卒業研究中間報告書

クラスタ数推定に用いる最適な情報量基準の探求

指導教員

藤田 一寿

津山工業高等専門学校

情報工学科

萩原 涼介

平成 29 年 9 月 27 日

目次

1. はじめに

クラスタリングとはデータを教師なし学習により任意の数のクラスタに分ける手法である．クラスタリングはデータ解析，データマイニング，パターン認識など様々な分野で用いられる．K-means を始めとする多くのクラスタリング手法では，予めクラスタ数がわかっているものとして，クラスタ数を指定しクラスタリングを行う．しかし，データに対し最適なクラスタ数を指定しなければ，最適なクラスタリング結果を得ることはできないが，一般にクラスタ数が事前にわかっているデータは少ない．その為，クラスタ数が未知である場合にも，適切にクラスタ数を推定することは重要な課題となっている．

既存の手法の多くは，データが確率分布関数から生成されたと想定して，その確率分布を生成するモデルを推定することにより，クラスタ数推定を行う．クラスタ数推定を行う際，よく用いられるのが情報量規準と呼ばれれる指標である．情報量規準とは簡単に言えば確率分布とデータの分布の当てはまり具合を表す．その情報量基準は多くの研究者により様々なものが提案されている．例えば，1973 年に赤池が提案した AIC (Akaike Information Criterion) や，Bayes の定理によって算出される事後確率を用いる BIC (Bayesian Information Criterion) が有名である．

しかし，どの情報量規準がどのようなデータに対し有効かは分かっていない．そこで本研究では，クラスタ数推定に用いる情報量規準として最適なものを数値実験を通し明らかにする．

前期は，混合等方 Gauss 分布から生成されたデータを X-means によりクラスタ数推定およびクラスタリングを行い，AIC, cAIC, BIC と呼ばれる情報量規準によるクラスタリングの性能の評価を行った．

第 2 章では，既存のクラスタリング手法である K-means のアルゴリズムの紹介を行う．

第 3 章では，本研究で利用する X-means の理論の説明を行う．まず，モデルと真の確率分布との近さを計る指標である Kullback-Leibler 情報量について述べ，それと最尤推定との情報量規準の関係性について詳しく述べる．その後，X-means の手法について述べる．

第 4 章では，本研究により得られた実験結果について述べる．

第 5 章では，本研究を通してのまとめおよび今後の課題について述べる．

2. 先行研究

2.1 クラスタリング

クラスタリングとはデータを教師なし学習により任意の数のクラスタに分ける手法である．多くのクラスタリング手法においては，データの類似度をユークリッド距離やマンハッタン距離などの距離尺度によって定義し，それによってクラスタを抽出する．クラスタリングはデータ解析，データマイニング，パターン認識など様々な分野で用いられる．

2.2 K-means

K-means[1] は，多次元空間上のデータ点集合について，各データが属するクラスタを同定する最もよく使われるクラスタリング手法の一つである．

D 次元空間上の確率変数 x の N 個の観測点で構成されるデータ集合 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ があると仮定する．このデータ集合を K 個のクラスタに分割することを考える．直感的には，クラスタとは，その内部のデータ点間の距離が外部のデータとの距離と比較して小さいテータのグループであるとみなせる．ここで，セントロイド $\boldsymbol{\mu}_k$ を導入する．K-means はベクトルの集合 $\{\boldsymbol{\mu}_k\}$ だけでなく，全データ点をうまくクラスタに対応させ，各データ点から $\boldsymbol{\mu}_k$ をへの二乗距離の総和を最小にすることで，クラスタリングを行う．

ここで，各データ点 $\boldsymbol{x}_n$ に対し，対応する 2 値指示変数 $r_{nk} \in \{0, 1\}$ $(k = 1, \cdots, K)$ を定める．これは，そのデータ点 $\boldsymbol{x}_n$ がクラスタ k に割り当てられるかを表す変数である．すなわち，データ点 $\boldsymbol{x}_n$ がクラスタ k に割り当てられる場合は $r_{nk} = 1$ とし，$j \neq k$ については，$r_{nk} = 0$ とする．これは，1-of-K 符号化法として知られている．

次に，次の目的関数 J を定義する．

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2 \tag{1}$$

これは，歪み尺度と呼ばれ，各データ点からそれらが割り当てられたベクトル $\boldsymbol{\mu}_k$ までの二乗距離の総和を表している．K-means によるクラスタリングは，J を最小にする $\{r_{nk}\}$ と $\{\boldsymbol{\mu}_k\}$ の値を求めることである．

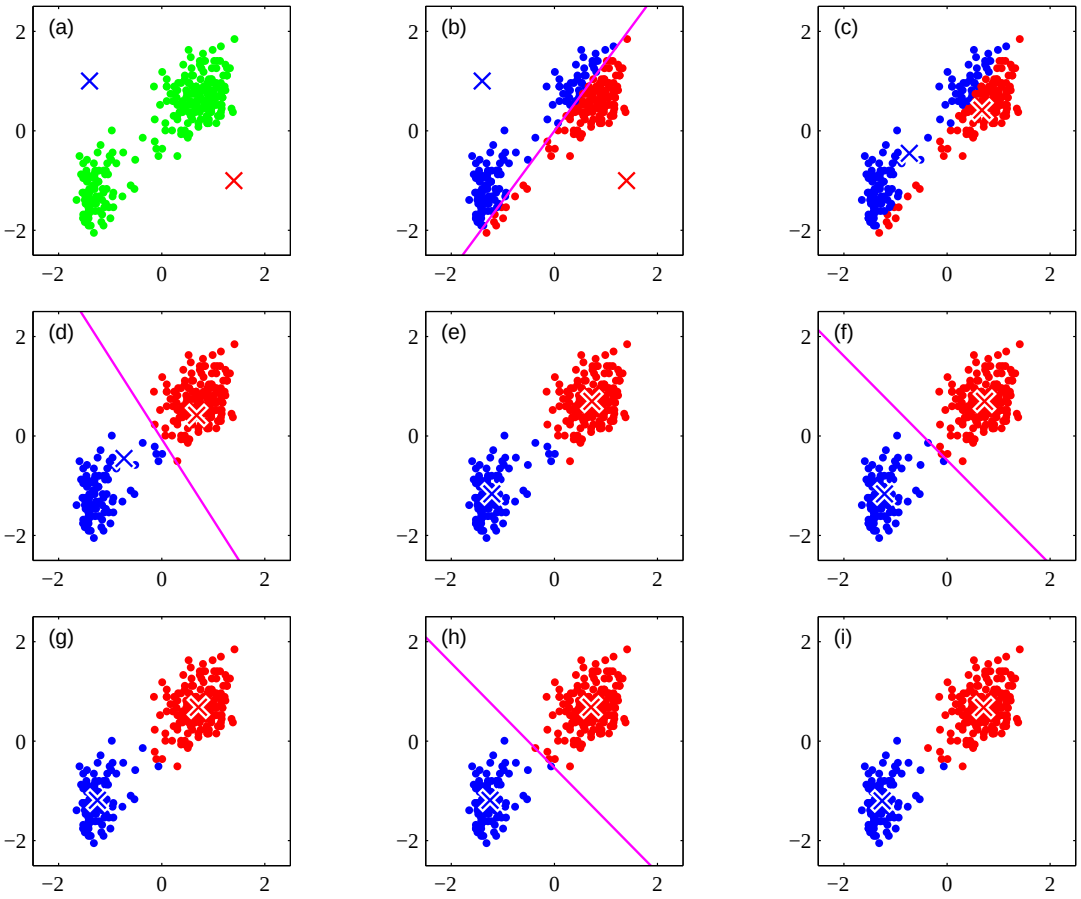

図 1　K-means の動作

図 1 に K-means によるクラスタリングの具体例を示す．K-means では，r_{nk} と $\boldsymbol{\mu}_k$ をそれぞれ最適化する 2 つのステップを交互に繰り返す手続きでクラスタリングを実現する．

最初に，$\boldsymbol{\mu}_k$ の初期値を選ぶ (a)．次に，最初のフェーズで $\boldsymbol{\mu}_k$ を固定しつつ，r_{nk} について J を最小化する (b)．第二フェーズでは，r_{nk} を固定しつつ，$\boldsymbol{\mu}_k$ について J を最小化する (c)．そして，このような二段階最適化を収束するように繰り返す．

まず r_{nk} の決定を考える．(1) 式における J は r_{nk} についての線形関数なので，最適化は代数的に解くことができる．異なる n を含む項は互いに独立である．よって，各 n について別々に $r_{nk} = 1$ としたときに，$||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2$ が最小になるような k の値に対して r_{nk} を選んで 1 とおけばよい（(2) 式）．つまり，単純に n 番

目のデータ点の中心がそれに最も近いクラスタ中心に割り当てるのである．

$$r_{nk} = \begin{cases} 1 & k = \arg\min_j ||\boldsymbol{x}_n - \boldsymbol{\mu}_j|| \text{のとき} \\ 0 & \text{それ以外} \end{cases} \tag{2}$$

次に，r_{nk} を固定したもとで $\boldsymbol{\mu}_k$ の最適化を考える．対象関数 J は $\boldsymbol{\mu}_k$ の二次関数であり，次のように $\boldsymbol{\mu}_k$ に関する偏微分を 0 とおく事で最小化できる（(3) 式）．

$$2\sum_{n=1}^{n} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{3}$$

これを $\boldsymbol{\mu}_k$ についてとくと，(4) 式を得る．

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\boldsymbol{x}_n}{\sum_n r_{nk}} \tag{4}$$

この式の分母は k 番目のクラスタに割り当てられたデータの数に等しい．それゆえこの式は $\boldsymbol{\mu}_k$ は k 番目のクラスタに割り当てられた全てのデータ点 $\boldsymbol{x}_n$ の平均値と置いているものとして単純に解釈することができる．

データ点のクラスタへの再割り当てと，クラスタ平均の再計算という 2 つのフェーズは，再割当てが起こらなくなるまで（もしくはあらかじめ定めた最大繰り返し数を超えるまで）繰り返される．各フェーズは，対象関数 J の値を減少させるので，このアルゴリズムの収束は保証されている．しかしながら，大域的最小点ではなく曲商店に収束する可能性はある．

なお，K-means によるクラスタリングは，事前にクラスタ数を指定することによりクラスタリングを行うためクラスタ数が未知の場合，K-means を用いることはできない．

3. 手法

3.1 Kullback-Leibler 情報量

偶然を伴う現象は，ある確率分布に従う確率変数の実現値であると考えることができる．この確率分布を近似するモデル（以後「モデル」）は，データを生成する真の確率分布にどの程度近いかによって評価することができる．また，データにモデルを当てはめることは，データから真の確率分布を推定しているものとみなすことができる．このようにモデルと真の分布が共に確率分布であると見なし，モデルの評価や推定を行う．

真の分布とモデルの近さを図る客観的な規準として Kullback-Leibler 情報量（以後「K-L 情報量」）がある．連続型の確率分布のとき，$g(x)$ を真の確率密度関数，$f(x)$ をモデルが定める確率密度関数とすると，モデルに関する真の分布の K-L 情報量は $\log\{g(X)/f(X)\}$ の期待値を取り (5) 式で表される．

$$\begin{aligned} I(g \mid f) &= E_X\left(\log\left\{\frac{g(X)}{f(X)}\right\}\right) \\ &= \int_{-\infty}^{\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx \end{aligned} \tag{5}$$

ただし，log は自然対数で，注記がない限り一貫してこの意味で用いる．

このように，真の分布がわかっている場合には K-L 情報量によってモデルの良し悪しを比較できた．しかし，通常は真の分布が未知で，真の分布から得られたデータだけが与えられていることがい．したがって，データから K-L 情報量を推定する必要がある．(5) 式を展開すると

$$\begin{aligned} I(g \mid f) &= \int_{-\infty}^{\infty} \left\{\log\frac{g(x)}{f(x)}\right\} g(x)dx \\ &= -\int_{-\infty}^{\infty} \{\log f(x)\}g(x)dx - \left(-\int_{-\infty}^{\infty} \{\log g(x)\}g(x)dx\right) \end{aligned}$$

となるが，右辺の第 2 項は定数であり，右辺第 1 項が大きいほど K-L 情報量 $I(g \mid f)$ は小さくなることがわかる．すなわち，K-L 情報量の大小比較のためには，本質的には $\int_{-\infty}^{\infty}\{\log f(x)\}g(x)dx$ だけを推定すれば良いことがわかる．右辺第 1 項の $\int_{-\infty}^{\infty}\{\log f(x)\}g(x)dx$ は，確率密度関数 $\log f(x)$ の期待値 $E(\log f(x))$ であり，平均対数尤度と呼ばれている．ここで，

$$\sum_{i=1}^{n}\log f(x_i)$$

を対数尤度と呼ぶことにすると，n 個の独立な観測値 $\{x_1, x_2, \cdots, x_i\}$ が得られると，この平均対数尤度は，対数尤度の n 分の 1

$$\frac{1}{n}\sum_{i=1}^{n}\log f(x_i)$$

で近似される．したがって，符号に注意すると，対数尤度が大きいほど，そのモデルは真の分布に近いと考えられる．このようにして，対数尤度を K-L 情報量の推定値を考えることにすると異なったタイプのモデルの良し悪しも比較できるのである．

ところで，確率変数 $(X_1, X_2, \cdots, X_n)$ の同時密度関数が $f(x_1, x_2, \cdots, x_n \mid \theta)$ で与えられているものとする．θ は確率密度関数を規定するパラメータである．この時，観測値 $(x_1, x_2, \cdots, x_n)$ は与えられたものとして固定し，f を θ の関数と考える時，この関数を**尤度**と呼び，$L(\theta)$ で表す．すなわち，

$$L(\theta) = f(x_1, x_2, \cdots, x_n \mid \theta)$$

である．特に，確率変数が独立な場合には $(X_1, X_2, \cdots, X_n)$ の確率密度関数は，各 $X_i(i = 1, \cdots, n)$ の確率密度関数の積に等しいことから，

$$L(\theta) = f(x_1 \mid \theta)f(x_2 \mid \theta)\cdots f(x_n \mid \theta)$$

となる．この両辺の対数をとると，すでに求められた対数尤度関数

$$l(\theta) = \sum_{i=1}^{n}\log f(x_i \mid \theta)$$

が導かれる．

ここでは，平均対数尤度の推定量から対数尤度を直接導入した．しかし，モデルが確率分布の形で与えられている場合には，まず観測値の同時分布から尤度を定義し，その対数として対数尤度を求めるほうが都合が良い．$(X_1, X_2, \cdots, X_n)$ が独立でない場合にも，尤度の対数として対数尤度

$$l(\theta) = \log f(x_1, \cdots, x_n \mid \theta)$$

が定義できる．

3.2 最尤法

ここまで，データに基づいて K-L 情報量の大小を比較するためには対数尤度を比較すれば良いことを示した．あらかじめ与えられたいくつかのモデルがある場合には，対数尤度が最大となるモデルを選択することによって，近似的には真の分布にいちばん近いモデルが得られることになる．したがって，モデルがいくつかの調整できるパラメータを保つ場合には，対数尤度を最大とするようにパラメータの値を選ぶことによって良いモデルが得られることがわかる．この推定を最大尤度法，略して**最尤法**と呼ばれている．また，最尤法で導かれた推定量は**最尤推定量**と呼ばれ，この最尤推定量によって定められるモデルが最尤モデルである．最尤モデルの対数尤度を**最大対数尤度**という．

3.3 情報量規準

情報量規準とは，最尤推定によって当てはめられたモデルが複数個あるときに，その中の一つを選択する規準である．

多くの情報量規準は，

(モデルの最大尤度対数) – (モデルの自由パラメータ数などの罰則項)

という形をしている．

以下に，いくつか情報量規準の例を示す．なお，モデル M_j の p_j 変量 Gauss 分布の対数尤度関数を $\hat{l}_j$ と表し，モデルのデータ数を R と表す．

AIC (Akaike Information Criterion; **赤池情報量規準**)

1973 年に赤池によって提案された情報量規準である [2]．

$$\mathrm{AIC}(M_j) = \hat{l}_j(D) - p_j \tag{6}$$

cAIC (Conditional Akaike Information Criterion; **条件付き赤池情報量規準**)

AIC は導出に漸近理論を使っているため，標本サイズが無限に大きいことを想定している．したがって，標本サイズが小さい場合はその過程が成り立たず，AIC によるモデル決定はパラメータ数を過大に見積もってしまう．

そこで，N. Sugiura は漸近理論を使わない不偏推定量である cAIC を導出した [3]．

$$\mathrm{cAIC}(M_j) = \hat{l}_j(D) - \frac{p_j \cdot R}{R - p_j - 1} \tag{7}$$

BIC (Beyesian Information Criterion; **ベイズ情報量規準**)

BIC は 1978 年に Schwartz によって提案された [4]．AIC とは異なり，罰則項に事後確率を利用する．

$$\mathrm{BIC}(M_j) = \hat{l}_j(D) - \frac{p_j}{2}\ln R \tag{8}$$

3.4 X-means

先に紹介した K-means は，クラスタ数を事前に指定する必要が生じる．しかし，実際にデータのクラスタリングを行う際，クラスタ数が事前に与えられることは少ない．

X-means[5] は，データ分布が混合等方 Gauss 分布から生成されたと想定してクラスタ数推定及びクラスタリングを行う手法である．K-means の逐次繰り返しと，BIC による分割停止規準を用いることで，クラスタ数を推定しクラスタリングを実行する．

具体的には以下の手順で行われる．

(1) クラスタ数 k を初期化する (通常は $k = 2$) .
(2) K-means を実行する．
(3) 次の処理を $j = 1$ から $j = k$ まで繰り返す．
    (a) クラスタ j の BIC_j を計算する．
    (b) クラスタ j に所属するデータに対し，クラスタ数 2 として K-means を行う．
    (c) クラスタ数 2 としてクラスタリングした結果に対し $\mathrm{BIC'}_j$ を計算する．
    (d) BIC_j と $\mathrm{BIC'}_j$ を比較し，$\mathrm{BIC'}_j$ が大きければクラスタ数 k に 1 を足す．
(4) 前の処理で k が増加した場合は処理 2 へ戻る．そうでない場合は終了する．

X-means で用いる BIC は次のように求められる．d 次元のデータ $\boldsymbol{D} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_d)$ を K 個のクラスタに分割することを考える．モデル M_j の評価に用いる BIC は (8) 式で与えられる．p_j はモデル M_j のパラメータ数であり，R は M_j のデータ数，$\hat{l}_j(D)$ は p 変量 Gauss 分布の対数尤度関数である．

等方 Gauss 分布を考えると分散 σ^2 は (9) 式により表される．

$$\hat{\sigma}^2 = \frac{1}{R-K} \sum_i \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}\right)^2 \tag{9}$$

すると，確率は次で表される．

$$\hat{P}(x_i) = \frac{R_{(i)}}{R} \frac{1}{\sqrt{2\pi}\hat{\sigma}^d} \exp\left(-\frac{1}{2\hat{\sigma}^2}||\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}||^2\right) \tag{10}$$

ここで $\boldsymbol{\mu}_i$ は d 次元の平均ベクトルである．したがって対数尤度関数は

$$\begin{aligned} l(D) &= \log \prod_i \hat{P}(x_i) \\ &= \sum_i \left(\log \frac{1}{\sqrt{2\pi}\sigma^d} - \frac{1}{2\sigma^2}||\boldsymbol{x}_i - \boldsymbol{\mu}_{(i)}||^2 + \log \frac{R_{(i)}}{R}\right) \end{aligned} \tag{11}$$

となる．ここでクラスタ $n(1 \leq n \leq K)$ のデータ D_n に着目する．クラスタ n のデータ数を R_n と表記すると，(11) 式は以下で表される．

$$\begin{aligned} \hat{l}(D_n) = &-\frac{R_n}{2}\log(2\pi) - \frac{R_n \cdot d}{2}\log(\hat{\sigma}^2) - \frac{R_n - K}{2} \\ &+ R_n \log R_n - R_n \log R \end{aligned} \tag{12}$$

一般的には，分割停止規準として BIC を用いるが，本実験においては，BIC 以外の情報量規準を用いたクラスタリングも行い，クラスタ数推定の精度を検証する．

4. クラスタリング実験

4.1 実験環境

実験には Python3.5 を用い，機械学習のライブラリとして TensorFlow を用いてアルゴリズムを実装した．

4.2 精度の評価

クラスタリング精度の評価は Python のライブラリである scilit-learn を用い，以下の 3 項目により行う．

ARI; Adjusted Rand Index, **調整ランド指数**

: クラスタの正解ラベルに対してクラスタリング結果の一致度を評価する指標．1 に近づくほどよい結果．

NMI; Normalized Mutual Information, **正規化相互情報量**

: 相互情報量を正規化した尺度．

Purity

: 生成されたクラスタがどれだけ多数派で占められているかを表す尺度

4.3 X-means によるクラスタリング

4.3.1 2 次元のクラスタリング

まず，2 次元のデータのクラスタリング結果を比較する．2 次元空間に分散 $\sigma^2 = 1$ の Gauss 分布により生成した，クラスタあたりのサンプル数を 500 として 5 つのクラスタを生成し，対数尤度関数，BIC, AIC, cAIC を分割停止規準として採用してクラスタリングを行った．

表 1 に 100 回クラスタリングを行ったときの，推定されたクラスタ数，ARI, NMI, Purity の平均値およびクラスタ数の分散値を示す．また，図 2 に 2 次元空間におけるクラスタリングの例を示す．

表 1　2 次元空間におけるクラスタリング結果

分割停止規準	クラスタ数（分散）	ARI	NMI	Purity
BIC	4.58 (0.9836)	0.84458792	0.88281495	0.84458792
cAIC	4.55 (0.6475)	0.85329139	0.89992544	0.85329139
AIC	4.69 (3.8739)	0.83642236	0.88147442	0.83642236
対数尤度関数	5.32 (10.236)	0.85699618	0.91572100	0.85699618

表 1 より，2 次元空間において Xmeans の分割停止規準として BIC と cBIC の間にはクラスタリング結果に大きな差がないことが読み取れる．クラスタ数の推定もおおよそ適当であり，推定したクラスタ数の分散もあまり大きな値とはなっていない．

また，AIC を分割停止規準として採用した場合に着目すると，推定したクラスタ数の分散が非常に大きくなっている．これは，3.3 節で述べたように，AIC はパラメータ数を過大に見積もってしまうことに起因するものと思われる．実際に，推定されたクラスタ数を見ると，クラスタ数を 20 や 22 と推定しているものが多く存在する．しかし，素の対数尤度関数を分割停止規準として採用した場合のクラスタリング結果と比較すると，比較的安定したクラスタ数推定を行っていることが見て取れる．

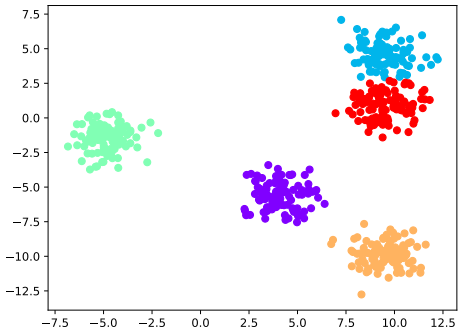

図 2　2 次元空間のクラスタリング例

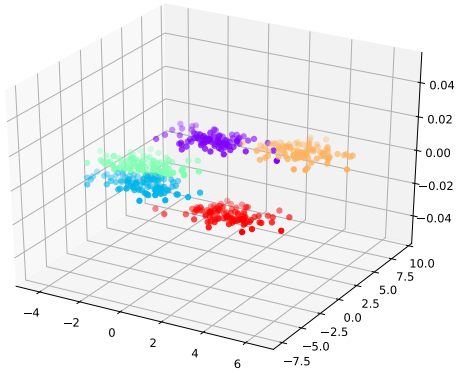

図 3　3 次元空間のクラスタリング例

4.3.2　3 次元のクラスタリング

3 次元空間に分散 $\sigma^2 = 1$ の Gauss 分布により生成した，クラスタあたりのサンプル数を 500 として 5 つのクラスタを生成し，対数尤度関数，BIC, AIC, cAIC によりクラスタリングを行った．

表 2 に 100 回クラスタリングを行ったときの，推定されたクラスタ数，ARI，NMI，Purity の平均値およびクラスタ数の分散を示す．

また，図 3 に 2 次元空間におけるクラスタリングの例を示す．

2 次元のクラスタリングとは異なり，3 次元空間においては BIC を分割停止規準として採用した場合の精度が良くなっている．分散値も非常に小さいため，安定して精度の高いクラスタリングを行っていることが伺える．

cAIC と AIC の場合を比較した場合，cAIC のほうがクラスタリングの精度が高いことがわかる．しかし，AIC におけるクラスタリングでは，2 次元空間ほど推定されたクラスタ数の分散が大きくないことがわかる．

表 2　3 次元空間におけるクラスタリング結果

分割停止規準	クラスタ数（分散）	ARI	NMI	Purity
BIC	4.95 (0.0669)	0.97179074	0.97913818	0.97179074
cAIC	4.92 (0.2313)	0.96312702	0.97023920	0.96312702
AIC	4.88 (0.1443)	0.95216819	0.96855698	0.95216819
対数尤度関数	5.12 (4.1443)	0.95731637	0.96541468	0.95731637

2 次元空間では，クラスタ数を過大に見積もってしまう問題があったが，3 次元空間においてその問題は発生していなかった．

5. おわりに

参考文献

1) James MacQueen et al.: Some methods for classification and analysis of multivariate observations, Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14, pp. 281–297 (1967).
2) Akaike, H.: Information theory and an extension of the maximum likelihood principle, Proceedings of the 2nd International Symposium on Information Theory, pp. 267-281 (1973).
3) Sugiura, N.: Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's, Communications in Statistics-Theory and Methods, 7(1), pp. 13-26 (1978).
4) Gideon Schwarz et al.: Estimating the dimension of a model., The annals of statistics Vol. 6, No. 2, pp. 461-464 (1978).
5) Dan Pelleg, Andrew W Moore, et al.: Xmeans: Extending K-means with Efficient Estimation of the Number of Clusters., ICML, Vol. 1, pp. 727–734 (2000).