修士論文 要旨

Cluster Structure Analysis in Finite Mixture Models With Component Overlap
(分布間に重なり合いを持つ有限混合モデルのクラスター構造の解析)

数理情報学専攻 48196212　京谷 駿希
指導教員　山西 健司 教授

1　はじめに

有限混合モデルは，データの出方を確率分布の足し合わせで説明するモデルであり，それぞれの要素を 1 つのクラスターとみなすことによってクラスタリングに広く応用されてきた．しかし，要素間に重なり合いがある場合，それらを別々のクラスターとみなせるかどうかは不明確になってしまう．そこで本研究では，重なり合いを持つ混合分布に対して「クラスター」の意味を問い直し，曖昧さを持つクラスター構造を定式化する方法について考察した．さらに，その応用として「クラスター構造の連続的な変化検知」と「クラスター構造の要約」の 2 つの問題に対して新たな手法を提案した．

2　記法

有限混合モデルの確率分布 f を以下のように書く：

$$f(x) = \sum_{k=1}^{K} \rho_k g_k(x) \ ,$$

ただし，K は要素数，ρ_k は混合割合，g_k は各要素の確率分布を表す．データに対応する確率変数 X を**観測変数**と呼び，X がどの要素からサンプルされたのかを表す確率変数 Z を**潜在変数**と呼ぶ．これらについて，

$$P(Z = k) = \rho_k \ , \quad P(X|Z = k) = g_k(x)$$

が成り立つ．さらに，潜在変数の事後確率を

$$\gamma_k(x) \coloneqq P(Z = k|X) = \frac{\rho_k g_k(x)}{f(x)}$$

と書く．

3　クラスター数の連続的な測定

本研究では，2 つの方法で曖昧なクラスター構造の解析を行った．はじめに，**Mixture Complexity (MC)** と呼ばれる新たな指標を導入し，混合分布中のクラスター数を連続値として捉える方法を提案する．まず，情報理論の観点から，潜在変数と観測変数の間の相互情報量がクラスター数 (の対数値) の連続的な拡張と言えることを示す．そして，そのデータ $x^N = x_1, \ldots, x_N$ による近似として MC を以下のように定める：

$$\begin{aligned} &\mathrm{MC}\Big(\{\gamma_k(x_n)\}_{k,n}\Big) \\ &\coloneqq -\sum_{k=1}^{K} \widetilde{\rho}_k \log \widetilde{\rho}_k + \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(x_n) \log \gamma_k(x_n) \ , \end{aligned}$$

ただし，$\widetilde{\rho}_k = \sum_n \gamma_k(x_n)/N$．これは，図 1 に示すように要素間の重なり合いと重みの偏りを考慮しながらクラスター数を連続的に定式化した指標になっている．

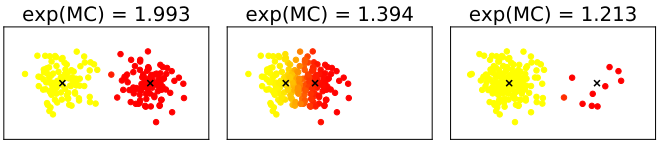

図 1. 2 要素の混合分布と MC の例

次に，MC の理論的な性質を整理する．まず，基本的な性質として，要素が完全に重なるときに最小値 0 を取ること，要素が完全に離れていて重みも等しいときに最大値 $\log K$ を取ること，そして MC の値が混合分布の表記に依らないことを確認した．次に，混合分布が部分的な構造の階層的な組み合わせで書ける場合に，MC の値も部分構造ごとの和に分解できることを示した．これは，全体だけでなく細部の構造まで定式化するのに有用である．最後に，推定された混合分布 $\hat{f}$ を用いて MC を計算する際に，$N \to \infty$ につれて $\hat{f}$ の分布が f に収束するならば，$\hat{f}$ の要素数に関わらず，MC の値は真の分布に対する MC に収束することを示した．これは，MC が要素数の違いに影響されないクラスター構造の本質的な量であることを示唆している．

そして，MC をクラスター構造の連続的な変化検知へと応用した．従来，クラスター構造の変化は要素数またはクラスター数の変化と考えられ，突然のものとされてきた [4]．本研究では，[4] の手法を用いて要素数を推定しつつ MC の値を追跡することで，クラスター数の連続的な変化を説明できることを示した．さらに，MC の階層的な分解値を追跡することで，変化の場所や内容の詳細を分析する手法も提案した．

最後に，人工データと実データを用いて MC による変化検知の性能を確認した．ここではある家庭の電力

消費量データ [2] に適用した結果を示す．に示す．実験では，あらゆる日における同時刻の消費量をまとめたものを 1 時点のデータとみなし，時刻ごとの変化を観察した．そのため，クラスター数が多いほど，その時刻での活動が多様で活発であると考えられる．実験結果を図 2 に示す．図から，MC の値の推移は要素数/クラスター数の変化を滑らかに繋いだものとなっており，手法間の差異を減らすことができていることも分かる．論文中では，この後 MC の分解値についてもその推移を列挙し，変化の詳細についてさらに分析した．

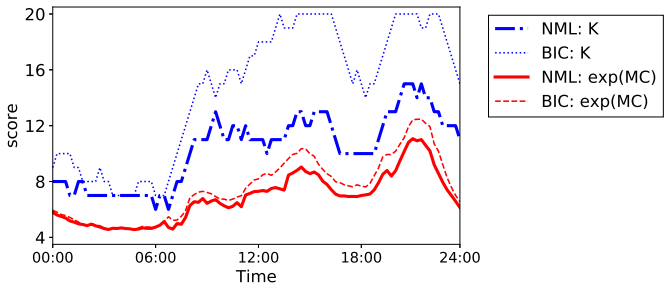

図 2. 電力消費量データに対する要素数/クラスター数 K と MC の推移．BIC と NML は混合モデルを選択する手法名．

4　クラスターの結合と要約

MC は，混合分布中のクラスター数を連続的に表現するが，具体的なクラスターを抽出することはしない．そこで 2 つ目の方法として，混合分布の要素を結合し，複数の要素を 1 つのクラスターとみなす手法について考察した．これは最も重なり合いが最も大きい 2 つの要素を結合する操作を繰り返していく手法で，重なり合いの度合いをどのように測るかが大きな問題となる．そのためにこれまで様々な指標が提案されてきたが，それらについて十分に比較検討されてこなかった．

本研究ではまず，重なり合いを測る指標が持っているべき必要条件を提示した．これは各指標に

- 要素が完全に重なる時に最良値を取ること
- 要素が完全に離れている時に最悪値を取ること
- 混合割合のスケールに対して不変であること

という自然で最小限の要求を課すものである．その後，指標としてエントロピー [1]，誤分類確率 [3]，MC を用いた手法を挙げ，上記の条件をみたすように指標を修正した．そして，修正によってより良いクラスター構造が得られることを実験的に確認した．

さらに，結合されたクラスター構造を定量的に解釈する方法も提案した．要素の結合後，クラスター構造は結合された上位の要素間の関係と各上位要素に属する下位要素の 2 つの構造に分けられる．それらについて，各要素がクラスターと見なせるかを MC とそれを標準化した **NMC** によって定式化する．この度合いを，MC は重なり合いと重みの観点から，NMC は重なり合いのみの観点から評価することができる．ここでは，乳がんのデータ [5] に対してクラスター要約を行った例を紹介する．クラスターの要約を表 1 に，予測されたクラスターを図 3 に示す．この要約から，クラスター構造に関する有用な解釈が得られる．例えば，2 つの結合された要素 1 と 2 を比べると，要素 2 の方は MC が小さいが，NMC は大きくなっている．ここから，要素 2 の方がより単一のクラスターに近いが，他とは外れた小さな要素を含んでいることが分かる．また，実験では各要素内の MC と NMC が真のラベルをもとにした分類の精度とも相関があり，クラスター構造の確信度とも関連付けられることを明らかにした．

表 1. 乳がんのデータに対するクラスター要約

Upper-components			
MC (exp):		0.509 (1.66)	
NMC:		0.763	
Component 1		**Component 2**	
Weight:	0.387	Weight:	0.613
MC (exp):	0.714 (2.04)	MC (exp):	0.270 (1.31)
NMC:	0.613	NMC:	0.676

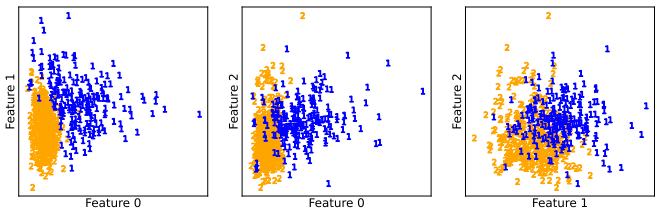

図 3. 乳がんのデータに対するクラスター結合の結果

参考文献

[1] JP. Baudry, A. E. Raftery, G. Celeux, K. LO and R. Gottardo. Comnbining mixture components for clustering. *J. Comput. Graph. Stat.* 19(2): 332–353, 2010.
[2] D. Dheeru and G. Casey. UCI machine learning repository. 2017. URL `http://archive.ics.uci.edu/ml`
[3] C. Hennig. Methods for merging Gaussian mixture components. *Adv. Data. Anal. Classif.* 4: 3–34, 2010.
[4] S. Hirai and K. Yamanishi. Detecting changes of clustering structures using normalized maximum likelihood coding. in *Proc. 18th KDD*, 343–351, 2012.
[5] O. L. Mangasarian, W. N. Street and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Oper. Res.* 43(4): 570–577, 1995.