早稲田大学大学院 創造理工学研究科

博士論文審査報告書

論　文　題　目

正則化に基づく計量距離学習手法と
自動分類への応用に関する研究

Distance Metric Learning Based on Regularization and Its Application for Automatic Classification

申　請　者

三川	健太
Kenta	MIKAWA

経営システム工学専攻　情報数理応用研究

2016 年 2 月

近年，多様かつ大規模なデータの蓄積や取得が容易になり，その自動分析手法が広く研究されると共に，多方面で活用され始めている．これらの自動分析手法の多くは機械学習に基づく手法がベースとなっており，非常に幅広い応用分野と将来可能性の高さから，非常に多くの研究がなされるようになった．これらのうち，予め正解カテゴリが付与された学習データを基に，新規データのカテゴリを予測するための規則を構築する教師あり学習の基本的な手法として，k-NN 法やテンプレートマッチング法と呼ばれる方法が良く知られている．これらの手法では用いる距離尺度によりその性能が左右されることから，学習データの統計的性質を反映させて，適応的に優れた距離尺度を得ようとするメトリックラーニングと呼ばれる手法が提案されている．

メトリックラーニングは，距離尺度としてマハラノビス距離を仮定し，任意の制約条件のもと，新規入力データに対する予測精度を向上させるような距離構造（マハラノビス距離における計量行列）を学習するという方法であり，既に様々な手法が提案されている．通常，教師ありメトリックラーニングでは全ての学習データのペア間の類似，非類似の情報が事前に与えられているという問題設定のもと，それらの情報，並びに分析の精度を向上させると想定される制約条件を用いて最適な計量行列を学習する．この際，計量行列を用いたマハラノビス距離が距離の公理を満たすことを保証するため，計量行列の半正定値性を制約条件とし，繰り返し処理を用いることでその学習が行われる．他方，最適な計量行列を解析的に得ることができる手法として持橋らの手法が存在する.この手法では,カテゴリの代表元を用いることで，最適な計量行列の数値解を導出することが可能となる．このため，繰り返し処理を用いた手法と比較し，学習計算量の面で優れていることが知られている．本論文では，多様なデータへの適用可能性を考え，学習計算量の面で他の手法と比較し優位性のある持橋らの手法に着目している．しかしながら，この手法には扱うデータの特性により，以下に示す問題点がある．

(1) 学習データの次元数に対し，使用可能な学習データ数が著しく少ない場合，最適な計量行列の存在が保証されていない．
(2) 学習データの成分間の一部にしか統計的関係性がないと想定されるデータを対象とした場合，導出された最適な計量行列は多くの要素が誤差による値になってしまう．
(3) 学習データのカテゴリ毎に統計的特徴が異なるデータを対象とした場合，その局所的構造を表現できない．

上記の問題点に対し，本論文ではデータの特性により生じる前述の問題点を改善する手法を構築し,学習計算量,予測計算量のバランスを考慮しつつ，データの自動分類における分類精度を向上させることを目的としている．具体的には，以下の手法を提案し，その性質の解析と評価実験を通じて，その有効性を検証している．

1. 正則化項を付与した最適化問題の定式化と計量行列の学習方法．

2. スパースな計量行列の学習方法.

3. 複数の計量行列の学習方法と分類時の結合方法.

上記の 1.は，計量行列の存在が常に保障されていないという持橋らの手法に対し，正則化を用いることで計量行列の数式解が解析的に求められることを示すと共に,その解が必ず一意に定まることを示している.上記の 2.では，不要な計量行列の要素を 0 とするスパースな計量行列の導出手法を提案している．具体的には，計量行列の l_1 ノルムを正則化項として付与した最適化問題を解くことにより，スパースな計量行列が導出できることを示している．また，得られた計量行列を用いることで，新規入力データの分類時に必要な予測計算量を大幅に削減できることが示されている．さらに 3.では，カテゴリ毎に計量行列の存在を仮定し，各カテゴリの統計的特徴を考慮した複数の距離構造の学習方法を提案している．ここで提案されている手法も，従来手法同様に解析的に最適解を得ることが可能である．さらに，学習した複数の計量行列を用いたデータ間の距離算出法と識別規則が提案されている．

本論文は，以下に示す 7 章から構成されている．

まず，第 1 章では研究の背景を述べ，本研究の目的を示している．

第 2 章では，本研究の準備として，取り扱う問題の設定と従来研究について述べ，マハラノビス距離を始めとする距離尺度や用いる変数の定義を与えている．加えて，本研究の提案に対する従来研究の概要と現状の問題点についてまとめている．

第 3 章では，代表元を用いたメトリックラーニングに対し，最適な計量行列が解析的に得られるという特性を活かしたまま，正則化を行う手法を提案している．この手法によって得られる計量行列が持つ特性についても解析を行うと共に，新聞記事データを用いた分類実験によって提案手法の有効性を示している．

第 4 章では，上記の正則化手法とは異なる，スパースな計量行列を導出する手法について述べている．スパースな計量行列の導出の際には，通常のメトリックラーニング手法と同様に繰り返し処理を用いた計算が必要であるが，ADMM (Alternating Direction Method for Multiplier)と呼ばれる最適化手法を用いることでその最適化が可能となることを示している．加えて，このスパースな計量行列の導出が統計学の一手法である sparse inverse covariance selection（スパース共分散選択）の特殊な場合であることを示しており，本手法によって予測精度を維持しつつ予測計算量を大幅に削減することが可能とされている．

第 5 章では，カテゴリ毎の局所的な統計的特徴を表現可能な計量行列の導出方法について述べると共に，提案した手法を用いることで従来の代表元を用いたメトリックラーニング手法と同様に，解析的に最適な計量行列を求められることを示している．さらに，導出した複数の計量行列を用いることで k-NN 法と同等のデータ間の距離測定を行う手法が構成できることを示し，

分類精度の面で優れていることを明らかにしている．

第 6 章では本研究で得られた考察を述べ，第 7 章では得られた成果をまとめ結論を示すと共に，今後の課題と展望について述べている．

本論文が提案しているメトリックラーニングの手法によって，現実的な学習計算量のもとで，高予測精度に繋がる計量距離学習の方法論が与えられている．提案されている手法は，正則化手法に基づくアプローチにより，非常に自由度が高いという点で過学習を起こし易く学習計算量も大きいメトリックラーニングにおいて,学習の結果が数式解として明示的に与えられる手法,並びに計量行列の不要な要素が 0 となってモデルの説明能力が高まるというメリットを有した手法など，適用場面によって使い分けが可能な学習手法,予測手法を提案している．正則化という技法自体は新しいものではないが,計量行列の最適化問題に適合した正則化の方法を独自に構築し，分類精度や計算量の観点から適用対象によって使い分けの可能な方法を示している．これらの手法は，高次元のテキストデータ等への適用実験を通じて有効性が検証されており,今後,経営工学分野の様々な応用事例への適用も期待できる．今後のさらなる研究の発展性にも期待が持てるという点で，本研究の価値は高い．以上より，本論文は今後のメトリックラーニングの基礎技術に加え,機械学習を用いた大規模データの分析技術に大きく寄与することが期待でき,博士（工学）早稲田大学の学位論文として価値あるものと認める．

2016 年 2 月

審査員

主査	早稲田大学教授	博士（工学）早稲田大学	後藤 正幸
	早稲田大学教授	理学博士（東京工業大学）	高橋 真吾
	早稲田大学教授	工学博士（大阪大学）	永田 靖
	早稲田大学名誉教授	工学博士（大阪大学）	平澤 茂一
	早稲田大学教授	博士（工学）早稲田大学	松嶋 敏泰