不確実性を考慮した
クラスタリングアルゴリズムの理論的考察

２０１６年　３月

木下　尚彦

不確実性を考慮した
クラスタリングアルゴリズムの理論的考察

木下　尚彦

システム情報工学研究科

筑波大学

２０１６年　３月

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　研究背景

コンピュータが一般に普及して以来，情報通信技術やコンピュータの性能は飛躍的に発達した．通信速度や記憶容量の向上はこのような技術発展の代表例であるといえ，これらはコンピュータ上に蓄積されるデータの大規模化をもたらしてきた．このようにして蓄積されたデータには Twitter でのつぶやき，クチコミ，商品レビュー等を代表とする，ユーザーの嗜好や考えを含んだものが多々ある．そのためこれらを解析し有益な情報を引き出すことは，我々の社会をより良いものへと変化させるためには必要不可欠であるといえる．しかし，大規模なデータから人の手で直接情報を処理することはほぼ不可能であるため，近年データを自動的に処理し有益な知見を抽出するデータマイニングへの関心が飛躍的に高まってきている．

このようなデータマイニング手法の代表的なものの一つとしてクラスタリングが挙げられる [1]．クラスタリングは，外的基準なしにデータを自動分類する手法であり教師なし分類の一種である．教師データなどの外的要因を必要としないクラスタリングは，外的要因を必要とする手法よりも簡便であり，自然言語処理・データマイニング・パターン認識・イメージ解析・バイオインフォマティクスといった幅広い分野で用いられている．クラスタリングはこれらの分野において有益な結果をもたらしており，今後も様々な発展が期待される．

クラスタリング手法はハードクラスタリングとソフトクラスタリングとに大別することができる．ハードクラスタリングは各データが一つのクラスタに一意に所属する手法のことであり，ハード c-平均法 (HCM) [2,3]，k-median [4,5]，最短距離法，最長距離法，群間平均法，郡

内平均法，重心法，ウォード法，スペクトラルクラスタリング [6]，DBSCAN [7] などの手法が該当する．ソフトクラスタリングは各データが複数のクラスタに属することを認めたクラスタリング手法であり，ファジィc-平均法 (FCM) [8,9]，possibilistic c-means (PCM) [10]，EM アルゴリズムに基づくクラスタリング [11,12]，relational clustering [13] などの手法が該当する．

ハードクラスタリングの中で最も代表的な手法は，ハード c-平均法 [2,3] や最短距離法・最長距離法に代表される階層併合的クラスタリングである．ハード c-平均法はデータ集合を c 個のクラスタに分類する手法であり，各データは目的関数と呼ばれる最適化関数が最も小さくなるように一つのクラスタに分類される．より詳しく述べると，ハード c-平均法はクラスタ分類とそのクラスタの代表点であるクラスタ中心を交互に最適化する手法である．クラスタ分類の最適化は，各データとクラスタ中心と呼ばれるクラスタの代表点との非類似度が最も小さいクラスタへの分類によってなされ，クラスタ中心の最適化は，各クラスタに属するデータ群の重心を求めることで実現される．階層併合的クラスタリング (agglomerative hierarchical clustering, AHC) はあらかじめすべてのデータをクラスタとみなし，それらを各手法で定義される類似度，非類似度に基づいて逐一結合していく方法であり，結合したクラスタ同士は一つのクラスタとみなされクラスタ数が徐々に減少していくために，任意のタイミングでのクラスタ分割を知ることができる．

ソフトクラスタリングの中で最も代表的な手法は Bezdek によって提案されたファジィc-平均法 [8] や EM アルゴリズムに基づくクラスタリングである．ファジィc-平均法はファジィ理論 [14] を援用しハード c-平均法を拡張した手法である．ハード c-平均法では，先程述べたように各データはクラスタに属すか，属さないかの $\{0,1\}$ でクラスタへの帰属度が定まる．一方ファジィc-平均法では，目的関数がクラスタへの帰属度に関して非線形化されているため，クラスタへの帰属度が $[0,1]$ の連続値で定まる．これにより，各データの複数のクラスタへの帰属が実現された．EM アルゴリズムに基づくクラスタリングでは，最尤推定法の一種である EM アルゴリズム [11,12] を援用する．EM アルゴリズムは対数尤度の条件付き期待値である

尤度関数を計算するステップ (E ステップ) と，その関数を最大化するステップ (M ステップ) との繰り返しによって最適解を求める方法である．EM アルゴリズムに基づくクラスタリングにおいて，クラスタは確率密度関数で表現され，各データのクラスタへの帰属度は個々の密度分布の混合密度分布に占める割合で表される．

またハード c-平均法やファジィ c-平均法，EM アルゴリズムに基づくクラスタリングといった諸手法は，分類をおこなう上で予め初期クラスタ分割，もしくは初期クラスタ中心を設定しておかなくてはならない．そのためこれらの手法は，設定した初期値によって得られる結果が大きく変わってしまういわゆる初期値依存性が非常に強い手法となっている．初期値依存性を防ぐために一般的に取られる対策は，予め複数の初期値を準備しておき，それらから得られる結果の中で目的関数値が最小もしくは最大となる場合の分類を最適解として採用する方法である．

1.2 不確実性について

不確実性とは事象の発生が確実でないことを指す概念であり，物事の正確な予測・解析をおこなう上では無視できないものとなっている．クラスタリングにおいてもそれは同様であり，クラスタリングにおける不確実性には以下の要素が考えられる．

- クラスタ表現に含まれる不確実性.
- データ自身に含まれる不確実性.
- クラスタリングアルゴリズムにおける不確実性.

クラスタ表現に含まれる不確実性は，クラスタ分割の不明瞭さから引き起こされる不確実性である．この不確実性は先程述べたようにソフトクラスタリング手法によって扱うことができる．ソフトクラスタリング手法は，帰属度表現へのあいまいさの導入によって帰属度合いを柔軟に表現し，クラスタ表現に含まれる不確実性を扱っている．だが一方で，ファジィ c-平均

法に代表されるファジィ集合論を用いた手法におけるクラスタ分類の表現は細かすぎるという批判もなされてきた．そこで Lingras らはラフ集合論 [15, 16] を援用したラフクラスタリング rough k-means (RKM) [17, 18] を提案し，ファジィ集合論に基づくクラスタ分類よりも荒い帰属度表現のあいまいさを実現した．この rough k-means を基に，これまで様々なラフクラスタリング手法が提案されてきた [19–25] が，これらの手法には二つの共通した問題点が存在した．それはクラスタリングによって得られた解が目的関数最適化という指標から得られたものではないため，解の妥当性が不明瞭であること，異なる初期値から得られた分類の良し悪しの判別がおこなえないため，初期値依存性が強いという欠点を補うことができないという点である．これらの問題は既存のラフクラスタリング手法が階層併合的手法ではなく，かつハード c-平均法やファジィ c-平均法のような目的関数を定義していないことに起因している．

データ自身に含まれる不確実性には，データが実空間からデータ化される際に生じるモデル化誤差や欠損が挙げられる．このようなモデル化誤差は主に以下の要因によって生じる．

- **実空間からパターン空間への写像時に生じる誤差** : 最も一般的に起こる誤差であり，実空間上の点をパターン空間上に写像するために丸めをおこなうことによって生じる誤差である．

- **実空間上で生じる誤差や欠損** : 実空間上のデータそのものに不確実性が含まれる場合がこれに該当する．前者は対象の大きさがあまりにも巨大なため厳密な計測がおこなえない場合や，対象に関する知識の欠乏によって生じる．後者は対象が化石の遺伝子データなどのように観測時にすでに不完全である場合に生じる．

- **元データに幅があることにより生じる誤差** : りんごを観測し実空間から RGB 空間に写像する際，ある一点を写像する場合においても実際のりんごの色は均一ではない．しかしそのような色の違いは考慮されず，写像する際には一つの色として表現される．

クラスタリング手法において一般的にはこれらの不確実性は無視される．一方不確実性を考慮する場合，代替データを設定する [26]，各データを不確実性を伴う区間データとして扱う [27,28]，各データを不確実性を含んだ確率密度関数として扱う [29] のが一般的である．不確実性を伴うデータを区間データとして扱った場合，実際に用いられるのはその端点のみであるため，十分に不確実性を考慮しているとはいえない．区間データと同様にデータに幅を持たせて不確実性を解析する概念として，許容範囲 [30–32] や不確実性ベクトル [33,34] がある．これらは区間の端点を扱い解析する区間データとは異なり，区間の領域すべてを対象としており最適化問題の枠組みで不確実性を論じることが可能な概念である．一方で各データを不確実性を含んだ確率密度関数として扱う手法は非常に少なく，十分に確立されていないのが現状である．

クラスタリングアルゴリズムにおける不確実性は，初期値依存性やパラメータ設定によって生じる不確実性である．特に初期値依存性は前節で述べたように階層的併合法以外の多くのクラスタリング手法において避けられぬ障害であり，クラスタリングアルゴリズム自体の初期値依存性をいかに減らし，計算量を減らすかがクラスタリング研究における課題の一つともなっている．この問題に対しては，Arthur らがハード c-平均法の初期値を確率的に選択することで，ハード c-平均法の目的関数の期待値がその最適解に対して一定の範囲で抑えられることを理論的に示した k-means++ [35] を提案している．また，k-means++の考え方を球面上クラスタリングに導入した Spherical k-means++ [36] も遠藤によって提案されている．

1.3 本研究の目的

本研究では，正確な解析をおこなう上で無視することのできない不確実性を扱うクラスタリング手法の構築を目的としている．ただし本研究では前節で述べたクラスタリングにおける不確実性のうち，クラスタ表現に含まれる不確実性，及びデータ自身に含まれる不確実性のみを対象とし，クラスタリングアルゴリズムにおける不確実性については扱わない．クラ

スタ表現に含まれる不確実性として，本研究ではラフクラスタリングに着目した新たなクラスタリング手法を構築する．rough k-means を基にしたラフクラスタリング手法は前節で述べたように，目的関数を定義していないがゆえの問題をいくつか抱えている．そこで本研究ではそれらの問題を解決するために，目的関数を陽に記述した，rough k-means とハード c-平均法に基づいた二種類のラフクラスタリング手法を構築する．それにより従来手法ではおこなえなかった解の妥当性の保証，初期値依存性に対する対策が期待できる．そして，それらの手法の制約を緩和した派生系として，一部分にファジィ化を導入したラフクラスタリング手法についても同様に構築する．データ自身に含まれる不確実性として，本研究では各データに対して不確実性ベクトルを導入した手法と各データをガウス分布として扱う手法のニ種類を構築する．不確実性ベクトルを導入したクラスタリング手法はすでにいくつか提案されているが，確率分布に基づいた EM アルゴリズムに基づくクラスタリングと不確実性ベクトルとの親和性についての議論は十分になされていない．また EM アルゴリズムに基づくクラスタリングではクラスタ分布が確率密度関数として表されるため，各データを確率密度関数と見なした場合の親和性がクラスタリング諸技法の中で非常に高いと考えられる．そこで，本手法はともに EM アルゴリズムに基づいたクラスタリング手法をベースとして構築する．

1.4 本論文の構成

本論文では第 2 章でクラスタ表現に含まれる不確実性を扱うラフクラスタリング手法，第 3 章でデータ自身に含まれる不確実性を確率密度関数として扱う EM アルゴリズムに基づくクラスタリング手法，第 4 章でデータ自身に含まれる不確実性を不確実性ベクトルとして扱う EM アルゴリズムに基づくクラスタリング手法を述べる．

第 2 章ではまずはじめに，ラフクラスタリングの基となる概念であるラフ集合論について述べた後に，関連研究として Lingras らによって提案されたラフクラスタリング rough k-means を紹介する．その後，提案手法である目的関数最適化に基づいたラフクラスタリング手法を

構築する．構築する手法は全部で五種類あり，その内の一つが rough k-means の最適解に準じた目的関数を定義した手法であり，二つがその手法の一部の制約を緩和した，Bezdek [8] による標準的なファジィ化と宮本ら [9] のエントロピー正則化によるファジィ化を導入したファジィラフクラスタリング手法である．また残りの二つの内の一つはハード c-平均法の目的関数を基に構築した手法であり，最後の一つは Bezdek らによるファジィ c-平均法の目的関数を基に構築したファジィラフクラスタリング手法となっている．第 2 章の最後では数値例として人工データ，実データを用いた場合の提案手法同士の比較をおこなう．この比較によって，各アルゴリズムの分類特性を示すことを目的とする．

第 3 章では関連研究として EM アルゴリズムに基づくクラスタリング手法を一次元の場合と多次元の場合それぞれについて説明する．その後，提案手法であるデータ自身を確率密度関数として扱う EM アルゴリズムに基づくクラスタリング手法を構築する．最後に数値例実験を通して，提案手法によるデータに課した不確実性の効果を既存手法と比較することで確認する．

第 4 章では関連研究として KL 情報量正則化によるファジィ c-平均法 [37]，ペナルティベクトル正則化に基づく標準型ファジィ c-平均法 [33] について述べる．その後 KL 情報量正則化と不確実性ベクトルを導入した提案手法を構築し，EM アルゴリズムとの関連性について述べる．最後に数値例実験を通して，各提案手法の比較検討をおこない，不確実性ベクトルの効果について議論する．

第2章　クラスタ表現における不確実性ー目的関数最適化に基づくラフクラスタリング

クラスタ表現における不確実性を取り扱う概念として，クラスタリングでよく応用されているのはファジィ集合 [14] である．ファジィ集合を用いることで，データのクラスタへの帰属は従来の $\{0,1\}$ の二値分類から，$[0,1]$ への細かい表現が可能になった．しかしこのような細かい表現は表現過多であるという批判が起こり，よりおおまかな分類をおこなうラフ集合 [15,16] を援用したラフクラスタリング [17,18,38,39] が提案された．ラフ集合はデータをクラスタに確実に含まれる下近似，含まれるかどうか不明である上近似，含まれないの三値に分類する，クラスタの帰属における不確実性を扱う新たな概念である．

これまで Lingras らによって提案されたラフクラスタリング (rough k-means) はファジィ化やラフ回帰など様々な派生手法 [19–25] が研究されてきた．しかしこれらの手法はハード c-平均法やファジィc-平均法に代表される，目的関数と呼ばれる評価関数の最小化あるいは最大化に基づいたクラスタリングアルゴリズムではなく，与えた初期値に対してクラスタの代表点であるクラスタ中心と，各データのクラスタへの帰属関係の交互最適化を繰り返す手法となっている．そのため複数回初期値を与えた際の各初期値同士の結果の判別や，最適解更新の判断がおこなえないため，極端に初期値依存性が強い手法となっている．

そこで本章ではこの問題点を解決すべく目的関数を明確にした新たなラフクラスタリング手法を構築する．ここで構築する手法は大まかに分けて二種類あり，一つは rough k-means で得られるクラスタ中心の最適解に基づいた手法であり，もう一つは最も基本的なクラスタリング手法であるハード c-平均法 [2] の目的関数を基にした手法である．また，一部の制約をゆ

るめた派生系として，一部分にファジィ化を導入した手法についても合わせて構築する．クラスタリング手法のファジィ化をおこなう方法として，ここでは Bezdek [8] によるファジィ化の方法と，宮本ら [9] によるエントロピー項導入によるファジィ化を用いる．

本章ではまずラフ集合について説明し，その後関連手法である rough k-means，及び提案手法について述べたい．

2.1 ラフ集合

ラフ集合はファジィ集合に変わるあいまいさを扱う概念として Pawlak [15] によって提案された．ラフ集合の特徴としては，ファジィ集合では $[0,1]$ の連続値として表していた集合への帰属性を，必ず属している，属しているか不明，属さないの三つに分類することが挙げられる．

U を空間上の集合とし，$R \subseteq U \times U$ を U の同値関係もしくは識別不能関係とする．このとき $X = (U, R)$ を近似空間と呼び，U は R によって c 個の部分空間 $A_1, \ldots, A_c$ に分割される．各 A_i は R の同値類とみなせる．つまり，$U/R = \{A_1, \ldots, A_c\}$ である．$x, y \in A_i \subset U/R$ であるなら，x と y は識別不能関係である．

同じ同値類に含まれる要素は識別不能なので，任意の部分集合 $A_i \subset U$ の正確な表現を得ることはできない．代わりに，任意の A_i は下近似と上近似を用いて表現される．下近似 $\underline{A_i}$ は A_i の部分集合であるすべての要素集合の集合である．上近似 $\overline{A_i}$ は A_i の空でない共通部分であるすべての要素集合の集合である．そして $(\underline{A_i}, \overline{A_i})$ は X のラフ集合を表す．特徴としては，$\underline{A_i}$ に含まれる要素は A_i に確実に含まれるが，$\overline{A_i}$ に含まれる要素は A_i に含まれるとは限らないということが挙げられる．また，$\mathrm{Bnd}(A_i) = \overline{A_i} - \underline{A_i}$ を A_i の境界と呼ぶ．図 2.1 がラフ集合の概要を表した図となっている．

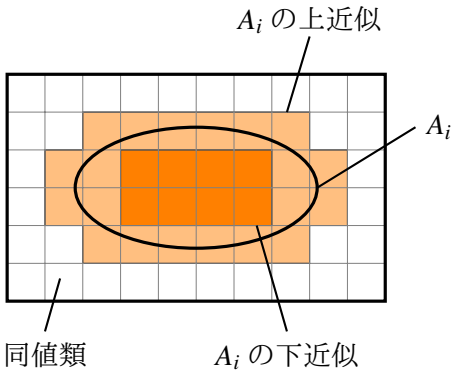

図 2.1: ラフ集合の概要

2.2 関連手法

この節では関連手法として，ラフ集合の考え方を援用したクラスタリング手法 [17,18,38,39] の一つとして，Lingras ら [17,18] によって提案された rough k-means (RKM) について述べる．Rough k-means はラフクラスタリングの中で非常に代表的な手法であり，様々な拡張がなされている．rogh k-means はハード c-平均法を基に構築されており，クラスタ中心の最適化と，データのクラスタへの帰属関係である下近似上近似を交互最適化するクラスタリング手法である．rough k-means ではクラスタ中心とデータ間の非類似度に基づき，各データが属する下近似と上近似が決定される．

2.2.1 Rough k-means

$U = \{x_k \mid x_k = (x_k^1, \ldots, x_k^p)^T \in \mathfrak{R}^p, k = 1, \ldots, n\}$，$U/R = \{A_i \mid i = 1, \ldots, c\}$ をそれぞれデータ集合，クラスタ集合とし，$v_i = (v_i^1, \ldots, v_i^p)^T \in \mathfrak{R}^p$ $(i = 1, \ldots, c)$ をクラスタ A_i のクラスタ中心とする．

Lingras らはラフ集合の下近似，上近似の性質からクラスタリングにおける制約を次のように設けた．

(C1) データ x_k は多くとも一つの下近似にしか属さない．

(C2) $x_k \in \underline{A_i}$ ならば $x_k \in \overline{A_i}$．

(C3) データ x_k が下近似に属さないならば，少なくとも二つの上近似に属す．

クラスタ中心の最適解は，ハード c-平均法ではクラスタに含まれるデータの重心を最適解としている．Rough k-means でも同様にクラスタに含まれるデータの重心を最適解としているが，rough k-means では上近似と下近似という二種類の集合が存在するため，クラスタの重心はそれらの集合の重心の内分点として最適解を与えている．

$$v_i = \begin{cases} \dfrac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|}, & (\mathrm{Bnd}(A_i) = \emptyset) \\ \dfrac{\sum_{x_k \in \overline{A_i}} x_k}{|\overline{A_i}|}, & (\underline{A_i} = \emptyset) \\ \underline{w} \times \dfrac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|} + \overline{w} \times \dfrac{\sum_{x_k \in \mathrm{Bnd}(A_i)} x_k}{|\,\mathrm{Bnd}(A_i)|}. & (\text{otherwise}) \end{cases} \tag{2.1}$$

制約条件は以下のとおりである．

$$\underline{w} + \overline{w} = 1. \quad (\underline{w} > 0,\ \overline{w} > 0)$$

$\underline{w}$ と $\overline{w}$ は重心の位置を決めるパラメータとなっている．

下近似と上近似の最適解は次の手順によって求められる．まず任意のデータ x_k と各クラスタとの非類似度を求める．Rough k-means では非類似度としてユークリッド距離の二乗を用いている．次に，x_k と最近隣のクラスタと，それ以外のクラスタとの非類似度との差を取り，それが一定値以下となるクラスタラベルの集合を作成する．その集合が空集合であったなら，x_k は最近隣クラスタの下近似に属し，空集合でなければ，x_k はその集合に含まれているクラスタすべてと最近隣クラスタの上近似に属する．数式として記述すると以下のように記述できる．

$$d_{ki} = \|x_k - v_i\|^2.$$

$$d_{km} = \min_{1 \le i \le c} d_{ki}.$$

$$T = \{i \mid d_{ki} - d_{km} \leq \text{threshold}\} \quad (i \neq m).$$

- $T \neq \emptyset$ ならば $x_k \in \overline{A_m}$ かつ $x_k \in \overline{A_i}$ $(\forall i \in T)$．これは制約 (C3) に該当する．
- $T = \emptyset$ ならば $x_k \in \underline{A_m}$．これは制約 (C1) に該当し，制約 (C2) より同時に $x_k \in \overline{A_m}$ も満たす．

Rough k-means は上記の式，手順を用いてクラスタ中心の変位が一定値以下になるまで交互最適化が繰り返される．最後に rough k-means のアルゴリズムを以下に示す．

Algorithm 1 RKM

RKM1 初期近似関係を与え，初期クラスタ中心を計算する．

RKM2 上記の手順を用いて下近似と上近似を更新する．

RKM3 (2.1) を用いてクラスタ中心を更新する．

RKM4 収束条件を満たしたら終了．そうでなければ **RKM2** に戻る．

2.3 提案手法

先に紹介した rough k-means は最適化に基づいたクラスタリング手法ではあるが，以下に挙げる問題点が存在する．

- Rough k-means は目的関数という最適化関数に基づいたクラスタリング手法でないため，得られた解の妥当性が不明瞭である．
- 異なる初期値から得られた分類の良し悪しの判別がおこなえないため，初期値依存性問題に対応できない．
- threshold で表されるしきい値を適切に設定するのが困難である．

そこで本節ではこれらの問題を解決するために，目的関数を陽に示し数理計画法によって構築した新たなラフクラスタリング手法を構築する．提案する五種類の手法と概要は以下のとおりである．

- **目的関数に基づくラフクラスタリング** : Rough k-means のクラスタ中心の最適解に基づいた目的関数を定義した手法.
- **目的関数に基づく標準型ファジィラフクラスタリング** : 目的関数に基づくラフクラスタリングの境界領域をべき乗によってファジィ化した拡張手法.
- **目的関数に基づくエントロピー型ファジィラフクラスタリング** : 目的関数に基づくラフクラスタリングの境界領域をエントロピー正則化によってファジィ化した拡張手法.
- **ラフハード c-平均法** : ハード c-平均法の目的関数を基に目的関数を定義した手法.
- **ラフファジィ c-平均法** : ラフハード c-平均法の境界領域をべき乗によってファジィ化した拡張手法.

2.3.1 目的関数に基づくラフクラスタリング

目的関数に基づくラフクラスタリング (rough c-means, RCM) は，rough k-means のクラスタ中心の最適解を基に既存のラフクラスタリングの問題点を解決した手法である．Rough k-means と同様に RCM もクラスタ中心と，下近似上近似への帰属関係を交互最適化する最適化手法となっている．ただし上近似は下近似をも含んでしまうため，都合上，RCM では上近似の最適解ではなく境界領域の最適解を導出している．

$x_k = (x_k^1, \ldots, x_k^p)^T \in \mathfrak{R}^p$ $(k = 1, \ldots, n)$ を任意のデータ，$N = (v_{ki})$ $(1 \leq k \leq n,\ 1 \leq i \leq c)$ をクラスタ A_i の下近似 $\underline{A_i}$ に対する x_k の帰属度行列，$U = (u_{ki})$ を A_i の境界 $\mathrm{Bnd}(A_i)$ に対する x_k の帰属度行列，$V = \{v_1, \ldots, v_c\}$ をクラスタ中心の集合とする．

このとき RCM の目的関数は以下のように定義される．

$$J_{\mathrm{RCM}}(N,U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{l=1}^{n}\left(v_{li}u_{ki}(\underline{w}d_{li}+\overline{w}d_{ki})+(v_{ki}v_{li}+u_{ki}u_{li})D_{kl}\right). \tag{2.2}$$

$$\underline{w}+\overline{w}=1,$$

非類似度はユークリッド距離の二乗を用いて次のように定義される．

$$d_{ki}=\|x_k-v_i\|^2,\ D_{kl}=\|x_k-x_l\|^2.$$

RCM のクラスタリングにおける制約は以下のとおりである．

$$v_{ki},u_{ki}\in\{0,1\},\ \forall k,i$$
$$\sum_{i=1}^{c}v_{ki}\in\{0,1\},\ \sum_{i=1}^{c}u_{ki}\neq 1,\ \forall k$$
$$\sum_{i=1}^{c}v_{ki}=1\iff\sum_{i=1}^{c}u_{ki}=0.\ \forall k$$

これらの制約をまとめると，次の制約が得られる．

$$\sum_{i=1}^{c}v_{ki}=0\iff\sum_{i=1}^{c}u_{ki}>1.\ \forall k$$

これらの制約は 2.2.1 で述べた (C1) – (C3) の制約と同等である．RCM の最適化問題は，この制約のもとでの (2.2) の最小化である．

次にアルゴリズムを構築する上で必要な最適解を導出していく．まず最初にクラスタ中心の最適解を導出する．(2.2) を v_i に関して偏微分すると，

$$\frac{\partial J_{\mathrm{RCM}}}{\partial v_i}=\sum_{k=1}^{n}\sum_{l=1}^{n}v_{li}u_{ki}\left(\underline{w}(x_l-v_i)+\overline{w}(x_k-v_i)\right)=0.$$

$\underline{w}+\overline{w}=1$ であることも考慮すると，上式は次のように変形できる．

$$\sum_{k=1}^{n}\sum_{l=1}^{n}v_{li}u_{ki}v_i=\sum_{k=1}^{n}\sum_{l=1}^{n}v_{li}u_{ki}(\underline{w}x_l+\overline{w}x_k).$$

この式を変形して，

$$\sum_{l=1}^{n} v_{li} \sum_{k=1}^{n} u_{ki} v_i = \underline{w} \sum_{k=1}^{n} u_{ki} \sum_{l=1}^{n} v_{li} x_l + \overline{w} \sum_{l=1}^{n} v_{li} \sum_{k=1}^{n} u_{ki} x_k.$$

ここで先ほどの制約より以下の関係が成り立つことがわかる．

$$|\underline{A_i}| = \sum_{l=1}^{n} v_{li},$$
$$|\operatorname{Bnd}(A_i)| = \sum_{k=1}^{n} u_{ki}.$$

この関係を先ほどの式に適用し式変形をおこなうと，

$$|\underline{A_i}||\operatorname{Bnd}(A_i)|v_i = \underline{w}|\operatorname{Bnd}(A_i)| \sum_{x_k \in \underline{A_i}} x_k + \overline{w}|\underline{A_i}| \sum_{x_k \in \operatorname{Bnd}(A_i)} x_k.$$

従って，クラスタ中心の最適解は以下の式によって求めることができる．

$$v_i = \underline{w} \times \frac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|} + \overline{w} \times \frac{\sum_{x_k \in \operatorname{Bnd}(A_i)} x_k}{|\operatorname{Bnd}(A_i)|}.$$

ただし，$\underline{A_i} = \emptyset$ ならば $\overline{w} = 1$ であり，$\operatorname{Bnd}(A_i) = \emptyset$ ならば $\underline{w} = 1$ である．まとめると，

$$v_i = \begin{cases} \dfrac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|}, & (\operatorname{Bnd}(A_i) = \emptyset) \\ \dfrac{\sum_{x_k \in \overline{A_i}} x_k}{|\overline{A_i}|}, & (\underline{A_i} = \emptyset) \\ \underline{w} \times \dfrac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|} + \overline{w} \times \dfrac{\sum_{x_k \in \operatorname{Bnd}(A_i)} x_k}{|\operatorname{Bnd}(A_i)|}. & (\text{otherwise}) \end{cases} \tag{2.3}$$

(2.3) は rough k-means のクラスタ中心の最適解 (2.1) と一致する．

次に近似関係の最適解を求める．しかし，u_{ki}，v_{ki} ともに線形項なので，クラスタ中心の最適解の導出のように偏微分による導出はできない．また，クラスタリングにおける制約 (C1) – (C3) より，各データはクラスタの下近似もしくは境界領域のどちらかにしか所属しない．そのためそれぞれに所属した場合の比較をおこない，どちらに所属するのがより最適となるかを調べる必要がある．そこでデータ x_k に着目し，x_k が下近似に含まれるべきか境界領域に含まれるべきかを目的関数から判断することにする．これらをまとめると考慮すべき条件は以下のようにまとめられる．

- データ x_k がクラスタ A_{p_k} の下近似 $\underline{A}_{p_k}$ に属する場合.
- データ x_k が二つのクラスタ A_{p_k} と A_{q_k} の境界 $\mathrm{Bnd}(A_{p_k})$，$\mathrm{Bnd}(A_{q_k})$ に属する場合.

p_k と q_k は次式で定義される.

$$p_k = \arg\min_i d_{ki}, \quad q_k = \arg\min_{i \neq p_k} d_{ki}.$$

x_k が $\underline{A}_{p_k}$ に属するならば，目的関数は以下のように表される.

$$\underline{J}^k_{\mathrm{RCM}} = \sum_{l=1, l\neq k}^{n} (u_{lp_k}(\underline{w} d_{kp_k} + \overline{w} d_{lp_k}) + 2\nu_{lp_k} D_{kl}). \tag{2.4}$$

x_k が $\mathrm{Bnd}(A_{p_k})$ と $\mathrm{Bnd}(A_{q_k})$ に属するならば，目的関数は以下のように表される.

$$\overline{J}^k_{\mathrm{RCM}} = \sum_{i=p_k, q_k} \sum_{l=1, l\neq k}^{n} (\nu_{li}(\underline{w} d_{li} + \overline{w} d_{ki}) + 2u_{li} D_{kl}). \tag{2.5}$$

本手法では目的関数の最小化によって最適解を得る手法であるため，(2.4) と (2.5) で得られる値を比較し，その値がより小さくなるようにデータを帰属させる．最終的に N と U についての最適解は以下のように求められる.

$$\nu_{ki} = \begin{cases} 1, & (\underline{J}^k_{\mathrm{RCM}} < \overline{J}^k_{\mathrm{RCM}} \wedge i = p_k) \\ 0, & (\text{otherwise}) \end{cases}$$

$$u_{ki} = \begin{cases} 1, & (\underline{J}^k_{\mathrm{RCM}} > \overline{J}^k_{\mathrm{RCM}} \wedge (i = p_k \vee i = q_k)) \\ 0. & (\text{otherwise}) \end{cases}$$

得られた最適解を基にアルゴリズムを構築すると，RCM のアルゴリズムは以下のようになる.

Algorithm 2 RCM

RCM1 初期近似関係を設定し，初期クラスタ中心を計算する．

RCM2 上記の手順を用いて下近似と境界を更新する．

RCM3 (2.3) を用いてクラスタ中心を更新する．

RCM4 収束条件を満たしたなら終了．そうでなければ **RCM2** に戻る．

2.3.2 目的関数に基づく標準型ファジィラフクラスタリング

目的関数に基づく標準型ファジィラフクラスタリング (rough c-means with fuzzy upper approximations, RCM-FU) は RCM の境界領域のみをファジィ化した手法である．ここでのファジィ化は Bezdek による標準的なファジィ化を指す．RCM では前節の近似関係の最適解の導出からもわかるように，三つ以上のクラスタの境界領域にデータが属することはない．なぜなら，データがクラスタの境界領域に属する場合は二つ以上のクラスタの境界領域に属することになるが，目的関数の最小化の観点から，二つより多くのクラスタの境界領域に属することはありえないからである．しかし境界領域をファジィ化することによって，RCM ではおこなえなかったデータの三つ以上のクラスタの境界領域への帰属が期待でき，より精密なクラスタ分類を可能にすると考えられる．

RCM-FU の目的関数は次のように定義される．

$$J_{\text{RCM-FU}}(N,U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{l=1}^{n}\left(v_{li}u_{ki}^{m}(\underline{w}d_{li}+\overline{w}d_{ki})+(v_{ki}v_{li}+u_{ki}^{m}u_{li}^{m})D_{kl}\right). \tag{2.6}$$

$$\underline{w}+\overline{w}=1.$$

$$d_{ki}=\|x_k-v_i\|^2,\ D_{kl}=\|x_k-x_l\|^2.$$

m はファジィ化パラメータを表し一般的には $m > 1$ である．Lingras らの考案したクラスタリングにおける制約は以下のように表される．

$$v_{ki} \in \{0, 1\}, \quad u_{ki} \in [0, 1], \ \forall k$$

$$\sum_{i=1}^{c} (v_{ki} + u_{ki}) = 1. \ \forall k$$

RCM-FU の最適化問題は，この制約のもとでの (2.6) の最小化である．

RCM-FU は RCM と同様にクラスタ中心と下近似と境界領域を最適化する手法なので，アルゴリズムを構築するために，それらの最適解を導出していく．まずクラスタ中心の最適解を得るために，RCM と同様に (2.6) を v_i に関して偏微分する．

$$\frac{\partial J_{\text{RCM-FU}}}{\partial v_i} = \sum_{k=1}^{n} \sum_{l=1}^{n} \nu_{li} u_{ki}^m \left(\underline{w}(x_l - v_i) + \overline{w}(x_k - v_i) \right) = 0.$$

この式を $\underline{w} + \overline{w} = 1$ であることも考慮して整理すると，

$$\sum_{k=1}^{n} \sum_{l=1}^{n} \nu_{li} u_{ki}^m v_i = \sum_{k=1}^{n} \sum_{l=1}^{n} \nu_{li} u_{ki}^m (\underline{w} x_l + \overline{w} x_k).$$

上式を変形すると，

$$\sum_{l=1}^{n} \nu_{li} \sum_{k=1}^{n} u_{ki}^m v_i = \underline{w} \sum_{k=1}^{n} u_{ki}^m \sum_{l=1}^{n} \nu_{li} x_l + \overline{w} \sum_{l=1}^{n} \nu_{li} \sum_{k=1}^{n} u_{ki}^m x_k.$$

ここで先ほどの制約より，以下の関係が成り立つ．

$$|\underline{A_i}| = \sum_{l=1}^{n} \nu_{li}.$$

この関係を先ほどの式に代入していくと次式が得られる．

$$|\underline{A_i}| \sum_{k=1}^{n} u_{ki}^m v_i = \underline{w} \sum_{k=1}^{n} u_{ki}^m \sum_{x_k \in \underline{A_i}} x_k + \overline{w} |\underline{A_i}| \sum_{k=1}^{n} u_{ki}^m x_k.$$

本式より，クラスタ中心の最適解は，

$$v_i = \underline{w} \times \frac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|} + \overline{w} \times \frac{\sum_{k=1}^{n} u_{ki}^m x_k}{\sum_{k=1}^{n} u_{ki}^m}.$$

$\underline{A_i} = \emptyset$，$\mathrm{Bnd}(A_i) = \emptyset$ である場合をそれぞれ $\overline{w} = 1$，$\underline{w} = 1$ とすると，最適解は次のようにまとめられる．

$$v_i = \begin{cases} \dfrac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|}, & (\mathrm{Bnd}(A_i) = \emptyset) \\ \dfrac{\sum_{k=1}^{n} u_{ki}^m x_k}{\sum_{k=1}^{n} u_{ki}^m}, & (\underline{A_i} = \emptyset) \\ \underline{w} \times \dfrac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|} + \overline{w} \times \dfrac{\sum_{k=1}^{n} u_{ki}^m x_k}{\sum_{k=1}^{n} u_{ki}^m}. & (\text{otherwise}) \end{cases} \tag{2.7}$$

次に近似関係の最適解を求める．RCM と異なり v_{ki} は線形であるが，u_{ki}^m は非線形であるため u_{ki} の最適解はラグランジュの未定乗数法を用いて導出できる．しかし，下近似と境界領域どちらに属した方がより最適になるかを決めなくてはいけないため，RCM と同様にデータ x_k がどちらに属したほうが最適となるかを目的関数から判断する．考慮すべき条件は以下のとおりである．

- データ x_k がクラスタ A_{p_k} の下近似 $\underline{A}_{p_k}$ に属する場合．
- データ x_k が任意のクラスタ A_i の境界 $\mathrm{Bnd}(A_i)$ $\forall i$ に属する場合．

$$p_k = \arg\min_i d_{ki}.$$

x_k が $\underline{A}_{p_k}$ に属するならば，目的関数は以下のように表される．

$$\underline{J}_{\text{RCM-FU}}^k = \sum_{l=1, l\neq k}^{n} (u_{lp_k}^m (\underline{w} d_{kp_k} + \overline{w} d_{lp_k}) + 2 v_{lp_k} D_{kl}). \tag{2.8}$$

x_k が $\mathrm{Bnd}(A_i)$ に属するならば，目的関数は以下のように表される．

$$\overline{J}_{\text{RCM-FU}}^k = \sum_{i=1}^{c} \sum_{l=1, l\neq k}^{n} (u_{ki}^m v_{li} (\underline{w} d_{li} + \overline{w} d_{ki}) + 2 u_{ki}^m u_{li}^m D_{kl}). \tag{2.9}$$

(2.8) と (2.9) で得られる値を比較し，その値が小さくなる方に各データを帰属させるため，N

と U についての最適解は以下のように求められる．

$$\nu_{ki} = \begin{cases} 1, & (J_k^{\nu} < J_k^{u} \wedge i = p_k) \\ 0, & (\text{otherwise}) \end{cases}$$

$$u_{ki} = \begin{cases} 0, & (J_k^{\nu} < J_k^{u} \wedge i = p_k) \\ \dfrac{\left(\frac{1}{\alpha_i}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\frac{1}{\alpha_j}\right)^{\frac{1}{m-1}}}. & (\text{otherwise}) \end{cases}$$

ただし，

$$\alpha_i = \sum_{l=1, l\neq k}^{n} (\nu_{li}(\underline{w}d_{li} + \overline{w}d_{ki}) + 2u_{li}^{m}D_{kl}).$$

境界領域 u_{ki} の最適解はラグランジュの未定乗数法を用いて求めている．(2.6) より，ラグランジュ関数は以下のように定義される．

$$L_{\text{RCM-FU}} = \sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{l=1}^{n}\left(\nu_{li}u_{ki}^{m}(\underline{w}d_{li} + \overline{w}d_{ki}) + (\nu_{ki}\nu_{li} + u_{ki}^{m}u_{li}^{m})D_{kl}\right) - \sum_{k=1}^{n}\lambda_k\left(\sum_{i=1}^{c}u_{ki} - 1\right).$$

この関数を u_{ki} に関して偏微分する．境界領域にデータが帰属する場合，そのデータが同時に下近似に属することはありえないので $\nu_{ki} = 0$ である．このことも考慮すると，

$$\frac{\partial L_{\text{RCM-FU}}}{\partial u_{ki}} = \sum_{l=1, l\neq k}^{n} m u_{ki}^{m-1}\left(\nu_{li}(\underline{w}d_{li} + \overline{d}_{ki}) + 2u_{li}^{m}D_{kl}\right) - \lambda_k = 0.$$

この式より，

$$u_{kj} = \left(\frac{\lambda_k}{\sum_{l=1, l\neq k}^{n}\left(\nu_{lj}(\underline{w}d_{lj} + \overline{w}d_{kj}) + 2u_{lj}^{m}D_{kl}\right)}\right)^{\frac{1}{m-1}}. \tag{2.10}$$

この式の両辺を $j = 1, \ldots, c$ について加え，制約 $\sum_{j=1}^{c} u_{kj} = 1$ を用いると，

$$\sum_{j=1}^{c}\left(\frac{\lambda_k}{\sum_{l=1, l\neq k}^{n}\left(\nu_{li}(\underline{w}d_{lj} + \overline{d}_{kj}) + 2u_{lj}^{m}D_{kl}\right)}\right)^{\frac{1}{m-1}} = 1,$$

$$\lambda_k^{\frac{1}{m-1}} = \frac{1}{\sum_{j=1}^{c}\left(\frac{1}{\sum_{l=1, l\neq k}^{n}\left(\nu_{li}(\underline{w}d_{lj} + \overline{d}_{kj}) + 2u_{lj}^{m}D_{kl}\right)}\right)^{\frac{1}{m-1}}}.$$

この式を (2.10) に代入することで境界領域の最適解は得られる．

最後に RCM-FU のアルゴリズムを以下に示す．

Algorithm 3 RCM-FU

RCM-FU1 初期近似関係を設定し，初期クラスタ中心を計算する．

RCM-FU2 上記の手順を用いて下近似と境界を更新する．

RCM-FU3 (2.7) を用いてクラスタ中心を更新する．

RCM-FU4 収束条件を満たしたなら終了．そうでなければ **RCM-FU2** に戻る．

2.3.3 目的関数に基づくエントロピー型ファジィラフクラスタリング

目的関数に基づくエントロピー型ファジィラフクラスタリング (entropy rough c-means with fuzzy upper approximations, ERCM-FU) も RCM の境界領域のみをファジィ化した手法である．標準型との違いは u_{ki} を非線形化してファジィ化を実現するのではなく，正則化項であるエントロピー項を加えることでファジィ化を実現している点である．

ERCM-FU の目的関数は次のように定義される．

$$J_{\text{ERCM-FU}}(N,U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{l=1}^{n}\left(v_{li}u_{ki}(\underline{w}d_{li}+\overline{w}d_{ki})+(v_{ki}v_{li}+u_{ki}u_{li})D_{kl}\right)+\lambda\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ki}\log u_{ki}. \tag{2.11}$$

$$\underline{w}+\overline{w}=1.$$

$$d_{ki}=\|x_k-v_i\|^2,\ D_{kl}=\|x_k-x_l\|^2.$$

クラスタリングにおける制約は RCM-FU と同様で，以下のとおりである．

$$v_{ki} \in \{0, 1\}, \quad u_{ki} \in [0, 1], \ \forall k$$
$$\sum_{i=1}^{c} (v_{ki} + u_{ki}) = 1. \ \forall k$$

ERCM-FU の最適化問題は，この制約のもとでの (2.11) の最小化である．

最適解の導出に関する考え方は今までの手法と同じである．クラスタ中心の最適解を求めるために，(2.11) を v_i に関して偏微分する．

$$\frac{\partial J_{\text{ERCM-FU}}}{\partial v_i} = \sum_{k=1}^{n} \sum_{l=1}^{n} \nu_{li} u_{ki} \left(\underline{w}(x_l - v_i) + \overline{w}(x_k - v_i) \right) = 0.$$

この式を $\underline{w} + \overline{w} = 1$ であることも考慮して整理すると，

$$\sum_{k=1}^{n} \sum_{l=1}^{n} \nu_{li} u_{ki} v_i = \sum_{k=1}^{n} \sum_{l=1}^{n} \nu_{li} u_{ki} (\underline{w} x_l + \overline{w} x_k).$$

上式を変形すると，

$$\sum_{l=1}^{n} \nu_{li} \sum_{k=1}^{n} u_{ki} v_i = \underline{w} \sum_{k=1}^{n} u_{ki} \sum_{l=1}^{n} \nu_{li} x_l + \overline{w} \sum_{l=1}^{n} \nu_{li} \sum_{k=1}^{n} u_{ki} x_k.$$

ここで RCM-FU と同様に先ほどの制約より，

$$|\underline{A_i}| = \sum_{l=1}^{n} \nu_{li}.$$

この関係を先ほどの式に代入して次式を得る．

$$|\underline{A_i}| \sum_{k=1}^{n} u_{ki} v_i = \underline{w} \sum_{k=1}^{n} u_{ki} \sum_{x_k \in \underline{A_i}} x_k + \overline{w} |\underline{A_i}| \sum_{k=1}^{n} u_{ki} x_k.$$

従ってクラスタ中心の最適解は，

$$v_i = \underline{w} \times \frac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|} + \overline{w} \times \frac{\sum_{k=1}^{n} u_{ki} x_k}{\sum_{k=1}^{n} u_{ki}}.$$

$\underline{A_i} = \emptyset$，$\mathrm{Bnd}(A_i) = \emptyset$ である場合をそれぞれ $\overline{w} = 1$，$\underline{w} = 1$ とすると，最適解は次のようにまとめられる．

$$v_i = \begin{cases} \dfrac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|}, & (\mathrm{Bnd}(A_i) = \emptyset) \\ \dfrac{\sum_{k=1}^{n} u_{ki} x_k}{\sum_{k=1}^{n} u_{ki}}, & (\underline{A_i} = \emptyset) \\ \underline{w} \times \dfrac{\sum_{x_k \in \underline{A_i}} x_k}{|\underline{A_i}|} + \overline{w} \times \dfrac{\sum_{k=1}^{n} u_{ki} x_k}{\sum_{k=1}^{n} u_{ki}}. & (\text{otherwise}) \end{cases} \tag{2.12}$$

次に近似関係の最適解を導出する．今までの手法と同様に，データ x_k が下近似に属した場合と，境界領域に属した場合との目的関数を比較することで最適解を求める．考慮すべき条件は以下のとおりである．

- データ x_k がクラスタ A_{p_k} の下近似 $\underline{A}_{p_k}$ に属する場合．
- データ x_k が任意のクラスタ A_i の境界 $\mathrm{Bnd}(A_i)$ $\forall i$ に属する場合．

この条件は RCM-FU で最適解を求める際の条件と同じである．

$$p_k = \arg\min_i d_{ki}.$$

x_k が $\underline{A}_{p_k}$ に属するならば，目的関数は以下のように表される．

$$\underline{J}^k_{\text{ERCM-FU}} = \sum_{l=1, l\neq k}^{n} \left(u_{lp_k} (\underline{w} d_{kp_k} + \overline{w} d_{lp_k}) + 2\nu_{lp_k} D_{kl} + \lambda u_{lp_k} \log u_{lp_k} \right). \tag{2.13}$$

x_k が $\mathrm{Bnd}(A_i)$ に属するならば，目的関数は以下のように表される．

$$\overline{J}^k_{\text{ERCM-FU}} = \sum_{i=1}^{c} \sum_{l=1, l\neq k}^{n} \left(u_{ki} \nu_{li} (\underline{w} d_{li} + \overline{w} d_{ki}) + 2u_{ki} u_{li} D_{kl} \right) + \lambda \sum_{i=1}^{c} u_{ki} \log u_{ki}. \tag{2.14}$$

(2.13) と (2.14) の目的関数値を比較し，その値が小さくなる方に各データを帰属させると，N

と U についての最適解は以下のように求められる．

$$\nu_{ki} = \begin{cases} 1, & (J_k^{\nu} < J_k^{u} \wedge i = p_k) \\ 0, & (\text{otherwise}) \end{cases}$$

$$u_{ki} = \begin{cases} 0, & (J_k^{\nu} < J_k^{u} \wedge i = p_k) \\ \dfrac{\exp\left(-\frac{\beta_i}{\lambda}\right)}{\sum_{j=1}^{c} \exp\left(-\frac{\beta_j}{\lambda}\right)}. & (\text{otherwise}) \end{cases}$$

ただし，

$$\beta_i = \sum_{l=1}^{n} \left(\nu_{li} (\underline{w} d_{li} + \overline{w} d_{ki}) + 2 u_{li} D_{kl} \right).$$

境界領域 u_{ki} の最適解はエントロピー項が追加されているため，ラグランジュの未定乗数法を用いて導出できる．(2.11) より，ラグランジュ関数は以下のように定義される．

$$\begin{aligned} L_{\text{ERCM-FU}} = & \sum_{i=1}^{c} \sum_{k=1}^{n} \sum_{l=1}^{n} \left(\nu_{li} u_{ki} (\underline{w} d_{li} + \overline{w} d_{ki}) + (\nu_{ki} \nu_{li} + u_{ki} u_{li}) D_{kl} \right) \\ & + \lambda \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki} \log u_{ki} - \sum_{k=1}^{n} \gamma_k \left(\sum_{i=1}^{c} u_{ki} - 1 \right). \end{aligned}$$

この関数を u_{ki} に関して偏微分する．境界領域にデータが帰属する場合，そのデータが同時に下近似に属することはありえないので $\nu_{ki} = 0$ である．このことも考慮すると，

$$\frac{\partial L_{\text{ERCM-FU}}}{\partial u_{ki}} = \sum_{l=1, l \neq k}^{n} \left(\nu_{li} (\underline{w} d_{li} + \overline{d}_{ki}) + 2 u_{li} D_{kl} \right) + \lambda (\log u_{ki} + 1) - \gamma_k = 0.$$

この式より，

$$u_{ki} = \exp\left(\frac{- \sum_{l=1, l \neq k}^{n} \left(\nu_{li} (\underline{w} d_{li} + \overline{d}_{ki}) + 2 u_{li} D_{kl} \right) + \gamma_k}{\lambda} - 1 \right). \tag{2.15}$$

制約 $\sum_{j=1}^{c} u_{kj} = 1$ より，

$$\sum_{j=1}^{c} \exp\left(\frac{- \sum_{l=1, l \neq k}^{n} \left(\nu_{li} (\underline{w} d_{lj} + \overline{d}_{kj}) + 2 u_{lj} D_{kl} \right) + \gamma_k}{\lambda} - 1 \right) = 1,$$

$$\exp\left(\frac{\gamma_k}{\lambda} - 1 \right) = \frac{1}{\sum_{j=1}^{c} \exp\left(\frac{- \sum_{l=1, l \neq k}^{n} \left(\nu_{lj} (\underline{w} d_{lj} + \overline{d}_{kj}) + 2 u_{lj} D_{kl} \right)}{\lambda} \right)}.$$

この式を (2.15) に代入することで境界領域の最適解は得られる．

最後に ERCM-FU のアルゴリズムを示す．

Algorithm 4 RCM-FU

ERCM-FU1 初期近似関係を設定し，初期クラスタ中心を計算する．

ERCM-FU2 上記の手順を用いて下近似と境界を更新する．

ERCM-FU3 (2.12) を用いてクラスタ中心を更新する．

ERCM-FU4 収束条件を満たしたなら終了．そうでなければ **ERCM-FU2** に戻る．

2.3.4 ラフハード c-平均法

ラフハード c-平均法 (rough hard c-means, RHCM) は今までの RCM，RCM-FU，ERCM-FU のような rough k-means の最適解を基にした目的関数に基づくクラスタリング手法ではなく，ハード c-平均法の目的関数に下近似，境界領域という概念を盛り込んだラフクラスタリング手法である．RHCM も目的関数の最小化に基づき，クラスタ中心の最適化と近似関係の最適化を繰り返す交互最適化手法となっている．

RHCM の目的関数は次のように定義される．

$$J_{\mathrm{RHCM}}(N, U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\underline{w} v_{ki} + \overline{w} u_{ki}) d_{ki}. \tag{2.16}$$

$$\underline{w} + \overline{w} = 1.$$

$$d_{ki} = \|x_k - v_i\|^2.$$

RHCM のクラスタリングにおける制約は以下のとおりである．

$$
\begin{aligned}
v_{ki}, u_{ki} &\in \{0,1\}, \quad \forall k, i \\
\sum_{i=1}^{c} v_{ki} &\in \{0,1\}, \ \sum_{i=1}^{c} u_{ki} \neq 1, \ \forall k \\
\sum_{i=1}^{c} v_{ki} = 1 &\iff \sum_{i=1}^{c} u_{ki} = 0. \ \forall k
\end{aligned}
$$

これらの制約より，次の制約が得られる．

$$
\sum_{i=1}^{c} v_{ki} = 0 \iff \sum_{i=1}^{c} u_{ki} > 1. \ \forall k
$$

これらの制約は 2.2.1 で述べた (C1) – (C3) の制約と同等である．RHCM の最適化問題は，この制約のもとでの (2.16) の最小化である．

アルゴリズムを構築するために，まずクラスタ中心の最適解を導出していく．クラスタ中心の最適解は (2.16) を v_i に関して偏微分することで求められる．

$$
\frac{\partial J_{\mathrm{RHCM}}}{\partial v_i} = \sum_{k=1}^{n} (\underline{w} v_{ki} + \overline{w} u_{ki})(x_k - v_i) = 0.
$$

この式を整理すると，

$$
\sum_{k=1}^{n} (\underline{w} v_{ki} + \overline{w} u_{ki}) v_i = \sum_{k=1}^{n} (\underline{w} v_{ki} + \overline{w} u_{ki}) x_k.
$$

よって，クラスタ中心の最適解は，

$$
v_i = \frac{\sum_{k=1}^{n} (\underline{w} v_{ki} + \overline{w} u_{ki}) x_k}{\sum_{k=1}^{n} (\underline{w} v_{ki} + \overline{w} u_{ki})}. \tag{2.17}
$$

近似関係の最適解は今までの提案手法と同様の考え方を用い，データ x_k が下近似に属した場合と，境界領域に属した場合の目的関数値の比較をおこなう．考慮すべき条件は以下のとおりである．

- データ x_k がクラスタ A_{p_k} の下近似 $\underline{A}_{p_k}$ に属する場合.

- データ x_k が二つのクラスタ A_{p_k} と A_{q_k} の境界 $\mathrm{Bnd}(A_{p_k})$，$\mathrm{Bnd}(A_{q_k})$ に属する場合．

$$p_k = \arg\min_i d_{ki}, \quad q_k = \arg\min_{i \neq p_k} d_{ki}.$$

x_k が $\underline{A}_{p_k}$ に属するならば，目的関数は以下のように表される．

$$\underline{J}^k_{\mathrm{RHCM}} = \underline{w} d_{kp_k}. \tag{2.18}$$

x_k が $\mathrm{Bnd}(A_{p_k})$ と $\mathrm{Bnd}(A_{q_k})$ に属するならば，目的関数は以下のように表される．

$$\overline{J}^k_{\mathrm{RHCM}} = \sum_{i=p_k,q_k} \overline{w} d_{ki}. \tag{2.19}$$

(2.18) と (2.19) で得られる値を比較し，その値が小さくなる方に各データを帰属させる．N と U についての最適解は以下のように求められる．

$$v_{ki} = \begin{cases} 1, & (\underline{J}^k_{\mathrm{RHCM}} < \overline{J}^k_{\mathrm{RHCM}} \wedge i = p_k) \\ 0. & (\text{otherwise}) \end{cases}$$

$$u_{ki} = \begin{cases} 1, & (\underline{J}^k_{\mathrm{RHCM}} \geq \overline{J}^k_{\mathrm{RHCM}} \wedge (i = p_k \vee i = q_k)) \\ 0. & (\text{otherwise}) \end{cases}$$

最後に RHCM のアルゴリズムを示す．

Algorithm 5 RHCM

RHCM1 初期クラスタ中心を設定する．

RHCM2 上記の手順を用いて下近似と境界を更新する．

RHCM3 (2.17) を用いてクラスタ中心を更新する．

RHCM4 収束条件を満たしていたら終了．そうでなければ **RHCM2** へ戻る．

2.3.5 ラフファジィc-平均法

RHCMもRCM同様にハードラフクラスタリングであるため，データが三つ以上のクラスタの境界領域に属することはない．そこでより柔軟な分類がおこなえるように，本節ではRHCMの境界領域をファジィ化したラフファジィc-平均法 (rough fuzzy c-means, RFCM) を構築する．ここでのファジィ化はBezdekによる標準的なファジィ化を指す．

RFCMの目的関数は次のように定義される．

$$J_{\mathrm{RFCM}}(N,U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(\underline{w}v_{ki}+\overline{w}u_{ki}^{m})d_{ki}. \tag{2.20}$$

$$\underline{w}+\overline{w}=1.$$

$$d_{ki}=\|x_k-v_i\|^2.$$

クラスタリングにおける制約は以下のとおりである．

$$v_{ki},u_{ki}\geq 0,\ \forall k,i$$

$$\sum_{i=1}^{c}(v_{ki}+u_{ki})=1.\ \forall k$$

RFCMの最適化問題は，この制約のもとでの (2.20) の最小化である．

これまでのアルゴリズムと同様に，クラスタ中心と近似関係の最適解を導出していく．クラスタ中心の最適解は (2.20) の v_i に関する偏微分によって得られる．

$$\frac{\partial J_{\mathrm{RFCM}}}{\partial v_i}=\sum_{k=1}^{n}(\underline{w}v_{ki}+\overline{w}u_{ki}^{m})(x_k-v_i)=0.$$

この式を整理すると，

$$\sum_{k=1}^{n}(\underline{w}v_{ki}+\overline{w}u_{ki}^{m})v_i=\sum_{k=1}^{n}(\underline{w}v_{ki}+\overline{w}u_{ki}^{m})x_k.$$

以上より，クラスタ中心の最適解は，

$$v_i=\frac{\sum_{k=1}^{n}(\underline{w}v_{ki}+\overline{w}u_{ki}^{m})x_k}{\sum_{k=1}^{n}(\underline{w}v_{ki}+\overline{w}u_{ki}^{m})}. \tag{2.21}$$

これまでの手法と同様に，目的関数の比較より近似関係の最適解を求めるための条件を考える．

- データ x_k がクラスタ A_{p_k} の下近似 $\underline{A}_{p_k}$ に属する場合．
- データ x_k が任意のクラスタ A_i の境界 $\mathrm{Bnd}(A_i)$ $\forall i$ に属する場合．

$$p_k = \arg\min_i d_{ki}.$$

x_k が $\underline{A}_{p_k}$ に属するならば，目的関数は以下のように表される．

$$\underline{J}^k_{\mathrm{RFCM}} = \underline{w} v_{ki} \|x_k - v_{p_k}\|^2. \tag{2.22}$$

x_k が $\mathrm{Bnd}(A_i)$ に属するならば，目的関数は以下のように表される．

$$\overline{J}^k_{\mathrm{RFCM}} = \sum_{i=1}^{c} \overline{w} u^m_{ki} \|x_k - v_i\|^2. \tag{2.23}$$

(2.22) と (2.23) の目的関数値を比較し，その値が小さくなる方に各データを帰属させ最適解を求める．最終的に N と U についての最適解は以下のように求められる．

$$v_{ki} = \begin{cases} 1, & (\underline{J}^k_{\mathrm{RFCM}} < \overline{J}^k_{\mathrm{RFCM}} \wedge i = p_k) \\ 0. & (\text{otherwise}) \end{cases}$$

$$u_{ki} = \begin{cases} \dfrac{\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c} \left(\frac{1}{d_{kj}}\right)^{\frac{1}{m-1}}}, & (\underline{J}^k_{\mathrm{RFCM}} \geq \overline{J}^k_{\mathrm{RFCM}}) \\ 0. & (\text{otherwise}) \end{cases}$$

境界領域 u_{ki} の最適解はラグランジュの未定乗数法を用いて求められる．RFCM のラグランジュ関数は (2.20) より以下のように表される．

$$L_{\mathrm{RFCM}} = \sum_{k=1}^{n} \sum_{i=1}^{c} \left(\underline{w} v_{ki} + \overline{w} u^m_{ki} \right) d_{ki} - \sum_{k=1}^{n} \lambda_k \left(\sum_{i=1}^{c} u_{ki} - 1 \right).$$

この関数を u_{ki} に関して偏微分する．境界領域にデータが帰属する場合，そのデータが同時に下近似に属することはありえないので $v_{ki} = 0$ であることを考慮すると，

$$\frac{\partial L_{\mathrm{RFCM}}}{\partial u_{ki}} = m u^{m-1}_{ki} d_{ki} - \lambda_k = 0.$$

この式より，

$$u_{kj} = \left(\frac{\lambda_k}{d_{kj}}\right)^{\frac{1}{m-1}}. \tag{2.24}$$

この式の両辺を $j = 1, \ldots, c$ について加え，制約 $\sum_{j=1}^{c} u_{kj} = 1$ を用いると，

$$\sum_{j=1}^{c}\left(\frac{\lambda_k}{d_{kj}}\right)^{\frac{1}{m-1}} = 1,$$

$$\lambda_k^{\frac{1}{m-1}} = \frac{1}{\sum_{j=1}^{c}\left(\frac{1}{d_{kj}}\right)}.$$

この式を (2.24) に代入することで境界領域の最適解は得られる．

RFCM のアルゴリズムは以下のとおりである．

Algorithm 6 RFCM

RFCM1 初期クラスタ中心を設定する．

RFCM2 上記の手順を用いて下近似と境界を更新する．

RFCM3 (2.21) を用いてクラスタ中心を更新する．

RFCM4 収束条件を満たしていたら終了．そうでなければ **RFCM2** へ戻る．

2.4 数値例

本節では，提案手法と関連手法である rough k-means との比較を数値実験を通しておこない，提案手法の有効性の検討と特徴把握について論じる．実験データには人工データ一種類と実データ二種類を用いる．人工データとしては図 2.2 に示した二次元データを用い，提案手法の分類特徴について考察する．実データは UCI Repository より，ベンチマークデータである Iris データ [40] と Breast Cancer データ [41] を用いている．各実データの詳細は下記の表に記載されている．Breast Cancer データは本来 699 のデータからなるデータセットであるが，

本実験では重複データや欠損データをすべて削除した上で用いているので，データ数が減少している．これらのベンチマークデータは予めいくつかの正しいクラスラベルが与えられているため，クラスラベルと，実際にクラスタリングをおこなった際の結果とを比較することで，各手法の優劣を決めることができる．そこでこれらの実データを用いて，関連研究である rough k-means と提案手法との分類精度について考察する．

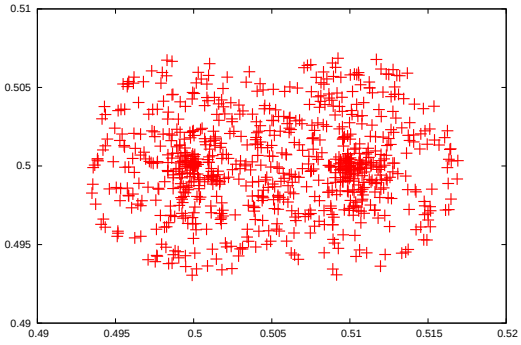

図 2.2: 人工データ

表 2.1: Iris データ

データ数	150
属性数	4
クラス数	3

表 2.2: Breast Cancer データ

データ数	449
属性数	9
クラス数	2

本数値例実験ではすべてのデータセットをすべての軸について $[0, 1]$ の空間上のデータ点として扱っている．本来，実データである Iris データと Breast Cancer データはこの範囲に収まるデータセットではない．しかし，rough k-means ではパラメータ threshold を設定する必要があり，データセットごとに適切なスケールのパラメータを設定するのは困難である．そこで本数値例実験では予めこれらのデータを正規化して用い，パラメータの設定を簡略化している．

また rough k-means の初期値依存性を減らすために，本研究ではハード c-平均法の目的関数

を基に評価関数 (2.25) の構築をおこない，その評価関数が最も小さくなる解を，すべての初期値から得られた解の中での最適解として用いている．

$$J_{\mathrm{RKM}}(U,N,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}\left(v_{ki}\|x_k-v_i\|^2+\frac{u_{ki}}{\sum_{i=1}^{c}u_{ki}}\|x_k-v_i\|^2\right). \tag{2.25}$$

2.4.1 人工データに対するクラスタリング結果

ここでは，人工データに対して提案手法を用いて分類をおこなった結果を示す．各結果ではこのデータセットを二つのクラスタに分割するよう設定し，クラスタリングをおこなっている．

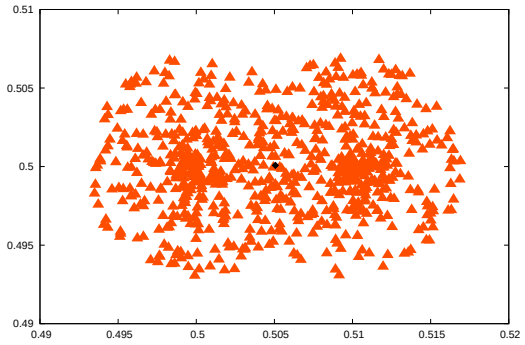

図 2.3: 人工データに対して RCM を用いた結果 ($\underline{w}$=0.05)

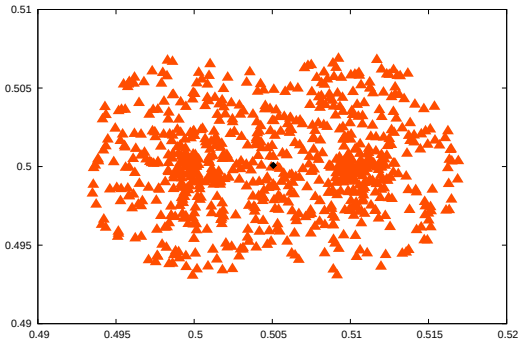

図 2.4: 人工データに対して RCM を用いた結果 ($\underline{w}$=0.25)

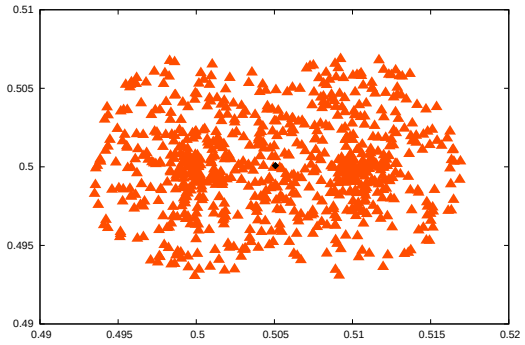

図 2.5: 人工データに対して RCM を用いた結果 (w=0.55)

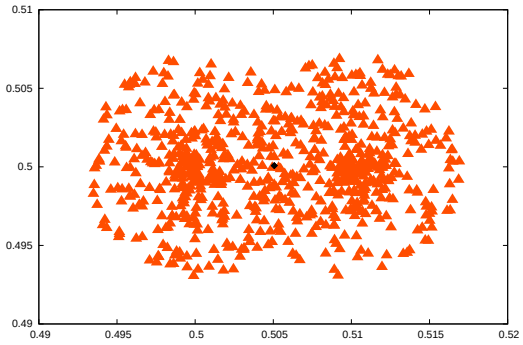

図 2.6: 人工データに対して RCM を用いた結果 (w=0.75)

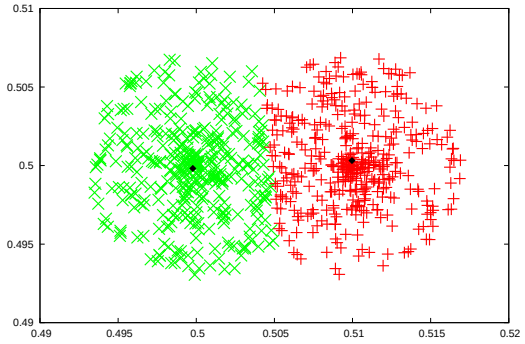

図 2.7: 人工データに対して RCM を用いた結果 (w=0.9)

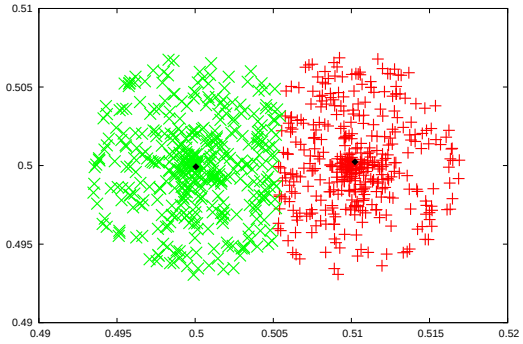

図 2.8: 人工データに対して RCM を用いた結果 (w=0.95)

図 2.3 – 2.8 は人工データに対して RCM を用いてクラスタリングをおこなった結果である．図で黒い菱型で描かれている点がクラスタ中心であり，それ以外の点がデータ点となっている．図 2.3 – 2.6 の結果はデータ点がすべて橙の三角で描かれているが，これは二つのクラスタの境界に帰属する点を示しており，これらの結果ではすべてのデータ点が境界に属すことがわかる．また，二つのクラスタ中心もほとんど一致しており，全体が一つのクラスタのようになってしまっている．図 2.7，図 2.8 では，各データは赤い十字か緑のクロスで描かれて

いる．これは赤い点，緑の点それぞれがクラスタの下近似に属することを意味している．これらのことから，RCM では下近似係数を大きくしていくことで下近似に含まれるデータ数も増加していく傾向があることが予想できる．

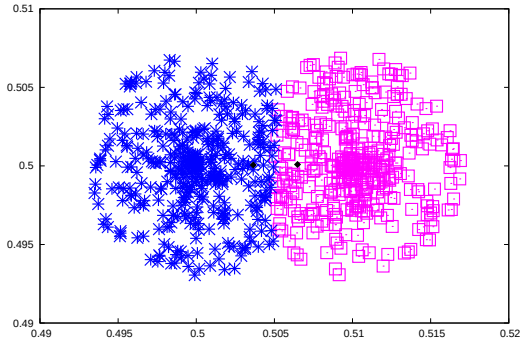

図 2.9: 人工データに対して RCM-FU を用いた結果 ($\underline{w}$=0.05)

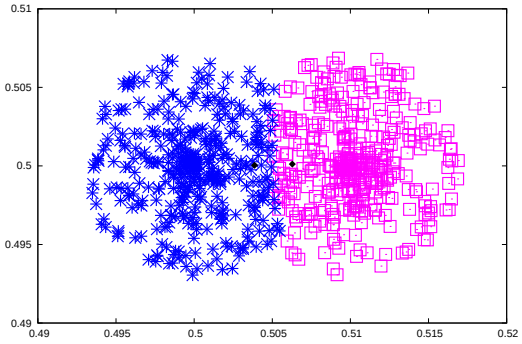

図 2.10: 人工データに対して RCM-FU を用いた結果 ($\underline{w}$=0.3)

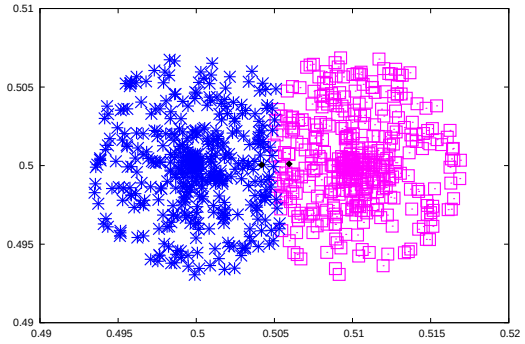

図 2.11: 人工データに対して RCM-FU を用いた結果 ($\underline{w}$=0.5)

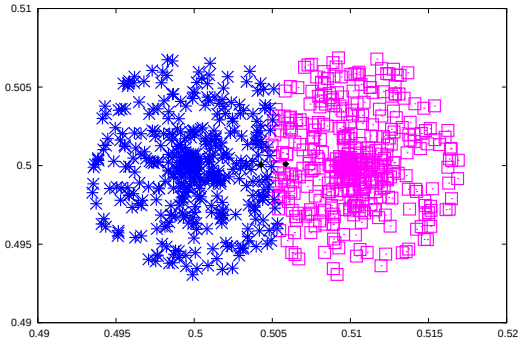

図 2.12: 人工データに対して RCM-FU を用いた結果 ($\underline{w}$=0.55)

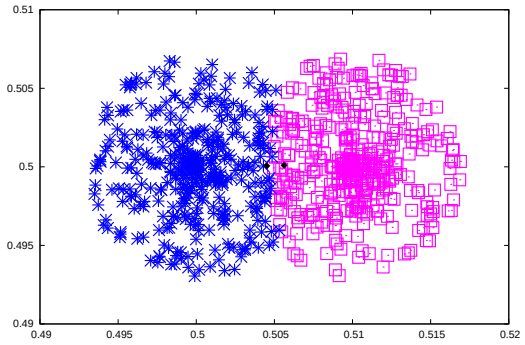

図 2.13: 人工データに対して RCM-FU を用いた結果 ($\underline{w}$=0.7)

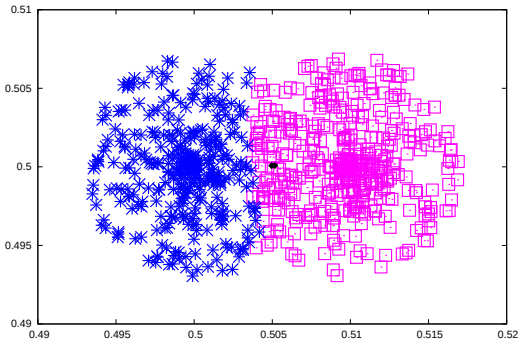

図 2.14: 人工データに対して RCM-FU を用いた結果 ($\underline{w}$=0.95)

図 2.9 – 2.14 は人工データに対して RCM-FU を用いてクラスタリングをおこなった結果である．すべての図において，各データ点は青い星もしくはピンクの四角で描かれている．これらの点はすべて二つのクラスタの境界に属する点である．境界に属する点の中でも，左側のクラスタへの帰属度合いが強い点が青い星，右側のクラスタへの帰属度合いが強い点がピンクの四角で描かれている．$\underline{w}$ の値を変化させても，すべての点が境界に分類され，クラスタ中心にもほとんど変化がない．このことから，クラスタ同士が近いデータを用いる場合，RCM-FU では各データ点を境界に分類しやすいと考えられる．

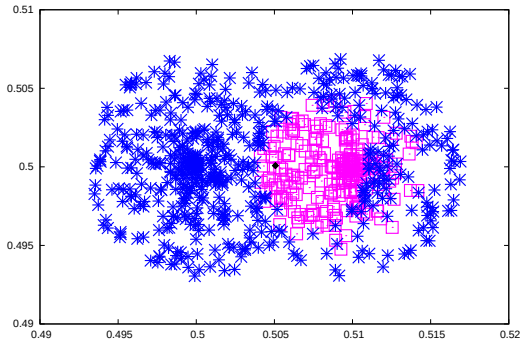

図 2.15: 人工データに対して ERCM-FU を用いた結果 ($\underline{w}$=0.05)

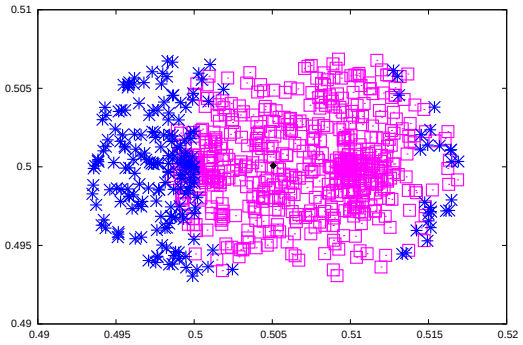

図 2.16: 人工データに対して ERCM-FU を用いた結果 ($\underline{w}$=0.3)

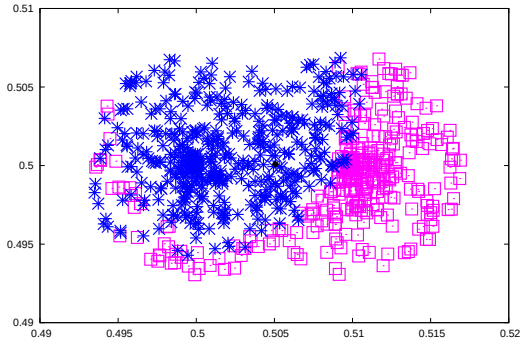

図 2.17: 人工データに対して ERCM-FU を用いた結果 ($\underline{w}$=0.55)

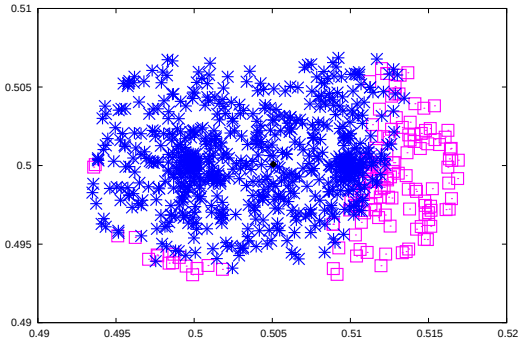

図 2.18: 人工データに対して ERCM-FU を用いた結果 ($\underline{w}$=0.7)

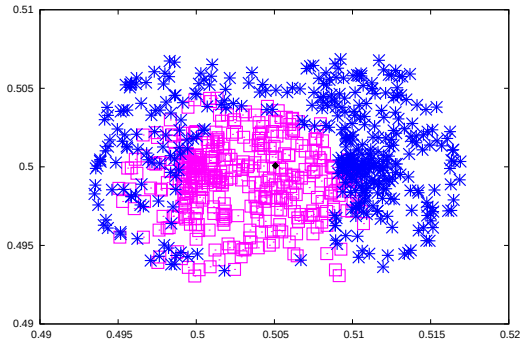

図 2.19: 人工データに対して ERCM-FU を用いた結果 ($\underline{w}$=0.85)

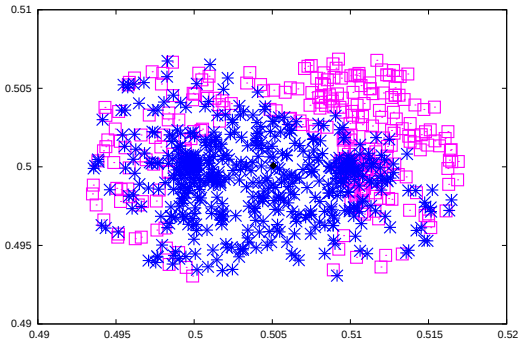

図 2.20: 人工データに対して ERCM-FU を用いた結果 ($\underline{w}$=0.95)

図 2.15 – 2.20 は人工データに対して ERCM-FU を用いてクラスタリングをおこなった結果である．RCM-FU の結果と同様にすべてのデータ点が境界に分類されていることがわかる．RCM-FU と異なり，不自然な分類になっているように見えるが，これはクラスタ中心が RCM と同様にほぼ一致しているためである．実際データ点の各クラスタへの帰属度はどれもほぼ 0.5 であり，帰属度による差はほとんど無い．このことから，人工データに対しては ERCM-FU はほぼ一つのクラスタを生成しているといえ，下近似係数の変化による分類結果の違いもほとんど見られない．

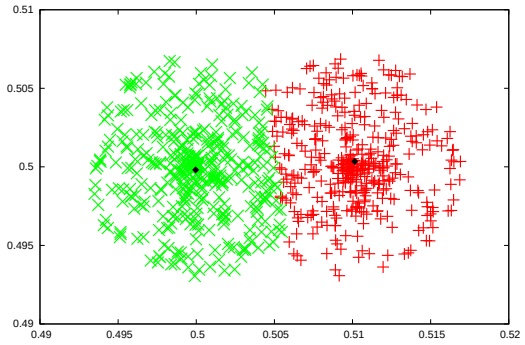

図 2.21: 人工データに対して RHCM を用いた結果 ($\underline{w}$=0.3)

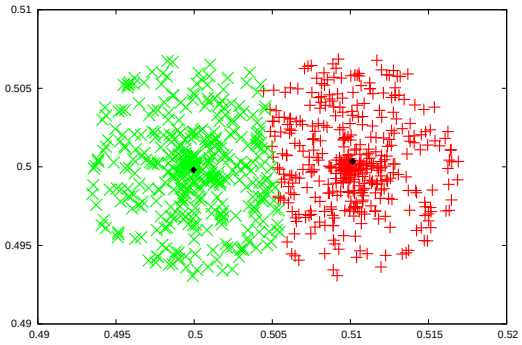

図 2.22: 人工データに対して RHCM を用いた結果 ($\underline{w}$=0.55)

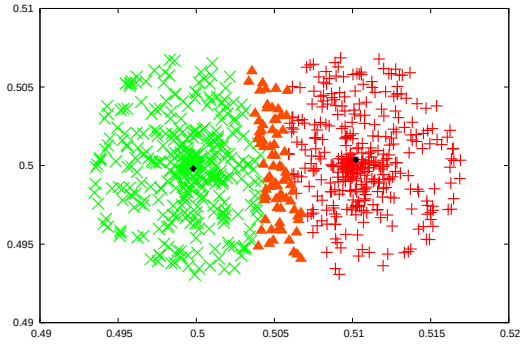

図 2.23: 人工データに対して RHCM を用いた結果 ($\underline{w}$=0.75)

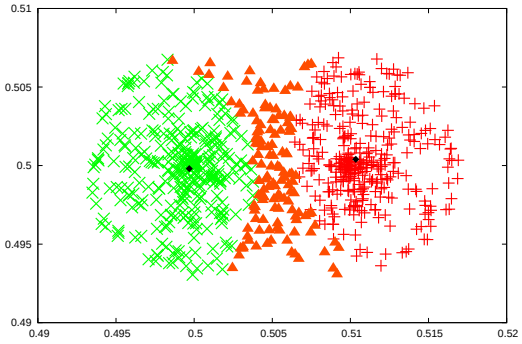

図 2.24: 人工データに対して RHCM を用いた結果 ($\underline{w}$=0.8)

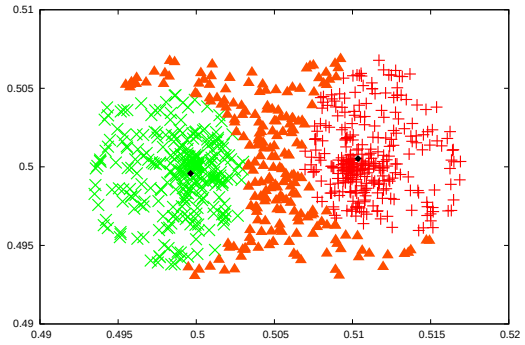

図 2.25: 人工データに対して RHCM を用いた結果 (w=0.85)

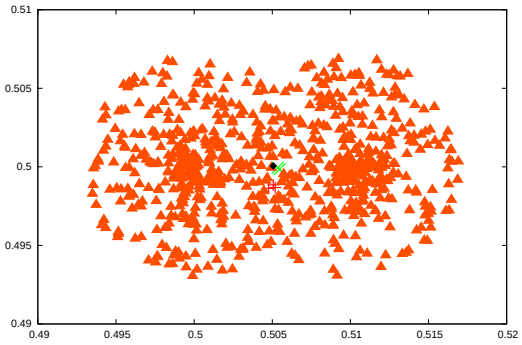

図 2.26: 人工データに対して RHCM を用いた結果 (w=0.9)

図 2.21 – 2.26 は人工データに対して RHCM を用いてクラスタリングをおこなった結果である．これまでの手法と異なり下近似係数を徐々に増加させていくと，境界領域に含まれるデータ点が多くなっていく様子が見て取れる．そして下近似係数を増加させると，境界に含まれるデータは下近似に含まれるデータを囲むように広がっていき，最終的にはほぼすべての点が境界に属する．

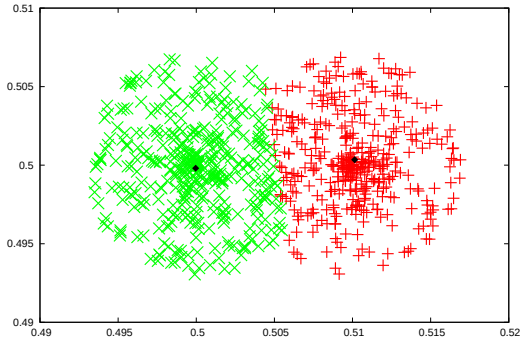

図 2.27: 人工データに対して RFCM を用いた結果 (w=0.3)

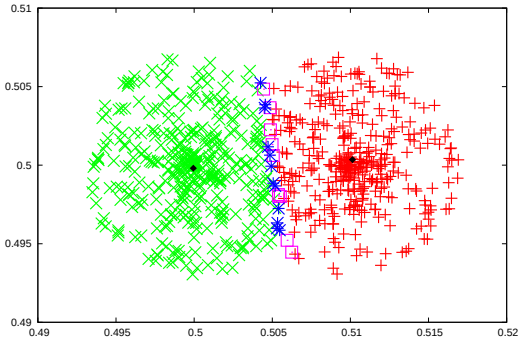

図 2.28: 人工データに対して RFCM を用いた結果 (w=0.35)

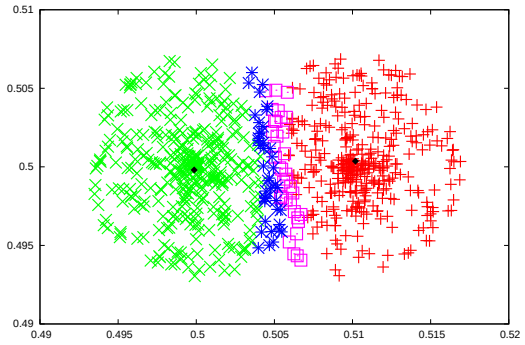

図 2.29: 人工データに対して RFCM を用いた結果 ($\underline{w}$=0.4)

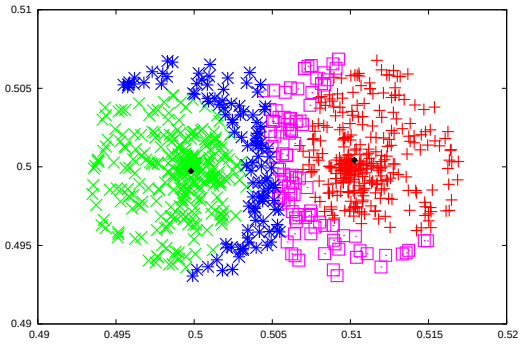

図 2.30: 人工データに対して RFCM を用いた結果 ($\underline{w}$=0.45)

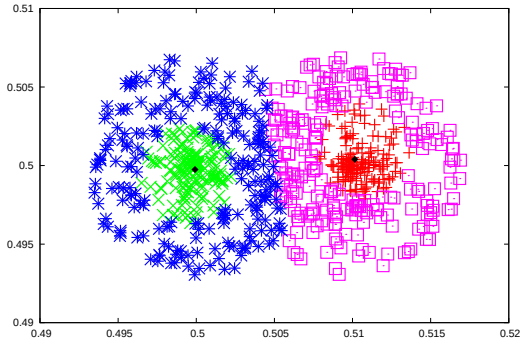

図 2.31: 人工データに対して RFCM を用いた結果 ($\underline{w}$=0.48)

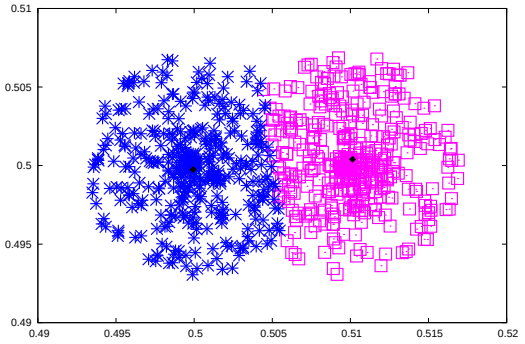

図 2.32: 人工データに対して RFCM を用いた結果 ($\underline{w}$=0.5)

図 2.27 – 2.32 は人工データに対して RFCM を用いてクラスタリングをおこなった結果である．RHCM の結果と同様に下近似係数 $\underline{w}$ を徐々に増加させていくと，境界領域に含まれるデータ点が多くなっていく様子が見て取れる．また RHCM と同様に下近似係数を増加させると，境界に含まれるデータ点が下近似に含まれるデータ点を囲むように広がっていく．しかしながら境界領域のデータ点の増加量は RHCM よりも大きく，下近似係数が 0.5 以上になるとすべてのデータ点が境界に属するという結果になった．すべてのデータ点が境界に属する

ようになっても，左右のクラスタは明確に分離されていることがわかる．

2.4.2 Iris データに対するクラスタリング結果

次に実データの一つである Iris データに対してクラスタリングをおこなった結果を二種類載せる．一つ目は提案手法と関連手法である rough k-means との比較を正答率の観点からおこなった表であり，もう一つは各提案手法について下近似係数 $\underline{w}$ を 0.05 ずつ変化させた際の，下近似及び境界に含まれるデータ数の推移を表したグラフである．

正答率の計算は Iris データに元々割り振られているラベルと一致している場合のデータ数をカウントし，全データ数で割るという方針を取る．ただしラフクラスタリングには下近似と境界という概念があり，データが境界に属する場合そのデータは複数のクラスタの境界に属することとなる．そのため，データが境界に属する場合の扱いを別に考えなくてはいけない．

そこで本研究では，ハードラフクラスタリング手法であればデータが元々のラベルのクラスタを含む境界に属する場合，ファジィラフクラスタリング手法であればそのラベルのクラスタの境界への帰属度が最も高い場合，正しい分類がおこなわれたと判断する．

表 2.3: Iris データに対するクラスタリング結果の正答率

Algorithm	下近似			境界			全体	
	データ数	正答数	割合	データ数	正答数	割合	正答数	割合
RCM ($\underline{w}$ = 0.75)	150	135	0.9	0	0	—	135	0.9
RCM-FU ($\underline{w}$ = 0.75, m = 2.0)	74	54	0.73	76	65	0.855	119	0.793
ERCM-FU ($\underline{w}$ = 0.75, λ = 2.0)	150	130	0.867	0	0	—	130	0.867
RHCM ($\underline{w}$ = 0.75)	139	128	0.921	11	11	1.0	139	0.923
RFCM ($\underline{w}$ = 0.75, m = 2.0)	0	0	—	150	134	0.893	134	0.893
RKM ($\underline{w}$ = 0.75, threshold = 0.01)	145	131	0.903	5	5	1.0	136	0.907
RKM ($\underline{w}$ = 0.75, threshold = 0.1)	120	115	0.958	30	30	1.0	145	0.967
RKM ($\underline{w}$ = 0.75, threshold = 0.5)	114	50	0.439	36	28	0.778	78	0.52

表 2.3 は Iris データに対して各クラスタリング手法を用いてクラスタリングをおこなった場合の結果である．この表から関連手法である rough k-means はパラメータ threshold の値により結果が大きく変化することがわかり，データセット毎にふさわしいパラメータ値を発見することが大きな課題であることもわかる．今回の実験では $\underline{w} = 0.75$ とした場合における，正答率の比較をおこなったが，これは，Lingras らの論文 [17] での実験において，最適な結果を得たパラメータ設定であることを参考にしている．提案手法ではほとんどの結果が，下近似のみにデータが分類されるという形となった．正答率という観点から述べると，このパラメータ設定では RHCM での結果が提案手法の中では最も良い結果を示している．

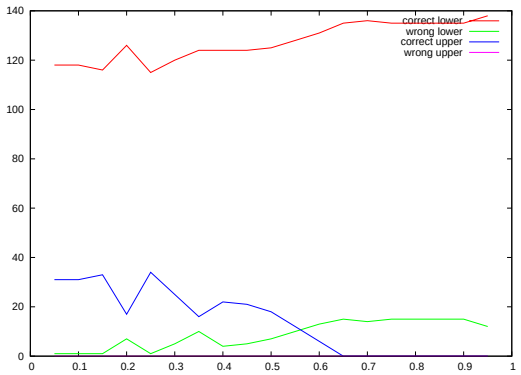

図 2.33: Iris データに対して RCM を用いた場合の下近似係数の変化による分類結果の変遷

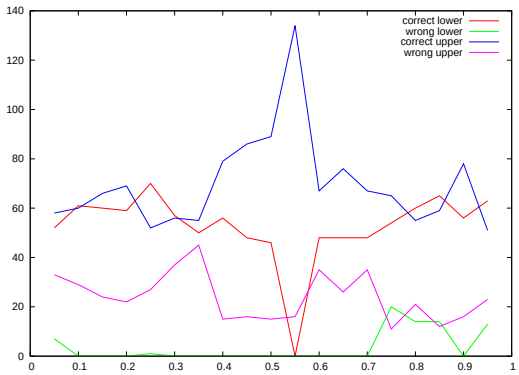

図 2.34: Iris データに対して RCM-FU を用いた場合の下近似係数の変化による分類結果の変遷

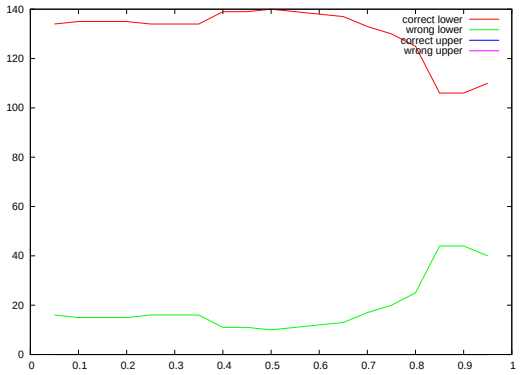

図 2.35: Iris データに対して ERCM-FU を用いた場合の下近似係数の変化による分類結果の変遷

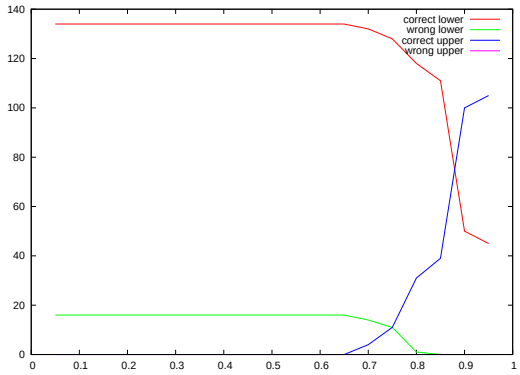

図 2.36: Iris データに対して RHCM を用いた場合の下近似係数の変化による分類結果の変遷

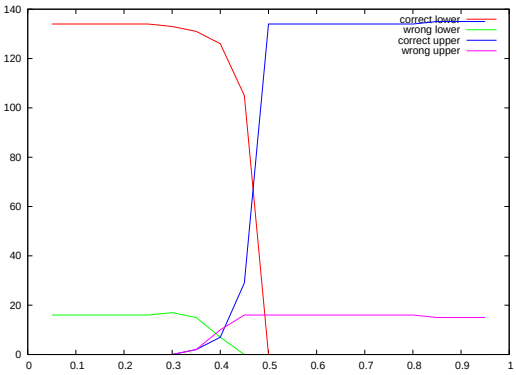

図 2.37: Iris データに対して RFCM を用いた場合の下近似係数の変化による分類結果の変遷

図 2.33 – 2.37 は，下近似係数 $\underline{w}$ を 0.05 から 0.95 まで 0.05 刻みに変化させた際の，各提案手法における正解分類数と誤分類数の変遷である．各図において，赤で示される線は下近似に所属するデータの内正しい下近似に分類されたデータ数，緑で示される線は下近似に帰属するデータの内間違った下近似に分類されたデータ数，青で示される線は境界に帰属するデータの内正しい境界に分類されたデータ数，ピンクで示される線は境界に帰属するデータの内間違った境界に分類されたデータ数をそれぞれ示している．

図 2.33 より RCM を用いた場合，下近似係数を大きくする程下近似に含まれるデータ数は増加していくことがわかる．この傾向は，人工データを用いた場合の傾向と同様であることも伺える．RCM-FU と ERCM-FU を用いた際の結果は，図 2.34，図 2.35 である．RCM-FU の結果からは下近似係数の変化による傾向は特に見いだせず，人工データの結果であるすべてが境界に含まれるといった傾向も見られない．このことから，用いるデータセットに分類結果が依存すると考えられる．ERCM-FU の結果からは下近似係数が大きくなると正答率が下がる傾向が見て取れる．しかし人工データを用いた際の結果ではすべてのデータが境界に分類されたのに対して，Iris データを用いた場合の結果はすべてのデータが下近似に分類された．このことから，RCM-FU 同様分類結果は用いるデータセットに分類結果が大きく依存すると考えられる．図 2.36，図 2.37 は RHCM と RFCM による分類結果である．これらの結果は RCM の結果と異なり，下近似係数が大きくなると境界へ分類されるデータが多くなることを示している．特に，RHCM は下近似係数が 0.65 を超えたあたりから境界に分類されるデータが出始め，RFCM は下近似係数が 0.3 を超えたあたりから境界に分類されるデータが出始めている．

2.4.3 Breast Cancer データに対するクラスタリング結果

最後に Breast Cancer データに対して提案手法と，関連手法を用いて分類をおこなった結果を示す．

表 2.4: Breast Cancer データに対するクラスタリング結果の正答率

Algorithm	下近似			境界			全体	
	データ数	正答数	割合	データ数	正答数	割合	正答数	割合
RCM ($\underline{w}$ = 0.75)	0	0	—	449	449	1.0	449	1.0
RCM-FU ($\underline{w}$ = 0.75, m = 2.0)	0	0	—	449	417	0.929	417	0.929
ERCM-FU ($\underline{w}$ = 0.75, λ = 2.0)	328	315	0.96	121	107	0.884	422	0.94
RHCM ($\underline{w}$ = 0.75)	388	375	0.966	61	61	1.0	436	0.971
RFCM ($\underline{w}$ = 0.75, m = 2.0)	0	0	—	449	414	0.922	414	0.922
RKM ($\underline{w}$ = 0.75, threshold = 0.01)	449	416	0.927	0	0	—	416	0.927
RKM ($\underline{w}$ = 0.75, threshold = 0.1)	431	411	0.954	18	18	1.0	429	0.955
RKM ($\underline{w}$ = 0.75, threshold = 0.5)	294	167	0.568	155	155	1.0	322	0.717

表 2.4 は各手法を用いて，Breast Cancer データを分類した結果の正答率を示している．Iris データセットに対する結果と同様に，rough k-means はパラメータ threshold の値によって結果が大きく変化することがわかる．この表から提案手法の中では，RCM はすべてのデータを正しく分類しているという結果となっている．Breast Cancer データはニクラスタからなるデータセットである．そのため，本実験においてすべてのデータが境界に属するということは，必然的にすべてのデータを正しく分類していることを意味する．

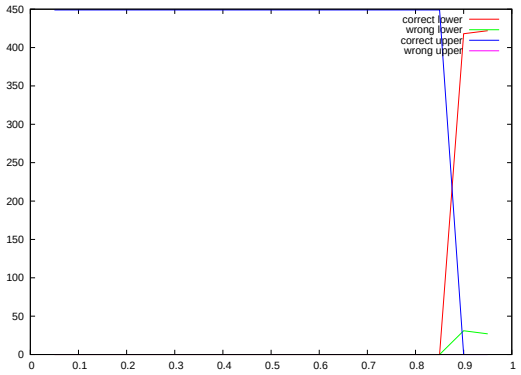

図 2.38: Breast Cancer データに対して RCM を用いた場合の下近似係数の変化による分類結果の変遷

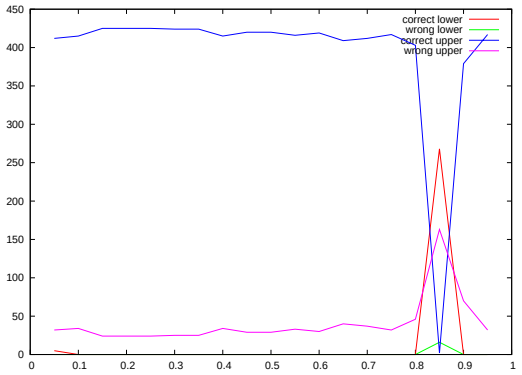

図 2.39: Breast Cancer データに対して RCM-FU を用いた場合の下近似係数の変化による分類結果の変遷

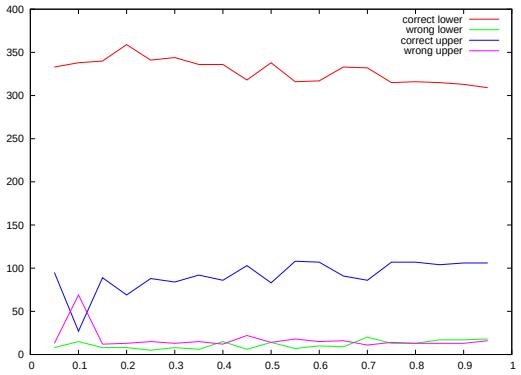

図 2.40: Breast Cancer データに対して ERCM-FU を用いた場合の下近似係数の変化による分類結果の変遷

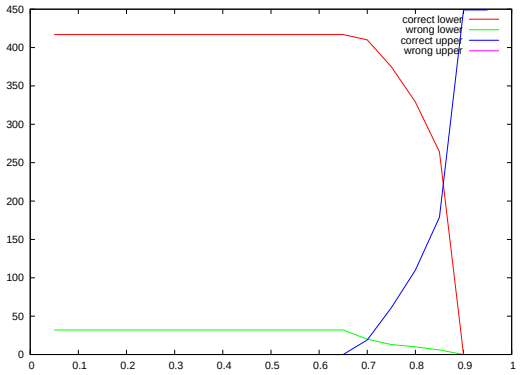

図 2.41: Breast Cancer データに対して RHCM を用いた場合の下近似係数の変化による分類結果の変遷

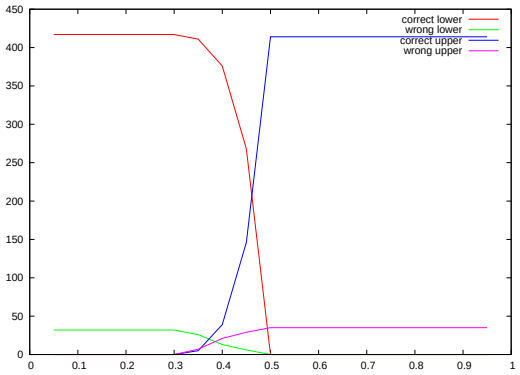

図 2.42: Breast Cancer データに対して RFCM を用いた場合の下近似係数の変化による分類結果の変遷

図 2.38 – 2.42 は，下近似係数 $\underline{w}$ を 0.05 から 0.95 まで 0.05 刻みに変化させた際の，各提案手法における正解分類数と誤分類数の変遷である．各図における線の意味するところは，Iris データの場合と同じである．

図 2.38 より RCM による分類では，下近似係数を大きくすることで下近似に含まれるデータ数が増加することがわかる．この傾向は程度の差こそあるが人工データ，Iris データにもあてはまることであり，RCM では下近似係数が増加するにつれデータの下近似への分類数も増加するという特徴を持つと考えられる．RCM-FU と ERCM-FU による結果は，図 2.39，図 2.40 である．これらの結果はほとんどの下近似係数において，正答数が一定の値を保っていることが伺える．RCM と異なり，パラメータの変化による分類の傾向はこれらの手法ではどのデータセットを用いた場合にも顕著ではなく，データセット毎に異なる．そのためこれらの手法における分類結果は，パラメータよりもデータセットによって定まると考えられる．図 2.41，図 2.42 は RHCM と RFCM による分類結果である．これらの結果は Iris データに対しておこなった結果である図 2.36，図 2.37 と同じ傾向を示している．またこの傾向は人工データを用いた場合の結果とも同様であり，これら二手法と RCM に関しては分類結果のデータセットへの依存性は低く，下近似係数に大きく依存すると考えられる．

第3章　データ自身に含まれる不確実性ーデータ自身を確率密度関数として扱うEMアルゴリズムに基づくクラスタリング

クラスタリングにおいてデータに含まれる不確実性を扱う方法としては無視する，代替データを設定する [26]，データを区間データとして扱う [27,28]，データ自身に不確実性の確率密度関数を仮定する [29] などが挙げられる．これらの考え方について述べていくと，無視するというのは不確実性をそもそも考慮しなかったりそのようなデータを削除してしまう方針である．クラスタリング手法の多くでは不確実性の扱いを考慮していないので，このような方針が取られる．

次に不確実性を考慮する方法について述べていく．まず代替データを設定する方法である．これは Hathaway ら [26] によって提案された欠損データを扱う場合の方法であり，初期値として欠損データに予め値を設定した上で，ファジィc-平均法を用いてクラスタ中心と帰属度を求める．通常のファジィc-平均法ではクラスタ中心と帰属度の最適化を解が収束するまで繰り返し，目的関数が最小となる場合を最適解として用いるが，この方法ではそれらの最適化の後に，欠損データが属すクラスタのクラスタ中心の値を用いて欠損データを補完する段階が付け加えられている．つまりクラスタ中心と帰属度，欠損データの 3 つを交互に最適化する手法となっている．

データを区間データとして扱う代表的な手法として，高田らの手法 [27,28] が挙げられる．この手法では各データに不確実性としての上限と下限を定め区間データと定義し，それをファ

ジィc-平均法に適用している．区間データを使用した際のクラスタ中心の最適化はファジィc-平均法の導出過程と異なるため，高田らは区間データとクラスタ中心との非類似度として最短距離と最長距離を用い，クラスタ中心の最適解を導出するニ種類の手法を提案している．村田ら [30] によって提案された許容範囲もデータの不確実性を扱うことのできる概念であり，村田らはこの許容範囲をファジィc-平均法に導入した新たなクラスタリング手法を提案している．許容範囲も区間データと同様にデータに含まれる不確実性の範囲に制限があり，このような範囲を持つデータは許容範囲つきデータと呼ばれる．区間データではデータを集合として表しているため，非類似度は集合間の距離を用いて定義される．一方，許容範囲つきデータでは許容ベクトルと呼ばれるベクトルを新たに定義し，そのベクトルの変動領域を，不確実性の範囲の限界という制約条件を満たすように最適化問題より導出する．つまり，目的関数に許容ベクトルを制約条件に不確実性の範囲の限界を導入している．そのため，許容ベクトルを導入したファジィc-平均法ではクラスタ中心，帰属度，許容ベクトルの交互最適化から成り立つアルゴリズムとなっている．この概念はハードc-平均法 [31] やファジィ回帰 [32] など様々な手法に適用されている．許容範囲と類似した概念に遠藤らによって提案された不確実性ベクトル [33] がある．この概念は，データの持つ不確実性をベクトルとして定義しそれを不確実性ベクトルと呼ぶ．しかし許容範囲と異なり，不確実性ベクトルは範囲に制限を設けない．代わりに目的関数に不確実性ベクトルの大きさを制御する正則化項を導入する．この正則化項は，不確実性ベクトルが大きくなり過ぎないようにペナルティとして働く．以上のことより，不確実性ベクトル導入によって考慮すべきことは，目的関数への不確実性ベクトルと正則化項の導入のみである．そのため，最適解導出のプロセスが許容範囲よりも簡素なので，許容範囲よりも応用性や親和性が高いと考えられる．

本章ではデータ自身に含まれる不確実性を扱うために，まずデータ自身に確率密度関数を考慮した EM アルゴリズムに基づくクラスタリング手法を構築する．次章では，不確実性ベクトルを導入した EM アルゴリズムに基づくクラスタリング手法を構築する．

一般的なEMアルゴリズムに基づくクラスタリング手法では，各データを各データがその点で必ず生起するデルタ関数であるとみなすことができる．一方本章で提案する手法では，各データはその点を平均とした一定の分散を含んだガウス分布として表現することを特徴とする．本章ではまずはじめに関連手法であるEMアルゴリズムに基づくクラスタリングについて紹介し，その後提案手法について述べていく．

3.1 関連手法

3.1.1 EMアルゴリズムに基づくクラスタリング

EMアルゴリズム [11] は代表的な最尤推定法である．EMアルゴリズムに基づくクラスタリングは各データが複数のクラスタに属することを可能としたソフトクラスタリング手法であり，各クラスタ表現に確率密度関数を仮定して用いる．そしてその確率密度関数のパラメータを尤度関数の最適化によって推定する．クラスタ表現にはガウス分布を用いることが多く，本論文でも混合ガウス分布を用いた場合のクラスタリングについて紹介する．ガウス分布の形状を決定するパラメータには平均と分散があるが，ガウス分布の平均値はクラスタリングにおけるクラスタ中心を示す．また，混合ガウス分布はガウス分布の重み付き和で表現される．EMアルゴリズムに基づくクラスタリングは，パラメータである平均値，分散値，各密度関数の混合率の推定を行いガウス分布の形状を決定しクラスタ分類を定める手法であり，ファジィc-平均法などと異なり，データのクラスタへの帰属度の最適化は直接は行わない．

$U = (u_{ki})$ をデータ x_k のクラスタ A_i への帰属度行列とする．また，クラスタ A_i を表す確率密度関数を $p_i(x_k|\phi_i)$ とし，混合密度分布を確率密度関数の重み付き和として次式で定義する．

$$p(x_k|\Phi) = \sum_{i=1}^{c} \pi_i p_i(x_k|\phi_i).$$

ここで ϕ_i は密度関数に含まれるパラメータをまとめて表記したものであり，本論文ではクラスタが従う確率密度関数としてガウス分布を用いるため，平均と分散のベクトルを表す．π_i

は混合密度関数に対する密度関数 $p_i(x_k|\phi_i)$ の占める割合であり，混合比と呼ばれる．また，推定すべきパラメータのすべてを $\Phi = (\pi_1, \ldots, \pi_c, \phi_1, \ldots, \phi_c)$ で表す．

EM アルゴリズムに基づくクラスタリングの最適化問題は，次の尤度関数を最大にする密度関数のパラメータを推定することである．

$$Q_{\mathrm{EM}}(\Phi|\Phi') = \sum_{k=1}^{n} \sum_{i=1}^{c} \log[\pi_i p_i(x_k|\phi_i)] \frac{\pi'_i p_i(x_k|\phi'_i)}{p(x_k|\Phi')}. \tag{3.1}$$

制約条件は以下のとおりである．

$$\sum_{i=1}^{c} \pi_i = 1.$$

Φ' はパラメータの推定値である．

(3.1) の $\dfrac{\pi'_i p_i(x_k|\phi'_i)}{p(x_k|\Phi')}$ はデータ x_k が与えられた時の，クラスタ A_i が生起する事後確率であるため，

$$u_{ki} = \frac{\pi'_i p_i(x_k|\phi'_i)}{p(x_k|\Phi')} \tag{3.2}$$

とすると，(3.2) はデータ x_k のクラスタ A_i への帰属度と捉えることができる．もちろん

$$\sum_{i=1}^{c} u_{ki} = 1$$

である．各推定値についてはファジィ c-平均法などと同様にパラメータに関する (3.1) の偏微分やラグランジュの未定乗数法を用いて最適解を得ることができる．ラグランジュ関数は制約条件より以下の式で定義される．

$$L_{\mathrm{EM}} = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki} \log[\pi_i p_i(x_k|\phi_i)] - \tau \left(\sum_{i=1}^{c} \pi_i - 1 \right). \tag{3.3}$$

これより確率密度関数が一次元である場合と多次元である場合とに分け，パラメータ推定値の最適解を導出していく．

一次元の場合

まず確率密度関数が一次元である場合の EM アルゴリズムに基づくクラスタリングの説明をおこなう．v_i をクラスタ A_i を表す確率密度関数 $p_i(x_k|\phi_i)$ の平均，σ_i^2 をクラスタ A_i を表す確率密度関数 $p_i(x_k|\phi_i)$ の分散，π_i を確率密度関数 $p_i(x_k|\phi_i)$ の混合比とする．このとき一次元のガウス分布 $p_i(x_k|\phi_i)$ は次式で定義される．

$$p_i(x_k|\phi_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_k - v_i)^2}{2\sigma_i^2}\right).$$

ただし，$\phi_i = (v_i, \sigma_i)$ である．

ここから，アルゴリズムを構築するために必要な最適解を導出していく．求める最適解は，確率密度関数を構築するパラメータの最適解であり，ここでは混合比 π_i，分散 σ_i^2，平均 v_i がこれに該当する．

まず最初に混合比の最適解を導出する．混合比の最適解はラグランジュの未定乗数法を用いて求めるので，(3.3) を π_i に関して偏微分する．

$$\begin{aligned}\frac{\partial L_{\mathrm{EM}}}{\partial \pi_i} &= \frac{1}{\pi_i}\sum_{k=1}^{n} u_{ki} - \tau = 0,\\ \tau\pi_i &= \sum_{k=1}^{n} u_{ki}.\end{aligned} \tag{3.4}$$

ここで，制約条件 $\sum_{i=1}^{c} \pi_i = 1$ 及び，$\sum_{i=1}^{c} u_{ki} = 1$ より，両辺で i についての総和をとると，

$$\tau \sum_{i=1}^{c} \pi_i = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki},$$

$$\tau = n.$$

この関係を (3.4) 代入すると，混合比の最適解は次のように定まる．

$$\pi_i = \frac{1}{n}\sum_{k=1}^{n} u_{ki}. \tag{3.5}$$

次に分散の最適解を導出する．確率密度関数を一次元のガウス分布とした場合の尤度関数 (3.1) を具体的に記すと次式となる．

$$\begin{aligned} Q_{\mathrm{EM}}(\Phi|\Phi') &= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} \log\left(\frac{\pi_i}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_k - v_i)^2}{2\sigma_i^2}\right)\right) \\ &= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} \left(\log\frac{\pi_i}{\sqrt{2\pi}} - \log\sqrt{\sigma_i^2} - \frac{(x_k - v_i)^2}{2\sigma_i^2}\right). \end{aligned} \tag{3.6}$$

(3.6) を σ_i^2 に関して偏微分すると，

$$\frac{\partial Q_{\mathrm{EM}}}{\partial \sigma_i^2} = -\sum_{k=1}^{n} u_{ki}\left(\frac{1}{\sigma_i^2} - \frac{(x_k - v_i)^2}{\sigma_i^4}\right) = 0.$$

この式を整理すると，分散の最適解は以下のように求まる．

$$\sigma_i^2 = \frac{\displaystyle\sum_{k=1}^{n} u_{ki}(x_k - v_i)^2}{\displaystyle\sum_{k=1}^{n} u_{ki}}. \tag{3.7}$$

同様にクラスタ中心を意味する平均の最適解は，(3.6) を v_i に関して偏微分して，

$$\frac{\partial Q_{\mathrm{EM}}}{\partial v_i} = \sum_{k=1}^{n} u_{ki}\frac{x_k - v_i}{\sigma_i^2} = 0.$$

この式を整理すると，平均の最適解は以下のように求まる．

$$v_i = \frac{\displaystyle\sum_{k=1}^{n} u_{ki} x_k}{\displaystyle\sum_{k=1}^{n} u_{ki}}. \tag{3.8}$$

多次元の場合

次に確率密度関数が多次元の場合の EM アルゴリズムに基づくクラスタリングについて説明する．$v_i = (v_i^1, \ldots, v_i^p)^T$ をクラスタ A_i を表す確率密度関数 $p_i(x_k|\phi_i)$ の平均，$R_i = \left(r_i^{jl}\right)$ $(1 \le j, l \le p)$

クラスタ A_i を表す確率密度関数 $p_i(x_k|\phi_i)$ の分散共分散行列，π_i を確率密度関数 $p_i(x_k|\phi_i)$ の混合比とする．このとき多次元のガウス分布は以下の式で定義される．

$$p_i(x_k|\phi_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|R_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right).$$

ここで，ϕ_i は密度関数に含まれるパラメータであり，$\phi_i = (v_i, R_i)$ である．

目的関数は (3.1) と同一であり，制約条件も等価なので，ラグランジュ関数も (3.3) と表せる．帰属度と事後確率を求める式は同一なので，(3.2) と同じ式で求められる．(3.1) および (3.3) を用いて混合比，分散共分散行列，平均の最適解を求めていく．

まず混合比の最適解を導出する．混合比の最適解は一次元の場合と同様にラグランジュの未定乗数法を用いて求めるので，(3.3) を π_i について偏微分する．

$$\frac{\partial L_{\mathrm{EM}}}{\partial \pi_i} = \frac{1}{\pi_i}\sum_{k=1}^{n} u_{ki} - \tau = 0,$$

$$\tau\pi_i = \sum_{k=1}^{n} u_{ki}.$$

一次元の場合と同様に制約条件 $\sum_{i=1}^{c} \pi_i = 1$ 及び，$\sum_{i=1}^{c} u_{ki} = 1$ より，両辺で i についての総和をとると，

$$\tau \sum_{i=1}^{c} \pi_i = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki},$$

$$\tau = n.$$

この関係を代入すると，混合比の最適解は次のように定まる．

$$\pi_i = \frac{1}{n}\sum_{k=1}^{n} u_{ki}. \tag{3.9}$$

(3.9) は (3.5) と等価である．

次に分散共分散行列の最適解を導出する．確率密度関数を多次元のガウス分布とした場合

の尤度関数 (3.1) を具体的に記すと次式となる．

$$
\begin{aligned}
Q_{\mathrm{EM}}(\Phi|\Phi') &= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} \log\left(\frac{\pi_i}{(2\pi)^{\frac{p}{2}}|R_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right)\right) \\
&= \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} \left(\log\frac{\pi_i}{(2\pi)^{\frac{p}{2}}} - \log|R_i|^{\frac{1}{2}} - \frac{1}{2}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right). \qquad (3.10)
\end{aligned}
$$

(3.10) を R_i について偏微分して最適解を得る．しかし，(3.10) は複雑なので項に分けて偏微分していく．

s^{jl} をある行列 S の (j,l) 成分とし，その余因子を $\mathrm{Cof}\, s^{jl}$ とする．このとき，(3.10) の第二項 $\log|R_i|$ の R_i に関する偏微分は以下のように求められる．

$$
\begin{aligned}
\frac{\partial}{\partial R_i}\log|R_i| &= \left[\frac{\partial}{\partial r_i^{jl}}\log|R_i|\right] \\
&= \left[\frac{1}{|R_i|}\frac{\partial}{\partial r_i^{jl}}|R_i|\right] \\
&= \left[\frac{1}{|R_i|}\mathrm{Cof}\, r_i^{jl}\right] \\
&= R_i^{-1}. \qquad (3.11)
\end{aligned}
$$

第三項の偏微分を求めるためにまず R_i^{-1} の偏微分を求める．E^{jl} を (j,l) 成分のみ 1 でその他の成分が 0 の行列とすると，

$$
\begin{aligned}
\frac{\partial}{\partial r_i^{jl}}\left(R_i R_i^{-1}\right) &= \left(\frac{\partial R_i}{\partial r_i^{jl}}\right)R_i^{-1} + R_i\left(\frac{\partial R_i^{-1}}{\partial r_i^{jl}}\right) \\
&= E^{jl}R_i^{-1} + R_i\left(\frac{\partial R_i^{-1}}{\partial r_i^{jl}}\right) \\
&= 0, \\
\frac{\partial R_i^{-1}}{\partial r_i^{jl}} &= -R_i^{-1}E^{jl}R_i^{-1}. \qquad (3.12)
\end{aligned}
$$

(3.12) より，第三項の偏微分は次のように求められる．

$$\begin{aligned}
\frac{\partial}{\partial R_i}\left((x_k - v_i)^T R_i^{-1}(x_k - v_i)\right) &= \left[\frac{\partial}{\partial r_i^{jl}}\left((x_k - v_i)^T R_i^{-1}(x_k - v_i)\right)\right] \\
&= \left[-(x_k - v_i)^T R_i^{-1} E^{jl} R_i^{-1}(x_k - v_i)\right] \\
&= \left[-\left(R_i^{-1}(x_k - v_i)\right)^T E^{jl}\left(R_i^{-1}(x_k - v_i)\right)\right] \\
&= \left[-\left(R_i^{-1}(x_k - v_i)\right)^j \left(R_i^{-1}(x_k - v_i)\right)^l\right] \\
&= -\left(R_i^{-1}(x_k - v_i)\right)\left(R_i^{-1}(x_k - v_i)\right)^T \\
&= -R_i^{-1}(x_k - v_i)(x_k - v_i)^T R_i^{-1}. \qquad (3.13)
\end{aligned}$$

(3.11)，(3.13) より，

$$\begin{aligned}
\frac{\partial Q_{\mathrm{EM}}}{\partial R_i} &= \sum_{k=1}^{n} u_{ki}\left(R_i^{-1} - R_i^{-1}(x_k - v_i)(x_k - v_i)^T R_i^{-1}\right) \\
&= 0.
\end{aligned}$$

以上より分散共分散行列の最適解は，

$$R_i = \frac{\sum_{k=1}^{n} u_{ki}(x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^{n} u_{ki}}. \qquad (3.14)$$

最後に平均の最適解を求める．b^j をあるベクトル b の j 成分とする．

$$\begin{aligned}
\frac{\partial Q_{\mathrm{EM}}}{\partial v_i} &= \left[\frac{\partial Q_{\mathrm{EM}}}{\partial v_i^j}\right] \\
&= \left[\frac{\partial}{\partial v_i^j}\left(-\frac{1}{2}\sum_{k=1}^{n} u_{ki}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right)\right] \\
&= \left[\frac{1}{2}\sum_{k=1}^{n} u_{ki}\left(\sum_{l=1}^{p}\left(r_i^{-1}\right)^{jl}(x_k - v_i)^l + \sum_{l=1}^{p}\left((x_k - v_i)^T\right)^l\left(r_i^{-1}\right)^{lj}\right)\right] \\
&= \left[\sum_{k=1}^{n} u_{ki}\sum_{l=1}^{p}\left(r_i^{-1}\right)^{jl}(x_k - v_i)^l\right] \\
&= \sum_{k=1}^{n} u_{ki} R_i^{-1}(x_k - v_i) = 0.
\end{aligned}$$

以上より平均の最適解は，

$$v_i = \frac{\sum_{k=1}^{n} u_{ki} x_k}{\sum_{k=1}^{n} u_{ki}}. \tag{3.15}$$

このことからクラスタ中心は一次元の場合と同一の式により求まることがわかる．

最後に，EM アルゴリズムに基づくクラスタリングのアルゴリズムを載せる．

Algorithm 7 EM アルゴリズムに基づくクラスタリング

EM1 初期パラメータを設定する．

EM2 一次元 : (3.5)，多次元 : (3.9) を用いて混合比を更新．

EM3 一次元 : (3.8)，多次元 : (3.15) を用いてクラス中心を更新．

EM4 一次元 : (3.7)，多次元 : (3.14) を用いて分散共分散行列を更新．

EM5 各パラメータが収束すれば終了．そうでなければ **EM2** に戻る．

3.2 提案手法

3.2.1 不確実データに対する EM アルゴリズムに基づくクラスタリング

本節では本研究の提案手法の一つである不確実データに対する EM アルゴリズム (EMU, EM Algorithm for Uncertain Data) に基づくクラスタリングについて述べる．EM アルゴリズムに基づくクラスタリングにおける各データは不確実性を含まないデルタ関数として捉えることができる．そのため，この手法ではデータ自身の不確実性は考慮されない．そこで本手法では，各データをその値を平均とした一定の分散を持つガウス分布として表現し，データ自身の持つ不確実性を考慮する．以下の図が通常の EM アルゴリズムに基づくクラスタリングと提案手法の概念図である．

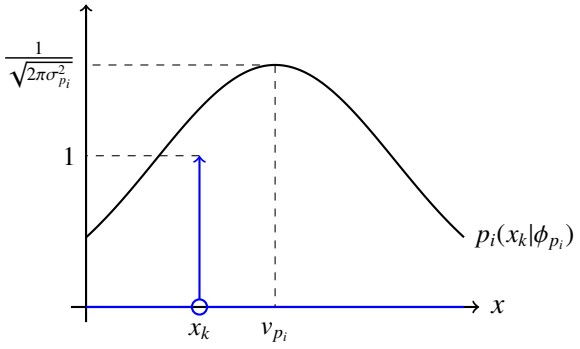

図 3.1: EM アルゴリズムに基づくクラスタリングの概念図

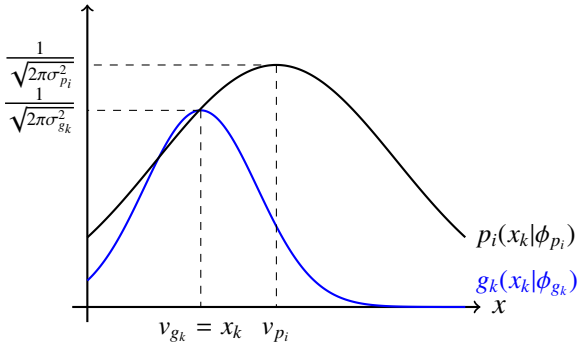

図 3.2: データの不確実性を考慮した EM アルゴリズムに基づくクラスタリングの概念図

データ x_k の不確実性を考慮した確率密度関数を $g_k(x_k|\phi_{g_k})$ とおき，クラスタ A_i を表す確率密度関数は $p_i(x_k|\phi_{p_i})$ とする．ここで，ϕ_{g_k} と ϕ_{p_i} は密度関数に含まれるパラメータである．このとき，データ x_k の不確実性を考慮したクラスタ A_i に対応する密度関数は

$$q_i(x_k|\phi_{p_i}, \phi_{g_k}) = \int_{-\infty}^{\infty} p_i(x|\phi_{p_i}) g_k(x|\phi_{g_k}) dx.$$

混合分布は

$$q(x_k|\Phi) = \sum_{i=1}^{c} \pi_i q_i(x_k|\phi_{p_i}, \phi_{g_k}).$$

また各パラメータをまとめて，$\Phi = (\pi_1, \ldots, \pi_c, \phi_{p_1}, \ldots, \phi_{p_c}, \phi_{g_1}, \ldots, \phi_{g_n})$ で表す．

EMU の最適化問題は，次の尤度関数の最大にする密度関数のパラメータを推定することである．

$$Q_{\mathrm{EMU}}(\Phi|\Phi') = \sum_{k=1}^{n} \sum_{i=1}^{c} \log[\pi_i q_i(x_k|\phi_{p_i}, \phi_{g_k})] \frac{\pi_i' q_i(x_k|\phi_{p_i}', \phi_{g_k})}{q(x_k|\Phi')}. \tag{3.16}$$

制約条件は以下のとおりである．

$$\sum_{i=1}^{c} \pi_i = 1.$$

ただし，$\Phi' = (\pi'_1, \ldots, \pi'_c, \phi'_{p_1}, \ldots, \phi'_{p_c}, \phi_{g_1}, \ldots, \phi_{g_n})$ である．(3.16) の $\dfrac{\pi'_i q_i(x_k|\phi'_i)}{q(x_k|\Phi')}$ はデータ x_k が与えられたときの，クラスタ A_i が生起する事後確率であるため，

$$u_{ki} = \frac{\pi'_i q_i(x_k|\phi'_{p_i}, \phi_{g_k})}{q(x_k|\Phi')} \tag{3.17}$$

とすると，(3.17) はデータ x_k のクラスタ A_i への帰属度とみなせる．また，EM アルゴリズムに基づくクラスタリングと同様に，

$$\sum_{i=1}^{c} u_{ki} = 1.$$

各推定値についてはパラメータに関する (3.16) の偏微分やラグランジュの未定乗数法を用いて最適解を得ることができる．ラグランジュ関数は制約条件より以下の式で定義される．

$$L_{\mathrm{EMU}} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} \log[\pi_i q_i(x_k|\phi_{p_i}, \phi_{g_k})] - \tau\left(\sum_{i=1}^{c} \pi_i - 1\right). \tag{3.18}$$

一次元の場合

まず確率密度関数が一次元である場合の説明をおこなう．v_{p_i} をクラスタ A_i を表す確率密度関数 $p_i(x_k|\phi_{p_i})$ の平均，$\sigma^2_{p_i}$ をクラスタ A_i を表す確率密度関数 $p_i(x_k|\phi_{p_i})$ の分散，v_{g_k} をデータ x_k の不確実性を表す確率密度関数 $g_k(x_k|\phi_{g_k})$ の平均，$\sigma^2_{g_k}$ をデータ x_k の不確実性を表す確率密度関数 $g_k(x_k|\phi_{g_k})$ の分散とし，$\phi_{p_i} = (v_{p_i}, \sigma^2_{p_i})$，$\phi_{g_k} = (v_{g_k}, \sigma^2_{g_k})$ とする．

クラスタとデータの不確実性の確率密度関数はそれぞれ次のように表わされる．

$$p_i(x_k|\phi_{p_i}) = \frac{1}{\sqrt{2\pi\sigma^2_{p_i}}} \exp\left(-\frac{(x_k - v_{p_i})^2}{2\sigma^2_{p_i}}\right).$$

$$g_k(x_k|\phi_{g_k}) = \frac{1}{\sqrt{2\pi\sigma^2_{g_k}}} \exp\left(-\frac{(x_k - v_{g_k})^2}{2\sigma^2_{g_k}}\right).$$

これらから合成密度関数 $q_i(x_k|\phi_{p_i}, \phi_{g_k})$ を求める．

$$
\begin{aligned}
q_i(x_k|\phi_{p_i}, \phi_{g_k}) &= \int_{-\infty}^{\infty} p_i(x|\phi_{p_i}) g_k(x|\phi_{g_k}) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_{p_i}\sigma_{g_k}} \exp\left(-\frac{\sigma_{g_k}^2 (x - v_{p_i})^2 + \sigma_{p_i}^2 (x - v_{g_k})^2}{2\sigma_{p_i}^2 \sigma_{g_k}^2} \right) dx \\
&= \frac{1}{2\pi\sigma_{p_i}\sigma_{g_k}} \exp\left(-\frac{(v_{g_k} - v_{p_i})^2}{2(\sigma_{p_i}^2 + \sigma_{g_k}^2)} \right) \\
&\quad \int_{-\infty}^{\infty} \exp\left(-\frac{(\sigma_{p_i}^2 + \sigma_{g_k}^2)}{2\sigma_{p_i}^2 \sigma_{g_k}^2} \left(x - \frac{\sigma_{g_k}^2 v_{p_i} + \sigma_{p_i}^2 v_{g_k}}{\sigma_{p_i}^2 + \sigma_{g_k}^2} \right)^2 \right) dx \\
&= \frac{1}{2\pi\sigma_{p_i}\sigma_{g_k}} \sqrt{\frac{2\pi\sigma_{p_i}^2 \sigma_{g_k}^2}{(\sigma_{p_i}^2 + \sigma_{g_k}^2)}} \exp\left(-\frac{(v_{g_k} - v_{p_i})^2}{2(\sigma_{p_i}^2 + \sigma_{g_k}^2)} \right) \\
&= \frac{1}{\sqrt{2\pi(\sigma_{p_i}^2 + \sigma_{g_k}^2)}} \exp\left(-\frac{(v_{g_k} - v_{p_i})^2}{2(\sigma_{p_i}^2 + \sigma_{g_k}^2)} \right).
\end{aligned} \tag{3.19}
$$

ここで，一次元のガウス積分が

$$
\int_{-\infty}^{\infty} \exp\left(-a(x - b)^2 \right) dx = \sqrt{\frac{\pi}{a}}
$$

であることを用いている．

これより，アルゴリズムを構築するために必要な最適解を導出していく．求める最適解は確率密度関数を構築するパラメータの最適解であり，ここでは混合比 π_i，分散 $\sigma_{p_i}^2$，平均 v_{p_i} がこれに該当する．

まずはじめに混合比の最適解を導出する．混合比の最適解はラグランジュの未定乗数法を用いて求めるので，(3.18) を π_i について偏微分する．このとき，合成密度関数は (3.19) によって定義されることも考慮すると，

$$
\frac{\partial L_{\mathrm{EMU}}}{\partial \pi_i} = \frac{\sum_{k=1}^{n} u_{ki}}{\pi_i} - \tau = 0.
$$

EM アルゴリズムに基づくクラスタリングのときと同様に，制約条件 $\sum_{i=1}^{c} \pi_i = 1$ 及び，$\sum_{i=1}^{c} u_{ki} =$

1 より，両辺で i についての総和をとると，

$$\tau \sum_{i=1}^{c} \pi_i = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki},$$

$$\tau = n.$$

この関係を代入すると，混合比の最適解は次のように定まる．

$$\pi_i = \frac{1}{n} \sum_{k=1}^{n} u_{ki}. \tag{3.20}$$

次に分散の最適解を導出する．尤度関数 (3.16) に合成密度関数 (3.19) を代入して整理すると，

$$\begin{aligned} Q_{\mathrm{EMU}}(\Phi|\Phi') &= \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki} \log \left(\frac{\pi_i}{\sqrt{2\pi(\sigma_{p_i}^2 + \sigma_{g_k}^2)}} \exp\left(-\frac{(v_{g_k} - v_{p_i})^2}{2(\sigma_{p_i}^2 + \sigma_{g_k}^2)} \right) \right) \\ &= \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki} \left(\log \frac{\pi_i}{\sqrt{2\pi}} - \log \sqrt{\sigma_{p_i}^2 + \sigma_{g_k}^2} - \frac{(v_{g_k} - v_{p_i})^2}{2(\sigma_{p_i}^2 + \sigma_{g_k}^2)} \right). \end{aligned} \tag{3.21}$$

(3.21) を $\sigma_{p_i}^2$ に関して偏微分すると，

$$\frac{\partial Q_{\mathrm{EMU}}}{\partial \sigma_{p_i}^2} = -\sum_{k=1}^{n} u_{ki} \left(\frac{1}{\sigma_{p_i}^2 + \sigma_{g_k}^2} - \frac{(v_{g_k} - v_{p_i})^2}{(\sigma_{p_i}^2 + \sigma_{g_k}^2)^2} \right) = 0. \tag{3.22}$$

(3.22) を満たす $\sigma_{p_i}^2$ が最適解となるが，この式は解析的に解けないので，ニュートン法を用いて近似的に最適解を求める．

$$f(\sigma_{p_i}^2) = \frac{\partial Q_{\mathrm{EMU}}}{\partial \sigma_{p_i}^2} = \sum_{k=1}^{n} u_{ki} \frac{-(\sigma_{p_i}^2 + \sigma_{g_k}^2) + (v_{g_k} - v_{p_i})^2}{(\sigma_{p_i}^2 + \sigma_{g_k}^2)^2} = 0.$$

とおくと，漸化式より分散の近似解が得られる．

$$\begin{aligned} \sigma_{p_i}^{2\,(t+1)} &= \sigma_{p_i}^{2\,(t)} - \frac{f(\sigma_{p_i}^{2\,(t)})}{f'(\sigma_{p_i}^{2\,(t)})} \\ &= \sigma_{p_i}^{2\,(t)} - \frac{\displaystyle\sum_{k=1}^{n} u_{ki} \frac{-(\sigma_{p_i}^{2\,(t)} + \sigma_{g_k}^2) + (v_{g_k} - v_{p_i})^2}{(\sigma_{p_i}^{2\,(t)} + \sigma_{g_k}^2)^2}}{\displaystyle\sum_{k=1}^{n} u_{ki} \frac{(\sigma_{p_i}^{2\,(t)} + \sigma_{g_k}^2) - 2(v_{g_k} - v_{p_i})^2}{(\sigma_{p_i}^{2\,(t)} + \sigma_{g_k}^2)^3}}. \end{aligned} \tag{3.23}$$

クラスタ中心を意味する平均の最適解は，(3.21) を v_{p_i} に関して偏微分して，

$$\frac{\partial Q_{\mathrm{EMU}}}{\partial v_{p_i}} = \sum_{k=1}^{n} u_{ki} \frac{v_{g_k} - v_{p_i}}{(\sigma_{p_i}^2 + \sigma_{g_k}^2)} = 0.$$

この式を整理すると，平均の最適解は以下のように求まる．

$$v_{p_i} = \frac{\sum_{k=1}^{n} u_{ki} v_{g_k}}{\sum_{k=1}^{n} u_{ki}}. \tag{3.24}$$

多次元の場合

次に確率密度関数が多次元の場合の説明をおこなう．$v_{p_i} = (v_{p_i}^1, \ldots, v_{p_i}^p)^T$ をクラスタ A_i を表す確率密度関数 $p_i(x_k|\phi_{p_i})$ の平均，$R_{p_i} = (r_{p_i}^{jl})$ $(1 \leq j, l \leq p)$ をクラスタ A_i を表す確率密度関数 $p_i(x_k|\phi_{p_i})$ の分散共分散行列とし，$\phi_{p_i} = (v_{p_i}, r_{p_i})$ とする．また，$v_{g_k} = (v_{g_k}^1, \ldots, v_{g_k}^p)^T$ をデータ x_k の不確実性を表す確率密度関数 $g_k(x_k|\phi_{g_k})$ の平均，$R_{g_k} = (r_{g_k}^{jl})$ $(1 \leq j, l \leq p)$ をデータ x_k の不確実性を表す確率密度関数 $g_k(x_k|\phi_{g_k})$ の分散共分散行列とし，$\phi_{g_k} = (v_{g_k}, R_{g_k})$ とする．

クラスタとデータの不確実性の確率密度関数はそれぞれ次のように表わされる．

$$p_i(x|\phi_{p_i}) = \frac{1}{(2\pi)^{\frac{p}{2}} |R_{p_i}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - v_{p_i})^T R_{p_i}^{-1}(x - v_{p_i})\right).$$

$$g_k(x|\phi_{g_k}) = \frac{1}{(2\pi)^{\frac{p}{2}} |R_{g_k}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - v_{g_k})^T R_{g_k}^{-1}(x - v_{g_k})\right).$$

これらから合成密度関数 $q_i(x_k|\phi_{p_i}, \phi_{g_k})$ を求める．

$$\begin{aligned} q_i(x_k|\phi_{p_i}, \phi_{g_k}) &= \int_{-\infty}^{\infty} p_i(x|\phi_{p_i}) g_k(x|\phi_{g_k}) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^p |R_{p_i}|^{\frac{1}{2}} |R_{g_k}|^{\frac{1}{2}}} \exp\left(-\frac{(x - v_{p_i})^T R_{p_i}^{-1}(x - v_{p_i}) + (x - v_{g_k})^T R_{g_k}^{-1}(x - v_{g_k})}{2}\right) dx. \end{aligned} \tag{3.25}$$

ここで，x について整理すると，

$$
\begin{aligned}
&(x-v_{p_i})^T R_{p_i}^{-1}(x-v_{p_i}) + (x-v_{g_k})^T R_{g_k}^{-1}(x-v_{g_k}) \\
&= x^T R_{p_i}^{-1} x - 2v_{p_i}^T R_{p_i}^{-1} x + v_{p_i}^T R_{p_i}^{-1} v_{p_i} + x^T R_{g_k}^{-1} x - 2v_{g_k}^T R_{g_k}^{-1} x + v_{g_k}^T R_{g_k}^{-1} v_{g_k} \\
&= x^T (R_{p_i}^{-1} + R_{g_k}^{-1}) x - 2(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}) x + v_{p_i}^T R_{p_i}^{-1} v_{p_i} + v_{g_k}^T R_{g_k}^{-1} v_{g_k}. \qquad (3.26)
\end{aligned}
$$

得られた (3.26) を (3.25) に代入する．

$$
\begin{aligned}
q_i(x_k|\phi_{p_i}, \phi_{g_k}) &= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^p |R_{p_i}|^{\frac{1}{2}} |R_{g_k}|^{\frac{1}{2}}} \exp\left(-\frac{(x-v_{p_i})^T R_{p_i}^{-1}(x-v_{p_i}) + (x-v_{g_k})^T R_{g_k}^{-1}(x-v_{g_k})}{2} \right) dx \\
&= \frac{1}{(2\pi)^p |R_{p_i}|^{\frac{1}{2}} |R_{g_k}|^{\frac{1}{2}}} \exp\left(-\frac{v_{p_i}^T R_{p_i}^{-1} v_{p_i} + v_{g_k}^T R_{g_k}^{-1} v_{g_k}}{2} \right) \\
&\quad \int_{-\infty}^{\infty} \exp\left(-\frac{x^T (R_{p_i}^{-1} + R_{g_k}^{-1}) x - 2(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}) x}{2} \right) dx \\
&= \frac{1}{(2\pi)^{\frac{p}{2}} |R_{p_i}|^{\frac{1}{2}} |R_{g_k}|^{\frac{1}{2}} |R_{p_i}^{-1} + R_{g_k}^{-1}|^{\frac{1}{2}}} \exp\left(-\frac{v_{p_i}^T R_{p_i}^{-1} v_{p_i} + v_{g_k}^T R_{g_k}^{-1} v_{g_k}}{2} \right. \\
&\quad \left. + \frac{(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1})(R_{p_i}^{-1} + R_{g_k}^{-1})^{-1} (v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1})^T}{2} \right). \qquad (3.27)
\end{aligned}
$$

ここで，n 次元のガウス積分が

$$
\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} x^T A x + b^T x \right) dx = \sqrt{\frac{(2\pi)^n}{|A|}} \exp\left(\frac{1}{2} b^T A^{-1} b \right),
$$

であることを用いている．

これより，アルゴリズムを構築するために必要な最適解を導出していく．求める最適解はクラスタの確率密度関数を構築するパラメータの最適解であり，ここでは混合比 π_i，分散共分散行列 R_{p_i}，平均 v_{p_i} がこれに該当する．

まずはじめに混合比の最適解を導出する．混合比の最適解はラグランジュの未定乗数法を用いて求めるので，(3.18) を π_i について偏微分する．このとき，合成密度関数は (3.27) によって定義されることも考慮すると，

$$
\frac{\partial L_{\mathrm{EMU}}}{\partial \pi_i} = \frac{\sum_{k=1}^{n} u_{ki}}{\pi_i} - \tau = 0.
$$

制約条件 $\sum_{i=1}^{c} \pi_i = 1$ 及び，$\sum_{i=1}^{c} u_{ki} = 1$ より，両辺で i についての総和をとると，

$$\tau \sum_{i=1}^{c} \pi_i = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki},$$

$$\tau = n.$$

この関係を代入すると，混合比の最適解は次のように定まる．

$$\pi_i = \frac{1}{n} \sum_{k=1}^{n} u_{ki}. \tag{3.28}$$

次に分散共分散行列の最適解を求める．尤度関数 (3.16) に合成密度関数 (3.27) を代入して整理すると，

$$\begin{aligned} Q_{\mathrm{EMU}}(\Phi|\Phi') &= \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki} \log \left(\frac{\pi_i}{(2\pi)^{\frac{p}{2}} |R_{p_i}|^{\frac{1}{2}} |R_{g_k}|^{\frac{1}{2}} |R_{p_i}^{-1} + R_{g_k}^{-1}|^{\frac{1}{2}}} \exp \left(-\frac{v_{p_i}^T R_{p_i}^{-1} v_{p_i} + v_{g_k}^T R_{g_k}^{-1} v_{g_k}}{2} \right. \right. \\ &\qquad \left. \left. + \frac{(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1})(R_{p_i}^{-1} + R_{g_k}^{-1})^{-1} (v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1})^T}{2} \right) \right) \\ &= \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki} \left(\log \frac{\pi_i}{(2\pi)^{\frac{p}{2}} |R_{g_k}|^{\frac{1}{2}}} - \log \left(|R_{p_i}|^{\frac{1}{2}} |R_{p_i}^{-1} + R_{g_k}^{-1}|^{\frac{1}{2}} \right) - \frac{v_{p_i}^T R_{p_i}^{-1} v_{p_i} + v_{g_k}^T R_{g_k}^{-1} v_{g_k}}{2} \right. \\ &\qquad \left. + \frac{(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1})(R_{p_i}^{-1} + R_{g_k}^{-1})^{-1} (v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1})^T}{2} \right). \end{aligned} \tag{3.29}$$

(3.29) を R_{p_i} に関して偏微分することで分散共分散行列の最適解は求められる．

$$\begin{aligned} \frac{\partial Q_{\mathrm{EMU}}}{\partial R_{p_i}} = \frac{\partial}{\partial R_{p_i}} \Bigg(&- \frac{1}{2} \sum_{k=1}^{n} u_{ki} \Big(\log |R_{p_i}| + \log |R_{p_i}^{-1} + R_{g_k}^{-1}| + v_{p_i}^T R_{p_i}^{-1} v_{p_i} \\ &- \left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1} \right) \left(R_{p_i}^{-1} + R_{g_k}^{-1} \right)^{-1} \left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1} \right)^T \Big) \Bigg). \end{aligned} \tag{3.30}$$

しかし (3.30) の第ニ項目及び第四項目は R_{p_i} に関して偏微分が行えないため，(3.30) は解析的に解くことができない．そこで提案手法は交互最適化に基づくアルゴリズムであることを利用して，第二項目及び第四項目の一部の R_{p_i} を更新前の R_{p_i} とみなし，変数ではなく定数とし

て扱う．このことを考慮して (3.30) を新たに書き直すと，

$$\frac{\partial Q_{\mathrm{EMU}}}{\partial R_{p_i}} = \frac{\partial}{\partial R_{p_i}}\Bigg(-\frac{1}{2}\sum_{k=1}^{n} u_{ki}\Big(\log|R_{p_i}| + \log|\hat{R}_{p_i}^{-1} + R_{g_k}^{-1}| + v_{p_i}^T R_{p_i}^{-1} v_{p_i} \\ - \left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}\right)\left(\hat{R}_{p_i}^{-1} + R_{g_k}^{-1}\right)^{-1}\left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}\right)^T \Big)\Bigg). \tag{3.31}$$

ただし，$\hat{R}_{p_i}$ は定数として扱われる t 回目の更新における分散共分散行列であり，ここで求めたいのは $t+1$ 回目の更新における分散共分散行列 R_{p_i} である．

このように置くことで，分散共分散行列の最適解は近似的に求めることができる．最適解を求めるために，(3.31) を項毎に偏微分していく．(3.11) より第一項目は，

$$\frac{\partial}{\partial R_{p_i}} \log|R_{p_i}| = R_{p_i}^{-1}. \tag{3.32}$$

(3.13) より第三項目は，

$$\frac{\partial}{\partial R_{p_i}}\left(v_{p_i}^T R_{p_i}^{-1} v_{p_i}\right) = -R_{p_i}^{-1} v_{p_i} v_{p_i}^T R_{p_i}^{-1}. \tag{3.33}$$

最後に第四項目であるが，まず $B = \left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^T\right)\left(\hat{R}_{p_i}^{-1} + R_{g_k}^{-1}\right)^{-1}\left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k} R_{g_k}^{-1}\right)^T$ と置くと，

$$\begin{aligned}
\frac{\partial B}{\partial R_{p_i}} &= \left[\frac{\partial B}{\partial r_{p_i}^{jl}}\right] \\
&= \Big[-\left(v_{p_i}^T R_{p_i}^{-1} E^{jl} R_{p_i}^{-1}\right)\left(\hat{R}_{p_i}^{-1} + R_{g_k}^{-1}\right)^{-1}\left(R_{p_i}^{-1} v_{p_i} + R_{g_k}^{-1} v_{g_k}\right) \\
&\quad - \left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}\right)\left(\hat{R}_{p_i}^{-1} + R_{g_k}^{-1}\right)^{-1}\left(R_{p_i}^{-1} E^{jl} R_{p_i}^{-1} v_{p_i}\right)\Big] \\
&= -R_{p_i}^{-1} v_{p_i}\left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}\right)\left(\hat{R}_{p_i}^{-1} + R_{g_k}^{-1}\right)^{-1} R_{p_i}^{-1} \\
&\quad - R_{p_i}^{-1}\left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}\right)\left(\hat{R}_{p_i}^{-1} + R_{g_k}^{-1}\right)^{-1} v_{p_i}^T R_{p_i}^{-1} \\
&= -R_{p_i}^{-1} v_{p_i} \alpha R_{p_i}^{-1} - R_{p_i}^{-1} \alpha^T v_{p_i}^T R_{p_i}^{-1}.
\end{aligned} \tag{3.34}$$

ただし，

$$\alpha = \left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}\right)\left(\hat{R}_{p_i}^{-1} + R_{g_K}^{-1}\right)^{-1}.$$

(3.32)，(3.33)，(3.34) より，

$$\frac{\partial Q_{\mathrm{EMU}}}{\partial R_{p_i}} = -\frac{1}{2}\sum_{k=1}^{n} u_{ki}\left(R_{p_i}^{-1} - R_{p_i}^{-1}\left(v_{p_i} - \alpha^T\right)\left(v_{p_i}^T - \alpha\right)R_{p_i}^{-1} + R_{p_i}^{-1}\alpha^T\alpha R_{p_i}^{-1}\right) = 0.$$

以上より，

$$R_{p_i} = \frac{\displaystyle\sum_{k=1}^{n} u_{ki}\left(\left(v_{p_i} - \alpha^T\right)\left(v_{p_i}^T - \alpha\right) - \alpha^T\alpha\right)}{\displaystyle\sum_{k=1}^{n} u_{ki}}. \tag{3.35}$$

また (3.29) を v_{p_i} に関して偏微分することで，クラスタ中心の最適化解は求められる．

$$\begin{aligned}\frac{\partial Q_{\mathrm{EMU}}}{\partial v_{p_i}} &= \left[\frac{\partial Q_{\mathrm{EMU}}}{\partial v_{p_i}^j}\right] \\ &= \left[\frac{\partial}{\partial v_{p_i}^j}\left(-\frac{1}{2}\sum_{k=1}^{n} u_{ki}\left(v_{p_i}^T R_{p_i}^{-1} v_{p_i} - \left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}\right)\left(R_{p_i}^{-1} + R_{g_k}^{-1}\right)^{-1}\left(v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}\right)^T\right)\right)\right].\end{aligned}$$

分散と同様に項毎に偏微分していく．

$$\begin{aligned}\frac{\partial}{\partial v_{p_i}}\left(v_{p_i}^T R_{p_i}^{-1} v_{p_i}\right) &= \left[\frac{\partial}{\partial v_{p_i}^j}\left(v_{p_i}^T R_{p_i}^{-1} v_{p_i}\right)\right] = \left[\sum_{l=1}^{p}(r_{p_i}^{-1})^{jl} v_{p_i}^l + \sum_{l=1}^{p} v_{p_i}^l (r_{p_i}^{-1})^{lj}\right] \\ &= 2\left[\sum_{l=1}^{p}(r_{p_i}^{-1})^{jl} v_{p_i}^l\right] = 2R_{p_i}^{-1} v_{p_i}.\end{aligned} \tag{3.36}$$

ここで $W = (R_{p_i}^{-1} + R_{g_k}^{-1})^{-1}$ と置き，

$$\begin{aligned}&\frac{\partial}{\partial v_{p_i}}\left((v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}) W (v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1})^T\right) \\ &= \left[\frac{\partial}{\partial v_{p_i}^j}\left((v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1}) W (v_{p_i}^T R_{p_i}^{-1} + v_{g_k}^T R_{g_k}^{-1})^T\right)\right] \\ &= \left[\sum_{l=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p}(r_{p_i}^{-1})^{jl} w^{ls}\left((r_{p_i}^{-1})^{ts} v_{p_i}^t + (r_{g_k}^{-1})^{ts} v_{g_k}^t\right) + \sum_{l=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p}(v_{p_i}^t (r_{p_i}^{-1})^{ts} + v_{g_k}^t (r_{g_k}^{-1})^{ts}) w^{sl} (r_{p_i}^{-1})^{jl}\right] \\ &= 2\left[\sum_{l=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p}(r_{p_i}^{-1})^{jl} w^{ls}\left((r_{p_i}^{-1})^{ts} v_{p_i}^t + (r_{g_k}^{-1})^{ts} v_{g_k}^t\right)\right] \\ &= 2\left[\left((r_{p_i}^{-1})^{j1}, \ldots, (r_{p_i}^{-1})^{jl}, \ldots, (r_{p_i}^{-1})^{jp}\right) W\left(R_{p_i}^{-1} v_{p_i} + R_{g_k}^{-1} v_{g_k}\right)\right] \\ &= 2R_{p_i}^{-1} W\left(R_{p_i}^{-1} v_{p_i} + R_{g_k}^{-1} v_{g_k}\right) = 2R_{p_i}^{-1}\left(R_{p_i}^{-1} + R_{g_k}^{-1}\right)^{-1}\left(R_{p_i}^{-1} v_{p_i} + R_{g_k}^{-1} v_{g_k}\right).\end{aligned} \tag{3.37}$$

(3.36)，(3.37) より最適解は，

$$\frac{\partial Q_{EMU}}{\partial v_{p_i}} = -\sum_{k=1}^{n} u_{ki}\left(2R_{p_i}^{-1}v_{p_i} - R_{p_i}^{-1}\left(R_{p_i}^{-1}+R_{g_k}^{-1}\right)^{-1}\left(R_{p_i}^{-1}v_{p_i}+R_{g_k}^{-1}v_{g_k}\right)\right) = 0.$$

$$v_{p_i} = \frac{\sum_{k=1}^{n} u_{ki}v_{g_k}}{\sum_{k=1}^{n} u_{ki}}. \tag{3.38}$$

最後に，EMU アルゴリズムに基づくクラスタリングのアルゴリズムを載せる．

Algorithm 8 EMU アルゴリズムに基づくクラスタリング

EMU1 初期パラメータを設定する．

EMU2 一次元 : (3.20)，多次元 : (3.28) を用いて混合比を更新．

EMU3 一次元 : (3.23)，多次元 : (3.35) を用いて分散共分散行列を更新．

EMU4 一次元 : (3.24)，多次元 : (3.38) を用いてクラス中心を更新．

EMU5 各パラメータが収束すれば終了．そうでなければ **EMU2** に戻る．

3.3 数値例

本節では，提案手法と関連手法である EM アルゴリズムに基づくクラスタリングとの比較を数値実験を通しておこない，提案手法の有効性の検討をおこなう．実験データには一次元の実データである 2015 年の世界 30 カ国の男性の平均身長データ [42]，2014 年の世界 30 カ国の名目 GDP データ [43] のニ種類を使用した．実データの詳細は表 3.1，3.2 に記載されている．

3.3.1 身長データに対するクラスタリング結果

表 3.1 のデータを，不確実データに対する EM アルゴリズムに基づくクラスタリングと EM アルゴリズムに基づくクラスタリングを用いて，二つもしくは三つのクラスタに分類した場合の比較をおこなう．数値実験ではデータの分散 $\delta_{g_k}^2$ は適宜変化させているが，データの平均 v_{g_k} にはデータ値をそのまま用いている．

表 3.3 – 3.10 は身長データに対して，EM アルゴリズムに基づくクラスタリングと EMU アルゴリズムに基づくクラスタリングを用いた際の，各データのクラスタへの帰属度とその分類結果である．表 3.3 と表 3.5 – 3.7 は身長データを二つのクラスタに分類した結果であり，表 3.4 と表 3.8 – 3.10 は身長データを三つのクラスタに分類した結果である．また表 3.5 – 3.6，表 3.8 – 3.9 ではデータの分散値として一律の分散値を用いているが，表 3.7，表 3.10 では一部のデータの分散値が異なっている．

二つのクラスタに分類した場合，既存手法と提案手法どちらも分類結果は変わらず，わずかに帰属度が異なるのみであることが表から見て取れる．しかし三つのクラスタに分類した場合，一律の分散値を用いた結果である表 3.8 – 3.9 は EM アルゴリズムを用いた分類結果と変わらないが，表 3.10 の分類結果はフランスのデータのみ分類が異なることがわかる．これは元々フランスのデータが二つのクラスタに同程度属していたものを，フランスのデータの持つ不確実性の分散値を他のデータよりも小さくすることで，クラスタの帰属関係が変化したためである．また提案手法の結果同士は帰属度の差がほとんど見られない結果が多い．このことから提案手法は，不確実性に対してロバスト性を持つことが予想される．

表 3.1: 世界 30 カ国の男性の平均身長

国名	平均身長 (cm)
オランダ	183.8
デンマーク	182.6
スウェーデン	181.5
ドイツ	181
クロアチア	180.5
チェコ	180.3
オーストリア	179.2
フィンランド	179
アメリカ	178.9
ベルギー	178.6
スペイン	178
オーストラリア	177.8
イギリス	177.6
フランス	175.6
ブラジル	175
韓国	173.7
アルゼンチン	173.48
イラン	173.4
メキシコ	172
チリ	171
日本	170.7
タイ	170.3
マレーシア	170.2
ガーナ	169.5
インド	166.3
ベトナム	165.7
イラク	165.4
ペルー	164
フィリピン	163.4
インドネシア	158

表 3.2: 世界 30 カ国の名目 GDP

国名	名目 GDP (10 億 US ドル)
アメリカ	17348.08
中国	10356.51
日本	4602.37
ドイツ	3874.44
イギリス	2950.04
フランス	2833.69
ブラジル	2346.58
イタリア	2147.74
インド	2051.23
ロシア	1860.6
カナダ	1785.39
オーストラリア	1442.72
韓国	1410.38
スペイン	1406.54
メキシコ	1291.06
インドネシア	888.65
オランダ	880.72
トルコ	798.33
サウジアラビア	746.25
スイス	703.85
ナイジェリア	574
スウェーデン	570.59
ポーランド	547.89
アルゼンチン	543.06
ベルギー	534.23
台湾	529.6
ノルウェー	499.82
オーストリア	437.58
イラン	416.49
タイ	404.82

表 3.3: 身長データに対して EM アルゴリズムに基づくクラスタリングを用いた結果 1

国名	$\delta^2_{g_k}$	各クラスタへの帰属度	
		クラスタ 1	クラスタ 2
オランダ	—	0.895148	0.104852
デンマーク	—	0.92388	0.07612
スウェーデン	—	0.930098	0.069902
ドイツ	—	0.928047	0.071953
クロアチア	—	0.922754	0.077246
チェコ	—	0.9196	0.0804
オーストリア	—	0.88772	0.11228
フィンランド	—	0.878342	0.121658
アメリカ	—	0.873103	0.126897
ベルギー	—	0.854865	0.145135
スペイン	—	0.804282	0.195718
オーストラリア	—	0.782227	0.217773
イギリス	—	0.757095	0.242905
フランス	—	0.335599	0.664401
ブラジル	—	0.202444	0.797556
韓国	—	0.043641	0.956359
アルゼンチン	—	0.032046	0.967954
イラン	—	0.028556	0.971444
メキシコ	—	0.00303	0.99697
チリ	—	0.000483	0.999517
日本	—	0.000268	0.999732
タイ	—	0.00012	0.99988
マレーシア	—	0.000097	0.999903
ガーナ	—	0.000022	0.999978
インド	—	0	1.000000
ベトナム	—	0	1.000000
イラク	—	0	1.000000
ペルー	—	0	1.000000
フィリピン	—	0	1.000000
インドネシア	—	0	1.000000

表 3.4: 身長データに対して EM アルゴリズムに基づくクラスタリングを用いた結果 2

国名	$\delta^2_{g_k}$	各クラスタへの帰属度		
		クラスタ 1	クラスタ 2	クラスタ 3
オランダ	—	0.999999	0	0.000001
デンマーク	—	0.999997	0.000001	0.000003
スウェーデン	—	0.999986	0.000008	0.000006
ドイツ	—	0.999968	0.000022	0.00001
クロアチア	—	0.999921	0.000062	0.000017
チェコ	—	0.999885	0.000093	0.000022
オーストリア	—	0.999044	0.000879	0.000077
フィンランド	—	0.998589	0.001312	0.000099
アメリカ	—	0.998286	0.001601	0.000113
ベルギー	—	0.996933	0.002901	0.000166
スペイン	—	0.990272	0.009353	0.000375
オーストラリア	—	0.985777	0.013728	0.000496
イギリス	—	0.979278	0.020064	0.000658
フランス	—	0.530668	0.460498	0.008833
ブラジル	—	0.276912	0.709614	0.013474
韓国	—	0.037484	0.940346	0.02217
アルゼンチン	—	0.025968	0.950035	0.023997
イラン	—	0.022714	0.952558	0.024728
メキシコ	—	0.00219	0.949202	0.048607
チリ	—	0.000419	0.904683	0.094898
日本	—	0.000255	0.880782	0.118963
タイ	—	0.00013	0.837141	0.162729
マレーシア	—	0.00011	0.823673	0.176218
ガーナ	—	0.000032	0.693576	0.306392
インド	—	0	0.024057	0.975943
ベトナム	—	0	0.008404	0.991596
イラク	—	0	0.004818	0.995182
ペルー	—	0	0.000282	0.999718
フィリピン	—	0	0.000074	0.999926
インドネシア	—	0	0	1.000000

表 3.5: 身長データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 1

国名	$\delta^2_{g_k}$	各クラスタへの帰属度	
		クラスタ 1	クラスタ 2
オランダ	1.0	0.898636	0.101364
デンマーク	1.0	0.925988	0.074012
スウェーデン	1.0	0.931758	0.068242
ドイツ	1.0	0.929662	0.070338
クロアチア	1.0	0.924412	0.075588
チェコ	1.0	0.921302	0.078698
オーストリア	1.0	0.890002	0.109998
フィンランド	1.0	0.88081	0.11919
アメリカ	1.0	0.875675	0.124325
ベルギー	1.0	0.857806	0.142194
スペイン	1.0	0.808239	0.191761
オーストラリア	1.0	0.786615	0.213385
イギリス	1.0	0.761959	0.238041
フランス	1.0	0.34397	0.65603
ブラジル	1.0	0.20929	0.79071
韓国	1.0	0.045947	0.954053
アルゼンチン	1.0	0.033837	0.966163
イラン	1.0	0.030184	0.969816
メキシコ	1.0	0.003263	0.996737
チリ	1.0	0.000527	0.999473
日本	1.0	0.000294	0.999706
タイ	1.0	0.000132	0.999868
マレーシア	1.0	0.000108	0.999892
ガーナ	1.0	0.000024	0.999976
インド	1.0	0	1.000000
ベトナム	1.0	0	1.000000
イラク	1.0	0	1.000000
ペルー	1.0	0	1.000000
フィリピン	1.0	0	1.000000
インドネシア	1.0	0	1.000000

表 3.6: 身長データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 2

国名	$\delta^2_{g_k}$	各クラスタへの帰属度	
		クラスタ 1	クラスタ 2
オランダ	0.01	0.896974	0.103026
デンマーク	0.01	0.925094	0.074906
スウェーデン	0.01	0.931121	0.068879
ドイツ	0.01	0.929061	0.070939
クロアチア	0.01	0.923805	0.076195
チェコ	0.01	0.920679	0.079321
オーストリア	0.01	0.88912	0.11088
フィンランド	0.01	0.879838	0.120162
アメリカ	0.01	0.874652	0.125348
ベルギー	0.01	0.856596	0.143404
スペイン	0.01	0.806485	0.193515
オーストラリア	0.01	0.784618	0.215382
イギリス	0.01	0.759688	0.240312
フランス	0.01	0.339083	0.660917
ブラジル	0.01	0.205092	0.794908
韓国	0.01	0.044415	0.955585
アルゼンチン	0.01	0.032634	0.967366
イラン	0.01	0.029087	0.970913
メキシコ	0.01	0.003098	0.996902
チリ	0.01	0.000495	0.999505
日本	0.01	0.000275	0.999725
タイ	0.01	0.000123	0.999877
マレーシア	0.01	0.0001	0.9999
ガーナ	0.01	0.000022	0.999978
インド	0.01	0	1.000000
ベトナム	0.01	0	1.000000
イラク	0.01	0	1.000000
ペルー	0.01	0	1.000000
フィリピン	0.01	0	1.000000
インドネシア	0.01	0	1.000000

表 3.7: 身長データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 3

国名	$\delta^2_{g_k}$	各クラスタへの帰属度	
		クラスタ 1	クラスタ 2
オランダ	1.0	0.895095	0.104905
デンマーク	1.0	0.925699	0.074301
スウェーデン	1.0	0.932663	0.067337
ドイツ	1.0	0.930878	0.069122
クロアチア	1.0	0.925844	0.074156
チェコ	0.01	0.934062	0.065938
オーストリア	1.0	0.891357	0.108643
フィンランド	1.0	0.881999	0.118001
アメリカ	1.0	0.876755	0.123245
ベルギー	1.0	0.858425	0.141575
スペイン	1.0	0.807098	0.192902
オーストラリア	1.0	0.784544	0.215456
イギリス	1.0	0.758744	0.241256
フランス	1.0	0.323257	0.676743
ブラジル	1.0	0.189558	0.810442
韓国	1.0	0.037858	0.962142
アルゼンチン	1.0	0.027416	0.972584
イラン	1.0	0.024307	0.975693
メキシコ	1.0	0.002347	0.997653
チリ	1.0	0.000347	0.999653
日本	1.0	0.000188	0.999812
タイ	0.01	0.000005	0.999995
マレーシア	1.0	0.000065	0.999935
ガーナ	1.0	0.000014	0.999986
インド	1.0	0	1.000000
ベトナム	1.0	0	1.000000
イラク	1.0	0	1.000000
ペルー	1.0	0	1.000000
フィリピン	1.0	0	1.000000
インドネシア	1.0	0	1.000000

表 3.8: 身長データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 4

国名	$\delta^2_{g_k}$	各クラスタへの帰属度		
		クラスタ 1	クラスタ 2	クラスタ 3
オランダ	1.0	0.999999	0	0.000001
デンマーク	1.0	0.999997	0.000001	0.000003
スウェーデン	1.0	0.999986	0.000008	0.000006
ドイツ	1.0	0.999968	0.000022	0.00001
クロアチア	1.0	0.999921	0.000062	0.000017
チェコ	1.0	0.999885	0.000093	0.000022
オーストリア	1.0	0.999044	0.000879	0.000077
フィンランド	1.0	0.998589	0.001312	0.000099
アメリカ	1.0	0.998286	0.001601	0.000113
ベルギー	1.0	0.996933	0.002901	0.000166
スペイン	1.0	0.990272	0.009353	0.000375
オーストラリア	1.0	0.985777	0.013728	0.000496
イギリス	1.0	0.979278	0.020064	0.000658
フランス	1.0	0.530668	0.460498	0.008833
ブラジル	1.0	0.276912	0.709614	0.013474
韓国	1.0	0.037484	0.940346	0.02217
アルゼンチン	1.0	0.025968	0.950035	0.023997
イラン	1.0	0.022714	0.952558	0.024728
メキシコ	1.0	0.00219	0.949202	0.048607
チリ	1.0	0.000419	0.904683	0.094898
日本	1.0	0.000255	0.880782	0.118963
タイ	1.0	0.00013	0.837141	0.162729
マレーシア	1.0	0.00011	0.823673	0.176218
ガーナ	1.0	0.000032	0.693576	0.306392
インド	1.0	0	0.024057	0.975943
ベトナム	1.0	0	0.008404	0.991596
イラク	1.0	0	0.004818	0.995182
ペルー	1.0	0	0.000282	0.999718
フィリピン	1.0	0	0.000074	0.999926
インドネシア	1.0	0	0	1.000000

表 3.9: 身長データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 5

国名	$\delta^2_{g_k}$	各クラスタへの帰属度		
		クラスタ 1	クラスタ 2	クラスタ 3
オランダ	0.01	0.999999	0	0.000001
デンマーク	0.01	0.999997	0.000001	0.000003
スウェーデン	0.01	0.999986	0.000008	0.000006
ドイツ	0.01	0.999968	0.000022	0.00001
クロアチア	0.01	0.999921	0.000062	0.000017
チェコ	0.01	0.999885	0.000093	0.000022
オーストリア	0.01	0.999044	0.000879	0.000077
フィンランド	0.01	0.998589	0.001312	0.000099
アメリカ	0.01	0.998286	0.001601	0.000113
ベルギー	0.01	0.996933	0.002901	0.000166
スペイン	0.01	0.990272	0.009353	0.000375
オーストラリア	0.01	0.985777	0.013728	0.000496
イギリス	0.01	0.979278	0.020064	0.000658
フランス	0.01	0.530668	0.460498	0.008833
ブラジル	0.01	0.276912	0.709614	0.013474
韓国	0.01	0.037484	0.940346	0.02217
アルゼンチン	0.01	0.025968	0.950035	0.023997
イラン	0.01	0.022714	0.952558	0.024728
メキシコ	0.01	0.00219	0.949202	0.048607
チリ	0.01	0.000419	0.904683	0.094898
日本	0.01	0.000255	0.880782	0.118963
タイ	0.01	0.00013	0.837141	0.162729
マレーシア	0.01	0.00011	0.823673	0.176218
ガーナ	0.01	0.000032	0.693576	0.306392
インド	0.01	0	0.024057	0.975943
ベトナム	0.01	0	0.008404	0.991596
イラク	0.01	0	0.004818	0.995182
ペルー	0.01	0	0.000282	0.999718
フィリピン	0.01	0	0.000074	0.999926
インドネシア	0.01	0	0	1.000000

表 3.10: 身長データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 6

国名	$\delta^2_{g_k}$	各クラスタへの帰属度		
		クラスタ 1	クラスタ 2	クラスタ 3
オランダ	1.0	0.999964	0.000031	0.000005
デンマーク	1.0	0.999865	0.000126	0.000009
スウェーデン	1.0	0.999497	0.000483	0.00002
ドイツ	1.0	0.999059	0.00091	0.00003
クロアチア	1.0	0.998217	0.001735	0.000048
チェコ	1.0	0.997688	0.002254	0.000059
オーストリア	1.0	0.990006	0.009803	0.000191
フィンランド	1.0	0.986889	0.01287	0.000241
アメリカ	1.0	0.984976	0.014752	0.000271
ベルギー	1.0	0.977365	0.022244	0.000391
スペイン	1.0	0.948618	0.050549	0.000833
オーストラリア	1.0	0.932657	0.066264	0.001079
イギリス	1.0	0.912027	0.086574	0.001399
フランス	0.01	0.261092	0.727326	0.011583
ブラジル	1.0	0.165208	0.817084	0.017709
韓国	1.0	0.023254	0.945461	0.031285
アルゼンチン	1.0	0.016209	0.949422	0.034369
イラン	1.0	0.014197	0.950216	0.035587
メキシコ	1.0	0.001274	0.928198	0.070528
チリ	1.0	0.000205	0.874295	0.1255
日本	1.0	0.000116	0.849188	0.150695
タイ	1.0	0.000053	0.80701	0.192936
マレーシア	1.0	0.000044	0.794706	0.20525
ガーナ	0.01	0	0.676544	0.323456
インド	1.0	0	0.075263	0.924737
ベトナム	1.0	0	0.037332	0.962668
イラク	1.0	0	0.0257	0.9743
ペルー	1.0	0	0.003821	0.996179
フィリピン	1.0	0	0.00157	0.99843
インドネシア	1.0	0	0	1.000000

3.3.2 GDP データに対するクラスタリング結果

表 3.2 のデータを，不確実データに対する EM アルゴリズムに基づくクラスタリングと EM アルゴリズムに基づくクラスタリングを用いて，二つもしくは三つのクラスタに分類した場合の比較をおこなう．身長データのときと同様にデータの分散 $\delta_{g_k}^2$ は適宜変化させているが，データの平均 v_{g_k} にはデータ値をそのまま用いている．

表 3.11 – 3.18 は GDP データに対して，EM アルゴリズムに基づくクラスタリングと EMU アルゴリズムに基づくクラスタリングを用いた際の，各データのクラスタへの帰属度とその分類結果である．表 3.11 と表 3.13 – 3.15 は身長データを二つのクラスタに分類した結果であり，表 3.12 と表 3.16 – 3.18 は身長データを三つのクラスタに分類した結果である．また表 3.13 – 3.14，表 3.16 – 3.17 ではデータの分散値として一律の分散値を用いているが，表 3.15，表 3.18 では一部のデータの分散値が異なっている．

二つのクラスタに分類した場合，先ほどの身長データでは分類結果には差が見られなかった．GDP データでは提案手法を用いた場合，ドイツのデータが既存手法とは異なるクラスタに分類された．これは先ほどの身長データと同様に，二つのクラスタに同程度属していたドイツのデータに不確実性を与えたためである．身長データと異なり均一の分散値を与えた場合でもデータ分類が変化した理由としては，身長データに対して GDP データがより疎であることに原因があると考えられる．三つのクラスタに分類した場合，既存手法との分類結果に差は生じていない．しかし帰属度を比べてみると，提案手法ではほとんどのデータの所属クラスタへの帰属度の占める割合が上昇し，より明確に分類される結果となった．この傾向も身長データでは見られなかったため，データの疎密性が影響していると考えられる．また身長データのときと同様に，提案手法間の帰属度の差はほとんど見られないこともわかる．これらのことから，提案手法は不確実性に対してロバスト性を持つこと，疎なデータにより有効に働く可能性が高いことが考えられる．

表 3.11: GDP データに対して EM アルゴリズムに基づくクラスタリングを用いた結果 1

国名	$\delta^2_{g_k}$	各クラスタへの帰属度	
		クラスタ 1	クラスタ 2
アメリカ	—	1.000000	0
中国	—	1.000000	0
日本	—	0.926565	0.073435
ドイツ	—	0.427856	0.572144
イギリス	—	0.049802	0.950198
フランス	—	0.038813	0.961187
ブラジル	—	0.01595	0.98405
イタリア	—	0.011989	0.988011
インド	—	0.010614	0.989386
ロシア	—	0.008621	0.991379
カナダ	—	0.008038	0.991962
オーストラリア	—	0.006367	0.993633
韓国	—	0.006275	0.993725
スペイン	—	0.006264	0.993736
メキシコ	—	0.00601	0.99399
インドネシア	—	0.005899	0.994101
オランダ	—	0.005909	0.994091
トルコ	—	0.006035	0.993965
サウジアラビア	—	0.006143	0.993857
スイス	—	0.006247	0.993753
ナイジェリア	—	0.006665	0.993335
スウェーデン	—	0.006679	0.993321
ポーランド	—	0.00677	0.99323
アルゼンチン	—	0.00679	0.99321
ベルギー	—	0.006827	0.993173
台湾	—	0.006847	0.993153
ノルウェー	—	0.006981	0.993019
オーストリア	—	0.007294	0.992706
イラン	—	0.007411	0.992589
タイ	—	0.007478	0.992522

表 3.12: GDP データに対して EM アルゴリズムに基づくクラスタリングを用いた結果 2

国名	$\delta^2_{g_k}$	各クラスタへの帰属度		
		クラスタ 1	クラスタ 2	クラスタ 3
アメリカ	—	1.000000	0	0
中国	—	1.000000	0	0
日本	—	0.024448	0.975551	0.000001
ドイツ	—	0.007364	0.99245	0.000186
イギリス	—	0.004059	0.932584	0.063357
フランス	—	0.00388	0.879932	0.116188
ブラジル	—	0.002041	0.360501	0.637458
イタリア	—	0.001208	0.176013	0.82278
インド	—	0.000918	0.119536	0.879547
ロシア	—	0.000539	0.054293	0.945168
カナダ	—	0.000443	0.039756	0.959801
オーストラリア	—	0.00021	0.010123	0.989667
韓国	—	0.000198	0.008951	0.990851
スペイン	—	0.000197	0.008822	0.990982
メキシコ	—	0.000164	0.005744	0.994092
インドネシア	—	0.000114	0.001461	0.998425
オランダ	—	0.000114	0.001425	0.998461
トルコ	—	0.000112	0.001105	0.998784
サウジアラビア	—	0.000111	0.000945	0.998944
スイス	—	0.000111	0.000834	0.999055
ナイジェリア	—	0.000115	0.000577	0.999308
スウェーデン	—	0.000115	0.000572	0.999313
ポーランド	—	0.000117	0.000537	0.999346
アルゼンチン	—	0.000117	0.00053	0.999353
ベルギー	—	0.000117	0.000518	0.999365
台湾	—	0.000118	0.000511	0.999371
ノルウェー	—	0.00012	0.000472	0.999408
オーストリア	—	0.000125	0.000401	0.999474
イラン	—	0.000127	0.00038	0.999493
タイ	—	0.000129	0.000369	0.999503

表 3.13: GDP データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 1

国名	$\delta^2_{g_k}$	各クラスタへの帰属度	
		クラスタ 1	クラスタ 2
アメリカ	1.0	1.000000	0
中国	1.0	1.000000	0
日本	1.0	0.998505	0.001495
ドイツ	1.0	0.924167	0.075833
イギリス	1.0	0.218687	0.781313
フランス	1.0	0.161865	0.838135
ブラジル	1.0	0.049936	0.950064
イタリア	1.0	0.033569	0.966431
インド	1.0	0.028289	0.971711
ロシア	1.0	0.021074	0.978926
カナダ	1.0	0.019071	0.980929
オーストラリア	1.0	0.013628	0.986372
韓国	1.0	0.013338	0.986662
スペイン	1.0	0.013306	0.986694
メキシコ	1.0	0.012509	0.987491
インドネシア	1.0	0.012035	0.987965
オランダ	1.0	0.012059	0.987941
トルコ	1.0	0.01239	0.98761
サウジアラビア	1.0	0.012678	0.987322
スイス	1.0	0.012962	0.987038
ナイジェリア	1.0	0.014135	0.985865
スウェーデン	1.0	0.014173	0.985827
ポーランド	1.0	0.014433	0.985567
アルゼンチン	1.0	0.014491	0.985509
ベルギー	1.0	0.014598	0.985402
台湾	1.0	0.014656	0.985344
ノルウェー	1.0	0.015043	0.984957
オーストリア	1.0	0.015963	0.984037
イラン	1.0	0.016311	0.983689
タイ	1.0	0.016513	0.983487

表 3.14: GDP データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 2

国名	$\delta^2_{g_k}$	各クラスタへの帰属度	
		クラスタ 1	クラスタ 2
アメリカ	0.01	1.000000	0
中国	0.01	1.000000	0
日本	0.01	0.998505	0.001495
ドイツ	0.01	0.924167	0.075833
イギリス	0.01	0.218687	0.781313
フランス	0.01	0.161865	0.838135
ブラジル	0.01	0.049936	0.950064
イタリア	0.01	0.033569	0.966431
インド	0.01	0.028289	0.971711
ロシア	0.01	0.021074	0.978926
カナダ	0.01	0.019071	0.980929
オーストラリア	0.01	0.013628	0.986372
韓国	0.01	0.013338	0.986662
スペイン	0.01	0.013306	0.986694
メキシコ	0.01	0.012509	0.987491
インドネシア	0.01	0.012035	0.987965
オランダ	0.01	0.012059	0.987941
トルコ	0.01	0.01239	0.98761
サウジアラビア	0.01	0.012678	0.987322
スイス	0.01	0.012962	0.987038
ナイジェリア	0.01	0.014135	0.985865
スウェーデン	0.01	0.014173	0.985827
ポーランド	0.01	0.014433	0.985567
アルゼンチン	0.01	0.014491	0.985509
ベルギー	0.01	0.014598	0.985402
台湾	0.01	0.014656	0.985344
ノルウェー	0.01	0.015043	0.984957
オーストリア	0.01	0.015963	0.984037
イラン	0.01	0.016311	0.983689
タイ	0.01	0.016513	0.983487

表 3.15: GDP データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 3

国名	$\delta^2_{g_k}$	各クラスタへの帰属度	
		クラスタ 1	クラスタ 2
アメリカ	1.0	1.000000	0
中国	1.0	1.000000	0
日本	1.0	0.998505	0.001495
ドイツ	1.0	0.924167	0.075833
イギリス	0.01	0.218688	0.781312
フランス	1.0	0.161865	0.838135
ブラジル	1.0	0.049936	0.950064
イタリア	1.0	0.033569	0.966431
インド	1.0	0.028289	0.971711
ロシア	1.0	0.021074	0.978926
カナダ	1.0	0.019071	0.980929
オーストラリア	1.0	0.013628	0.986372
韓国	1.0	0.013338	0.986662
スペイン	1.0	0.013306	0.986694
メキシコ	1.0	0.012509	0.987491
インドネシア	1.0	0.012035	0.987965
オランダ	1.0	0.012059	0.987941
トルコ	1.0	0.01239	0.98761
サウジアラビア	1.0	0.012678	0.987322
スイス	1.0	0.012962	0.987038
ナイジェリア	1.0	0.014135	0.985865
スウェーデン	1.0	0.014173	0.985827
ポーランド	1.0	0.014433	0.985567
アルゼンチン	1.0	0.014491	0.985509
ベルギー	1.0	0.014598	0.985402
台湾	1.0	0.014656	0.985344
ノルウェー	1.0	0.015043	0.984957
オーストリア	1.0	0.015963	0.984037
イラン	1.0	0.016311	0.983689
タイ	1.0	0.016513	0.983487

表 3.16: GDP データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 4

国名	$\delta^2_{g_k}$	各クラスタへの帰属度		
		クラスタ 1	クラスタ 2	クラスタ 3
アメリカ	1.0	1.000000	0	0
中国	1.0	1.000000	0	0
日本	1.0	0.012923	0.987075	0.000002
ドイツ	1.0	0.003695	0.99593	0.000374
イギリス	1.0	0.001981	0.900329	0.097691
フランス	1.0	0.001854	0.826344	0.171802
ブラジル	1.0	0.000806	0.268823	0.730371
イタリア	1.0	0.000451	0.121479	0.878069
インド	1.0	0.000338	0.080283	0.919379
ロシア	1.0	0.000195	0.034929	0.964876
カナダ	1.0	0.00016	0.025187	0.974653
オーストラリア	1.0	0.000074	0.005957	0.993969
韓国	1.0	0.00007	0.005226	0.994704
スペイン	1.0	0.00007	0.005146	0.994785
メキシコ	1.0	0.000058	0.003252	0.99669
インドネシア	1.0	0.000039	0.000729	0.999232
オランダ	1.0	0.000038	0.000709	0.999252
トルコ	1.0	0.000037	0.000533	0.99943
サウジアラビア	1.0	0.000037	0.000447	0.999516
スイス	1.0	0.000037	0.000388	0.999576
ナイジェリア	1.0	0.000037	0.000254	0.999709
スウェーデン	1.0	0.000037	0.000251	0.999711
ポーランド	1.0	0.000037	0.000234	0.999729
アルゼンチン	1.0	0.000037	0.00023	0.999732
ベルギー	1.0	0.000037	0.000224	0.999738
台湾	1.0	0.000038	0.000221	0.999742
ノルウェー	1.0	0.000038	0.000201	0.999761
オーストリア	1.0	0.000039	0.000166	0.999795
イラン	1.0	0.00004	0.000156	0.999804
タイ	1.0	0.00004	0.000151	0.999809

表 3.17: GDP データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 5

国名	$\delta^2_{g_k}$	各クラスタへの帰属度		
		クラスタ 1	クラスタ 2	クラスタ 3
アメリカ	0.01	1.000000	0	0
中国	0.01	1.000000	0	0
日本	0.01	0.012913	0.987085	0.000002
ドイツ	0.01	0.003691	0.995937	0.000373
イギリス	0.01	0.001977	0.900636	0.097387
フランス	0.01	0.00185	0.826822	0.171328
ブラジル	0.01	0.000805	0.269432	0.729762
イタリア	0.01	0.000451	0.121812	0.877737
インド	0.01	0.000338	0.080516	0.919146
ロシア	0.01	0.000195	0.035039	0.964766
カナダ	0.01	0.00016	0.025269	0.974571
オーストラリア	0.01	0.000074	0.005979	0.993947
韓国	0.01	0.00007	0.005245	0.994685
スペイン	0.01	0.000069	0.005165	0.994766
メキシコ	0.01	0.000058	0.003265	0.996678
インドネシア	0.01	0.000038	0.000733	0.999229
オランダ	0.01	0.000038	0.000713	0.999249
トルコ	0.01	0.000037	0.000536	0.999427
サウジアラビア	0.01	0.000037	0.000449	0.999514
スイス	0.01	0.000036	0.00039	0.999574
ナイジェリア	0.01	0.000037	0.000256	0.999707
スウェーデン	0.01	0.000037	0.000253	0.99971
ポーランド	0.01	0.000037	0.000235	0.999727
アルゼンチン	0.01	0.000037	0.000232	0.999731
ベルギー	0.01	0.000037	0.000225	0.999737
台湾	0.01	0.000037	0.000222	0.99974
ノルウェー	0.01	0.000038	0.000202	0.99976
オーストリア	0.01	0.000039	0.000167	0.999794
イラン	0.01	0.00004	0.000157	0.999804
タイ	0.01	0.00004	0.000151	0.999809

表 3.18: GDP データに対して EMU アルゴリズムに基づくクラスタリングを用いた結果 6

国名	$\delta^2_{g_k}$	各クラスタへの帰属度		
		クラスタ 1	クラスタ 2	クラスタ 3
アメリカ	1.0	1.000000	0	0
中国	1.0	1.000000	0	0
日本	1.0	0.012911	0.987087	0.000002
ドイツ	1.0	0.00369	0.995938	0.000372
イギリス	1.0	0.001976	0.900703	0.097321
フランス	0.01	0.00185	0.826926	0.171224
ブラジル	0.01	0.000805	0.269565	0.72963
イタリア	1.0	0.000451	0.121884	0.877665
インド	1.0	0.000337	0.080567	0.919096
ロシア	1.0	0.000195	0.035063	0.964742
カナダ	1.0	0.00016	0.025287	0.974554
オーストラリア	1.0	0.000074	0.005984	0.993942
韓国	1.0	0.00007	0.00525	0.99468
スペイン	1.0	0.000069	0.005169	0.994761
メキシコ	1.0	0.000058	0.003268	0.996675
インドネシア	1.0	0.000038	0.000733	0.999228
オランダ	1.0	0.000038	0.000713	0.999248
トルコ	1.0	0.000037	0.000537	0.999426
サウジアラビア	1.0	0.000037	0.00045	0.999514
スイス	1.0	0.000036	0.00039	0.999573
ナイジェリア	1.0	0.000037	0.000256	0.999707
スウェーデン	1.0	0.000037	0.000253	0.99971
ポーランド	1.0	0.000037	0.000236	0.999727
アルゼンチン	1.0	0.000037	0.000232	0.999731
ベルギー	1.0	0.000037	0.000226	0.999737
台湾	1.0	0.000037	0.000222	0.99974
ノルウェー	1.0	0.000038	0.000203	0.999759
オーストリア	1.0	0.000039	0.000167	0.999793
イラン	1.0	0.00004	0.000157	0.999803
タイ	1.0	0.00004	0.000152	0.999808

第4章　データ自身に含まれる不確実性ー不確実性ベクトルを用いたEMアルゴリズムに基づくクラスタリング

これまで，不確実性ベクトルや許容ベクトルは様々なクラスタリング手法 [30–34] に導入されてきた．しかし，混合確率分布の観点からクラスタ分割を論じる EM アルゴリズムに基づくクラスタリングに関しては，不確実性ベクトル，許容範囲ともに議論がなされてこなかった．そこで本研究では不確実性ベクトルを導入した EM アルゴリズムに基づくクラスタリングを提案し，EM アルゴリズムと不確実性ベクトルとの親和性を議論する．

前章で述べたように，不確実性ベクトルは許容ベクトルと類似した性質を持つ．しかし理論的扱いやすさと拡張性の高さで考えるならば，不確実性ベクトルの方が許容ベクトルよりも簡易であり，幅広い手法に応用することができる．本章の提案手法が許容ベクトルではなく，不確実性ベクトルを用いるのも同様の理由である．宮岸ら [37] は Kullback-Leibler 情報量 (KL 情報量) 正則化ファジィc-平均法 (KFCM, fuzzy c-means clustering with regularization by K-L information) を提案している．この研究では EM アルゴリズムとの関連性についても議論されており，パラメータを一定の値に設定することで，KL 情報量正則化ファジィc-平均法が EM アルゴリズムに基づくクラスタリングと等価となることが示されている．またこのことから，KL 情報量正則化ファジィc-平均法は EM アルゴリズムに基づくクラスタリングを内包する手法であることも明らかにされている．本研究でも宮岸らの研究方針を踏襲し，データの不確実性を考慮した EM アルゴリズムに基づくクラスタリングを直接構築するのではなく，

まずデータの不確実性を考慮した KL 情報量正則化ファジィc-平均法を構築する．その後，この手法と等価となる EM アルゴリズムに基づくクラスタリングを構築する．ここで構築された手法は，上記の関連性からデータの不確実性を考慮した EM アルゴリズムであるといえる．そのため，本研究では目的関数の制約条件が増える許容ベクトルではなく，正則化項を付け加えるだけで実現できる不確実性ベクトルを利用する．

以上より本章ではまず関連手法として，宮岸らの提案した KL 情報量正則化ファジィc-平均法，及び遠藤らの提案した不確実性ベクトルを導入したファジィc-平均法を紹介し，その後提案手法について述べる．

4.1 関連手法

4.1.1 KL 情報量正則化ファジィc-平均法

KL 情報量正則化ファジィc-平均法 (KFCM，fuzzy c-means clustering with regularization by K-L information) [37] は宮岸らによって提案された手法であり，ファジィc-平均法に Kullback-Leibler 情報量による正則化項を導入した手法である．KL 情報量正則化ファジィc-平均法では，データとクラスタ間の非類似度をマハラノビス距離として定義している．また，KL 情報量正則化ファジィc-平均法特定の条件のもとで EM アルゴリズムに基づくクラスタリングと等価であることが明らかとなっている．

KL 情報量正則化ファジィc-平均法の目的関数は次式で定義される．

$$J_{\mathrm{KFCM}}(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ki}(x_k-v_i)^T R_i^{-1}(x_k-v_i)+\lambda\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ki}\log\frac{u_{ki}}{\pi_i}+\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ki}\log|R_i|. \quad (4.1)$$

ただし，$R_i=(r_i^{jl})\ (1\le j,l\le p)$ はクラスタ A_i の分散共分散行列である．KL 情報量正則化ファ

ジィc-平均法の制約条件は以下のとおりである．

$$\sum_{i=1}^{c} u_{ki} = 1,$$
$$\sum_{i=1}^{c} \pi_i = 1.$$

ここで，π_i は全クラスタ内におけるクラスタ A_i の比重を表す．(4.1) の第一項はデータとクラスタ間の非類似度に関してクラスタの分散を考慮したマハラノビス距離であり，第ニ項は π_i と u_{ki} との値が近いほど小さくなる u_{ki} と π_i との分布の近さを測る KL 情報量である．すなわち，u_{ki} と π_i ができるだけ等しくなるように制御する正則化項である．KL 情報量正則化ファジィc-平均法の最適化問題は，この制約のもとでの (4.1) の最小化である．

次にアルゴリズムを構築するために必要な最適解を導出していく．各変数の最適解はその変数に関する偏微分やラグランジュの未定乗数法を用いて導出する．KL 情報量正則化ファジィc-平均法のラグランジュ関数は次式で定義される．

$$\begin{aligned} L_{\mathrm{KFCM}} = &\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}(x_k - v_i)^T R_i^{-1}(x_k - v_i) + \lambda \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} \log \frac{u_{ki}}{\pi_i} + \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} \log |R_i| \\ &- \tau\left(\sum_{i=0}^{c} \pi_i - 1\right) - \sum_{k=1}^{n} \gamma_k \left(\sum_{i=1}^{c} u_{ki} - 1\right). \end{aligned} \tag{4.2}$$

はじめに，クラスタ中心の最適解を導出する．(4.1) を v_i に関して偏微分すると，

$$\begin{aligned} \frac{\partial J_{\mathrm{KFCM}}}{\partial v_i} &= \left[\frac{\partial J_{\mathrm{KFCM}}}{\partial v_i^j}\right] \\ &= \left[\frac{\partial}{\partial v_i^j}\left(-\frac{1}{2}\sum_{k=1}^{n} u_{ki}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right)\right] \\ &= \left[\frac{1}{2}\sum_{k=1}^{n} u_{ki}\left(\sum_{l=1}^{p}\left(r_i^{-1}\right)^{jl}(x_k - v_i)^l + \sum_{l=1}^{p}\left((x_k - v_i)^T\right)^l \left(r_i^{-1}\right)^{lj}\right)\right] \\ &= \left[\sum_{k=1}^{n} u_{ki}\sum_{l=1}^{p}\left(r_i^{-1}\right)^{jl}(x_k - v_i)^l\right] \\ &= \sum_{k=1}^{n} u_{ki} R_i^{-1}(x_k - v_i) = 0. \end{aligned}$$

上式の変形から，クラスタ中心の最適解は得られる．

$$v_i = \frac{\sum_{k=1}^{n} u_{ki} x_k}{\sum_{k=1}^{n} u_{ki}}. \tag{4.3}$$

次に帰属度の最適解を導出する．帰属度には制約条件があるため，ラグランジュの未定乗数法を用いて最適解を導出する．(4.2) を u_{ki} に関して偏微分すると，

$$\frac{\partial L_{\mathrm{KFCM}}}{\partial u_{ki}} = (x_k - v_i)^T R_i^{-1}(x_k - v_i) + \lambda\left(\log\frac{u_{ki}}{\pi_i} + 1\right) + \log|R_i| - \gamma_k = 0.$$

この式を u_{ki} に関して整理していく．

$$\begin{aligned}
\log u_{ki} &= -\frac{1}{\lambda}(x_k - v_i)^T R_i^{-1}(x_k - v_i) + \log\pi_i - 1 - \frac{1}{\lambda}\log|R_i| + \frac{\gamma_k}{\lambda}, \\
u_{ki} &= \pi_i |R_i|^{-\frac{1}{\lambda}} \exp\left(-\frac{1}{\lambda}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right)\exp\left(\frac{\gamma_k}{\lambda} - 1\right).
\end{aligned} \tag{4.4}$$

制約条件 $\sum_{i=1}^{c} u_{ki} = 1$ より，

$$\begin{aligned}
\sum_{i=1}^{c} u_{ki} &= \exp\left(\frac{\gamma_k}{\lambda} - 1\right)\sum_{i=1}^{c} \pi_i |R_i|^{-\frac{1}{\lambda}} \exp\left(-\frac{1}{\lambda}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right) = 1, \\
\exp\left(\frac{\gamma_k}{\lambda} - 1\right) &= \frac{1}{\sum_{i=1}^{c} \pi_i |R_i|^{-\frac{1}{\lambda}} \exp\left(-\frac{1}{\lambda}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right)}.
\end{aligned}$$

この式を，(4.4) に代入すると帰属度の最適解は次式で得られる．

$$u_{ki} = \frac{\pi_i |R_i|^{-\frac{1}{\lambda}} \exp\left(-\frac{1}{\lambda}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right)}{\sum_{j=1}^{c} \pi_j |R_j|^{-\frac{1}{\lambda}} \exp\left(-\frac{1}{\lambda}(x_k - v_j)^T R_j^{-1}(x_k - v_j)\right)}. \tag{4.5}$$

クラスタの重みの最適解も帰属度と同様に，(4.2) を π_i に関して偏微分して，

$$\begin{aligned}
\frac{\partial L_{\mathrm{KFCM}}}{\partial \pi_i} &= -\lambda\sum_{k=1}^{n} u_{ki}\frac{1}{\pi_i} - \tau = 0, \\
-\frac{\lambda}{\tau}\sum_{k=1}^{n} u_{ki} &= \pi_i.
\end{aligned} \tag{4.6}$$

ここで，制約条件 $\sum_{i=1}^{c} \pi_i = 1$ 及び，$\sum_{i=1}^{c} u_{ki} = 1$ より，両辺で i についての総和をとると，

$$-\frac{\lambda}{\tau} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki} = \sum_{i=1}^{c} \pi_i,$$
$$-\frac{\lambda}{\tau} = \frac{1}{n}.$$

この関係を (4.6) に代入すると，クラスタの重みの最適解は次のように定まる．

$$\pi_i = \frac{1}{n} \sum_{k=1}^{n} u_{ki}. \tag{4.7}$$

分散共分散行列の最適解は，(4.1) を R_i に関して偏微分して求める．ここで (3.11)，(3.13) より，

$$\begin{aligned} \frac{\partial J_{\mathrm{KFCM}}}{\partial R_i} &= \left[\frac{\partial}{\partial r_i^{jl}} \sum_{k=1}^{n} u_{ki} \left((x_k - v_i)^T R_i^{-1} (x_k - v_i) + \log |R_i| \right) \right] \\ &= \sum_{k=1}^{n} u_{ki} \left(-R_i^{-1} (x_k - v_i) (x_k - v_i)^T R_i^{-1} + R_i^{-1} \right) = 0. \end{aligned}$$

上式左辺の各項に左右から R_i を掛け，R_i に関して整理すると分散共分散行列の最適解を得る．

$$R_i = \frac{\sum_{k=1}^{n} u_{ki} (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^{n} u_{ki}}. \tag{4.8}$$

KL 情報量正則化ファジィ c-平均法の最適解を EM アルゴリズムに基づくクラスタリングの最適化解 (3.2)，(3.9)，(3.14)，(3.15) と比較すると，$\lambda = 2.0$ とした際に解が一致することがわかる．そのため，KL 情報量正則化ファジィ c-平均法は EM アルゴリズムを内包した手法で

あるといえる．

$$u_{ki} = \frac{\pi'_i p_i(x_k|\phi'_i)}{p(x_k|\Phi')} = \frac{\pi_i |R_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x_k - v_i)^T R_i^{-1}(x_k - v_i)\right)}{\sum_{j=1}^{c} \pi_j |R_j|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x_k - v_j)^T R_j^{-1}(x_k - v_j)\right)}, \tag{3.2}$$

$$\pi_i = \frac{1}{n}\sum_{k=1}^{n} u_{ki}, \tag{3.9}$$

$$R_i = \frac{\sum_{k=1}^{n} u_{ki}(x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^{n} u_{ki}}, \tag{3.14}$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ki} x_k}{\sum_{k=1}^{n} u_{ki}}. \tag{3.15}$$

最後に，KL 情報量正則化ファジィc-平均法のアルゴリズムを載せる．

Algorithm 9 KFCM

KFCM1 初期クラスタ分割を与え，初期帰属度を決定する．

KFCM2 (4.7) を用いてクラスタの重みを更新．

KFCM3 (4.3) を用いてクラスタ中心を更新．

KFCM4 (4.8) を用いて分散共分散行列を更新．

KFCM5 (4.5) を用いて帰属度を更新．

KFCM6 各パラメータが収束すれば終了．そうでなければ **KFCM2** に戻る．

4.1.2 ペナルティベクトル正則化に基づく標準型ファジィc-平均法

ペナルティベクトル正則化に基づく標準型ファジィc-平均法 (sFCMQ，standard fuzzy c-means clsutering for uncertain data using quadratic penalty-vector regularization) [33] は遠藤らによって提案された手法で，許容範囲を用いたファジィc-平均法 [30] と同様に，データの不確実性を考慮した手法であるが，後者では，データの不確実性を許容ベクトルで表現し，その限界値を制約として最適化問題に取り入れている．そのため許容ベクトルの最適化には KKT 条件を用いなくてはならない．一方，前者ではデータの不確実性を不確実性ベクトルの正則化項として取り入れているため，不確実性ベクトルの大きさは自動的に調整され，限界を必要としない．そのため，不確実性ベクトルを用いた方が許容ベクトルより簡素な最適化問題に帰着できる．

新たなノーテーションとして，データ x_k に含まれる不確実性ベクトルを $\delta_k = (\delta_{k1}, \ldots, \delta_{kp})^T$ で表し，その集合を $\Delta = \{\delta_k \mid k = 1, \ldots, n\}$ とする．

ペナルティベクトル正則化に基づく標準型ファジィc-平均法の目的関数は次式で定義される．

$$J_{\mathrm{sFCMQ}}(U, V, \Delta) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ki}^{m} \|x_k + \delta_k - v_i\|^2 + \sum_{k=1}^{n} \delta_k^T W_k \delta_k. \tag{4.9}$$

ペナルティベクトル正則化に基づく標準型ファジィc-平均法の制約条件は以下のとおりである．

$$\sum_{i=1}^{c} u_{ki} = 1.$$

ペナルティベクトル正則化に基づく標準型ファジィc-平均法の最適化問題は，この制約のもとでの (4.9) の最小化である．ただし第ニ項は，

$$\sum_{k=1}^{n} \delta_k^T W_k \delta_k = \sum_{k=1}^{n} \sum_{j=1}^{p} \sum_{l=1}^{p} w_{k_{lj}} \delta_{kl} \delta_{kj}$$

であり，W_k $(k = 1, \ldots, n)$ は

$$W_k = \begin{pmatrix} w_{k_{11}} & \cdots & w_{k_{1p}} \\ \vdots & \ddots & \vdots \\ w_{k_{p1}} & \cdots & w_{k_{pp}} \end{pmatrix}$$

を満たす正定値対称行列で，不確実性 δ_k についてのペナルティとなる正則化項である．

これらを考慮した上で，アルゴリズムを構築するために必要な最適解を導出していく．各変数の最適解はその変数に関する偏微分やラグランジュの未定乗数法を用いて導出する．ペナルティベクトル正則化に基づく標準型ファジィc-平均法のラグランジュ関数は次式で定義される．

$$L_{\mathrm{sFCMQ}} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}\|x_k + \delta_k - v_i\|^2 + \sum_{k=1}^{n} \delta_k W_k \delta_k - \sum_{k=1}^{n} \lambda_k \sum_{i=1}^{c}(u_{ki} - 1). \tag{4.10}$$

はじめに，クラスタ中心の最適解を導出する．(4.9) を v_i に関して偏微分すると，

$$\frac{\partial J_{\mathrm{sFCMQ}}}{\partial v_i} = -2\sum_{k=1}^{n} u_{ki}^m (x_k + \delta_k - v_i) = 0.$$

上式の変形から，クラスタ中心の最適解は得られる．

$$v_i = \frac{\displaystyle\sum_{k=1}^{n} u_{ki}^m (x_k + \delta_k)}{\displaystyle\sum_{k=1}^{n} u_{ki}^m}. \tag{4.11}$$

次に帰属度の最適解を導出する．帰属度には制約条件があるため，ラグランジュの未定乗数法を用いて最適解を導出する．(4.10) を u_{ki} に関して偏微分すると，

$$\frac{\partial L_{\mathrm{sFCMQ}}}{\partial u_{ki}} = m u_{ki}^{m-1}\|x_k + \delta_k - v_i\|^2 - \lambda_k = 0,$$

この式を u_{ki} に関して整理していく．

$$u_{ki}^{m-1} = \frac{\lambda_k}{m\|x_k + \delta_k - v_i\|^2},$$
$$u_{ki} = \left(\frac{\lambda_k}{m\|x_k + \delta_k - v_i\|^2}\right)^{\frac{1}{m-1}}. \tag{4.12}$$

制約条件 $\sum_{i=1}^{c} u_{ki} = 1$ より，

$$\sum_{j=1}^{c} u_{kj} = \sum_{j=1}^{c}\left(\frac{\lambda_k}{m\|x_k + \delta_k - v_j\|^2}\right)^{\frac{1}{m-1}} = 1,$$
$$\lambda_k^{\frac{1}{m-1}} = \left(\sum_{j=1}^{c}\left(\frac{1}{m\|x_k + \delta_k - v_j\|^2}\right)^{\frac{1}{m-1}}\right)^{-1}.$$

この式を，(4.12) に代入すると帰属度の最適解は次式で得られる．

$$u_{ki} = \frac{\left(\frac{1}{\|x_k + \delta_k - v_i\|^2}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\frac{1}{\|x_k + \delta_k - v_j\|^2}\right)^{\frac{1}{m-1}}}. \tag{4.13}$$

不確実性の最適解は，クラスタ中心と同様に，(4.9) を δ_k に関して偏微分する．

$$\frac{\partial J_{\text{sFCMQ}}}{\partial \delta_k} = \sum_{i=1}^{c} u_{ki}^m (x_k + \delta_k - v_i) + W_k \delta_k = 0.$$

δ_k に関して整理すると，

$$\left(\sum_{i=1}^{c} u_{ki}^m I + W_k\right)\delta_k = -\sum_{i=1}^{c} u_{ki}^m (x_k - v_i).$$

上式の両辺に左から $\left(\sum_{i=1}^{c} u_{ki}^m I + W_k\right)^{-1}$ を掛け，分散共分散行列の最適解が求まる．

$$\delta_k = -\left(\sum_{i=1}^{c} u_{ki}^m I + W_k\right)^{-1}\left(\sum_{i=1}^{c} u_{ki}^m (x_k - v_i)\right). \tag{4.14}$$

最後に，ペナルティベクトル正則化に基づく標準型ファジィc-平均法のアルゴリズムを載せる．

Algorithm 10 sFCMQ

sFCMQ1 初期クラスタ分割を与え初期帰属度を決定し，初期不確実性を与える．

sFCMQ2 (4.11) を用いてクラスタ中心を更新．

sFCMQ3 (4.13) を用いて帰属度を更新．

sFCMQ4 (4.14) を用いてデータの不確実性を更新．

sFCMQ5 収束判定条件を満たせば終了．そうでなければ **sFCMQ2** に戻る．

4.2 提案手法

4.2.1 KL 情報量を用いたペナルティベクトル正則化ファジィ c-平均法

ここでは，データの不確実性を考慮した EM アルゴリズムに基づくクラスタリングを導くために必要な，提案手法の一つである KL 情報量を用いたペナルティベクトル正則化ファジィ c-平均法 (KLFCMQ, fuzzy c-means with quadratic penalty-vector regularization using Kullback-Leibler information) について述べる．KLFCMQ は KL 情報量正則化ファジィ c-平均法と同様に，データとクラスタ間の非類似度としてマハラノビス距離を用い，KL 情報量によって正則化している．また，不確実性ベクトルを制御するために，ペナルティベクトル正則化に基づく標準型ファジィ c-平均法と同様に，データの不確実性に関するペナルティ正則化項も合わせて導入している．

KLFCMQ の目的関数は次式で定義される．

$$J_{\mathrm{KLFCMQ}}(U,V,\Delta) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}(x_k+\delta_k-v_i)^T R_i^{-1}(x_k+\delta_k-v_i) + \lambda\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}\log\frac{u_{ki}}{\pi_i} + \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}\log|R_i| + \sum_{k=1}^{n}\delta_k^T W_k\delta_k. \tag{4.15}$$

KLFCMQ の制約条件を以下のとおりである．

$$\sum_{i=1}^{c} u_{ki} = 1,$$
$$\sum_{i=1}^{c} \pi_i = 1.$$

KLFCMQ の最適化問題は，この制約のもとでの (4.15) の最小化である．

次にアルゴリズムを構築するために必要な最適解を導出していく．各変数の最適解はその変数に関する偏微分やラグランジュの未定乗数法を用いて導出する．KLFCMQ のラグランジュ関数は次式で定義される．

$$\begin{aligned} L_{\mathrm{KLFCMQ}} = &\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}(x_k + \delta_k - v_i)^T R_i^{-1}(x_k + \delta_k - v_i) + \lambda \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} \log \frac{u_{ki}}{\pi_i} \\ &+ \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} \log |R_i| + \sum_{k=1}^{n} \delta_k^T W_k \delta_k - \sum_{k-1}^{n} \gamma_k \left(\sum_{i=1}^{c} u_{ki} - 1 \right) - \tau \left(\sum_{i=1}^{c} \pi_i - 1 \right). \end{aligned} \tag{4.16}$$

はじめに，クラスタ中心の最適解を導出する．(4.15) を v_i に関して偏微分すると，

$$\begin{aligned} \frac{\partial J_{\mathrm{KLFCMQ}}}{\partial v_i} &= \left[\frac{\partial J_{\mathrm{KLFCMQ}}}{\partial v_i^j} \right] \\ &= \left[\frac{\partial}{\partial v_i^j} \left(-\frac{1}{2} \sum_{k=1}^{n} u_{ki}(x_k + \delta_k - v_i)^T R_i^{-1}(x_k + \delta_k - v_i) \right) \right] \\ &= \left[\frac{1}{2} \sum_{k=1}^{n} u_{ki} \left(\sum_{l=1}^{p} \left(r_i^{-1} \right)^{jl} (x_k + \delta_k - v_i)^l + \sum_{l=1}^{p} \left((x_k + \delta_k - v_i)^T \right)^l \left(r_i^{-1} \right)^{lj} \right) \right] \\ &= \left[\sum_{k=1}^{n} u_{ki} \sum_{l=1}^{p} \left(r_i^{-1} \right)^{jl} (x_k + \delta_k - v_i)^l \right] \\ &= \sum_{k=1}^{n} u_{ki} R_i^{-1}(x_k + \delta_k - v_i) = 0. \end{aligned}$$

上式の変形から，クラスタ中心の最適解は得られる．

$$v_i = \frac{\sum_{k=1}^{n} u_{ki}(x_k + \delta_k)}{\sum_{k=1}^{n} u_{ki}}. \tag{4.17}$$

次に帰属度の最適解を導出する．帰属度には制約条件があるため，ラグランジュの未定乗数法を用いて最適解を導出する．(4.16) を u_{ki} に関して偏微分すると，

$$\frac{\partial L_{\mathrm{KLFCMQ}}}{\partial u_{ki}} = (x_k + \delta_k - v_i)^T R_i^{-1}(x_k + \delta_k - v_i) + \lambda\left(\log\frac{u_{ki}}{\pi_i} + 1\right) + \log|R_i| - \gamma_k = 0.$$

この式を u_{ki} に関して整理していく．

$$\begin{aligned}
\log u_{ki} &= -\frac{1}{\lambda}(x_k + \delta_k - v_i)^T R_i^{-1}(x_k + \delta_k - v_i) + \log\pi_i - 1 - \frac{1}{\lambda}\log|R_i| + \frac{\gamma_k}{\lambda}, \\
u_{ki} &= \pi_i |R_i|^{-\frac{1}{\lambda}} \exp\left(-\frac{1}{\lambda}(x_k + \delta_k - v_i)^T R_i^{-1}(x_k + \delta_k - v_i)\right)\exp\left(\frac{\gamma_k}{\lambda} - 1\right). \qquad (4.18)
\end{aligned}$$

制約条件 $\sum_{i=1}^{c} u_{ki} = 1$ より，

$$\begin{aligned}
\sum_{i=1}^{c} u_{ki} &= \exp\left(\frac{\gamma_k}{\lambda} - 1\right)\sum_{i=1}^{c}\pi_i|R_i|^{-\frac{1}{\lambda}}\exp\left(-\frac{1}{\lambda}(x_k + \delta_k - v_i)^T R_i^{-1}(x_k + \delta_k - v_i)\right) = 1, \\
\exp\left(\frac{\gamma_k}{\lambda} - 1\right) &= \frac{1}{\sum_{i=1}^{c}\pi_i|R_i|^{-\frac{1}{\lambda}}\exp\left(-\frac{1}{\lambda}(x_+\delta_k - v_i)^T R_i^{-1}(x_k + \delta_k - v_i)\right)}.
\end{aligned}$$

この式を，(4.18) に代入すると帰属度の最適解は次式で得られる．

$$u_{ki} = \frac{\pi_i|R_i|^{-\frac{1}{\lambda}}\exp\left(-\frac{1}{\lambda}(x_k + \delta_k - v_i)^T R_i^{-1}(x_k + \delta_k - v_i)\right)}{\sum_{j=1}^{c}\pi_j|R_j|^{-\frac{1}{\lambda}}\exp\left(-\frac{1}{\lambda}(x_k + \delta_k - v_j)^T R_j^{-1}(x_k + \delta_k - v_j)\right)}. \qquad (4.19)$$

クラスタの重みの最適解も帰属度と同様に，(4.16) を π_i に関して偏微分して，

$$\begin{aligned}
\frac{\partial L_{\mathrm{KFLCMQ}}}{\partial \pi_i} &= -\lambda\sum_{k=1}^{n} u_{ki}\frac{1}{\pi_i} - \tau = 0, \\
\pi_i &= -\frac{\lambda}{\tau}\sum_{k=1}^{n} u_{ki}. \qquad (4.20)
\end{aligned}$$

ここで，制約条件 $\sum_{i=1}^{c}\pi_i = 1$ 及び，$\sum_{i=1}^{c} u_{ki} = 1$ より，両辺で i についての総和をとると，

$$\begin{aligned}
-\frac{\lambda}{\tau}\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki} &= \sum_{i=1}^{c}\pi_i, \\
-\frac{\lambda}{\tau} &= \frac{1}{n}.
\end{aligned}$$

この関係を (4.20) に代入すると，クラスタの重みの最適解は次のように定まる．

$$\pi_i = \frac{1}{n}\sum_{k=1}^{n} u_{ki}. \tag{4.21}$$

分散共分散行列の最適解は，(4.15) を R_i に関して偏微分して求める．ここで (3.11)，(3.13) より，

$$\begin{aligned}\frac{\partial J_{\mathrm{KLFCMQ}}}{\partial R_i} &= \left[\frac{\partial}{\partial r_i^{jl}}\sum_{k=1}^{n} u_{ki}\left((x_k+\delta_k-v_i)^T R_i^{-1}(x_k+\delta_k-v_i)+log|R_i|\right)\right] \\ &= \sum_{k=1}^{n} u_{ki}\left(-R_i^{-1}(x_k+\delta_k-v_i)(x_k+\delta_k-v_i)^T R_i^{-1}+R_i^{-1}\right)=0.\end{aligned}$$

上式左辺の各項に左右から R_i を掛け，R_i に関して整理すると分散共分散行列の最適解を得る．

$$R_i = \frac{\sum_{k=1}^{n} u_{ki}(x_k+\delta_k-v_i)(x_k+\delta_k-v_i)^T}{\sum_{k=1}^{n} u_{ki}}. \tag{4.22}$$

不確実性の最適解は，クラスタ中心と同様に，(4.15) を δ_k に関して偏微分する．

$$\frac{\partial J_{\mathrm{KLFCMQ}}}{\partial \delta_k} = \sum_{i=1}^{c} u_{ki}R_i^{-1}(x_k+\delta_k-v_i)+W_k\delta_k = 0.$$

δ_k に関して整理すると，

$$\left(\sum_{i=1}^{c} u_{ki}R_i^{-1}+W_k\right)\delta_k = -\sum_{i=1}^{c} u_{ki}R_i^{-1}(x_k-v_i).$$

上式の両辺に左から $\left(\sum_{i=1}^{c} u_{ki}R_i^{-1}+W_k\right)^{-1}$ を掛け，分散共分散行列の最適解が求まる．

$$\delta_k = -\left(\sum_{i=1}^{c} u_{ki}R_i^{-1}+W_k\right)^{-1}\left(\sum_{i=1}^{c} u_{ki}R_i^{-1}(x_k-v_i)\right). \tag{4.23}$$

最後に，KLFCMQ のアルゴリズムを載せる．

Algorithm 11 KLFCMQ

KLFCMQ1 初期クラスタ分割を与え初期帰属度を決定し，初期不確実性を与える．

KLFCMQ2 (4.21) を用いてクラスタの重みを更新．

KLFCMQ3 (4.17) を用いてクラスタ中心を更新．

KLFCMQ4 (4.22) を用いて分散共分散行列を更新．

KLFCMQ5 (4.23) を用いてデータの不確実性を更新．

KLFCMQ6 (4.19) を用いて帰属度を更新．

KLFCMQ7 各パラメータが収束すれば終了．そうでなければ **KLFCMQ2** に戻る．

4.2.2 正則化 EM アルゴリズムに基づくクラスタリング

この節では KL 情報量正則化ファジィ c-平均法でおこなわれたような，KLFCMQ と EM アルゴリズムに基づくクラスタリングとの関連性について述べる．

クラスタ A_i を表す確率密度関数を次式で定義する．

$$p_i(x_k+\delta_k|\phi_i)=\frac{1}{(2\pi)^{\frac{p}{2}}|R_i|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x_k+\delta_k-v_i)^T R_i^{-1}(x_k+\delta_k-v_i)\right).$$

この確率密度関数を用いて EM アルゴリズムに基づくクラスタリングの尤度関数を次式で定義する．

$$Q_{\mathrm{REM}}(\Phi|\Phi')=\sum_{k=1}^{n}\sum_{i=1}^{c}\log[\pi_i p_i(x_k+\delta_k|\phi_i)]\frac{\pi_i' p_i(x_k+\delta_k|\phi_i')}{p(x_k+\delta_k|\Phi')}-\sum_{k=1}^{n}\delta_k^T R_k\delta_k. \tag{4.24}$$

本手法の最適化問題は，この尤度関数を最大にする密度関数のパラメータを推定することである．この目的関数は第ニ項に正則化項を含んでいるため，便宜的に正則化 EM アルゴリズム (REM，Regularized EM Algorithm) と呼ぶ．(4.24) の第ニ項は $x_k+\delta_k$ がすべて v_i に一致す

ることで失われる，分布の滑らかさを制御する正則化項であり，正則化 EM アルゴリズムは正則化最尤法 [44] の一種であるといえる．EM アルゴリズムに基づくクラスタリングと同様に (4.24) の $\frac{\pi'_i p_i(x_k+\delta_k|\phi'_i)}{p(x_k+\delta_k|\Phi')}$ はデータ x_k が与えられた時のクラスタ A_i が生起する事後確率であるため，

$$u_{ki} = \frac{\pi'_i p_i(x_k+\delta_k|\phi'_i)}{p(x_k+\delta_k|\Phi')} \tag{4.25}$$

とできる．各推定値についても同様に求めていく．正則化 EM アルゴリズムに基づくクラスタリングのラグランジュ関数は次式で定義される．

$$L_{\mathrm{REM}} = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki}\log[\pi_i p_i(x_k+\delta_k|\phi_i)] - \sum_{k=1}^{n}\delta_k^T R_k \delta_k - \tau\left(\sum_{i=1}^{c}\pi_i - 1\right). \tag{4.26}$$

はじめに，クラスタ中心を意味する平均の最適解を導出する．(4.24) を v_i に関して偏微分すると，

$$\begin{aligned}
\frac{\partial Q_{\mathrm{REM}}}{\partial v_i} &= \left[\frac{\partial Q_{\mathrm{REM}}}{\partial v_i^j}\right] \\
&= \left[\frac{\partial}{\partial v_i^j}\left(-\frac{1}{2}\sum_{k=1}^{n} u_{ki}(x_k+\delta_k-v_i)^T R_i^{-1}(x_k+\delta_k-v_i)\right)\right] \\
&= \left[\frac{1}{2}\sum_{k=1}^{n} u_{ki}\left(\sum_{l=1}^{p}\left(r_i^{-1}\right)^{jl}(x_k+\delta_k-v_i)^l + \sum_{l=1}^{p}\left((x_k+\delta_k-v_i)^T\right)^l\left(r_i^{-1}\right)^{lj}\right)\right] \\
&= \left[\sum_{k=1}^{n} u_{ki}\sum_{l=1}^{p}\left(r_i^{-1}\right)^{jl}(x_k-v_i)^l\right] \\
&= \sum_{k=1}^{n} u_{ki}R_i^{-1}(x_k+\delta_k-v_i) = 0.
\end{aligned}$$

上式の変形から，クラスタ中心の最適解は得られる．

$$v_i = \frac{\sum_{k=1}^{n} u_{ki}(x_k+\delta_k)}{\sum_{k=1}^{n} u_{ki}}. \tag{4.27}$$

混合比の最適解はラグランジュの未定乗数法を用いて求めるので，(4.26) を π_i に関して偏微分する．

$$\frac{\partial L_{\mathrm{REM}}}{\partial \pi_i} = \frac{1}{\pi_i}\sum_{k=1}^{n} u_{ki} - \tau = 0,$$
$$\tau\pi_i = \sum_{k=1}^{n} u_{ki}. \tag{4.28}$$

ここで，制約条件 $\sum_{i=1}^{c}\pi_i = 1$ 及び，$\sum_{i=1}^{c} u_{ki} = 1$ より，両辺で i についての総和をとると，

$$\tau\sum_{i=1}^{c}\pi_i = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ki},$$
$$\tau = n.$$

この関係を (4.28) に代入すると，クラスタの重みの最適解は次のように定まる．

$$\pi_i = \frac{1}{n}\sum_{k=1}^{n} u_{ki}. \tag{4.29}$$

分散共分散行列の最適解は，(4.24) を R_i に関して偏微分して求める．ここで (3.11)，(3.13) より，

$$\begin{aligned}\frac{\partial Q_{\mathrm{REM}}}{\partial R_i} &= \left[\frac{\partial}{\partial r_i^{jl}}\sum_{k=1}^{n} u_{ki}\left((x_k+\delta_k-v_i)^T R_i^{-1}(x_k+\delta_k-v_i) + \log|R_i|\right)\right] \\ &= \sum_{k=1}^{n} u_{ki}\left(-R_i^{-1}(x_k+\delta_k-v_i)(x_k+\delta_k-v_i)^T R_i^{-1} + R_i^{-1}\right) = 0.\end{aligned}$$

上式左辺の各項に左右から R_i を掛け，R_i に関して整理すると分散共分散行列の最適解を得る．

$$R_i = \frac{\displaystyle\sum_{k=1}^{n} u_{ki}(x_k+\delta_k-v_i)(x_k+\delta_k-v_i)^T}{\displaystyle\sum_{k=1}^{n} u_{ki}}. \tag{4.30}$$

最後に不確実性の最適解は，クラスタ中心と同様に，(4.24) を δ_k に関して偏微分する．

$$\frac{\partial Q_{\mathrm{REM}}}{\partial \delta_k} = \sum_{i=1}^{c} u_{ki}R_i^{-1}(x_k+\delta_k-v_i) + W_k\delta_k = 0.$$

δ_k に関して整理すると，

$$\left(\sum_{i=1}^{c} u_{ki} R_i^{-1} + W_k\right)\delta_k = -\sum_{i=1}^{c} u_{ki} R_i^{-1}(x_k - v_i).$$

上式の両辺に左から $\left(\sum_{i=1}^{c} u_{ki} R_i^{-1} + W_k\right)^{-1}$ を掛け，分散共分散行列の最適解が求まる．

$$\delta_k = -\left(\sum_{i=1}^{c} u_{ki} R_i^{-1} + W_k\right)^{-1}\left(\sum_{i=1}^{c} u_{ki} R_i^{-1}(x_k - v_i)\right). \tag{4.31}$$

これらの最適解を (4.17)，(4.19)，(4.21)，(4.22)，(4.23) と比較すると，$\lambda = 2.0$ の場合において，すべての解が一致していることがわかる．このことから，KLFCMQ は (4.24) で表わされる正則化 EM アルゴリズムに基づくクラスタリングを内包していることがわかる．

4.3 数値例

本節では提案手法と関連手法との比較を数値例実験を通じておこない，提案手法の有効性の検討と，不確実性ベクトルの効果について論じる．実験には人工データと一種類と実データ二種類を用いる．人工データとしては図 4.1 に示した二次元データを用いて，提案手法における不確実性ベクトルの扱いと，関連手法との関係について考察する．実データは 2.4 と同じく Iris データと Breast Cancer データを用いる．実データに関する詳細は 2.4 で記載した通りなので，本節では省略する．

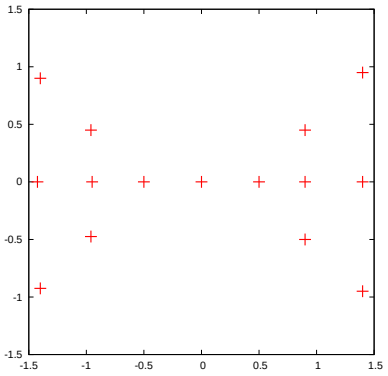

図 4.1: 人工データ

4.3.1 人工データに対するクラスタリング結果

ここでは人工データに対して，提案手法及び関連手法を用いて分類をおこなった結果を示す．各結果ではこのデータセットを二つのクラスタに分割するよう設定し，クラスタリングをおこなっている．

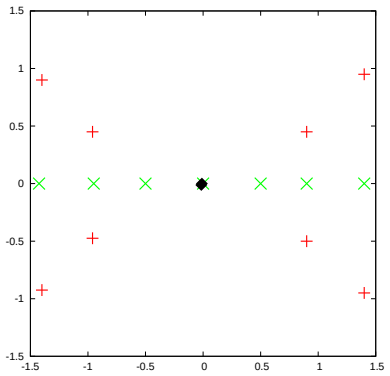

図 4.2: 人工データに対して KFCM を用いた結果 ($\lambda = 2.0$)

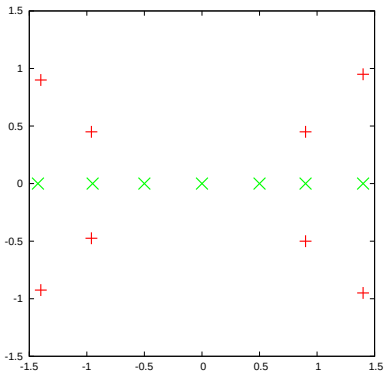

図 4.3: 人工データに対して EM アルゴリズムに基づくクラスタリングを用いた結果

図 4.2 – 4.3 は人工データに対して関連手法である KL 情報量正則化ファジィc-平均法，及び EM アルゴリズムに基づくクラスタリングを用いてクラスタリングをおこなった結果である．黒の菱型の点はクラスタ中心でありそれ以外の点はデータ点である．データ点は属するクラスタ毎に赤い十字，もしくは緑のクロスで描かれている．この結果では KL 情報量正則化ファジィc-平均法と EM アルゴリズムに基づくクラスタリングとの結果は同じである．

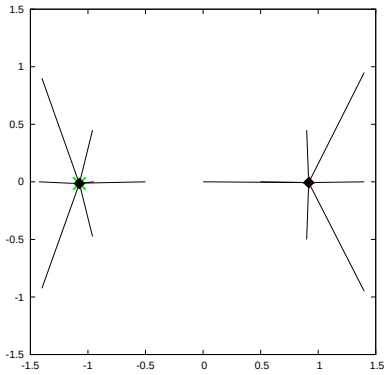

図 4.4: 人工データに対して KLFCMQ を用いた結果 ($W_k = 10E$, 初期 $\delta_k = 0$)

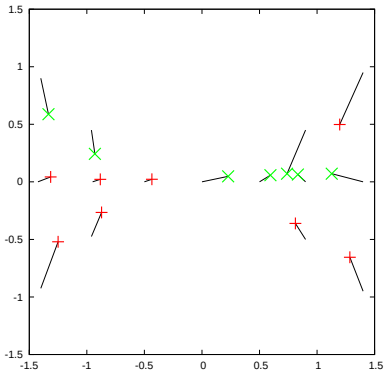

図 4.5: 人工データに対して REM アルゴリズムに基づくクラスタリングを用いた結果 ($W_k = 10E$, 初期 $\delta_k = 0$)

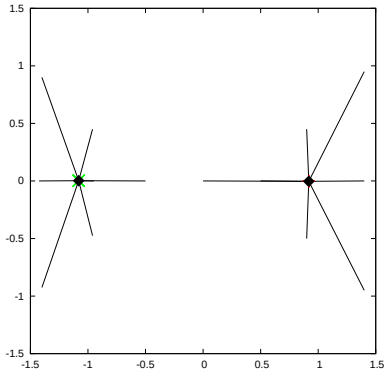

図 4.6: 人工データに対して KLFCMQ を用いた結果 ($W_k = 10E$, 初期 $\delta_k = 0.01$)

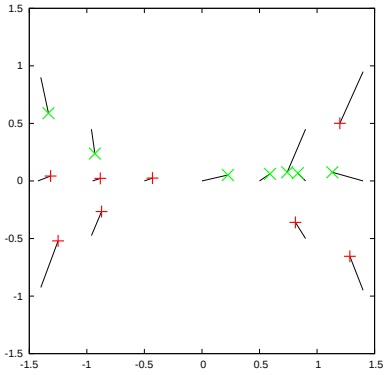

図 4.7: 人工データに対して REM アルゴリズムに基づくクラスタリングを用いた結果 ($W_k = 10E$, 初期 $\delta_k = 0.01$)

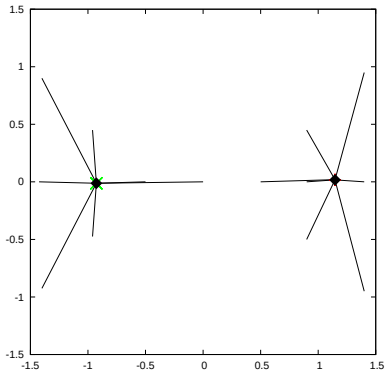

図 4.8: 人工データに対して KLFCMQ を用いた結果 ($W_k = 10E$, 初期 $\delta_k = 0.1$)

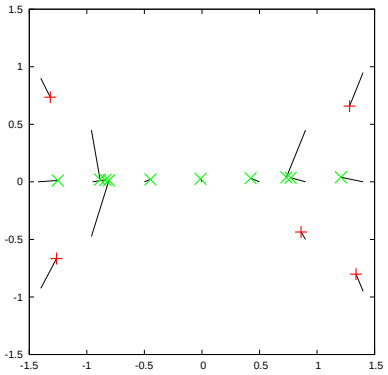

図 4.9: 人工データに対して REM アルゴリズムに基づくクラスタリングを用いた結果 ($W_k = 10E$, 初期 $\delta_k = 0.1$)

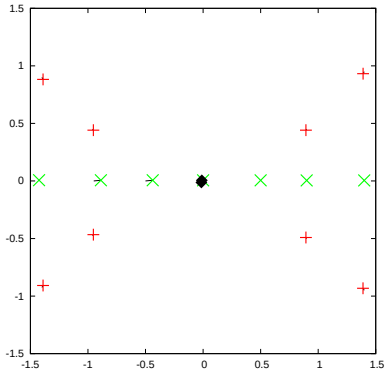

図 4.10: 人工データに対して KLFCMQ を用いたクラスタリング結果 ($W_k = 100E$, 初期 $\delta_k = 0$)

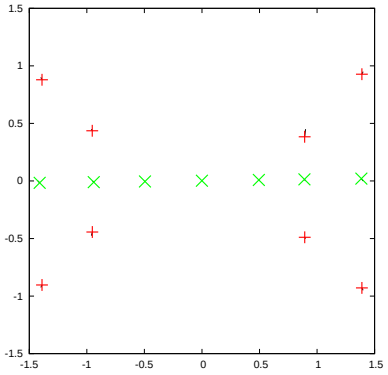

図 4.11: 人工データに対して REM アルゴリズムに基づくクラスタリングを用いたクラスタリング結果 ($W_k = 100E$, 初期 $\delta_k = 0$)

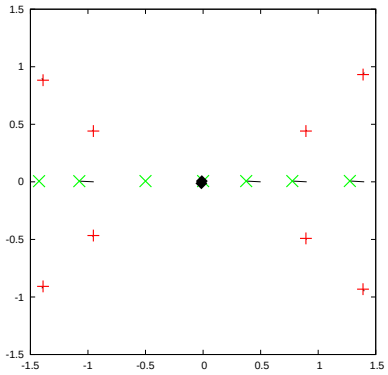

図 4.12: 人工データに対して KLFCMQ を用いたクラスタリング結果 ($W_k = 100E$, 初期 $\delta_k = 0.01$)

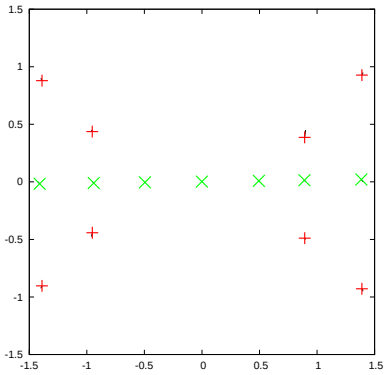

図 4.13: 人工データに対して REM アルゴリズムに基づくクラスタリングを用いたクラスタリング結果 ($W_k = 100E$, 初期 $\delta_k = 0.01$)

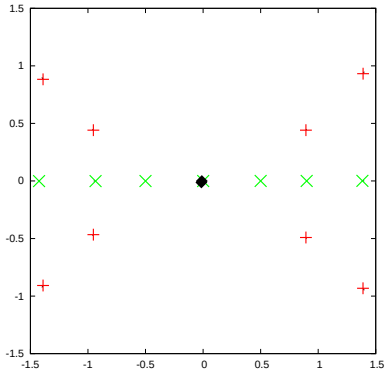

図 4.14: 人工データに対して KLFCMQ を用いたクラスタリング結果 ($W_k = 100E$, 初期 $\delta_k = 0.1$)

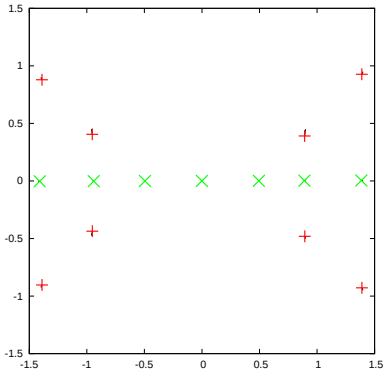

図 4.15: 人工データに対して REM アルゴリズムに基づくクラスタリングを用いたクラスタリング結果 ($W_k = 100E$, 初期 $\delta_k = 0.1$)

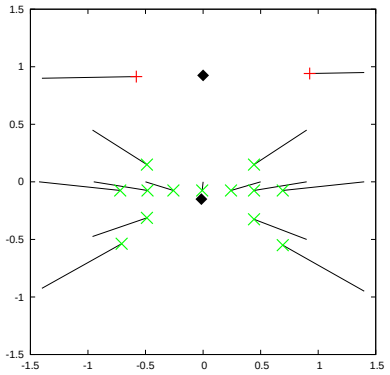

図 4.16: 人工データに対して KLFCMQ を用いたクラスタリング結果 ($W_k = R^{-1}_{\arg\max_j u_{kj}}$, 初期 $\delta_k = 0$)

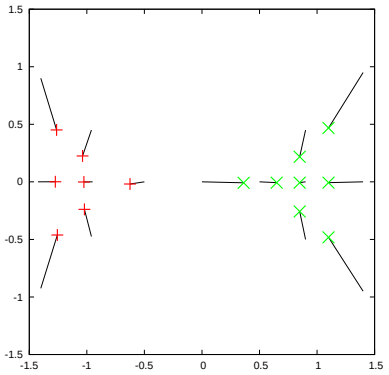

図 4.17: 人工データに対して REM アルゴリズムに基づくクラスタリングを用いたクラスタリング結果 ($W_k = R^{-1}_{\arg\max_j u_{kj}}$, 初期 $\delta_k = 0$)

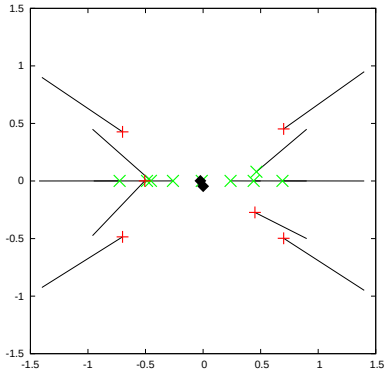

図 4.18: 人工データに対して KLFCMQ を用いたクラスタリング結果 ($W_k = R^{-1}_{\arg\max_j u_{kj}}$, 初期 $\delta_k = 0.01$)

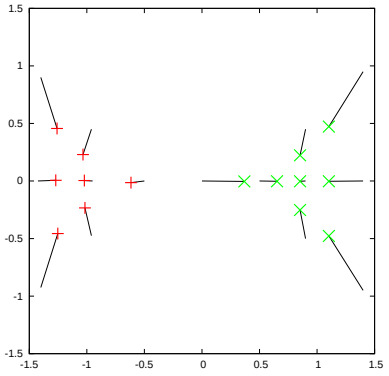

図 4.19: 人工データに対して REM アルゴリズムに基づくクラスタリングを用いたクラスタリング結果 ($W_k = R^{-1}_{\arg\max_j u_{kj}}$, 初期 $\delta_k = 0.01$)

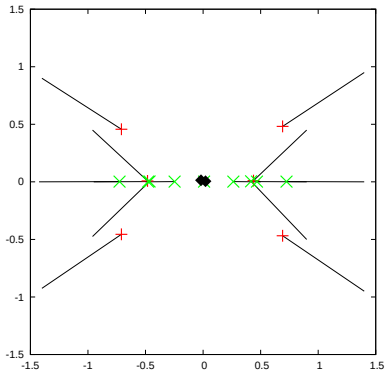

図 4.20: 人工データに対して KLFCMQ を用いたクラスタリング結果 ($W_k = R^{-1}_{\arg\max_j u_{kj}}$, 初期 $\delta_k = 0.1$)

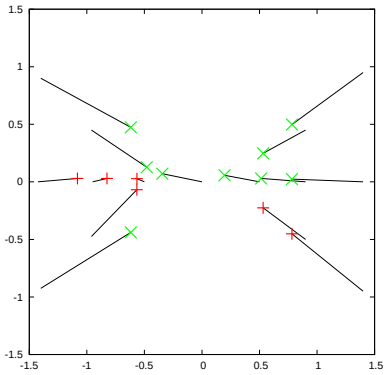

図 4.21: 人工データに対して REM アルゴリズムに基づくクラスタリングを用いたクラスタリング結果 ($W_k = R^{-1}_{\arg\max_j u_{kj}}$, 初期 $\delta_k = 0.1$)

図 4.4 – 4.21 は，人工データに対して KLFCMQ，及び REM アルゴリズムに基づくクラスタリングを用いた結果である．図の黒い直線は不確実性ベクトルを表しており，描かれているデータ点は元々黒い直線の先にあったことを意味している．これらの結果は不確実性ベクトルのペナルティ項の重み W_k と初期不確実性ベクトルが異なっている．W_k を大きくするとペナルティ項の大きさが大きくなるので，最終的に得られる不確実性ベクトルも小さくなることがわかる．ただしペナルティ項の目的関数と尤度関数に占める割合はそれぞれ異なるため，同じ重み，初期不確実性ベクトルを与えた結果同士でも，不確実性ベクトルの大きさは異なっている．また W_k を大きくしたそれぞれの結果は，関連手法である二手法の結果，図 4.2 – 4.3 と同じであることもわかる．そのためこの人工データにおいては，不確実性ベクトルは元々の手法の分類特徴を崩すことなく組み込まれているといえる．また図 4.4，図 4.6，図 4.8 から，ペナルティ項の重みが小さすぎるとクラスタ中心とすべてのデータが一致してしまう．図 4.16 – 4.21 はペナルティ項の重みとしてデータの属するクラスタの分散共分散行列の逆行列を用いており，ペナルティ項をマハラノビス距離として計算している．これらの結果では，すべてのデータがクラスタの中心に集まるように不確実性ベクトルが求められており，不確

実性ベクトルが正確に働いていることが確認できる．

4.3.2 Iris データに対するクラスタリング結果

次に実データの一つである Iris データに対して実験をおこなった結果を示す．Iris データの詳細は 2.4 のとおりである．元々の Iris データのラベルとクラスタリング結果との一致性は評価指標である Rand Index [45,46] を用いて評価している．

表 4.1 は Iris データに対して，提案手法及び関連手法を用いてクラスタリングをおこなった結果である．この表から KL 情報量正則化ファジィ c-平均法と EM アルゴリズムに基づくクラスタリングの結果は人工データの場合と異なり，同様の結果となっていないことがわかる．これは目的関数 (4.1) と尤度関数 (3.1) の違いからもたらされるものであることがわかっている．提案手法について見ていくと，初期の不確実性ベクトルを 0 とした結果よりも予め不確実性を与えたときの結果の方が改善されている．提案手法では W_k の値を大きくすると関連手法の結果と同等となると考えられるが，本実験結果からは見いだせない．また W_k の値にデータの属するクラスタの分散共分散行列の逆行列を用いた場合，KLFCMQ では関連手法よりも高い精度を得ることができたが，REM アルゴリズムに基づくクラスタリングでは関連手法よりも精度が下がる結果となった．

表 4.1: Iris データに対するクラスタリング結果の正答率

Algorithm	λ	初期 δ_k	W_k	正答数	Rand Index 値
KLFCMQ	2.0	0	$10000I$	77	0.614
KLFCMQ	2.0	0	$100I$	98	0.727
KLFCMQ	2.0	0	$R^{-1}_{\arg\max_j u_{kj}}$	108	0.782
KLFCMQ	2.0	0.01	$10000I$	93	0.678
KLFCMQ	2.0	0.01	$100I$	99	0.762
KLFCMQ	2.0	0.01	$R^{-1}_{\arg\max_j u_{kj}}$	147	0.974
KLFCMQ	2.0	0.1	$10000I$	86	0.627
KLFCMQ	2.0	0.1	$100I$	101	0.772
KLFCMQ	2.0	0.1	$R^{-1}_{\arg\max_j u_{kj}}$	143	0.942
REM アルゴリズムに基づくクラスタリング	—	0	$10000I$	95	0.654
REM アルゴリズムに基づくクラスタリング	—	0	$100I$	60	0.446
REM アルゴリズムに基づくクラスタリング	—	0	$R^{-1}_{\arg\max_j u_{kj}}$	77	0.593
REM アルゴリズムに基づくクラスタリング	—	0.01	$10000I$	115	0.791
REM アルゴリズムに基づくクラスタリング	—	0.01	$100I$	100	0.691
REM アルゴリズムに基づくクラスタリング	—	0.01	$R^{-1}_{\arg\max_j u_{kj}}$	90	0.66
REM アルゴリズムに基づくクラスタリング	—	0.1	$10000I$	141	0.927
REM アルゴリズムに基づくクラスタリング	—	0.1	$100I$	85	0.602
REM アルゴリズムに基づくクラスタリング	—	0.1	$R^{-1}_{\arg\max_j u_{kj}}$	84	0.654
KFCM	2.0	—	—	98	0.759
EM アルゴリズムに基づくクラスタリング	—	—	—	145	0.957

4.3.3 Breast Cancer データに対するクラスタリング結果

最後に Breast Cancer データに対して実験をおこなった結果を示す．Breast Cancer データの詳細は 2.4 のとおりである．元々の Breast Cancer データのラベルとクラスタリング結果との一致性は Iris データ同様，Rand Index を用いて評価している．

表 4.2 は Breast Cancer データに対して，提案手法及び関連手法を用いてクラスタリングをおこなった結果である．これらの結果から Iris データに対する結果と同様に，最適解が等しくなる KLFCMQ と REM アルゴリズムに基づくクラスタリング，KFCM と EM アルゴリズムに基づくクラスタリングはそれぞれ目的関数と尤度関数の違いから結果が異なることが見て取れる．また W_k の値にデータの属するクラスタの分散共分散行列の逆行列を用いた場合も，KLFCMQ は KL 情報量正則化ファジィ c-平均法で得られる結果より精度が上がり，REM アルゴリズムに基づくクラスタリングでは関連手法よりも精度が下がる結果となる点も同様である．しかしながら，初期の不確実性ベクトルの与え方に関しては Iris データのときと異なり，予め不確実性を与えた方が必ずしも結果が改善されるわけではない．これらのことから，初期の不確実性の与え方はデータセットに依存すると考えるのが妥当である．

表 4.2: Breast Cancer データに対するクラスタリング結果の正答率

Algorithm	λ	初期 δ_k	W_k	正答数	Rand Index 値
KLFCMQ	2.0	0	$10000I$	332	0.614
KLFCMQ	2.0	0	$100I$	396	0.791
KLFCMQ	2.0	0	$R^{-1}_{\arg\max_j u_{kj}}$	395	0.788
KLFCMQ	2.0	0.01	$10000I$	325	0.599
KLFCMQ	2.0	0.01	$100I$	267	0.517
KLFCMQ	2.0	0.01	$R^{-1}_{\arg\max_j u_{kj}}$	395	0.788
KLFCMQ	2.0	0.1	$10000I$	326	0.601
KLFCMQ	2.0	0.1	$100I$	356	0.671
KLFCMQ	2.0	0.1	$R^{-1}_{\arg\max_j u_{kj}}$	400	0.805
REM アルゴリズムに基づくクラスタリング	—	0	$10000I$	421	0.883
REM アルゴリズムに基づくクラスタリング	—	0	$100I$	263	0.514
REM アルゴリズムに基づくクラスタリング	—	0	$R^{-1}_{\arg\max_j u_{kj}}$	284	0.534
REM アルゴリズムに基づくクラスタリング	—	0.01	$10000I$	419	0.875
REM アルゴリズムに基づくクラスタリング	—	0.01	$100I$	306	0.565
REM アルゴリズムに基づくクラスタリング	—	0.01	$R^{-1}_{\arg\max_j u_{kj}}$	327	0.603
REM アルゴリズムに基づくクラスタリング	—	0.1	$10000I$	421	0.883
REM アルゴリズムに基づくクラスタリング	—	0.1	$100I$	292	0.544
REM アルゴリズムに基づくクラスタリング	—	0.1	$R^{-1}_{\arg\max_j u_{kj}}$	312	0.575
KFCM	2.0	—	—	306	0.565
EM アルゴリズムに基づくクラスタリング	—	—	—	422	0.887

第5章　おわりに

5.1　まとめ

本論文では，クラスタリングにおける不確実性を体系化した上でそれらの内クラスタ表現に含まれる不確実性とデータ自身に含まれる不確実性に着目し，それらに対応する新たなクラスタリング手法を構築した．

クラスタ表現に含まれる不確実性は，一般的に帰属度表現へのあいまいさの導入によって考慮される．本論文では，帰属度表現のあいまいさを扱う新たな概念であるラフ集合論に基づいたクラスタリング手法を構築した．この提案手法は全部で五種類あるが大別すると二種類に分けられる．その内の一つが既存のラフクラスタリング手法である rough k-means の最適解を基に構築したラフクラスタリング手法であり，もう一つが代表的なクラスタリング手法であるハード c-平均法の目的関数を基に構築したラフクラスタリング手法である．残りの三種類は，これらの手法の制約の一部を緩和したファジィラフクラスタリング手法となっている．そして構築した提案手法の分類特性の把握とクラスタ分類の精度を，数値例実験をおこない検証した．数値例実験では人工データと実データである Iris データ，Breast Cancer データを用いた．実データを用いた検証の結果，提案手法の多くは既存手法に劣らぬ結果を得ることがわかった．また，人工データと実データを用いた実験を通して，提案手法の分類特性をある程度明らかにすることができた．提案手法の特に大きな利点としては，既存手法の問題点であったパラメータ選択と最適解選択の問題点を解消した上で，クラスタ表現に含まれる不確実性を考慮している点である．

データ自身に含まれる不確実性は，データを不確実性を伴った区間データや確率密度関数として扱うことで考慮される．そこで本論文では，各データをガウス分布として扱う EM アルゴリズムに基づくクラスタリング手法と，各データに不確実性ベクトルを導入した EM アルゴリズムに基づくクラスタリング手法を構築した．

データをガウス分布として扱う EM アルゴリズムでは，データを示す確率密度関数とクラスタを示す確率密度関数の合成関数を用いている．提案手法では一次元の場合と多次元の場合に関して合成関数と最適解を求め，それぞれのアルゴリズムを構築した．数値例実験では二つの一次元実データを用いて，提案手法と既存手法との比較をおこなった．データに不確実性を導入したことによる既存手法との大きな差は出なかったものの，既存手法の分類と同等の分類結果を得ることができた．また提案手法の分類結果は，パラメータであるデータの不確実性の分散値に対してロバスト性を持つことが期待できる．

不確実性ベクトルを EM アルゴリズムに導入するために，EM アルゴリズムと関連が深い KL 情報量正則化に基づくファジィc-平均法に不確実性ベクトルを導入し，不確実性ベクトルを導入した EM アルゴリズムを構築した．そして人工データ及び Iris データ，Breast Cancer データを用いた数値例実験をおこない，不確実性ベクトルの働きとその効果を既存手法との比較によって検証し，従来の不確実性ベクトルを導入した手法と同じように働いていることを確認した．不確実性ベクトルの初期値や不確実性ベクトルのペナルティ項に関しては今回の数値例実験を通した検証しかおこなえておらず，それらのパラメータが結果にどのような影響を及ぼすのかより深く調べる必要がある．

5.2 今後の展望

本論文で提案したラフクラスタリングは既存手法の問題点を大方取り除くことに成功したが，下近似係数値の選択問題が残っている．数値例実験では下近似係数を変化させた場合の検証をおこなっているが，何をもって最もふさわしい値であるかを判断するのは難しく，現

状は使用者依存となってしまっている．そのためより多くの数値例実験を通して，各手法の下近似係数による分類特性の検証をおこなう必要がある．

データをガウス分布として扱う EM アルゴリズムの数値例実験では，一次元の場合の数値例実験しかおこなえていない．そのため多次元の場合の数値例実験をおこなうことは必要不可欠である．また，一次元多次元の場合ともに一部の最適解を近似解として導出している．そのためその近似解の導出方法を変化させることで得られる結果も異なることが予想される．そこで様々な近似解の導出方法を試し，それらの解による数値例実験の結果の違いや，不確実性の効果について検証し，提案手法に最も適した近似解の導出法を模索する必要がある．

不確実性ベクトルを導入した手法では，初期値として与える不確実性ベクトルの大きさとペナルティ項の影響力を決定するのが最も難しい．特に提案手法では通常のクラスタリング手法と異なり，ユークリッド距離ではなくマハラノビス距離を用いている．不確実性ベクトルを導入した既存手法は様々なものがあるが，マハラノビス距離を扱っているものはなく，ペナルティ項の影響力を予想するのが困難である．そのため数値例実験を用いて，ペナルティ項のパラメータを様々に変化させた場合のクラスタ分類に及ぼす影響を調べ，本手法に最も適したペナルティ項を定めるのが課題である．

謝辞

本論文は，著者が筑波大学大学院システム情報工学研究科に在学中におこなった研究成果をまとめたものです．この期間中，ご指導，ご助言並びにご協力を頂いた方々に心から御礼申し上げます．

卒業研究から始まり博士論文作成までの五年間という長きに渡り，ご指導を賜りました筑波大学大学院 システム情報系 遠藤靖典 教授に深く感謝致します．研究の素養を持ち合わせていなかった私を，博士論文を完成させるまでにご指導していただきました．そして，研究の方針だけでなく私生活に至るまでの様々なアドバイスを通して，研究姿勢や生活態度など非常に多くのことを学ばせていただきました．

また，システム情報系の宮本定明 教授，イリチュ (佐藤) 美佳 教授にはソフトコンピューティング基礎グループの研究発表会，学会発表への参加，そして論文審査に至るまであらゆる機会を通じてご指導いただきました．さらに，システム情報系の亀山啓輔 教授，芝浦工業大学大学院 理工学研究科の神澤雄智 准教授からの論文審査での厳しいご講評・ご指摘は，本論文を作成するに当たり非常に有益なものとなりました．ここに感謝の意を表します．

さらに，ソフトコンピューティング基礎グループの皆様には，楽しく快適な研究室環境を提供していただき，研究室内やゼミでの議論を通してたくさんの貴重な意見をいただきました．ここに感謝の意を表します．

最後になりましたが，博士論文作成に至るまで学生生活を支え暖かく見守っていただいた両親に心から感謝致します．

参考文献

[1] 宮本定明, クラスター分析入門, 森北出版 (1999).

[2] J. MacQueen，*Some methods for classification and analysis of multivariate observations*，Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability， Vol.1，pp.281–297 (1967).

[3] E. W. Forgy, *Cluster analysis of multivariate data: Efficiency versus interpretability of classification*, Biometrics, Vol.21, p.768–769 (1965).

[4] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice-Hall (1988).

[5] P. S. Bradley, O. L. Mangasarian, and W. N. Street, *Clustering via Concave Minimization*, Advances in Neural Information Processing Systems, Vol.9, pp.368–374 (1997).

[6] U. Luxburg, *A Tutorial on Spectral Clustering*, Statistics and Computing, Vol.17, No.4, pp.395–416 (2007).

[7] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, Proceedings of 2nd International Conference on Knowledge Discovery, pp.226–231 (1996).

[8] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press (1981).

[9] Z.-Q. Liu and S. Miyamoto, *Soft Computing and Human-Centered Machines*, pp.85–129, Springer–Verlag (2000).

[10] R. Krishnapuram and J. M. Keller, *A Possibilistic Approach to Clustering*, IEEE Transactions on Fuzzy Systems, Vol.1, No.2, pp.98–110 (1993).

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society. Series B (Methodological), Vol.39, No.1, pp.1-38 (1977).

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, pp.423–455, Springer–Verlag (2006).

[13] R. J. Hathaway and J. W. Davenport, *Relational duals of the c-means clustering algorithms*, Pattern Recognition, Vol.22, No.2, pp.205–212 (1989).

[14] L. A. Zadeh, *Fuzzy Sets*, Information and Control, Vol.8, pp.338–353 (1965).

[15] Z. Pawlak, *Rough Sets*, International Journal of Computer and Information Sciences, Vol.11, No.5, pp.341–356 (1982).

[16] 乾口雅弘, ラフ集合による情報の解析, Institute of Systems, Control and Information Engineers, Vol.49, No.5, pp.165–172 (2005).

[17] P. Lingras, C. West, *Interval Set Clustering of Web Users with Rough K-Means*, Journal of Intelligent Information Systems, Vol.23, No.1, pp.5-16 (2004).

[18] P. Lingras, G. Peters, *Rough clustering*, WIREs Data Mining and Knowledge Discovery, Vol.1, No.1, pp.64-72 (2011).

[19] S. Mitra, *An evolutionary rough partitive clustering*, Pattern Recognition Letters, Vol.25, No.12, pp.1439–1449 (2004).

[20] G. Peters, *Outliers in Rough k-Means Clustering*, Proceedings of PReMI 2005, LNCS 3776, pp.702-707 (2005).

[21] S. Mitra, H. Banka, and W. Pedrycz, *Rough-Fuzzy Collaborative Clustering*, IEEE Transactions Systems，Man and Cybernetics Part B, Vol.36, No.4, pp.795–805 (2006).

[22] G. Peters, *Some refinements of rough k-means clustering*, Pattern Recognition, Vol.39, No.8, pp.1481–1491 (2006).

[23] G. Peters, *Rough Clustering and Regression Analysis*, Proceedings of RKST 2007, LNAI 4481, pp.292–299 (2007).

[24] P. Maji, S. K. Pal, *Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices*, IEEE Transactions Systems，Man and Cybernetics Part B, Vol.37, No.6, pp.1529–1540 (2007).

[25] S. Mitra, B. Barman, *Rough-Fuzzy Clustering: An Application to Medical Imagery*, Proceedings of RKST 2008, LNAI 5009, pp.300–307 (2008).

[26] R. J. Hathaway and J. C. Bezdek, *Fuzzy c-Means Clustering of Incomplete Data*, IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, Vol.31, No.5, pp.735–744 (2001).

[27] 高田治, 宮本定明, 区間データに基づく不確定性を含むデータのファジィクラスタリング, Japan Society for Fuzzy Theory and Systems, Vol.12, No.5, pp.686–695 (2000).

[28] 高田治, 宮本定明, ファジィデータに対する L_1 距離を用いたファジィクラスタリング, Japan Society for Fuzzy Theory and Systems, Vol.13, No.6, pp.689–698 (2001).

[29] T. Denoeux, *Maximum likelihood estimation from Uncertain Data in the Belief Function Framework*, IEEE Transactions on Knowledge and Data Engineering, Vol.25, No.1, pp.119–130 (2011).

[30] R. Murata, Y. Endo, H. Haruyama, and S. Miyamoto, *On Fuzzy c-Means for Data with Tolerance*, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.10, No.5, pp.673–681 (2006).

[31] Y. Hamasuna, Y.Endo, Y. Hasegawa, and S. Miyamoto, *Two Clustering Algorithms for Data with Tolerance based on Hard c-Means*, Proceedings of 2007 IEEE International Conference on Fuzzy Systems, pp.688–691 (2007).

[32] Y. Endo, K. Kurihara, S. Miyamoto, and Y. Hamasuna, *Hard and Fuzzy c-Regression Models for Data with Tolerance in Independent and Dependent Variables*, Proceedings of The 2010 IEEE World Congress on Computational Intelligence, pp.1842–1849 (2010).

[33] Y. Endo, Y. Hasegawa, Y. Hamasuna, and Y. Kanzawa, *Fuzzy c-Means Clustering for Uncertain Data Using Quadratic Penalty-Vector Regularization*, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.15, No.1, pp.76–82 (2011).

[34] Y. Hamasuna, Y. Endo, and S. Miyamoto, *On Mahalanobis Distance Based Fuzzy c-Means Clustering for Uncertain Data Using Penalty Vector Regularization*, Proceedings of IEEE International Conference on Fuzzy Systems, pp.810–815 (2011).

[35] D. Arthur and S. Vassilvitskii, *k-means++: the advantages of careful seeding*, Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms, pp.1027–1035 (2007).

[36] Y. Endo, *A Note on Spherical k-Means++ Clustering*, Proceedings of 2015 Conference of the International Federation of Classification Societies, (2015).

[37] 宮岸聖高, 市橋秀友, 本多克宏, K-L 情報量正則化 FCM クラスタリング法, Japan Society for Fuzzy Theory and Systems, Vol.13, No.4, pp.64–75 (2001).

[38] 平野章二, 津本周作, ラフクラスタリングによる医療データの類型化の試み, The 19th Annual Conference of the Japanese Society for Artificial Intelligence (2005).

[39] 加藤功記, 村井哲也, 工藤康生, 佐藤義治, 平野らのラフ・クラスタリング法の非対称データへの適用 (第 2 報), The 23rd Fuzzy System Symposium, pp.260–265 (2007).

[40] R. A. Fisher, Iris Data Set, http://archive.ics.uci.edu/ml/datasets/Iris.

[41] W. H. Wolberg, Breast Cancer Wisconsin (Original) Data Set, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original).

[42] 世界各国の平均身長 (男性), http://www.suku-noppo.jp/data/world_average_height_boy.html.

[43] IMF World Economic Outlook Database, 世界の名目 GDP (US ドル) ランキング, http://ecodb.net/ranking/imf_ngdpd.html.

[44] 川野秀一, 廣瀬慧, 立石正平, 小西貞則, 回帰モデリングと L_1 型正則化法の最近の展開, 日本統計学会, Vol.39, No.2, pp.211–242 (2010).

[45] W. M. Rand, *Objective Criteria for the Evaluation of Clustering Methods*, Journal of the American Statistical Association, Vol.66, No.336, pp.846–850 (1971).

[46] L. Hubert and P. Arabie, *Comparing Partitions*, Journal of Classification, Vol.2, No.1, pp.193–218 (1985).

参考論文

- 査読付き学術論文
  - Naohiko Kinoshita, Yasunori Endo, Yukihiro Hamasuna, *Fuzzy c-Means with Quadratic Penalty-Vector Regularization Using Kullback-Leibler Information for Uncertain Data*, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.19, No.5, pp.624–637 (2015.9).
  - Naohiko Kinoshita, Yasunori Endo, Ken Onishi, *On Objective-based Rough Clustering with Fuzzy-Set Representation*, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.19, No.5, pp.632–638 (2015.9).
  - Naohiko Kinoshita, Yasunori Endo, *On Objective-based Rough Hard and Fuzzy c-Means Clustering*, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.19, No.1, pp.29–35 (2015.1).
  - Yasunori Endo, Naohiko Kinoshita, *Objective-Based Rough c-Means Clustering*, International Journal of Intelligent Systems, Vol.28, No.9, pp.907–925 (2013.9).
- 査読付き国際会議論文
  - Naohiko Kinoshita, Yasunori Endo, Yukihiro Hamasuna, *Fuzzy c-Means with Quadratic Penalty-Vector Regularization Using Kullback-Leibler Information for Uncertain Data*, Proceedings of The 11th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2014), USB, pp.1-13 (Tokyo, Japan, 2014.10.29).
  - Naohiko Kinoshita, Yasunori Endo, Sadaaki Miyamoto, *On Some Models of Objective-based Rough Clustering*, The 2014 IEEE/WIC/ACM International Conference on Web Intelligence (WIC204) (Warsaw, Poland, 2014.8.11–14).

– Naohiko Kinoshita, Yasunori Endo, *EM-based Clustering Algorithm for Uncertain Data*, Proceedings of The 5th International Conference on Knowledge and Systems Engineering (KSE 2013), Advances in Intelligent and Soft Computing 245, Springer, Vol.2, pp.69–81 (Hanoi, Vietnam, 2013.10.17).

– Naohiko Kinoshita, Yasunori Endo, Sadaaki Miyamoto, *Regularized Rough c-Means and Applications to Risk Analysis*, Proc. of The Fifth International Symposium on Computational Intelligence and Industrial Applications (ISCIIA 2012), 6p. (Sapporo, Japan, 2012.8.22).

– Yasunori Endo, Naohiko Kinoshita, *On Objective-Based Rough c-Means Clustering*, Proc. of The 2012 IEEE International Conference on Granular Computing (GrC 2012), #SA002, pp.123–128 (Hangzhou, China, 2012.8.12).

– Ken Onishi, Yasunori Endo, Naohiko Kinoshita, *A Note on Objective-based Rough Clustering with Fuzzy-Set Representation*, The 11th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2014), LNAI 8825, Springer, pp.122–134 (Tokyo, Japan, 2014.10.29).

- その他

– 木下 尚彦, 遠藤 靖典, 目的関数最適化に基づくラフクラスタリングについて, 第 38 回ファジィワークショップ, pp.27–30 (東京, 2012.3.16).

その他の論文

- 査読付き学術論文

– Naohiko Kinoshita, Yasunori Endo, Akira Sugawara, *On Hierarchical Linguistic-Based Clustering*, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.19, No.6, pp.900–906 (2015.11).

– Akira Sugawara, Yasunori Endo, Naohiko Kinoshita, *On Objective-based Rough c-Regression*, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.19, No.1, pp.36–42 (2015.1).

- 査読付き国際会議論文

  – Naohiko Kinoshita, Yasunori Endo, Yuchi Kanzawa, Sadaaki Miyamoto, *A Note on Non-Hierarchical Linguistic-based Clustering*, Proceedings of the 12th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2015), USB, pp.25–35 (Skoevde, Sweden, 2015.9.21).

  – Yasunori Endo, Tomoyuki Suzuki, Naohiko Kinoshita, Yukihiro Hamasuna, Sadaaki Miyamoto, *Fuzzy Non-metric Model for Data with Tolerance and Its Application to Incomplete Data Clustering*, Proceedings of 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2015), #15249 (Istanbul, Turkey, 2015.8.2-5).

  – Tsubasa Hirano, Yasunori Endo, Naohiko Kinoshita, Yukihiro Hamasuna, *On Even-sized Clustering Algorithm Based on Optimization*, Proceedings of Joint 7rd International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on advanced Intelligent Systems (SCIS & ISIS 2014), TP4–3–5–(3), #69 (Kitakyushu, Japan, 2014.12.4).

  – Yasunori Endo, Naohiko Kinoshita, Kuniaki Iwakura, Yukihiro Hamasuna, *Hard and Fuzzy c-means Algorithms with Pairwise Constraints by Non-metric Terms*, The 11th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2014), LNAI 8825, Springer, pp.145–157 (Tokyo, Japan, 2014.10.30).

  – Akira Sugawara, Naohiko Kinoshita, Yasunori Endo, *On Linguistic-based Clustering*, Proceedings of The 2014 IEEE International Conference on Granular Computing (GrC

2014), G273 (Noboribetsu, Hokkaido, Japan, 2014.10.23).

– Yasunori Endo, Akira Sugawara, Naohiko Kinoshita, *Rough c-Regression based on Optimization of Objective Function*, The 10th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2013), LNAI 8234, Springer, pp.272–283 (Barcelona, Spain, 2013.11.20).

- その他

  – Tsubasa Hirano, Naohiko Kinoshita, Yasunori Endo, Yukihiro Hamasuna, *Even-sized Clustering Algorithm Based on Optimization*, Doctoral Consortium Proceedings of The 11th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2014), USB, pp.16–18 (Tokyo, Japan, 2014.10.30).

  – Yu Shiraishi, Akira Sugawara, Naohiko Kinoshita, Yasunori Endo, *A Note on Visualization of Asymmetric Data*, Doctoral Consortium Proceedings of The 11th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2014), USB, pp.19–21 (Tokyo, Japan, 2014.10.30).

  – Kuniaki Iwakura, Yasunori Endo, Naohiko Kinoshita, Sadaaki Miyamoto, *On Relation Between Kernelization and Quadratic-Regularization in Hard Non-Metric Model*, Doctoral Consortium Proceedings of The 11th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2014), USB, pp.22–24 (Tokyo, Japan, 2014.10.30).

  – 栢分 雄基, 菅原 彬, 木下 尚彦, 遠藤 靖典, 神澤 雄智, 力学モデルに基づく階層型言語ベースクラスタリング, 第 41 回ファジィワークショップ講演論文集, pp.5–8 (八王子市南大沢, 2015.3.6-7).

  – 白石 遊, 遠藤 靖典, 菅原彬, 木下 尚彦, 非対称データの可視化に関する一考察, 第 30 回ファジィシステムシンポジウム (FSS2014) (高知, 2014.9.2 (1-3)).