2T-04

ホタルのアルゴリズムによるデータクラスタリング

高松　志帆†　　水野　一徳†　　西原　清一‡

†拓殖大学工学部情報工学科　　‡筑波大学大学院コンピュータサイエンス専攻

1　はじめに

近年，ビッグデータを用いて，社会的・経済的問題を解決する試みが注目されている．また，巨大なデータ集合に対して，機械学習や自然言語処理など，人工知能の各分野においても，データを効率的に処理する方法に関する研究が盛んに行なわれている [1].

本研究では，データ解析手法の 1 つであるクラスタリングに注目する．代表的な方法として *k*-means 法が挙げられるが，初期値に依存してしまうという問題点がある．本報告では，群知能モデルの Firefly Algorithm (FA) [2, 3] と *k*-means 法を組み合わせたクラスタリング手法を提案する．

2　研究分野の概要

2.1　Firefly Algorithm

FA[2] は，ホタルの点滅光に応じて移動する現象にヒントを得た群知能モデルの 1 つであり，最適化問題等で広く利用されている．その主な特徴を以下に示す.

- 各ホタルの位置は，状態空間中の各状態に相当し，その状態の評価値（適応度）は，光の強度（魅力）に比例する．
- 光の強度はホタル間の距離の大きさに応じて減少する．
- 各ホタルの移動方向と移動量は，近隣のホタルが発する明るい光に魅了され引きつけられるように確率的に決定される．
- 魅力のあるホタルが近隣に存在しない場合は，ランダムに飛び回る（移動する）.

2.2　クラスタリング

クラスタリングは，データ集合をある共通の特徴をもつような部分集合（クラスタ）に切り分けるものであり，各データに適切なクラスタが割り当てられた組合せを見つけることに相当する．代表的な手法である *k*-means 法は，下記の手順により，各クラスタにおける代表点（中心）から各データへの距離を最小化するように，各データを適切なクラスタへの再割当てを繰り返すものである．

手順 1: 各データ x_i $(i = 1 \ldots N)$ をランダムにクラスタ C_k $(k = 1 \ldots K)$ に割り振る

手順 2: 割り振ったデータをもとに各クラスタの代表点 x_k を計算し，各 x_i と各 x_k との距離を求めて，x_i を最も近い代表点をもつクラスタに割り当て直す

手順 3: もし，割当てが変化しないなら終了．そうでなければ手順 2 へ戻る

k-means 法は，山登り法のような局所的な探索を進めていくため高速ではあるが，初期値によっては局所最適解に陥ってしまうという問題点がある．

3　提案する手法

3.1　基本方針

本研究では，大規模なデータに対するクラスタリングを達成するために，FA と *k*-means 法を組み合わせた方法を提案する．その基本方針は以下の 2 つである.

i) FA における各ホタルは，すべてのデータに何らかのクラスタが割り当てられた組合せを 1 つの状態（位置）として保持する.

ii) FA による探索を行なった結果の最良解を初期値として与えて *k*-means 法を実行することによりクラスタリングを収束させる.

3.2　アルゴリズム

本手法のアルゴリズムを図 1 に示す．本手法は，まず FA による探索を実行し，その最良解を *k*-means 法の初期値として与えてクラスタリングを行なう.

3.3　距離の定義（適応度関数）

本研究では，入力したデータを多次元空間の点（次元数はデータのもつ属性の数）と捉え，式 (1) のように，各データ x_i とクラスタ C_k の代表点 x_k とのユーク

Data Clustering by Firefly Algorithms
Shiho Takamatsu†　Kazunori Mizuno†　Seiichi Nishihara‡
†Department of Computer Science, Faculty of Engineering, Takushoku University
‡Department of Computer Science, University of Tsukuba

```
t = 0, s* = φ, γ = 1.0;    //世代カウンタ，最良解，
                             誘引度の初期化
p^(0) = InitializeFA();    //ホタルの初期集団の生成

while (t < T) do
   α^(t) = AlphaNew();
   Evaluate(p^(t), f(s));    //解の評価
   OrderFA(p^(t), f(s));    //適応度順に解候補をソート
   s* = FindTheBestFA(p^(t), f(s));    最良解の保持
   p^(t+1) = MoveFA(p^(t));    // ホタルの位置の更新

   t = t + 1;
end while

ApplyKmeans(s*);    //最良解 s* を初期値として
                      k-means 法を実行
```

図 1: 本手法のアルゴリズム

リッド距離の総和を各ホタル s の適応度と定義し，これを最小化するように探索を進める．

$$f(s) = \sum_{k=1}^{K} \sum_{x_i \in C_k} D(x_i, x_k) \tag{1}$$

$$D(x_i, x_k) = \sqrt{\sum_{j=0}^{N-1} (x_{ij} - x_{kj})^2} \tag{2}$$

$$x_{kj} = \sum_{x_i \in C_k} \frac{x_{ij}}{|C_k|} \tag{3}$$

ただし，x_{ij} は，データ x_i の第 j 成分を表わしている．

4 評価実験

本手法の性能を評価するために実験を試みた．ここでは，UCI Machine Learning Repository の代表的なベンチマークデータセットである Iris data set，Glass data set，Cancer-Int data set を用いた．これらのデータセットの詳細を表 1 に示す．本手法の FA のパラメータは，ホタルの数（N_f）= 20, 世代数（T) = 100, ランダム化係数（α）= 0.5, 誘引度（β）= 1.0, 光吸収係数（γ）= 1.0 とした．

表 2 に実験結果を示す．表 2 は，各ベンチマークデータに対して，k-means 法のみおよび本手法について，それぞれ 100 回適用した場合の平均値を表わしている．また，表 2 の正答率は，入力したすべてのデータに対して，適切なクラスタに割り当てることができたデータ数の割合を示している．表 2 より，Iris データについてはランダムな初期値の方が正答率が高いものの，よりデータ数や属性数が多い Glass や Cancer-Int では，FA による結果を初期値として与えた方が適切に分けられているということが分かる．

表 1: テストに使用したベンチマークデータ

	データ数	属性数	クラスタ数
Iris	150	4	3
Glass	214	9	6
Cancer-Int	699	9	2

表 2: 実験結果

(a) k-means 法の実験結果

	Iris	Glass	Cancer-Int
正答率 (%)	85.1	21.7	50.98
time(*sec*.)	0.069	0.024	0.058

(b) 本手法の実験結果

	Iris	Glass	Cancer-Int
正答率 (%)	83.7	22.2	51.8
time(*sec*.)	0.25	0.3	0.72

5 おわりに

本報告では，データ解析手法の 1 つであるクラスタリングに対して，FA と k-means 法を組み合わせた方法を提案した．本手法は，まず FA による探索を行ない，その結果を k-means 法の初期値として利用するものである．また，実験により FA による初期値生成が有効である可能性を示した．今後は，より多様なデータおよびより大規模なデータによる実験を行ない，本手法のさらなる性能向上のための改良を行なう予定である．

参考文献

[1] 神嶌敏弘: データマイニング分野のクラスタリング手法 (1), 人工知能学会誌, Vol. 18, pp. 59–65 (2003).

[2] Fister, I., Fister Jr., I., Yang, X., and Brest, J.: A comprehensive review of firefly algorithms, *Swarm and Evolutionary Computation*, Vol. 13, pp. 34–46 (2013).

[3] Senthilnath, J., Omkar, S. N., and Mani, V.: Clustering using firefly algorithm: Performance study, *Swarm and Evolutionary Computation*, Vol. 1, pp. 164–171 (2011).