

京都大学	博士 (情報学)	氏名	松原 靖子
論文題目	Statistical Data Mining for Time-series Dataset (時系列データのための統計的データマイニング)		
(論文内容の要旨)			
本論文では、気象、科学、金融をはじめとする様々な分野で発生している大規模な時系列データのための統計的学習とデータマイニングの問題を対象とし、主に、大規模時系列データを多次元時系列データ、時系列分布データ、時刻印付き複合事象の三つの視点から考察し、理論的な分析とモデルの提案、スケーラブルなアルゴリズムおよび応用の提案を行った。 第一に、多次元時系列データに関し隠れマルコフモデルのための階層的学習方法を提案した。複数の部分系列から構成されるデータストリームが与えられたときに、その典型的な特徴を表す任意の数の隠れマルコフモデルのパラメータを推定する。本研究では、主にデータストリームを扱うため、学習手法は入力に対し増分的かつ高速であり、必要とする記憶領域が少ないことが要求される。提案するモデルは、これらの点を満足し、従来手法と比較し精度を犠牲にすることなく大幅な性能向上を実現した。また、隠れマルコフモデルに基づき、与えられた問合せモデルの特徴に類似した部分系列をデータストリームから検出する手法を開発した。様々な実データを用いた実験を行い、本手法がデータストリームから正確に部分系列を検出し、従来の手法と比較し大幅な性能向上を達成していることを明らかにした。 第二に、時系列分布データとしての大規模分布データのための高速学習手法を開発した。大規模な分布データは産業や経済を初めとする様々な分野で発生しており、これらの解析をすることは大変有用である。しかし、この大規模分布データを対象とする研究はあまり注目されていない。本研究では、大規模分布データを対象とし、大量に発生する分布データを効率よく学習し、モデル化する手法を提案した。本モデルでは、数値データおよびカテゴリカルデータ両方に対し、任意のモデルを用いて学習を行うことが可能であり、従来研究と比較し大幅な性能向上を実現した。また、与えられた問合せデータに対して、複数の分布データの中から類似データを高速に探索する手法を提案した。本手法は、主成分分析を利用したKL情報量の近似計算を高速に行い、分布データを様々な近似の粒度で探索し類似度の低い分布を高速に枝刈りする。また、この手法をさらに拡張し、時系列分布データの類似探索手法を開発した。 第三に、時刻印付き複合事象の解析と予測アルゴリズムの開発を行った。まず、時系列軌跡データを対象とし、時間変化に伴う軌跡の特徴の変化をモデル化した。具体的には、シンガポールのタクシーの一日の利用数のデータを用い、これらのデータを、{GPSの緯度、経度、タクシーの状態(実車、空車、故障)、時刻印}の四つ組の情報からなる列と捉え、それぞれ、個別レベル(フラクタル)と集約レベル(テンソルの主要成分分解)の二つの視点からマイニングを行った。さらに、ウェブのアクセス履歴を対象とし、(URL、ユーザID、時間)の三つの要素から構成される組を解析し、統計的手法を用いることで、潜在的なユーザのアクセスパターンを発見する手法を提案			

した．さらに，本手法の有用な応用として，ユーザの成りすまし発見（異常アクセスの発見）や予測問題（将来どのユーザがどのURLをアクセスするか）に取り組み，従来手法に比べ，高い精度でこれらの問題を実現することを示した．

注) 論文内容の要旨と論文審査の結果の要旨は1頁を38字×36行で作成し、合わせて、3,000字を標準とすること。

論文内容の要旨を英語で記入する場合は、400～1,100wordsで作成し
審査結果の要旨は日本語500～2,000字程度で作成すること。

(論文審査の結果の要旨)

本論文は時系列データを対象とした、統計的学習とマイニングを扱っている。時系列データは、経済や科学の様々な分野で大量に発生しているデータであり、これらの時系列データを対象とするマイニングは、将来の予測を始めとするアプリケーションへの発展が可能である。本論文で提案する技術は、マイニングの分野の中核となる要素技術であり、本論文の対象である技術の適用範囲は広い。

本論文の基本的な貢献として時系列分布データを対象とした理論的な分析を行うと同時に、自然現象を的確に表現するモデルの提案を行っている点を挙げることができる。具体的には、時系列データを、多次元時系列データ、時系列分布データ、時刻印付き複合事象の三つの視点から捉え、それぞれを考察し適切なモデルの提案を行っている。隠れマルコフモデルのための階層的学習方法では、与えられた複数の時系列データを表現する任意個数のモデルを与えるための階層的データ構造を提案している。また、隠れマルコフモデルに基づき、与えられた問合せモデルの特徴に類似した部分系列をデータストリームから検出する手法を提案した。また、大量の時刻印分布データに対し主成分分析を利用したKL情報量の近似計算を行うことにより、精度を犠牲にすることなく高速に与えられたものと類似する分布データを求める手法を提案した。さらに、時刻印付き複合事象の解析として時系列軌跡データを対象とし時間変化に伴う軌跡の特徴の変化をモデル化した。これらはいずれも問題設定とモデルの新規性を認めることができる。

また、上記の理論的分析とモデルをふまえ、実世界で有用な様々な応用とアルゴリズムを提案している。具体的には、先に述べた三つの視点について、それぞれ、パターン発見、探索問題、予測問題等の、実社会で非常に重要とされる問題について取り組んでいる。これらは、モーションキャプチャデータ、ウェブページアクセスデータ、タクシー移動軌跡データなど実際のデータであり、提案モデルおよび手法を実装し有効性を実データによって検証している点も評価できる。また、実社会におけるアプリケーションとして最も重要な問題である、大規模データへの適用に関しても考慮している。より具体的には、本論文で提案したアルゴリズムが既存手法と比較して、スケーラブルである点を示している。この点において、本論文で提案されたモデル及びアルゴリズムの有用性は高いといえる。

以上のように、本研究は、時系列データを対象とした、統計的マイニングを行ったものであり、理論的な分析を行うと同時に、大規模なデータを用いた実証実験結果に裏付けられたスケーラブルなアルゴリズムの可能性を示した点で、この分野における新規性と有用性が高く評価される。

よって、本論文は博士(情報学)の学位論文として価値あるものと認める。

また、平成24年2月10日に実施した論文内容とそれに関連した試問の結果合格と認めた。

注) 論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨(例えば「平成 年 月 日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」)を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日: 年 月 日以降