修士論文 要旨

変化検知とグラフ埋め込みのための潜在変数モデル選択の研究

数理情報学専攻 48-226206　浦野 健人
指導教員　山西 健司 教授

1　はじめに

潜在変数モデルは，観測変数と潜在変数の同時分布で表現された確率モデルであり，潜在変数モデル選択は，データの背後に隠れた構造を発見することに繋がる．本研究では，2 つの問題設定に対して，潜在変数に明示的に着目することによる潜在変数モデル選択に関する手法を提案する．1 つ目は，有限混合モデルにおけるクラスタ構造変化の予兆検知で，2 つ目は，グラフ埋め込みにおける次元とクラスタ数の選択である．

2　クラスタ構造変化予兆検知

クラスタ構造の変化は，現実の重要なイベントと対応するため，クラスタ構造変化を検知する意義は大きい．構造変化検知の従来の手法は，離散的な構造の突発的な変化を扱っていた．しかし現実には，図 1 のように構造の変化も徐々に起きると考えた方が自然な場合があり，徐々に起こる変化を早期に検知することを目的とした，クラスタ構造変化予兆検知の研究が進んでいる．

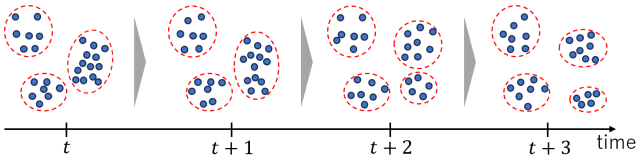


図 1: クラスタ数が徐々に 3 から 4 に変化する様子.

クラスタの比率の偏りやクラスタ間の重なりを考慮して，連続値でクラスタサイズを定義した概念として，Mixture Complexity (MC) [2] がある．MC では，混合数 k の有限混合モデル f：

$$f(x) = \sum_{i=1}^{k} \pi_i g(x; \mu_i)$$

に対して，観測変数 X に加えて，観測変数 X がどのクラスタから生成されたのかを表す，潜在変数 $Z \in \{1, \ldots, k\}$ を導入する．そして，潜在変数と観測変数の相互情報量を，クラスタサイズを連続的に評価した指標としている．

本研究では，モデルの記述次元を連続値で定義した Descriptive Dimensionality (Ddim) [3] で用いられていた Model fusion のアイデアを用いることで，MC の改良を試みる．Model fusion では，複数の混合数の有限混合モデルが混ざり合う状況を仮定する．$\mathcal{K} = \{k_1, \ldots, k_s\}$ をとり得る混合数の集合とし，混合数 $k \in \mathcal{K}$ の有限混合モデルからデータが生成されている確率が $p(K = k)$ であるとする．この仮定の下では，MC を以下のように拡張でき，これを計算したものを MC-fusion とする：

$$I(Z; X|K) = \sum_{k \in \mathcal{K}} p(K = k) \cdot I(Z; X|K = k).$$

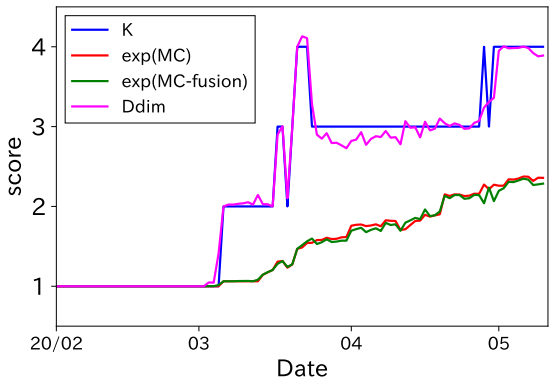
各 $k \in \mathcal{K}$ に対して $I(Z; X|K = k)$ は MC と同様に求める．$p(K = k)$ については，NML 符号長とモデルの推移確率を考えることで，時系列データに対して各時刻 t における，$k \in \mathcal{K}$ に対する $p(K_t = k)$ を推定する方法を与える．

そして，MC-fusion をクラスタ構造変化予兆検知に応用し，人工データと実データを用いて有効性を実証した．人工データの解析では，MC-fusion は，構造変化の早期検知の観点で，既存手法の MC および Ddim よりも良い結果を示した．MC-fusion は，特に MC では早期検知に失敗したようなクラスタの分裂を検知する際に効果的であった．また，Ddim はクラスタの消滅を捉えることができなかったが，MC-fusion はクラスタの偏りの変化を早期に捉えられた．人工データの実験で見られた傾向は，実データの実験でも見られた．COVID-19 感染症データの解析結果の一部を図 2 に示す．MC-fusion と MC は，感染初期に流行国が増えてクラスタの偏りが変化するのを滑らかに捉えられた．感染中期には，流行が進んだ国のクラスタが，感染が継続する国と収まりつつある国に分裂する変化の予兆を，MC では捉えられなかったが MC-fusion で早期に捉えられた．

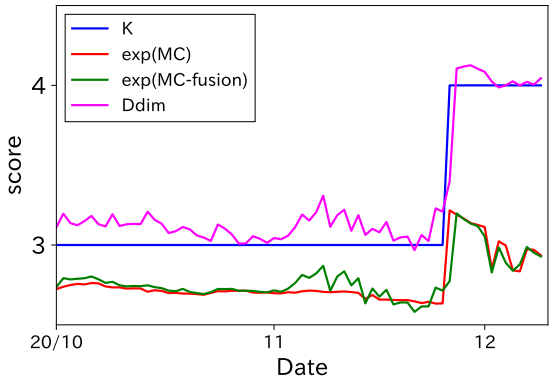
3　グラフ埋め込みの次元とクラスタ数選択

グラフのノード埋め込みによって得られたベクトルらに，クラスタ構造が見られることがある．本研究では，埋め込みベクトルらをガウス混合モデルでクラスタリングする，ガウス混合埋め込みを考える．

コミュニティ埋め込み (Community Embedding, ComE) [1] は，最適化計算の損失関数に，「コミュニティを意識した近接性」を保持するための項を加え，埋



(a) 2020 年 2 月 1 日から 2020 年 5 月 10 日.



(b) 2020 年 10 月 1 日から 2020 年 12 月 9 日.

図 2: COVID-19 感染症データのクラスタ構造の解析.

め込みの更新と，ガウス混合モデルのパラメータの更新を繰り返す手法である．ガウス混合埋め込みにおけるクラスタ数は，適切なコミュニティ構造を捉えるために重要なハイパーパラメータであり，埋め込み次元は，リンク予測等の応用タスクの精度や訓練時間に影響を与えるが，ComE ではクラスタ数や埋め込み次元を事前に設定する必要がある．

本研究では，ガウス混合埋め込みにおいて，埋め込み次元とクラスタ数の両方を選択する手法を提案する．埋め込み空間の次元を D，クラスタ数を K とする．n をノード数とする．隣接行列を表す $\boldsymbol{y} = \{y_{i,j}\}_{(i,j)\in\Lambda_{[n]}}$ を観測変数とする．このとき，潜在変数として，各ノードの埋め込みベクトルを表す $\boldsymbol{\phi} = \{\phi_i\}_{i\in[n]}$ と，各ノードの属するコミュニティを表す $\boldsymbol{z} = \{z_i\}_{i\in[n]}$ を導入する．これらに確率構造を入れることで，ガウス混合埋め込みを潜在変数モデルとして定式化し，完全変数 $(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{z})$ の DNML 符号長 [4] を計算する：

$$L_{\mathrm{DNML}}(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{z}) = L_{\mathrm{NML}}(\boldsymbol{y}|\boldsymbol{\phi}) + L_{\mathrm{NML}}(\boldsymbol{\phi}|\boldsymbol{z}) + L_{\mathrm{NML}}(\boldsymbol{z}).$$

また，潜在変数を推定する最適化アルゴリズムも与える．最適化計算では，次の 2 つのステップを繰り返す：

1. GMM のパラメータを固定して，確率的勾配降下法によって $\boldsymbol{\phi}$ と $p(\boldsymbol{y}|\boldsymbol{\phi})$ に含まれるパラメータ β,γ を最適化する．
2. $\boldsymbol{\phi},\beta,\gamma$ を固定して，EM アルゴリズムによって GMM のパラメータを最適化する．

そして，提案手法の有効性を人工データを用いて実証した．特にクラスタ数の選択に注目すると，他の情報量規準に基づいた手法に比べて，より高い割合で真のクラスタ数を選択できた．また，提案手法は，埋め込みの計算と埋め込みベクトルらのクラスタリングを同時に扱って計算しているため，埋め込みの計算とは独立に埋め込みベクトルらのクラスタリングを行う手法と比較しても，クラスタ数をよりよく捉えられた．

4 まとめ

クラスタ構造変化予兆検知では，潜在変数に明示的に着目してクラスタサイズを連続的に評価する手法を提案した．グラフのガウス混合埋め込みでは，埋め込みベクトルを潜在変数とみなして，観測変数と潜在変数の両方を符号化することによるモデル選択手法を提案した．いずれも，潜在変数に注目した，潜在変数モデル選択に関する手法であり，それぞれ有効性を実証した．今後の展望としては，動的なグラフデータに対して，ガウス混合埋め込みの次元とクラスタ数選択を動的に行い，構造変化の予兆検知に繋げるというように，本研究で提案した 2 つの手法を結びつけることもできるだろう．

参考文献

[1] S. Cavallari, V. W. Zheng, H. Cai, K. C. Chang, and E. Cambria: Learning Community Embedding with Community Detection and Node Embedding on Graphs. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*, pp. 377–386. ACM, 2017.

[2] S. Kyoya and K. Yamanishi: Mixture Complexity and Its Application to Gradual Clustering Change Detection. *Entropy*, 24, 10, pp. 1407. MDPI, 2022.

[3] K. Yamanishi and S. Hirai: Detecting Signs of Model Change with Continuous Model Selection Based on Descriptive Dimensionality. *Applied Intelligence*, 53, 22, pp. 26454–26471. Springer, 2023.

[4] K. Yamanishi, T. Wu, S. Sugawara, and M. Okada: The Decomposed Normalized Maximum Likelihood Code-length Criterion for Selecting Hierarchical Latent Variable Models. *Data Mining and Knowledge Discovery*, 33, 4, pp. 1017–1058. Springer, 2019.