論文メモ

文献番号	0016
日付	2021 年 01 月 21 日
名前	武川海斗

文献情報

著者	R. R. Yager and D. P. Filev
英文タイトル	Approximate Clustering Via the Mountain Method
和文タイトル	マウンテン法による近似クラスタリング
書誌情報	IEEE TRANSACTIONS ON SYSTEM，Vol. 24，No. 8，pp. 1279–1284（ページ），1994
キーワード	

1　論文のトピック

本論文は，ファジィ c-means の中心を推定する手法として，マウンテンクラスタリングを提案している．マウンテンクラスタリングとは，データ空間を格子状で与え，格子状の座標とデータの距離に基づき値を与える手法である．

2　ベースとなった手法

2.1　ファジィ c-means

ファジィ c-means はクラスター中心と帰属度について交互最適化を行うことでクラスタリングを行う．

$$J_m = \sum_{k=1}^{n} \sum_{i=1}^{q} v_{ik}^{m} \left| x_k - x_i \right|^m \tag{1}$$

データ数を n，クラスタ数を q とし，v_{ik} はクラスタ i に属するデータ k を表す．また，x_i はクラスタ中心である．

3　提案手法のコア要素

3.1　マウンテンクラスタリング

マウンテンクラスタリングは，データ空間を格子状で表し格子の頂点とデータの距離から算出した値を与える．この値を M 値とし，以下の式で与える．M 値が高いほど，データの格子の頂点にデータが集まっていることを表す．

$$M\left(N_i\right) = \sum_{k=1}^{n} e^{-\alpha d\left(x_k, N_i\right)} \tag{2}$$

ここで，N_i は i 個目の格子の頂点を表し，α はパラメータである．(x_k, N_i) は，データ k と格子の頂点 N_i との距離を表す．$M(N_i)$ は N_i の周りにデータがあるほど高い値を示す．マウンテンクラスタリングでは，M 値が大きい格子の頂点をクラスタの中心とする．最も大きい M 値は，以下の式で選択される．

$$M_1^* = \mathrm{Max}_i \left[M\left(N_i\right) \right] \tag{3}$$

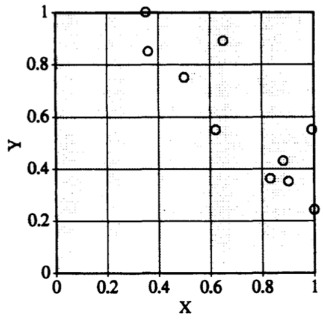

図 1　実験データ

ここで，k 番目に大きな M 値は以下の式で表される．

$$\widehat{M}^k(N_i) = \widehat{M}^{k-1}(N_i) - M^*_{k-1} \sum_{k=1}^{n} e^{\beta d\left(N^*_{k-1}, N_i\right)} \tag{4}$$

子の頂点を表す．なお，β はパラメータである．一度選択された格子の頂点の影響を考慮し，第二項で k − 1 番目に大きいと選択された格子の頂点と N_i の M 値を引く．例えば，2 番目に大きな M 値を持つ格子の頂点は，一番大きな格子の頂点の影響を取り除いて選択される．大きな M 値を持つ格子の頂点の周りも大きな M 値を持っており，各クラスタ中心の推定に影響があるため，前回選択された格子の頂点の影響を取り除く．

4　実験デザイン・結果と考察

実験では,2 次元で表される 10 個のデータに対して mountain クラスタリングを用いた図 (4)．格子の数は各次元でそれぞれ 6 個に分けられる．パラメータは，$\alpha = \beta = 5.4$ とした．クラスタ中心の選択は，最も大きな M 値を持つ格子の頂点と k − 1 番目に大きな M 値の比率を計算し，一定の値以下になるまで行なった．実験の結果，良好なクラスタリングを行うことができた．

5　手法の限界・今後の課題

この手法は，高次元なデータに対して分類を行うことが困難であることが述べられている．高次元になれば，格子点を求めることが困難になるからで，計算量も激増する．