J. Soc. Korea Ind. Syst. Eng Vol. 40, No. 4 : 203-210, December 2017
https://doi.org/10.11627/jkise.2017.40.4.203
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Combined Artificial Bee Colony for Data Clustering

Bum-Su Kang · Sung-Soo Kim†

Department of System & Management Engineering, Kangwon National University

융합 인공벌군집 데이터 클러스터링 방법

강범수 · 김성수†

강원대학교 시스템경영공학과

Data clustering is one of the most difficult and challenging problems and can be formally considered as a particular kind of NP-hard grouping problems. The K-means algorithm is one of the most popular and widely used clustering method because it is easy to implement and very efficient. However, it has high possibility to trap in local optimum and high variation of solutions with different initials for the large data set. Therefore, we need study efficient computational intelligence method to find the global optimal solution in data clustering problem within limited computational time. The objective of this paper is to propose a combined artificial bee colony (CABC) with K-means for initialization and finalization to find optimal solution that is effective on data clustering optimization problem. The artificial bee colony (ABC) is an algorithm motivated by the intelligent behavior exhibited by honeybees when searching for food. The performance of ABC is better than or similar to other population-based algorithms with the added advantage of employing fewer control parameters. Our proposed CABC method is able to provide near optimal solution within reasonable time to balance the converged and diversified searches. In this paper, the experiment and analysis of clustering problems demonstrate that CABC is a competitive approach comparing to previous partitioning approaches in satisfactory results with respect to solution quality. We validate the performance of CABC using Iris, Wine, Glass, Vowel, and Cloud UCI machine learning repository datasets comparing to previous studies by experiment and analysis. Our proposed KABCK (K-means+ABC+K-means) is better than ABCK (ABC+K-means), KABC (K-means+ABC), ABC, and K-means in our simulations.

Keywords : Data Clustering, Combined Artificial Bee Colony (CABC), K-means

1. 연구의 배경과 목적

컴퓨터 정보화와 사물인터넷의 보급에 따라 많은 양의 정보와 데이터가 수집되고 저장할 수 있게 되었다. 이런 결과로 수집된 빅데이터를 관리하고 분석하는 것이 기업 경쟁력의 핵심이 되었다[1]. 또한, 최근 빅데이터 기반 다양한 분야에 활용방안에 대하여 연구가 진행되고 있다[5]. 빅데이터 분석 방법 중 데이터 클러스터링은 기업의 마케팅 분석 등 여러 분야에 사용될 수 있고 사회 각 분야에서의 필요로 인해 추가적인 기술 개발의 필요성이 증대되고 있다. 한편, 기존 다양한 데이터 클러스터링 방법의 한계와 연구 배경은 다음과 같다.

데이터 클러스터링 방법은 비계층적 방법과 계층적 방법으로 나누어진다[2]. 비계층적 방법으로 K-means는 가장 유명하고 많이 사용되었으나 지역해에 빠질 가능성이 있다[16]. 이와 같이, 초기값에 따라 지역해에 빠질 가능성 있는 K-means는 최근 데이터 양이 많을 때에는 더 큰 문제가 될 수 있다. 이러한 문제점을 해결하기 위해

Received 12 October 2017; Finally Revised 12 December 2017; Accepted 13 December 2017
† Corresponding Author : kimss@kangwon.ac.kr

기존 연구에서는 K-means의 다양한 변형과 혼합 방법론이 제안되었다. 최근에도 데이터 클러스터링 문제에 휴리스틱알고리즘을 적용해야 할 필요성이 주장되었고[4] 다음과 같은 다양한 클러스터링 방법이 제안되었다.

Selim[15]은 클러스터링 문제에 시뮬레이티드 어닐링(simulated annealing, SA)을 제안하였고, Sun[18]과 Perim[13]은 K-means로 구한 해를 SA의 초기값으로 하는 클러스터링 방법을 제안하였다. Gungor[3]은 K-harmonic means 클러스터링 문제를 해결하기 위해 SA를 적용하였다. Kim[8]은 혼합 SA 데이터 클러스터링 방법인 KSAK를 제안하였다. KSAK는 K-means로 해를 구한 후 이 해를 초기값으로 한 SA로 해 탐색 후 각각의 해를 K-means로 더 좋은 해를 탐색하는 방법이다. Maulik[12]는 유전자 알고리즘(Genetic algorithm, GA) 방법을 제안하고 전역 해를 탐색할 수 있음을 검증하였다. Krishna[10]는 유전자 알고리즘의 교배과정 대신에 K-means 알고리즘을 적용한 Genetic K-means 방법을 제안하였다. Kao[7]는 K-means, Nelder-Mead와 파티클 군집최적화(particle swarm optimization, PSO)가 혼합된 데이터 클러스터링 방법을 제안하였다. Van der Merwe[25]는 K-means로 구한 해를 PSO 방법의 초기해로 사용하여 혼합 방법을 제안하였다. 또한 데이터의 평균을 중심점으로 하는 일반적인 인공벌군집(artificial bee colony, ABC) 방법[6], Deb의 규칙을 적용한 ABC 방법[28], ABC 방법의 해 생성 시 유전자알고리즘의 교배 방법을 적용하는 방법[27] 등이 제안되었다. 또한, Kumar[11]는 초기값들을 랜덤하게 선택하지 않고 K-means로 초기해 군을 선택하는 2단계 ABC 방법을 제안하였다. Tran[19]는 기존 ABC 방법에는 없는 돌연변이 기능을 추가하여 다양한 해 탐색을 추구하고 K-means로 수렴적 해탐색을 추구하는 클러스터링 방법을 제안하였다. Sithara[17]는 K-harmonic means와 ABC가 결합된 클러스터링 방법을 제안하였다. Reisi[14]은 ABC 방법을 이용하고 데이터의 특성에 가중치를 주는 클러스터링 방법을 제안하였다. 이와 같이 다양한 클러스터링 방법이 최근까지도 개발되고 있다

본 논문의 목적은 데이터 분석(클러스터링)의 성능을 최대화 하고 안정적인 전역해를 탐색할 수 있도록 K-means와 융합한 인공벌군집(Artificial Bee Colony, ABC) 클러스터링 방법을 개발하는 것이다. 즉, 수렴적 탐색 성능이 뛰어난 K-means를 사용하여 초기해를 생성하여 ABC 방법의 초기해 군으로 사용하고 다양한 탐색을 추구하여 best p개의 해를 추천하고 다시 이 해들을 K-means로 수렴시켜 해 탐색 효과가 최대화 될 수 있는 효과적인 KABCK(K-means+ABC+K-means) 조합의 융합 방법을 새롭게 제안하였다.

본 논문의 제 2장에서는 데이터 클러스터링 문제와 데이터 클러스터링 해 평가를 위한 유클리드 거리함수 식에 대하여 설명하였다. 제 3장에서는 새로운 클러스터링 방법을 제안하기 위한 인공벌군집(ABC) 방법을 소개하고 본 논문에서 제안하는 융합 ABC 방법을 구체적으로 설명하였다. 제 4장에서는 실험과 분석을 통하여 새롭게 제안하는 데이터 클러스터링 방법의 성능을 분석 검증하였다.

2. 데이터 클러스터링 문제와 해 평가

데이터 클러스터링은 데이터마이닝, 머신러닝과 패턴분류를 하기 위한 기본적인 방법인데, 데이터를 적절한 평가기준으로 그룹화하는 NP-hard 문제이고 다음과 같이 데이터 클러스터링 문제로 설명된다[4]. 데이터의 개수를 N개라 할 때 데이터 집합 $\mathrm{X}=\{x_1, x_2, \cdots, x_i, \cdots x_N\}$는 데이터 $x_i=\{x_{i1}, x_{i2}, \cdots, x_{ia}\}$로 a개의 특징(attribute)으로 구성되고 겹치지 않는 K개의 클러스터 서브 집합 $\mathrm{C}=\{\mathrm{C}_1, \mathrm{C}_2, \cdots, \mathrm{C}_\mathrm{K}\}(K<N)$로 구성된다. 각 클러스터 집합은 적어도 한 개의 데이터가 존재하고 $\mathrm{C_i}\neq\phi(\ \mathrm{C}_1\cup\mathrm{C}_2\cup\cdots\cup\mathrm{C_k}=\mathrm{X}$ 그리고 $C_i\cap C_j=\phi, i\neq j$)로 나타낼 수 있다.

클러스터링의 타당성 지표는 데이터의 속성에 따라 달라질 수 있고, 어떤 지표를 선택하는가에 따라 클러스터링의 품질에 중요한 영향을 끼칠 수 있다. 또한, 적절한 지표를 선택하지 못하면 전혀 다른 클러스터링 결과를 제시할 수도 있다. 따라서, 가장 좋은 클러스터링 결과를 얻기 위해서는 데이터의 속성에 가장 적합한 클러스터링 타당성 지표를 선택하는 것이 매우 중요하다. 본 논문에서는 ABC 방법에서 생성된 모든 데이터 클러스터링 해를 평가하기 위해 각 클러스터 데이터의 평균을 중심점으로 클러스터 내의 다른 데이터까지의 거리의 합(intra-cluster distance)을 계산하기 위해 유클리드 거리 함수식 (1)을 사용하였다.

n개의 데이터를 K개의 클러스터로 그룹핑 할 때, x_{ij}는 데이터 $i(i=1, 2, \cdots, n)$의 특징 $j(j=1, 2, \cdots, a)$를 나타낸다. 본 논문 제 4장 실험에 사용한 Iris 데이터의 경우 150개 데이터 각각이 4개의 특징으로 이루어지고 특징 a는 sepal length, sepal width, petal length, petal width로 이루어진다. 식 (1)은 각 클러스터링 해에 대하여 중심점(클러스터 k에 소속된 데이터들의 평균) C_k와 데이터 x_i까지의 거리를 계산할 때 사용된다. 식 (2)는 데이터 i와 그 데이터가 소속된 클러스터 중심점까지의 거리의 총합을 나타낸다.

$$d_i=\sqrt{\sum_{j=1}^{a}(x_{ij}-C_{kj})^2},\ i=1, 2, \cdots, n \qquad (1)$$

$$d = \sum_{i=1}^{n} d_i \quad (2)$$

3. 융합 ABC 데이터 클러스터링 방법

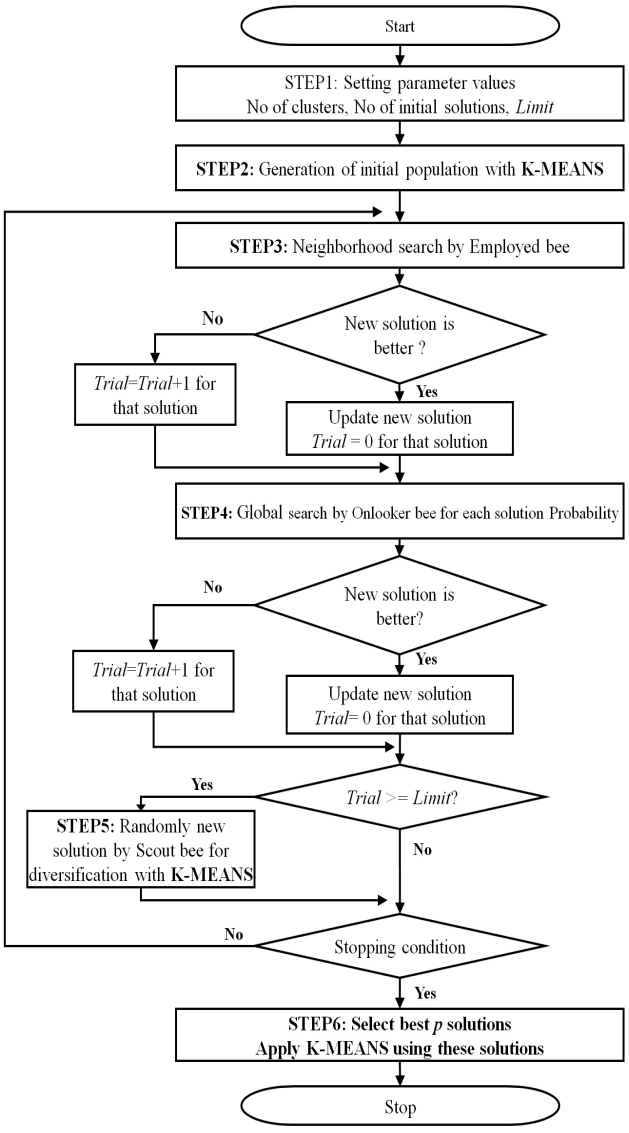
본 장에서는 새로운 데이터 클러스터링 방법을 제안하기 위해 제 3.1절에서는 인공벌군집(ABC) 방법의 일반적인 메커니즘에 대하여 설명하였고 제 3.2절에서는 ABC와 K-means를 융합하여 어떻게 데이터 클러스터링 방법을 개발하였는지 설명하였다.

3.1 인공벌군집 방법

군집 지능(swarm intelligence)은 구현이 편리하고 병행 능력과 전역해 탐색 능력이 뛰어나 유망한 클러스터링 방법으로 부상하고 있다[26]. 집단 지능 중에 인공벌군집(ABC) 방법은 벌들이 꽃을 찾아 꿀을 효과적으로 채취하는 방법을 알고리즘화 한 것으로 다른 알고리즘과 비교하여 파라미터 수가 적어 파라미터 값의 영향이 적어 안정적인 결과를 제시할 수 있다[9]. 이 방법에서는 벌을 Employed Bee(EB), Onlooker Bee(OB), Scout Bee(SB)로 역할에 따라 구분한다. EB는 현재 위치의 꽃에서 가까이 있는 꽃을 탐색하여 더 많은 꿀을 가지고 있는 꽃을 찾아내듯이 현재의 해의 이웃 해를 탐색하여 더 좋은 해를 탐색한다. OB는 각각의 EB들이 탐색하고 있는 꽃들의 위치 정보를 참조하여 찾아낸 꽃의 꿀을 채취하듯이 여러 개의 해들로부터 찾아낸 해를 확률적으로 탐색하여 더 좋은 해를 탐색해 간다. SB는 EB와 OB가 더 이상 좋은(꿀이 많은) 꽃을 탐색해 내지 못할 때 새로운 위치의 꽃을 찾아 꿀을 채취하듯이 기존의 해들과 다른 임의의 해를 탐색하는 역할을 한다[6]. 본 논문에서 제안하는 ABC 방법의 장점은 사전에 결정해야 할 파라미터의 수가 적어 상대적으로 안정적인 해 탐색이 가능하다는 것이다. 새롭게 제안하는 ABC 데이터 클러스터링 방법은 클러스터링 수 결정과 파라미터와 종료조건 설정, 초기 데이터 클러스터링 해 생성과 해 평가, EB가 데이터 클러스터링 해들의 이웃 해 탐색 후 해 평가, OB가 클러스터링 평가값에 비례하는 확률로 이웃 해 탐색 후 평가, EB와 OB가 더 이상 좋은 해를 탐색해 내지 못할 때 새로운 해를 탐색하는 SB의 해 생성을 제 3.2절에서 단계별로 설명하였다.

3.2 융합 ABC 데이터 클러스터링 방법

본 논문에서 제안하는 융합 인공벌군집(ABC) 방법은 각 클러스터의 중심점(평균)을 기준으로 하여 데이터 클러스터링 해를 평가하고 K-means와 ABC 방법이 다양한 조합으로 융합하여 전역해를 탐색할 수 있는 방법들을 비교분석 하였다. 아래 첫 번째와 두 번째 방법은 기존 방법이고 세 번째와 네 번째 방법은 본 논문에서 새롭게 제안하는 방법이다. 첫 번째는 일반적인 ABC 방법이다. 두 번째는 K-means로 초기값을 구하여 ABC 적용 시 EB의 초기해로 사용하여 수행하는 방법(K-means+ABC, KABC, [11])이다. 세 번째는 ABC를 수행 후 best p개의 해에 대하여 K-means를 적용하는 방법(ABC+K-means, ABCK)이다. 네 번째는 K-means로 해를 생성하고 이들 해를 ABC 적용 시 EB의 초기해들로 사용하고 SB로 해 생성 할 때마다 그 해에 대하여 K-means를 적용하여 해를 개선하고 ABC 종료 후 best p개의 해에 대하여 K-means를 수행하는 방법(K-means+ABC+K-means, KABCK)이다.



<Figure 1> Flowchart of CABC Data Clustering Method

본 논문에서 제안하는 KABCK는 제 3.1절 ABC의 EB, OB, SB의 역할과 연관시켜 제 3.2절 핵심 단계 3~단계 5를 설명하였다. 또한, <Figure 1>에 각 단계를 STEP으로 표시하여 설명하고 ABC와 융합된 K-means 부분을 대문자로 표시하였다.

단계 1 : [클러스터 수 K, 파라미터와 종료조건 설정]

가장 중요한 데이터 클러스터링의 수, K를 결정하고 다음과 같이 인공벌군집 ABC의 파라미터를 결정한다. 각 데이터 클러스터링 해의 이웃해를 찾는 역할을 하는 EB(Employed bee)의 수와 한 세대의 해의 수를 POP으로 표시한다. 또한, 각 해의 더 좋은 이웃해를 탐색할 때까지 시도 횟수(*Trial*)의 최대 시도 횟수(*Limit*), 종료 조건(예 : 사용자가 제시한 시간제한 등)을 결정한다.

단계 2 : [초기 해들을 POP만큼 생성하고 K-means를 실행하여 해 개선하고 해 평가]

데이터 클러스터링 해를 2차원 행렬로 표현하였고 클러스터 $k(k=1, 2, \cdots, K)$에 소속된 데이터는 1, 소속되지 않은 데이터는 0으로 표시하였다. 초기 가능 해군을 랜덤하게 생성 후 K-means를 적용함으로써 해를 초기에 개선하여 EB의 해로 사용한다. 각각의 개선된 해들은 2절의 식 (1)~식 (2)를 적용하여 해를 평가한다.

단계 3 : [EB가 해들의 이웃해 탐색 후 해 평가]

각각의 EB의 지역탐색을 위하여 이웃해를 탐색하고 2절의 식 (1)~식 (2)를 적용하여 해를 평가한다. 평가값이 더 좋은 해가 생성되면 더 좋은 해로 업데이트 하고 이웃해 탐색 시도 횟수 카운터(*Trial*)를 0으로 설정한다. 그렇지 않을 경우(평가값이 더 좋은 해를 탐색하지 못했을 경우) 해당 해의 탐색 시도 횟수(*Trial*)를 증가시킨다.

단계 4 : [OB가 평가값에 비례하는 확률로 이웃해 탐색 후 해 평가]

단계 3의 EB에 의한 모든 해의 이웃해 탐색 후 갱신된 해 군의 모든 해들의 평가값을 계산한다. 식 (2)의 d를 사용하여 해 i의 평가값 $f(i)=1/d$에 비례하는 확률 P_i 식 (3)으로 평가값이 좋은 해를 확률값으로 선택한다. 이 선택된 해의 이웃 해를 생성하고 평가값을 비교하여 더 좋은 해가 생성되면 더 좋은 해로 업데이트 하고 이웃해 탐색 시도 횟수 카운터(*Trial*)를 0으로 설정한다. 그렇지 않을 경우(평가값이 더 좋은 해를 탐색하지 못했을 경우) *Trial*을 증가 시킨다. 이와 같은 전역 해 탐색은 OB (Onlooker bee)가 수행한다.

$$P_i = \frac{f(i)}{\sum_{n=1}^{POP} f(n)} \tag{3}$$

단계 5 : [*Trial*이 *Limit* 이상일 경우 SB가 새로운 해 생성하고 K-means로 해 개선]

해 탐색 시도 횟수 카운터(*Trial*)가 탐색 최대 수(*Limit*) 보다 같거나 클 경우 해당 해는 더 이상 탐색하지 않고 그에 해당하는 데이터 해를 탈락 시킨다. 탈락시킨 해 수만큼의 해들을 SB(Scout bee)를 통하여 임의적으로 해 생성 하고 다시 K-means를 적용하여 해를 추가적으로 개선한다. 이런 과정을 통하여 지역 해에 빠지지 않고 다양한 해 탐색을 추구함으로써 전역 해를 탐색할 수 있도록 한다. 단계 5 수행 후 ABC 종료 조건에 맞으면 단계 6을 수행하고 그렇지 않으면 단계 3의 EB의 이웃해 생성부터 다시 반복 수행한다.

단계 6 : [ABC 클러스터링을 마친 해들 중 베스트 p개의 해에 대하여 K-means 실행]

인공벌군집 ABC 수행 종료 후 가장 좋은 해 p개의 해 각각에 대하여 K-means를 수행하여 해를 개선하고 좋은 해들을 업데이트한다.

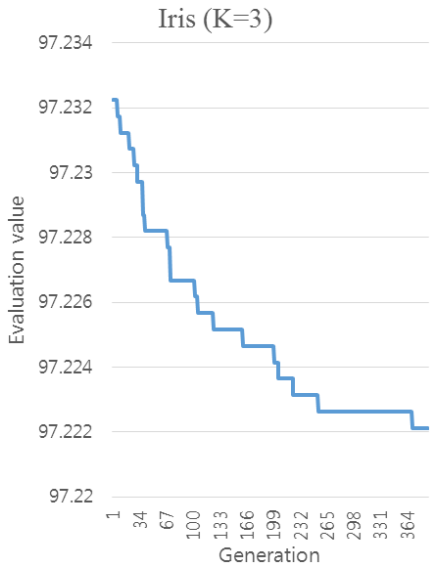
4. 실험 및 분석

본 논문에서 제안한 융합 ABC 데이터 클러스터링 방법의 성능을 검증하기 위해서 윈도우10 프로세서 : Intel(R) Core Tm i5-4590 CPU @ 3.30GHz 3.30GHz 메모리(RAM) : 4GB, C++ 환경에서 실험하였다. Kim[8] 논문에서 사용한 Iris, Wine, Glass, Vowel, Cloud 데이터(UCI machine learning repository[20~24])를 사용하여 각 20회 실험 분석하였고, 각 데이터는 <Table 1>과 같이 데이터의 클러스터 수(No of Clusters, K), 각 데이터의 특징 수(No of features), 데이터 수(No of data)로 구성되어 있다.

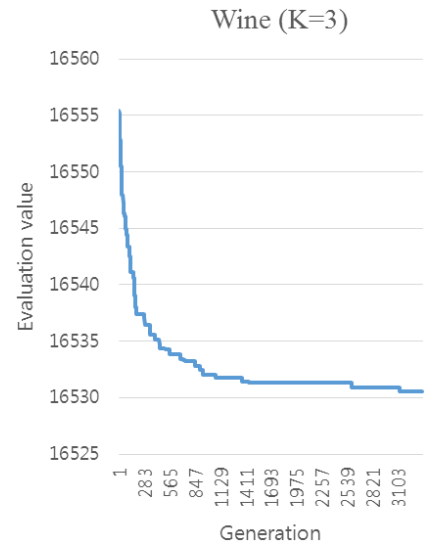
<Table 1> Data for Experiment(UCI Machine Learning Repository)-[20~24]

Name of dataset	No of clusters, K	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

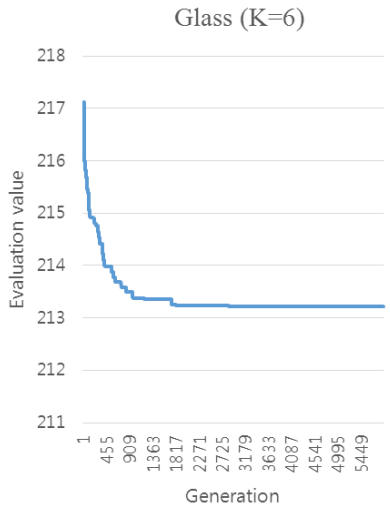
<Figure 2>는 5가지 데이터에 대하여 융합 ABC 방법 중 본 논문의 제 3장에서 제안한 KABCK 클러스터링 방법을 적용했을 때 수렴 경향을 나타낸 것이다. X축은 세대 수를 나타내고 Y축은 유클리드 거리 평가값으로 본 논문 제 2장의 식 (2)로 계산되는데 해당 값들은 각 세대마다 측정한 데이터 클러스터링 20회 평가값의 평균을 나타낸 것이다. 인공벌군집의 중요한 파라미터인 Limit (각 해의 더 좋은 이웃해 탐색 최대 수)는 여러 번의 실험을 통하여 적절한 값 25로 실험하였다. 종료조건은 현재까지 탐색한 최선해(best solution)가 미리 정한 특정횟수 이상의 세대가 진행되어도 개선이 되지 않을 때 종료하도록 설정 하였다.



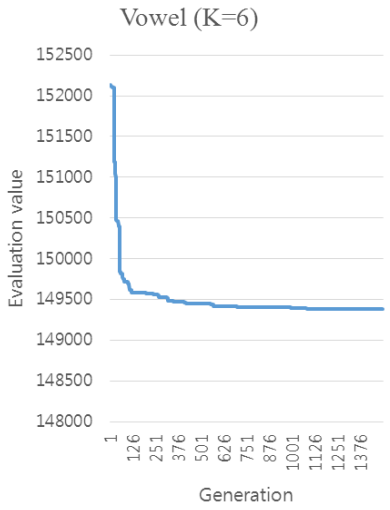
<Figure 2>(A) Convergence Trend of Data Iris



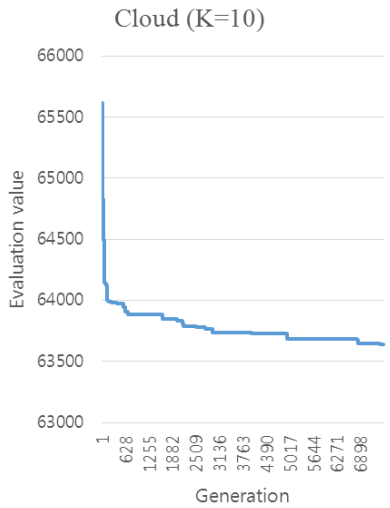
<Figure 2>(B) Convergence Trend of Data Wine



<Figure 2>(C) Convergence Trend of Glass



<Figure 2>(D) Convergence Trend of Vowel



<Figure 2>(E) Convergence Trend of Data Cloud

<Table 2> Comparative Results of K-means, HSA[8] and CABC

		K-means	Hybrid SA(HSA)[8]				Combined ABC(CABC)			
			SA[15]	KSA[13]	SAK[8]	KSAK[8]	ABC[6]	KABC[11]	ABCK	KABCK
IRIS	mean	102.003	97.4150	97.6803	97.2723	*97.2312*	97.22261	97.2729	97.2267	**97.2221**
	SD	11.3788	0.2105	0.9062	0.0534	*0.0031*	0.0023	0.0913	0.0052	**0**
	best	97.3259	97.2221	97.2221	97.2322	*97.2221*	97.2221	97.2221	97.2221	**97.2221**
WINE	mean	16934.6	16564.5	16530.5	16530.5	*16530.5*	16530.5	16530.5	16530.5	**16530.5**
	SD	1651.63	151.90	0	0	0	0	0	0	**0**
	best	16555.7	16530.5	16530.5	16530.5	*16530.5*	16530.5	16530.5	16530.5	**16530.5**
GLASS	mean	225.202	231.319	223.156	222.052	*217.867*	219.927	219.530	218.969	**213.22**
	SD	10.6686	14.5685	2.4894	10.5507	*1.2949*	1.3283	0.6796	0.4154	**0.0050**
	best	215.678	221.69	214.727	218.476	*214.669*	217.991	218.774	218.45	**213.204**
VOWEL	mean	159251	149685	150412	*149431*	149759	150816.9	151930	149626	**149376.3**
	SD	9794.83	283.41	880.17	*81.15*	533.73	411.24	996.89	216.30	**4.19**
	best	149384	149407	149405	*149375*	149380	150161	150366	149377	**149369**
CLOUD	mean	67055.4	64638.7	63214.1	*64338.6*	**63132.7**	69019.6	65938.2	63866.3	*63642.6*
	SD	648.09	755.63	406.59	*737.73*	**417.62**	1262.37	1510.36	862.31	*339.14*
	best	66194.6	62889.9	62938.0	*62834.7*	**62856.9**	65980.1	63966.2	63048.3	*63145.4*

<Table 2>는 K-means, 시뮬레이티드 어닐링(Simulated annealing, SA[15]), KSA[13], SAK[8], KSAK[8] 방법과 ABC[6]와 ABC 융합(KABC[11], 본 논문에서 제안하는 ABCK, KABCK) 방법의 실험 결과를 비교한 것이다. 각 방법에 대하여 20회의 실험을 통하여 평균(mean), 표준편차(standard deviation, SD), 가장 좋은값(best)을 비교하였다. K-means는 임의로 선택한 초기값에 따라 클러스터링 결과의 표준편차가 매우 크고 초기값에 따라 최종 탐색 결과가 다르고 안정적인 해 탐색을 할 수 없다.

상대적으로 탐색 공간이 작은 Iris와 Wine 데이터로는 융합 방법 간의 차별성이 크지 않으나, 상대적으로 해 탐색 공간과 복잡도가 높은 Glass와 Vowel 데이터로 실험한 결과 본 논문에서 제안하는 KABCK 방법의 성능이 가장 우수하였다. 그러나, 가장 사이즈가 큰 Cloud 데이터의 경우, 평가값 측면에서는 KSAK가 KABCK보다 약간 좋은 결과를 탐색할 수 있었으나, 평균계산시간 측면에서는 KSAK는 73.2초, KABCK는 59.6초로 본 논문에서 제안하는 KABCK의 계산시간이 우수하였다.

이와 같이, 일반적으로 휴리스틱 알고리즘은 해결하려고 하는 문제에 따라 또는 적용하려고 하는 알고리즘을 어떻게 설계 했는가에 따라 성능이 달라질 수 있기 때문에 절대적으로 특정 알고리즘이 다른 알고리즘보다 우수하다고 단정적으로 주장할 수 없다. 기존 SA 방법의 주요 파라미터는 초기 온도(T), 온도의 감소분($\triangle T$), 일정 온도가 고정된 상태에서 이웃해 탐색 횟수(t)의 조합을 어떻게 선택하는가에 따라 탐색 결과에 큰 영향을 준다. 따라서, 이 주요 파라미터의 조합 중 가장 최적의 조합을 선택하기 쉽지 않다. 그러나, ABC 방법의 주요 파라미터는 각 해의 더 좋은 이웃해를 탐색할 때까지 최대 시도 횟수(*Limit*) 하나이기 때문에 적절한 파라미터 값을 정하는 것이 SA 방법보다 ABC 방법이 상대적으로 유리하다.

5. 결 론

K-means는 데이터 클러스터링을 위해 현재까지도 널리 사용되고 효율적이나 초기값에 매우 민감하여 탐색한 데이터 클러스터링 결과가 지역해에 머물 가능성이 높다. 이러한 문제점은 과거 수집된 데이터의 양이 상대적으로 적었을 때보다 최근처럼 엄청난 양의 빅데이터를 분석하고자 할 때에는 큰 문제가 될 수 있다. 이런 문제점을 극복하기 위해 본 논문에서는 K-means와 융합한 인공벌군집(artificial bee colony, ABC) 데이터 클러스터링 방법을 제안하였다. 본 논문에서 제안한 KABCK는 일반적인 ABC 방법을 적용할 때 임의로 초기해를 선택하는 대신에 K-means로 해를 개선 한 후 이 해들의 이웃해를 Employed bee(EB)로 더 좋은 해를 탐색한다. 또한, Scout bee(SB)로 랜덤하게 해를 생성할 때도 그 해를 K-means로 향상시켜 적용하였다. ABC 적용 후 탐색한 Best p개의 해들의 각각을 K-means로 한번 더 해를 탐색하여 추가적으로 해를 향상시킨다. 즉, 인공벌군집의 EB와 SB의 해들을 임의로 선택하기 보다는 K-means를 적용하여 좋은 해로 향상 시켰다. 또한, 인공벌군집(ABC) 적용 후 탐색한 해 집단을 K-means로 한 번 더 향상시킴으로써 탐색 기능을 강화하

였다. 결국, 본 논문에서 제안한 KABCK는 K-means의 수렴적 탐색과 ABC의 전역 탐색 기능이 조화롭게 융합된 것이고 이 방법의 성능과 효과가 우수함을 UCI 데이터를 활용하여 검증하였다. 또한, 기존 SA 방법과 새롭게 제안한 ABC 방법을 비교 분석하였다. 본 논문에서 제안하는 융합 ABC(CABC) 방법은 파라미터 수가 혼합 SA(HSA) 방법의 수보다 작아 안정적이고 유리한 해 탐색이 가능하여 유리하였다.

Acknowledgements

This study is supported by 2017 Research Grant from Kangwon National University (Grant No. 520170157). This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP)(No.R0250-17-1002, Education Project for SW Convergence Business Data Analysis).

References

[1] Assuncao, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A., and Buyya, R., Big Data computing and clouds : Trends and future directions, *Journal of Parallel and Distributed Computing*, 2015, Vol. 79, pp. 3-15.

[2] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S., and Bouras, A., A survey of clustering algorithms for big data : Taxonomy and empirical analysis, *IEEE transactions on emerging topics in computing*, 2014, Vol. 2, No. 3, pp. 267-279.

[3] Gungor, Z. and Unler, A., K-harmonic means data clustering with simulated annealing heuristic, *Applied Mathematics and Computation*, 2007, Vol. 184, No. 2, pp. 199-209.

[4] Hruschka, E.R., Campello, R.J., and Freitas, A.A., A survey of evolutionary algorithms for clustering, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2009, Vol. 39, No. 2, pp. 133-155.

[5] Jeon, S.Y., Lee, D.H., and Bae, M.J., A study on the Application Method of Munition's Quality Information based on Big Data, *Journal of the Korea Academia-Industrial cooperation Society*, 2016, Vol. 17, No. 6, pp. 315-325.

[6] Karaboga, D. and Ozturk, C., A novel clustering approach : Artificial Bee Colony (ABC) algorithm, *Applied soft computing*, 2011, Vol. 11, No. 1, pp. 652-657.

[7] Kao, Y.T., Zahara, E., and Kao, I.W., A hybridized approach to data clustering, *Expert Systems with Applications*, 2008, Vol. 34, No. 3, pp. 1754-1762.

[8] Kim, S.S., Baek, J.Y., and Kang, B.S., Hybrid Simulated Annealing for Data Clustering, *Journal of Society of Korea Industrial and Systems Engineering*, 2017, Vol. 40, No. 2, pp. 92-98.

[9] Kim, S.S. and Byeon, J.H., Cell Grouping Design for Wireless Network using Artificial Bee Colony, *Journal of Society of Korea Industrial and Systems Engineering*, 2016, Vol. 39, No. 2, pp. 46-53.

[10] Krishna, K. and Murty, M.N., Genetic K-means algorithm, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 1999, Vol. 29, No. 3, pp. 433-439.

[11] Kumar, Y. and Sahoo, G., A two-step artificial bee colony algorithm for clustering, *Neural Computing and Applications*, 2017, Vol. 28, No. 3, pp. 537-551.

[12] Maulik, U. and Bandyopadhyay, S., Genetic algorithm-based clustering technique, *Pattern recognition*, 2000, Vol. 33, Issue. 9, pp. 1455-1465.

[13] Perim, G., Wandekokem, E., and Varejao, F., K-Means Initialization Methods for Improving Clustering by Simulated Annealing, *11th Ibero-American Conference on AI*, 2008, Lisbon, Vol. 5290, pp. 133-142.

[14] Reisi, M., Moradi, P., and Abdollahpouri, A., A feature weighting based artificial bee colony algorithm for data clustering, *In Information and Knowledge Technology (IKT), 2016 Eighth International Conference on*, 2016, Hamedan, Iran, pp. 134-138.

[15] Selim, S.Z. and Alsultan, K., A simulated annealing algorithm for the clustering problem, *Pattern recognition*, 1991, Vol. 24, No. 10, pp. 1003-1008.

[16] Singh, S.S. and Chauhan, N.C., K-means v/s K-medoids: A Comparative Study, *National Conference on Recent Trends in Engineering & Technology*, 2011, Vol. 13.

[17] Sithara, E.P. and Nazeer, K.A.A, A Hybrid K Harmonic Means with ABC Clustering Algorithm using an Optimal K value for High Performance Clustering, *International Journal on Cybernetics & Informatics*, 2016, Vol. 5, No. 2.

[18] Sun, L.X., Xu, F., Liang, Y.Z., Xie, Y.L., and Yu, R.Q., Cluster analysis by the K-means algorithm and simulated annealing, *Chemometrics and intelligent laboratory systems*, 1994, Vol. 25, No. 1, pp. 51-60.

[19] Tran, D.C., Wu, Z., Wang, Z., and Deng, C., A Novel Hybrid Data Clustering Algorithm Based on Artificial Bee Colony Algorithm and K-Mean, *Chinese Journal of Electronics*, 2015, Vol. 24, No. 4, pp. 694-701.

[20] *UCI machine learning repository Cloud datasets*, https://archive.ics.uci.edu/ml/datasets/cloud.

[21] *UCI machine learning repository Glass datasets*, https://archive.ics.uci.edu/ml/datasets/glass.

[22] *UCI machine learning repository Iris datasets*, https://archive.ics.uci.edu/ml/datasets/iris.

[23] *UCI machine learning repository Vowel datasets*, https://archive.ics.uci.edu/ml/datasets/vowel.

[24] *UCI machine learning repository Wine datasets*, https://archive.ics.uci.edu/ml/datasets/wine.

[25] Van der Merwe, D.W. and Engelbrecht, A.P., Data clustering using particle swarm optimization, *In Evolutionary Computation, 2003, CEC'03. The 2003 Congress on, IEEE*, 2003, Vol. 1, pp. 215-220.

[26] Xu, R., Xu, J., and Wunsch, D.C., A comparison study of validity indices on swarm-intelligence-based clustering, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2012, Vol. 42, No. 4, pp. 1243-1256.

[27] Yan, X., Zhu, Y., Zou, W., and Wang, L., A new approach for data clustering using hybrid artificial bee colony algorithm, *Neurocomputing*, 2012, Vol. 97, pp. 241-250.

[28] Zhang, C., Ouyang, D., and Ning, J., An artificial bee colony approach for clustering, *Expert Systems with Applications*, 2010, Vol. 37, No. 7, pp. 4761-4767.

ORCID

Bum-Su Kang | http://orcid.org/0000-0003-0507-3658

Sung-Soo Kim | http://orcid.org/0000-0002-8765-1193