修士論文 要旨

ネットワーク上を伝播するイベント時系列データのクラスタリング

数理情報学専攻 48206210　金井 亮雅
指導教員　山西 健司 教授

1　序論

1.1　動機

イベントデータのクラスタリングにおいて従来考えられてきた時空間的な近接性に着目する手法では，ネットワーク上の伝播を通じて一連のイベントが発生する場合に伝播構造を反映したクラスタが得られるわけではない．そこで本研究は伝播構造に適合するクラスタを構成するためのクラスタリング手法を提案する．

1.2　関連研究

情報拡散モデルとして代表的なモデルに Independent Cascade (IC) モデル [1] がある．この IC モデルの伝播遅れ時間に連続的な分布を導入したモデルが連続時間 IC モデル [2] があり，EM アルゴリズムを通じてモデルのパラメータ推定を行うことができる．

2　提案手法

2.1　問題設定

グラフ $G = (V, E)$ とイベントデータ $X = \{x_1, \cdots, x_N\}$ が与えられており，各イベントデータはイベント発生場所 $v_j \in V$ と発生時刻 t_j のペア $x_j = (v_j, t_j)$ で表現されている．そのうち K 個のイベントが発生源 (root event) であり，残りのイベントが伝播により発生していると仮定する場合に，同一発生源から生成された一連のイベントが同一クラスタに含まれるようなクラスタリング手法を構成する．

ここで提案手法の新規性を列挙する．

- 連続時間 IC モデル [2] の推定問題をベースに，伝播構造を持つイベントに対するクラスタリング問題を定式化し，その推定アルゴリズムを構築した．
- 提案手法に基づくクラスタリングが伝播構造を反映したクラスタを得ることができ，さらに伝播特性の差異に基づいて伝播系列を識別できる．

2.2　定式化

まず発生源を除く各イベント x_j に親イベントの存在を仮定し，その親イベントの候補集合 F_j^+ を定義する．ここで x_j の親イベント候補 $i \in F_j^+$ を 1 つ取る．このとき x_i から x_j 間で k 番目の系列が伝播する場合の確率密度を

$$a_{ij}^k = p_k r_k \exp\left(-r_k(t_j - t_i)\right), \tag{1}$$

伝播しない場合の確率を

$$b_{ij}^k = p_k \exp\left(-r_k(t_j - t_i)\right) + (1 - p_k) \tag{2}$$

と定める．ここでパラメータ p_k，r_k が系列ごとの伝播特性を定めるパラメータとなっている．

ここで各イベントのクラスタ所属確率 π_{jk} を導入し，伝播系列の混合モデルにおける周辺分布の尤度関数を書き下すと次式になる．

$$\begin{aligned} l(X) = \prod_j \Big\{ \sum_{i \in F_j^+} \Big(\sum_k \pi_{ik} a_{ij}^k \prod_{l \in F_j^+, l \neq i} \sum_k \pi_{lk} b_{lj}^k \Big) \\ \times \Big(\sum_k \pi_{jk} (1 - p_k) \Big)^{N_j} \Big\}. \end{aligned} \tag{3}$$

ただし N_j は x_j からイベントが伝播しない隣接ノード数を表す．

2.3　推定アルゴリズム

混合モデルのパラメータ推定を行うために，EM アルゴリズムの枠組みに基づくアルゴリズムを構築する．

2.3.1　E ステップ

データが与えられたもとでの，イベント x_j の親イベント候補 $i \in F_j^+$ が親イベントである確率 α_{ij} と x_j のクラスタ割り当て確率 γ_{jk} を推定する．

ここで α_{ij} の更新において x_j のみならず，将来のデータの発生パターンを考慮して近似的に親イベントの推定を行う点が従来の IC モデルベースの手法と技術的に異なる部分である．

2.3.2　M ステップ

Q 関数をもとに，パラメータ p_k，r_k，π_{jk} の推定値を更新する．

3　数値実験

3.1　人工データへの適用

2 次元格子上で 2 つのイベント系列が伝播する場合を考える．これらのイベントに対し，横軸に時刻，縦軸にノード番号をとりプロットした図を図 1 に示す．

図 1 から伝播の時間スケールが異なる系列が混合している様子が分かる．ここで各イベントがどちらの系

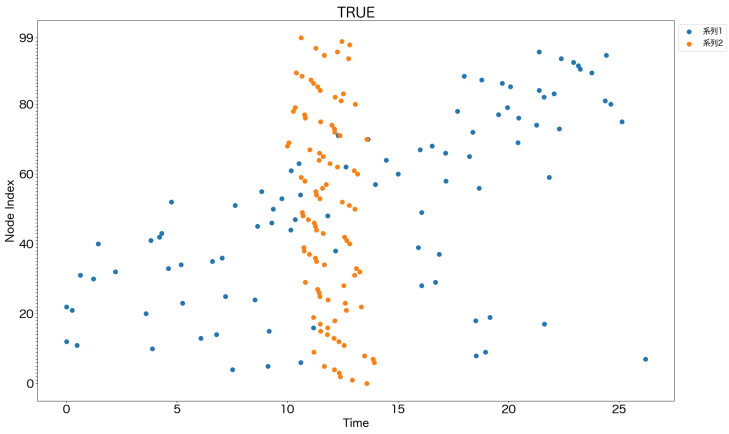

図 1. 2 次元格子上のイベント伝播のプロット表現.

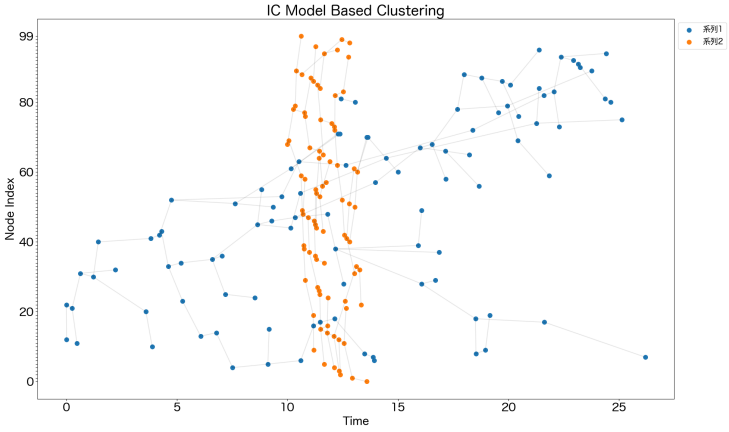

図 2. 提案手法に基づくクラスタリング結果.

列に含まれるか不明な場合に，そのクラスタの推定を行うのが提案手法であり，root event が既知の場合に推定アルゴリズムを実行した結果を図 2 に示す．

まず各イベントにおいて π_{jk} が最大となるクラスタを割り当てて色分けしている．また図中の線分は，各イベントに対して親イベント確率が最大となるイベントと結んだものとなっており，伝播確率が高い伝播経路を表現している．この結果を見ると伝播特性を考慮したクラスタが形成されていることが分かり，実際パラメータ r の推定値は $\hat{r}_1 = 0.52$, $\hat{r}_2 = 3.57$ であり，系列ごとの時間スケールの違いが反映されている．

3.2 COVID-19 データ解析への応用

各都道府県の COVID-19 日別感染者数に対して，SCAW[3] を適用することで変化点検知を行い，その結果得られた変化点からなるイベントデータと，都道府県間流動表をもとに往来が多い都道府県間にエッジを与えた移動ネットワークを作成した．

このデータに関して提案手法を適用する．伝播遅れ時間の閾値を $\Delta = 35$ と設定し，親イベント候補が存在しないイベントを root event とした場合の実験結果を図 3 に示す．

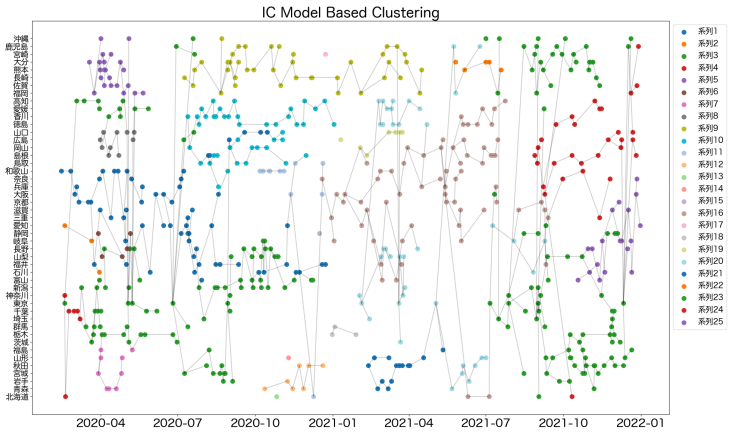

図 3. 変化点に対する提案手法の適用結果.

図 3 より，変化点の発生時期と地域ごとに応じて伝播系列が形成されていることが観察され，さらに伝播経路から都市圏から周辺地域に感染が広まっている可能性が示唆される．以上の考察から提案手法が変化点間の関連を考察し，変化に対する解釈を与える方法論になっていることが分かる．

4 結論

本研究では連続時間 IC モデル [2] をベースにイベント伝播系列の混合モデルを定式化し，その推定を通じてクラスタリングを行う手法を提案した．数値実験を通じて異なる伝播特性を持つイベント系列の識別性能を確かめるとともに，COVID-19 データを対象に変化点からなるイベントに対して提案手法を適用し，変化の波及構造を解析することが感染症流行を捉えるために有効であることを確認した．

展望としては root event の選択方法の確立，アルゴリズムの改良，伝播系列の識別可能性の理論的考察，より広い実データへの適用が挙げられる．

参考文献

[1] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146, 2003.

[2] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda. Learning continuous-time information diffusion model for social behavioral data analysis. In *Proceedings of the 1st Asian Conference on Machine Learning*, pp. 322–337. Springer, 2009.

[3] Ryoya Kaneko, Kohei Miyaguchi, and Kenji Yamanishi. Detecting changes in streaming data with information-theoretic windowing. In *Proceedings of 2017 IEEE International Conference on Big Data (Big Data)*, pp. 646–655. IEEE, 2017.