15-2

Parameter Setting of K-Means Clustering used Firefly Algorithm Modified Random Component

Masaki TAKEUCHI[1], Thomas Ott[2], Haruna MATSUSHITA[3], Yoko UWATE[1] and Yoshifumi NISHIO[1]
(Tokushima University[1], Zurich University of Applied Sciences[2], Kagawa University[3])

1. Introduction

A clustering is a popular data analysis technique and is used for data mining, image analysis, etc. The goal of clustering is to find cluster, which is homogeneous groups of objects in data set. One of the most famous clustering methods is K-means algorithm. This algorithm finds K cluster centers and each object is assigned to the closest cluster center. This means the K-means clustering is an optimization problem. It is well-known that the performance of this algorithm depend on initial conditions.

In 2011, an algorithm that used Firefly Algorithm for K-means clustering (KMFA) was proposed. Firefly Algorithm is one of the Swarm Intelligence algorithm. In this study, we introduce a new clustering algorithm that combined K-means clustering and improved Firefly Algorithm (KMIFA). Numerical experiment indicates KMIFA is more efficient than K-means algorithm and KMFA.

2. Proposed Method

In KMFA, each firefly contains the positions of all cluster centers. The brightness of each firefly is determined by the sum of the squared distances between the objects and the corresponding cluster centers. The less brighter firefly i moves towards the brighter firefly j and is defined by

$$\boldsymbol{x}_i^{new} = \boldsymbol{x}_i + \beta(\boldsymbol{x}_j - \boldsymbol{x}_i) + \alpha\boldsymbol{\epsilon}, \tag{1}$$

$$\alpha(t) = \alpha_0 \left(\frac{10^{-4}}{0.9}\right)^{t/t_{max}}, \tag{2}$$

where $\boldsymbol{x}_i$ is the position vector of firefly i, β is the attractiveness: attractiveness stared above decrease as their distance increases, ϵ is a uniform random number and t is the number of iteration.

In KMIFA, each firefly has its own value of $\alpha(t)$:

$$\alpha(t)_i = \lambda_i^t \left(\frac{10^{-4}}{0.9}\right)^{t/t_{max}}. \tag{3}$$

We set all initial values of λ to λ_0 when initializing the population of fireflies and define the minimum value of λ is 0. The new parameter λ_i changes whether the assignment of firefly i change or not:

$$\lambda_i^{t+1} = \begin{cases} \lambda_i^t - \theta \text{ (the assignment changes)} \\ \lambda_i^t + \theta \text{ (otherwise)} \end{cases} \tag{4}$$

where θ is predefined parameter.

In the case of $\lambda >> 0$, a firefly moves with a relatively strong random influence. This makes the firefly easier to escape from a local optimum. In the case of $\lambda = 0$, a firefly does not move randomly, which leads to a faster convergence. Therefore, the concept of our proposed method is at the beginning of the search, fireflies easily escape from local optima. Then, as the number of iteration increases, fireflies tend to converge.

Table 1: Information about data toy model we used

cluster	ideal center	object number	ball of radius
1	(50, 50, 70)	50	15
2	(20, 20, 40)	30	10
3	(20, 80, 40)	30	10
4	(80, 20, 40)	30	10
5	(80, 80, 40)	30	10
6	(50, 50, 40)	20	5

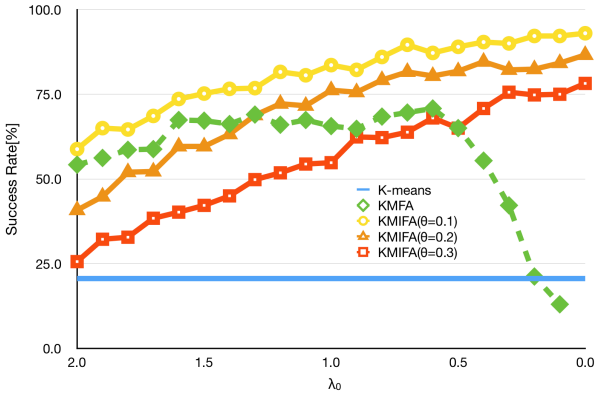

Figure 1: Comparison of each algorithm.

3. Numerical Experiment

We compare K-means algorithm, KMFA and KMIFA using data toy model that is summarized in Tab. 1. In this model, the range of each dimension is $[0, 100]$. The data objects were generated randomly around the ideal centers within each ball of radius. We used all the same data set in each numerical experiment. We set the number of fireflies is 20 and t_{max} is 100. Each experiment was run 500 times and we compared the success rate of each algorithm, which the success rate is defined as the fraction of objects that are assigned to the correct center.

Figure 1 shows the graph of the success rate of each algorithm. In the case of $\theta = 0.1$, we obtain the best results at all setting initial values. Though as the value of θ increases, the range that we can obtain the better results is gradually reduced. When the value of θ is less than 0.5, the graph of KMFA decreases, while the graph of KMIFA increases.

4. Conclusions

In this study, we have introduced a new clustering method that utilizes an improved Firefly Algorithm for K-means clustering. Our algorithm is based on the idea that the randomization parameter is changed whether the assignment changes or not. Numerical experiment has indicated our proposed method is more efficient method than the conventional method and another improved method.