クラスタサイズ調整変数を導入した クラスタリング手法の特性比較及び精度評価

Characteristic Comparison and Accuracy Evaluation of Clustering Method with Cluster Size Adjustment Variable

AF16009 池辺 颯一 Soichi Ikebe
指導教員 神澤 雄智 Yuchi Kanzawa

1 はじめに

近年，情報通信社会の発展に伴いデータ量が増大し，日々多様なデータがコンピュータに蓄積されている．この大量のデータから有益な情報を抽出する手法として，データを類似度に基づきグループ化するクラスタリングに注目が集まっている．既存の手法における課題として，各クラスタのサイズに差がある場合，クラスタリングから有意な結果が得られないというものがある．そこで，各クラスタのサイズを考慮してクラスタリングを行う手法が複数提案されており，本研究はそれらの手法について各手法の特性を把握するとともに，最も有効な手法を発見することを目的とする．

2 実験内容

各クラスタのサイズを考慮するために，既存の手法にクラスタサイズ調整変数を導入した sFCMA [1], eFCMA [2], qFCMA [3] の 3 手法について実験を行う．

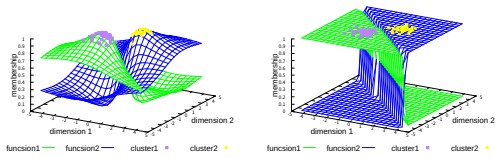
まず，これらの手法についてそれぞれの特性を把握するため，人工データを用いて実験を行う．複数のパラメータで実験を行い，それぞれで算出された分類関数 [4] から比較及び評価を行う．分類関数は，各クラスタに対する帰属度を座標空間上に可視化したもので，分類関数により，データがどのクラスタに属するかということが調べることができ，新たに与えられたデータ点についても，その帰属度を計算することができる．また，分類関数の曲面が滑らかであればその手法がファジィであり，平面に近ければクリスプであるということが分かる．

次に，これらの手法から最も有用なものを発見するために，実データを用いて Adjusted Rand Index(ARI) [5] を算出しする．ARI は，分類結果の精度評価を -1 から 1 までの範囲で行う指標で，1 の時に完全一致，0 の時にランダムであることを表す．本研究では ARI の値が最も高いものを有効な手法と評価する．

3 人工データの実験結果

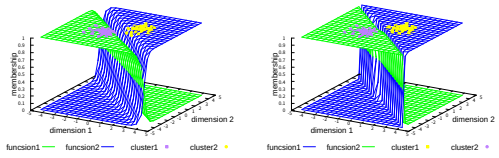
人工データとして，クラス数 2，各クラスのデータ数 50，合計データ数 100 のデータを平均値 $(-1, -1)$，標準偏差 $(0.5, 0.5)$ 及び平均値 $(1, 1)$，標準偏差 $(0.5, 0.5)$ のガウスサンプリングで生成したデータを用いた．また，初期値として，クラスタ中心にランダムな座標を与え，クラスタサイズ調整変数にクラスタ数の逆数を与えた．実験結果の図における垂直軸は帰属度を，底面はデータ空間を表す．網掛けで示されるのが分類関数であり，各点がデータを表している．また，分類関数とデータ点はそれぞれ 2 色に別れており，各色がそれぞれのクラスタに属することを表している．

sFCMA の実験結果を図 1a, 1b に示す．パラメータ m を 2.00 から 1.01 に変化させたところ，分類関数は m の値が大きいほどファジィになり，小さいほどクリスプになることが分かった．次に，eFCMA の実験結果を図 2a, 2b に示す．パラメータ λ を 1 から 10 に変化させたところ，分類関数は λ の値が小さいほどファジィになり，大きいほどクリスプになることが分かった．qFCMA の実験結果を図 3a, 3b, 3c に示す．こちらは，パラメータ (m, λ) の組み合わせとして，$(2.00, 1)$, $(1.01, 1)$, $(1.01, 10)$ の 3 通りでクラスタリングを行った．図 3a 及び図 3b の分類関数より，m の値が大きいほどファジィになり，小さいほどクリスプになることが分かった．また，図 3b 及び図 3c の分類関数より，λ の値が小さいほどファジィになり，大きいほどクリスプになることが分かった．そして，図 1 及び図 3a, 3b の分類関数より，qFCMA において $m - 1 \to +0$ とすると sFCMA と同じ特性が得られ，図 2 及び図 3b, 3c より，$\lambda \to \infty$ とすると eFCMA と同様の特性を示すことがわかった．これらの実験結果より qFCMA は sFCMA と eFCMA の特性を併せ持つと言える．



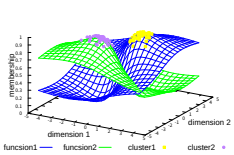
(a) m=2.00 **(b)** m=1.01

図 1: sFCMA の人工データの実験結果

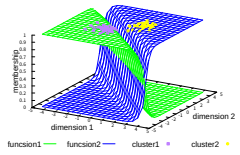


(a) λ=1 **(b)** λ=10

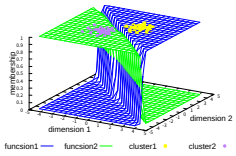
図 2: eFCMA の人工データの実験結果



(a) $m = 2.00, \lambda = 1$



(b) $m = 1.01, \lambda = 1$



(c) $m = 1.01, \lambda = 10$

図 3: qFCMA の人工データの実験結果

4　実データの実験結果

実データとしては，個体数 403，クラス数 4 の，被験者の勉強時間や試験結果などの 5 属性を収録した “User Knowledge Modeling Dasta Set” を用いた．また，初期値として，それぞれのデータの帰属度に正解の帰属度を与え，クラスタサイズ調整変数にクラスタ数の逆数を与えた．sFCMA, eFCMA, qFCMA の実データ実験の結果について，それぞれ図 4, 5, 6 に示す．sFCMA では m の値を 1.1 から 3.0 まで 0.1 刻み，eFCMA では λ の値を 1 から 100 まで 1 刻み，qFCMA では m の値を 1.1 から 3.0 まで 0.1 刻み，λ の値を 1 から 100 まで 1 刻みで変化させた．

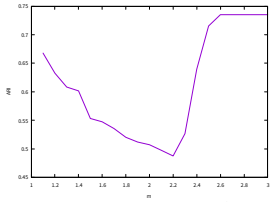


図 4: sFCMA の実データの実験結果

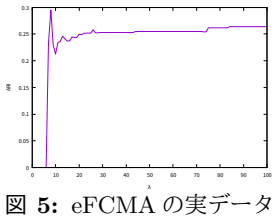


図 5: eFCMA の実データの実験結果

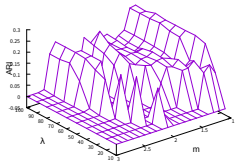


図 6: qFCMA の実データの実験結果

それぞれの手法の最高 ARI を表 1 に示す．最も高い ARI を示した手法は sFCMA であり，他の 2 手法と比較して ARI に 0.4 以上の差が見られた．

表 1: 各手法の ARI の最高値とパラメータ

手法名	ARI の最高値	パラメータ値
sFCMA	0.73515	$m = 3$
eFCMA	0.29500	$\lambda = 8$
qFCMA	0.26286	$\lambda = 80$, $m = 1.1$

5　まとめと今後の課題

既に提案されていた 3 種のクラスタリング手法の特性と精度について現在に至るまで明らかになっていなかったため，本研究では，人工データを用いた特性比較及び実データを用いた精度比較を行った．その結果として，sFCMA は m が大きくなるとファジィになり，eFCMA は λ が大きくなるほどクリスプになることが分かった．また，qFCMA は sFCMA と eFCMA の両方の特性を併せ持つということが分かった．精度は sFCMA が最も高評価となった．要因として，この手法の最適化問題にエントロピー項が含まれないということが考えられる．sFCMA の精度には，エントロピー項が含まれる eFCMA, qFCMA の 2 手法と比較して大きな差が見られた．今後の課題は，今回用いなかった他の実データで 3 手法の比較を行い，精度についての裏付けを行うことである．

参考文献

[1] Miyamoto, S., Kurosawa, N.: “Controlling Cluster Volume Sizes in Fuzzy c-means Clustering”, Proc. SCIS&ISIS2004, pp. 1–4, (2004).

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

[3] Miyamoto, S., Ichihashi, H., and Honda, K.: Algorithms for Fuzzy Clustering, Springer (2008).

[4] 宮本 定明, 馬屋原 一孝, 向殿 政男:“ファジイ c-平均法とエントロピー正則化法におけるファジィ分類関数”, 日本ファジィ学会誌 Vol. 10, No. 3 pp. 548–557, (1998).

[5] Hubert, L., and Arabie, P.:“Comparing Partitions”, Journal of Classifcation, Vol. 2, No. 1, pp. 193–218, (1985).