한국 퍼지 및 지능시스템 학회 2003, Vol. 13, No. 4, pp. 416-426

새로운 초기치 선정 방법을 이용한 향상된 EM 알고리즘

Improved Expectation and Maximization via a New Method for Initial Values

김성수 · 강지혜

Sung-Soo Kim and Jee-Hye Kang

충북대학교 전기전자 및 컴퓨터 공학부

요 약

본 논문은 시스템 공학의 인식에 관련된 여러 분야에서 널리 쓰이는 클러스터링 기법인 Expectation-Maximization의 초기값 설정문제에 관하여 새로운 방법을 제안한다. 기존의 임의로 지정하는 랜덤한 초기치 선정 문제점을 지적하고, 새로이 제안하는 균등 영역 분할과 분할 된 데이터의 통계적 특성을 이용한 초기치 설정 방법을 사용한 새로운 EM 알고리즘을 제안한다. 일반적으로 EM에서 초기값 설정 방법으로 랜덤한 설정 방식의 약점을 보완하기 위하여 K-means 방법을 많이 사용하고 있다. 하지만, K-means 초기치 설정 방법도 근본적인 문제는 해결하지 못하고 있다. 이러한 문제의 하나의 해결 방안으로 논문이 제안한 균등 분할 및 통계적 특성을 이용한 초기치 선정의 방법을 EM 알고리즘에 적용하였다. 제안된 방법은 기존보다 EM 알고리즘의 특성을 극대화하는 방향으로 더 좋은 결과를 가져온다. 본 논문에서 제안된 알고리즘의 우수성을 제안한 초기치 선정 방법을 적용한 EM과 기존 EM의 시뮬레이션 결과를 비교 분석하여 그 우수성을 제시하였다.

Abstract

In this paper we propose a new method for choosing the initial values of Expectation-Maximization(EM) algorithm that has been used in various applications for clustering. Conventionally, the initial values were chosen randomly, which sometimes yields undesired local convergence. Later, K-means clustering method was employed to choose better initial values, which is currently widely used. However the method using K-means still has the same problem of converging to local points. In order to resolve this problem, a new method of initializing values for the EM process. The proposed method not only strengthens the characteristics of EM such that the number of iteration is reduced in great amount but also removes the possibility of falling into local convergence.

Key Words : Expectation-Maximization, K-means, Uniform Partitioning, Initial Values.

1. 서 론

현재, 패턴 인식, 자동제어 및 영상처리 등의 많은 분야에서 응용되고 있는 클러스터링의 기법[1]은 날로 그 중요성이 더해지고 있다. 이처럼 그 필요성이 증가됨에 따라 여러 분야의 연구자들은 그 성능을 향상시키기 위한 많은 이론과 기법을 연구해왔다. 일반적으로, 클러스터링 과정 중에서, 초기값 설정의 문제[2]는 이미 오래 전부터 인식되어 왔고, 그 해결 방안의 필요성이 점차 증대되고 있다. 본 논문에서는 이러한 Expectation-Maximization (EM)의초기값 선정에 대한 문제[3]의 해결 방안의 하나로서 클러스터링의 향상된 결과를 얻을 수 있는 알고리즘을 제시하였다. 이것은 EM과정[4]에서 가장 중요한 초기값을 설정하는데 있어서, 일반적으로 K-means을 이용하는데, K-means의 초기값 설정 방법을 기존과 다른 알고리즘으로 개선하여 전체적인 EM과정의 초기치 선정의 문제를 해결하고자 하였다.

우선, 기존의 K-means 방법의 문제점은, 단순히 주어진 데이터 중에서 임의로 지정된 초기값으로부터[5-7], 점차 중심값을 갱신해 나가는 것이다. 이와 같은 방법은 랜덤하게 설정되는 초기값에 따라 여러 다른 값으로 수렴하는 불안정한 결과[8-9]로서, 제대로 클러스터링을 하지 못 할 경우 바람직하지 않은 방향의 중심값을 찾게 된다. 그러한 문제의 해결방안으로써 본 논문이 제안하는 것은 클러스터링의 대상이 되는 주어진 데이터의 분포 특성을 이용한다는 것이다. 이러한 데이터의 분석을 통하여 초기값을 설정함으로써 효율적인 K-means의 알고리즘을 수행하고 향상된 결과 값을 얻게 된다. 새로운 알고리즘의 장점으로는, 기존 방법에서의 막대한 반복 계산량을 줄이고, 타당하지 않은 방향으로의 클러스터링 오차를 줄일 뿐만 아니라, 안정되고 향상된 클러스터링의 결과를 가져온다. 본 논문이 제안한 알고리즘의 우수성을 보여주기 위하여, 제안한 방법과 기존의 방법이 각각 EM에 적용하여 미치는 결과를 시뮬레이션을 통해 비교 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 K-means

접수일자 : 2003년 2월 15일
완료일자 : 2003년 6월 30일
본 논문은 2002년도 한국학술진흥재단의 신진연구과제 지원사업(D00343)의 지원에 의하여 연구된 것입니다.

와 EM의 알고리즘의 개념과 과정에 대해서 설명하였고, 3장에서는 제안된 새로운 EM 시스템의 초기치 설정 방법에 대하여 설명하였고, 4장에서는 앞에서 제시된 이론적 정립을 바탕으로 EM에 적용한 실험 결과를 보여주고 설명함으로써 본 논문의 객관적인 타당성을 보였다. 특히, 기존의 알고리즘과 제안된 알고리즘에 대해 동일한 데이터의 시뮬레이션 결과를 비교를 하였는데, 각각 특성이 다른 두 개의 데이터 모델에 적용하였다. 마지막으로 5장에서는 본 논문의 총체적 결론을 맺었다.

2. 기존의 K-means와 Fuzzy 알고리즘

2.1. K-means 알고리즘

K-means 알고리즘은 적용하고자 하는 데이터인 각 개체들을 다음과 같이 벡터 $x=[x_1, \cdots, x_n]$ 로 표현된다. 알고리즘이 적용되는 각 개체(벡터)들의 Euclidean norm과 같은 차원의 두 벡터 x와 y의 차를 나타내는 식들은 다음과 같이 정의된다.

$$\|x\| = \left[\sum_{i=1}^{n} x_i^2\right]^{1/2}$$

$$\|\mathbf{x}-\mathbf{z}\| = \left[\sum_{i=1}^{n} (x_i - z_i)^2\right]^{1/2} \tag{1}$$

전체 K-means 알고리즘은 초기화 단계, 개체분산단계, 새로운 클러스터의 중심설정단계로 이루어져있다. 간략히 각 단계를 정리하고 수렴성에 관한 사항을 알아보면 다음과 같다. 첫째, 초기화 단계에서는 생성할 클러스터의 개수 K를 정하고, K개의 각 클러스터에 대하여 클러스터의 중심을 초기화한다. 이때의 초기 중심을 지정하는 방법은 전체 데이터(n개) 중에서 임의로 K개만큼 선택한다.

$$\{z_1(l), z_2(l), \ldots, z_K(l) \in S_i \quad i=1,2,3,\ldots,n \tag{2}$$

둘째, 개체분산 단계에서는 각 개체들과 각 클러스터의 중심과의 유클리디언 거리를 구한다. 여기서 각 개체는 각 클러스터의 중심에서 객체까지의 거리가 가장 최소가 되는 클러스터에 속하게 된다. 모든 데이터 개체들은 매번, 각 클러스터 중심과의 거리를 계산한 후, 최소가 되는 클러스터에 각 개체들이 할당되게 된다. 즉, 다음을 최소화시키는 클러스터로 그룹 지어지는 것이다.

$$J_j = \|x_i - z_j(l)\|^2 \quad \text{for } i=1,2,\ldots,n,\ j=1,2,\ldots,K$$
$$\text{if } J_j(l) < J_i(l) \quad \text{for all } i, j=1,2,\ldots,K,\ i \neq j$$
$$\text{then } x_i \in S_j^{(l)} \tag{3}$$

여기서 개체간의 거리는 개체간의 유사성과 비유사성을 측정하는데 사용된다. 개체들 간의 거리측정을 위해서 Minkowski distance를 사용하기도 하는데, 일반적으로, 유클리디안 거리측정 방법을 이용하는 것이 상례이다. 아래 식들은 Minikowski 거리의 표현과 Euclidean 거리와 특성을 식으로 나타낸 것으로, Metric의 특성을 잘 만족시키고 있음을 알 수 있다.

Minkowski distance :

$$d(i,j) = \sqrt[q]{(|x_{i1}-x_{j1}|^q + |x_{i2}-x_{j2}|^q + \ldots + |x_{ip}-x_{jp}|^q)}$$
$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}),\ x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$$

$q=2$, Euclidean distance

$$d(i,j) = \sqrt{(|x_{i1}-x_{j1}|^2 + |x_{i2}-x_{j2}|^2 + \ldots + |x_{ip}-x_{jp}|^2)}$$

Properties :

$$d(i,j) \geq 0$$
$$d(i,i) = 0$$
$$d(i,j) = d(j,i)$$
$$d(i,j) \leq d(i,k) + d(k,j) \tag{4}$$

세 번째 단계로는 새로운 클러스터의 중심 계산이 이루어진다. 이전 단계에서 개체들이 할당되어 재구성된 K개의 각 클러스터에서, 기존의 중심 값에서 새로운 클러스터의 중심을 계산한다. 이는 각 클러스터에 속한 데이터 개체들의 평균값을 구함으로써, 곧 새로운 중심이 된다. 따라서 이 과정에서 기존의 클러스터의 중심이 새롭게 바뀌는 것이다.

$$z_j(l+1) = \frac{1}{N_j} \sum_{x_i^{(j)} \in S_j^{(l)}} x_i^{(j)} \tag{5}$$

여기서, N_j는 두 번째 단계에서 만들어진 각 클러스터에 할당된 총 데이터 개체들의 수를 나타내고, $x_i^{(j)}$는 j 클러스터에 속해진 각 개체들을 의미한다. K-means 알고리즘의 수렴여부에 관해서는 더 이상 각 클러스터의 중심에 변화가 생기지 않을 때 종료된다. 만일 클러스터의 중심에 변화가 생겼다면 두 번째 단계로 돌아가서 반복을 계속한다.

$$\text{if } z_j(m) = z_j(m+1) \quad \textit{then} \quad \textit{end} \tag{6}$$

위의 방법 이외에도, 종료의 조건으로 앞의 두 번째 단계에서 구한 각 클러스터링 중심과 각 개체와의 거리계산 값 중 최소의 값을 모두 더한 값을, 각 반복단계에서의 오차로 한다. 따라서 허용 오차의 임계치를 설정해 주어서 각 반복과정이 수행되기 전에 조건으로 지정하여 수렴 여부를 확인할 수도 있다. 이러한 과정을 식으로 표시하면, 다음과 같다.

$$M_i = \min(J_j) \quad j=1,2,\ldots,K \quad , \quad i=1,2,\ldots,n$$
$$\varepsilon_{iter} = \sum_{i=1}^{n} M_i$$
$$\textit{if} \quad \varepsilon_{\min} \leq \varepsilon_{iter} \quad \textit{then} \quad \textit{end} \tag{7}$$

여기서 M_j는 두 번째 단계에서 구한 각 클러스터 중심과 개체의 유클리디언 거리(J_j) 중 최소 값들의 총 합이고, ε_{iter}는 각 단계에서의 최소 값(M_j)을 모두 더한 값으로 각 단계에서의 오차 값을 의미한다. 그리고, $\varepsilon_{\min}$은 설정해준 에러 값의 임계치를 나타낸다.

2.2 EM 알고리즘

많은 추정 문제들에 있어서 알고자 하는 파라미터는 잡음이 있는 환경에서의 관심의 대상이 되는 정보의 평균값을 추정하는 것이다. EM(Expectation-maximization) 알고리즘은 이러한 문제에 적합하도록 파라미터들의 최대의 가능성

(Maximum-likelihood : ML)을 추정한다[10]. 정보가 직접적으로 얻어지지 않고, 다른 관측 가능한 변수를 통하여 획득 할 수 있는 경우이므로, 관심의 대상이 되는 정보를 관측 가능한 변수의 공간을 통하여 추정하는 통계적 방법이다. 여기서 파라미터들은 관측된 값들에서 다수 대 일(many-to-one) 대응관계의 분포를 갖는다. 바로 EM의 장점은 관측 가능한 변수의 공간에 일대일 대응으로 정보가 관계되어 있지 않더라도, 원하는 정보를 추정할 수 있다는 점이다. 예를 들면, 관측 가능한 샘플 공간을 Y라 놓고, Y의 부분 공간인 관측된 정보 공간을 $y \in R^n$ 라 하면, 샘플 공간 Y와 대응 관계를 갖는 공간을 X라 놓는다. 여기서도 마찬가지로, 관측된 정보 공간과의 관계를 갖는 부분 공간을 $x \in R^m$, $m < n$, 라고 한다. 여기서 x 는 직접 관측되지 않으며 관측 가능한 y와 $y = y(x)$의 다 대 일 매핑의 관계를 통하여 얻어 진다. 관측된 y는 X의 부분 공간 $x(y)$를 결정한다. 여기서 관측되는 공간 $y \in R^n$을 불완전한 데이터 공간이라 하고, 이 불완전한 공간을 구성하는 랜덤변수의 확률 밀도 $g(y)$를 불완전한 공간의 확률밀도함수라 한다. 이에 반하여, 구하려는 정보 공간 x를 완전한 데이터라 하고, 이 완전한 데이터를 구성하는 랜덤변수들의 확률밀도 $f(x)$를 완전한 데이터의 확률 밀도함수라고 한다. 일반적으로 이 두 공간의 확률밀도 함수는 임의의 관련된 변수를 통하여 연결되어있다. 위의 관계를 매개변수 θ를 사용하여 나타내면 다음과 같다. 완전한 데이터의 확률밀도함수는 $f_X(x|\theta) = f(x|\theta)$이고 여기서 $\theta \in \Theta \subset R^n$ 는 밀도의 파라미터 집합이다. 불완전한 데이터의 확률밀도함수는 $g(y|\theta) = \int_{x(y)} f(x|\theta)dx$ 로 나타낸다. 여기서, 매개 변수의 값을 구하기 위해서 매개변수를 조건으로 하는 대수-추정값(Log-likelihood)함수를 이용하여, 주어진 데이터 y와 앞에서 살펴본 파라미터의 ML 추정값으로 L(Log-likelihood)의 최대화를 통해서 찾는 것으로 요약할 수 있다.

우선, 이용할 혼합 모델의 확률 밀도 함수는 기저함수의 선형조합으로 표현할 수 있다고 가정하자. M개의 요소들의 조합으로 이루어진 모델을 식(8)와 같은 형태로 나타낸다.

$$p(x) = \sum_{j=1}^{M} P(j)\, p(x|j) \tag{8}$$

위 식에서의 $P(j)$를 혼합 계수(mixing coefficients)라고 하며, $p(x|j)$은 요소의 가능성을 나타내는 "활성화(activations)"라 한다. 이것들은 다음과 같은 특성을 만족한다.

$$\sum_{j=1}^{M} P(j) = 1, 0 \le P(j) \le 1$$
$$\int p(x|j)\, dx = 1 \tag{9}$$

일반적으로, $P(j)$는 j 번째 요소의 Prior(선행) 확률이다. 또한, $p(x|j)$은 j 번째 요소에 속하는 경우를 조건으로 하는 확률 밀도함수이므로, Baye's 이론에 의하여 다음의 식(10)과 같은 Posterior(후행) 확률을 구할 수 있다.

$$P(j|x) = \frac{p(x|j)\, P(j)}{p(x)} \tag{10}$$

이것도 마찬가지로, 다음의 성질을 만족한다.

$$\sum_{j=1}^{M} P(j|x) = 1, 0 \le P(j|x) \le 1 \tag{11}$$

앞으로 위에서 제시한, 가우시안 혼합 모델을 가지고 알아본다. 일반적으로, ML 추정은 그 분포에 따라 유도되는 관측 데이터에 근본을 두고 파라미터를 추정하는 방법이라고 언급한 바와 같이, ML추정에서의 주요 개념은 $x_1, x_2, \cdots, x_N$을 관측할 확률이 가능한 높도록 파라미터를 결정하는 것으로서, 최대 가능성(ML)의 해를 식 (12)과 같이 Negative log-likelihood를 만족하도록 한다.

$$E = -\ln L = -\sum_{n}^{N} \ln p(\mathbf{x}^n) = -\sum_{n}^{N} \ln \left\{ \sum_{j}^{M} p(\mathbf{x}^n|j)\, P(j) \right\} \tag{12}$$

위 식은 에러 함수로도 사용된다. 그것은 최대의 가능성(ML)의 L은 최소의 E와 같기 때문이다. ML추정의 예로서, $X_1, X_2, \cdots, X_N$이 미지의 평균 $\hat{\mu}$과 분산 $(\hat{\sigma})^2$을 갖는 독립 가우시안 랜덤 변수라 하고 $x_1, x_2, \cdots, x_N$은 이들 랜덤 변수의 샘플이라 하면, 평균과 분산 그리고 Prior(선행) 확률의 ML 추정값이 다음 식(13)과 같이 표현될 수 있다.

$$\hat{\mu}_j = \frac{\sum_n P(j|\mathbf{x}^n)\mathbf{x}^n}{\sum_n P(j|\mathbf{x}^n)}$$
$$(\hat{\sigma}_j)^2 = \frac{1}{d} \frac{\sum_n P(j|\mathbf{x}^n) \|\mathbf{x}^n - \hat{\mu}_j\|^2}{\sum_n P(j|\mathbf{x}^n)}$$
$$\hat{P}(j) = \frac{1}{N} \sum_{n}^{N} P(j|\mathbf{x}^n) \tag{13}$$

위의 주어진 식들은 최대가능성(Maximum likelihood)의 해의 본질을 파악하는데 유용하게 제공되지만, 직접적인 파라미터들의 계산 값을 얻지는 못한다. 사실상 그것들은 높은 차원의 비선형 결합 방정식으로 재해석되는데, 그것은 파라미터들이 전적으로 식 (10)의 우변 항과 같은 결과로 나타나기 때문이다. 우리는 가우시안 혼합 모델의 파라미터를 위한 초기 추정 값을 설정해줌으로써 시작한다고 가정하고, 그것을 과거(old)의 파라미터 값이라고 부른다. 그런 후, (13)의 우변을 계산할 수 있고, 이것은 에러 방정식을 더 작게 하기 위한 값인 새로운(new) 파라미터로 불려질 파라미터들을 위해 갱신된 추정을 준다. 이러한 파라미터들은 다음 과정에서는 과거의 값으로 되고 결국, EM 알고리즘은 가우시안 혼합 모델의 파라미터들의 값이 에러함수 E가 작아지도록, 즉 최대의 가능성 L을 만족하도록 수렴하기까지 그러한 과정은 반복되어진다. 만약 k번째 반복에서의 파라미터 예측 값을 $\theta^{[k]}$라고 할 때 전체적으로 다음과 같은 두 단계로 표현된다. 우선 E (Expectation)단계는 식(14)와 같이 현재 추정된 파라미터와 관측된 데이터를 조건으로 하여 Posterior(후행) 확률의 기대값을 계산 할 수 있다.

$$Q(\theta|\theta^{[k]}) = E[\log f(\mathbf{x}|\theta)\,|\,\mathbf{y}, \theta^{[k]})] \tag{14}$$

그 다음, 식(15)과 같이 파라미터 θ가 $Q(\theta|\theta^{[k]})$를 최대화하게 선택한다.

$$\theta^{[k+1]} = \arg\max Q(\theta|\theta^{[k]}) \tag{15}$$

바로, M (maximization) 단계에서는 식 (15)처럼, E 단계에서 구한 값을 이용하여 L(Log-likelihood)의 최대화를 만족하기 위한 새로운 추정값, 파라미터 θ를 갱신한다. 이러한 두 단계는 수렴조건을 만족할 때까지 반복된다. 이때의 반복은 M 단계의 갱신 값들이 실행되는 방향으로 우세하게 일어나는 조건으로, 각 반복되는 시점에서의 에러 함수를 작게

하여 국부 최소점(Local minimum point)이 발견될 때까지 계속된다. 이것은 비선형 최적화 알고리즘의 복잡성을 피한 혼합 파라미터들의 추정을 위한 단순하고 실제적인 방법을 제공한다. 과거의 파라미터들이 새로운 값으로 교체될 때 에러 값 E의 변화를 다음과 같이 쓸 수 있다.

$$E^{new} - E^{old} = -\sum_n \ln\left\{\frac{p^{new}(x^n)}{p^{old}(x^n)}\right\} \quad (16)$$

여기서 $p^{new}(x^n)$ 은 파라미터들을 위한 새로운 값들로 계산되어지도록 사용된 확률 밀도함수로 나타내고, 반면에 $p^{old}(x^n)$은 과거의 값들로 계산되도록 사용된 밀도 값이다. 식(8)의 주어진 혼합 분포의 정의를 이용하여, 다음과 같이 쓸 수 있다.

$$E^{new} - E^{old} = -\sum_n \ln\left\{\frac{\sum_j P^{new}(j)p^{new}(x^n|j)}{p^{old}(x^n)}\frac{P^{old}(j|x^n)}{P^{old}(j|x^n)}\right\} \quad (17)$$

위 (17)식의 대 괄호 안의 마지막 인자는 같도록 일치 시켜준다. 우리는 $\sum_j \lambda_j = 1$를 만족하는 $\lambda_j \geq 0$인 수들의 집합에 대해서 다음과 같은 Jensen's inequality(젠센의 부등식)을 이용한다.

$$\ln(\sum_j \lambda_j x_j) \geq \sum_j \lambda_j \ln(x_j) \quad (18)$$

확률 $P^{old}(j|x^n)$의 합은 1이므로, 그것은 위 식의 λ_j 처럼 쓰일 수 있으므로 다음과 같이 전개된다.

$$E^{new} - E^{old} \leq -\sum_n \sum_j P^{old}(j|x^n) \ln\left\{\frac{\sum_j P^{new}(j)p^{new}(x^n|j)}{p^{old}(x^n)P^{old}(j|x^n)}\right\} \quad (19)$$

새로운 파라미터를 고려한 E^{new}를 최소로 만들기 위해서, 만일 위 식의 우변을 Q 로 놓는다면, $E^{new} \leq E^{old} + Q$ 를 가지고 $E^{old} + Q$ 은 E^{new}의 상위 범위로 재해석된다. 그러므로 파라미터의 새로운 값을 고려한 이 범위의 최소 값을 찾을 수 있다. 최소의 Q는 E^{new}가 이미 국부 최소 값이 아니라면 반드시 E^{new}값의 감소를 초래하게 된다. 만일, 과거의 파라미터에 의지하는 요소들을 제거해버리면, 식(19)을 다음과 같이 쓸 수 있다.

$$\tilde{Q} = -\sum_n \sum_j P^{old}(j|x^n) \ln\{P^{new}(j)p^{new}(x^n|j)\} \quad (20)$$

그리고 상위 범위의 최소 값은 식(20)을 최소화함으로써 구해진다. 만약 우리가 가우시안 혼합 모델의 특별한 경우로 여긴다면, 다음과 같이 쓸 수 있다.

$$\tilde{Q} = -\sum_n \sum_j P^{old}(j|x^n) \ln\{\triangle_{new}\} + \text{const.} \quad (21)$$

$$\triangle_{new} = P^{new}(j) - d\ln\sigma_j^{new} - \frac{\|x^n - \mu_j^{new}\|^2}{2(\sigma_j^{new})^2}$$

이제 새로운 파라미터들을 고려한 위 함수를 최소화 할 수 있다. 최소화를 위한 파라미터 μ_j 와 σ_j은 곧장 계산해 갈 수 있지만, 혼합 파라미터 $P^{new}(j)$은 반드시 제약조건 $\sum_j P^{new}(j) = 1$ 을 고려해야만 한다. 이것은 라그랑지 곱셈기 λ 를 도입함으로써 쉽게 처리되고, 최소 함수는 다음과 같다.

$$\tilde{Q} + \lambda\left(\sum_j P^{new}(j) - 1\right) \quad (22)$$

다음은, $P^{new}(j)$를 고려한 (22)의 도함수를 0으로 놓음으로써, 아래와 같은 식을 얻는다.

$$0 = \sum_n \frac{P^{old}(j|x^n)}{P^{new}(j)} + \lambda \quad (23)$$

λ의 값은 (23)의 양변을 $P^{new}(j)$로 곱하고 전체 j 에 대하여 합하면 구할 수 있다. $\sum_j P^{new}(j) = 1$ 와 $\sum_j P^{old}(j|x^n) = 1$ 를 이용하여 $\lambda = N$ 것을 구한 후, 최종적으로 혼합 모델의 파라미터들을 위한 다음의 갱신 방정식들을 얻는다.

$$\mu_j^{new} = \frac{\sum_n P^{old}(j|x^n)x^n}{\sum_n P^{old}(j|x^n)} \quad (24)$$

$$(\sigma_j^{new})^2 = \frac{1}{d}\frac{\sum_n P^{old}(j|x^n)\|x^n - \mu_j^{new}\|^2}{\sum_n P^{old}(j|x^n)} \quad (25)$$

$$P^{new}(j) = \frac{1}{N}\sum_n P^{old}(j|x^n) \quad (26)$$

위 식들의 우변 항 중 어느 위치에 새로운 파라미터들 과 과거의 파라미터들이 나타나는지 주의 깊게 관찰하여, 이에 대응되는 최대 추정(ML)값의 결과와 비교되어만 한다.

3. EM에서의 새로운 초기값 설정 방법

앞장에서 살펴본 EM알고리즘에서 가장 중요한 것은 알고리즘을 수행할 수 있는 파라미터 θ 값의 초기값이다. 주어진 관측 데이터와 파라미터 θ값으로부터 구해지는 E 단계의 Posterior(후행) 확률을 가지고 M 단계에서 새로운 파라미터를 갱신하게 되기 때문이다. 이렇듯, 초기 파라미터값의 설정은 전체적인 EM 알고리즘에서 막대한 영향력을 발휘하게된다. 그리하여, 일반적으로 EM의 초기 파라미터 θ값을 설정할 때 널리 쓰는 방법은 K-means를 이용하는 것이다. 그러나 1장의 서론에서 살펴봤듯이, 기존의 K-means 방법은 여러 가지 문제점들을 가지고 있다. K-means 알고리즘 역시, 그 초기값에 따라서 결과값이 좌우됨으로 근본적으로 K-means의 초기값 문제부터 해결해야된다. 랜덤하게 관측된 데이터들 가운데에서 초기값을 지정하는 경우에는 막대한 반복 횟수량과 불안정한 결과값을 수반하기 때문이다.

본 논문에서 제안하는 초기치 선정의 방법은 클러스터링을 하고자 하는 데이터의 통계적 특성에 그 기반을 두고 있다. 데이터를 구성하는 차원의 공간을 균등 영역 분할 기법을 이용하여 데이터 밀도분포에 따른 영역의 평균값을 구해 초기 중심값으로 선정하는 방법이다. 이 방법을 데이터 공간의 적응적인 균등분할법이라 정의한다. 이 방법은 일단 구성된 데이터의 상대적 위치는 데이터가 회전이 되거나, 이동이 되거나, 또는 확대나 축소가 된다 하더라도, 데이터를 구성하는 각 요소들의 상호 위치는 변하지 않는다는 사실로부터 유도된다. 일반적으로 데이터 공간을 균등 분할하는 경우, 분할된 각 정보 공간의 단위공간에는 서로 다른 개수의 데이터가 포함되어 있다고 가정할 수 있다. 즉, 균등 분할된 정보공간의 데이터의 개수를 빈도수로 갖는 정보공간 차원의 데이터 밀도 분포를 갖게 된다.

첫째로 원하는 개수만큼의 초기 중심값을 구하는 방법으로, 우선 데이터 밀도 분포에서 밀도가 높은 개수만큼의 단위 정보 공간을 택한다. X를 클러스터링에 적용하려는 데이터로 놓고, 이 데이터가 $X = x_i, i=1,\cdots,M$의 M개의 개체들로 이루어져있다. 각 개체는 $x_i = x_{i,1}, x_{i,2}\cdots, x_{i,N}, i=1,2,\cdots,M$으로 N차원의 데이터이다. 이 때, 원하는 클러스터의 개수를 C 라하면, N차원의 데이터를 각 차원마다 C개의 균등분할을 실시한다. 결과적으로 C^N개의 균등분할 된 N차원의 단위공간들이 형성된다. 각각의 단위정보공간 내에 속해 있는 데이터의 개수를 밀도라 보고, C개의 밀도가 가장 큰 순서로 균등 분할된 공간을 선택한다. C개 이상의 단위공간들이 존재할 때는, 분포의 특성을 나타내는 분산이 작은 단위정보공간을 우선적으로 택한다. 이유는 분산이 작을수록 데이터의 밀집도가 높아지기 때문이다.

둘째로 원하는 클러스터개수 만큼 선택된 단위정보공간에서 초기 중심값은 선택된 단위정보공간에 포함되어 있는 데이터들의 평균값을 취한다. 즉, 임의의 단위정보인 S_k, $k=1,\cdots,C$ 공간에 L_k개의 개체가 존재할 때 초기 중심값은 다음과 같이 얻는다.

$$C_k = C_{k,1}, C_{k,2}, \cdots, C_{iN}, \quad k=1,2,\cdots,C$$

여기서, $C_{k,1} = \frac{1}{L_k}\sum_{i=1}^{L_k} x_{i,1}$는 k 번째 중심값으로 N차원의 요소 중에 m 번째 요소이다. 중심치가 포함될 가능성을 갖은 C^N개의 단위정보 공간들 중에서 C개만을 초기 중심값으로 선택하게 되는 것이다. 이러한 방법으로 초기 중심값을 선정하는 것은 다음과 같은 타당성을 뒷받침하고 있다. 바로, 데이터의 클러스터링과정은 데이터의 통계적 특성 중에 평균치와 유사성을 소유한 사실을 이용하기 때문이다. 이제, 본 논문이 제시한 균일한 영역 분할법을 이용하여 K-means의 초기값을 설정해 줌으로써, 기존보다 단축된 반복 과정을 통해서 안정된 클러스터링 결과 값을 얻게된다. 이러한 타당성 있는 값을 EM의 초기 파라미터값으로 설정함으로써, 에러 함수를 최소화시키는 국부 최소점(Local minimum point)에 도달하기까지 EM의 반복과정을 감소시키고, 또한 더욱 향상된 결과 값을 얻기 위해 제안된 K-means 방법을 EM의 초기값으로 적용한다. 그림 1의 블록도는 EM의 초기값을 제안한 방법의 K-means를 도입한 전체적인 알고리즘 과정을 나타내고 있다.

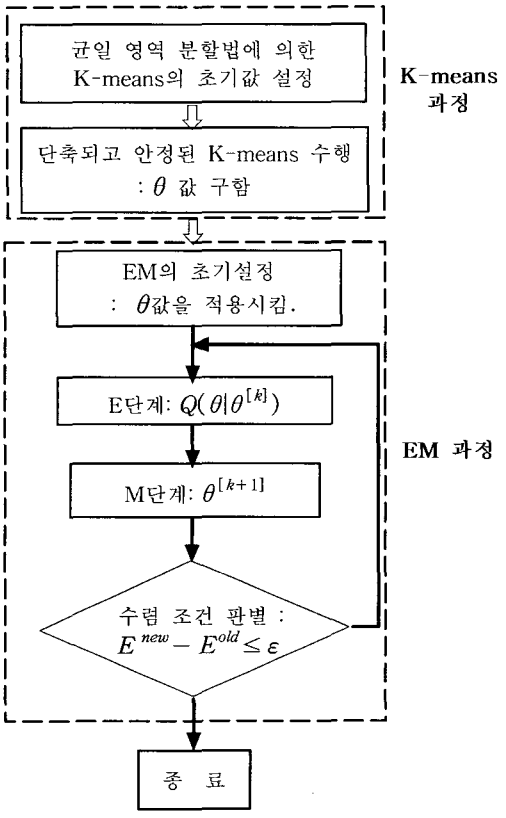

그림 1. 제안된 K-means를 적용한 EM 알고리즘 과정
Fig. 1 Flowchart of the proposed EM algorithm via the newly defined K-means

4. 시뮬레이션 및 결과 고찰

본 논문에서 제안한 알고리즘의 우수성을 보여주기 위해서, 여러 가지 데이터를 가지고 시뮬레이션의 결과를 가지고 기존 방법과의 차이를 비교하였다. 시뮬레이션 1에 사용된 데이터의 각 클러스터의 중심값은 그림 2에서 삼각형모양 분포인 C1, 원모양 분포인 C2, 별모양 분포인 C3로, 십자가 표시로 나타내었다. 각각은 분산이 1인 가우시안 정규분포로 다음과 같이 C1:(-3, 2), C2:(0, 0), C3:(6,-3) 3개의 중심값을 가지는 데이터를 보여준다. 그림2를 보면, 각 데이터 그룹은 3개로 구분되어 있지만, 그 중 2개의 그룹은 서로의 중심값이 얼마 떨어져 있지않다. 따라서, 다음과 같은 데이터는 아무런 사전정보 없이 수행되는 클러스터링의 결과가 매우 중요하다. 앞에서 살펴보았듯이, 기존의 k-means는 초기값을 랜덤하게 생성하므로, 그 초기값을 통한 클러스링이 제대로 이루어지기가 어렵다. 또한, 매번 클러스터링을 수행할 때마다 그 값이 바뀌므로, 안정한 클러스터링의 중심값을 얻기란 쉽지가 않다. 반면에 본 논문이 제시하는 k-means는 클러스터링이 수행하기 전에 주어진 데이터의 분포 특성을 이용해줌으로써, 안정하고 각 데이터와의 분산이 가장 작은 최적의 중심을 얻을 수 있다. 본 논문에서는 제시한 k-means의 방법이 EM의 결과에 얼마나 큰 영향을 미치는 가를 알아보고자 한다.

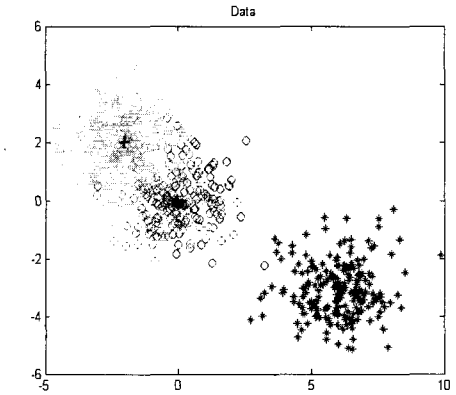

그림 2. 시뮬레이션1에 쓰인 실제 데이터 모델의 분포와 중심값
Fig. 2 Actual data distribution and their center points in simulation 1.

표 1. 시뮬레이션 1: 기존의 방법과 제안된 방법의 K-means 결과 비교.

Table 1. Comparison between the proposed method and the conventional methods in simulation 1

Kmeans	초기 중심값	최종 중심값	반복횟수
기존의 방법 (Random)	-0.1773 1.9747 4.6663 -2.5155 9.8782 -1.8896	-0.9990 0.9673 5.7976 -3.2499 6.9724 -2.6202	5
제안된 방법 (Uniform partition)	-2.1897 2.0942 -0.0081 0.0232 5.9984 -3.0531	-2.0820 2.0410 -0.0183 -0.0125 5.9983 -3.0531	5

EM 알고리즘 상에서 EM을 수행하기 전에 클러스터링을 위한 초기값을 얻기 위해서 k-means를 이용한다. 여기서 얻어진 값들이 EM의 초기값으로 쓰이게 되는데, 이는 향상된 EM의 결과를 얻거나, 알고리즘의 반복 과정을 줄이기 위한 전처리 과정으로 볼 수 있다. K-means을 사용하는 이유는 임의로 선택된 초기치에 따라 EM의 결과가 많은 부분에서 서로 달라지고, 또한, 장시간의 알고리즘 수행시간이 필요하기 때문이다. 그러나, 이러한 목적으로 이용되는 K-means에서도 근본적인 문제점이 존재하고 있다. 일반적으로, EM에서 K-means를 수행할 때에는 그 반복 횟수를 지정한 다음, 그 결과를 초기값으로 이용한다. 여기서, 기존 K-means의 랜덤한 초기값이 충분한 알고리즘 과정을 거치지 않았을 때, K-means의 결과값의 신뢰도는 매우 낮아지게 된다. 또한 EM에 적용되는 초기값은 매번 바뀌게 되므로, 안정된 결과를 얻지 못한다. 반면에, 제안된 방법의 K-means는 기존의 방법보다도 적은 횟수의 반복과정을 거치고, 각 데이터 분포 특성에 맞는 중심값을 찾아주게 된다. 표 1에는 동일한 반복 횟수를 거친 기존의 방법과 제시된 방법에 의한 K-means를 수행한 결과이다. 여기서 최종 중심값이 바로 EM의 초기값으로 쓰이게 된다. 표1에서 제시된 값은 각 클러스터 C1, C2, C3에 따라 차례로 기록된 것이다. 표 1을 통해서도 쉽게 알 수 있듯이, 동일한 반복 수행 과정을 거친 K-means의 결과는 매우 다르다는 것을 알 수 있다. 실제의 데이터 모델과의 각 클러스터의 중심과 비교해 볼 때 기존 방법에 의한 것은 2개의 그룹이 뭉쳐있는 것을 클러스터링 하지 못하고, 랜덤하게 선택된 초기 값에 의한, 5번의 충분하지 못한 알고리즘 수행의 결과이다. 반면에, 동일한 과정을 수행하였지만, 제안된 방법에 의한 클러스터링 결과는 실제 모델과 거의 완벽할 정도의 각 그룹의 중심값을 찾아내었다. 이것은 클러스터링 개수에 따른 균일한 영역분할을 수행한 뒤, 데이터 밀도가 높은 영역에 속한 각 데이터들의 평균값을 초기값으로 선택해 클러스터링을 실행하므로, 짧은 반복과정을 거치고도 좋은 결과를 얻을 수 있다. 이제 위의 결과를 EM의 초기값으로 이용하였을 때의 영향에 대해서 살펴본다.

EM과정을 수행하는 동안, 수렴이 되는 조건을 위해서, 최대 반복 횟수는 100회로 지정하였고, 허용오차는 E의 단계에서 구한, Posterior값을 이용하여, M단계에서 계산된 평균값과 분산값, Prior값은 매 반복 과정을 거치면서 갱신되므로, 이전 반복과정에서의 파라미터값들과 현재 M단계에서 갱신된 파라미터들과의 차이로 지정한다. 본 논문에서 제시하는 모든 시뮬레이션의 허용오차는 0.0001로 설정하여 더 이상의 파라미터 갱신이 없을 정도의 충분한 반복과정을 거치도록 하였다.

그림 3는 표1에서 보여준 기존의 K-means의 결과를 초기값으로 이용한 EM의 반복수행을 거치면서 중심값을 찾는 과정이다. 예상한대로, 초기값이 제대로 얻어지지 않은 알고리즘의 수행은 최대 반복횟수 100번을 넘도록 종료조건을 만족시키지 못한다. 그것은 파라미터들의 초기값을 기반으로 하여 E단계에서 계산되는 Posterior값이 정확하지 않으므로, 이 값을 이용하여 M단계에서 갱신되는 파라미터 값 역시, 제대로 구해 질 수가 없다. 따라서, 처음부터 잘못 얻어진 값을 가지고, 반복적으로 수행하는 EM과정은 악순환을 거듭할 뿐이다. 그림 3에서 볼 수 있듯이, K-means결과로 초기값을 취했으므로, 제대로 클러스터링이 되지 않은 상태에서 각각의 파라미터값들은 오랜 반복 과정을 통해서 갱신되더라도, 실제의 모델과는 상당히 거리가 먼 값들을 얻게된다.

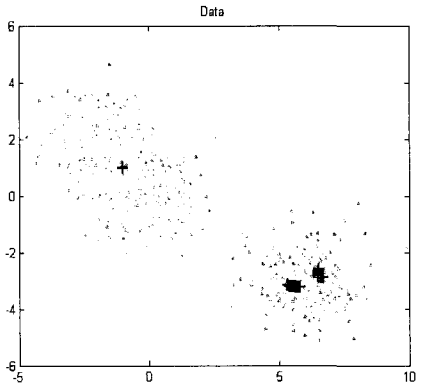

그림 3. 기존 K-means를 이용한 EM의 반복과정

Fig 3 EM Iteration procedures by the conventional K-means.

다음의 그림 4는 EM결과의 확률 밀도를 보여주는 3차원 그림이다. E과정에서 구한 Posterior값을 이용하여 그려준 결과이다. 앞에서 살펴보았듯이, 실제 모델에서는 3개의 그룹으로 분포되었는데, EM결과로는 2개의 클러스터링으로 구분하였다. 비록 지정된 개수만큼의 중심값을 찾았다 하더라도, 그림 4와 같은 확률 밀도 분포를 보면, 2개의 그룹으로 나누어짐을 알 수 있다.

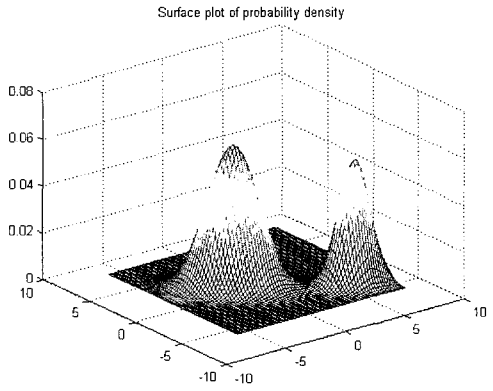

그림 4. 기존 K-means를 이용한 EM의 결과의 3차원 확률 밀도 분포도.

Fig 4 EM results represented in 3-D probability density function by the conventional K-means.

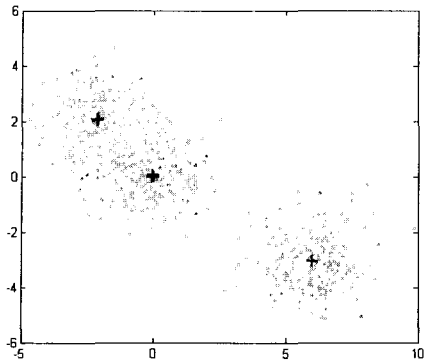

그림 5. 제안된 K-means방법을 이용한 EM 반복과정.
Fig 5. EM iteration procedure by using proposed K-means

반면에, 본 논문에서 제안된 방법에 의한 K-means의 결과를 EM의 초기값으로 이용한 경우는 기존의 방법에서는 얻지 못했던, 실제 데이터 모델과 거의 일치하는 결과값을 얻었다. 그림 5는 균일 영역분할을 수행한 뒤, 얻게된 K-means의 결과를 초기값으로 이용하여 EM을 수행하는 과정이다. 그림 5에서 보듯이, 매번 EM의 반복 과정을 수행하는 경우에도 초기값으로부터 크게 벗어나지 않음을 알 수 있다. 이것은 기존의 파라미터 값들과 갱신된 값들과의 차이가 매우 적기 때문이다. 따라서, 오랜 반복 과정이 불필요하기 때문에 짧은 시간 안에 모든 파라미터 값들을 구하고, 알고리즘을 종료하게 된다. 앞의 경우에는 최대 반복 지정 횟수 100를 초과하지만, 제안된 방법에 의한 EM과정은 단지 18번의 반복과정을 거치게 된다. 뿐만 아니라, 실제 이용한 모델과 거의 완벽한 파라미터 값들을 찾게됨으로써, 수행시간 단축과 정확한 결과해석까지, 두 가지 모두를 만족하게 된다. 제안된 방법에 의한 EM의 확률 밀도 분포를 그림 6에서 보여주고 있다. 그림 5에서 살펴본 EM과정에서 짐작한 바와 같이 3개의 데이터 분포를 제대로 찾는 결과를 보여준다. 이제, 기존 K-means를 이용한 EM과 제안된 K-means를 사용한 EM과의 결과로 얻은 여러 가지 파라미터값을 표와 그림을 통해서, 비교 분석한 결과를 제시한다.

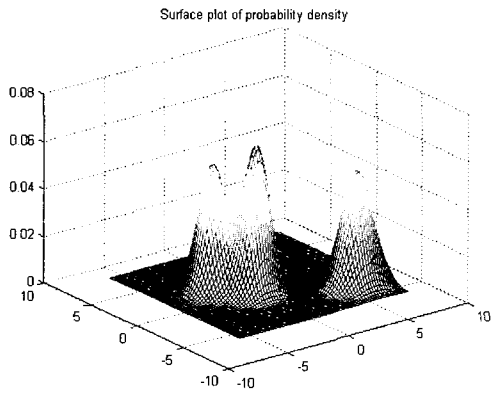

그림 6. 제안된 K-means를 이용한 EM 결과의 3차원 확률 밀도 분포도.
Fig 6 EM results represented in 3-D probability density function by the proposed K-means

우선, 앞장의 EM 알고리즘에서 살펴본 바와 같이 매번 EM과정을 거치면서 계산되는 Log likelihood Error 값을 그림 7에서 기존 방법과 제안된 방법으로 나타내었다. 에러값은 전체 EM 반복 계산량 만큼 얻게되므로, 그림 7을 통해서 전체 반복 횟수 대 에러값의 수렴 정도를 쉽게 파악할 수 있다. 그림 7에서 기존 K-means에 의한 EM의 반복 횟수는 지정해준 최대 반복 횟수를 초과하고, 그 값이 거의 일정하기 때문에 비교를 자세히 보이기 위해서 30회 이상은 나타내지 않았다. 이것은 허용 오차 범위를 매우 낮게 지정하였기 때문에 충분히 수렴하도록 반복 과정을 거치도록 해주었기 때문이다. 제안된 방법에 의한 EM 결과, 초기 에러값도 훨씬 작을 뿐만 아니라, 짧은 반복 횟수를 거치면서 수렴하게 된다. 반면에, 기존 방법에 의한 EM은 초기의 높은 에러 값을 보이면서 최종 반복 회수를 초과하도록 수렴하지 않게 된다. 또한, 전체적인 에러값도 기존의 방법보다 제안된 방법이 매우 낮은 값으로 수렴됨을 알 수 있다.

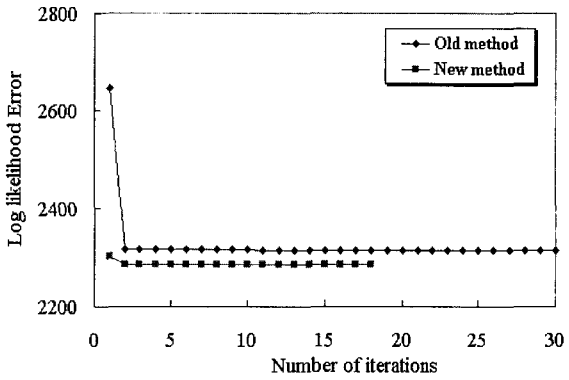

그림 7. 기존 방법과 제안된 방법의 반복 횟수 대 에러 값의 비교.
Fig 7 Comparison between error and the number of iterations.

다음의 표2는 기존의 방법과 제안된 방법에 의한 EM 결과로 얻은 파라미터 값들을 분석하여 표를 통해서 비교한 것이다. 시뮬레이션을 통한 전체 반복 횟수와 각 파라미터 값들 즉, EM 알고리즘 종료후의 평균값과 분산값을 제시하였다. 우선, 표 2에서 제시한 EM의 결과 중에서 기존의 K-means 방법에 의한 EM 결과를 그림 8에서 보여주고 있다. 그림 8에서 보여주고 있는 것은 타원 안의 두 선의 교차점이 각 클러스터의 중심값이고, 두 선의 길이는 표준편차를 나타낸다. 따라서 각 클러스터에 속한 데이터들의 분산 정도를 표준편차를 이용하여 그려줌으로써, 중심값으로부터의 분산정도를 쉽게 알아 볼 수 있다. 표 2에서 제시된 것처럼, 첫 번째 클러스터에서 중심값과 데이터들의 분산정도가 매우 크다는 것을 알 수 있고, 이것은 2개로 나누어져야할 클러스터를 제대로 찾아내지 못했기 때문이다. 그 결과 나머지 두 개의 클러스터는 두 중심값이 비슷한 곳에서 나누어 졌으므로, 그 분산이 작다고 하여도, 제대로 클러스터링이 되지 않음을 파악할 수 있다. 이같은 모든 문제점은 바로, EM의 초기값을 처음부터 잘못 설정한 결과이기 때문에, 그 초기값의 중요성을 다시 한번 생각하게 된다. 반면에 제안된 K-emans를 이용한 EM의 결과를 그림 9에서 보이고 있다. 앞에서 살펴본 바와 같이, 시뮬레이션에 이용한 실제 데이터 모델과 거의 일치하는 파라미터 값을 얻게 된다. 표준 편차를 나타

표 2. 기존의 방법과 제안된 방법의 의한 시뮬레이션1의 결과 비교.
Table 2. Comparison between the proposed and the conventional methods

	반복 횟수	평균값	분산값
기존의 Kmeans를 이용한 EM 결과	100회 초과	-0.9990 0.9673 5.7976 -3.2499 6.5143 -2.6202	2.1359 1.9238 1.22425 0.6690 1.0612 1.1082
제안된 Kemans를 이용한 EM 결과	18회	-2.0820 2.0410 -0.0183 -0.0125 5.9983 -3.0531	0.9763 0.9343 1.0446 0.8490 1.3326 0.8314

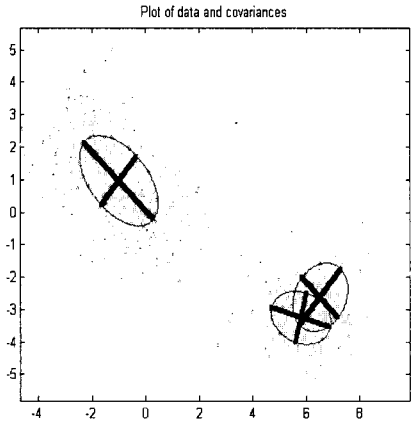

그림 8. 기존 K-means를 이용한 EM 결과.
Fig 8. EM results by the conventional K-menas

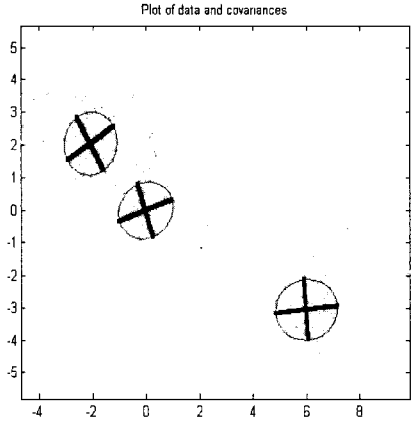

그림 9. 제안된 K-means를 사용한 EM의 결과
Fig 9. EM results by the proposed K-means.

내는 두 직선의 교차점을 각 클러스터의 중심값으로 표현하고, 마찬가지로 각 데이터와의 분선 정도를, 표준편차를 이용하여 원으로 표현하였다. 그림 8에서 보았던 EM의 결과와는 다르게, 각 중심으로부터의 클러스터의 분산 정도가 실제 모델과 비슷하게 나타났다. 이렇게 향상된 결과를 얻고, 또한 전체 반복 횟수를 줄일 수 있는 것은 제대로 설정된 초기값의 영향 때문이다. 동일한 데이터를 가지고, 동일한 EM 과정을 수행한 결과, 서로 다른 초기값의 영향이 위와 같은 많은 차이를 가져온 다는 것으로도 그 중요성을 쉽게 파악할 수 있고, 아울러, 본 논문이 제안한 알고리즘의 우수성을 증명해 보인다.

일반적으로, 데이터 분포의 특성에 따라서 K-means나 EM의 결과가 많이 다르다는 것은 잘 알려진 사실이다. 데이터들의 상관 정도가 매우 높은 경우에는 클러스터링이 그만큼 쉽기 때문에 K-means의 영향이 크게 발휘되지는 않는다. 그러나, 클러스터링이 힘든, 즉 데이터들의 분산이 큰 경우는 상관관계가 작기 때문에 향상된 K-means의 결과가 EM에서도 그대로 적용될 수 있다.

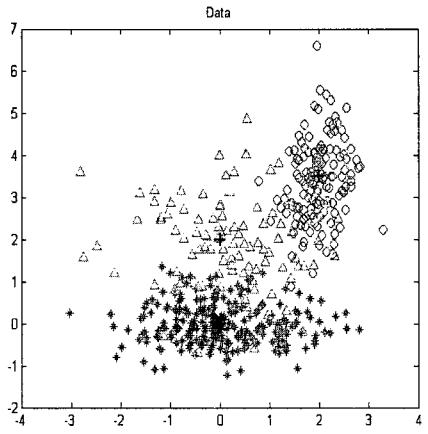

그림 10. 시뮬레이션 2의 실제데이터의 분포와 중심값
Fig 10. Distribution and Center points of actual data in simulation 2.

표 3. 시뮬레이션2에 쓰인 실제모델의 파라미터값
Table 3. Parameters of actual model in simulation2.

	중심값	분산값
실제 모델의 파라미터값	C1 : (0.0 0.0) C2 : (0.0 2.0) C3 : (2.0 3.5)	1.1553 1.1155 0.9739 0.1941 0.1183 0.8236

다음에 보여줄 시뮬레이션2의 경우는 비교적 상관관계가 높은 데이터 분포를 가지는 경우의 K-means와 EM이다. 각 클러스터의 분산정도가 이전에 제시한 경우처럼 가우시안 분포의 일정한 분산이 아닌, 서로 다른 분산을 가진 데이터들이다. K-means의 반복 계산횟수 역시 5회를 지정해 주었고, EM의 수렴 조건도 동일하다. 단지, 시뮬레이션에 이용한 데이터의 특성이 다를 뿐이다. 표 3에서는 시뮬레이션2에 사용된 실제 모델의 중심값과 분산값을 제시해주고 있고, 그림 10에서는 각 클러스터의 분포와 중심값을 보여준다. 이것은 앞에서와 마찬가지로, 두 가지 다른 초기값을 사용한 경우 EM의 결과 비교를 위한 지표로 쓰인다. 표3의 각 클러스터는 아래의 그림에서 별 모양의 분포인 C1, 삼각형의 분포인 C2, 원모양의 분포인 C3로 나타내었고, 각 클러스터의 중심값은 십자가 표시로 나타내었다. 표4에서는 기존의 K-means와 제안된 방법의 K-means를 수행할 경우에, 초기 중심값과 5회 반복과정을 거친 최종 중심값이다. 여기서도, 최종 중심값이 EM의 초기값으로 사용된다. 기존의 K-means 방법처럼, 랜덤하게 초기 중심값을 설정한 뒤, 충

분하지 않은 반복과정을 거친, 중심값은 실제 모델과 많이 다르다. 반면에, 제안된 방법처럼 데이터의 분포특성을 이용한 균일 영역 할당법이 적용된 K-means의 결과는 짧은 반복과정을 수행하였더라도, 실제 모델과 비슷한 결과값을 얻게 된다. 이제 앞에서 제시된 값을 EM의 초기값으로 이용한 결과를 분석해본다. 그림 11은 기존의 방법에 의한 EM과정을 보여주고 있다. 빨간색 십자가 표시가 K-means로부터 얻어진 EM의 초기값을 표현한 것이다. 이처럼, 실제 데이터 모델과 차이가 큰 초기값을 사용했을 때는 많은 반복 과정을 거치게된다. 그림 11에서는 매번 갱신되는 각 클러스터의 중심값을 표시해주었으므로, EM이 수행되는 과정을 보여준다. 아울러 그림 12는 제안된 K-means를 이용한 경우의 EM과정을 나타내었다. 앞에서 보인 것과 비교해보면, K-menas를 통해서 얻은 EM의 초기값으로부터 거의 벗어나지 않은 상태에서 수렴이 된다. 특히 C1과 C3의 두 클러스터는 거의 초기값으로부터 최종 갱신된 값까지 변화가 매우 적다.

다행히도, 시뮬레이션 2에 사용된 데이터들은 서로 상관관계가 비교적 높기 때문에 충분한 EM의 반복과정을 거치면, 어느 정도 수렴이 되는 Local-minimum점에 도달되게 된다. K-means와 마찬가지로, EM에서도 충분한 반복 과정이 없었다면, 그림 11에서 보인 EM 과정은 Local minimum점에

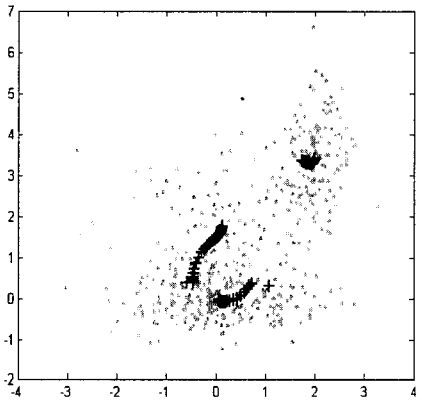

그림 11. 기존 K-means를 사용한 EM의 과정
Fig 11. EM procedures by the conventional K-means

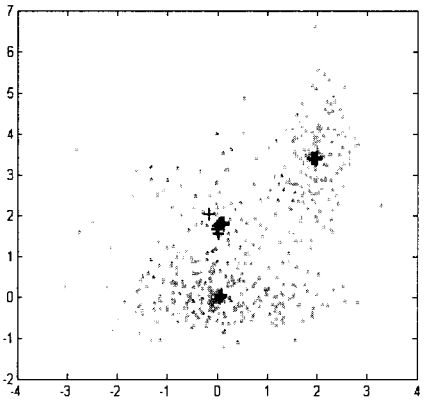

그림 12. 제안된 K-means를 사용한 EM의 과정.
Fig 12. EM procedures by the proposed K-means

표 4. 시뮬레이션 2 : 기존의 방법과 제안된 방법의 K-means 결과 비교.
Table 4. Comparison between the proposed method and the conventional methods in simulation 2.

Kmeans	초기 중심값	최종 중심값	반복횟수
기존방법 (Random)	1.2114 0.4549 -0.0619 -0.0996 -0.0484 2.0427	-0.5856 0.3876 1.0825 0.3071 1.7619 3.3651	5
제안된 방법 (Uniform partition)	0.0007 0.1126 0.1839 2.3192 1.9658 2.9877	0.0861 0.0029 -0.1576 2.0047 1.9543 3.3915	5

도달하지 못하게 된다. 앞의 시뮬레이션1 결과 비교와 마찬가지로, 기존 K-means를 이용한 방법과 제안된 방법을 적용한 EM의 결과를 여러 파라미터 값들을 가지고 비교 분석하여 표와 그림을 통해서 제시한다. 우선, 그림 13은 EM 알고리즘을 수행하면서 계산된 Log-likelihood Error값을 전체 반복 횟수를 가지고 비교하여 나타내었다.

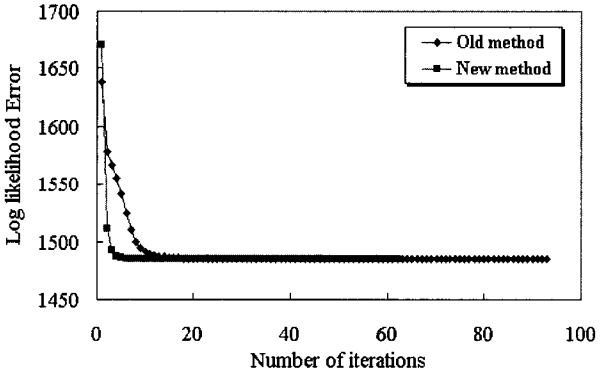

그림 13. 기존 방법과 제안된 방법의 반복 횟수 대 에러 값의 비교.
Fig 13. Comparison of errors between the conventional and the proposed methods

앞의 시뮬레이션1에서 보인 그림 7과 다른 것은, 무엇보다도 이용한 데이터들의 특성이 상관관계가 비교적 높기 때문에 위와 같은 결과를 보이게된다. 기존 방법에 의한 것과 제안된 방법에 의한 EM의 에러 값의 수렴과정을 보면, 제안된 방법이 조금 높은 값으로부터 시작하지만, 두 번째 반복 과정을 거치면서 급격히 줄어들어 기존 방법보다는 훨씬 빠르게 수렴함을 볼 수 있다. 반면에 기존의 방법에 의한 EM은 최종 수렴까지 오랜 반복과정을 거치면서 서서히 수렴됨을 보이고 있다. 그림 7처럼, 더 낮은 에러값으로 수렴되지는 않지만 기존의 방법에 비해 제안된 방법이 적은 반복 횟수를 거치는 것을 알 수 있다. 그림 14는 기존의 K-means를 이용한 전체 93회의 EM 반복 과정을 종료한 결과이고, 그림 15는 제안된 방법에 의한 EM의 결과이다. 시뮬레이션2에서는 사용한 데이터의 상관 관계가 비교적 높은 특성을 지녔기 때문에 앞의 시뮬레이션1과 같이 뚜렷하게 향상된 결과를 얻지는 못하지만, EM이 수렴하는 반복 과정의 횟수를 단축시켰다. 다음에 제시한 표 5는 시뮬레이션2에 의한 EM 결과를 비교한 것이다.

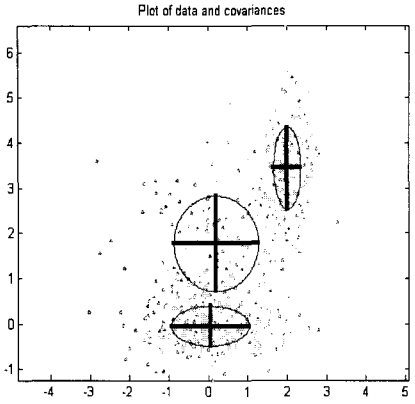

그림 14. 기존 K-means에 의한 EM의 결과
Fig 14. EM results by the conventional K-means

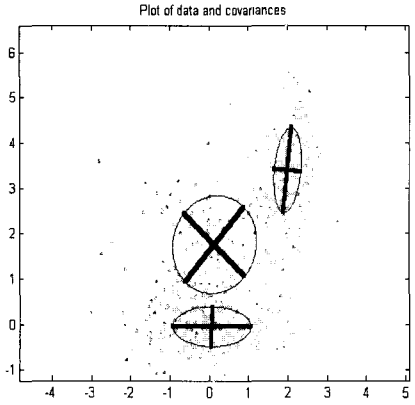

그림 15. 제안된 K-means에 의한 EM의 결과.
Fig 15. EM results by the proposed K-means

표 5. 기존방법과 제안된 방법을 이용한 EM의 결과 비교(평균, 분산, 수렴까지의 반복횟수).
Table 5. Comparison between the proposed and the conventional methods w.r.t. mean, variance, and the number of iterations

	반복 횟수	평균값		분산값	
기존의		0.0621	-0.0495	0.9758	0.1925
Kmeans를	93회	0.1373	1.7577	1.1479	1.1640
이용한EM		1.9923	3.3989	0.1227	0.8748
제안된		0.0620	-0.0492	0.9757	0.1928
Kmeans를	62회	0.1379	1.7605	1.1485	1.1616
이용한EM		1.9923	3.3990	0.1226	0.8748

지금까지 두 시뮬레이션을 통해서 살펴본 것처럼, 본 논문이 제안한 Kmeans를 이용한 EM의 결과는 기존의 방법보다는 향상된 결과를 가져온다. 그것은 EM 알고리즘 자체가 그 초기값에 따라 영향이 지대한 만큼, 당연한 결론이다.

4. 결 론

본 논문에서는 여러 분야에서 널리 이용되고 있는 클러스터링 방법 중의 하나인 EM 알고리즘의 향상된 결과를 얻기 위한 방법으로서 초기치 선정의 새로운 방법을 제시하였다. 아무런 사전 정보 없이, 데이터 모집단을 임의의 그룹으로 분류하는 클러스터링 알고리즘에서 기존의 랜덤하게 초기값을 설정하는 방법을 데이터의 통계적 특성을 이용하는 균등 분할법으로 개선하였다. 임의로 랜덤하게 초기값을 선정하는 방법 대신에 주어진 데이터의 분포 특성을 이용함으로서, 적은 횟수의 반복 과정을 거치면서도 각 클러스터의 중심값을 제대로 찾을 수 있는 시스템을 제안하였다.

이렇게 개선된 초기치 선정 방법이 적용된 새로운 EM 과정은 향상된 결과를 가져왔다. 기존의 초기값으로 얻어진 파라미터들을 대상으로 E-단계와 M-단계를 거치는 동안 갱신되는 값들은 초기값에 상당히 민감하다. 따라서 본 논문에서 제시한 초기치 선정 방법을 이용한 시스템에서는, 클러스터링하고자 하는 데이터의 특성이 잘 반영되어 중심 값으로의 빠른 수렴을 하는 장점을 보여 주고 있다. 이러한 향상된 결과를 시뮬레이션을 통하여 제안된 방법의 우수성을 살펴보았다.

참 고 문 헌

[1] Wong Ching-Chang and Chen Chia-Chong, "K-means-based fuzzy classifier design," *IEEE Fuzzy Systems International Conference,* vol. 1, pp. 48-52, May 2000

[2] K. K. Paliwal and V. Ramasubramainan, "Modified K-means algorithm for vector quantizer design," *IEEE Image Processing Trans,* vol. 9 pp. 1964-1967, Nov 2000

[3] A. P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statisticla Soc., Set. B,* vol. 39, no. 1, pp.1-38, 1977.

[4] For an extensive list of references to papers describing applications of the EM-algorithm, see, *http://www.engineering/usu.edu/Departments/ece/Publications/Moon*

[5] R. Redner and H. F. Walker, "Mixture densities, maximum-likelihood estimation and the EM algorithm (review)," *SIAM Rev.*, vol. 26, no. 2, pp. 195-237, 1984.

[6] S. Zabin and H. Poor, "Efficient estimation of class- A noise parameters via the EM algorithm," *IEEE Trans. Info T.*, vol. 37, no. 1, pp. 60-72, 1991.

[7] H. Chen, R. Perry, and K. Buckley, "Direct and EM-based map sequence estimation with unknown time-varying channels," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 2129-2132, 2001.

[8] R. A. Boyles, "On the convergence of the EM algorithm," *J. Roy. Sta. B.,* vol. 45, no. 1, pp. 47-50, 1983.

[9] C. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11. 1, pp. 95-103, 1983.

[10] R. J. Kozick, B. M. Sadler, "Maximum-likelihood array processing in non-Gaussian noise with Gaussian mixtures," *IEEE Trans. on Signal Processing*, vol. 48, No. 12, pp. 3520-3535, 2000..
[11] H. Takagi and M. Sugeno, "Fuzzy Identification of Systems and Its Application to Modeling and Control," IEEE Trans. on Sys. Man and Cybern., Vol. 15, pp. 116-132, 1985.
[12] J. H. Holland, "Genetic Algorithms in Search, Optimization and Machine Learning," Addison-Wesley, Reading, MA, 1989.

저 자 소 개

김성수(Kim Sung Soo)
1983.2 : 충북대 전기공학과(B.S).
1989.2 : University of Arkansas-Payetteville(M.S)
1997.12 : University of Central Florida (Ph.D)
1998.2～1999.3 : 시스템 공학연구소/전자통신연구원
1999.3～2001.8 우석대학교 전기공학과 조교수.
2001년～현재 충북대학교 전기공학과 조교수.

관심분야 : 퍼지 이론, 신경회로망, 인공지능
Phone : 043-261-2421
E-mail : sungkim@cbucc.chungbuk.ac.kr

강지혜(Kang Jee Hye)
2003년 : 충북대 전기전자컴퓨터공학부과 졸업.
2003년～현재 : 동 대학원 전기학과 석사 과정

관심분야 : 신경회로망, 인공지능, 신호 처리

Phone : 011-9407-1571
E-mail : k23511181@hotmail.com