研　究　報　告　書

「統計的潜在意味解析によるデータ駆動インテリジェンスの創発」

研究タイプ：通常型
研究期間：　平成26年10月～平成29年3月
研 究 者：　佐藤　一誠

1.　研究のねらい

個人が日常的に触れる情報量が膨大になり、一個人に関係する情報に限ったとしても、その全容を知ることがすでに不可能な状況に我々は直面している。しかし、創造的な活動は、普段組み合わせて考えない物事の組み合わせや日常注目していなかった物事に目を向けるなど、多様な情報に対して横断的な思考を行うことが重要である。本研究では、科学技術によって人の知の営みを拡張することを目的とする。この目的を達成するアプローチとして、「データ駆動インテリジェンス(英：Data-driven Intelligence)」という概念により1つの見通しを与える。「データ駆動インテリジェンス」は、「人工知能(英: Artificial Intelligence)」や「人間の知能(英:human Intelligence)」に並ぶ第 3 の知のカテゴリとして期待されている。 本研究では、人の知能と計算機による知能の交差する領域ではなく、データが誘発する計算機特有の知能である「データ駆動インテリジェンス」を創発するアルゴリズム開発を目的とした。データ駆動インテリジェンスを創発する基盤技術として統計的潜在意味解析に着目する。統計的潜在意味解析が利用する基本情報は「共起性」である。事象間の共起性を捉えることで、単純な頻度の数え上げでは発見できない事象間の関係性を抽出することができる。特に、データ上実際に共起した表層的なパターンだけではなく、潜在的に共起するパターンも含まれる。この潜在的共起情報が新たな知の創発につながると考えられる。本研究では特に「カウントデータ」および「医用画像データ」という異なる2つの性質を持つデータに着目し研究を進めた。多くのカウントデータで共通に現れる問題として、特定のデータが高頻度で出現し、その他の多様なデータが低頻度であるという、いわゆるロングテール現象を扱うアルゴリズムの開発を行った。医用画像データ解析では、解析対象である病変データ数は正常データ数よりも圧倒的に少ない状況で効率的に学習するためのアルゴリズム開発を行った。

2.　研究成果

（1）概要

研究成果として主に以下の3つのテーマについて説明する。
研究テーマ A 「大規模ロングテールデータにおける LDA の学習アルゴリズム開発」
研究テーマ B 「ロングテールデータにおける Population bias を緩和するアルゴリズム開発」
研究テーマ C 「医用画像読影支援システム開発」
研究テーマ AB はカウントデータにおけるロングテール現象に対する学習アルゴリズムに関する研究であり、研究テーマ C は医用画像解析に関する研究である。共通する問題として興味に対象とする自称の頻度が非常に低いことが挙げられる。したがって、データ全体としては大規模であるものの重用であると考えられるデータは非常に少ない状況下での学習をテーマと

している。

（2）詳細

研究テーマ A 「大規模ロングテールデータにおける LDA の学習アルゴリズム開発」

カウントデータにおいて統計的潜在意味解析を行う最も重用なモデルの1つとして Latent Dirichlet Allocation (LDA)がある。LDA は、文書や閲覧履歴をカウントデータとして扱ったときに、単語やアイテムの共起現象をモデル化したものである。潜在変数と呼ばれる確率変数をモデルに組み込むことで、潜在的な共起を扱うことができる。このような潜在的な共起は、文書の潜在的なトピックやユーザの潜在的な嗜好を表現する情報としてさまざまな応用分野で用いられている。

LDA の大規模データからの学習アルゴリズムとして、データ集合から部分集合をサンプリングし、部分データからの学習を繰り返すことで、実質的にデータ数に依存しない学習手法が注目を集めている。しかし、従来の研究では、部分データを用いたとしても、部分データ毎の学習において、パラメータの更新に必要な計算量が全データの次元に依存してしまうという問題がある。そのため、カウントデータにロングテールの性質がある場合、テール部分に該当する単語やアイテムを削除することで計算量を予め減らし学習するのが主流であった。しかし、このような前処理をしてしまうと、テール部分に該当するデータが分析対象であったときに、その潜在意味解析を行うことができず、有用な情報を得ることができない可能性がある。

本研究では、部分データにおける学習時に、部分データに現れるデータの種類数のみに依存したアルゴリズムを開発した。これによりテール部分を前処理として削除せずに大規模ロングテールデータにおいて LDA を学習することができる。さらに提案手法は、カウントデータにおける統計モデルの汎化（予測）能力を測る指標である Perplexity において、state-of-the-art である学習アルゴリズムにおいて 5 倍程度効率的なアルゴリズムであることも実験的にわかった。具体的には、state-of-the-art のアルゴリズムが達成する Perplexity をを20%程度の学習データで達成することができた。

研究テーマ B 「ロングテールデータにおける Population bias を緩和するアルゴリズム開発」

リクナビ 2015 のユーザの閲覧行動履歴を解析すると、企業に対する閲覧頻度にロングテール現象があり、特に就職活動初期においては、特定の企業のみを閲覧し企業に関する知識が偏っている可能性があることがわかった。その結果として、最終的に就職する企業に行き着くまでに必要以上にコストがかかり、就職活動が長期化する原因の1つと考えられる。実際に就職活動を円滑に進めているユーザは、就職活動初期においても様々な企業の情報を取得することに長けており、そのユーザに合った企業選びができる傾向にある。このような閲覧履歴にみられる閲覧数の偏りは、カウントデータにしばしばみられるロングテール現象であり、Population bias と呼ばれている。このような偏りを解消するシステムが必要である。

就職活動初期において、ユーザに合った企業を推薦する推薦システムが有用であると考えられるが、推薦システムを構築する際にもこのデータの偏りは問題となる。推薦システムを構築するためには、過去の閲覧履歴データから学習する必要があるが、通常推薦システムで使われる学習アルゴリズムを用いるとデータ数の多いアイテムの推薦精度を重要視することになる。したがって、本研究で対象とする閲覧頻度の低いアイテムを推薦する目的には使うこと

ができない。また、通常、購買情報や5つ星評価のようにユーザの興味が明示的に与えられたデータにおける推薦システムとは異なり、閲覧履歴はユーザの嗜好が明示的には得られない。前者は explicit feedback、後者は implicit feedback と一般的に呼ばれている。本研究では、implicit feedback である閲覧履歴を用いて、explicit feedback である将来的なエントリーを予測する問題であり、この問題設定も従来とは異なる。

本研究で開発したアルゴリズムは、ユーザ・企業それぞれの嗜好を表す潜在変数を学習する際に Popularity effect を表現するコスト関数を設計し、コスト付き最適化問題として学習アルゴリズムを導出する。さらに、ユーザ・企業がもつ様々な情報（出身地、大学、学部、心理テスト結果、TOEIC の点数、留学の有無、所在地、企業規模などなど）を事前知識として組合せることで、このような偏りを緩和するモデルを構築した。構築したモデルは、5つのハイパーパラメータを保持することから、閲覧履歴を用いた交差検証をガウス過程によるベイズ的最適化で自動化した。問題となるのは、今回の目標は最終的にはエントリー予測であるが、エントリーデータをシステムの構築段階では使うことができないため交差検証で用いることができない。そこで、閲覧履歴における1回以上の閲覧予測によってベイズ的最適化交差検証を行ったところ、エントリー予測の精度と高い相関があることがわかった。つまり、エントリーデータを用いること無く閲覧履歴のみを用いてシステムを構築すればエントリー予測の精度を高めることができることがわかった。従来手法にくらべ、2倍程度の精度向上が確認された。

研究テーマ C 「医用画像読影支援システム開発」

医用画像診断装置が多くの医療現場に導入され、検査数はここ数年で増加している。

検査で撮影した画像から病変の有無を判断し診断を行う読影は、読影医と呼ばれる医師によって目視により行われ、東大病院では一検査あたり10～30 分で読影を行い、診断を行うことが現状求められている。しかし、医用画像診断装置の技術進展によって、1 検査あたり300～500 枚の画像を撮影できるようになっており、その時間内で読影するのは、時間的制約が非常に厳しくなりつつあるというのが問題として認識されている。また、医用画像診断では、病変あり（陽性）の判断が行われる件数は、病変なし（陰性）の判断が行われる件数に比べ非常に少ないという性質があるため、医師は、画像をより正確に診断し、数少ない重大な病変を見落とすことのないようにすることが求められる。つまり、医師は短時間で、重大な病変を見落とすことなく診断するという、非常に難しい要求を迫られている。

このような背景の下、我々は機械学習を用いた Computer-Assisted Detection （CAD） ソフトウェア と呼ばれる医用画像読影支援システムを開発している。

医用画像読影支援システムでは、読影後にシステムが病変部位の特定を行い、医師へ特定箇所を提示する支援システムである。医師は提示された病変部位を元に、最終的な診断をする。このような支援システムは、医師の負担を軽減するだけでなく、病変を見落とす危険性も低減させることができる。本研究では、東大病院で臨床実用中の医用画像読影支援システムを他の病院で運用する際の制約や問題点を指摘し、その解決策を提示する。

機械学習を用いて支援システムを構築する際に問題となるのが病変データの希少性である。医療の現場で用いられるシステムを作るためには性能のよい学習アルゴリズムが必須であるが、そのための学習データを作るコストが高い。本研究では、日常診療の読影過程で病

変ラベル付きデータを蓄積することができる機能を支援システムに構築した。医師が日常診療で読影する際に、病変を見つけた場合、その部分を簡単にマーキングすることで、病変ラベルの定義が容易にできる機能を構築した。これにより他施設が支援システムを導入することで、そこで新たに病変ラベル付きデータが蓄積され学習に反映される機能を追加することができる。

他の病院が支援システムを導入する場合、導入初期に十分な病変ラベル付きデータを作成するのは時間的にもコスト的にも難しいため、既存のシステムで用いられているデータを学習に利用することが運用上求められる。しかし、病院間における医用画像診断装置の違いによりデータの性質が変化し、そのままデータを用いたのでは性能悪化が起こる可能性がある。また、そもそも医療におけるデータは個人情報であるため病院間でデータを共有することは難しい。

このように学習データの少ない領域で学習する際に、他の利用可能な領域における学習データを用いることで学習効率を上げる手法は転移学習と呼ばれている。

転移学習を読影支援システムで利用する場合、病院間における医用画像診断装置や設定の違いによりデータの性質が変化し、そのまま学習データを用いたのでは性能劣化が起こる可能性がある。このように転移学習させた結果、学習器の性能を劣化させてしまう現象は「負の転移」と呼ばれている。

本研究では、各々の施設での学習器の出力を転移することで病院間でデータの共有をせずに、「負の転移」に関する性能保証も与える学習する手法を開発した。具体的には、読影支援システムの性能評価に用いられる症例ベースの AUC(Area Under Curve)において、転移したことによる性能の劣化の上限を抑えることで「負の転移」に関する性能保証付き転移学習を可能にした。 また、我々の手法は出力を転移することから異なる学習器間の転移も可能である。例えば、現在の読影支援システムでは Adaboost という学習器が動いているが、畳込みニューラルネットを用いた場合でも Adaboost と畳込みニューラルネット間において転移が可能である。

現在の読影支援システムでは、脳動脈瘤の読影支援を中心としており、脳動脈瘤に関する特徴量を専門家の知見をもとに作成し Adaboost を用いて識別を行っている。この性能自体は現在臨床実験中で臨床に耐えうるものだと判断できるが、将来的に異なる症例に関して読影支援システムを構築する際には、症例ごとに特徴量設計することは非現実的である。そこで、畳込みニューラルネットに基づく読影支援システムの開発も行った。畳込みニューラルネットではボクセルデータそのものを入力として学習するため特徴量設計の必要がない。開発したアルゴリズムは、現行の Adaboost の性能を超えるものであることがわかった。畳込みニューラルネットでは Adaboost とは異なり数多くのハイパーパラメータが存在するため、本研究ではベイズ的最適化による自動チューニングシステムの開発も行った。通常の畳込みニューラルネットを用いる問題設定とは異なり病変検知は、病変データと正常データとのデータ数の偏りが大きいため、単純な適用ではうまくいかない。本研究ではデータ拡張、AUC 最適化に基づく学習、ベイズ的最適化を組み合わせることでこのような問題を解決した。

3. 今後の展開

LDA は応用範囲が広く様々な研究分野で使われているモデルである。近年で1細胞解析など全く事なる分野でも用いられるためこの領域の研究者が大規模なデータを解析する際のアルゴリズムとしてコラボレーションすることが考えられる。東大病院と共同開発しているシステムは、「日常診療におけるデータの取得」と「取得されたデータから機械学習によって病変検知を行うシステムの実験自動化（ハイパーパラメータ自動チューニング）」から成り立つが，この枠組みは病理診断にも応用できると考えている．医用画像読影支援システムは、現在脳動脈瘤に特化しているが、症例の種類数を増やすとともに、細胞診断へ広げるプロジェクトを立ち上げている。

4．　評価

（1）自己評価

統計的潜在意味解析を可能にする代表的なモデルである LDA は引用数が15000件を超えるモデルであり、実に様々な応用研究がなされている。今回基礎研究として開発したアルゴリズムは計算量および汎化能力の点でも最も優れたアルゴリズムであり、応用研究への波及効果は高いものと考えられる。また、応用研究として開発した読影支援システムでは、システムを様々な施設で用いる際の問題点を数理的な背景のある手法を提案することで解決することができた。基礎研究としても負の転移に関する理論保証付きのアルゴリズムは無いため、理論面・応用面で大きな貢献ができたと考えられる。今回開発したシステムやアルゴリズムをより多くの症例・多くの施設で展開することで、国内における医療診断の様々な問題の解決につながると考えられる。

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

ビッグデータを機械学習アルゴリズムの訓練データとして用い，有用な「知能」を人工的に創出する研究が，現在，世界中で活発に行われている．本研究もこのトレンドに沿っているが，「人間のような知能」ではなく，「機械独特の知能」であるデータ駆動インテリジェンスを志向している．そして，人知を持ってしては全体像を掴みきれないビッグデータの分析手法を提案し，医用画像診断などの応用分野も切り拓いている．

論文をトップカンファレンスなどで多数発表しており，優れた学術成果をあげている．特に，統計的潜在意味解析の重要なモデルの一つである Latent Dirichlet Allocation に関しては，ビッグデータのロングテールを捨てることなく，比較的少ないメモリ量で学習可能な方式を提案した．また，民間企業や大学病院と共同で応用研究も行っている．このうち医用画像診断では，学習に使えるデータが少ない病院が，他病院の学習結果を補正利用するために，転移学習の方式を提案し，その有効性を検証している．データの持ち出しが難しい状況でも適用可能な方式として有望である．

要素技術と応用の両面での優れた成果をベースに，今後，体系化や方法論の深化とさらなる応用分野の開拓を行うことを期待したい．

5．　主な研究成果リスト

（1）論文（原著論文）発表

1. Issei Sato, Hiroshi Nakagawa. Stochastic Divergence Minimization for Online Collapsed Variational Bayes Zero Inference of Latent Dirichlet Allocation. The 19th ACM International Conference on Knowledge Discovery and Data Mining (KDD2015). 2015, pp.1035-1044.
2. Issei Sato, Hiroshi Nakagawa, Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Ito Process. The 31st International Conference on Machine Learning (ICML 2014). 2014, pp. 982-990.
3. Issei Sato, Hisashi Kashima, Hiroshi Nakagawa. Latent Confusion Analysis by Normalized Gamma Construction. The 31st International Conference on Machine Learning (ICML 2014). 2014. pp. 1116-1124.
4. Masahiro Kazama, Issei Sato, Haruaki Yatabe, Tairiku Ogihara, Tetsuro Onishi, Hiroshi Nakagawa. Company Recommendation for New Graduates via Implicit Feedback Multiple Matrix Factorization with Bayesian Optimization. The 2016 IEEE International Conference on Big Data. 2016.
5. 佐藤 一誠, 野村 行弘, 林 直人. オンライン転移学習と医用画像読影支援への応用. 日本応用数理学会. 2016.

（2）特許出願

研究期間累積件数：1件

1.
発　明　者: 佐藤 一誠、 石黒 勝彦
発明の名称: 収束判定装置、方法、及びプログラム
出　願　人: 東京大学
出　願　日: 2014/02/28
出 願 番 号: 2014-039036

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

受賞

日本データベース学会 2014 年度　上林奨励賞

著作物

・佐藤一誠(著)、 奥村学(監修):トピックモデルによる統計的潜在意味解析、 コロナ社、 2015 年 3 月

・佐藤一誠(著):ノンパラメトリックベイズ、講談社、 2016 年 4 月