様式 C-19

科学研究費補助金研究成果報告書

平成 22 年　6 月 10 日現在

研究種目：基盤研究（C）
研究期間：2007～2009
課題番号：１９５００１３６
研究課題名（和文）　分散・統合データ解析に対する情報幾何学的アプローチ

研究課題名（英文）　Information geometrical approach for distribution-integration data analysis

研究代表者
赤穂 昭太郎　　（AKAHO SHOTARO）
独立行政法人産業技術総合研究所・脳神経情報研究部門・情報数理研究グループ長
研究者番号：４０３５６３４０

研究成果の概要（和文）：情報幾何の枠組みから新たな分散・統合データ解析法の構築を行った．まず，基礎となる指数分布族における次元圧縮法に基づいて，指数分布族には属さない混合分布族にも適用可能な手法への拡張，確率モデルとしての定式化によるベイズ推定の枠組みへの拡張,次元圧縮とクラスタリングの同時最適化への拡張という３つの拡張を行った．また，新たな学習パラダイムとして，少数の高品質データと大量の低品質データの統合を行う飼い慣らしという枠組みを転移学習に基づき開発した．

研究成果の概要（英文）：Novel methods of distribution-integration data analysis were constructed based on information geometrical framework. First, three extensions of the exponential family dimension reduction have been developed: application to mixture distributions that are not exponential family, Bayesian inference by probabilistic formulation, and simultaneous optimization of dimension reduction and clustering. Next, new machine learning paradigm “taming” which integrates small number of high-quality data and large number of low-quality data based on transfer learning was proposed.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,300,000	390,000	1,690,000
2008 年度	1,100,000	330,000	1,430,000
2009 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：情報幾何学，次元縮約，クラスタリング，ベイズ推定，協調フィルタリング

１．研究開始当初の背景

インターネット上の購買履歴などに基づいた協調フィルタリングにより，個人個人の嗜好に適応した推薦システムの構築が可能になってきた．こうしたパーソナライゼーションに向けたサービスはインターネットに留まらず，医療診断や創薬といった，個人ごとの差が大きく，かつ，個人データを収集可能な他の対象分野にも広がりつつある．

ただし，それらが無条件に進まない理由の一つに，個人情報の取り扱いがある． きめの細かい個人適応をするためには，それ相応の個人情報を収集する必要があるが，個人の側からすれば秘密の情報を知られたくないとか，個人を同定されると困る事情などがあり，かならずしもすべての個人情報を利用できるわけではない．

近年秘密情報を秘匿しながらデータ解析の質をできるだけ落とさないようにするという要請からプライバシー保護データマイニングと呼ばれる枠組みが提案され，暗号や統計といった分野で研究が始まっている． ただし，まだ技術や制度などの面で確立されていない部分が多かった．

一方，協調フィルタリングを大規模に行うために分散データマイニングという枠組みが注目を集めている． これは分散した環境下でデータを収集し，それらを効率的に統合してデータマイニングを行うというもので，大量のセンサーから情報を得るセンサネットワークなどとも関連が深く，社会的なインフラ整備も進みつつあった． 分散データマイニングにおいてもプライバシー保護の観点は重要であり，最終的な統合結果の質をできるだけ落とさないように個人情報を扱う枠組みが求められていた．

２．研究の目的

個人の嗜好や特性に合わせた推薦などのサービスを提供する際にプライバシー保護が問題となる． 本研究課題では，情報幾何的データ解析法を「全体」と「個」という二つの相反する規準が統一的に扱えるように拡張し，プライバシー保護の問題に情報理論的な視点から新たな規準を提案するとともに，その規準を満たしながら精度の高い推薦システムを構築するための手法の開発を目的とする．

３．研究の方法

情報幾何的データ解析をプライバシー保護や協調フィルタリング等に適用可能とするための複数の規準の統一的な規準などの理論的な側面は主に代表者の赤穂が担当し，実際のアルゴリズム構築等は赤穂と分担者の神嶌が密接に連携して開発を行う．

(1) 情報幾何により複数の最適化規準の統一的な記述法を開発（赤穂）
現在カルバックダイバージェンス規準に基づいて行われている情報幾何的データ解析手法をプライバシー保護に適用可能な形にするために，U-ダイバージェンスあるいはTsallis エントロピーを用いたものに拡張する．

(2) 情報幾何的データ解析のプライバシー保護協調フィルタリングへの拡張（赤穂・神嶌）
現在クラスタリングや次元圧縮といった基本的な問題に開発されている情報幾何的データ解析手法を，プライバシー保護を考慮した協調フィルタリングに拡張する．

(3) 実験環境の整備と数値実験による検証（神嶌・赤穂）
前二項に際してシミュレーション環境を構築するための準備(プロトタイプ作成，小規模データでの数値実験）を進め，プライバシーを扱うためのセキュアな環境を構築する．

(4) 成果の学会等での発表（赤穂・神嶌）
得られた成果を国際誌，国際会議，研究会等で発表し，情報収集・情報交換に努める．

４．研究成果

(1) まず分散・統合データ解析の問題設定については，従来研究されてきた転移学習の枠組みを発展させ，分散・統合データ解析のための基本的な数理モデルを「飼い慣らし学習」と名付け，インターネットで収集したデータに対してアルゴリズムを適用し，神嶌と赤穂がいくつかの研究会において発表し，人工知能学会での発表に対して人工知能学会優秀賞を受賞した．具体的には，少数の高品質データと大量の低品質データがあったときにそれらを分散環境下で効率的に統合し，質の高い統計的推定を行うというものである． 集団学習の１手法であるバギングを応用したアルゴリズムを構築し，ソーシャルブックマークのタグ付け問題への適用を行い，実用的にも有効であることを示し，理論面でも単純な確率モデルでアルゴリズムの性能を補償した． これは今後，協調フィルタリングなどの推薦システムを大規模化，高精度化する上で重要な基盤となるものである．

(2) 次に，理論面では，従来指数型分布族に属さないため取り扱いの困難であった混合分布モデルのモデルパラメータを情報幾何学的アプローチを用いて圧縮する手法について研究を行った． 具体的には，混合分布モデルに潜在変数を導入することによって指数型分布族の空間に埋め込むことによって従来法に帰着させるとい

う方針をとる．ただし，そのままでは組み合わせ的な自由度を残すので，それを解消するために，線形計画問題によるネットワーク最適化と組み合わせて準最適化するアルゴリズムを考案し，国際会議等で発表を行った．これは，顧客分析を行う際に，各サイトの情報を混合分布によってクラスタリングし，プライバシーを保ちつつセンターに送って処理できる手法の基礎となるものである．

(3) また，指数分布族のベイズ的な取り扱いについて文字認識など実際のベンチマークデータに対して有効性を確認し，国際誌で発表した．さらに，次元圧縮だけではなくクラスタリングと次元圧縮を同時に最適化することによってよりデータに隠れている構造をきめ細かく抽出できるような拡張も行った．　その推定は一般に複雑な非線形最適化問題となり解くのが困難な形となるが，近年研究が進んでいる変分ベイズ法やラプラス近似法といった近似法を適用することによって実用的な計算量で高品質な推定を行う手法を確立した．　ベイズ的な取り扱いをすることにより，確率的な推論を行うための汎用性をもたせることができ，U-ダイバージェンスやTsallis エントロピーといった規準を導入してロバスト性などの規準を導入することが容易となる．　また，ベイズ推定は特にサンプル数がパラメータ数に比べて少ないときに有効であり，飼い慣らし学習や混合分布モデルといったテーマについても拡張することが重要である．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　5 件）

① 赤穂昭太郎，渡辺一帆，岡田真人，指数型分布族の空間におけるデータ解析法について，統計数理，2010 (印刷中)
② N. Matsumoto, S. Akaho, Y. Sugase-Miyamoto, M. Okada, Visualization of Multi-Neuron Activity by Simultaneous Optimization of Clustering and Dimension Reduction, Neural Networks (印刷中)
③ K. Watanabe, S. Akaho, S. Omachi, M. Okada, Variational Bayesian Mixture Model on a Subspace of Exponential Family Distributions, IEEE Transactios on Neural Networks, 査読有, Vol.20, pp.1783-1796, 2009.
④ S. Akaho, Dimension Reduction for Mixtures of Exponential Families, Artificial Neural Networks (Lecture Notes in Computer Science 5163), 査読有, pp.1-10, 2008.
⑤ K. Watanabe, S. Akaho, M. Okada, Clustering on a Subspace of Exponential Family Using Variational Bayes Method, Proceedings of International Conference on Information Theory and Statistical Learning, 査読有, pp.10-16, 2008.

〔学会発表〕（計　9 件）

① 赤穂昭太郎，神嶌敏弘，品質の異なる二つのデータ集合間の転移学習の解析，情報論的学習理論ワークショップ，2009
② 神嶌敏弘，赤穂昭太郎，転移学習を利用した集団協調フィルタリング，第 23 回人工知能学会全国大会，2009
③ T. Kamishima, M. Hamasaki, S. Akaho, BaggTaming --- Learning from Wild and Tame Data, ECML/PKDD2008 Workshop: Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop, 2008.
④ 神嶌敏弘，濱崎雅弘，赤穂昭太郎，飼い慣らし，人工知能学会全国大会，2008.
⑤ 神嶌敏弘，濱崎雅弘，赤穂昭太郎，飼い慣らしを用いた協調タグ付けのタグ予測，統計関連学会連合大会，2008.
⑥ 渡辺一帆，赤穂昭太郎，大町真一郎，岡田真人，非ガウスデータに関する次元圧縮とクラスタリングの同時最適化と工学的パターン認識への応用，日本神経回路学会全国大会，2008.
⑦ 神嶌敏弘，赤穂昭太郎，参加システムの嗜好パターンが異なる場合の集団協調フィルタリング，人工知能基本問題研究会，2007.
⑧ 渡辺一帆，赤穂昭太郎，岡田真人，変分ベイズ法による混合指数型分布を用いたクラスタリング法，電子情報通信学会ニューロコンピューティング研究会，2007.
⑨ 渡辺一帆，赤穂昭太郎，岡田真人，指数型分布族の部分空間上での変分ベイズ的クラスタリング，情報論的学習理論ワークショップ，2007.

６．研究組織
(1)研究代表者

赤穂 昭太郎　（AKAHO SHOTARO）
独立行政法人産業技術総合研究所・脳神経情報研究部門・情報数理研究グループ長
研究者番号：４０３５６３４０

(2)研究分担者

神嶌 敏弘　(KAMISHIMA TOSHIHIRO)
独立行政法人産業技術総合研究所・脳神経情報研究部門・研究員
研究者番号：５０３５６８２０

(3)連携研究者
（　　）

研究者番号：