様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 06 月 17 日現在

機関番号：２５４０３
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：２０５００１３９
研究課題名（和文）
時系列データの分類，類似検索，およびラベルづけ
研究課題名（英文）
Classification, Similarity Search, and Labeling of Time Series Data
研究代表者
林　朗　　（HAYASHI AKIRA）
広島市立大学・情報科学研究科・教授
研究者番号：６０２４０９０９

研究成果の概要（和文）：
1．動的時間伸縮（DTW）距離から，時系列データのためのカーネル行列を開発した．半正定値計画法（SDP）を用いて，カーネル行列の正定値性を保証する．二つの応用問題，すなわち時系列データの分類，類似検索のための埋め込みにより，我々のアプローチの妥当性を示した．
2．HHCRF（階層隠れ CRF）を提案した。実験により、パラメータ学習時の訓練集合サイズが大きくなり、かつデータ生成源が非一次マルコフモデルに近づくについて、状態系列推定における HHCRF の性能が HHMM（階層隠れマルコフモデル）より高くなることを示した。

研究成果の概要（英文）：
1. We have developed kernels for time series data using dynamic time warping(DTW) distances. We use semidefinite programming to guarantee the positive definiteness of a kernel matrix. We use two applications, time series classification and time series embedding for similarity search to validate our approach.
２．We have proposed HHCRFs (hierarchical hidden CRFs). In the experiment, we show that HHCRFs perform better than HHMMs (hierarchical hidden Markov models) in state sequence estimation, as the training set size becomes larger and the data source becomes non-Markovian.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,400,000	420,000	1,820,000
２００９年度	1,200,000	360,000	1,560,000
２０１０年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

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：時系列データ，カーネル法，DTW，正定値計画法，隠れマルコフモデル，CRF，ラベルづけ

１．研究開始当初の背景

1.1 DTW カーネルの研究

ビデオ（動画），音声，オーディオ，テキスト，遺伝子情報などの（時）系列データを対象にした分類，類似検索，あるいはラベルづけなどの解析の必要性が高まっている．属性値からなるベクトルで表されるデータと異なり，時系列データはデータによって長さの異なる系列データであり，その扱いは困難である．

機械学習，パターン認識の分野では，過去１０年間でカーネル法の研究が大いに進んできた．カーネル法の長所の一つは，半正定値性カーネルと呼ばれるオブジェクト間の類似度さえ定義されれば，グラフなどの非ベクトルデータも扱える点にある．しかし．これまで時系列データへのカーネル法への適用はほとんどされてこなかった．

1.2 HHCRF の研究

時系列データのラベルづけ問題とは，与えられた一本の時系列データに時々刻々ラベルを付ける問題であり，音声認識では，与えられた連続音声データから音素列あるいは単語列を求める問題に対応する．時系列データのラベルづけには，これまで隠れマルコフモデル(HMM)やそれを階層化した階層隠れマルコフモデル(HHMM)が用いられてきた．近年，テキストマイニング分野で，生成モデルである HMM に代わり，識別モデルである条件付確率場(CRF)が提案され，多くのタスクにおいて HMM の性能を上回ることが確認されている．

２．研究の目的

2.1 DTW カーネルの研究

本研究では，時系列データのマッチングによく用いられる動的時間伸縮（DTW）距離を用いて，時系列データ間のカーネル（DTW カーネルと呼ぶ）を導く．DTW 距離は三角不等式を満たさない擬距離であるため，DTW 距離に基づいて計算されたカーネル行列は半正定値性を必ずしも満たさない．本研究では，半正定値計画法（SDP）を用いてカーネル行列の半正定値性を保証する．なお，SDP は近年パターン認識や機械学習の分野で注目されている最適化手法であり，線形計画法（LP）の拡張と見なすことができる．

2.2 HHCRF の研究

われわれは，HHMM に対応する識別モデル，階層隠れ CRF(HHCRF)，を開発し，実験にてその有効性を確認する．HHCRF は観測系列が与えられた時に，上層状態系列の条件つき確率を表現するモデルである．HHCRF においては，下層状態は隠れ状態であり，確率計算において周辺化される．

３．研究の方法

3.1 DTW カーネルの研究

サンプル時系列データ集合のカーネル行列を求める問題を，近傍保存埋め込み（NPE）とよぶ半正定値計画問題として定式化する．また，新たな時系列データが与えられたとき，NPE で求めたカーネル行列を拡張する問題をサンプル外拡張(OSE)とよぶ半正定値計画問題として定式化する．さらに NPE，OSE により求めた DTW カーネルの有効性を SVM を用いた分類問題，類似検索に関する実験にて検証する．Haasdonk は多次元尺度法(MDS)を援用し，擬距離からカーネルを求めている．求まったカーネルは必ずしも半正定値ではないが，SVM を用いた分類では k 近傍法よりも高い精度を得ている．この手法と，提案手法を実験により比較する．

3.2 HHCRF の研究

Liao らは HHMM を基に階層 CRF を提案した．しかし，階層 CRF は隠れ状態を持たないので，モデルのパラメータ学習時に観測系列とそれに対応する全階層の状態系列を必要とする．

Gunawardana らは HCRF（Hidden CRF，隠れ CRF）を示した． HCRF は２階層のモデルであり，２つの階層のうち下層の状態を隠れ状態と見なす．つまり，観測系列とそれに対応する上層の状態系列のみでパラメータ学習を行える．しかしながら，HCRF の問題点として，上層の状態が固定であり，時間とともに変化しないことが挙げられる．そのために，HCRF は時系列データの分類問題には応用できるが，セグメンテーション問題や状態系列推定問題には応用できない．

これらの研究を参考にして，HHCRF を開発する．まず，HHCRF は HHMM や階層 CRF のように状態を２階層以上の階層構造で表すことができ，HCRF のように下層に隠れ状態を持つものとする．さらに，HCRF の上層の状態が固定であるのに対して，HHCRF の上層の状態は時間とともに変化しなければならない．それにより，HHCRF は時系列データのセグメンテーション問題や状態系列推定問題に応用することが可能となる．

さて，生成モデルと識別モデルの比較のために，Lafferty らは実データに対するシミュレーションとして，非一次マルコフモデルからの人工データをモデル化する実験を行った．

また，Ng らはパラメータ学習時の訓練集合サイズに着目した比較実験を行った．

本研究では，これら二編の論文で得られた生成モデルと識別モデルに関する知見が，生成モデル HHMM と対応する識別モデル HHCRF にも当てはまることを実験により確認する．

４．研究成果

4.1 DTW カーネルの研究

カーネル法の一つの長所として，固定された次元の特徴量を持つベクトルデータだけでなく，様々な種類のデータを扱える事があげられる．本論文では，動的時間伸縮（DTW：

Dynamic Time Warping) 距離を用いて時系列データのためのカーネル行列を提案する．DTW 距離は三角不等式を満たさない擬距離なので，一般的にそれらの距離を用いたカーネル行列は半正定値行列にならない．

我々は，半正定値計画法（SDP：Semidefinite Programming）を用いカーネル行列の半正定値性を保証する．そして，時系列データの局所的幾何を最もよく保存する為の SDP 定式化である近傍保存埋め込み(NPE：Neighborhood Preserving Embedding)，および,NPE に対するサンプル外拡張法(OSE：Out-of-Sample Extension)を提案する．

DTW 距離はパターンマッチングスコアである．よって，値の小さい DTW 距離は信頼できるが，値の大きい DTW 距離は信頼できない．それゆえに，近傍の距離だけに注目した写像を用いた方が良い結果を得られる事が予想される． NPE は最適な近傍間の二乗距離を保存する様な半正定値対称カーネル行列を学習する．そして，最適化されたカーネル行列を固有値分解する事で埋め込み座標を得る．

新規データが与えられた時にカーネル行列を NPE を使って計算し直す事は直感的に自然である．しかし，新規データが得られるたびにカーネル行列を計算すると重い計算コストがかかる． 以前 NPE によって求めたカーネル行列を計算しなおすことなく，拡張する事によって，拡張カーネル行列を得るための手法が OSE である．

提案手法の有効性を示す為に DTW カーネルを用いた時系列分類実験を行う．時系列データを DTW カーネルを用いてベクトル空間に埋め込んだ後，サポートベクトルマシン(SVM：Support Vector Machines)を用いて分類する．テストデータとして，UNIPEN-DTW を使う．UNIPEN-DTW は UNIPEN Train-R01/V07 オンライン手書き文字データ集合から計算された DTW 距離の行列によって構成されている．

我々は，ラベル付きデータとラベルなしデータが最初に同時に与えられる問題設定（以下，transductive setting という）と，ラベル付きデータ与えられた後，順次，ラベルなしデータが与えられる問題設定（以下，sequential setting という）の二つの問題設定のもとで，多クラス分類実験を行った．多クラス分類問題の SVM として，この実験では one-versus-the-rest SVM を用いた．

この実験では， Haasdonk らの Distance Substitution (DS) kernels の実験結果と比較する． DS-Kernels は必ずしも半正定値でない (Not Positive semiDefinite , NPD) カーネルである．この NPD カーネル行列を半正定値性を満たすようにするために２種類のアドホックな変換方法が提案されている．一つは，Cutting off Negative Eigenvalues (CNE)で，負の固有値を切り捨て，その代わりに０とするものである．もう一つの方法は，Reflecting Negative Eigenvalues (RNE)で負の固有値の絶対値をとるものである．なお，これらの CNE と RNE は transductive setting における実験にしか用いる事ができない．

分類実験の結果を，leave-one-out (LOO) errors で評価する．Transductive setting では我々の提案手法の Polynomial Kernel と RBF kernel は，比較対象である DS-Kernels の CNE と RNE の分類結果よりも全体的に良い結果である事がわかる．その例外として，我々の RBF Kernel は 2 つ目のデータにおいて分類結果が悪くなっている．そして Sequential setting では，我々の提案手法は常に DS-Kernel の NPD Kernel よりも良い結果を得る事ができた．それに加え，提案手法は 1-nn と k-nn 分類器よりも良い結果を得ている．

4.2 HHCRF の研究

HMM (Hidden Markov Model) は時系列データの生成モデルとして良く知られている．しかし，近年，HMM に対応する識別モデルである CRF (Conditional Random Field) が提案され，多くの応用問題で有効性が示されている．HHMM (Hierarchical HMM) は HMM を一般化した生成モデルであり，時系列データの状態を階層的に表現する．我々は HHMM に対応する識別モデルとして， HHCRF (Hierarchical Hidden CRF, 階層隠れ CRF) を提案する．

生成モデルと識別モデルの比較のために，Lafferty らは実データに対するシミュレーションとして，非一次マルコフモデルからの人工データをモデル化する実験を行った．また，Ng らはパラメータ学習時の訓練集合サイズに着目した比較実験を行った．本研究では，これら二編の論文で得られた生成モデルと識別モデルに関する知見が，生成モデル HHMM と対応する識別モデル HHCRF にも当てはまることを実験により確認する．

データ生成モデルは一次マルコフモデルと非一次マルコフモデルの混合モデルとする．一次マルコフモデルと非一次マルコフモデルは，階層数を 2，第１層および第２層の状態数を 2 とし，出力は単一ガウス分布に従うとする．また，一次／非一次マルコフモデルの違いとして，非一次マルコフモデルの第１層における横遷移は二次マルコフ過程に従うとする．つまり，次時刻の状態が現時刻の状態だけでなく前時刻の状態にも依存する．また，非一次マルコフモデルの出力は観測値の条件付き独立性を仮定しないとする．具体的には，非一次マルコフモデルの出力確率は,出力平均が前時刻の観測値 ot−1 に依存する式とする．

一次マルコフモデルと非一次マルコフモ

デルの混合比率を(1-α):αとする(0<=α<=1)．すなわち，α=0で完全に一次マルコフモデル，α=1で非一次マルコフモデルとなる．性能を公平に評価するために，データ生成モデルの状態遷移確率と出力確率をそれぞれランダムに設定する．ただし，実験結果の正解率を安定させるための制約として，第２層の終了確率を 0.1，ガウス分布出力の分散を 1 に固定する．また，ガウス分布出力の平均は閉区間[0, 5]からランダムに選択する．

データ集合は観測系列と第１層の状態系列，第２層の指標変数系列から成る集合とする．以下に，訓練集合とテスト集合の詳細を示す．

- 訓練集合：T=100 の N 本のデータ集合から成る．訓練集合サイズの大小を考慮するため，N={1, 10, 100} とする．
- テスト集合：T=100 の 100 本のデータ集合．

訓練集合を用いてモデルパラメータを学習した後，性能評価としてテスト集合に対する状態系列推定を行う．具体的には，Forward-Backward アルゴリズムを用いて事後確率分布を計算し，その値を用いて各時刻 t における第１層（最上層）の状態変数系列の推定値を求める．すべての時刻のうち状態が正しく推定された時刻の比率を計算し，正解率とする．

データ生成モデルを 10 個作成し，作成した各モデルと各αの値に対して 20 回実験を行った結果を以下に示す．まず，訓練集合サイズ 1 の場合，いずれの混合比率α=0.00, 0.25, 0.50, 0.75, 1.00 の値に対しても HHMM の正解率が HHCRF の正解率を上回っている．次に，訓練集合サイズ 10 の場合，α=0, 0.25, 0.5 では HHMM の正解率の方が高く，α=0.75, 1.0 では HHCRF の正解率の方が高い．そして，訓練集合サイズ 100 の場合，全てのαの値に対して HHCRF が HHMM を上回っている．また，訓練集合サイズ 100 の結果からは，αが 1 に近づくにつれて，HHCRF と HHMM の正解率の差が広がることが分かる

実験結果から，パラメータ学習時の訓練集合サイズが大きくなり，かつデータ生成源が非一次マルコフモデルに近づくにつれて，状態系列推定における HHCRF の性能が HHMM のそれよりも，より高くなることが示された．

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）

玉田寛尚，林朗，末松伸朗，岩田一貴，階層隠れ CRF，電子情報通信学会誌，査読あり，J93-D/ 12, 2610-2619，2010

大内悠、岩田一貴、末松信朗、林朗，形状認識のための核関数を用いた形状表現，電子情報通信学会論文誌 D，査読有り，J-92D/ 11, 2011-2021，2009

Hiroyuki Narita, Yasumasa Sawamura, and Akira Hayashi, DTW-Distance Based Kernel for Time Series Data, IEICE Transactions on Information and Systems, 査読有り，E92-D/ 1, pp.51-58,2009

Kumiko Maebashi, Nobuo Suematsu, Akira Hayashi, Component Reduction for Gaussian Mixture Models, IEICE Transactions on Information and Systems, 査読有り，E91-D/ 12, 2846-2853，2008

〔学会発表〕（計 7 件）

Satoshi Kaneko, Hierarchical Hidden Conditional Random Fields for Information Extraction, Learning and Intelligent Optimization (LION5), 2011/01, Rome.

玉田寛尚， HHMMs と HHCRFs の状態系列推定性能に関する比較, 信学技報, 109/ 461, 101-106，2010/03，東京.

小野兼嗣，ソフトマックス行動選択のパラメータ調整の手間を省くための新たな関数の導入, 信学技報, 109/ 461, 107-112, 2010/03, 東京.

鷲頭祐樹，混合ディリクレ過程モデルを用いた ARMA モデルベース時系列クラスタリング，信学技報，109/ 461, 279-284，2010/03，東京.

Kenji Ono, An Action-Selection Strategy Insensitive to Parameter-Settings in Reinforcement Learning, ICROS-SICE International Joint Conference 2009 (ICCAS-SICE 2009) , 2009/08, Fukuoka.

Hirotaka Tamada, Sports Video Segmentation Using a Hierarchical Hidden CRF, 15th International Conference on Neural Information Processing,, 2008/11, Auckland.

Kazunori Iwata and Akira Hayashi, Sampling Curve Images to Find Similarities among Parts of Images, 15th International Conference on Neural Information Processing, 2008/11, Auckland.

〔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：
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
林　朗　（HAYASHI AKIRA）
広島市立大学・情報科学研究科・教授
研究者番号：６０２４０９０９

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：