Technical Report of the 3rd Workshop on Latent Dynamics (Sep 24, 2012, Tokyo, Japan)

正規化最尤符号を用いたクラスタリング構造変化点検出

平井　聡*　　山西　健司†
So Hirai　　Kenji Yamanishi

Abstract: 本研究では，多次元ガウス混合分布に対する正規化最尤符号 (NML) の計算法を提案し，クラスタリングへの応用を考える．そこで MDL (Minimum Description Principle) 原理に基づいたクラスタリングの構造変化を検出するアルゴリズムを提案する．これは MDL 原理に基づいて，逐次的に記述長が最小となるクラスタリング構造変化を抽出するものである．これによりクラスターの統合・分割・生成・消滅などを符号長の観点から統一的に扱うことができる．さらに人工データを用いた検証において AIC や BIC といった既存の規準を用いた方法と比較し，クラスタリングの変化をより正確にとらえることができることを示す．また，以上の手法を実際のマーケティングデータに適用し，商品の購買パターンのクラスタリングや購買パターンの時間的変化の検出において本手法が有効であることを示す．

1　序論

1.1　問題設定

この予稿は我々の直近の論文 (KDD2012) の要点を抽出したものである [3]．本研究では多次元データのクラスタリング問題において，データが逐次的に変化するときにクラスタリングの構造の変化を検出する問題 (クラスタリング構造変化点検出) を考える．本研究ではクラスタリングのモデルとしてガウス混合分布 (GMM:Gaussian Mixture Model) を導入し，記述長最小化 (MDL:Minimum Description Length) 原理を用いたクラスタリング構造変化点検出アルゴリズムを提案する．このアルゴリズムでは，クラスタリングの記述長と変化の記述長の合計を最小とするような構造を逐次的に追跡する．

1.2　本研究の意義

本研究の新規性は以下の 3 点に要約される：

- DMS アルゴリズムの逐次的クラスタリング構造変化点検出への応用

  静的なモデルの変化点検出手法として，動的モデル選択 (DMS:Dynamic Model Selection) が Yamanishi and Maruyama [10, 11] によって提案されている．このアルゴリズムを逐次的なモデル選択問題に応用した．データが逐次的に生成するとき，データの記述長とクラスタリング構造変化の記述長の総和が最小となるようなモデルを逐次的に導出していくことを考える．これによって動的にクラスタリングの追跡が可能となり，また「クラスターの結合」や「クラスターの分解」,「クラスターの消失」や「クラスターの生成」を MDL 原理に基づいて統一的に扱うことが可能となる．

- 正規化最尤符号 (NML:Normalized Maximum Likelihood) を逐次的 DMS に新しく適用した：

  逐次的 DMS アルゴリズムでは，クラスタリングの指標の選択が大変重要な問題である．そこで，Shtarkov の minimax [8] の意味で最適な NML を用いた逐次的 DMS アルゴリズムを提案する．NML を GMM に適用すると正規化項が無限に発散してしまう，また計算量が非常に多いという問題があるが，Hirai and Yamanishi は NML を解析的・効率的に計算する方法を提案している [2]．これは Kontkanen and Myllim”aki [5] が提案している手法を応用したものである．しかし NML は符号長がパラメータに大きく依存してしまうという問題があるため，再正規化のテクニック [6] を用いてパラメータ依存度の小さい再正規化最尤符号 (RNML:Re-normalized Maximum Likelihood) を提案した [4].

- 人工データ・マーケティングデータを用いた提案

*東京大学 (現在は NTT データに所属), 〒 113-0033 東京都文京区本郷 7 丁目 3-1, e-mail So.Hirai.16@gmail.com,
The University of Tokyo (Currently in NTT DATA corporation), 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan
†東京大学, 〒 113-0033 東京都文京区本郷 7 丁目 3-1, e-mail yamanishi@mist.i.u-tokyo.ac.jp,
The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan

手法と既存手法との比較

人工データを用いた実験においては，Song's and Wang's の手法 [9]，AIC (Akaike's information criteria) [1] / BIC (Bayesian information criteria) [7] に基づく変化点検出手法との比較を行った．また，ビールの消費行動のマーケティングデータを用いて消費者の嗜好の変化などの構造変化の検出を行った．

2 RNML を用いたクラスタリング構造変化点検出

クラスタリング構造変化点検出手法として，動的モデル選択 (Dynamic Model Selection： DMS)[11] をクラスタリング構造変化点検出に適用し，RNML に基づく逐次的な DMS 手法を提案する．RNML の詳細は次節で説明する．これは MDL 原理に基づいて，逐次的に記述長が最小になるクラスタリング構造変化を抽出するものである．本手法の特徴は以下のようになる：

1. 各時刻で以下のようなクラスタリングの符号長と変化の符号長の合計を最小化するようなクラスタリング構造を選択する：

$$\begin{aligned} &L(X_t, Z_t, K_t | X^{t-1}, Z^{t-1}, K^{t-1}) \\ &\quad = l(X_t, Z_t | X^{t-1}, Z^{t-1} : K_t \cdot \hat{K}^{t-1}) \\ &\qquad\qquad + l(K_t | \hat{K}^{t-1}). \end{aligned}$$

ただし，X_t, Z_t, K_t は各時刻 t でのデータ，クラスターインデックス，クラスター数を表す．

2. 各時刻 t におけるクラスターインデックスは時刻 $t-1$ のクラスターインデックスを初期値として EM アルゴリズムによって推定する．

3. 1 時刻経過につき高々1 個のクラスターの増減 (クラスターの分割・統合を含む) しかないという制限を与えている．この限定は，クラスター構造にある程度の時間的継続性を考慮したものになっている．

3 連続値分布における正規化最尤符号

この節ではガウス分布に対する正規化最尤符号の近似的計算法を提案し，ガウス混合分布の混合数決定問題に適用する．

3.1 ガウス分布における NML

以下ではデータ $\mathbf{x}^n = (\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathcal{X}^n$ ，$\mathbf{x}_i = (x_{i1}, \cdots, x_{im})^{\mathrm{T}}$ $(i = 1, \cdots, n)$ に対するガウス分布を扱うこととする．このガウス分布に対する NML 分布は

$$f_{\mathrm{NML}}(\mathbf{x}^n) \stackrel{\mathrm{def}}{=} \frac{f(\mathbf{x}^n; \hat{\theta}(\mathbf{x}^n))}{\int_{\mathbf{y}^n \in \mathcal{X}^n} f(\mathbf{y}^n; \hat{\theta}(\mathbf{y}^n)) \mathrm{d}\mathbf{y}^n} \qquad (1)$$

のように表せる．しかし，式 (1) の分母の正規化項はこのまま計算すると無限に発散してしまうという問題がある．そこで本研究では，パラメータ $\eta = (R, \lambda_{\min})$ $(\lambda_{\min} = \lambda_{\min}^{(1)}, \cdots, \lambda_{\min}^{(m)})$ を用いて積分範囲を制限することで，NML の正規化項を近似的に計算できる以下のような定理を導いた：

定理 3.1. 正規化項の値は，

$$\int_{Y(\eta)} f(\mathbf{y}^n; \hat{\theta}(\mathbf{y}^n)) \mathrm{d}\mathbf{y}^n \quad = \quad B(m, \eta) \times \left(\frac{n}{2\mathrm{e}}\right)^{\frac{mn}{2}} \frac{1}{\Gamma_m(\frac{n-1}{2})}$$

となる．ただし，$Y(\eta)$ は以下のように定義される：

$$\begin{aligned} Y(\eta) \quad \stackrel{\mathrm{def}}{=} \quad \{\mathbf{y}^n | \quad & ||\hat{\mu}(\mathbf{y}^n)||^2 \le R,\ \lambda_{\min}^{(j)} \le \hat{\lambda}_j(\mathbf{y}^n), \\ & j = 1, \cdots, m,\ \mathbf{y}^n \in \mathcal{X}^n\} \end{aligned}$$

3.2 ガウス分布における RNML

前項の正規化最尤符号はパラメータ依存度が大きいという問題がある．この項では，パラメータ依存度を小さくした再正規化最尤符号 (RNML) を提案する．

ここでは $R, \lambda_{\min}$ を最尤推定し，ハイパーパラメータとして $\gamma = (\lambda_1, \lambda_2, R_1, R_2)$ を導入して再正規化最尤分布を以下のように定める：

$$f_{\mathrm{RNML}}(\mathbf{x}^n; \gamma) = \frac{f_{\mathrm{NML}}(\mathbf{x}^n; \hat{\eta}(\mathbf{x}^n))}{\int_{Y(\gamma)} f_{\mathrm{NML}}(\mathbf{y}^n; \hat{\eta}(\mathbf{y}^n)) \mathrm{d}\mathbf{y}^n}.$$

この正規化項に対して以下のような定理を示した：

定理 3.2. 正規化項は以下のように計算できる：

$$\int_{Y(\gamma)} f_{\mathrm{NML}}(\mathbf{y}^n; \hat{\eta}(\mathbf{y}^n)) \mathrm{d}\mathbf{y}^n \quad = \quad \left(\frac{m}{2}\right)^{m+1} \cdot \log \frac{R_2}{R_1} \cdot \left(\log \frac{\lambda_2}{\lambda_1}\right)^m.$$

3.3 GMM における RNML の効率的計算法

ここではガウス混合分布 (GMM) における RNML の効率的計算法について述べる．本研究では，クラスターインデックス $z^n = z_1 \cdots z_n$ がデータ $\mathbf{x}^n$ と共に与えられた下で以下のように RNML 符号長を計算する：

定理 3.3. *GMM での RNML 符号長は以下のようになる:*

$$\begin{aligned} &SC(\mathbf{x}^n, z^n;\ \gamma, \mathcal{M}(K)) \\ &\quad = -\log f(\mathbf{x}^n, z^n; \mathcal{M}(K), \hat{\theta}(\mathbf{x}^n, z^n)) + \log \mathcal{C}_1(\mathcal{M}(K), n) \\ &\qquad + \log \mathcal{C}_2(\mathcal{M}(K), n) + \log B(\mathbf{x}^n, z^n) + K \log I(m, \gamma), \\ &\mathcal{C}_2(\mathcal{M}(K), n) \\ &\quad = \sum_{h_1 + \cdots + h_K = n} \frac{n!}{h_1! \cdots h_K!} \prod_{k=1}^{K} \left(\frac{h_k}{n}\right)^{h_k} \cdot J(h_k). \end{aligned} \qquad (2)$$

また，式 (2) で表される $\mathcal{C}_2(\mathcal{M}(K), n)$ を直接計算しようとすると $\mathrm{O}(n^K)$ かかってしまうという問題がある．そこで以下の定理は $\mathcal{C}_2(\mathcal{M}(K), n)$ が $\mathrm{O}(n^2 \cdot K)$ で計算可能であることを示す：

定理 3.4. 式 *(2)* で与えられる $\mathcal{C}_2(\mathcal{M}(K), n)$ に対して，以下のような漸化式が成り立つ：

$$\mathcal{C}_2(\mathcal{M}(K+1), n) = \sum_{r_1+r_2=n} {}_n\mathrm{C}_{r_1} \left(\frac{r_1}{n}\right)^{r_1} \left(\frac{r_2}{n}\right)^{r_2} \mathcal{C}_2(\mathcal{M}(K), r_1) J(r_2).$$

これによって，式 *(2)* は $\mathrm{O}(n^2 \cdot K)$ で計算可能である．

3.4 実験

クラスター数を人工的に変化させたデータを生成し，クラスタリング構造を追跡する実験を行った．クラスター数を $K = 3$ $(t = 1, \cdots, 50)$, $K = 4$ $(t = 51, \cdots, 100)$ と変化させたときの実験結果が図 1 となる．この図では，縦軸が複数回実験を行ったときの平均クラスター数を表している．このグラフから，RNML によるクラスタリング構造変化点検出ではクラスター数を正確に追跡できていることが分かる．

4 マーケティングデータへの応用

ここでは提案手法をマーケティングデータ (ビールの購買行動データ QPR，株式会社マクロミル提供) のデータに適用した場合の結果について述べる．このデータでは，各消費者が購買した商品のブランドや商品を購買した日時や流通チャネルなどが示されている．

このデータに対してクラスタリング構造変化点検出実験を行った結果が図 2 となる．このグラフから，RNML を用いた実験では年末需要の変化をクラスター数の変化として捉えられていることが分かる．また，BIC では購買の変化に過剰に反応してしまう傾向があると言える．

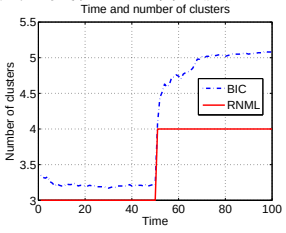

図 1: 時間と平均クラスター数

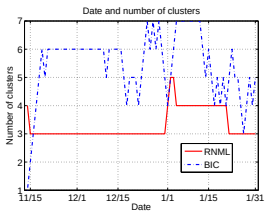

図 2: クラスター数の変化

5 結論

本研究では，ガウス混合分布に対する正規化最尤符号 (NML) の近似的計算法を示し，クラスタリングにおけるクラスター数決定問題に適用した．また，動的モデル選択に正規化最尤符号を適用し，クラスタリング構造が動的に変化する場合に構造変化点をオンラインで検出する新たな手法を提案した．そしてこれらの手法を実際のマーケティングデータに適用し，有効性を検証した．

6 謝辞

本研究にあたり，議論していただきました（株）博報堂とデータをご提供いただいた株式会社マクロミルに深謝します．本研究の一部は, 科研費基盤研究 23240019（A）, NTT に助成されたものである．また, 本研究の一部は, 総合科学技術会議により制度設計された最先端研究開発支援プログラム（FIRST 合原最先端数理モデルプロジェクト）により．日本学術振興会を通して助成されたものである．

参考文献

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Proceeding of the Second International Symposium on Information Theory*, pp. 267–281, 2011.

[2] S. Hirai and K. Yamanishi. Efficient computation of normalized maximum likelihood coding for gaussian mixtures with its applications to optimal clustering. *The IEEE International Symposium on Information Theory*, pp. 1031–1035, 2011.

[3] S. Hirai and K. Yamanishi. Detecting changes of clustering structures using normalized maximum likelihood coding. *Proc. of KDD2012*, 2012.

[4] S. Hirai and K. Yamanishi. Normalized maximum likelihood coding for exponential family with its application to optimal clustering. *arXiv 0474364*, 2012.

[5] P. Kontkanen and P. Myllymäki. A linear time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, Vol. 103, pp. 227–233, 2007.

[6] J. Rissanen. MDL denoising. *IEEE Transactions on Information Theory*, Vol. 46, No. 7, pp. 2537–2543, November 2000.

[7] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics 6 (2)*, pp. 461–464, 1978.

[8] Yu. M. Shtarkov. Universal sequential coding of single messages. *Translated from Problems of Information Transmission*, Vol. 23, No. 3, pp. 3–17, July-September 1987.

[9] M. Song and H. Wang. Highly efficient incremental estimation of gaussian mixture models for online data stream clustering. *Intelligent Computing: Theory and Application*, 2005.

[10] K. Yamanishi and Y. Maruyama. Dynamic syslog mining for network failure monitoring. *Proc. of KDD2005*, pp. 499–508, 2005.

[11] K. Yamanishi and Y. Maruyama. Dynamic model selection with its applications to novelty detection. *IEEE Transactions on Information Theory*, Vol. 53, No. 6, pp. 2180–2189, June 2007.