고차원 응용에서의 군집 유효성 평가 기법

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

Cluster Validity Assessment Techniques for High-Dimensional Applications

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

요 약

군집 유효성은 다양한 입력 변수에 따라 변하는 군집화 알고리즘의 결과들을 평가하는 것이다. 본 논문에서는 고차원의 데이터 집합에 대한 군집 유효성의 문제점에 대한 새로운 해결책을 제시한다. 고차원 군집화 결과들을 평가할 때 발생하는 기존의 군집 유효성 지수들의 적용성의 문제점을 살펴보고, 고차원으로 인해 발생하는 문제를 효과적으로 다룰 수 있는 다양한 새로운 군집 유효성 지수들을 제안한다. 제안된 군집 유효성 지수들은 본 논문에 제공된 실험에서 최적의 군집 유효성 결과를 제공한다.

1. 서 론

군집화는 분류 (Classification)와는 다른 비감독 학습 (Unsupervised Learning)이다. 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 등과 같은 알고리즘은 정확한 군집 수 k를 입력해야 한다. 잘 알려져 있듯이, 그 군집 결과의 품질은 이 입력 변수에 의해 크게 좌우된다. 이 입력 변수를 최적화시키기 위해서는 결국 클래스 정보를 이용해야만 하는데, 이것은 진정한 비감독 학습의 근본 취지와는 모순된 부분이다. 이러한 문제를 해결하기 위한 방법으로써 최근 군집 유효성 지수 (Cluster Validity Index, CVI)에 대한 관심이 높아지고 있다 [4] [5] [6] [7]. CVI는 각 군집화 알고리즘의 입력 변수들을 변화시켰을 때 최적의 군집 결과를 낳을 경우 최소값 또는 최대값을 가짐으로써 입력 변수의 최적성을 지시하게 된다.

현존하는 대부분의 CVI는 낮은 차원의 데이터 집합에서 합리적인 결과를 보여주는 군집화 알고리즘들의 입력 변수를 최적화하기 위해 개발되었다. 그리고, 이들 CVI는 데이터 객체들에 대한 거리 함수를 기반으로 정의 되어 있다.

그런데, [3]에서 특성 벡터가 고차원의 공간에 존재할 때 수 많은 데이터 분포들과 거리 함수들에 대해 데이터 객체들의 모든 쌍에 대한 거리는 거의 같은 값을 가지게 되는 것으로 증명되었다. 이것은 고차원의 데이터 집합에 대해서는 기존의 거리 함수들을 통해 얻을 수 있는 객체들 사이의 유사도 측정 능력을 잃게 됨을 의미한다. 이것이 바로 잘 알려진 차원의 저주 (Curse of Dimensionality)이다. 다시 말해서, 기존의 거리 함수에 기반을 둔 많은 군집 유효성 지수들로는 고차원의 데이터 집합에 대해 좋은 결과를 기대하는 것은 요원하다 할 수 있다.

따라서, 본 논문에서는 이러한 문제점을 해결하기 위해 고차원 군집화에서 우수한 성능을 보인 투영 군집화의 기반 요소 기법을 이용하여 다양한 새로운 CVI들을 제안한다. 본 논문에서 제시한 다양한 실험은 제안한 CVI들의 우수성을 증명한다.

본 논문의 구성은 다음과 같다. 2절에서는 고차원의 데이터 집합에 대한 군집화에 대해 논의하고, 3절에서는 군집 유효성 지수에 대해 다룰 것이다. 실험 결과와 결론은 각각 4절과 5절에서 제시한다.

2. 고차원 데이터 집합에 대한 군집화

고차원의 데이터 집합에서 모든 차원이 임의의 한 군집과 연관되어 있을 가능성은 매우 적다. 이것은 연관된 차원 이외의 차원들은 군집화에 방해될 수도 있음을 의미한다. 실제로 [3]에서 고차원의 데이터 집합이 특정한 조건과 분포를 가진 다면 데이터 객체의 모든 쌍들에 대한 거리가 거의 같은 값을 가진다라고 증명해 보였다.

이 문제를 처음으로 다룰 당시에는 군집화를 행하기 이전에 군집화에 유용한 차원만을 미리 선택하는 특성 선정 (Feature Selection)에 대한 다양한 방법들이 제시 되었다. 하지만, 특성 선정은 군집화 이전에 일부 특성만을 미리 골라내기 때문에 정보의 손실을 유발할 수도 있다. 실제 많은 실세계의 전형적인 고차원 데이터 집합들의 경우, 특정 데이터 객체들은 임의의 한 차원 집합에 대해 밀접하게 군집화되고 다른 데이터 객체들은 다른 차원 집합에 대해

밀접하게 군집화된다. 그러므로, 모든 군집에 대해 만족하는 소수의 차원들만으로 구성된 단일 집합을 찾는 것은 거의 불가능 할 수 있다.

이러한 특성 선정을 통한 군집화와 전체 차원을 원시적으로 이용하는 군집화의 문제점에 대한 해결책은 앞에서 언급했던 고차원 데이터 집합에서 서로 다른 그룹들은 차원에 대해 서로 다른 상관성을 가진다는 사실에서 찾을 수 있다 [1]. 이를 이용하면 다음과 같은 투영된 군집(Projected Cluster)을 정의할 수 있으며, 고차원의 데이터 집합에 대한 군집화 문제는 이 정의를 바탕으로 해결할 수 있다. *투영된 군집*은 전체 차원의 한 부분 집합에 대해 전체 데이터 집합의 일부 데이터 객체들이 밀접하게 연관되어 있을 때, 차원의 부분 집합 D와 함께 정의된 데이터 객체의 부분 집합 C를 의미한다.

이것을 실제 군집화에 응용한 알고리즘이 PROCLUS(PROjected CLUStering) [1] 이며, 핵심 요소인 거리 함수로써 다음과 같은 Manhattan 부분(Segmental) 거리 함수를 정의하였다:

$$d_D(x_1, x_2) = \sum_{m \in D} | x_{1,m} - x_{2,m} | / | D | \quad (1)$$

여기에서, x_1과 x_2는 d차원의 데이터 객체이며, D는 전체 d 차원집합의 부분집합을 의미한다 (즉, $|D| \le d$).

여기서는 PROCLUS 알고리즘의 핵심만을 제시하였으며, 좀 더 자세한 설명은 [1]에서 구할 수 있다.

3. 군집 유효성 지수 (Cluster Validity Indices, CVIs)

많은 군집화 알고리즘들의 결과 품질은 군집화되는 데이터 집합의 특성과 입력 변수에 대해 많은 영향을 받는다. 이러한 요소에 따라 군집화 결과들의 변화를 비교 평가하여 최적 결과를 지시할 수 있는 기법이 요구되는데, 그 대표적인 도구가 군집 유효성 지수이다.

많은 군집 유효성 지수들은 일반적으로 두 가지 평가 기준, 즉, *군집 내부의 응집성*, *군집 사이의 분리도*를 조합하여 정의 된다. 왜냐하면, 좋은 군집화는 각 군집의 멤버 데이터 객체들 사이는 서로 가깝도록 해야 하며, 군집들 사이는 충분히 멀리 떨어져 있도록 할 수 있어야 하기 때문이다.

대표적인 기존의 CVI 들은 다음과 같다.

$$I(nc) = \left(\frac{1}{nc} \times \frac{E_1}{E_{nc}} \times D_{nc} \right)^p, (p = 2) \quad (2)$$

$$E_{nc} = \sum_{k=1}^{nc} \sum_{j=1}^{N} u_{kj} d(x_j, c_k), D_{nc} = \max_{i,j=1}^{nc} d(c_i, c_j)$$

$$v_{xv}(nc) = v_{uN}(nc) + v_{oN}(nc)$$

$$v_u(nc) = \frac{1}{nc} \sum_{i=1}^{nc} \left(\frac{1}{n_i} \sum_{x \in C_i} d(x, c_i) \right), v_o(nc) = \frac{nc}{\min_{i \neq j} d(c_i, c_j)} \quad (3)$$

$$XB(nc) = \sum_{k=1}^{nc} \sum_{j=1}^{N} u_{kj}^2 d(x_j, c_k)^2 \Big/ N \cdot \min_{i,j} d(c_i, c_j)^2 \quad (4)$$

$$DB(nc) = \frac{1}{nc} \sum_{i=1}^{nc} \left(\max_{j=1, nc, j \neq i} \frac{S_i + S_j}{d(c_i, c_j)} \right), S_i = \frac{1}{n_i} \sum_{x \in C_i} d(x, c_i) \quad (5)$$

위의 수식들에서 N은 전체 객체의 수, nc는 군집수, n_i는 군집 C_i의 객체의 수, c_i는 군집 C_i의 center를 각각 의미한다. 그리고, v_{uN}과 v_{oN}은 v_u와 v_o의 min-max 정규화된 값이다. 제시한 각 CVI 들에 대한 자세한 설명은 I [6], vsv [5], XB [7], DB [4] 에서 각각 찾을 수 있다.

3.1 고차원 군집화의 유효성 평가를 위한 새로운 CVI들

기존의 CVI들은 저차원의 응용에 적합한 거리함수를 통해 정의되었기 때문에 이들을 기반으로 정의된 CVI들 또한 2절에선 소개한 차원의 저주의 영향으로 인하여 고차원의 데이터 집합에 대한 군집화 알고리즘의 결과를 제대로 평가할 수 없다. 이것은 4절의 실험에서도 실제로 확인할 수 있다.

그러므로, 고차원의 응용에서 군집 결과가 주어진 데이터 집합에 얼마나 잘 부합하는지를 측정할 수 있는 새로운 CVI를 정의함은 자명하다. 이를 위해서는 고차원의 데이터 집합에 대해서도 [2]에서 정의한 군집 내부의 응집도와 군집 사이의 분리도의 의미를 새로운 CVI가 제대로 표현할 수 있어야 한다. 응집도나 분리도 또한 일반적으로 거리 함수에 의해서 정의된 것을 감안 할 때, 이 문제를 해결하기 위해서는 고차원에서 무용지물이 된 거리 함수를 재정의함으로써 가능할 것이다.

새로운 CVI를 위한 거리 함수 대한 힌트는 PROCLUS에서 사용된 $d_{D_i}(\cdot,\cdot)$에서 찾을 수 있다. [2]에 따르면 바람직한 군집은 그 군집에 있는 멤버 데이터 객체들이 서로 가까워야 한다. 그런데, $d_{D_i}(\cdot,\cdot)$를 적용한 응집도 계측은 이것을 효과적으로 표현할 수 있다. 왜냐하면 군집 집합 D_i가 해당 군집 C_i에 속한 데이터 객체들에 대해 상호 연관성을 가지는 차원들의 집합이며 $d_{D_i}(\cdot,\cdot)$는 이들 차원만을 반영한 거리 함수이기 때문이다. 분리도에 대해서도 $d_{D_i}(\cdot,\cdot)$가 효과적으로 이용될 수 있다. 그 이유는 다음과 같다. 먼저 매운 유사한 두 군집이 있을 경우 그 군집들은 동일하거나 매우 비슷한 구성의 차원 집합을 가진다고 볼 수 있기 때문에 $d_{D_i}(c_i, c_j)$가 상대적으로 작은 값을 가져서 두 군집이 병합 (merge)되어야 함을 표시할 수 있을 것이다. 다음으로 서로 다른 두 군집에 대해서는 한 군집의 차원 집합이 다른 군집의 차원 집합과는 그 구성이 많은 부분 다를 가능성이 높기 때문에 $d_{D_i}(c_i, c_j)$가 큰 값을 가져서 두 군집 사이의 분리도를 표시할 수 있을 것이다.

4. 실험 결과

본 절에서는 고차원 응용을 위해 제안된 CVI 들의 성능을 실험을 통해 알아 본다. 성능 측정에 사용된 군집화 결과들은 PROCLUS [1]를 통해서 얻어졌다. 군집 수는 군집 결과의 정확성과 밀접한 관계를 가지고 있기 때문에 예측된 최적 군집 수와 데이터 집합의 실제 군집 수를 비교함으로써 CVI 들의 성능을 평가하였다.

성능을 평가하기 위한 데이터 집합은 [1]에서 주어진 방법을 이용하여 생성하였다. 데이터 집합들은 그 신뢰성

을 높이기 위해 다양한 환경을 대표하도록 생성되었다. 예를 들면, 군집 차원 집합이 서로 완전히 다른 데이터 집합에서부터, 일부 군집 차원을 공유하거나 군집 차원 집합 전체를 공유하는 데이터 집합들도 포함하였다.

기존 및 새로운 CVI들에 대한 실험 결과가 표 1과 2에 각각 주어져 있다. 표 1에서 알 수 있듯이, t2에 대한 V_{SV}를 제외한 모든 기존의 CVI들은 각 데이터 집합의 실제 군집 수와 틀린 군집 수를 예측하였다. 즉, 실제의 최적 군집 수를 찾는데 실패했다. 표의 결과상으로는 V_{SV}가 일부 성공한 것처럼 보일 수도 있으나, 대부분의 다른 실험 데이터 집합들에서는 데이터의 실제 군집 수를 찾는데 실패했기 때문에 이것은 단지 우연이라 할 수 있다. 결과적으로 기존의 CVI들은 고차원의 데이터 집합에 대한 군집 수를 찾는데 적합하지 않다고 할 수 있다. 물론 그 원인은 앞에서 분석했던 것처럼 차원의 저주로 인한 것이다. 기존의 CVI들은 표 1의 t1과 t2는 가장 간단한 데이터 집합에 대해서도 실제 군집수를 찾는데 실패하였기 때문에 점점 더 복잡한 데이터 집합들에 해당하는 t3~t8 에 대한 실험 결과는 지면 관계상 기술하지 않았다.

표 1. 기존 CVI 들에 의해 예측된 군집 수와 그들의 정확성

	I	V_{SV}	XB	DB
t1	2 (X)	8 (X)	2 (X)	2 (X)
t2	2 (X)	5 (O)	3 (X)	4 (X)

기존의 CVI 들과 달리 본 논문에서 제안된 새로운 CVI들은 t1과 t2에 대해 완벽한 결과를 보여주었다 (표 2 참조). 이것은 새로운 CVI들에 적용된 d_{D_i} 가 고차원의 데이터들 사이의 거리는 측정하는데 매우 효과적임을 암시한다.

표 2에서는 좀 더 복잡한 데이터집합에 대한 새로운 CVI들의 성능 평가도 보여주고 있다. 이 표에서 알 수 있듯이, I_{HD}가 가장 좋은 성능을 보여주고 있다. 그 다음으로 $V_{SV,HD}$가 좋은 성능을 보였고, XB_{HD}가 가장 나쁜 성능을 보여 주고 있다.

표 2. 제안된 CVI 들에 의해 예측된 군집 수와 그들의 정확성

	t1	t2	t3	t4	t5	t6	t7	t8
I_{HD}	5 (O)	5 (O)	6 (O)	6 (O)	8 (O)	8 (O)	6 (O)	6 (O)
$V_{SV,HD}$	5 (O)	5 (O)	6 (O)	6 (O)	8 (O)	8 (O)	5 (X)	6 (O)
XB_{HD}	5 (O)	5 (O)	7 (X)	5 (X)	4 (X)	2 (X)	5 (X)	5 (X)
DB_{HD}	5 (O)	5 (O)	6 (O)	6 (O)	7 (X)	7 (X)	5 (X)	5 (X)

XB_{HD}의 문제점을 분석하기 위해서는 그 설계 원리를 살펴볼 필요가 있다. XB_{HD}와 같은 군집 내부 거리 ($dW = \sum_{k=1}^{nc}\sum_{j=1}^{N} u_{kj}^2 d_{D_i}(x_j, c_k)^2 / N$)와 군집 사이 거리 ($dB = \min_{i,j} d(c_i, c_j)^2$) 의 비의 형태로 정의된 CVI는 dW와 dB가 각각 $nc^{optimal}$와 $nc^{optimal}$+1 에서 급격한 값의 증가가 있을 때, $nc^{optimal}$에서 XB가 최소값을 가질 수 있다. 하지만, 실패한 경우의 결과들을 분석해 본 결과 dW가 $nc^{optimal}$에이 아닌 곳에서 급격한 값의 증가를 보여주고 있었다 (지면상 그래프 생략). 즉, 군집 내부 거리의 정의가 적절하지 못함이 나쁜 결과의 원인으로 작용했다고 할 수 있다.

5. Conclusions

본 논문에서는 고차원의 데이터 집합에서 발생하는 군집 유효성의 문제를 고찰해보고 그에 대한 해결책을 제시하였다. 기존의 군집 유효성 지수 (CVI)에 대한 차원의 저주의 영향을 고찰하였고, 이러한 분석을 통해 고차원 응용에서의 군집 유효성을 위한 새로운 CVI들을 제안하였다. 실험에서 기존의 CVI들은 차원에 의해 완벽하게 분리되어 있는 간단한 고차원의 데이터 집합에 대해서도 PROCLUS의 군집 결과를 정확히 평가하는 데 모두 실패 하였다. 하지만, 본 논문에서 제안한 새로운 CVI들은 이들 데이터 집합의 실제 군집 수를 찾는데 성공하였다. 추가로 다양한 환경을 대표하는 좀 더 복잡한 데이터 집합들에 대해서 새로운 CVI들의 성능을 평가하였는데, 이 실험을 통해 XB_{HD}가 가장 나쁜 결과를 보여 주었고 I_{HD}가 가장 좋은 결과를 보여 주었다. 본 논문에서 제안한 CVI들은 PROCLUS뿐만 아니라 다양한 고차원 군집화 알고리즘들에 대해서도 그들의 군집 결과의 품질을 평가하는데 결정하는데 유용할 것으로 기대한다.

References

[1] C.C. Aggarwal, C. Procopiuc, J.L. Wolf, P.S. Yu, and J.S. Park, " Fast Algorithms for Projected Clustering," Proc. ACM SIGMOD, pp. 61-72, 1999.

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

[3] K.S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, " When Is 'Nearest Neighbor' Meaningful?," Proc. ICDT, pp.217-235, 1999.

[4] D.L. Davies and D.W. Bouldin, " A Cluster Separation Measure," IEEE Trans. PAMI, vol. 1, no. 2, pp. 224-227, 1979.

[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] U. Maulik and S. Bandyopadhyay, " Performance Evaluation of Some Clustering Algorithms and Validity Indices," IEEE Trans. PAMI, vol. 24, no. 12, pp. 1650-1654, 2002.

[7] X.L. Xie and G.A. Beni, " A Validity Measure for Fuzzy Clustering," IEEE Trans. PAMI, vol. 3, no. 8, pp. 841-846, 1991.