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

令和　４ 年　 ６ 月 １５ 日現在

機関番号：３２６８９
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１４４８

研究課題名（和文）様々な低品質データに対応するロバストな分類アルゴリズムの開発

研究課題名（英文）Develop robust classification algorithms for a variety of low-quality data

研究代表者
須子　統太（Suko, Tota）

早稲田大学・社会科学総合学術院・准教授

研究者番号：４０４０９６６０

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では，蓄積されたデータをもとにある特徴量に対応するラベルを予測する分類アルゴリズムについて扱った．分類アルゴリズムの実用では，ノイズ等を含む低品質なデータを用いる場合が多々ある．本研究では様々なノイズを統一的なモデルで表現したもとで高性能な分類アルゴリズムを提案し，理論的な性能限界を導くとともに，性能限界と実アルゴリズムの性能差の解析を行った．

研究成果の学術的意義や社会的意義
現在，画像認識やテキスト分類などの分類アルゴリズムは広く普及しており，一般にも実用されている．しかしながら，実用の場面ではノイズを含む低品質なデータが用いられる事も多く，分類アルゴリズムの本来の性能が発揮できていない場合があり場合によっては十分な分類精度が得られない事がある．本研究の成果を発展させることで，より多くの場面で高性能な分類アルゴリズムが開発できる可能性がある．

研究成果の概要（英文）：This study deals with classification algorithms that predict labels corresponding to certain features based on accumulated data. In practical applications of classification algorithms, low-quality data including noise is often used.
In this study, we proposed a high-performance classification algorithm that uses a unified model to represent various types of noise. Theoretical performance limits of the proposed algorithm are derived. We analyzed the performance difference between the performance limits and the proposed algorithm.

研究分野：機械学習

キーワード：パターン認識　ラベルノイズ　EMアルゴリズム　漸近解析

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

近年，統計学や機械学習，人工知能の活用に対するニーズが急速に広がっている．統計学や機械学習，人工知能の対象とする重要な問題のひとつに分類問題がある．これは，蓄積されたデータをもとに，ある特徴量に対応するラベル（離散値）を予測する問題である．例えば，手書き文字認識や，新聞記事の文章情報から記事内容の表すトピック（政治，経済，スポーツ，など）の予測，EC サイトにおける顧客の購買履歴から対象商品の購買の有無を予測するときなどに用いられる．従来，この分類問題に対しては，機械学習の分野で，サポートベクトルマシンやランダムフォレストといった，非常に高精度な分類アルゴリズムが開発されており，様々な場面で実用化されている．

その一方，データ活用の場面が広がるなか，必ずしも理想的な状況でデータが得られるとは限らなくなっており，低品質なデータに対する分類アルゴリズムの必要性が増してきている．例えば，得られたラベルに誤ったラベルが付与される「ノイズを含むラベルからの学習」や，ラベルありデータが一部しか得られず，残りは全てラベルがない「半教師あり学習」，識別したい２種類のラベルのうち片方のラベルが一部得られるのみで，残りは全てラベルなしデータとなる，「正例とラベルなしデータからの学習」，更には「外れ値を含むデータからの学習」などである．現在ではそれぞれのデータの得られる状況に対し，個別に分類アルゴリズムが研究されており，どのような手法が有効なのかはそれぞれの状況ごとに議論されている．

そのため，実問題に対し分類アルゴリズムを活用しようとした場合，データの得られる状況ごとにアルゴリズムの性能を理解し実装しなければならず，データ分析者の知識やスキルに依存するところが大きい．また，ノイズを含むラベルとラベルなしデータが両方得られるような複合的な状況に対応するのは困難であるという問題もある．

そこで，分類問題におけるこれら低品質データの得られる状況を単一のモデルで表現し，そのもとで高性能な分類アルゴリズムを構成できれば，様々な状況で得られる実データに対し統一的なアプローチで分析が可能となる．

２．研究の目的

当初研究目的としては，データ生成のモデルとして代表的な生成モデルと識別モデルそれぞれに対し，基本的な線形モデルを仮定したもとでの高性能な分類アルゴリズムの開発とそのアルゴリズムを非線形モデルへ拡張することを目的としていた．

しかしながら，研究の進捗に対応するため，計画段階での想定の範囲内の変更として，生成モデルに着目したもとで，目的①「ラベルノイズモデルの更なる一般化および，そのもとでの高性能な分類アルゴリズムの提案」と目的②「提案アルゴリズムの性能の限界に関する理論評価」の 2 つの目的について研究を実施した．

３．研究の方法

目的①については，本研究課題の準備段階で先行的に実施した研究成果である[1]で提案した，一般化ラベルノイズモデルを更に一般化したうえで，[1]で提案されている生成モデルにおける分類アルゴリズムを改良することで，より現実的な状況において有効な分類アルゴリズムを提案する．

目的②については，[1]で提案された一般化ラベルノイズモデルに対し，漸近解析を用いた分類アルゴリズムの性能評価を行うことで，分類性能の理論限界を定理の形で示す．そのもとで，理論限界と提案アルゴリズムの性能を比較することで，提案アルゴリズムの有効性の評価を行う．

４．研究成果

主な研究成果について説明する．

[成果 1]遷移確率未知の一般化ラベルノイズモデルに対する分類アルゴリズムの提案

一般化ラベルノイズモデルとして図１のようなモデルを仮定した．これは，真のラベルが M 値，観測ラベルが L 値を取るとし，真のラベルに対しある遷移確率で観測ラベルが得られるモデルである．この時，遷移確率のパラメータ Λ の値を変えることで，「ノイズを含むラベルからの学習」「半教師あり学習」「正例とラベルなしデータからの学習」「外れ値を含むデータからの学習」を表現できる．

先行的な研究[1]では，遷移確率パラメータ Λ が既知の場合の分類アルゴリズムを提案していたが，遷移確率パラメータ Λ が未知の場合にアルゴリズムを拡張し EM アルゴリズムを

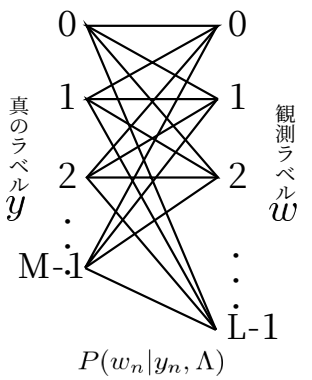

図 1. 一般化ラベルノイズモデル

用いた分類アルゴリズムを提案した．分類精度に関する実験の結果，Λが既知の場合と遜色のない分類精度がでる事が分かった．（図2）

[成果 2]潜在クラスを含むラベルノイズモデルの提案とその分類アルゴリズムの提案

成果 1 のモデルをさらに拡張し，ラベルノイズが特徴量に依存するモデルとして，潜在クラスごとにラベルノイズの遷移確率が異なるモデルを提案した．

そのもとで，EM アルゴリズムを用いた分類アルゴリズムを提案し，実験によりその性能の評価を行った．実験の結果から，真の潜在クラスの数やアルゴリズムで仮定する潜在クラスの数によって提案アルゴリズムの性能にばらつきがでる事が分かった．（図 3）

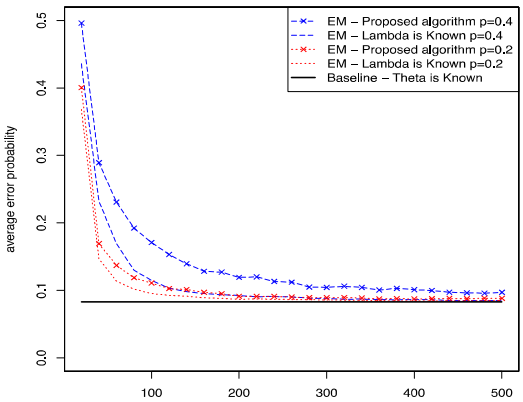

図 2．提案アルゴリズムの分類精度の例

[成果 3]一般化ラベルノイズモデルに対する最尤推定法を用いた分類アルゴリズムの漸近解析

成果 2 で仮定した一般化ラベルノイズモデルに対し，最尤推定（厳密には一致性を満たす尤度方程式の解）によりパラメータの推定を行った後，分類する手法を仮定した場合の漸近的な分類精度の理論解析を行った．成果 1 の EM アルゴリズムは近似最尤推定を行うアルゴリズムであるため，最尤推定による分類手法は成果 1 のアルゴリズムの理論限界を示していると考えられる．

平均分類誤差を特徴量と観測ラベルに関するフィッシャー情報行列の関数として示し，o(1/n)の項を含む厳密な評価を行った．

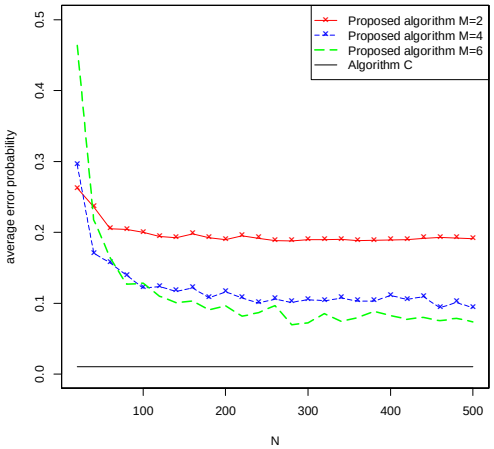

図 3．真の潜在クラスが 6 の場合の提案アルゴリズムの性能の例

[成果 4]理論限界と EM アルゴリズムの差に関する数値解析

成果 1 で提案した EM アルゴリズムによる分類法と成果 3 で示した理論限界とが，実際にどれほど差があるのかについて数値解析を用いた評価を行った．

いくつかの条件のもとで数値解析を行った結果，ラベルノイズの量が比較的少ない範囲では EM アルゴリズムによる分類はほぼ理論限界と同等の性能を示すのに対し，ラベルノイズの量が一定量を超えると理論限界との乖離が大きくなることが分かった．また，EM アルゴリズムの初期値を変えることでこの乖離を小さくできることが分かった．

<引用文献>

[1]須子統太，堀井俊佑，”一般化ラベルノイズモデルにおける分類問題について，” 電子情報通信学会技術研究報告 IBISML2017，pp.377-382，2017 年 11 月．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 安田豪毅，須子統太，小林学，松嶋敏泰
２．発表標題 消失を含むラベルノイズの下での分類に関する性能の限界について
３．学会等名 電子情報通信学会技術研究報告
４．発表年 2020年

１．発表者名 須子統太，安田豪毅，堀井俊佑，小林学
２．発表標題 潜在クラスを含むラベルノイズモデルにおける分類アルゴリズム
３．学会等名 第42回情報理論とその応用シンポジウム
４．発表年 2019年

１．発表者名 安田豪毅，須子統太，小林学，松嶋敏泰
２．発表標題 ラベルノイズ下での分類に関する性能評価について
３．学会等名 第42回情報理論とその応用シンポジウム
４．発表年 2019年

１．発表者名 沢田拓也，西尾和恭，石澤由宇輔，須子統太
２．発表標題 アンケートデータにおける選択バイアスの補正手法の選択方法についての一考察
３．学会等名 情報処理学会第82回全国大会
４．発表年 2020年

１．発表者名
成川広貴，村木鴻介，須子統太

２．発表標題
アンケート分析における不良回答の影響に関する一考察

３．学会等名
情報処理学会第82回全国大会

４．発表年
2020年

１．発表者名
G. Yasuda, T. Suko and T. Matsushima

２．発表標題
Asymptotoic Analysis of Classification in the Presence of Generalized Label Noise

３．学会等名
2018 International Symposium on Information Theory and its Applications（国際学会）

４．発表年
2018年

１．発表者名
須子統太，安田豪毅，堀井俊佑，小林学

２．発表標題
パラメータ未知の一般化ラベルノイズモデルにおける分類法について

３．学会等名
第21回 情報論的学習理論ワークショップ（IBIS 2018）

４．発表年
2018年

１．発表者名
安田豪毅，須子統太，小林学，松嶋敏泰

２．発表標題
一般化ラベルノイズの下での分類に関する漸近評価

３．学会等名
第21回 情報論的学習理論ワークショップ（IBIS 2018）

４．発表年
2018年

１．発表者名 村木鴻介，髙橋朋治，小田陽平，丹生谷英子，須子統太
２．発表標題 不良回答を含むアンケートデータの信頼性向上手法について
３．学会等名 情報処理学会第81回全国大会
４．発表年 2019年

１．発表者名 西尾和恭，岡野拓巳，芝千尋，戸次陸，須子統太
２．発表標題 アンケートデータにおける選択バイアス補正法に関する一考察
３．学会等名 情報処理学会第81回全国大会
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関