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

平成　２９　年　　６　月　１８　日現在

機関番号：　３２７０６
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７５０１１８

研究課題名（和文）メトリックラーニングに基づく大規模実データの分析手法と理論評価に関する研究

研究課題名（英文）A Study on Large Scale Data Analysis Method and Theoretical Evaluation based on Distance Metric Learning

研究代表者

三川　健太（Mikawa, Kenta）

湘南工科大学・工学部・講師

研究者番号：４０７０７７３３

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究課題では，近年その蓄積が容易かつ大規模となっている電子データに対する分析手法として，機械学習の一手法であるメトリックラーニングに着目し，大規模実データへの適用のための各種手法について検討を行った．具体的には，正則化を用いたロバストな解の導出法の提案，データ選択による計算コストの低減法の提案，複数の局所的距離計量を仮定，統合した分析の方法について検討し，その有効性を示した．これらの手法は分析対象データの規模や分析時に重点を置く内容（計算に要するコストや分析精度など）により使い分けることが可能であり，大規模実データへの適用可能性を示した．

研究成果の概要（英文）：The development in information technology highlighted the importance of knowledge discovery from enormous electronic data. We focus on the distance metric learning which is one of the methods of machine learning. In this study, we propose the regularization methods and the way to select the suitable training data in order to reduce computational complexity of distance metric learning. In addition, we propose the way to obtain multiple distance metrics and integrate those in order to improve the classification accuracy. Consequently, we clarify the effectiveness of our proposed methods.
These methods can be used properly according to the characteristics of the analysis (e.g. to gain high performance, low computational complexity and so on). By using these method properly, it can be implemented various types of analysis.

研究分野：　経営情報学

キーワード：　計量距離学習　機械学習　パターン認識　正則化

１．研究開始当初の背景

近年の情報化社会の発展に伴い，多様な種類の電子データが容易かつ大規模に蓄積されるようになった．このような大規模データの分析は，単に情報工学や通信技術の分野に留まらず，Web マーケティングや経営戦略の立案支援など，経営情報工学の分野での重要性がますます高まっている．一方，このようなデータに対し，電算機を用いることで大量のデータを扱い，有益な情報を抽出するための機械学習手法が数多く提案されている．本研究ではこれら機械学習手法のうち，距離構造をデータから学習する手法であるメトリックラーニングに着目する．メトリックラーニングはデータの特徴を考慮した距離構造（マハラノビス距離）を任意の制約条件の下で学習するための手法であり，データの分類やクラスタリングなどのパターン認識手法で主に用いられている．本研究では，メトリックラーニングの手法を EC サイトの購買履歴データ分析や経営業務で蓄積されるテキストデータなど，経営情報分野の諸問題に適用可能な手法として改良し，有効な分析ツールとして整備することを考える．

しかし．これらのメトリックラーニング手法のほとんどは，凸最適化を用いた繰り返し演算を用いて最適解を導出しているため，データが大規模になった場合，計算量の面でその適用が難しい．加えて，その評価は比較的小規模なベンチマークデータを対象としており，実データを用いた場合の有効性，ならびに実データへ特化した手法を検討している研究は少ない．

２．研究の目的

本研究では，近年の高度情報化に伴い，その蓄積が容易かつ大規模となっている電子データに対し，これらへの適用のためのメトリックラーニング手法の構築，ならびに分類誤差の理論的評価を行うことを目的とする．

具体的には，(1) 統計的学習理論，確率論の知見を援用したメトリックラーニング手法の汎化誤差の理論解析，(2) 凸最適化手法を用いた大規模データに対する効率的な計算手法の開発，(3) メトリックラーニング手法の実問題への適用手法の開発，などを研究の軸とし，統計的学習理論や情報数理，統計学などの基盤技術を援用した基礎研究と応用研究の融合を図る．

３．研究の方法

本研究では，実データへの適用のためのメトリックラーニング手法の構築と，そのための分類誤差の理論評価を目的とし，前述の 3 点のテーマについてそれぞれ検討を行った．その研究プロセスとしては主に，正則化手法を用いたロバストなパラメータ推定と学習時のデータ選択アルゴリズムの検討から構成されている．

具体的な内容として，前者の正則化を用いたロバストなパラメータ推定については，既に提案されているメトリックラーニング手法に対し，計量行列のトレースを用いたもの，その l_1 ノルムを用いたものの 2 種類の正則化を行うことにより異なる特徴を持つ計量行列の学習法について検討を行った．前者のトレースに基づく正則化では学習データ数が少ない場合でも計量行列が学習できること，後者の l_1 ノルムに基づく正則化では疎な構造を持つ計量行列を得ることによって分析時の計算コスト削減を検討した．また，計量行列の l_1 ノルムを正則化項とした最適化を行う際には，効率的な計算法として近年注目されている Alternating Direction Method for Multiplier（以下，ADMM）を用いることで効率的に入力データの分類タスクに適した距離構造を学習することが可能となった．

また，研究プロセスの 2 点目の学習時のデータ選択アルゴリズムの検討については，入力データ数が増加してしまった場合，その計算コストは大幅に増大するというメトリックラーニングの問題点に着目したものである．このような点の解決のため，学習時に不要なデータを適切に選択し，計算量の削減を行うための方法について検討を行った．これにより，従来用いられている手法のうち，その性能が優れていると言われている 2 手法において，分析性能を保ったまま学習に必要なデータ数を削減することが可能となった．メトリックラーニング手法の理論解析に関しては，仮定したモデルのパラメータを変化させた場合の挙動についての解析を行うことで新たな知見を得ることができた．

また，研究進行の途中では，入力データの持つ局所的な統計的特徴を把握，抽出するために局所的な距離計量の学習，ならびに得られた局所的距離計量を統合的に用いた新たなメトリックラーニング手法の構築方法についての着想を得ることができ，並行して研究を行った．これらにより，より高性能なメトリックラーニング手法について知見を得ることができた．

４．研究成果

本研究では，3 年間の研究期間を通じて以下に示す通りの研究成果を得ることができた．

(1) 大規模実データへのメトリックラーニング手法の適用において，従来手法に対し計量行列のトレース，ならびにその l_1 ノルムをベースとした正則化項を導入することにより様々な知見を得ることができた．前者の計量行列のトレースに基づく正則化については，従来手法では，学習データ数が少なくパラメータが推定できないような問題でも従来手法以上の性能を達成できることを示した．当該手法は多くのメトリックラーニング手法で用いている繰り返し処理を行わないため，比較的大規模なデータに対し

ても適用可能であり，結果的にその適用範囲を大きく広げることが可能となった．また，後者の l_1 ノルムをベースとした正則化項導入に関しては，得られる最適な計量行列がスパースな構造となるため，分析実施時に必要な計算コストを大幅に低減することが可能となった．前述のトレースに基づく正則化とは異なり，最適解の導出に繰り返し処理が必要となってしまうものの，ADMM を用いることによって効率的な計算ができることを示した．これらにより，従来手法をそのまま適用したのでは分析が難しい問題に対しても従来手法以上の分析性能を達成できることを示した．これらの研究成果は日本経営工学会の学術論文として既に掲載されている．

(2) 大規模なデータを扱う際に生じる計算量増大の問題解決のため，学習時に必要なデータを分析精度の低下を生じさせることなく選択する手法について成果を得ることができた．従来用いられるメトリックラーニング手法では学習データ数が増加してしまうと制約条件数が大幅に増加し，学習に必要な計算コストもまた大幅に増加してしまうことが問題点の一つとして指摘されている．電子商取引サイトなどで蓄積される購買履歴などのデータは増加の一途を辿り，これらの中には特異なデータが含まれていることが多くある．このようなデータを使用せず，従来手法同等の精度を得るためのデータ選択を行うことは大きな意味をなすと言える．この点の解決のため，学習に必要なデータを効率的に選択する方法を提案し，その有効性についていくつかの知見を得ることができた．このようなデータ選択の方法は今後さらに増大する各種データの分析に対して有効となる可能性があり，現在も研究が進行中である．

(3) データの多様化に伴い，一つのデータセットに対し唯一の距離構造を仮定することは分析精度向上の観点から適切ではない可能性がある．この点に着目し，入力データに対して複数の局所的な距離計量を仮定，得られた複数の距離計量を統合的に扱うための方法論について検討を行った．これにより，学習データの局所的な特性を考慮した距離構造を得ることができた．また，これらを用いることで，従来手法と比較しその分析精度を大きく向上させることが可能となった．距離計量を複数仮定することは，導出するための計算コストが仮定した距離計量の数に線形に比例してしまう．本研究で提案した手法ではそれぞれの計算を独立に行うことにより計算コストの増加を低減させることが可能であり，複数台の計算機を必要とするものの，従来手法との計算コストの差はわずかであり，現実的な時間で分析を行うことが可能である．

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. Takashi Maga, Kenta Mikawa, Masayuki Goto : "Data pair selection for accurate classification based on information-theoretic metric learning" , Asian J. Management Science and Applications, 査読有, Vol.3, No.1, pp.61-74, 2017 年 4 月
2. 三川健太, 後藤正幸："カテゴリ毎に異なる計量行列を用いた計量距離学習手法に関する一考察", 日本経営工学会論文誌, 査読有, Vol.66, No.4, pp.335-347, 2016 年 1 月
3. 三川健太, "計量距離学習手法によるパターン認識とその知識発見への応用", 経営システム誌, 査読無, Vol. 25, No. 4, pp. 213-222, 2016 年 1 月
4. 三川健太, 小林 学, 後藤正幸："教師あり学習に基づく l1 正則化を用いた計量行列の学習法に関する一考察", 査読有, 日本経営工学会論文誌, Vol.66, No.3, pp.230-239, 2015 年 10 月
5. Kenta Mikawa, Masayuki Goto : "Regularized Distance Metric Learning for Document Classification and Its Application", Journal of Japan Industrial Management Association, 査読有, Vol.66, No.2E, pp.190-203, 2015 年 7 月

〔学会発表〕（計 16 件）

1. Kenta Mikawa, Manabu Kobayashi, Masayuki Goto, Shigeichi Hirasawa: "A Study on Distance Metric Learning using Distance Structure among Category Centroids", The 17th Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS 2016), Taipei, Taiwan, 2016 年 12 月
2. Kenta Mikawa, Manabu Kobayashi, Masayuki Goto, Shigeichi Hirasawa: "Distance Metric Learning based on Different l1 Regularized Metric Matrices in Each Category", 2016 International Symposium on Information Theory and its Applications, Monterey, California, USA, 2016 年 10 月
3. Shuhei Nakano, Seiya Nagamori, Kenta Mikawa, Masayuki Goto: "A study of Improving Classification Accuracy of

k-nearest Neighbor Based on Local Metric Learning and Adaptive Weighted Ensemble", The 7th Forum for Council of Industrial Engineering and Logistics Management Department Heads (CIEDH2016) & The 5th Institute of Industrial and Systems Engineering Asian Conference (IISEAsia2016), Hong Kong, China, 2016年7月

4. Kenta Mikawa, Manabu Kobayashi, Masayuki Goto, and Shigeichi Hirasawa: "A Study of Distance Metric Learning by Considering the Distances between Category Centroids", 2014 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC2015), CYB-03, City University of Hong Kong, Hong Kong, China, 2014年10月
5. Takashi Maga, Kiichiro Yukawa, Kenta Mikawa, Masayuki Goto: "Data Pair Selection for Improving Classification Accuracy of Information-Theoretic Metric Learning", 2015 Asian Conference of Management Science & Applications (ACMSA2015), Dalian, China, 2015年9月
6. Kenta Mikawa, Manabu Kobayashi, Masayuki Goto, Shigeichi Hirasawa: "A Proposal of l1 Regularized Distance Metric Learning for High Dimensional Sparse Vector Space", 2014 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC2014), San Diego, USA, 2014年10月
7. 中野修平, 永森誠矢, 三川健太, 後藤正幸："局所的距離学習と適応的重み付け和に基づく k 最近傍法の分類精度向上に関する一考察", 日本経営工学会春季大会予稿集, pp.138-139, 東京, 2016年5月
8. 齋藤 洋, 三川健太, 後藤正幸："複数の局所的距離の学習法とその統合による分類手法に関する一考察", 電子情報通信学会 技術研究報告 人工知能と知識処理研究会(AI), Vol.115, No.381, AI2015-50, pp.143-148, 沖縄, 2015年12月
9. 三川健太, 小林 学, 後藤正幸, 平澤茂一："代表元間の距離構造を用いた計量距離学習に関する一考察", 日本経営工学会秋季大会予稿集, pp.236-237, 金沢, 2015年11月
10. 齋藤 洋, 三川健太, 後藤正幸："局所的構造をモデル化可能な計量距離学習に関する一考察", 日本経営工学会秋季大会予稿集, pp.266-267, 金沢, 2015年11月
11. 馬賀嵩士, 湯川 輝一朗, 三川健太, 後藤 正幸："Information-Theoretic Metric Learning の分類精度向上を目的とした学習データペアの選択法", 日本経営工学会 平成27年春季大会 予稿集, pp.160-161, 東京, 2015年5月
12. 山崎史博, 三川健太, 後藤正幸:"Large Margin Nearest Neighbor の分類精度向上を目的とした学習データの重み付けに関する一考察", 第37回情報理論とその応用シンポジウム, SITA2014, 富山, 2014年12月
13. 齋藤洋, 山崎史博, 三川健太, 後藤正幸："低次元計量行列の学習とその結合による計量行列学習法", 計測自動制御学会 システム・情報部門学術講演会, SSI2014, SS27-9, 岡山, 2014年11月
14. 山崎史博, 三川健太, 後藤正幸："Large Margin Nearest Neighbor の分類精度向上を目的とした学習データの重み付けに関する一考察", 計測自動制御学会 システム・情報部門学術講演会, SSI2014, 岡山, 2014年11月
15. 三川健太, 後藤正幸:“カテゴリの統計的特徴を利用した適応的計量距離学習に関する一考察”, 日本経営工学会 平成26年秋季大会予稿集, pp.232-233, 広島, 2014年11月
16. 齊藤 洋, 山崎史博, 三川健太, 後藤正幸:“低次元計量行列の学習とその結合による計量行列学習の計算量削減法”, 電子情報通信学会 技術研究報告 情報理論(IT),Vol.114, No.138, pp.7-12, 兵庫, 2014年7月

〔その他〕
ホームページ等
http://www.info.shonan-it.ac.jp/mikawa-lab/

６．研究組織
(1)研究代表者
三川　健太　（MIKAWA Kenta）
湘南工科大学・工学部・講師
研究者番号：40707733