Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 2, pp. 301 ~ 309

Improvement on Fuzzy C-Means Using Principal Component Analysis

Hang-Suk Choi[1] · Kyung-Joon Cha[2]

Abstract

In this paper, we show the improved fuzzy c-means clustering method. To improve, we use the double clustering as principal component analysis from objects which is located on common region of more than two clusters. In addition we use the degree of membership (probability) of fuzzy c-means which is the advantage. From simulation result, we find some improvement of accuracy in data of the probability 0.7 exterior and interior of overlapped area.

Keywords : Clustering, Fuzzy C-Means, Principal Component Analysis

1. 서론

다양한 정보가 복합적으로 구성된 자료에서 유의미한 정보를 파악하기 위한 방법으로 통계적 기법(statistical method)과 기계 학습(machine learning) 중 신경망(neural network), 의사 결정 나무(decision tree), 연관성 규칙(association rule), 유전자 알고리즘(genetic algorithm), 그리고 본 연구에서 사용된 군집화 기법(clustering method) 등이 활용되고 있다.

이와 같은 방법 중 군집화(clustering)는 방대한 양의 자료에서 다양한 특성을 지닌 관찰대상의 유사성을 바탕으로 동질적인 집단으로 분류하는 기법이다. 이러한 군집화 기법이 Tryon(1939)에 의해 소개된 후 Tryon과 Bailey(1970), Jardine과 Sibson(1968), Anderberg(1973), Hartigan(1975), Jain과 Dubes(1988) 등에 의해 발전되었다. 그리고 현재 유전자 탐색, 질병 진단, 패턴인식, 영상처리 등의 의학, 자연과학, 공학 분야에 널리 적용되고 있다(Dembele and Kastner, 2003).

현재 알려진 군집화 방법은 K-means, fuzzy c-means, possibilistic c-means 등이

1) First Author : Ph.D student, Dept. of Mathematics, Hanyang Univ., 17 Haengdang-dong, Seongdong-Gu, Seoul, 133-791, Korea.
E-mail : neuldol@ihanyang.ac.kr
2) Professor, Dept. of Mathematics, Hanyang Univ., 17 Haengdang-dong, Seongdong-Gu, Seoul, 133-791, Korea.

있다(Krishnapuram and Nasraoui, 1995). 이 중 K-means는 주어진 군집(cluster) 간의 관계가 명확한 경우 분류가 정확한 방법으로, 만약 분류 대상인 군집의 경계가 명확하지 않을 경우 군집을 묘사하기에 부적절하며, 주어진 군집 분포의 손실을 초래할 수 있다(Pal and Bezdek, 1995).

이를 개선하기 위해 Bezdek(1980)는 각 군집에 소속된 정도(확률)를 이용하여 군집화하는 fuzzy c-means 알고리즘을 제안하였다. 그러나 fuzzy c-means 알고리즘은 자료로부터 각 군집에 대한 소속 정도(확률)의 합이 1이 되는 확률적 제약 조건(probabilistic constraint)을 이용하기 때문에 소속 함수 값이 직관적인 개념과 항상 일치하지는 않는다. 따라서 최근에 belief theory와 possibility theory 등이 이와 같은 문제점을 개선하기 위해 연구되고 있다.

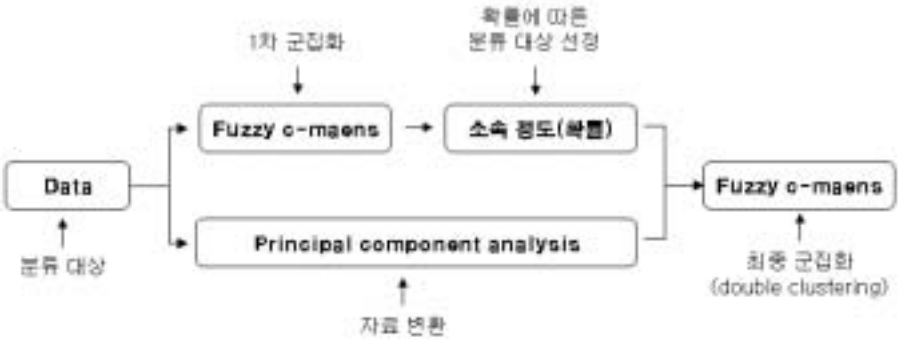
한편 fuzzy c-means 알고리즘은 각 객체(object)들이 군집의 중심에 소속하는 정도를 부여해 군집화 하는데, 중심 탐색에 있어 군집의 크기와 상관성이 다른 경우 분류에 문제점이 발생한다. 이러한 문제점을 해결하기 위해 본 연구에서는 다변량 분석 기법 중 상관성 제거와 변수의 차원 축소가 가능한 주성분 분석(principal component analysis)을 이용한 fuzzy c-means double clustering 방법을 제안하였다.

2. 연구 절차 및 분석 기법 소개

군집화는 군집의 수 혹은 군집의 구조에 대한 가정 없이, 오직 자료들 사이의 유사성을 기준으로 군집을 형성하고, 형성된 군집의 특성을 파악하여 군집 사이의 관계를 분석한다.

군집화는 자료들 간의 유사성의 정도를 측정하는 기준(measure)을 정의하여 군집화한다. 한편 군집화 알고리즘은 의미 없는 변수 제거 과정이 없으므로 선택된 변수들이 모두 동일한 가중치를 갖는다.

본 연구는 <그림 1>과 같은 절차에 따라 fuzzy c-means 알고리즘을 개선하였다. 우선 분류 대상의 자료를 이용하여 fuzzy c-means 알고리즘으로 1차 군집화 과정을 거친다. 1차 군집화 결과에서 도출된 소속 정도(확률)를 이용하여 오분류 가능성이 높은 자료를 재분류 대상으로 선정한다. 재분류에 선택된 자료는 주성분 분석 방법으로 변환 후 다시 fuzzy c-means 알고리즘으로 군집화 하는 double clustering 방법을 적용하여 최종 군집을 결정하였다. 한편 double clustering에 이용되는 재분류 자료 집합은 두 군집의 중심 간의 평균과 합동 분산(pooled variance)을 이용하여 오분류 가성이 높은 경계 지역의 분포를 추정한 자료이다.



<그림 1> double clustering 절차 및 과정

2.1 Fuzzy C-Means 알고리즘

Fuzzy c-means 알고리즘은 자료와 군집 중심과의 거리를 고려한 목적함수 (1)을 최소화 할 수 있도록 자료 집합을 분류한다.

$$J_m(u, c) = \sum_{k=1}^{K}\sum_{i=1}^{N}(u_{ki})^m d^2(x_i, c_k), \quad \sum_{k=1}^{K} u_{ki} = 1 \quad \text{for all } i = 1, \cdots, n, \tag{1}$$

여기서, u_{ki}는 객체 i가 군집 k에 속할 확률이며 K와 N은 자료 집합에서 군집의 수와 객체의 수이고, $m(\in[1,\infty])$은 퍼지정도(fuzziness)를 나타내는 변수로 일반적으로 $m=2$인 퍼지정도로 군집화 한다(Dembele and Kastner, 2003). $d^2(x_i, c_k)$는 군집 중심 c_k에서 객체 x_i의 거리이고, 자료로부터 각 군집에 대한 소속정도의 합이 1이 되는 확률적 제약조건(probabilistic constraint)이 주어진다.

$X=\{x_1, x_2, \cdots, x_n\}$인 자료 집합과 $C=\{c_1, c_2, \cdots, c_k\}$인 군집 중심들 사이의 소속정도를 $k\times n$인 행렬 U로 나타낸다. 이때, 행렬 U의 각 원소들은

$$u_{ki} = \frac{1}{\sum_{a=1}^{K}\left(\frac{d^2(x_i, c_k)}{d^2(x_i, c_a)}\right)}, \qquad c_k = \frac{\sum_{i=1}^{N} u_{ki}^m x_i}{\sum_{i=1}^{N} u_{ki}^m}, \tag{2}$$

로 나타내고, 여기서 m이 1보다 큰 경우에 모든 i, j에 대해서 $c_i \neq x_j$를 만족한다고 가정하면 위의 식 (2)을 만족할 때만 (u, c)가 J_m의 최소화를 가능하게 한다. Fuzzy c-means 알고리즘은 위 식의 과정을 반복하므로 J_m이 정해진 값에 수렴하게 된다.

이와 같은 fuzzy c-means 알고리즘의 전체 과정의 유도는 Bezdek(1980)가 소개하였으며, 다음과 같이 4단계로 진행된다.

Step 1. K, m, ε의 수렴 조건을 결정.
$\sum_{k=1}^{K} u_{ki}=1$ 인 소속 행렬(membership matrix) U를 임의로 초기화.
Step 2. 각 군집의 중심 좌표 c_k를 계산.
Step 3. 객체 i가 군집 k에 속할 확률 u_{ki}를 계산.
Step 4. 객체 i를 u_{ki}가 가장 큰 군집 k에 속하게 만들고 군집화를 수행.
앞의 군집 결과와 비교하여 동일하거나 변동이 ε보다 작으면 정지하고 그렇지 않으면 Step 1을 반복 수행.

2.2 주성분 분석(Principal Component Analysis)

주성분 분석은 서로 상관관계가 있는 변수들 사이의 복잡한 구조를 단순화하고, 이해하기 쉽게 설명하기 위해 사용되는 분석 기법으로, 변수들의 선형결합을 통하여 변수들이 갖는 전체 정보를 최대한 설명할 수 있는 서로 독립인 새로운 인공변수(artificial variable)들을 유도하여 해석하는 기법이다. 이러한 인공변수를 주성분(principal component)이라 한다(Richard and Dean, 2002).

m차원을 갖는 n개의 자료 X가 존재할 때, 이 자료에 대한 평균행렬 $\overline{X}$와 공분산(covariance) 행렬 S는 다음과 같다.

$$\overline{X}=\begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_m \end{bmatrix},\quad S=\begin{bmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \ddots & \vdots \\ s_{m1} & \cdots & s_{mm} \end{bmatrix},\quad \bar{x}_l=\sum_{k=1}^{n} x_{lk},\quad s_{ij}=\frac{n\sum_{k=1}^{n} x_{ik}x_{ik}-\sum_{k=1}^{n} x_{jk}\sum_{k=1}^{n} x_{jk}}{n(n-1)}.$$

이와 같은 자료의 공분산 행렬 S에 대한 고유치(eigen value)와 고유벡터(eigen vector)는 다음과 같이 구할 수 있다.

$$U^T S U = L\ ,$$

여기서 대각행렬 L의 원소 $l_1, l_2, \cdots, l_m$을 S의 고유치라 하고, 행렬 U의 종행렬 $u_1, u_2, \cdots, u_m$을 S의 고유벡터라 하며, 각 고유벡터들은 다음과 같은 조건 (3)을 만족하며 자료의 특성을 가장 잘 나타내는 직교 좌표계를 대표한다.

$$u_i^T u_j=\begin{cases} 1, & \text{if} \quad i=j \\ 0, & \text{otherwise} \quad i\neq j \end{cases}. \tag{3}$$

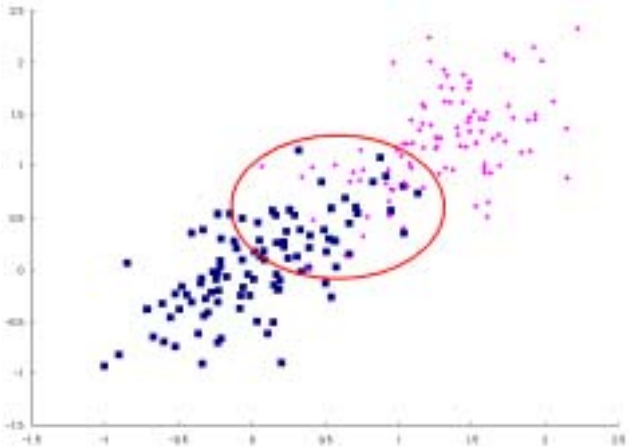
그리고 좌표축 변환으로 m차원의 연관된 자료 $X=[x_1, x_2, \cdots, x_m]^T$는 새롭게 연관되지 않은 m차원의 자료 $Z=[z_1, z_2, \cdots, z_m]^T$를 생성하게 된다. 이때,

$$Z=U^T[X-\overline{X}]\ .$$

3. 시뮬레이션 과정 및 결과

주성분 분석을 이용한 fuzzy c-means double clustering은 분류할 자료를 fuzzy c-means 알고리즘으로 1차 군집화 하여 각 객체의 소속 정도(확률)인 소속 행렬 U의 수치에 따라 군집을 정의한다. 다음으로 소속 정도(확률)가 불확실한 객체를 선정, 이들을 주성분 분석으로 변환하여 다시 fuzzy c-means 알고리즘을 적용하는 방법이다. 이에 대한 시뮬레이션 과정은 다음 절차를 거친다.

Step 1. 이변량 정규분포(bivariate normal density)에서 두 변량 사이에 상관성이 존재하고 두 군집 경계에 자료가 공통으로 위치하게 자료 집합 생성.
Step 2. 1차 fuzzy c-means 알고리즘을 이용한 자료 분류와 이에 대한 소속 행렬 U 계산.
Step 3. 두 군집의 중심(c_1, c_2)에 대한 평균과 합동 분산(pooled variance)을 계산하여 새로운 분포를 추정 공통 영역에 속하는 자료 집합 선정.
Step 4. 공통 영역에 속하는 자료에 대해 2차 fuzzy c-means 알고리즘을 적용하여 double clustering 수행.



<그림 2> 이변량 정규분포에서 생성된 공통 영역이 존재하는 자료 plotting

이와 같은 과정을 거쳐 <그림 2>에서와 같이 두 자료의 공통 영역에 위치한 타원 안의 자료를 분류하는 과정에서 정확도를 높이기 위해 주성분 분석을 이용한 fuzzy c-means double clustering 기법을 적용했다.

3.1 이변량 정규분포(Bivariate Normal Density)

주성분 분석을 이용한 fuzzy c-means double clustering의 시뮬레이션을 위해 두

변량 사이에 연관성이 있고 공통 영역에 자료가 분포하는 이변량 정규분포(bivariate normal density) 자료를 생성하여 시뮬레이션 하였다. 이변량 정규분포 자료의 생성 절차는 다음과 같다.

Step 1. $z \sim N(0, 1^2)$ 인 두 개의 표준 정규 변수 z_1과 z_2를 생성.
Step 2. $x = \mu_1 + \sigma_1 z_1$ 이고 $y = \mu_2 + \sigma_2\left[z_1\rho_{z_2}\sqrt{1-\rho^2}\right]$ 인 x와 y를 계산.
Step 3. x와 y는 평균이 각각 μ_1, μ_2이고 표준편차 σ_1, σ_2 그리고 상관관계 ρ.

<표 1> 시뮬레이션 이변량 정규 분포의 평균, 표준편차 그리고 상관계수

		μ_1	σ_1	μ_2	σ_2	ρ
Simulation 1	$X_1(N=100)$	0.0	0.5	0.0	0.5	0.7
	$X_2(N=100)$	1.0	0.5	1.0	0.5	0.5
Simulation 2	$X_1(N=100)$	0.0	0.5	0.0	0.5	0.7
	$X_2(N=100)$	1.2	0.5	1.2	0.5	0.5
Simulation 3	$X_1(N=100)$	0.0	0.5	0.0	0.5	0.7
	$X_2(N=100)$	1.4	0.5	1.4	0.5	0.5
Simulation 4	$X_1(N=100)$	0.0	0.5	0.0	0.5	0.7
	$X_2(N=100)$	1.6	0.5	1.6	0.5	0.5

이와 같은 과정을 거쳐 두 군집에 소속된 자료를 각각 100개씩 생성하며, 생성에 사용된 초기치인 평균(μ_1, μ_2), 표준편차(σ_1, σ_2) 그리고 상관계수(ρ)는 <표 1>과 같다. 본 연구에서는 두 군집 중심(c_1, c_2) 변화에 따라 공통 영역(경계 지역)에 위치한 자료에 대해 double clustering하여 시뮬레이션 결과를 확인하였다.

3.2 시뮬레이션을 위한 공통 영역의 설정

시뮬레이션을 위해 두 군집 사이에 존재하는 공통 영역에 대한 선택도 필수적인 과정이다. 이를 위해 두 군집의 중심(c_1, c_2)의 평균과 합동분산을 이용한 새로운 분포를 정규분포로 가정한다. 다음으로 두 군집 중심 (c_1, c_2)의 평균 $\bar{c} = (c_1 + c_2)/2$ 와 합동분산(pooled variance) $\overline{sd} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ 를 이용 신뢰구간 $\bar{c} \pm z_{a/2} \cdot \overline{sd}$ 을 <그림 3>처럼 선정해 이 영역에 포함된 자료의 소속 정도(확률) 중 가장 큰 값(최대 소속 정도)을 구한다. 신뢰구간 계산을 위한 $\bar{c}$에 대한 신뢰수준은 5%, 10%, 32%, 50%, 68%, 90% 그리고 95%로 하였다.

신뢰 구간에 따라 구간에 포함되는 최대 소속 정도(확률)를 구한 결과는 <표 2>와 같이 나타났으며, 시뮬레이션 결과 신뢰 구간 50%인 즉, 최대 소속 정도 0.70 내외에

서 가장 좋은 double clustering 결과를 보였다. 신뢰구간 90%와 95%의 경우는 최대 공통 영역을 초과하여 분석 과정에서 제외하였다.

<그림 3> 두 군집에 가정한 새로운 분포(왼쪽)와 평균에 대한 신뢰 구간(오른쪽)

<표 2> 분포의 신뢰 구간에 따른 최대 소속 정도(확률)

	5%	10%	32%	50%	68%
최대 소속 정도	0.52	0.54	0.63	0.74	0.88

3.3 시뮬레이션 결과

최종 시뮬레이션은 두 군집의 중심간 거리의 변화에 따른 4가지 경우에 대해 이변량 정규분포에서 생성한 두 군집에 속하는 자료 각각 100개씩 생성하고, 신뢰 구간을 이용한 결과에서 최대 소속 정도 0.70을 중심으로 0.65와 0.75도 함께 fuzzy c-means double clustering에 이용하였다.

<표 3>은 1차 fuzzy c-means 오분류와 소속 정도(확률)에 따른 double clustering 오분류 비교 결과로 simulation 1의 경우 1차 분류에서 23.23개의 mis-cluster로 11.62% 오분류 되었다. 최대 소속 정도에 따른 double clustering에서 0.65에서 20.92개로 10.46% 오분류 되어 개선되었으며, 세 소속 정도(0.75, 0.70 그리고 0.65)에서 분류한 결과 평균적으로 1.78개를 정확히 분류하였다.

Simulation 2에서 200개의 자료 중 1차 fuzzy c-means 알고리즘으로 군집화 한 결과 24.13개로 12.07%, 최대 소속 정도 0.70에서 19.76개로 9.88% 오분류 되었다. $N_1 \sim (0, 0.5^2)$, $N_2 \sim (0, 0.5^2)$ 그리고 $\rho = 0.7$ 인 자료 100개와 $N_1 \sim (1.2, 0.5^2)$, $N_2 \sim (1.2, 0.5^2)$ 그리고 $\rho = 0.5$ 인 초기치로 생성된 자료인 simulation 2의 경우가 나머지 3개의 시뮬레이션보다 좋은 결과를 보였다.

다음으로 simulation 3의 결과 최대 소속 정도 0.65에서 7.93개로 3.97% 오분류 되어 1차 분류 10.68개보다 2.75개 정도 정확히 분류하여 개선된 결과를 보였다. 마지막으로 simulation 4의 경우 최대 소속 정도 세 개의 평균 0.5개가 개선되었으나 1차 결과에 비해 큰 변화가 없는 것으로 분석되었다.

이와 같은 시뮬레이션 결과 1차 분류된 군집과 최대 소속 정도에 따른 그리고 두 군집 사이 거리에 따라 차이를 보이고 있으나 전체적으로 fuzzy c-means double clustering 결과가 정확도 측면에서 효용성을 보여 주었다.

<표 3> 1차 fuzzy c-means 오분류(original)와 소속 정도(p)에 따른 오분류 결과

	original	$p=0.75$	$p=0.70$	$p=0.65$
	mis-cluster	mis-cluster	mis-cluster	mis-cluster
simulation 1	23.23	22.11	21.31	20.92
simulation 2	24.13	20.98	19.76	21.06
simulation 3	10.68	9.22	8.47	7.93
simulation 4	6.57	7.13	6.18	4.89

3.4 붓꽃(Iris) 자료를 이용한 군집화

실제적인 주성분 분석을 이용한 fuzzy c-means double clustering을 위해 패턴 분류에서 가장 많이 이용되는 Fisher의 붓꽃 자료에 제안된 방법을 적용하였다. 150개에 대한 붓꽃 자료는 꽃받침의 길이, 꽃받침의 너비, 꽃잎의 길이, 그리고 꽃잎의 너비의 변수이다. 4가지 변수로 분류되는 붓꽃들은 setosa, versicolour 그리고 virginica의 3종류이다.

<표 4>에서 cluster 1에 소속된 자료의 경우 1차 fuzzy c-means와 주성분 분석을 이용한 fuzzy c-means double clustering의 경우 모두 100% 정확도로 같은 결과를 보여 차이를 보이지 않았으나, cluster 2, 3의 경우 각각 78%, 90%가 88%, 94%로 정확도의 증가를 보여주었다(최대 소속 정도 0.75).

그러므로 시뮬레이션뿐만 아니라 실제 다양한 기법의 성능 평가에 이용되고 있는 붓꽃 자료에서도 본 연구에서 제한한 방법의 정확도의 증가를 보여 주는 것을 확인하였다.

<표 4> 붓꽃 자료의 1차 fuzzy c-means(왼쪽)와 제안된 방법(오른쪽)의 결과

	Clu 1	Clu 2	Clu 3	accuracy
Clu 1	50	0	0	100%
Clu 2	0	39	11	78%
Clu 3	0	5	45	90%

	Clu 1	Clu 2	Clu 3	accuracy
Clu 1	50	0	0	100%
Clu 2	0	44	6	88%
Clu 3	0	3	47	94%

* Clu 1(cluster 1): setosa, Clu 2(cluster 2): versicolour, Clu 3(cluster 3): virginica

4. 결론 및 토의

Fuzzy c-means 알고리즘은 각 군집 중심까지의 거리가 같은 객체 또는 소속 정도(확률)가 유사할 때, 그리고 군집의 크기와 상관성이 다른 경우 군집화 과정에서 오분류 가능성을 갖고 있다. 이를 개선하기 위해, 상관관계 제거와 자료의 정보를 그대로 유지하고 변수의 수를 감소시키는 주성분 분석을 이용한 fuzzy c-means double clustering 방법을 제안하였다.

Fuzzy c-means double clustering 시뮬레이션 결과 최대 소속 정도(확률)에 따라 정확도가 다르게 나타나고 있으나 전체적으로 0.5~4.37개의 정확도 향상을 보였다. 그리고 실제 분류에 대표적으로 사용되는 붓꽃자료에서는 1차 fuzzy c-means 알고리즘의 경우 전체 150개 중 16개로 오분류 되어 89%의 정확도를 보였고, 제안된 fuzzy c-means double clustering 방법으로 9개를 오분류 되어 94%의 정확도로 분류되었다. 결론적으로 5%의 정확도가 향상된 것을 확인 하였다.

본 연구에서 제안된 주성분 분석을 이용한 fuzzy c-means double clustering은 시뮬레이션과 붓꽃 자료에서 기존의 fuzzy c-means 알고리즘보다 정확도의 개선을 보여 주었다. 그러나 소속의 정도의 기준에 대한 정확한 분류 방법의 해소와 두 군집 사이에 분포하는 자료의 관계에 따라 다른 결과가 나타나는데, 이에 대한 정확한 기준 제시에 대한 연구가 앞으로 필요할 것으로 보인다.

참고문헌

1. Anderberg, M.R. (1973). *Cluster Analysis for Applications*, Academic Press.
2. Bezdek, J.C. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithms, IEEE Trans. on Patt. Anla. and March Intell, Vol. PAMI-2, No. 1, pp. 1-8.
3. Dembele, D. and Kastner, P. (2003). Fuzzy C-means method for clustering microarray data, *Bioinformatics*, 19: 973-980.
4. Hartigan, J.A. (1975). *Clustering Algorithms,* John Wiley & Sons.
5. Jain, A. K. and Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.
6. Jardine, N. and Sibson, R. (1968). Construction of Hierarchinc and Non-hierarchic classifications, Computer Journal 11, pp 177-184.
7. Krishnapuram, H.F and Nasraoui, O. (1995). Fuzzy and Possibilistic shell Clustering algorithms and their application to boundary detection and surface approximation, IEEE Trans. on Fuzzy Sys, Vol. 3, No. 1, pp. 29-60.
8. Pal, N. and Bezdek, J.C. (1995). On Cluster Validity for the Fuzzy C-Means Model, IEEE Trans. on Fuzzy Sys., Vol. 3, No. 3.
9. Richard A.J. and Dean W.W. (2002). *Applied Multivariate Statistical Analysis*, Prentice Hall.
10. Tryon, R.C. (1939). *Cluster analysis*, Edwards Brothers, Ann Arbor, MI.
11. Tryon, R.C. and Bailey, D.E. (1970). *Cluster analysis*, McGraw-Hill, New York.

[2006년 2월 접수, 2006년 4월 채택]