戦略的創造研究推進事業　CREST
研究領域「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」
研究課題「複雑データからのディープナレッジの
発見と価値化」

研究終了報告書

研究期間　2013年10月～2019年3月

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）

§１ 研究実施の概要

（1）実施概要

当チームはBigDataの変動性、多様性に注目し、巨大なデータの背後に在る潜在的知識（ディープナレッジ）を発見し、利活用することを目的として研究を行った。そのために、チームをディープナレッジ理論基盤と活用基盤に分け、前者は山西グループ（東大）と増田グループ（ブリストル大）が、後者はIBMグループと大澤グループ（東大）が担当して進めた。

理論基盤では、山西グループが、潜在的知識の変化と構造最適化の観点から、増田グループがテンポラル・ネットワークの観点からディープナレッジの基礎理論を構築した。活用基盤では、IBM グループが行動と意思決定のモデリングの観点から、大澤グループがデータ市場の観点からディープナレッジの活用基盤を構築した。

山西グループではBigDataからその背後にある潜在的な構造変化を検知するための「潜在的ダイナミクス」の研究と、多様なデータの間の潜在的関係性を抽出するための「潜在構造最適化」の研究を二大テーマとして据え、それぞれの理論を確立した。これを、主に緑内障進行予測、交通リスクマイニングの応用に展開して、その実際的な有効性を実証した。

顕著な成果として、潜在的ダイナミクスでは、「MDL(Minimum Description Length)変化統計量に基づく漸進的変化検知」の方法論を確立した。これは、従来の突発的な変化だけではなく、漸進的な変化を情報理論的原理に基づいて検知する手法である。また、本方式を、モデル（潜在変数の数等）の変化検知にも拡張し、変化予兆検知の基礎を築いた。また、潜在構造最適化では、潜在変数モデルに含まれる潜在変数の最適数を決定するためのモデル選択規準として、「分解型正規化最尤符号長規準」を提案した。それにより従来手法に比べて高精度、高効率、高普遍的な潜在変数モデル選択を実現した。緑内障進行予測では、視野感度データからの進行予測手法を確立した。また、視野感度データより低コストで採取できる網膜厚データからの緑内障診断を実現するために、網膜厚から視野感度を高精度に推定する「パターン正則化学習技術」を開発した。交通リスクマイニングでは、事故情報から道路地形情報まで含むヘテロな因子を掛け合わせて、道路スポットの潜在的危険度を定量化することに成功した。

増田グループでは、潜在的知識の表現として「テンポラル・ネットワーク」と呼ばれる時間的に変動するネットワークをとりあげ、ディープナレッジのダイナミクスを研究した。

顕著な成果として、テンポラル・ネットワークの一表現形式として、「エネルギー地形」を新しく提案した。これをfMRIデータから推定して、脳機能の知覚現象を解析することに成功した。また、ネットワーク上のイベント発生間隔の分布を少数の指数分布の重ね合わせとしても表現する効率的シミュレーションアルゴリズム「新ギレスピーアルゴリズム」を開発した。従来は、イベント発生間隔は裾野の長い「べき則」に従うと考えられていたが、これと異なる全く新しい見方を与えた。

IBM グループでは、人間が行動を選択し、意思決定を行うプロセスの新しいモデル化に取り組んだ。これを通じて、行動の中に潜むディープナレッジを活用するための方法論を構築した。

顕著な成果として、人間の選択行動を制約付きボルツマンマシンと呼ばれる潜在変数表現を用いてモデル化することに成功した。また、「動的ボルツマンマシン」と呼ばれる新しい時系列モデルを提案し、これを行動解析の新しいモデルとして展開した。動的ボルツマンマシンの学習原理に基づいて神経生理学的における時間依存可塑性（STDP）の理論的根拠を与えることに成功した。動的ボルツマンマシンのオープンソースを開発し、幅広い応用の道を開拓した。

大澤グループでは、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を与えた。顕著な成果として、データの概要（データジャケット）だけから潜在的なデータの活用シナリオを策定する発想支援技術であるIMDJ（Innovators Marketplace on Data Jacket）を確立した。その中で、ユーザの要求に合うデータジャケット群を代表する変数を発見する技術Variable Questを開発した。また、IMDJ を通じて得られたデータ活用シナリオの中からデータの可視化解析技術が幾つか生まれた。IMDJ は医療、安全都市計画、マーケティングなどに産業界に広く活用されると共に、各種データ活用コンソーシアムでも採用された。

チーム全体としては、ダイナミックでヘテロなデータからディープナレッジを抽出するという共通の目的に向かって、分担しながら独自の科学的方法論を完結させ、体系化することができた。

(2)顕著な成果

＜優れた基礎研究としての成果＞

1．潜在変数モデル選択規準「分解型正規化最尤符号長」を提案、高精度と高普遍性を実現

概要:

潜在変数モデルは深い知識を表現するモデルである。潜在変数モデルが含む潜在変数の最適な数をデータから選択することは重要であるが、その数学的な特異性ゆえに、従来の情報量規準を直接適用することは難しかった。そこで、記述長最小原理に基づいて「分解型正規化最尤符号長」というモデル選択規準を新たに提案した。本規準は、広いクラスの潜在変数モデルに適用可能であり、他の規準を上回る精度と効率計算を実現した。本成果をデータマイニングトップの国際会議 KDD2017 で発表し、フルバージョンは Data Mining and Knowledge Discovery 誌に採択された。

2．MDL 変化統計量に基づく漸進的な潜在的構造変化検知手法を開発

概要:

従来の変化検知は突然起こる変化の検知を対象にしていた。しかし、変化は徐々に起こる場合が多く、これに対応するために、世界初の MDL(Minimum Description Length)変化統計量に基づく漸進的変化検知手法を開発した。本手法が漸進的変化を高精度に検知できることを理論的及び実験的に示した。また、本手法をモデル（潜在変数の数等）の変化検知にも拡張し、潜在構造の変化検知理論の基礎を築いた。本成果は BigData 2016, BigData2017 で発表し、IEEE Trans. Information Theory 誌に掲載された。

3． エネルギー地形を用いて脳機能を解明

概要:

人間の脳の知覚状態のダイナミクスを説明する新しい方法として「エネルギー地形」を開発した。実際の fMRL データに基づき、エントロピー最大法によって脳のエネルギー地図を構成し、錯視等の知覚現象を説明することに成功した。その成果は Nature Communications に掲載された。また、イメージングデータを用いて、若年者と老人の脳のエネルギー地形を比較し、老人の方が若者よりも状態遷移が起こりにくいことを明らかにした。本成果は Human Brain Mapping に掲載された。

＜科学技術イノベーションに大きく寄与する成果＞

1．パターン正則化に基づく網膜厚からの新しい緑内障診断の手法を開発

概要:

緑内障の診断は、従来、Humphrey Field Analyzer (HFA)で測定された視野感度データを用いて行われてきた。しかし、HFA 検査は時間的にも労力的にもコストがかかり、高雑音を伴うという問題があった。一方で、光干渉断層計により網膜厚データを低コストかつ低雑音で測定できるようになってきた。そこで、網膜厚から視野感度を高精度に推定する技術として、「パターン正則化学習法」と「アフィン構造化非負値行列因子分解」を開発し、網膜厚から緑内障を診断する方法を世界で初めて開発した。本成果を KDD2017, KDD2018, American Jr. of Ophthalmology 誌で発表し、二件特許出願した。

2．新時系列モデル動的ボルツマンマシンを発明

概要:

新しい時系列モデルとして動的ボルツマンマシン(Dynamic Boltzmann Machine：DyBM)を発明した。DyBM の学習則が神経科学におけるスパイク時間依存可塑性（STDP）の理論的根拠と

なることを示した。また、DyBM は時系列データを一定時間で効率的にオンライン学習できることを示し、医療や金融データの解析に応用して有効性を検証した。それらの成果は Scientific Reports 誌に掲載され、AI のトップ会議である AAAI17,AAAI18,ICML2017 等で発表した。DyBM のオープンソース・ライブラリを開発して幅広い応用への道を切り開いた。

3. データ活用発想支援技術 Innovators Marketplace on Data Jackets を確立、産業界に展開

概要:

データの概要(データジャケット:DJ)だけからその活用シナリオを策定する発想支援技術としてIMDJ(Innovators Marketplace on Data Jacket)を確立した。IMDJ ではデータの「提供者」「分析者」「ユーザ」が DJ の関係を基に取引条件を交渉することで活用シナリオを生み出す。さらに、ユーザの要求に合う DJ 群を代表する変数を発見する技術 Variable Quest を開発し、Advances in Knowledge Discovery and Data Mining 誌に掲載された。IMDJ は各種企業や地方自治体に実地適用された。

＜代表的な論文＞

・Takamitsu Watanabe, Naoki Masuda, Fukuda Megumi, Ryota Kanai, and Geraint Rees, “Energy landscape and dynamics of brain activity during human bistable perceptio”, Nature Communications, vol. 5, 4765, 2014.
・Takayuki Osogami and Makoto Otsuka, “Seven neurons memorizing alphabetical images via spike-timing dependent plasticity”, Scientific Reports, 5, 14149, 2015.
・Tianyi Wu, Shinya Sugawara, Kenji Yamanishi: ″Decomposed normalized maximum likelihood codelength criterion for selecting hierarchical latent variable models″, Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD 2017), pp :1165-1174, 2017.

§2 研究実施体制

(1)研究チームの体制について

① 山西グループ
- 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)
- 研究項目
  ディープナレッジのモデル論、推定論の構築
  - ✧ 潜在的ダイナミクスの研究
  - ✧ 関係データ統合予測の研究
  - ✧ 緑内障進行予測の研究
  - ✧ 交通リスクマイニングの研究
  - ✧ その他応用研究

② 増田グループ
- 主たる共同研究者:増田 直紀 (University of Bristol, Department of Engineering Mathematics, Senior Lecturer)
- 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
  - ✧ エネルギー地形を用いたテンポラル・ネットワークの解析手法の開発と脳データへの応用
  - ✧ イベント時間間隔が長い裾野を持つ場合についてのギレスピーアルゴリズムの開発
  - ✧ ランダム・ウォークの理論解析
  - ✧ 時間的に変動するソーシャル・ネットワーク上の感染症動態記述のための個体ベース近似理論の開発
  - ✧ 合意行動を速くする:合議順の最適化

③ IBM グループ

- 主たる共同研究者：恐神　貴行（日本アイ・ビー・エム（株）東京基礎研究所　シニア・リサーチ・スタッフ・メンバー）
- 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

  ✧ 行動モデルの学習
  ✧ 意思決定の最適化
  ✧ 行動モデルと意思決定最適化の融合

④ 大澤グループ

- 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）
- 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

  ✧ Innovators Marketplace on Data Jackets（IMDJ）のフレームワーク構築
  ✧ データジャケット（DJ）収集手法:
  ✧ データ価値認知モデリング
  ✧ 異種データ結合、表出化
  ✧ 要素の関係性の可視化

（2）国内外rの研究者や産業界等との連携によるネットワーク形成の状況について

山西グループでは、情報理論およびその機械学習応用の基礎に関してフィンランドの Helsinki University 及び Tampere University と協力関係の下で研究を行っており、WITMSE（Worshop on Information Theoretic Methods for Science and Enginering）などの開催で協力した実績をもつ。また、緑内障進行予測の研究を通じて、東大病院、東北大、金沢大、山梨大との協力関係を結び、緑内障患者の日本最大のデータベースの構築を目指してデータ収集を行った。また、University of Tennessee Health Science Center や島根大学との共同研究を開始した。また、交通リスクマイニング研究を通じて、博報堂、ホンダ、国際興業、ITARDA 等と協働関係を構築し、国内の道路の交通リスク推定の研究を展開した。他にも、変化予兆検知の実証実験を東レ（株）など複数の企業と実施した。

増田グループでは、エネルギー地形法に関し、イギリスの Cardiff 大学（Jiaxiang Zhang 上級講師）、Reading 大学（榊美知子博士）、理化学研究所（渡部喬光副チームリーダー）、JST さきがけ（江崎貴弘博士） などとのネットワークを形成し、共同研究を推進した。また、ギレスピー法や他のテンポラル・ネットワーク解析手法に関し、国内外の多くの研究者と共同研究を行っている。具体的には、Luis E C Rocha 講師（イギリス・Greenwich 大学）、Petter Holme 教授（東工大）、Victor M Eguiluz 准教授（スペイン・University of Balearic Islands）、Konstantin Klemm 博士（同）、Leo Speidel 氏（イギリス・Oxford 大学）、Renaud Lambiotte 准教授（イギリス・Oxford 大学）、James Gleeson 教授（アイルランド・Limerick 大学）などである。

増田グループは、2018 年 7 月に Cookpad（欧州）と、共同研究を開始した。そのプロジェクトの中では、本研究で開発された手法を同社のデータに適用する可能性がある。このような連携も、本 CREST の研究成果が可能にしたものである。また、増田グループは、イギリスの Bath に本拠を置くリスクサービスである CheckRisk 社から、経済データの供与を受け、エネルギー地形法を適用した。このミニプロジェクトは、増田が指導する学生が修士論文として行った。本連携のスケールアップ（博士課程学生をリクルートして本研究を行うことなど）は、同社と協議中である。

IBM グループでは、九州大学・東京工業大学・早稲田大学・同志社大学・藤田保健衛生大学・東京大学・理化学研究所・（株）KDDI 総合研究所・第一生命保険（株）・日産自動車（株）の研究者らと論文を共著するなどして連携・協働を進めた

大澤グループは、経済産業省「データ駆動型イノベーション創出戦略協議会」「先端課題に対応したベンチャー事業化支援等事業（データ利活用促進支援事業）」、データ・エクスチェンジ・コ

ンソーシアム（DXC）、国土交通省「国土交通行政に資するビックデータの活用に関する調査業務」などに IMDJ を実地適用した。参加者として経済産業省の場合には最終的に58社が残ったほか、多数の企業に DJ 利用推進のネットワークが拡がった。

また、大澤グループの開発した IMDJ は 2017 年度末に発足したデータ流通協議会において、データ利活用の手法として Web 版が導入された。例えば「データ流通推進協議会 利活用委員会におけるワークショップ」の一度だけで、参加者が企業から約80名となり、彼らは自社に同技術を持ち帰って実務に導入している。また基盤技術に関しても、大澤グループの開発したデータジャケットが我が国のメタデータの標準仕様として導入の検討が進められている。また、三菱地所（株）、富士通（株）、ソフトバンク（株）、東京大学からなる幹事組織など10社（2018 年 8 月 1 日現在）とともに、データジャケットを利用して、日本の代表的な経済活動拠点である大丸有地区の再開発における価値の付与をめざす実験を開始した。2017 年からは、データジャケット推進協議会を発足させた。富士通、構造計画研究所、日本リサーチセンター、トッパン・フォームズ、データセクションなど10社が参加し、データジャケットおよび周辺技術の普及と伝承活動を行っている。

2018 年からスタンフォード大学 Larry Leifer らとコラボレーションの連絡を開始し、同氏の編集する Springer の Understanding Innovation シリーズに Innovators Marketplace on Data Jackets を発刊することとなった。