早稲田大学大学院 先進理工学研究科

博 士 論 文 概 要

論 文 題 目

データ分布空間における
距離構造の学習に関する研究

A Study on Distance Metric Learning of
Data Distribution Spaces

申 請 者

日野	英逸
Hideitsu	Hino

電気・情報生命専攻 情報学習システム研究

２０１０年 ５ 月

統計的機械学習やデータマイニング手法はその適用範囲を広げ続けており、多種多様なデータから有用な情報を抽出する課題が日々生まれてきている。多くの学習アルゴリズムの性能は、入力データから抽出する情報の性質と、データの間に定義される距離構造に大きく依存している。例えば、類似したデータをグループに分類するクラスタリング問題を考えるとき、通常はデータに処理を行って分類に有用と考えられる特徴量を抽出し、その特徴量の間に距離を定義する。その上で、定義した距離に従って類似したデータが同一のグループに属するように分類を行う。このとき望ましい結果を得るためには、有用な特徴量を生のデータから取り出し適切な距離を定義しなければならない。具体的な例として、クラスタリングによって文書の分類を行うことを考えると、文書が扱っているトピックの類似度に基づいて分類を行った場合と、文書の文体の類似度に基づいて分類を行った場合では、得られる結果は全く異なるものになることが予想される。また別な例として、個人の嗜好に応じた推薦サービスを考える。多くのオンラインショップサイトでは、過去の購買履歴やアンケートなどから、ユーザが好みそうなアイテムを推薦するサービスを提供している。あるいは、ニュース記事のポータルサイトでも、過去の閲覧履歴や傾向によって、ユーザが関心を持ちそうな記事を推薦するサービスがある。こうした情報推薦を実現するシステムは、一般に推薦システムと呼ばれている。特に、過去の購買履歴に基づきユーザ同士の類似度を定義し、類似度の高いユーザが高評価をしているアイテムを、そのアイテムをまだ購入していないユーザに推薦するシステムを、協調フィルタリングと呼ぶ。協調フィルタリングにおける推薦の良し悪しを決定づけるのは、アイテムの購買履歴という形式で表現されるユーザデータ同士の類似度である。

これまで多くのクラスタリング手法や判別手法が提案され、実際的なデータに適用されて成果をあげているが、精度良く所望の結果を得られる手法は、与えられた問題・データから有用な情報を抽出し、データ間に適切な距離構造を定めることに成功している手法であると考えられる。所与のデータに基づき与えられた課題に応じてデータから情報を抽出し､データ同士の距離構造を学習する手法は、教師無し学習、教師付き学習それぞれの枠組みで数多く提案されている。教師無し学習とは、学習用のデータとしてデータのクラスラベルや応答変数に関する情報が与えられずに、説明変数のみから特徴量を抽出し、その特徴をよく表現する様にデータ間の距離構造を学習する枠組みである。例えば、多変量解析の分野でよく知られている主成分分析は教師無しの距離構造の学習手法の代表例である。主成分分析においては、所与のデータの分散が最も大きくなるという意味でデータの特徴的な構造を捉えた部分空間への射影を学習する。主成分分析によって学習された部分空間にデータを射影した上でデータの比較を行うことで、データの主たる特徴とは無関係なノイズを低減してデータの類似度を比較することが可能となる。最近では、生のデータに対する近傍関係からデータをグラフとして表現

して、スペクトルグラフ理論に基づきデータを低次元空間に埋め込む手法が盛んに研究されている。こうした研究は、データの近傍関係を保存するように低次元空間における距離構造を学習する手法として理解できる。一方、教師付き学習とは、判別や回帰といった課題における入力データに対応する出力例が複数与えられた上で、望ましい入出力関係を記述する写像を学習する枠組みである。教師付き学習の最も代表的な例の一つは、Fisherの判別分析と呼ばれる手法である。これは主成分分析と同様にデータの分散構造に着目した手法であり、データ全体の分散と、データをクラス別に見た場合のクラス内分散の比を最小化することで、各クラスのデータを分離するのに最も適した部分空間への射影を学習する。また、距離構造の学習とはデータをある空間において最適配置する問題と捉えることも出来るため、高度な最適化の手法を用いたアプローチも数多く提案されている。

本論文では、主に教師付き学習の枠組みで、データからの特徴抽出と距離構造の学習問題を扱う。特徴抽出の方法としては、低次元空間への写像、つまり次元削減による少数の特徴抽出のみを考える。距離構造の学習としては、データの変換により直接的に適切な特徴空間を学習するアプローチの他に、データ同士の内積を学習するアプローチや、あるいはデータの生成モデルを学習することで生成モデルに基づく自然な距離構造を学習するというアプローチがある。本論文では、特徴空間の学習及びデータ同士の内積を学習するための統一的な手法として、情報理論に基づく条件付きエントロピー最小化基準による方法を提案する。これにより、従来の距離構造の学習問題では十分に論じられていなかった、学習対象の情報論的な意味が明確になる。また、モデルに基づき自然な距離構造を学習するアプローチの一つとして、近年重要性を増している、映画や書籍等のアイテム評価データの生成モデルを提案し、提案モデルに基づくデータ間の距離を導出する。

本論文の構成を以下に示す。

まず第1章では導入として、本論文で対象とする距離構造の学習という問題の背景と動機を述べ、論文全体の構成を示す。

第2章、第3章では、本論文で用いる情報理論及び統計的機械学習理論のレビューを行う。第2章では、エントロピーや相互情報量といった情報理論における基本的な量を定義する。特に、Shannonの微分エントロピーと、その効率的な推定アルゴリズムを紹介する。第3章では、教師付きの次元削減及び距離構造の学習に関する研究を紹介する。また、データ同士の距離の学習と特徴空間における内積の学習が等価であることを述べた上で、特徴空間における内積を間接的に定めるカーネル関数を用いた手法の総称であるカーネル法の理論的な背景を説明する。ここで、第4章で考察するMultiple Kernel Learningによる特徴空間の距離構造の最適化の理論的背景となる事実を示す。

第4章では、条件付きエントロピー最小化という情報論的な規準に基づくデー

タ処理の枠組みを提案する。この枠組みの中で、データの線型変換による次元削減、距離構造の学習手法を具体的に構成する。さらに、カーネル関数に付随する非線型特徴空間における距離構造の学習を条件付きエントロピー最小化規準の枠組みで行い、近年盛んに研究されている Multiple Kernel Learning のための一手法として定式化する。提案する線型次元削減・距離学習及び Multiple Kernel Learning 手法を、人工データ及び実データに適用し、既存手法との性能比較を行う。標準的な 2 クラス判別問題のベンチマークデータセットに提案手法を適用したところ、線型次元削減手法及び Multiple Kernel Learning 手法ともに、従来の手法と同等以上の性能を示した。

第 5 章では少し視点を変えて、データの生成モデルに基づく類似度及びモデルのパラメタ空間へのデータ配置の学習手法を考察する。まず、映画や書籍、レストランなどへの評価データの新しい生成モデルを提案する。多数のアイテムに対して多数のユーザが比較をおこなったりランキングを与えたりすることで得られるデータを解析して、一つ一つのデータが有する本質的な価値を推定する問題は、心理学や経済学の分野で古くから研究が行われているが、近年機械学習の分野でも注目されている。従来、こうした比較データやランキングデータは Bradley-Terry モデルあるいは Plackett-Luce モデルと呼ばれる確率モデルによってモデル化されてきた。本章では、ランキングデータの生成モデルである Plackett-Luce モデルの自然な一般化として、Grouped Ranking モデルを提案する。このモデルはアイテムの持つ価値パラメタによって特徴付けられるが、尤度関数の直接評価が困難である。そこで、効率的に評価可能な尤度関数の近似を与え、さらに情報幾何学的な考察を通してモデルのパラメタ推定方法を提案する。提案する確率モデルを現実の映画評価データ及び書籍評価データに適用し、Fisher カーネルと呼ばれるカーネル関数を用いてユーザ同士の類似度を定義する。この類似度を、協調フィルタリングにおけるユーザ間類似度として用いたアイテム推薦システムを提案する。従来のアイテム推薦手法では、ユーザによるアイテム評価値をベクトル表現し、その内積や相関を類似度とするが、こうしたデータの扱いは理論的妥当性が十分とは言い難い。一方、提案手法では生成モデルのパラメタ空間にデータを埋め込んだ上で、モデルに基づく自然な内積を定義したことになる。提案する手法による推薦の精度は従来の手法と同等であるが、評価データの生成モデルに基づく手法であることから、ユーザ同士・アイテム同士の関係性を確率分布のパラメタ空間において可視化することが可能であり、推薦の根拠が自然に提示できるという特長を持つ。

第 6 章では本論文の内容をまとめ、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　　日野英逸　　印

（２０１０年　４　月　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	Hideitsu Hino, Yu Fujimoto, Noboru Murata, “A Grouped Ranking Model for Item Preference Parameter”, Neural Computation, Vol.22, Issue 9, 2010
○論文	Hideitsu Hino, Noboru Murata, “Conditional Entropy Minimization Criterion for Dimensionality Reduction and Multiple Kernel Learning”, Neural Computation, (掲載決定)
○論文	Hideitsu Hino, Nima Reyhani, Noboru Murata “Multiple Kernel Learning by Conditional Entropy Minimization”, International Conference on Machine Learning and Application, 2010, （投稿中）
○論文	Yu Fujimoto, Hideitsu Hino, Noboru Murata, “Item-User Preference Mapping with Mixture Models - Data Visualization for Item Preference-”, Proc. International Conference on Knowledge Discovery and Information Retrieval, pp.105-111, Madeira, Portugal, October, 2009.
○論文	Hideitsu Hino, Noboru Murata, “An Information Theoretic Perspective of the Sparse Coding”, Proc. 6th International Symposium on Neural Networks, pp.84-93, Wuhan, China, May, 2009.
○論文	Hideitsu Hino, Yu Fujimoto, Noboru Murata, “Item Preference Parameter from Grouped Ranking Observations”, Proc. 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.875-882, Bangkok, Thailand, April, 2009.
講演	Yu Fujimoto, Hideitsu Hino, Noboru Murata, “An Estimation Method for Bradley-Terry and its Related Models based on the Bregman Divergence”, Learning Workshop, Utah, USA, April, 2010.
講演	日野英逸、村田昇、”条件付きエントロピー最小化に基づく教師付き次元削減手法”、IBIS2009 第１２回情報論的学習理論ワークショップ、福岡、2009 年 10 月
講演	日野英逸、藤本悠、村田昇、”Grouped Ranking モデル: Plackett-Luce モデルの一般化とその応用”、IBIS2008 第１１回情報論的学習理論ワークショップ、仙台、2008 年 10 月
講演	日野英逸、”アイテムの選好度のモデルとパラメタ推定法”, 統計数学セミナー、東京、2008 年 6 月

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）
その他	（論文）Yoshio Uwano, Hideitsu Hino, Yasue Ishiwatari, “Certain Integrable System on a Space Associated with a Quantum Search Algorithm”, Physics of Atomic Nuclei, vol.70, No.4, pp.784-491, 2007.
その他	（論文）Jun Fujiki, Hideitsu Hino, Yumi Usami, Shotaro Akaho, Noboru Murata, “Self-calibration of Radially Symmetric Distortion by Model Selection”, Proc. 20th International Conference on Pattern Recognition, August,2010, （掲載決定）.
その他	（論文）Hideitsu Hino, Yumi Usami, Jun Fujiki, Shotaro Akaho, Noboru Murata, “Calibration of Radially Symmetric Distortion by Fitting Principal Component”, Proc. 13th International Conference on Computer Analysis of Images and Patterns, September, 2009
その他	（講演）吉田朋広，その他７名（申請者は第３著者） “YUIMA プロジェクトについて”、日本数学会 2010 年度年会、 2010 年 3 月
その他	（講演）藤木淳、日野英逸、宇佐見由美、赤穂昭太郎、村田昇、“極射影平面を利用した放射対称歪曲の較正”、パターン認識・メディア理解研究会、2010 年 3 月
その他	（講演）藤木淳、日野英逸、宇佐見由美、赤穂昭太郎、村田昇、“歪曲関数のモデル選択を利用した放射対称歪曲の較正”、画像の認識・理解シンポジウム、2009 年 7 月
その他	（講演）藤木淳、日野英逸、村田昇、赤穂昭太郎、“頑健なヤコビ核主成分分析に向けて”、コンピュータビジョンとイメージメディア研究会、2009 年 3 月
その他	（講演）藤木淳、赤穂昭太郎、日野英逸、村田昇、“主成分曲線のあてはめによる放射対称歪曲の較正”、 パターン認識・メディア理解研究会、2009 年 12 月
その他	（講演）高橋健太、日野英逸、村上隆夫、“生体情報の情報量に関する一考察”、コンピュータセキュリティ研究会、2007 年 7 月
その他	（講演）日野英逸、高橋健太、磯部義明、“入退室管理のための存在確率計算モデル”、数理モデル化と問題解決研究会、2006 年 9 月
その他	（講演）上野嘉夫、日野英逸、石渡康恵、“順序つきデータに対する量子探索アルゴリズムの幾何と力学(1)”、応用数理学会、2005 年 9 月
その他	（講演）上野嘉夫、石渡康恵、日野英逸、 “順序つきデータに対する量子探索アルゴリズムの幾何と力学(2)”、応用数理学会、2005 年 9 月

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）
その他	（特許）日野英逸、高橋健太、磯部義明、“入退室・所在管理システム”、特開 2007-280083
その他	（特許）高橋健太、比良田真史、日野英逸、“生体認証システム、登録端末、認証端末、及び認証サーバ”、特 10-0911594
その他	（特許）高橋健太、日野英逸、“生体情報の特徴量変換方法、その装置およびそれを用いたユーザ認証システム”、特開 2008-129743
その他	（特許）高橋健太、比良田真史、日野英逸、三村昌弘、“生体認証方法およびシステム”、特開 2007-293807
その他	（特許）Kenta Takahashi, Shinji Hirata, Hideitsu Hino, Masahiro Mimura, “Method, system and program for authenticating a user by biometric information”, EP20070005963
その他	（特許）Kenta Takahashi, Shinji Hirata, Hideitsu Hino, “Biometric Authentication System, Enrollment Terminal, Authentication Terminal, and Authentication Server”, 11/862240