최적의 군집을 찾기 위한 상대적 군집 평가 방법

김영옥[0] 이수원
숭실대학교 컴퓨터학과 데이터마이닝 연구실
smash@valentine.ssu.ac.kr[0], swlee@computing.ssu.ac.kr

Clustering Validity Assessment Using Relative Criteria for Finding Optimal Clusters

Young-Ok Kim[0] Soo-Won Lee
Dept. of Computing, Soongsil University

요 약

군집 분석은 데이터의 속성을 분석하여 서로 유사한 패턴을 가진 데이터를 묶는 방법이다. 군집 분석은 많은 응용 분야에서 쓰이고 있으나, 수행된 군집 분석 결과가 과연 정확한 결과이고 의미 있는 결과인지를 평가하는데 어려움이 있다. 본 논문에서는 군집이 형성된 데이터를 분석하여 군집 분석 결과를 평가하는 상대적 군집 평가 방법을 제안한다. 본 논문에서는 상대적 군집 평가 방법의 인덱스를 정의하고 형성된 군집 분석 결과에 적용해 최적의 군집, 의미 있는 군집을 찾을 수 있음을 보인다. 또한 실험을 통해 제안한 인덱스의 적합성을 보이며, 제안한 인덱스가 기존의 인덱스에 비해 최적의 군집, 의미 있는 군집을 더 잘 찾을 수 있음을 보인다.

1. 서론

군집 분석은 데이터의 속성을 분석하여 서로 유사한 패턴을 가진 데이터를 묶는 방법이다. 현재 많은 연구 분야에서 군집 분석에 대한 연구가 진행되고 있으며, 데이터마이닝, 인공 지능, 정보 검색, 패턴 인식 등 여러 분야에서 응용되고 있다.

군집 분석을 위한 알고리즘은 응용 분야, 목적, 데이터 용량에 따라 다양한 알고리즘이 있으며 각각의 알고리즘 별로 다양한 입력 변수와 매개 변수를 가진다[1][2][3]. 따라서, 같은 데이터에 대해 알고리즘 별로 상이한 결과가 나올 수 있고, 같은 알고리즘을 사용해도 서로 다른 입력 변수에 의해 상이한 결과가 나올 수 있다. 이와 같은 점 때문에 형성된 군집이 의미 있고 유사한 데이터로 그룹화 되었는지, 형성된 군집 중 어느 군집이 가장 최적의 군집이며 실제 데이터 군집과 유사한지를 평가하는 것이 중요하다.

군집 평가 방법은 군집 분석 결과를 입력으로 받아 정량화 된 인덱스 값으로 만들고 그 인덱스 값을 기준으로 형성된 군집의 적합도(Cluster Validity)를 평가한다. 군집 평가 방법은 평가 기준에 따라 외부 군집 평가(External Criteria) 방법, 내부 군집 평가(Internal Criteria) 방법, 상대적 군집 평가(Relative Criteria) 방법으로 나뉘어진다[1][2][4]. 외부, 내부 군집 평가 방법을 위한 인덱스는 상대적 군집 평가 방법의 인덱스에 비해 계산 복잡도가 커 효율성이 떨어진다[1].

상대적 군집 평가 방법을 위한 인덱스에는 *Dunn* 인덱스, *S_Dbw* 인덱스 , *CD* 인덱스, *DB* 인덱스, *RS* 인덱스 , *SPR* 인덱스, *RMSSDT* 인덱스 등이 있다[1].

이중 *S_Dbw*(*Scatter, Density between Clusters*) 인덱스는 여러 논문에서 최적의 군집을 찾고 의미 있는 군집을 찾는다고 평가되고 있다[1][4]. 그러나 *S_Dbw* 인덱스는 군집 모형이 일정치 않을 때 군집 분석 결과에 대해 정확한 평가를 수행하는데 문제점이 있으며 세분화 된 군집에 더 좋은 평가를 해 실제 데이터 군집을 찾는데도 어려움이 있다.

본 논문에서는 *S_Dbw* 인덱스를 수정한 *S_Dbw** 인덱스를 제안한다. *S_Dbw** 인덱스는 *S_Dbw* 인덱스의 문제점을 보완한 인덱스로 군집 분석 결과에 대한 정확한 평가를 수행한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 군집 평가 방법에 대해 설명한다. 3장에서는 *S_Dbw* 인덱스의 문제점을 제시하고 그 문제점을 보완한 *S_Dbw** 인덱스에 대해 설명한다. 4장에서는 실험을 통해 *S_Dbw* 인덱스와 *S_Dbw** 인덱스를 비교 분석하고, 5장에서는 결론을 내린다.

2. 관련 연구

2.1 군집 평가 방법

군집 평가 방법은 평가 기준에 의해 세 가지로 접근할 수 있다. 세 가지 군집 평가 방법은 첫째, 군집 분석의 대상이 되는 데이터가 사람의 직관 또는 외부 정보를 이용해 묶일 수 있다는 가정에 기반을 두고 형성된 군집의 복원도를 측정하는 외부 군집 평가 방법, 둘째, 데이터 자체의 유사도 벡터를 이용한 내부 군집 평가 방법, 마지막으로 같은 군집 분석 알고리즘을 이용해 입력 변수들을 달리하면서 형성된 군집 분석 결과를 비교 평가해 최적의 군집을 찾는 상대적 군집 평가 방법이다[1][4].

각각의 군집 평가 방법은 일반적으로 다음의 두 가지 척도를 측정하여 군집의 적합성을 평가한다.

i) Inter-Cluster Similarity
: 서로 다른 군집간의 결합도를 평가
ii) Intra-Cluster Similarity
: 각각의 군집 내에 소속 된 데이터의 밀집도를 평가

2.2 S_Dbw 인덱스

S_Dbw 인덱스는 상대적 군집 평가 방법을 위한 인덱스이다. *S_Dbw* 인덱스에서는 군집간의 밀집도를 정의하여 Inter-Cluster Similarity를 평가하고, 군집 내의 분산도를 정의하여 Intra-Cluster Similarity를 평가한 다음 두 척도의 합을 인덱스 값으로 하여 최적의 군집을 찾고 형성된 군집의 적합성을 평가한다. *S_Dbw* 인덱스의 정의는 다음과 같다.

전체 데이터 집합을 S, 군집의 수를 c, S의 전체 데이터 수를 n이라 할 때, 데이터 m의 이웃 데이터를 찾는 $density(m)$은 다음과 같다.

$$density(m) = \sum_{i=1}^{n} f(x_i, m) \qquad 식(1)$$

$f(x,m)$은 데이터 x가 데이터 m의 이웃 데이터인지를 판별하는 역할을 한다. 데이터 m의 이웃 데이터는 중심이 m이고 반지름이 $stdev$인 원안에 포함된 데이터를 말한다. $f(x,m)$의 정의는 다음과 같다.

$$f(x,m) = \{0 : d(x,m) > stdev, 1 : otherwise\} \qquad 식(2)$$

$stdev$는 각 군집의 중심 벡터에 대한 평균 표준 편차를 말한다. 군집 c_i 의 중심 벡터를 v_i 라 할 때, $stdev$는 다음과 같다.

$$stdev = \frac{1}{c}\sqrt{\sum_{i=1}^{c} \|\sigma(v_i)\|} \qquad 식(3)$$

여기서 $\|x\|$는 $\|x\| = (x^T x)^{1/2}$ 이고 x 는 벡터이다.

Inter-Cluster Similarity를 평가하는 $Dens_bw(c)$는 다음과 같다.

$$Dens\_bw(c) = \frac{1}{c(c-1)} \sum_{i=1}^{c} \left[\sum_{\substack{j=1 \\ i \neq j}}^{c} \frac{density(m_{ij})}{\max\{density(v_i), density(v_j)\}} \right] \qquad 식(4)$$

v_i, v_j 는 군집 c_i, c_j 의 중심 벡터이고, m_{ij} 는 v_i, v_j 의 연결선의 중심 벡터이다.

m_{ij}에 대한 $density(m_{ij})$는 군집 c_i, c_j의 데이터 집합 ($x_l \in c_i \cup c_j \subseteq S$)을 대상으로 이웃 데이터를 판별한다.

Intra-Cluster Similarity를 평가하는 $Scat(c)$는 다음과 같다.

$$Scat(c) = \frac{1}{c}\sum_{i=1}^{c} \|\sigma(v_i)^2\| \Big/ \|\sigma(S)^2\| \qquad 식(5)$$

$\sigma(S)^2$ 은 전체 데이터 집합 S의 분산을 의미하며, p차원에서는 다음과 같이 정의된다.

$$\sigma_x^{p^2} = \frac{1}{n}\sum_{k=1}^{n}\left(x_x^p - \bar{x}^p\right)^2 \qquad 식(6) \qquad \bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k, \forall x_k \in S \qquad 식(7)$$

$\bar{x}^p$ 는 식(7)의 p차원 값을 말한다. $\sigma(v_i)^2$은 군집 c_i 의 분산을 말한다. p차원에서 정의는 다음과 같다.

$$\sigma_{v_i}^{p^2} = \sum_{k=1}^{n_i}\left(x_k^p - v_i^p\right)^2 \Big/ n_i \qquad 식(8)$$

S_Dbw 인덱스의 정의는 다음과 같다.

$$S\_Dbw(c) = Dens\_bw(c) + Scat(c) \qquad 식(9)$$

$Dens_bw(c)$의 값이 작으면 군집간의 밀집도가 작다는 것이고, $Scat(c)$의 값이 작으면 군집내의 분산도가 작다는 것이므로, $S_Dbw(c)$의 값을 가장 작게 가지는 군집이 최적의 군집이라 할 수 있다[4].

3. S_Dbw* 인덱스

3.1 S_Dbw 인덱스의 문제점

S_Dbw 인덱스에서는 Inter-Cluster Similarity를 군집의 중심과 평균 표준 편차를 이용한 이웃 탐색에 의해 평가하고, Intra-Cluster Similarity를 분산을 이용해 평가한다. *S_Dbw* 인덱스에서는 Inter-Cluster Similarity 평가 시 군집 모형에 상관없이 평균 표준 편차를 이웃 탐색 범위로 적용해 원 또는 타원 모형이 아닌 군집에는 정확한 이웃 탐색을 할 수 없게 된다. 군집 모형이 일정치 않은 군집에는 적합하지 않은 문제점이 있다. 또한, Intra-Cluster Similarity 평가 시 군집에 소속된 데이터 수를 무시해 더 세분화 된 군집에 좋은 평가를 하게 되고 잡음을 인식하지 못하는 문제점이 있다. 따라서, 이와 같은 문제점을 해결하고자 수정된 인덱스를 제안한다.

3.2 S_Dbw* 인덱스의 정의

제안된 인덱스는 Inter-Cluster Similarity 평가의 이웃 탐색 시 각 변수의 신뢰 구간을 정의하여 탐색하고, Intra-Cluster Similarity 평가의 분산 측정 시 군집에 소속된 데이터 비율로 가중치를 준다.

$Dens_bw^*(c)$의 정의는 식(4)와 같고, $density(m)$은 식(1)과 같다. 이웃 탐색 시 이용되는 $f(x,m)$은 다음과 같다.

$$f(x,m) = \{1 : CI^p \le d(x^p, m^p) \le CI^p (1 \le p \le k), 0 : otherwise\} \qquad 식(10)$$

CI^p는 p차원에서의 신뢰 구간을 의미하며 95%의 신뢰 구간에서 CI^p는 아래와 같다.

$$CI^p = u^p \pm \left(1.96 \times \frac{\sigma^p}{\sqrt{n}}\right) \qquad 식(11)$$

u^p, σ^p, n 은 각각 p차원에서의 평균, 표준 편차, 해당 군집의 데이터 수를 말한다. 군집 c_i, c_j의 m_{ij}에서 u_{ij}^p, σ_{ij}^p, n_{ij}는 다음과 같다.

$$u_{ij}^p = \frac{u_i^p + u_j^p}{2}, \sigma_{ij}^p = \frac{\sigma_i^p + \sigma_j^p}{2}, n_{ij} = n_i + n_j \qquad 식(12)$$

$Scat^*(c)$는 다음과 같다.

$$Scat^*(c) = \frac{1}{c}\sum_{i=1}^{c} \|\sigma(v_i)^2\| \frac{n_s - n_i}{n_s} \Big/ \|\sigma(S)^2\| \qquad 식(13)$$

n_s는 전체 데이터 수이고 n_i는 군집 c_i의 데이터 수이다.

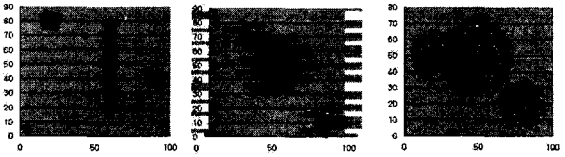
수정된 군집 평가 인덱스 $S_Dbw^*(c)$는 다음과 같다.

$$S\_Dbw^*(c) = Dens\_bw^*(c) + Scat^*(c) \qquad 식(14)$$

4. 실험

4.1 실험 데이터

실험 데이터는 좌표로 직접 확인 할 수 있는 2차원 가상 데이터를 사용하였다. 데이터 수는 1000개이며, 각 데이터의 분포는 그림[1], 그림[2], 그림[3]과 같다.



그림[1] 그림[2] 그림[3]

데이터1(그림[1])은 4개의 분명한 군집이 있는 데이터이고, 데이터2(그림[2])는 3개의 군집이 있다고 볼 수 있고, 2개의 군집이 붙어 있는 데이터이다. 데이터3(그림[3])은 크게 2개의 군집이 있으나 데이터가 전체적으로 퍼져 있는 데이터이다.

4.2 실험 평가 방법

본 연구에서는 실험 데이터에 대해 군집의 수를 조절하면서 군집 분석 알고리즘을 수행한다. 형성된 군집 데이터를 통해 *S_Dbw*, *S_Dbw** 인덱스 값을 얻는다. 2차원 좌표를 통해 실제 군집 분석 알고리즘이 형성한 군집을 확인하고 두 인덱스 값을 비교함으로써 최적의 군집, 의미 있는 군집을 찾은 인덱스를 평가한다. 사용된 군집 분석 알고리즘은 거리 기반의 K-Means 알고리즘과 확률 기반의 EM(Expectation & Maximization) 알고리즘이다. 군집의 수는 2개부터 10개로 한정하며 인덱스 값이 가장 낮은 군집 순으로 정렬하여 각 인덱스에서 가장 좋은 군집이 어떤 것인지 확인한다.

4.3 실험 결과

다음은 실험 데이터에 대해 군집의 수를 변화하면서 K-Means 및 EM 알고리즘를 수행하고 형성된 군집 데이터를 입력으로 *S_Dbw*, *S_Dbw** 인덱스 값을 얻어 가장 작은 인덱스 값을 얻은 군집 순으로 정렬한 결과이다.

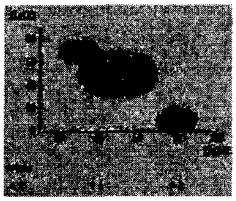
[표1] 최적의 군집 수

Data	Algorithm	Index	Number of Clusters								
			Optimal <--								Not Optimal -->
Data1	K-Means	*S_Dbw*	**10**	9	7	8	6	4	5	3	2
		*S_Dbw**	**4**	3	5	7	2	10	8	9	6
	EM	*S_Dbw*	**4**	5	8	6	7	9	10	3	2
		*S_Dbw**	**4**	5	6	8	3	2	10	7	9
Data2	K-Means	*S_Dbw*	**10**	3	9	5	6	2	4	8	7
		*S_Dbw**	**3**	9	2	4	8	7	10	6	5
	EM	*S_Dbw*	**10**	9	8	6	7	4	5	2	3
		*S_Dbw**	**3**	4	10	6	2	9	8	7	5
Data3	K-Means	*S_Dbw*	**10**	9	8	7	6	5	4	3	2
		*S_Dbw**	**5**	10	9	2	7	6	8	3	4
	EM	*S_Dbw*	**10**	9	8	6	7	5	4	3	2
		*S_Dbw**	**9**	2	10	8	5	7	3	6	4

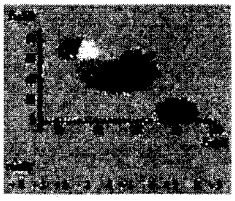
4.4 실험 결과 분석

실험 데이터1에 대해 *S_Dbw* 인덱스는 K-Means에서 10개, EM에서 4개의 군집을 최적의 군집으로 찾았다. 이에 비해, *S_Dbw** 인덱스는 K-Means와 EM 모두에서 4개의 군집이 최적의 군집임을 보여주고 있다. 실험 데이터2에 대해 *S_Dbw* 인덱스는 K-Means 및 EM에서 10개의 군집을 최적의 군집으로 찾았고, *S_Dbw** 인덱스는 K-Means와 EM 모두에서 3개의 군집이 최적임을 보여준다. 실험 데이터3에 대해 *S_Dbw* 인덱스는 K-Means 및 EM에서 10개의 군집을 최적의 군집으로 찾았고, *S_Dbw** 인덱스는 K-Means에서 5개, EM에서 9개의 군집을 최적의 군집으로 찾았다.

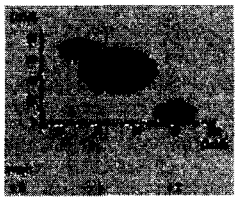
데이터1에 대해 *S_Dbw* 인덱스는 실제 데이터 군집 모형인 4개의 군집이 K-Means에서 여섯 번째 순위를 가진 것을 볼 수 있다. 또한 데이터2에 대해 *S_Dbw* 인덱스는 실제 데이터 군집인 3개의 군집을 K-Means에서는 두 번째, EM에서는 가장 적합하지 않은 군집으로 찾았다. *S_Dbw** 인덱스는 K-Means, EM 알고리즘에 상관없이 데이터1에 대해 4개, 데이터2에 대해 3개의 군집을 최적의 군집으로 찾아 극명한 대조를 보였다. 데이터3에 대해 *S_Dbw*, *S_Dbw** 인덱스가 비슷한 결과를 보였지만, EM에서 *S_Dbw* 인덱스는 두 번째 순위로 9개의 군집이 있는 것을 볼 수 있고 *S_Dbw** 인덱스는 2개의 군집이 있음을 볼 수 있다. 그림[4], 그림[5]는 데이터2에 K-Means를 적용한 군집 3개, 군집 10개의 결과이고, 그림[6],그림[7]은 EM을 적용한 군집 3개, 군집 10개의 결과이다.



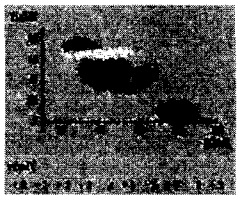
그림[4]



그림[5]



그림[6]



그림[7]

전체적인 결과에서 *S_Dbw* 인덱스는 더 세분화된 군집을 최적의 군집으로 평가하는 것을 확인 할 수 있었다. 따라서, 원 데이터와 유사한 실제 군집을 찾는 데는 좋은 성능을 보이지 못했다. 이에 비해 *S_Dbw** 인덱스는 실제 데이터 군집과 거의 유사한 군집을 최적의 군집으로 찾는 것을 볼 수 있었다.

S_Dbw 인덱스는 군집간의 밀집도 평가 시 나뉘어진 군집의 평균 표준 편차를 가지고 이웃을 탐색해 군집이 많이 나뉘어도 군집간의 밀집도가 크게 증가되지 않는 것을 볼 수 있었다. 또한, 세분화된 군집의 분산 값이 더 낮게 나오므로 세분화된 군집에 더 좋은 평가를 하게 되는 것을 알 수 있었다. 이에 비해 *S_Dbw** 인덱스는 이웃 탐색 시 각 변수의 신뢰 구간을 이용해 군집간의 밀집도에 좀 더 민감한 반응을 한다는 것을 알 수 있었고, 군집에 속한 데이터의 수에 가중치를 줌으로써 실제 데이터 군집과 유사한 군집을 찾을 수 있었다.

5. 결론

본 논문에서는 최적의 군집을 찾기 위한 상대적 군집 평가 방법의 *S_Dbw** 인덱스를 제안하였다. *S_Dbw* 인덱스의 *Dens_bw*와 *Scat*을 수정하여 실제 데이터 군집과 유사한 최적의 군집을 찾았고 좀 더 의미 있는 군집을 찾는 방법을 제시하였다. 실제 데이터 군집과 다른 군집 분석 결과는 분석가로 하여금 잘못된 분석을 하게 할 수 있다. 이와 같은 관점에서 볼 때 본 논문에서 제안한 *S_Dbw** 인덱스는 실제 군집과 유사한 군집을 최적의 군집으로 찾음으로써 군집 분석 알고리즘 평가에 많은 효용성을 가진다.

6. 참고 문헌

[1] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis "On Clustering Validation Techiniques", Intelligent Information Systems Journal. Vol17. No2-3. Pages 107-145, 2001

[2] A.K Jain, M.N.Murty, P.J.Flyn "Data Clustering : A Review", ACM Computing Surverys, Vol.31,No3, 1999

[3] D. Fasulo "An Analysis of Recent Work on Clustering Algorithms" Technical Report, 1999

[4] Maria Halkidi, Michalis Vazirgiannis "Clustering Validity Assessment : Finding the optimal partitioning of a data set", ICDM, 2001