筑波大学大学院博士課程

システム情報工学研究科修士論文

サイズ均等クラスタリングの
ファジィ化に関する研究

北島　慧

修士（工学）

（リスク工学専攻）

指導教員　遠藤　靖典

2019年3月

概要

クラスタリングの一手法として，最適化に基づくサイズ均等クラスタリング (ECBO) が提案されている．通常のクラスタリング手法はクラスタサイズに制約を設けていないため，クラスタサイズに大きな差があるような結果を返すことがある．それに対し，ECBO は全てのクラスタのサイズを均等とするように分割を行う．そのため，配送計画問題やタスク分配問題といったクラスタサイズが均等である分割が好ましい場合に有用である．しかし，ECBO で設定されているクラスタサイズの制約は非常に厳しいため，クラスタサイズをある程度揃えたい場合に不便であった．

この問題に対し，2 方面から問題の解決がなされた．1 つ目は，クラスタサイズの制約に幅を持たせる方法である．クラスタサイズの制約に幅を持たせ，それを調整することでクラスタサイズの制約の強さを調整できるようになる．このような発想から，最適化に基づくマージン付きサイズ均等クラスタリング (COCBO) が提案されている．もう 1 つの方法は，帰属度をファジィ化することである．この方法では，帰属度がより柔軟な値をとることができるため，ECBO と比較してクラスタサイズの制約を緩和することができる．この観点から，ファジィサイズ均等クラスタリング (FECBO) が提案された．

さて，上にあげたアルゴリズムはいずれも HCM や FCM に制約条件を加えた手法と見做すことができる．HCM や FCM は非常に多く利用されているアルゴリズムであるが，外れ値やノイズに対して強く影響を受けることが問題である．このような問題はサイズ均等クラスタリングにおいても同様に見受けられる．しかし，サイズ均等クラスタリングのロバスト性についての検討は十分になされていない．本研究では，これらのサイズ均等クラスタリングのロバスト性の検討を行うとともに，ノイズクラスタリングを用いた新たな手法の提案を行う．

また，HCM のもう一つの短所として，クラスタリング結果が初期値によって大きく異なることが挙げられる．初期値の選択方法を変えることで，局所解への収束を減らすことができれば，クラスタリング結果を安定して得られる，計算コストを削減できるというメリットがある．そこで，本研究では ECBO を用いて HCM の初期値依存の改善を行う手法を提案する．

目 次

第 1 章　序論 **1**
1.1 背景 1
1.2 目的 2
1.3 本論文の構成 3

第 2 章　主要なクラスタリング手法 **4**
2.1 クラスタリングについて 4
2.2 Hard c-means 6
2.3 Fuzzy c-means 7
2.4 ノイズクラスタリング 8
2.5 最適化に基づくサイズ均等クラスタリング 9

第 3 章　サイズ均等クラスタリングの発展 **10**
3.1 マージン付きサイズ均等クラスタリング 10
3.2 ファジィサイズ均等クラスタリング 11

第 4 章　サイズ均等クラスタリングのロバスト性についての検討 **13**
4.1 クラスタリングアルゴリズムのノイズに対するロバスト性について 13
4.2 既存のアルゴリズム 13
4.2.1 Medoid COCBO 13
4.2.2 L_1-COCBO 14
4.3 提案手法 16
4.3.1 COntrolled-sized Noise Clustering 16
4.3.2 Fuzzy Even-sized Noise Clustering 16

第 5 章　サイズ均等クラスタリングを用いた HCM の初期値依存の改善 **18**
5.1 クラスタリングアルゴリズムの初期値依存性について 18
5.2 k-means++ 18
5.3 サイズ均等クラスタリングに基づく HCM の初期値設定法 19

第 6 章　数値例 **20**
6.1 Adjusted Rand Index 20
6.2 ロバスト性の評価 21

6.2.1 外れ値を含む 2 クラスのデータ 21
6.2.2 一様ノイズを含む 2 クラスのデータ 24
6.2.3 ノイズを加えた Fisher's Iris Dataset 27
6.3 初期値依存性の評価 28
6.3.1 人工データ：クラスタ間距離が大きいデータ 28
6.3.2 人工データ：クラスタ間距離が小さいデータ 30
6.3.3 Fisher's Iris Dataset 31
6.3.4 Breast Cancer Wisconsin Data Set 32

第 7 章 結論 **34**

謝辞 **35**

参考文献 **36**

図 目 次

6.1 外れ値を含む 2 クラスのデータ 21
6.2 FENC による分類結果 ($\delta = 4.0$) 22
6.3 CONC による分類結果 ($\delta = 4.0$) 22
6.4 Medoid COCBO による分類結果 22
6.5 L_1-COCBO による分類結果 22
6.6 FECBO による分類結果 23
6.7 COCBO による分類結果 23
6.8 一様なノイズを含む 2 クラスのデータ 24
6.9 FENC による分類結果 ($\delta = 3.0$) 25
6.10 CONC による分類結果 ($\delta = 3.0$) 25
6.11 Medoid COCBO による分類結果 25
6.12 L_1-COCBO による分類結果 25
6.13 FECBO による分類結果 26
6.14 COCBO による分類結果 26
6.15 クラスタ間距離が大きいデータ 28
6.16 大域解への収束例 29
6.17 局所解への収束例 29
6.18 クラスタ間距離が小さいデータ 31

第1章　序論

1.1　背景

近年のスマートフォンの普及や Internet of Things (IoT) の浸透によってデータの収集，蓄積が非常に容易となった．また，Web 上の行動だけでなく，GPS を利用した位置情報やポイントカード，クレジットカードに紐づいた購買情報など，実空間における行動のデータについても収集，蓄積が盛んに行われている．これらのデータからなる大規模なデータはビッグデータと呼ばれ，様々なサービスを提供する基盤となっている．ストリーミング配信サービスにおけるレコメンドや，カーナビの所要時間の予測，渋滞予測などはビッグデータを用いて行われているサービスの代表といえる．

このようにデータが大規模化，複雑化してきたなかで，人間がデータを詳細に理解しその中から有用な知見を得ることは難しい．そこで，膨大なデータから自動的に有用な知見を取り出す，データマイニング技術が非常に注目を集めている．近年のデータを用いたサービスは，データを収集する基盤が整ったこと，それを扱うデータマイニング技術が進歩したことの 2 つから成り立っている．

クラスタリング [1] はそのようなデータマイニング手法の一つである．これは，データをクラスタと呼ばれる個体の集合に分割する手法であり，その特徴として，データの集合を外的基準を用いずに自動的に分割を行うことが挙げられる．代表的な手法として，Hard c-means (HCM)[2] がある．HCM は各個体とそれが所属するクラスタ中心との距離を最小化することで最適な分割を求める．このように，HCM をはじめ，多くのクラスタリングアルゴリズムが目的関数と制約条件からなる最適化問題として定式化されている [3, 4]．K-member クラスタリング (KMC) はクラスタサイズに注目した手法の一つである．KMC はデータセットをクラスタサイズが少なくとも K であるようなクラスターへ分割する．KMC は k-匿名化やタスク分配問題への応用が期待されている．greedy k-member clustering (GKC)[5] や one-pass k-means (OKA)[6]，clusteirng based k anonymity (CBK)[7] が KMC の代表的な手法であるが，これらの手法は目的関数の最適化に基づく手法でないため，分類精度があまり高くない．

そこで，目的関数に基づく手法として，最適化に基づくサイズ均等クラスタリング (ECBO)[8] が提案された．ECBO で考慮されている制約条件は各クラスタのサイズが K 以上かつ $K+1$ 以下であることである．ECBO は HCM にクラスタサイズの制約を加えた手法であり，帰属度とクラスタ中心の更新を繰り返すことで分割を行う．

しかしながら，ECBO で設定されているクラスタサイズの制約が厳しいため，ECBO はクラスタサイズを厳密に均等にする必要がない際には不便である．例えば 100 個の個体を含むデータの分割を行うとき，ECBO によって分割を行うと，クラスタサイズが 33 個または 34

個となるような分割が行われる．しかし，クラスタサイズをおよそ 30 個とする分割で充分である場合も存在する．ECBO ではそのような分割を行うことは不可能であるが，このクラスタサイズの制約を緩和することによって，より幅広い場面にサイズ均等クラスタリングを適用できる．

この問題を解決するため，主に 2 つの方法が考えられる．1 つ目は，クラスタサイズの制約に幅を持たせる方法である．クラスタサイズの制約に幅を持たせ，それを調整することで，クラスタサイズの制約の強さを調整できるようになる．このような発想から，最適化に基づくマージン付きサイズ均等クラスタリング (COCBO)[9] が提案されている．もう 1 つの方法は，帰属度をファジィ化することである．この方法では，帰属度がより柔軟な値をとることができるため，ECBO と比較してクラスタサイズの制約を緩和することができる．この観点から，ファジィサイズ均等クラスタリング (FECBO)[10] が提案された．これらの手法によって，ECBO のクラスタサイズに関する制約条件の緩和がなされた．

さて，上にあげた 3 つのサイズ均等クラスタリングはいずれも HCM やそのファジィ化である FCM に制約条件を加えた手法と見做すことができる．HCM や FCM は非常に多く利用されているアルゴリズムであるが，外れ値やノイズに対して強く影響を受けることが問題である．これらの手法はクラスタ中心に重心を用いているため，分割を行う対象のデータに外れ値が含まれていると，その外れ値がクラスタの重心に影響を与える．その結果，クラスタ中心が外れ値に近い位置に設定される．このような問題はサイズ均等クラスタリングにおいても同様に見受けられる．しかし，サイズ均等クラスタリングの外れ値に対するロバスト性についての検討は十分になされていない．

また，HCM のもう一つの短所として，クラスタリング結果が初期値によって大きく異なることが挙げられる．HCM は目的関数の最小化を行うが，初期値によっては局所解へと収束してしまうためである．この問題を回避するため，異なる初期値を何通りか与えてクラスタリングを行い，その中で最良の結果を採用するという手続きをとらなければならない．初期値依存の改善には，クラスタリング結果を安定して得られる，計算コストを削減できるというメリットがある．

1.2 目的

本論文では，サイズ均等クラスタリングについて，ロバスト性についての検討を行うとともに，問題に対する手法の提案を行う．サイズ均等クラスタリングにおいて，ノイズに対するロバスト性を示す方法として，Medoid ECBO[11]，L_1-ECBO[11]，Medoid COCBO[12]，L_1-COCBO[12] 等が提案されている．一方で，提案手法ではノイズに対処する新たなアプローチとして，ノイズクラスタリングの手法を取り入れる．ノイズクラスタリングは，ノイズの可能性が高いオブジェクトをクラスタに帰属させないという特徴を持つ．

また，サイズ均等クラスタリングの応用として，HCM の初期値設定に ECBO を利用した手法を提案する．HCM の初期値依存は広く知られている問題であり，k-means++[13] 等，様々な手法が提案されてきた．ECBO はクラスタサイズが同じであるような分割を行うため，デー

タ内の個体の密度にあまり偏りがない場合，偏りのない初期値を与えることが期待できる．

1.3 本論文の構成

本論文では，第 2 章において関連するクラスタリング手法及び最適化に基づくサイズ均等クラスタリングについて述べる．第 3 章では ECBO の発展手法，主に COCBO と FECBO に関して述べる．第 4 章でこれらのサイズ均等クラスタリングにおけるロバスト性を改善するため，ノイズクラスタリングの手法を導入した手法の提案を行う．第 5 章では HCM の初期値依存を改善するため，HCM の初期値設定に ECBO を用いた手法の提案を行う．第 6 章では数値実験によって，手法の評価及び考察を行う．最後に第 7 章において，本研究で得られた結論を述べる．

第2章　主要なクラスタリング手法

本章では，まずクラスタリングについての概要を説明し，種々のアルゴリズムの紹介を行う．

2.1　クラスタリングについて

クラスタリングとは，データを外的基準を用いずに自動的に分類を行う手法である．個体間に定義される類似度，非類似度を用いて分割を行う．

クラスタリングアルゴリズムはクリスプクラスタリングとファジィクラスタリングの二つに大別される．クリスプクラスタリングにおいて，ある個体は，クラスタに「属す」か「属さない」か 2 値のどちらかをとる．そのため，基本的には一つの個体はただ一つのクラスタに属することとなる．それと異なりファジィクラスタリングは，ファジィ理論を取り入れた手法である．ファジィクラスタリングは，あるオブジェクトがクラスタに所属する度合いに曖昧さを認め，0 から 1 の連続値を用いてクラスタに所属する度合いを表現する．本論文中の ECBO，COCBO はクリスプクラスタリングであり，FECBO はファジィクラスタリングである．

ファジィとクリスプという分類とは別に，クラスタリング手法は階層的手法と非階層的手法の 2 つに大別される．階層的手法は最も類似するクラスタを逐次結合するアルゴリズムである．この手法では，分割を行う前にクラスタ数を指定する必要がないこと，樹形図等を用いてクラスタが生成される過程を可視化できるといった利点がある．一方で，非階層的手法では，単にデータセットの分割を行う．非階層的手法は一般に，階層的手法と比較して計算量が少ないという特徴を持つ．

本論文中で扱う紹介するアルゴリズムは非階層的手法であり，さらに目的関数の最適化に基づく手法である．そのため，以降の手法の紹介においては目的関数及び制約条件と，アルゴリズムを中心とした説明を行う．また，本論文中での記号の定義を以下に示す．

- $x_k \in \Re^p \ (k = 1, \ldots, n)$: クラスタリング対象の個体
- $X = \{x_1, \ldots, x_n\}$: 個体の集合
- n : データセット X に含まれる個体数
- K : クラスタサイズ
- c : クラスタ数

- $C_i\ (i = 1, \ldots, c)$: i 番目のクラスタ
- $v_i \in \Re^p\ (i = 1, \ldots, c)$: クラスタ C_i の中心
- $V = \{v_1, \ldots, v_c\}$: クラスタ中心の集合
- u_{ki} : x_k のクラスタ C_i への帰属度
- $U = \{u_{11}, \ldots, u_{nc}\}$: 帰属度の集合

2.2 Hard c-means

本節では最も基本的なクラスタリングアルゴリズムである Hard c-means (HCM)[2] について述べる．HCM において各個体はただ一つのクラスタに帰属するため，帰属度は $u_{ki} \in \{0, 1\}$ として表される．HCM では，各個体とそれが所属するクラスタ中心との距離の総和を最小にする分割を求める．そこで，目的関数 J_{HCM} を以下のように定義し，J_{HCM} の最小化を行う．

$$J_{\text{HCM}}(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki} d_{ki} \tag{2.1}$$

本章では，d_{ki} はユークリッド距離の自乗，即ち $d_{ki} = \|x_k - v_i\|^2$ を表す．この目的関数を U と V の交互最適化によって最小化することにより，最適なクラスタ分割を求める．また，U に関する制約として，

$$\sum_{i=1}^{c} u_{ki} = 1 \ (k = 1, \ldots, n) \tag{2.2}$$

を与える．この制約は任意の個体 x_k がただ一つのクラスタ C に帰属することを意味している．

各個体の帰属度の最適化とクラスタ中心の最適化を交互に行うことで，目的関数の最小化を行う．帰属度 U の最適解は，

$$u_{ki} = \begin{cases} 1 & (v_i = \arg\min\limits_{l} d_{kl}) \\ 0 & (\text{otherwise}) \end{cases} \tag{2.3}$$

となる．また，クラスタ中心 V の最適解は，以下の式で与えられる．

$$v_i = \frac{\sum_{k=1}^{n} u_{ki} x_k}{\sum_{k=1}^{n} u_{ki}} \tag{2.4}$$

Algorithm 1 に HCM のアルゴリズムを示す．

Algorithm 1 HCM

Step 1. 定数 c を設定する．

Step 2. ランダムに c 個のクラスタ中心を選択する．

Step 3. U の更新：(2.3) より U を求める．

Step 4. V の更新：(2.4) より V を求める．

Step 5. 収束すれば終了．そうでなければ Step 3 に戻る．

2.3 Fuzzy c-means

Fuzzy c-means (FCM)[14] は HCM をファジィ化した手法であり，ファジィクラスタリングの代表的手法である．ファジィクラスタリングは，帰属度にあいまいさを認める手法であり，これらの手法では帰属度は $u_{ki} \in [0,1]$ として表される．FCM では，HCM の目的関数を変更する必要がある．そこで，目的関数 J_{FCM} を以下のように定義し，J_{FCM} の最小化を行う．

$$J_{\text{FCM}}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m d_{ki} \tag{2.5}$$

ここで，m は $m > 1$ と定義されたファジィ化パラメータである．この目的関数を HCM と同様に U と V の交互最適化によって最小化することにより，最適なクラスタ分割を求める．U に関する制約として，

$$\sum_{i=1}^{c} u_{ki} = 1 \ (k = 1, \ldots, n) \tag{2.6}$$

を与える．この制約は任意の個体 x_k のそれぞれのクラスタ C に対する帰属度の総和が 1 であることを意味している．

HCM と同様に FCM においても目的関数の交互最適化によって分割を行う．U の最適化については，ラグランジュ未定乗数法を用いて，

$$u_{ki} = \frac{\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\frac{1}{d_{kj}}\right)^{\frac{1}{m-1}}} \tag{2.7}$$

と求めることができる．V の最適化については，J_{FCM} を v_i について偏微分することで求まる．

$$v_i = \frac{\sum_{k=1}^{n}(u_{ki})^m x_k}{\sum_{k=1}^{n}(u_{ki})^m} \tag{2.8}$$

Algorithm 2 に FCM のアルゴリズムを示す．

Algorithm 2 FCM

Step 1. 定数 c を設定する．

Step 2. ランダムに c 個のクラスタ中心を選択する．

Step 3. U の更新：(2.7) より U を求める．

Step 4. V の更新：(2.8) より V を求める．

Step 5. 収束すれば終了．そうでなければ Step 3 に戻る．

2.4 ノイズクラスタリング

ノイズクラスタリング [15] とは，データにノイズが含まれていることを考慮し，ノイズの影響を除くことを目的に作られたアルゴリズムである．様々なアルゴリズムが提案されているが，ここでは Dave らによって提案されたアルゴリズムを用いる．このアルゴリズムは分割を行うクラスタに加えてノイズクラスタを持ち，ノイズである可能性が高いオブジェクトをノイズクラスタに分類する．ここで，分割を行うクラスタを $C_i\ (i = 1, \ldots, c)$，ノイズクラスタを $C_i\ (i = 0)$ とする．また，パラメータ δ を設定する．δ はノイズの識別を行うパラメータであり，$d_{k0} = \delta$ とする．即ち

$$\min_{i=1,\ldots,c} d_{ki} > \delta \tag{2.9}$$

である場合，x_k はノイズクラスタとして分類される．この δ を用いて，ノイズクラスタリングの目的関数は，

$$J_{\text{NOISE}}(U, V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m d_{ki} + \sum_{k=1}^{n}(u_{k0})^m \delta^2 \tag{2.10}$$

と表せる．

δ は定数であるため，(2.10) は FCM と同様に解くことができる．即ち，U の最適化については，

$$u_{ki} = \sum_{j=0}^{c}\left\{\left(\frac{d_{ki}}{d_{kj}}\right)^{\frac{1}{m-1}}\right\}^{-1} \tag{2.11}$$

と求められ，V の最適化については，

$$v_i = \frac{\sum_{k=1}^{n}(u_{ki})^m x_k}{\sum_{k=1}^{n}(u_{ki})^m}\ (i = 1, \ldots, c) \tag{2.12}$$

と求まる．

Dave らによるノイズクラスタリングのアルゴリズムを Algorithm 3 に示す．

Algorithm 3 ノイズクラスタリング

Step 1. 定数 c を設定する．δ を設定する．

Step 2. ランダムに c 個のクラスタ中心を選択する．

Step 3. U の更新：(2.11) により U を求める．

Step 4. V の更新：(2.12) により V を求める．

Step 5. 収束したら終了．そうでなければ Step 3 に戻る．

2.5 最適化に基づくサイズ均等クラスタリング

最適化に基づくサイズ均等クラスタリング (ECBO)[8] はクラスタサイズに注目したアルゴリズムの一つであり，クラスタサイズが均等となるような分割を行う．ECBO の目的関数は HCM のものと同様であるが，制約条件にクラスタサイズを K 以上 $K+1$ 以下にするという制約式を追加することでサイズの制約を実現する．したがってクラスタリングは (2.13)-(2.16) の最適化問題として定式化される．

$$\text{minimize} \quad J_{\text{ECBO}}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}d_{ki} \tag{2.13}$$

$$\text{s.t.} \sum_{i=1}^{c} u_{ki} = 1 \ (k = 1,\ldots,n) \tag{2.14}$$

$$\sum_{k=1}^{n} u_{ki} \geq K \ (i = 1,\ldots,c) \tag{2.15}$$

$$\sum_{k=1}^{n} u_{ki} \leq K+1 \ (i = 1,\ldots,c) \tag{2.16}$$

クラスタ数 c を設定した場合，クラスタサイズ K は (2.17) のように設定される．

$$K = \lfloor n/c \rfloor \tag{2.17}$$

ECBO の目的関数及び制約式は線形であるため，U の最適化はシンプレックス法を用いて解くことができる．また，V の最適化については HCM と同様に，クラスタの重心を用いる．Algorithm 4 に ECBO のアルゴリズムを示す．

Algorithm 4 ECBO

Step 1. 定数 K, c を設定する．

Step 2. ランダムに c 個のクラスタ中心を選択する．

Step 3. U の更新：シンプレックス法により U を求める．

Step 4. V の更新：(2.4) より V を求める．

Step 5. 収束したら終了．そうでなければ Step 3 に戻る．

第3章　サイズ均等クラスタリングの発展

本章では，ECBO の発展手法としてマージン付きサイズ均等クラスタリングアルゴリズムとファジィサイズ均等クラスタリングアルゴリズムの説明を行う．ECBO において，サイズの制約が非常に厳しいため，不自然な分割を得る場合があった．それを解決する方法として，クラスタサイズの制約に幅を持たせる方法，帰属度をファジィ化する方法の 2 つが考えられる．マージン付きサイズ均等クラスタリングは，前者の考えに基づいて構築されており，クラスタサイズ制約の強さをマージンを使い制御できる．また，ファジィサイズ均等クラスタリングは後者の方法を用いており，ECBO のファジィ化を行うことでサイズ制約の緩和を実現した．

3.1　マージン付きサイズ均等クラスタリング

マージン付きサイズ均等クラスタリング (COCBO)[9] は，ECBO のサイズ制約にマージンを持たせることで，制約の強さを制御できるようにしたアルゴリズムである．ECBO は各クラスタのサイズが K 以上 $K+1$ 以下であるというクラスタサイズの制約を持っていた．一方で，COCBO では各クラスタのサイズは $K-\alpha'$ 以上 $K+\alpha$ 以下と設定される．ここで，K, α', α は以下の式を満たすものとする．

$$K = \lfloor n/c \rfloor \tag{3.1}$$

$$\alpha' \geq 0 \tag{3.2}$$

$$\alpha \geq 1 \tag{3.3}$$

したがって，COCBO の目的関数および制約条件は (3.4)-(3.7) で表される．

$$\text{minimize} \quad J_{\text{COCBO}}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} d_{ki} \tag{3.4}$$

$$\text{s.t.} \sum_{i=1}^{c} u_{ki} = 1 \ (k = 1, \ldots, n) \tag{3.5}$$

$$\sum_{k=1}^{n} u_{ki} \geq K - \alpha' \ (i = 1, \ldots, c) \tag{3.6}$$

$$\sum_{k=1}^{n} u_{ki} \leq K + \alpha \ (i = 1, \ldots, c) \tag{3.7}$$

ここでECBOと同様にCOCBOの目的関数及び制約式はUについて線形であるため，Uの最適化はシンプレックス法を用いて解くことができる．また，Vの最適化についてはクラスタの重心を用いる．Algorithm 5にCOCBOのアルゴリズムを示す．

Algorithm 5 COCBO

Step 1. 定数c, Kを設定する．

Step 2. マージンα, α'を設定する．

Step 3. ランダムにc個のクラスタ中心を選択する．

Step 4. Uの更新：シンプレックス法によりUを求める．

Step 5. Vの更新：(2.4)よりVを求める．

Step 6. 収束したら終了．そうでなければStep 4に戻る．

3.2 ファジィサイズ均等クラスタリング

ファジィサイズ均等クラスタリング(FECBO)[10]はECBOをファジィ化したアルゴリズムである．帰属度を$\{0, 1\}$のように2値で表現するのでなく，連続値$[0, 1]$とすることで，より柔軟な帰属度の表現が可能となる．その結果，クラスタサイズの制約が緩和されることを期待したものである．FECBOはECBOのファジィ化とも，FCMにサイズ制約を付加した手法とも見做すことができる．

FECBOにおいては全てのクラスタのサイズがKであるような分割を行う．Kは(3.8)のように設定される．

$$K = n/c \tag{3.8}$$

ECBO及びCOCBOにおいて，クラスタサイズはそれぞれのクラスタに含まれる個体の数と定義されている．しかしFECBOにおいて，帰属度は連続値であるため同様の定義を用いることができない．そこでFECBOにおいては，クラスタサイズを「それぞれのクラスタに対する帰属度の総和」と定義する．そのため，Kは整数であるとは限らないことに留意する．FECBOの目的関数と制約条件は以下の(3.9)-(3.11)で表される．

$$\text{minimize} \quad J_{\text{FECBO}}(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ki})^2 d_{ki} \tag{3.9}$$

$$\text{s.t.} \sum_{i=1}^{c} u_{ki} = 1 \quad (k = 1, \ldots, n) \tag{3.10}$$

$$\sum_{k=1}^{n} u_{ki} = K \quad (i = 1, \ldots, c) \tag{3.11}$$

FECBO の帰属度行列 U を求める問題は凸 2 次計画問題となり，主双対パス追跡法 [16] によって解くことができる．また，V の最適化についてはクラスタの重心を用いる．FECBO のアルゴリズムを Algorithm 6 に示す．

Algorithm 6 FECBO

Step 1. 定数 K, c を設定する．

Step 2. ランダムに c 個のクラスタ中心を選択する．

Step 3. U の更新：主双対パス追跡法により U を求める．

Step 4. V の更新：クラスタ中心を求め，新たな V とする．

Step 5. 収束したら終了．そうでなければ Step 3 に戻る．

第4章　サイズ均等クラスタリングのロバスト性についての検討

4.1　クラスタリングアルゴリズムのノイズに対するロバスト性について

HCM や FCM をはじめとするクラスタリングアルゴリズムは，クラスタ中心にクラスタの重心を用いている．そのため，データが外れ値やノイズを含む場合，クラスタ中心がノイズや外れ値に大きく影響を受けるという問題がある．この問題への対応が長年にわたり考えられてきた．外れ値やノイズに対する頑健性を示す方法は 2 つに大別される．

一つ目は，クラスタ中心及び類似度の算出において外れ値の影響を軽減する方法である．k-medoids 法 [17] や，L_1 ノルムを用いたクラスタリング [18] がこの例として挙げられる．これらの手法では，ノイズはいずれかのクラスタに帰属するものの，外れ値への頑健性が高いことが示されている．遠藤らはこの手法をサイズ均等クラスタリングへ応用し，外れ値の影響を軽減することを示した．

二つ目は，ノイズの可能性が高いオブジェクトをノイズクラスタとして分類する手法である．この手法は，上の手法と比較して，ノイズの可能性が高いオブジェクトをクラスタに帰属させないという特徴を持つ．

さて，本論文中で扱うサイズ均等クラスタリングは HCM や FCM と同様に，クラスタ中心に重心を利用している．そのため，サイズ均等クラスタリングにおいてもノイズや外れ値への対応が必要である．本章では，サイズ均等クラスタリングのロバスト性を改善した手法として Medoid COCBO，L_1-COCBO を取り上げる．また，ノイズクラスタリングの手法を導入した新たな手法の提案を行う．

4.2　既存のアルゴリズム

本節では，ノイズの影響を考慮したサイズ均等クラスタリングとして，Medoid COCBO 及び L_1- COCBO について説明する．

4.2.1　Medoid COCBO

クラスタリング結果がノイズの影響を強く受ける要因はクラスタ中心に重心を用いているためだと前節で述べた．そこで，k-medoids 法ではクラスタ中心に距離の総和を最小とするようなクラスタ中の個体を用いる．これをサイズ均等クラスタリングに適用したものが Medoid

COCBO である．非類似度として，k-medoids 法と同様にユークリッド距離を用いる．したがって，Medoid COCBO の目的関数および制約条件は (4.1)-(4.4) で表される．

$$\text{minimize} \quad J_{\text{Medoid COCBO}}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}\|x_k - v_i\| \tag{4.1}$$

$$\text{s.t.} \sum_{i=1}^{c} u_{ki} = 1 \ (k = 1,\ldots,n) \tag{4.2}$$

$$\sum_{k=1}^{n} u_{ki} \geq K - \alpha' \ (i = 1,\ldots,c) \tag{4.3}$$

$$\sum_{k=1}^{n} u_{ki} \leq K + \alpha \ (i = 1,\ldots,c) \tag{4.4}$$

Algorithm 7 に Medoid COCBO のアルゴリズムを示す．

Algorithm 7 Medoid COCBO

Step 1. 定数 K, c を設定する．

Step 2. マージン α, α' を設定する．

Step 3. ランダムに c 個のクラスタ中心を選択する．

Step 4. U の更新：シンプレックス法により U を求める．

Step 5. V の更新：各クラスタ内で距離の総和を最小とする個体を求め，新たなクラスタ中心とする．

Step 6. 収束したら終了．そうでなければ Step 4 に戻る．

4.2.2 L_1-COCBO

k-means 等のアルゴリズムでは，非類似度として個体間のユークリッド距離の自乗を利用している．それに対し，L_1 ノルムを非類似度として利用すると，外れ値の影響を受けづらくなるということが知られている．そこで，COCBO の非類似度として，L_1 ノルムを用いた L_1-COCBO が提案された．制約条件は COCBO と同様であり，非類似度として L_1 ノルムを

用いるため，以下の式で定式化される．

$$\text{minimize} \quad J_{L_1 COCBO}(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}\|x_k - v_i\|_1 \tag{4.5}$$

$$\text{s.t.} \sum_{i=1}^{c} u_{ki} = 1 \ (k = 1,\ldots,n) \tag{4.6}$$

$$\sum_{k=1}^{n} u_{ki} \geq K - \alpha' \ (i = 1,\ldots,c) \tag{4.7}$$

$$\sum_{k=1}^{n} u_{ki} \leq K + \alpha \ (i = 1,\ldots,c) \tag{4.8}$$

ここで，U の最適化については COCBO と同様にシンプレックス法で求めることができる．一方，V の最適化については各クラスタの各成分ごとに最適化を行う．具体的には，各成分毎に Algorithm 8 の手続きに従って導出する．

Algorithm 8 最適な V の導出

Step1. 個体 x_{kj} を各成分 j ごとに昇順でソートし，$x_{q(k)j}$ とする．

Step2. $S = -\frac{1}{2}\sum_{k=1}^{n}(u_{q(k)j})^m, r = 0$ とする．

Step3. S が負から正となるまで $S = S + (u_{q(k)j}), r = r + 1$ を計算する．

Step4. Step 3 で得られた r を用いて最適解 $v_{ij} = x_{q(r)j}$ とする．

Algorithm 9 に L_1-COCBO のアルゴリズムを示す．

Algorithm 9 L_1-COCBO

Step 1. 定数 K, c を設定する．

Step 2. マージン α, α' を設定する．

Step 3. ランダムに c 個のクラスタ中心を選択する．

Step 4. U の更新：シンプレックス法により U を求める．

Step 5. V の更新：Algorithm 8 によって V を求める．

Step 6. 収束したら終了．そうでなければ Step 4 に戻る．

4.3 提案手法

従来手法はクラスタ中心及び類似度の計算法を変更することで外れ値の影響を軽減していた．これに対し，ノイズクラスタリングでは，ノイズの可能性が高い個体をノイズクラスタに分類することでその影響を軽減する．ノイズクラスタリングは外れ値だけでなく，一様に分布するようなノイズに強いという特徴を持つ．本論文ではノイズクラスタリング手法をサイズ均等クラスタリングに取り入れた 2 つの手法の提案を行う．提案手法はノイズ識別フェーズとクラスタリングフェーズの 2 フェーズで構成される．ノイズ識別フェーズでは，ノイズクラスタリングを用いてノイズクラスタと非ノイズクラスタに分割を行う．クラスタリングフェーズでは，非ノイズクラスタに対してサイズ均等クラスタリングを用いた分割を行う．

4.3.1 COntrolled-sized Noise Clustering

COntrolled-sized Noise Clustering (CONC) はクラスタリングフェーズに COCBO を用いたアルゴリズムである．最初にノイズ識別フェーズでノイズクラスタリングによるノイズ識別を行う．得られた結果を $C_i\ (i = 0, \ldots, c)$ とすると，非ノイズクラスタは，$C_i\ (i = 1, \ldots, c)$ と表せる．この非ノイズクラスタに対して，クラスタリングフェーズでサイズが均等となるよう，再度分割を行う．このとき，クラスタリングフェーズで用いるデータセットを X' とすると，X' は

$$X' = \{x | x \in C_i\ (1, \ldots, c)\} \tag{4.9}$$

と表せる．また，X' に含まれるデータを $x_k\ (k = 1, \ldots, n')$ とする．したがって，CONC の目的関数および制約条件は (4.10)-(4.13) で表される．

$$\text{minimize} \quad J_{\text{CONC}}(U, V) = \sum_{k=1}^{n'} \sum_{i=1}^{c} u_{ki} d_{ki} \tag{4.10}$$

$$\text{s.t.} \sum_{i=1}^{c} u_{ki} = 1\ (k = 1, \ldots, n') \tag{4.11}$$

$$\sum_{k=1}^{n'} u_{ki} \geq K - \alpha'\ (i = 1, \ldots, c) \tag{4.12}$$

$$\sum_{k=1}^{n'} u_{ki} \leq K + \alpha\ (i = 1, \ldots, c) \tag{4.13}$$

提案手法のアルゴリズムを Algorithm 10 に示す．

4.3.2 Fuzzy Even-sized Noise Clustering

Fuzzy Even-sized Noise Clustering (FENC) はクラスタリングフェーズに FECBO を用いたアルゴリズムである．CONC と同様に，最初にノイズ識別フェーズでノイズクラスタリングによるノイズ識別を行う．得られた非ノイズクラスタに対して FECBO を用いる．

Algorithm 10 CONC

Step 1. 定数 c, δ を設定する．

Step 2. マージン α, α' を設定する．

Step 3. ノイズクラスタリングアルゴリズムによってノイズ識別を行う．

Step 4. クラスタサイズ K を $K = \lfloor n'/c \rfloor$ と設定する．

Step 5. c 個のクラスタ中心を選択する．

Step 6. U の更新：シンプレックス法により U を求める．

Step 7. V の更新：クラスタ中心を求め，新たな V とする．

Step 8. 収束したら終了．そうでなければ Step 6 に戻る．

FENC の目的関数と制約条件は以下の (4.14)-(4.16) で表せる．

$$\text{minimize} \quad J_{\text{FENC}}(U, V) = \sum_{k=1}^{n'} \sum_{i=1}^{c} (u_{ki})^2 d_{ki} \tag{4.14}$$

$$\text{s.t.} \sum_{i=1}^{c} u_{ki} = 1 \quad (k = 1, \ldots, n') \tag{4.15}$$

$$\sum_{k=1}^{n'} u_{ki} = K \quad (i = 1, \ldots, c) \tag{4.16}$$

提案手法のアルゴリズムを Algorithm11 に示す．

Algorithm 11 FENC

Step 1. 定数 c, δ を設定する．

Step 2. ノイズクラスタリングアルゴリズムによってノイズ識別を行う．

Step 3. クラスタサイズ K を $K = n'/c$ と設定する．

Step 4. c 個のクラスタ中心を選択する．

Step 5. U の更新：主双対パス追跡法により U を求める．

Step 6. V の更新：クラスタ中心を求め，新たな V とする．

Step 7. 収束したら終了．そうでなければ Step 5 に戻る．

第5章　サイズ均等クラスタリングを用いたHCMの初期値依存の改善

5.1　クラスタリングアルゴリズムの初期値依存性について

代表的なクラスタリングアルゴリズムであるHCMやFCMは目的関数を設定し，その目的関数をクラスタ中心と帰属度の交互最適化を行うことで分割を行う．クラスタ中心の更新の際は，クラスタに割り当てられたすべての点の重心がクラスタ中心となり，帰属度の更新の際は，データ点から最も近いクラスタ中心にクラスタの割り当てを行う．これらの2ステップを収束するまで繰り返すことで分割を行う．

この手続きを行う際に，初期値の設定が必要となる．一般的な手法では，初期値としてランダムに初期クラスタ中心を設定する．HCMやFCMは，この初期クラスタ中心から目的関数を単調減少させるようなアルゴリズムである．ランダムに決定された初期状態から解の探索を行うため，初期値によって分割結果が異なる．

このような問題は以前から研究されており，初期値選択にクラスタ中心からの距離に基づいた確率密度を用いたk-means++等が提案されている．k-means++では，既存のクラスタ中心から遠いデータ点が次のクラスタ中心として選ばれる可能性が高い．このような初期値選択を行うことで，初期クラスタ中心間の距離が近くなることを防いでいる．このように，初期クラスタ中心の距離が大きければクラスタ間距離が大きいような分割結果が得られることが期待できる．

ところで，ECBOはクラスタサイズが同じであるような分割を行う．データ内に個体の密度の偏りがあまりない場合，クラスタ中心はある程度離れた値が得られることが期待できる．また，クラスタサイズの制約によって，大きくクラスタリング結果が変動することがないということも利点である．ある程度安定した結果が得られることが期待できる．以上のような理由より，本章では，HCMの初期値設定にECBOを使用する手法の提案を行う．

5.2　k-means++

k-means++[13]はHCMの初期値選択手法を改善した手法である．HCMが初期クラスタ中心をランダムに選択するのに対し，k-means++では，クラスタ中心からの距離に基づいた確率密度を用いて選択を行う．初期クラスタ中心が非常に近く選択されるような確率を小さくすることで，分類結果の改善を行う．手続きとしては，最初に1つのクラスタ中心をランダムに選択し，次のクラスタ中心の決定の際にクラスタ中心からの距離に基づいた確率密度を用

いる．ここで，あるオブジェクトからすでに選択された最も小さいクラスタ中心までの距離を $D(x)$ とする．次のクラスタ中心の算出は以下の確率に基づいて算出される．

$$\frac{D(x^2)}{\sum_{x \in X} D(x)^2} \tag{5.1}$$

このような初期値選択を行うことで，初期クラスタ中心の偏りを小さくしている．以下にk-means++のアルゴリズムを示す．

Algorithm 12 k-means++

Step 1. 1 つのクラスタ中心 v_1 をデータセット X 中からランダムに選択する．

Step 2. 新たなクラスタ中心 v_i をデータセット X 中から，(5.1) の確率に基づき選択する．

Step 3. c 個のクラスタ中心が求まるまで Step 2 を繰り返す．

Step 4. 通常の HCM を行う．

5.3 サイズ均等クラスタリングに基づく HCM の初期値設定法

本節では，HCM の初期値依存性の改善手法として，サイズ均等クラスタリングに基づく HCM の初期値設定法 (HCM+E) を提案する．HCM+E では，ECBO を HCM の初期値選択に利用する．ECBO はサイズの制約のため，クラスタリング結果が大きく変動しない，そのためある程度安定した初期値が得られることが期待できる．

Algorithm 13 にアルゴリズムを示す．

Algorithm 13 ECBO に基づく初期値選択法を用いた HCM

Step 1. 定数 c を設定する．

Step 2. ランダムに c 個のクラスタ中心を選択する．

Step 3. U の更新：シンプレックス法により U を求める．

Step 4. V の更新：(2.4) より V を求める．

Step 5. 収束したら終了．そうでなければ Step3 に戻る．

Step 6. 得られたクラスタ中心を初期クラスタ中心として通常の k-means を行う．

第6章　数値例

6.1　Adjusted Rand Index

クラスタリング手法の評価を行うため，Adjusted Rand Index (ARI)[19] という指標を用いる．ARI とは，2 つのクラスタリング間の類似性を測るもので，次のように定義される．

n 個のデータからなるデータセット X に対して，2 通りの手法でクラスタ分割を行う．この結果を，$G = \{g_1, \ldots, g_n\}, H = \{h_1, \ldots, h_n\}$ とする．

このデータセットから，2 つのデータを取り出し，2 通りの手法で同じクラスタに帰属しているかどうかを全てのデータのペアについて数え上げる．ここで，2 つのデータが G で同じクラスタに属し，かつ H でも同じクラスタに帰属しているペアの数を δ とする．同様に，G で同じクラスタに帰属しているペアの数を ϵ, H では同じクラスタに帰属している場合の数を ζ とする．ここで, ペアの数を数え上げると，${}_nC_2$ と表せる．このとき，ARI は以下の式によって示される．

$$\text{ARI} = \frac{\delta - (\epsilon\zeta/{}_nC_2)}{1/2(\epsilon + \zeta) - (\epsilon\zeta/{}_nC_2)} \tag{6.1}$$

ARI は，2 つのクラスタリングに相関がないときの値が 0, クラスタリング結果が一致するときに 1 をとる値である．

今回は各手法の性能を測るため，入力データの正解ラベルを G，各手法で得られたラベルを H として，ARI を用いて各手法の比較を行う．

6.2　ロバスト性の評価

本節では，第 4 章において提案を行った FENC および CONC について数値実験によって評価を行う．

6.2.1　外れ値を含む 2 クラスのデータ

実験に使用したデータは 2 つの円状のクラスタに 15 個ずつ外れ値を加えたものである．詳細を以下に示す．

- クラスタ 0：中心 (0.0, 0.0), 半径 1.0 の円内に 150 個の個体を配置
- クラスタ 1：中心 (3.0, 3.0), 半径 3.0 の円内に 100 個の個体を配置
- ノイズ：30 個のノイズを 2 つに分けて配置

この人工データを図 6.1 に示す．

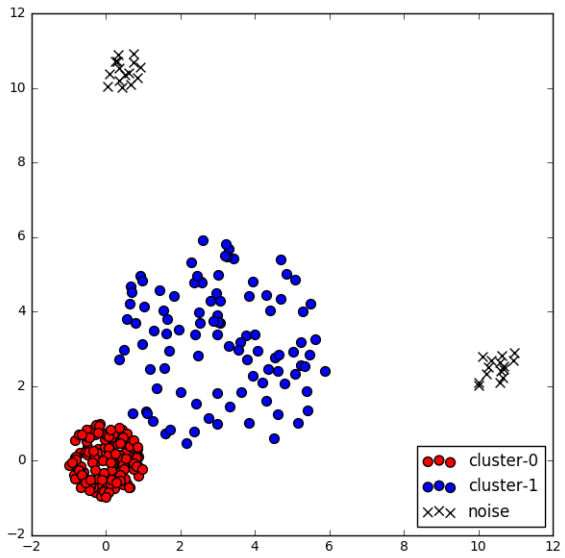

図 6.1: 外れ値を含む 2 クラスのデータ

この人工データに対して各手法で初期値を 100 回変えて，クラスタリングを行った．与えた条件は，$c = 2, \alpha = \alpha' = 25$ とした．

最も目的関数の値が小さくなった時の分割を図 6.2-6.7 に示す．また，各手法での分割結果及び ARI を表 6.1 に示す．

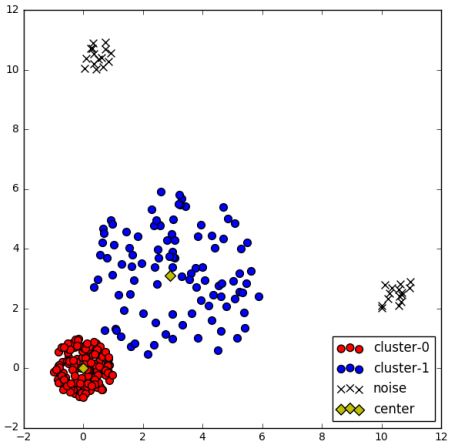

図 6.2: FENC による分類結果 ($\delta = 4.0$)

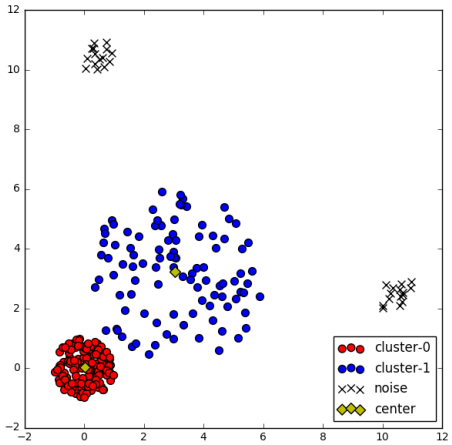

図 6.3: CONC による分類結果 ($\delta = 4.0$)

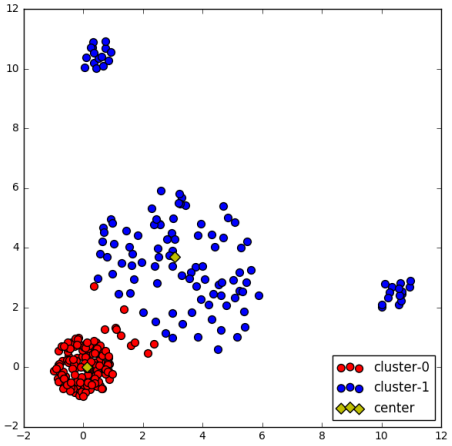

図 6.4: Medoid COCBO による分類結果

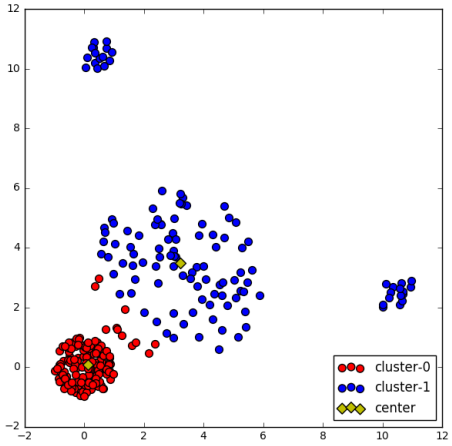

図 6.5: L_1-COCBO による分類結果

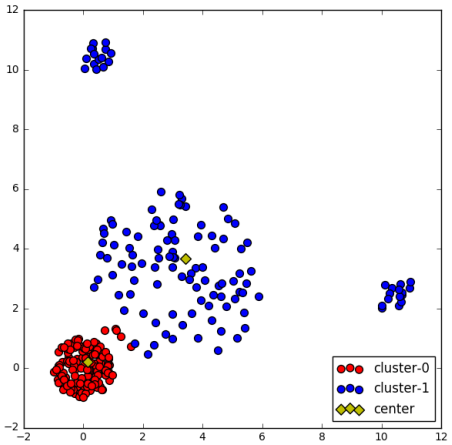

図 6.6: FECBO による分類結果

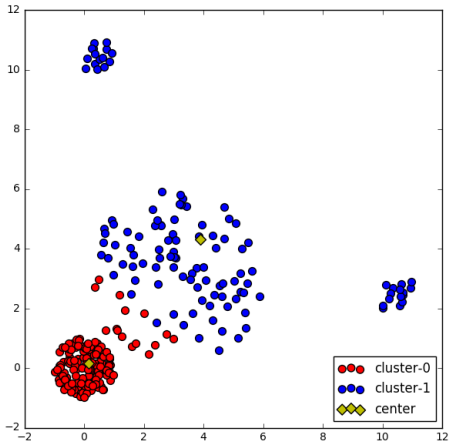

図 6.7: COCBO による分類結果

表 6.1: 外れ値を含む 2 クラスのデータに対する各手法の実行結果

手法	δ	分割結果	ノイズ数	ARI
FECBO	-	(155,125)	-	0.79
COCBO	-	(165,115)	-	0.69
Medoid COCBO	-	(160,120)	-	0.74
L_1-COCBO	-	(161,119)	-	0.73
FENC	3.00	(150,97)	33	0.98
FENC	4.00	(150,100)	30	1.00
FENC	7.00	(151,97)	22	0.94
FENC	10.0	(155,125)	0	0.79
CONC	3.00	(149,98)	33	0.97
CONC	4.00	(150,100)	30	1.00
CONC	7.00	(154,104)	22	0.90
CONC	10.0	(165,115)	0	0.69

COCBO による分割では，クラスタ 1 の中心が外れ値に引き寄せられ，分類境界も右上に位置している．従来手法の中で COCBO が最も強く外れ値の影響を受けたといえる．それと比較して従来手法である，L_1-COCBO，Medoid COCBO においては，ノイズの影響が軽減されていることが見て取れる．ARI も COCBO より大きく，より正解ラベルに近い分類が行えている．FECBO は外れ値を考慮した手法でないのにも関わらず，これらの手法より高い ARI を示した．外れ値がどちらのクラスタ中心からもある程度遠いため，ノイズである個体がど

ちらのクラスタにも大きな帰属度をとらなかったためだと考えられる．帰属度の柔軟な表現が外れ値に対しても有用であったことがわかる．

提案手法による分割では，δ を適切に設定できた時に正解ラベルと同じ分割を行うことができている．また，δ の値を大きくするにつれ，ノイズと判定されるデータの数が少なくなることがわかる．δ の値を非常に大きくとった時は，FENC は FECBO と同様の分類結果を示す．CONC についても同様に，COCBO と同様の分類結果を示す．δ の値を小さくとった時は，ノイズでない個体までノイズであると判定されてしまう．このため，δ はある程度大きな値をとることが望ましいと考えられる．

6.2.2 一様ノイズを含む 2 クラスのデータ

実験に使用したデータは 2 つの円状のクラスタに 50 個の一様ノイズを加えたものである．詳細を以下に示す．

- クラスタ 0：中心 (0.0, 0.0), 半径 1.0 の円内の円内に 150 個の個体を配置
- クラスタ 1：中心 (3.0, 3.0), 半径 3.0 の円内の円内に 100 個の個体を配置
- ノイズ：50 個のノイズをランダムに配置

この人工データを図 6.8 に示す．

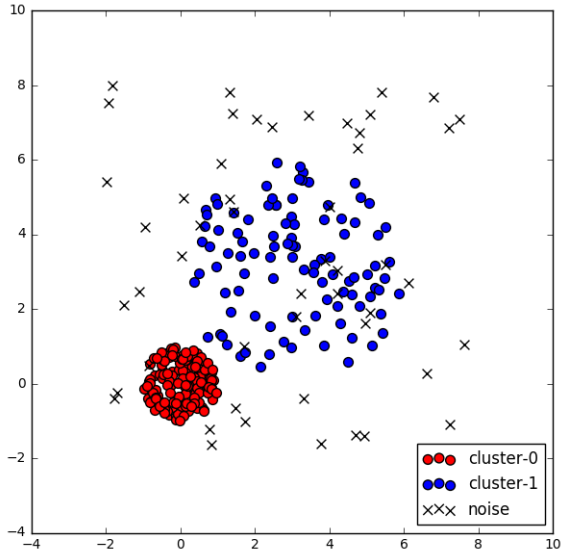

図 6.8: 一様なノイズを含む 2 クラスのデータ

この人工データに対して各手法で初期値を 100 回変えて，クラスタリングを行った．与えた条件は，$c = 2, \alpha = \alpha' = 25$ とした．

最も目的関数の値が小さくなった時の分割を図 6.9 - 6.14 に示す．また，各手法での分割結果及び ARI を表 6.2 に示す．

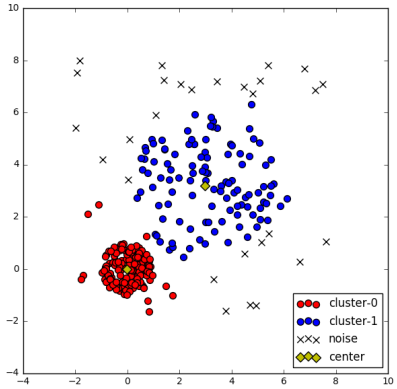

図 6.9: FENC による分類結果 ($\delta = 3.0$)

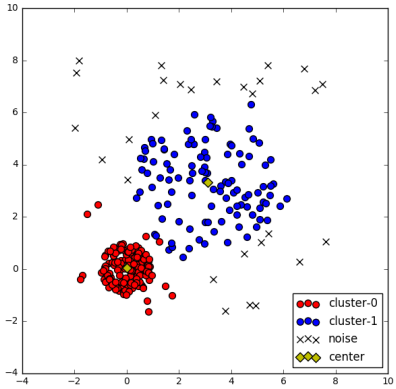

図 6.10: CONC による分類結果 ($\delta = 3.0$)

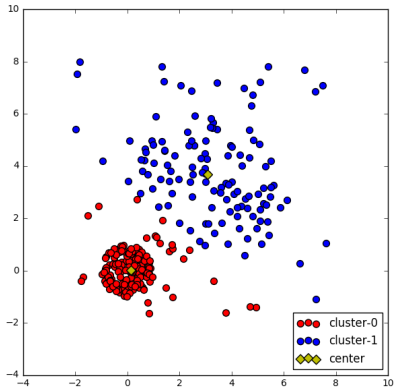

図 6.11: Medoid COCBO による分類結果

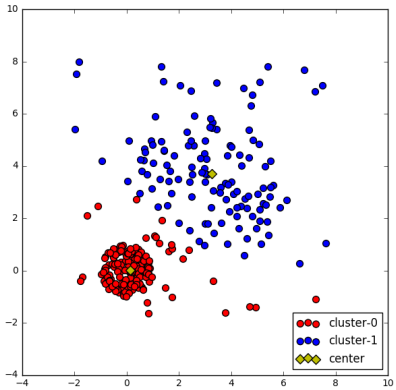

図 6.12: L_1-COCBO による分類結果

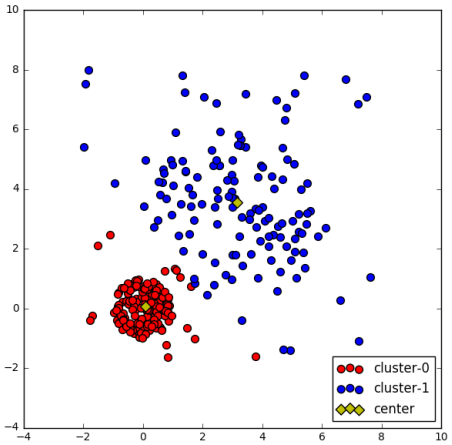

図 6.13: FECBO による分類結果

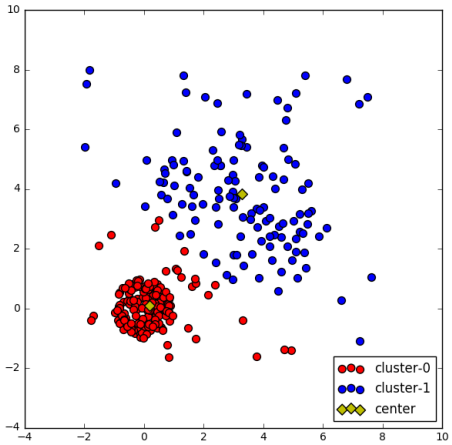

図 6.14: COCBO による分類結果

表 6.2: 一様なノイズを含む 2 クラスのデータに対する各手法の実行結果

手法	δ	分割結果	ノイズ数	ARI
FECBO	-	(165,135)	-	0.69
COCBO	-	(175,125)	-	0.62
Medoid COCBO	-	(160,120)	-	0.63
L_1-COCBO	-	(161,119)	-	0.62
FENC	3.00	(160,111)	29	0.81
FENC	4.00	(160,124)	16	0.78
FENC	5.00	(164,129)	7	0.72
FENC	10.0	(165,135)	0	0.69
CONC	3.00	(161,110)	29	0.80
CONC	4.00	(167,117)	16	0.71
CONC	5.00	(172,121)	7	0.66
CONC	10.0	(175,125)	0	0.62

前節の結果と同様に，従来手法では COCBO が最も大きくノイズの影響を受けた．L_1-COCBO，Medoid COCBO においても，COCBO と大きな差はみられなかった．これらの手法は外れ値の影響を小さくするが，一様なノイズには COCBO と同程度の影響を受けることがわかる．FECBO においては，一様ノイズを与えた場合にでも，L_1-COCBO，Medoid COCBO より ARI が大きくなった．外れ値に対する結果と同様，クラスタ中心から遠いデータの影響を小さくできている．

提案手法において，δ を適切に設定できた場合，ある程度ノイズの識別ができている．しかし，クラスタ中心からの距離に基づいてノイズの識別を行うため，クラスタ内に存在するノイズの識別を行うことができていない．前節と同様に δ を大きくするにつれ，元の手法の結果と同様の結果に収束していくことが確認できた．

6.2.3 ノイズを加えた Fisher's Iris Dataset

Fisher's Iris Dataset[20] はアヤメの品種の分類に関するデータである．3 種類のアヤメのがく辺の幅と長さ，花弁の幅の長さを測定した 4 次元のデータセットであり，それぞれの種類が 50 個体ずつ含まれている．これに一様なノイズを 30 個与えた．このデータセットに対して $c = 3, \alpha = \alpha' = 15$ と設定して 100 回クラスタリングを行い，分類結果の比較を行う．各手法での分割結果及び ARI を表 6.3 に示す．

表 6.3: ノイズを含む iris データに対する各手法の実行結果

手法	δ	分割結果	ノイズ数	Max(ARI)
FECBO	-	(60,63,57)	-	0.57
COCBO	-	(60,75,45)	-	0.54
Medoid COCBO	-	(60,69,51)	-	0.59
L_1-COCBO	-	(60,71,49)	-	0.55
FENC	2.0	(54,52,49)	25	0.78
FENC	3.0	(56,54,52)	18	0.70
CONC	2.0	(54,63,38)	25	0.72
CONC	3.0	(56,66,40)	18	0.65

従来手法では，Medoid COCBO の ARI が最も大きくなった．COCBO と比較して，Medoid COCBO，L_1-COCBO は分類結果が大きく異なる結果となった．提案手法では，$\delta = 2.0$ と設定したときに ARI が最も大きくなった．提案手法ではノイズの識別が行われたため，ARI が従来法と比較して大きくなっている．

6.3 初期値依存性の評価

提案したアルゴリズム (HCM+E) が HCM の初期値依存を改善できているかの評価を行う．本節では，提案手法に加え，HCM，k-means++でクラスタリングを行った結果を示す．

6.3.1 人工データ：クラスタ間距離が大きいデータ

100 個の個体数を持つクラスタを中心 (0.0, 0.0), (8.0, 0.0), (8.0, 0.0), (8.0, 8.0) として 4 つ配置したデータセットを用いて各手法でクラスタリングを行う．各個体の成分は標準正規分布に従う乱数より生成した．データセットを図 6.15 に示す．このデータセットに対し，各手法について $c = 4$ として初期値を 100 回変えて実験を行った．

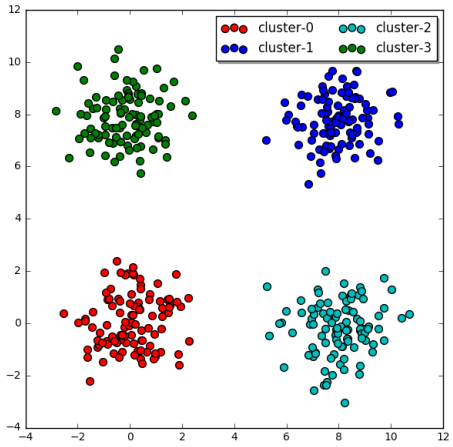

図 6.15: クラスタ間距離が大きいデータ

このデータセットは HCM の初期値によってクラスタリング結果が異なる．100 回の試行の中で，各手法とも図 6.16 のように大域解に収束した．一方で，初期値選択が上手くいかなかった場合，図 6.17 のような局所解へと収束してしまう．各手法の目的関数及び ARI の統計量を表 6.4 に示す．表中 J は目的関数の値を示している．

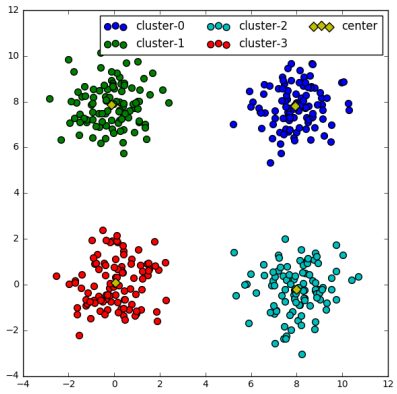

図 6.16: 大域解への収束例

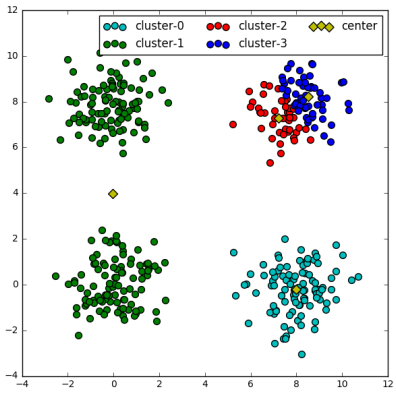

図 6.17: 局所解への収束例

表 6.4: クラスタ間距離が大きいデータの分割結果

	HCM+E	k-means++	HCM
Min(J)	7.86×10^2	7.86×10^2	7.86×10^2
Max(J)	7.86×10^2	3.98×10^3	3.98×10^3
Ave(J)	7.86×10^2	1.03×10^3	1.56×10^3
Var(J)	0.00	7.86×10^5	1.80×10^6
Min(ARI)	1.00	0.62	0.62
Max(ARI)	1.00	1.00	1.00
Ave(ARI)	1.00	0.97	0.91
Var(ARI)	0.00	0.010	0.026

各手法について，ARI の最大値は 1.00 となった，これは大域解へと収束したとき，正解ラベルと同じ分割を行うことができることを示す．HCM+E の結果では，すべての試行で大域解へと収束し，ARI の平均が 1 となった．ECBO のサイズの制約が働いた結果，初期クラスタ中心が各クラスタに割り当てられたためである．HCM 及び k-means++では J の分散が非常に大きくなってしまう．また，HCM による分割では平均 ARI が 0.91 と局所解へ多く収束したことがわかる．

さて，このデータセットは全てのクラスタサイズが 100 であり，ECBO による分割に適したデータセットである．そこで，図中右下，(0.0, 8.0) を中心としたクラスタ (サイズ調整クラスタ) に含まれるデータ数を 100 個ずつ増やして数値実験を行う．クラスタサイズが異なるようなデータセットに対して，提案手法の初期値依存性を検証するためである．クラスタサイズを変化させた際の平均 ARI 及び ARI の標準偏差を表 6.5-6.6 に示す．表中の size はサイズ調整クラスタのクラスタサイズを示している．クラスタサイズを 100 から 500 まで変化させ

て数値実験を行った．

表 6.5: クラスタ間距離が大きいデータの平均 ARI

size	100	200	300	400	500
HCM+E	1.00	1.00	0.82	0.75	0.52
k-means++	0.97	0.99	0.96	0.94	0.93
HCM	0.91	0.93	0.85	0.79	0.74

表 6.6: クラスタ間距離が大きいデータの ARI の標準偏差

size	100	200	300	400	500
HCM+E	0.000	0.000	0.210	0.228	0.003
k-means++	0.101	0.060	0.117	0.150	0.166
HCM	0.160	0.147	0.187	0.217	0.237

どの手法においてもサイズ調整クラスタのクラスタサイズを増やしていくにつれ，平均 ARI が小さくなっていく傾向がみられる．クラスタサイズが 200 までは HCM+E の平均 ARI は 1 であり，すべて大域解に収束する．しかし，クラスタサイズが 300 以上の時，平均 ARI が HCM よりも小さくなってしまう．これは，データ数が 600 以上となったことが問題であると考えられる．データ数が 600 の時，ECBO は各クラスタのサイズが 150 となるような分割を行う．そのため，サイズ調整クラスタ内に 2 つのクラスタ中心が集まりやすい状況が作られる．偏った初期クラスタ中心を多く与えてしまうため HCM+E の平均 ARI が HCM よりも小さくなってしまう．

6.3.2 人工データ：クラスタ間距離が小さいデータ

100 個の個体数を持つクラスタを中心 (0.0, 0.0), (0.0, 5.0), (5.0, 0.0), (5.0, 5.0) として 4 つ配置したデータセットを用いて各手法でクラスタリングを行う．各個体の成分は標準正規分布に従う乱数より生成した．データセットを図 6.18 に示す．

このデータについても，(0.0, 5.0) を中心としたクラスタ (サイズ調整クラスタ) に含まれるデータ数を 100 個ずつ増やして数値実験を行った．クラスタサイズを変化させた際の平均 ARI 及び ARI の標準偏差を表 6.7-6.8 に示す．

前節での結果と同様に，どの手法においてもサイズ調整クラスタのクラスタサイズを増やしていくにつれ，平均 ARI が小さくなっていく傾向がみられる．サイズ調整クラスタに初期値が偏ることが増えるためではないかと考えられる．また，このデータセットではクラスタサイズを変化させても，HCM+E の平均 ARI は HCM 及び k-means++より大きくなった．クラスタ間距離が小さいため，サイズ調整クラスタ内に初期値が複数存在しても，局所解に収束しなかったためだと考えられる．その結果，サイズ調整クラスタのクラスタサイズは HCM+E

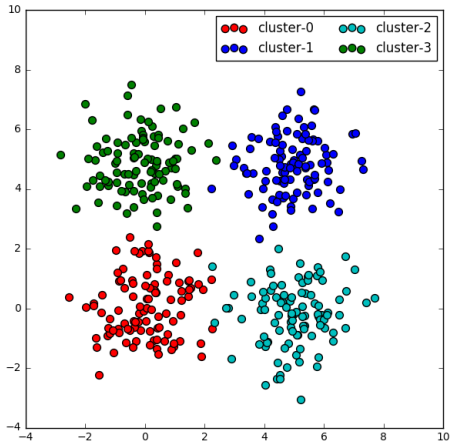

図 6.18: クラスタ間距離が小さいデータ

表 6.7: クラスタ間距離が小さいデータの平均 ARI

size	100	200	300	400	500
HCM+E	0.98	0.96	0.95	0.95	0.95
k-means++	0.98	0.95	0.95	0.93	0.94
HCM	0.98	0.95	0.95	0.91	0.93

の分割結果に大きく影響を与えなかった．HCM や HCM++が 100 回の試行のうち何度か局所解へと収束したのに対し，HCM+E はクラスタサイズが 400 の場合を除いて，100 回すべてで同じ値に収束した．HCM+E は HCM や k-means++と比較して，安定して正確な分割を行うことができている．

6.3.3 Fisher's Iris Dataset

Fisher's Iris Dataset について，各手法での分割を行った. 各手法について $c = 3$ として初期値を 100 回変えて実験を行った．実験結果を表 6.9 に示す.

HCM+E では 100 回とも同じ値に収束したのに対し，HCM 及び k-means++では目的関数の分散が大きい値を示した．提案手法が安定した結果を出していることがわかる．しかし，最大 ARI は HCM，k-means++よりも小さくなっている．HCM+E は，このデータセットについては大域解へと収束していないが，ある程度よい結果を安定して示した．

表 6.8: クラスタ間距離が小さいデータの ARI の標準偏差

size	100	200	300	400	500
HCM+E	0.00	0.00	0.00	6.32×10^{-2}	0.00
k-means++	3.85×10^{-2}	3.56×10^{-2}	2.44×10^{-3}	9.89×10^{-2}	7.86×10^{-2}
HCM	4.34×10^{-2}	4.83×10^{-2}	2.23×10^{-3}	6.32×10^{-1}	1.01×10^{-1}

表 6.9: Iris Data($c = 3$) の分類結果

	HCM+E	k-means++	HCM
Min(J)	78.95	78.94	78.94
Max(J)	78.95	145.28	145.28
Ave(J)	78.95	87.36	98.97
var(J)	0.00	474.29	893.32
min(ARI)	0.72	0.42	0.42
max(ARI)	0.72	0.73	0.73
ave(ARI)	0.72	0.68	0.63
var(ARI)	0.000	0.010	0.019

6.3.4 Breast Cancer Wisconsin Data Set

Breast Cancer Wisconsin Data Set[21] はウィスコンシン大学病院において，胸に腫瘍が見られた患者の検診結果をまとめたデータである．腫瘍の状態や患者の状態などの 9 次元の指標と，その患者の腫瘍が良性のもの (A) か悪性のもの (B) かの診断結果を含むデータである．オリジナルのデータセットは 699 個の個体からなるデータであるが，今回は重複する個体や欠損値を含む個体を削除した 449 個のデータセットに対して分割を行った．各手法について $c = 2$ として初期クラスタ中心を 100 回変えてクラスタリングを行った．結果を以下に示す．各手

表 6.10: Breast Cancer Wisconsin Data Set ($c = 2$) の分類結果

	提案手法	k-means++	HCM
Min(J)	179.34	179.34	179.34
Max(J)	179.34	179.34	179.34
Ave(J)	179.34	179.34	179.34
Var(J)	0.00	0.00	0.00
Min(ARI)	0.73	0.73	0.73
Max(ARI)	0.73	0.73	0.73
Ave(ARI)	0.73	0.73	0.73

法ともに 100 回の試行でほぼ同じ値に収束する結果となった．このデータセットでは，初期

値がクラスタリング結果にあまり関係ないことがわかる.

第7章　結論

本研究では，既存手法である FECBO 及び COCBO に対し，ノイズの影響を考慮したアルゴリズムである FENC 及び CONC の提案を行った．また，サイズ均等クラスタリングの応用例として，HCM の初期値の決定に ECBO を用いた手法を提案した．

COCBO にノイズクラスタを取り入れた CONC では，外れ値やノイズに対してロバスト性を持たせることができた．先行研究である L_1-COCBO や Medoid COCBO では一様分布のノイズに対応できないという問題があったが，CONC では，ノイズにも外れ値にもロバスト性を持つような手法であることを確認できた．

FECBO について，ロバスト性の検討を行った．FECBO についてはこれまでロバスト性の検討がなされておらず，関連した手法も提案されていない．数値実験の中で，クリスプなクラスタリングと比較して，外れ値に強いということが分かった．外れ値はいずれのクラスタ中心からも遠くなるため，特定のクラスタに対する帰属度が大きくなりづらいためだと考えられる．本研究では FECBO のノイズに対するロバスト性を向上させるため，ノイズクラスタリングの手法を取り入れたが，COCBO と同様に medoid をクラスタ中心として採用することや L_1 距離を非類似度として採用する手法でもロバスト性の改善を行うことが可能である．

また，サイズ均等クラスタリングの応用として，ECBO を HCM の初期値選択に利用する手法の提案を行った．クラスタ間の距離が大きい場合，クラスタサイズが提案手法の結果に大きく影響を与えることが分かった．クラスタ間の距離がそれほど大きくない場合には，クラスタサイズに関係なく，k-means++と同じかそれを上回る結果を安定して得ることができた．今回は ECBO を用いて初期値選択を行ったが，FECBO 等を用いて初期値選択を行うことも考えられる．

本研究ではサイズ均等クラスタリングのロバスト性，初期値依存性に着目して手法の提案を行った．今後の研究として，FECBO にカーネルを導入し，線形分離ではうまく分割できないデータに対応することや FECBO にマージンを与え，制約の強さを制御できるようにすることが挙げられる．

謝辞

本論文の作成にあたり，指導教員の遠藤靖典教授には，工学システム学類の 1 年次から 6 年間に渡りお世話になりました，心より感謝致します．研究指導のみでなく，学習指導や就職活動等，多岐に渡り丁寧なご指導を頂きました．

また，ソフトコンピューティング基礎グループの高安亮紀助教には，発表の指導や達成度評価，中間発表で研究に関する助言を頂きました．深く感謝致します．

ソフトコンピューティング基礎グループのメンバーには，数多くの助言を頂き，そのおかげで充実した研究室生活を送ることができました．本当にありがとうございました．

参考文献

[1] 宮本 定明, “クラスター分析入門”, 森北出版株式会社, 1999.

[2] J.B. MacQueen, “Some Methods of Classification and Analysis of Multivariate Observations”, Proceedings of the Fifth Berkeley Symposium on Mathematical Statictics and Probability, pp.281–297, 1967.

[3] S. Miyamoto, H. Ichihashi, and K. Honda, “Algorithms for Fuzzy Clustering”, Springer, Heidelberg, 2008.

[4] H. Ichihashi, K. Honda, and N. Tani, “Gaussian Mixture PDF Approximation and Fuzzy c-means Clustering with Entropy Regularization”, Proc. of the 4th Asian Fuzzy Systems Symposium, pp. 217–221, 2000.

[5] J.-W. Byun, A. Kamra, E. Bertino, N. Li, “Efficient k-Anonymization Using Clustering Techniques”, Proc. of the 12th International Conference on Database Systems for Advanced Applications (DASFAA 2007), pp.188–200, 2007.

[6] J.-L. Lin, M.-C. Wei, “An efficient clustering method for k-anonymization”, Proc. of the 2008 international workshop on Privacy and anonymity in information society (PAIS ’08), pp.46–50, 2008.

[7] X. He, H. H. Chen, Y. Chen, Y. Dong, P. Wang, Z. Huang, “Clustering-Based k-Anonymity”, Proc. of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2012), pp.405–417, 2012.

[8] T. Hirano, Y. Endo, N. Kinoshita, Y. Hamasuna, “On Even-sized Clustering Algorithm Based on Optimization”, Proc. of Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on advanced Intelligent Systems (SCIS & ISIS 2014), TP4-3-5-(3), #69, 2014.

[9] Y. Endo, S. Ishida, N. Kinoshita, “Controlled-sized Clustering Based on Optimization”, Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS 2017), 2017.

[10] K. Kitajima, Y. Endo, Y. Hamasuna, "Fuzzified Even-Sized Clustering Based on Optimization", Journal of Advanced Computational Intelligence and Intelligent Informatics, 22-4, pp.537–543, 2018.

[11] Y. Endo, T. Hirano, N. Kinoshita, Y. Hamasuna, "On Various Types of Even-sized Clustering Based on Optimization", The 13th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2016), Springer, LNAI 9880, pp.165–177, 2016.

[12] Y. Endo, S. Ishida, N. Kinoshita, Y. Hamasuna, "On Various Types of Controlled-sized Clustering Based on Optimization", 2017 IEEE International Conference of Fuzzy Systems (FUZZ-IEEE 2017), 2017.

[13] D. Arthur, S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding", Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics Philadelphia, PA, USA, pp.1027–1035, 2007.

[14] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum, New York, 1981.

[15] R. Dave, "Characterization and Detection of Noise in Clustering", Pattern Recognition Letters, 112.11 ,pp–657–664 1991.

[16] M. Kojima, S. Mizuno, A. Yoshise, "A Primal-Dual Interior Point Algorithm for Linear Programming", Progress in Mathematical Programming, Interior Point and Related Methods (ed. N. Megiddo), Springer, New York, pp.29–47, 1989.

[17] H.S. Park, C.H. Jun, "A simple and fast algorithm for K-medoids clustering", Expert Systems with Applications, 36.2, pp.3336–3341, 2009.

[18] H. Kashima, et al, "K-means clustering of proportional data using L1 distance", Pattern Recognition, 2008, ICPR 2008, 19th International Conferenceon IEEE, 2008.

[19] L. Hubert, P. Arabie, "Comparing Partitions", Journal of Classification 2, pp.193–218, 1985.

[20] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", Annals of Eugenics, vol.7,No.2,pp.179–188,1936.

[21] O. Mangasarian, W.Wolberg, "Cancer diagnosis via linear programming", SIAM News, Volume 23, pp.1–18, September 1990.