4M4-1

高次元時系列データにおけるチャンス発見

Chance Discovery in High-Dimensional Time Series

春日瑛 *1　大澤幸生 *1
Akira Kasuga　Yukio Ohsawa

*1東京大学大学院 工学系研究科
Graduate School of Engineering, The University of Tokyo

In recent years, high-dimensional time series data have been accumulated according to development of brand-new device. It is desired to extract important information from the accumulated data and support humans to make their decision. However, the traditional statistical methods where only a few variables are considered are not always applicable when analyzing a dataset of several dozens or more variables. In this paper, we propose a method for detection of structural change based on clustering in high-dimensional time series so as to detect change points in real time and at high speed from the data. In the experiment, the proposed method is applied to the test data, which are 2 types of data including 2 change points and including 2 change points with noise. As a result, the proposed method can detect change point robustly with high precision. In future works, we will apply the method to real business data.

1. はじめに

近年は様々なデバイスの発展によって, 日々大量のデータが蓄積されていく環境となっている. これに伴って, 蓄積したデータから新しい価値を発見し, 意思決定に役に立てることが重要となってきた. 特に, 人間にとって解釈が難しい高次元時系列データから有益な情報を抽出し, 意思決定に用いるための新たな手法が必要とされている. 伝統的な時系列分析では数種類の変数に基づいた分析を行うが, 数十から数百種類の変数を持つ高次元時系列データを分析する際に, 伝統的な統計手法はそのまま適用できないからである [早川 14].

本論文では, 高次元時系列データから高速かつリアルタイムに変化点を検知する手法として, 高次元時系列データにおけるクラスタリング構造変化点検知を提案する. 提案手法は, 変化点を検出するために潜在的な構造として確率分布を仮定する既存の手法と異なり, 顕在データのみを用いた構造的な変化点検出を行っている. これは, 計算機による情報抽出と人間による思考の相互作用によって新たな価値を生み出すこと, 及び抽出された情報から人間がその意味を容易に理解できることを目的としているためである.

以下では, 初めに関連する先行研究を示したのち, 提案手法の具体的なアルゴリズムを示す. 実験においては人工データを用いることで提案手法の検知の精度を検証する.

2. 先行研究

2.1 チャンス発見

チャンスとは意思決定に重要な影響を与える可能性のある事象, 状況, またはそれらについての情報であると大澤によって定義されている [大澤 03]. チャンス発見が従来の統計, パターン認識, データマイニングにおける予測手法と大きく異なるのは3つの特徴による. モデルや変数の発見と生成, 稀な事象への着目, 人間と計算機の協調である. さらに, チャンス発見に基づく価値創造は二重らせんプロセスモデルによって行われる [Ohsawa and Nara 03]. 実証実験では, 顧客購買データからスーパーマーケットの購入金額増加の鍵となる商品を発見し [臼井 03], 繊維会社の生地展示会データに適用すると, それまで注意の払われていなかった生地のニーズを掘り起こすことができた [大澤 02] という結果が得られている. これらの技術によって, データに基づく意思決定支援を行うことがチャンス発見の意義である.

連絡先: 春日瑛, 東京大学大学院　工学系研究科　システム創成学専攻, me7te7or.sai.dsw@gmail.com

2.2 構造変化検知

時系列データにおける構造変化検知において, 例えば平井らによって再正規化最尤符号（RNML）に基づく逐次的な動的モデル選択（DMS）手法が提案されている [Hirai and Yamanishi 12]. これは, 記述長最小原理に基づいて, 逐次的に記述長が最小になるクラスタリング構造変化を抽出するものであり, 符号化の観点から構造的変化を捉えることができる. AIC, BIC といった情報量規準を用いた手法より, 高速かつ正確に構造変化を検出できている. また, 上田らによって Latent Dirichlet Allocation（LDA）を用いた潜在的構造変化検知も提案されている [上田 12]. この手法は, 購買行動を LDA によりモデル化し, その分布の距離を潜在的な嗜好の変化としてスコアリングしたのちに, 動的閾値法によって変化点を検知している. 実験結果によると, 多項分布モデルと比較して, 検出性能が高いことが示されている. これらの変化点検知のアプローチとしては, 確率分布間の距離の比較によって状態変化を定義する手法が一般的である [杉山 13]. しかし, このアプローチには推定という複雑な問題を扱うこととなり, 大規模なデータかつ高度な計算理論モデルが必要となる.

3. クラスタリング構造変化検知

3.1 本手法の構想

本手法は, 意思決定において重要となる変化点を高次元時系列データにおいて検知することを目的としている. 上記の変化点検知アプローチにおいては, 出力に対しての解釈が容易であるとはまだ言えない. 実務における意思決定に活用するには, 出力結果の解釈が容易であるという点が重要視されるため, 我々は人々の意思決定を支援するチャンスを発見するためのツールとして本手法を提案する.

変化の定義は, 非同調点に基づいて行う. 非同調点とは, 周囲と比較して例外的な挙動をする点として定義した点である. 例外的な挙動を示す点は, 一般に外れ値として取り除かれる. しかし, 非同調点が時間的推移に伴って, 周囲の点と同調を示すという現象を示すことがある. このような, 非同調点から同調点への変遷を本論文では変化として定義する. 変化度の定量化は 2 段階のステップによって行う. STEP1 においては, 部分時系列におけるクラスタリングによって非同調点及び同調点を特定する. まず, スライド窓によって入力時系列データを分割した後に, 各部分時系列ごとにクラスタリングを行う. 非同調点は小さなクラスタに含まれる点であり, 構造的に他の点と比較して特異である点と定義する. 同調点は大きなクラスタに含まれる点であり, 構造的に普遍性を持つ点であると定義する. クラスタリングは, クラスタ数の最適化を行うために, Affinity Propagation [Frey and Dueck 07] [Frey and Givoni 09] を用いる. Affinity Propagation はデータ間の類似度のみを用いてクラスタリングを行う手法であり, Availability と Responsibility によってクラスタ数を最適化するアルゴリズムである. STEP2 においては, 最適化されたクラスタに基づいてスコアリングを行う. 各部分時系列内の各点について, 前後の部分時系列と比較して構造的な変化を定量化する. 構造的な変化とは非同調点が同調点へと変遷することであり, この状態変化をスコアリングする（図 1）. このように 2 段階のステップによって, 時系列内の各点の変化度を定量化することができる.

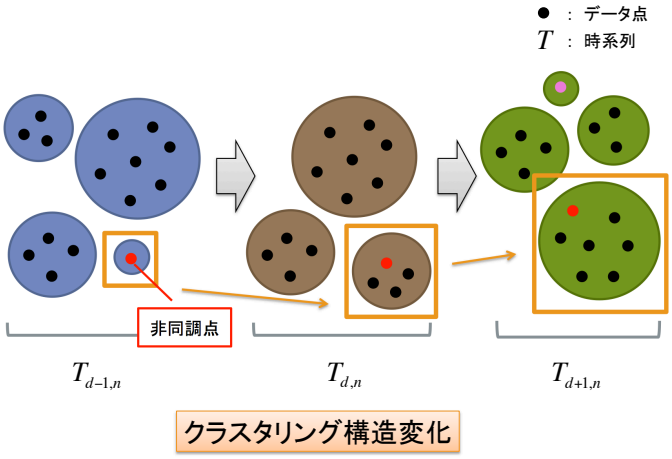

図 1: 提案手法の全体像

3.2 アルゴリズム概要

本節では, 本手法の具体的なアルゴリズムについて述べる. アルゴリズムは Affinity Propagation による部分時系列のクラスタリングとクラスタ構造変化のスコアリングの 2 段階のステップ: STEP1, STEP2 によって構成される. アルゴリズムの概要チャートを図 2 に示す.

3.2.1 STEP1

初めに, 行が時刻を表し, 長さ m の高次元行列 $T = (t_1, t_2, ..., t_m)^{\mathrm{T}}$ があるとする. このとき, 各データ点を $t_l(1 \leq l \leq m)$ とする. 部分時系列の長さを n とすると, 各部分時系列は $T_{d,n}$, where $1 \leq d \leq m-n+1, n \leq m$ と表される. $T_{d,n}$ に含まれる各データ点を $t_{l,d}(d \leq l \leq d+n-1)$ とする.

Affinity Propagation によるクラスタリングでは, Responsibility と Availability と呼ばれる 2 種類のメッセージを互いに

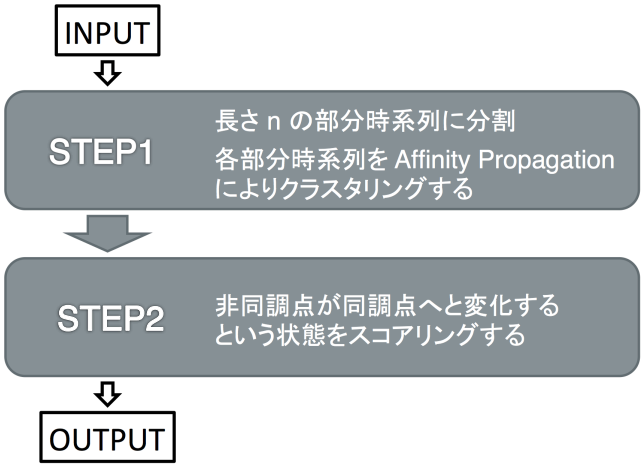

図 2: アルゴリズム概要フローチャート

更新し合って, クラスタの中心となる exemplar を最適化する (図 3). 入力データは各データ点 $t_{l,d}(d \leq l \leq d+n-1)$ 間の類似度 $s(i,j)$, where $i, j \in \{d, d+n-1\}$ であり, 例えば負のユークリッド距離などで定義される. Affinity Propagation は, 類似度に基づいてクラスタを代表とする点である exemplar を一意に定める. 各データ点において Responsibility と Availability を交互に更新し合い, 一定の値に収束するまで繰り返すことによって exemplar を定める. この結果がクラスタリング結果に反映される. Responsibility $r(i,j)$ は, データ点 j がデータ点 i を含むクラスタを代表とする exemplar としてどれほど適切かを示す値である. この値をデータ点 i からデータ点 j へメッセージとして送る. Availability $a(i,j)$ は, データ点 i がデータ点 j によって代表されるクラスタに含まれる適切さを示す値である. データ点 j からデータ点 i へ, この値が同様に送信される. この 2 種類のメッセージである $r(i,j)$ と $a(i,j)$ を以下の更新式 (1)(2)(3) に従って再帰的に計算する. まず初めに $r(i,j)$ と $a(i,j)$ は 0 で初期化された後に, responsibility $r(i,j)$ が計算され, この値を利用して availability $a(i,j)$ が計算される. 以下同様に, responsibility $r(i,j)$, availability $a(i,j)$ の順で繰り返し計算される.

$$
\begin{aligned}
r(i,j) &\leftarrow (1-\lambda)\rho(i,j) + \lambda r(i,j) \\
a(i,j) &\leftarrow (1-\lambda)\alpha(i,j) + \lambda a(i,j)
\end{aligned}
\tag{1}
$$

更新式 (1) において, λ は damping factor と呼ばれるもので, 繰り返し計算における振動を防ぐことを目的としたパラメータである. $\rho(i,j)$ は responsibility の計算における伝搬値であり, $\alpha(i,j)$ は availability の計算における伝搬値である. 仮に $i \neq j$ であるならば, $\rho(i,j)$ と $\alpha(i,j)$ は以下の更新式 (2) に従って計算される.

$$
\begin{aligned}
\rho(i,j) &\leftarrow s(i,j) - \max_{j' s.t. j' \neq j} \left[\alpha(i,j') + s(i,j') \right] \\
\alpha(i,j) &\leftarrow \min \left[0, r(j,j) + \sum_{i' s.t. i' \notin \{i,j\}} \max\left[0, r(i',j)\right] \right]
\end{aligned}
\tag{2}
$$

一方, $i = j$ であるならば, 以下の更新式 (3) に従って計算される.

$$\rho(i,j) \leftarrow s(i,j) - \max_{j's.t.j'\neq j}\left[s(i,j')\right]$$
$$\alpha(i,j) \leftarrow \sum_{i's.t.i'\neq i} \max\left[0, r(i',j)\right] \quad (3)$$

$\rho(i,j)$ と $\alpha(i,j)$ が収束した後, 最適化された exemplar は以下の式 (4) に従って定められる.

$$\arg\max_i \left[r(i,j) + a(i,j) : i,j \in \{d, d+n-1\}\right] \quad (4)$$

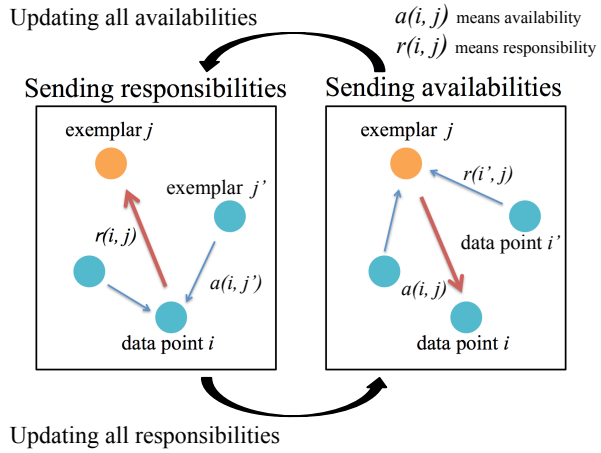

図 3: Affinity Propagation [Frey and Dueck 07]

3.2.2 STEP2

STEP2 では, STEP1 で最適化されたクラスタリング結果に基づいて, 各点の変化の度合いを定量化する. 各点の近傍で前後に属するクラスタの変化の度合いが大きな時に, その時点の変化を捉える. 長さ n の各部分時系列 $T_{d,n}(1{\leq}d{\leq}m-n+1)$ における各データ点は $t_{l,d}(d{\leq}l{\leq}d+n-1)$ である. 各データ点はクラスタ $C_{\omega,d} = \{C_{1,d}, C_{2,d}, \cdots, C_{K,d}\}(1{\leq}\omega{\leq}K)$ のいずれかに含まれるとし, その各集合を $C_{\omega,d}$ で示す. ただし, K は Affinity Propagation によって最適化された exemplar の数である.

次に, ある時点 l が含まれる各部分時系列 $T_{d,n}$ においての変化の度合いを重み $W_{l,d}$ として, データ点 $t_{l,d}(\frac{d+n-1}{2} < l{\leq}d+n-1, l \in \mathcal{N})$ について計算を行う. $\frac{d+n-1}{2} < l{\leq}d+n-1, l \in \mathcal{N}$ という制約を設けるのは, 本手法の変化度合いはクラスタリング結果に基づいて算出するため, $d{\leq}l < \frac{d+n-1}{2}, l \in \mathcal{N}$ において変化の有無に関わらず変化度合いの重みが増加することを防ぐためである. 各データ点 $t_{l,d}$ はクラスタ $C_{\omega,d}$ に含まれることから, $C_{\omega,d}$ の集合において $t_{l,d}$ の前後の要素の数を比較する. つまり, $t_{l,d}$ の前についての要素の数を $N(t_p \in C_{\omega,d} | d{\leq}p < l)$, 後ろの要素の数を $N(t_q \in C_{\omega,d} | l < q{\leq}d+n-1)$ と表すと, 変化度合いの重み $W_{l,d}$ は $W_{l,d} = \left[N(t_q \in C_{\omega,d} | l < q{\leq}d+n-1) - N(t_p \in C_{\omega,d} | d{\leq}p < l)\right]$, where $N(t_q \in C_{\omega,d} | l < q{\leq}d+n-1){\geq}N(t_p \in C_{\omega,d} | d{\leq}p < l)$ と定義する. これを全ての部分時系列 $T_{d,n}(1{\leq}d{\leq}m-n+1)$ に関して同様に計算を行う. ある時点 l の変化の度合い $Y(t_l)$ は, 累積和 $Y(t_l) = \sum_{d=l-n+1}^{l-\frac{n}{2}} W_{l,d}$ として定義する. この $Y(t_l)$ が本手法における変化の度合いであり, これによって定量化を行う.

4. 実験

本手法の有効性を確かめるべく, 人工データを用いて実験を行った. 用いた人工データは 2 通りで, 一方は 2 点の変化点を持つ高次元時系列データ（以下, Data1 とする）であり, 他方は 2 点の変化点を同様に持つがランダムなノイズを付与した高次元時系列データ（以下, Data2 とする）である. 人工データは, Python3.4.3 を用いた擬似乱数によって生成した. Data1 に関して, 平均 μ が 0, 分散 σ が 1 となる正規分布に基づく乱数を 500 回生成させたのちに, 同様に平均 μ が 10, 分散 σ が 1, 及び平均 μ が 20, 分散 σ は 1 となる乱数をそれぞれ 500 回生成させた. これを 100 回繰り返すことによって, 1500 行 100 列の高次元行列 Data1 を生成する. これはつまり, 長さ 1500 の高次元行列 $T = (t_1, t_2, ..., t_{1500})^{\mathrm{T}}$ であり, 各データ点は $t_l(1{\leq}l{\leq}1500)$ で表される. 一方で Data2 に関しては, Data1 と同様の乱数を 50 回繰り返し発生させ, 加えて 0 以上 25 以下となる範囲でランダムに乱数を 1500 回発生させたのちにこれを 50 回繰り返すことによって, Data1 と比較して半数のノイズを付与した 1500 行 100 列の高次元行列 Data2 を生成する. どちらも変化点は, t_{501} 及び t_{1001} であることは明らかである. 生成した人工データのうち, 変化点を含むデータとノイズデータを以下の図 4 に示す. この人工データに対して, 本手法を適用した.

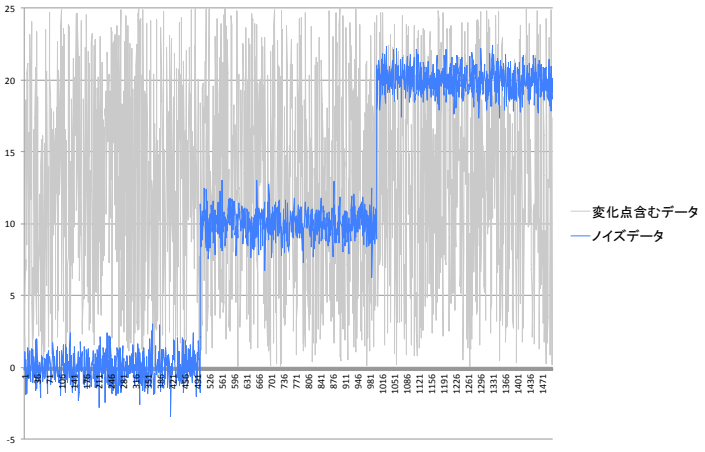

図 4: 人工データの例

4.1 結果

本手法を 2 通りの人工データに適用した結果は以下の図 5 の通りである. 実験はランダムに生成された 2 通りの人工データに対して 5 回ずつ行った. 以下に, その結果を Data1 と Data2 ごとに各試行における $Y(t_l)$ の上位 5 点を表 1,2 にまとめて示す.

表 1: Data1

1 回目	2 回目	3 回目	4 回目	5 回目
1001	501	502	501	501
1002	502	1002	502	502
501	1001	501	1002	503
502	1002	1001	1004	1001
1003	1003	1003	503	504

表 2: Data2

1 回目	2 回目	3 回目	4 回目	5 回目
502	501	501	1001	501
501	502	1001	1002	504
504	1001	1004	502	1003
1001	1002	502	501	502
1002	1003	504	1003	1001

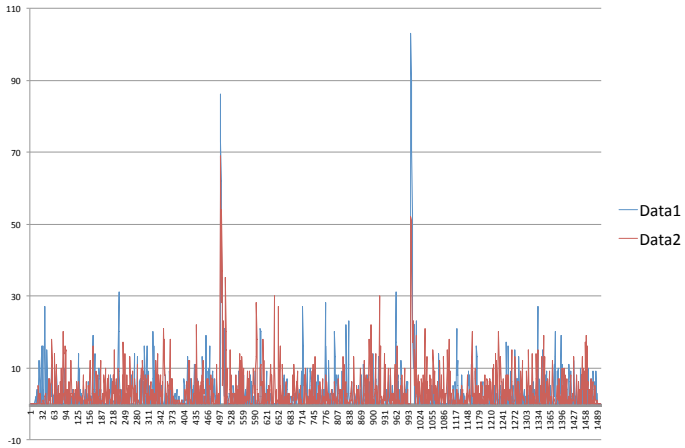

図 5: 実験結果（1 回目）

4.2 考察

本実験において, 抽出した上位 5 点に関する評価を行う. 評価を行う上で, 検知すべき変化点を t_{501},t_{502},t_{1001},t_{1002} の 4 点であるとする. これらの点が各試行の上位 5 点においてどの程度の割合で含まれたかを評価値 E, その平均値を E_{ave} とする. Data1 において, 5 回の試行における評価値の平均は $E_{ave} = 90$ である. 一方, Data2 において, 5 回の試行における評価値の平均は $E_{ave} = 90$ である. 従って, 本手法を人工データに適用することによって, 90%の精度で変化点を検知できたと考えられる. さらに, データの半分にランダムな値によるノイズが付与されていた場合においても精度は変わらず 90%であり, 本手法の頑健性が示唆された. ただし, 今回は評価における精度の指標として t_{501},t_{502},t_{1001},t_{1002} の 4 点のみを代表として用いたが, これ以外に抽出された点は t_{503} や t_{1003} であり, 完全に誤った抽出点ではないことを示しておきたい. つまり, 本手法を高次元時系列データに適用することによって高精度かつ高い頑健性で変化点検出ができることが示された. しかしながら, t_{504} や t_{1004} が上位 5 点で検知されるという t_{501} からやや遅れた点が検知される点に関しては改善の余地がある.

5. 結論及び今後の課題

本論文では, 高次元時系列データから高速かつリアルタイムに変化点を検知する手法として, 高次元時系列データにおけるクラスタリング構造変化点検知を提案した. 本手法は, 計算機による情報抽出と人間による思考の相互作用によって新たな価値を生み出すことを目的としたものであり, 高次元時系列における顕在的データから意思決定に重要な影響を与える可能性のある変化点の抽出を行うアルゴリズムである. 実験では, 2 点の変化点を持つ人工データとノイズを付与した 2 点の変化点を持つ人工データの 2 通りのデータに対して本手法を適用した. この結果, 高い精度かつ高い頑健性による変化点の検出が可能であることが示唆された. 今後の課題としては, 人工データではなく実ビシネスにおけるデータに適用し, 抽出した変化点が実際に意思決定に重要な影響を与えたかどうかを評価する必要がある. 実証実験を通して, 本手法の精緻化を図っていきたいと考えている.

謝辞 本研究は JST CREST の助成を受けたものである.

参考文献

[早川 14] 早川和彦: 高次元時系列データ分析の最近の展開 (<特集> 計量経済学におけるファクター・モデルの諸問題と展望), 日本統計学会誌, 43(2), pp275-292, 2014

[大澤 03] 大澤幸生: チャンス発見の情報技術, 東京電気大学出版局, 2003

[Ohsawa and Nara 03] Ohsawa, Y. and Nara, Y.: Decision process modeling across internet and real world by double helical model of chance discovery, New Generation Computing, 21(2), pp109-121, 2003

[臼井 03] 臼井優樹, 大澤幸生: 生地メーカーにおける暗黙的顧客ニーズの発見, ファジィ学会論文誌, 15(3), 2003

[大澤 02] 大澤幸生, 臼井優樹, 福田寿, 松尾豊, 松村真宏, 高山美和, 相馬浩隆, 佐橋官: 二重螺旋モデルを用いたスーパーの顧客行動変化の予兆発見, 情報処理学会第 128 回知能と複雑系研究会 人工知能学会第 56 回知識ベースシステム研究会 (SIG-KBS) 合同研究会, 2002

[Hirai and Yamanishi 12] Hirai, S. and Yamanishi, K.: Detecting changes of clustering structures using normalized maximum likelihood coding, ACM, pp343-351, 2012

[上田 12] 上田真士, 冨岡亮太, 山西健司, 石黒勝彦, 澤田宏, 上田修功: Latent Dirichlet Allocation を用いた潜在的構造変化検知, 電子情報通信学会技術研究報告, 111(480), pp15-20, 2012

[杉山 13] 杉山将: 確率分布間の距離推定 : 機械学習分野における最新動向, 日本応用数理学会論文誌, 23(3), pp439-452, 2013

[Frey and Dueck 07] Frey, B.J. and Dueck, D.: Clustering by passing messages between data points, Science, 315(5814), pp972-976, 2007

[Frey and Givoni 09] Frey, B.J. and Givoni, I.E.: A Binary Variable Model for Affinity Propagation, Neural computation, 21(6), pp1589-1600, 2009