様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　５　月　１１　日現在

機関番号：　３２６６０
研究種目：　若手研究(B)
研究期間：　2013～2014
課題番号：　２５７３０１２７

研究課題名（和文）巨大行動事例集合からの知識発見基盤の構築

研究課題名（英文）Development of Knowledge Discovery Basis for Massive Behavioral Data

研究代表者
安藤　晋（ANDO, SHIN）

東京理科大学・経営学部・講師

研究者番号：７０４０１６８５

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究課題では巨大事例集合に関する探索的解析基盤の構築に取り組み，行動センシングデータを具体的な対象として事例集合の非均質性・高速性に由来する共通の問題を克服する成果を挙げた．まず，データ隣接値間の相関が強い系列構造化での判別モデルの学習のためプリミティブパターンを用いた表現とインデクシングによって性能を改善し，実世界データ実験で学習したモデルの可読性と判別性能を同時に向上した．また，
非均質な時間的スケールのパターンを持つデータに対し，平面切断法による多スケール特徴空間での最適化，異常検出のためのクラスタを用いたメタ特徴生成法を提案し，判別時間の短縮や多スケールでの異常検出を可能にした．

研究成果の概要（英文）：This project devoted to building the foundation of the exploratory analysis for very large data sets. It achieved concrete results on behavior sensing data addressing the problems originating from velocity and variety common over very large data sets.
For learning discriminative models under sequence structured data with strong correlations between adjacent observations, we developed indexing based on primitive patterns which improved model interpretability and precision simultaneously in real-world data experiments. Furthermore, we developed cutting-plane method optimization in a multi-scale feature space for temporal data with heterogeneous time-scale and meta-feature generation method for anomaly detection in a multi-scale feature space. These developments made possible the reduction of prediction time and detection of multi-scale anomalies.

研究分野：　知識発見・データマイニング

キーワード：　巨大情報資源マイニング　系列パターンインデクシング　時間非均質性

１．研究開始当初の背景

近年，画像処理，センサ，生体計測技術の進展と普及により実世界空間内のユーザやエージェントの行動を詳細に観測する環境が整備されつつあり，スマートホーム，オフィス，ビルディング等の設備ではトラッキング，センサ，モバイルエージェントを利用して多面的，継続的に行動を観測可能にした．多数の観測情報源を集約した行動事例は高次の時系列であり，継続的な観測によって容易にテラオーダを超える巨大事例集合が蓄積される．蓄積された事例集合は自立生活支援，作業支援，インテリジェントインターフェイス等の人間情報サービスのための重要な知識資源となり得る．

当初，多くの研究エフォートはオンライン認識やコンテクスト利用といった知識活用の方法に向けられているが，実世界行動のより複雑な事象に関する新たな知識，例えばサービスエージェントとユーザのインタラクションが各行動に与える影響等，を発見するには，行動事例集合を悉皆に扱えるデータマイニング・探索的解析の手法に取り組む必要があった．また，巨大事例集合を知識資源として活用する問題についてデータ科学分野での関心が非常に高まっていた．米国 NSF の公募（通称 BigData）は大きな注目を集め，国際会議でのパネルやワークショップも活発化した．

この問題における最大の困難はデータの巨大性であり，従来よりも６桁以上大きい標本数を扱うことからくる．このような事例集合を扱う上では部分標本群を利用した学習が必須であり，そのため分割標本間のバイアス要因，すなわち非均質性を踏まえた学習が必須である．例えば，実世界行動事象ではユーザやエージェントの単位行動と高次意図に関する確率モデルに時間経過や位置の違いによる変化が見られた．本課題ではそのような部分標本毎の特異性を踏まえた学習方法のベースとして申請者の提案による非 IID データ学習，転移学習手法が有効と考え，利用方法を検討した．

２．研究の目的

本研究課題では巨大集合に対する探索的解析の基礎として分割統治を行った際の問題点を評価し，特に部分標本群での判別モデルや類型化を集約する方法を検討した．集約学習の効果に影響すると考えられる巨大事例集合内での分割や経時変化に依存する非均質性及び大域的に学習可能な共通性を明示的にモデル化する．さらに，時系列データにおける特殊な構造を考慮した際の索引法による実行時間減少の効果を検証した．

各技法の単体での効果と限界を検証すると同時に，相互の影響，例えばクエリ処理における近似が集約学習・転移学習に与える影響や非均質性が集約学習に与える影響等を明らかにすることを目指した．そして，ユーザやエージェントの実世界行動データを具体的な対象とした知識発見の応用に取り組んだ．

３．研究の方法

本研究計画では巨大情報資源のマイニング・知識発見の基盤構築のため，巨大事例集合を扱うことが可能な探索的解析手法：分割統治・非均質性分析・索引付けの手法の拡張や統合に取り組んだ．初年度はプロトタイプ手法の理論的解析により各技法の問題を整理し，実世界行動データを対象としたベンチマーク問題を整備した．単体での効果と限界を実験的に検証する．第２年度は複数技法を統合した枠組みを実装し，総合的な効果と相互への影響を定量的に検証した．実世界データへの応用実験を実施し，得られた知見を随時国内外の研究者との議論やサーベイを通じて洗練し，知識発見のための原理的なアプローチとして国内外の会議や学術雑誌にて発表した．

４．研究成果

本研究課題では巨大事例集合に関する探索的解析基盤の構築に取り組み，行動センシングデータを具体的な対象として事例集合の非均質性・高速性に由来する共通の問題を克服する成果を挙げた．

まず，データ隣接値間の相関が強い系列構造化での判別モデルの学習のためプリミティブパターンを用いた表現とインデクシングによって性能を改善し，実世界データ実験で学習したモデルの可読性と判別性能を同時に向上した．

また，非均質な時間的スケールのパターンを持つデータに対し，平面切断法による多スケール特徴空間での最適化，異常検出のためのクラスタを用いたメタ特徴生成法を提案し，判別時間の短縮や多スケールでの異常検出を可能にした．

本研究課題は巨大事例集合に関する探索的解析基盤の構築を目的とし，様々なドメインの巨大事例集合に共通する非均質性・高速性に由来する問題を克服する手法の開発に取り組んだ．本年度は行動センシングにから得られる実世界データを具体的な対象とする成果を挙げた．

まず，行動データにおける高速性に由来する問題として，隣接する値間の相関が強い系列

構造により，一般的な判別モデルの利用が難しいという課題があった．本課題で系列テンプレートと呼ぶプリミティブパターンを用いた表現形式とインデクシング方法を開発した．従来用いられていた Time Series Shapelet と呼ばれるプリミティブパターンでは，データに含まれるパターンを個別のものとしてインデクシングに有用なものを抽出していたが，提案手法ではパターンがばらつきを持つことを踏まえ，半教師付きクラスタリング手法を用いた抽出方法，特徴生成方法，マージン最大化学習を実装し，非均質な部分標本集合間での転移が可能な枠組みを構築した．

距離センサデータを用いた系列データの教師付き学習の実験において，学習したモデルの可読性と判別性能を同時に向上することを実現した．上記の成果についてデータマイニング分野の国際会議において発表した他，和文誌（情報処理学会トランザクション・数理モデル化と問題解決）に採録された．

さらに，行動データの非均質性としてパターンの時間的スケールの違い，すなわち長期のパターンや短期のパターンが含まれる場合において，判別分析や異常検出が十分な効果を得らない点が課題となっていた．これに対し，われわれは平面切断法法による多スケールの特徴空間での最適化手法と，異常検出のためのクラスタを用いたメタ特徴空間生成方法を提案した．これにより，部分標本集合間で長期のパターンや短期のパターンが含まれる場合にも判別分析や異常検出において十分な効果が挙げられる枠組みを構築した．具体的な効果として，判別に要する時間の短縮や多スケールでの異常検出を可能にした．以上の成果についてデータマイニング・知識発見分野の英文誌２誌にて発表した．

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　4　件）

1) Shin Ando; Einoshin Suzuki, “Minimizing Response Time in Time Series Classification,” Knowledge and Information Systems, Online First, 2015 (DOI:10.1007/s10115-015-0826-7)

2) Shin Ando, Theerasak Thanomphongphan, Youichi Seki, Einoshin Suzuki, “Ensemble Anomaly Detection from Multi-resolution Trajectory Features,” Data Mining and Knowledge Discovery, 29, 2015, pp. 39–83 (DOI: 10.1007/s10618-013-0334-x)

3) Shin Ando, “Classifying Imbalanced Data in Distance-based Feature Space,” Knowledge and Information Systems, （2015 掲載決定）

4) 須賀佑太朗，関庸一，安藤晋 「特徴的部分系列に基づく時系列及び形状系列の判別分析」情報処理学会トランザクション誌（数理モデル化と問題解決）（2015 年掲載決定）

〔学会発表〕（計　2　件）

1) Shin Ando, “Discriminative Learning on Exemplary Patterns in Sequential Numerical Data,” 2014 IEEE International Conference on Data Mining, Shengzhen, China (December, 2014)

2) 須賀佑太朗「特徴的部分系列に基づく時系列及び形状系列の判別分析」数理モデル化と問題解決（MPS）研究会（情報処理学会）2015 年 3 月

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究代表者
安藤晋（ANDO, Shin）
東京理科大学・大学院経営学専攻・講師
研究者番号：70401685

(2)研究分担者
なし　　（　　）

研究者番号：

(3)連携研究者

なし（　　）

研究者番号：