Bayesian Cluster Ensembles におけるベースクラスタ生成法に関する研究

1X09C024-8　加藤 裕樹
指導教員　後藤 正幸

1　研究背景・目的

近年，情報技術の発展と共に大規模データからの知識発見手法の重要性が高まっている．その手法の 1 つであるクラスタリングは，データをいくつかのグループ (以下 クラスタ) に自動分割する手法である．k-means 法などの代表的なクラスタリング手法には，探索の初期値に依存してクラスタリング結果の精度が異なるという問題がある．この問題を解決するため，Cluster Ensembles [1] (以下 CE) という枠組みが考案されている．CE は，ベースクラスタと呼ばれる複数回のクラスタリング結果を混合すること (以下 アンサンブル) で，コンセンサスクラスタと呼ばれる一意なクラスタリング結果を得る手法である．

この手法の 1 つとして，Wang ら [2] はベイズ統計の枠組みに基づき，ベースクラスタに潜在クラスを導入した Bayesian Cluster Ensembles (以下 BCE) を提案している．ここで，BCE はベースクラスタのクラスタ数 k を一定としているが，その決定法は示されていない．もし k が過度に大きい場合，学習データに対して過適合したクラスタが出来てしまい，これらをアンサンブルしてもその効果が望めない．一方，ベースクラスタのクラスタ数 k が小さい場合，クラスタリング結果が類似したベースクラスタをアンサンブルしてしまう可能性が高い．しかし，一般的にアンサンブル学習では，多様な学習器を混合した方が高性能な結果を得られることが知られており [3]，BCE においても多様なベースクラスタを用意した方が望ましい．そのため，ベースクラスタの多様性を保ちつつ，適切なクラスタ数 k を選択することが重要となる．

そこで本研究では，モデル選択基準を用いてデータにあてはまりの良いクラスタ数 k を選びつつ，複数の k によって得られる多様なベースクラスタを生成してアンサンブルする BCE を提案する．提案手法をベンチマークデータに適用し，その有効性を示す．

2　Cluster Ensembles

CE は，クラスタリング対象である p 次元のデータ集合 $\boldsymbol{\mathcal{O}} = \{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_N\}$, $\boldsymbol{o}_i \in \mathbb{R}^p$ を K 個のクラスタにクラスタリングする際，M 回の適当なクラスタ数 k を用いたベースクラスタをアンサンブルする手法である．ここで，j 回目のベースクラスタを $\boldsymbol{\lambda}_j = (x_{1j}, \cdots, x_{ij}, \cdots, x_{Nj})^T$ で表し，$\boldsymbol{\lambda}_j$ の要素 $x_{ij} \in \{1, 2, \cdots, k\}$ は，i 番目のデータに対する j 回目のベースクラスタの番号を意味するものとする．これら M 回のベースクラスタの結果 $\boldsymbol{\lambda}_j$ を各列とする $N \times M$ 行列をベースクラスタ行列 $\boldsymbol{B} = (\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_M)$ と定義する．この $\boldsymbol{B}$ の i 行目を取り出した横ベクトルを $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{iM})$ とすると，これは i 番目のデータに対する M 回のベースクラスタの結果を表している．CE は，ベースクラスタ行列 $\boldsymbol{B}$ において行または列に着目した上でアンサンブルし，コンセンサスクラスタ数 K となる単一のコンセンサスクラスタ $\boldsymbol{\lambda}^* = (\lambda_1^*, \cdots, \lambda_N^*)^T$ を得る手法である．ここで，λ_i^* は i 番目のデータ $\boldsymbol{o}_i$ のコンセンサスクラスタ番号であり，$\lambda_i^* \in \{1, 2, \cdots, K\}$ である．

CE を行うことで，個々のクラスタリングでは得がたいクラスタリング結果が得られ，データに対する頑健性が高まり，より高精度なクラスタリングが可能となることが知られている [1].

3　従来研究

3.1　Bayesian Cluster Ensembles

BCE では，各ベースクラスタの結果 x_{ij} が真のコンセンサスクラスタ番号を意味する潜在クラス z_{ij} から確率的に生起するという構造を仮定している．具体的には，ベースクラスタの結果 x_{ij} は，潜在クラスが $z_{ij} = h$ のとき β_{hj} をパラメータとする多項分布に従って生成される．また，その潜在クラス z_{ij} は，ベクトル $\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{ih}, \cdots, \theta_{iK})^T$ をパラメータとする多項分布に従う．さらに，$\boldsymbol{\theta}_i$ はハイパーパラメータ α によるディリクレ分布に従って生成される．以上の仮定のもと，i 番目のデータに対するベースクラスタの結果である $\boldsymbol{x}_i$ の生起確率は式 (1) で与えられる．

$$P(\boldsymbol{x}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int_{\boldsymbol{\theta}_i} P(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) \prod_{j=1}^{M} \sum_{h=1}^{K} P(z_{ij} = h|\boldsymbol{\theta}_i) P(x_{ij}|\beta_{hj}) d\boldsymbol{\theta}_i \qquad (1)$$

最終的に，$\boldsymbol{\theta}_i$ と $\boldsymbol{z}_i$ の事後確率を式 (2) で求め，$\boldsymbol{\theta}_i$ の事後確率が最大となるコンセンサスクラスタ番号 h を，$\boldsymbol{\theta}_i$ のコンセンサスクラスタ λ_i^* とする．

$$P(\boldsymbol{\theta}_i, \boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\boldsymbol{\theta}_i, \boldsymbol{z}_i, \boldsymbol{x}_i|\boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\boldsymbol{x}_i|\boldsymbol{\alpha}, \boldsymbol{\beta})} \qquad (2)$$

ここで未知パラメータである $\boldsymbol{\alpha}, \boldsymbol{\beta}$ は解析的に求めることが困難なため，変分ベイズ法によって推定する．BCE のグラフィカルモデルは図 1 のようになる．

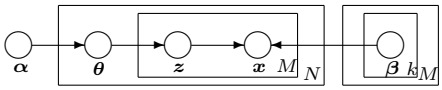

図 1.BCE のグラフィカルモデル

3.2　従来手法の BCE アルゴリズム

従来手法の BCE アルゴリズムを以下に示す．

Step1) クラスタ数を k とし，入力データ $\boldsymbol{\mathcal{O}}$ を k-means 法によりクラスタリングを M 回行い，ベースクラスタ行列 $\boldsymbol{B}$ を生成する．

Step2) Step1 で生成したベースクラスタ行列 $\boldsymbol{B}$ を用いて，式 (1)，(2) により，$P(\boldsymbol{\theta}_i, \boldsymbol{z}_i)$ を求める．$P(\theta_{ih})$ が最大となる h を λ_i^* とする．□

ここで，従来研究では，ベースクラスタのクラスタ数 k は k=K で全て一定としている．

4　提案手法

4.1　提案への展開

従来研究ではベースクラスタに用いるクラスタ数 k を全て一定としてアンサンブルを行っている．一般的にアンサンブル学習では，多様な学習器を混合した方が高性能な結果を得られることが知られており [3]，BCE においても多様なベースクラスタを用いた方がアンサンブルの効果が期待できる．

そのため，本研究では異なるクラスタ数 k に対するクラスタリング結果をベースクラスタとして用いる方法を考える．ここでベースクラスタに用いるクラスタ数 k が過度に大きい場合，データに対して過適合したクラスタを生成してしまい，アンサンブルによる精度向上が望めない．一方，小さい k を選択した場合，ベースクラスタが類似してしまい多様性を保つことができない．

そこで，ベースクラスタの多様性を保ちつつ，適切なクラスタ数 k を複数選択するためにモデル選択基準を用いる．本研究では，ベイズ統計の枠組みに基づく BCE と相性の良いモデル選択基準として，ベイズ情報量規準 (以下 BIC) [4] を用いる．BIC はベイズ統計に基づき，モデルの事後確率を漸近近似したモデル選択基準である．ベースクラスタに用いるクラスタ数の絞り込みに BIC を適用し，ランダム抽出と組み合わせることで，多様なベースクラスタを生成してアンサンブルする BCE を提案する．

4.2 ベイズ情報量規準

各クラスタに属するデータがそれぞれ正規分布に従うと仮定すると BIC は式 (3) で与えられる．ここで，n_l は l 番目のクラスタに含まれるデータ数，$\boldsymbol{\mu}_l$ は l 番目のクラスタの平均，$\boldsymbol{\Sigma}_l$ は l 番目のクラスタの分散共分散行列を示す．ここで BIC は周辺尤度とパラメータ数の和で表され，BIC の値が低いモデルの方がデータに対する当てはまりが良いモデル (クラスタ数) であると言える．

$$\mathrm{BIC} = \sum_{l=1}^{k} \left\{ \frac{1}{2} \sum_{i=1}^{n_l} (\boldsymbol{o}_i - \boldsymbol{\mu}_l)^T \boldsymbol{\Sigma}_l^{-1} (\boldsymbol{o}_i - \boldsymbol{\mu}_l) + \frac{n_l p}{2} \log 2\pi + \frac{n_l}{2} \log \Sigma_l \right\} + \frac{\log N}{2} \left(\frac{p\,(p+1)}{2} + p \right) \tag{3}$$

ただし，$\boldsymbol{\Sigma}_l$ が特異行列であると式 (3) における尤度の計算ができなくなるため，$\boldsymbol{\Sigma}_l$ が全て特異とならない範囲の k が選定対象となる．

4.3 提案手法の BCE アルゴリズム

提案手法の BCE アルゴリズムを以下に示す．

Step1) クラスタ数を k=2 とする．

Step2) 入力データ $\boldsymbol{\mathcal{O}}$ に対し，k-means 法によりクラスタリングを行い，各クラスタ l に対して $\boldsymbol{\mu}_l$ と $\boldsymbol{\Sigma}_l$ を推定する $(l = 1, 2, \cdots, k)$.

Step3) Step2 で推定した $\boldsymbol{\Sigma}_l$ のうちに特異行列があった場合 $k_{max} = k$ とし，Step4 へ．さもなければ BIC を式 (3) により求め，$k = k + 1$ とし，Step2 へ戻る．

Step4) $k = 2, 3, \cdots, k_{max}$ のうち BIC が低い上位 T 個を $\tau_1, \tau_2, \cdots, \tau_T$ とする．この集合からランダムに τ を選択し，$k = \tau$ とした k-means 法でクラスタリングを行う操作を $(M - T)$ 回繰り返す．

Step5) Step4 で得られた結果と Step2 で生成した BIC が低い上位 T 個の結果を合わせてベースクラスタ行列 $\boldsymbol{B}$ を生成する．

Step6) Step5 で生成したベースクラスタ行列 $\boldsymbol{B}$ を用いて，式 (1)，(2) により，$P(\boldsymbol{\theta}_i, \boldsymbol{z}_i)$ を求める．$P(\theta_{ih})$ が最大となる h を λ_i^* とする． □

5 実験

提案手法の有効性を検討するため，UCI 機械学習レポジトリを用いたシミュレーション実験を行った．予め，正解クラスが分かっているデータに対してクラスタリングを適用し，クラスタリング結果と正解クラスの一致度を測ることで，従来手法とのクラスタリング精度の比較を行う．

5.1 実験条件

実験には，以下の UCI データセットの Iris，Bal，Digits を用いた．データセットの内容は表 1 の通りである．

表 1. 実験データセット

	データ数	特徴量	クラス
Iris(アヤメ)	150	4	3
Bal(風船)	625	4	3
Digits(手書き文字)	9298	50	10

比較手法として，アンサンブルを行わない k-means 法と従来手法の BCE を用いた．本実験では，ベースクラスタ数を 20 個 (M=20)，提案手法のベースクラスタに用いるクラスタ数 k は BIC が低い上位 5 件 (T=5) とした．評価指標は，Micro-Precision (以下 MP) を用いた．MP は式 (4) で表される．ここで，a_h は h 番目のコンセンサスクラスタ内の正解数を示す．また，α, β の初期値を変えて，実験を 100 回行い平均をとった．

$$\mathrm{MP} = \frac{1}{N} \sum_{h=1}^{K} a_h \tag{4}$$

5.2 実験結果と考察

図 2 の実験結果より，従来手法に比べて提案手法のクラスタリング精度が向上しており，提案手法が有効であることが示された．

提案手法では，クラスタリング結果が異なる多様なベースクラスタをアンサンブルすることができたため，クラスタリング精度が向上したと考えられる．また，データ数が少ないデータセットに対しては，クラスタリング精度が大きく向上しなかった．これはデータ数が少ない場合，モデル選択基準を用いてクラスタ数 k を選択しても，小さい k が選択され易く，ベースクラスタが類似しやすいためと考えられる．

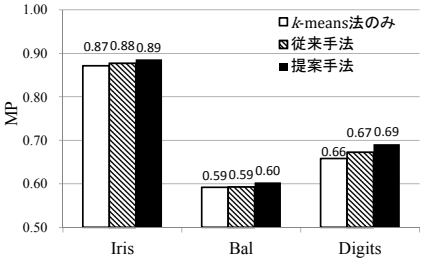

図 2. 提案手法の実験結果

6 まとめと今後の課題

本研究では，BCE において高精度なコンセンサスクラスタを得るためのベースクラスタの生成法を提案し，実験によりその有効性を示した．今後の課題は，ベースクラスタに複数用いているクラスタ数の最適な組み合わせを一意に定める手法の検討が挙げられる．

参考文献

[1] J. Ghosh, A. Acharya, “Cluster ensembles,” *WIREs Data Mining and Knowledge Discovery 1*, pp.1–12, 2011.

[2] H. Wang, H. Shan, A. Banerjee, “Bayesian Cluster Ensembles,” *the Ninth SIAM*, pp.211–222, 2009.

[3] R. Schapire, “The strength of weak learnability,” *Machine Learning,* Vol.5, pp.197–227, 1990.

[4] 松嶋敏泰,“ 統計モデル選択の概要,” オペレーションズ・リサーチ学会誌, Vol.14, pp.369–374, 1996.