The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

深層混合モデルによるクラスタリング

Clustering by Deep Mixture Models

林　楓 *1　岩田具治 *2　谷口忠大 *3
Kaede Hayashi　Tomoharu Iwata　Tadahiro Taniguchi

*1*3立命館大学　*2NTT コミュニケーション科学基礎研究所
Ritsumeikan University　NTT Communication Science Laboratories

Clustering is an important task in the field of machine learning and artificial intelligence. Since probabilistic generative models have strong assumptions on data, ‘ feature engineering ’ has been required for clustering with Gaussian mixture models (GMM). In the last few years, research on clustering complicated data with a model combining Variational Autoencoder (VAE) and GMM has attracted attention. In this paper, we propose Deep Mixture Models (DMM). In DMM, a latent vector is first generated by GMMs, then latent vector is transfered into an observation. DMMs are trained by maximizing the lower bound of the marginal likelihood. In the experiment, our proposal model shows the best performance compared to baseline methods for the data which are difficult to obtain clusters by GMMs.

1. はじめに

近年のセンシング技術や情報技術の発展により，多様かつ大量のデータ得ることが可能となりつつある．それに伴い，膨大なデータから有益な情報を機械的に発見するための情報処理技術が期待されている．大量の未知データから有益な情報を発見するためにクラスタリングがよく行われる．クラスタリングとは関連するオブジェクト同士をグループ化する手法であり，機械学習や人工知能分野において最も基本的なタスクのひとつである．クラスタリングの典型的な生成モデルは混合ガウスモデル (Gaussian mixture model: GMM) である．

GMM は様々な拡張モデルが開発され広く用いられてきたが，効率的に推論するためには分布に共役性などを仮定する必要があり，その仮定を満たすようにデータの特徴を設計しなければならない．本研究の試みは，Variational Autoencoder (VAE)[Kingma 13] を用いてデータの統計的な構造を捉える良い特徴表現を学習すると同時に GMM を学習するモデルとその近似推論のフレームワークを提案し，柔軟にデータをクラスタリングすることである．本稿では提案法の生成過程とその近似推論を説明し，GMM ではクラスタリングできないデータに対してもクラスタリング可能であることを示す．

2. 関連研究

近年，GMM の学習によるクラスタ割り当てと深層ニューラルネットワークによる特徴表現学習を同時に行うことで，単純な GMM ではクラスタリングできないデータに対しても柔軟にクラスタリングする手法が研究されている [Johnson 16, Jiang 16, Dilokthanakul 16]．これらの手法の生成モデルは，図 1 であり，次のようにサンプルが生成される．まず GMM により潜在ベクトルを生成する．そして，深層ニューラルネットワークにより，潜在特徴量を観察可能なデータに変換する．Structured Variational Autoencoder(SVAE)[Johnson 16] はこのモデルの推論に確率的変分推論を用いてオンライン化したものである．Variational Deep Embedding(VaDE)[Jiang 16](VaDE) と Gaussian Mixture Variational Autoencoder(GMVAE)[Dilokthanakul 16] は VAE と GMM を訓練することでクラスタと特徴表現の分布を繰り返し学習する．特に VaDE は VAE と GMM を交互に訓練するのではなく，特徴表現とクラスタリングフレームワークを同時に学習するという点で，提案モデルの深層混合モデルと近い位置付けにあると言える．VaDE では近似推論で変分下限を計算する際，クラスタ割り当ての真の分布とその近似分布のカルバックライブラー距離を 0 に近似している一方で，提案法では二段階の下限を用いて近似している．

連絡先: 林楓，立命館大学情報理工学研究科，滋賀県草津市野路東 1-1-1，k.hayashi@em.ci.ritsumei.ac.jp

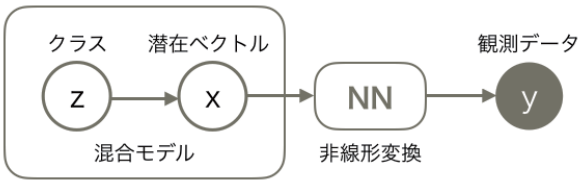


図 1: 生成過程

3. 深層混合モデル

本章では，VAE によるデータの潜在空間へのマッピングと GMM のクラスタリングを同時に行うことができる深層混合モデルである提案モデルと近似推論を説明する．

3.1 生成過程

深層混合モデルはクラスタリング問題の生成アプローチのひとつであるため，まず深層混合モデルの生成過程について説明する．具体的に，K 個のクラスタが存在するとすると観測データ $\mathbf{y}_n \in \mathbb{R}^D$ は以下の過程で生成される．ここで，

$$z_n \overset{\text{iid}}{\sim} \text{Categorical}(\boldsymbol{\pi}), \tag{1}$$

$$\boldsymbol{x}_n \overset{\text{iid}}{\sim} \text{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n}), \tag{2}$$

$$\boldsymbol{y}_n|\boldsymbol{x}_n, \gamma \overset{\text{iid}}{\sim} \text{N}(\mu(\boldsymbol{x}_n;\gamma), \Sigma(\boldsymbol{x}_n;\gamma)) \tag{3}$$

まず，混合比 $\boldsymbol{\pi}, \sum_k^K \pi_k = 1$ をパラメータとする Categorical 分布でクラスタ z_n が選ばれ，ガウス分布の平均と分散 $\boldsymbol{\mu}_k \in \mathbb{R}^d, \boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ をもつガウス分布から潜在ベクトル $\boldsymbol{x}_n \in \mathbb{R}^d$ が選ばれる．そして，潜在ベクトル $\boldsymbol{x}_n$ をパラメータ γ のニューラルネットに通すことで $\mu(\boldsymbol{x}_n;\gamma), \Sigma(\boldsymbol{x}_n;\gamma)$ を決定し，ガウス分布から観測データ $\mathbf{y}_n$ をサンプルする．

3.2 変分下限の最大化

提案法の推論では，Jensen の不等式により下限を 2 段階で求めることで，GMM の下限と VAE の下限を含んだ変分下限を導出する．変分下限は VAE で用いられている確率的勾配変分ベイズと reparametrization trick により最適化する．

まず深層混合モデルの尤度は次式のとおりである．

$$p(\mathbf{Y}) = \prod_{n=1}^{N} p(\mathbf{y}_n) = \prod_{n=1}^{N}\sum_{k=1}^{K} p(z_n)\int p(\boldsymbol{x}_n|z_n)p(\mathbf{y}_n|\mathbf{x}_n)d\boldsymbol{x}_n \tag{4}$$

N は観測データ数，$p(z_n) = \pi_{z_n}$ である．深層混合モデルの学習では，対数尤度 $\log p(\mathbf{Y}) = \sum_{n=1}^{N}\log p(\mathbf{y}_n)$ の下限を最大化するように混合ガウスモデル及びニューラルネットのパラメータを最適化する．ひとつの観測データ $\boldsymbol{y}_n$ の対数尤度の下限は Jensen の不等式を用いると，次のように求められる．

$$\begin{aligned}
&\log p(\boldsymbol{y}_n|\gamma)\\
&= \log\int p(\boldsymbol{y}_n|\boldsymbol{x}_n,\gamma)\sum_{k=1}^{K}\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)d\boldsymbol{x}_n\\
&= \log\int q(\boldsymbol{x}_n|\boldsymbol{y}_n)\frac{p(\boldsymbol{y}_n|\boldsymbol{x}_n,\gamma)\sum_{k=1}^{K}\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{q(\boldsymbol{x}_n|\boldsymbol{y}_n)}d\boldsymbol{x}_n\\
&\geq \int q(\boldsymbol{x}_n|\boldsymbol{y}_n)\log\frac{p(\boldsymbol{y}_n|\boldsymbol{x}_n,\gamma)\sum_{k=1}^{K}\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{q(\boldsymbol{x}_n|\boldsymbol{y}_n)}d\boldsymbol{x}_n\\
&\geq \int q(\boldsymbol{x}_n|\boldsymbol{y}_n)\left[\sum_{k=1}^{K}Q_k(\boldsymbol{x}_n)\log\frac{\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{Q_k(\boldsymbol{x}_n)}d\boldsymbol{x}_n\right.\\
&\left.+\log\frac{p(\boldsymbol{y}_n|\boldsymbol{x}_n,\gamma)}{q(\boldsymbol{x}_n|\boldsymbol{y}_n)}\right\}d\boldsymbol{x}_n
\end{aligned} \tag{5}$$

変分事後分布 $q(\boldsymbol{x}_n|\boldsymbol{y}_n)$ はニューラルネットによりモデル化されている．下限の γ, ϕ に関する確率的勾配を計算するため，まず各項の $q(\boldsymbol{x}_n|\boldsymbol{y}_n)$ に関する期待値をモンテカルロ近似する．ここではこれを $\mathcal{L}$ とする．

$$\begin{aligned}
\log p(\mathbf{Y}) \geq\approx \frac{1}{L}\sum_{l=1}^{L}\left[\sum_{k=1}^{K}Q_k(\boldsymbol{x}_n^{(l)})\log\frac{\pi_k p(\boldsymbol{x}_n^{(l)}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{Q_k(\boldsymbol{x}_n^{(l)})}\right.\\
\left.+\log\frac{p(\boldsymbol{y}_n|\boldsymbol{x}_n^{(l)},\gamma)}{q(\boldsymbol{x}_n^{(l)}|\boldsymbol{y}_n)}\right] \equiv \mathcal{L}
\end{aligned} \tag{6}$$

$$\boldsymbol{x}_n^{(l)} = h(\boldsymbol{y}_n;\phi) + \epsilon^{(l)}J(\boldsymbol{y}_n;\phi)^{\frac{1}{2}} \tag{7}$$

ここで，$h(\boldsymbol{y}_n;\phi), J(\boldsymbol{y}_n;\phi)$ は観測データ $\boldsymbol{y}_n$ を入力としてニューラルネットワークを通して計算されるガウス分布の平均と分散である．$\boldsymbol{\epsilon}^{(l)}$ は標準ガウス分布からサンプルされる．

$$\boldsymbol{\epsilon}^{(l)} \sim \quad \mathrm{N}(0, I) \tag{8}$$

$\mathcal{L}$ を最大化する $Q_k(\boldsymbol{x}_n^{(l)})$ は，$\frac{\partial \log p(\boldsymbol{y}_n)}{\partial Q_k(\boldsymbol{x}_n^{(l)})} = 0$ より次のように求められる．

$$Q_k(\boldsymbol{x}_n^{(l)}) = \frac{\pi_k p(\boldsymbol{x}_n^{(l)}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j p(\boldsymbol{x}_n^{(l)}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} \tag{9}$$

式 (9) を用いて $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ は次の式で更新される．

$$\pi_k^* = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L}Q_k(\boldsymbol{x}_n^{(l)})}{LN} \tag{10}$$

$$\boldsymbol{\mu}_k^* = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L}Q_k(\boldsymbol{x}_n^{(l)})\boldsymbol{x}_n^{(l)}}{\sum_{n=1}^{N}\sum_{l=1}^{L}Q_k(x_n^{(l)})}, \tag{11}$$

$$\boldsymbol{\Sigma}_k^* = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L}Q_k(\boldsymbol{x}_n^{(l)})(\boldsymbol{x}_n^{(l)}-\boldsymbol{\mu}_k)(\boldsymbol{x}_n^{(l)}-\boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N}\sum_{l=1}^{L}Q_k(x_n^{(l)})} \tag{12}$$

近似推論のアルゴリズムを Algorithm1 に示す．

Algorithm 1 提案手法の近似推論

パラメータ γ, ϕ, $\boldsymbol{\pi}$, $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ を初期化する
$\boldsymbol{Y} \leftarrow$ 全観測データ
repeat
  for $n = 1$ to N **do**
    for $l = 1$ to L **do**
      $\boldsymbol{\epsilon}^{(l)} \sim N(0, I)$
      データ $\boldsymbol{y}_n$ から $\boldsymbol{x}_n^{(l)} = h(\boldsymbol{y}_n;\phi) + \epsilon J(\boldsymbol{y}_n;\phi)^{\frac{1}{2}}$ で $\boldsymbol{x}_n^{(l)}$ を生成する
      $\log p(\boldsymbol{y}_n|\boldsymbol{x}_n^{(l)},\gamma), \log q(\boldsymbol{x}_n^{(l)}|\boldsymbol{y}_n)$ を計算する
      for $k = 1$ to K **do**
        式 (9) を用いて $\boldsymbol{x}_n^{(l)}$ から負担率 $Q_k(\boldsymbol{x}_n^{(l)})$ を計算する
      end for
      式 (11)(12) の分子をそれぞれ計算する
    end for
  end for
  式 (10)(11)(12) を用いて $\pi_k^*, \boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*$ を計算する
  $\log \pi_k^* p(\boldsymbol{x}_n^{(l)}|\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)$ を計算する
  式 (6) を用いて $\tilde{\mathcal{L}}(\gamma, \phi; \boldsymbol{Y})$ を計算し，最大化する

until converge

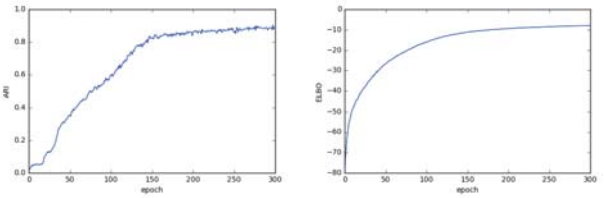
4. 実験

4.1 実験設定

本実験では，単純な GMM ではクラスタリングすることが難しいデータ (pinwheel data) に対する提案手法のクラスタリング性能を確認する，なお，評価指標には調整ランド指数 (Adjusted Rand Index: ARI)[Hubert 85] を用いた．ARI はクラスタの正解ラベルと推定されたクラスタリング結果の相関度を評価する指数であり，1 に近づくほど高い相関があることを示す．本実験で示す ARI および変分下限の値はいずれも 10 試行の平均を取っている．VAE には 2 層各 40 次元のネットワークを使用し，活性化関数は $f(x) = \tanh(x)$ を用いた．比較手法は 5-means と通常の GMM を用いたクラスタリング，VAE を用いて抽出した特徴を GMM でクラスタリングする手法 (GMM+VAE) の 3 つである．

	5-means	GMM	GMM+VAE	提案手法
ARI	0.48	0.51	0.72	**0.94**
生成	×	○	○	○

表 1: 他手法との比較



(a) epoch ごとの ARI の推移　(b) epoch ごとの変分下限の推移

図 2: VAE を用いた実験の結果

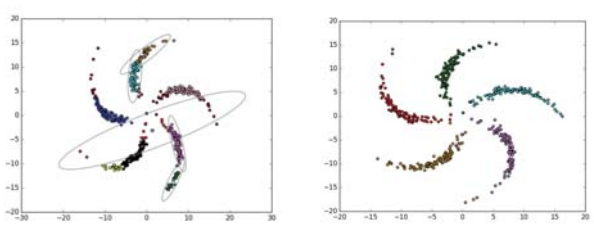
4.2 実験結果

まず，提案法と 5-means，GMM，GMM+VAE の ARI を比較した (表 4.2)．実験では提案法が最も高い ARI を示した．図 2 は提案手法は epoch ごとの ARI および変分下限の推移である．図 3 の左図は GMM によるクラスタリング結果，右図は提案手法によるクラスタリング結果であり，図 4 は提案手法の潜在空間におけるクラスタリングの過程である．収束に向かうにつれて，クラスタが GMM で表現できるような形に分離していく様子が確認できる．

5. 結論

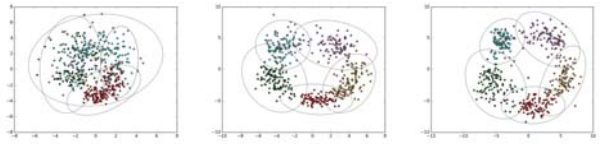
本稿では，VAE による特徴表現学習と GMM でのクラスタ割り当てを同時に行うクラスタリングのフレームワークを提案した．提案法の近似推論では，2 段階の下限を取ることで VAE の確率的勾配変分ベイズと reparameterization trick および GMM の EM アルゴリズムを組み合わせることで変分下限を導出した．実験では，深層混合モデルが比較手法よりも良い結果を示した．

今後の課題は，まず GMM ではクラスタリングが困難な音声などの実データでの実験をすることである．実データのクラスタリングを行う際，クラスタ数があらかじめわかる事例は少ない上に，オンライン学習に対応していないとスケーラビリティの観点からも難しいことが多い．提案法は GMM の部分がモデル上で定式化されているため，infinit GMM[Rasmussen 00] などのノンパラメトリックベイズに拡張しやすいと考えられる．また，オンライン化についても EM アルゴリズムを online



(a) GMM(ARI=0.68)　(b) 提案法 (ARI=0.93)

図 3: クラスタリング結果の比較



(a) 50 epochs　(b) 150 epochs　(c) 300 epochs

図 4: 潜在空間におけるクラスタリングの過程

EM[Sato 00] にすることでオンライン化に対応できる．

参考文献

[Dilokthanakul 16] Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., and Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders, *arXiv preprint arXiv:1611.02648* (2016)

[Hubert 85] Hubert, L. and Arabie, P.: Comparing partitions, *Journal of classification*, Vol. 2, No. 1, pp. 193–218 (1985)

[Jiang 16] Jiang, Z., Zheng, Y., Tan, H., Tang, B., and Zhou, H.: Variational deep embedding: An unsupervised and generative approach to clustering, *arXiv preprint arXiv:1611.05148* (2016)

[Johnson 16] Johnson, M., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R.: Composing graphical models with neural networks for structured representations and fast inference, in *Advances in neural information processing systems*, pp. 2946–2954 (2016)

[Kingma 13] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013)

[Rasmussen 00] Rasmussen, C. E.: The infinite Gaussian mixture model, in *Advances in neural information processing systems*, pp. 554–560 (2000)

[Sato 00] Sato, M.-A. and Ishii, S.: On-line EM algorithm for the normalized Gaussian network, *Neural computation*, Vol. 12, No. 2, pp. 407–432 (2000)