# 퍼지 성능 측정자를 결합한 최적 클러스터 분석방법

# An Optimal Cluster Analysis Method with Fuzzy Performance Measures

이 현 숙*, 오 경 환**
Hyun Sook Rhee*, Kyung Whan Oh**

※본 논문은 1995년도 한국과학재단 핵심전문연구(951-0906-042-2) 지원에 의해 연구되었음.

요 약

클러스터링은 주어진 데이타 집합의 패턴을 비슷한 성질을 가지는 그룹으로 나누어 패턴 상호간의 관계를 정립하기 위한 방법론이다. 이러한 클러스터링 기법을 위하여 많은 알고리즘이 개발되었고, 패턴인식과 영상처리 등의 여러 공학 영역에 적용되어왔다. 대부분의 실세계 데이타는 그 경계가 명확하지 않으므로 그 특성을 보다 정확히 반영하기 위하여 퍼지이론이 도입되었다. 이와 같은 클러스터 분석 방법은 보다 적절히 응용하기 위하여 클러스터링의 적절성을 평가하기 위한 방법론과 함께 연구되어야 한다. 이를 위하여 각 데이타 패턴이 얼마나 잘 분류되었는지를 수학적으로 계산하기 위한 함수들이 제안되었다. 그러나 클러스터 타당성 문제는 주어진 클러스터링 방법론의 특성, 그 알고리즘에서 사용한 파라메터의 성질, 주어진 입력 데이타 집합의 특성 등 여러 복잡한 상황을 포함하고 있으므로 기존의 연구에서와 같이 하나의 함수를 이용하여 해결하기는 어렵다. 그러므로 본 논문에서는 기존에 연구되어온 타당성 측정 함수를 조사하고 그의 단점을 고찰하여 이를 해결하기 위한 방법으로 4가지 성능 측정자를 제안하고 이의 결합에 의하여 형성된 클러스터 타당성의 정도를 구하는 방법론을 제시하고자 한다. 또한 이러한 방법은 퍼지 클러스터링을 위한 학습 알고리즘과 결합하여 클러스터의 수나 데이타의 분포에 대한 정보없이 최적 클러스터를 찾아주는 방법에 응용될 수 있음을 보인다.

ABSTRACT

Cluster analysis is based on partitioning a collection of data points into a number of clusters, where the data points in side a cluster have a certain degree of similarity and it is a fundamental process of data analysis. So, it has been playing an important role in solving many problems in pattern recognition and image processing. For these many clustering algorithms depending on distance criteria have been developed and fuzzy set theory has been introduced to reflect the description of real data, where boundaries might be fuzzy. If fuzzy cluster analysis is to

*서강대학교 전자계산학과 박사과정
** Journal of Intelligent and Fuzzy Systems Associate Editor

make a significant contribution to engineering applications, much more attention must be paid to fundamental questions of cluster validity problem which is how well it has identified the structure that is present in the data. Several validity functionals such as partition coefficient, claasification entropy and proportion exponent, have been used for measuring validity mathematically. But the issue of cluster validity involves complex aspects, it is difficult to measure it with one measuring function as the conventional study. In this paper, we propose four performance indices and the way to measure the quality of clustering formed by given learning strategy.

## Ⅰ. 서 론

클러스터링은 주어진 데이타 집합의 패턴을 비슷한 성질을 가지는 그룹으로 나누어 패턴 상호간의 관계를 정립하기 위한 방법론이다[1]. 이러한 클러스터링 기법을 위하여 많은 알고리즘이 개발되었고, 패턴인식과 영상처리 등의 여러 공학 영역에 적용되어 왔다. 기존의 클러스터링 방법론은 대부분 hard partitioning에 의한 방법으로 주어진 데이타 상호간의 경계가 명확하다는 가정에서 각 패턴을 하나의 클래스에 소속시키는 방법이다. 그러나 이 모델은 우리가 다루는 데이타의 경계가 대부분 불명확하므로 실제 데이타 상호간의 군집성을 묘사하기에 부적절하며 주어진 데이타 분포의 성질을 잃어버리는 결과를 가져온다.

이를 개선하기 위하여 Bezdek은 Fuzzy c-Means (FCM) 알고리즘[2, 3]이라고 불리우는 퍼지분할에 의한 방법을 고안하였다. FCM 알고리즘은 최소자승 기준 함수(least square criterion function)에 퍼지 이론을 적용한 목적함수의 반복 최적화(iterative optimization)에 기반을 둔 방식이다. 이 알고리즘은 hard partitioning에 의한 기존의 클러스터링 방법이 승자 독점(winner take all)형태의 전략을 취하는데 비하여, 각 패턴이 특정 클러스터에 속하는 소속정도를 줌으로서 보다 정확한 정보를 형성하도록 도와준다. 이러한 FCM 알고리즘은 최적 퍼지 분할, 패턴분류와 영상 분할등의 여러 응용에 적용되어 유용한 결과를 얻었으며, 이 방법을 변형한 여러 알고리즘이 개발, 응용되었다[2, 4].

그러나 입력 데이타의 분포나 조밀도 등 데이타의 특성이 다양하므로 클러스터링 방법론이 더욱 유용하게 되기 위해서는 클러스터링의 결과가 주어진 데이타의 구조를 얼마나 잘 반영하였는지를 측정하는 척도가 필요하다. 이와 같은 척도를 "cluster validity problem"이라 정의하여 연구되어 왔다. 이러한 연구는 주어진 클러스터링 방법론의 타당성을 입증하는 방법이 되기도 하므로 비교사 학습방법을 채택하는 클러스터링 방법론의 연구에 중요한 부분이다. 특히 퍼지 클러스터링의 경우는 클러스터링의 결과를 평가하는 방법이 더욱 중요하며 이를 클러스터링 알고리즘에 접목시키면 클러스터의 수를 미리 정의하지 않아도 적절한 클러스터를 형성할 수 있으므로 필요한 연구분야이다.

퍼지 클러스터의 타당성(cluster validity)문제를 위하여 각 데이타 패턴이 얼마나 잘 분류되었는지를 수학적으로 계산하기 위한 함수들이 제안되었다[2, 5, 6]. 그러나 클러스터 타당성 문제는 주어진 클러스터링 방법론의 특성, 그 알고리즘에서 사용한 파라메터의 성질, 주어진 입력 데이타 집합의 특성 등 여러 복잡한 상황을 포함하고 있으므로 기존의 연구에서와 같이 하나의 함수를 이용하여 해결하기는 어렵다. 그러므로 본 논문에서는 기존에 연구되어온 타당성 측정 함수를 조사하고 그의 단점을 고찰하여 이를 해결하기 위한 방법으로 4가지 성능 측정자를 제안한다. 또한 이러한 방법은 퍼지 클러스터링을 위한 학습 알고리즘[7]과 결합하여 클러스터의 수나 데이타의 분포에 대한 정보없이 최적 클러스터를 찾아주는 방법에 응용될 수 있음을 보인다.

## Ⅱ. FCM 알고리즘과 타당성 측정 함수

### 2.1 FCM 알고리즘

퍼지 이론은 1965년 Zadeh에 의하여 처음으로 도입되었다[8]. 퍼지이론은 0이나 1중에 어느 하나만을 선택하는 이분법에서의 정보 손실을 막기 위하여, 0과 1사이의 값으로 소속정도를 표현하도록 하는 접근방법으로 주어진 데이타 구조에 대하여 더욱 정확한 표현방법을 제공해 준다.

Bezdek[2]은 이러한 퍼지개념을 데이타 클러스터링 방법론에 적용하여 식(1)과 같은 최소 자승 오류 함수(least square error functional) $J_m$의 반복 최적화(iterative optimization)에 기반을 둔 Fuzzy $c$-Means(FCM) 알고리즘을 개발하였다.

$$J_m = \sum_{j=1}^{n} \sum_{i=1}^{c} (u_{ij})^m (d_{ij})^2 \qquad (1)$$

여기서 $u_{ij}$는 주어진 입력 데이타 집합 $X=\{x_1, ..., x_n\}$에 대한 퍼지 c 분할을 $n \times c$의 벡터 U로 나타낼때 그의 한 요소로 데이타 $x_j$의 클러스터 i에 속하는 소속정도를 표현한다. 또한 $(d_{ij})^2 = \|x_j - v_i\|^2$이고 $\|\cdot\|$은 유클리드 놈을, $v_i$는 클러스터 i의 중심점을 나타내며 $m \in [1, \infty)$은 퍼지정도를 표시하는 파라메타를 나타낸다. 이때 Bezdek은 $m>1$인 경우 $J_m$의 국소적 최소점이 되기 위한 필요조건(충분조건은 아니지만)으로 다음의 식(2)와 식(3)를 유도하였다.

$$v_i = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j}{\sum_{j=1}^{n} (u_{ij})^m}, \qquad 1 \le i \le c \qquad (2)$$

$(d_{ij})^2 = \|x_j - v_i\|^2 > 0$이면 모든 클러스터 i에 대하여

$$u_{ij} = \frac{1}{\sum_{t=1}^{c} \left(\frac{d_{ij}}{d_{tj}}\right)^{2/(m-1)}} \qquad (3a)$$

로 정의하고, $(d_{ij})^2 = \|x_j - v_i\|^2 = 0$인 경우 다음의 조건을 만족하도록 모든 클러스터 i에 대하여 $u_{ij}$를 정의한다.

$$u_{ij} = 0 \ \text{ if } \ d_{ij}^2 \neq 0 \ \text{ and } \ \sum_{i \in I_k} u_{ij} = 1 \qquad (3b)$$

FCM 알고리즘은 단지 식(2)와 식(3)의 반복에 의하여 수렴점을 찾아가는 과정이다[2, 3].

위의 유도된 식(3)을 이용하여 입력데이타와 중심점 사이의 거리를 통한 퍼지 소속 함수값(fuzzy membership value)을 결정하게 된다. 식(3b)를 통하여 알 수 있는 바와 같이 임의의 한 클래스의 중심점과의 거리가 0인 입력 데이타에 대한 그 클래스안의 퍼지 소속함수값은 1이 될 것이다. 그리고 식(3)을 고찰하여 알 수 있는 바와 같이 그를 통하여 결정된 퍼지 소속 함수값은 형성된 각 클래스에 대하여 상대적인 값을 가지며 확률적인 제약을 준수하여 그 클래스에 대한 소속의 확률치나 공유의 정도로 해석된다. 그러나 퍼지 이론에서 이용하는 소속함수는 그 클래스에 대한 일치도나 전형성의 정도로서 해석되는 절대적인 값이므로 이를 보완하기 위한 연구도 진행되고 있다[9].

### 2.2 타당성 측정 함수

클러스터링을 특정 응용에 적용하기 위해서는 그 결과가 주어진 입력 데이타의 구조를 얼마나 잘 반영하고 있는가를 측정하는 척도가 필요하다. 이에 관련된 연구 영역을 "cluster validity problem"이라하며 이를 위하여 다음과 같이 클러스터 타당성 측정 함수를 정의하였다[2, 5, 6].

- Partition Coefficient(PC):

$$PC(U;c) = \frac{\sum_{j=1}^{n} \sum_{i=1}^{c} (u_{ij})^2}{n} \qquad (4)$$

- Classification Entropy(CE):

$$CE(U;c) = \frac{-\sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij} \log_a u_{ij}}{n} \qquad (5)$$

- Proportion Exponent(PE):

$$PE(U;c) = \log_e \left[ \prod_{j=1}^{n} \left\{ \sum_{l=1}^{\lceil u_l - 1 \rceil} (-1)^{j+1} \binom{c}{j} (1 - j u_l)^{c-1} \right\} \right]$$

$$\text{where } u_l = \bigvee_i u_{ij} \qquad (6)$$

위와같이 주어진 측정 함수들은 클러스터링의 결과 형성된 소속값안에 포함된 정보를 하나의 값으로 요약해 준다. Partition Coefficient PC(U;c)는 식(4)에 주어진 정의로 부터 다음과 같이 그 성질을 정의할 수 있다.

(i) $(1/c) \le PC(U;c) \le 1$

(ii) $PC(U;c) = 1$ iff fuzzy c-partition U is hard.

(iii) $PC(U;c) = 1/c$ iff $U = [1/c]$

위의 성질로 부터 PC(U;c)가 1에 가까운 값을 가질수록 데이타는 더 잘 분류된 것임을 나타냄을 알 수

있다. 그러므로 임의의 퍼지 클러스터링 알고리즘을 적용하고자할 때, 각 c의 값, 2, 3, ..., n−1에 대하여 입력 데이타 X의 최적 c 분할, $\Omega_c$를 구한 후 PC(U;c)의 값을 최대로 갖는 c를 찾아내는 방법으로 최적의 클러스터를 얻게된다.

Classification Entropy, CE(U;c)는 Shannon의 에너지에 대한 정보이론 개념으로 부터 정의된다. 식(6)에 주어진 정의로 부터 CE(U;c)는 다음과 같은 성질을 가짐을 알 수 있다.

(i) $(1/c) \le CE(U;c) \le \log_a(c)$

(ii) $CE(U;c) = 0$ iff fuzzy c-partition U is hard.

(iii) $CE(U;c) = \log_a(c)$ iff $U = [1/c]$

위의 성질로 부터 퍼지 c 분할, U의 각 요소가 모두 같은 값 1/c을 가지는 경우, 형성된 분할은 가장 불확실한 상태를 나타내고, 반대로 각 데이타가 어느 한 클래스에만 소속되게 되는 hard c-partition의 경우는 확실하게 데이타를 분류한 경우이다. 그러므로 클러스터의 타당성 관점에서 생각하면 CE(U;c)의 값이 적을수록 잘 분류되었음을 의미한다. 이를 임의의 퍼지 클러스터링 알고리즘을 적용하고자할 때, 각 c의 값, 2, 3, ..., n−1에 대하여 입력 데이타 X의 최적 c 분할 $\Omega_c$를 구한 후 CE(U;c)의 값을 최소로 갖는 c를 찾아내는 방법으로 최적의 클러스터를 얻게된다.

Proportion Exponent PE(U;c)는 퍼지 c 분할 U로 부터의 n개의 최대값만을 고려하여 타당성의 값을 구한다. Windham은 n개의 최대값이 각 입력데이타의 소속을 명확히 말해주기 때문에 그 n개의 값만을 고려하는 것도 의미 있다고 주장하고 있다[6]. 그는 또한 분석을 통하여 U의 각 요소가 1/c인 경우, PE(U;c)=0이며 PE(U;c)는 X의 hard c-partition에서는 정의되지 않음을 보였다. PE(U;c)를 임의의 퍼지 클러스터링 알고리즘에 적용하고자할 때, 각 c의 값, 2, 3, ..., n−1에 대하여 입력 데이타 X의 최적 c 분할, $\Omega_c$를 구한 후 PE(U;c)의 값을 최대로 갖는 c를 찾아내는 방법으로 최적의 클러스터를 얻게된다.

이미 기술한 바와 같이 PC(U;c)와 CE(U;c)는 U에 나타난 cn개의 값으로 부터 유도되지만 PE(U;c)는 U의 n개의 값으로 부터 유도되므로 PC(U;c)와 CE(U;c)는 PE(U;c) 보다 더 큰 도메인 위에서 고려되며 그런 의미에서 더욱 일반적인 것으로 알려져 있다. 또한 PC(U;c)와 CE(U;c)는 U안의 값들이 1에 가까운 값을 가짐에 따라 U의 값의 작은 변화에도 더욱 민감하게 반응하는 것처럼 보인다.

이러한 차이점에도 불구하고 위에서 기술한 3가지 타당성 측정 함수들은 그들이 산출하는 값이 주어진 데이타가 가지는 기하학적인 성질을 직접 반영하고 있지 않으며 c의 값이 증가함에 따라 클러스터링의 질과 관련없이 단조 감소하는 값을 유도해 내는 단점을 가지고 있다. 특히 이와같은 측정 함수를 클러스터링 알고리즘에 결합시킨 방법은 경험적 성질을 가지며 임의의 임계치를 설정해 주어야 한다. 타당성 측정을 위한 함수가 단조 감소하는 경향을 극복하기 위하여 그 함수에 정규화(normalization)나 통계적 표준화(statistical standardization)를 적용하는 방법이 제안되었다[2]. 또한 Gunderson은 주어진 데이타의 기하학적인 고려를 염두해 두고 separation coefficient를 타당성 측정 함수로서 제안하기도 하였다[10]. Separation coefficient는 기하학적으로 같은 클래스의 데이타는 군집해 있고 다른 클래스 사이의 데이타는 멀리 위치하고 있는지를 알아내기 위하여 고안되었으나 이 방법은 퍼지 클러스터링 알고리즘에 직접 적용할 수 없다. 우선 클러스터링의 결과를 hard한 것으로 변경한 후에 적용해야 하는데 그 변경과정은 여러 방법이 있을 수 있으므로 하나의 결과를 얻을 수 없는 단점을 가지고 있다. Separation coefficient는 최악의 경우를 고려하므로, 평균적인 상태를 고려하기 위한 방법으로 Xie[11]는 타당성 함수 S를 제안하였다. S는 퍼지 c 분할의 클래스안에서의 밀접성과 클래스 사이의 분리정도의 평균을 구하기 위한 측정 함수로서 FCM 알고리즘의 목적함수와도 밀접한 관계가 있음이 고찰되었다.

## Ⅲ. 제안된 퍼지 성능 측정자와 이를 결합한 최적 클러스터 분석 알고리즘

클러스터 타당성 문제를 고려하고자할 때 클러스터 형태의 다양성, 클러스터안의 데이타의 조밀도, 클러스터에 속하는 데이타의 수 등과 같은 복잡한 면을 고려해야 한다. 그러므로 2.2절에서 기술한 바와 같이 하나의 측정함수를 가지고 여러 복잡한 면을 고려한다는 것은 어려운 일이다. 이를 고려하여 본 논문에서는 타당성 측정를 위한 4가지 성능 측정자를 고안

하고 주어진 알고리즘에 의하여 형성된 클러스터의 적합성 정도를 측정하기 위하여 사용된다.

### 3.1 제안된 퍼지 성능 측정자

n은 주어진 데이타 집합의 데이타 수이며, c는 정해진 클러스터의 수이다. 이때 $u_{ij}$는 임의의 데이타 $x_j$가 클러스터 $C_i$에 속하는 퍼지 소속 함수값이다. 또한 $d^2(x, y) = \|x-y\|^2$이고 $\|\cdot\|$은 유클리드 놈을 나타낸다.

(1) Uniformness($I_u$)

$n_i$는 각 데이타를 가장 소속 함수값이 큰 클러스터에 대응시켰을때 클러스터 $C_i$에 속하는 데이타의 수이며 이들로 구성된 벡터를 $N=(n_1, n_2, ..., n_c)$이라 하자. 이때 $\sigma$는 벡터 N안의 값 사이의 표준편차이며 $\sigma_{max}$는 벡터 $N'=(n, 0, 0, ..., 0)$의 표준편차이다. 이때 각 클러스터안에 속하는 데이타 수의 유사성을 측정하기 위하여 다음의 식(7)을 정의한다.

$$I_u = (1 - \frac{\sigma}{\sigma_{max}}) \tag{7}$$

여기서 $\sigma$는 각 클러스터에 속하는 데이타의 수 사이의 표준 편차이며, $\sigma_{max}$는 데이타의 수 사이의 표준편차의 최대값을 나타낸다. 식(7)을 통하여 알 수 있듯이 가능한한 클러스터들이 비슷한 수의 데이타를 가질때 1에 가까운 값을 가진다. 즉 데이타의 분포가 고른 경우 1에 가까운 값이된다.

(2) Fuzziness($I_f$)

퍼지 클러스터링 알고리즘에서 산출한 퍼지 c 분할, U안의 값들이 가지는 퍼지한 정도를 측정해 주는 방법으로 식(8)과 같이 정의한다.

$$I_f = \frac{\sum_{j=1}^{n} [\vee_i u_{ij} - \wedge_i u_{ij}]}{n} \tag{8}$$

각 입력 데이타에 대하여 가장 큰 소속값과 가장 작은 소속값 사이의 차이가 1에 가까울 수록 그 데이타는 특정 클러스터에 정확히 속하는 것임을 나타내준다. 그러므로 $I_f$의 측정값은 클러스터 사이의 평균 중첩정도와 반비례하게 된다.

(3) 입력데이타의 표준편차에 대한 클러스터 중심점의 표준편차의 비율($I_r$)

입력 데이타의 표준 편차를 xsd라 하고 클러스터 중심점에 대한 *wsd*라 할때, $I_r$은 다음의 식(9)와 같이 정의된다.

$$I_r = 1 - \frac{|wsd - xsd|}{\max\{wsd, wsd\}} \tag{9}$$

클러스터 중심점은 특정 클러스터링 알고리즘에 의하여 얻어진 각 클러스터에 속하는 데이타의 대표값이다. 그러므로 $I_r$은 알고리즘에 의하여 찾아낸 클러스터의 중심점이 주어진 입력 데이타의 특성을 얼마나 잘 반영하고 있는지를 측정한다.

(4) 데이타 사이의 거리에 대한 intraclass distance의 비율($I_d$)

Intraclass distance는 하나의 클래스안에서의 데이타점들 사이의 거리로서 식(10)의 Idist로서 정의된다.

$$Idis = \frac{\sum_{i=1}^{c} \left\{ \frac{2}{n(n-1)} \left[ \sum_{j=1}^{n} \sum_{j'=1}^{n} d^2(x_j, x_{j'}) \times \min(u_{ij}, u_{ij'}) \right] \right\}}{c} \tag{10}$$

식(10)에서 $\min(u_{ij}, u_{ij'})$는 두 데이타 Xj와 Xj'이 모두 클러스터 i에 소속되는 정도를 나타낸다. 또한 데이타사이의 거리는 식(11)의 *dis*로 정의된다.

$$dis = \frac{2}{n(n-1)} \left[ \sum_{j=1}^{n} \sum_{j'=1}^{n} d^2(x_j, x_{j'}) \right] \tag{11}$$

이때 $I_d = Idis/dis$로서 정의되며 $I_d$의 값이 0에 가까울수록 타당한 최적 분할을 이루었음을 지시한다. 또한 $I_d$는 클러스터 사이의 분리 정도와 클러스터안에서의 결합정도를 측정해 준다.

### 3.2 최적 클러스터 분석 알고리즘

전체 알고리즘의 구조는 목적함수 $J_m$을 비교사 신경망에 학습시킨 학습방법[7]에 의하여 주어진 클러스터의 수 c에 대한 퍼지 소속정도와 중심점을 구한다. 이때 3.1절에서 제안한 4가지 성능 측정자를 이용하여 계산된 타당성 정도가 주어진 임계치를 넘지 않으면 c를 증가시켜 다시 클러스터링을 시도하는 과정

을 타당성 정도가 임계치를 넘을때 까지 반복한다. 여기서 사용된 퍼지 클러스터링 알고리즘의 학습 규칙은 식(1)의 목적함수 $J_m$에 대한 최급 하강법(gradient descent method)에 의하여 다음과 같이 유도된다.

$$\Delta v_{i,t} = -\eta_i \frac{\partial J_m}{\partial v_{i,t}}$$

$$= -\frac{m}{m-1} \sum_{j=1}^{n} \sum_{i=1}^{c} \left\{ (u_{ij})^{m+1} \left( \frac{d_{ij}}{d_{ij}} \right)^{\frac{m}{m-1}} \right\} (x_j - v_i)$$

이때

$$\alpha_{ij} = \frac{m}{m-1} \left\{ \sum_{i=1}^{c} (u_{ij})^{m+1} \left( \frac{d_{ij}}{d_{ij}} \right)^{\frac{m}{m-1}} \right\}$$

라 하면 중심점의 변화량 델타는 다음과 같이 유도된다.

$$\Delta v_{i,t} = \eta_i \sum_{j=1}^{n} \alpha_{ij} (x_j - v_{i,t})$$

이러한 학습규칙에 기반을 둔 최적 클러스터 분석 알고리즘의 절차를 요약하면 다음과 같다.

[단계1] init_c, ε, $\theta$의 값을 설정한다. 여기서 init_c는 최소의 클러스터 수를, ε은 수렴 임계치를, $\theta$는 형성된 클러스터의 타당성 테스트를 위한 임계치를 말한다. 클러스터의 수 c를 init_c의 값으로 초기화 gkrh,. m의 값을 정한다.

[단계2] 학습 횟수를 지시하는 t를 1으로 한다. 중심점 V의 초기치를 0과 1 사이의 난수로서 초기화 한다.

[단계3] 식(3)를 사용하여 퍼지 c 분할 U를 생성한다. 이때 식(3)은 각 입력 데이타의 c개의 클러스터 중심점 각각에 대한 0과 1사이의 소속값을 만들어낸다.

[단계4] 다음의 규칙을 이용하여 c개의 중심점을 변경한다.

$$v_{i,t+1} = v_{i,t} + \Delta v_{i,t}$$

$$= v_{i,t} + \eta_i \sum_{j=1}^{n} \alpha_{ij} (x_j - v_{i,t})$$

[단계5] $diff = \sum_{i=1}^{c} \| v_{i,t+1} - v_{i,t} \|^2$ 를 계산한다.

[단계6] $diff > \varepsilon$이면 t를 증가시키고, [단계3]으로 간다. 그렇지 않으면 [단계7]로 간다.

[단계7] 타당성 성능 측정자, $I_u$, $I_f$, $I_r$, $I_d$의 값을 계산하고 식(12)을 이용하여 클러스터 타당성의 정도, CV($I_u$, $I_f$, $I_r$, $I_d$)를 산출한다.

$$CV(I_u, I_f, I_r, I_d) = (I_u + I_f + I_r + (1 - I_d))/4 \qquad (12)$$

[단계8] CV($I_u$, $I_f$, $I_r$, $I_d$)$> \theta$이면 알고리즘을 끝낸다. 그렇지 않으면 c를 증가시키고 [단계2]로 간다.

위의 알고리즘에서 [단계2] 부터 [단계6]의 과정은 FCM 모델을 비교사 학습망(unsupervised learning network)으로 결합한 것이다. 그들은 특정한 클러스터의 수 c에 대하여 c개의 클러스터 중심점을 학습하고 퍼지 c 분할 U를 계산한다. [단계7]의 과정은 [단계2] 부터 [단계6]의 과정을 반복적으로 수행하여 수렴한 후의 학습결과를 성능 측정자, $I_u$, $I_f$, $I_r$, $I_d$를 이용하여 평가하는 단계이다. 식(12)은 클러스터 타당성, CV($I_u$, $I_f$, $I_r$, $I_d$)을 구하기 위하여 고안되었으며, 클러스터링의 여러면을 고려하는 각 측정값의 값을 평균한 것이다. CV($I_u$, $I_f$, $I_r$, $I_d$)의 값이 1에 가까울 수록 클러스터링이 잘 이루어짐을 나타낸다. 그러므로 제안된 알고리즘의 타당성 전략은 CV($I_u$, $I_f$, $I_r$, $I_d$)이 어떤 임계치 이상이 되며 최소의 클러스터의 수 c를 갖는 클러스터를 찾아내는 것이다.

## Ⅳ. 실험 및 결과

제안된 방법의 타당성을 보이기 위하여 4가지 테스트 데이타의 집합-butterfly data, Anderson의 Iris data, cubic data, quadratic data-을 준비하였다. 본 실험에서는 $\theta$는 0.85, init_c는 2로, m은 2로 정하였다,

### 4.1 Butterfly data

Butterfly data set은 두개의 클러스터를 가진 15개의 2차원 벡터로 구성되어있다. 이 데이타는 여러 실험에서 테스트용으로 사용되었으며 대칭적인 데이타 분포를 이루고 있다. 이 데이타를 가지고 실험한 결과는 Table 1에 요약되어 있다. c=2인 경우, CV($I_u$, $I_f$, $I_r$, $I_d$)의 값이 임계치를 넘으므로 이 경우를 최적의 클러스터로 결정한다. 이때 학습한 클러스터의 중심점은 (5.14, 2.00), (0.85, 2.00) 이다.

Table 1. 성능 측정자의 계산값과 클러스터 타당치(Butterfly Data)

| Number of clusters | $I_u$ | $I_f$ | $I_r$ | $I_d$ | CV($I_u$, $I_f$, $I_r$, $I_d$) |
|---|---|---|---|---|---|
| c=2 | 0.93 | 0.82 | 0.97 | 0.10 | 0.91 |

### 4.2 Iris data

Anderson의 Iris data set은 각 클래스는 iris plant의 한 형태를 나타내며, 각각 50개의 데이타로 구성된 세개의 클러스터로 구성되어 있다[2]. 그 중 하나의 클래스는 다른것들로 부터 선형분리 가능하고 나머지 두 클래스는 선형분리 가능하지 않다. 이 데이타 집합은 비교사 기법과 교사학습 분류 기법의 성능을 평가하기 위하여 여러 논문에서 사용되어 왔다. 이 데이타집합을 가지고 교사학습 분류 기법의 경우 보통 0-5개의 오류를 나타내며, 비교사학습기법에 대하여 15개정도의 오류를 나타내는 것으로 보고되었다. 이 데이타를 가지고 실험한 결과는 Table 2에 요약되어 있다. c=3인 경우를 최적의 클러스터로 결정하였고, 이때 학습된 클러스터의 중심점은 (6.74, 3.04, 5.61, 2.04), (5.86, 2.75, 4.32, 1.38), (6.74, 3.04, 5.61, 2.04)이다. 또한 잘못 분류한 경우는 평균적으로 14개정도이다.

Table 2. 성능 측정자의 계산값과 클러스터 타당치(Iris Data)

| Number of clusters | $I_u$ | $I_f$ | $I_r$ | $I_d$ | CV($I_u$, $I_f$, $I_r$, $I_d$) |
|---|---|---|---|---|---|
| c=2 | 0.71 | 0.87 | 0.90 | 0.15 | 0.83 |
| c=3 | 0.93 | 0.75 | 0.98 | 0.07 | 0.90 |

### 4.3 Cubic data

Cubic data는 4개의 클러스터로 구성된 120개의 3차원 벡터로 구성되어 있다. 각 클러스터의 데이타는 각각 (0~10, 0~10, 0~10), (10~20, 30~40, 10~20), (30~40, 10~20, 20~30), (30~40, 2~12, 30~40)으로 부터 임의로 30개씩을 선정하였다. 여기서 세번째 클러스터와 네번째 클러스터는 중첩된 부분을 가지고 있음을 알 수 있다. 이 데이타에 대한 실험 결과는 Table 3에 요약되어 있으며, c=4인 경우를 최적 클러스터로 결정해 준다. 이때 학습된 클러스터의 중심점은 (4.59, 4.14, 4.83), (15.75, 35.76, 16.37), (34.16, 14.88, 24.60), (33.93, 8.06, 35.28)이며, 잘못 분류한 데이타의 갯수는 평균적으로 3개이다.

Table 3. 성능 측정자의 계산값과 클러스터 타당치(Cubic Data)

| Number of clusters | $I_u$ | $I_f$ | $I_r$ | $I_d$ | CV($I_u$, $I_f$, $I_r$, $I_d$) |
|---|---|---|---|---|---|
| c=2 | 0.67 | 0.73 | 0.97 | 0.17 | 0.80 |
| c=3 | 0.65 | 0.70 | 0.99 | 0.11 | 0.81 |
| c=4 | 0.93 | 0.70 | 0.99 | 0.05 | 0.89 |

### 4.4 Quadratic data

Quadratic data는 5개의 클러스터로 구성된 15개의 2차원 벡터이다. 각 클러스터의 데이타는 각각 (0~10, 0~10), (10~20, 20~30), (10~20, −10~−20), (30~40, 0~10), (30~40, 10~20)으로 부터 임의로 30개씩을 선정하였다. 이때 네번째 클러스터와 다섯번째 클러스터는 인접해 있음을 알 수 있다. 이 데이타에 대한 결과는 Table 4에 요약되어 있으며, c=5인 경우를 최적 클러스터로 결정해 준다. 이때 학습된 클러스터의 중심은 (4.43, 4.72), (14.60, 25.04), (16.31, −16.99), (34.54, 4.85), (36.20, 15.55)이며, 잘못 분류한 데이타의 갯수는 평균적으로 2 이다.

Table 4. 성능 측정자의 계산값과 클러스터 타당치(Quadratic Data)

| Number of clusters | $I_u$ | $I_f$ | $I_r$ | $I_d$ | CV($I_u$, $I_f$, $I_r$, $I_d$) |
|---|---|---|---|---|---|
| c=2 | 0.75 | 0.70 | 0.90 | 0.21 | 0.79 |
| c=3 | 0.72 | 0.63 | 0.93 | 0.16 | 0.78 |
| c=4 | 0.65 | 0.75 | 0.99 | 0.13 | 0.82 |
| c=5 | 0.96 | 0.72 | 0.99 | 0.07 | 0.90 |

## V. 결 론

퍼지 클러스터링은 패턴 인식과 의사 결정 영역에서 많은 문제를 해결하는데 중요한 역할을 수행하는 분야이다. 그러나 퍼지 클러스터링 기법이 효과적으로 적용되기 위해서는 클러스터 타당성의 측정에 대한 연구와 병행되어야 한다. 본 논문에서는 클러스터링 문제의 복잡성을 고려하여 4가지 성능 측정자를 제안하고 이를 이용하여 타당성의 정도를 계산하였다. 또한 퍼지 클러스터링을 위한 학습 알고리즘과 결합하여 클러스터의 수나 데이타 분포에 대한 정보 없이 최적 클러스터를 찾아주는 방법에 적용될 수 있음을 보였다. 본 논문에서 제안한 최적 클러스터 분

석방법은 제안된 클러스터의 타당성 정도가 주어진 임계치 이상이 되는 최소의 클러스터 수를 가지는 분할을 찾는 것으로 요약할 수 있다.

이와 같은 클러스터 타당성의 문제는 여러 복잡한 면을 포함하므로 보다 많은 데이타에 적용하여 실험, 분석되어야 한다. 또한 제안된 타당성 측정자들의 성질은 면밀히 분석되어야 하며, c의 증가에 따른 측정값의 변화를 조사하여 좀 더 체계적인 기반을 마련해야 한다.

## 참 고 문 헌


1. Duda, R. and Hart, P. *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
2. J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum press, New York, 1981.
3. R. L. Cannon, J. V. Dave and J. C. Bezdek, "Efficient Implememtation of the Fuzzy c-Means Clustering Algorithms", *IEEE Trans. on Pattern Anal. Machine Intell.*, vol.PAMI-8, no.2, 1986.
4. T. L. Huntsberger, C. L. Jacobs and R. L. Cannon, "Iterative Fuzzy Image Segmentation", *Pattern Recognition*. 18, no.2, 1986.
5. M. P. Windham, "Cluster Validity for the Fuzzy c-Means Clustering Algorithm", *IEEE Trans. on Pattern Anal. Machine Intell.*, vol.PAMI-4, no.4, 1982.
6. Windham, M. P., "Cluster Validity for Fuzzy Clustering Algorithms", *J. Fuzzy Sets and Systems*, vol.3, 1980.
7. H.S. Rhee and K.W. Oh, "Unsupervised Learning Network Based on Gradient Descent Procedure of Fuzzy Objective Function", *Proceeding of International Conference on Neural networks*, 1996, Washington, DC.
8. L. A. Zadeh, "Fuzzy Sets", *Information and Control* 8, 1965.
9. R. Krishnapuram and J. M. Keller, "A Possibilistic Approach to Clustering", IEEE Transctions on Fuzzy Systems, vol.1, no.2, 1993.
10. R. Gunderson, "Application of Fuzzy ISODATA Algorithms to Star Tracker Pointing Systems", *Proc, 7th Triennial World IFAC Cong.* (Helsinki, Finland), 1978.
11. 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.



이 현 숙(Hyun-Sook Rhee)정회원

1989년:서강대학교 전자계산학과(학사)

1991년:포항공과 대학 전자계산학과(석사)

1991년~현재:한국전자통신(ETRI) 연구원

1993년~현재:서강대학교 전자계산학과 박사과정

※주관심분야:Fuzzy Cluster Analysis, Neural Network Modelling, Fuzzy-Neural Information System

오 경 환(Kyung-Whan Oh) 정회원

1978년:서강대학교 이공대학 수학과(학사)

1985년 Florida:State University 전산학과(석사)

1988년 Florida:State University 전산학과(박사)

1985년~1988년:Supercomputer Computations Research Institute 연구원

1992년~현재:Journal of Intelligent and Fuzzy Systems Associate Editor

※주관심분야:Fuzzy Expert System, Neural Networks, Computer Vision and Face Recognition