大規模時系列テンソルによる多角的イベント予測

本田 崇人、松原 靖子、川畑 光希、櫻井 保志
大阪大学 産業科学研究所　産業科学 AI センター

1. はじめに

近年、製造業において製造工場のスマート化が推し進められている。大量のセンサを使用して生産ラインの稼働状況をつねに監視し、その様子を時系列データとして蓄積、分析することにより、機器の異常検知 [1]、[2] や品質管理 [3] 等、あらゆる側面から生産性を向上する取り組みが行われている。これらの取り組みに共通する重要な課題は、収集した大規模データからの効果的な知見獲得と、それに基づく将来予測技術の開発である。

特に、製造工場から得られる時系列データは複数のドメイン（設備、センサ、時間等）を持つ複雑なデータであり、複数の作業工程（パターン）の時間遷移や複数ラインでの並列作業によって生まれる作業ラインごとに共通/相違なパターンなど多角的な特徴を持つことが多い。

加えて、スマート工場で想定されるタスクでは、故障や不具合、加工精度の低下等、各イベントの発生を事前に把握することで、対策の選択肢が広がる。つまり、大規模センサデータの将来予測技術は、より長期的な予測能力を有することが望ましい [4]。

本論文では、大規模時系列センサデータのためのイベント予測手法である SplitCast について述べる。SplitCast は、時系列データに含まれる典型的なパターン（本研究では、“レジーム”と呼ぶ）の数と変化点を多角的にとらえ、システムの稼働状況を正確に把握することで、将来発生するイベントを予測する。

2. 問題定義

本研究で扱う工場設備センサデータは、(facility、sensor、time)の三つ組で表現され、それぞれ、w 個の設備、d 種のセンサ、n の期間（5 秒単位）からなる。このセンサデータは、3 階のテンソル $X \in \mathbb{R}^{w\times d\times n}$ として表現することができ Xの要素 $x_{ij}(t)$ は時刻 t における i 番目の設備の j 番目のセンサの計測値を示す。本論文では、このセンサデータを多次元時系列テンソルと呼ぶ。本研究の最終目的は、与えられた時系列テンソル X から l_s ステップ先の設備アラートを予測することであり、取り組む問題を以下のように定義する。

問題 1：時系列テンソル $X(t_s{:}\,t_e)$ が与えられたとき、l_s ステップ先のアラートラベル $Y(t_e+l_s)$ を次式に基づいて予測する。

$$Y(t_e+l_s) \approx F\big(X(t_s{:}\,t_e)\big) \qquad (1)$$

ここで、$(t_s{:}\,t_e)$ は予測に使用するシーケンスのウインドウを表し、F を提案モデルとする。

3. 提案モデル

提案モデルは次の 3 つの能力を有する。各章において、それぞれの概要と必要な定義について述べる。

3.1 （P1）潜在的な動的パターンの多角的な検出

多次元時系列テンソル X が与えられたとき、提案手法はまず、X を m 個のセグメント集合 $S=\{s_1,\dots,s_m\}$ に分割してその特徴をとらえる。S_iは i 番目のセグメントの開始点 t_s、終了点 t_e、設備番号で構成され（つまり、$s_i=\{t_s,t_e,facilityID\}$）、各セグメントは重複がないものとする。そして、発見したセグメント集合を類似セグメントのグループに分類する。本論文ではこれらのグループをレジームと呼ぶ。

定義 1（レジーム）r を最適なセグメントグループの個数とする。それぞれのセグメント s はセグメントグループの 1 つに割り当てられる。

さらに、各セグメントが所属するレジームを表現するため、新たにセグメントメンバシップを定義する。

定義 2（セグメントメンバシップ）多次元時系列テンソル X が与えられたとき、$F=\{f_1,\dots,f_m\}$ を、m 個の整数列とし、f_i を i 番目のセグメントが所属する

レジームの番号とする ($1 \le f_i \le m$)。

これにより、多次元時系列テンソルを m 個のセグメントと r 個のレジームで $\{m, r, S, \Theta, F\}$ として表現することができる。次に提案手法は、得られたレジーム情報に基づき、多次元時系列テンソルを統計モデル化し重要な特徴を抽出する。

3.2 (P2) 動的パターンに基づく特徴抽出

それぞれのレジームは統計モデル $\Theta = \{\theta_1, \dots, \theta_r, \Delta_{r\times r}\}$ として表現される。本研究では、多次元時系列テンソルの振舞いを表現するため、隠れマルコフモデル(HMM : Hidden Markov Model)を用いる。HMM は隠れ状態を持つマルコフ過程を仮定した確率モデルの 1 種であり、音声認識を含む様々な分野において、時系列処理手法として広く利用されている。HMM は初期確率 $\pi = \{\pi_i\}_{i=1}^{k}$
遷移確率 $A = \{a_{ij}\}_{i,j=1}^{k}$、出力確率 $B = \{b_i(x)\}_{i=1}^{k}$ 三つ組で表現される (すなわち、$\theta = \{\pi, A, B\}$)。ここで、k は HMM の潜在状態数を示す。本論文では出力確率 B が多次元ガウス分布から生成されるものとする。HMM のモデルパラメータ $\theta = \{\pi, A, B\}$ と、入力データとしてあるユーザのシーケンス X が与えられたとき、X の尤度 $P(X|\theta)$ は次のように計算される：

$$P(X|\theta) = \max_{1\le i \le k}\{p_i(n)\}$$

$$p_i(t)\begin{cases} \pi \cdot b_i(x_1) & (t=1) \\ \max\limits_{1\le i\le k}\{p_j(t-1)\cdot a_{ji}\}\cdot b_i(x_t) & (2 \le t \le n) \end{cases}$$

ここで、$p_i(t)$ は時刻 tにおける潜在状態 i の最大確率を示し、n は X のシーケンス長である。この尤度は動的計画法の 1 種であるビタビアルゴリズム [5] を用いて計算される。ここでさらに、新たな概念としてレジーム遷移行列 $\Delta_{r\times r}$ を導入する。

定義 3 (レジーム遷移行列) $\Delta_{r\times r}$ を r 個のレジーム群の遷移行列と呼ぶ。ここで、要素 $\delta_{ij} \in \Delta$ は i 番目のレジームから j 番目のレジームへの遷移確率を示す。すなわち、$0 \le \delta_{ij} \le 1,\ \Sigma_j \delta_{ij} = 1$ という条件を持つ。

上記のモデルを用いて、多次元時系列テンソル X を、以下に示す HMM の潜在状態系列 Z とモデル化した際の誤差 E で要約し、特徴量化することで、高精度かつ長期的な予測を実現する。

定義 4 (潜在状態テンソル) 各設備ごとの HMM の潜在状態系列 $Z = \{Z_1, \dots, Z_w\}$ を潜在状態テンソルと呼ぶ。ここで、$Z_i = \{z_{ij}(1), \dots, z_{ij}(n)\}_{j=1}^{d}$ であり、$z_{ij}(t)$ は自身と同じ潜在状態に属するデータ集合 xの平均と分散の組 $\{\mu, \sigma\}$ で構成される。

定義 5 (誤差テンソル) 多次元時系列テンソル X を潜在状態テンソル Z でモデル化した際の誤差 $E = \{E_1, \dots, E_w\}$ を誤差テンソルと呼ぶ。

3.3 (P3) ls ステップ先の長期予測

上述の定義を用いて、式(1) は、

$$Y(t_e + l_s) \approx F(\{Z(t_e{:}\,t_e), E(t_s{:}\,t_e)\}) \qquad (2)$$

と表現できる。ここで F は予測モデルを表す。すなわち、時系列テンソル X が与えられたとき、提案手法は X を潜在状態テンソル Z と誤差テンソル E で要約することで重要な特徴を抽出し、それらに提案モデル F を適用することで l_s ステップ先の長期的な予測を行う。

4. アルゴリズム

問題 1 を解決するためのアルゴリズムについて説明を行う。ここで問題となるのは、どのようにレジームやセグメントの数を決定するかである。提案手法は、最小記述長(MDL : Minimum Description Length)の概念に基づき、適切なモデルを生成するための基準となる符号化スキームを導入する。

$$Cost_T(X; M) = Cost_M(M) + \alpha \cdot Cost_C(X|M) \qquad (3)$$

ここで、$Cost_M(M)$ はモデル M を表現するためのモデルコストを示し、$Cost_C(X|M)$ は M が与えられたときのテンソル X の符号化コストを示す。α は符号化コストに対する重みであり、α の値が大きいほどより実データに正確なモデルを生成する (すなわち、セグメントの数 m、レジームの数 r が大きくなる)。

提案手法である SplitCast は、次のアルゴリズムで構成される。

・RegimeGenration (P1) : テンソル X に含まれる時系列パターンの種類と変化点を検出する。各時系列パターンのダイナミクスをモデルパラメータ Θ として表現し、モデルパラメータ集合 $\{m, r, S, \Theta, F\}$ を

得る。

・FeatureExtraction (P2)：時系列パターンの要約情報 $\{m, r, S, \Theta, F\}$ を用いて、オリジナルテンソル X を潜在状態テンソル Z と誤差テンソル E で表現する。

・SplitCast (P3)：$\{Z, E\}$ のうち、あるウィンドウ $t_s{:}\,t_e$ の部分テンソル $\{Z(t_s{:}\,t_e), E(t_s{:}\,t_e)\}$ から故障の予兆となる特徴を抽出し、l_s 先の故障ラベル $Y(t_s{:}\,t_e)$ を予測する。

4.1 RegimeGeneration (P1)

本研究では、与えられた時系列テンソルの根底にある構造を簡潔に要約した、多角的なパターン発見とグループ化を同時に行う。ここで、時系列テンソルの多角的解析のためのアルゴリズムである V-Split と H-Split を提案する。V-Split は時間方向の観点からレジームを推定し、H-Split は設備ごとの特性をレジームとして表現する。これら 2 つのアルゴリズムを任意方向に行うことで効率的かつ効果的に重要なパターンを多角的に発見しレジームとして要約する。具体的には、式(3)に基づき、以下の 2 つのアルゴリズムを繰り返す。

・V-Split：テンソル X から時間遷移するパターンとその変化点を検出し 2 つのグループ（すなわちレジーム）に分割する。それら 2 つのレジームに対し、モデルパラメータ $\{\theta_1, \theta_2, \Delta\}$ を推定する。

・H-Split：テンソル X に表れるある 1 つのレジームから設備ごとの特徴を抽出し、2 つのレジームに分割後、それらのレジームのモデルパラメータを推定する。

上記のアルゴリズムにより、$r = 1,2,\ldots$ とレジーム数が変化していく。もしレジーム θ_0 を 2 つのレジーム $\{\theta_1, \theta_2\}$ に分割した際、コスト関数（式 (3)）の値が大きくなれば θ_0 は最適と見なし、これ以上分割しない。生成されたすべてのレジームについて同様にコスト計算を繰り返し、コストが下がらなくなるまで上記の分割アルゴリズムを繰り返す。最終的に、コストが収束したときのセグメント、レジーム、モデルパラメータ $\{m, r, S, \Theta, F\}$ を出力し、RegimeGeneration を終了する。

4.2 FeatureExtraction (P2)

時系列テンソル X とモデルパラメータ集合 $\{m, r, S, \Theta, F\}$ が与えられたとき、X を時系列パターンに基づく潜在状態テンソル Z とモデル化した際の誤差テンソル E に分割する。今、r 個のレジーム集合 $\Theta = \{\theta_1, \ldots, \theta_r\}$ が与えられたとすると、各時刻 t における、設備 i のデータ $x_i(t) = \{x_{ij}(t)\}_{j=1}^{d}$ は Θ 内のレジームのいずれかの状態 $z_i(t)$ に変換される。ここで、$z_i(t)$ は自身と同じ状態に属する全データポイントの平均と分散の組 $\{\mu, \sigma\}$ を示す。つまり、潜在状態テンソルの次元は $Z \in \mathbb{R}^{w \times 2d \times n}$ となる。続いて、Θ が与えられたときの、時刻 t における設備 i のセンサ j の計測値 $x_{ij}(t) \in X$ の符号化誤差を事後確率 $p(x_{ij}(t)|\theta)$ で表現する。すなわち、X 全体の符号化誤差は $E \in \mathbb{R}^{w \times d \times n}$ である。最終的に、2 つの特徴を結合した系列 $X' = \mathbb{R}^{w \times 3d \times n}$ を出力する。以上の処理により、入力データの情報を失うことなく、学習モデル推定の際に時系列方向の潜在的な振舞いを考慮することができる。

4.3 SplitCast (P3)

確率モデルに基づく特徴抽出手法と深層学習手法を組み合わせ、実データから抽出された特徴的な時系列パターンを学習することで、より小さなネットワークで学習でき、過学習の問題を軽減しながら効率的かつ効果的なラベル予測を実現する。具体的には、テンソル $X' = \{Z, E\}$ の時間発展の様子をモデル化し、特徴ベクトル h_t を得るために、Long-short term memory（LSTM）[6] を適用する。また、本研究では、時刻 t における最新の部分シーケンスからの l_s 先故障予測を 2 クラス分類タスクとして扱い、出力を時刻 $t + l_s$ における故障発生確率とする。したがって、SplitCast の最終的な出力は、

$$y_{t+l_s} = sigmoid(W^{y^h} h_t + b^y)$$

となる。

5. 評価実験

5.1 提案手法の有効性

図 1 は、スマート工場データにおける SplitCast の出力結果を示している。図 1 (a)はオリジナルのセン

サデータを示しており、5 つの設備における 3 つのセンサ値（回転速度、稼働電圧、設備温度）で構成される。黒矩形で塗られた箇所は対応する設備が非常停止中であることを示す。図 1 (b)は SplitCast によるオリジナルデータからのパターン検出結果である。図中の縦線は時系列パターンが変化した時刻を示し、同一レジームに属するセグメントは同一色で塗られている。提案手法は、複数の設備から得られた時系列データを同時に解析することにより、多角的なパターン、すなわち、各設備内のパターンの時間遷移だけでなく、設備間で共通、あるいは相違なパターンを検出することが可能である。図 1 (c) に、オリジナルデータの中から、l_s= 200 ステップ（約 17 分）後に非常停止した場合とそうでない場合の典型的な例を図示した。各図の左はセグメンテーション結果を示す。右の $\theta_1 \sim \theta_5$ はそれぞれ共通の時系列パターン（すなわちレジーム）を表し、それらの遷移の様子を可視化したものである。p_{200} の値は、左図にある部分シーケンスとそのパターン検出結果が与えられたとき、提案手法が出力した 200 ステップ先での非常停止確率である。右図において、より多くの遷移が検出されたレジーム間には太い矢印が表示される。また、円の大きさはレジームの発生期間の大きさを示す。図 1 (c-ii) を見ると、設備が非常停止する前に回転速度が上昇 (θ_5) しており、その傾向はレジーム θ_4、θ_5の遷移が現れることによって表現されている。実際に、提案手法は非常停止を正確に予測し、p_{200} が高い値を示している。つまり、データに含まれる潜在的なパターンを検出することで、非常停止に至る過程を多角的に分析できるだけでなく、それらの要約情報を用いることで長期的かつ高精度な予測が可能となる。

5.2 提案手法の予測精度

次に、与えられた時系列テンソルに対する提案手法の故障予測精度について検証する。比較手法には、一般的な 2 値予測モデルであるロジスティック回帰(LR : Logisticregression)[7]と再帰型ニューラルネットワークモデルである RNN(Recurrent neural network)、GRU(Gated recurrent unit)、LSTM と比較した。評価指標には Accuracy を使用し、5 分割交差検証を行った際の平均値を比較する。使用したデータセットは、三菱重工エンジン&ターボチャージャ株式会社で 2017 年 10 月から 3 カ月間実際に稼働し、ベアリング・ハウジング加工を行っていた 55 の工場設備に取り付けられた、回転速度(Speed)、稼働電圧(Load)、設備温度(Temp)の 3 つのセンサによって 5 秒間隔で取得されたものである。スライディングウインドウで学習用サンプルを生成しており、設備自体が稼働していないときのサンプルは省いている。正常稼働時のサンプル数が 62、983、非常停止前のサンプル数が 1069 あり、学習に偏りが生じるため、非常停止時のサンプル数に正常稼働時のサンプル数を揃え、結果として 1069× 2 サンプルを用い実験を行った。

図 2 は予測先のステップ数 ls を変化させたときの精度の比較である。比較手法は、ランダムに予測した場合と同程度(Accuracy = 0.5)の予測精度を示す一方で、SplitCast はいずれの条件下でも優れた性能を示している。この結果から、非常停止の要因は温度の上昇や稼働電圧の低下といった単純なものではなく、非線形性を有する複雑な事象だと考えられる。提案手法は実データに含まれる時系列パターンを考慮して各時刻のダイナミクスをとらえることができるため、他の再帰型モデルと比べて効果的に非常停止の要因を抽出することに成功した。図 3 は、ネットワーク学習時に使用するミニバッチのウインドウ幅を変化させたときの予測精度の比較である。提案手法は、異なるウインドウ幅のデータに対しても安定して高い性能を示している。

6. おわりに

本研究では大規模 IoT センサデータのための予測アルゴリズムとして SplitCast を提案した。SplitCast は、与えられた時系列テンソルに含まれる特徴的なパターンやその変化点を多面的に抽出し、レジームとして要約することで、長期的な故障予測を実現する工場設備で得られた実データを用いて実

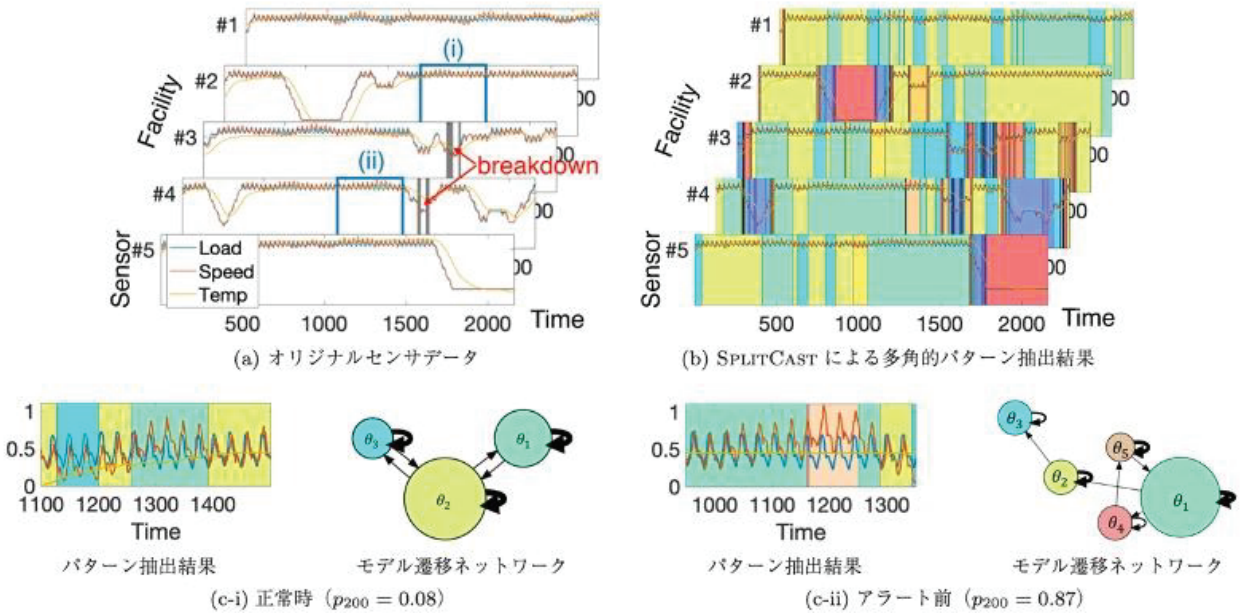

図 1　工場設備センサデータに対する SplitCast の出力結果

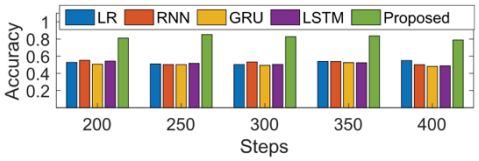

図 2　予測先のステップ数ごとの予測精度

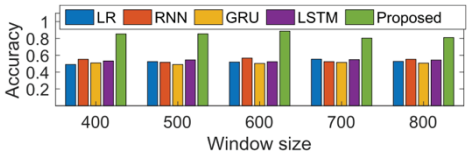

図 3　入力ウインドウサイズごとの予測精度

験を行い、SplitCast が複雑な時系列パターンを適切にモデル化し、長期的な故障予測を高精度に行うことを確認した。また、既存手法と比較して大幅な精度と性能の向上を達成していることを示した。提案手法から得られる、高精度の異常予測とその要因分析によって、工場稼働率の向上や設備保全などの実社会での応用が期待できる。

参考文献

[1] Sakurai, Y., Papadimitriou, S. and Faloutsos, C.: Braid: Stream mining through group lag correlations, SIGMOD, pp.599–610 (2005).

[2] Zhou, Y., Zou, H., Arghandeh, R., Gu, W. and Spanos, C.J.: Non-parametric outliers detection in multiple time series A case study: Power grid data analysis, AAAI (2018).

[3] Li, Y., Wang, J., Ye, J. and Reddy, C.K.: A multitask learning formulation for survival analysis, KDD, pp.1715–1724 (2016).

[4] Matsubara, Y. and Sakurai, Y.: Regime shifts in streams: Real-time forecasting of co-evolving time sequences, KDD (2016).

[5] Forney, G.D.: The viterbi algorithm, Proc. IEEE, pp.268–278 (1973).

[6] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., Vol.9, No.8, pp.1735–1780 (Nov. 1997).

[7] Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics), Springer (2006).