様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月１７日現在

研究種目：若手研究（Ｂ）
研究期間：2008～2009
課題番号：２０７００１３２
研究課題名（和文）　レート歪み理論に基づく学習とその高度なクラスタ解析への応用に関する研究

研究課題名（英文）　A Study on Rate-distortion Theory-based Learning and its Application for Advanced Cluster Analyses

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）：
本課題の成果として，レート歪み理論に基づく正規化学習手法（RD 学習）の拡張により，機械学習における先端的かつ実用的な問題への応用手法を開発した．
まず，RD 学習を時系列データの主要なモデル（線形回帰モデル・多変量自己回帰モデル）に拡張する定式化を行った．これにより，多項回帰モデルやマルコフモデルで記述される動的システムに対する異常状態の検出手法を実現した．この手法の効果として遺伝子ネットワークで観測される発現時系列を対象として従来手法よりも有意に高い精度・再現率で活性化状態を検出できることを示した．この成果は機械学習・データマイニングの主要国際会議である KDD にて発表した．

時系列データへの応用に関してはさらにマイクロアレーデータ，金融時系列データ，自律移動ロボット軌跡データ等を整備し，効果的なインスタンスベース異常検出手法を開発している．これは時系列データに内在する多粒度・多視点を同時に考慮した手法であり，データスカッシング・アンサンブル学習等の組み合わせによって効率的なオンラインの異常発見手法を開発した．この成果は国内の研究会にて発表し，データマイニングの主要国際会議に投稿した．

一方，RD 学習の理論を分布が独立かつ均一でないデータを翻訳学習問題設定において利用する拡張を行い，教師無し学習に対する翻訳学習の方法論を提案した．主要な成果として，異種情報源から収集した文書集合に対するクラスタリングを実装し，従来手法よりも有意に高い精度や再現率で文書クラスタを発見できることを示した．さらに，RD 学習の枠組みの自然な拡張により幾何的な構造を正規化に取り入れる方法論を示した．これを検証するために学術文献データをもとにタイトル・著者・トピックに関するグラフ構造情報を付随したベンチマークを準備し，提案手法ＩｔＧＡを適用した．ItGA はトピック発見に関して代表的なＰＬＳＡ，ＬＤＡといった従来手法を上回る性能を示したほか，次元縮約法としても文書に関する重要な特徴を抽出できることを示した．これらの成果はデータマイニングの主要国際会議である ICDM で発表し，さらに最新の成果についても主要国際会議に投稿している．

研究成果の概要（英文）：
Our study achieved the extension of Rate-Distortion (RD) theoretically-principled learning method for practical and leading-edge problems in data mining and machine learning.

One of our concrete achievement is formalizingRD learning for time series data described bymultivariate polynomial regression models and Markov chains. As a result, we developeda methodology for anomaly detection of dynamic systems. We validated our methodswith microarray time series data.We were able to detect the active state of the network with significantly higher precision and recall than the conventional methods.These results were published in KDD,which is one of the major conferences in

Machine Learning/Data Minining.

With respect to the time series data mining, we constructed benchmark datasets for different domains: including microarray data, financial time series, and robot trajectory. We developed an efficient instance-based method for online outlier detection method based on multi-perspective ensemble learning. This results is presented at a Japanese workshop and submitted for an international conference.

We also extended the RD formalization fortransfer learning problems, addressingmultiple, heterogeneous data sourcesand developed a methodology for regularized learning for unsupervised transfer learning.The main concrete result is the clustering of the heterogeneous text data, where significantly higher precision and recall was achieved in comparison to conventional methods.We showed further extension of the RD learningfor integrating geometric structures intoregularization framework. For validating theproposed approach, we prepared a benchmark data from bibliographical data annotated withco-author graph information.We applied ItGA, an information-theoreticGeo-topico analysis, and discovered better topics than popular PLSA and LDA methods.ItGA were significantly better as a dimensionality reduction method to extract important featuresof text. These results are published at ICDMand are in submission for other major Data Mining conferences.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	2,100,000	630,000	2,730,000
２００９年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：知識発見とデータマイニング

１．研究開始当初の背景

機械学習分野の最近の重要な流れとして情報ボトルネック(IB)法や最小 Bregman 情報原理といったレート歪み(RD)理論を基礎とする学習が研究され，情報圧縮の理論的枠組みを利用した学習の解釈が進展した．その結果正規化学習としての有効性が認識されイスラエルや米国の大学がこのテーマの研究をリードし，主に ICDM，NIPS，ICML 等北米の主要国際会議やジャーナルで発表される，一方で国内の理論応用研究は申請者のマイノリティ検出(MD)法のほかは数少ない状況であった．また，機械学習・データマイニングにおいて，データの大規模化と同時に多様な情報源から得られる情報や時間的に分布が変化するの扱いに関する従来手法問題が認識され，多様な手法が提案されつつあった．

２．研究の目的

不可逆情報圧縮の理論に基づくレート歪み（RD）学習は正規化学習としては最小記述長（MDL）原理よりもモデル選択が容易である一方，ベイズ的正規化よりも効率的である．このことからモデル選択が困難な異種混合データや時間的に変化する系列データにおいて従来手法よりも効果的・効率的な学習を可能にすると本給課題では上記の視点から RD 学習を翻訳学習問題設定や時系列データからの異常検出に拡張することを主要な目標とした．

３．研究の方法

異種情報源の学術文献データや金融・自律ロボット軌跡・生物情報時系列データをベンチマークとして整備し，それぞれの問題設定において RD 学習手法と従来手法やモデルベース・マージンベース・インスタンスベース等異なるアプローチにおける特性を比較し，ドメイン固有の知見を蓄積した．その上で，知見を一般化して理論的枠組みを拡張した．

４．研究成果

本課題の成果として，レート歪み理論に基づく正規化学習手法（RD 学習）の拡張により，機械学習における先端的かつ実用的な問題への応用手法を開発した．

まず，RD 学習を時系列データの主要なモデル（線形回帰モデル・多変量自己回帰モデル）に拡張する定式化を行った．これにより，多項回帰モデルやマルコフモデルで記述される動的システムに対する異常状態の検出手法を実現した．この手法の効果として遺伝子ネットワーク（図１）で観測される発現時系列を対象として従来手法よりも有意に高い精度・再現率で活性化状態を検出できることを示した．この成果は機械学習・データマイニングの主要国際会議である KDD にて発表した．

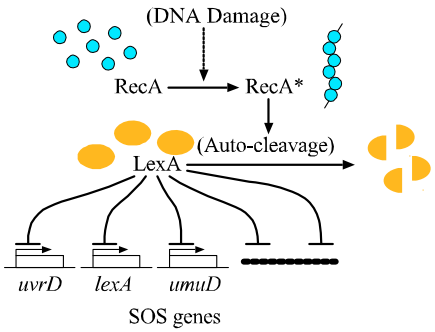

図 １：遺伝子ネットワークの例

時系列データへの応用に関してはさらにマイクロアレーデータ，金融時系列データ，自律移動ロボット軌跡データ等を整備し，効果的なインスタンスベース異常検出手法を開発している．これは時系列データに内在する多粒度・多視点を同時に考慮した手法であり，データスカッシング・メタ特徴表現によるアンサンブル学習の効率的な組み合わせによりオンラインの異常発見手法を開発した（図２）．この成果は国内の研究会にて発表し，データマイニングの主要国際会議に投稿した．

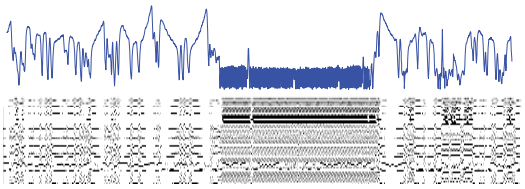

図２：時系列のメタ特徴量表現

さらに，RD 学習の理論を分布が独立かつ均一でないデータを翻訳学習問題設定において利用する拡張を行い，教師無し学習に対する翻訳学習の方法論を提案した（図３）．主要な成果として，異種情報源から収集した文書集合に対するクラスタリングを実装し，従来手法よりも有意に高い精度や再現率で文書クラスタを発見できることを示した．

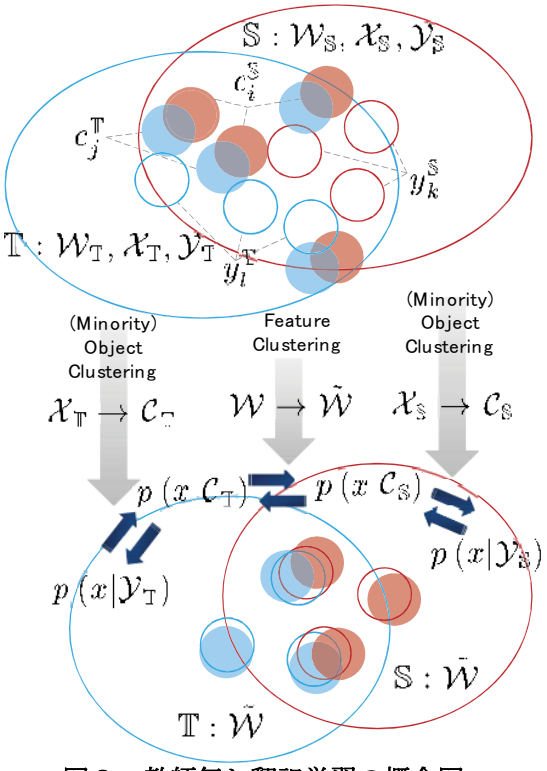

図３：教師無し翻訳学習の概念図

さらに，RD 学習の枠組みの自然な拡張により幾何的な構造を正規化に取り入れる方法論を示した．これを検証するために学術文献データをもとにタイトル・著者・トピックに関するグラフ構造情報を付随したベンチマークを準備し，提案手法ＩｔＧＡを適用した．ItGA はトピック発見に関して代表的なＰＬＳＡ，ＬＤＡといった従来手法を上回る性能を示したほか，次元縮約法としても文書に関する重要な特徴を抽出できることを示した．

これらの成果はデータマイニングの主要国際会議である ICDM で発表し，さらに最新の成果についても主要国際会議に投稿している．

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）
①Shin Ando, Einoshin Suzuki, “Detection of unique temporal segments by information theoretic meta-clustering,” Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.59-68, 2009 査読有

〔学会発表〕（計２件）
① 星野 大祐, Theerasak Tanomphongphang, 安藤 晋, 関 庸一, Swagat, 鈴木 英之進, 異常クラスタのアンサンブルによる特異行動の検出, 第７８回数理モデル化と問題解決研究会（ＭＰＳ７８）, ２０１０．５．２１ 群馬大学荒牧キャンパス情報処理センター

②多賀谷 侑史, 安藤 晋, 関 庸一, サンプルの所属度に応じた可変自己組織化マップ, 第７７回数理モデル化と問題解決研究会（ＭＰＳ７７）, ２０１０．３．５ 伊豆高原ルネッサ赤沢

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
（ ）

研究者番号：

(3)連携研究者
（ ）

研究者番号：