Journal of Industrial Technology, Kangwon Natl. Univ., Korea, No. 44:1 (2024) 17~23
ISSN 1229-9588
eISSN 1598-1371
https://doi.org/10.22805/JIT.2024.44.1.017

전역 최적화를 위한 강건한 K-means

Robust K-means for Global Optimization

장시환[a], 이준[b], 엄재현[b], 김성수[c*]
Si-Hwan Jang[a], Joon Lee[b], Jae-Hyeon Eom[b], Sung-Soo Kim[c*]

[a] ETRI, Senior Researcher, 218, Gajeong-ro, Yuseong-gu, Daejeon, 34129, Republic of Korea
[b] Division of Energy Resource and Industrial Engineering, Kangwon National University, Student, Chuncheon, 24341, Republic of Korea
[c] Division of Energy Resource and Industrial Engineering, Kangwon National University, Professor, Chuncheon, 24341, Republic of Korea

Received 31 May 2024; Revised 11 June 2024; Accepted 18 November 2024

Abstract

K-means is a popular and efficient data clustering method which is one of the most important technique in data mining. K-means is sensitive for initialization and has the possibility to be stuck in local optimum because of hill climbing clustering method. Therefore, we need a robust K-means (RK-means) not only to reduce this possibility but also to increase the probability to search the global optimal clustering solution. The objective of this paper is to propose RK-means with best initial solution from good solutions with good central data for each cluster. The central data of each cluster is selected based on Roulette wheel probabilistic selection using sum of relative distance rate of each data. They have a problem in high density data because they deterministically select the central data for just one initial solution of K-medoid. Our proposed initial solution is the good starting point to find the robust solution by K-means with reducing the possibility being stuck in local optimal solutions. The performance of proposed RK-means data clustering is validated using machine learning repository datasets (Iris, Wine, Glass, Vowel, Cloud) comparing to original K-means by experiment and analysis. Our simulation shows that RK-means using probabilistically relative distance rate are better than K-means with random initialization. The minimum squared distance by RK-means with smaller deviation is lower than that by K-means with higher deviation. RK-means is competitive comparing to data clustering methods based on simulated annealing (SA) and hybrid K-means with SA (KSA & KSAK).

Keywords: Data Clustering, Robust K-means(RK-means), Initialization

1. 서 론

1.1 연구의 배경 및 목적

데이터 분석을 하기 위해 널리 사용되는 K-means는 초기 해를 임의적으로 선택하여 각 클러스터에 소속된 데이터들의 평균과 소속 데이터 간의 거리의 합을 평가 기준으로 해를 평가하여 최소화할 수 있는 해를 탐색한다. 모든 데이터를 각 클러스터의 평균을 기준으로 재할당하여 새로운 클러스터링 해를 탐색하고 다시 평가한다. 이러한 과정들을 더 좋은 해 결과가 나오지 않을 때까지 반복하여 해를 탐색한다. 이와 같이 K-means방법은 초기 해를 선택할 때 임의적으로 선택하기 때문에 초기 해를 잘못 선택할 경우, 지역 해에 머무를 확률이 높다. 예를 들어 Fig. 1은 X축이 특성(attribute)을 나타내는 독립 변수이고 Y축은 K-means에서 사용되는 각 클러스터에 소속된 데이터들의 평균과 소속 데이터 간의 거리(intra-cluster distance)의 합을 나타내는 종속변수이며, 이 값을 최소화하는 것이 목적이다. K-means가 초기 해를 임의적으로 4개를 선택할 경우 ①~

* Corresponding author. Tel.: +82-33-250-6283
fax: +82-504-135-6768
E-mail address: kimss@kangwon.ac.kr (Sung-Soo Kim).

④가 선택될 수 있다. ①~③의 경우 지역 해를 탐색하게 되고 4개의 초기 해 중 3개이므로 확률이 75%, 초기해 ④의 경우 전역 해를 탐색하게 되어 4개의 초기해 중 1개이므로 확률이 25%가 된다. 이와 같이 지역 해 탐색 확률이 높고 임의로 선택한 초기 해에 따라 최종 해가 달라지고 탐색 해의 편차가 크기 때문에 이러한 K-means의 문제를 개선해야 한다.

기존 연구에서도 K-means는 적용하기 쉽고 효율적이나, 초기해를 임의로 선택하기 때문에 해 탐색 결과가 초기해 선택에 따라

민감하고 탐색 해의 표준 편차가 크며 지역 해에 빠질 가능성이 높다고 지적하고 안정적인 해를 탐색할 수 있는 데이터 클러스터링 방법의 개발 필요성을 지적하였다 [Arthur et al. 2007, Celebi et al. 2013, Fränti et al. 2019, Khan et al. 2004, Likas et al. 2003, Selim et al. 1991, Xie et al. 2011]. 특히, 지역 해 탐색 확률을 감소시키고 전역 해 탐색을 위해 Selim & et al.[13]은 Simulated annealing (SA), Perim & et al.[11]은 K-means와 SA 혼합 방법, Kim & et al. [7]은 K-means와 SA를 혼합 한 KSAK 혼합한 방법을 제안하였다.

따라서, K-means의 문제를 극복하고 성능을 개선하기 위해 초기 해를 임의적으로 선택하지 않고 데이터 간의 거리 비율을 확률적으로 고려하여 각 클러스터의 중심 데이터를 결정한다. 이렇게 선택된 중심 데이터를 기준으로 좋은 해들을 생성하고 그 해들 중 가장 좋은 해를 초기 해로 사용하는 강건한 K-means (Robust K-means, RK-means)를 제안하는 것이 본 논문의 목적이다. RK-means는 기존 K-means보다 탐색한 최종 해들의 표준편차가 작아 안정적인 해 탐색이 가능하고 지역 해에 빠질 확률이 낮아져 전역 해 탐색이 가능하다.

2절에서는 데이터 클러스터링 문제의 수학적 모델을 설명하였다. 3절에서는 거리 비율을 적용한 데이터 클러스터링 초기해 선택을 적용하여 전역 최적화가 가능한 강건한 RK-means 방법을 설명하였다. 4절에서는 제안하는 RK-means의 성능을 검증하기 위해 Machine Learning Repository 데이터(Iris, Wine, Glass, Vowel, Cloud)를 사용하여 K-means, SA, K-means와 SA가 혼합된 방법(KSA와 KSAK)과 비교 실험 분석하였다.

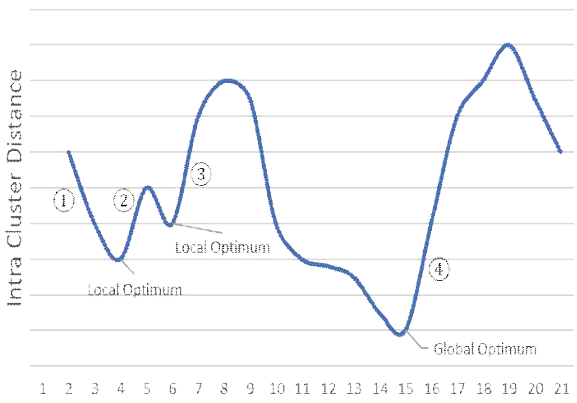

Fig. 1 Initial Solution Selection for K-means

2. 데이터 클러스터링 문제

데이터 클러스터링 문제 (즉, n개의 데이터를 k개의 그룹으로 클러스터링하는 문제)를 수리적모델로 정립화 하였다[8]. 데이터 집합 $X=\{x_1,x_2,\ldots,x_n\}$데이터 $i\,(x_i)$로 구성된다($i=1,2,\cdots,n$). 또한, k $(k=1,2,\cdots,K)$개의 클러스터 서브 집합 $C=\{C_1,C_2,\ldots,C_K\}$로 서로 겹치지 않는 클러스터로 구성된다. 각 클러스터 집합은 적어도 한 개의 데이터가 존재한다. 데이터 클러스터링 해를 평가할 때 평가기준은 일반적으로 클러스터 내의 거리와 클러스터 간의 거리를 고려할 수 있다. 위 두 요소를 모두 고려할 수 있는 실루엣(Silhouette) 평가 기준은 데이터양이 많아질수록 계산시간이 많이 소요되어 효율성이 떨어진다[4]. 따라서, 본 연구에서는 클러스터링 해의 평가 기준인 클러스터 내의 거리만을 적용하였다. 즉, 각 클러스터 내에서 평균과 소속된 데이터 사이의 거리의 총합 식(1)을 목적식으로 사용하여 최소화하고 제한식 (2)~(6)을 적용하였다.

$$Minimize \sum_{k=1}^{K} S(k) \tag{1}$$

$$s.t\, w_{ik} = \begin{cases} 1 & \text{데이터 } i\text{가 클러스터 } k\text{에 포함된 경우} \\ 0 & \text{그렇지 않을 경우} \end{cases} \tag{2}$$

$$\sum_{k=1}^{K} w_{ik} = 1,\ \ i=1,2,\cdots,n \qquad w_{ik} \in 0,1 \tag{3}$$

$$\sum_{i=1}^{n} w_{ik} \geq 1,\ \ k=1,2,\cdots,K \tag{4}$$

$$C_{kj} = \frac{\sum_{i=1}^{n} w_{ik} x_{ij}}{\sum_{i=1}^{n} w_{ik}} \tag{5}$$

$$S(k) = \sum_{i=1}^{n} w_{ik} \sqrt{\sum_{j=1}^{d} (x_{ij} - c_{kj})^2} \tag{6}$$

만약 데이터 i (x_i)가 클러스터에 포함되었을 경우, 의사 결정 변수 w_{ik}를 1로 표시하고 그렇지 않으면 0으로 표시하여 식(2)와 같이 정의할 수 있다. 데이터 클러스터링 해 표현 매트릭스를 $W=\{w_{ik}\}$로 나타낼 수 있고 x_i가 하나의 클러스터 k에 포함되는 여부에 따라 식(3)과 같이 표현할 수 있다. 식(4)는 클러스터 k에 적어도 하나 이상의 데이터 $i(x_i)$가 포함된 것을 표현하였다. 식(5)의 C_{kj}는 $C_k=(C_{k1}, C_{k2}, \cdots, C_{kd})$의 클러스터 k에서 특징(feature) 데이터 j의 평균값을 나타낸다. 식(6)은 클러스터 k의 데이터 x_{ij}와 특징 데이터 j의 평균값과의 거리의 합을 나타낸 것이다.

3. 강건한 RK-means

본 절에서는 데이터 클러스터링을 할 때 어떻게 좋은 초기해를 선택하여 K-means에 적용할 수 있는지를 서술하였다. 즉, 데이터의 상대적인 거리 비율의 합 (V_j)을 이용하여 어떻게 좋은 초기해를 선택하고 K-means의 문제점(초기해 선택에 따라 해 탐색이 민감하고 지역 해에 빠질 가능성이 높아 안정적인 해 탐색이 어렵다는 단점)을 해결할 수 있는지 서술하였다. 데이터i에서 데이터 j까지 거리의 상대적인 비율의 합을 계산한다. 이 값을 기준으로 확률적 선택 방법을 적용하여 클러스터 수만큼 클러스터 중심 데이터를 선택하여 해들을 생성한다. 이 해들 중 가장 좋은 해를 초기해로 결정한다. 3.1절에서는 확률적 거리 비율을 적용한 초기해 선택 방법을 설명한다. 3.2절에서는 초기 해 선택 후 K-means방법을 적용하여 안정적인 해 탐색을 할 수 있는 강건한 K-means 방법을 제안한다.

3.1 확률적 거리 비율을 적용한 초기해 선택

본 절에서는 K-means를 수행하기 전에 모든 데이터에서 데이터 j까지 거리의 상대적인 비율의 합을 계산하고 확률적인 중심 데이터 선택으로 해를 생성하여 초기 해를 선택한다.

각각의 데이터 $i(x_i)$는 p차원(특징, attribute)으로 구성되는데 x_{ip}는 데이터 i의 특징 데이터 p 의 값을 표현한다. 데이터 i에서 데이터 j까지의 거리(d_{ij})를 식(7)로 계산하고, 식(8)을 사용하여 데이터 j를 기준으로 하여 모든 데이터의 상대적인 거리 비율의 합 V_j를 계산한다. V_j가 작을수록 해당 데이터 j가 클러스터링을 할 때 중심 데이터 역할을 할 경우 다른 데이터와의 거리가 상대적으로 짧아서 유리하다 [6, 9]. Park & Jun (2009) [9]은 식(7)~(8)로 K-medoid의 초기값을 선택하는 것을 제안하였다. 이들이 제안 방법에서 식(8)의 확정적인 값을 사용하여 K-medoid의 중심 데이터를 결정할 때, 분석 데이터에 밀도가 높은 영역이 존재하면 중심데이터가 이 영역에 몰려 선택될 가능성이 높다. 이 문제를 해결하기 위해서는 식(8)을 고려하여 각 클러스터의 중심데이터를 확정적으로 선택하지 않고 확률적으로 선택하여 여러 가능해를 생성하고 이들 중에서 가장 좋은 해를 초기해로 사용한다.

$$d_{ij} = \sqrt{\sum_{a=1}^{p}(x_{ia}-x_{ja})^2} \tag{7}$$

$$V_j = \sum_{i=1}^{n}\frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \tag{8}$$

모든 데이터에 대한 거리 비율의 합(V_j), 식(8)을 기준으로 룰렛 휠 선택(Roulette wheel selection) 방법을 적용하여 클러스터 수만큼 클러스터 중심 데이터를 선택하여 해를 생성한다. 즉, 데이터j의 V_j가 작을수록 (역수 값$(1/V_j)$) 데이터가 중심 데이터로 선택될 확률이 높게 하여 클러스터 수만큼 선택하고 해를 생성한다. Fig. 2는 확률적 거리 비율의 합(V_j)을 사용한 초기해 선택 과정이고 각 단계별 과정은 다음과 같다.

단계 1. 확률적으로 몇 개(N)의 다양한 해를 생성할 것인지 결정한다.

단계 2. 모든 데이터 간의 거리 d_{ij}를 식(7)로 계산하고 저장한다.

단계 3. 모든 데이터 j의 V_j를 식(8)로 계산한다.

단계 4. 모든 V_j의 역수 값 $(1/V_j)$에 대하여 룰렛 휠 선택 방식을 적용한다.

단계 5. 각 데이터 j의 선택확률 $P_j=\frac{1/V_j}{\sum_{j=1}^{n}1/V_j}$ 을 계산한다.

단계 6. 각 데이터 별로 랜덤 확률 값을 할당해 선택 확률 구간에 포함되는지 확인하여 K개의 중심 데이터 선택한다.

단계 7. K개의 중심 데이터를 기준으로 나머지 $(n-K)$개의 데이터가 어느 중심점에 가까운지 계산하여 해당 클러스터 소속으로 재구성하여 해를 생성한다.

단계 8. 생성된 해의 평가값(Valid index: intra-cluster distance) 값을 계산한다.

단계 9. N회만큼 실시하였다면 단계 10으로 넘어가고 그렇지 않으면 단계6부터 다시 실행한다.

단계 10. 생성된 해들 중 평가값이 가장 좋은 해를 초기 해로 선택한다.

본 절에서 제안한 초기해 선택 방법을 적용할 경우, 임의의 초기해를 선택하는 K-means의 문제점을 상당히 감소시킬 수 있다. 예를 들어, Fig. 1에서 초기 해를 선택할 경우, 초기해 ①~③의 선택 확률은 낮아져 지역 해 탐색 확률은 낮아지고, 초기해 ④의 선택 확률은 높아져 전역해 탐색 확률은 높아진다.

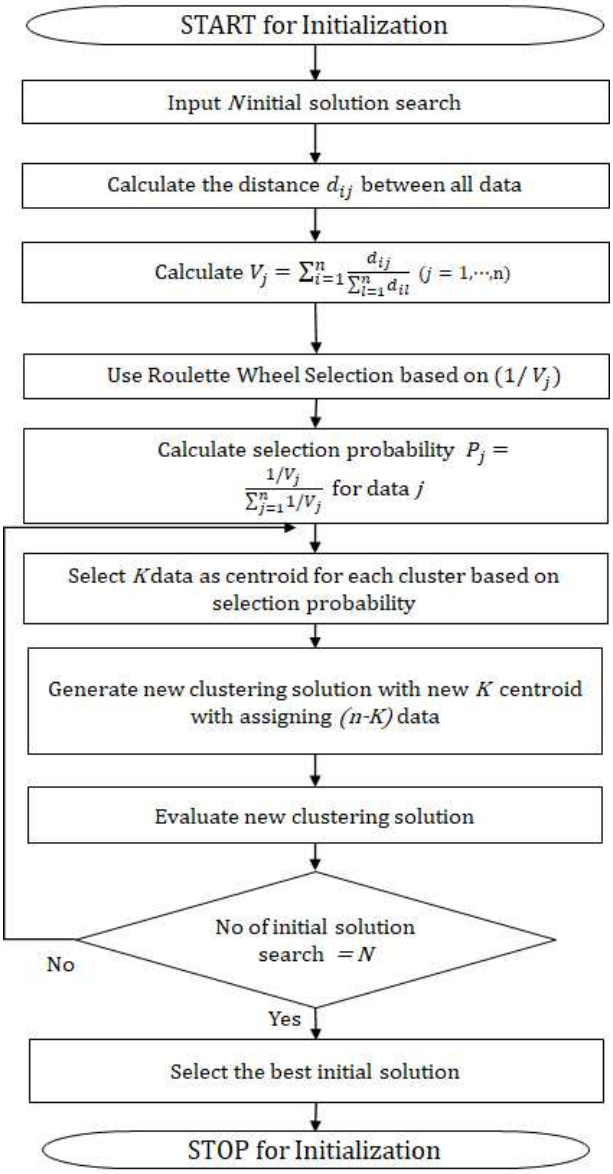

Fig. 2 Selection of Good Initial Solution

3.2 전역해 탐색 확률이 높은 RK-means

초기해에 민감하고 지역해를 탐색할 가능성이 높다는 K-means의 문제점을 해결하기 위해 3.1절의 초기해 선택 방법을 적용하여 안정적인 해 탐색을 위한 초기해를 생성한다. 이 초기해를 K-means에 적용하면 임의로 초기해를 선택하는 일반적인 K-means방법보다 지역해 탐색 확률을 감소시켜 안정적인 해 탐색이 가능하여 강건해진 RK-means를 만들어 낼 수 있다. Fig. 3은 3.1 절에서 제안하는 확률적 초기해 선택 방법을 적용하여 강건성 (robustness)이 높은 RK-means 과정이다.

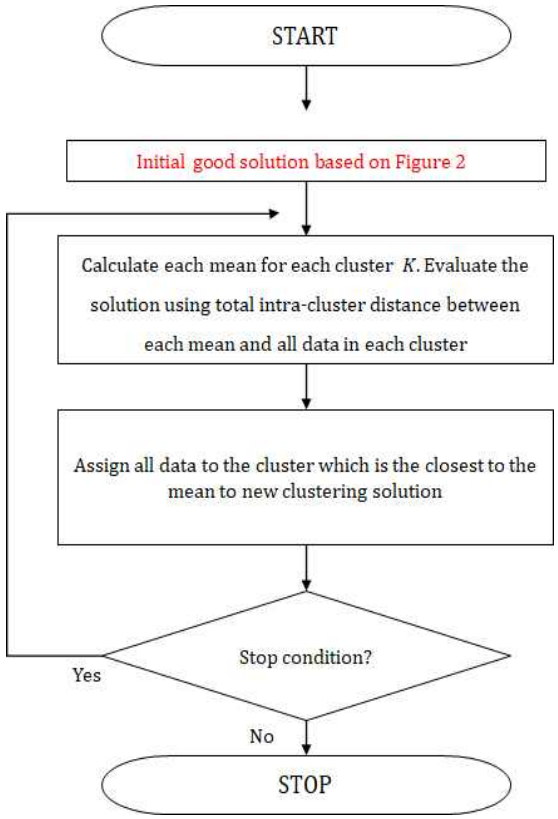

Fig. 3 RK-means with Initial Good Solution

4. 실험결과 및 분석

본 절에서의 실험은 윈도우10 Pro 프로세서: Intel(R) CoreTM i7-8700K CPU @ 3.70GHz 3.70 GHz 메모리(RAM): 16GB, 64비트 운영 체제, x64 기반 프로세서 운영체제, Visual C++ 환경에서 실험하였다. 본 논문에서 제안하는 RK-means(제안한 방법으로 선택한 초기해+K-means) 데이터 클러스터링 방법의 성능을 검증하기 위해서 UCI machine learning repository [11]의 Iris, Wine, Glass, Vowel, Cloud, 데이터를 각 20회 실험하였다.

Table 1은 실험 분석을 위한 각 데이터의 클러스터 수, 특징 수, 데이터 수를 표로 정리한 것이다.

Table 1 UCI Data for Experiment[11]

Name of dataset	No. of classes	No. of features	No. of data
Iris	3	4	150
Wine	3	13	178
Glass	6	9	214
Vowel	6	3	871
Cloud	10	10	1024

Fig. 4 ~ Fig. 8은 Table 1의 5개의 실험 데이터에 대하여 K-means와 RK-means의 수렴 그래프를 비교한 것이다. 수렴 그래프의 한 점은 20회 실행한 값의 평균을 의미한다. 예를 들어, Iris 97.6242, Wine 16642.7715, Glass 230.6866, Vowel 156482.5, Cloud 71922.15 등에서 시작한다.

RK-means 적용 시 Table 2에서 Iris 97.2783, Wine 16551.3435, Glass 218.64, Vowel 152,348.3, Cloud 64,325.55 등으로 서서히 수렴한다.

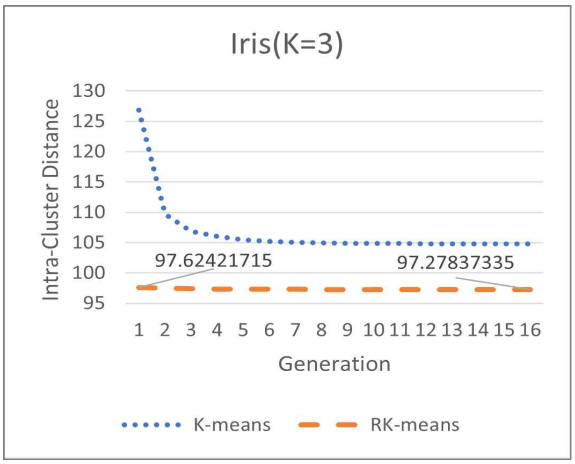

Fig. 4 Trend of Convergence for Iris

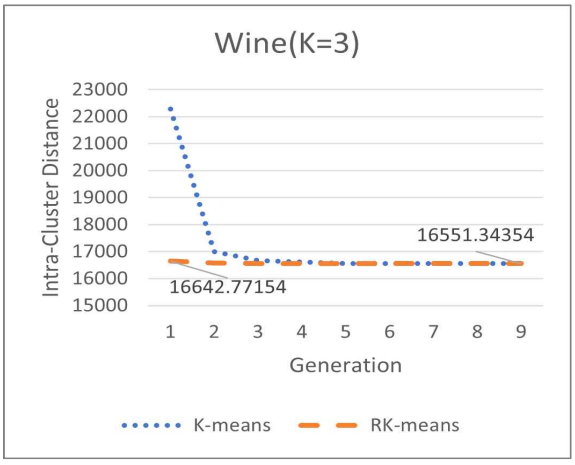

Fig. 5 Trend of Convergence for Wine

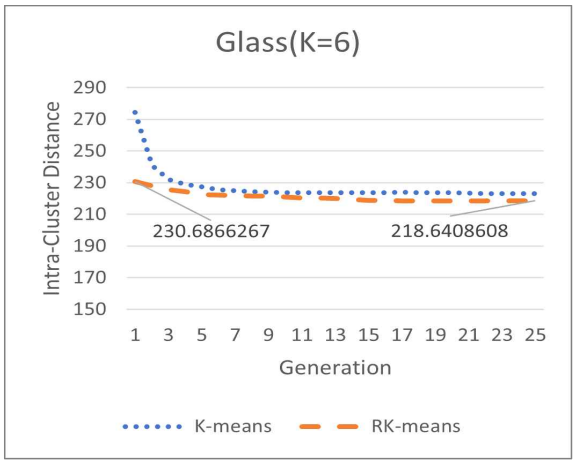

Fig. 6 Trend of Convergence for Class

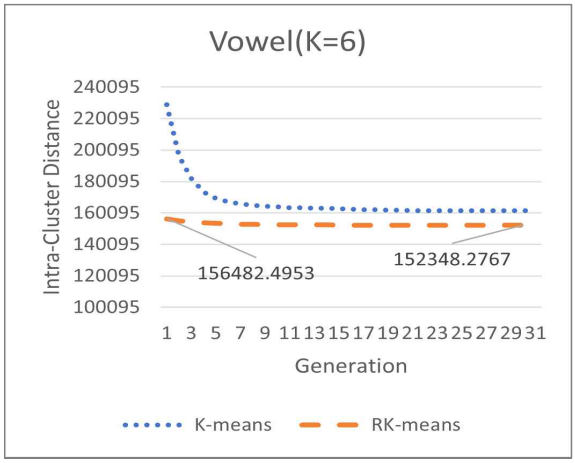

Fig. 7 Trend of Convergence for Vowel

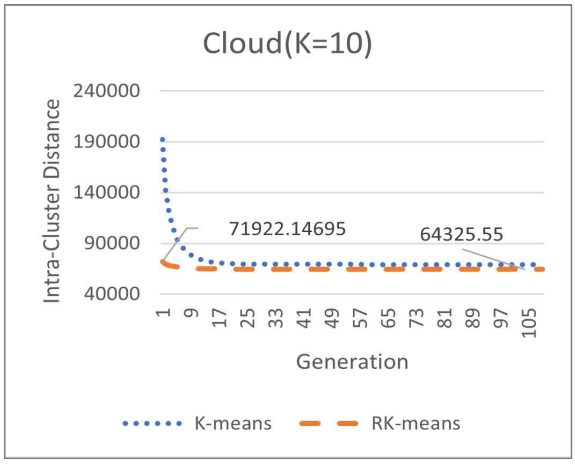

Fig. 8 Trend of Convergence for Cloud

Table 2 Comparisons of RK-means & previous methods

		K-means	SA[13]	KSA [11]	KSAK[7]	RK-means
IRIS	Mean	104.79	97.41	97.68	97.23	97.27
	S.D.	13.47	0.21	0.90	0.00	0.04
	Best	97.23	97.22	97.22	97.22	97.23
WINE	Mean	16554.42	16564.47	16530.50	16530.50	16551.34
	S.D.	5.62	151.89	0	0	6.74
	Best	16530.53	16530.50	16530.50	16530.50	16530.53
GLASS	Mean	223.21	231.31	223.15	217.86	218.64
	S.D.	8.37	14.56	2.48	1.29	2.17
	Best	215.51	221.69	214.72	214.66	215.64
VOWEL	Mean	161485.20	149685.30	150412.10	149758.80	152348.30
	S.D.	10250.05	283.40	880.17	533.72	3485.64
	Best	149902.00	149407.00	149405.00	149380.00	149384.00
CLOUD	Mean	69165.49	64638.70	63214.06	63132.68	64325.55
	S.D.	3080.98	755.63	406.59	417.61	974.18
	Best	64075.46	62889.88	62937.95	62856.85	63141.89

Table 2는 본 논문에서 제안하는 RK-means와기존 K-means, Selim & et al.[12]가 제안한 Simulated annealing (SA), Perim & et al.[11]이 제안한 K-means와 SA의 혼합 방법 (KSA), Kim & et al.[7]이 제안한 K-means와 SA에 다시 K-means를 혼합한 방법 (KSAK)을 5개의 실험 데이터 결과로 비교한 것이다. 각 방법에 대하여 20회의 실험을 통하여 탐색한 최종해의 평균, 편차, 가장 좋은 값 Best를 비교하였다.

우선, RK-means와 K-means를 비교했을 때, RK-means가 K-means보다 평균 평가 값과 표준편차가 5개의 실험 데이터에서 더 우월한 해(평가 값 개선률 Iris 7.17%, Glass 2.04%, Vowel 5.65%, Cloud 6.99%)를 더 안정적(표준편차 개선률 Iris 99.62%, Glass 71.07%, Vowel 65.99%, Cloud 68.38%) 탐색이 가능하여 RK-means의 성능 우수성이 뚜렷하게 나타났다. Iris, Glass, Vowel, Cloud 데이터의 경우 초기해를 임의로 선택하는 K-means는 최종해의 평균값과 표준편차가 상대적으로 매우 크다. 반면, RK-means는 최종해의 평균값과 표준편차가 상대적으로 매우 작아 RK-means의 성능이 우수하였다. Wine 데이터 경우, 두 방법의 평균값과 표준 편차의 차이는 크지 않은 것으로 분석되었다. 일반적인 K-means의 약점인 임의의 초기해 선택으로 인한 지역 해 탐색 위험률을 본 논문에서 제안하는 RK-means를 사용하여 상당히 개선하였다.

RK-means와 SA를 비교했을 때, 5개의 데이터 중 Vowel을 제외한 Iris, Wine, Glass는 RK-means의 평균과 표준편차가 SA의 결과보다 우수하였다. Cloud는 RK-means의 평균값이 우수하였다. RK-means방법이 5개의 데이터 중 Iris, Glass는 RK-means의 평균과 표준편차가 KSA의 결과보다 우수하였다. RK-means방법이 5개의 데이터 중 Iris, Glass는 RK-means의 평균과 표준편차가 가장 우수한 KSAK의 결과에 근접한 결과를 탐색할 수 있었다.

5. 결 론

데이터 클러스터링을 할 때 가장 보편적으로 사용되는 K-means 방법은 초기해를 임의적으로 선택하기 때문에 지역해에 빠질 가능성이 높고 민감하여 안정적인 해 탐색을 할 수 없어 탐색한 해 편차가 크다는 한계성이 있다.

본 논문에서는 이 문제를 해결하기 위해 데이터 간 상대적인 거리 비율의 합이 작은 데이터일수록 각 클러스터의 중심데이터로 선택되면 유리하기 때문에 이 중심데이터를 기준으로 초기해를 생성하였다. 즉, 이 비율의 역수 값에 따른 룰렛 휠 선택을 하여 해를 생성하고, 생성된 해 중 가장 좋은 해를 초기해로 사용하는 강건한 K-means (RK-means) 방법을 제안하였다. 임의의 초기해를 적용한 K-means보다 본 논문에서 제안한 강건한 RK-means가 실험 데이터에서 평가값 평균이 상당히 향상되었고 표준편차는 상대적으로 매우 감소되어 안정적인 해 탐색이 가능하였다. 또한, RK-means는 기존 휴리스틱 SA 데이터 클러스터링 방법보다 우수하였고, SA를 발전시킨 KSA 및 KSAK와 비교해서도 경쟁력이 있는 방법으로 분석되었다.

6. 감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2024년도 문화체육관광 연구개발사업으로 수행되었음(과제명 : 중소 게임 기업의 게임 제작 검증 효율화를 위한 AI 기반의 대규모 게임 자동검증 기술 개발, 과제번호 : RS-2024-00393500, 기여율: 100%)

References

[1] Arthur, D., Vassilvitskii, S., 2007, k-means++: the advantages of

careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, New Orleans.

[2] Celebi, M.E, Kingravi, H.A., Vela, P.A., 2013, A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Syst. Appl. 40:1 200-210.

[3] Fränti, Pasi, Sieranoja, Sami, 2019, How much can k-means be improved by using better initialization and repeats?, Pattern Recognition 93:9 95-112.

[4] Kang, B., Kim, S., 2018, Two-Step Artificial Bee Colony Data Clustering Based on Silhouette, Management Science Review 42:2 1-9.

[5] Khan, S. S., Ahmad, A., 2004, Cluster center initialization algorithm for K-means clustering, Pattern Recognition Letters 25:11 1293-1302.

[6] Kim, S., Kang, B., 2018, Efficient Data Clustering using Fast Choice for Number of Clusters, J. Soc. Korea Ind. Syst. Eng. 41:2 1-8.

[7] Kim, S., Baek, J., Kang, B., 2017, Hybrid Simulated Annealing for Data Clustering, Journal of Society of Korea Industrial and Systems Engineering 40:2 92-98.

[8] Krishna K, Narasimha Murty M., 1999, Genetic K-means algorithm, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 29:3 433-439.

[9] Likas, A., et al., 2003, The global k-means clustering algorithm, Pattern Recognition 36:2 451-461.

[10] Park, Jun, 2009, A simple and fast algorithm for K-medoids clustering, Expert Systems with Applications 36 3336-3341.

[11] Perim, G. T., Wandekokem, E. D., Varejao, F. M., 2008, K-Means Initialization Methods for Improving Clustering by Simulated Annealing, 11th Ibero-American Conference on AI 5290 133-142.

[12] Singh, Chauhan, 2011, K-means v/s K-medoids: A Comparative Study, National Conference on Recent Trends in Engineering & Technology.

[13] Selim, S. Z., Alsultan, K., 1991, A simulated annealing algorithm for the clustering problem, Pattern Recognition 24:10 1003-1008.

[14] UMass Machine Learning Laboratory, n.d., UCI machine learning repository datasets, <http://mlr.cs.umass.edu/ml/datasets.html>.

[15] Xie, J., et al., 2011, An Efficient Global K-means Clustering Algorithm, Journal of Computers, 6:2.