경영정보학연구
제16권 제1호
2006년 3월

K-means 알고리즘 기반 클러스터링 인덱스 비교 연구*

심 요 성**, 정 지 원***, 최 인 찬****

A Performance Comparison of Cluster Validity Indices based on K-means Algorithm

Yo-Sung Shim, Ji-Won Chung, In-Chan Choi

The K-means algorithm is widely used at the initial stage of data analysis in data mining process, partly because of its low time complexity and the simplicity of practical implementation. Cluster validity indices are used along with the algorithm in order to determine the number of clusters as well as the clustering results of datasets. In this paper, we present a performance comparison of sixteen indices, which are selected from forty indices in literature, while considering their applicability to nonhierarchical clustering algorithms. Data sets used in the experiment are generated based on multivariate normal distribution. In particular, four error types including standardization, outlier generation, error perturbation, and noise dimension addition are considered in the comparison. Through the experiment the effects of varying number of points, attributes, and clusters on the performance are analyzed. The result of the simulation experiment shows that Calinski and Harabasz index performs the best through the all datasets and that Davis and Bouldin index becomes a strong competitor as the number of points increases in dataset.

Keywords : Data Mining, Cluster Analysis, Nonhierarchical Clustering, K-means, Cluster Validity Index

---

* 이 논문은 2004년도 한국학술진흥재단 지원에 의하여 연구되었음(KRF-2004-041-D00813).
** 고려대학교 산업시스템정보공학과 석사과정
*** 고려대학교 산업시스템정보공학과 박사과정
**** 교신저자, 고려대학교 산업시스템정보공학과 교수

Ⅰ. 서 론

데이터 마이닝은 대용량 데이터에서 의미 있는 규칙이나 패턴을 찾기 위한 데이터 탐색 및 분석 과정이다[Berry and Linoff, 2000]. 데이터 마이닝은 분류, 추정, 예측, 유사 그룹 및 연관 규칙, 군집화, 그리고 의사결정나무 등을 비롯한 다양한 기법들이 사용될 수 있으며, 마케팅을 비롯한 제품 개발, 프로세스 개선, 의사 결정 시스템 등에 폭 넓게 활용된다.

데이터 마이닝에서 사용되는 자율학습 기법의 하나인 클러스터링(Clustering)은 주어진 데이터를 유사한 성질을 갖는 그룹(클러스터)으로 나누는 기법이다. 분류(Classification) 기법이 미리 정의된 집합으로 데이터를 구분하는 것에 반해, 클러스터링은 자기유사성에 기반한 데이터 구분 기법으로 데이터가 갖는 속성값에 따라 각 그룹의 범위나 성격이 다르게 정의될 수 있다.

클러스터링 알고리즘은 크게 계층형과 비계층형 알고리즘으로 구분된다. 비계층형 알고리즘의 하나인 K-means 알고리즘은 우수한 시간 효율성과 구현의 용이성으로 인해 데이터 마이닝, 데이터베이스 지식탐색(Knowledge Discovery in Database) 등에 널리 활용되고 있다[Kurita, 1991; Day, 1992; Jain, Murty, and Flynn, 1999]. 그러나 K-means 알고리즘의 현실 적용에서 가장 큰 문제점은 분할될 클러스터 수(k)를 사전에 설정해주어야 한다는 것이다. 클러스터 수의 산정은 클러스터링 문제의 해 공간을 결정해주는 매우 중요한 문제로서, 주어진 대용량 다속성 데이터에 존재하는 클러스터 수를 사전에 인지하는 것은 사실상 불가능하기 때문에 클러스터링 알고리즘과 함께 클러스터링 인덱스(Cluster Validity Index, Clustering Index)를 사용하게 된다.

클러스터링 인덱스는 클러스터링 결과의 유효성을 평가하기 위해 사용되는 평가기준(Measure)이다. 주어진 범위의 다양한 클러스터 수에 대하여 클러스터링을 구하고, 그 결과로부터 계산된 인덱스 값에 근거하여 가장 선호되는 클러스터 수를 선택하게 된다. 현재까지 통계적 기법 또는 직관적·경험적 사실에 기반한 다양한 인덱스들이 제시되어 있으나, 모든 데이터 구조와 알고리즘에 대해 절대적으로 통용되는 인덱스는 존재하기 어렵다. 따라서, 사용자는 데이터 구조에 따라 적절한 인덱스를 선택적으로 사용하거나 하나 이상의 인덱스를 함께 사용하기도 한다. 이와 같은 현실 적용의 문제점을 극복하기 위하여 특정 데이터 구조와 알고리즘에 대한 인덱스의 성능 평가는 매우 중요하다. 특히, 데이터 마이닝에서와 같이 대용량 다속성의 데이터를 대상으로 하는 경우 인덱스의 부적절한 선택으로부터 발생된 결과를 인지하기 매우 어려울 뿐 아니라 다양한 인덱스의 시도 자체가 현실적으로 어렵다.

Milligan and Cooper[1985]는 계층형 알고리즘에 기반한 클러스터링 인덱스 성능 비교에 대한 연구 결과를 제시하였다. 실험에서는 4가지 계층형 클러스터링 알고리즘과 Milligan[1985]의 시뮬레이션 데이터 생성 알고리즘을 사용하였으며, 30개의 기존 클러스터링 인덱스들의 클러스터 복원 성능을 비교하였다. 실험 결과 Calinski and Harabasz[1974]와 Duda and Hart[1973]의 인덱스가 다른 인덱스들에 비해 좋은 결과를 나타냈으나 데이터 의존성이 상대적으로 강한 Duda and Hart 인덱스보다 Calinski and Harabasz 인덱스를 계층형 알고리즘에서 가장 적합한 인덱스로서 평가하였다. 그러나 데이터 개수 50개, 속성 개수 4, 6, 8개, 그리고 클러스터 개수 5개 이하인 표준데이터(Error Free Datasets) 만을 사용하였을 뿐 아니라, 현실 문제에서 발생할 수 있는 다양한 에러유형을 고려하지 않음으로써 제한적인 실험에 그쳤다.

Dimitriadou, Dolnicar, and Weingessel[2002]은 K-means를 이용하여 14개 인덱스에 대한 비교 연구 결과를 제시하였다. 이진 데이터만을 포

함한 16개의 데이터 셋을 사용하였으며 각 인덱스들의 성능이 클러스터 크기 등의 데이터 셋 특성에 따라 서로 상이한 결과를 나타냈다. Dimitriadou *et al.*[2002]의 실험은 이진 데이터만을 대상으로 하므로 연구 결과를 수치형 데이터를 포함한 일반적인 데이터 유형에 대하여 일반화하기는 어렵다. 이 외에도 클러스터링 인덱스 관련 기존 연구들이 각기 제안하는 인덱스의 유효성 검증을 위해 소규모의 비교 연구를 실시하고 있으나, 제한된 소수의 인덱스만을 비교함으로써 실험결과의 일반화에는 무리가 따른다[Rohlf, 1974; Ratkowsky and Lance, 1978; Davis and Bouldin, 1979; Milligan, 1980; Milligan, 1981; Ray and Turi, 1999].

이와 같이 K-means에 기반한 포괄적인 인덱스 성능비교 연구가 미흡하여 현실 문제에서 K-means 알고리즘을 적용하는 경우 적절한 인덱스 선택에 대한 가이드 라인을 찾기 어렵다. 따라서, 본 연구에서는 비계층형 기법에 적용이 가능한 인덱스를 대상으로 정규분포를 따르는 데이터에 대한 동질 클러스터링을 위하여 K-means 알고리즘을 이용한 성능 비교 실험을 수행한다. 실험을 통해 K-means 기반의 동질 클러스터링(Homogeneous Clustering)에서 인덱스들의 장단점을 살펴보고, 인덱스 선택에 대한 일반적인 가이드 라인을 제시한다. 또한, 실험 요인 변화에 따른 인덱스 성능 변화를 관찰함으로써, 시뮬레이션 데이터 생성 기법의 제한점과 함께 확장된 실험에서의 인덱스 수행도 변화를 분석한다.

본 논문은 다음과 같이 구성된다. Ⅱ장에서는 비계층형 클러스터링 알고리즘과 연구에 사용된 인덱스를 설명하며, Ⅲ장에서는 실험 데이터 생성 기법과 이에 따른 인덱스 성능 비교 실험을 수행하고 그 결과를 분석한다. 마지막으로 Ⅳ장에서는 결론 및 연구 결과의 적용 한계와 추후 연구 방향을 제시한다.

Ⅱ. 클러스터링 알고리즘과 인덱스

2.1 비계층형 클러스터링 알고리즘

클러스터링 알고리즘은 크게 계층형과 비계층형 알고리즘으로 구분될 수 있다. 계층형 클러스터링 알고리즘은 새로 병합되거나 분할될 대상이 되는 데이터 또는 부분 데이터 셋만을 이용하여 관련된 데이터들을 두 개의 그룹으로 분할할 것인지 또는 하나의 그룹으로 병합할 것인지를 판단하여 점진적으로 분할 또는 병합함으로써 전체 데이터를 클러스터링 하게 된다. 반면, 비계층형 클러스터링 알고리즘은 전체 데이터를 대상으로 사전에 주어진 클러스터 수로 분할하는 과정을 반복적으로 수행하여, 매 반복 단계마다 전체 데이터의 분할을 재설정하게 된다.

MacQueen[1967]에 의해 제시된 K-means 알고리즘은 동질클러스터링을 위한 대표적인 비계층형 알고리즘으로서, 주어진 클러스터 수(k)에 대하여 데이터의 그룹내분산(Within-Group-Sum-of-Squares, WGSS)이 가장 적은 클러스터링 결과를 도출한다.

1) k개 클러스터 중심점을 설정한다.
2) 각 개체를 k개의 중심점을 기준으로 거리가 가장 가까운 클러스터에 할당한다.
3) 모든 개체를 클러스터에 할당한 후, 각 클러스터에 할당된 개체 속성 값의 평균으로 클러스터 중심점을 이동한다.
4) 클러스터 중심이 더 이상 이동하지 않을 때까지 2), 3)을 반복한다

<그림 1> K-means 알고리즘

<그림 1>은 K-means 알고리즘의 절차를 나타낸다. K-means는 초기 중심값에 따라 클러스터링 결과에 크게 영향을 받는다는 단점이 있으나, 절차의 단순성으로 인해 구현이 용이하고 다른 알고리즘, 특히 계층형 알고리즘에 비해 상대적

인 계산효율성이 우수하여 대용량 데이터 클러스터링에 널리 사용되고 있다[Kurita, 1991; Day, 1992; Jain *et al.*, 1999].

2.2 클러스터링 인덱스

본 연구에서는 Milligan and Cooper[1985]의 실험에서 사용된 30개 인덱스 중에서 비계층형 기법에 적용이 가능한 13개의 인덱스와 1985년 이후 비교적 최근에 발표된 인덱스들 중에서 상대적으로 널리 사용되는 3개의 인덱스를 선택하여 모두 16개 클러스터링 인덱스에 대한 성능을 비교한다. 병합되는 두 클러스터 또는 분할되는 하나의 클러스터에 속하는 부분 데이터 셋만을 이용하는 인덱스들[Duda and Hart, 1973; Beale, 1969; Mojena, 1977; Johnson, 1967; Wolfe, 1970; Gnanadesikan, Kettenring, and Landwehr, 1977; Sneath, 1977; Frey and Groenewoud, 1972; Bock, 1977; Lingoes and Cooper, 1971; Day, 1969; Mountford, 1970], 또는 전체 데이터 셋에 대한 정보를 이용하더라도 데이터 셋으로 만들 수 있는 모든 가능한 조합에 대하여 값을 계산해야 하는 인덱스들[Hubert and Levin, 1976; Baker and Hubert, 1975; Rohlf, 1974], 그리고 특정 분포를 가정하는 인덱스 [Ray, 1982]는 실험에서 제외되었다. 선택된 인덱스들의 명칭과 수식, 이에 대한 설명, 그리고 최적 클러스터 수 산정 전략은 <표 1>과 같다.

<표 1> 실험에 사용된 클러스터링 인덱스

순번	인덱스 명	수 식
	수식 설명, 최적 클러스터 개수 산정 전략 및 참고 문헌	
1	**Calinski and Harabasz (CH)**	$\frac{BGSS}{WGSS} \cdot \frac{n-k}{k-1}$
	WGSS는 그룹내분산(Within-Group-Sum-of-Squares), BGSS는 그룹간분산(Between-Group-Sum-of-Squares), n은 데이터 개수, 그리고 k는 클러스터 개수를 의미한다. 수식의 최대값을 주는 k를 클러스터 개수의 추정치로 선택한다[Calinski and Harabasz, 1974].	
2	**Ray and Turi (RT)**	$\frac{1}{n}\sum_{i=1}^{K}\sum_{i \in C_i} \| x - z_i \| / \min(\| z_i - z_j \|^2)$
	n은 데이터 개수, K는 클러스터 개수, C_i는 클러스터 i에 속한 데이터 집합, x는 데이터, z_i와 z_j는 각각 i, j번째 클러스터의 중심점을 의미한다. 수식의 최소값을 주는 k를 클러스터 개수의 추정치로 선택한다[Ray and Turi, 1999].	
3	**Davis and Bouldin (DB)**	Avg(max(pairwise comparison))
	pairwise comparison은 두 클러스터간 WGSS의 합을 BGSS로 나눈 값이다. 수식의 최소값을 주는 k를 클러스터 개수의 추정치로 선택한다[Davis and Bouldin, 1979].	
4	**G(+)**	$2s(-)/n_d(n_d-1)$
	s(-)는 같은 클러스터에 속한 불특정 두 데이터 사이 거리보다 거리가 더 가까움에도 불구하고 다른 클러스터에 속한 데이터 쌍의 개수(Number of inconsistent outcomes)를, n_d는 같은 클러스터에 속한 데이터 쌍의 개수를 의미한다. 수식의 최소값을 주는 k를 클러스터 개수의 추정치로 선택한다[Rohlf, 1974].	
5	**Point-Biserial (PB)**	$Cov(x,y)/sd(x)sd(y)$
	$Cov(\cdot,\cdot)$, $sd(\cdot)$는 각각 공분산(Covariance)과 표준편차(Standard Deviation)를 의미한다. 수식의 최대값을 주는 k를 클러스터 개수의 추정치로 선택한다[Milligan, 1981].	

순번	인덱스 명	수 식
	수식 설명, 최적 클러스터 개수 산정 전략 및 참고 문헌	
6	$\bar{c}/k^{0.5}$	$\bar{c}/k^{0.5}$
	$\bar{c}$는 각 속성에서 구해지는 $(BGSS/WGSS)^{1/2}$의 평균값이며, k는 클러스터 개수이다. 수식의 최대값을 주는 k를 클러스터 개수의 추정치로 선택한다[Ratkowsky and Lance, 1978].	
7	$k^2\|\mathbf{W}\|$	$k^2\|\mathbf{W}\|$
	W는 각 클러스터의 공분산 행렬의 합이며, k는 클러스터 개수이다. 연속된 두 k에 대해 수식 값에 가장 큰 변화를 주는 k를 선택한다[Marriot, 1975].	
8	Ball and Hall (BH)	*Avg WGSS*
	Avg는 평균(Average)을 의미하며 WGSS는 그룹내분산을 의미한다. 연속된 두 k에 대해 수식 값에 가장 큰 변화를 주는 k를 선택한다[Ball and Hall, 1965].	
9	*Trace* **W**	*Trace* **W**
	W는 각 클러스터의 공분산 행렬의 합이다. 연속된 두 k에 대해 수식 값에 가장 큰 변화를 주는 k를 선택한다[Edwards and Cavalli-Sforza, 1965].	
10	log(*SSB*/*SSW*)	log(*SSB*/*SSW*)
	SSB는 그룹간 분산(Sum-of-Squares Between Cluster Distances), SSW는 그룹내분산(Sum-of-Squares Within Cluster Distances)를 의미한다. 연속된 두 k에 대해 수식 값에 가장 큰 변화를 주는 k를 선택한다[Hartigan, 1975].	
11	**GDI(31)**	$\min_{1\le s\le c}\left\{\min_{\substack{1\le t\le c\\ t\ne s}}\left\{\frac{\delta_i(X_s,X_t)}{\max_{1\le k\le c}\{\Delta_j(X_k)\}}\right\}\right\}$
	$\delta_i(X_s,X_t)$는 클러스터 s, t 간 중심 거리, $\Delta_j(X_k)$는 클러스터 k의 WGSS이다. 수식의 최소값을 주는 k를 클러스터 개수의 추정치로 선택한다[Bezdek and Pal, 1998].	
12	$n\log(\|\mathbf{T}\|/\|\mathbf{W}\|)$	$n\log(\|\mathbf{T}\|/\|\mathbf{W}\|)$
	T는 전체 공분산 행렬, **W**는 각 클러스터 공분산 행렬의 합을 의미하며, n은 데이터 개수이다. 연속된 두 k에 대해 수식 값에 가장 큰 변화를 주는 k를 선택한다[Scott and Symons, 1971].	
13	*Trace* $\mathbf{W}^{-1}\mathbf{B}$	*Trace* $\mathbf{W}^{-1}\mathbf{B}$
	W는 각 클러스터 공분산 행렬의 합을 의미하며, **B**는 **T**에서 **W**를 뺀 값이다. 연속된 두 k에 대해 수식 값에 가장 큰 변화를 주는 k를 선택한다[Friedman and Rubin, 1967].	
14	**McClain and Rao (MR)**	*Avg WGSS*/*Avg BGSS*
	수식의 최소값을 주는 k를 클러스터 개수의 추정치로 선택한다[McClain and Rao, 1975].	
15	$\|\mathbf{T}\|/\|\mathbf{W}\|$	$\|\mathbf{T}\|/\|\mathbf{W}\|$
	연속된 두 k에 대해 수식 값에 가장 큰 변화를 주는 k를 선택한다[Friedman and Rubin, 1967].	
16	**S_Dbw**	$\frac{1}{c}\sum_{i=1}^{c}\|\sigma(v_i)\|/\|\sigma(S)\| + \frac{1}{c\cdot(c-1)}\sum_{i=1}^{c}\left(\sum_{\substack{j=1\\ i\ne j}}^{c}\frac{density(u_{ij})}{\max\{density(v_i),density(v_i)\}}\right)$
	c는 클러스터 개수, v_i는 클러스터 i의 중심, $\sigma(v_i)$는 클러스터 i의 WGSS, $density(v_i)$는 클러스터 i의 중심에서 1-표준편차 내에 들어오는 데이터 개수, 그리고 $density(u_{ij})$는 클러스터 i, j의 중심의 평균에서 두 클러스터에 속한 데이터들의 분포를 고려하여 1-표준편차 내에 들어오는 데이터 개수를 의미한다. 수식의 최소값을 주는 k를 클러스터 개수의 추정치로 선택한다[Halkidi and Vazirgiannis, 2001].	

선정된 16개의 인덱스는 최악의 상황을 피하려는 노력(Worst-case improvement concept)의 명시적 사용 여부에 따라 두 개의 그룹으로 나뉠 수 있다. RT, DB, G(+), 그리고 GDI(31) 인덱스는 그룹 내 데이터간 최대 거리를 최소화하는 방법이나 거리가 가까운 데이터를 다른 클러스터에 할당하는 오류를 최소화시키는 방법 등을 통해 최대-최소 전략(Min-Max Strategy)을 명시적으로 사용한다. 그 외 인덱스들은 WGSS나 BGSS 등을 이용하여 최대-최소 전략을 직접 사용하지는 않는다

다른 측면에서는 최적 클러스터 개수 산정 전략에 따라 인덱스들을 두 그룹으로 나눌 수 있다. CH, RT, DB, G(+), PB, $\bar{c}/k^{0.5}$, MR, GDI(31), 그리고 S_Dbw 인덱스는 최대값 혹은 최소값을 주는 k를 선택하는 반면, 나머지 인덱스들은 연속되는 두 개의 그룹수(k)에 대하여 인덱스 값에서 가장 큰 차이를 보이는 그룹수(k)를 최적 클러스터 개수로서 선택한다.

Ⅲ. 실험 및 결과

본 연구에서는 클러스터링 인덱스의 성능 비교를 위하여 Milligan[1985]의 데이터 생성 알고리즘을 이용한 기본실험과 알고리즘의 실험요인을 확장시킨 확장실험을 수행한다.

기본실험은 K-means 알고리즘에 기반한 클러스터링 인덱스의 성능이 기존 계층형 알고리즘을 이용한 연구 결과와 어떤 차이를 보이는지 관찰하는 것을 목적으로 하였으며, 확장실험은 데이터 수, 속성 수, 그리고 클러스터 수 등 기본실험에서 사용된 실험 요인의 크기 확장에 따라 클러스터링 인덱스의 상대적 성능에서 어떠한 변화를 보이는지를 관찰하는 것을 목적으로 하였다. 또한, 기본실험과 확장실험 결과를 바탕으로 정규분포를 따르는 데이터에 대하여 K-means 알고리즘을 사용할 때의 인덱스 선택 기준을 제시하고자 하였다.

각 인덱스, 데이터 셋에 대해

1) 클러스터 개수 k에 대하여 초기해 개수만큼 K-means를 실행한다.
2) 가장 작은 WGSS 값을 주는 클러스터링 결과에 근거해 인덱스 값을 계산한다.
3) 1), 2)를 k=k_{min}에서 k_{max}까지 반복한다.
4) 최적 클러스터 개수 산정 전략에 따라 특정 클러스터 개수를 선택한다.

<그림 2> 실험 절차

기본실험과 확장실험은 <그림 2>의 절차에 따라 진행되며, 각 인덱스의 성능은 데이터 생성 알고리즘에 의해 주어진 클러스터 수를 정확히 추정한 횟수로서 측정된다. 클러스터 개수 k를 k_{min}에서 k_{max}까지 변경시키면서 각각의 k 값에서 임의의 초기 중심점에 대하여 K-means 알고리즘을 실행한다. 또한, K-means 알고리즘이 초기해에 민감한 단점을 보완하기 위해 Forgy[1965]의 초기화 기법에 근거하여 각 k에 대해 다수의 초기 중심점 셋을 생성하여 알고리즘을 실행한 후 WGSS가 가장 작은 클러스터링 결과에 대하여 인덱스 값을 계산한다. 같은 방식으로 k_{min}에서 k_{max}까지 구한 인덱스 값들을 이용하여 각 인덱스의 전략에 따라 최적 클러스터 수를 선택한다.

3.1 시뮬레이션 데이터 생성 기법

Milligan[1985]은 클러스터링 알고리즘 및 인덱스의 시뮬레이션 연구를 위하여 시뮬레이션 데이터 생성 알고리즘을 제안하였다. Milligan의 알고리즘은 데이터 생성 요인으로 데이터 개수, 속성 개수, 클러스터 개수, 에러유형(Error Type), 및 클러스터의 밀집도(Discrepancy Type)를 사용하였다. 클러스터 밀집도는 세 가지(Equal, 10%, 60%) 유형으로 구분되었으며, 모든 클러스터에 속한 데이터의 수를 가능한 동일하게 하거나(Equal), 첫 번째 클러스터에 전체 데이터 수의 10%, 또는 60%에 해당하는 데이터를 포함시킨

후, 나머지 클러스터들에 남은 데이터를 균등하게 배분함으로써 클러스터 밀집도에 변화를 주었다. 이와 같이 생성된 데이터는 다변량 정규분포를 따르도록 하였으며, 속성별 정규분포의 난수 초기값(Random Seed)을 달리함으로써 동일한 분포를 따르는 세 개의 유사 데이터 셋(Replication)을 만들었다. 세부 알고리즘은 다음과 같다.

먼저 첫 번째 속성이 서로 중첩되지 않도록 각 속성에 대하여 클러스터 경계값을 클러스터 숫자만큼 임의로 생성한다. 경계의 중심을 평균으로, 중심에서 경계까지의 길이를 1.5-표준편차로 하는 정규분포를 가정하여 각 클러스터 별로 데이터를 생성한다. 이 때, 경계를 벗어나는 데이터는 제외시킴으로써 생성된 클러스터는 잘려진 정규분포를 갖게 된다. 이와 같은 표준데이터 셋에 대하여 세 가지 에러유형 즉, 특이값(Outlier), 오차반영(Error Perturbation), 차폐속성(Noise Dimension)을 추가하여 최종 데이터 셋을 얻는다. Milligan의 데이터 생성 알고리즘에 사용된 입력 변수 값은 데이터 개수(50, 100, 150, 200), 속성 개수(4, 6, 8), 클러스터 개수(2, 3, 4, 5), 그리고 특이값, 오차반영, 차폐속성의 반영 여부이다.

표준데이터가 평균으로부터 1.5-표준편차 범위 내에서 데이터를 생성하는 반면, 특이값 데이터는 3-표준편차까지의 범위에서 임으로 생성된다. 또한, 오차반영 데이터는 생성된 표준데이터의 속성값에 평균을 0으로 한 1-표준편차 범위의 오차값을 더해주는 방법으로 생성된다. 또한, 차폐속성은 정규분포가 아닌 균등분포에 따라 속성 데이터를 생성하여 표준데이터에 속성을 추가한다. 이와 같이 생성된 데이터는 첫 번째 속성으로 인해 클러스터가 비교적 명확히 구분되어 클러스터링 알고리즘의 성능에 데이터 자체가 주는 영향을 최소화할 수 있다[Milligan, 1985].

3.2 기본실험

기본실험에서는 Milligan의 데이터 생성 알고리즘을 변형 없이 적용하여 데이터 개수 50개, 속성 개수 4, 6, 8개, 그리고 클러스터 개수 2, 3, 4, 5개인 데이터 셋을 생성한다. 에러유형과 클러스터 밀집도 유형, 그리고 유사 데이터 셋 등도 적용되었다. <표 2>는 기본실험의 실험 계획을 요약한 것이다. 총 432개(3x4x3x3x4)의 데이터 셋이 실험에 사용되었으며, 사용된 데이터 셋에 포함된 클러스터 개수는 최대 5 이하이므로 k_{min}과 k_{max}는 2와 20으로 제한하였다.

<표 2> 기본실험 계획

데이터 개수	50
속성 개수	4, 6, 8
클러스터 개수	2, 3, 4, 5
기타	클러스터 밀집도 (평균, 60%, 10%) 유사 데이터 셋 3개 에러유형: 1, 2, 4, 6 K-means 초기해: 20개

<표 3>은 108개의 표준데이터 셋에 대한 개별 인덱스들의 성능을 요약한 기본실험 결과이다. '≤-2', '-1', '0', '+1', '+2', 및 '≥+3' 열은 데이터 생성 알고리즘에 의해 주어진 클러스터 수와 인덱스에 의한 추정치와의 차이를 나타내며, +/-는 각 인덱스가 클러스터 개수를 해당 숫자만큼 많게 또는 적게 추정한 것을 의미한다. '0'은 데이터 셋에 존재하는 클러스터 개수를 정확히 추정한 횟수를 의미하고, '≤-2'와 '≥+3'는 각 인덱스가 추정한 클러스터 개수가 알고리즘에 의해 주어진 클러스터 개수보다 두 개 이상 적은 경우와 세 개 이상 많은 경우의 합계이다. Milligan and Cooper[1985]의 순위결정 방법을 그대로 유지하기 위하여, '순위'는 '0'열의 값만을 기준으로 결정하였다. 'H' 열에 나타낸 Milligan and Cooper [1985]의 실험 결과는 4개의 계층형 알고리즘에서 클러스터 수를 정확히 추정한 횟수를 합한 것이므로, 직접 비교를 위하여 '0'열을 4로 곱한 값을 'NH' 열에 나타냈다.

표준데이터 셋에 대한 실험 결과 Milligan and Cooper[1985]의 실험에서와 마찬가지로 CH 인덱스가 K-means에서도 가장 좋은 성능을 보였으며, G(+) 인덱스와 DB 인덱스는 계층형 알고리즘에서보다 K-means 알고리즘에서 향상된 결과를 보이는 것으로 'NH'열과 'H'열을 통해 알 수 있다. CH와 RT 인덱스의 경우, 정확히 추정한 횟수에서는 86과 81로서 CH 인덱스가 상대적으로 높은 성능을 보였지만, '≤-2', '-1', '0', '+1', '+2', '≥+3' 열에 해당하는 값을 살펴보면 CH 인덱스는 '0' 열을 중심으로 넓게 분포된 반면, RT 인덱스의 잘못 추정된 횟수는 한 개 정도 적은 '-1' 열에 집중되는 경향을 보였다.

Milligan and Cooper[1985] 이후 발표된 인덱스들 중, RT 인덱스는 CH 인덱스 다음으로 우수한 성능을 보였으나, GDI(31) 인덱스와 S_Dbw 인덱스는 K-means 알고리즘에서 상대적으로 열악한 성능을 나타냈다. MR, |T|/|W|, S_Dbw 등 최하위에 위치한 인덱스들은 실험에 사용된 대부분의 데이터 셋에 대해 클러스터 개수를 k_{max}로 추정하려는 경향을 보이고 있으며, 이러한 결과는 새로 포함된 S_Dbw를 제외하고 기존 Milligan and Cooper[1985]의 연구 결과와 유사하게 나타났다.

최대-최소 전략을 명시적으로 사용한 인덱스들은 그렇지 않은 인덱스들에 비해 평균적으로 좋은 결과를 보이고 있다. GDI(31) 인덱스를 제외하고 RT 인덱스가 2위, DB 인덱스가 3위, G(+) 인덱스가 4위를 차지하는 등 최악의 상황을 방지하려는 명시적 노력이 동질클러스터링에서의 K-means 알고리즘에 유효한 결과를 나타내고 있는 것으로 추정된다.

<표 3> 기본실험 결과 - 표준데이터 셋(Error Free Datasets)

순위	인 덱 스	Group Type #1[1]	Group Type #2[2]	클러스터 개수 추정						NH ('0' x 4)	H[3]
				≤-2	-1	0	+1	+2	≥+3		
1	Calinski and Harabasz (CH)	X	M	0	8	86	9	4	1	344	390
2	Ray and Turi (RT)	O	M	4	21	81	0	0	2	324	
3	Davis and Bouldin (DB)	O	M	0	12	79	1	0	16	316	287
4	G(+)	O	M	3	21	77	6	1	0	308	297
5	Point Biserial (PB)	O	M	13	28	65	2	0	0	260	308
6	$\overline{c}/k^{0.5}$	X	M	18	47	33	9	1	0	132	200
7	$k^2\|\mathbf{W}\|$	X	D	50	26	27	3	1	1	108	146
7	Ball and Hall (BH)	X	D	54	27	27	0	0	0	108	128
7	*Trace* **W**	X	D	54	27	27	0	0	0	108	120
10	log(*SSB*/*SSW*)	X	D	32	48	26	2	0	0	104	212
11	GDI(31)	O	M	37	13	15	1	3	39	60	
11	$n\log(\|\mathbf{T}\|/\|\mathbf{W}\|)$	X	D	42	38	15	4	0	9	60	149
13	*Trace* $\mathbf{W}^{-1}\mathbf{B}$	X	D	3	2	4	0	0	99	16	84
14	McClain and Rao (MR)	X	M	0	0	0	0	0	108	0	25
14	$\|\mathbf{T}\|/\|\mathbf{W}\|$	X	D	0	0	0	0	0	108	0	0
14	S_Dbw	X	M	0	0	0	0	0	108	0	

주) [1] 최대 최소 전략 사용 여부 - 사용(O), 미사용(X)
[2] 최적 클러스터 개수 산정 전략 - 최대/최소 값 취함(M), 최대 차이 취함(D).
[3] Milligan and Cooper[1985]에 제시된 결과

<표 4> 기본실험 결과 - 에러유형 별 요약

순위	인 덱 스	Group Type #1[1]	Group Type #2[2]	ET#1 (표준)	ET#2 (특이값)	ET#4 (오차)	ET#6 (차폐)	합계 (총 432개)
1	Calinski and Harabaaz	X	M	86	67	81	41	275
2	Ray and Turi	O	M	81	32	74	57	244
3	G(+)	O	M	77	22	71	50	220
4	Point-Biserial	O	M	65	20	60	56	201
5	Davis and Bouldin	O	M	79	15	62	35	191
6	$\bar{c}/k^{0.5}$	X	M	33	30	34	33	130
7	$k^2\|\mathbf{W}\|$	X	D	27	28	24	28	107
7	Ball and Hall	X	D	27	27	27	27	107
7	*Trace* **W**	X	D	27	27	27	27	107
10	log(*SSB*/*SSW*)	X	D	26	24	26	28	104
11	GDI(31)	O	M	15	25	10	14	64
12	$n\log(\|\mathbf{T}\|/\|\mathbf{W}\|)$	X	D	15	9	21	15	60
13	*Trace* $\mathbf{W}^{-1}\mathbf{B}$	X	D	4	1	4	9	18
14	McClain and Rao	X	M	0	0	0	0	0
14	$\|\mathbf{T}\|/\|\mathbf{W}\|$	X	D	0	0	0	0	0
14	S_Dbw	X	M	0	0	0	0	0

주) [1], [2] <표 3>과 동일

한편, 최적 클러스터 수 산정 전략에서 클러스터 개수 변화에 대한 최대 차이를 찾는 점진적 방법(Incremental Strategy)을 사용하는 $k^2|\mathbf{W}|$, BH, *Trace* **W**, log(*SSB*/*SSW*), $n\log(|\mathbf{T}|/|\mathbf{W}|)$, *Trace* $\mathbf{W}^{-1}\mathbf{B}$, $|\mathbf{T}|/|\mathbf{W}|$ 인덱스는 최대값이나 최소값을 취하는 인덱스들에 비해 평균적으로 낮은 성능을 나타냈다. 이는 클러스터 수 k의 증가에 따른 WGSS의 단조 감소 경향성 때문인 것으로 판단된다. 즉, WGSS가 단조 감소하면서 해당 인덱스 값에도 일정한 경향성을 만들며, 데이터 셋의 구조와는 무관하게 특정 클러스터 수에서 인덱스 값의 가장 큰 차이를 발생시키게 된다는 것이다. 예를 들어, *Trace* **W** 인덱스의 경우, WGSS 값이 가장 크게 변하는 시점은 클러스터 개수가 많은 경우보다 적은 경우에 발생하므로, 많은 클러스터가 존재하는 데이터 셋에 좋지 않은 결과를 나타내고 있으며, 실제로 정확히 클러스터 수를 추정한 27개는 모두 k=3인 데이터 셋이었다.

<표 4>는 기본실험 결과를 에러유형별로 요약한 것이다. 열 제목의 'ET#1, ET#2, ET#4, ET#6'은 Milligan[1985]에 의해 제시된 에러유형을 의미하며 괄호 안에 간략한 설명을 병기하였다. '순위'는 '합계' 열을 기준으로 결정되었으며, 4가지 에러유형이 고려된 상황에서도 표준데이터 셋을 사용한 결과와 마찬가지로 CH 인덱스가 가장 우수한 성능을 보였다. 또한, 여전히 최대 최소 전략을 명시적으로 고려하는 인덱스들이 그렇지 않은 인덱스들에 비해 평균적으로 좋은 성능을 나타냈다.

ET#1와 ET#4 열은 각각 표준데이터와 오차반영 데이터를 이용한 실험 결과이며, 두 결과에 큰 차이가 없다. 따라서 데이터 생성 알고리즘 [Milligan, 1985]의 적용에서 오차반영 자체 또는 1-표준편차 정도의 오차반영은 인덱스 성능 비교

에 큰 영향을 주지 못하는 것으로 보인다.

ET#2 열은 표준데이터에 특이값이 20% 포함된 데이터 셋을 이용한 결과이며, 모든 인덱스들에서 성능 저하 현상이 명확하게 나타났다. 특히, 최대-최소 전략을 명시적으로 사용한 인덱스들 가운데 GDI(31)를 제외한 RT, G(+), PB, DB 인덱스가 그렇지 않은 인덱스들에 비해 상대적으로 특이값에 큰 영향을 받는 것으로 나타났다. 인덱스 별로 살펴보면, RT와 DB 인덱스는 점간 최대거리, 그룹간 최소 거리 등에 특이값이 직접적인 영향을 줄 수 있으며, G(+) 인덱스는 분자인 s(-)가 작을수록 좋은 값을 주게 되는데 특이값은 이를 증가시키는 역할을 하게 된다. 또한 특이값은 PB 인덱스의 분자를 감소시키고 분모를 증가시키려는 경향이 있어 최대값을 찾으려는 해당 인덱스의 전략에 반하는 효과를 가져온다.

ET#6 열은 표준데이터에 차폐속성 1개가 추가된 데이터 셋에 대한 실험 결과를 나타내며, CH, RT, G(+), DB 인덱스가 표준데이터인 ET#1 열의 값과 비교하여 큰 차이를 보이고 있다. 특히, CH 인덱스와 DB 인덱스의 성능이 ET#1 열과 비교하여 절반 이하로 감소하는 것으로 나타나 두 인덱스가 차폐속성에 가장 큰 영향을 받는 것으로 나타났다.

6, 7위에 위치한 $\bar{c}/k^{0.5}$, $k^2|\mathbf{W}|$, BH, *Trace* $\mathbf{W}$ 인덱스는 실험에 사용된 모든 에러유형에 걸쳐서 실제 클러스터 수와는 무관하게 2 또는 3으로 클러스터 수를 추정하는 강한 경향을 보였다. 또한 MR, $|\mathbf{T}|/|\mathbf{W}|$, S_Dbw 인덱스 등은 K-means 알고리즘에서 매우 낮은 성능을 나타냈다.

Milligan and Cooper[1985] 이후 새로 추가된 인덱스들 중에서는 RT 인덱스가 에러유형이 반영된 데이터 셋에 대해서도 좋은 결과를 보이고 있으며, 표준데이터, 오차반영, 차폐속성이 추가된 데이터에서는 CH 인덱스와 유사한 성능을 보였으나 특이값이 포함된 데이터에서는 그 성능이 크게 저하되었다.

기본실험 결과 다변량 정규분포를 따르는 데이터 셋에 대하여 K-means 알고리즘을 이용하는 경우 모든 데이터 구조에 대해 일관된 인덱스 선택 기준을 적용하는 것은 무리가 따르는 것으로 판단된다. CH 인덱스가 전반적으로 좋은 성능을 나타내고는 있지만, 일부 인덱스가 낮은 또는 높은 수의 클러스터 개수를 선호한다든지 등의 클러스터 개수 추정 경향 또는 차폐속성과 같은 에러의 영향 등에 따라 RT 또는 PB 인덱스 등이 CH 인덱스보다 좋은 결과를 나타내는 경우가 발생하였다. 따라서 K-means 알고리즘을 사용하는 경우 데이터에 대한 사전지식 유무에 따라 데이터의 특성에 맞는 적절한 선택 기준을 적용해야 한다고 판단된다.

3.3 확장실험

본 연구에서는 확장실험을 통해 실험요인의 확장에 따른 인덱스 성능 변화를 관찰하고, 실험요인별 성능 분석을 바탕으로 기본실험 결과를 보완하고자 하였다. 확장실험에서는 기본 실험에서 상대적으로 우수한 성능을 보인 CH, RT, G(+), PB, DB 인덱스만을 대상으로 실험을 진행하였다. 즉, 기본실험 결과 클러스터 수 추정에 일정한 경향을 보이는 $\bar{c}/k^{0.5}$, $k^2|\mathbf{W}|$, BH, *Trace* $\mathbf{W}$, $|\mathbf{T}|/|\mathbf{W}|$, MR, S_Dbw 인덱스와 표준데이터를 기준으로 상위 인덱스들과 뚜렷한 성능 차이를 보이는 $\log(SSB/SSW)$, GDI(31), *Trace* $\mathbf{W}^{-1}\mathbf{B}$ 인덱스는 확장실험에서 제외되었다. 확장실험에 사용된 인덱스들은 모두 최적 클러스터 개수 산정에서 최대값 또는 최소값을 선택하는 전략을 사용하며, CH를 제외한 DB, RT, PB, G(+) 인덱스는 최대 최소 전략을 명시적으로 사용한다.

확장실험에서는 <표 5>에 나타난 것과 같이 기본실험의 모든 실험요인을 2배, 4배 8배로 대칭적으로 확장시킨 데이터 셋과 함께 (데이터 수, 속성 수, 클러스터 수)가 각각 (100, 48, 20) 또는 (400, 12, 3) 등 한 개 또는 두 개 실험요인을 비대칭적으로 늘린 데이터 셋을 동시에 적용

하였다. 이를 통해 각 요인의 대칭적 또는 비대칭적 변화가 클러스터링 인덱스 성능에 미치는 영향을 관찰하였다. 실험에서는 총 2268개(3×3×7×3×4×3)의 데이터 셋이 사용되었으며, Fortran으로 작성된 Milligan[1985]의 원시코드(Source Code)를 C 언어로 변환한 후 확장된 데이터 생성이 가능하도록 수정하였다. <표 5>의 각 항목에 대한 설명은 <표 2>를 참조하면 된다.

<표 5> 확장실험 계획

요 인	Low	Mid	High
데이터 개수	100	200	400
속성 개수	12	24	48
클러스터 개수	3, 5, 7	10, 14	20, 28
기타	클러스터 밀집도(평균, 60%, 10%) 유사 데이터 셋 3개 에러유형: 1, 2, 4, 6 K-means 초기해: 100개		

확장실험은 <그림 2>에 나타난 것과 같이 기본실험과 동일한 절차에 따라 진행된다. 다만, 실험요인 확장에 따른 문제의 해 공간 확장을 고려하여 K-means 알고리즘의 초기 중심점 셋을 20개에서 100개로 증가시켰으며, 클러스터 개수 범위는 k_{min}=2와 k_{max}=(2*실험데이터의 클러스터 개수)로 확대하였다. 초기 중심점 셋의 개수를 100으로 설정한 것은 CH 인덱스를 이용하여 데이터 개수가 100, 200, 400개인 540개의 데이터 셋에 대해 초기해 수를 20, 50, 100, 200, 400, 800개로 변경하여 실험한 결과, 초기 중심점 셋의 개수가 증가함에 따라 클러스터 숫자를 정확히 측정한 횟수가 증가하다가 100개 이상에서는 더 이상의 별다른 차이를 보이지 않았기 때문이다.

<표 6>은 확장실험을 통해 추정된 클러스터 수를 기준으로 요약한 것이며, 열 제목은 <표 3>과 동일하다. 순위는 기본실험과 마찬가지로 표준데이터 셋 결과의 '0' 열을 기준으로 결정하였다. 표준데이터인 '0' 열을 기준으로 가장 좋은 결과를 나타낸 CH 인덱스의 경우에도 기본실험의 86/108에서 확장실험에서는 185/567로 성능이 감소하는 등 기본실험 결과에 비해 인덱스들의 성능이 크게 저하되었다.

표준데이터만을 비교하거나 에러유형별 추정치의 합계를 비교해도 모두 CH, DB, RT 인덱스가 G(+)와 PB 인덱스에 비해 좋은 성능을 보였다. 특히, DB 인덱스의 경우 다른 인덱스 보다 좋은 성능을 보였다. 또한 '0'열을 기준으로 CH 인덱스가 DB 인덱스와 근소한 차이를 보이며 가장 좋은 성능을 나타냈으나, 클러스터 개수 추정 경향이 서로 다르게 나타나는 등 실험 결과만으로는 두 인덱스간 절대적 성능 우위를 단정하기 어렵다.

<표 6> 확장실험 결과 - 클러스터 개수 추정 요약

순위[1]	인 덱 스	표준데이터-567[2] 개 데이터 셋						에러유형 합계-2268[2] 개 데이터 셋					
		≤-2	-1	0	1	2	≥3	≤-2	-1	0	1	2	≥3
1	Calinski and Harabasz (CH)	169	28	185	41	15	129	898	178	560	134	47	451
2	Davis and Bouldin (DB)	106	60	184	19	12	186	377	254	556	94	60	927
3	Ray and Turi (RT)	156	49	141	16	16	189	654	263	474	69	66	742
4	G(+)	216	34	98	24	37	158	1114	198	289	83	95	489
5	Point Biserial (PB)	311	53	73	27	32	71	1279	221	283	83	95	307

주) [1] 순위는 표준데이터 셋 '0'열 기준
[2] 테스트 데이터 셋 개수

<표 7> 확장실험 결과 - 에러유형 별 요약

순위[1]	인 덱 스	Group Type #1[2]	Group Type #2[2]	ET#1 (표준)	ET#2 (특이값)	ET#4 (오차)	ET#6 (차폐)	합계 (총 2268개)
1	Calinski & Harabaaz (CH)	X	M	185	95	176	104	560
2	Davis and Bouldin (DB)	O	M	184	107	178	87	556
3	Ray and Turi (RT)	O	M	141	92	145	96	474
4	G(+)	O	M	98	36	110	45	289
5	Point Biserial (PB)	O	M	73	49	69	92	283

주) [1] 순위는 합계 기준
[2] <표 3>과 동일한 의미

<표 8> 실험요인별 인덱스 성능

인덱스	#pts[1] / #dim[2] / #cls[3]	100				200				400			
		12	24	48	합계	12	24	48	합계	12	24	48	합계
Calinski and Harabasz (CH)'	3	20	18	13	**51**	25	16	11	52	25	26	18	**69**
	5	32	16	10	**58**	34	19	18	71	33	25	22	**80**
	7	23	14	2	**39**	28	11	2	41	25	16	9	**50**
	10	11	0	0	**11**	7	1	2	10	11	7	3	**21**
	14	1	0	0	**1**	5	0	0	5	1	0	0	**1**
	20	0	0	0	**0**	0	0	0	0	0	0	0	**0**
	28	0	0	0	**0**	0	0	0	0	0	0	0	**0**
	합계	**87**	**48**	**25**	**160**	**99**	**47**	**33**	179	**95**	**74**	**52**	**221**
Davis and Bouldin (DB)	3	23	9	15	**47**	31	14	7	52	27	26	15	**68**
	5	25	4	3	**32**	29	15	16	60	31	28	20	**79**
	7	21	1	1	**23**	22	7	9	38	29	21	12	**62**
	10	13	0	0	**13**	11	5	3	19	17	16	5	**38**
	14	1	0	0	**1**	3	0	0	3	6	4	1	**11**
	20	1	0	0	**1**	0	0	0	0	4	1	0	**5**
	28	0	0	0	**0**	2	0	0	2	2	0	0	**2**
	합계	**84**	**14**	**19**	**117**	**98**	**41**	**35**	174	**116**	**96**	**53**	**265**
Ray and Turi (RT)	3	21	8	11	**40**	27	13	6	46	25	23	9	**57**
	5	29	6	1	**36**	27	10	15	52	31	19	15	**65**
	7	27	1	0	**28**	21	5	8	34	25	8	13	**46**
	10	18	2	0	**20**	14	1	2	17	19	2	2	**23**
	14	3	0	0	**3**	2	2	0	4	1	1	0	**2**
	20	1	0	0	**1**	0	0	0	0	0	0	0	**0**
	28	0	0	0	**0**	0	0	0	0	0	0	0	**0**
	합계	**108**	**22**	**16**	**146**	**96**	**39**	**35**	170	**109**	**67**	**46**	**222**

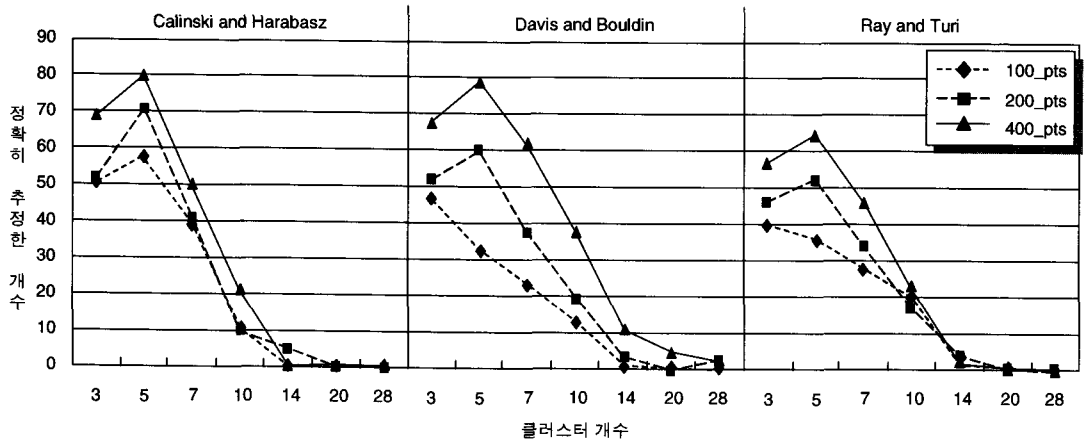
주) [1], [2], [3]의 #pts, #dim, #cls는 각각 데이터 개수, 속성 개수, 클러스터 개수를 의미한다.

<표 7>은 확장실험 결과를 에러유형별로 나타낸 것으로, 각 열에 대한 설명은 <표 6>과 동일하다. <표 6>의 기본실험 결과와 마찬가지로 오차반영 데이터 셋은 인덱스 성능에 큰 영향을 주지 못하였으나 특이값과 차폐속성이 반영된 경우에는 표준데이터 셋에 비해 뚜렷한 성능 저하를 관찰할 수 있다. <표 6>과 <표 7>을 비교해 볼 때 CH 인덱스의 경우 확장실험에서 특이값 데이터의 영향도를 비교할 수 있는 ET#2/ET#1 비율이 기본실험에 비해 감소한 반면 다른 인덱스들은 비율이 다소 증가하였다. 즉, 최대 최소 전략을 사용하는 인덱스들이 실험요인 확장으로 인해 특이값의 영향을 적게 받는 것으로 나타났다. 오차데이터나 차폐속성은 기본실험과 비교해 인덱스에 따라 그 영향이 다르게 나타났다.

<표 6>과 <표 7>에서는 확장된 데이터에 대해 CH와 DB 인덱스가 다른 인덱스에 비해 상대적으로 우수한 성능을 나타내고 있다. CH가 DB 인덱스보다 전반적으로 다소 높은 성능을 보여주지만, ET#2의 경우, 즉, 데이터에 특이값이 일정 수준 포함된 경우에는 CH보다 DB 인덱스가 상대적으로 우수한 성능을 나타냈다. 이러한 결과는 기본실험에서 DB 인덱스가 특이값에 영향을 많이 받는 것과 상반된 결과이나 실험요인 확장으로 인해 특이값의 영향이 상대적으로 감소한 것으로 추정된다.

<표 8>은 확장실험의 결과로 나타난 CH, DB, RT 인덱스의 성능을 실험요인 별로 나타낸 것이며, #pts는 데이터 개수, #dims는 속성 개수, 그리고 #cls는 클러스터 개수를 의미한다. 일반적으로 알려진 바와 같이, 데이터 개수가 증가함에 따라 인덱스의 성능이 향상되며, 속성 개수와 클러스터 개수의 증가는 인덱스 성능에 부정적인 영향을 주는 것으로 나타났다.

데이터 개수의 증가에 따라 세 인덱스 가운데 DB 인덱스 성능이 꾸준히 개선되는 것을 관찰할 수 있으며, 특히 데이터 개수가 400개인 경우 DB 인덱스가 CH 인덱스보다 더 나은 성능을 보였다. 결과의 경향성을 고려해볼 때, 현실에서 빈번히 발생하는 400개 이상의 데이터에 대하여 CH 인덱스보다 DB 인덱스가 우선적으로 고려될 수 있으며, 두 인덱스가 함께 사용될 경우 DB 인덱스의 결과에 보다 큰 가중치가 할당되어야 할 것으로 보인다. 본 연구의 실험 결과로부터 실험과 유사한 조건에 해당하는 경우에는 데이터 개수가 200 이하이면 CH 인덱스, 데이터 개수가 200개 이상 400개 이하이면 DB 인덱스가 우선적으로 고려될 수 있을 것으로 보인다.



<그림 3> 클러스터 개수/데이터 개수 변화에 따른 인덱스 성능

<그림 3>은 <표 8>의 '합계' 열 즉, 에러 유형별로 정확히 클러스터 수를 추정한 횟수를 합산한 수를 클러스터 개수와 데이터 개수의 증가에 대한 변화를 그래프로 나타낸 것이다. 클러스터 개수가 14개 이상에서는 어떤 인덱스도 데이터 개수 증가에 관계없이 거의 모든 데이터 셋에 대하여 클러스터 개수를 효과적으로 추정하지 못하고 있으며, 세 인덱스 모두 클러스터 개수가 3개 또는 5개 일 때 가장 좋은 결과를 보이는 경향이 있고, 데이터 개수가 증가할수록 클러스터 수가 5인 경우에서 가장 좋은 결과를 보였다.

Ⅳ. 결론 및 향후 연구 방향

본 연구에서는 K-means 알고리즘에 기반한 클러스터링 인덱스의 성능 비교 연구를 수행하였다. 모든 인덱스를 실험에 포함시키려 하기보다는 상대적으로 많이 인용되면서 비계층형 클러스터링에 적용 가능하다고 판단되는 16개 인덱스에 대하여 객관적인 비교 결과를 제시하고자 하였으며, 기존 비교 연구의 틀 속에서의 인덱스간 성능 비교를 위한 기본실험과 함께 실험요인 확장에 따른 영향도 분석을 위한 확장실험을 통해 보다 심도 있는 분석을 제공하고자 하였다.

기본실험 결과 Milligan and Cooper[1985]의 실험 결과와 마찬가지로 CH 인덱스가 가장 좋은 결과를 보였으나 차폐속성, 특이값 등 상황에 따라 RT 또는 DB 인덱스가 보다 나은 결과를 보이기도 했다. 최악의 상황을 방지하기 위한 최대/최소 전략을 명시적으로 사용한 RT, DB, G(+) 인덱스 등은 그렇지 않은 인덱스에 비해 상대적으로 더 좋은 성능을 나타냄을 실험적으로 보였다. 또한, 에러유형이 반영된 데이터를 이용한 실험에서는 모든 인덱스가 특이값과 차폐속성의 영향으로 급격한 성능 저하를 보였으며, 특히 CH 인덱스가 다른 에러유형에 비해 차폐속성에 가장 큰 영향을 받는 것으로 나타났다. 따라서 특이값 또는 차폐속성이 다수 포함될 수 있는 현실 문제에서는 인덱스의 성능 개선을 위한 추가 연구가 필요할 것이다.

확장실험에서는 기본실험에서와 마찬가지로 CH 인덱스가 평균적으로 가장 좋은 결과를 보였지만, 실험요인의 확장에 따라 부분적으로 DB 인덱스가 더 나은 결과를 보이기도 했다. 실험을 통해 K-means 알고리즘의 실행 시 설정해야 하는 모수 값의 범위와 데이터에 대한 사전 지식 유무에 따른 인덱스 선택 기준을 제안하였으며, 속성수의 증가가 클러스터링 결과에 미치는 부정적인 영향을 실험적으로 확인할 수 있었다. 이에 따라 데이터 마이닝의 주요 대상인 대용량 다속성 데이터(Large-Scale Multidimensional Data)를 클러스터링 하는 경우, 적절한 속성 선택(Feature Selection) 기법을 사용함으로써 보다 우수한 품질의 클러스터링 결과 도출이 가능할 것으로 예상된다.

본 연구는 '다변량 정규분포를 따르는 데이터 셋'에 한정되었다. 데이터의 분포에 따라 인덱스의 성능 또한 다르게 나타날 수 있으므로, 데이터 분포에 대한 정규 가정을 완화시킨 대용량 시뮬레이션 데이터 생성 알고리즘 등 데이터의 다양성과 현실성을 반영한 실험 데이터 생성 기법과 그에 따른 추가적인 인덱스 성능 비교 연구가 필요할 것으로 판단된다.

〈참 고 문 헌〉

[1] Baker, F.B. and Hubert, L.J., "Measuring the Power of Hierarchical Cluster Analysis," *Journal of the American Statistical Association*, Vol. 70, 1975, pp. 31-38.

[2] Ball, G.H. and Hall, D.J., "ISODATA, A Novel Method of Data Analysis and Pa-

ttern Classification," Menlo Park: Stanford Research Institute. (NTIS No. AD 699616), 1965.

[3] Beale, E.M.L., *Cluster Analysis*, London: Scientific Control Systems, 1969.

[4] Berry, M.J.A. and Linoff, G.S., *Mastering Data Mining - The Art and Science of Customer Relationship Management*, John Wiley and Sons, Inc. 2000.

[5] Bezdek, J.C. and Pal, N.R., "Some New Indexes of Cluster Validity," *IEEE Transactions on Systems, Man, and Cybernetics – PART B: CYBERNETICS*, Vol. 28, No. 3, 1998.

[6] Bock H.H., "On Tests Concerning the Existence of a Classification," In *First International Symposium on Data Analysis and Informatics*, Vol. 2, 1977, pp. 449-464, Rocquencourt, France: IRIA.

[7] Calinski T. and Harabasz, J., "A Dendrite Method for Cluster Analysis," *Communications in Statistics*, Vol. 3, No. 1, 1974, pp. 1-27.

[8] Davies D.L. and Bouldin, D.W., "A Cluster Separation measure," *IEEE Transactions on Pattern analysis and Machine Intelligence*, Vol. PAMI 1, No. 2, 1979, pp. 224-227.

[9] Day, N.E., "Estimating the Components of a Mixture of Normal Distributions," *Biometrika*, Vol. 56, 1969, pp. 463-474.

[10] Day, W.H.E., Complexity Theory: An Introduction for Practitioners of Classification, *Clustering and Classification*, P. Arabie and L. Hubert, Eds. World Scientific Publishing Co., Inc., River Edge, NJ., 1992.

[11] Dimitriadou, E., Dolnicar, S. and Weingessel, A., "An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets," *Psychometrika*, Vol. 67, No. 1, 2002.

[12] Duda, R.O. and Hart, P.E., *Pattern Classification and Scene Analysis*, New York: Wiley, 1973.

[13] Edwards, A.W.F. and L. Cavalli Sforza, "A Method for Cluster Analysis," *Biometrika*, Vol. 56, 1965, pp. 362-375.

[14] Forgy, E., "Cluster Analysis of Multivariate Data: Effciency vs. Interpretability of Classifications," *Biometrics*, Vol. 21, 1965, 768.

[15] Frey, T. and Groenewoud, H.V., "A cluster Analysis of the D-squared Matrix of White Spruce Stands in Saskatchewan based on the Maximum Minimum Principle," *Journal of Ecology*, Vol. 60, 1972, pp. 873-886.

[16] Friedman, H.P. and Rubin, J., "On Some Invariant Criteria for Grouping Data," *Journal of the American Statistical Association*, Vol. 62, 1967, pp. 1159-1178.

[17] Gnanadesikan, R., Kettenring, J.R. and Landwehr, J.M., "Interpreting and Assessing the Results of Cluster Analyses," *Bulletin of the International Statistical Institute*, Vol. 47, 1977, pp. 451-463.

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

[19] Hartigan, J.A., *Clustering Algorithms*, New Work, Wiley, 1975.

[20] Hubert, L.J. and Levin, J.R., "A General Statistical Framework for Assessing Categorical Clustering in Free Recall," *Psychological Bulletin*, Vol. 83, 1976, pp. 1072-1080.

[21] Jain, A.K., Murty, M.N. and Flynn, P.J., "Data Clustering: A Review," *ACM Computing Surveys*, Vol. 31, No. 3, 1999.

[22] Johnson, S.C., "Hierarchical Clustering Schemes," *Psychometrika*, Vol. 32, 1967, pp. 241-254.

[23] Kurita, T., "An Efficient Agglomerative Clustering Algorithm using a Heap," *Pattern Recognition*, Vol. 24, No. 3, 1991, pp. 205-209.

[24] Lingoes, J.C. and Cooper, T., "PEP-I: A FORTRAN IV (G) program for Guttman-Lingoes Nonmetric Probability Clustering," *Behavioral Science*, Vol. 16, 1971, pp. 259-261.

[25] MacQueen, J.B., "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of 5 th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, Vol. 1, 1967, pp. 281-297.

[26] Marriot, F.H.B., "Practical Problems in a Method of Cluster Analysis," *Biometrics*, Vol. 27, 1975, pp. 456-460.

[27] McClain, J.O. and Rao, V.R., "CLUSTISZ: A Program to Test for the Quality of Clustering of a Set of Objects," *Journal of Marketing Research*, Vol. 12, 1975, pp. 456-460.

[28] Milligan, G.W. and Cooper, M.C., "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, Vol. 50, No. 2, 1985, pp. 159-179.

[29] Milligan G.W., "A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis," *Psychometrika*, Vol. 46, 1981, pp. 187-199.

[30] Milligan, G.W., "An Algorithm for Generating Artificial Test Clusters," *Psychometrika*, Vol. 50, No. 1, 1985, pp. 123-127.

[31] Milligan, G.W., "An Examination of the Effect of six Types of Error Perturbation on Fifteen Clustering Algorithms," *Psychometrika*, Vol. 45, No. 3, 1980, pp. 325-342.

[32] Mojena, R., "Hierarchical Grouping Methods and Stopping Rules: An Evaluation," *The Computer Journal*, Vol. 20, 1977, pp. 359-363.

[33] Mountford, M.D., "A Test for the Difference between clusters, In G.P. Patil, E.C. Pielou, and W.E. Waters(EDs.)," *Statistical Ecology*, Vol. 3, 1970, pp. 237-257, University Park, Pa.: Pennsylvania State University Press.

[34] Ratkowsky, D.A. and Lance, G.N., "A Criterion for determining the number of groups in a classification," *Australian Computer Journal*, Vol. 10, 1978, pp. 115-117.

[35] Ray, A.A., *SAS user's guide: Statistics*, Cary, North Carolina: SAS Institue, 1982.

[36] Ray, S. and Turi, R.H., "Determination of Number of Clusters in k-means Clustering and Application in Colour Image Segmentation," in *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, 1999, pp. 137-143.

[37] Rohlf, F.J., "Methods of Comparing Classifications," *Annual Review of Ecology and Systematics*, Vol. 5, 1974, pp. 101-113.

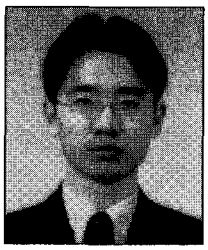
[38] Scott, A.J. and Symons, M.J., "Clustering Methods based on Likelihood Ratio Criteria," *Biometrics*, Vol. 27, 1971, pp. 387-397.

[39] Sneath, P.H.A., "A Method for Testing the Distinctness of Clusters: A Test of the Dis-

junction of two Clusters in Euclidean Space as Measured by their Overlap," *Mathematical Geology*, Vol. 9, 1977, pp. 123-143.

[40] Wolfe, J.H., "Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research*, Vol. 5, 1970, pp. 329-350.

◆ 저자소개 ◆



심요성 (Shim, Yo-Sung)
현재 고려대학교 산업시스템정보공학과 석사과정에 재학 중이다. 고려대학교에서 산업공학사를 취득하였으며, ㈜Nextwave와 ㈜로커스에서 금융권 CRM 관련 업무를 수행하였다. 주요 연구분야는 데이터 마이닝, 최적화, 기계 학습 및 인공 지능 등이다.

정지원 (Chung, Ji-Won)
현재 고려대학교 산업시스템정보공학과 박사과정에 재학 중이다. 인하대학교를 졸업하고 University of Nebraska-Lincoln에서 공학석사를 취득하였다. 주요 관심분야는 데이터마이닝 및 전자상거래, 물류정보시스템 등이다.

최인찬 (Choi, In-Chan)
현재 고려대학교 산업시스템정보공학과 교수로 재직 중이다. 고려대학교 산업공학사, Columbia University에서 산업공학석사/박사 학위를 취득하였다. 주요 연구분야는 최적화 이론 및 응용연구(Optimization Theory and Applications), 데이터 마이닝(Data Mining), 기계 학습(Machine Learning) 등이다.

이 논문은 2005년 9월 18일 접수하여 1차 수정을 거쳐 2005년 12월 5일 게재확정되었습니다.