戦略的創造研究推進事業　AIP 加速課題
研究課題「潜在空間を高度活用した
ディープナレッジの発見」

研究終了報告書

研究期間　2019年4月～2022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

ビッグデータ時代において、データがますます多様で動的になり、複雑性を増すにつれて、データから深い知識や意味を発見して説明することが重要になってきている。研究代表者は、前身となる JST CREST のプロジェクトにおいて、データの背後にある「**潜在空間**」に着目した研究を行った。そこでは、潜在空間の構造を最適化し、変化を捉える方法を確立し、体系化した。近年では、潜在空間を用いるデータの表現方法、並びにその活用方法は急速な勢いで進展している。。そこで、本 AIP 加速課題においては、潜在空間をさらに高度に活用し、データからより深い知識を発見する技術を開発することを目指した。具体的には以下の課題を設定した。

（Ⅰ）潜在空間高度活用のための潜在空間表現学習理論の研究
（Ⅱ）潜在空間高度活用に基づく潜在構造変化検知の研究
（Ⅲ）潜在空間を高度活用した AI 眼科学の創出

(I)と(II)には山西グループが、(III)には朝岡グループと山西グループが貢献した。

(I)については大きく、(I-1)潜在空間埋め込みの方法と、(I-2)潜在変数モデル選択 にフォーカスした。(I-1)の**潜在空間埋め込み**とは、離散データを連続値空間に写像することで、データをより把握しやすく、活用しやすくするための方法である。特に、近年では Euclid 空間に埋め込むのではなく、**双曲空間**と呼ばれる、曲がった空間に埋め込むことが注目されている。本研究プロジェクトでは、双曲空間に順序関係やグラフ構造をもったデータを埋め込むことの効能と限界について、世界で初めて理論的に明らかにした。その結果、順序関係やグラフ構造を持つデータについては、サンプル数が少ない時には、Euclid 空間に埋め込む方が期待リスクが小さく、サンプル数が多くなると双曲空間の方がより期待リスクが小さくなるという性質を理論的に明らかにした。ここで、期待リスクとは、埋め込まれたデータが互いに識別されなくなるリスクの期待値である。本成果は ICML2021, NeuLIPS2021 にて発表した。

また、埋め込みの次元は高ければ高いほど、データの識別はできるようになるが、高すぎると過学習を起こしてしまう。そこで、適切な次元を選ぶことが必要であるが、これまではシステマティックに決める方法が存在しなかった。そこで、本研究では、世界で初めて埋め込み次元をシステマティックに決定する方法を提案した。我々は、これを**記述長最小原理**（Minimum Description Length: MDL）原理に基づき、**逐次的正規化最尤符号長**(Sequential Normalized Maximum Likelihood: SNML)を規準とすることで実現し、その優位性を実験的に検証した。本成果は Entropy 誌に掲載された。

(I-2)の**潜在変数モデル選択**とは、データから潜在変数の数を自動決定する方法である。ここで、潜在変数の数とは、例えば、クラスタリングモデルのクラスター数や、ネットワークモデルのコミュニティ数などに相当する。潜在変数の数を推定することは、データの内在的構造を把握する上で本質的な問題である。この問題に対して当研究チームは MDL 原理に基づき、**分解型正規化最尤符号長規準**(Decomposed Normalized Maximum Likelihood: DNML)を提案してきた（KDD2017）。今回新たに、その理論的性質として、推定モデルがミニマックス最適性をもつこと、計算効率性をもつこと、広い潜在変数モデルの族に適用可能であることを示すとともに、他手法と比べた優位性を実験的に検証した。本成果は Data Mining and Knowledge Discovery 誌に掲載された。また、DNML 規準を大規模グラフの要約、階層的変化検知、ネットワークのイベント発生区間分布の指数混合モデルによる近似応用して効果を実証した。これらは Complex Networks2021、ICDM2021 及び Royal Society Open Science 誌に発表した。

(II)については大きく、(II-1)潜在構造変化検知の研究と(II-2)潜在構造変化予兆検知の研究にフォーカスした。ここでは後者を主に報告する。

当チームは CREST の活動を通じて、データの背後にある確率モデルのパラメータの変化検知、及び潜在構造の変化検知を研究してきた。その場合、変化は突発的に起こるものと仮定していた。しかしながら、変化は漸進的に進むことが多く、その開始点である変化予兆点を早期に正しく検知することが課題であった。これを**変化予兆検知**と呼ぶ。当チームは、変化予兆検知の指標として**微分的MDL変化統計量**を提案した。これは突発的変化検知のために我々がCRESTで開発した

MDL 変化統計量の時間微分を計算したものである。ここで、微分的 MDL 変化統計量の有効性を統計的検定の枠組みを用いて理論的に証明し、予兆アラートを上げるための閾値を理論的に導出した。本手法を COVID-19 の第一波の感染爆発予兆検知に応用した、2020 年 4 月 30 日までに累積感染者数が 1 万人を超えた 37 か国に対して適用したところ、106 個の感染爆発が変化点として検知できた。そのうち64%については、平均 6 日前に変化予兆アラートを上げていた。本手法は、従来の感染症流行を予測する数理モデル（SIRモデル）のような予測・シミュレーションモデルとは異なる角度から感染爆発の予兆を捉える点が革新的である。本成果は Scientific Reports にて掲載された。

また、離散的な潜在構造が変化する際の予兆検知手法を研究した。例えば、クラスタリングを行うときのクラスタ数が変化する、ネットワークコミュニティの数が変化するといった場合の予兆を検知することを問題にした。そのような離散構造の変化は、突発的な変化と捉えられがちであるが、実際には漸進的に変化している場合が多い。そこで、離散構造を連続的に緩和した連続的指標を定式化し、その漸進的変化を追跡することにより離散構造の変化予兆を検知する手法を開発した。その連続的指標の1つとして**記述次元（Descriptive Dimensionality: Ddim）**を提案した。これはモデルの記述長にフラクタル次元の考え方を適用して定量化したものであり、パラメータ次元の一般化に相当する。本研究では、Ddim の理論的な性質を明らかにし、実際にクラスター数変化の予兆を捉えることを実験的に検証した。もう一つが **Kernel Complexity（KC）**である。これは分布の偏りを測るジニ係数の考え方を、カーネル密度関数の正規化最尤符号長曲線に適用して得られたもので、ノンパラメトリックな分布の複雑さを測る指標である。KC により、ノンパラメトリックな潜在構造の変化予兆を検知できることを実証し、その成果は IEEE BigData 2021 に採択された。

（III）については、網膜層厚と視野感度といったヘテロ情報から緑内障進行の診断及び予測を行う方法を開発した。緑内障の診断は、主に視野感度を用いて行われる。しかしながら、この検査には時間がかかる、測定誤差が入りやすいという問題があった。一方で、近年では光干渉断層計を用いて短時間で網膜各層厚を計測する検査が普及してきた。そこで、以下の課題に取り組んだ。

① 網膜層厚データから現時点の視野感度をいかに精度良く推定できるか（推定問題）
② 視野感度と網膜層厚の双方を用いて「将来の」視野の欠損具合をいかに精度よく予測できるか（予測問題）

①では、**テンソル回帰法**（AmericanJr.Ophthalmol.誌掲載）、**パターン正則化法**（BritishJr.Ophthalmol.誌掲載）という手法を提案して世界最高の精度を達成した。また、②では、**深層潜在空間線形回帰**（KDD2019 発表）という手法を提案して世界最高の精度を達成した。さらには、「**マルチタスク潜在空間統合学習**」という新しい機械学習技術を開発することにより、①と②の課題を同時に解決し、それぞれの精度を増強することに成功した。本技術は、視野感度と網膜層厚のデータの時空間的特徴を、低次元に圧縮して表現した「**潜在空間**」の中で統合して学習することを特長とする。また、学習の際に推定に用いた情報と予測に用いた情報を共有する（**マルチタスク学習**）ことで、①の推定誤差と、②の予測誤差について、平方根平均二乗誤差が従来手法の世界最良の結果をそれぞれ 6.33%と 3.48%上回る高精度化を実現した。この成果は実用化に向けた着実な一歩を示すものである。本研究成果は KDD2021 にて発表し、その後、医学的再検証を行った結果が Ophthalmology Science に掲載された。

更に、マルチタスク学習モデルの更なる予測精度改善のために、眼球生体力学特性の計測を行うことが有用であることを英文科学誌（Invest Ophthalmol Vis Sci 2021 など）に報告した。その上、得られた計測結果を数理的に処理して特徴量抽出を行ってから組み込むことが殊更有用であることを英文科学誌（Graefes Arch Clin Exp Ophthalmol 2020、SciRep 2020）に報告した。このことから、これらの結果を統合活用したマルチタスク学習モデルが、緑内障進行予測に有用であることが示唆される。またスマートフォンで得られた眼底写真から緑内障性視神経障害の特徴を自動で判定するアルゴリズムの開発も完了しており、将来的にこの結果も利用することが可能である。

(2)顕著な成果
＜優れた基礎研究としての成果＞

1. 双曲空間埋め込みの性能保証と Euclid 空間埋め込みの次元決定
概要:
離散データを低次元の連続値空間に埋め込む方法を発展させた。1つは双曲空間と呼ばれる曲がった空間に順序構造やグラフ構造を持つデータを埋め込むことの評価の枠組みを開発し、汎化損失解析によって、双曲空間に埋め込むことがなぜ良いのか？を世界で初めて定量的に明らかにした。さらに、Euclid 空間に埋め込む際の空間の次元を選択する手法を世界で初めて開発した。これは逐次的正規化最尤符号長を規準にして、最適な次元を決定する手法であり、自然言語の分野にて手動で決定されていた次元に近い値を自動的に決定できることを示した。

2. 変化予兆手法の開発
概要:
時系列データから変化予兆点を早期に正確に検知するための手法を開発した。データの生成分布が連続的に変化する場合には、微分的MDL変化統計量と呼ばれる情報論的尺度に基づいて変化予兆点を検知する手法を提案した。これをCOVID－19の感染数時系列データに適用したところ、実際の感染爆発に先立つ予兆を平均 6 日前に検知できた。データの生成分布の離散構造(クラスター数等)が変化する場合には、記述次元およびカーネルコンプレキシティとよばれる離散構造を連続緩和した量を発明し、これを追跡することで変化予兆を検知する手法を世界で初めて開発した。これにより市場動向の変化の予兆を検知できることを検証した。

3. 潜在変数モデル選択の理論構築と複雑データ解析への応用
概要:
データから潜在変数モデルにおける潜在変数の数(クラスター数、コミュニティ数など)を自動的に推定する手法を開発した。これは、新しいモデル選択規準として分解型正規化最尤符号長(DNML)規準を提案することによって実現できた。DNMLの優位性を、推定最適性、計算効率性、汎用性という観点から理論的に示した。さらに、DNMLをグラフ要約、階層的変化検知、ネットワーク上のイベント発生区間分布(べき分布)の混合指数分布近似などの問題に応用することで、幅広い応用可能性と有効性を実証した。

＜科学技術イノベーションに大きく寄与する成果＞

1. 網膜層厚と視野感度からの緑内障診断及び予測
概要:
従来、緑内障の診断と予測には、視野感度の測定データが用いられてきたが、より測定が容易な網膜層厚データも利用する以下の 2 つの問題を扱った。①網膜層厚を視野感度に変換する推定問題、②両者の時系列を用いて将来の視野感度の値を予測する予測問題。これらを同時に解決する方法として「マルチタスク潜在空間回帰法」を開発した。そのポイントは、ヘテロな情報を潜在空間で統合するともに、マルチタスク学習により①と②の情報を互いに共有し、活用することである。これにより①と②の双方の問題で世界最高水準の精度を達成した。①は緑内障の診断の、②は緑内障予測の高精度化に貢献する。本成果でプレスリリースを行った。

2. 網膜厚から視野感度の推定
概要:
当チームは網膜層厚データから視野感度に変換する手法として畳み込みニューラルネット(CNN) の最終出力層の出力ベクトル値の線形回帰として推定する方法を構築してきた(KDD2017)。今回さらに CNN の最終出力層の出力値のテンソル回帰を基に視野の値を推定する方式(CNN-TR)を提案した。多施設から得られた 591 眼(86 正常眼、304 緑内障眼)を

用いて、対抗手法（データ多重線形回帰、サポートベクトル回帰、CNN-PR）と比較評価したところ。有意な差をもって CNN-TR が最も優れた予測精度を達成した。本成果は American Journal of Ophthalmology に掲載され、日本眼科学会雑誌の外国誌要覧セクションでも紹介された。

3. 緑内障における高速視野計測

概要：

CREST 課題で開発した変分近似ベイズ線形回帰法による視野感度予測（実際の診断補助ソフトウエアとして実装され臨床現場で既に使用されている；GlaPre®、ビーライン株式会社、東京および、NAVIS®、ニデック株式会社、東京）を利用した高速視野測定アルゴリズムを構築した。この結果、73 例 122 眼の緑内障例を用いて実証検証では、既存の標準手法に比べ、測定の正確性を全く損なうことなしに、視野計測時間を 15％程度短縮することに成功した（British Journal of Ophthalmology）。この測定アルゴリズムは国産視野計に組み込まれて販売され、実際の臨床現場で頻用されている（AP7700 視野、興和株式会社、名古屋）。

＜代表的な論文＞

1．K.Yamanishi, T.Wu, S.Sugawara, M.Okada: ″The Decomposed Normalized Maximum Likelihood Code-length Criterion for Selecting Hierarchical Latent Variable Models″, *Data Mining and Knowledge Discovery*, 33(4): 1017-1058, 2019.

概要:

新しい潜在変数モデル選択規準として分解型正規化最尤符号長規準（DNML）を提案した。これは潜在変数モデル特有の非正則性の問題を克服し、広いクラスの潜在変数モデルに対して潜在変数の数（クラスター数やコミュニティ数等）をデータから決定するための規準である。その優れた性質（ミニマックス推定最適性、計算効率性）を理論的に証明し、実験的にも既存の情報量規準を凌駕することを示した。また、この論文がもとになって、グラフ要約、階層的変化検知、ネットワーク次数分布の指数混合分布近似などの研究も生まれた。

2. L.Xu, R.Asaoka, T.Kiwaki, H.Murata, Y.Fujino, M.Matsuura, Y.Hashimoto, S.Asano, A.Miki, K.Mori, Y.Ikeda, T.Kanamoto, J.Yamagami, K.Inoue, M.Tanito, K.Yamanishi: Predicting the glaucomatous central 10 degrees visual field from optical coherence tomography using deep learning and tensor regression. *American Journal of Ophthalmology*, 2020 Oct;218:304-313.

概要:

CREST 時に当チームは網膜層厚データから視野感度に変換する手法として畳み込みニューラルネット（CNN）の最終出力層の出力ベクトル値の線形回帰として推定する方法を構築していた（KDD2017）。本論文では、さらに CNN の最終出力層の出力値のテンソル回帰を基に視野の値を推定する方式（CNN-TR）を提案した。今回多施設から得られた 591 眼（86 正常眼、304 緑内障眼）を用いて、対抗手法（データ多重線形回帰、サポートベクトル回帰、CNN-PR）と比較評価したところ。有意な差をもって CNN-TR が最も優れた予測精度が達成された。本成果は日本眼科学会雑誌の外国誌要覧セクションでも紹介された。

3. Y. Hashimoto, R. Asaoka, T. Kiwaki, H. Sugiura, S. Asano, H. Murata, M. Matsuura, A. Miki, K.Mori, Y. Ikeda, T.Kanamoto, J. Yamagami, K. Inoue, M. Tanito, K. Yamanishi: ″A deep learning model to predict visual field in central 10 degrees from optical coherence tomography measurement in glaucoma″ *the British Journal of Ophthalmology*, 105(4):bjophthalmol-2019-315600 2020.

概要：

我々は光干渉断層計（OCT）から緑内障の中心 10 度視野を予測するためのパターン正則化付き深層学習モデルを提案していた（KDD2018）。本研究は当モデルの外部データへの汎用性の検証を目的とした。訓練データは、OCT と視野の両方を有する開放隅角緑内障眼および健

常眼 591 眼と 7715 枚の視野単独データ、検証データは開放隅角緑内障 160 眼を用いた。当モデルの平均絶対誤差は 5.5 dB で、従来法に比べ有意に小さかった。予測精度は臨床応用レベルに近づいており、今後の改善が期待される。

§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:

(Ⅰ)潜在空間高度活用のための潜在空間表現学習理論の研究

・ネットワーク埋め込みアルゴリズムの研究

・潜在構造表現最適化の研究

(Ⅱ)潜在空間高度活用に基づく潜在構造変化検知の研究

・潜在構造変化検知の研究

・潜在構造変化予兆検知の研究

(Ⅲ)潜在空間を高度活用した AI 眼科学の創出

・ヘテロ時系列データからの緑内障進行予測

・視野感度推定および視野進行予測包括モデルの構築

②朝岡グループ

主たる共同研究者:朝岡　亮(静岡大学電子工学研究所　特任准教授/聖隷浜松病院眼科主任医長/聖隷クリストファー大学　臨床准教授/光産業創成大学院大学　光産業創成研究科客員准教授)

研究項目:

(Ⅲ)潜在空間を高度活用した AI 眼科学の創出

・緑内障データ取得

・ヘテロ時系列データからの緑内障進行予測

・視野感度推定および視野進行予測包括モデルの構築

(2)国内外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

・Prof. J.Vreeken (Max Plank Inst.):潜在空間の情報論的学習理論について協力した。。KDD2019 では共同で、チュートリアル“Modern MDL meets Data Mining”を開催した。

・Prof. L.Xu, Prof. J.Cao, (Poly Tech. HongKong): ネットワークマイニングと緑内障進行予測について共同研究を実施した。

・Prof.J. Oliver(Iniv. Lisbon) Attention Network 関連の理論と応用に関して共同研究を開始。

・Prof. David Garway-Heath (英国 Moorfields Eye Hospital): 緑内障患者データを取得し、データコンソーシアムを構築中。

・Prof. Linda Zangwill (University of California San Diego): 緑内障患者データを取得し、データコンソーシアムを構築。緑内障診断アルゴリズムを共同解析の上、解析結果を論文報告した(Asaoka R, Zangwill L et al. Transl Vis Sci Technol　2020)。