論文メモ

文献番号	0007
日付	2021 年 10 月 19 日
名前	武川海斗

文献情報

著者	Rajesh N. Dave
英文タイトル	Characterization and detection of noise in clustering
和文タイトル	クラスタリングにおけるノイズの特性と検出
書誌情報	Pattern Recognition Letters，Vol. 12，pp. 657–664，1991
キーワード	Clustering, noise cluster, classification amon t noisy data, K-means algorithms, fuzzy K-means algorithms

1　論文のトピック

本論文では，ノイズクラスタリングの新規手法について提案を行う．

2　ベースとなった手法

2.1　Fuzzy K-means

Fuzzy K-means については，目的関数と制約条件を載せるだけの紹介とする．以下の式 (1)(2) はそれぞれ目的関数と制約条件である．ここで，データ集合を $\boldsymbol{X} = \{x_k \mid k = 1, \cdots n\}$，クラスタ数を c，帰属度を $\boldsymbol{U} = \{u_{ki} \mid 0 < u_{ki} < 1\}$，クラスタ中心の集合を $\boldsymbol{v} = \{v_i \mid i = 1, \cdots c\}$ とする．ここで二乗誤差は，$(d_{ik})^2 = (x_k - v_i)^T A_i (x_k - v_i)$ と表され，$\boldsymbol{A_i}$ は，正定値行列である．

$$J(\boldsymbol{U}, \boldsymbol{v}) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (d_{ik})^2 \tag{1}$$

$$\sum_{i=1}^{c} (u_{ik}) = 1 \tag{2}$$

しかし，既存の K-means には，ノイズ値の影響を受けやすいという欠点がある．これは，式 (2) の制約条件が原因で，ノイズ値も無理矢理にクラスタに割り当てられるからである．

3　提案手法のコア要素

3.1　ノイズクラスタリング

ノイズクラスタ距離 σ を導入することで，K-means のノイズ値の影響を受ける問題点を解決した．

3.2　σ の初期値の求め方

本節は，自身の研究につながる最も重要な節である．ノイズクラスタ距離 σ は，自身で決定する必要のあるハイパーパラメータである．適切な σ を求めるには，クラスタの分布を知る必要があり，ノイズクラスタリングの趣旨と異なる．

そこで，適切な σ を求める方法として，平均点間距離から求める方法がある．平均点間距離はクラスタの構造を反映するためである．

$$\delta^2 = \lambda \left[\frac{\sum_{i=1}^{c-1} \sum_{k=1}^{n} (d_{ik})^2}{n(c-1)} \right] \tag{3}$$

4　実験デザイン・結果と考察

人工データを基に，ノイズクラスタリングの検証実験を行った．ノイズ点がかなり多いデータで，k-means との比較で性能を確かめた (データの詳細については載っていない)．細長いデータに対しては適切にクラスタリングをされなかったので，Gustafson and Kessel (1979) による適応クラスタリングを基にクラスタリングを行った．その結果，適切なクラスタリングを行えている．

5　手法の限界・今後の課題

本論文では，k-means を基にノイズクラスタリングを行ったが，回帰分析にも応用可能である．それは，ノイズクラスタリングの概念はあらゆる二乗誤差に適応可能であり，残差の二乗に対応できるからである．画像のエッジ処理などへの応用が考えられる．

δ のパラメータ値決定についても問題が残っている．本論文では，平均点間距離を基に初期値を決定した．これは，クラスタの分布が球状か楕円状かなどでうまくいくかが決まる．一つの可能性として，偏差を考慮することで，適切な δ を決定できるかもしれないと示唆している．

6　特に重要な関連研究

ノイズクラスタリングの関連研究として以下に二つ挙げる．両者とも，手順の異なるノイズクラスタリングを行っており，私の研究に役立つものだと考える．

論文 [1] は，各データ点に密度に比例した重みを加えることにより，ノイズ点を検出する手法である．つまり，重みが低いデータ点は相対的に重要度の低いと見なすことができ，重みの値によってノイズ点を検出することができる．

論文 [2] は，最尤法を利用して，ランダムノイズから目的のデータを分類する手法である．この手法はノイズが多いデータの中から少数の目的データを抽出する場合に適している．

次に読むべき論文のリスト

[1] J. Jolion and A. Rosenfeld, Cluster detection in background noise, Pattern Recognition, Vol. 22, No. 5, pp 603–pp 607, 1988

[2] I. Weiss ,Straight line fitting in a noisy image, Computer Vision and Pattern Recognition, pp. 647–pp 652, 1988