1M3-OS-24b-3

多次元数値観測量の事象系列に対するクラスタ系列パターンの抽出

Extracting Cluster Sequence Patterns from Numerical Event Sequence

岡田 佳之 *1　福井 健一 *2　沼尾 正行 *2
Yoshiyuki Okada　Ken-ichi Fukui　Masayuki Numao

*1大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

*2大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

In this work, we propose a novel pattern mining algorithm, called Cluster Sequence Mining (CSM), for discovering cluster sequence patterns from multi-dimentional numerical event sequence. In such data, some rules are often hidden as sequence patterns that are strongly related to causes of the events. The CSM can extract such patterns with probability density of time intervals, where a cluster refers to similar events represented by multi-dimensional data. We applied the CSM to synthetic data for validation, and then applied to acoustic emission event sequence from damages of a fuel cell and hypocenter list of eathquakes, and discovered some interaction mechanisms.

1. はじめに

データ空間中で類似の事象集合を求めるクラスタリング [Everitt 11] と，Apriori[Agrawal 94] に代表される記号（アイテム）の系列から頻出するアイテム集合もしくは系列を求める頻出（系列）パターン抽出は，データマイニングにおいて基本的なタスクであるが，両者は別々の文脈で発展してきた．本研究では，両者の特徴を併せ持つ新たなマイニングアルゴリズムとして，クラスタ系列マイニング (CSM：Cluster Sequence Mining)*1 を提案する．

本手法は，波形データや位置情報の様に各事象が多次元の特徴量で表される事象の系列データから，系列上で近接し，かつ頻出して発生している事象集合（クラスタ）の系列をパターン（$\mathbf{A} \rightarrow \mathbf{B}$，ここで $\mathbf{A}, \mathbf{B}$ はクラスタ）として抽出する．そうしたパターンには，対象となる事象系列の発生メカニズムが現れていると考えられ，それらを解明することは機械の故障防止や災害予測等，様々な分野で役立つと期待される．

我々は以前，数値観測量の事象系列に対する共起パターン抽出法として，共起クラスタマイニング (CCM: Co-occurring Cluster Mining)[Inaba 12] を提案した．これは，データ空間上のクラスタペアの候補から，系列上のクラスタ間の共起性と頻出性，およびデータ空間上のクラスタ内の類似性を同時に考慮し，共起パターンとして抽出する手法である．しかし，CCM では抽出された共起パターン (クラスタ $\mathbf{A}, \mathbf{B}$) において，発生の順序や時間間隔は考慮されていない．そこで，本研究ではクラスタ間の順序や時間間隔を加えた**クラスタ系列パターン**を抽出する新たなアルゴリズムを構築した．ここで，候補クラスタペア間の時間間隔は，不確実性を考慮してベイズ推定によりその分布を推定した．

提案手法に対し，まず，人工データを用いてクラスタ系列パターン抽出精度の検証を行った．そして，実応用例として，燃料電池損傷時に得られる破壊音の弾性波系列データ，ならびに 2011 年東日本大震災後の震源系列のデータに本手法を適用した．燃料電池の損傷系列データからは部材間の損傷因果性，また地震系列データからは，異なる地域間の地震因果性に関する推定パターンを得た．

連絡先: 連絡先：大阪大学産業科学研究所沼尾研究室
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1,
E-mail: fukui@ai.sanken.osaka-u.ac.jp

*1　詳しくは [Okada 15] を参照

2. クラスタ系列マイニング

2.1 定義と要件

まず初めに，対象とする事象 (データ) の性質を定義し，その後，本手法が抽出するパターンの要件を述べる．

定義 1（事象系列データ） v 次元から成る数値観測量の事象 N 個：$\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,v}), (i = 1, \cdots, N)$ が，それぞれの観測時刻 $t(\mathbf{x}_i)$ をもって，時間順に $\mathbf{x}_1, \cdots, \mathbf{x}_N$ として得られているとき，$\mathcal{D} = \{\mathbf{x}_i, t(\mathbf{x}_i)\}_{i=1}^{N}$ を事象系列データと呼ぶ．

次に抽出パターンについては，以下の 3 つの要件を満たすこととする．

要件 1（時間的近接性） $\mathbf{x}^{(\mathbf{A})} \in \mathbf{A}$ が発生し，それに対応する $\mathbf{x}^{(\mathbf{B})} \in \mathbf{B}$ の発生に対して，それらの時間差 $t_{\mathrm{AB}} \equiv t(\mathbf{x}^{(\mathbf{B})}) - t(\mathbf{x}^{(\mathbf{A})})$ が小さいこと *2．

要件 2（頻出性） 要件 1 の順序（$\mathbf{A} \rightarrow \mathbf{B}$）で出現する回数が多いこと．

要件 3（空間的類似性） 事象の集合 $\mathbf{A}, \mathbf{B}$ それぞれにおいて，集合内の各事象が類似していること．

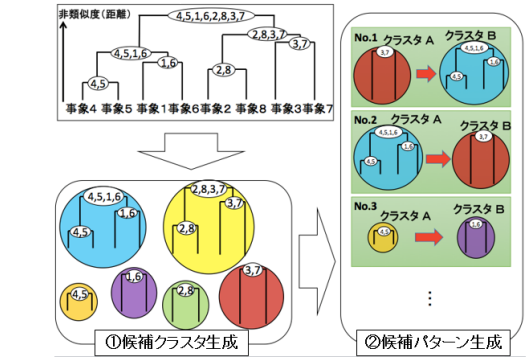
本研究では事象間の時系列上での前後関係や時間間隔を抽出することで，事象間の因果関係を導くことを目的としている．要件 1 と 2 はクラスタ間の時系列上での因果性に関する要件であり，要件 3 はクラスタ内の類似性に関する要件である．

定義 2（クラスタ系列パターン） 要件 1〜3 を満たすクラスタ $\mathbf{A}, \mathbf{B}$ に対し，t_{AB} が特定のパラメータ θ で示される確率分布 $\psi(t_{\mathrm{AB}}|\theta)$ に従う場合，$\mathrm{P}_{(\mathbf{A}\rightarrow\mathbf{B})} = \{\mathbf{A}, \mathbf{B}, \psi(t_{\mathrm{AB}}|\theta)\}$ をクラスタ系列パターン（以降，パターンと表記）と呼ぶ．

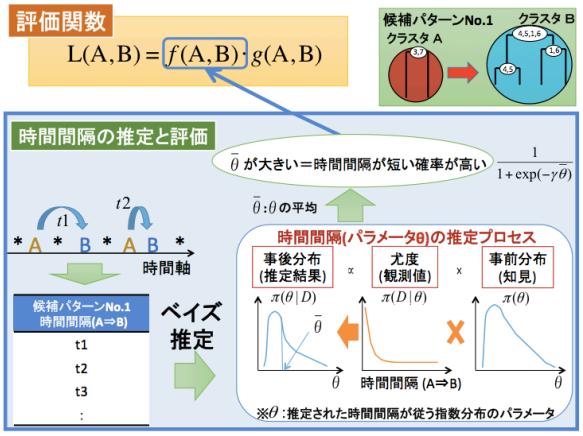
2.2 アルゴリズム

本手法ではまず，クラスタの探索空間の削減のため，階層型クラスタリングを用いる．そして，包含関係を除く全てのクラ

*2　より一般的には，時間間隔が従う分布の分散が小さいほど因果性が強いと考えられるが，本稿では，指数分布に従うと仮定したため，分布の分散が小さいことと，時間間隔が短いことは同義である．



(a) 候補パターンの生成



(b) 候補パターンの評価

図 1: クラスタ系列マイニングのアルゴリズム概念図

スタを順序を区別してペア化し，クラスタ系列パターンの候補クラスタペアを作成する（図 1(a)）．各候補パターンに対して，それぞれの候補内のクラスタ $\mathbf{A}$, $\mathbf{B}$ に属する各事象について，時間的近接性の評価 $f(\mathbf{A},\mathbf{B})$ と，空間的類似性の評価 $g(\mathbf{A},\mathbf{B})$ を合わせた次式に示す評価関数 $\mathcal{L}(\mathbf{A},\mathbf{B})$ により評価値を算出する（図 1(b)）．

$$\mathcal{L}(\mathbf{A},\mathbf{B}) = f^{\alpha}(\mathbf{A},\mathbf{B}) \cdot g^{(1-\alpha)}(\mathbf{A},\mathbf{B}) \tag{1}$$

ここで，$\alpha \in [0,1]$ は時間的近接性と空間的類似性のどちらを重視するかを調整するパラメータである．

評価値が一定の閾値以上であり，かつ一定回数以上の頻出性を満たしたパターン $\mathrm{P}_{(\mathbf{A}\to\mathbf{B})}$ を列挙する．しかし，このままでは類似パターンが大量に抽出されてしまうため，パターン間のクラスタ包含関係をチェックし，包含関係にある場合は評価値が高いパターンを採用し，類似パターンの除去を行うことで多様なクラスタ系列パターンの集合 $\{\mathrm{P}_{(\mathbf{A}\to\mathbf{B})}\}$ を得る．

2.3 時間的近接性の評価関数

$\mathrm{P}_{(\mathbf{A}\to\mathbf{B})}$ の時間間隔が従う確率分布 $\psi(t_{\mathrm{AB}}|\theta)$ のパラメータ推定にはベイズ推定を用いた．これは観測事象の個体差や，事象数の不充分，ノイズの内包といった問題に対して頑健に推定できるためである．具体的には，まず候補パターン $\mathrm{P}_{(\mathbf{A}\to\mathbf{B})}$ において，時間間隔の集合 $\{t_{\mathrm{AB}}\}$ を得る．本研究では単純化のため，全ての $\mathbf{x}^{(\mathbf{A})} \in \mathbf{A}$ に対して，その次に現れる $\mathbf{x}^{(\mathbf{B})} \in \mathbf{B}$ について，それらの時間差 t_{AB} を算出した．

そして，それら時間間隔は指数分布に従うと仮定して，尤度関数は $\psi(t_{\mathrm{AB}}|\theta) = \theta\exp(-\theta t_{\mathrm{AB}})$ とした．ベイズ推定では，次式のベイズの定理に基づいて事前分布 $\pi(\theta)$ と尤度関数 $\psi(t_{\mathrm{AB}}|\theta)$ からパラメータ θ の事後分布 $\pi(\theta|t_{\mathrm{AB}})$ を求める．

$$\pi(\theta|t_{\mathrm{AB}}) \propto \psi(t_{\mathrm{AB}}|\theta) \times \pi(\theta) \tag{2}$$

ここで，事前分布と事後分布には自然共役分布としてガンマ分布 $\mathrm{Ga}(\alpha_{prior},\beta_{prior})$，$\mathrm{Ga}(\alpha_{post},\beta_{post})$ を採用し，以下のベイズの更新式に従って事後分布のパラメータを求める（式中の n は $\{t_{\mathrm{AB}}\}$ の総数，$\bar{t}_{\mathrm{AB}}$ は t_{AB} の平均値をそれぞれ表す）．

$$\alpha_{post} = \alpha_{prior} + n, \qquad \beta_{post} = \beta_{prior} + n\bar{t}_{\mathrm{AB}} \tag{3}$$

時間的近接性の評価 $f(\mathbf{A},\mathbf{B})$ について，事後分布の平均値 $\bar{\theta}_{\mathrm{AB}} = \alpha_{post}/\beta_{post}$ を用いて，指数分布の傾きが急であるほど，すなわち時間間隔が短い確率が高いほど，評価値が高くなるように設定した．なお，値の規格化のためにシグモイド関数を掛けている．

$$f(\mathbf{A},\mathbf{B}) = \frac{1}{1+\exp(-\gamma\bar{\theta}_{\mathbf{AB}})}, \qquad (\gamma > 0) \tag{4}$$

2.4 空間的類似性の評価関数

一方，空間的類似性の評価 $g(\mathbf{A},\mathbf{B})$ は，クラスタ重心からの各事象の平均二乗誤差 $V(\mathbf{A})$，$V(\mathbf{B})$ をガウス関数により [0,1] に規格化して以下のように定めた．

$$g(\mathbf{A},\mathbf{B}) = \exp\left(-\frac{V(\mathbf{A})^2 + V(\mathbf{B})^2}{2\sigma^2}\right), \qquad (\sigma > 0) \tag{5}$$

これはクラスタ $\mathbf{A},\mathbf{B}$ ともにデータ空間上で密になっていることを要請している．

3. 人工データによる評価実験

3.1 人工データの生成

正解のクラスタ系列パターン生成にあたり，まず 2 次元の正規乱数によって 2 つのクラスタ $\mathbf{A},\mathbf{B}$ に対応する事象集合を生成する．次にクラスタ $\mathbf{A},\mathbf{B}$ 内の各事象に対し，その内の任意数を正解事象の集合（$\mathbf{A}_{true},\mathbf{B}_{true}$）としてクラスタ中心から順に選択し，$\mathbf{A}_{true} \to \mathbf{B}_{true}$ の時間間隔が指数乱数に従うように対応付けた．残りの事象集合（$\mathbf{A}_{false},\mathbf{B}_{false}$）はノイズとし，$\mathbf{A}_{false} \to \mathbf{B}_{false}$ の時間間隔が一様分布に従うように対応付けた．

3.2 クラスタ抽出精度に関する評価

各クラスタ 500 点，全 1000 点のデータを生成した．そのうち，ノイズの数を 200〜600 点まで変化させて評価用データセットを複数設定した．それぞれのノイズ数に対して異なる乱数により 30 セット用意した．正解クラスタと抽出クラスタに関する F 値の平均を正解パターンの抽出精度として評価した．なお，評価関数内のパラメータは事前調整に基づき $\gamma = 100$, $\sigma = 1$ とした．

図 2 より，ノイズを半数以上含むような場合でも，提案法 CSM は F 値 0.7 以上と精度良く正解事象を抽出できていることを確認した．さらに，従来法 CCM[*3] を同一のデータに適用した結果も掲載している．CCM では時系列上の事象間の順序を考慮せずにパターン抽出を行うため，クラスタ $\mathbf{A},\mathbf{B}$ は可換である．そのため，抽出されたクラスタペアを入れ替えた場合

*3 CCM では共起性の評価に系列を区間に分割する必要があるが，正解事象の時間間隔が指数分布に従うことから時系列順に 2 事象毎に分割した．

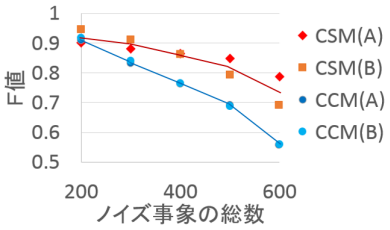


図 2: F 値によるクラスタ系列パターン抽出精度評価

表 1: 抽出精度評価 (F 値) と指数分布のパラメータ推定値 ($\bar{\theta}_{\mathrm{AB}}$)

θ_{true}	F 値	$\bar{\theta}_{\mathrm{AB}}^{\mathrm{CSM}}$	$\bar{\theta}_{\mathrm{AB}}^{\mathrm{CCM}}$
0.01	0.917	0.0048	0.0017
0.05	0.923	0.0312	0.0097
0.10	0.918	0.0526	0.0173

の F 値も計算し，高い方を採用した．ノイズの割合が少ない場合は 2 つの手法の結果に差はほとんど見られないが，ノイズの割合が 6 割を占める場合では F 値のおいて約 0.2 の差が生じている．この結果から，ノイズの割合が増加するに従い，本手法の方が堅調に正解パターンの抽出が行われているといえる．

3.3 時間間隔の分布推定に関する評価

次に，正解事象間の時間間隔が従う指数分布のパラメータ θ_{true} を 0.01，0.05，0.10 と変化させ，それぞれの場合で作成した人工データに対して同様にパターン抽出を試みた．表 1 の各 F 値は，クラスタ $\mathbf{A}, \mathbf{B}$ の平均 F 値のさらに 30 試行分の平均値である．なお，ノイズ事象数は 200 に固定した．

表 1 より，まず θ_{true} の値によらず F 値が約 0.9 と堅調に正解クラスタを抽出できていていることが示された．また，表 1 には CSM，CCM によって抽出されたそれぞれのパターンについて，θ の平均推定値を掲載している．なお，従来法おいて時間間隔の推定は，CCM で抽出されたクラスタペアについて，提案法と同じベイズ推定を適用して求めた．表より，CSM の推定値は，CCM と比較すれば θ_{true} に近い結果ではあるが，θ_{true} の半分程度の推定値であった．

CSM は正解クラスタを高精度に抽出できているにも関わらず，θ の推定の精度が低い結果であった．原因として，現状，t_{AB} の算出において，直近の 1 対 1 の対応関係しか見ていないため，ノイズ事象の混入に対して脆弱であることが考えられる．今後の対策として，弾性 (DP) マッチングによる多対多の柔軟な対応関係から t_{AB} の集合を算出することを検討する．

4. 適用例 1：燃料電池の損傷部材間の因果推定

4.1 背景

固体酸化物燃料電池（SOFC）は高い発電効率を有し期待される一方で，セラミックスで構成されているため，熱膨張や還元膨張により物理的な劣化が確認されている [佐藤 05]．福井らは，図 3 に示すような物理的損傷時に生じる微弱な弾性波事象（Acoustic Emission: AE）に対して，その周波数スペクトルの類似性に基づいて，損傷部材がある程度特定できることを示した [福井 10]．

さらに，稲場らは CCM を用いてこれらの AE 事象系列から，各構成部材間の相互作用を推定した [Inaba 12]．燃料電池の専門家からは，安定的な運転制御や部材の交換時期の推定に部材間の影響関係を把握しておくことが望まれている．そこで，本研究では提案法により，構成部材間の損傷因果性や時間間隔の抽出を試みた．

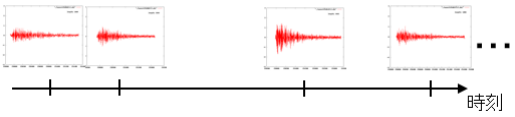


図 3: 燃料電池の損傷によって生じる AE 事象系列

4.2 データと前処理

本研究では，[Inaba 12] と同じ 60 時間分の SOFC 損傷試験において得られた AE 事象系列データを用いた．AE 計測のサンプリングレートは 1MHz である．常時得られる信号から Kleinberg のバースト抽出法 [Kleinberg 02] を適用して抽出した 1429 個の AE 事象を実験対象とした．そして各 AE 事象は周波数パワースペクトルに変換し，スペクトルの離散点を各 AE 事象の特徴とした．

4.3 適用結果

CSM のパラメータは，$\gamma = 2$，$\sigma = 0.5$，$\alpha_{prior} = 0, \beta_{prior} = 0$ とした [*4]．これに加え，最小評価関数値 0.70，最小支持度 10 回としたところ，計 29 パターンを抽出した．CSM で抽出したパターンの可視化・解釈のため，カーネル SOM（Self-Organizing Map）による分類結果 [福井 10] 上に図示した（図 4）．

図中のパターン例 1 は，初期欠陥の進展による損傷波形が生じると約 5 分以内に電解質の亀裂に関する波形が生じる確率が高いということを示している．この部材間については稲場らの結果でも相互関係を指摘されており，本研究により発生順序や時間間隔を明らかにされた．パターン例 2 は構成するクラスタの一方が福井らの分類には当てはまらないものとなった．この結果について専門家と波形の形状や発生状況について議論を重ねることにより，このクラスタは電極材の亀裂に関する AE 事象である可能性が新たに示唆された．他にも，パターン例 3 は両方向のクラスタ系列パターンが抽出されたため，電解質とガラスシール間では相互に影響を及ぼしあっていることが確認された．燃料電池の専門家の評価によれば，これらの結果は妥当であるだけでなく，構成部材間の力学的影響関係を示す興味深い結果であると評価された．

5. 適用例 2：2011 年東日本大震災の余震活動

5.1 背景

地震は他の自然災害と比較しても，特に大きな被害を生じさせる．これまで地震学者を中心として，主に地震の中長期予測について数多くの研究がなされている [尾形 98, Geller 97]．本研究では，地域間の地震発生の順序や時間間隔を考慮することにより，局所的な地殻構造のみならず別の地域との連動性を探ることで，日本を取り巻く地殻全体を系としてマクロに理解することや，誘発地震の可能性を知ることができると考える．

5.2 適用結果

今回用いた震源データは 2011〜2012 年の 2 年間の計 5954 回（M4.0 以上）である．本研究では，M4.0 以上の地震の発生時刻，ならびに数値観測量として緯度と経度を用いて地域間

*4 これらのパラメータは経験的であるが，これらの値付近では安定して同様なパターンが得られることを確認している．

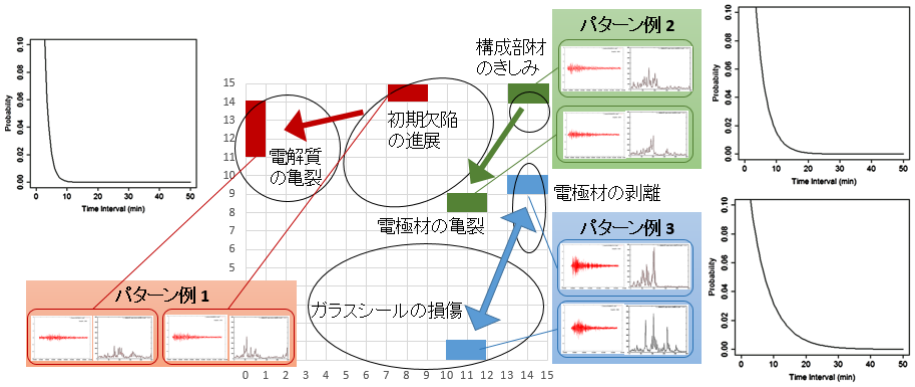
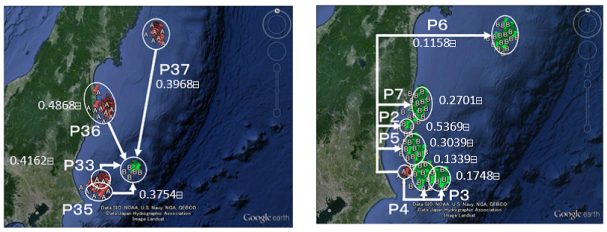


図 4: 抽出された燃料電池の損傷パターンの代表例．中央はカーネル自己組織化マップ（SOM）による分類結果 [福井 10] を示す．

における地震発生の因果性を推定した．CSM のパラメータは，$\gamma = 1$，$\sigma = 0.03$，$\alpha_{prior} = 0$,$\beta_{prior} = 0$ とした．これに加え，最小評価関数値 0.80，最小支持度 8 回としたところ，計 37 パターンを抽出した．

図 5 に今回の結果で得られた特徴的なパターンを示した．図上の白線で囲まれた各領域はパターンを構成するクラスタ，赤は先行する地震の震源，緑はその後で起きた事象を示している．ここで，複数のパターン間で先行クラスタもしくは後部クラスタが共通のパターンを連結し，影響を受けやすい地域（図 5(a)）や影響を与えやすい地域（図 5(b)）を特定した．また，推定された発生時間間隔が従う指数分布の平均値（単位：日）も併記している．地震学ではアスペリティ間の連動性 [Ariyoshi 09] について示唆されているものの，本研究のように網羅的に調べた研究は存在せず，地震発生の新たなメカニズムの発見が期待される．



(a) 影響を受けやすい地域　　(b) 影響を与えやすい地域

図 5: 複数のパターンから示された影響度の高い地域

6. まとめ

本研究では，数値観測量の事象系列から順序や発生間隔を考慮したクラスタ系列パターンの抽出アルゴリズムを提案した．人工データを用いた定量評価では，ノイズ事象の増加に対して頑健に抽出できることを確認したものの，時間間隔のパラメータ推定精度については改善の余地が残されている．燃料電池の損傷部材因果性の推定においては，物理現象として妥当と考えられる結果と共に，これまで発見できなかった損傷タイプを示すクラスタが抽出された．地震の発生因果性の推定においては，本手法により特定された影響度の高い地域はアスペリティとの関連性が示唆されるものの，さらなる裏付けが必要である．

謝辞

本研究は JSPS 科研費 24650068 の助成を受けたものです．適用例 1 では，東北大学工学研究科の佐藤一永准教授にご助言を賜りました．適用例 2 では気象庁の一元化震源リストを使用しました．ここに感謝の意を表します．

参考文献

[Agrawal 94] Agrawal, R. and Srikant, R.: Fast algorithms for mining association rules, in *Proc. of 20th International Conference on Very Large Databases (ICVLD)*, pp. 487–499 (1994)

[Ariyoshi 09] Ariyoshi, K., Hori, T., Ampuero, J.-P., Kaneda, Y., Matsuzawa, T., Hino, R., and Hasegawa, A.: Influence of Interaction between Small Asperities on Various Types of Slow Earthquakes in a 3-D Simulation for a Subduction Plate Boundary, *Gondwana Research*, Vol. 16, pp. 534–544 (2009)

[Everitt 11] Everitt, B. S., Landau, S., Leese, M., and Stahl, D.: *Cluster Analysis, 5th Edition*, Wiley (2011)

[Geller 97] Geller, R. J.: Earthquake prediction: a critical review, *Geophysical Journal International*, Vol. 131, No. 3, pp. 425–450 (1997)

[Inaba 12] Inaba, D., Fukui, K., Sato, K., Mizusaki, J., and Numao, M.: Co-occurring Cluster Mining for Damage Patterns Analysis of a Fuel Cell, in *Proc. the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-12)*, pp. 49–60 (2012)

[Kleinberg 02] Kleinberg, J.: Bursty and hierarchical structure in streams, in *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pp. 91–101 (2002)

[Okada 15] Okada, Y., Fukui, K., Moriyama, K., and Numao, M.: Cluster Sequence Mining: Causal Inference with Time and Space Proximity under Uncertainty, in *Proc. The 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-15)(accepted)* (2015)

[佐藤 05] 佐藤 一永, 橋田 俊之, 八代 圭司, 湯上 浩雄, 川田 達也, 水崎 純一郎：模擬作動環境下における固体酸化物燃料電池の機械的損傷評価法の開発, *Journal of the Ceramic Society of Japan*, Vol. 113, pp. 562–564 (2005)

[尾形 98] 尾形 良彦：ETAS モデルによる地震活動静穏化現象の解析, 地震, Vol. 50, pp. 115–127 (1998)

[福井 10] 福井 健一, 赤崎 省悟, 佐藤 一永, 水崎 純一郎, 森山 甲一, 栗原 聡, 沼尾 正行：固体酸化物燃料電池における損傷過程の可視化, 日本機械学会論文集 A 編, Vol. 76, No. 762, pp. 223–232 (2010)