様式Ｃ－１９

# 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月２０日現在

機関番号：１２３０１
研究種目：基盤研究（Ｃ）
研究期間：2010～2012
課題番号：２２５００１１９
研究課題名（和文）　不可逆圧縮原理に基づく非ＩＩＤデータ学習手法に関する研究

研究課題名（英文）　Learning from Non-IID Samples based on the Principles of Lossy Compression

研究代表者
安藤　晋（ANDO SHIN）
群馬大学・大学院工学研究科・助教
研究者番号：70401685

研究成果の概要（和文）：

本研究課題では大規模な情報源にみられる非 IID 性データからの学習手法の設計を行い，具体的なエージェント・人間の物理的行動データに関する成果を２件の国際会議で発表した．１件目は異常行動を認識する問題，２件は目的文脈の違いによる行動分布の違いを検出し，非 IID 性を補正する問題を扱い.さらに,非 IID 性学習を一般的に適用するための表現形式を提案し，学会発表(2)にて発表した.経時変化による分布の変化をともなう時系列の分類問題に取り組み，その成果を学会発表 (1)にて発表した.

研究成果の概要（英文）：

In this project, we have developed a principle approach for learning from non-IID data of large-scale information sources. The algorithms developed from this principle were applied to the concrete subjects of physical behaviors of people and autonomous agents. The details of the principles and the algorithms have been published in two international conferences proceedings and with an international journal paper (1). The developed algorithms and benchmark datasets are made public on our website. The algorithms can address the problems of detecting anomalous behaviors and detecting context-specific distributions of behavior patterns, respectively. Furthermore, we developed a general representation model for conducting non-IID data learning presented at oral presentation (2), and a classification model for addressing time-sensitive classification problems presented at oral presentation (1).

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2010 年度 | 1,300,000 | 390,000 | 1,690,000 |
| 2011 年度 | 1,100,000 | 330,000 | 1,430,000 |
| 2012 年度 | 900,000 | 270,000 | 1,170,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 3,300,000 | 990,000 | 4,290,000 |

研究分野：情報学・データ科学
科研費の分科・細目：情報学・知能情報学
キーワード：非 IID データ，不可逆圧縮原理学習

## １．研究開始当初の背景

分散環境で低コスト・大規模にデータを集めるシステムが整備され，その利用法への注目が高まった．
このような情報源では多くの場合独立かつ同一な分布(IID)という統計的学習の仮定をおくことが困難であるため，非均質または複数の情報源(ドメイン)を扱える一般的な方法論が必要とされる．
これまでの問題では．分類，クラスタリング，半教師付き学習といった従来問題で非 IID 情報源を利用する際の負の効果を定量的に評価する理論的背景が無く，ある問題設定で高い性能を示してもその効果がどの範囲まで及ぶのか明確にできなかった．
この問題は知識の転用という人工知能，機械学習の基本テーマに関連し，ドメイン間学習や翻訳学習など複数のトピックで扱われる．国外の研究動向は関連セッションやワークショップが ICML/ECML といった主要会議で開かれ，JMLR 等主要論文誌での発表も増加する等活発であった．また，国内では杉山の共変量シフト（科研番号：20680007），神嶌の協調フィルタリングへの応用（科研番号 21500154）や鈴木のマルチタスク学習（科研番号 21300053)等の取組みがあった．ただし，これまでの研究では有効性の範囲や限界の定量的検証の取り組みは少なかった．
申請者は本課題の事前の成果としてクラスタ構造のドメイン間相違を評価する手法を提案し，さらにドメイン内の非 IID 性について定式化し，協調フィルタリングデータで実験的に検証した．本課題ではこれらの研究を一般化し，データをより多く用意できるメリットと非 IID 性による歪みが増大するデメリットのトレードオフを評価・最適化する原理的な非 IID 学習のアプローチを構築することを目指した．

## ２．研究の目的

本研究課題では，独立かつ同一な分布(IID)という従来学習の前提から外れる非均質な情報源を扱う上での負の効果や有効性の限界を定量的に検証できる枠組は構築することを目的とした．
申請者の過去の成果からクラスタ構造のドメイン間相違の情報量的評価やドメイン内の非 IID 性に関する定式化，および協調フィルタリングにおける検証結果を一般化し，データをより多く用意でき
多くのデータを用意するメリットと非 IID 性により歪みが増大するデメリットを情報量的に評価，最適化する原理的な非 IID 学習アプローチを構築する．
本課題ではさらに，広範囲もしくは長時間に渡って収集されたデータにおける非 IID 性を検証するため軌跡・系列データを中心に新たな応用対象を検討した．
情報源の非 IID 性を記号的・視覚的な提示方法を検討し，データからの”気付き”を促し，意思決定や設計に活かせるような利用方法を検討した．
例として設計変数から生じる非均質性を抽出・可視化するといったユーザ支援の方法論を検討し，効果を検証した．より具体的な目的として下記の内容の検証を目指した．
非 IID データ学習問題のタクソノミーを整理し，各々に情報理論的学習を拡張したプロトタイプ手法を適用した場合の効果を明らかにするため，従来の翻訳学習における分類や特徴選択等目的の違いとデータ・ラベル有無の構成により区別に加え，非 IID 性の形態にも着目して問題クラスを整理することを目標とした．
同時に，個別の設定に関して得られた知見を定式化した上でボトムアップ的に汎用的枠組みを構築することを目指した．
さらに，統一的な知見を踏まえて新しい応用対象を探る．特に情報源の相違を視覚化手法等と組合せてユーザ支援に利用する方法を開発し，その効果を明らかにすることを目指した．

## ３．研究の方法

本研究計画では不可逆データ圧縮の枠組みに基づく学習手法を非 IID データを扱う問題に応用し，理論的・経験的検証を行った．このため，現在個別に扱われている問題クラスを整理した上で，それぞれに適したプロトタイプアプローチを設計した．具体的には機械学習・人工知能の関連トピック（翻訳学習，マルチタスク学習）で個別に扱われている問

題を目的（分類・特徴選択・教師無し学習等）や問題設定により分類し，各々をプロトタイプ手法により経験的に検証した．
問題設定の大枠として(a)連続的な非 IID 性，(b)離散的な非 IID 性を検討した．(a)については申請者の従来成果であるメタクラスタリングを拡張し，ユーザ個別の評価基準を非常に多数のサブクラスとして扱うプロトタイプ手法を用いた．(b)については申請者の従来性かである相互作用情報量クラスタリングをベースとし，目的毎に学習の定式化と歪み評価の実験的検証を行った．
翻訳学習をユーザ支援に利用する方法論の準備として，データの収集・整理を主に 2 次元群ロボット制御プログラムの設計問題を対象として実施した．また，設計過程でのユーザの意図を調査し，設計変数や環境変数，および問題点や成功の評価を実際の軌跡パターンとの関連付けを行った．
経験的検証により得た知見を国内外の研究者とのディスカッションやサーベイを通じて洗練し，ボトムアップ的に汎用的な定式化を導いた．さらに，汎用的な理解に基づき新たな応用対象を系列・軌跡データを中心に検討した．開発した枠組みを応用を指向した時系列データマイニングに手法を開発した．
獲得した知見および開発した手法を成果として国内外の会議や学術ジャーナルにて発表した他，ツールやデータセットを WWW を通じて提供した．

## ４．研究成果

本研究課題の低コストで大規模なデータ収集が可能な情報源を扱うための一般的な方法論を構築することを目的とし，独立かつ同一な分布からの抽出という統計学的前提に従わないサンプル群を扱う非 IID 性データ学習の問題の定式化および原理的手法の設計を実施した．直接的な対象として，センサ機器や画像認識技術の普及により応用が期待される物理的行動データ，エージェントや人体部位を表す多変量時系列を扱った．
本計画では，まず行動データにおける非 IID 性のカテゴリを個人差による行動パターンの分布の違いのようなドメイン内非 IID 性と異なる行動文脈におけるパターン分布の違いのような不連続な非 IID 性の２つに大きく分け，それぞれに対する問題を定式化した．これらを基礎として多変量時系列分類において有効とされる事例ベース・距離ベース手法を設計し，それらのユーザ行動認識，行動データマイニングにおける応用成果を示した．
開発の直接的な対象として，センサ機器や画像認識技術の普及により応用が期待される物理的行動データ，エージェントや人体部位を表す多変量時系列を取り上げベンチマークを整備した．
行動データにおける非 IID 性のカテゴリを個人差による行動パターンの分布の違いのようなドメイン内非 IID 性と異なる行動文脈におけるパターン分布の違いのような不連続な非 IID 性の２つに大きく分け，それぞれに対する問題を定式化した．これらを基礎として多変量時系列分類において有効とされる事例ベース・距離ベース手法を設計し，それらの成果を米応用数学学会国際データマイニング会議(2011)および IEEE 国際データマイニング会議(2011)にて発表した．前者は非 IID 性の一例としてエージェント行動の異常性を認識する問題を扱い，後者では追跡と探索行動といった目的文脈の違いによる確率密度分布の違いを検出し，非 IID 性を補正する方法を示した．この成果をふまえ，人行動の日常的なの動作と危険性事故性のある動作を分類するといった応用を行い．この成果はさらに国際ジャーナル(雑誌論文(1))において発表された．また，行動時系列データ学習において非 IID 性学習，ドメイン交差学習の手法を一般的で適用可能にする表現形式を提案し，IEEE 国際データマイニング会議附設時空間データマイニングワークショップ(2012)において発表した．これを基礎とした，経時変化による分布の変化をともなう時系列の分類問題に取り組み，その成果を米応用数学学会国際データマイニング会議(2013)において発表した．

## 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 1 件）

abc Ando, S.; Thanomphongphan, T.; Seki, Y. & Suzuki, E. Ensemble Anomaly Detection from Multi-resolution Trajectory Features *Data Mining and Knowledge Discovery*, 2013, Online First, pp. 1-45　査読有

〔学会発表〕（計 2 件）

(1)Ando, S., Suzuki, E. Time-sensitive Classification of Behavioral Data Proceedings of the 13th SIAM International Conference on Data Mining, 2013,

pp.458-466,　Austin,Texas,USA

(2)Ando, S. Performance-Optimizing Classification of Time-Series Based on Nearest Neighbor Density Approximation IEEE 12th International Conference on Data Mining Workshops (ICDMW), 2012, pp.759-764, Brussels, Belgium

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
須賀佑太朗，安藤晋，関庸一：人行動分類のための類型パターンに基づく最近傍法．情報処理学会研究報告，2013，2013-MPS-93，pp. 1-5

６．研究組織

(1)研究代表者
安藤　晋（ANDO SHIN）
群馬大学・大学院工学研究科・助教
研究者番号：70401685

(2)研究分担者

(3)連携研究者
（　　）

研究者番号：