퍼지 클러스터 타당성 척도를 이용한 최적 클러스터 수의 선택방법

이 현숙, 오 경환
서강대학교 전자계산학과 인공지능 연구실
서울시 마포구 신수동 1, 서강대학교(121-742)
Email : hsrhee@ailab6.sogang.ac.kr

A Selection Method of an Optimal Number of Clusters Using a Fuzzy Cluster Validity Measure

Hyun-Sook Rhee, Kyung-Whan Oh
A.I. Lab., Dept. of Computer Science, Sogang University,
Sinsudong 1, Mapoku, Seoul (121-742)
Email : hsrhee@ailab6.sogang.ac.kr

요약

클러스터의 타당성 정도를 계산하기 위한 측정자로서, 퍼지 분할된 데이타의 서로 다른 클래스 사이의 분리성과 한 클래스안에서의 밀접성의 비율, G를 정의하였다. 본 논문에서는 이렇게 정의된 G로 부터, 각 클러스터가 가지는 데이타 수의 차이점을 고려하여 하나의 데이타 집합에 대하여 서로 다른 분할들을 비교할 수 있도록 하기 위하여, I_G를 재정의하였다. 기존의 클러스터 타당성 전략은 클러스터 수의 함수로서, 주어진 척도의 값을 계산하여 기록한 후 그 값의 변화가 가장 큰 경우를 최적의 클러스터의 수로서 선택하였다. 이때 그 값의 변화를 고려하기 위한 주관적인 해석이 필요하게 된다. 본 논문에서는 주관적인 해석없이 I_G를 이용하여 최적의 클러스터 수를 결정하기 위한 방법을 제안하고자 한다. 제안된 방법은 널리 알려진 Iris data와 서로 다른 클러스터 인구수를 가지는 가상의 데이타 집합에 적용하여 그 타당성을 보인다.

서론

클러스터의 분석 과정은 데이타 군집사이의 유사성을 결정하므로 패턴인식이나 영상처리등의 연구분야에서 벡터 양자화나 데이타 분류를 위한 방법에 적용되어 왔다[1]. 특히 기존의 클러스터링 방법론은 주어진 데이타 군집사이의 관계가 명확하다는 가정에서 각 패턴을 하나의 클러스터에 소속시키는 방법이다. 그러나 우리가 다루는 대부분의 데이타는 그 경계가 불명확하므로 기존의 클러스터링 방법에 퍼지 집합이론[2]을 적용한 퍼지 클러스터링 알고리즘이 고안되어 널리 사용되어왔다[3].

이와 같은 클러스터링 방법론이 더욱 유용하게 되기 위해서는 클러스터링의 결과가 주어진 데이타의 구조를 얼마나 잘 반영하였는지를 측정하는 척도가 필요하다. 이와 같은 척도는 "클러스터 타당성(cluster validity) 문제"로 정의되어 연구되어 왔다. 퍼지 클러스터의 타당성 문제를 위하여 각 데이타 패턴이 얼마나 잘 분류되었는지를 수학적으로 계산하기 위한 함수들이 제안되었다[3,4]. 이러한 대부분의 방법은 클러스터링의 결과인 퍼지 c 분할 정보를 하나의 값으로 요약하는 방법으로 partition coefficient(F), classification entropy(H), proportion exponent(P) 등이 있다[2]. 그러나 이들이 산출하는 값은 데이타가 가지는 기하학적인 성질을 직접 반영하고 있지 않으며 c의 값이 증가함에 따라 클러스터링의 결과와 관계없이 단조 감소하는 값을 산출하는 단점을 가지고 있다. 또한 Xie[4]는 데이타 분포의 기하학적 특성을 고려한 함수 S를 고안하였다. S도 또한 c가 증가함에 따라 감소하는 경향을 나타내고 있으며, 분리성을 반영한 d^2_{min}은 최악의 경우를 측정한 것이다. PC와 S를 소개하면, 다음과 같다.

$$F(U;c) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^2}{n} \qquad (1)$$

$$S = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \|V_i - X_j\|^2}{n\, d^2_{min}} \quad (2)$$
$$, \text{where} \quad d^2_{min} = (\min \|V_i - V_j\|^2)$$

위의 정의로부터 PC와 S는 0과 1 사이의 값을 가짐을 알 수 있다. 또한 PC의 값은 1에 가까운 값일 수록 타당한 클러스터임을 나타내며, S는 0에 가까운 값일 수록 타당한 클러스터임을 나타낸다.

[1]에서는 mini-max 필터 개념[5]과 퍼지이론을 적용한 새로운 퍼지 클러스터 타당성 척도 G를 제안하였다. G는 퍼지 분할된 데이타의 서로 다른 클래스 사이의 분리성과 한 클래스안에서의 밀집성의 비율로서 정의되었다.

본 논문에서는 이렇게 정의된 G로 부터, 각 클러스터가 가지는 데이타 수의 차이점을 고려하여 하나의 데이타 집합에 대하여 서로 다른 분할들을 비교할 수 있도록 하기 위하여, I_G를 재정의하였다. 기존의 클러스터 타당성 전략은 클러스터 수의 함수로서, 주어진 척도의 값을 계산하여 기록한 후 그 값의 변화가 가장 큰 경우를 최적의 클러스터의 수로서 선택하였다. 이때 그 값의 변화를 고려하기 위한 주관적인 해석이 필요하게 된다. 본 논문에서는 주관적인 해석없이 I_G를 이용하여 최적의 클러스터 수를 결정하기 위한 방법을 제안하고자 한다. 제안된 방법은 Iris data와 서로 다른 클러스터 인구수를 가지는 가상의 데이타 집합에 적용하여 그 타당성을 보인다.

서로 다른 퍼지분할의 비교를 위한 클러스터 타당성 척도

n을 주어진 데이타 집합의 데이타 수라 하고, c는 정해진 클러스터의 수이며 각 클러스터는 $C_1, C_2, \ldots, C_c$로 나타낸다. 또한 $d^2(X, Y) = \| X - Y \|^2$ 이며 $\| \|$은 유클리드 놈을 나타낸다.

정의1: intraclass distance는 같은 클러스터 안에서의 임의의 두 데이타 사이의 거리로서 전체 데이타에 대한 평균은 다음의 식 (3)과 같이 정의된다. 이때 ω_1은 임의의 두 데이타가 모두 같은 클러스터 C_i에 속하는 소속정도를 나타낸다. 이와 같이 정의된 C는 퍼지 분할의 밀집성을 나타낸다.

$$C = \frac{2}{n(n-1)} \sum_{j_1=1}^{n-1} \sum_{j_2=j_1+1}^{n} d^2(X_{j_1}, X_{j_2})\, \omega_1 \quad (3)$$
$$\omega_1 = \min\{u_{ij_1}, u_{ij_2}\}$$

정의2: interclass distance는 서로 다른 클러스터에 속하는 임의의 두 데이타 사이의 거리로서 전체 데이타에 대한 평균은 다음의 식 (4)과 같이 정의된다. 이때 ω_2는 임의의 두 데이타가 각각 서로 다른 두 클러스터에 속하는 소속정도를 나타낸다. 이와 같이 정의된 D는 퍼지 분할의 분리성을 나타낸다.

$$D = \frac{1}{n^2} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} d^2(X_{j_1}, X_{j_2})\, \omega_2 \quad (4)$$

$$\omega_2 = \min\left\{ \max_{i_1} \{u_{i_1 j_1}\},\ \max_{i_2 \neq i_1} \{u_{i_2 j_2}\} \right\}$$

정의3: 퍼지 분할의 밀집성 C에 대한 분리성 D의 비율을 클러스터 타당성 척도 G로서 정의한다. 즉 G = D/C.

이렇게 정의된 G는 밀집성과 분리성의 비율로 나타내므로 데이타 분포와 클러스터 사이의 관계를 반영한 값을 산출하며 데이타 집합 사이의 상대적인 비교가 가능하게 하였다. G의 정의로 부터 G의 값이 클수록 클러스터 안에서의 밀집성과 클러스터 사이의 분리성이 더욱더 뚜렷한 경우임을 알 수 있다.

이때 <그림 1>에 주어진 데이타 집합과 같이 주어진 클러스터안의 데이타 수가 현저히 차이나는 경우를 생각해 보자. <그림 1>의 데이타 집합은 각각 81개, 9개의 데이타로 구성된 두개의 클러스터를 가지고 있다. Table I은 이 데이타 집합에 대하여 클러스터의 수 c를 2,3,4로 가정하고 퍼지 클러스터링한 결과에 대한 G의 측정값을 보여준다. 이때 c=2인 경우는 육안으로 보이는 것처럼 분류되며, c=3인 경우는 더 큰 클러스터가 둘로, c=4인 경우는 더 큰 클러스터가 셋으로 나누어지게된다. 이 경우 c=2인 경우의 측정치가 가장 클것으로 예상할 수 있다. 그러나 Table I의 결과는 우리의 예상과 다름을 알 수 있다. 그 이유는 (식 4)의 D의 정의에서 전체 데이타의 평균을 고려하기 때문에 클러스터사이의 인구수가 다른 경우 커다란 클러스터 안에 속하는 임의의 두 데이타는 서로 다른 클러스터에 속할 정도가 작아지기 때문이다. 즉 같은 상황에서 클러스터 사이의 인구수가 같은 경우 G의 값은 최대가 될 것이다.

이와 같은 고찰을 고려하여 클러스터안의 인구 수의 차이를 고하여 G를 수정한다면 하나의 데이타 안에서 서로 다른 퍼지 분할을 비교할 수 있을 것이다.

정의 4: $n_i = \sum_{j=1}^{n} u_{ij}$는 클러스터 Ci에 속하는 퍼지 수(fuzzy number)라 하고, 이들로 구성된 벡터 $N=(n_1, n_2, \ldots, n_c)$의 표준편차를 σ라 하고, σ_{max}는 $N'=(n, 0, \ldots, 0)$의 표준편차라고 하자. 이때 클러스터 인구수 사

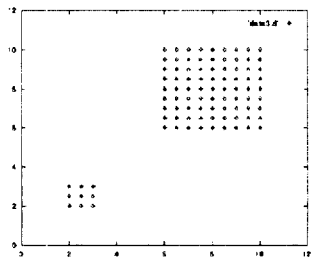
이의 유사성을 측정하기 위한 P는 다음과 같이 정의된다.

$$P = 1 - \frac{\sigma}{\sigma_{max}}$$

(정의 4)에서 정의된 P는 모든 클러스터가 같은 인구수를 가지는 경우 1이 되고, 모든 데이타가 특정 클러스터 한 곳에만 소속된 경우 0이된다. 이를 이용하여 인구수 사이의 차이를 고려한 클러스터 타당성 척도 I_G는 다음과 같이 정의된다.

$$I_G = \frac{G}{P}$$

Table I은 I_G의 측정치가 우리의 직관에 더욱 적합하다는 것을 확인할 수 있다.



<그림 1> 예제 데이타

<표 1> 예제 데이타에 대한 P, G, I_G의 측정값
(c = 2, 3, 4 인 경우)

	P	G	I_G
c = 2	0.32	1.95	6.32
c = 3	0.68	3.29	4.83
c = 4	0.82	4.70	5.76

I_G를 이용한 클러스터 타당성 전략

일반적으로 주어진 함수의 최소값이나 최대값을 가지는 경우보다는 그 값의 상대적인 변화가 큰 경우를 최적 클러스터의 수로서 선택한다. 그러므로 상대적인 변화를 측정하기 위한 방법으로 정의 5에서 R(c)를 정의한다. 이때 $I_G(i)$는 클러스터의 수 c=i인 가정에서 클러스터링 한 결과에 대한 I_G의 측정값이다.

정의 5 : π_1은 $I_G(i-1)$과 $I_G(i)$ 사이의 상대적인 변화의 양을 π_2는 $I_G(i)$와 $I_G(i+1)$ 사이의 상대적인 변화의 양을 측정하기 위하여 다음과 같이 정의한다.

$$\pi_1 = \frac{I_G(i) - I_G(i-1)}{I_G(i)}, \quad \pi_1 = \frac{I_G(i) - I_G(i+1)}{I_G(i)}$$

이때 R(c)는 과 의 합에 의하여 다음과 같이 정의된다.

$$R(c) = H(\alpha(\pi_1 + \pi_2))$$

이때 H는 비선형함수로서 임계치 이상의 값은 모두 1의 값을 내도록 하였고 $\alpha = \frac{2(c-1)}{c}$ 으로 정하여 1이상 2 미만의 값을 가지도록 하였다.

이렇게 정의된 R(c)는 c에서의 상대적인 측정값으로 0과 1사이의 값을 가지도록 하였다. 또한 절대적인 측정값을 반영하기 위하여 이를 0과 1사이로 만든 A(c)를 이용하여 c에서의 클러스터 타당성 정도는 다음과 같이 계산한다.

$$CV(c) = \frac{1}{2}\{R(c) + A(c)\}$$
$$, \text{where} A(c) = \frac{I_G(c)}{\max_{c \in 1}\{I_G(c)\}}$$

이때 CV(c)는 0과 1 사이의 값을 가지며 1에 가까울수록 타당한 클러스터를 형성한 것으로 가정할 수 있다. 그러므로 본 논문에서는 주어진 구간 I에서 최대의 CV(c) 값을 가지는 경우를 최적의 클러스터로 선택하는 타당성 전략을 세운다.

실험 및 고찰

본 절에서는 앞에서 제안한 타당성 전략을 이용하여 주어진 데이타의 최적 클러스터의 수를 찾는 예제를 보인다. 이를 위한 타당성 전략은 주어진 구간에서 클러스터의 수 c의 각각에 대하여 OFUNN으로 불리워진[6] 클러스터링 알고리즘에 적용한 결과를 이용한다. 클러스터 타당성을 계산하기 위한 CV(c)는 사용된 클러스터링 알고리즘과는 별도로 정의되었으므로 퍼지 c-means 알고리즘[3]등과 같은 안정된 다른 알고리즘을 사용하여도 무관하다. 마찬가지로 partition coefficient(PC)를 측정하여 최대값을 가지는 경우를, S를 측정하여 최소값을 가지는 경우를 최적의 클러스터의 수로서 선택하여 제안된 방법과 비교한다. 이를 위한 예제 데이타 집합으로 Iris data와 (그림 1)에 나타난 90개의 2차원 데이타를 준비한다.

Iris data는 클러스터링 알고리즘의 성능 비교 등을 위하여 널리 사용되어왔다. 4차원의 150개의 데이타로 3개의 클러스터를 형성하고 있으며 그 중 두개의 클러스터는 그 경계가 퍼지하며 나머지 하나는 분명하게 구별되는 것으로 알려져 있다. 그러므로 효과적인 클러스터 타당성 전략은 클러스터의 수가 3인 경우를 최적으로 선택할 것으로 기대할 수 있다. <표 2>은 c

의 값을 2 부터 6까지 고려하여 퍼지 클러스터링 한 결과에 대한 각 측정함수의 값을 보여준다. 이때 제안된 방법은 최적의 클러스터의 수를 3으 로, F와 S를 사용한 경우는 최적의 클러스터의 수를 2로 선택함을 알수 있다.

마찬가지 방법으로 <그림 2> 의 데이타에 대해서도 c의 값을 2부터 4까지 고려하여 같은 실험을 하였다. 그림에서 보여지는 것처럼 최적의 클러스터 수를 2로서 선택함이 타당함을 알 수 있다. <표3> 에서 보여지는 실험 결과는 세가지 방법 모두 클러스터의 수를 2로 선택함을 알 수 있다.

이상의 실험으로 부터 제안된 방법은 그 경계가 퍼지한 데이타 집합에 대하여 효과적인 결과를 낸다는 것을 알 수 있다.

<표 2> Iris 데이타에 대한 각 클러스터 수에 대한 타당성 척도 CV, F, S의 계산값

	CV	F	S
c=2	0.63	0.89*	0.05*
c=3	0.71*	0.78	0.14
c=4	0.54	0.69	0.61
c=5	0.67	0.63	0.40
c=6	0.64	0.58	0.74

<표 2> <그림 1>의 데이타에 대한 각 클러스터 수에 대한 타당성 척도 CV, F, S의 계산값

	CV	F	S
c=2	1.00*	0.95*	0.02*
c=3	0.38	0.74	0.24
c=4	0.77	0.70	0.60

결론

퍼지 클러스터 분석 알고리즘은 영상분할이나 벡터 양자화등의 분야에 적용되어왔다. 그러나 이와 같은 방법론이 적절히 적용되기 위해서는 클러스터링의 결과를 핑가하는 방법과 함께 연구되어야 한다. 퍼지클러스터링의 결과는 각 패턴에 대한 각 클러스터의 소속함수 값을 가지는 퍼지 c-분할 벡터에 의하여 대표된다. 그러므로 퍼지 c-분할 벡터가 가지는 퍼지 소속함수값을 하나의 값으로 요약하여 형성된 클러스터의 타당성 정도를 계산하였다.

본 논문에서는 클러스터의 타당성 정도를 계산하기 위한 측정자로서 퍼지 분할된 데이타의 서로 다른 클래스 사이의 분리성과 한 클래스안에서의 밀접성의 비율을 정의하였다. 이는 퍼지이론과 mini-max filter concept를 적용하여 정의되었다. 이러한 정의를 바탕으로 클러스터 인구수의 차이를 고려하여 클러스터 타당성 전략에 사용할 수 있도록 하는 측정자 I_G를 정의하였다. 또한 I_G의 상대적인 값과 절대적인 값을 모두 고려하여 클러스터의 타당성 정도를 계산하기 위한 방법을 제안하고 이를 이용한 타당성 전략을 세웠다.

제안된 타당성 전략을 이용하여 주어진 데이타의 최적 클러스터 수를 찾는데 이용됨을 보였다. 또한 그 결과를 기존의 측정자 F와 S를 사용한 경우와 비교하였다. 이때 주어진 데이타가 퍼지한 경계를 가지는 경우 제안된 방법이 더욱 효과적인 결과를 얻을 수 있음을 확인하였다..

참고문헌

[1] Duda, R. and Hart, P. *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[2] L. A. Zadeh, "Fuzzy Sets", *Information and Control* 8, 1965.

[3] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum press, New York, 1981.

[4] Xuanli Lisa Xie and Gerado Beni, "A Validity Measure for Fuzzy Clustering", *IEEE Trans. on Pattern Anal. Machine Intell.*, vol. PAMI-13, no.8, 1991.

[5] H. Szu, "Reconfigurable Neural Nets by Energy Convergency Learning Principle based on extended McCulloch-Pitts Neurons and Synapses, *Proceedings of International Joint Conference On Neural Networks*, June, 1989.

[6] H.S. Rhee and K.W. Oh, "A Design and Analysis of Objective Function Based Unsupervised Neural Networks for Fuzzy Clustering", *Neural Processing Letters*, vol. 4., no. 2, 1996.