「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

H26 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研究している。

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、1）潜在的ダイナミクス と 2）関係データ統合予測 の研究を行っている。1）潜在的ダイナミクスとは、観測時系列データの背後にあるディープナレッジの変化を検知することにより、重大なイベントの兆候を検知する研究である。2）の関係データ統合予測とは、多様なデータを関係づけるディープナレッジを抽出し、欠損したデータの予測に活用する研究である。今年度は、関係データ統合予測を緑内障進行予測に応用した。従来の予測では、対象患者の緑内障の進行を予測するのにデータが十分多くなく、高精度予測ができないことが問題であった。そこで、今年度はマルチタスク学習と呼ばれる機械学習の手法を用いて、似ている患者のデータを有効活用して高精度予測を行う方式を新しく開発した(図1)。東大病院が集めた緑内障患者の視野測定データに対し、対象患者だけを用いる方式に比べ、検査回数2回にて 80%以上の予測精度を改善した。

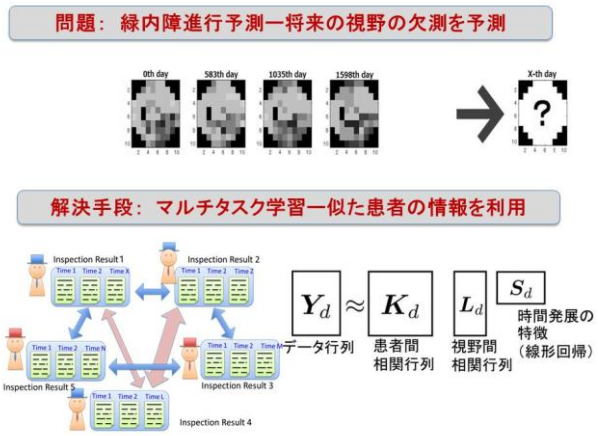

図１． 緑内障進行予測

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の研究を行っている。今年度は、脳のネットワークが生み出すテンポラルなダイナミクスを解析した。脳の状態を空間内の一点と見なし、その一点が動的に変化するというダイナミクスを、最大エントロピー法の原理に基いて適応度地形を構築することによって解析した。特に、錯視の一種で 2 つの安定な知覚を引き起こすような視覚刺激を被験者に提示して、脳から fMRI によって取得されたデータを用いて解析した。結果、脳の状態は、3 つの適応度地形のくぼみ（エネルギーが極小となる状態に対応）に分類され、知覚の安定化や知覚の遷移に関わることが示唆された。特に、視覚に関わる脳領野の活動が支配的である状態や、前頭野の活動が支配的である状態が、本手法によって同定された。

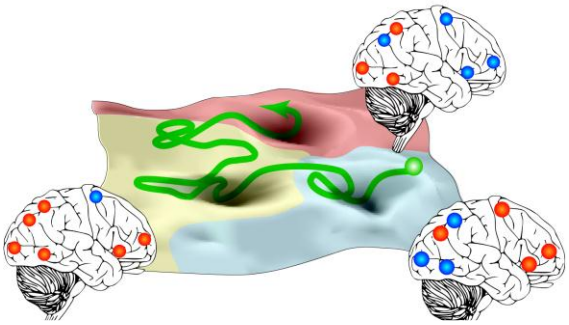

図２． 脳の潜在的ダイナミクス．脳の適応度地形の模式図。赤は視覚に関わる脳領野の活動が支配的である状態、青は前頭野の活動が支配的である状態、黄はそれらの中間的である状態に対応する。

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、人の選択行動のモデルを新しく構築し、その有効性を実証した。人の選択は、選択肢集合に強く依存し、魅力効果、妥協効果、類似効果などの現象を生む。この選択肢集合への依存性というディープナレッジをビッグデータから発見し、人の選択を高精度で予測する「制限付きボルツマンマシンに基づく選択モデル（RBM 選択モデル）」を開発した(図3)。これは上記効果を同時にすべて説明できる世界初のモデルである。本モデルを交通手段の選択に関する実データに適用し、従来の多項ロジットモデルの予測誤差約 0.12 に対して、予測誤差を約 0.02 まで小さくできることを示した。

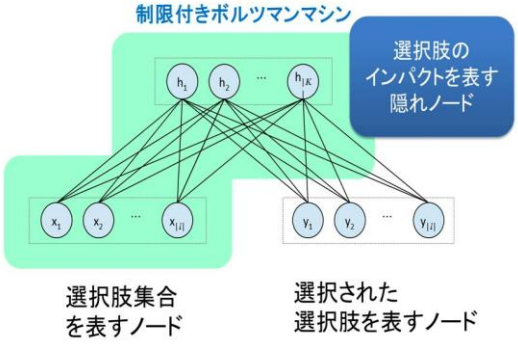

図３　RBM 選択モデル

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仕組み Innovators Marketplace on Data Jackets（IMDJ）の技術的基礎を確立し、その効果を実験的施行により評価している。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、データ間・分析ツール間・データとツールの結合を伴う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。各 DJ にはデータ概要は記載されるが、データの内容そのものは提示する必要がない。IMDJ では、DJ 間の関係を可視化したマップを参照しながらデータの「提供者」「分析者」「利用者」が取引条件を交渉する。今年度は、IMDJ の実験例として、文京区の街灯位置データと Google Map のデータを結合することにより夜道の安全なガイドを可能とするシステムを実現した。また、冷蔵庫においてビールの残量を求める製品開発案が企業において検討されるなどの事例が発生した。さらに、IMDJ において示された要求にこたえるための時系列データ可視化技術として Tangled String を開発した。これは、単語や事象などの要素が並ぶ系列データを、過去の要素との一致が発生したらそこに糸が戻る「絡まり」をそのまま可視化させる方法をとる。これにより、潜在的なダイナミクスを仮説化し表出させる効果を検証し、改良を進めている。

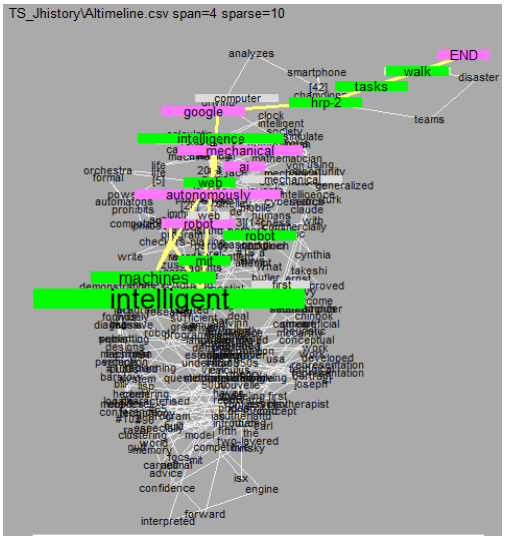

図４　Tangled Sting

代表論文

・Shigeru Maya, Kai Morino, and Kenji Yamanishi, “Predicting glaucoma progression using multi-task learning with heterogeneous features”, Proceedings of IEEE International Conference on BogData, pp: 261 - 270, 2014.

・Takamitsu Watanabe, Naoki Masuda, Fukuda Megumi, Ryota Kanai, Geraint Rees, “Energy landscape and dynamics of brain activity during human bistable perception,” Nature Communications, 5, 4765 (2014).

・Takayuki Osogami and Makoto Otsuka, ″Restricted Boltzmann machines modeling human choice,″ in Advances in Neural Processing Systems, vol. 27, pages 73-81, December 2014

・Yukio Ohsawa, Chang Liu, Teruaki Hayashi, Hiryuki Kido, ″Data Jackets for Externalizing Use Value of Hidden Datasets,″ 18th International Conference on Knowledge-Based and Intelligent Information and Engineering System, pp.946-953, Procedia Computer Science 35, 2014.