비형식의 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 Ratio Type

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

요 약

군집 유효화 평가는 군집화 알고리즘을 진정한 의미의 비감독 학습이 가능하도록 만든다는 의미에서 그 중요성이 더해지고 있다. 본 논문에서는 이 군집 유효화 평가에 일반적으로 이용되는 군집 유효화 지수들의 설계 원리를 분석하고 기존 지수들의 부합성을 분석한다. 우리는 제 (I) 부에서 합 형식의 지수들을 다루었으며, 본 논문에서는 비 형식의 지수들을 다룬다. 합형식의 CVI에서처럼 저역 필터링의 문제점을 해결하였으며, 또한, 부작용 없이 비형식의 지수들의 성능을 향상시킬 수 있는 새로운 기법을 제시한다. 새로운 지수들의 성능은 실험 학습을 통해 제시된다.

1. 서론

많은 군집화 알고리즘들의 결과 품질은 군집화되는 데이터 집합의 특성과 입력 변수에 대해 결정적인 영향을 받는다. 이것은 임의의 군집화 알고리즘에 대해 특정 데이터 집합의 특성을 제대로 고려하지 않은 부적합한 입력 변수를 사용하면 실제 데이터 집합과는 틀린 군집 결과를 낳을 수 있음을 의미한다. 즉, 어떠한 입력 변수가 실제 데이터 집합에 가장 적합한 군집 결과를 낳는지를 찾아내기 위해서는 군집 결과 자체를 실제 데이터 집합에 비추어 최적성을 평가하는 기술이 요구되며, *군집 유효화 지수* (*Cluster Validity Index, CVI*)가 바로 이러한 군집 결과의 평가에 널리 사용되고 있다 [1] [2] [3] [4] [5].

제 (I) 부에서는 ' 합형식의 CVI' 에 대한 기본 설계 원리와 기존 CVI 들의 부합성을 분석하였다. 이 때, *군집 내부 거리* (*Distance Within a Cluster, dW*) 계산에 있어 평균화로 인한 번짐 효과의 문제점을 제기하였으며 대안을 제시하였다. 하지만, 합형식의 CVI 들은 dW 와 *dB* (*군집 사이 거리, Distance Between Clusters*)의 결합할 때 발생하는 가중화라는 문제점을 가지고 있다.

본 논문에서는 가중화의 문제가 없는 비형식의 CVI를 다룰 것이다. 현존하는 대부분의 비형식의 CVI들도 합형식의 CVI에서처럼 평균화의 문제점을 가지고 있다. 따라서, 본 논문에서는 대표적인 비형식의 CVI인 XB [5]에 대해 이 문제를 해결한 새로운 CVI들을 제시한다. CVI들의 설계 원리는 실세계 응용의 다양하고 복잡한 모든 상황을 고려하기에는 역부족일 수 밖에 없다. 이러한 단점을 부작용 없이 보완할 수 있는 기법을 제안하며, 실제로 이것을 적용하여 XB의 변이한 새로운 CVI를 제안한다. 실험 학습에서 새로 제안된 CVI들은 향상된 성능도 제시한다.

2. 기존의 CVI

비형식을 취하는 CVI는 그 이름에서도 알 수 있듯이 dW와 dB의 비를 통해서 얻어진다. 대표적인 지수로써 *XB* [5]를 들 수 있다. XB의 정의는 Eq. (1)와 같다. XB는 $\sum_{k=1}^{nc}\sum_{j=1}^{N} u_{kj}^2 d(x_j,c_k)^2 / N$ 와 $\min_{i,j} d(c_i,c_j)^2$ 로 분리할 수 있으며, 각각 dW와 dB로 정의될 수 있다. 수식에서 u_{kj}는 데이터 객체 x_j의 군집 C_k에 대한 포함 정도 (degree of membership)를 나타내는 값으로써 [0,1]의 범위를 가질 수 있으나 본 논문에서는 강(Hard) 군집화로 그 범위를 한정할 것이므로 {0, 1}의 값을 가질 수 있다.

$$XB(nc) = \frac{\sum_{k=1}^{nc}\sum_{j=1}^{N} u_{kj}^2 d(x_j,c_k)^2}{N \cdot \min_{i,j} d(c_i,c_j)^2} \quad (1)$$

3. 분석과 새로운 CVI 들

비형식을 취하는 CVI는 다음과 같은 가정에서 설계되었다:

1) *dW*는 $nc = nc_{optimal}$에서 $nc = nc_{optimal} - 1$로 감소할 때 값의 급격한 증가가 발생한다.
2) *dB*는 $nc = nc_{optimal} + 1$에서 $nc = nc_{optimal}$로 감소할 때

값의 급격한 증가가 발생한다 (Fig. 1).

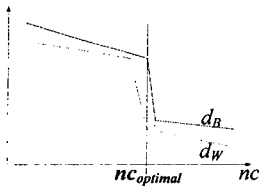


Fig. 1. 비형식의 CVI 에 대한 설계 원리

nc = $nc_{optimal}$에서 dW는 nc > $nc_{optimal}$에서보다 상대적으로 매우 큰 비율로 값의 증가가 있는 반면에 dB는 그 증가 비율이 상대적으로 적다. 하지만, 그래프에서도 예상이 되듯이 $nc_{optimal}$ 이외의 다른 nc에서는 dW와 dB 모두 상대적으로 비슷한 비율의 변화가 발생하기 때문에 이와 같은 큰 차가 발생하지는 않는다. 따라서, nc = $nc_{optimal}$에서 CVI가 최소값을 가지게 된다. 물론, CVI를 dB/dW로 정의하면 nc = $nc_{optimal}$에서 최대값을 가진다.

비형식의 CVI에서도 합형식 CVI의 dW에 발생했던 것과 동일한 문제가 발생한다. 즉, dW에서 각 군집의 분산도 (decompactness)를 평균화시킬 경우 하나의 군집에서 발생한 분산도의 급격한 변화가 묻혀지게 된다. XB에서도 이런 현상을 발생할 수 있는데, 이 문제를 해결하기 위해 XB를 Eq. (2)과 같이 XB*으로 재정의 한다.

$$XB^*(nc) = \frac{\max_{k=1\ldots nc}\{\frac{\sum_{j=1}^{N} u_{kj}^2 \|x_j - c_k\|^2}{n_k}\}}{\min_{i,j}\|c_i - c_j\|^2} \quad (2)$$

Fig. 1에는 암묵적인 가정이 하나 더 있다. 그것은 nc = $nc_{optimal}$ 주위에서 나타나는 dW와 dB의 패턴이 단 한번 일어날 것이라는 가정이다. 하지만, 실세계의 응용에서는 이것이 보장되지 않는다. 단적으로 비슷한 크기의 다수의 군집이 비슷한 거리로 분리되어 있을 경우 nc < $nc_{optimal}$에서 그러한 패턴이 여러 번 나타날 수 있으며 nc = $nc_{optimal}$에서보다 더 급격한 변화를 가질 수도 있다. 즉, $dW'(nc)$ = $dW(nc)/dW(nc_{optimal})$와 $dB'(nc)$ = $dB(nc)/dB(nc_{optimal})$라 했을 때 $dW'(nc)$ < $dB'(nc)$가 되어 CVI(nc) < CVI($nc_{optimal}$)인 상황이 나타날 수도 있다 (CVI(nc) = $dW(nc)/dB(nc)$ 일 때). 이 문제를 완화시키기 위한 대안을 제시하기 전에 몇 가지 관찰을 해 보도록 하자. nc < $nc_{optimal}$에서 dW 값은 크면 클수록 CVI 값을 증가시키기 때문에 바람직한 현상이라 할 수 있다. 여기에서 $diff_{dW}$ = $dW(nc)$ - $dW(nc-1)$이라 하고 $maxDiff = \max_{nc_{max}\ldots nc} diff_{dW}$라 정의해 보자. $maxDiff$ 값의 특징은 다음과 같다.

i) $nc \geq nc_{optimal}$에서 비교적 작은 값을 가진다.
ii) nc < $nc_{optimal}$에서만 큰 값을 가진다.

이러한 특성은 dW에서 요구되는 바람직한 특성과 동일한 것이다. 따라서, $maxDiff$는 부작용이 없이 dW를 보강할 수 있음을 알 수 있다. $maxDiff$를 이용하여 XB*를 향상시킨 새로운 CVI, XB**를 Eq. (3)과 같이 제안한다.

$$XB^{**}(nc) = \frac{\max_{k=1\ldots nc}\{\sum_{j=1}^{N} \frac{u_{kj}^2 \|x_j - c_k\|^2}{n_k}\} + maxDiff}{\min_{i,j}\|c_i - c_j\|^2} \quad (3)$$

4. 실험 결과

이번 절에서는 기존의 CVI 들과 본 연구에서 제안한 CVI 들의 성능을 평가할 것이다. 이를 위해 군집화 알고리즘으로는 ' 제 (I) 부: 합형식의 CVI' 논문에서 채택했던 것과 동일한 VALCLU [5]를 사용할 것이다.

실험에 사용된 데이터 집합은 총 4개이며, Fig. 2에서는 Dataset 4를 보여주고 있다(17개 군집). Dataset 1~3은 ' 제 (I) 부: 합형식의 CVI' 논문에 찾을 수 있다.

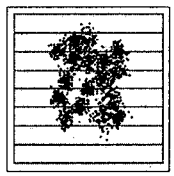
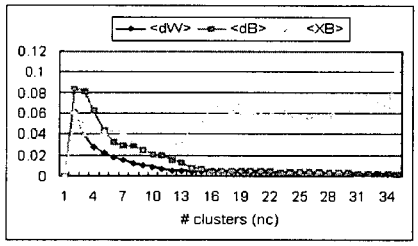


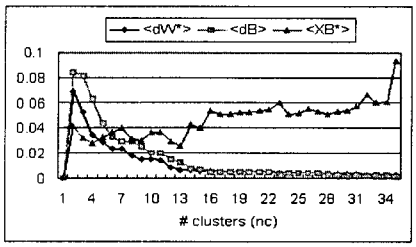
Fig. 2. 합성 데이터 집합 (Dataset 4)

먼저 Dataset 3에 대한 XB와 XB*의 비교 실험을 살펴보자 (Fig. 3). Fig. 3 (과 Fig. 4)의 그래프들은 세 그래프 (dW, dB, XB)를 그림 하나에 모두 집어넣기 위해 비율을 조절하였는데, 이것은 원본과 비교하여 그래프의 원활한 해석을 위한 것이며 실제 유효화 평가를 방해하지는 않는다. 이번 실험에서 XB* 역시 dW의 평균화의 문제를 효과적으로 처리하고 있음을 보여주고 있다. dW와 dW* 그래프 상으로는 알아차리기 어려울지 모르지만, 실제로는 nc = 13 ($nc_{optimal}$)에서 nc = 12로 변할 때 XB의 경우 12.8%의 증가가 발생했던 반면 XB*의 경우 34.2%의 증가가 발생했다. 역으로 계산했을 경우 XB와 XB*에 대해 각각 11.4%와 25.5%의 감소가 있었다. 즉, 제안된 방법에 의해 nc = 13에서 dW 이 다른 곳 보다 상대적으로 작은 값을 가지도록 함으로써 dW' < dB' 가 되어, XB*가 최소값을 가질 수 있도록 하고 있음을 알 수 있다.

다음 실험은 Dataset 4에 대해 XB*와 XB**를 비교한 것이다. Fig. 4에 이를 보이고 있다. Fig. 4 (a)에서 보면 nc = 13 ($nc_{optimal}$)에서 Fig. 1에서와 같은 dW와 dB값의 상대적인 급격한 변화가 발생했음에도 불구하고 XB가 최소값을 가지는데 실패했다. 그 이유는 이러한 패턴이 nc < $nc_{optimal}$에서 여러 번 발생했으며 그들 중 하나 (nc = 4)가 nc = 13 일 때 보다 더 작은 값을 가지고 있기 때문이다. 다시 말해서 실세계의 데이터 집합은 Fig. 1의 이상적인 경우처럼 단 한 번만 그러한 패턴을 보여주는 것이 아님을 알 수 있다. 다른 데이터집합에서도 이런 문제를 확인할 수 있다. XB**는 $maxDiff$를 이용하여 XB*를 보강한 것인데 설계 원리와 동일하게 nc < $nc_{optimal}$에서만 큰 값을 가짐으로써 $dW'(nc)$ > $dB'(nc)$로 만들어 이 문제점을 해결한 것을 볼 수 있다.

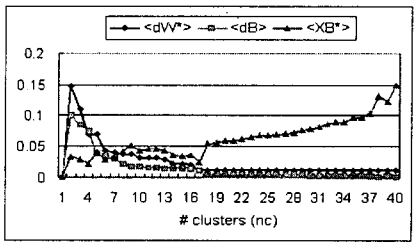


(a) XB

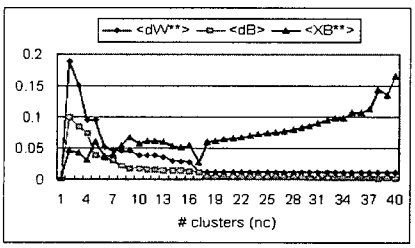


(b) XB*

Fig. 3. XB 와 XB*의 비교(Dataset 3)



(a) XB*



(b) XB**

Fig. 4. XB* 와 XB**의 비교 (Dataset 4)

표 1은 다양한 데이터 집합들에 대한 CVI (' 제 (I) 부: 합형식의 CVI' 포함)들의 실험 결과를 요약한 것이다. 표의 결과에서 주지할 점은 비형식을 취하는 CVI들보다 합형식을 취하는 CVI들이 더 나쁜 결과를 보여 주었다는 것이다. 이러한 현상은 합형식의 CVI들은 *dW*와 *dB*를 결합할 때 요구되는 가중화가 적절하지 못했기 때문으로 분석된다. ' 제 (I) 부: 합형식의 CVI' 논문에서도 이러한 점을 실험을 통해 제시하였다. 비형식의 CVI 중에서는 당연히 예상했던 것처럼 XB**가 가장 좋은 결과를 보여 주었는데, 이것은 dW에 대해 평균화의 문제점을 제거하였을 뿐만 아니라 *maxDiff*에 의해 보강까지 되었기 때문이다.

표. 1.각 CVI 에 의해 제안된 군집수와 성공여부

	V_{SV}	V_{SV}*	SD	SD*	XB	XB*	XB**
Dataset 1	15 (X)	18 (X)	4 (O)	4 (O)	4 (O)	4 (O)	4 (O)
Dataset 2	13 (O)	13 (O)	7 (X)	13 (O)	13 (O)	13 (O)	13 (O)
Dataset 3	11 (X)	13 (O)	6 (X)	11 (X)	11 (X)	13 (O)	13 (O)
Dataset 4	15 (X)	16 (X)	7 (X)	6 (X)	4 (X)	4 (X)	17 (O)

5. 결론

본 논문에서는 ' 제 (I) 부: 합형식의 CVI' 에 이어 비형식의 CVI에 대한 기본 설계 원리 분석하고 분석을 통해 밝혀진 이들의 단점을 보완할 수 있는 다양한 CVI들을 제시하였다. 대표적인 비형식의 CVI인 XB에서도 기존 합형식의 CVI에서처럼 dW의 저역 필터링 (Low-Pass Filtering)으로 인한 문제점을 확인하였다. 이 문제점을 V_{SV}*, SD*와 유사한 방법으로 XB*을 정의함으로써 해결하였다. 또한, 복잡한 데이터 집합에 대해서도 비형식 CVI들의 설계 원리를 효과적으로 보완할 수 있는 *maxDiff*를 제안하였으며, 이것을 XB* 에 적용하여 XB**를 제안하였다. 실험 학습에서 합 형식의 CVI들보다 비 형식의 CVI들이 더 좋은 성능을 보여 주었으며, 그 중에서도 XB**가 가장 좋은 결과를 보여주었다. ' 제 I 부: 합형식' 의 논문과 더불어 본 논문에서 제시한 설계 원리 및 보완 기법은 이들 논문에서 제시한 CVI들 뿐만 아니라 다른 CVI에 대해서도 적용이 가능할 것으로 기대되며, 향후 이에 대한 연구를 보강하고자 한다.

[참고문헌]

[1] D.L. Davies and D.W. Bouldin, " A Cluster Separation Measure, " IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), vol. 1, no. 2, pp. 224-227, 1979.

[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] 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, Feb. 2001.

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

[5] X.L. Xie and G.A. Beni, " A Validity Measure for Fuzzy Clustering, " IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), vol. 3, no. 8, pp. 841-846, 1991.