様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　３０　日現在

機関番号：　１４４０１
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１６０５２

研究課題名（和文）事象系列データからの因果性マイニングと地震および損傷間の因果発見への応用

研究課題名（英文）Causality Mining from Event Sequence Data and Its Applications to Causality Discovery in Earthquakes and Damages

研究代表者

福井　健一（Fukui, Ken-ichi）

大阪大学・産業科学研究所・准教授

研究者番号：８０４１８７７２

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究では，多次元の事象系列データから事象間の発生相関を抽出する新規データマイニングアルゴリズム，クラスタ系列マイニング（Cluster Sequence Mining: CSM）を考案した．さらに，事象間の時間間隔を算出する際の対応関係を1対多もしくは多対1に拡張し，最小コスト弾性マッチング問題として定式化し対応事象対を一意に求める方法を考案した．これによりベイズ推定の精度向上を図った．人工データを用いた評価実験の結果，提案法は従来法に比べて特に時間軸上で事象が密に存在する場合に精度向上が確認された．さらに本手法を，燃料電池の損傷相関分析や地震間の発生相関分析に応用した例を示した．

研究成果の概要（英文）：In this research, we proposed a new data mining algorithm, called Cluster Sequence Mining (CSM), which extracts occurrence correlation between events from multidimensional event series data. Furthermore, we extended the correspondence relation when calculating the time interval between events to one-to-many or many-to-one, by formulating as a minimum cost elastic matching problem, and devised a method to uniquely obtain corresponding event pairs. This aimed at improving the accuracy of Bayesian estimation. As a result of the evaluation experiment using artificial data, accuracy improvement was confirmed in the proposed method, especially when the event densely exists on the time axis as compared with the conventional method. Furthermore, we showed examples of applying this method to damage correlation analysis of fuel cells and occurring correlation analysis between earthquakes.

研究分野：　データマイニング

キーワード：　データマイニング　クラスタリング　頻出パターン　ベイズ推定　弾性マッチング　燃料電池　地震発生相関

１．研究開始当初の背景

近年のセンシングデバイス・通信・記録デバイスの発展に伴い，科学の様々な分野において，現象が電子的に容易に記録され分析できる状況になってきた．大規模な電子データを整理し，その中から計算機により法則性を発見するデータマイニングを基盤とした帰納的方法論はデータ中心科学と呼ばれ，第 4 の科学的方法論として期待される．膨大な観測データに内在する法則性・パターンを抽出することは，現象のメカニズムを探るひとつの有力な手段である．そのような法則の中で，事象間の因果性は現象の基本的なダイナミクスを理解するサイエンスの上でも，工学的な応用の上でも重要である．

これまでデータマイニング研究において，事象の因果性は，発生相関として相関ルール抽出もしくは頻出パターン抽出として研究が行われてきたが，基本的には記号化されたアイテム集合を対象としている．一方，画像，音波をはじめ各種センサの観測量のような実数値で特徴付けられる実数値データにおいて，類似する事象のクラスタリングは数多く提案されているが，両者は独立に研究がなされてきた．数値観測量をクラスタリングにより事象化し，クラスタ ID を記号列として頻出パターン抽出を適用したとしても，事象化において時系列上の発生相関は考慮されていないため，適切なパターンである保証はない．

２．研究の目的

本研究は，事象の類似性と時系列上の発生相関を両者同時に抽出する新たな問題領域を開拓する．両者は相互に依存し合っているため，高精度なパターン抽出には両者を考慮する必要がある．

この問題に対して，申請者が以前提案した共起クラスタマイニング（Co-occurring Cluster Mining: CCM）では，系列上の事象間の共起性の判定において，区間毎の事象の発生順序や時間間隔は一切考慮されていなかった．そのため，得られるパターンは相互作用関係を表すものの，時間的な発生順序や時間間隔は表していなかった．本研究では，従来法 CCM を拡張し，事象の発生順序や時間間隔を取り入れた事象系列データからの知識発見アルゴリズムの創出を目的とする．

３．研究の方法

(1) 事象の発生順序および時間間隔を考慮した発生相関パターン抽出法の開発

① 不確実性への対処

多次元の観測量で特徴付けられる事象からなる系列データに対して，類似事象のクラスタ対の発生順序および時間間隔を取り入れた発生相関パターン抽出法を創出する．その際，計測誤差や個別差などの不確実性を考慮した時間間隔の推定が課題となる．従来法の CCM は，パターンを構成するクラスタ候補の生成と候補パターンの評価から成る．ここで，パターンの評価はクラスタ毎の密集度合い f と，2 つのクラスタに属す事象の時系列上の共起度合い g の積 L=f×g により評価していた．本研究では，時系列上の共起度合い g を改良する．時間間隔がある確率分布に従うと仮定し，ベイズ推定によりパラメータに関する事後分布を推定する．

② 事象間の時間間隔算出における対応付け問題

事象間の時間間隔算出において，事象対の対応付け問題を解く必要がある．これは本質的には解くことができない逆問題であるが，本研究では時間間隔が指数分布に従うと仮定の下で，時間差の総和が最小となる対応関係をコスト最小マッチング問題として定式化し，動的計画法によってある種の解を得ることを考える．

(2)提案法の性能特性

人工的にデータを生成し目標とする正解パターンを埋め込み，データの性質を制御することで，不確実性に対する頑健性などの性能特性を明らかにする．

(3)実問題への応用

以下の実データに提案法を適用し，提案法の汎用性を示す．

①燃料電池の損傷相関分析

燃料電池の運転中に発生する部材の損傷に由来する弾性波の事象系列から，構成部材間の力学的相関性を推定する．

②地震間の発生相関分析

震源地の発生系列データから，異なる地域間の地震の発生相関性を推定する．

４．研究成果

(1)事象の発生順序および時間間隔を考慮した発生相関パターン抽出法の開発

①不確実性への対処

本研究では，数値特徴量からなる事象系列データから時間的に発送相関を持つクラスタ対を抽出する新規アルゴリズム，クラスタ系列マイニング(Cluster Sequence Mining: CSM)を考案した．その際に，ベイズ推定により事象間の発生時間間隔を推定し，不確実性への対処を行った．CSM アルゴリズムの概要を以下に示す．

Step 1: 候補パターンの生成　まず，時間情報は用いずに特徴空間で階層型クラスタリングによってクラスタの候補を生成する．階層型クラスタリングによって得られるクラスタの併合過程の全ての部分クラスタを生成し，包含関係を除いた全てのクラスタペアを候補パターン集合とする（図 1(a)）．

Step 2: 候補パターンの評価　次に，各候補パターンについて評価関数 L により評価値を得る．その評価値が指定された最小評価関数値以上，かつ最小支持度以上の場合，該当候補パターンを出力パターン集合に追加する（図 1(b)）．

Step 3: 類似パターンの除去　最後に，出力パターン集合の中で包含関係にある類似パターンを除去して残ったパターン集合を**クラスタ系列パターン**として出力する（図 1(c)）．

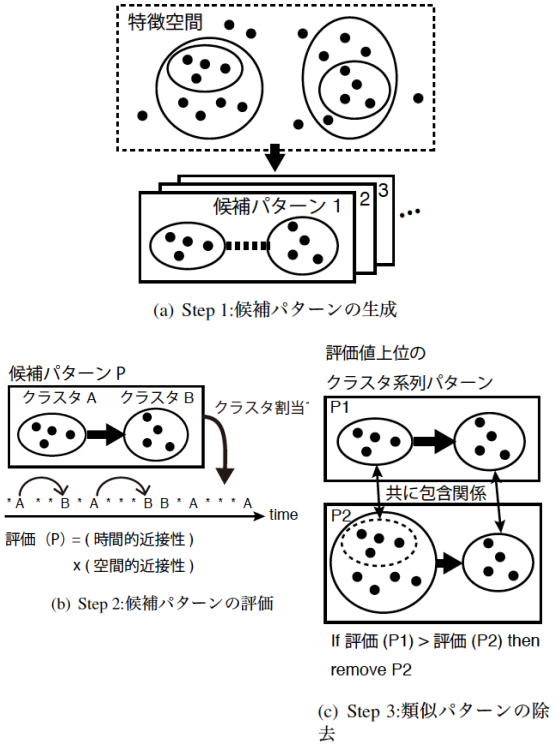


図 1 提案 Cluster Sequence Mining の流れ

②事象間の時間間隔算出における対応付け問題

提案手法では，2 種類のラベルが与えられた事象系列データに対し複数の発生作用をもつ組を対応付ける問題として捉え，以下の 3 つの条件を設定する．

1. 組を作成する時，事前事後のどちらか一方が一つ以上の事象である．
2. 全ての事象をいずれかの組に入れる．
3. 各組の事象間の時間差の総和が最小となるように組にする．

条件 1 及び 2 は組の個数を増やすことでより高い確信度を得るためである．条件 3 は，クラスタ系列マイニングの時間的近接性の要件を満たすためである．両者を満たすことで，時間間隔の推定精度の向上を図っている．

前述した条件に基づき，提案手法では組の作り方をコスト最小弾性マッチング問題として定式化した．

本拡張により，1. ハイパーパラメータの設定なしに，時間差の総和を最小とする組を一意に決定することができ，2. 1 対 1 に限らず複数対 1，もしくは 1 対複数の対応関係を求めることができる（図 2）．

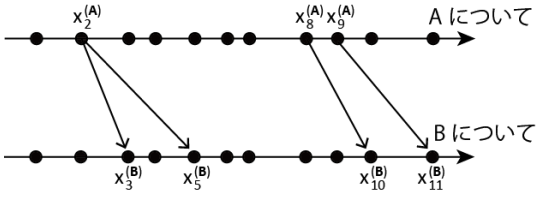


図 2　コスト最小弾性マッチングによる複数対応付け

本手法により得られた事象間の対応から時間差を算出，(1)①の Step 2 における時間的近接性の評価を行い，クラスタ系列パターンを抽出する．

(2) 提案法の性能特性

データ空間上で 2 次元ガウス分布から 2 クラス生成し，2 クラスのデータの時間間隔は指数分布に従って生成した．さらに，ノイズデータはガウス分布の周辺に配置し，時間軸上では一様分布に従って生成した（図 3）．

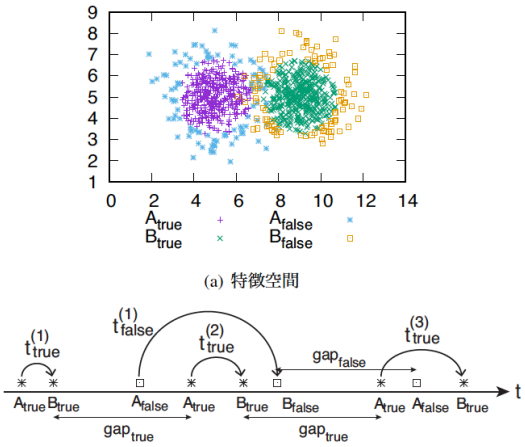


図 3　検証用人工データの生成過程

この人工データに対して，F 値によるクラスタの抽出性能の評価，ならびに推定した指数分布のパラメータにより時間間隔が従う確率密度関数の推定精度を評価した．表 1 より，F 値の観点では，いずれの設定においてもおよそ 0.9 以上の結果となり，目標とするクラスタを漏れも少なく精度良く抽出できていることを示している．一方，指数分布のパラメータ推定について提案法 (CSM) と従来法 (CCM)（文献①）を比較すると，提案法により大きく推定精度が向上することを確認した．しかし，それでも真の値 (λ_true) の半分程度であったため十分であるとはいえない．

表 1 人工データによる CSM の評価結果

λ_{true}	F-measure for CSM Cluster A	F-measure for CSM Cluster B	$\hat{\lambda}^{CSM}_{AB}$	$\hat{\lambda}^{CCM}_{AB}$
0.01	0.897 (0.015)	0.937 (0.018)	0.0048 (0.0032)	0.0017 (0.0007)
0.05	0.900 (0.015)	0.945 (0.016)	0.0312 (0.0186)	0.0097 (0.0040)
0.10	0.892 (0.014)	0.944 (0.013)	0.0526 (0.0369)	0.0172 (0.0074)

表 1 の結果は (1)②の弾性マッチングによる複数事象対の導入はしておらず，単純に最も近い事象同士を対応付けたときの結果であった．弾性マッチングの導入した CSM_DP では，抽出される組の個数が大幅に増えるため Recall（真の対の被覆率）が大幅に改善され，

その結果，ベイズ推定による指数分布のパラメータ推定が大幅に改善することが確認された（表 2）．

表 2 弾性マッチングによる複数対応を許容した CSM_DP との比較

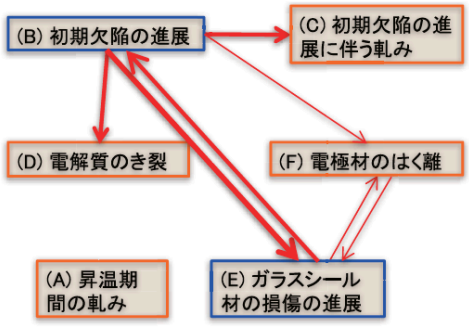
	$\hat{\lambda}$	組の個数	Presicion	Recall
CSM	0.0356 ± 0.0032	255.8	0.8643	0.4600
CSM_DP	**0.0446** ± 0.0038	**359.5**	**0.8684**	**0.6327**

(3) 実問題への応用

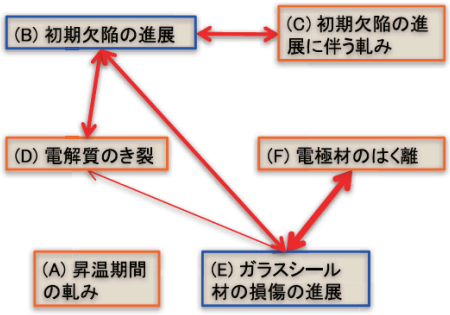
①燃料電池の損傷相関分析の概略

まず，燃料電池の損傷評価データに提案法を適用した例を示す．本研究では，固体酸化物燃料電池を対象として，運転中に生じるき裂やはく離などによって発生する弾性波を計測した Acoustic Emission(AE)事象の系列データから部材間の力学的関係の推定を試みた．

損傷試験は 60 時間分の運転中に得られた AE 事象約 1400 事象の系列データを解析した．前処理は先行研究（文献①）に基づき，カーネル SOM（自己組織化マップ）により AE 事象同士の類似性を可能な限り保存するように構成された 2 次元平面上で CSM を適用した．そして，抽出された全 29 パターンを損傷タイプ間の関係グラフに落とした図を図 4 に示す．



(a) CSM による結果



(b) CCM による結果

図 4 燃料電池の AE 事象系列から得られた損傷相関パターンの比較

CSM による抽出結果（図 4(a)）は，概ね CCM による抽出結果（図 4(b)）と同様の傾向を示しており，「(B) 初期欠陥の進展-(C) 初期欠陥の進展に伴う軋み」間や「(B) 初期欠陥の進展-(D) 電解質のき裂」間で順序を特定できる一方で，「(B) 初期欠陥の進展-(E) ガラスシール材の損傷の進展」間や「(E) ガラスシール材の損傷の進展-(F) 電極材のはく離」間のように双方向の関係性も抽出できている．CSM により，事象間の発生順序を考慮しない従来法との整合性を保ちつつ，新たに発生順序やその時間間隔の推定を可能になっていることが確認された．

②地震間の発生相関分析の概略

次に，2011 年東北地方太平洋沖地震以降の日本周辺の震源リストデータに提案法を適用した結果を示す．本研究では，2011～2015 年に発生した M4.0 以上の地震約 10,000 事象を対象とした．

図 5,6 に CSM により抽出されたクラスタ系列パターンの例を示す．矢印はひとつのクラスタ系列パターンを示しており，複数のパターンで事前もしくは事後クラスタを共有するパターンをまとめて描画している．図中の数字は推定された平均時間間隔（単位：日）を示している．

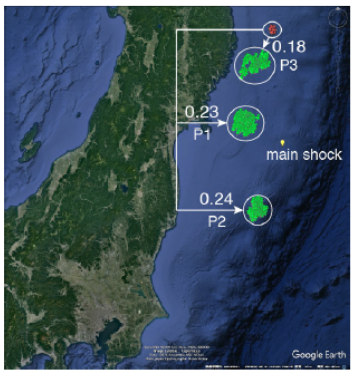


図 5 CSM により抽出された地震発生相関パターン 1（他の地域へ影響を与えやすい領

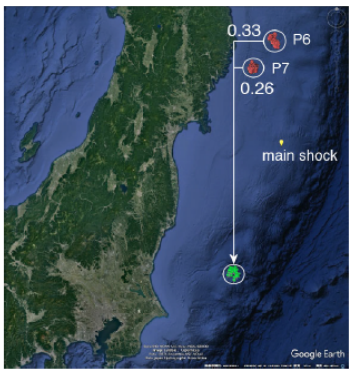


図 6 CSM により抽出された地震発生相関パターン 2（他の地域から影響を受けやすい領域）

図 5 の赤色で示されたクラスタは複数のパターン間で共有された事前クラスタであり，他の地域に影響を与えやすい地域と解釈できる．逆に図 6 のように緑で示された事後クラスタの場合は，他の地域から影響を受けやすい地域と解釈できる．これらのクラスタは CSM により発生順序が特定できるようになったた

め，初めて抽出することに成功した．
一方，海溝型地震のメカニズムは，プレート境界に分布する固着面からなるアスペリティモデルにより説明されており，アスペリティ同士の相互作用や連動性も確認されている．CSM による地震発生相関パターンの分布はアスペリティの分布を表している可能性があり，さらなる調査が必要である．

なお，本成果の一部はデータマイニングにおける主要国際会議にて発表し，その査読付き論文は Springer から出版されている．さらに，弾性マッチングによる事象間対応付けの導入と地震発生相関分析の結果について，現在ジャーナル論文として投稿準備中である．

＜引用文献＞

① 稲場大樹，福井健一，佐藤一永，水崎純一郎，沼尾正行．“燃料電池における損傷パターン抽出のための共起クラスタマイニング”，*人工知能学会論文誌 特集「データマイニングとシミュレーション」*，Vol. 27, No. 3, pp. 121-132, 2012.

５．主な発表論文等

〔雑誌論文〕（計１件）

① Yoshiyuki Okada, Ken-ichi Fukui, Koichi Moriyama, and Masayuki Numao, “Cluster Sequence Mining: Causal Inference with Time and Space Proximity under Uncertainty”, *Lecture Notes in Artificial Intelligence*, Vol. 9078, pp. 293-304, with review, Springer, 2015.
DOI:10.1007/978-3-319-18032-8_23

〔学会発表〕（計１件）

① Yoshiyuki Okada, Ken-ichi Fukui, Koichi Moriyama, and Masayuki Numao.“Cluster Sequence Mining: Causal Inference with Time and Space Proximity under Uncertainty”, *Proc. The 19th Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD2015)*, May 2015.

６．研究組織

(1)研究代表者
福井　健一（Ken-ichi Fukui）
大阪大学・産業科学研究所・准教授
研究者番号：80418772

(2)研究協力者
岡田　佳之（Yoshiyuki Okada）
大阪大学・大学院情報科学研究科・大学院生

佐藤　和輝（Kazuki Sato）
大阪大学・大学院情報科学研究科・大学院生