J. Soc. Korea Ind. Syst. Eng Vol. 41, No. 2 : 1-8, June 2018
https://doi.org/10.11627/jkise.2018.41.2.001
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Efficient Data Clustering using Fast Choice for Number of Clusters

Sung-Soo Kim† · Bum-Su Kang

Department of Industrial Engineering, Kangwon National University

빠른 클러스터 개수 선정을 통한 효율적인 데이터 클러스터링 방법

김성수† · 강범수

강원대학교 산업공학과

K-means algorithm is one of the most popular and widely used clustering method because it is easy to implement and very efficient. However, this method has the limitation to be used with fixed number of clusters because of only considering the intra-cluster distance to evaluate the data clustering solutions. Silhouette is useful and stable valid index to decide the data clustering solution with number of clusters to consider the intra and inter cluster distance for unsupervised data. However, this valid index has high computational burden because of considering quality measure for each data object. The objective of this paper is to propose the fast and simple speed-up method to overcome this limitation to use silhouette for the effective large-scale data clustering. In the first step, the proposed method calculates and saves the distance for each data once. In the second step, this distance matrix is used to calculate the relative distance rate (V_j) of each data j and this rate is used to choose the suitable number of clusters without much computation time. In the third step, the proposed efficient heuristic algorithm (Group search optimization, GSO, in this paper) can search the global optimum with saving computational capacity with good initial solutions using V_j probabilistically for the data clustering. The performance of our proposed method is validated to save significantly computation time against the original silhouette only using Ruspini, Iris, Wine and Breast cancer in UCI machine learning repository datasets by experiment and analysis. Especially, the performance of our proposed method is much better than previous method for the larger size of data.

Keywords : Data Clustering, Heuristic Algorithm, Number of Clusters, Silhouette

1. 연구의 배경 및 목적

데이터 클러스터링 분석에서 널리 사용되는 K-means는 빠르게 해를 탐색할 수 있지만, 지역해에 머물 가능성이 있다. Kang et al.[5]이 제안한 인공벌군집(Artificial Bee Colony)방법은 지역해에 머물지 않고 전역해를 탐색할 수 있으나, 해 평가 시 K-means와 같이 클러스터 내의 유클리드 거리만을 평가기준으로 고려하기 때문에 적절한 클러스터 수를 결정할 수 없다. 데이터 클러스터링 할 때 실루엣(Silhouette)은 적절한 클러스터 수를 결정할 수 있는 적절한 평가기준(valid index)이다[11]. 여러 가지 데이터 클러스터링 평가기준 중 Arbelaitz et al.[1]는 실루엣이

Received 21 February 2018; Finally Revised 25 April 2018; Accepted 26 April 2018
† Corresponding Author : kimss@kangwon.ac.kr

가장 우수하다고 서술하였고, Xu et al.[18]도 실루엣 평가기준이 안정적이고 신뢰성이 높다고 서술하였으나, 실루엣 값을 얻기 위해서는 모든 데이터간의 거리를 계산해야 하기 때문에 계산 부담이 크다고 지적하였다.

데이터 클러스터링의 해에 대한 평가기준이 결정되면 수많은 해들 중 가장 좋은 해를 탐색해 내기 위해 최적화 방법의 적용이 필요하다. Hruschka et al.[4]는 NP-hard 그룹핑 문제인 데이터 클러스터링 문제에 효율적인 진화알고리즘 등 휴리스틱 알고리즘 적용 필요성을 주장하였다. Ng et al.[9]는 클러스터 분석을 할 때 클러스터 수를 결정하는 것이 매우 어려운 문제라고 설명하고 이를 해결하기 위해 휴리스틱 방법을 적용하였다. Lleti et al.[8]는 데이터 클러스터링을 할 때 가장 중요한 파라미터는 그룹 수이고 K-means 클러스터링 분석을 하기 위해 데이터의 그룹 수를 유전자 알고리즘과 실루엣을 이용해 찾고자 하였다. Hruschka et al.[3]은 실루엣 평가기준을 적용하고 유전자알고리즘(Genetic Algorithm, GA)을 사용하여 적절한 클러스터 수를 결정하고 해를 탐색하였다. Kim et al.[6]도 데이터에 대한 사전 정보가 없을 때, 적절한 클러스터 수를 결정하고 이에 맞는 최적의 데이터 클러스터링 해를 탐색할 수 있는 그룹탐색최적화(Group Search Optimization, GSO) 방법을 제안하였다. 이와 같이 실루엣은 적절한 클러스터 수와 해를 동시에 탐색할 수 있는 유용한 평가기준이지만, 모든 데이터 사이의 거리를 계산하여 평가하기 때문에 데이터의 크기가 커질수록 계산 시간이 많이 소요된다. 따라서, 빅데이터 분석 시 계산 시간을 단축할 수 있는 방법의 개발이 필요하다.

본 논문의 목적은 데이터 클러스터링의 해 평가 척도인 실루엣의 계산 부담을 극복할 수 있는 빠른 클러스터 수 선택 방법과 효율적인 휴리스틱 데이터 클러스터링 방법을 제안하는 것이다. 제안하는 방법은 일차적으로 거리의 상대적인 비율이 작은 순서대로 클러스터 수와 클러스터 중심 데이터를 정하고, 각 클러스터 수에 따른 실루엣 값 중에서 가장 큰 값에 해당하는 클러스터 수를 가장 적절한 클러스터 수로 빠르게 선택한다. 또한, 적절한 클러스터 수가 선택된 상태에서, 거리의 상대적인 비율의 크기가 작을수록 클러스터의 중심 데이터 역할을 할 수 있도록 확률적으로 중심 데이터를 선택하고 해를 생성한다. 이렇게 생성한 해들을 휴리스틱 알고리즘의 초기해로 적용하여 최적해를 효율적으로 탐색한다. 제 2장에서는 데이터 클러스터링 문제를 설명하고 본 논문에서 제안한 클러스터 수 선택 방법과 적용된 휴리스틱 알고리즘 방법을 상세히 설명하고 제 3장에서는 제안하는 방법의 효율성을 검증한다.

2. 빠른 클러스터 수 선택과 휴리스틱 알고리즘

데이터 클러스터링 평가기준인 실루엣의 계산 시간이 많이 소요된다는 문제점을 해결하기 위해 모든 데이터간의 거리를 한 번 계산하여 저장 후 필요 시 재사용한다. 분석하고자 하는 모든 데이터에서 데이터 j까지 거리의 상대적인 비율 V_j를 고려하여 적절한 클러스터 수를 간단하고 빠르게 찾아낼 수 있는 방법을 설명한다. 또한, 클러스터 수가 선택된 상태에서 휴리스틱 알고리즘에 적용하여 최적해를 탐색 할 수 있는 방법을 설명한다.

2.1 데이터 클러스터링 문제와 빠른 클러스터 수 선택 방법

각각의 데이터는 $\{x_i, i = 1, 2, \cdots, n\}$이라하고 n개의 데이터 집합으로 구성된다. 각 데이터 $i(x_i)$는 P차원(특징, feature)으로 구성되는데 x_{ip}는 데이터 i의 특징 p의 값을 표현할 수 있고 n개의 데이터를 K개의 그룹으로 클러스터링 하는 문제를 수리적으로 정립할 수 있다[7]. 간단하고 빠른 클러스터링 해 탐색 방법을 만들어 내기 위해 P개의 특징을 가진 데이터 i에서 데이터 j까지의 거리(d_{ij})를 식 (1)을 사용하여 한번 계산하여 저장하고 필요 시 재사용한다.

본 연구에서는 실루엣 평가기준을 효율적으로 적용하기 위해 식 (2) V_j(모든 데이터에서 데이터 j까지 거리의 상대적인 비율)를 활용하는데, 본래, Park et al.[10]이 제안한 V_j는 K-medoid의 계산 문제점을 개선하기 위해 제안한 효율적인 아이디어이다. V_j가 작을수록 해당 데이터 j가 클러스터링을 할 때 중심 데이터 역할을 하는 것이 유리하고 실루엣 값이 좋은 해 탐색에 효과적으로 적용할 수 있다.

즉, 가장 작은 V_j값 순서대로 클러스터 수 K개의 데이터 중심을 결정하고 나머지 데이터는 가장 가까운 데이터 중심에 소속시켜 클러스터링 했을 경우 해당 해에 대하여 실루엣 값을 계산하고 1에 가장 가까운 K를 적절한 클러스터 수로 결정할 수 있다. 클러스터링 해를 평가할 때 식 (3)~식 (5)를 사용하여 데이터 $i(x_i)$에 대한 실루엣 $S(x_i)$를 계산한다[6, 11]. 실루엣 값 계산을 위한 식 (3)~식 (5)를 구체적으로 설명하면 다음과 같다. 클러스터 A에 포함되어 있는 데이터 $i(x_i)$와 클러스터 A에 속한 다른 데이터들과의 평균 거리를 $a(x_i)$라 할 때, 이 값이 작을수록 클러스터 내의 데이터들이 조밀하게 모여 있다는 것을 의미한다. 클러스터 A와 다른 클러스터 B와 C가 존재하고 클러스터 A에 속한 데이터 i에서 다른 클러스터 B와 C의 각 데이터들과의 평균거리를 각각

$d(x_i, B)$와 $d(x_i, C)$라 할 때, $b(x_i)$는 데이터 i에서 다른 클러스터에 포함된 데이터간의 평균거리 중 가장 작은 값을 의미한다. 만약, $d(x_i, B)$가 $d(x_i, C)$보다 작다면 $b(x_i) = d(x_i, B)$가 된다. $b(x_i)$는 식 (3)과 같이 나타낼 수 있는데, 이 값이 클수록 클러스터 간의 구별이 뚜렷하다고 판단할 수 있다. 따라서, 식 (4)와 같이 $S(x_i)$를 계산할 수 있고 모든 데이터 $i(x_i)$에 대하여 $S(x_i)$를 구하여 합한 값 $\sum_{i=1}^{n} S(x_i)$을 데이터 수 n으로 나누면 평균 실루엣 값 $SIL(S(x_i))$을 구할 수 있고 식 (5)로 나타낸다.

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

$$V_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \quad (2)$$

$$b(x_i) = \min\{d(x_i, B), d(x_i, C)\} \quad (3)$$

$$S(x_i) = \{b(x_i) - a(x_i)\} / \max\{a(x_i), b(x_i)\} \quad (4)$$

$$SIL(S(x_i)) = (1/n)\sum_{i=1}^{n} S(x_i) \quad (5)$$

본 논문의 목적은 데이터에 대한 사전 정보가 없을 때, 실루엣 평가기준을 적용하여 적절한 클러스터 수와 클러스터링 해를 제시할 수 있는 간단하고 빠른 방법을 제안하는 것이다. 클러스터 수 K가 2라면 V_j가 가장 작은 데이터 2개를 선택하고 해당 데이터를 데이터의 중심으로 설정하고 다른 데이터들은 이 2개의 데이터에서 가장 가까운 데이터 중심의 클러스터로 포함시켜 해를 구성하고 이 해에 대한 실루엣 값을 계산한다. 추가적으로 클러스터 수가 3~K를 수행하여 해당 실루엣 값을 계산하고 실루엣 값이 가장 큰 해당 클러스터 수 K를 가장 적절한 클러스터 수로 결정한다.

이와 같이 적절한 클러스터 수와 클러스터링 해를 결정할 수 있는 간단하고 빠른 방법은 <Table 1> 10개의 데이터 예제(각 데이터 x_i는 특징 a_1과 a_2로 표시)로 다음과 같이 설명할 수 있다. 먼저, 식 (1)로 모든 데이터간의 거리를 1회 계산하여 저장하여 필요 시 재사용하고 식(2)로 10개의 데이터의 V_j를 <Table 1>과 같이 계산한다.

만약, K가 2라면 V_j값이 가장 작은 x_7과 x_5가 데이터 중심으로 선택되고 나머지 데이터 x_1, x_2, x_3, x_4는 데이터 중심 x_5에 가까워 한 개의 클러스터가 형성되고 다른 나머지 데이터 x_6, x_8, x_9, x_{10}은 데이터 중심 x_7에 가까워 다른 한 개의 클러스터가 형성된다. 이와 같이 2개의 클러스터로 구성된 클러스터링 해의 실루엣 값은 식 (3)~식 (5)를 사용하여 0.272938로 계산된다. 만약, K가 3이라면 V_j 값이 가장 작은 x_7, x_5, x_6이 데이터 중심으로 선택되고 나머지 데이터 x_1, x_2, x_3, x_4는 데이터 중심 x_5에 가까워 한 개의 클러스터가 형성되고 데이터 x_8, x_9, x_{10}은 데이터 중심 x_7에 가까워 다른 한 개의 클러스터가 형성되고 나머지 x_6는 1개의 데이터로 하나의 클러스터가 형성된다. 이와 같이 3개의 클러스터로 구성된 클러스터링 해의 실루엣 값은 0.181605로 계산된다. 만약, K가 4라면 V_j 값이 가장 작은 x_7, x_5, x_6, x_4이 데이터 중심으로 선택되고 나머지 데이터 x_1, x_2, x_3는 데이터 중심 x_5에 가까워 한 개의 클러스터가 형성되고 데이터 x_8, x_9, x_{10}은 데이터 중심 x_7에 가까워 다른 한 개의 클러스터가 형성되고 나머지 x_4와 x_6는 각각 1개의 데이터로 하나의 클러스터가 형성된다. 이와 같이 4개의 클러스터로 구성된 클러스터링 해의 실루엣 값은 0.390931로 계산된다.

이런 과정을 통하여 가능한 클러스터 수 K의 모든 경우를 계산 한 후, 실루엣 값이 가장 클 때(1에 가까울 때)의 K 값을 적절한 클러스터 수로 선택할 수 있다. 제 3.1절에서 UCI 데이터를 활용하여 이 예제와 같이 실험을 통하여 적절한 클러스터 수를 결정하는 과정을 설명한다.

2.2 거리의 상대적인 비율을 적용한 휴리스틱 알고리즘

본 절에서는 제 2.1절에서 결정된 클러스터 수 K를 고정한 상태에서 다음 과정을 진행한다. V_j값에 반비례하는 확률로 K개의 중심 데이터를 선택하고, 이 데이터를 중심으로 다른 모든 데이터가 클러스터링 된 해를 초기해로 사용한 휴리스틱 알고리즘을 적용하여 최적해를 탐색한다.

휴리스틱 알고리즘 중에 하나인 그룹탐색최적화(Group search optimization, GSO)는 He et al.[2]가 동물의 탐색 행태를 모방하여 개발하였다. GSO는 Producer, Scrounger 및 Ranger의 세 가지 역할을 담당하는 구성원(member)들의

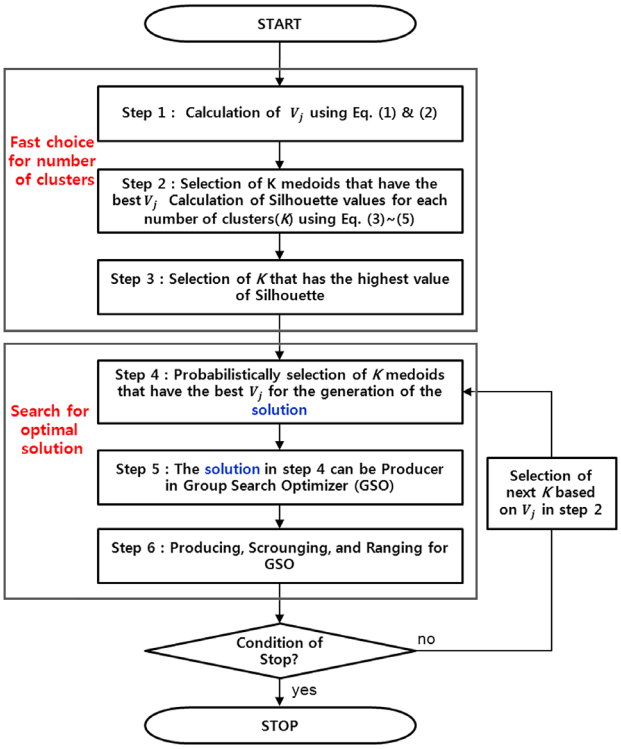
<Table 1> 10 Data (a_1, a_2) Example for V_j

object	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
a_1	1	1	2	5	5	6	6	9	10	10
a_2	8	7	8	1	2	1	2	7	6	7
V_j	1.2111	1.1429	1.1166	0.8819	0.7930	0.8768	0.7871	1.0154	1.0625	1.1127

해로 이루어진다. Kim et al.[6]은 기본 실루엣 평가기준을 적용한 GSO를 다음과 같이 제안하였다. 단계 1에서는 초기 구성원(해) 표현과 해 생성 후 실루엣 평가하였다. 단계 2에서는 Producer(가장 좋은 해) 선택과 producing을 진행하였다. 단계 3에서는 Scrounger 선택과 scrounging을 진행하였다. 단계 4는 Ranger 선택과 ranging을 진행하였다. 단계 5는 모든 해에 대하여 실루엣 평가 후 해를 업데이트하였다. 단계 6에서는 종료 조건 확인 후 최적해 탐색 진행하였다. 기존 GSO 방법을 사용하여 데이터 수 75개 Ruspini[12]와 UCI 데이터[14, 15, 16, 17] 중 데이터 수 각각 150, 178, 683, 214인 Iris, Wine, Breast cancer, Glass에 대하여 최적해를 탐색하였다. 상대적으로 큰 사이즈의 데이터인 Breast cancer, Glass는 탐색한 실루엣 값이 0에 가까워 실루엣 값이 0.5 이상인 적절한 클러스터링 해를 탐색할 수 없다는 한계가 있음을 <Table 3>으로 확인할 수 있다. 이런 문제점을 극복하기 위해 본 논문의 제 2.1절에서 제안하는 간단하고 빠른 클러스터링 해 탐색 방법이 적용된 효율적인 알고리즘인 GSO 방법은 다음과 같이 설명할 수 있다.

제 2.1절 식 (2)를 사용한 V_j 값으로 얻은 해는 효율적 GSO 방법의 중요한 초기 정보가 된다. 본 논문에서 제안하는 방법의 단계 1에서는 GSO의 초기해 군을 생성할 때, 제 2.1절의 방법을 적용하여 초기해 1개를 생성한다. 만약, V_j 값이 가장 작은 K개의 데이터를 선택하고 그 데이터를 중심으로 다른 데이터를 클러스터링 할 경우 중심점으로 선택한 데이터가 고르게 분포하지 못하고 한쪽으로 쏠릴 경우가 발생할 가능성이 있다. 따라서, GSO 방법에 적용 할 때는 중심데이터가 다양하게 분포하여 위치될 수 있도록 각 데이터의 V_j 값이 작을수록 선택확률을 높도록 하여 확률적으로 중심데이터를 선택한다. 이렇게 생성된 해 1개를 초기해 군에 포함시키고 나머지 해들은 랜덤하게 생성한다.

위의 방법으로 얻은 초기해는 랜덤하게 생성한 해들보다 실루엣 값이 높아(제 3.1절에 Ruspini, Iris, Wine, Breast cancer, Glass 데이터에 대하여 빠른 해 탐색 방법 적용하여 실루엣 값 계산 결과로 확인 할 수 있음), GSO를 적용하여 시작할 때 Producer가 될 확률이 높고 producing 할 때도 현재의 Producer보다 더 좋은 해를 생성할 수 있다. 나머지 scrounger들은 V_j를 고려하여 만들어진 현재의 Producer를 모방하여 닮아 가면서 새로운 해들을 생성하고 더 좋은 해들을 탐색하게 되는 scrounging을 진행한다. 나머지 Ranger들은 다양한 해 탐색을 수행한다. V_j값이 적용된 해가 포함된 초기해 군의 해들은 GSO를 적용하여 최적해를 탐색할 때 효율적으로 해를 탐색할 수 있고 계산시간을 상당히 줄일 수 있다는 것을 제 3.2절의 실험을 통하여 검증하고자 한다. <Figure 1>은 본



<Figure 1> Flow Chart of Algorithm

논문의 제 2장에서 제안한 간단하고 빠른 클러스터 수 선택과 효율적인 데이터 클러스터링 방법의 흐름도를 나타낸 것이다.

3. 실험 및 분석

제 2장에서 제안하는 실루엣 기반 간단하고 빠른 클러스터 수 선택 방법과 효율적인 휴리스틱 알고리즘 적용에 대한 실용성을 검증하기 위해 본 절의 실험 및 분석은 윈도우10 프로세서 : Intel(R) Core™ i5-4590 CPU @ 3.30GHz 3.30GHz 메모리(RAM) : 4.00GB, 64비트, x64 기반 프로세서 운영체제, Visual C++ 환경에서 수행하였다.

3.1 빠른 클러스터 수 선택

간단하고 빠른 클러스터 수 선택 방법을 적용하면 제 2.1절의 예제와 같이 어느 정도 실루엣 평가값 이상의 해를 <Table 2>와 같이 찾을 수 있다. 즉, Ng et al.[9]와 Struyf et al.[13]가 제시한 실루엣 값 해석에 따르면 실루엣 값이 0.5 이상이면 클러스터링이 적절히 된 것으로 해석하는데, 데이터 Ruspini, Iris, Wine는 최적화 방법과 추가적인 계산시간을 사용하지 않더라도 0.5에 가깝거나 그 이상인 해를 찾아낼 수 있다.

<Table 2> Silhouette Value using Fast Silhouette Searching(Number of Clusters, K)

K	Ruspini	Iris	Wine	Breast cancer	Glass
2	0.38318	0.53287	0.48853	0.36937	0.086498
3	0.61647	0.12788	0.09405	0.24926	0.060450
4	0.48667	0.11664	-0.02716	0.07732	0.026130
5	0.47028	0.17305	-0.03463	-0.03118	0.066948
6	0.45159	0.15499	-0.04593	0.04022	0.008745
7	0.26412	-0.04515	-0.04521	0.08293	-0.12970
8	0.24077	-0.06618	-0.02384	0.05984	-0.23469
9	0.33001	-0.07287	0.11723	0.05288	-0.25344

<Table 2>의 4가지 UCI 데이터의 경우 제 2.1절의 방법으로 얻을 수 있는 실루엣 값을 각각의 K에 대하여 계산한다. 실루엣 값이 가장 클 때의 K를 적절한 클러스터 수 K로 결정할 수 있다. 즉, Ruspini 데이터는 K = 2일 때 실루엣 값 0.383177, K = 3일 때 실루엣 값 0.616466 … K = 9일 때 실루엣 값 0.330007을 본 논문 제 2.1절의 방법만으로도 얻을 수 있고, 이중 가장 좋은 해(<Table 2>의 K = 3일 때 실루엣 값 0.616466)를 탐색 해 낼 수 있다. 다른 데이터에 대하여 제 2.1절의 방법만을 적용했을 경우 Iris의 경우 K = 2일 때 실루엣 최대값 0.532872, Wine의 경우 K = 2일 때 실루엣 최대값 0.488526, Breast cancer의 경우 K = 2일 때 실루엣 최대값 0.369373과 이에 해당하는 해를 빠르고 간단하게 탐색해 낼 수 있다. 그러나, Glass의 경우 빠른 클러스터 수 선택 방법만으로는 실루엣 값이 0에 가까워 적절한 클러스터 해를 탐색해 내지 못하였다.

3.2 거리의 상대적인 비율을 적용한 휴리스틱 알고리즘 분석

제 3.1절의 간단하고 빠른 방법만으로는 최적의 클러스터링 해를 탐색할 수 없고 추가적인 수렴과 해 개선이 필요할 때, 휴리스틱 알고리즘을 적용할 수 있다. 본 논문에서 제안하는 간단하고 빠른 방법을 그룹탐색최적화(GSO) 방법에 적용하여 계산시간 부담을 줄이면서 더 좋은 실루엣 값의 해를 탐색한다. 본 절에서는 제 2.2절에서 제안한 방법의 성능을 검증하기 위해 UCI 데이터에 대하여 실험 분석하였다. 각 데이터의 실루엣 평가값이 수렴 할 때 일정 세대가 진행되어도 실루엣 값이 더 이상 추가적으로 수렴되지 않을 때 종료하고 그 때까지의 계산시간을 측정하였다.

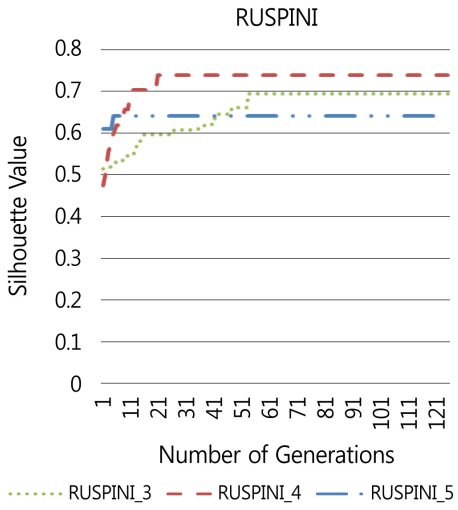
<Figure 2>는 간단하고 빠른 클러스터 수 선택과 효율적인 해 탐색 방법을 GSO에 적용 했을 때, 각 클러스터 수 K에서 가장 좋은 실루엣 값을 탐색하는 과정의 수렴 경향을 나타낸 것이다. 예를 들어, Ruspini 경우 <Table 2>에서 실루엣 값이 큰 클러스터 수는 3~5이고, Ruspini_3, 4, 5로 나타낸다. 클러스터 수가 4일 때 가장 좋은 값 0.737657을 탐색함을 나타낸다. Iris, Wine과 Breast cancer 경우 클러스터 수가 각각 2일 때 가장 좋은 값 0.686232, 0.660087, 0.595527을 Glass 경우 클러스터 수가 5일 때 가장 좋은 값 0.597413을 수렴 탐색함을 나타낸다.

<Table 3>은 각각의 데이터에 대하여 빠른 해 탐색 방법을 적용한 GSO와 기존 GSO[6]의 최적해 실루엣 평가값과 계산시간을 비교한 것이다. <Table 3>의 경우 Ruspini는 K = 3~5, Iris와 Wine의 경우 K = 2와 이와 근접한 K = 3~4, Breast cancer는 K = 2~4, Glass는 K = 5~7을 실험하였다.

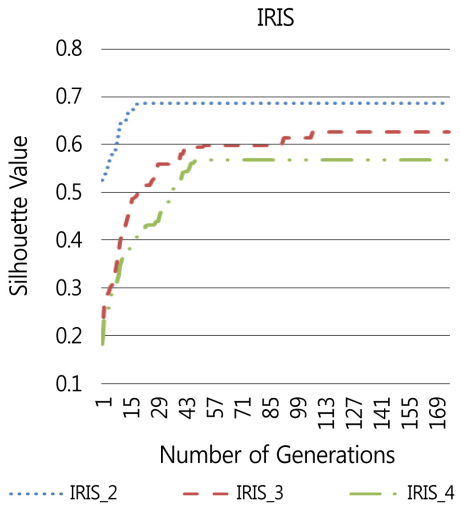
<Table 3>의 결과에 따르면 제 3.2절에서 제안하는 간단하고 빠른 해 탐색 방법을 적용한 GSO 방법의 성능이 월등한 것으로 분석되었다. Ruspini, Iris, Wine에 간단하고 빠른 해 탐색 방법의 GSO 방법을 적용했을 경우 비슷한 실루엣 평가값의 해를 탐색하는데 소요되는 계산시간이 기존 방법과 비교하여 각각 24~27%, 16~21%, 10~13% 수준으로 상당히 줄어들었다.

다른 데이터보다 상대적으로 큰 사이즈의 Breast cancer와 Glass는 클러스터 수 K가 증가할수록 가능해의 경우의 수가 기하급수적으로 증가하여 기존 GSO 방법의 한계로 인하여 클러스터링 해 탐색이 어렵고 탐색한 실루엣 평균값들이 0에 가까워 적절한 클러스터링 해를 탐색해 낼 수 없었다.

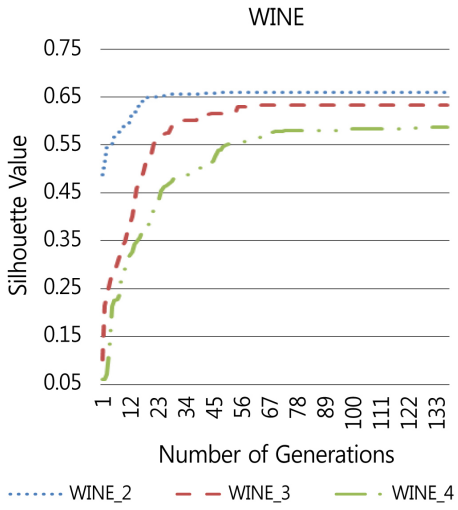
그러나, 새롭게 제안한 빠른 클러스터 수 선택과 GSO 데이터 클러스터링 방법은 실루엣 평균값이 0.5에 근접하거나 그 이상의 해를 탐색해낼 수 있어 적절한(reasonable) 클러스터링 해를 탐색해낼 수 있다. 새로운 방법의 계산시간이 줄어들지 않고 기존 방법과 유사하거나 다소 길게 소요된 이유는 더 좋은 해를 탐색하기 위해 소요되는 계산시간이 필요하기 때문이다.



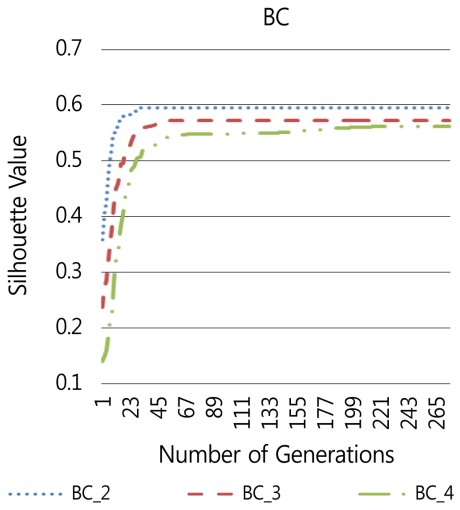
<Figure 2(A)> Trend of Convergence for Data Ruspini



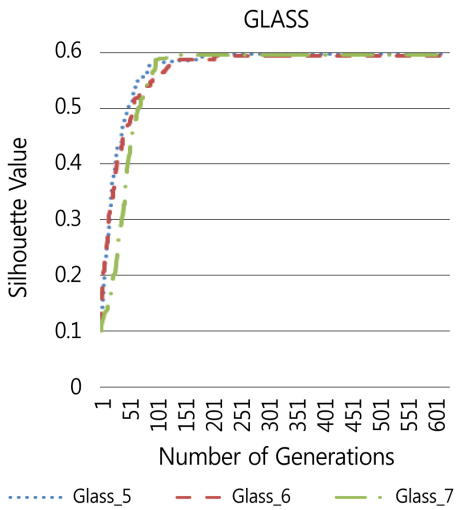
<Figure 2(B)> Trend of Convergence for Data Iris



<Figure 2(C)> Trend of Convergence for Data Wine



<Figure 2(D)> Trend of Convergence for Data BC



<Figure 2(E)> Trend of Convergence for Data Glass

<Table 3> Comparative Study between GSO with Fast Silhouette Search and GSO[6]

			GSO[6]		GSO with Fast Silhouette Search	
			Silhouette Value	Computation time(sec)	Silhouette Value	Computation time(sec)
RUSPINI	3	AVG	0.569368	23.6138	0.604836	5.7368
		S.D	0.086715		0.068472	
		Best	0.641392		0.641392	
	4	AVG	0.686717	26.844	0.591366	7.0372
		S.D	0.314109		0.142899	
		Best	0.737657		0.737657	
	5	AVG	0.463924	31.29495	0.584433	8.5155
		S.D	0.314109		0.083701	
		Best	0.695847		0.694735	
IRIS	2	AVG	0.686232	136.1451	0.647952	22.2502
		S.D	0		0.121052	
		Best	0.686232		0.686232	
	3	AVG	0.621266	161.0614	0.558428	29.6126
		S.D	0.016069		0.047904	
		Best	0.630200		0.627294	
	4	AVG	0.128177	174.7001	0.522702	37.1478
		S.D	0.258531		0.02131	
		Best	0.568458		0.568475	
WINE	2	AVG	0.660087	373.7161	0.660087	36.03
		S.D	0		0	
		Best	0.660087		0.660087	
	3	AVG	0.573892	444.9382	0.591475	47.7696
		S.D	0.161261		0.045098	
		Best	0.636046		0.632456	
	4	AVG	0.046863	448.8067	0.554223	60.3295
		S.D	0.154076		0.03061	
		Best	0.382772		0.586593	
BREAST CANCER	2	AVG	0.595527	1164.92	0.595527	1169.484
		S.D	0		0	
		Best	0.595527		0.595527	
	3	AVG	0.110928	1697.525	0.572383	1755.485
		S.D	0.257866		0.001095	
		Best	0.572212		0.573471	
	4	AVG	-0.00716	2130.97	0.503771	2282.981
		S.D	0.00034		0.130727	
		Best	-0.00691		0.563126	
GLASS	5	AVG	0.03913	113.2165	0.41996	133.6945
		S.D	0.196318		0.278344	
		Best	0.597824		0.597413	
	6	AVG	-0.02312	107.6907	0.528216	152.1379
		S.D	0.022772		0.195883	
		Best	0.04151		0.593695	
	7	AVG	-0.0264	165.5666	0.473177	186.2114
		S.D	0.025024		0.234095	
		Best	0.043428		0.59522	

4. 결 론

기존 연구에 따르면 데이터 클러스터링 할 때 해 평가 기준은 매우 중요하며, 여러 평가기준 중에서 실루엣이 여러 측면에서 유용하지만 계산 부담(특히, 데이터 크기가 커지거나 클러스터 수가 커질수록)으로 적용하는 것에 한계가 있다.

본 논문에서는 실루엣 기반 간단하고 빠른 클러스터 수 선택 방법과 효율적인 휴리스틱 데이터 클러스터링 방법을 제안하였다. 이 방법은 1차적으로 거리의 상대적인 비율을 사용하여 빠르게 클러스터 수를 선택한다. 클러스터 수가 결정된 상태에서 2차적으로 거리의 상대적인 비율을 확률적으로 적용하여 초기해를 생성하고, 이 해들을 활용하여 휴리스틱 데이터 클러스터링 방법으로 해를 효율적으로 탐색한다. 실제로 UCI 데이터에 본 논문에서 제안하는 방법을 적용하여 최적해 탐색이 매우 효과적임을 검증하였다.

본 논문에서 제안한 방법을 적용할 경우, 기존 실루엣만을 적용한 방법과 비교했을 때 성능(평가값과 계산시간)이 상당히 향상 되었다. 상대적으로 작은 크기의 데이터 경우 유사한 실루엣 평가값을 탐색 해 내는데 제안하는 방법을 적용할 경우 계산시간이 크게 향상되었다. 상대적으로 큰 크기의 데이터 일 때 기존 방법의 경우 해 탐색 시 전역해를 탐색하지 못할 경우가 많아 추가적인 수렴에 한계가 있었다. 제안하는 방법을 적용할 경우 해 탐색 시 전역해 탐색이 가능하고 추가적인 수렴이 가능하게 되어 실루엣 평가값은 상당히 개선되어 본 논문에서 제안한 방법이 매우 효용성이 높다는 것을 검증하였다.

References

[1] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J.M., and Perona, I., An extensive comparative study of cluster validity indices, *Pattern Recognition*, 2013, Vol. 46, No. 1, pp. 243-256.

[2] He, S., Wu, Q.H., and Saunders, J.R., Group search optimizer : an optimization algorithm inspired by animal searching behavior, *IEEE transactions on evolutionary computation*, 2009, Vol. 13, No. 5, pp. 973-990.

[3] Hruschka, E.R. and Ebecken, N.F., A genetic algorithm for cluster analysis, *Intelligent Data Analysis*, 2003, Vol. 7, No. 1, pp. 15-25.

[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] Kang, B.S. and Kim S.S., Combined Artificial Bee Colony for Data Clustering, *Journal of Society of Korea industrial and Systems Engineering*, 2017, Vol. 40, No. 4, pp. 203-210.

[6] Kim, S.S., Baek, J.Y., and Kang, B.S., Group Search Optimization Data Clustering Using Silhouette, *Journal of the Korean Operations Research and Management Science Society*, 2017, Vol. 42, No. 3, pp. 25-34.

[7] 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.

[8] Lleti, R., Ortiz, M.C., Sarabia, L.A., and Sánchez, M.S., Selecting variables for k-means cluster analysis by using a genetic algorithm that optimizes the silhouettes, *Analytica Chimica Acta*, 2004, Vol. 515, No. 1, pp. 87-100.

[9] Ng, R.T. and Han, J., Efficient and Effective Clustering Methods for Spatial Data Mining, *In Proceedings of VLDB*, 1994, pp. 144-155.

[10] Park, H.S. and Jun, C.H., A simple and fast algorithm for K-medoids clustering, *Expert systems with applications*, 2009, Vol. 36, No. 2, pp. 3336-3341.

[11] Rousseeuw, P.J., Silhouettes : a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, 1987, Vol. 20, pp. 53-65.

[12] Ruspini, E.H., Numerical methods for fuzzy clustering, *Information Sciences*, 1970, Vol. 2, No. 3, pp. 319-350.

[13] Struyf, A., Hubert, M., and Rousseeuw, P., Clustering in an object-oriented environment, *Journal of Statistical Software*, 1997, Vol. 1, No. 4, pp. 1-30.

[14] *UCI machine learning repository Breast Cancer datasets*, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29.

[15] *UCI machine learning repository Glass datasets*, https://archive.ics.uci.edu/ml/datasets/Glass+Identification.

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

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

[18] 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.

ORCID

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

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