한국경영과학회지
제41권 제2호
2016년 5월
http://dx.doi.org/10.7737/JKORMS.2016.41.2.017

경영사례를 이용한 군집화 유효성 지수의 성능비교*

이수현[1] · 정영선[2] · 김재윤[3†]
[1]전남대학교 기후변화특성화대학원, [2]전남대학교 산업공학과, [3]전남대학교 경영학과

Performance Comparison of Clustering Validity Indices with Business Applications

Soo-Hyun Lee[1] · Youngseon Jeong[2] · Jae-Yun Kim[3]
[1]The Graduate Program on Climate Change, Sustainability and Business, Chonnam National University
[2]Dept. of Industrial Engineering, Chonnam National University
[3]Dept. of Business Administration, Chonnam National University

■ Abstract ■

Clustering is one of the leading methods to analyze big data and is used in many different fields. This study deals with Clustering Validity Index (CVI) to verify the effectiveness of clustering results. We compare the performance of CVIs with business applications of various field. In this study, the used CVIs for comparing performance are DU, CH, DB, SVDU, SVCH, and SVDB. The first three CVIs are well-known ones in the existing research and the last three CVIs are based on support vector data description. It has been verified with outstanding performance and qualified as the application ability of CVIs based on support vector data description.

Keywords : Clustering, Clustering Validity Index, Support Vector Data Description, Business Application

논문접수일 : 2015년 10월 28일 논문게재확정일 : 2015년 12월 03일
논문수정일(1차 : 2015년 12월 02일)
* 한국경영과학회 제1회 박사학위논문 우수상을 수상한 제1저자의 학위논문 일부를 수정·보완하였으며, 심사과정에서 유익한 논평을 해주신 심사위원들께 감사의 인사를 드립니다.
† 교신저자, jaeyun@jnu.ac.kr

1. 서 론

빅 데이터(big data)는 대용량 데이터를 저장, 수집, 발굴, 분석, 비즈니스화하는 일련의 과정을 말한다[4]. 빅 데이터를 분석하면 지금까지 이해할 수 없었던 정보를 이해하고, 시스템이나 그 구성요소들의 움직임을 예측하여 새로운 가치를 창출하는데 활용할 수 있다. 따라서 빅 데이터의 분석에 대한 산업계와 학계의 관심이 증대되고 있다[4, 5].

군집분석(clustering or cluster analysis)은 빅 데이터의 대표적인 분석기술 중 하나로 데이터의 패턴과 분포를 찾아 유사한 속성을 갖는 몇 개의 그룹으로 데이터를 나누는 기술이다[6, 7]. 군집분석은 데이터 내에 특정한 구조가 없는 상황에서도 군집화 알고리즘이 주어지면 군집화 결과를 찾아낼 수 있기 때문에 군집분석의 결과를 실무에 활용하기 위해서는 군집화 결과의 유효성을 검증하는 과정이 요구된다. 일반적으로 군집화 결과의 유효성을 검증하는데 군집화 유효성 지수(clustering validity index : CVI)가 이용되며, CVI를 이용하여 군집화 결과를 검증하면 군집분석에 대한 신뢰도를 측정할 수 있다고 알려져 있다[15]. 그리고 유효성이 검증된 군집화 결과는 수많은 데이터 속에서 목적에 부합하는 정보를 찾아내어, 효과적인 분석 및 그 결과를 제공할 수 있다. 이렇게 얻게 되는 군집분석의 결과는 공공 및 민간부문의 다양한 영역에 활용된다. 군집분석의 중요성과 연구영역이 증가하고 있지만, 상대적으로 CVI에 대한 연구는 특정 데이터에 한정된 연구가 일반적이며, 다양한 영역에 적용 가능한 CVI의 개발 및 성능비교 연구는 상대적으로 부족하다. 이러한 학술적 배경하에서 본 연구는 경영학의 여러 분야에서 다루어지는 군집분석의 사례문제들을 이용하여 CVI들의 성능을 비교하고자 한다. 이 연구는 다양한 특성과 데이터 형태를 갖는 경영사례문제를 이용하여 군집분석하고 그 결과를 통해 CVI의 성능을 비교함으로써, 군집분석 및 CVI의 실무적 활용 측면에서 시사점들을 도출할 수 있을 것으로 기대된다.

빅 데이터의 핵심은 데이터의 비즈니스화에 있다[4]. 특히, 경영학 분야에서 군집분석은 기업의 성과를 높일 수 있는 핵심 자원의 탐색에 중요한 역할을 하고 있다. 경영학의 군집분석 활용사례들을 살펴보면 다음과 같다. 재무분야에서는 고객의 우·불량 여부를 판별하거나 고객신용등급을 관리하는 신용평가(credit score), 부정행위를 적발하여 불량채권 발생을 방지하는 고객관계관리(customer relation management), 주가예측(stock prediction), 위험관리, 포트폴리오 분석(portfolio analysis), 비용예측(forecasting of cost) 등에 군집분석을 사용하고 있다[14, 16]. 마케팅 분야에서는 유사한 고객군을 도출하고 각 고객군에 따라 특화된 마케팅 전략을 수립하는 고객세분화(segmentation) 및 목표 마케팅(target marketing), 이탈고객을 사전에 방지하기 위하여 고객 데이터 중에서 이탈 확률이 높은 고객을 선별하는 고객성향 변동 분석(churn analysis), 그리고 관련 제품을 판매할 수 있는 교차판매(cross-selling) 가능성을 식별하는 연관분석(association analysis) 등에 군집분석을 활용하고 있다[12, 14]. 생산관리 분야에서는 결함의 위치 및 형태에 기초하여 유사한 특성을 지닌 결함으로 군집화하여 공정시스템에 반영하는 품질개선(quality improvement)에 군집분석이 이용되고 있다[14].

본 연구는 경영학중 재무, 마케팅, 생산관리 분야의 중·소규모 사례문제를 이용하여 기존 연구들에서 제안된 DU(Dunn), CH(Calinski-Harabasz), DB(Davies-Bouldin), SVDU(Dunn based on SVDD), SVCH(Calinski-Harabasz based on SVDD), SVDB(Davies-Bouldin based on SVDD) 지수 등 6개 CVI의 성능을 비교함으로써, 군집분석을 활용하는 경영사례문제에서 CVI의 특성 및 군집화 결과의 유효성을 확인하고자 한다. 이를 통해, 다양한 특성이나 데이터의 형태가 서로 다른 문제별로 적용 가능한 CVI를 제안할 수 있을 것이다. CVI에 의해 검증된 군집분석의 결과는 빅 데이터와 같이 복잡한 형태 및 다양한 특성을 지닌 데이터로부터 목적에 부합하는 패턴을 추출하여 기업의 경쟁력 강화 및 생산성 향상을 촉진시킬 수 있는 핵심자원으로 활용될 수 있다.

본 논문은 다섯 부분으로 구성되어 있다. 제1장에서는 연구배경 및 연구내용을 언급하였고, 제2장에서는 군집분석에 대한 이론적 내용을 설명하고, 제3장에서는 본 연구에서 성능비교에 사용된 6가지 CVI의 개념, 특징, 수리모형 등을 제시한다. 제4장에서는 경영사례문제들을 설명하고 CVI의 성능 실험 및 실험결과를 해석하며, 제5장에서 연구의 기여도와 시사점, 그리고 연구의 한계 및 향후연구 방향을 제시한다.

2. 군집분석[1)]

군집분석의 학술적 연구분야는 군집하는 방법인 군집화 알고리즘(clustering algorithm) 개발, 사전에 결정된 군집화 알고리즘에 의해 얻어진 군집화 결과의 유효성을 판단하는 지수 개발, 최적 군집화 결과 도출 등으로 구분할 수 있다. 본 연구는 군집분석의 연구분야 중 군집화 결과의 유효성을 검증하는 CVI의 활용에 관하여 다룬다.

군집화 알고리즘은 사전에 군집의 수를 정하지 않고 단계적으로 서로 다른 군집화 결과를 제시해 주는 계층적 군집화(hierarchical clustering) 알고리즘과 사전에 군집의 수를 정하고 각 데이터를 군집에 배정하는 비계층적 군집화(nonhierarchical clustering) 알고리즘이 있다[34]. 비계층적 군집화 알고리즘은 계층적 군집화 알고리즘에 비해 계산 소요시간 측면에서 유리하기 때문에 다양한 분야에서 많이 사용되고 있으며 다양한 기법들이 소개되어 왔다[14, 29]. 대표적인 알고리즘으로는 K-means, K-medoids, Fuzzy K-means, 모형기반(model-based) 알고리즘 등이 있다[23]. 본 연구에서는 유연성이 높다고 알려져 있는 Fuzzy K-means 알고리즘[19]을 이용하여 비계층적 군집화 알고리즘의 결과를 검증할 수 있는 CVI들의 성능비교 연구를 수행하고자 한다. Fuzzy K-means 알고리즘은 주어진 K개 군집의 중심좌표(centroid)와 모든 객체(데이터)간 거리를 계산하고, 객체와 중심좌표간 거리가 가장 가까운 군집에 각 객체를 배정하는 반복적 알고리즘이다. 대표적인 군집화 알고리즘인 K-means 알고리즘과 달리, Fuzzy K-means 알고리즘은 객체를 군집에 배정할 때, 한 객체가 여러 군집에 속할 가능성을 허용하는 확률 또는 퍼지개념을 이용하여 소속군집을 결정한다는 점에서 그 활용도가 높아 다양한 문제에 널리 쓰이고 있다[1, 2, 3].

앞에서 언급한 바와 같이, 군집화 알고리즘을 통해 얻은 군집결과는 데이터 내의 구조확인을 통한 신뢰성 확보가 요구된다. 일반적으로 군집화한 후, 주어진 데이터 집합이 구조를 지니고 있는지에 대한 평가는 다음 기준들을 이용하여 가능하다. 첫째, 데이터의 군집화 경향(clustering tendency)을 살펴봄으로써 실제로 데이터에 구조가 존재하는지를 확인한다. 둘째, 정확한 군집의 개수를 결정하였는지 살핀다. 셋째, 외부정보 없이 군집화 결과가 데이터에 얼마나 잘 맞는지를 평가한다. 넷째, 외부적으로 잘 알려져 있는 결과와 군집분석의 결과를 비교하여 더 좋은 군집들의 집합으로 결정되어 있는지를 판단한다. 군집화 유효성 지수(CVI)는 군집화 결과를 평가하는 도구로, 계산된 지수 값을 근거로 군집화 결과에 대한 신뢰성을 검증한다[10, 17, 24, 35]. 일반적으로 CVI는 앞에서 언급한 데이터 집합의 내부구조를 확인할 수 있는 4가지 평가기준들의 일부를 고려하여 개발한다. 따라서 CVI를 이용하여 군집화 결과를 검증하면 군집분석에 대한 신뢰도를 측정할 수 있을 뿐 아니라, 군집결과의 품질을 향상 시킬 수 있어 CVI에 대한 관심이 높아지고 있다[4, 15, 27].

3. 군집화 유효성 지수

데이터의 형태(예 : 수치형, 명목형, 순위형 등)에 따라 다양한 CVI가 존재할 수 있다. 본 연구에서는 수치형 데이터의 군집분석에 따른 CVI를 다룬다. 수치형 데이터를 위한 CVI는 군집화 결과의 유효성을 응집도(compactness : C)와 분리도(separability :

1) '전치혁, 「데이터마이닝 기법과 응용」, 서울 : 한나래, 2012'의 내용을 요약하여 정리함.

S)라는 2가지 평가기준을 이용하여 판단한다[14]. 여기서 응집도는 한 군집에 존재하는 두 객체의 닮은 정도에 대한 수치인 유사도(similarity)를 나타내는 척도이고, 분리도는 서로 다른 군집 간 객체들의 다른 정도를 나타내는 비유사도(dissimilarity)에 대한 척도이다. 유사도와 비유사도는 인접성(proximity)을 계산하여 정량적으로 평가된다. 따라서 응집도와 분리도는 인접성 함수에 의해 측정되며, 이때 응집도는 인접성이 클수록 좋은 것으로 판단되고, 분리도는 인접성이 작을수록 좋은 것으로 판단된다.

일반적으로 CVI는 응집도와 분리도 중 한가지만을 고려하여 설계하거나, 응집도와 분리도를 동시에 고려하여 설계하되 각각을 나누어 고려한 후 이를 결합하는 방식으로 설계한다[3, 14]. 이때 응집도와 분리도 중 한 가지만을 고려하면, 군집의 수에 따라 CVI값이 단조 증가하거나 감소하여 최적 군집수를 구하기 어렵다는 단점을 갖는다[14, 25]. 따라서 본 연구는 응집도와 분리도를 동시에 고려한 CVI들을 살펴보고자 한다.

본 연구는 응집도와 분리도를 동시에 고려한 CVI 중에서 기존 연구[18, 23, 25, 27, 28]를 바탕으로 사용 빈도가 가장 높거나 성능이 우수하다고 알려진 Dunn(DU) 지수[22], Calinski-Harabasz(CH) 지수[20], Davies-Bouldin(DB) 지수[21] 등 3개와 이수현[11]이 DU, CH, DB 지수에 서포트 벡터 데이터 표현(support vector data description : SVDD) 개념을 반영하여 수정 개발한 SVDU, SVCH, SVDB 지수 등 3개를 합하여 총 6개 CVI의 성능을 비교 분석하고자 한다.

SVDD는 기계학습 알고리즘(machine learning algorithm) 중 최근에 등장하여 여러 가지 문제에서 우수한 해결능력을 보여주는 비선형 SVM(non-linear support vector machine)의 응용 형태이다[8, 34]. SVDD의 기본 아이디어는 학습 데이터가 주어졌을 때, 학습 데이터와 중심 간의 거리가 반지름을 벗어나면 벌점(penalty)을 부과하여, 최소한의 반지름을 갖는 구를 찾는 것이다. SVDD는 단일 클래스 분류기법이지만, 다중 클래스 분류에도 응용이 가능하다. 즉, 데이터 집합내에 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[9, 26, 30, 32, 33]. SVDD의 개념을 기반으로 하는 CVI인 SVDU, SVCH, SVDB는 SVDD의 원리를 응집도 계산에 반영한 것이다. 거리척도를 이용하여 응집도를 계산하는 DU, CH, DB 지수에서 응집도는 그 값이 작을수록 응집도가 좋다고 판단한다. 이는 구의 반지름이 작을수록 좋다고 판단하는 SVDD의 최적판단기준과 유사하다고 생각한 것이다.

일반적으로 CVI들은 군집분석을 위해 주어진 데이터 특성을 사전에 알고, 이에 적합한 알고리즘의 성능을 판단할 수 있도록 설계되어 있기 때문에 성능이 우수한 CVI라 할지라도 그 적용이 제한적이다. 따라서 데이터의 특성을 사전에 알고, 데이터에 적합한 CVI를 적용하여야만 군집결과의 타당성을 평가할 수 있다. 그러나 빅 데이터는 그 양이 방대하고 정보의 변화속도가 빠르기 때문에 데이터에 대한 특별한 정보가 주어져 있지 않는다. 따라서 빅 데이터를 이용하여 군집분석하고 군집결과로부터 의사결정에 활용 가능한 정보를 도출하기 위해서는 다양한 데이터의 구조 및 패턴에 이용할 수 있는 CVI가 요구된다.

기존 연구에 따르면 일반적으로 CVI는 노이즈(noise) 및 임의형상(arbitrary shape) 그리고 부분군집(sub-cluster) 등의 데이터 특성에 특히 취약하다고 알려져 있다[20, 21, 22, 23, 25, 27, 28]. CVI의 응집도에 SVDD를 반영하면 노이즈 및 임의형상의 특성을 갖는 데이터에서 성능이 향상된다고 알려져 있다[11]. 본 논문에서 다루는 경영사례문제들도 이러한 형태적 특성을 갖는 문제들을 주로 선별하고자 하였다. 이를 통해, CVI에 대한 실무적 활용성 검증뿐 아니라, CVI의 데이터 형태별 성능 분석도 함께 이루어질 수 있을 것이다. 본 연구의 성능비교분석을 위해 선별된 6개 CVI들에 대한 수리모형과 응집도와 분리도의 결합방법 및 특성은 <표 1>에 제시하였다.

〈표 1〉 비교대상 CVI들

번호	명칭(표기)	수리모형(기호)
	정의, 응집도와 분리도의 결합방법, 최적 군집수 산정 방법, 장·단점	
1	Dunn (*DU*)	$DU_K = \min_{i,j=1,\cdots,K, i\neq j}\{\min_{x\in c_i, y\in c_j} d(x,y)\}/\max_{i=1,\cdots,K}\{\max_{x_1,x_2\in c_i} d(x_1,x_2)\}$ • x, y : 군집 C_i와 C_j의 객체 • x_i : 군집 C_i의 임의 객체
	- 서로 다른 군집에 속한 객체 간 거리(분리도)를 군집 내 객체 간 거리(응집도)로 나눈 값 - 응집도에 대한 분리도의 가중비에 의해 결합 - '응집도에 대한 분리도의 비' 중에서 최대가 되는 군집수 K가 최적 군집 수(MAX S/C) - 부분군집과 임의형상에 민감	
2	Calinski-Harabasz (*CH*)	$CH_K = \sum_{i=1}^{K} n_i \cdot d(z_i, z_{tot})^2/(K-1) \cdot (N-K)/\sum_{i=1}\sum_{x\in c_i}^{K} d(x, z_i)^2$ • N : 전체 데이터의 개수 • z_i : 군집 C_i의 중심점 • z_{tot} : 전체 데이터의 중심점
	- 군집 중심점과 전체 중심점 간 거리(분리도)를 군집 내 객체와 중심점 간 거리(응집도)로 나눈 값 - 응집도에 대한 분리도의 가중비에 의해 결합 - '응집도에 대한 분리도의 비' 중에서 최대가 되는 군집수 K가 최적 군집 수(MAX S/C) - 노이즈와 비대칭분포에 민감	
3	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
	- 군집 내 객체와 중심점 간 거리(응집도)의 합을 군집들의 중심점 간 거리(분리도)로 나눈 값 - 분리도에 대한 응집도의 가중비에 의해 결합 - '분리도에 대한 응집도의 비' 중에서 최소가 되는 군집수 K가 최적 군집 수(MIN C/S) - 부분군집과 임의형상에 민감	
4	Dunn based on SVDD (*SVDU*)	$SVDU_K = \min_{i,j=1,\cdots,K, i\neq j}\{\min_{x\in C_i, y\in C_j} d(x,y)\}/\max_{i=1,\cdots,K}\{J(C_i)\}$ • $J(C_i) = 2\left(\sum_{p_i\in SV_i}\frac{R_{p_i}^2}{|\{SV_i\}|}\right)$ • $R_{p_i}^2 = K(S_{p_i}, S_{p_i}) - 2\sum_k \alpha_k K(S_{p_i}, X_k) + \sum_{k,l}\alpha_k\alpha_l K(X_k, X_l)$
	- 서로 다른 군집에 속한 객체 간 거리(분리도)를 SV의 평균 반지름(응집도)로 나눈값 - 응집도에 대한 분리도의 가중비에 의해 결합 - '응집도에 대한 분리도의 비' 중에서 최대가 되는 군집수 K가 최적 군집 수(MAX S/C) - 부분군집에 민감, Dunn 지수에 비해 임의형상에 미약한 성능향상	
5	Calinski-Harabasz based on SVDD(*SVCH*)	$SVCH_K = \sum_{i=1}^{K} n_i \cdot d(z_i, z_{tot})^2/K-1 \cdot N-K/\sum_{i=1}^{K} J(C_i)$
	- 군집 중심점과 전체 중심점을 이용한 군집 간 거리(분리도)를 SV의 평균 반지름(응집도)로 나눈값 - 응집도에 대한 분리도의 가중비에 의해 결합 - '응집도에 대한 분리도의 비' 중에서 최대가 되는 군집수 K가 최적 군집 수(MAX S/C) - 비대칭분포에 민감, CH 지수에 비해 노이즈 및 임의형상에 성능향상	
6	Davies-Bouldin based on SVDD (*SVDB*)	$SVDB_K = \frac{1}{K}\sum_{i=1}^{K}\max_{i,j=1,\cdots,K,\ i\neq j}\{(J(C_i)+J(C_j))/d(z_i,z_j)\}$
	- SV의 평균 반지름(응집도)의 합을 군집들의 중심점 간 거리(분리도)로 나눈 값 - 분리도에 대한 응집도의 가중비에 의해 결합 - '분리도에 대한 응집도의 비' 중에서 최소가 되는 군집수 K가 최적 군집 수(MIN C/S) - 부분군집에 민감, DU 지수에 비해 임의형상에 성능향상	

4. 경영사례를 이용한 성능비교 실험

4.1 실험설계

CVI의 성능비교는 어떤 특성을 갖는 데이터를 사용하여 분석할 것이며, 어떤 방법으로 그 성능을 비교할 것인지가 중요하다. 앞에서 언급하였듯이, 본 연구에서는 Fuzzy K-means 알고리즘을 사용하여 재무, 마케팅 및 생산관리 분야의 수치형 데이터를 갖는 군집분석 사례문제에서 6개 CVI가 주어진 군집수를 정확하게 맞추는지 확인하는 방식으로 CVI의 성능을 비교한다.

CVI들의 성능비교를 위한 구체적인 절차는 다음과 같다. 첫 번째는 군집화 단계로 군집수를 2부터 10까지 순차적으로 변화시키면서 Fuzzy K-means 알고리즘에 의해 군집분석을 실행하고 데이터를 군집한다. 두 번째는 CVI 값의 측정단계로 도출된 군집결과를 바탕으로 각 CVI의 지수값을 <표 1>에 제시된 수식들에 의하여 계산한다. 이때, 초기해에 민감한 Fuzzy K-means 알고리즘의 단점을 보완하기 위하여 군집화를 100회 반복 실행하고, 그 평균값을 이용하여 각 CVI의 지수값을 산출하였다. 즉, 첫 번째 단계와 두 번째 단계를 100회 반복하여 실행하고, 이들이 평균값을 각 CVI의 최종 지수값으로 확정한다. 또한, SVDU, SVCH, SVDB 등 응집도에 SVDD의 개념이 반영된 3개 CVI들의 지수값을 계산하기 위해서는 커널(kernel) 함수에 이용되는 매개변수와 조절상수의 값을 결정해야 한다. 본 연구에서는 비교적 성능이 우수하다고 알려진 RBF(radial basis function) 커널함수를 사용하였고[9], RBF 커널함수에 사용되는 매개변수와 조절상수의 최적값을 선정하기 위하여 다양한 매개변수 값의 조합을 대입한 시행착오적 방법을 선택하였다. 특히, 시행착오적 방법을 시행할 때 Tay and Cao[31]가 제안한 매개변수 범위를 참고하여 시그마(σ)는 1에서 10사이의 값을 1 간격으로, 에러수준(f)은 0.01에서 0.1사이의 값을 0.01 간격으로 세분화하여 문제별로 최적의 매개변수 값을 선정하였다. 이에 대한 구체적인 값들은 '4.3 실험결과'에서 제시한다. 마지막으로 세 번째 단계는 최적 군집수의 판단 단계로 각 CVI의 군집수 판단기준(최대값 또는 최소값)에 근거하여 실험문제별로 각 지수가 판단한 군집수를 확정하고, 실험문제에서 주어진 군집수와 각 지수의 확정된 최종 군집수를 비교하여 정확하게 군집수를 맞추었느냐로 CVI의 성능을 비교한다.

4.2 실험문제

4.2.1 재무분야 사례문제

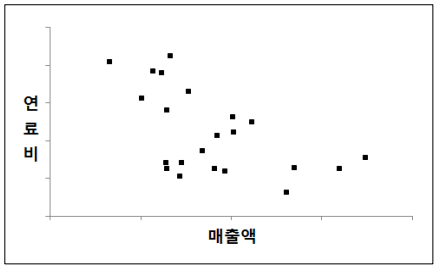
재무분야에서는 주식과 같은 투자기회에 대한 데이터를 이용하여 균형있는 투자항목을 선택하는 균형포트폴리오(balanced portfolio)와 기업의 성장률, 수익성, 시장크기, 생산규모 등과 같은 기업의 특징을 나타내는 데이터에 근거하여 같은 군집에 속한 기업들의 산업구조를 이해하는 산업분석(industry analysis)에 군집분석을 활용하고 있다[16].

본 연구에서는 산업분석에 활용되는 소규모 사례문제[2]를 이용하여 CVI의 성능을 비교하고자 한다. 이 데이터는 미국에서 전기와 가스 등을 공급하는 22개 공공전력기업체의 연간전력판매 매출액과 총연료비에 관한 데이터로, 이 사례문제는 22개 기업체를 군집분석하고 각 군집의 대표 기업체를 결정하여 '군집별 규제완화에 대한 비용예측 연구'에 사용하는 것을 목표로 한 것이다. 물론, 개별 기업체별로 비용예측 모형을 구축할 수도 있지만, 군집별로 대표 기업체를 선정하고, 이를 대상으로 비용 예측 모형을 구축하면 시간과 노력을 줄일 수 있다.

[그림 1]은 22개 공공전력기업의 매출액과 연료비의 데이터 형상(shape)을 나타낸 것이다. 이 데이터는 길죽한 임의형상의 형태로, 각 군집은 군집의 크기 및 형태에 차이가 존재하는 비대칭 분포의 특성이

2) 이 사례는 '조재희, 조성배, 이성임, 신현정, 김성범 공역, [비지니스 인텔리전스를 위한 데이터 마이닝], 서울 : 이앤비플러스, 2판, 2012.'의 pp.330-332에서 인용함. 이 사례문제는 각 기업별로 고정비용부담율, 투자수익률 등 8개 변수로 구성되어 있으나, 본 논문에서는 이중 2개의 변수(연간 전력판매 매출액과 총 연료비)만을 사용하기로 함.

존재하나 비교적 군집의 분리가 용이한 형태를 지니고 있으며 알려진 최적 군집 수는 3개이다. 이 데이터를 직관적으로 나누어 보면, 전체 기업체를 2개 그룹으로 볼 수도 있고, 3개 그룹으로 볼 수도 있다. 분석 대상이 되는 전체 기업체들을 몇 개의 군집으로 나누느냐에 따라 만들어야 하는 비용모형의 수도 달라지고, 모형의 구축을 위한 시간과 노력에도 차이가 발생한다. 대표할 기업체의 수를 선정하기 위해서는 군집분석을 통해 정확한 군집수 추정이 요구된다. 따라서 정확한 군집수를 추정하는 것은 산업분석에서 중요한 문제이다. 본 논문에서는 Fuzzy K-means 알고리즘을 통해 도출된 군집수가 최적인지의 판단을 앞에서 언급한 6개 CVI들을 통해 수행한다. 하나의 군집화 결과에 대해서 각 CVI별로 서로 다른 최적 군집수 판단이 이루어질 수 있으므로, 사전에 주어진 군집수와 CVI별로 판단된 최적 군집수를 비교함으로써 CVI들의 유효성과 성능을 확인해 볼 수 있다.

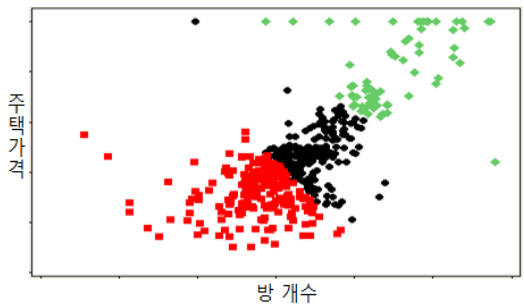


[그림 1] 공공전력기업문제의 데이터 형상

4.2.2 마케팅 분야 사례문제

마케팅 분야에서는 경쟁 유사도와 관련한 데이터를 이용하여 유사한 제품을 식별하는 시장구조분석(market structure analysis)과 고객 정보에 기초하여 고객 집단을 세분화하는 고객세분화(segmentation) 등에 군집분석을 활용하고 있다[13, 16]. 시장구조분석에 의해 도출된 결과는 제품이나 서비스의 잠재적인 요구를 추정하는 데 사용될 수 있고, 고객 세분화에 의한 결과는 성향이 유사한 고객군을 도출하여 고객 집단별로 특화된 마케팅 전략을 수립하는데 이용될 수 있다.

본 연구에서는 시장구조분석과 고객세분화를 위한 2가지 사례문제를 이용하여 CVI의 성능을 비교하고자 한다. 시장구조분석에 활용되는 사례 데이터[3]는, 보스턴 근교 구역별 주택들의 평균 방개수와 주택가격의 중앙값에 관한 데이터를 이용한다. 이 사례문제는 506개 구역을 군집분석하고 유사 가격대를 형성하는 구역을 식별하여 '미래주택가격을 예측하는 연구'에 사용하는 것을 목표로 한 것이다.



[그림 2] 주택가격문제의 데이터 형상

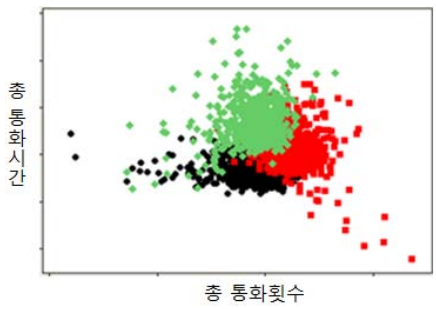
[그림 2]는 시장구조분석을 위한 보스턴 근교의 506개 구역별주택의 평균 방개수와 해당 구역에 있는 주택가격의 중앙값에 대한 데이터 형상을 나타낸 것으로, 길쭉한 타원형의 형태를 지닌다. 이 데이터는 군집 간의 경계가 모호하고, 부분군집 특성과 함께 소속 군집에서 멀리 떨어진 노이즈 객체가 존재하는 특성을 지니고 있어 정확한 군집수를 추정하기 어려운 특성을 지니고 있으며, 군집수는 3개라고 본다. 데이터의 특성을 바탕으로 거리상으로는 가깝지만 서로 다른 그룹에 존재하는 객체를 구분하고, 거리상으로 멀리 떨어져 있지만 유사한 특성을 지닌 그룹에 객체를 배정함으로써, 정확한 군집수를 추정하는 것은 시장구조분석에서 중요한 문제이다.

또 하나의 마케팅 분야 사례문제로 사용하는 고객세분화의 사례문제[4]는 국내 한 이동통신회사의 3,237명

3) 이 사례는 '조재희, 조성배, 이성임, 신현정, 김성범 공역, [비지니스 인텔리전스를 위한 데이터 마이닝], 서울 : 이앤비플러스, 2판, 2012.'의 pp.46-52에서 인용함. 이 사례문제는 각 구역별로 범죄율, 저소득층 비율 등 14개 변수로 구성되어 있으나, 본 논문에서는 이중 2개의 변수(평균 방 개수와 주택가격의 중앙값)만을 사용하기로 함.

고객들의 주간 총통화횟수와 주간 총통화시간에 관한 데이터를 이용하고자 한다. 이 사례문제는 국내 이동통신 가입자의 통화횟수와 통화시간을 고객들의 통화행태(behavior)로 규정하고, 이를 기준으로 유사한 통화행태를 보이는 고객들을 군집하여 '이동통신요금제개발 연구'의 기초 자료로 사용하는 것을 목표로 한다.

[그림 3]은 휴대전화 가입자들의 통화횟수와 통화시간의 데이터 형상을 나타낸 것으로 군집간의 분리도가 약하며, 소속 군집에 속한 객체간의 거리보다 다른 군집에 속한 객체와의 거리가 가까운 특이점 객체가 존재하는 부분군집 특성을 지니고 있고, 군집수는 3개라고 본다. 데이터의 특성을 바탕으로 분리도가 약한 군집들을 구분하고, 다른 군집의 영역내에 존재하는 이상치 객체를 구분하여, 정확한 군집수를 추정하는 것은 고객세분화에서 중요한 문제이다.



[그림 3] 휴대전화고객문제의 데이터 형상

4.2.3 생산관리분야 사례문제[5)]

생산관리분야에서는 불량품을 유발시키는 조업패턴 및 공정변수에 대한 데이터를 바탕으로 결함의 종류를 결정하여 공정시스템에 반영하는 품질개선(quality improvement)에 군집분석을 활용하고 있다[14]. 본 연구에서는 품질개선을 위한 반도체 웨이퍼 결함 검사(wafer defect test) 사례문제를 이용하여 CVI의 성능을 비교하고자 한다. 품질개선에 활용되는 사례문제는 웨이퍼 표면에 나타나는 결함의 위치 및 형태에 관한 데이터를 이용한다. 이 사례문제는 결함의 형태를 추출하고 결함들이 군집을 형성하는지를 판단하여, 불량의 원인이 유사한 결함을 식별하고 결함과 관련된 공정 및 시스템을 인식함으로써 '불량품을 유발시키는 조업개선 및 수율증대를 위한 공정변수 발견 연구'에 사용하는 것을 목표로 한 것이다.

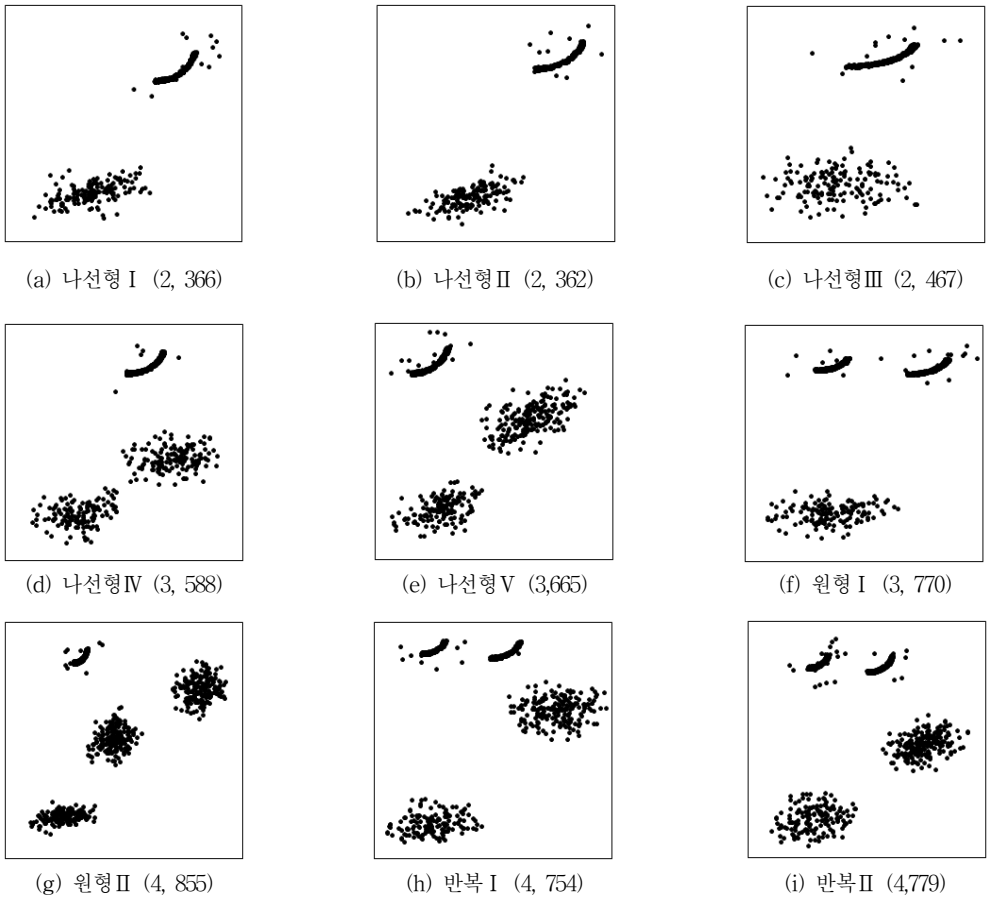
[그림 4]는 본 연구에서 사용된 9가지 웨이퍼의 결함 데이터 형상을 나타낸 것으로, 그림에서 x축과 y축은 웨이퍼 표면의 가로와 세로 위치를 나타낸 것이며, 그 값이 의미를 갖지 않는다. 웨이퍼 결함은 크게 나선형과 원형 그리고 반복의 형태로 나뉜다. [그림 4]의 (a)부터 (e)는 나선형이고, (f)와 (g)는 원형, (h)와 (i)는 반복의 결함형태를 지니고 있고, 9개의 문제 모두 노이즈 특성을 지니고 있다. 노이즈 특성 외에도 나선형의 결함형태를 가진 데이터 중에서 (a), (b)와 (e)는 군집간의 분리양호 특성을 지니고 있고, (c)는 군집간의 분리는 양호하나 비대칭 분포 특성을 지니며, (d)는 부분군집 특성을 지닌다. 원형의 결함형태를 가진 데이터 중에서 (g)는 분리가 양호한 특성을 지닌다. 마지막으로 반복형태를 가진 데이터에서 (i)는 대칭의 특성을 지니고 있다. 각 문제에서 알려진 군집수와 데이터 수는 <표 2>에 제시되어 있다. 예를 들어 [그림 4]의 실험문제 (a)는 <표 2>의 P4로 표시되어 있고, 결함수 366개의 데이터로 이루어진 문제이고 알려진 군집수는 2이다. 데이터의 특성을 바탕으로 결함이 군집을 형성하는지를 판단하고, 정확한 군집수를 추정하는 것은 품질개선에서 중요한 문제이다.

4.3 실험결과

SVDD 개념이 반영된 SVDU, SVCH, SVDB 지

4) 이 사례는 '조성배(2012)의 미니탭 교육자료(http://sclab.yonsei.ac.kr/index.php)'에서 제공하는 교육자료를 인용함. 이 사례문제는 각 고객별로 요금제, 주야간별 통화횟수와 통화시간 등 32개 변수로 구성되어 있으나, 본 논문에서는 이중 2개의 변수(주간 총통화횟수와 주간 총통화시간)만을 사용하기로 함.

5) 이 사례는 'Shin, K.S., Y.S. Jeong, M.K. Jeong, "A two-leveled symbiotic evolutionary algorithm for clustering problems," *Applied Intelligence*, Vol.36, No. 4(2012), pp.788-799'에서 인용함.



(a) 나선형Ⅰ (2, 366) (b) 나선형Ⅱ (2, 362) (c) 나선형Ⅲ (2, 467)

(d) 나선형Ⅳ (3, 588) (e) 나선형Ⅴ (3,665) (f) 원형Ⅰ (3, 770)

(g) 원형Ⅱ (4, 855) (h) 반복Ⅰ (4, 754) (i) 반복Ⅱ (4,779)

[그림 4] 웨이퍼 결함문제의 데이터 형상

수값을 계산하기 위해서는 각 군집별로 최적의 경계선을 구하기 위한 매개변수의 최적화 과정이 필요하다. 실험문제별로 예비실험을 통해 선정된 시그마(σ)와 에러수준(f)의 매개변수 값은 <표 2>에 제시되어 있다. 한편, 나머지 세 개의 CVI들은 매개변수가 존재하지 않는다. <표 2>에는 각 문제별 데이터수와 군집수도 함께 제시되어 있다.

공공전력기업문제의 실험결과는 <표 3>에 제시되어 있다. 표에서 음영으로 표시된 열(column)의 군집수는 각 문제별로 제시된 최적 군집수를 의미하고, '*'가 표시된 셀(cell)의 군집수는 실험결과 각 CVI가 판단한 최적 군집수를 의미한다. 따라서 각 표의 음영으로 표시된 열에 '*'가 표시된 셀이 존재하는 CVI는 그 문제에서 정확한 군집수를 추정하였다고 판단된다. 이 사례문제의 경우 알려져 있는 군집수는 3이기 때문에 DU, SVDU, SVCH, SVDB 지수는 정확한 군집수를 추정하였다고 해석할 수 있다.

[그림 1]에서 보는 바와 같이, 이 사례문제의 각 군집은 비대칭 분포와 임의형상의 형태를 지니고 있다. 일반적으로 SVDD 개념이 반영되면 CH 지수는 노이즈와 임의형상 특성을 지닌 데이터에서 성능이 향상되고, DB 지수는 임의형상 특성을 지닌 데이터에서 성능이 향상된다고 알려져 있고[11], 실험결과를 통해 이를 확인할 수 있었다. 본 사례문제는 비대칭 분포 특성을 지니고 있지만 분리가 양호하다는 특성을 동시에 지니고 있어, 비대칭분포 특성을 지니고

〈표 2〉 사례문제별 설명과 매개변수

실험문제				매개변수	
번호	설명	데이터수	군집수	σ	f
P1	공공전력기업(임의형상, 비대칭, 분리양호)	22	3	4	0.1
P2	주택가격(타원형, 부분군집, 노이즈)	506	3	5	0.07
P3	휴대전화고객(원형, 특이점, 부분군집)	3,237	3	4	0.07
P4	나선형 웨이퍼 결함Ⅰ(분리양호, 노이즈)	366	2	5	0.03
P5	나선형 웨이퍼 결함Ⅱ(분리양호, 노이즈)	362	2	3	0.02
P6	나선형 웨이퍼 결함Ⅲ(분리양호, 비대칭, 노이즈)	467	2	6	0.04
P7	나선형 웨이퍼 결함Ⅳ(부분군집, 노이즈)	588	3	8	0.01
P8	나선형 웨이퍼 결함Ⅴ(분리양호, 노이즈)	665	3	3	0.02
P9	원형 웨이퍼 결함Ⅰ(노이즈)	770	3	3	0.01
P10	원형 웨이퍼 결함Ⅱ(분리양호, 노이즈)	855	4	2	0.02
P11	반복 웨이퍼 결함Ⅰ(노이즈)	754	4	4	0.02
P12	반복 웨이퍼 결함Ⅱ(대칭, 노이즈)	779	4	7	0.01

〈표 3〉 공공전력기업 사례문제의 결과

CVI	군집수(최적 군집수 : 3)								
	2	3	4	5	6	7	8	9	10
DU	2.542	2.789*	2.257	1.378	1.920	1.165	1.310	0.435	0.000
CH	30.39	38.38	45.53	39.52	45.92	38.58	38.68	33.86	60.17*
DB	0.626	0.620	0.568	0.636	0.495	0.437	0.481	0.433	0.313*
SVDU	210.7	411.5*	105.7	30.4	36.7	33.3	27.5	6.1	0.0
SVCH	64714	519666*	77904	16573	14014	28446	12919	4606	5619
SVDB	0.007	0.001*	0.005	0.010	0.013	0.012	0.021	0.020	0.022

있음에도 불구하고 SVDD 개념이 반영에 의해 CH 지수의 성능이 향상된 것으로 보인다.

주택가격문제의 실험결과는 <표 4>에 제시되어 있다. 실험결과, CH와 DB 지수는 군집수를 맞추지 못했으나, SVCH와 SVDB 지수는 최적 군집수를 맞추었다. 또한, DU와 SVDU 지수는 모두 최적 군집수를 맞추지 못하였으며, 두 지수가 최적이라고 판단한 군집수도 동일하다. 따라서 CH와 DB 지수는 SVDD 개념이 반영된 후에 그 성능이 향상되었으며, DU 지수는 SVDD 개념의 반영 전후에 성능의 차이를 보이지 않는다고 말할 수 있다. [그림 2]에서 보는 바와 같이, 사례문제의 각 군집에는 집중적으로 데이터가 분포되어 있는 공간에서 멀리 떨어져 있는 데이터들이 존재한다. 이 객체들을 각 군집에 정확하게 배정하면, 개별 군집의 모양은 매끄러운 원형이 아닌, 길죽하고 불규칙한 모양이 된다. SVDD 개념은 이러한 군집 형태를 갖는 군집화 결과의 유효성 검증에 적합하다고 알려져 있고[11], 이는 본 사례를 통해 재확인되었다.

휴대전화 통화행태문제의 실험결과는 <표 5>에 제시되어 있다. 실험결과는 앞의 두 사례문제와 유사하게 나왔다. 즉, CH와 DB 지수는 SVDD 개념을 반영하여 성능이 향상되었고, DU 지수는 SVDD 개념의 반영 전후에 성능의 차이를 보이지 않았다. 그러나 이 사례 문제의 결과는 다음과 같은 시사점들을 포함하고 있다. 첫째, 휴대전화통화 행태의 사례문제

〈표 4〉 주택가격 사례문제의 결과

CVI	군집수(최적 군집수 : 3)								
	2	3	4	5	6	7	8	9	10
DU	2.496*	1.167	1.385	1.015	1.102	0.972	0.724	0.595	0.687
CH	675.1*	606.5	644.4	582.7	618.8	576.9	533.3	498.6	624.4
DB	0.673	0.73	0.711	0.733	0.652*	0.715	0.749	0.836	0.782
SVDU	115.1*	84.44	63.18	27.74	35.64	30.56	22.56	23.48	25.62
SVCH	28861	59618*	54014	26395	28976	30108	26891	30160	34138
SVDB	0.014	0.007*	0.008	0.014	0.017	0.014	0.016	0.019	0.017

는 본 연구에서 지금까지 다루었던 실험문제들 중에서 가장 데이터의 수가 많은 문제에 해당한다. 이 실험문제에서도 CVI에 SVDD 개념을 적용한 것이 효과적임을 확인하였으므로, 데이터 수가 증가하더라도 SVDD 개념을 적용한 CVI들은 좋은 성능이 보장됨을 예상할 수 있다. 물론, 이러한 결론이 일반화되기 위해서는 더욱 많은 데이터양의 실험 문제를 이용한 성능 실험이 추가되어야 할 것으로 생각된다. 둘째, [그림 3]에서 보는 바와 같이, 이 사례문제는 데이터들간의 밀집도가 매우 높다. 또한, 실제로 군집화된 결과를 살펴보면, 2차원 평면상에서 서로 다른 군집에 배정되어 있는 데이터들이 같은 군집에 배정되어 있는 것처럼 보이는 문제로써 군집화 유효성을 판단하는데 어려운 문제에 해당한다. 그럼에도 불구하고 SVDD의 개념을 반영하면 CH와 DB 지수의 성능이 향상되는 것을 확인할 수 있다. 본 사례처럼 현실 세계에서의 군집화 문제는 선형 분리보다는 비선형 분리에 해당되는 경우가 더욱 많다. SVDD의 커널함수는 선형 분리가 가능하지 않은 문제를 선형 분리 문제로 만든다는 장점을 갖는다. 이러한 SVDD의 특징이 휴대전화 통화 행태의 사례문제에서 SVDD 개념을 반영한 CVI가 좋은 성능을 보이는데 기여한 것으로 생각된다.

웨이퍼 결함 문제의 실험결과는 <표 6>부터 <표 14>에 제시되어 있다. 실험결과, CH 지수는 SVDD 개념이 반영된 후에 성능이 크게 향상되었고, DB 지수는 SVDD 개념을 반영한 전후에 성능의 변화가 거의 없었다. 또한, DU 지수는 SVDD 개념을 반영한 후에 성능의 변동이 있었다.

<표 6>부터 <표 14>를 요약하면, <표 15>와 같다. <표 15>는 알려져 있는 문제의 군집수와 각 지수가 추정한 군집수 간의 차이를 정리한 것으로 '0'은 정확한 군집수 추정을 의미하고, 오차 값의 양과 음은 알려진 군집수에 비해 더 많은 군집수로 추정했는지 적게 추정했는지를 의미한다. 군집수 추정 정확도 및 추정 오차 측면에서 살펴보면 SVCH, SVDB, DB 지수의 성능이 우수하였고, CH 지수가 가장 낮은 성능을 보였다. SVDU 지수는 원형의 결함형태에서

〈표 5〉 휴대전화 통화행태 사례문제의 결과

CVI	군집수(최적 군집수 : 3)								
	2	3	4	5	6	7	8	9	10
DU	2.057*	1.320	1.045	0.941	0.792	0.790	0.702	0.647	0.641
CH	4400	4291	4033	4151	4146	4361	4437	4457	5741*
DB	0.745	0.760	0.835	0.791	0.752	0.704	0.666	0.703*	0.689
SVDU	52.95*	31.75	27.53	36.72	22.97	24.38	18.73	16.03	14.38
SVCH	6988	7315*	8365	11926	9026	9967	8036	7550	7311
SVDB	0.022	0.024*	0.026	0.024	0.025	0.021	0.024	0.028	0.027

〈표 6〉 나선형 웨이퍼 결함 I 의 결과

CVI	군집수(최적 군집수 : 2)								
	2	3	4	5	6	7	8	9	10
DU	6.706*	2.167	1.945	2.259	1.229	1.723	1.518	1.022	1.043
CH	4218	3222	3508	4216	3630	4085	3668	3454	4412*
DB	0.250*	0.440	0.505	0.489	0.492	0.541	0.590	0.523	0.551
SVDU	101.1*	21.39	16.42	17.89	8.160	7.068	6.798	4.282	4.129
SVCH	1303*	8102	5396	5313	4164	2189	1850	1473	1382
SVDB	0.011*	0.034	0.047	0.047	0.060	0.077	0.104	0.129	0.130

〈표 7〉 나선형 웨이퍼 결함 II 의 결과

CVI	군집수(최적 군집수 : 2)								
	2	3	4	5	6	7	8	9	10
DU	7.777*	2.117	2.022	1.6291	1.637	1.542	1.319	1.056	0.860
CH	5354	4097	4544	4758	4910	5299	4704	4543	5632*
DB	0.223*	0.422	0.508	0.474	0.576	0.556	0.581	0.580	0.606
SVDU	36.87*	7.171	5.010	4.160	3.557	2.676	2.136	1.611	1.345
SVCH	1945*	9014	518	498	462	293	213	209	219
SVDB	0.033*	0.124	0.173	0.187	0.216	0.253	0.341	0.384	0.373

〈표 8〉 나선형 웨이퍼 결함 III 의 결과

CVI	군집수(최적 군집수 : 2)								
	2	3	4	5	6	7	8	9	10
DU	4.447*	2.093	1.764	1.212	1.323	1.236	1.456	0.741	0.787
CH	2575	1972	2188	2305	2095	2230	2336	2063	2615*
DB	0.365*	0.502	0.569	0.499	0.608	0.592	0.672	0.582	0.582
SVDU	141*	43	37	23	19	17	19	11	8
SVCH	27912*	17980	19912	17768	10407	11189	8353	12902	5915
SVDB	0.007*	0.016	0.018	0.016	0.030	0.024	0.033	0.028	0.037

〈표 9〉 나선형 웨이퍼 결함 IV 의 결과

CVI	군집수(최적 군집수 : 3)								
	2	3	4	5	6	7	8	9	10
DU	2.857	4.299*	1.882	1.864	1.695	1.720	1.083	0.932	1.129
CH	1042	3727	3163	3132	3576	3825	3566	3631	4548*
DB	0.633	0.346*	0.475	0.509	0.542	0.548	0.591	0.583	0.598
SVDU	164.3	164.7*	68.9	45.7	31.7	25.4	13.5	11.2	11.0
SVCH	39581	57456*	51206	41404	23678	16516	12694	13085	11275
SVDB	0.007	0.006*	0.010	0.014	0.016	0.008	0.022	0.023	0.023

〈표 10〉 나선형 웨이퍼 결함Ⅴ의 결과

CVI	군집수(최적 군집수 : 3)								
	2	3	4	5	6	7	8	9	10
DU	1.72	4.80*	1.76	1.44	1.32	1.57	1.38	1.29	1.42
CH	787	3821	3472	3242	3472	3754	3506	3465	4655*
DB	0.694	0.33*	0.42	0.50	0.51	0.52	0.61	0.62	0.59
SVDU	41.86	48.84*	13.10	11.20	7.28	7.84	6.58	6.64	5.86
SVCH	3098	3375*	2157	2010	1194	1244	1266	1265	994
SVDB	0.028	0.026*	0.041	0.059	0.073	0.075	0.099	0.090	0.099

〈표 11〉 원형 웨이퍼 결함Ⅰ의 결과

CVI	군집수(최적 군집수 : 3)								
	2	3	4	5	6	7	8	9	10
DU	2.16	3.91*	1.39	1.41	1.23	1.34	1.18	0.96	1.26
CH	907	7752	6486	8165	7330	8398	9430	8654	11081*
DB	0.66	0.23*	0.33	0.39	0.44	0.47	0.46	0.50	0.45
SVDU	67.52*	42.36	9.47	7.08	6.51	4.66	2.98	3.77	2.50
SVCH	5272	6661*	2667	2295	2049	1178	947	1368	604
SVDB	0.02	0.02*	0.05	0.06	0.07	0.11	0.12	0.14	0.13

〈표 12〉 원형 웨이퍼 결함Ⅱ의 결과

CVI	군집수(최적 군집수 : 4)								
	2	3	4	5	6	7	8	9	10
DU	1.91	2.50	6.06*	1.35	1.41	1.37	1.25	1.40	1.55
CH	793	1661	9878	8489	7838	7633	7157	7116	10204*
DB	0.948	0.500	0.238*	0.453	0.549	0.555	0.566	0.579	0.572
SVDU	19.33*	17.17	15.29	3.27	2.98	2.32	2.22	2.32	2.36
SVCH	465.4	467.5	564.9*	422.3	343.5	256.5	299.8	214.1	269.0
SVDB	0.08	0.06	0.06*	0.13	0.18	0.22	0.21	0.27	0.24

〈표 13〉 반복 웨이퍼 결함Ⅰ의 결과

CVI	군집수(최적 군집수 : 4)								
	2	3	4	5	6	7	8	9	10
DU	3.30*	2.95	3.23	1.66	0.88	0.87	0.97	0.95	1.12
CH	1206	2577	5886	5810	5009	5165	5309	5313	6608*
DB	0.42	0.42	0.27*	0.40	0.46	0.49	0.56	0.55	0.59
SVDU	79.25*	46.66	34.57	9.40	4.42	4.37	4.33	3.54	3.96
SVCH	6446	6180	7102*	1921	1629	1660	1118	945	1013
SVDB	0.019	0.018	0.017*	0.061	0.077	0.096	0.112	0.122	0.132

〈표 14〉 반복 웨이퍼 결함II의 결과

CVI	군집수(최적 군집수 : 4)								
	2	3	4	5	6	7	8	9	10
DU	1.67	3.30*	3.07	1.41	1.58	0.99	1.17	1.12	1.32
CH	801	2383	5595	5457	5560	5619	5457	5197	6045*
DB	0.92	0.40	0.28*	0.37	0.45	0.47	0.54	0.57	0.63
SVDU	109.1	117.7*	79.3	29.3	28.1	12.2	15.6	19.0	14.9
SVCH	21052	27348	34834*	16480	21895	11590	11302	14742	10074
SVDB	0.011	0.009	0.007*	0.018	0.019	0.026	0.031	0.030	0.042

〈표 15〉 웨이퍼 결함 검사 문제의 CVI별 군집수 추정 오차

실험문제	CVI					
	DU	CH	DB	SVDU	SVCH	SVDB
P4	0	8	0	0	0	0
P5	0	8	0	0	0	0
P6	0	8	0	0	0	0
P7	0	7	0	0	0	0
P8	0	7	0	0	0	0
P9	0	8	0	-1	0	0
P10	0	6	0	-2	0	0
P11	-1	6	0	-2	0	0
P12	-1	6	0	-1	0	0
오차 절대값의 총합	2	63	0	6	0	0

DU 지수보다 근소하게 성능의 하락을 보였고, 반복 결함형태에서는 DU와 SVDU 지수 모두 군집수 추정에 실패하였다. 이 사례 문제는 크게 웨이퍼 결함의 유형에 따라 전체 데이터의 형상을 나선형, 원형, 반복으로 분류하고, 각 결함형상 별로 군집 간의 형태와 개별 데이터에 따라 비대칭, 타원, 부분군집, 대칭, 분리양호 등의 다양한 데이터 특성을 지니고 있다. 이는 전체 데이터의 형상, 군집 간의 형태, 객체 특성으로 보다 세밀하게 실험문제의 특성을 구분한 것이다. 본 실험문제에 따르면 SVCH 지수는 전체 데이터의 형상, 개별 군집의 형태 및 데이터 특성에 관계없이 CH 지수과 비교하였을 때, 성능이 향상됨을 확인하였다. 이는 노이즈 특성에 강한 SVDD의 특징이 반영된 결과로 보인다. 다만, 앞의 사례문제 실험결과에서 나타났듯이, DU 지수에서는 SVDD 개념 반영에 의한 성능 변화가 뚜렷하지 않다. 이를 밝히기 위해서는 CVI의 수리적 측면이나 군집분석을 위한 데이터들의 형상적 측면을 좀 더 면밀하게 분석할 필요가 있다.

5. 결론 및 토의

최근 빅 데이터의 활용 영역이 크게 확대되고 있다는 측면에서 빅 데이터의 분석기술은 학술적으로나 실무적으로 충분히 연구의 가치가 있다고 판단되는 분야이다. 본 연구는 빅 데이터의 분석기술 중 군집분석에서 필요한 군집화 결과의 유효성을 검증하는 CVI의 실무적 활용을 다루었다. 일반적으로 빅 데이터는 그 양이 방대하고 정보의 변화속도가 빠르기 때문에 데이터에 대한 특별한 정보가 주어져 있지

않다. 따라서 빅 데이터를 이용하여 군집분석하고 군집결과로부터 의사결정에 활용 가능한 정보를 도출하기 위해서는 다양한 데이터의 구조 및 패턴에 이용할 수 있는 CVI가 요구된다. 본 연구는 복잡한 형태 및 다양한 특성을 지닌 데이터를 이용하여 CVI의 성능비교 및 실무적 활용을 다루었으며, 연구결과들을 정리하면 다음과 같다.

첫째, 성능이 좋다고 알려진 CVI들이 경영학 분야의 실무사례 문제에서도 적용 가능한지 살펴보기 위하여, 재무분야(산업구조분석), 마케팅 분야(시장구조분석, 고객세분화), 생산관리 분야(품질개선)의 사례문제를 선정하고, 유연성이 높은 Fuzzy K-means 알고리즘을 사용하여 CVI들의 성능을 비교분석한 결과, SVCH와 SVDB 지수, SVDU과 DU 지수 순으로 성능이 우수함을 확인하였다. 특히, SVCH와 SVDB 지수는 매끄러운 원형이 아닌 길죽한 모양, 그리고 서로 불규칙하며 특별한 형태로 지정하기 어려운 경우에서 CH와 DB 지수에 비해 성능이 우수하였다. 즉, 전체 데이터의 형상 및 군집간의 형태 그리고 데이터 특성을 구분한기 어려운 경우에 적용이 가능함을 확인하였다. 따라서 SVDD 기반의 CVI들은 본 연구에서 다루었던 문제에서 군집화 결과의 유효성을 판단하는데 보다 적합하다고 판단된다. 둘째, SVDD 기반의 CVI들은 데이터 수가 증가하더라도 좋은 성능이 보장됨을 말할 수 있었으며, 데이터들 간의 밀집도가 매우 높은 문제에서도 좋은 성능을 보였다. 셋째, SVDD 기반의 CVI들은 노이즈 특성을 지닌 데이터 문제에서 높은 성능향상을 보였다. 위의 결론들을 바탕으로 SVDD 기반의 CVI들은 형태가 알려져 있지 않고, 데이터의 양이 방대하다고 알려진 빅 데이터에 적용이 가능할 것으로 기대된다.

본 논문은 수치형 자료와 다양한 경영사례의 군집화 결과를 효과적으로 검증할 수 있는 CVI를 선별하고자 시작하였다. 그리고 본 연구의 결과로부터 SVDD 기반의 CVI와 기존 CVI의 성능을 비교함으로써 복잡한 형태의 데이터에 활용 가능한 CVI를 제안한 연구라는 점에서 의의가 있다. 그러나 빅 데이터의 활용이라는 측면에서 보았을 때, 본 연구에서 성능비교 실험에 사용한 데이터 양은 많지 않다. 이는 분석결과를 명확하게 확인하기 위하여, 데이터나 군집의 형태, 그리고 군집수가 주어진 수치형 데이터를 대상으로 실험문제를 추출하였기 때문이다. 추가적으로 데이터 양이 많고, 군집수가 주어지지 않은 문제에서도 새롭게 제안한 CVI들이 좋은 성능을 보이는지 면밀하게 살펴볼 필요가 있다.

참고문헌

[1] 강지혜, 김성수, “적응적인 초기치 설정을 이용한 Fast K-means 및 Fuzzy c-means 알고리즘”, 「정보과학회논문지」, 제31권, 제4호(2004), pp. 516-524.

[2] 고정원, 최병인, 이정훈, “노이즈에 강한 밀도를 이용한 Fuzzy C-means 클러스터링 알고리즘”, 「한국퍼지 및 지능시스템학회 추계학술대회」, 제16권, 제2호(2006), pp.211-214.

[3] 김민호, R.S. Ramakrishna, “비형식의 군집 유효화 지수의 분석과 새로운 지수 개발”, 「한국컴퓨터종합학술대회」, 제32권, 제1(B)호(2005), pp.601-603.

[4] 김상락, 강만모, 박상무, “빅데이터가 여는 미래의 세상, 「정보과학회지」, 제30권 제6호(2012), pp.18-24

[5] 김성우, 김각규, 윤봉규, “국방분야 빅 데이터 분석의 활용가능성에 대한 고찰”, 「한국경영과학회지」, 제39권, 제2호(2014), pp.1-19.

[6] 김정숙, “빅 데이터 활용과 관련기술 고찰”, 「한국콘텐츠학회지」, 제10권, 제1호(2012), pp.34-40.

[7] 민재형, 송영민, “국내 생명보험회사의 재무건전성 평가 : ELECTREⅡ, 단순가중합 모형, 군집분석의 비교”, 「한국경영과학회지」, 제28권, 제4호(2003), pp.39-60.

[8] 민재형, 이영찬, “Support Vector Machine을 이용한 부도예측 모형의 개발 : 격자탐색을 이용한 커널 함수의 최적 모수 값 선정과 기존 부도예측 모형과의 성과 비교”, 「한국경영과학회지」, 제30

권, 제1호(2005), pp.55-74.

[9] 안현철, 김경재, 한인구, “Support Vector Machine을 이용한 고객구매예측 모형”, 「한국지능정보시스템학회논문지」, 제11권, 제3호(2005), pp.69-81.

[10] 오은녕, 이희상, “클러스터링 기법을 이용한 이동통신의 고객 세분화 연구”, 「한국경영과학회 추계논문집」, (2002), pp.421-424.

[11] 이수현, “빅 데이터의 군집분석을 위한 군집화 유효성 지수 개발과 응용”, 「전남대학교 일반대학원」, 박사학위논문(2015).

[12] 이준호, 박광호, “군집분석을 통한 중소기업 온라인 마케팅 지원 수혜기업의 세분화 전략에 관한 연구”, 「e-비지니스연구」, 제13권, 제4호(2012), pp.169-194.

[13] 전현치, 신영근, 박상성, 김명훈, 장동식, “신경망 기법을 이용한 온라인 서점 이용자들의 고객유형 분석”, 「환경콘텐츠학회논문지」, 제7권, 제9호(2007), pp.127-138.

[14] 전치혁, 「데이터마이닝 기법과 응용」, 서울 : 한나래, 2012.

[15] 정윤경, 백장선, “고차원(유전자 발현) 자료에 대한 군집 타당성분석 기법의 성능 비교”, 「응용통계연구」, 제20권, 제1호(2007), pp.167-181.

[16] 조재희, 조성배, 이성임, 신현정, 김성범 공역, 「비지니스 인텔리전스를 위한 데이터 마이닝」, 서울 : 이앤비플러스, 2판, 2012.

[17] 황인수, “데이터 마이닝에서 그룹 세분화를 위한 2단계 계층적 클러스터링 알고리듬”, 「경영과학」, 제19권, 제1호(2002), pp.189-196.

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

[19] Bezdek, J.C., “Pattern Recognition with Fuzzy Objective Function Algorithm,” *Plenum Press*, Vol.13(1981), pp.367-373.

[20] Calinski, R.B. and J. Harabasz, “A Dendrite Method for Cluster Analysis,” *Communications in Statistics*, Vol.3, No.1(1974), pp.1-27.

[21] Davies, D. and D. Bouldin, “A Cluster Separation Measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, No.2(1979), pp.224-227.

[22] Dunn, J.C., “Well Separated Clusters and Optimal Fuzzy Partitions,” *Journal of Cybernetics*, Vol.4, No.1(1974), pp.95-104.

[23] Halkidi, M., Y. Batistakis, M. Vazirgiannis, “On Clustering Validation Techniques,” *Journal of Intelligent Information Systems*, Vol.17, No.2-3(2001), pp.107-145.

[24] Hruschka, E.R., R.G.B. Campello, A.A. Freitas, and A.P.L. Carvalho, “A Survey of Evolutionary Algorithms for Clustering,” *IEEE Transactions on Systems, Man, and Cybernetics-Part C : Applications and Reviews*, Vol.39, No.2(2009), pp.133-155.

[25] Liu, Y., Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, “Understanding and Enhancement of Internal Clustering Validation Measures,” *IEEE Transactions on Cybernetics*, Vol.43, No.3(2013), pp.982-994.

[26] MacQueen, J.B., “Some Methods for classification and Analysis of Multivariate Observations,” *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, (1967), pp.281-297.

[27] Maulik, U. and S. Bandyopadhyay, “Performance Evaluation of Some Clustering Algorithms and Validity Indices,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.24, No.12(2002), pp.1650-1654.

[28] Saitta, S., B. Raphael, and I.F.C. Smith, “A Comprehensive Validity Index for Clustering,” *Intelligent Data Analysis*, Vol.12, No.6(2008), pp.529-548.

[29] Shim, Y., J. Chung, and I.-C. Choi, “A Com-

parison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm," *Proceedings of the International Conference on Computational Intelligence for Modeling, Control and Automation, and International Conference Intelligent Agents, Web Technologies and Internet Commerce*, (2005), pp.199-204.

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

[31] Tay, F.E.H. and L.J. Cao, "Modified support vector machines in financial time series forecasting," *Neurocomputing*, Vol.48, No.1-4(2006), pp.847-861.

[32] Theodoridis, S. and K. Koutroumbas, *Pattern Recognition*, Academic Press, 2006.

[33] Tou, J.T. and R.C. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley, 1974.

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

[35] Xu, R. and D.II. Wunsch, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, Vol.16, No.3(2005), pp.645-678.