# 합형식의 군집 유효화 지수의 분석과 새로운 지수 개발

김민호°, R.S. Ramakrishna
광주과학기술원 정보통신공학과
{mhkim°, rsr}@gist.ac.kr

# Analysis and New Indices of Cluster Validity Indices in Summation Type

Minho Kim°, R.S. Ramakrishna
Dept. Information and Communications, Gwangju Institute of Science and Technology (GIST)

요 약

군집 유효화 평가란 기본적으로 클래스 (Class)에 대한 정보가 주어지지 않은 상태에서 다양한 입력 변수에 의해 발생되는 군집화의 결과들을 평가하여 그들 중에서 주어진 데이터 집합의 자연적인 분할 상태에 가장 적합한 결과를 찾는 기법을 말한다. 군집 유효화 평가에서 그 척도로 사용되는 것이 군집 유효화 지수이다. 본 논문에서는 우선 현존하는 다양한 군집 유효화 지수들 중에서 합 형식을 가지는 지수들을 다룬다. 구체적으로 이 지수들의 설계 원리와 각 지수들의 부합성 (Compliance) 분석한다. 다음으로 분석을 통해 밝혀진 그들의 단점을 보완할 수 있는 새로운 군집 유효화 지수들을 제안한다. 마지막으로 기존의 군집 유효화 지수들을 포함한 새로이 제안한 지수들의 성능을 실험 학습을 통해 평가한다.

## 1. 서론

군집화 (Clustering)는 전체 데이터 집합을 비슷한 패턴을 가지는 데이터 객체들의 그룹으로 나누는 작업을 의미한다 [3] [4] [6]. 이러한 군집화는 분류 (Classification)와는 다른 비감독 학습 (Unsupervised Learning)으로 알려져 있다. 이것은 학습을 하는 동안 각 데이터 객체에 대한 클래스 정보를 사용하지 않음을 의미한다. 하지만, 대부분의 군집화 알고리즘은 최적의 결과를 위해 입력 변수의 조절을 요구한다. 대표적인 예로써 K-means 와 같은 알고리즘은 군집 수 $k$ 를 입력해야 한다. 잘 알려져 있듯이, 그들의 군집 결과의 품질은 이 입력 변수에 의해 크게 좌우된다. 이러한 입력 변수를 최적화 시키기 위해서 일반적으로 클래스 정보를 이용하게 되는데, 이것은 진정한 비감독 학습의 근본 취지와는 모순된 부분이다. 이러한 문제를 해결하기 위한 방법으로써 최근 *군집 유효화 지수* (*Cluster Validity Index, CVI*)에 대한 관심이 높아지고 있다 [1] [2] [5] [6]. CVI 는 각 군집화 알고리즘의 입력 변수들을 변화시켰을 때 최적의 군집 결과를 낳을 경우 최소값 또는 최대값을 가짐으로써 입력 변수의 최적성을 지시하게 된다.

본 논문에서는 기존의 다양한 CVI들의 설계시 가정된 기저 원리를 분석한다. CVI들은 두 계열로 나눌 수 있는데, 그 기본 요소인 *군집 내부 거리* (*Distance Within a Cluster*)와 *군집 사이 거리* (*Distance Between Clusters*)의 결합 방식에 따라 합의 형식과 비의 형식으로 나뉜다. 본 논문은 두 계열중에서 합형식의 CVI들만을 다룬다. 많은 CVI들의 군집 내부 거리에 대한 정의에서는 그 기저 설계 원리에 위배된 평균화 (Averaging) 기법이 사용되고 있다. 평균화는 잘 알려져 있듯이 번짐 (Blurring) 효과의 문제점을 가지고 있다. 이 문제점을 내포하고 있는 $v_{sv}$ [5], SD [4] 에 대한 대안으로써 각각에 대한 새로운 CVI 들을 제안한다.

## 2. 군집 유효화 지수

많은 군집 유효화 지수들은 일반적으로 다음과 같은 두 가지 평가 기준을 조합하여 정의 되며, 이러한 기준은 [6]에서 제안되었다.

1. *응집도* (*Compactness*): 군집의 멤버 데이터 객체들이 서로 얼마나 가까운가를 나타내는 척도이다. 대표적인 예로써, variance를 들 수 있다.
2. *분리도* (*Separability*): 군집들이 서로 얼마나 멀리 떨어져 있는가를 나타내는 척도다. 군집 중심 사이의 거리가 이에 대한 대표적인 예라 할 수 있다.

즉, 좋은 군집화는 군집의 멤버 데이터 객체들 사이는 서로 가까워야 하며, 군집들 사이는 충분히 멀리 떨어져 있도록 하는 군집화를 의미한다.

## 3. 기존의 CVI 들

합의 형식으로 된 지수들은 실제로는 군집 내부 거리와 군집 사이 거리에 대해 적절한 가중치 (weight)를 준 다음

에 합을 하게 된다.

최근 제안된 합의 형식을 취하는 지수들로써 $v_{SV}$ [5], SD [4]를 들 수 있다. 먼저 Eq. (1)에 $v_{SV}$가 정의 되어 있다. Eq. (1)에서 $v_U$는 군집 내부 거리에 해당하는 값이고, $v_O$는 군집 사이 거리에 해당하는 값이다. $v_{SV}$계산에서 사용된 $v_{UN}$와 $v_{ON}$은 $v_U$와 $v_O$를 [0,1]의 범위로 각각 min-max 정규화 (Normalization)시킨 값으로써 가중치에 대한 해법으로 채택되었다.

$$v_{sv}(nc) = v_{uN}(nc) + v_{oN}(nc)$$
$$v_u(nc) = \frac{1}{nc}\sum_{i=1}^{nc}\left(\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i)\right) \quad (1)$$
$$v_o(nc) = \frac{nc}{d_{\min}}, \quad d_{\min} = \min_{i\neq j} d(c_i,c_j)$$

*SD* 지수는 Eq. (2)에서처럼 *Scat*와 *Dis*의 합으로써 정의 되어 있다. 이 때, *Scat*에 대해 가중치 *a*가 주어졌다. 수식에서 $\sigma$는 각 차원에 대한 표준 편차들로 구성된 벡터로써 $\sigma = (\sigma_1, \cdots, \sigma_d)$가 된다.

$$SD(nc) = a \cdot Scat(nc) + Dis(nc), \quad a = Dis(nc_{\max})$$
$$Scat(nc) = \frac{1}{nc}\sum_{i=1}^{nc}\frac{\|\sigma(c_i)\|}{\|\sigma(X)\|} \quad (2)$$
$$Dis(nc) = \frac{D_{\max}}{D_{\min}}\sum_{i=1}^{nc}\left(\sum_{j=1}^{nc} d(c_i,c_j)\right)^{-1}$$
$$D_{\max} = \max_{i\neq j} d(c_i,c_j), \quad D_{\min} = \min_{i\neq j} d(c_i,c_j)$$

## 4. 분석과 새로운 CVI 들

CVI 들이 어떻게 최적의 군집 결과를 찾을 수 있는지를 이해하기 위해서는 먼저 그들의 설계 원리를 이해해야 한다. Fig. 1 의 예시 (Illustration)를 보면서 합의 형식을 취하는 CVI 들의 설계 원리를 살펴 보도록 하자. 그림의 예는 최적 군집 수를 결정하기 위해 군집 수 (*nc*)를 변화시켜 가면서 얻은 군집 내부 거리 (*dW*)와 군집 사이 거리 (*dB*)에 대한 그래프들이다. 물론 이 그래프들의 형태는 이상적인 상황을 가정한 것이다. 즉, 그림에서 보는 것처럼 *dW*는 $nc = nc_{optimal}$에서 $nc = nc_{optimal} - 1$ 로 감소할 때 값의 급격한 증가가 발생하고, *dB*는 $nc = nc_{optimal} + 1$ 에서 $nc = nc_{optimal}$ 로 감소할 때 값의 급격한 감소가 발생한다고 가정한다. 이렇게 함으로써 두 그래프의 합으로 형성된 CVI 에 대한 그래프는 $nc = nc_{optimal}$에서 최소값을 가지게 된다. 여기에 숨겨져 있는 또 하나의 가정이 있는데, *dW* 와 *dB* 의 값의 범위가 그림에서 보는 것처럼 서로 비슷하거나 동일하다는 것이다. 다시 말해서, *dW* 와 *dB* 에 적절한 가중치 (weight)가 주어졌다라고 가정한 것이다. [5]에서도 이와 비슷한 설계 원리가 설명되어 있다.

CVI 가 최적 군집 결과를 찾는데 성공하기 위해서는 당연히 앞에서 설명한 가정들이 모두 잘 지켜졌을 때일 것이다. 하지만, 기존의 CVI 들이 위의 가정들을 충실하게 만족시키도록 설계된 것으로 보이지는 않는다.

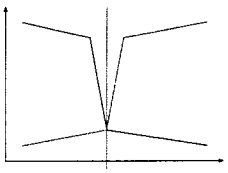


Fig. 1. 합 형식의 CVI 에 대한 설계 원리

가장 두드러진 문제가 바로 *dW*에 해당하는 부분이다. Eq. (1)의 $v_{SV}$에서 그 문제점을 살펴 보자. *dW*는 $nc <= nc_{optimal}$에서는 *nc*를 줄이게 되면 불필요한 군집의 병합으로 인해 병합된 군집의 *dW*값의 급격한 증가를 유발해야 것이다. 그런데, $v_{SV}$에서 *dW*에 해당하는 $v_U$는 각 군집의 centroid와 객체사이의 거리를 평균한 것이다. 이 방법은 불필요한 병합이 일어난 군집에서 발생한 효과 (응집도의 급격한 감소)가 평균을 취함으로 인해 묻혀져 버리게 되는 문제를 가져 올 수 있다. 따라서 이러한 문제에 대한 대안으로 $v_U$*를 Eq. (3)와 같이 제안한다. 또한 $v_U$*를 사용하는 $v_{SV}$의 대안으로써 $v_{SV}$*도 같이 정의 된다.

$$v_u^*(nc) = \max_{i=1\ldots nc}\{v_{u,i}(nc)\} = \max_{i=1\ldots nc}\left\{\frac{\sum_{j=1}^{N} u_{ij}^2 \|x_j - c_i\|^2}{n_i}\right\} \quad (3)$$
$$v_{sv}^*(nc) = v_{uN}^*(nc) + v_{oN}(nc)$$

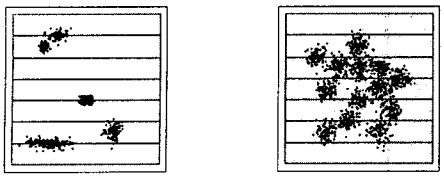
SD 의 *Scat*도 $v_U$에서처럼 평균화를 사용하고 있으므로 동일한 문제가 예상된다. 따라서 $v_U$*와 비슷한 방법으로 *Scat*에 대한 대안으로써 *Scat**를 Eq. (4)와 같이 정의한다. 이에 따라 *SD** 도 새롭게 정의한다.

$$Scat^*(nc) = \max_{i=1\ldots nc}\left\{\frac{\|\sigma(c_i)\|}{\|\sigma(X)\|}\right\} \quad (4)$$
$$SD^*(nc) = a \cdot Scat^*(nc) + Dis(nc), \quad a = Dis(nc_{\max})$$

## 5. 실험 결과

이번 절에서는 기존의 CVI 들과 본 연구에서 제안한 CVI 들의 성능을 평가할 것이다. 이를 위해 군집화 알고리즘으로는 [6]에서 제안한 VALCLU 를 사용할 것이다. VALCLU 알고리즘은 계층적 군집화 알고리즘의 일종으로써 CVI 를 통해 계층적 군집화 결과의 여러 단계 (Level)중에서 최적의 군집 결과를 가지는 단계를 효과적으로 찾을 수 있는 알고리즘이다. 실험에 사용된 데이터 집합은 총 3 개로써 그 중 둘을 Fig. 2 에 나타냈다. Dataset 3 은 Dataset 2 에서 군집들 사이의 거리를 조절함으로써 파생된 데이터 집합이다.

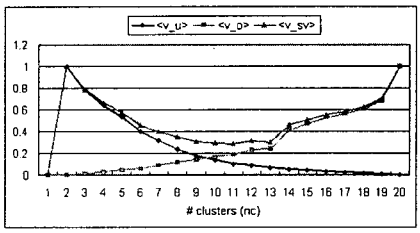
첫 번째 실험은 $v_{SV}$와 $v_{SV}$*을 비교한 실험이다 . Fig. 3 는 Dataset 3에 대한 $v_{SV}$와 $v_{SV}$*에 대한 실험 결과를 보인 것이다. Fig. 1에서 우리는 *dW*가 $nc = nc_{optimal}$에서 $nc = nc_{optimal} - 1$로 감소할 때 값이 급격하게 증가해야 한다고 가정했다. 하지만, Fig. 3 (a)의 *dW*, 즉, $v_U$ 그래프를



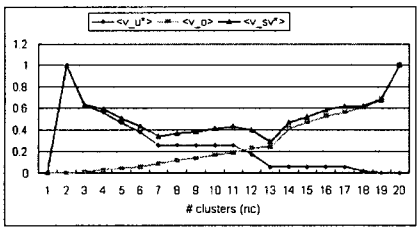
(a) Dataset 1 (b) Dataset 2

Fig. 2. 합성 데이터 집합

살펴보면 $nc$ = 13 ($nc_{optimal}$)에서 $nc$ = 12로 감소하는 동안 그러한 증가를 볼 수가 없다. 이로 인해 비록 $dB$, 즉, $v_o$가 $nc$ = 13에서 CVI의 설계 원리에 적합한 특성 ($v_o$값의 급격한 감소)을 보여 주었음에도 불구하고 $v_{sv}$가 $nc_{optimal}$을 지칭하는데 실패하였다. 다시 말해서 $v_{sv}$가 $nc$ = 13 ($nc_{optimal}$)가 아닌 $nc$ = 11에서 최소값을 가졌다. 하지만, Fig. 3 (b)의 $v_u$*는 $v_u$와는 달리 $nc$ = 13에서 급격한 증가를 보여 주어 $v_{sv}$*가 정확히 $nc_{optimal}$을 찾았다. 이것은 4절에서 설명했듯이 $v_u$에 내포된 평균화의 문제점때문이며 $v_u$*의 새로운 설계에서 효과적으로 해결되었다.




(a) $v_{sv}$




(b) $v_{sv}$*

Fig. 3. $v_{sv}$ 와 $v_{sv}$*의 비교 (Dataset 3)

지면 관계상 보이지는 못했지만 (Dataset 2 에 대해) SD 와 SD* 에서도 $v_{sv}$ 와 $v_{sv}$*에서와 유사한 결과를 확인할 있었다.

Fig. 4는 Dataset 1에 대한 $v_{sv}$*의 결과이다. $v_u$*와 $v_o$ 그래프들을 살펴 보면 $nc$ = 4 ($nc_{optimal}$)에서 적절한 패턴이 나타났음에도 불구하고 최소값을 가지는데 실패했다. 이 문제점은 부분적으로 $v_u$*가 $nc$ = 4에서 $v_o$에 비해 상대적으로 큰 값을 가지고 있어서 발생한 것이라 할 수 있다. 즉, $v_u$*와 $v_o$에 대한 가중화 (weighting)가 부적절했음을 의미한다. 합의 형식을 취하는 CVI들($v_{sv}$, SD)은 항상 적절한 가중화를 시켜 주어야 하는 단점을 가지고 있다. 실제로 이것은 또 하나의 최적화 문제를 야기시킨다. (참고로, ' 제 (II)부 비 형식의 CVI' 논문에서 다루는 지수들은 가중화에 대한 문제점가 없는 장점이 있다.)

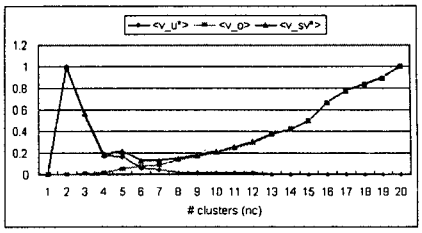



Fig. 4. $v_{sv}$* 의 약점 (Dataset 1)

## 6. 결론

본 연구에서는 기존의 다양한 군집 유효화 지수 (CVI)들에 대한 분석과 분석을 통해 밝혀진 이들의 단점을 보완할 수 있는 CVI들을 제시하였다. 분석과 실험 학습에서 많은 CVI들이 그 구성 요소의 하나인 군집 내부 거리에 해당하는 요소에서 평균화로 인한 번짐 효과로 인해 불필요한 병합으로 나타나는 한 군집의 내부 거리 값의 급격한 변화가 저역 필터링 (Low-Pass Filtering)되는 문제점을 확인하였다. 즉, CVI의 기저 설계 원리에 부합하지 않는 부분을 발견하였다. 이 문제점을 해결하기 위해 각 군집의 내부 거리 중에서 최대값을 전체 군집 내부 거리로 정의하는 새로운 방법을 제안하였으며, $v_{sv}$, SD 에 적용하여 $v_{sv}$*, SD*을 새롭게 정의하였다. 실험 결과에서도 그 이전 버전들보다 향상된 성능을 보여 주었다. 본 연구에서 제안된 새로운 CVI들과 ' 제 (II)부 비 형식' 논문에서 제안된 CVI 들은 군집화 알고리즘의 복잡한 특성 및 입력 변수 최적화에 익숙하지 않은 실세계 사용자 (Domain Expert)들에게 훨씬 더 신뢰성 있는 도구로써 사용될 수 있을 것이다.

## [참고문헌]

[1] M.J.A. Berry, G. Linoff, " Data Mining Techniques: For Marketing, Sales, and Customer Support," John Wiley & Sons, 1997

[2] M. Halkidi and M.Vazirgiannis, " Quality scheme assessment in the Clustering process," Proc. PKDD (Principles and Practice of Knowledge Discovery in Databases), Lyo, France, 2000.

[3] J. Han and M. Kamber, " Data Mining: Concepts and Techniques," Morgan Kaufmann, 2001.

[4] A.K. Jain, M.N. Murty, and P.J. Flynn, " Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[5] D.-J. Kim, Y.-W. Park, and D.-J. Park, " A Novel Validity Index for Determination of the Optimal Number of Clusters," IEICE Trans. Inf. & Syst., Vol.E84-D, No. 2, 2001.

[6] Minho Kim, R. S. Ramakrishna, " A New Clustering Algorithm Based On Cluster Validity Indices," LNAI/LNCS (DS2004), vol. 3245, pp. 322-329, 2004.