유효성 기반 군집화 알고리즘

김민호^{O} R.S. Ramakrishna
광주과학기술원 정보통신공학과
{mhkim^{O}, rsr}@kjist.ac.kr

Validation-based Clustering Algorithm

Minho Kim^{O} and R.S. Ramakrishna
Dept. of Info. and Comm., Kwangju Institute of Science and Technology (K-JIST)

요 약

본 논문에서는 군집화의 가장 중요한 2가지 문제에 대한 새로운 해결책을 제시한다. 첫 번째 문제는 두 객체가 하나의 군집내에 포함될 수 있는지를 결정하는 유사 결정으로써, 이를 해결하기 위해 군집 유효화 지수에 기반한 유사 결정 기법을 제안한다. 이 기법은 정성적인 인지 과정을 정량적인 비교 결정 과정으로 바꾼다. 이 기법은 본 논문에서 제안한 랜덤 군집화와 전체 군집화의 두 부분으로 구성된 유효성 기반 군집화 알고리즘의 핵심을 이루며, 기존의 많은 군집화 알고리즘에서 요구되는 복잡한 파라미터를 결정할 필요가 없어지도록 한다. 두 번째 문제는 최적 군집 수 (optimal number of clusters)를 찾는 것으로써, 이것 또한 앞에서 제안한 기법에 의해서 전체 군집화에서 찾을 수 있다. 마지막으로 제안한 기법과 군집화 알고리즘의 효용성 및 효율성을 보여주는 실험 결과가 제시 된다.

1. 서 론

군집화 (clustering)에서 핵심적인 문제는 유사 결정이다. 유사 결정은 두 객체가 어느 정도 비슷할 때 이를 비슷하다고 할 수 있느냐를 결정하는 문제이다. 다른 말로 표현하면, 두 객체가 어느 정도 가까울 때 동일한 그룹, 즉 동일한 군집 (cluster)에 포함될 수 있느냐를 결정하는 것이다. 유사 결정의 가장 보편적인 형태는 문턱값 비교일 것이다 [1]. 문턱값 비교의 가장 간단한 형태로써 두 객체 사이의 거리와 사용자에 의해 정해진 문턱값과 비교해서 유사도를 결정하는 것이다. 많은 군집화 알고리즘들은 위와 유사한 방법을 이용하여 최적의 군집 구조를 찾는데 매우 중요한 역할을 하는 문턱값 또는 문턱값 연관 변수들을 사용하고 있다. 그런데 이들 문턱값 또는 문턱값 연관 변수들은 수 많은 시행 착오를 통해서 결정되거나, 복잡한 과정을 통해 결정되는 경우가 많았다. 그러므로, 이러한 과정 자체는 특히나 일반 사용자에게는 대단히 어려운 과정이 아닐 수 없다.

최근 분할형 (partitional) 또는 계층적 (hierarchical) 군집화 알고리즘에서 최적 군집 수를 결정하기 위한 군집 유효화 지수 (cluster validity index)에 대한 연구가 활발히 진행되고 있다 [4, 3, 5, 6]. 하지만, 군집 유효화 지수는 두 객체 자체에 대한 유사도를 결정하기 위한 것은 아니었다.

본 논문에서는 군집 유효화 지수와 두 객체에 대한 유사 결정과의 연관성에 대해 제기할 것이며, 이 연관성에 기반을 둔 유효성 기반 군집화 알고리즘을 제안한다. 또한, 유사 결정을 위해 새로운 군집 유효화 지수 *VI*를 제안한다. 제안된 알고리즘은 앞서 언급했던 것과 같은 유사 결정을 위해 필요한 복잡한 파라미터 결정이 필요하지 않을 뿐만 아니라, 최적 군집 수 또한 결정할 수 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 유사 결정에 대해 논의하고, 3 장에서는 제안된 유효성 기반 군집화 알고리즘을 기술할 것이다. 실험 결과와 결론은 각각 4 장과 5 장에서 제시한다.

2 유사성 결정

두 객체 사이의 유사도는 어떻게 결정할 수 있을까? 이 문제에 대한 해결책을 그림 1에 주어진 예제를 보며 생각해 보자.

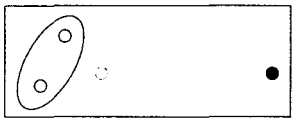

그림 1. 두 흰색 객체에 대한 유사도 인지 과정

위의 예에서 우리가 군집을 인지할 때 중요하게 작용하는 요소는 바로 상대성 (상대적인 거리)이다. 다시 말해, 어떠한 두 객체 (두 흰색 객체)가 또 다른 객체 (검정색 객체)에 비해 상대적으로 비슷한 패턴을 가질 (가깝게 위치할) 때 두 객체가 동일한 군집에 속한다고 결정할 수 있는 것이다.

두 객체를 그룹화할 때 개념적으로 어떠한 인지의 변화가 일어나게 되는 걸까? 우선 두 객체를 그룹화함으로써 새로 생성된 군집의 intra-cluster distance - 군집 자체의 compactness, 예를 들어, 동일한 군집에 속하는 객체 사이의 거리의 평균 - 는 이전의 두 객체가 독립적으로 있을 때의 intra-cluster distance보다 큰 값을 가지게 된다. 즉, compactness를 희생해야 한다. 하지만, 두 객체를 하나로 보면 되기 때문에 객체 사이의 구분 - 각 그룹 (군집) 사이의 separability - 이 용이해지게 된다. 즉, 시스템 전체의 separability가 증가하게 된다. 여기에서 주목해야 할 사실은 이러한 인지의 변화의 두 정성적인 요소가 군집 유효화 지수의 이론적 근거와 동일하다는 것이다. 그리고, 군집 유효화 지수는 정량적으로 나타낼 수 있다는 것이다. 그래서, 위의 같은

인지의 과정을 군집 유효화 지수를 이용하여 아래와 같이 정의할 수 있다.

Definition. 두 객체를 그룹화할 때, 병합 후의 군집 유효화 지수가 이전의 군집 유효화 지수보다 더 최적의 값을 가지면, 두 객체는 *유사*하며, 동일한 그룹에 포함될 수 있다.

위의 정의에 의해 두 객체 사이의 유사 결정이 군집 유효화 지수값을 비교함으로써 가능하게 되었으며, 이 정의는 3 장에서 제시될 유효성 기반 군집화 알고리즘의 주축을 이룬다.

3 유효성 기반 군집화 알고리즘

군집 유효화 지수를 이용하여 유사 결정을 하기 위해서는 병합될 두 객체 (또는 군집)과 비교 대상이 되어줄 최소한 하나 이상의 군집이 필요하다. 비교를 할 군집들을 선택할 때, 효율적인 처리를 위해 알고리즘은 크게 두 개의 부분으로 나누어진다. 첫 번째 부분은 최소한의 군집들만을 이용하는 랜덤 군집화이고, 두 번째 부분은 전체 군집들을 이용하는 전체 군집화이다.

3.1 랜덤 군집화

군집 유효화 지수를 이용하는 현존하는 모든 알고리즘들은 지수를 계산할 때, 각 단계에 존재하는 모든 군집을 그 대상으로 한다. 그 때문에 각 단계에서의 군집의 수가 많으면 많을수록 계산 복잡도 (computational complexity)가 높아진다. 그런데, 2장에서 제안된 유사 결정을 위해 필요한 군집의 수는 병합될 가능성이 있는 두 개의 군집과 비교 대상이 되는 하나 이상의 군집이 요구 된다. 다시 말해서, 단지 2개의 군집에 대한 유사 결정을 위해 현재 존재하는 모든 군집들을 대상으로 군집 유효화 지수를 계산할 필요는 없다는 것이다.

따라서, 본 논문에서는 전체 군집 집합으로부터 랜덤 하게 선택된 |CP|개의 군집들의 집합인 clustering pool (CP)을 구성하여, 그 pool 내부에서 유사 결정을 시도하는 *랜덤 군집화*를 제안한다.

3.2 군집 유효화 지수

현존하는 많은 군집 유효화 지수들은 크게 두 가지로 나눌 수 있다. 한 부류는 intra-cluster distance 와 inter-cluster distance 의 ratio 로 된 지수들이고 [6], 나머지 한 부류는 summation 으로 된 지수들이다 [3, 4, 5]. 이들 중에서 ratio 형의 지수는 하나의 member 데이터로 구성된 군집이 거의 대부분인 랜덤 군집화의 초기 상태에서는 사용될 수 없다. 왜냐하면, 군집이 하나의 member 데이터로 구성되어 있을 경우 intra-cluster distance 가 0 의 값을 가지게 되고 지수는 0 또는 ∞의 값을 가지게 되어 비교가 불가능해지기 때문이다. 따라서, summation 형의 지수가 랜덤 군집화를 위해 이용될 수 있다.

전체 군집화에서는 랜덤 군집화와 달리 ratio 형의 지수가 사용될 수 있는데, 본 논문에서는 ratio 형의 새로운 군집 유효화 지수, *VI*를 수식 (1)과 같이 제안한다.

$$VI(nc) = \frac{1}{nc}\sum_{i=1}^{nc}\left(\frac{\max_{k=1\ldots nc, k\neq i}\{S_i + S_k\}}{\min_{l=1\ldots nc, l\neq i}\{d_{i,l}\}}\right) \qquad (1)$$

$$S_i = \frac{1}{n_i}\sum_{x\in X_i}\|c_i - x\|,\ d_{i,l} = \|c_i - c_l\|$$

제안된 VI 는 크게 세 가지의 특징으로 이해할 수 있다. 첫 번째 특징은 $\min_{l=1\ldots nc, l\neq i}\{d_{i,l}\}$ 와 관련된 부분으로써, 이 값에 의해 병합될 필요가 있는 군집들이 아직 존재함을 지시할 수 있다. 두 번째 특징은 $\max_{k=1\ldots nc, k\neq i}\{S_i + S_k\}$ 에 의해 불필요한 병합이 일어났음 나타낼 수 있다는 것이고, 마지막으로 각 군집에서 얻은 값들을 평균하는 부분은 전체 정보를 결합하게 함으로써 강건성 (robustness)를 제공할 수 있게 된다. 4 장의 실험 결과에서 제안된 지수 VI의 효용성을 확인할 수 있다.

3.3 전체 군집화

Clustering pool을 이용한 랜덤 군집화는 두 가지 문제점을 가지고 있다. 첫 번째는 군집 유효화 지수를 계산할 때 전체 군집 집합을 이용하지 않는다는 것이다. 이것은 랜덤 군집화에서 잘못된 결정을 할 수도 있고, 이로 인해서 최적 *nc*를 찾지 못할 수도 있다. 그러므로, 이를 해결하기 위해서는 전체 군집 집합을 clustering pool로 이용해야 할 것이다. 랜덤 군집화의 또 다른 문제점은 지역 최소값 (local minimum)에 빠질 수 있는 가능성을 가지고 있다는 것이다. 이 문제점은 유사 결정과는 관계 없이 $nc = 1$이 될 때까지 강제적인 병합을 한 후에 지수가 최적의 값을 가지는 군집 구조를 찾음으로써 해결할 수 있다. 위의 2가지를 해결하기 위해 제안된 알고리즘이 유효성 기반 군집화 알고리즘의 두 번째 부분인 *전체 군집화*다.

4 Experimental Results

제안된 알고리즘의 효용성을 확인하기 위해 5 개의 합성 데이터집합과 1 개의 실세계 데이터집합을 사용하였으며, 5 개의 합성 데이터집합은 그림 2 에 제시 되었다. 그림 2 에서 dataset 4 는 13 개의 군집으로 구성되어 있으며, dataset 5 는 17 개의 군집으로 구성되어 있다. 실세계 데이터집합은 잘 알려진 Iris 데이터집합 [2]을 사용하였다. Iris 데이터집합은 4D 데이터로써 2 개의 군집으로 그룹화 되는 것으로 알려져 있다.

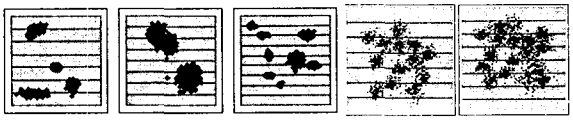

그림 2. 합성 데이터집합 (좌측에서부터 dataset 1 ~ 5)

먼저, 런덤 군집화의 효용성을 살펴보기 위해, 각 *nc*에 대해 두 군집을 병합하는데 성공할 때까지의 평균 유사 결정 test 횟수 (*n1*)를 측정해 보았다. 만약 응집형 계층적 군집화 (agglomerative hierarchical clustering)처럼 전역 최소 거리 (d_{min})을 가지는 두 군집을 찾으려면 각 *nc*에 대해 $nc\cdot(nc-1)/2$의 거리계산이 필요하다. 따라서, 랜덤 군집화와 동일한 범위인 $N \le nc \le \sqrt{N}$ 에 대한 총 계산량은 $n2 = (N+1)\cdot N\cdot(N-1)/6 - (\sqrt{N}+1)\cdot\sqrt{N}\cdot(\sqrt{N}-1)/6$이다. 이에 비해 clustering pool은 매번 테스트마다 $n3 = |CP|\cdot(|CP|-1)/2 + (|CP|-1)\cdot(|CP|-2)/2 = (|CP|-1)^2$의 거리계산이 필요하다. 그러므로 랜덤 군집화와 전역 d_{min}을 이용할 때의 군집화와의 효율성을 공정하게 비교하려면 *n1*과 *n2*/*n3* = *n4*를 비교 해주면 된다. 표 1에서 그 비교를 보여주고 있다. 랜덤 군집화에서 $|CP| \ge 30$이면 가능하다. 하지만, 본 논문의 실험에서는 $|CP| = 15$를 사용한다.

Table 1. 랜덤 군집화에서 다양한 데이터집합을 이용한 v_{sv} 와 SD 에 대한 거리계산 횟수 *n1*, *n4* 에 대한 비교

Data	index	*n1*	*N*	*n4*
Dataset 1	v_{sv}	28.215	500	106,283
	SD	1.033		
Dataset 2	v_{sv}	69.673	800	435,354
	SD	1.073		
Dataset 3	v_{sv}	34.663	550	141,464
	SD	1.000		
Dataset 4	v_{sv}	33.047	1300	1,868,156
	SD	1.004		
Dataset 5	v_{sv}	30.255	1300	1,868,156
	SD	1.014		
Real data	v_{sv}	13.225	150	562,189
	SD	1.040		

위의 결과에서 전역 d_{min} 을 사용할 때의 계산량인 *n4*에 비해 (랜덤) clustering pool을 이용할 때의 계산량인 *n1*의 값이 대단히 작은 값을 가짐을 알 수 있다. 즉, 랜덤 군집화가 전체 군집을 대상으로 하는 군집화 (예: 응집형 계층적 군집화)에 비해 매우 효율적임을 알 수 있다. 표 1의 결과를 보면 SD [3]가 랜덤 군집화를 위해 v_{sv} [4]보다 적은 횟수의 테스트가 필요했음을 알 수 있다. 지수 v_{sv}와 SD가 각각 평균 0.026과 0.024의 거의 동일한 군집화 오류율을 가진다는 사실을 감안해 보면 SD가 랜덤 군집화에서 더 효율적인 지수임을 알 수 있다. 여기에서 군집화 오류율은 각 군집에서 다수 (majority) class와 틀린 class를 가진 데이터의 비율을 의미한다. 그리고, 이 오류율은 유효성 기반 군집화 알고리즘의 끝에 가면 전체 정보를 사용해서 정제를 (refinement)를 하기 때문에 줄어들거나 없어지게 된다.

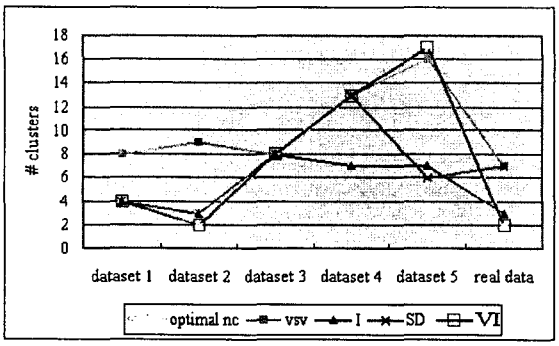

그림 3. 다양한 군집 유효화 지수를 이용하여 전체 군집화를 통해 찾은 군집 수 비교

다음으로 랜덤 군집화를 기반으로 전체 군집화가 얼마나 잘 최적 *nc*를 찾는지를 확인해 보자. 전체 군집화에서는 랜덤 군집화에서 사용되었던 지수들뿐만이 아니라 최근 제안된 지수 I [6]와 3.2절에서 제안한 지수 VI도 사용하였다. 전체 군집화는 랜덤 군집화에서처럼 summation 형의 지수를 사용해야 하는 제약이 없기 때문에 지수 I와 VI 같은 ratio 형의 지수도 적용될 수 있다. 본 실험에서는 앞의 실험 결과에서 SD가 랜덤 군집화에서 가장 좋은 결과를 보여주었기 때문에, 그 결과를 기반으로 하였다. 그림 3에서 각 지수에 대한 그래프와 최적 *nc*의 그래프와 비교함으로써, 각 지수의 성능을 확인할 수 있다. 그림 3의 결과를 살펴 보면 4개의 데이터집합에서 최적 *nc*와 틀린 값을 가진 v_{sv} 가 가장 나쁜 결과를 보여 주었음을 확인 할 수 있다. 그 다음으로 나쁜 결과는 3 개의 데이터집합에서 틀린 결과를 보여준 지수 I와 SD이다. 본 실험에서는 지수 VI가 최적 *nc*의 그래프와 정확히 일치한 그래프를 보여주었다. 따라서, 전체 군집화 에서 VI가 가장 좋은 결과를 보여주었다.

5 결론

본 논문에서는 군집 유효화 지수를 이용한 새로운 군집화 알고리즘인 유효성 기반 군집화 알고리즘을 제안하였고 그 효용성 및 효율성을 평가해 보았다. 본 논문에서 제안한 군집 유효화 지수를 기반으로 두 군집 사이의 유사도를 결정하는 방법은 유효성 기반 군집화의 주요 두 부분인 랜덤 군집화와 전체 군집화에 적용되어, 두 군집 사이의 유사도를 효과적으로 결정할 수 있었다. 다시 말해서, 두 군집이 병합되어 하나의 군집이 될 수 있는지를 정량적인 군집 유효화 지수 값의 변화로써 결정하였다. 또한 전체 군집화에서는 최적 군집 수도 찾을 수 있었다. 실험 결과에서 랜덤 군집화가 응집형 계층적 군집화보다 훨씬 더 적은 계산량으로 군집화를 하고 있음을 확인하였다. 유사 결정을 위해 여러 가지의 군집 유효화 지수가 평가되었는데, 그 중에서 랜덤 군집화를 위해서는 SD 가 가장 효율적이었으며, 전체 군집화에서는 VI 가 가장 정확하게 최적 군집 수를 찾았다. 향후 연구로써 현존하는 많은 군집 유효화 지수에 대한 이론적인 분석과 각 지수들과 많은 군집화 알고리즘 사이의 적합성 및 일반성에 대한 연구가 필요하다.

감사의 글

본 연구는 광주과학기술원 (K-JIST) 국제화 캠퍼스 사업단과 교육부 두뇌한국21(BK21) 정보기술사업단의 지원에 의한 것입니다.

참고문헌

[1] L. Kaufman and P.J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," John Wiley & Sons, New York, 1990.

[2] C.L. Blake and C.J. Merz, "UCI Repository of machine learning databases," http://www.ics.uci.edu/~mlearn/MLRepository.html, Univ. of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[3] 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.

[4] 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.

[5] M. Halkidi and M.Vazirgiannis, "Clustering Validity Assessment: Finding the Optimal Partitioning of a Dataset," Int'l Conf. Data Mining (ICDM) California, USA, Nov. 2001.

[6] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), vol. 24, no. 12, Dec. 2002.