様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ９ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（一般）
研究期間：2019～2023
課題番号：１９Ｈ０４０７１

研究課題名（和文）高次元・大規模・多ドメインデータの特徴抽出と情報統合による統計的学習

研究課題名（英文）Statistical Learning with feature extraction and information integration of High-dimensional, large-scale, multi-domain data

研究代表者

金森　敬文（Kanamori, Takafumi）

東京工業大学・情報理工学院・教授

研究者番号：６０３３４５４６

交付決定額（研究期間全体）：（直接経費）　　13,300,000円

研究成果の概要（和文）：本研究では、高次元・大規模な多ドメインデータを使った統計的学習のフレームワークの構築を目指す．ビッグデータ時代では，様々なドメインで異なるサイズや次元，表現形式の複雑なデータを収集できるが，ドメイン間の相互関係が不明瞭であり，データが増えれば増えるほど知識不足の状態になるパラドックスがある．これを打破するため，各ドメインの関連性を考慮しつつ，データの特徴を抽出し統合することが重要である．本研究ではこの課題を多ドメイン間の相互関係に焦点を当てて定式化する．ヘテロな構造を持つ多ドメインデータのモデリング技術と機械学習アルゴリズムを開発し，理論の深化を目指す．

研究成果の学術的意義や社会的意義
本研究では，異なるデータサイズ，次元，タイプなどの多様なデータを活用し，予測，推論，構造推定など複数のタスクを行う学習アルゴリズムを，数学的な知見に基づいて提案，開発する．理論的解析により予測精度向上のためのパラメーター調整などが容易になり，飛躍的な性能向上が期待できる．理論的知見に基づくアルゴリズムの実装により，画像，音声，タグその他の情報を含むヘテロなデータからの関連性分析などの精度が大きく向上し，機械学習システムの安全性や信頼性を高める基盤を提供する。

研究成果の概要（英文）：This study aims to construct a framework for statistical learning using high-dimensional and large-scale multi-domain data. In the era of big data, diverse and complex data with different sizes, dimensions, and representation formats can be collected across various domains. However, there exists a paradox wherein the relationships between domains are often unclear, leading to a knowledge deficit as data volume increases. To overcome this, it is crucial to extract and integrate features of data while considering the inter-domain relationships. This research focuses on formalizing this task with a focus on inter-domain relationships. It seeks to develop modeling techniques and machine learning algorithms for multi-domain data with heterogeneous structures, aiming to advance the theoretical understanding in this field.

研究分野：機械学習

キーワード：AI　データサイエンス

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

ネットワークが高度に発展している現代社会では，コストに応じてさまざまな種類のデータを収集することができます．例として画像認識のタスクを考えます．データを大別すると，写っている物を示すラベルが付いているデータと，ラベルの無い画像のみのデータがあります．ラベルありデータの収集では，各画像に人手でラベルを付けるコストが掛かります．一方，自然画像や人物画像などラベルなしデータは，インターネット上から大量かつ安価に収集できます．さらに，画像があるウェブサービスのユーザーによってインターネットにアップロードされている場合は，その属性なども付加的データとして収集できることもあります．各データドメインごとに，データ収集のためのコストや情報量が異なります．一般にコストをかけるほど，情報量の多いデータを獲得することができます．

コスト面の制約などから，目標とするタスクに関連するさまざまなタイプのデータを収集することがよくあります．その結果，大量のデータがあるにもかかわらず，データドメイン間の関連が不明瞭になり目標とするタスクにおいて期待したほどの統計的精度が得られないことがあります．端的な例として，ラベルあり・なしの両方のデータを用いるほうが，ラベルありデータのみを用いるときより，データ量が多いにもかかわらず予測精度が低くなる「負の情報転送」とよばれる現象があります．

上記のようなパラドックスを解決するためには，各データドメインにおける統計的学習の精度や信頼性を向上させるだけでは不十分です．さらに一歩進んで，各ドメイン間の関連を考慮しながら，個々のドメインにおける特徴量抽出や次元削減，スパース学習，多様体学習などの方法論を再設計することが求められます．加えて，多ドメインからなるビッグデータ(many domain big data)から，複数のタスクを同時並列的に学習する技術は，資源の効率的な利活用の観点から重要です．

このようなビッグデータ時代を取り巻く複雑なデータ環境を背景に，次のような「問い」が自然に現れてきます：異なるサイズや次元で表される複雑な多ドメインデータを用いて，予測や推論，構造推定など複数のタスクを一括して高精度かつ効率的に遂行するためには，どのような機械学習アルゴリズムが求められているか？ 本研究課題では，この問いの解決に向けて研究を推進することを目標とします．

２．研究の目的

本研究では，次の課題を解決することを目標とします．
(i) 各データドメインにおける特徴抽出とドメイン間の情報統合のための方法論を，多ドメイン間の相互関係を適切に捉えるという観点に立ってモデリングし，理論的に展開する．
(ii) 複雑な相互依存性をもつ多ドメイン・ビッグデータから，複数のタスクに対する統計的推論を効率的に実行するための方法論を構築する．

従来からの典型的な問題設定である単ソース・単タスクドメインだけでなく，多ソース・単タスク，さらには多ソース・タスクドメインのようなデータ環境での統計的学習までカバーする，汎用的方法論を構築することを目指します．ここでタスクドメインとは，回帰分析や判別分析における予測などのタスクを扱うために直接必要となるデータ(例えばラベル付きデータ)が得られるドメインを指します．これに対してソースドメインは，タスクドメインから見て補助的な，ラベルなしデータのようなデータが得られるドメインを指します．それぞれのドメインにおけるデータの統計的性質は一般に異なりますが，ドメイン間は緩やかに関連していることが，実際のデータではよくあります．例えば臨床データなどでは，各病院(すなわち各ドメイン)が扱う専門によって患者データの分布は一般に異なります．しかし，個々の患者に対する検査結果の条件付き分布は病院間で差はない，と仮定できることもあります．

各ドメインでの特徴抽出とドメイン間の情報統合を一体的に行うためには，個々のデータドメインにおける解析法を単純に合わせるだけでは限界があります．「負の情報転送」のような，情報統合におけるネガティブ効果を解決することは，多ドメイン・ビッグデータを扱う上で大きな課題です．この課題に対して，ドメイン間の関連を考慮しながら情報転送するフレームワークを構築し，各タスクドメインにおける統計的学習の信頼性を向上させるための研究を進めます．各データドメイン間の関連を明らかにすることで，従来の多ソース・単タスクドメインのフレームワークでは達成し得なかった，データの高度利活用の可能性を拓くことを目指します．

３．研究の方法

これまで，大別すると以下のような研究テーマについて成果を上げてきました.
(A) 別ドメインでデータの特徴を捉えるための研究：密度稜線推定による多様体学習，変数選択，次元削減といった特徴選択法の開発.
(B) 個別ドメインにおける大規模統計モデルに対する効率的学習アルゴリズムの開発：非正規化統計モデルによる推定法の提案.
(C) 単ソース・単タスクドメインの情報統合：半教師つき学習，共変量シフト下での密度比による回帰分析などに関するアルゴリズム開発と理論展開.
上記の研究に加えて，特徴抽出の方法としてカーネル法における変数選択，情報統合の方法としてドメインマッチングを用いた相互推薦，情報転送のための辞書学習アルゴリズムの理論展開など，本研究課題に関連するさまざまな研究を現在進めています.

これらの進捗状況を考慮し，本研究の成果を最大化するために，つぎに示す課題をそれぞれ以下の方法で順次推進していきます.
単方向型学習：それぞれのソースドメインで特徴抽出し，これらを情報統合しタスクドメインで活用する学習方法を構築します．その際，負の情報転送が生じないように学習フレームワークを設計し，信頼性を理論的に保証します.
双方向型学習：多ソース・多タスクドメイン間，また多数のタスクドメイン間において，それぞれ有益な情報を抽出し，近隣ドメインに情報転送をする設定を考えます．タスクドメインで適切に情報統合し，多ドメイン環境におけるグローバルな学習フレームワークを樹立します．その際，ドメイン間の情報は双方向に共有します．すなわち，タスクドメインにおける予測誤差や推定精度などの情報をソースドメインに逆伝播し，各タスクの精度に合わせて特徴量を再設計するなど，柔軟な学習アルゴリズムを構築します.

４．研究成果

初年度は個別ドメインや単ソース・単タスクドメインに関する研究で得られている上記の(A)，(B)，(C)などの成果を着実に深化，発展させることを計画しました．同時に単方向型学習に対する研究を開始し，２年次以降の研究のための準備を整えました.
・単方向型学習に関する研究では，多ソース・単タスクドメイン環境における「負の情報転送」のメカニズムを解明しました.
共同研究者らと，辞書学習を用いて単ソース・単タスク間の転移学習を行う研究を進め，
辞書学習を基底関数の学習と捉えることで，負の情報転送が生じ得る条件について理論的な成果が得られました.
これに基づき，より一般的なデータ環境において「ドメイン間の相互関係が不明であるために生じる，統計的学習の信頼性の低下」
を回避するための新しい学習アルゴリズムを構築しました.
・個別ドメインに対する研究として，高次元データに内在する複雑な低次元特徴を捉えるための方法である
密度稜線の推定(リッジ推定)を発展させました．そのために，応募者らがこれまでに提案した方法を改良し，推定精度を向上を実現しました.
共同研究者らと強力し，大規模問題に対して有効なアルゴリズムを開発し，多ドメイン環境においても有用な統計的ツールとして展開しました.

本研究の 2 年目以降は，データサイズ，次元，タイプなどが異なる多様なデータドメインにおける予測や推論のタスクを行うための学習アルゴリズムの開発を進めました.
ドメインに共通する不変なデータ構造を推定し抽出することで，適切なデータ解析を行うことができる．ここで問題となるのは「不変性」を適切に定義し，その特定，抽出を効率的に行うことです．特に適切な分布間距離に基づく不変構造の学習について研究を推進しました.
これまで培ってきた単ドメインにおける機械学習アルゴリズムを，不変構造を有する非一様な多ドメイン学習に適用するパラダイムについて，共同研究者らと議論し論文を出版した．さらに自己教師付き学習における表現学習と下流タスクとの関連についても理論的な性質を考察しました．自己教師付き学習を用いると，多数のドメインの集合に共通する不変な表現と，個別のドメインにおいて重要な表現を切り分けることができます．これにより，実データを扱うための多ドメイン学習フレームワークを提供することが可能になります．この観点から研究を推進し，自己教師つき学習の理論的な性質を解明した論文を国際会議にて発表しました．さらにこれまでの研究成果を実装し，広く応用に展開することが重要であることから，学習時には出現しなかったデータラベルに対処するための学習フレームワークであるオープンワールドの学習において，データセットシフト下で効果的に動作するアルゴリズムを開発し，国際会議論文として発表しました.

以上, 単ドメインにおけるロバスト学習, 多様なデータドメインにおける不変な特徴量学習に関する研究, 自己教師つき学習における表現学習の数理的研究などについて研究を推進し, 成果を発表しました.

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　8件／うち国際共著　3件／うちオープンアクセス　4件）

１．著者名 S. Liu, T. Kanamori, and D. J. Williams,	４．巻 23
２．論文標題 Estimating Density Models with Truncation Boundaries using Score Matching.	５．発行年 2022年
３．雑誌名 Journal of Machine Learning Research,	６．最初と最後の頁 1-38
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Mae Yuki、Kumagai Wataru、Kanamori Takafumi	４．巻 144
２．論文標題 Uncertainty propagation for dropout-based Bayesian neural networks	５．発行年 2021年
３．雑誌名 Neural Networks	６．最初と最後の頁 394～406
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.neunet.2021.09.005	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Y. Wada, S. Miyamoto, T. Nakagawa, L. ANDEOL, W. Kumagai, T. Kanamori,	４．巻 21
２．論文標題 Spectral Embedded Deep Clustering	５．発行年 2019年
３．雑誌名 Entropy Journal	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/e21080795	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 WADA Yuichiro、SU Siqiang、KUMAGAI Wataru、KANAMORI Takafumi	４．巻 E102.D
２．論文標題 Robust Label Prediction via Label Propagation and Geodesic k-Nearest Neighbor in Online Semi-Supervised Learning	５．発行年 2019年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 1537～1545
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2018edp7424	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Matsui Kota、Kumagai Wataru、Kanamori Kenta、Nishikimi Mitsuaki、Kanamori Takafumi	4．巻 31
2．論文標題 Variable Selection for Nonparametric Learning with Power Series Kernels	5．発行年 2019年
3．雑誌名 Neural Computation	6．最初と最後の頁 1718～1750
掲載論文のDOI（デジタルオブジェクト識別子） 10.1162/neco_a_01212	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 K. Sudo, N. Osugi, T. Kanamori	4．巻 2
2．論文標題 Numerical Study of Reciprocal Recommendation with Domain Matching	5．発行年 2019年
3．雑誌名 Japanese Journal of Statistics and Data Science	6．最初と最後の頁 221-240
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42081-019-00033-3	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Kanamori Takafumi、Osugi Naoya	4．巻 21
2．論文標題 Model Description of Similarity-Based Recommendation Systems	5．発行年 2019年
3．雑誌名 Entropy	6．最初と最後の頁 702～702
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/e21070702	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 W. Kumagai, T. Kanamori,	4．巻 108
2．論文標題 Risk Bound of Transfer Learning using Parametric Feature Mapping and Its Application to Sparse Coding	5．発行年 2019年
3．雑誌名 Machine learning	6．最初と最後の頁 1975-2008
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10994-019-05805-2	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計8件（うち招待講演　1件／うち国際学会　6件）

1．発表者名
Y. Sanada1, T. Nakagawa, Y. Wada, K. Takanashi, Y. Zhang, K. Tokuyama, T. Kanamori, T. Yamada,

2．発表標題
Deep Self-Supervised Learning of Speech Denoising from Noisy Speeches.

3．学会等名
INTERSPEECH 2022（国際学会）

4．発表年
2022年

1．発表者名
和井田博貴; 和田裕一郎; Andeol Leo; 中川匠; Zhang Yuhui; 金森敬文

2．発表標題
Unified surrogate bounds for kernel-based contrastive unsupervised representation learning.

3．学会等名
第25回情報理論的学習理論ワークショップ(IBIS2022)

4．発表年
2022年

1．発表者名
中川匠; 眞田雄太郎; 和井田博貴; Zhang Yuhui; 和田裕一郎; 髙梨耕作; 山田知典; 金森敬文

2．発表標題
Denoising Cosine Similarity: A Theory-Driven Approach for Efficient Representation Learning

3．学会等名
第25回情報理論的学習理論ワークショップ(IBIS2022)

4．発表年
2022年

1．発表者名
H. Sasaki, J. Hirayama, T. Kanamori,

2．発表標題
Mode estimation on matrix manifolds: Convergence and robustness

3．学会等名
The 25th International Conference on Artificial Intelligence and Statistics (AISTATS2022)（国際学会）

4．発表年
2022年

１．発表者名
H. Sasaki, T Sakai, T. Kanamori,

２．発表標題
Robust modal regression with direct gradient approximation of modal regression risk.

３．学会等名
The Conference on Uncertainty in Artificial Intelligence (UAI2020)（国際学会）

４．発表年
2020年

１．発表者名
M. Uehara, T. Kanamori, T. Takenouchi, T. Matsuda,

２．発表標題
A Unified Statistically Efficient Estimation Framework for Unnormalized Models

３．学会等名
The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)（国際学会）

４．発表年
2020年

１．発表者名
S. Liu, T. Kanamori, W. Jitkrittum, Y. Chen

２．発表標題
Fisher Efficient Inference of Intractable Models.

３．学会等名
The Neural Information Processing Systems (NeurIPS 2019), December 2019.（国際学会）

４．発表年
2019年

１．発表者名
K. Matsui, W. Kumagai, K. Kanamori, M. Nisikimi, S. Matsui, T. Kanamori

２．発表標題
Foundations of transfer learning and its application to multi-center prognostic prediction.

３．学会等名
2019 WNAR/IMS/JR Annual Meeting（招待講演）（国際学会）

４．発表年
2019年

〔図書〕　計1件

1．著者名 D.P.Kroese ほか著，金森 敬文 監訳	4．発行年 2022年
2．出版社 東京化学同人	5．総ページ数 416
3．書名 データサイエンスと機械学習	

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	熊谷　亘 (Kumagai Wataru) (20747167)	東京大学・大学院工学系研究科（工学部）・特任助教 (12601)	
研究分担者	竹之内　高志 (Takenouchi Takashi) (50403340)	政策研究大学院大学・政策研究科・教授 (12703)	
研究分担者	松井　孝太 (Matsui Kota) (50737111)	名古屋大学・医学系研究科・講師 (13901)	
研究分担者	川島　孝行 (Kawashima Takayuki) (60846210)	東京工業大学・情報理工学院・助教 (12608)	
研究分担者	武田　朗子 (Takeda Akiko) (80361799)	東京大学・大学院情報理工学系研究科・教授 (12601)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関