Fuzzy K-means 군집분석을 위한 다양한 유효성 지수 개발*

이수현(주저자)
전남대학교 기후특성화대학원, 박사후연구원
(lovingsh79@jnu.ac.kr)
김재윤(공저자)
전남대학교 경영학부, 교수
(jaeyun@jnu.ac.kr)
정영선(교신저자)
전남대학교 산업공학과, 조교수
(young.jeong@jnu.ac.kr)

경영학 분야에서는 군집분석을 이용하여 동질적인 특성을 지닌 집단을 도출하고 이를 재무, 마케팅, 생산관리 분야 등에서 다양하게 활용하고 있다. 따라서 군집분석에 의한 군집화 결과는 기업의 가치를 극대화 시킬 수 있는 핵심자원의 역할을 하고 있다. 본 연구에서는 군집분석에서 필요한 군집화 결과의 유효성을 검증하는 군집화 유효성 지수(clustering validity index)의 개발에 관한 이론적 연구를 다루고자 한다. 구체적으로, 다양한 형태의 데이터에서 군집화의 유효성 검증 성능이 우수하다고 알려진 Dunn(DU) 지수, Calinski and Harabasz(CH) 지수, 그리고 Davies-Bouldin(DB) 지수들을 응집도와 분리도의 개념으로 분해하고, 각 CVI의 응집도 계산에 서포트 벡터 데이터 표현(support vector data description) 개념을 반영하여 새로운 CVI들을 제안하였다. 그리고 Fuzzy K-means 알고리즘으로 다양한 속성을 갖는 벤치마크 문제를 군집분석한 결과의 유효성을 검증하였다. 새로운 CVI들은 CH와 DB 지수의 약점을 개선하였음을 확인하였다. CH 지수는 노이즈와 비대칭 데이터에서 약점을 가지고 있었고, DB 지수는 부분군집과 임의형상 데이터에서 약점을 가지고 있었다. 본 연구를 통해 SVDD 개념을 CVI의 응집도에 반영할 수 있으며, 이를 반영한 새로운 CVI들은 군집화 유효성 검증에 효과적임을 확인할 수 있었다. 본 연구에서 제안한 CVI의 응집도 계산방법은 기존에 알려진 다양한 CVI의 응집도에 적용이 가능할 것으로 기대된다. 이는 군집분석 대상이 확대되고 연구가 다양해지고 있는 상황에서 군집분석 및 CVI의 이론 확장, 그리고 SVDD 적용범위 확장에 공헌할 것으로 기대된다.

주제어: 군집분석, 유효성 지수, 서포트 벡터 데이터 표현, 응집도, 퍼지 K-평균

Ⅰ. 서 론

군집분석(cluster analysis)은 데이터(data)의 패턴과 분포를 찾아 유사한 속성을 갖는 몇 개의 그룹으로 데이터를 나누는 기술로써, 경영학과 경제학, 의학 등 다양한 학문에서 사용되고 있다(이만재, 2012; Theodoridis and Koutroumbas, 2006). 특히, 경영학 분야에서 군집분석은 기업 활동을 통해 수집된 많은 데이터들을 탐색하고 분석하여 내재적 구조를 지닌 그룹으로 분할하고, 각 그룹이 의미하는 정보를 기업의 의사결정 시스템에 반영함으로써 기업의 성과를 높이는데 이용되고 있다(오은녕・이희상, 2002; 황인수, 2002; Hruschka, Campello, Freitas and Carvalho, 2009; Xu and Wunsch, 2005). 따라서 데이터를 몇 개의 그룹으로 어떻게 군집화할 것인가는 매우 중요한 문제이다. 이와 관련하여 지금까지 많은 연구들에서 다양한 군집화 알

최초투고일: 2016. 12. 6 수정일: (1차: 2017. 5. 18) 게재확정일: 2017. 7. 7
* 본 연구는 제1저자의 박사학위 논문 일부를 재정리한 내용임

고리즘들이 개발되었고, 여러 응용문제에 대한 군집분석의 적용 사례들이 보고되고 있다. 그러나 대부분의 군집화 알고리즘은 데이터의 내재적 군집구조가 없는 상황에서도 군집을 찾아 내기 때문에(용환승 · 나연묵 · 박종수 · 승현우 · 이민수 · 이상준 · 최린, 2007), 좋은 군집화 알고리즘의 개발을 통해 데이터를 군집화하는 것 못지않게 그 결과가 타당한지에 대한 의문을 해결하는 것도 중요하다. 즉, 군집분석에서는 형성된 군집이 유사한 데이터로 잘 그룹화되어 있는지, 데이터에는 구조적 형태가 존재하는지 등을 확인할 수 있는 군집화 결과의 유효성(타당성 또는 적절성) 검증이 요구된다(김영옥 · 이수원, 2002). 군집분석에서 알고리즘 개발 및 사례 연구에 비하여, 군집화 결과의 유효성 검증에 관한 연구는 미흡하다. 본 연구는 군집화 결과의 유효성을 판단하는 기법에 초점을 맞추어 진행하고자 한다.

군집화 결과의 유효성을 판단하는 방법은 감독(supervised), 비감독(unsupervised), 상대적(relative) 기법의 3가지 유형으로 분류된다(Halkidi and Vazirgiannis, 2001; Jain, Murty and Flyn, 1999). 감독기법은 군집화된 구조가 외부 구조와 어느 정도 일치하는지를 측정하는 방법으로 외부 인덱스(external indices)라고도 부른다. 비감독기법은 외부 정보에 의존하지 않고 군집화된 구조의 좋은 정도를 평가하는 방법으로 내부 인덱스(internal indices)라고도 부른다. 마지막으로 상대적 기법은 서로 다른 군집화 결과를 비교하는 감독 또는 무감독 기법이다. 예로, K-means 알고리즘에 의한 군집화 결과를 오차제곱합(sum of the squared error) 또는 엔트로피(entropy)를 사용하여 비교하는 것이 여기에 속한다(용환승 외, 2007). 일반적으로 군집화 문제의 군집수는 사전에 정해진 것이 없고 정답도 없으며, 군집분석은 비감독 학습으로 데이터를 군집한다(김민호 · 유현진 · Ramakrishna, 2005). 따라서 본 연구에서는 외부 정보에 의존하지 않고, 군집화 결과의 유효성을 판단하는 비감독 방법에 관한 연구가 중요하다고 판단하였다. 또한, 최근 비감독 방법 중 하나로 군집화 결과의 타당성을 평가하는 지표인 군집화 유효성 지수(clustering validity index: CVI)에 대한 관심이 높아지고도 있다(김민호 외, 2005; Maulik and Bandyopadhyay, 2002; Xie and Beni, 1991).

지금까지 CVI들은 통계적 기법 또는 직관적이거나 경험적인 사실에 기반하여 개발되었을 뿐만 아니라(Halkidi *et al.*, 2001; Liu, Li, Xiong, Gao, Wu and Wu, 2013; Saitta, Raphael and Smith, 2008), 데이터의 특성을 사전에 알고 이에 적합한 군집화 알고리즘의 성능을 판단할 수 있도록 개발된 것이 주를 이루고 있어서 적용이 제한적이다. 따라서 주어진 데이터의 특성을 사전에 모르더라도, 군집화 유효성을 효과적으로 검증할 수 있는 CVI를 개발할 필요가 제기되고 있다. 본 연구는 적용성이 높은 CVI를 제안하는데 그 목적이 있다. 이를 위해 기존 CVI들이 군집화 유효성 검증에 취약하다고 알려져 있는 구조(예: 임의형상(arbitrary shape), 노이즈(noise), 부분군집(sub cluster), 비대칭분포(skewed distribution) 등)를 갖는 데이터에서 우수한 성능을 보일 수 있도록 SVDD(support vector data description: SVDD) 기반의 새로운 CVI들을 제안한다. 그리고 본 연구에서 개발한 CVI들은 다양한 벤치마킹(benchmarking) 문제를 이용하여 성능을 분석한다. 여기에서 언급한 각 데이터의 구조에 대한 정의 및 특성과 새로운 CVI들의 특징은 '5.3 시사점'에서 자세하게 설명한다.

본 연구는 다음과 같이 구성된다. 제2장에서는 CVI의 연구현황 및 특성에 대해 기술한다. 제3장에

서는 Fuzzy K-means 군집분석 알고리즘을 설명하고, 제4장에서는 새롭게 제안하는 CVI들의 개발 원리 및 특징을 소개한다. 제5장에서는 벤치마킹 문제를 Fuzzy K-means 알고리즘으로 군집화한 후, 이 결과에 대하여 본 연구에서 제안한 CVI들의 유효성 검증 성능 결과를 기존 CVI들과 비교 분석한다. 마지막으로 제6장에서는 연구결과 및 시사점, 그리고 연구의 한계점과 추후연구를 다룬다.

Ⅱ. CVI 연구현황 및 기존 지수 분석

CVI는 군집화 결과를 평가하는 도구이며, 계산된 지수 값을 근거로 군집화 결과에 대한 신뢰성을 검증한다. 일반적으로 CVI는 군집화 결과의 유효성을 응집도와 분리도라는 2가지 평가 기준을 이용하여 판단한다. 응집도(compactness)는 한 군집에 존재하는 두 객체의 닮은 정도에 대한 수치인 유사도(similarity)를 나타내는 척도이고, 분리도(separability)는 서로 다른 군집 간 객체들의 다른 정도를 나타내는 비유사도(dissimilarity)에 대한 척도를 말한다. 지금까지의 CVI 연구는 특정 데이터의 특성, 구조, 군집화 알고리즘에 적합한 지수를 개발하는 연구가 주를 이루고 있다. 본 연구에서는 선행연구에서 제안된 CVI들의 특징 및 장단점을 분석하고, 이를 바탕으로 적용의 유연성이 높은 CVI를 새롭게 제안하고자 한다. 이를 위하여 본 연구에서는 2000년 이후의 선행연구에서 소개된 CVI들을 선별하고 이들에 대하여 구체적으로 검토하였다.

본 연구의 선행연구로 적합하다고 판단된 최근 연구들은(Halkidi *et al.*, 2001; Liu *et al.*, 2013; Maulik and Bandyopadhyay. 2002; Saitta *et al.*, 2008)이다. 〈표 1〉은 이들 연구에서 성능비교에 사용된 14개 CVI들을 사용 빈도순으로 정리한 것이다. 이들은 대부분 비계층적 군집화(nonhierarchical clustering) 알고리즘에 적합하다고 알려져 있으며, 계층적 군집화(hierarchical clustering) 알고리즘에 적합한 CVI는 '*'로 표시하였다. 선별된 선행연구들에서 2회 이상 중복 사용된 CVI는 10개이며, 이를 다시 비계층적 알고리즘의 결과 검증에 적합하고, 응집도와 분리도를 동시에 고려한 지수들만을 추출하면 1번부터 7번까지의 7개 CVI가 된다. 본 연구에서는 7개 CVI들의 특성을 분석해 보고, 이들을 변형하여 새로운 개념이 포함된 CVI를 제안하고자 한다. 선별한 7개 CVI들의 정의, 수리모형 및 특성은 〈표 2〉와 같다. 표에서 알 수 있듯이, 모든 CVI들은 응집도와 분리도의 개념을 반영하여 계산한다는 공통점을 갖지만, 이를 어떻게 계산하느냐에 따라 지수의 정의나 수리모형은 모두 달라지게 된다.

기존 CVI의 응집도와 분리도는 인접성 함수에 의해 측정된다. 인접성 함수는 거리(distance)와 상관관계(correlation) 그리고 코사인(cosine) 유사도 등의 척도를 평균값, 중앙값, 최대값, 최소값, 로버스트(robust) 등의 계산 방법 중에서 어떤 방법으로 계산하였느냐에 따라 다양하게 정의된다. 각 척도에 따라 유사성을 평가하는 기준은 다음과 같다.

첫째, 거리척도는 거리가 멀수록(값이 클수록) 유사도가 낮아진다. 따라서 거리 값이 클 때 응집도는 나쁘고, 분리도는 좋다고 판단한다. 둘째, 상관관계 척도는 데이터의 속성간 상관성을 이용한 척도로, 그 값은 [-1, 1]의 범위에 존재한다. 이때 그 값이 '-1'과 '1'이면 데이터간에 음(-)과 양(+)의 유사성이 높음을 의미하고, 상관관계 값이 '0'이면 두 데이터 간에는 상관관계가 존재하지 않는 것을 의미하기 때문에 유사성이 낮다고 판단할 수 있다. 셋째, 코사

〈표 1〉 선행연구에서 소개된 CVI

	CVI	선행연구				사용 빈도
		Liu *et al.* (2013)	Saitta *et al.* (2008)	Maulik and Bandyopadhyay (2002)	Halkidi *et al.* (2001)	
1	Davies-Bouldin	√	√	√	√	4
2	Dunn	√	√	√	√	4
3	Calinski-Harabasz	√	√	√		3
4	Maulik-Bandyopadhyay	√	√	√		3
5	Silhoutte	√	√			2
6	Xie-Beni	√		√		2
7	SD	√			√	2
8	Modified Hubert Γ statistic	√			√	2
9	Root-mean-square std dev*	√			√	2
10	R-squared*	√			√	2
11	S_Dbw	√				1
12	Semi-partial R-squared*				√	1
13	Distance between two clusters*				√	1
14	Geometric		√			1

〈표 2〉 선별된 CVI의 정의와 수리모형 및 특징

	CVI에 대한 설명	
1	Davies-Bouldin (DB)	$DB_K = \frac{1}{K}\sum_{i=1}^{K}\max_{i=1,\cdots,K,\, i\neq j}\left\{\left(\sqrt{\frac{1}{n_i}\sum_{x\in c_i} d(x,z_i)^2}+\sqrt{\frac{1}{n_j}\sum_{y\in c_j} d(y,z_j)^2}\right)/d(z_i,z_j)\right\}$ n_i: 군집 C_i의 객체 수 z_i: 군집 C_i의 중심점
	- 군집 내 객체와 중심점을 이용한 군집 간 응집도의 합을 군집들의 중심점 간 분리도로 나눈 값 - 분리도에 대한 응집도의 가중비에 의해 결합 - 최소값을 주는 K가 최적 군집 수 (MIN C/S) - 부분군집과 임의형상에 민감	
2	Dunn (DU)	$DU_K = \frac{\min_{i,j=1,\cdots,K, i\neq j}\left\{\min_{x\in c_i, y\in c_j} d(x,y)\right\}}{\max_{i=1,\cdots,K}\left\{\max_{x_1,x_2\in c_i} d(x_1,x_2)\right\}}$
	- 서로 다른 군집에 속한 객체 간 거리를 이용한 분리도를 군집 내 객체 간 거리를 이용한 응집도로 나눈 값 - 응집도에 대한 분리도의 가중비에 의해 결합 - 최대값을 주는 K가 최적 군집 수 (MAX S/C) - 부분군집과 임의형상에 민감	

〈표 2〉 선별된 CVI의 정의와 수리모형 및 특징 (계속)

	CVI에 대한 설명	
3	Calinski-Harabasz (*CH*)	$CH_K = \frac{\sum_{i=1}^{K} n_i \cdot d(z_i, z_{tot})^2}{K-1} \cdot \frac{N-K}{\sum_{i=1}^{K}\sum_{x \in c_i} d(x, z_i)^2}$ N: 전체 데이터의 갯수 z_{tot}: 전체 데이터의 중심점
	- 군집 중심점과 전체 중심점을 이용한 군집 간 분리도를 군집 내 객체와 중심점을 이용한 군집 간 응집도로 나눈 값 - 응집도에 대한 분리도의 가중비에 의해 결합 - 최대값을 주는 K가 최적 군집 수 (MAX S/C) - 노이즈와 비대칭분포에 민감	
4	Maulik -Bandyopadhyay (*MB*)	$MB_K = (\frac{1}{K} \cdot \frac{\sum_{x \in C_1} d(x, z_1)}{\sum_{i=1}^{K}\sum_{x \in C_i} d(x, z_i)} \cdot \max_{i,j=1,\cdots,K} d(z_i, z_j))^p$ p: 서로 다른 군집 간 차이의 조절 계수 (주로 2로 선택)
	- 서로 다른 군집의 중심 간 거리를 이용한 분리도를 군집 내 객체와 중심점 간 거리를 이용한 응집도로 나눈 값 - 응집도에 대한 분리도의 가중비에 의해 결합 - 최대값을 주는 K가 최적 군집 수 (MAX S/C) - 밀집도와 임의형상에 민감	
5	Silhoutte (*S*)	$S_K = \frac{1}{N}\sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)}$ a_i: 군집 내 하나의 객체와 다른 객체들과의 평균거리 b_i: 한 객체와 다른 군집 내 객체들과의 평균거리
	- 서로 다른 군집에 속한 하나의 객체와 다른 모든 객체 간 평균거리를 이용한 분리도와 군집 내 한 객체와 다른 모든 객체 간 평균거리를 이용한 응집도의 차 - 분리도와 응집도의 차에 의해 결합 - 최대값을 주는 K가 최적 군집 수 (MAX S-C) - 부분군집과 임의형상에 민감	
6	Xie-Beni (*XB*)	$XB_K = \frac{\sum_{i=1}^{K}\sum_{x \in c_i} d(x, z_i)^2}{N \cdot \min_{i,j=1,\cdots,K} d(z_i, z_j)^2}$
	- 군집 내 객체와 중심점을 이용한 군집 간 응집도를 서로 다른 군집의 중심점 간 거리를 이용한 분리도로 나눈 값 - 분리도에 대한 응집도의 가중비에 의해 결합 - 최소값을 주는 K가 최적 군집 수 (MIN C/S) - 부분군집과 임의형상에 민감	
7	SD (*SD*)	$SD_K = a \cdot \frac{1}{K}\sum_{i=1}^{K} \frac{\|\sigma(C_i)\|}{\|\sigma(X)\|} + \frac{\max_{i,j=1,\cdots,K} d(z_i, z_j)}{\min_{i,j=1,\cdots,K} d(z_i, z_j)} \cdot \sum_{i=1}^{K}\left[\sum_{i=1}^{K} d(z_i, z_j)\right]^{-1}$ $\sigma(C_i)$, $\sigma(X)$: C_i와 데이터 집합 X의 군집 내 분산 벡터 a: 가중치, $\|X\|$: $(X^T, X)^{1/2}$
	- 군집 내 객체 간 상관계수를 이용한 응집도와 서로 다른 군집의 중심점 간 거리를 이용한 분리도의 합 - 응집도와 분리도의 합에 의해 결합 - 최소값을 주는 K가 최적 군집 수 (MIN C+S) - 부분군집과 임의형상에 민감	

인 유사도 척도는 각 데이터를 벡터(vector)로 표현하였을 때 벡터의 코사인 값을 측정한 것으로 '0'에서 '1'사이의 값을 갖는다. 이때 코사인 유사도 값이 '1'에 가까울수록 두 벡터 사이의 각도가 0°에 가깝기 때문에 데이터간 유사도는 높고, 그 값이 '0'에 가까우면 두 벡터 사이의 각도는 90°에 가깝기 때문에 유사도는 낮다고 판단한다.

한편, 각 척도는 평균값, 중앙값, 최대값, 최소값, 로버스트 등의 다양한 방법으로 계산되지만, 거리척도의 평균값을 이용한 인접성 함수가 주를 이루고 있다. 본 연구에서는 거리 척도를 이용한 CVI를 대상으로 한정하여 다룬다. 거리는 맨하탄(manhattan) 거리, 유클리디안(euclidean) 거리, 민코프스키(minkowski) 거리 등으로 측정된다. 군집분석에서는 유클리디안 거리가 많이 사용되므로, 본 논문에서는 유클리디안 거리를 사용한다.

일반적으로 CVI를 개발하기 위해서는 응집도와 분리도 중 한가지만을 고려할 것인지 동시에 고려할 것인지를 결정하고, 동시에 고려한다면 응집도와 분리도의 결합방식을 결정하는 것이 요구된다(김민호 · Ramakrishna, 2005). Liu *et al.*(2013)에 따르면 응집도와 분리도 중 한 가지만을 고려하면, 군집수에 따라 CVI값이 단조 증가하거나 감소하여 최적 군집수를 구하기 어렵다고 알려져 있다. 따라서 많은 연구에서 응집도와 분리도를 동시에 고려하고 있다.

본 연구에서는 응집도와 분리도를 동시에 고려한 CVI를 새롭게 제안하기 위하여, CVI를 개발하는 과정을 좀 더 세밀하게 분석한다. 기존 연구들을 분석한 결과, 응집도와 분리도를 동시에 고려한 CVI는 다음의 6단계를 거쳐 개발되는 것으로 확인되었다. ①한 군집의 군집별 응집도 계산방법을 결정한다. 군집별 응집도는 거리(Dist), 상관관계(CoR), 코사인 유사도(CoS)의 평균값(Av), 중앙값(Me), 최대값(Mx), 최소값(Mn), 로버스트(Rb) 등을 이용하여 계산한다. ②군집별 응집도 값의 평균(Avg), 최대(Max), 최소(Min), 분산(Var) 등을 이용하여 총 군집의 전체 응집도를 결정한다. ③서로 다른 2개의 군집 간 분리도의 계산방법을 결정한다. 군집간 분리도의 계산은 군집별 응집도 계산방법과 유사하나, 전체 데이터의 중심을 이용한 계산방법이 있다. ④총 군집의 전체 분리도를 어떻게 산출할 것인지 결정한다. 전체 분리도 계산방법은 2개 군집간 분리도의 평균, 최대, 최소, 분산 등이 있다. ⑤총 군집의 전체 응집도와 전체 분리도를 어떻게 결합하여 하나의 식으로 표현할 것인지를 결정한다. 전체 응집도와 전체 분리도의 결합은 응집도와 분리도의 곱(MulCS), 비(RatCS), 차(SubCS), 합(SumCS) 등이 있다. 마지막으로, ⑥완성된 CVI를 이용하여 계산한 지수 값의 최적 판단 기준을 결정한다. 이때, 최적 판단 기준은 주로 다섯번째 단계의 전체 응집도와 전체 분리도의 결합 방법에 의해 결정되게 된다. 예로, 거리척도로 계산된 전체 응집도와 전체 분리도를 응집도(분모)에 대한 분리도(분자)의 비로 결합하였다면, 분모값이 작고(응집도가 좋을수록), 분자값이 클수록(분리도가 좋을수록) 군집이 잘 되었다고 판단할 수 있기 때문에, 비의 최대값을 판단기준으로 결정한다.

새로운 CVI의 개발은 응집도나 분리도의 새로운 계산방법의 제안, 기존 응집도 또는 분리도 계산방법의 변형, 응집도와 분리도 조합의 변형 등에 의해 가능하다. 즉 각 단계에서 사용 가능한 응집도 및 분리도의 계산방법들의 조합을 새롭게 하거나, 새로운 응집도 및 분리도의 계산방법을 제안함으로써 가능하다. 본 연구에서는 〈표 2〉에서 언급한 7개의 기존 CVI들의 군집별 응집도 및 군집간 분리도를 계산방법과 전체 응집도와 전체 분리도의 결정 및 이들의

결합 방법이 무엇인지 살펴보고 이를 바탕으로 새로운 CVI를 제안하고자 한다. 〈표 3〉은 7개 CVI들의 단계별 인접성 계산방법을 나타낸 표이다. 표에서 거리척도를 이용하는 경우, 각 CVI의 상세한 구분을 위하여 거리의 기준에 따라 데이터와 데이터간의 거리는 'PP'로, 데이터와 중심점간의 거리는 'PC'로, 데이터와 대표점간의 거리는 'PM'로, 중심과 중심간의 거리는 'CC'로, 중심과 전체 데이터 중심간의 거리는 'CT'로 세분화하였다. 〈표 3〉의 분석결과, 기존 CVI들은 개별 응집도의 계산방법으로 객체와 군집 중심간의 평균거리를 많이 사용하였다. 또한, 응집도와 분리도의 인접성을 계산하는 방법이 매우 다양함에도 불구하고, 기존 연구들은 소수의 방법만을 사용하고 있었다. 본 연구에서는 '객체와 군집 중심간의 평균거리' 개념을 새로운 방법으로 변형하고, 기존 CVI의 성능을 향상시킬 수 있는 새로운 응집도 계산방법을 제안하고자 한다. 이를 위해 사용빈도가 높은 Dunn(DU), Calinski-Harabasz(CH),

〈표 3〉 기존 CVI들의 응집도와 분리도의 인접성 계산방법

응집도				분리도				결합	
개별	CVI	총	CVI	개별	CVI	총	CVI	방법	CVI
C_AvPP	⑤	C_Avg	③~⑦	S_AvPP	⑤	S_Avg	③⑤⑦	MulCS	
C_AvPC	①③④⑥	C_Max	①②	S_AvPC		S_Max	④	RatCS	①②③④⑥
C_AvPM		C_Min		S_AvPM		S_Min	①②⑥	SubCS	⑤
C_MePP	②	C_Var		S_MePP		S_Var		SumCS	⑦
C_MePC				S_MePC					
C_MePM				S_MePM					
C_MxPP				S_MnPP	②				
C_MxPC				S_MnPC					
C_MxPM				S_MnPM					
C_RbPP				S_RbPP					
C_RbPC				S_RbPC					
C_RbPM				S_RbPM					
C_AvCoR	⑦			S_AvCoR					
C_AvCoS				S_AvCoS					
C_MeCoR				S_MeCoR					
C_MeCoS				S_MeCoS					
C_MxCoR				S_MnCoR					
C_MxCoS				S_MnCoS					
C_RbCoR				S_RbCoR					
C_RbCoS				S_RbCoS					
				S_DistCC	①④⑥⑦				
				S_DistCT	③				
① DB		② DU		③ CH		④ MB		⑤ S	
⑥ XB		⑦ SD							

Davies-Bouldin(DB) 지수를 개선 대상으로 선정하였다. 왜냐하면 CVI의 사용빈도가 높다는 것은 다양한 데이터 및 알고리즘에 적용 가능함을 의미하기 때문이다. 선별된 CVI의 특징은 새로운 개념이 포함되어 개발된 지수들과 비교하여 다음 제4장에서 순차적으로 설명하도록 한다.

Ⅲ. Fuzzy K-means 군집분석

군집화 알고리즘은 사전에 군집 수를 정하지 않고 단계적으로 서로 다른 군집화 결과를 제시해 주는 계층적 군집화와 사전에 군집 개수를 정하고 각 데이터를 군집에 배정하는 비계층적 군집화가 있다(Xu and Wunsch, 2005). 계층적 군집화 알고리즘은 사전에 군집수를 입력할 필요는 없지만 알고리즘을 종료해야 할 기준을 결정해야 하는 문제점과 함께 대용량 데이터에서 처리속도가 느리다는 단점이 존재한다(이신원, 2012; 신경석 · 김재윤, 2011; 이신원 · 안동언 · 정성종, 2004). 비계층적 군집화 알고리즘은 계층적 군집화 알고리즘에 비해 계산 소요시간 측면에서 유리하기 때문에 다양한 분야에서 많이 사용되고 있으며 다양한 기법들이 소개되어 왔다(전치혁, 2012). 비계층적 군집화 알고리즘으로 사용빈도가 높은 대표적인 알고리즘으로는 K-means, K-medoids(PAM, CLARA, CLARANS), Fuzzy K-means, 모형기반 알고리즘 등이 있다(Halkidi *et al.*, 2001).

K-means 알고리즘은 이들 중에서 간단하면서 구현이 용이하여 여러 문제에 쉽게 적용된다. MacQueen (1967)에 의해 제안된 이 알고리즘은 각 데이터(이하, 군집화 알고리즘에서 일반적으로 사용하는 '객체(object)'라는 용어를 사용하기로 함)와 K개 군집의 중심좌표(centroid)와의 거리를 계산하고, 객체와 중심좌표간 거리가 가장 가까운 군집에 각 객체를 배정하는 반복적 알고리즘이다. K-means 알고리즘은 K개의 군집에 객체가 속하거나 속하지 않는 이산적인 값으로 나타내므로, 강한 군집화(hard clustering) 방법이라고도 불린다(허경용 · 서진석 · 이임건, 2011). 이 방법은 군집들이 중첩되어 나타나거나 노이즈가 포함된 경우에는 대처하기 어렵다는 단점이 있다.

Fuzzy K-means 알고리즘(또는 Fuzzy c-means 알고리즘이라고 부르기도 함)은 유연성이 높은 알고리즘으로 K-means 알고리즘과 유사하나, 하나의 객체가 여러 군집에 속할 가능성을 허용하는 확률 또는 이를 확장한 퍼지(fuzzy) 개념을 도입한 알고리즘이다(Bezdek, 1981). 약한 군집화(soft clustering) 방법이라고 불리는 이 알고리즘은 Raspini(1969)가 군집화에 처음 퍼지 분할(fuzzy partition)을 소개하고, 이를 이용하여 Dunn(1973)이 최초의 퍼지 군집화 알고리즘(fuzzy clustering algorithm)을 제안하였다. 이후, Bezdek(1981)에 의해 Dunn의 알고리즘이 일반화되어 현재 Fuzzy K-means 알고리즘은 일반적으로 Bezdek의 방법을 일컫는다. 본 연구에서는 Fuzzy K-means 알고리즘을 이용하여 군집분석을 실시하고, 그 결과에 대한 유효성을 검증할 수 있는 여러 CVI들을 제안하고자 한다. 따라서 Fuzzy K-means 알고리즘의 개념과 절차를 구체적으로 설명하면 다음과 같다(전치혁, 2012).

Fuzzy K-means 알고리즘은 다음의 최적화 문제를 해결하는 발견적 해법중 하나이다. 즉, P_{ij}를 객체 i가 군집 j에 속할 확률이라 할 때, 이 알고리즘은 다음과 같은 $\{P_{ij}\}$를 구하는 최적화 문제로 정식화된다.

$$\text{Min. } Z=\sum_{i=1}^{n}\sum_{j=1}^{K}P_{ij}^{m}d(x_i,c_j) \quad (3\text{-}1)$$

Subject to

$$\sum_{j=1}^{K}P_{ij}=1, \qquad i=1,\cdots,n \quad (3\text{-}2)$$

$$\sum_{i=1}^{n}P_{ij}>0, \qquad j=1,\cdots,K \quad (3\text{-}3)$$

$$P_{ij}\in[0,1], \quad i=1,\cdots,n; \quad j=1,\cdots,K \quad (3\text{-}4)$$

이때, 목적함수의 m은 퍼지상수(fuzziness index)로 불리는 1보다 큰 상수로서 퍼지정도를 나타낸다. m이 1에 가까울수록 강한 군집화 방법인 K-means 알고리즘에 가깝고, 반대로 큰 값을 가질수록 각 객체가 비슷한 확률로 군집에 배정되게 된다. 통상 $m=2$가 사용되고 있다. 한편, c_j는 군집 j의 중심좌표를 나타내는데 K-means의 경우와는 약간 달리 P_{ij}^{m}을 가중치로 사용하여 산출된다. 〈표 4〉는 Fuzzy K-means 알고리즘의 구체적인 절차를 단계별로 보인 것이다. 여기서, 단계 1에서 요구하는 초기 군집 개수 K는 일반적으로 문제 특성에 맞추어 임의로 선정하고, 단계 3에서 사용되는 P_{ij}는 식 (3-2)를 라그랑지안 방법과 편미분을 이용하여 유도할 수 있다. 구체적인 과정은 전치혁(2012)을 참고하면 된다.

〈표 4〉 Fuzzy K-means 알고리즘의 절차

단계 1:	초기 K개의 군집을 임의로 정한다. $P_{ij}=\begin{cases}1, & \text{객체 } i\text{가 군집 } j\text{에 속하면}\\ 0, & \text{그렇지 않으면}\end{cases}$
단계 2:	각 군집의 중심좌표를 다음 식에 의거하여 산출한다. $c_j=\dfrac{\sum_{i=1}^{n}P_{ij}^{m}x_i}{\sum_{i=1}^{n}P_{ij}^{m}}$
단계 3:	객체 i가 군집 j에 속할 확률 P_{ij}를 다음 식에 의거하여 산출한다. $P_{ij}=\dfrac{[d(x_i,c_j)]^{-1/(m-1)}}{\sum_{a=1}^{n}[d(x_i,c_a)]^{-1/(m-1)}}$
단계 4:	객체 i를 P_{ij}가 가장 큰 군집 j에 배정하여 군집결과를 얻는다. 이전의 군집결과와 동일하면 알고리즘을 종료하고, 그렇지 않으면 단계 2로 간다.

Ⅳ. SVDD 기반의 새로운 CVI

4.1 SVDD

서포트 벡터 데이터 표현(SVDD)은 기계학습 알고리즘(machine learning algorithm) 중 최근에 등장하여 여러 가지 문제에서 우수한 해결능력을 보여주는 비선형 SVM(support vector machine)의 응용 형태이다. 비선형 SVM은 기본적으로 두 범주를 갖는 객체들을 나눌 수 있는 초평면(hyper-plane) 함수를 이용하여 두 개의 클래스(군집)를 분류하는 방법이다(Vapnik, 1979). 일반적으로 입력 공간

(input space)에서 초평면 함수를 찾는 것은 쉽지 않다. 이를 해결하기 위하여 비선형 SVM은 입력 공간의 데이터(또는 객체)를 더 높은 차원의 특징 공간(feature space)으로 매핑(mapping)시킨다. 이때, 낮은 차원의 비선형 함수가 높은 차원으로 매핑되면 선형함수로 근사화되고, 그 결과 쉽게 특징 공간의 최적 초평면 함수를 구할 수 있게 된다. 이렇게 구한 특징 공간의 최적 초평면 선형함수는 입력 공간에서의 최적 비선형 함수라 할 수 있다.

SVDD는 분류 대상이 되는 목적 클래스(target class)에 속한 데이터만을 이용하여 학습을 수행함으로써 단일 클래스 분류 문제를 해결한다. SVDD의 기본 원리는 특이점(outlier)들을 검출하여 주어진 목적 데이터를 대부분 포함하고 특이점을 가장 적게 포함하는 원형의 경계선을 찾는 것이다(Tax and Duin, 2004). SVDD를 이용하여 입력 공간의 데이터를 특징 공간으로 매핑하면, 넓게 퍼져 있던 입력 공간의 데이터들이 특징 공간에서 가깝게 위치한다. 따라서 입력 공간에서는 데이터 전체를 둘러싸는 큰 타원형의 초평면이 그려지는 반면, 특징 공간에서는 반지름이 최소화된 원형의 초평면(초구)이 만들어진다. 이러한 방법을 통해 분류의 오류를 최소화하고, 특이점을 잘 탐지한다.

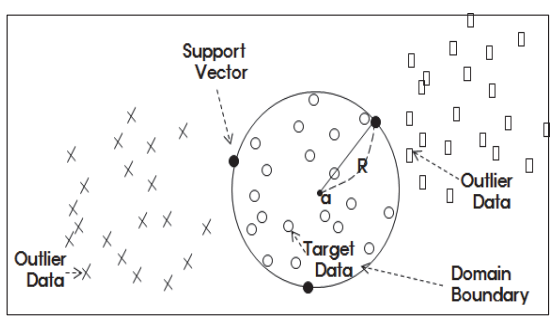
SVDD에 관한 기존 연구들은 SVDD 이론 자체에 관한 연구와 SVDD의 적용 확대에 관한 연구로 구분할 수 있다. 또한, 적용 확대에 관한 연구들은 통계학, 정보학, 전자 전기학 등 다양한 학문 분야에서 이루어지고 있다. 본 연구에서는 SVDD 개념을 이용하여 새로운 CVI를 개발하는 것이므로, SVDD 개념이 군집분석 문제에 활용된 최근 연구들만을 간략하게 살펴보고자 한다. Niazmardi, Homayouni and Safari(2013)은 SVDD 개념을 이용하여 군집의 중심을 산정하는 Fuzzy K-means 군집화 알고리즘을 새롭게 제안하였다. Xu, Yao and Nl(2011)은 K-means 알고리즘과 SVDD 개념을 이용한 고장 검출 방법을 제안하였다. Chang, Kim, Choi and Choi(2007)은 SVDD가 대용량 데이터를 다룰 때 문제되는 계산시간문제를 해결하기 위하여 군집화 기법을 이용한 새로운 SVDD를 제안하였다. Ji, Liu, Wu and Liu(2008)은 인자표현(gene expression) 데이터의 군집화를 위하여 SVDD를 적용하였다. 이상에서 살펴본 바와 같이, 군집분석의 CVI에 SVDD 개념이 접목된 연구는 아직까지 이루어지지 않은 것으로 조사되었다.

4.2 개발원리

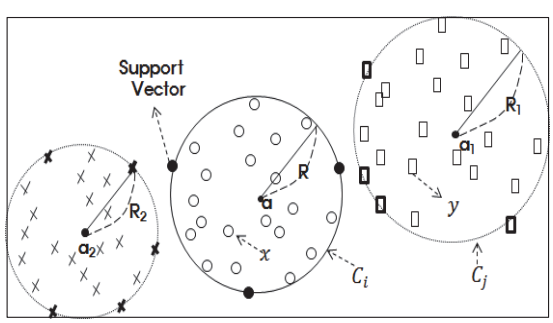
SVDD는 단일 클래스 분류기법이지만, 다중 클래스 분류에도 응용이 가능하다. 즉, 데이터 집합 안에 여러 개의 군집이 존재할 때 각각 대상 군집만을 고려하여 단일 클래스 문제처럼 학습을 한 뒤 얻어진 데이터 표현의 경계를 통합하여 여러 개의 군집을 지닌 분류문제에 활용할 수 있다(송동성 · 김표재 · 장형진 · 최진영, 2007). 〈그림 1〉은 3개의 군집으로 이루어진 데이터 집합을 SVDD 개념으로 분류한 것이다. 전체 데이터는 사각형, 원형, 십자형의 특성에 따라 3개 군집으로 분류된다. 다중 클래스 분류에 SVDD를 이용하기 위해서는 먼저 하나의 데이터와 나머지 데이터를 분류하기 위한 학습을 수행하고, 미분류된 데이터를 다시 같은 방법으로 반복 학습하여 분류하면 된다. 즉, 〈그림 1〉의 예에서는 먼저 원형 데이터를 기준으로 원형과 원형이 아닌(사각형 또는 십자형) 데이터를 분류하여 초구의 경계를 찾는다. 원형 특성 데이터에 대한 학습이 완료되면, 이와 같은 방식으로 사각형과 십자형의 특성 데이터에 대한 학습을 순차적으로 시행하여 서로 다른

특성의 데이터별로 초구 경계를 찾는다.

〈그림 2〉는 〈그림 1〉의 SVDD 개념을 반영하여 다중 클래스로 군집화한 결과를 형상화 한 것이다. 즉, 원형 데이터의 경계를 하나의 군집(C_i)이라 하면, 구의 중심(a)은 군집의 중심점(z_i)이 되고, 구의 학습 데이터(x_k)와 같이 군집에 속해 있는 목적 데이터는 군집(C_i)내의 데이터(x)가 된다. 따라서 특이점(예: 사각형)을 서로 다른 i, j에 대해 다른 군집(C_j)에 속한 임의의 데이터(y)로 생각할 수 있다.



〈그림 1〉 다중 클래스 분류에서 SVDD



〈그림 2〉 SVDD를 반영한 군집화

본 연구에서는 이러한 아이디어를 바탕으로 거리척도를 이용한 CVI의 응집도에 SVDD 개념을 반영함으로써 SVDD에 의해 그려지는 구를 개별군집으로 판단할 수 있다고 생각하였다. 일반적으로 거리척도를 이용한 CVI의 응집도는 그 값이 작을수록 응집도가 좋다고 판단하기 때문에, 초구의 반지름이 작을수록 응집도가 좋다고 해석한다. SVDD 기반의 응집도 계산 개념이 어떻게 기존 수리모형에 접목되어 새로운 CVI로 만들어지는지에 대해서는 다음 절에서 자세히 기술한다.

4.3 SVDD기반 CVI Ⅰ: SVDU

DU 지수는 Dunn(1974)이 개발하였고 현재까지 군집화의 유효성을 판단하는 평가지수로 많이 활용되고 있으며, 수리모형은 〈표 2〉에 제시되어 있다. DU 지수를 응집도와 분리도로 분해하고, 이들의 결합에 의해 최적 군집수를 판단할 수 있는 기준을 나타내면 〈표 5〉와 같다. DU 지수를 응집도와 분리도의 개념으로 분해하면 개별 응집도는 군집 내 객체간의 거리 척도로 정의되고, 총 응집도는 개별 응집도의 최대값을 이용하여 산출한다. 개별 분리도는 서로 다른 군집의 객체간 거리 척도로 정의하고, 총 분리도는 개별 분리도의 최소값으로 한다. 최종 CVI는 응집도에 대한 분리도의 비로 정의하고, 그 비가 최대인 군집수가 최적 군집수라고 판단한다.

DU 지수는 개념이 명료하고 계산이 용이하여 군집의 유효성을 손쉽게 판단할 수 있다는 장점을 가지고 있다. 그러나 응집도와 분리도를 계산할 때 최대값 또는 최소값을 사용하여 객체들의 거리를 계산하기 때문에 노이즈 특성을 지닌 데이터들의 군집화에 불리한 것으로 알려져 있다. 뿐만 아니라 임의형상이나 부분군집 등 공간적 구조가 복잡한 형태의 데이터도 민감하다(Liu *et al.*, 2013).

〈표 5〉에서 보인 각 군집의 개별 응집도 계산을 SVDD의 초구 경계면 추출 방법으로 대체하면 K개의 군집이 주어질 때 SVDD 개념이 반영된 새로운

〈표 5〉 Dunn(DU) 지수의 분해

응집도(C)		분리도(S)		결합 및 판단기준	
개별	총	개별	총	결합	기준
$\max_{x_1,x_2\in C_i} d(x_1,x_2)$	Max.	$\min_{x\in C_i,y\in C_j} d(x,y)$	Min.	S/C	최대값
군집 내 객체의 거리		군집 간 객체의 거리			

Dunn 지수(이하, SVDU라고 부르기로 한다)의 수리모형이 도출된다. 이는 식 (4-1)에 표현되어 있다. 식 (4-1)의 $J(C_i)$는 i번째 군집의 서포트 벡터(SV_i)에 의해 구해진 평균 반지름을 이용한 i번째 구의 지름을 의미하고, $R_{p_i}^2$는 i번째 구의 중심(a_i)과 SV_i의 한 점(p_i)사이의 거리 제곱을 의미하며, K는 커널함수를 의미한다. 이때 i번째 군집의 서포트 벡터(SV_i)는 구의 경계면을 그리는 방향에 따라 반지름 값을 달리하기 때문에, 반지름을 구하기 위해서는 i번째 군집의 여러 반지름들의 평균값을 구해야 한다. i번째 군집의 반지름들의 평균값을 구하는 방법은 SV_i의 한 원소인 p_i로 정해지는 반지름을 R_{p_i}라고 하였을 때, SV_i의 모든 원소에 대해 $R_{p_i}^2$의 모든 합을 SV_i 집합의 개수($|\{SV_i\}|$)로 나누어 계산한다. 이렇게 구한 반지름들의 평균값으로부터 $J(C_i)$를 계산할 수 있다.

$$SVDU_K = \frac{\min_{i,j=1,\cdots,K, i\neq j}\left\{\min_{x\in C_i,y\in C_j} d(x,y)\right\}}{\max_{i=1,\cdots,K}\{J(C_i)\}} \quad (4\text{-}1)$$

$$J(C_i) = 2\left(\sum_{p_i\in SV_i}\frac{R_{p_i}^2}{|\{SV_i\}|}\right)$$

$$R_{p_i}^2 = K(S_{p_i},S_{p_i}) - 2\sum_k \alpha_k K(S_{p_i},X_k) + \sum_{k,l}\alpha_k\alpha_l K(X_k,X_l)$$

4.4 SVDD기반 CVI II: SVCH

Calinski and Harabasz(1974)가 제안한 CH 지수도 DU 지수와 함께 현재까지 CVI의 성능비교 분석 대상으로 많이 활용되는 평가지수이며, 수리모형은 〈표 2〉에 제시되어 있다. 〈표 6〉은 CH 지수를 응집도와 분리도로 분해하고, 이들의 결합에 의해 최적 군집수를 판단할 수 있는 기준을 나타낸 것이다. CH 지수를 응집도와 분리도의 개념으로 분해하면 개별 응집도는 군집 내 중심점과 객체들의 거리의 총합으로 정의되고, 총 응집도는 개별 응집도들의 합에 의하여 산출한다. 개별 분리도는 한 군집의 중심점과 전체 데이터의 중심점의 거리로 계산하고, 총 분리도는 개별 분리도들의 합으로 계산한다. 최종 CVI는 응집도에 대한 분리도의 비로 정의하고, 그 비가 최대인 군집수가 최적 군집수라고 판단한다.

CH 지수에서 응집도는 군집의 중심점과 객체(점) 간의 거리를 이용하여 계산하고, 분리도는 군집의 중심점과 전체 데이터의 중심점을 이용하여 계산한다. 따라서 노이즈가 발생한 데이터의 군집 결과를 분석하면, 분리도에 비해 응집도가 상대적으로 많이 증가하게 된다. 즉, CH 지수는 노이즈 특성을 지닌 데이터에서 민감하다고 말할 수 있다. 또한 이 지수에서 응집도는 중심점과 모든 객체의 거리를 이용하여 계산하고, 분리도는 각 군집의 크기에 비례한 가중치를 부여한다. 따라서 CH 지수를 계산할 때에는

〈표 6〉 Calinski-Harabasz (CH) 지수의 분해

응집도(C)		분리도(S)		결합 및 판단기준	
개별	총	개별	총	결합	기준
$\Sigma_{x \in C_i} d(x, z_i)^2$	$\sum_i$	$n_i \cdot d(z_i, z_{tot})^2$	$\sum_i$	S/C	최대값
중심점과 객체 간 거리		중심점 간 거리			

군집의 객체 수 및 크기가 영향을 미칠 수 있다. 결국, CH 지수는 비대칭 분포 특성을 지닌 데이터를 군집분석 결과의 유효성 검증에 취약할 수 있다(Liu *et al.*, 2013).

〈표 6〉에서 보인 각 군집의 개별 응집도를 SVDD 개념으로 대체하면 K개의 군집이 주어질 때 SVDD 개념이 반영된 새로운 Calinski-Harabasz 지수(이하, SVCH라고 부르기로 한다)의 수리모형이 식 (4-2)와 같이 도출된다.

$$SVCH_K = \frac{\sum_{i=1}^{K} n_i \cdot d(z_i, z_{tot})^2}{K-1} \cdot \frac{N-K}{\sum_{i=1}^{K} J(C_i)} \quad (4\text{-}2)$$

$$J(C_i) = 2\left(\sum_{p_i \in SV_i} \frac{R_{p_i}^2}{|\{SV_i\}|} \right)$$

$$R_{p_i}^2 = K(S_{p_i}, S_{p_i}) - 2\sum_k \alpha_k K(S_{p_i}, X_k) + \sum_{k,l} \alpha_k \alpha_l K(X_k, X_l)$$

4.5 SVDD기반 CVI III: SVDB

Davies and Bouldin(1979)이 제안한 DB 지수는 응집도와 분리도의 상대적인 변화를 고려하는 평가지수이며, 수리모형은 〈표 2〉에 제시되어 있다. 일반적으로 응집도와 분리도를 동시에 고려하는 CVI들은 제2장에서 언급하였던 CVI 설계 6단계 중 1단계부터 4단계와 같이 군집별 응집도와 분리도를 계산한 후 전체 데이터(군집)의 총 응집도와 총 분리도를 산출한다. 그 후 산출된 총 응집도와 총 분리도를 결합하는 방법에 따라 CVI의 최적 군집수 판단기준을 결정한다. 그러나 DB 지수는 개별 군집의 응집도와 분리도를 계산하여 하나의 비율 값으로 결합한 후, 전체 데이터에서 조합 가능한 모든 두 개의 군집 쌍들의 결합 값들을 상대적으로 비교하여 최적 군집수를 찾아낸다. 〈표 7〉은 DB 지수를 응집도와 분리도로 분해하고, 이들의 결합에 의해 최적 군집수를 판단할 수 있는 기준을 나타낸 것이다. 개별 결합의 $C_{\cdot i}$는 i번째 군집 C_i의 개별 응집도를 의미하고, $S_{\cdot ij}$는 군집 C_i와 군집 C_j의 개별 분리도를 의미한다.

DB 지수는 계산이 빠르고 쉬우며 일관성 있는 값을 제시해 준다는 장점을 가지고 있다. 또한, 각 군집의 유효성을 개별 군집의 형태(응집도)와 군집 간의 위치(분리도)가 반영된 상대적 영향력으로 판단하기 때문에, 신뢰도 측면에서 우수하다고 알려져 있다(Saitta *et al.*, 2008). 그러나 분리도는 결합이 가능한 모든 경우 수의 군집 쌍들의 분리도를 모두 계산한 후에, 결합 값을 가장 유리하게 하는 군집 쌍을 선택하여 계산하기 때문에 부분군집의 특성을 지닌 데이터에 민감할 수 있다. 이는 두 개의 군집을 서로 다른 군집으로 간주할 때보다 하나의 군집으로 간주할 때, 전체 데이터 측면에서 분리도가 더 좋다

〈표 7〉 Davies-Bouldin (DB) 지수의 분해

개별 응집도(C.)	개별 분리도(S.)	결합		판단기준
		개별	총	
$\sqrt{\frac{i}{n_i}\sum_{x\in C_i} d(x,z_i)^2}$	$d(z_i,z_j)$	$\max_{i\neq j}\left(\frac{C_{\cdot i}+C_{\cdot j}}{S_{\cdot ij}}\right)$	$\sum_i$	최소값
중심점과 객체 간 거리	중심점 간 거리			

고 판단되기 때문에 각 부분군집을 하나의 독립된 군집으로 판단하기 어렵기 때문이다.

$$SVDB_K = \frac{1}{K}\sum_{i=1}^{K}\max_{i,j=1,\cdots,K,\ i\neq j}\{(J(C_i)+J(C_j))/d(z_i,z_j)\} \quad (4\text{-}3)$$

$$J(C_i) = 2\left(\sum_{p_i\in SV_i}\frac{R_{p_i}^2}{|\{SV_i\}|}\right)$$

$$R_{p_i}^2 = K(S_{p_i},S_{p_i}) - 2\sum_k \alpha_k K(S_{p_i},X_k) + \sum_{k,l}\alpha_k\alpha_l K(X_k,X_l)$$

새로운 DB 지수는 응집도 계산방법을 제외한 다른 설계원리를 DB 지수의 동일하게 사용한다. K개의 군집이 주어질 때 새로운 DB 지수(이하, SVDB라고 부르기로 한다)의 수리모형은 식 (4-3)과 같다. SVDB 지수에서 i번째 군집의 응집도와 분리도를 결합한 값은 i번째 군집을 기준으로 서로 다른 군집 C_i와 C_j의 SV_i와 SV_j에 의해 구해진 평균 반지름을 이용한 응집도의 합을 두 군집 간 분리도로 나누어 결합 값을 계산하고, 이렇게 나온 $K-1$개의 값 중에서 가장 큰 값을 군집 C_i의 결합 값(응집도와 분리도가 함께 계산된 값)으로 한다. K개의 군집별로 결합 값이 주어지면 평균을 이용하여 전체 데이터의 군집화 유효성을 판단하는 하나의 지수 값을 산출한다.

Ⅴ. 성능 분석

5.1 실험문제 및 설계

CVI의 성능은 어떤 특성을 갖는 데이터를 사용하여 분석할 것이며, 어떤 방법으로 그 성능을 보일 것인지가 중요하다. 본 연구에서는 새로운 CVI들의 특성을 살피기 위하여 기존 연구자들이 사용하였던 벤치마크 데이터를 이용하여 실험문제의 주어진 군집수를 정확히 맞추는지 확인하는 방식으로 기존 CVI들(DU, CH, DB 지수)과 새로운 CVI(SVDU, SVCH, SVDB 지수)들의 성능을 분석한다. 실험은 차원에 따라 2차원과 고차원 실험문제로 분리하여 수행하였다. 2차원 실험문제는 새로운 CVI들이 기존 CVI들의 단점을 보완하였는지를 평가하기 위하여, 기존 CVI들이 쉽게 최적의 군집화 결과를 판정하지 못했던 임의형상, 부분군집, 노이즈, 비대칭 분포 등의 특성을 지닌 문제[1]로 선정하였다. 고차원 실험

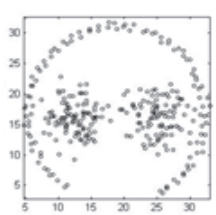
1) Speech and Image Processing Unit, University of Eastern Finland, "Clustering Dataset," http://cs.joensuu.fi/sipu/datasets.

문제는 2차원 문제보다 좀 더 복잡한 형상을 갖는 실험문제를 UCI Machine Learning Repository[2] 에서 선별하였다.

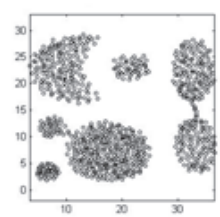
〈표 8〉은 실험에 사용한 2차원 데이터의 실험문제 번호, 이름, 데이터 수, 알려진 최적 군집수, 그리고 데이터의 형상 특징이다. 이들의 시각적 형상은 〈그림 3〉에서 확인할 수 있다. 〈표 9〉는 고차원 데이터의 실험문제 번호, 이름, 데이터 수, 차원 수, 그리고 알려진 최적 군집수이다. 고차원 실험문제의 경우 2차원 평면상에 사영된 형상을 이용하여, 원래 데이터의 형상을 추측하였으며 2차원 평면상에 사영된 형상을 〈그림 4〉에 제시하였으나, P12의 형상은

〈표 8〉 2차원 실험문제의 정보

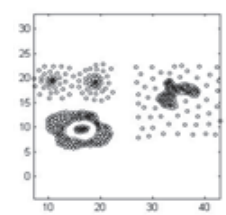
문제번호	문제이름	데이터 수	군집수	형상 특징
P1	Path-based 1	300	3	비대칭+임의형상
P2	Aggregation	788	7	부분군집
P3	Zahn's Compound	399	6	비대칭+부분군집
P4	Path-based 2	312	3	부분군집+임의형상
P5	A.K. Jain's Toy	373	2	비대칭+임의형상
P6	Flame	240	2	노이즈+부분군집



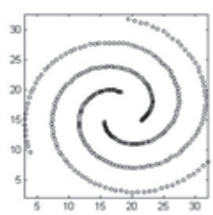
(a) P1: Path-based 1



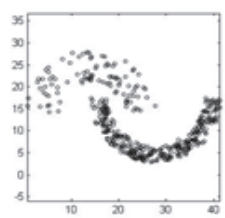
(b) P2: Aggregation



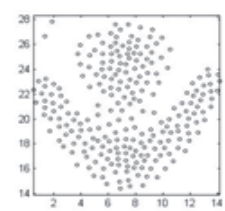
(c) P3: Zahn's Compound



(d) P4: Path-based 2



(e) P5: A.K. Jain's Toy



(f) P6: Flame

〈그림 3〉 2차원 실험문제의 형상

2) Center for Machine Learning and Intelligent Systems, UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.html.

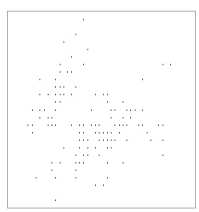
찾을 수 없었다. 참고로 고차원 문제는 데이터를 정규화(normalization)함으로써 데이터의 균일성을 고려하였다. 이는 사용한 고차원 문제 중에 차원(속성)에 따라서 데이터 스케일(scale)의 차이를 많이 보여 정규화를 이용한 데이터의 스케일을 통일할 필요가 있다고 판단하였기 때문이다.

본 연구는 Fuzzy K-means 알고리즘으로 군집분석한 결과를 검증함으로써 앞에서 설명한 6가지 CVI의 성능을 비교하였다. 이에 대한 구체적인 절차는 다음과 같다. 먼저 비계층적 군집화 알고리즘은 군집수가 사전에 지정되어야 하므로, 문제별로 $K_{\min}=2$부터 $K_{\max}=10$까지 순차적으로 군집수를 변화시키면서 각 K에 대하여 Fuzzy K-means 알고리즘으로 군집분석을 실행하면서 데이터를 군집화한다.

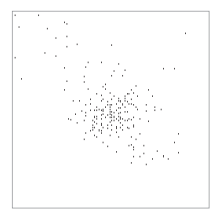
두 번째 단계에서는 도출된 군집화 결과를 바탕으로 각 CVI 지수 값을 계산한다. 이때, SVDD가 반영된 새로운 CVI의 지수 값들을 계산하기 위해서는

〈표 9〉 고차원 실험문제의 정보

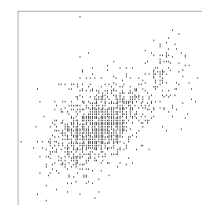
문제번호	문제이름	데이터 수	차원 수	군집수
P7	Iris	150	4	3
P8	Thyroid	215	5	2
P9	Yeast	1484	8	10
P10	Glass	214	9	7
P11	Wine	178	13	3
P12	Ecoil	336	8	8



(a) P7: Iris

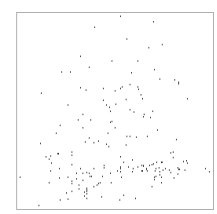


(b) P8: Thyroid



(c) P9: Yeast

(d) P10: Glass



(e) P11: Wine

〈그림 4〉 고차원 실험문제의 형상

커널함수에 따른 매개변수 값과 조절상수가 필요하다. 특히 커널함수는 문제의 종류 및 특성에 영향을 많이 받기 때문에, 커널함수를 선정하고 그에 맞는 매개변수를 결정하는 것은 SVDD를 이용하는데 있어 중요한 문제 중의 하나이다. 그러나 각 문제에 적합한 커널함수를 찾는 것은 또 다른 연구주제가 될 수 있을 만큼 다뤄야 할 연구내용이 많다. 따라서 본 연구는 커널함수에 따른 성능비교를 보인 선행연구에서 우수하다고 알려진 RBF 커널함수를 사용하고자 한다(안현철・김경재・한인구, 2005). RBF 커널함수의 매개변수 값은 Tay and Cao(2006)의 제안에 따라, 시그마는 1에서 10사이의 값을 1의 간격으로, 에러 수준은 0.01에서 0.1사이의 값을 0.01의 간격으로 세분화하여 시행착오적 방법에 의하여 실험문제별로 최적의 값을 선정하였다.

세 번째 단계는 군집화 결과를 이용하여 각 CVI의 수리모형에 의해 지수 값을 산출한다. Fuzzy K-means 알고리즘은 초기해에 민감하기 때문에, 매회 군집화 결과에 따라 CVI 값이 달라질 수 있다. 따라서 100회 반복 실행한 평균값을 이용하여 군집별로 CVI 값을 산출한다. 마지막으로, CVI별로 정해진 판단기준(최대값 또는 최소값)에 의해 실험문제별로 각 지수가 판단한 최적해(군집수)를 결정하고, 실험문제에서 주어진 군집수와 각 지수가 판단한 군집수를 비교하여 정확하게 군집수를 맞추었느냐로 CVI의 성능을 평가한다.

5.2 실험결과

CVI의 성능은 주어진 문제의 군집수를 얼마나 정확하게 판단하느냐로 보일 수 있다. 그런데 일반적인 군집분석 문제는 군집수가 사전에 결정되어 있지 않고 정답도 없으므로, 일반적인 문제를 이용하여 군집수를 정확하게 추정하였는지를 판단하는 것은 불가능하다. 따라서 CVI와 관련한 연구들은 〈표 8〉 및 〈표 9〉와 같이 사전에 최적 군집수가 알려진 벤치마킹 문제를 이용하고, 문제의 최적 군집수에 대한 추정의 정확도로 CVI의 성능을 평가한다(Liu *et al.*, 2013; Saitta *et al.*, 2008). 본 연구에서도 기존 연구들이 사용했던 방법으로 CVI의 성능을 평가하기로 한다. 이러한 실험결과를 통해 CVI의 성능을 일반화시키는 것은 또 다른 연구주제라고 생각되며, 이에 대해서는 마지막 장에서 다시 언급하기로 한다.

〈표 10〉은 분석 대상이 되는 6개 CVI들이 2차원 벤치마킹 문제에서 군집수를 정확하게 추정하였는지를 정리한 결과이다. 각 문제의 CVI별 구체적인 지수 값은 이수현(2015)의 연구를 참조할 수 있고, 여기에서는 지면의 제약상 생략한다. 〈표 10〉에서 각 문제별로 표시된 값은 'CVI가 제안한 군집수 - 알려져 있는 최적 군집수'를 계산한 것이다. 따라서 그 값이 '0'이면 CVI가 정확하게 최적 군집수를 추정한 것이고, 0에 가까울수록 군집수 추정의 정확도가 높다고 판단할 수 있다. CVI 지수 값의 특성에 따라 부호는 정확도에 영향을 미치지 않으며, 각 셀의 절대값 크기로 정확도를 판단한다. 또한, '군집수 추정 정확도'는 지수별로 '0'값의 개수이며, 이는 군집수를 정확하게 추정한 문제 수를 뜻하기 때문에 이 값이 클수록 추정의 정확도가 높다고 해석할 수 있다. 마지막으로 '오차 절대 값의 총합'은 지수별로 모든 문제들의 군집수 추정의 절대오차를 합산한 값으로, 이 값이 작을수록 추정의 정확도가 상대적으로 높다고 해석한다.

각 문제별 실험결과를 상세하게 설명하면 다음과 같다. 비대칭과 임의형상의 특성을 지닌 P1에 대하여 모든 CVI가 정확한 군집수를 추정하지 못하였다. 그러나 SVDD 개념의 반영 전의 DU, CH, DB

지수에 비하여 반영 후의 SVDU, SVCH, SVDB 지수에서는 정답과의 군집수 오차는 감소하였다. 이는 기존 CVI들의 비대칭 및 임의형상 특성에 대한 단점이 SVDD 개념의 반영에 의해 보완되었기 때문인 것으로 판단된다. 부분군집의 특성을 지닌 P2에서는 모든 지수들이 군집수를 맞추는데 실패하였으며, SVDD 반영 이후 오히려 DU, CH 지수의 최적 군집수와의 오차가 더욱 증가하였다. 이는 SVDD 개념이 부분군집 특성 데이터에 유리하지는 않기 때문인 것으로 판단된다.

부분군집과 비대칭 특징이 뚜렷한 문제 P3에서는 모든 지수들이 정확한 군집수 추정에 실패하였다. 이 문제는 앞에서도 언급한 바와 같이 시각적인 판단으로도 문제에서 제시된 최적 군집수 6을 판단하기 쉽지 않아 매우 어려운 문제로 생각된다. 특히, 문제 P3은 SVDD 개념을 반영함에 따라 오히려 정답과의 오차가 증가하였다. 이는 문제 P3이 부분군집의 특성이 강하기 때문인 것으로 판단된다. 약한 부분군집과 임의형상의 특성을 지닌 문제 P4에서는 SVCH와 SVDB 지수가 정확한 군집수를 추정하였다. 두 지수 모두 SVDD를 반영함에 따라 정확한 군집수를 추정하여 SVDD 반영에 따른 지수의 성능 향상을 확인할 수 있었다. 일반적으로 DU와 DB 지수는 임의형상 특성 데이터에 취약하다고 알려져 있고, CH 지수는 다른 지수들에 비해 임의형상 데이터에는 영향을 많이 받지 않는다고 알려져 있으나 SVDD 개념을 반영 한 이후에 DU 지수는 성능변화가 없고, 오히려 CH 지수는 성능이 향상되었다. 이로써 데이터의 형태 이외에 새로운 CVI의 성능에 영향을 미치는 다른 요인이 존재함을 확인할 수 있다. 문제 P4 뿐만 아니라 대부분의 문제에서 DU 지수는 SVDD 개념의 반영 전과 후의 성능의 차이가 존재하지 않는다.

임의형상과 비대칭의 특성을 지닌 또 다른 문제 P5에서는 DU, DB 그리고 SVDU 지수가 정확한 군집수를 추정하였다. DU 지수의 경우에는 SVDD 개념 반영 전과 후에 모두 군집수를 맞추었기 때문에 SVDD 개념 반영에 따른 영향력을 확인하기는 어렵다. CH 지수의 경우 SVCH가 정확한 군집수를 맞추지는 못하였으나 오답의 오차를 많이 감소하여 SVDD 개념을 반영함에 따라 성능이 향상되었다고 볼 수 있다. DB 지수는 SVDD 반영 후에 오히려 적확한 군집수를 맞추지 못하였으나 정답인 군집수 2와 군집수 3에 해당하는 지수 값의 차이가 다른 군

〈표 10〉 CVI별 군집수 추정 결과 (2차원 문제)

문제	DU	CH	DB	SVDU	SVCH	SVDB
P1	2	7	2	-1	-1	1
P2	-3	1	-3	-4	-4	-5
P3	-3	-3	-3	-4	-4	-4
P4	-1	7	7	-1	0	0
P5	0	8	0	0	1	1
P6	0	8	0	0	0	0
군집수 추정 정확도	2	0	2	2	2	2
오차 절대 값의 총합	9	34	15	10	10	11

집수들의 지수 값에 비해 그 차이가 적다. 이는 비계층적 알고리즘의 특성상 클러스터링 반복에 의해 군집결과가 달라진다면 정답을 맞출 수 있을 것으로 보인다. 노이즈와 부분군집 특성을 동시에 지닌 문제 P6은 DU, DB, SVDU, SVCH, SVDB 지수가 군집수를 맞추었다. 특히 CH 지수의 SVDD 개념 반영에 따른 성능이 뚜렷하다. 이는 노이즈에 취약한 CH 지수 성능이 SVDD 개념을 반영한 후에 향상된 것으로 보인다. 그러나 DU와 DB 지수는 SVDD 개념 반영 전과 후에 모두 군집수를 맞추어서 SVDU와 SVDB의 SVDD 반영에 따른 성능 변화를 확인하지는 못하였다.

〈표 11〉은 고차원 실험문제의 결과를 종합한 것이다. 문제 P7에서 CH와 SVCH 지수가 정확한 군집수를 추정하였으나 SVDD 반영 전과 후의 영향력은 확인하지 못하였다. 문제 P8에서는 DU 지수와 SVDD 개념을 반영한 지수들이 CH와 DB 지수에 비해 성능이 좋고, CH와 DB 지수의 경우 SVDD 개념의 반영에 따라 지수의 성능이 향상되었다. 문제 P9에서는 모든 지수가 정확한 군집수를 맞추는데 실패하였을 뿐만 아니라 SVDD 반영 이후의 성능의 차이는 확인되지 않았다. 문제 P10은 SVCH와 SVDB 지수만이 군집수를 맞추었다. 문제 P10은 9개의 속성을 갖는 문제로 다른 문제에 비해 속성이 많음에도 불구하고 SVCH와 SVDB 지수는 정확한 군집수를 추정하여, CH와 DB 지수가 SVDD 개념이 반영 후에 성능이 향상됨을 확인하였다. 문제 P11은 모든 지수들이 정확한 군집수 추정에 성공하여, 지수별 성능의 우열을 보이지 않았다. 문제 P11은 13개의 속성을 지닌 문제로 속성이 많음에도 불구하고 모든 지수가 정확한 군집수를 맞추었다. 따라서 이 문제는 군집 간의 경계가 뚜렷하거나, 분명한 문제일 것으로 추정된다.

마지막으로, 문제 P12에서는 모든 지수가 정확한 군집수를 바르게 추정하지 못하였으나, SVDD 개념을 반영한 이후 CH와 DB 지수의 CVI간 군집수 추정 오차의 차이가 많이 감소하였다. 고차원 성능실험 분석결과, SVCH, SVDB, DU, SVDU, DB, CH 지수 순으로 성능이 좋은 것으로 나타났다.

5.3 시사점

〈표 10〉과 〈표 11〉의 결과를 종합적으로 정리하면 CH와 DB 지수에 SVDD 개념을 반영하면 CVI

〈표 11〉 CVI별 군집수 추정 결과 (고차원 문제)

문제	DU	CH	DB	SVDU	SVCH	SVDB
P7	-1	0	-1	-1	0	-1
P8	0	8	6	0	0	0
P9	-8	-8	-8	-8	-8	-8
P10	-5	-5	-2	-5	0	0
P11	0	0	0	0	0	0
P12	-5	-5	-5	-6	-2	1
군집수 추정 정확도	2	2	1	2	4	3
오차 절대 값의 총합	19	26	22	20	10	10

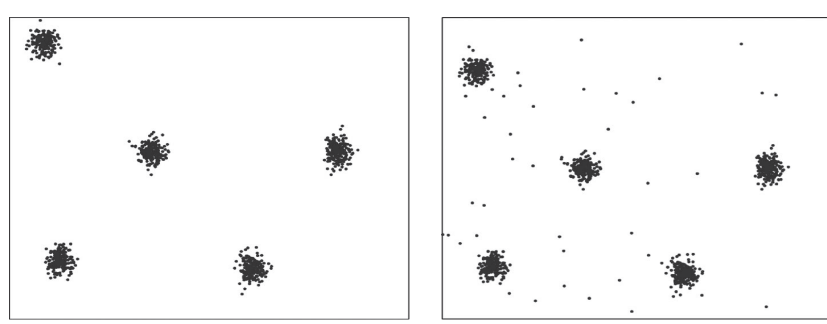
의 성능이 크게 향상됨을 확인할 수 있다. 즉, 본 연구에서 새롭게 제안한 SVDD의 개념을 이용한 CVI는 CH와 DB 지수가 갖는 약점을 보완하는데 큰 기여를 한다고 말할 수 있다. 앞에서 언급하였듯이, CH 지수는 노이즈와 비대칭분포 특성을 지닌 데이터, DB 지수는 부분군집과 임의형상 특성을 지닌 데이터에 대한 군집화 유효성 검증의 약점을 지닌다. 각 형상의 특징을 살펴보면 SVCH와 SVDB 지수가 갖는 강점을 이해하는데 도움이 될 것으로 판단되어, 이를 좀 더 상세하게 설명하기로 하자.

5.3.1 노이즈

〈그림 5〉의 (a)에서 보인 분리가 양호한 데이터는 군집 내 모든 객체들 간의 거리가 다른 군집의 객체와의 거리보다 가깝게(유사성이 크게) 나누어져 있다. (b)는 노이즈 특성을 표현하기 위하여 (a)의 데이터에 각 군집 수의 5%에 해당하는 데이터를 임의분포 되도록 한 것이다. 5%의 노이즈를 추가하면, (a)와 (b) 데이터들의 중심점 거리는 크게 변하지 않은 반면, 각 군집의 중심점과 모든 객체 간의 거리는 증가폭이 상대적으로 크다. 만약, CVI를 계산할 때 분리도에 비해 응집도의 변화가 크게 나타난다면, 노이즈 데이터의 군집화 유효성 검증에는 약점을 갖게 된다. CH 지수는 노이즈 데이터에서 분리도에 비해 응집도가 상대적으로 많이 증가한다는 특성을 지니므로, 이러한 특성의 데이터에서는 유효성 검증에 약점을 가지고 있다(Liu *et al.*, 2013). SVCH 지수는 노이즈에 대한 영향력을 직접적으로 감소시켜 주고, 벌점에 의해 의미있는 노이즈 데이터의 특성을 제대로 분석할 수 있다는 특징을 갖는다.

5.3.2 비대칭

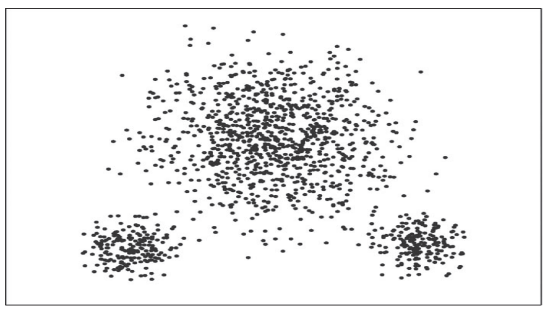
비대칭 데이터란 〈그림 6〉과 같이 군집 간 객체 수 또는 군집 크기가 현저히 다른 것을 말한다. CVI의 응집도를 계산할 때 중심점과 모든 객체의 거리를 이용하고, 분리도를 계산할 때에는 각 군집의 크기를 가중하여 계산하면, 군집의 객체 수 및 군집의 크기가 영향을 미친다. 이러한 개념으로 CVI를 계산하는 대표적인 것이 CH 지수이다. 그래서 CH 지수는 비대칭분포 특성을 지닌 데이터에도 민감한 것으로 알려져 있다(Liu *et al.*, 2013). 그러나 SVCH 지수는 개별 응집도에 단순히 구의 반지름만



(a) 분리양호 데이터 (b) 노이즈 데이터 (5%)

〈그림 5〉 노이즈 데이터의 군집

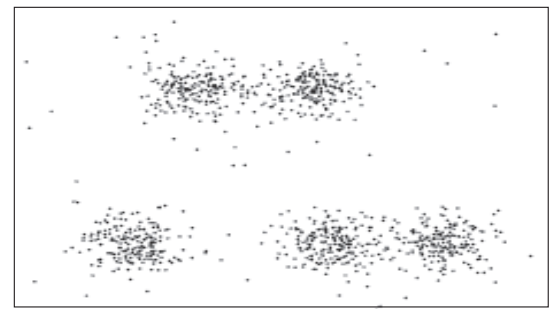
을 이용하기 때문에 비대칭 데이터에서 군집간의 객체수 차이에 따른 영향력이 완화될 수 있다.



〈그림 6〉 비대칭 데이터의 군집

5.3.3 부분군집

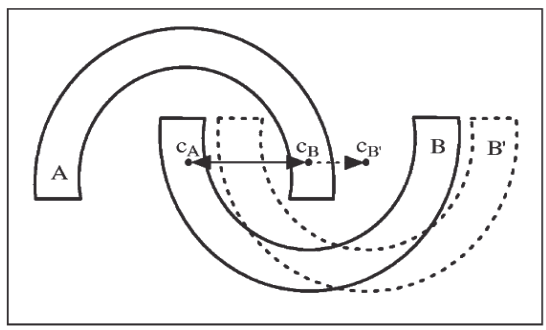
부분군집이란 서로 다른 군집에 속한 두 객체(데이터)간 거리가 동일한 군집내 객체간 거리보다 더 가까운 객체가 적어도 하나 이상 존재하는 데이터 형태를 의미한다. 〈그림 7〉은 전체 데이터가 5개의 군집으로 군집화되어 있고, 이들 중 4개의 군집은 이웃하는 군집과 부분군집의 형태를 지닌다. 전체 데이터 측면에서 결합이 가능한 모든 경우 수의 군집 쌍들의 중심점 간 거리로 분리도를 계산하면, 5개 군집으로 군집화 하는 것보다 3개 군집으로 군집화 하였을 때 분리도가 더 크게 산출된다. 따라서 최적 군집수가 3이 될 가능성이 높다. 따라서 DB 지수와 같이, 모든 경우 수의 군집 쌍의 분리도를 계산한 후, 결합에 유리한 분리도를 선택하는 방식으로 분리도를 계산하면 올바른 군집화 결과를 산출하기 어렵다. SVDB 지수는 커널함수를 이용하여 넓게 퍼져있던 입력공간의 데이터들을 특징 공간으로 서로 가깝게 매핑시키기 때문에 부분군집의 유효성 검증에 유리하게 된다.



〈그림 7〉 부분군집 데이터의 군집

5.3.4 임의형상

임의형상 데이터란 구모양의 군집이 아닌 임의적인 모양의 군집형태를 지닌 데이터를 말한다. DB 지수는 임의형상 특성을 지닌 데이터에서 군집의 중심점이 데이터가 집중 분포되어 있지 않은 곳에 존재할 수 있으므로, 실제로 군집 간 분리도는 낮음에도 불구하고, 중심점 간의 거리가 멀어져서 분리도 값을 크게 도출할 수 있다는 단점이 있다(Liu *et al.*, 2013). 〈그림 8〉에서 C_A와 C_B는 각각 군집 A와 B의 중심점이다. 이때, 군집 B가 오른쪽으로 평행이동하게 되면 B의 중심은 C_B에서 $C_{B'}$로 이동한다. 군집 B가 평행이동 함에 따라, 두 군집의 시각적인 위치는 더 가까워졌고, 두 군집이 접하게 되면 이는 분리도를 작게 할 수도 있다. 그러나 중심점 간의 거리로 분리도를 계산하면, C_A와 C_B의 거리보다는 C_A와 $C_{B'}$의 거리가 더 멀게 되어 분리도 값이 오히려 크게 도출된다. 이는 중심점을 이용하여 분리도를 계산하는 DB 지수에서 일어날 수 있는 현상이다. SVDB 지수의 응집도는 커널함수에 의한 특징 공간에서 '구의 중심과 SV사이의 거리'를 이용하므로, 임의형상 특성을 지닌 데이터에 적합하다고 판단된다.



〈그림 8〉 임의형상 데이터의 군집

Ⅵ. 결론 및 토의

본 연구에서는 Fuzzy K-means 알고리즘에 의하여 수치형 데이터의 군집분석한 결과를 검증할 수 있는 새로운 CVI들을 개발하고, 이들의 성능을 비교 및 분석하였다. 새롭게 개발한 CVI들은 기존 연구에서 많이 사용된 DU, CH, DB 지수의 응집도 계산에 SVDD 개념을 반영한 것이다. 분석결과, 2차원 실험문제에서 SVCH와 SVDB 지수가 가장 좋고, SVDU, DU, DB, CH 지수의 순서로 성능이 좋았다. DU 지수는 대부분의 문제에서 SVDD의 반영에 따른 영향력이 없었다. 반면, CH와 DB 지수는 SVDD 개념을 반영한 후, CVI의 성능이 향상됨을 확인할 수 있었다. 고차원 실험문제의 결과도 2차원 실험문제의 결과와 유사하게 CH와 DB 지수에서 SVDD 개념을 반영한 이후에 CVI의 성능이 향상되었다.

본 연구는 CVI의 응집도 계산방법에 SVDD 개념을 반영하여 새로운 응집도 계산방법을 제안한 연구이다. 앞에서 언급한 바와 같이, CVI에 SVDD 개념을 접목한 연구가 아직까지 이루어지지 않았다는 점에서 고려할 때, 본 연구는 학술적 의의가 크다고 판단된다. 그러나 다음과 같은 한계점을 갖고 있기 때문에, 한계점을 극복할 수 있는 추후 연구들이 반드시 이루어져야 할 것이다. 첫째, 최적 군집수와 데이터의 형상이 알려진 벤치마킹 문제만을 이용하여 CVI의 성능을 평가하였다. 그러나 군집분석을 해야 하는 실제 문제들은 군집수나 데이터의 형상을 모르는 경우가 대부분이다. 둘째, 수치형 데이터를 비계층적으로 군집화한 결과를 판단하는 것에 한정되어 있다. 셋째, 개발한 CVI들은 응집도 계산 방법만을 달리 하였고, 분리도 계산 방법은 기존 방법을 그대로 사용하였다. 이러한 접근이 CVI의 성능에 어떤 영향을 주는지에 대한 검토는 충분히 이루어지지 않았다. 따라서 본 연구결과를 일반화시키기 위해서는 보다 다양한 실험문제를 활용하여 분석해야 하고, 분석 결과는 다양한 기법들을 통해 검증되어야 할 것이며, 군집수가 알려지지 않은 문제에서 CVI의 성능을 판단하는 기준도 마련되어야 할 것이다. 또한, 분리도 계산의 개선을 통해 더욱 혁신적인 개념의 CVI를 제안할 수 있을 것이다. 이러한 과정을 통해 다양한 군집화 문제에 로버스트(robust)하게 적용될 수 있는 CVI를 개발할 수 있을 것으로 기대한다.

참고문헌

김민호 · Ramakrishna, R.S.(2005), "비형식의 군집 유효화 지수의 분석과 새로운 지수 개발," **한국컴퓨터종합학술대회**, 32, 601-603.

김민호 · 유현진 · Ramakrishna, R.S.(2005), "고차원 응용에서의 군집 유효성 평가 기법," **한국정보과학회 2005 가을 학술발표 문집(II)**, 32, 715-717.

김영옥 · 이수원(2002), "최적의 군집을 찾기 위한 상대적

군집 평가 방법," **한국정보과학회 2002 가을 학술발표논문집**, 29, 334-336.

송동성 · 김표재 · 장형진 · 최진영(2007), "Negative data를 고려한 K-means Support Vector Data Description," **대한전기학회 학술대회 논문집**, 310-312.

신경석 · 김재윤(2011), "클러스터 수가 주어지지 않는 클러스터링 문제를 위한 공생 진화알고리즘," **품질경영학회지**, 39, 98-108.

안현철 · 김경재 · 한인구(2005), "Support Vector Machine을 이용한 고객구매예측모형," **한국지능정보시스템학회논문지**, 11, 69-81.

오은녕 · 이희상(2002), "클러스터링 기법을 이용한 이동통신의 고객 세분화 연구," **한국경영과학회 추계학술대회논문집**, 421-424.

용환승 · 나연묵 · 박종수 · 승현우 · 이민수 · 이상준 · 최린(2007), "데이터마이닝," 서울, 인피니티북스.

이만재(2012), "빅 데이터 어낼리틱스와 공공 데이터 활용," **정보과학회지**, 30, 33-39.

이신원 · 안동언 · 정성종(2004), "K-Means 알고리즘을 이용한 계층적 클러스터링에서 클러스터 계층 깊이와 초기값 선정," **정보관리학회지**, 21, 173-185.

이신원(2012), "K-means 클러스터링에서 초기 중심 선정 방법 비교," **한국인터넷정보학회**, 13, 1-8.

이수현(2015), "빅 데이터의 군집분석을 위한 군집화 유효성 지수 개발과 응용," 전남대학교 대학원 박사학위 논문.

전치혁(2012), "데이터마이닝 기법과 응용," 서울, 한나래.

황인수(2002), "데이터 마이닝에서 그룹 세분화를 위한 2단계 계층적 클러스터링 알고리듬," **경영과학**, 19, 189-196.

허경용 · 서진석 · 이임건(2011), "Fuzzy c-means의 문제점 및 해결 방안," **한국컴퓨터정보학회 논문지**, 16, 39-46.

Bezdek, J. C.(1981), "Pattern Recognition with Fuzzy Objective Function Algorithm," *Plenum Press*, 13, 367-373.

Calinski, R. B., and J. Harabasz(1974), "A Dendrite Method for Cluster Analysis," *Communications in Statistics*, 3, 1-27.

Chang, H. J., P. J. Kim, J. H. Choi, and J.Y. Choi (2007), "Support Vector Data Description Using Clustering Method," *International Technical Conference on Circuits Systems, Computers and Communications*, 1132-1133.

Davies, D., and D. Bouldin(1979), "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, 224-227.

Dunn, J. C.(1973), "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well Separated Cluster," *Journal of Cybermetics*, 3, 32-57.

Dunn, J. C.(1974), "Well Separated Clusters and Optimal Fuzzy Partitions," *Journal of Cybernetics*, 4, 95-104.

Halkidi, M., and M. Vazirgiannis(2001), "Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set," *Proceedings of 2001 IEEE International Conference on Data Mining*, 187-194.

Halkidi, M., Y. Batistakis, and M. Vazirgiannis (2001), "On Clustering Validation Techniques," *Journal of Intelligent Information Systems*, 17, 107-145.

Hruschka, E. R., R. G. B. Campello, A. A. Freitas, and A. P. L. Carvalho(2009), "A Survey of Evolutionary Algorithms for Clustering," *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 39, 133-155.

Jain, A. K., M. N. Murty and P. J. Flyn(1999), "Data Clustering: A Review," *ACM Computing Surveys*, 31, 264-323.

Ji, R., D. Liu, M. Wu, and J. Liu(2008), "The Application of SVDD in Gene Expression Data Clustering," *The 2nd International Conference on Bioinformatics and Biomedical Engineering*, 371-374.

Liu, Y., Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu (2013), "Understanding and Enhancement of Internal Clustering Validation Measures," *IEEE Transactions on Cybernetics*, 43, 982-994.

MacQueen J. B.(1967), "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Maulik, U., and S. Bandyopadhyay(2002), "Performance Evaluation of Some Clustering Algorithms and Validity Indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 1650-1654.

Niazmardi, S., S. Homayouni, and A. Safari(2013), "An Improved FCM Algorithm Based on the SVDD for Unsupervised Hyperspectral Data Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6, 831-839.

Raspini, E. H.(1969), "A New Approach to Clustering," *Information and Control*, 16, 22-32.

Saitta, S., B. Raphael, and I. F. C. Smith(2008), "A Comprehensive Validity Index for Clustering," *Intelligent Data Analysis*, 12, 529-548.

Tax, D. M. J., and R. P. W. Duin(2004), "Support Vector Data Description," *Machine Learning*, 54, 45-66.

Tay, F. E. H., and L. J. Cao(2006), "Modified Support Vector Machines in Financial Time Series Forecasting," *Neurocomputing* ,48, 847-861.

Theodoridis, S. and K. Koutroumbas(2006), *Pattern Recognition*, Academic Press.

Vapnik, V.(1979), *Estimation of Dependences Based on Empirical Data*[in Russian], Nauka.

Xie, XL, Beni, G.(1991), "A Validity Measure for Fuzzy Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, 841-847.

Xu, J., J. Yao, and L. Ni(2011), "Fault Detection Based on SVDD and Cluster Algorithm," *International Conference on Electronics, Communications and Control*, 2050-2052.

Xu, R. and D. II Wunsch(2005), "Survey of Clustering Algorithms," *IIEEE Transactions on Neural Networks*, 16, 645-678.

Various Validity Indices for Fuzzy K-means Clustering

Soo-Hyun Lee* · Jae-Yun Kim** · Young-Seon Jeong***

Abstract

Cluster analysis (or Clustering) is used in many different fields such as finance, marketing, and operations management to draw homogeneous cases. Due to that reason, the result extracted from cluster analysis is stated to be the core element to maximize the firm's value. Because the number of clusters in clustering problems is usually unknown, it is significant to evaluate the clustering results produced by different parameter settings. After a range of possible number of clusters are evaluated, the best partition is selected based on the cluster validity analysis. Cluster validity index (CVI) is an indicator to provide a way of validating the quality of clustering algorithms and determine the correct number of clusters in datasets. A CVI is composed of the summation or ratio of compactness and separability measures in which compactness indicates the concentration of data in each cluster and separability refers to the inter-cluster distances. A good clustering result will have smaller compactness and larger separability values. This research will cover the theoretical research of CVI to verify the effectiveness of Fuzzy K-means clustering results among the analytical research methods. Depending on the different combination of compactness and separability measures, several CVIs have been developed. The CVIs calculated by the ratio of compactness to separability or vice versa such as Dunn index, DB index, and XB index were proposed, and the weighted sum of these two measurements was developed as SD index and S_Dbw index. In addition, several variants of conventional CVIs have been recently proposed. However, most of existing CVIs are sensitive to arbitrary shapes of clusters, sub-clusters, and outliers because the measure of compactness of those clusters is not obvious in the original domain. We suggest new CVIs by calculating the concept of Support Vector Data

* Post-Doc., The Graduate Program on Climate Change, Sustainability and Business, Chonnam National University, First Author
** Professor, Dept. of Business Administration, Chonnam National University, Co-Author
*** Assistant Professor, Dept. of Industrial Engineering, Chonnam National University, Corresponding Author

Description (SVDD) in each particular cluster calculation of CVI by separating the compactness and separability about some indices well known to prove effectiveness: Dunn (DU), Calinski and Harabasz (CH), and Davies-Bouldin (DB). By conducting efficiency comparisons utilizing Fuzzy K-means clustering algorithm and various benchmarking instances, the performance rate of new CVIs has been verified with outstanding performance. The performance of noise, skewed, sub-cluster, and arbitral shapes data in the new CVIs is promising in particular. The concept of SVDD has been applied to the compactness by this research and newly created CVIs were verified to be efficient in regards to cluster effectiveness. The compactness calculation method suggested in this research is expected to be widely applied in many different CVIs. As the research of cluster analysis become more expanded and the research follows the step of diversity, this research is expected to contribute the application scope of SVDD and the expansions of both cluster analysis and the concept of CVI.

Key words: Clustering, CVI, SVDD, Compactness, Fuzzy K-means

• 저자 이수현은 전남대학교 수학과에서 학사, 동 대학교에서 경영학 석사, 박사학위를 취득하였다. 현재 전남대학교 기후특성화대학원에서 박사후연구원으로 재직 중이다. 관심분야는 최적화, 비즈니스 애널리틱스, 환경경영, 지속가능경영, 생산설비의 수명주기분석 등이다.

• 저자 김재윤은 현재 전남대학교 경영학부 교수로 재직 중이다. 전남대학교 산업공학과에서 학사, 석사, 박사 학위를 취득하였다. 주요 연구분야는 AHP/DEA/BSC 기반 경영성과측정, 생산시스템의 분석과 설계, 진화연산기법을 이용한 조합최적화 문제 해결 등이다.

• 저자 정영선은 현재 전남대학교 산업공학과 조교수로 재직 중이다. 전남대학교 산업공학과를 졸업하였으며, 고려대학교에서 산업공학과 석사, Rutgers University Industrial and Systems Engineering에서 박사학위를 취득하였다. 주요 연구관심분야는 빅데이터 분석, 이상 공정시스템 탐지 방법론 개발, 지능형 교통시스템 구축 등이다.