「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

H27 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関係性を抽出して予測に役立てる研究を行っている。今年度は、東大病院の協力の下、緑内障診断の分野に適用して効果を実証した。 特に、緑内障進行予測の問題は重要であるが、従来は個人の視野データのみを用いていたので、そのデータ数が十分でないことから精度の良い予測ができなかった。本研究では、患者間の潜在的関係性を有効活用し、かつ眼圧データを世界で初めて活用した新しい緑内障進行予測手法を開発した。これによって従来の視野データのみを用いる予測に比べて有意に高い予測精度を実現した。鍵となる技術は、眼圧に基づく区分的線形回帰モデルを記述長最小（Minimum Description Length(MDL)原理に基づいて最適構成したことである（図1）。また、同様に MDL 原理に基づいて、視野欠損パタンを抽出し、臨床医学的にも知られる典型的な進行パタ-ンの存在をデータから裏付けることにも成功した。本成果はデータマイニング及びヘルスケア分野のトップ国際会議である KDD2015 と Healthinf2016 で発表した。

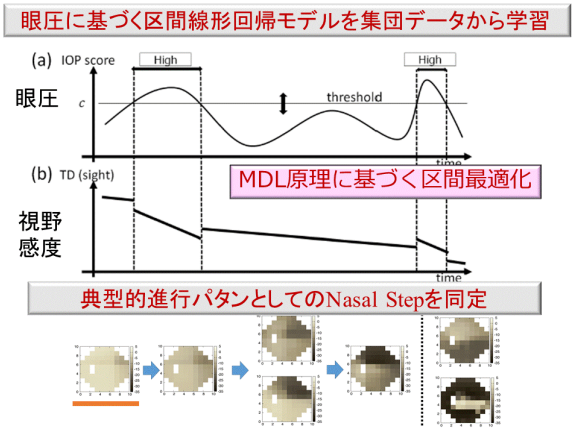

図 1．緑内障進行予測と緑内障パタン発見。

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。今年度は、TN 上の感染動態モデルに対する個体ベース近似理論と呼ばれる記述方法を開発し、感染源特定の問題に応用した。個体ベース近似は、各頂点が感染している確率や治癒した確率を変数とし、その時間発展を多次元の確率ベクトルの時間発展として定式化する半解析的手法である。この理論を、SIR モデルと呼ばれる、一過性、流行性の感染動態を表す標準的なモデルに対して開発した。提案手法は、単純に SIR モデルを数値計算した結果を精度よく近似することが明らかになった（図2）。また、提案手法は感染源特定問題において高い精度で感染源を特定することができることを示した。

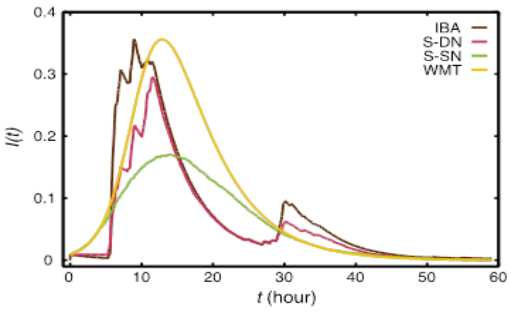

図 2． SIR モデルにおける感染人数の時間変化。茶（IBA）：個体ベース近似理論。赤（S-DN）：数値計算結果。黄緑（S-SN）：静的ネットワーク。黄：平均場理論。IBA が S-DN を最もよく近似している。

IBM G では、行動データを対象とするディープナレッジの抽出と活用の研究に取り組んでいる。今年度は、とくに、行動データのモデルとして動的ボルツマンマシン（DyBM）という新しい人工ニューラルネットワークモデルを提案した（図3）。従来、静的で複雑な情報を扱うモデルとして静的ボルツマンマシンが提案されており、画像のような情報をモデリングするのに用いられてきた。これに対して DyBM は動画のような動的な情報を取り扱えるモデルである。本研究では、DyBM のパラメータ更新規則として、学習データに対する尤度が大きくなるような学習則を導出した。従来のボルツマンマシンに対する学習則は神経回路のヘブ則に対応しており、スパイクの時刻に依存しない学習則であった。一方で、今回導出した学習則は、神経回路のスパイク時刻依存シナプス可塑性（STDP）に相当し、これが持つ長期増強や長期抑圧などの代表的な性質を有することが判明した。本成果は Scientific Reports 誌に掲載された。

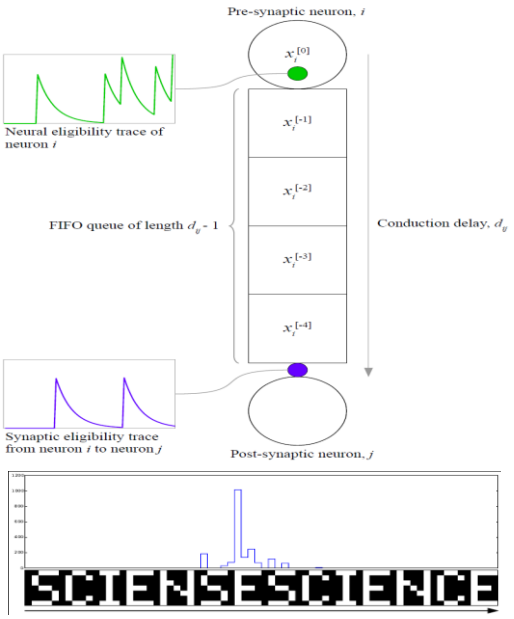

図３．動的ボルツマンマシン（DyBM）。先入れ先出し（FIFO）列とエリジビリティトレースを持つ DyBM の構造。DyBM が文字列”SCIENCE”のビットパターンを学習し異常を検知する。

大澤 G では、データ活用シナリオの提案と評価を行う議論の枠組み：Innovators Marketplace on Data Jackets（IMDJ）の技術的基礎を確立し、応用展開している。IMDJ とは、データの特色を端的にまとめて記載したデータジャケット（DJ）を共有し、DJ 間の関係を可視化することによって、データ利用指針を提案し評価し合うコミュニーションの仕組みである。今年度は、IMDJ における過去のデータ活用案と要求を登録し、容易に検索することで今後に生かすシステム（DJ ストア、ResourceFinder）を構築した。また、IMDJ で示されたデータ利用要求に応えるツールとして、昨年度より開発している Tangled String と呼ばれる時系列解析ツールを発展させ、それと連動して変化検知を行うツール及び、サプライズ現象を可視化する非同調変化検知アルゴリズムを開発した。さらに、応用領域としては、安全都市生活の設計、及びスーパーマーケットにおける食品のマーケティングで効果を実証した。

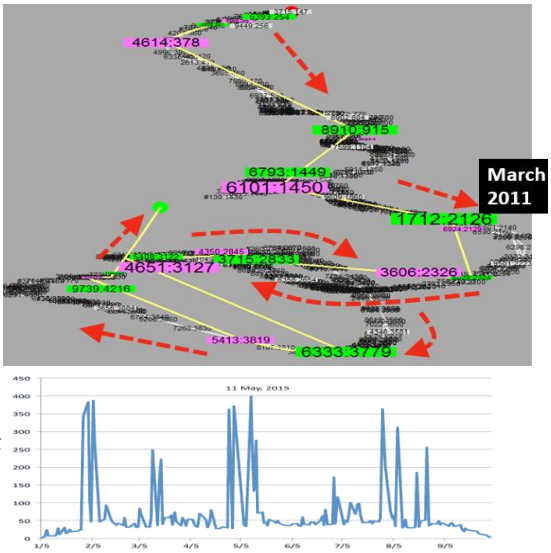

図４．（上段）Tangld String による時系列解析：市場変化の切れ目を検知（下段）非同調変化検知アルゴリズムによる不祥事検知　　ピークが不祥事の発生に対応

H27 年度の代表的原著論文

Takayuki Osogami and Makoto Otsuka, Seven neurons memorizing alphabetical images via spike-timing dependent plasticity, Scientific Reports, 5, 14149, 2015. (C1-1)

Naoki Masuda, ″Accelerating coordination in temporal networks by engineering the link order″, Scientific Reports, vol. 6, 22105, 2016. (B1-9)

Shigeru. Maya, Kai. Morino, Hiroshi Murata, Ryo. Asaoka, Kenji. Yamanishi: ″:Discovery of glaucoma progressive patterns using hierarchical MDL-based clustering.,″ in Proceedings of 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2015), pp:1979--1988, Aug. 2015.(A1-4)