早稲田大学大学院 創造理工学研究科

博 士 論 文 概 要

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

2015 年 10 月

近年の情報技術の発展に伴い,多様なデータが膨大に蓄積されるようになった.これらのデータを人手により分析することは物理的に不可能となっているため,その自動分析手法が広く研究されている.このような得られた膨大な量のデータから電算機を用いて知識発見を行う,もしくは有用な情報を抽出する為の手法はデータマイニング,テキストマイニングと呼ばれ,広く普及している.

これらの知識発見手法は機械学習に基づく手法がベースとなっている.本研究ではあらかじめデータに付与された正解(カテゴリ)を基に,新規データのカテゴリを予測するための規則を発見,予測を行う教師あり学習に着目する.加えて,本研究では,入力データ間の距離構造に着目し,その関係性を検討するベクトル空間モデルによるアプローチに着目する.このようなベクトル空間モデルを用いた手法には k-NN 法やテンプレートマッチング法などの手法が存在しているが,これらの手法では用いる距離尺度によりその性能が変化することが知られている.これらの性能を向上させるため,機械学習の分野では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）を学習するための方法であり,既に様々な手法が提案されている.メトリックラーニングは正解データが与えられたもとで予測を行う教師あり学習に基づくメトリックラーニングと,正解データが得られていない状況で新規入力データのクラスタリングや次元削減を行うための教師なし学習に基づくメトリックラーニングに大別される.本研究では,このうち教師あり学習に基づくメトリックラーニング手法に着目する.

一般的に,教師ありメトリックラーニングではすべ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,なら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.この際に,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を行う.

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,繰り返し処理を用いずに最適な計量行列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,計算量の面で優れていることが知られている.

一方で,今日の情報技術の発展に伴い,取り扱うデータも多様化している.特に,データの高次元化,大量化は著しく,従来のメトリックラーニング手法を直接的に用いては,現実的な時間で最適な計量行列を学習することは難しい.このような場合,計量行列の学習を現実的な計算量で行うことは重要な課題であると考えられる.

上記の議論より，多様なデータへの適用可能性を考え，本論文では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する．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存在する．

(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ではこれを適切に表現できない．

(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(1)は，ベクトル表現した学習データの次元数が学習データ数よりも多い時に，分散共分散行列の階数が落ちてしまうことに起因する．このため，特に高次元なデータ，もしくは学習データの次元数に対し著しくデータ数が少ない場合には最適な計量行列の存在が保証されない．

また，計量行列の各要素は学習に使用したデータの成分間の相関関係を表現していると解釈することができる．問題点(2)は，持橋らの手法で導出した計量行列が学習データの成分間に関係性がないものが存在している場合でも各要素へ微小な値を付与してしまうことにより生じる．これにより，学習データの成分間の関係性に過度に適合してしまうこととなり，上記のような構造を持つデータの関係性を適切に表現できない．

問題点(3)は，カテゴリ毎にその統計的特徴が大きく異なる場合，すなわちカテゴリ毎の学習データで特徴量の分布が大きく異なるような場合に大域的な計量行列を学習していることに起因する．これにより，学習データの局所的な構造を相殺してしまい，その特徴を正しく捉えられない可能性がある．

上記の議論より，本研究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以下の手法を構築し，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する．

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．

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が解析的に得られることを示すと共に，それが必ず一意に定まることを示す．

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について提案を行う．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す．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計算量を少なくできることを示す．

最後に 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距離構造の学習方法を検討する．ここで提案する手法についても，従来の持橋らの手法同様に解析的に最適解を得ることができることを示す．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，識別規則の提案を行う．

本論文は 6 章から構成されている．各章の内容は以下の通りである．

第 1 章では本研究の背景と目的について述べる．

第 2 章では本研究における問題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る．その後，本研究で対象とするメトリックラーニングについて説明を行うと共に，持橋らの手法，ならびに繰り返し処理を用いた計量行列学習手法のうち，代表的な手法である ITML (Information-Theoretic Metric Learning), LMNN (Large Margin Nearest Neighbor), Xing らの手法のそれぞれについて説明を行う．

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，正則化を行う手法について示す．また，この手法によって得られた計量行列が持つ特性についても解析を行う．提案した手法の有効性を新聞記事データを用いた分類実験により示す．

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す．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す．

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同様,解析的に最適な計量行列を求められることを示す．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す．

最後に第 6 章は結論であり，本研究により得られた考察を述べ，成果をまとめるとともに，今後の展望について述べる．

早稲田大学　博士（工学）　学位申請　研究業績書

氏 名　三川　健太　　印

（2016 年 2 月 5 日現在）

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
論文	（論文） [1] K. Mikawa and M.Goto, “Regularized Distance Metric Learning for the Document Classification and its Application,” 日本経営工学会論文誌, vol. 66, no. 2E, pp. 190-203, (2015-7) [2] 三川健太, 小林学, 後藤正幸,“教師あり学習に基づくl_1正則化を用いた計量行列の学習法に関する一考察,” 日本経営工学会論文誌, vol. 66, no. 3, pp. 230-239, (2015-11) [3] 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に関する一考察,” 日本経営工学会論文誌, vol.66, no.4, pp. 335-347, (2016-1)
講演	（国際会議） [1] K. Mikawa, T. Ishida, M. Goto and S. Hirasawa, “Regularized Distance Metric Learning and its Application to Knowledge Discovery,” 14th Asia Pacific Industrial Engineering and Management Society (14th APIEMS), (2013-12) [2] K. Mikawa, M. Kobayashi, M. Goto and S. Hirasawa, “A Proposal of l_1 Regularized Distance Metric Learning for High Dimensional Sparse Vector Space,” 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC2014), (2014-10) （講演） [1] 三川健太, 小林学, 後藤正幸, 平澤茂一,“高次元かつスパースなベクトル空間におけるl_1正則化に基づく計量距離学習に関する一考察,” 第 36 回情報理論とその応用シンポジウム予稿集, pp. 703-706, (2013-11) [2] 三川健太, 石田崇, 後藤正幸, 平澤茂一,“l_1正則化を用いた計量距離学習による特徴選択に関する一考察,” 日本経営工学会平成 25 年度秋季大会予稿集, pp. 194-195, (2013-11) [3] 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 年度秋季大会予稿集, pp. 232-233, (2014-11)
その他	（論文） [1] 三川健太, 高橋勉, 後藤正幸,“テキストデータに基づく顧客ロイヤルティの構造分析手法に関する一考察,” 日本経営工学会論文誌, vol. 58, No. 3, pp. 182-192, (2007-8) [2] 三川健太, 増井忠幸, 後藤正幸,“顧客ロイヤルティ構造図に基づく重要要因の定量化手法に関する一考察,” 日本経営工学会論文誌, vol. 59, No. 5, pp. 365-375, (2008-12)

早稲田大学　博士（工学）　学位申請　研究業績書

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
	[3] K. Mikawa, T. Ishida and M. Goto, “An Optimal Weighting Method in Supervised Learning of Linguistic Model for Text Classification,” Industrial Engineering & Management Systems, vol. 11, no. 1, pp. 87-93, (2012-1) [4] 荒川貴紀, 三川健太, 後藤正幸,“未観測カテゴリを含む文書データの自動分類手法に関する研究,”電子情報通信学会論文誌 D, vol.J96-D, no. 8, pp. 1955-1959, (2013-8) [5] 井沢祐介, 三川健太, 後藤正幸,“エージェントベースシミュレーションによる確率潜在空間モデルを用いた推薦システムの評価,”経営情報学会論文誌, vol. 22, no. 2, pp. 1-22, (2013-9) [6] T. Ogihara, K. Mikawa, G. Hosoya, and M. Goto, “Multi-valued Document Classication based on coding theory,” China-USA Business Reviw, vol. 12, no. 9, pp. 911-917, (2013-9) [7] T. Suzuki, G. Kumoi, K. Mikawa, and M. Goto, “A Design of Recommendation Based on Flexible Mixture Model Considering Purchasing Interest and Post-Purchase Satisfaction,” 日本経営工学会論文誌, vol. 64, no. 4E, pp. 570-578, (2014-1) [8] 下村良, 三川健太, 後藤正幸,“大規模テキストデータの分類体系化のための機械学習に基づく半自動分類法の提案,” 日本経営工学会論文誌, vol. 65, no. 2, pp. 51-60, (2014-7) [9] 大井貴裕, 三川健太, 後藤正幸,“評価と購買の両履歴データの学習による確率的潜在クラスモデルの推定精度向上に関する一考察,” 日本経営工学会論文誌, vol. 65, no. 4, pp. 286-293, (2015-1) （国際会議） [1] K. Mikawa, T. Ishida, and M. Goto, “A Proposal of Extended Cosine Measure for Distance Metric Learning in Text Classification,” 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC2011), pp. 1741-1746, (2011-10) [2] K. Mikawa, T. Ishida, and M. Goto, “An Optimal Weighting Method in Supervised Learning of Linguistic Model for Text Classification,” 12th Asia Pacific Industrial Engineering and Management Society (12th APIEMS), ID-141 (2011-10) [3] K. Mikawa, G. Kumoi, K. Suzuki, and M. Goto, “A Proposal of Extracting Unknown Information from Customer Review for SWOT Analysis,” 2011 Asian Conference of Management Science & Applications, ID-167 (2011-10)

早稲田大学　博士（工学）　学位申請　研究業績書

種 類 別	題名、　　発表・発行掲載誌名、　　発表・発行年月、　　連名者（申請者含む）
	[4] K. Mikawa, T. Ishida, M. Goto, and S. Hirasawa, “An Optimal Weighting Method by Using the Category Information in Text Classification based on Metric Learning,” 13th Asia Pacic Industrial Engineering and Management Society (13th APIEMS), No. 25-1 (2012-12) [5] K. Mikawa, T. Ishida, M. Goto, and S. Hirasawa, “A Proposal of Adaptive Metric Learning to Each Category Characteristics for Text Classification,” 2013 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, pp. 544-547, (2013-3) その他国際会議 38 件 （講演） [1] 三川健太，石田崇，後藤正幸，“満足度を考慮したユーザレビューの分析に関する一考察,” 日本経営工学会，平成 22 年度度秋季大会予稿集，pp. 206-207，(2010-11) [2] 三川健太，石田崇，後藤正幸，“拡張余弦尺度を用いた距離学習に関する一考察,” 日本経営工学会，平成 23 年度度春季大会予稿集，pp. 56-57，(2011-5) [3] 三川健太，石田崇，後藤正幸，“文書分類問題におけるカテゴリ情報を用いた適応的重み学習に関する一考察,” 日本経営工学会，平成 24 年度度秋季大会予稿集，pp. 206-207，(2012-10) [4] 三川健太，石田崇，後藤正幸，平澤茂一，“テキスト分類問題におけるカテゴリ情報を用いた適応的距離学習に関する一考察,”電子情報通信学会技術研究報告，情報論的学習理論と機械学習(IBISML)，pp. 83-88，(2012-11) [5] 三川健太，小林学，後藤正幸，平澤茂一，“代表元の距離構造に着目した計量距離学習に関する一考察,” 第 37 回情報理論とその応用シンポジウム(SITA2014) 予稿集，pp. 703-706，(2014-12) [6] 榮枝隼人，三川健太，後藤正幸，“宿泊施設を対象とした評価サイトにおけるユーザレビュー分析に関する一考察,”日本経営工学会　平成 22 年度秋季大会予稿集，pp. 192-193，(2010-10) [7] 井沢祐介，榮枝隼人，三川健太，後藤正幸，“アイテム評価値の高低を考慮した混合メンバーシップブロックモデルによる推薦システム,” 日本経営工学会　平成 23 年度秋季大会予稿集，pp. 36-37，(2011-5) その他講演 55 件