AIP 加速課題
2019 年度採択研究代表者

2019 年度
実績報告書

山西　健司

東京大学情報理工学系研究科
教授

潜在空間を高度活用したディープナレッジの発見

§1. 研究成果の概要

本研究では、データの背後にある潜在空間を読み解くことで、深い知識を発見するための方法論を確立し、医学に展開することを目標としている。潜在空間とは、データからは観測できない空間のことであり、潜在変数モデルという確率モデルで表現される。潜在変数モデルの歴史は長いが、近年、埋め込みや深層学習モデルなど新しい潜在変数モデルが出現し、データからこれらの最も適切な潜在構造を推定すること、及び潜在構造の時間的変化を検知すること、それらを高度活用して知識発見につなげることは発展途上である。そこで、我々は、上記目標に沿って「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」を 3 大テーマとして研究している。

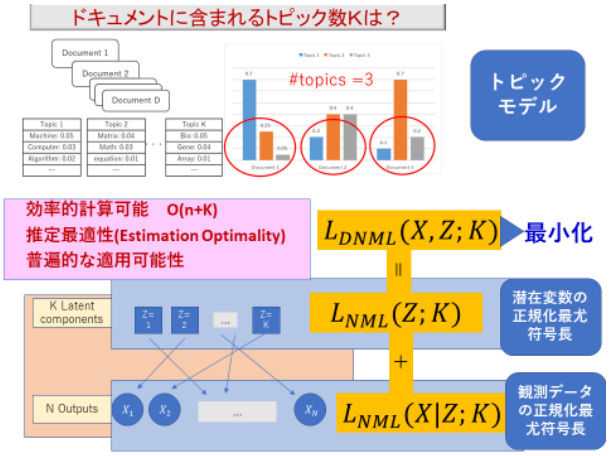

図１．分解型正規化最尤符号長規準とそのメリット

本年度は、「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」において、潜在構造モデル選択の問題を扱った。これは潜在変数モデルに必要な最小限のパラメータ数をデータから推定するという、統計学上重要な問題である。例えば、トピックモデルにおいてトピックの適切な数を推定するという問題が相当する。ここで、潜在変数モデルは一般に非同一性問題(パラメータと分布が 1 対 1 対応しない問題)を孕んでおり、従来の情報量規準を直接適用することは理論上適切でなかった。我々はこの問題を克服する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**(Decomposed Normalized Maximum Likelihood 略称、DNML)を 2017 年に提案

していた。DNML は、記述長最小原理に基づいて、潜在変数と観測データを分解して別々に正規化最尤符号化し、得られる総符号長を最小化するモデルを選択するという戦略である

本年度は A)DNML の理論的最適性の保証を与え、B)ガウス分布やテンポラルネットワークのモデル選択に適用し、その有効性を検証した。A)については、真の分布と推定分布の Kullback-Leibler 距離に関してミニマックスを考えるとき、最小を与えるモデル選択が DNML で実現できることを示した。これは DNML の最適性を保証する強力な結果である。実際、ガウス分布のモデル選択に適用して有効性を実証した。これらの結果はデータサイエンスのトップジャーナルである Data Mining and Knowledge Discovery 誌に掲載された。B)については、テンポラルネットワークと呼ばれる時間的に変化するネットワーク上のイベント(セキュリティインシデントの発生など)発生間隔の分布の推定問題へ適用した。従来、その発生間隔はべき分布に従うことが知られてきたが、本研究では、多くの場合、それが 3－4 個の指数分布の線形結合として近似できることを明らかにした。本成果はネットワーク科学にモデル選択の視点から重要な知見を与えたことが評価され、Royal Society Open Science 誌に掲載された。

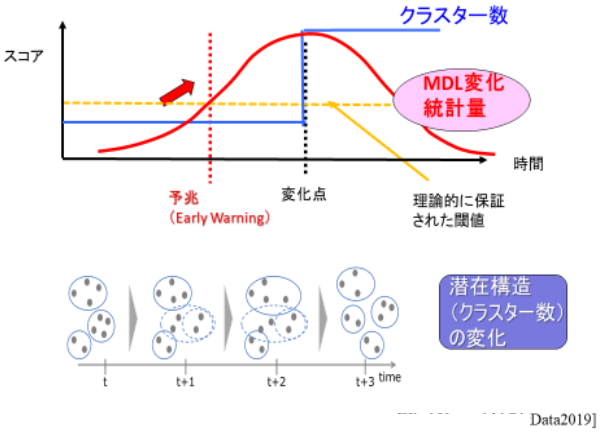

図 2. 潜在構造予兆変化検知

「Ⅱ. 潜在空間高度活用に基づく潜在構造変化検知の研究」においては、潜在構造変化予兆検知の問題を扱った。これは例えば、クラスタリングにおけるクラスター数が変化するときに、その予兆を検知することにつながるからである。本年度は、この問題に対して、「**MDL 変化統計量に基づく潜在構造変化予兆検知**」の手法を提案した。鍵となるアイデアは、潜在構造の変化度合いを MDL 変化統計量と呼ばれる連続値で定量化し、その閾値処理により変化予兆を検知することである(図 2)。ここで、閾値は誤警報確率を一定以下になるように理論的に保証された値を用いる。本手法を人工データに適用したところ、Structural Entropy や Fixed Share Algorithm などの従来技術よりも早期にかつ低い誤警報率で、潜在構造変化の開始点を検知できることを実験的に検証した。さらに、ビール購買記録や電力消費の実データに適用したところ、新しい消費クラスターの生成の予兆を検知できることを示した。本結果は 2019 IEEE International Conference on BigData に採択され発表した。

「Ⅲ.潜在空間を高度活用した AI 眼科学の創出」では、潜在空間高度活用の考え方を眼科学における**緑内障進行予測**の問題に適用した。従来の緑内障進行予測では、Humphrey Field Analyzer という機器を用いて計測された視野感度(Visual Sensitivity:VF)の時系列データを基に、将来の VF を予測していた。一方、光学干渉断層計の発達により網膜神経繊維層厚(Retinal Thickness: RT)を測定できるようになった。そこで、VF

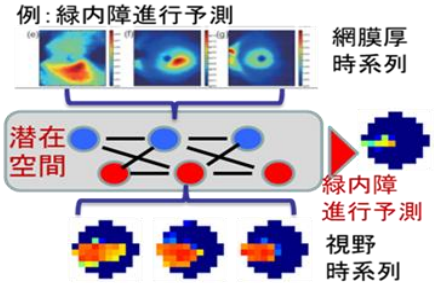

図 3. RT と VF からの緑内障進行予測

とRTの両方の時系列を用いて将来のVF値を予測するための手法を新たに構築した。その際に、VF と RT といったヘテロな情報をいかに統合するか？という大きな問題が生じる。そこで、本研究では、本問題を克服する手法として**深層正則化潜在線形回帰法**を開発した。本手法の特徴は、A) RTを深層学習で変換した後、RTと潜在空間の中で統合し、潜在的線形回帰モデルを構成することで、時間空間的なヘテロ性を克服したことにある(図 3)。この手法を広島記念病院、大阪大学付属病院、東京大学附属病院から提供された実データ(254 眼)に適用した結果、従来の VF のみを用いた予測手法に比べて 12%平方平均二乗誤差を小さくすることを検証した。本成果により光学干渉断層計の緑内障進行予測における可能性が大きく拡がった。本手法は特許出願を果たすと共に、データサイエンスのトップ国際会議 KDD2019 に採択され発表した。

【代表的な原著論文】

1. Kenji Yamanishi, Tianyi Wu, Shinya Sugawara, Makoto 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.
2. So Hirai, Kenji Yamanishi:“Detecting Model Changes and their Early Warning Signals Using MDL Change Statistics”, Proceedings of 2019 IEEE International Conference on BigData (BigData 2019), pp: 84-93 2019.
3. Yuhui Zheng, Linchuan Xu, Taichi Kiwaki, Jing Wang, Hiroshi Murata, Ryo Asaoka, Kenji Yamanishi:“Glaucoma Progression Prediction Using Retinal Thickness via Latent Space Linear Regression”, Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2019), pp: 2278-2286, 2019.

§2. 研究実施体制

(1)山西グループ

① 研究代表者:山西　健司 （東京大学情報理工学系研究科　教授）

② 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

(2)朝岡グループ

① 主たる共同研究者:朝岡　亮 （東京大学医学部附属病院　特任講師）

② 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