F-021

ガウス混合モデルを用いた時系列データのパターン抽出
Pattern Extraction of Time Series Data Using Gaussian Mixture Model

大城 海斗† 岡﨑 威生‡
Kaito Oshiro Takeo Okazaki

1. はじめに

クラスタリングは事前知識なしにデータを互いに似た性質をもつクラスタにまとめる教師なし学習手法であり、観測対象のメカニズムやそれらに共通するパターンの解明など様々な目的で用いられる。クラスタリングに関連した提案手法として、時系列データの次元圧縮やクラスタ数決定方法などが提案されている。一方で、プロトタイプをクラスタの代表的なパターンとみなした時、パターン抽出といった観点では既存のクラスタ数決定手法では必ずしもユニークなパターンを抽出できるとは限らない。

本研究では、ガウス混合モデルを用いて時系列データからいくつかの代表パターンの抽出を行い、その中でユニークな代表パターンが得られるクラスタ数を提示する手法の検討と評価を行う。

2. 時系列データのパターン抽出

2.1 離散ウェーブレット変換による時系列データの特徴抽出

Kim ら[1]は遺伝子発現データからパターンを発見しやすくするために、データの離散ウェーブレット変換（DWT）を行いノイズ除去及び次元圧縮を行っている。DWT によりレベルrのときの近似係数c^{-r}と詳細係数d^{-r}が得られる。このうち近似係数は時系列データの全体的な変動を表し、入力データの次元数よりも低い次元空間における近似データとして用いることにより、ノイズ除去と次元圧縮を可能とする。ここで、より元データを表現した近似データを得るためにレベルrの選択が課題となってくる。Zhang ら[2]は適切なレベルを決定するために元データと近似データとのエネルギー差に基づいたアルゴリズムを提案し、その有効性を示した。

2.2 ガウス混合モデルを用いたクラスタリング

Kim らはガウス混合モデルを用いて遺伝子発現データのクラスタリングを行い、データがどのクラスタに所属するかを確率により決定した。これにより、確率があるしきい値以下であればどのクラスタにも分類せず、獲得されるパターンへの外れ値による影響を少なくすることが可能となるため、本研究で用いることとした。

2.3 ユニークなプロトタイプが得られるクラスタ数の決定

ガウス混合モデルを用いたクラスタリングにおいてもk-means などのクラスタリングアルゴリズムと同様に、事前にクラスタ数を与えなければならない。AIC や BIC が最小となるクラスタ数を最適なクラスタ数として用いる場合があるが、ユニークなパターンの抽出を目的とした場合、統計モデルの良さを測る AIC や BIC が必ずしも最適なクラスタ数を提示できるとは限らない。本研究では AIC をベースにプロトタイプ間の距離の最小値を考慮したクラスタ数決定方法を提案し、AIC のみを考慮した指標よりもユニークなプロトタイプが得られるかを検証する。ここで、Pはプロトタイプ集合、距離関数distはユークリッド距離とした。

$$\mathrm{IC}_{|P|} = \frac{\mathrm{AIC}_{|P|}}{D} \tag{1}$$

$$D = \min_{\mathbf{p}_i \in P \setminus \mathbf{p}_j} \mathrm{dist}(\mathbf{p}_i, \mathbf{p}_j)$$

3. 検証実験

3.1 使用データ

実験に用いるデータは Phenotype Microarray（PM）データ[3]と小売店からの受注履歴のデータである。PM データは異なる 4 種の菌ごとに 96 個の代謝反応を数十時間に渡り観測した記録で、抽出されるパターン数は既知である。受注履歴データには 442 店舗、40 商品についての 1 年分の受注履歴が収められており、パターン数は未知である。いずれのデータセットも時系列データの長さは等しく欠損値は含まれていない。

3.2 PM データへの適用

パターン数が既知である PM データへ提案手法を適用し、AIC のみのクラスタ数決定方法と比較して既知のパターン数に近いクラスタ数と重複のないプロトタイプが得られるかを検証する。

AIC のみのクラスタ数決定方法において AIC 値が最小となるのはクラスタ数が 12 の時で、そのクラスタリング結果から得られるプロトタイプは図 1 の◇型の折れ線のように乱立した。それに伴い、プロトタイプと差がないにもかかわらず、直線で表されている時系列データの所属確率が散漫になり、いずれのクラスタにも属さないクラスタリング結果となった。一方で、プロトタイプ間の距離を考慮した $\mathrm{IC}_{|P|}$ の推移は図 2 のようになる。最小値をとるクラスタ数は 3 であり、このときのクラスタリング結果では図 3 中の黒線で示すようなユニークなプロトタイプが得られ、すべてのデータがいずれかのクラスタに属した。既知のパターン数であるクラスタ数 4 で最小値を取らなかった理由として、クラスタリング結果で得られた第 4 のプロトタイプが、図 3 のクラスタ V3 のプロトタイプと類似したものであったため $\mathrm{IC}_{|P|}$ 値が上昇したものと考えられる。

† 琉球大学大学院理工学研究科情報工学専攻, Graduate School of Engineering and Science, University of the Ryukyus
‡ 琉球大学工学部工学科知能情報コース, Computer Science and Intelligent Systems, University of the Ryukyus

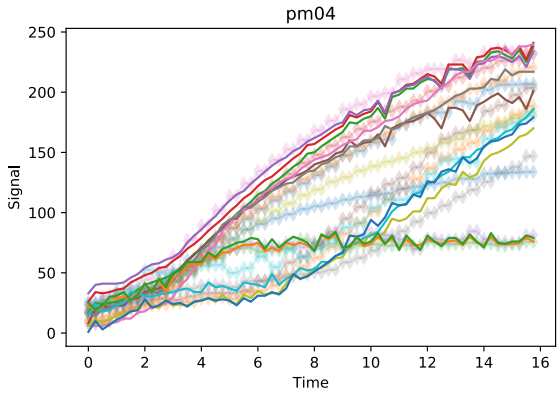

図 1　プロトタイプとクラスタ所属確率が 0.9 以下のデータ（クラスタ数 12）

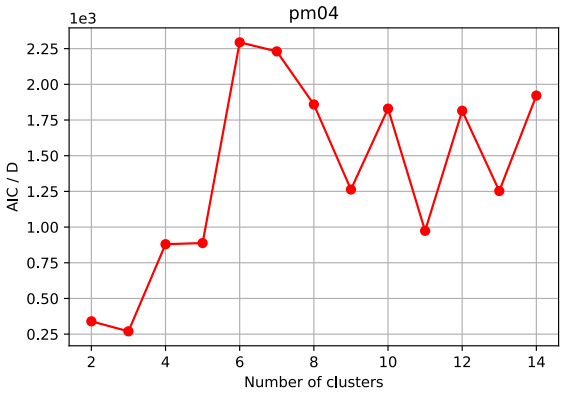

図 2　$IC_{|P|}$値

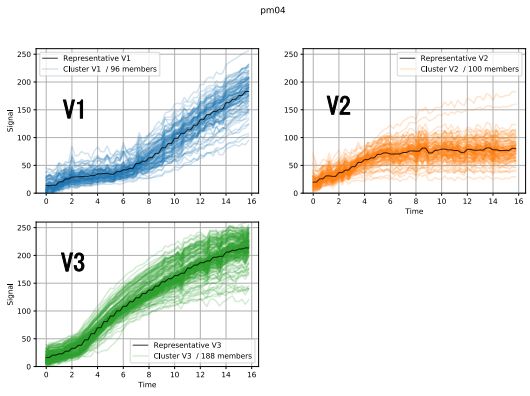

図 3　クラスタリング結果（クラスタ数 3）

3.3 受注履歴データへの適用

PM データへの適用結果から、ユニークなパターンを既知のパターン数と近いクラスタ数で抽出できることが示された。次に、パターン数が未知で性質の異なる受注履歴データに対してもユニークなパターン抽出が可能であるかを検証する。

クラスタ数に対する$IC_{|P|}$値は図 4 となる。AIC 値が最小となるのはクラスタ数が 2 の時で、そのクラスタリング結果は図 5 のようになる。2 つのプロトタイプは水準が異なり、クラスタ V2 は 3 月以降にかけて受注量がやや増えているように見えるものの、両者に属するデータの水準にはばらつきがありクラスタ数 2 は極端なクラスタ数であるように思われる。

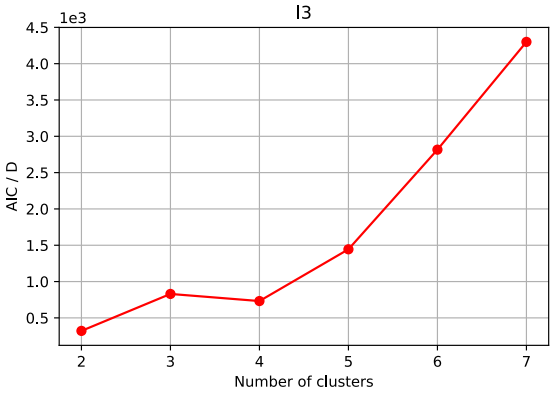

図 4　$IC_{|P|}$値

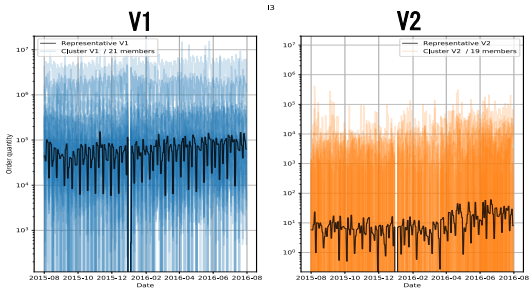

図 5　クラスタリング結果（クラスタ数 2）

4. おわりに

本稿では、ガウス混合モデルを用いて時系列データからいくつかの代表パターンの抽出を行い、その中でユニークな代表パターンが得られるクラスタ数を提示する手法の検討と評価を 2 つのデータセットを用いて行った。PM データへの適用では得られたパターン数が正解パターン数と異なるもののユニークなプロトタイプを獲得できた。一方で、値の上下が激しい性質を持つ受注履歴データではクラスタ数 2 という極端なクラスタリング結果が得られた。今後の課題として異なる性質をもつデータに対しても適当なクラスタ数を提示できる手法へと修正することが挙げられる。

参考文献

[1] Bong-Rae Kim, Timothy McMurry, Wei Zhao, Rongling Wu, and Arthur Berg, “Wavelet-Based Functional Clustering for Patterns of High-Dimensional Dynamic Gene Expression”, Journal of Computational Biology, (2010).

[2] Hui Zhang, Tu Bao Ho, Yang Zhang, and Mao Song Lin, “Unsupervised feature extraction for time series clustering using orthogonal wavelet transform”, Informatica (Slovenia), Vol. 30, No. 3, pp. 305–319, (2006).

[3] Anton Y. Peleg, Anna de Breij, Mark D. Adams, Gustavo M. Cerqueira, Stefano Mocali, Marco Galardini, Peter H. Nibbering, Ashlee M. Earl, Doyle V. Ward, David L. Paterson, Harald Seifert, and Lenie Dijkshoorn, “The success of acinetobacter species; genetic, metabolic and virulence attributes”, PLOS ONE, Vol. 7, No. 10, pp. 1–11, 10 (2012).