「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

H29 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

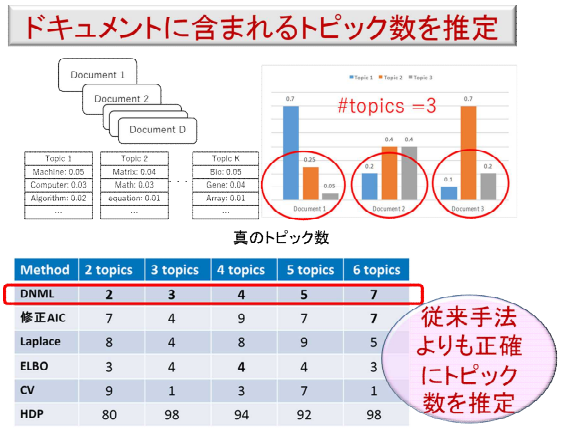
(3)「IBM」グループ
① 主たる共同研究者:恐神　貴行 （日本アイ・ビー・エム(株)社東京基礎研究所 リサーチスタッフメンバー）
② 研究項目
・ディープナレッジを価値につなげるための意思決定最適化技術

(4)「大澤」グループ(研究機関別)
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

本研究チームは、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在る潜在知識（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。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、データからディープナレッジの構造を推定する方法を研究している。本年度は、観測データからその背後にある潜在構造（潜在変数の数、トピック数等）を推定するための新たな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：DNML)を提案した。本規準では、記述長最小原理に基づいて、観測変数と潜在変数を分解して記述長を計算し、その和を最小化するモデルを選択する。従来のモデル選択規準は、潜在変数モデルに対しては、その数学的特異性ゆえに直接適用できなかったが、DNML は幅広い潜在変数モデルのクラス（（混合分布、トピックモデル、確率ブロックモデル等）に適用できる。また、データ数の線形時間で効率的に計算可能であり、真のモデルを従来規準(AIC,,BIC，HDP等)を上回る精度で推定できることを実証した。本成果はデータマイニングのトップ国際会議 KDD2017 で発表した。



Method	2 topics	3 topics	4 topics	5 topics	6 topics
DNML	2	3	4	5	7
修正AIC	7	4	9	7	7
Laplace	8	4	8	9	5
ELBO	3	4	4	4	3
CV	9	1	3	7	1
HDP	80	98	94	92	98

図1. DNMLに基づく潜在変数モデル選択

山西Gでは、ディープナレッジ発見の応用として緑内障進行予測の研究を行っている。従来の緑内障診断はハンフリー視野計（HFA）で測定した視野感度データに基づいて行われてきた。しかし、この測定は時間的・労力的にコストが大きく、雑音の影響を受けやすい等の問題があった。一方、光干渉断層計（OCT） を用いることにより、低コストで雑音の少ない網膜厚みデータを測定できるようになってきた。本研究では、OCT で測定した網膜厚みデータから視野感度を高精度に推定する手法を世界で初めて開発した。これにより、従来に比べて遥かに被験者の負担を軽減する緑内障診断の実現が期待できる。本手法は、アフィン構造化非負値行列分解手法(ASNMF)という新しい手法に基づいている（図2）。これは異なる領域のデータを潜在的な特徴を通じて変換する方法である。緑内障の潜在的特徴パターンも発見できる。また、畳み込みニューラルネット(CNN)を用いる方法も同時に提案し、

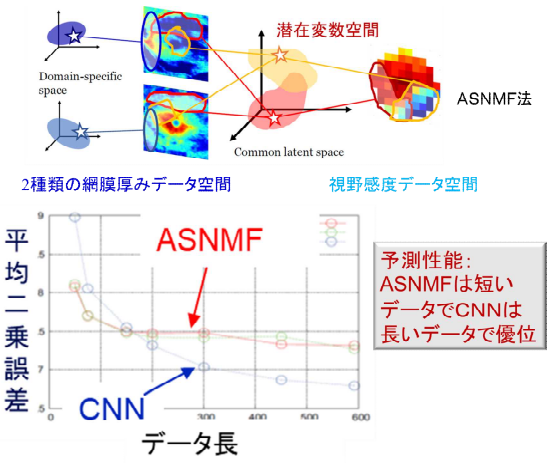


図 2. ASNMF:網膜厚みデータから視野感度の推定

データ長が比較的短い場合はASNMFが、長い場合はCNNが優れていることを、東大病院の協力の下検証した。結果をKDD2017で発表し、特許出願した。

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、そのようなネットワーク上の伝搬現象を、枝の同時性という観点から理論的に明らかにした。枝の同時性は、1990年代中盤に数理疫学で提案され、HIV/AIDSなどのフィールドデータに対しても計測されている量である。それは、テンポラル・ネットワークのひとつの性質と見なすことができ、ノードAとBを結ぶ枝ABとノードAとCを結ぶ枝ACが同時刻に存在する状況を指す。本研究では、枝の同時性が上がると、テンポラル・ネットワークにおいて感染が起こりやすくなることを理論的に明らかにした。主要な伝搬プロセスモデルであるSISモデルを用いて。感染力の閾値（それよりも大きい感染力になると、ネットワーク全体に感染が広まり始めるような感染力の値）を、枝の同時性の関数として、理論的に導出した。 本研究結果は、物理学のトップジャーナルの１つである Physical Review Letters 誌に発表された。

IBMグループは、時系列データを学習する動的ボルツマンマシン（DyBM）の研究開発を進めているが、潜在変数を持つDyBMの効率的な学習を可能とする双方向学習法を提案し、機械学習の国際会議ICML 2017で発表した。双方向学習法は、通常の「前向きのDyBM」に加えて、将来の値から過去の値を予測する「後ろ向きのDyBM」を同時に学習するが、これらの2つのDyBMでパラメータを共有しておく。これにより、従来のDyBMでは難しかった潜在変数に関わるパラメータの効率的な学習を可能とし、予測誤差を１０倍以上小さくできることを示した。

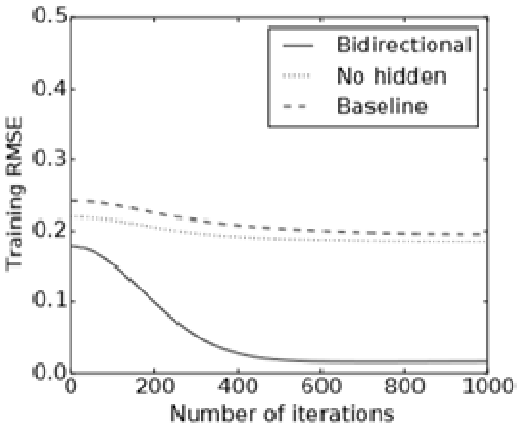


図 3. 双方向学習法の効果

大澤Gでは、社会的な問題解決要求とデータの概要情報（Data Jackets: DJ）を結びつける研究を行っている。本年度は、ユーザの社会的な問題解決要求に対して、関連するデータの変数セットを提供するシステムVariable Questを開発した。このシステムは、ユーザが漠然とした要求に沿ってデータを集めたい場合、これに対して①データの概要情報（データジャケット：DJと略） ②データ中の変数間の関連性ネットワークを提供する。そこで得られた情報を、データ提供者、分析者、利用者が参照しながら議論と検討を行うことにより、データを結合したり、必要に応じて設計し新規収集したりして利用するシナリオを作成することができるようになる。本システムは、実在するデータについて知識のない人にも利用できることから、データ流通推進協議会等のメンバーをはじめ、DJの利用者が増えるきっかけとなっている。本研究成果は Advances in Knowledge Discovery and Data Mining （Springer)に掲載された。

代表的原著論文

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.

Toshimitsu Uesaka, Kai Morino, Hiroki Sugiura, Taichi Kiwaki, Hiroshi Murata, Ryo Asaoka, Kenji Yamanishi, "Multi-view Learning over Retinal Thickness and Visual Sensitivity on Glaucomatous Eyes", Proceedings of ACM International Conference on Knowledge Discovery and Data Mining(KDD 2017), pp 2041-2050, 2017.

Tomokatsu Onaga, James P. Gleeson and Naoki Masuda, "Concurrency-induced transitions in epidemic dynamics on temporal networks", Physical Review Letters, vol. 119, 108301, 2017.

Takayuki Osogami, Hiroshi Kajino, Taro Sekiyama, "Bidirectional learning for time-series models with hidden units," Proceedings of the 34th International Conference on Machine Learning, PMLR 70:2711-2720, 2017.

Teruaki Hayashi, Yukio Ohsawa, “Matrix-based Method for Inferring Variable Labels Using Outlines of Data in Data Jackets,”Advances in Knowledge Discovery and Data Mining, pp.696-707, Vol.10235, Springer, 2017.