 研究項目
・ディープナレッジとしてのテンポラル・ネットワークの解析理論の構築推進

(3)「IBM」グループ
① 主たる共同研究者:恐神　貴行　(日本アイ・ビー・エム株式会社東京基礎研究所、リサーチスタッフメンバー)
② 研究項目
・ディープナレッジを価値につなげるための意思決定最適化技術

(4)「大澤」グループ
①主たる共同研究者:大澤　幸生(東京大学　大学院工学系研究科、 教授)
②研究項目
・ディープナレッジの利用価値を創造するデータ市場の構築手法

§2. 研究実施の概要

従来の BigData 研究はデータの大量性に関心が集中してきた。しかし、本研究では、BigData の複雑さ、多様性、変動性に注目し、巨大なデータの背後に眠る潜在知識（これを「ディープナレッジ」とよぶ）を発見し、価値を与えるための方法論を開発することを目的にしている。

本研究チームは、4 つのグループ（山西 G、増田 G、IBMG、大澤 G）に分かれて研究している。

山西 G では、データの背後に潜む変化の兆候を検知するための新たな方法論を開発した。従来の変化検知は、突発的変化を検知することを問題にしていた。しかし、実際には、変化は徐々に起きる場合が多い。そこで、本研究では、そのような**漸進的な変化の検知問題**に取り組んだ。これに対して、世界で初めて「**MDL（minimum description length）統計量**」に基づく変化検知手法を提案した。MDL 統計量は、ある時点の前後で分割してデータを記述した場合と、分割しないでデータを記述した場合の符号長の差として定義され、その時点の変化スコアと見なすことができる。固定長の窓枠で MDL 統計量を計算して、窓枠をスライドすることにより、逐次的な変化検知を実現した。仮説検定の枠組みを用いて MDL 変化統計量による変化検知の誤り確率を導いた。さらに、工場のセンサーデータから事故の予兆変化検知に適用して、その有効性を検証した。本研究成果はビッグデータのトップ国際会議 IEEE BigData 2016 にて採択され、発表した。

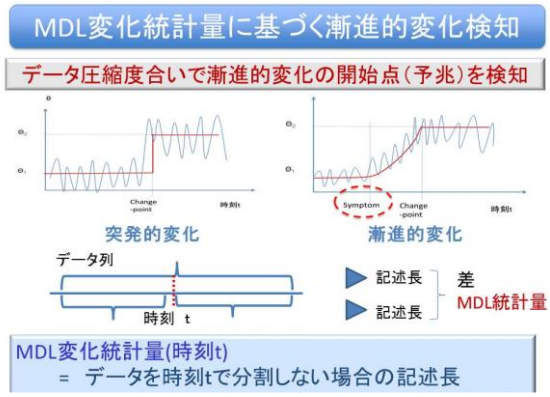

図1. MDL 変化統計量に基づく漸進的変化検知

増田 G では、時間的に構造変化するネットワークである**テンポラルネットワーク（TN）**の研究を行っている。TN 上の感染現象の理解することは、現実の感染症やオンライン上の情報伝播を予測・制御するために重要である。しかしながら、これまでは TN 上で感染現象を解析した研究は少なかった。今年度は、TN を隣接行列としてモデル化し、その上で SIS（susceptible infected susceptible）型の感染ダイナミクスを解析した。特に、臨界感染率（それより大きいと感染がネットワーク全体に広まりうるような感染率の閾値）を解析的に評価した。具体的には、各時刻でのネットワークが小さい構成要素に分離してしまっていないという条件のもとで、TN の臨界感染率は静止したネットワークの場合のそれよりも小さいことを示した。すなわち、テンポラルネットワーク上では、静的なネットワーク上よりも感染が広まりやすいことを明らかにした。本成果は New Journal of Physics 誌に掲載された。

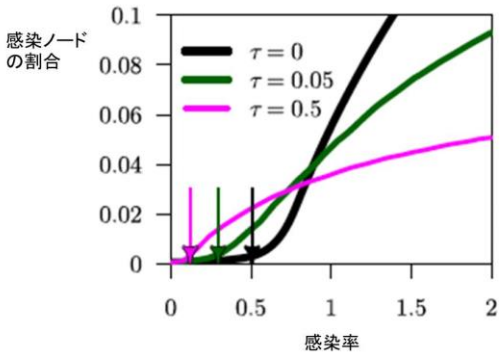

図２：SIS モデルの臨海感染率の評価： τ はネットワークの動的度合いを表し、τ が大きいほど小さい感染率でも感染が速い。矢印は臨海感染率の理論値を表す。

IBM G では、不確実な環境化における逐次的意思決定を支援するための意思決定最適化技術に取り組んでいる。2 値の時系列データの生成モデルを学習する**動的ボルツマンマシン(DyBM)**と呼ばれる人工ニューラルネットワークモデルを昨年度に提案したが、本年度は特に、実数値の時系列データを取り扱えるように DyBM を拡張し、強化学習による DyBM の逐次的意思決定への応用の可能性を切り開いた。実数値を取り扱えるように拡張した DyBM の性能を評価するために、複数の実データを用いて予測精度や学習時間を既存手法(現在標準的な再帰的ニューラルネットワークである LSTM など)と比較した(図3)。特に、DyBM の学習時間は LSTM の16分の 1 程度であり、バック・プロパゲーションを必要としない DyBM の学習効率の良さが顕著に確認された。本成果は人工知能のトップレベルの国際会議 AAAI-17 で発表した。

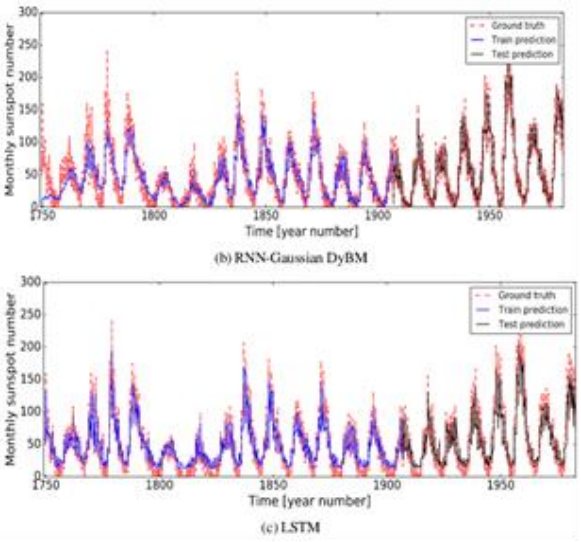

図3　太陽の黒点の数の DyBM(上段)と LSTM(下段)による予測結果。訓練データに対する予測を青線で、テストデータに対する予測を黒線で示す。DyBM の予測精度は LSTM と比べて遜色ないが、DyBM の学習時間は LSTM の 16 分の 1 程度である。

大澤 G では社会的要求とデータの概要情報(Data Jackets: DJ)の間をデータ利活用シナリオによって繋げる研究を行っている。従来はネットワーク状に DJ 間の関係を可視化する技術の基礎と位置づけてきたが、それでは DJ 間の局所的な線的関係しか表現できず、トップダウンに思考しようとするユーザには大局的な情報を与えないという問題があった。そこで、新たに地形図状の DJ マップを作成し、近接すべき DJ の共通の文脈を広がりのある面として表す方法として **Recut** を開発した。Recut は、関係性のマトリクスにおいて隣接する2列の内容の違いの大きな場合にその2列の間で行列を分離し、左右の順序を変えて再結合するという方法で、「隣り合う列は似ている」ようなマトリクスを生成する。Recut は変数の数を n として O(n log n)のオーダーで効率的に実行できることを示した。 また、Recut はネットワーク上の類似したものの配置を効果的に可視化するだけでなく、一般に、多く変数の集まった領域に高い標高を与える地形図状の可視化を与えることに成功した。本成果は Fundamenta Informaticae 誌に掲載された。

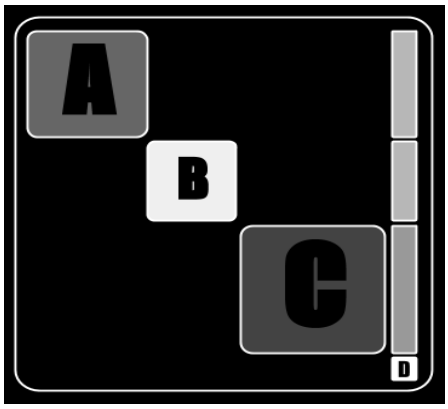

図4. A,B,C の部分が変数群のまとまりで、その接合点に変数の結合可能性が可視化される。データが増えると動的にこの図が変化してゆく。

代表原著論文

1. Kenji Yamanishi and Kohei Miyaguchi, ″Detecting gradual changes from data stream using MDL-change statistics,″. IEEE Iternational Conference on BigData 2016:(BigData2016), pp:156-163, 2016.

2. Leo Speidel, Konstantin Klemm, Víctor M. Eguíluz and Naoki Masuda, “Temporal interactions facilitate endemicity in the susceptible-infected-susceptible epidemic model”, New Journal of Physics, vol. 18, p.073013, 2016.

3. Sakyasingha Dasgupta and Takayuki Osogami, Nonlinear dynamic Boltzmann machines for time series prediction, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), pages 1833-1839, 2017.

4. Qi Ji and Yukio Ohsawa,Recut: a seriation algorithm balancing smooth display and aggregated features, Fundamenta Informaticae vol.146, no. 3, pp. 293-304 (2016)