새로운 고속 EM 알고리즘
(A New Fast EM Algorithm)

김 성 수 † 강 지 혜 ††
(Sung-Soo Kim) (Jee-Hye Kang)

요 약 본 논문은 여러 분야에서 활용될 수 있는 향상된 고속 Expectation-Maximization(FEM) 알고리즘을 제안한다. 첫째, EM의 초기값 설정의 방법으로 많이 사용되고 있는 클러스터링 기법인 K-means의 문제점을 해결하여 개선된 EM의 초기값 선정에 적용하였다. 이것은 기존 K-means 알고리즘에서 임의로 지정하던 랜덤한 초기값 선정을, 데이타 분포 특성을 이용한 균등 분할법을 사용하여 EM의 초기값 문제를 해결하였다. 둘째, EM 과정의 핵심을 이루는 후행 확률(Posterior)의 의미를 부각하여 최대 가능성 후행 확률(Maximum Likelihood Posterior: MLP)과정을 적용하였다.

최종적으로, 본 논문에서 제안한 고속 EM알고리즘(FEM)은 근본적으로 해결하기 못했던 기존의 EM 초기치 선정과 수렴에 대한 문제점을 개선함으로써, EM 알고리즘의 특성을 극대화하는 방향으로 상대적으로 빠른 수렴과 향상된 결과를 가져온다. 제안된 알고리즘의 객관적 타당성을 위해 기존의 방법과 제안된 방법에 의한 시뮬레이션의 결과를 여러 데이타들을 가지고 비교 분석하여 제안한 알고리즘의 우수성을 입증하였다.

키워드 : EM 알고리즘, K-means 클러스터링 알고리즘, 수렴성, 초기 값 선정 문제

Abstract In this paper, a new Fast Expectation-Maximization algorithm(FEM) is proposed. Firstly the K-means algorithm is modified to reduce the number of iterations for finding the initial values that are used as the initial values in EM process. Conventionally the initial values in K-means clustering are chosen randomly, which sometimes forces the process of clustering converge to some undesired center points. Uniform partitioning method is added to the conventional K-means to extract the proper initial points for each clusters. Secondly the effect of posterior probability is emphasized such that the application of Maximum Likelihood Posterior(MLP) yields fast convergence.

The proposed FEM strengthens the characteristics of conventional EM by reinforcing the speed of convergence. The superiority of FEM is demonstrated in experimental results by presenting the improvement results of EM and accelerating the speed of convergence in parameter estimation procedures.

Key words : Expectation-Maximization, K-means, Convergence, Initial Value Problem

1. 서 론

최근에 패턴 인식, 자동제어 및 영상처리 등의 많은 분야에서 응용되는 클러스터링, Expectation-Maximization(EM) 알고리즘은 점차 그 중요성이 더해지면서 그 성능을 향상시키기 위한 많은 이론과 기법이 연구되고 있다[1].

EM 알고리즘에서는 주어진 초기값을 가지고 가능성이 최대인 것으로부터 반복 과정을 통해 파라미터 값을 갱신함으로써 대가함수(cost function)가 최소로 수렴하게 된다[2]. 그러나 초기값에 민감하여 적절한 초기값 선정과 수렴속도의 문제점 등을 수반하고 있으므로, 이것의 해결정도에 따라서 EM의 성능이 좌우된다. 일반적으로, 클러스터링 과정에서 초기값 설정의 민감성은 이미 오래 전부터 인식되었고[3,4], 그 해결 방안의 필요성이 점차 증대되고 있다[5].

본 논문에서는 EM의 초기값 선정 문제에 대한 하나의 해결 방안을 제안함으로써 향상된 성능의 고속 EM(Fast Expectation-Maximization: FEM) 알고리즘을 제시하였다. 일반적인 EM의 초기값 설정은 K-means 알고리즘을 이용하는데, 본 논문에서는 K-means를 기존과 다른 알고리즘을 적용하여 개선하였다.

† 정 회 원 : 충북대학교 전기전자 및 컴퓨터공학부 교수
sungkim@cbucc.chungbuk.ac.kr
†† 비 회 원 : 충북대학교 전기전자 및 컴퓨터공학부
kk2351@nate.com
논문접수 : 2003년 2월 26일
심사완료 : 2004년 6월 29일

우선, 기존의 K-means은 주어진 관측 데이타 중에서 랜덤하게 초기값을 선정하므로 불안정한 수렴의 문제성을 수반한다[6,7]. 또한 충분한 반복 과정을 거치지 않으면 바람직하지 않은 방향으로 수렴하게 된다. 본 논문이 제안하는 K-means의 초기값 선정 문제의 해결 방법은 클러스터링의 대상이 되는 주어진 데이타 분포 특성을 이용하는 것이다. 따라서 새로운 K-means의 알고리즘을 통하여 향상된 EM의 초기값을 얻게 된다. 또한 본 논문이 제안한 FEM 알고리즘의 다른 특성은, EM 과정에서 핵심적인 역할을 하는 후행 확률(Posterior)의 효과를 극대화시킨 것이다. 파라미터 추정 과정을 보면, 확률에 근거하여 다음 추정값을 갱신하는데, 이러한 점은 대가 함수가 최소점에 도달하기 위한 반복의 수렴 여부가 후행 확률에 따라 좌우됨을 보여 준다. 따라서 적절한 초기값 선정과 후행 확률(Posterior)의 효과를 극대화시킨 FEM 알고리즘은 기존의 막대한 반복 계산량을 줄이고, 타당하지 않은 방향으로의 오차가 커지는 것을 막을 뿐만 아니라, 안정되고 향상된 수렴성을 가져온다. 본 논문이 제안한 FEM 알고리즘의 우수성을 보여주기 위하여, 시뮬레이션을 통해 기존의 EM의 성능과 비교 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 일반적인 K-means와 EM 알고리즘[8-11]에 대해서 설명하였고, 3장에서는 새롭게 제안한 K-means와 FEM 알고리즘을 설명하였다. 4장에서는 3장에서 제시된 이론적 정립을 바탕으로 한 실험 결과를 보였다. 특히, 기존의 알고리즘과 제안된 FEM의 시뮬레이션에서, 특성이 다른 여러 종류의 데이타 모델을 가지고 비교 분석하였다. 마지막으로 5장에서는 본 논문의 총체적 결론을 맺고 앞으로 연구되어야 할 방향을 제시하였다.

2. 기존 K-means와 EM의 알고리즘

2.1 K-means 알고리즘

K-means 알고리즘은 다음과 같이 벡터로 표현되는 각 개체 $\mathbf{x} = [x_1, ..., x_n]$ 사이의 거리측정을 Minkowski 거리를 사용하는데, 일반적으로 그 중, 유클리드 거리를 이용한다. 전체적인 알고리즘은 초기화 단계, 개체분산 단계, 새로운 클러스터의 중심단계로 나누어 볼 수 있는데, 각 단계의 역할과 수렴성에 관한 사항을 간략히 알아본다. 우선, 관측된 n 차원 데이타의 전체 데이타의 개수를 N 이라 가정한다.

첫째, 초기화 단계에서는 생성할 클러스터의 개수 K를 정하고, 각 클러스터에 대한 초기값을 설정하는데 특별한 조건 없이 전체 데이타 중에서 식 (1)과 같이 임의로 선택한다.

$$\{z_1, z_2, \ldots, z_k\} \subseteq S_i \quad i = 1, 2, \ldots, N \tag{1}$$

둘째, 둘째, 개체분산 단계에서는 각 개체들과 각 클러스터의 중심과의 유클리디안 거리(J)를 식 (2)와 같이 구하고, 이때 개체들은 계산된 거리가 식 (3)과 같이 가장 최소가 되는 클러스터($C_l, = 1, 2, \ldots, K$)에 속하게 된다. 식 (2)에서 l 와 m은 각각의 클러스터를 의미한다.

$$J_{il} = \|x_i - z_l\|^2 \qquad for\ i=1, 2,...,N\ ,\ l=1,2,...,K \tag{2}$$

$$if\ J_{il} < J_{im}\ for\ l,m = 1, 2, ..., K,\ l \neq m\ then\ x_i \in C_l \tag{3}$$

세 번째 단계에서는, 이전 단계에서 할당된 개체들로 재구성된 K개의 각 클러스터마다 새로운 클러스터의 중심을 계산한다. 이는 각 클러스터에 속한 데이타 개체들의 평균값을 구함으로써, 곧 새로운 중심(4)이 된다.

$$z_l(new) = \frac{1}{N_l} \sum (x_i \in C_l) \quad i = 1,..,N\ ,\ l = 1,..., K \tag{4}$$

여기서, N_l 는 각 클러스터에 새롭게 구성된 총 개체의 수를 나타내고, $x_i \in C_l$ 는 l 번째 클러스터에 속한 개체들을 의미한다. 이러한 클러스터의 중심값 $z_l(new)$이 반복적으로 갱신 되는데 그러한 반복에 대한 횟수와 전체 수렴성에 대한 조건이 최종 알고리즘의 결과를 좌우하게 된다.

K-means 알고리즘의 수렴여부에 관해서는 식 (5)와 같이 더 이상 각 클러스터의 중심에 변화가 생기지 않을 때 종료되는데, 만일 클러스터의 중심에 변화가 생겼다면 두 번째 단계로 피드백(feedback)되어 반복된다.

$$If \quad z_l(now) = z_l(new)\ \ then \quad End \tag{5}$$

위의 방법 이외도 수렴 조건은 앞의 두 번째 단계에서 구한 각 클러스터 중심과 개체들과의 거리계산에서 최소의 값들을 모두 합한 것을 각 반복단계에서의 오차(ε_{iter})로 여길 수 있으므로 식 (6)과 같이 지정된 허용오차 임계치($\varepsilon_{\min}$) 보다 작게 되도록 수렴 조건을 설정할 수도 있다.

$$M_i = \min(J_{il}) \quad l = 1, 2, ..., K \quad ,\ i = 1, 2, ..., N$$

$$\varepsilon_{iter} = \sum_{i=1}^{N} M_i$$

$$If \quad \varepsilon_{iter} \leq \varepsilon_{\min}\ then \quad End \tag{6}$$

2.2 EM 알고리즘

EM(Expectation-maximization) 알고리즘은 최소 대가함수(minimum cost function)를 만족하는 파라미터들의 최대의 가능성(Maximum-likelihood : ML)을 추정한다. 이는 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 파라미

터들은 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.

우선, 이용할 혼합 모델의 확률 밀도 함수는 기저함수의 선형조합으로 표현할 수 있다. M 개의 요소들의 조합으로 이루어진 모델을 식 (7)과 같은 형태로 나타낸다.

$$p(x) = \sum_{j=1}^{M} P(j)\, p(x|j) \tag{7}$$

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

$$\sum_{j=1}^{M} P(j) = 1, \quad 0 \le P(j) \le 1 \tag{8}$$

$$\int p(x|j)dx = 1$$

일반적으로, $P(j)$는 j번째 요소의 Prior(선행) 확률로 여겨진다. 또한, $p(x|j)$은 j 번째 요소에 속하는 경우를 조건으로 하는 확률 밀도함수이므로, Baye's 이론에 의하여 다음의 식 (9)와 같은 Posterior(후행) 확률을 구할 수 있다.

$$P(j|x) = \frac{p(x|j)P(j)}{p(x)} \tag{9}$$

이것도 마찬가지로, 다음의 성질을 만족한다.

$$\sum_{j=1}^{M} P(j|x) = 1, \quad 0 \le P(j|x) \le 1 \tag{10}$$

가우시안 혼합 모델의 경우는, 일반적으로, 최대 가능성(ML)추정은 그 분포에 따라 유도되는 관측 데이타에 근본을 두는 파라미터를 추정하는 방법이다. 그러므로 ML추정에서의 주요 개념은 $x_1, x_2, \ldots, x_L$을 관측할 확률이 가능한 높도록 파라미터를 결정하는 것이다. 따라서, 최대 가능성(ML)의 해를 식 (11)과 같이 Negative log-likelihood를 만족하도록 한다.

$$E = -\ln L = -\sum_{n}^{N} \ln p(\mathrm{x}^n) = -\sum_{n}^{N} \ln \left\{ \sum_{j}^{M} p(\mathrm{x}^n|j)\, P(j) \right\} \tag{11}$$

위 (11)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 변수이고, $x_1, x_2, \cdots, x_N$은 이들 랜덤 변수의 샘플이라 하면, 평균과 분산 그리고 Prior(선행) 확률의 ML 추정값이 다음 식 (12), 식 (13), 식 (14)와 같이 표현될 수 있다.

$$\hat{\mu}_j = \frac{\sum_{n} P(j|\mathrm{x}^n)\mathrm{x}^n}{\sum_{n} P(j|\mathrm{x}^n)} \tag{12}$$

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

$$\hat{P}(j) = \frac{1}{N} \sum_{n}^{N} P(j|\mathrm{x}^n) \tag{14}$$

위의 주어진 식들은 최대 가능성(ML)의 해의 본질을 파악하는데 유용하게 제공되지만, 직접적인 파라미터들의 계산 값을 얻지는 못한다. 사실상 그것들은 높은 차원의 비선형 결합 방정식으로 재해석되는데, 그것은 파라미터들이 전적으로 식 (9)의 우변 항과 같은 결과로 나타나기 때문이다. 그러나 그것은 Error(에러방정식)의 최소 값을 찾기 위한 반복적인 구조를 제시해준다. 우리는 가우시안 혼합 모델의 파라미터를 위한 초기 추정값을 설정해줌으로써 시작한다고 가정하고, 그것을 과거(old)의 파라미터 값이라고 부른다. 그런 후, 각 식 (11), (12), (13)의 우변을 계산할 수 있고, 이것은 에러 방정식을 더 작게 하기 위한 값인 새로운 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.

이러한 EM 알고리즘은 주어진 데이타 $\mathbf{y}$와 앞에서 살펴본 파라미터의 ML 추정값으로 L(Log-likelihood)의 최대화를 통해서 찾는 것으로 요약할 수 있다. 만약 k번째 반복에서의 파라미터 예측 값을 θ^k라고 할 때 전체적으로 다음과 같은 두 단계로 표현된다. 우선 E(Expectation)단계에서는 현재 추정된 파라미터와 관측된 데이타를 조건으로 하여 Posterior(후행) 확률의 기대값을 다음의 식 (15)으로 계산 할 수 있다.

$$Q(\theta|\theta^{[k]}) = E[\log \quad f(x|\theta)|y, \theta^{[k]}] \tag{15}$$

그 다음, 식 (16)과 같이 파라미터 θ가 $Q(\theta|\theta^{[k]})$를 최대화하게 선택한다.

$$\theta^{[K+1]} = \arg \max Q(\theta|\theta^{[k]}) \tag{16}$$

바로, M(maximization) 단계에서는, E 단계에서 구한 값을 이용하여 L(Log-likelihood)의 최대화를 만족하기 위한 새로운 추정값인 파라미터 θ를 갱신한다. 그리고 이러한 두 단계는 수렴조건을 만족할 때까지 반복된다.

M 단계에서의 갱신 값들이 실행되는 방향으로 우세하게 일어난다는 조건으로, 각 반복되는 시점에서의 에러 함수를 작게 하는 것을 보장한 국부 최소점(Local minimum point)이 발견될 때까지 실행된다. 이것은 비선형 최적화 알고리즘의 복잡성을 피한 혼합 파라미터들의 추정을 위한 단순하고 실제적인 방법을 제공한다 [14].

본 논문에서는 수렴성과 관련하여 주어진 관측 데이타와 파라미터 θ값으로부터 구해지는 Posterior(후행) 확률을 가지고 새로운 파라미터가 추정하는 방법을 제안함으로써 수렴의 속도를 가속화하는 알고리즘을 제안하였다. 이에 관련된 사항은 다음 장에서 설명한다.

3. EM에서의 새로운 초기값 설정 방법과 개선된 EM 과정

앞장에서 살펴본 EM알고리즘에서 가장 중요한 것은 파라미터 θ의 초기값이다. 이는 주어진 관측 데이타와 파라미터 θ값으로부터 구해지는 Posterior(후행) 확률을 가지고 새로운 파라미터가 추정되기 때문이다. 이렇듯, 초기 파라미터 값의 설정은 전체 EM 알고리즘 과정에서 막대한 영향력을 발휘한다. 일반적으로 EM의 초기 파라미터 θ값을 설정할 때 널리 쓰는 방법은 K-means 알고리즘이다. 그러나 1장의 서론에서 언급한 바와 같이, 기존의 K-means 방법은 여러 가지 문제점들을 가지고 있다. K-means 알고리즘 역시, 그 초기값에 따라서 결과값이 좌우됨으로 근본적으로 K-means의 초기값 문제부터 해결해야 된다. 관측된 데이타들 중에서 랜덤하게 초기값을 지정하는 경우에는 막대한 반복 횟수와 불안정한 결과 값을 수반하기 때문이다.

본 논문에서는 데이타를 구성하는 임의의 차원의 데이타 공간을 균등 영역 분할하여 클러스터의 초기값을 선정하는 새로운 방법을 제안한다. 제안하는 알고리즘은 임의의 데이타 공간상에 랜덤하게 분포되어 있는 데이타를 적응적으로 균등 분할하는 것을 말한다. 이 방법은 데이타 공간상에 분포되어 있는 각 데이타들의 상대적 위치가 데이타의 회전, 이동 또는 확대, 축소 등에 바뀌지 않는 공간상의 분할을 말한다. 이는 데이타를 구성하는 각 요소들의 상호 위치가 데이타의 회전, 이동, 확대(축소)에 따라 변하지 않는 특성에 기반을 두고 있다. 일반적으로 데이타 공간을 균등 분할하는 경우, 분할된 각 데이타 공간의 단위공간에는 서로 다른 개수의 데이타가 포함되어 있다고 가정할 수 있다. 다시 말해서, 균등 분할에 의해 나누어진 데이타 공간들은 각각의 공간에 속한 데이타 개수를 빈도수로 갖게 되어, 분할된 공간 차원의 데이타 밀도 분포를 이루게 된다. 이러한 일련의 과정을 그림 1의 블록도 나타내었다.

우선, 원하는 클러스터의 수 K가 결정된 후에 클러스터링을 하기 위해서는 각 클러스터의 초기값을 설정해야 한다. 본 논문이 제안한 초기값 설정 방법은 그림 1에 나타낸 순서대로 다음과 같이 크게 두 단계로 나누어 설명할 수 있다. 첫 번째로, 변수 X가 클러스터링 알고리즘에 적용하려는 데이타 집합이라면, 이 데이타

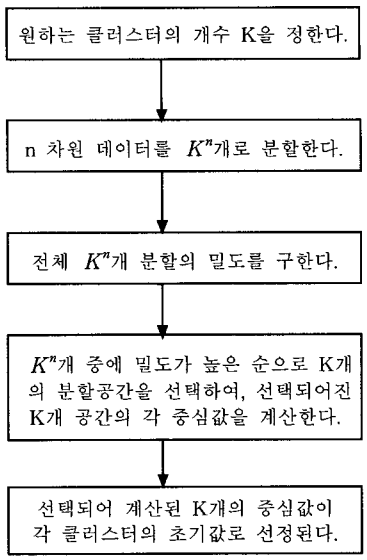


그림 1 제안된 균등 분할법에 대한 알고리즘 순서도

X는 N개의 개체로 이루어져 있다고 가정한다. 각 개체는 임의의 n차원의 데이타라면, 전체 집합 X를 이루는 N개의 각 개체는 $x_i=(x_{i,1},x_{i,2},\ldots,x_{i,n})$, $i=1,2,\ldots,N$으로 나타낼 수 있다.

물론 원하는 클러스터의 개수는 $K\leq N$라 가정한다. 만약, N개의 데이타를 표현하는데 필요치 않은 차원을 사용하는 경우가 발생할 때, 예를 들어 3차원 공간의 두 점은 2차원 평면상에 상관되는 정보를 가지고 충분히 나타낼 수 있으므로, 임의의 n차원의 데이타원소 N개로 이루어진 데이타 공간을 각 차원마다 K개의 클러스터로 균등 분할하여 나타낼 수 있다. 예를 들면, 2차원 평면상에 N개의 점들로 이루어진 데이타 $x_i=(x_{i,1},x_{i,2})$ $i=1,2,\ldots,N$는 $K\times K\,(=K^n)$개로 균등 분할된 임의의 공간상에 속하게 된다. 일반적으로 데이타가 n차원일 경우는 전체 데이타 공간을 K^n개의 n차원 균등분할 된 부분 공간들로 형성된다. 균등 분할을 하였으므로, 각 부분 공간들은 동일한 공간 면적을 가지고, 임의의 순서로 구분 지을 수 있다. 각각 구분되어진 분할된 공간 내에는 속해 있는 데이타의 개수를 가지고 각 부분 공간에서의 데이타 밀도를 구할 수 있다. 데이타 밀도가 높은 순서부터 K개만큼 선택된 부분 공간들이 전체 데이타 공간에서의 밀집도가 높은 순서로 선택되어진 부분 공간이 된다. 이렇게 선택된 K개의 부분 공간은 클러스터링 알고리즘이 시작될 초기값을 선정하기 위한 부분 공간으로 선택된 것이다. 만일 데이타 밀도가 동일한 여러 분할된 부분 공간들이 존재한다면 데이타 분포의 통계적 특성을 나타내는 분산이 작은 단위 데이타 공간을

우선적으로 선택한다. 그 이유는 분산이 작을수록 데이타의 밀집도가 높아지기 때문이다. 최종적인 클러스터링의 목표는 최종 클러스터의 분산값이 가장 작을 때 이루어지므로 정보공간의 밀집도가 높은 지점의 초기값에서 출발한 경우가 임의로 랜덤하게 지정한 초기값에서 시작된 경우보다 향상된 성능을 보이게 된다.

다음은, 두 번째 단계로 원하는 개수 K만큼 선택된 부분 공간상에서 클러스터링 알고리즘의 시작점인 초기값을 구하는 것이다. 이는 앞 단계에서 선택되어진 각각의 부분 공간에서의 중심값을 구하는 것으로써, 데이타 밀집도가 큰 선택된 K개의 부분공간상에 속해진 데이타들의 평균값을 구하면 곧 각 부분 공간에서의 중심값을 얻을 수 있다. 예를 들어, 첫 번째 과정에서 선택된 K개의 분할공간들의 중심값은 데이타가 n차원일 경우 $(C_{i,1}, C_{i,2}, \ldots, C_{i,n})$, $i=1,2,\ldots,K$ 로 나타낸다. 여기서 $C_{i,j}$ 는 임의로 선택된 i 번째 분할공간 내의 j번째 성분의 평균치이다. 따라서 K개의 데이타 밀도가 높은 분할공간으로 선택되어진 각각의 부분 공간에서의 중심값이 전체 클러스터링 알고리즘의 초기값으로 설정되어 알고리즘을 수행하게 된다. 이러한 방법으로 데이타 클러스터링의 초기 값을 선정하는 것은 데이타의 통계적 특성 중 평균치와 유사성을 소유한다는 사실을 근거로 이를 적절히 이용하는 방법이라 볼 수 있다.

이제, 본 논문이 제시한 균일한 영역 분할법을 이용하여 K-means의 초기값을 설정해 줌으로써, 기존보다 단축된 반복 과정을 통해서 최종 클러스터링 결과 값을 얻게 된다. 이러한 타당성 있는 값을 EM의 초기 파라미터 값으로 설정함으로써, 에러 함수를 최소화시키는 국부 최소점(Local minimum point)에 도달하기까지 EM의 반복과정을 감소시키고, 또한 더욱 향상된 결과값을 얻기 위해 제안된 K-means 방법을 EM의 초기값으로 적용한다.

본 논문이 제안하는 고속 EM(Fast Expectation Maximization: FEM)은 적절한 초기값 선정뿐만 아니라 EM의 과정에서 가장 큰 영향력을 발휘하는 후행 확률에도 초점을 맞추어 향상된 결과를 얻기 위한 새로운 알고리즘이다. 기존 EM 알고리즘 과정을 살펴보면 K-means로 구한 초기값을 가지고, E 단계에서 후행 확률(Posterior)을 구한다. 앞에서 언급했듯이 EM 과정에서의 초기 조건은 클러스터의 초기 중심값과 각 요소에 속할 확률 즉, Prior(선행)확률 값을 제시해준다. 이 값들로부터 주어진 데이타와 각 요소에 속할 선행 확률을 가지고, 베이스(Bayes's) 정리에 의해 후행 확률(Posterior)을 계산하게 된다. 그런 뒤, M 단계에서는 후행 확률을 가지고 대수-가능성(Log-likelihood)이 가장 큰 조건을 만족시키는 파라미터 값들을 갱신하게 된다. 이와 같이 후행 확률은 곧 EM 과정에서의 국부 최소점(Local minimum point)에 도달하기 위한 가장 핵심적인 역할을 한다. 2장에서 살펴본 바와 같이 EM 알고리즘에서 얻게 되는 후행 확률(Posterior)은 조건적 확률에 의한 값이다. 다시 말하면 주어진 데이타와 초기값에 의한 선행확률을 조건으로 계산된 확률값이다. 물론 주어진 조건에 의해 각 요소에 속할 후행 확률의 합은 1이 된다. 그런데 기존의 EM알고리즘에서는 조건에 의하여 계산되는 후행 확률값의 영향을 충분히 반영하지 않는다. M 단계에서 갱신되는 파라미터를 계산하는 과정에서 가중치의 역할을 하는 후행 확률 값은 각 요소별로 1보다 작은 가중치를 부여한다. 그 결과 E 단계에서 구한 값을 충분히 반영하지 못함으로써 많은 반복 횟수의 오랜 수렴 과정을 수반하게 된다. 그러한 문제점을 해결하기 위해서 본 논문이 제안하는 것은 주어진 조건을 기반으로 계산된 후행 확률의 값에 최대한의 가중치를 부과하는 것이다.

만약, 클러스터링을 하고자 하는 요소가 $j=1,2,\ldots,M$로 M개가 주어진다면, 각 요소에 속할 선행 확률 $P(j)=1,2,\ldots,M$와 주어진 관측된 데이타 값 $x(i)$, $i=1,2,\ldots,n$으로부터 후행 확률 값 $P(j|x)\,j=1,2,\ldots,M$ 이 식(9)와 같이 계산된다. E 단계에서 구한 후행 확률의 값은 각 요소들의 개수 M만큼 구해진다. 여기서, 현 단계 t에서 계산된 M개의 후행확률 값 중 가장 큰 값이 임의의 k번째 ($k<M$)의 후행 확률값 $P(j=k|x)$이라고 가정해본다. 주어진 전체 데이타와 각 요소에 속할 확률의 모든 조건을 고려할 때 $j=k$번째 요소에 속할 확률이 가장 크다고 계산된 값이다. 따라서 최대 가능성(Maximum-likelihood : ML) 추정에 입각하여 k번째 요소에 속할 후행 확률값에 전체 확률 값 1로 교체한다. 그러므로 나머지 요소들(j)에 속할 후행확률은 일어나지 않을 확률 0으로 교체된다. 다음 식 (30)과 같이 교체된 최대 후행확률의 값을 $P_{\max}(j|x)$로 한다.

$$P_{\max}(j=k|x)=\begin{cases}1 & \text{if } p(j=k),\ \max(\textit{Posterior probabilities})\\ 0 & \textit{elsewhere}\end{cases} \tag{17}$$

위와 같이 다시 교체된 후행 확률 값은 주어진 조건에 의한 최대 가능성을 부여하게 된다. 그런 뒤 M 단계에서 갱신되는 다음 단계 $t+1$의 추정 파라미터들은 최대 가능성의 후행 확률값에 의한 결과가 된다. M 단계에서 얻어지는 값들은 전적으로 E 단계에서 구한 후행 확률에 의해 계산되는 값들이다. 그러므로 최대 후행확률에 의해 계산된 다음 단계의 파라미터 추정값들 역시 최대의 가능성으로 얻어진다.

기존의 EM 알고리즘은 E 단계에서 구한 후행 확률

의 값을 그대로 M 단계에서 갱신될 파라미터 값들의 계산을 위해 사용한다. 만약 현 단계에서 구해진 후행 확률값들 중에서 $j=k$번째 값과 $j=k+1$번째의 값들이 비슷한 가능성의 확률로 주어졌다고 가정한다면, 이로부터 구해지는 M 단계의 갱신된 추정 값은 오히려 비슷한 후행 확률 값에 의해서 제대로 얻어지지 않는다. 즉 오차가 큰 추정 값을 구하게 됨으로써 바람직하지 않은 방향으로 반복 과정을 지속하게 된다.

EM 알고리즘은 오로지 초기 파라미터 값들과 주어진 관측된 데이타만을 가지고 반복 과정을 통하여 최종 추정치를 얻게 된다. 그러므로 초기 조건과 반복 과정 속에 갱신되는 값들의 영향이 매우 중요하다. 특히, 파라미터 추정 알고리즘에서는 그 초기값과 반복 과정에서의 가중치가 전체 추정의 방향을 결정하게 된다. 만약 적절한 초기 파라미터 값들에 의해 알고리즘이 시작되었다고 하여도 반복 과정 속에서 바람직하지 않는 길로

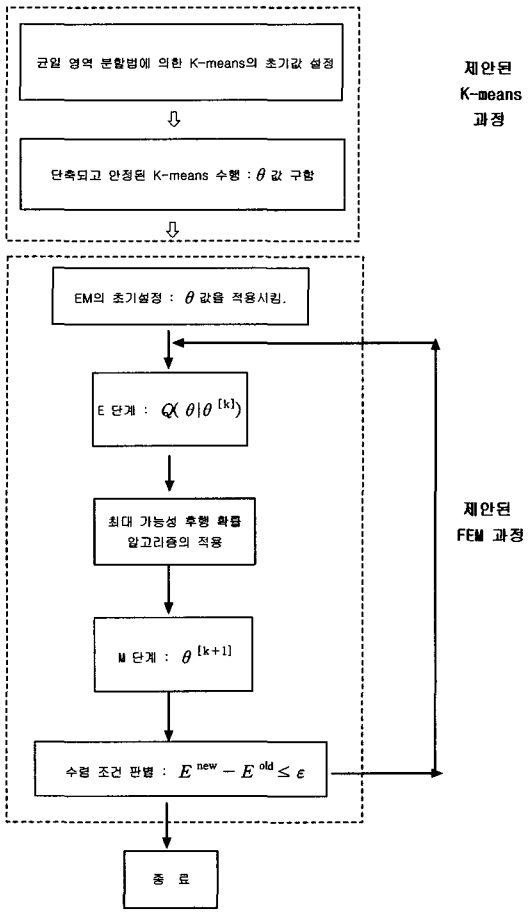


그림 2 제안된 초기값 선정과 EM 과정이 적용된 전체 EM 알고리즘 과정

추정된다면, 결국 국부 최소점으로의 수렴이 어렵기 때문에 최적의 파라미터 추정 값을 얻지 못하게 된다. 따라서 본 논문이 제안하는 것은 데이타 분포 특성을 고려한 적절한 초기값 선정과 최대 가능성의 근거를 바탕으로 한 후행 확률값을 적용한 Fast Expectation-Maximization(FEM) 알고리즘이다. 전체 FEM 알고리즘 과정을 그림 2의 블록도로 나타내었다.

4. 실험 및 고찰

본 실험에서는 서로 다른 특성을 지닌 데이타들을 가지고 제안된 알고리즘을 시뮬레이션을 하였다. 우선, 여러 분야에서 많이 사용되는 Iris 데이타를 기존의 EM 알고리즘과 제안한 FEM 알고리즘으로 비교 분석하였다.

그림 3에서는 시뮬레이션에 사용된 실제 Iris 모델의 데이타 분포와 각 클러스터의 중심값을 나타내고 있다.

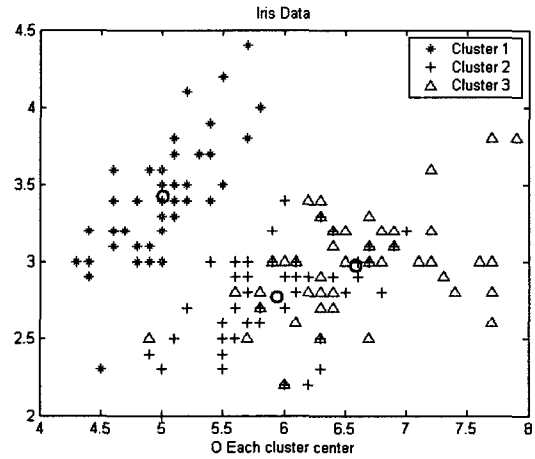
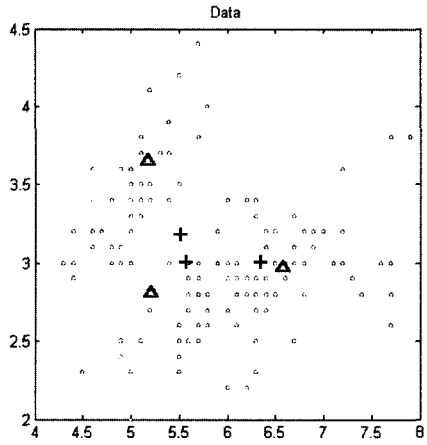


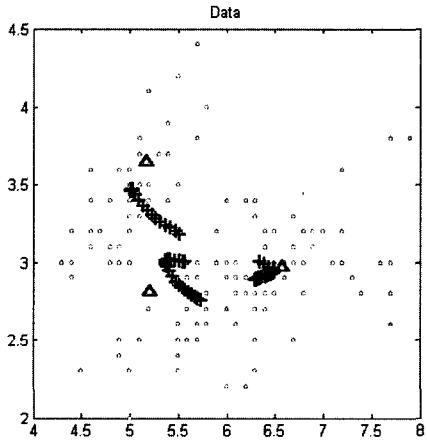
그림 3 시뮬레이션에 사용된 Iris 모델의 데이타 분포와 각 클러스터 중심값

본 논문이 제안하는 알고리즘의 핵심은 적절한 초기값 선정과 빠른 수렴을 통하여 최적의 클러스터의 중심점을 찾는 것이다. 따라서 기존의 알고리즘과의 타당한 비교를 위해서 초기 조건을 구하는 과정과 최대 가능성에 의한 후행 확률(MAP)을 구하는 부분을 제외하고는 동일한 파라미터 추정과정을 수행하였다. 초기 파라미터 값 중 중심값을 제외한 각 요소들에 포함될 선행 확률은 각 요소별로 균일한 확률값으로 설정하였다.

2장에서 살펴본 기존 EM 알고리즘 과정을 살펴보면, 임의로 주어진 초기값과 E 단계에서 구한 후행 확률값에 따라 현 단계의 중심값을 갱신한다. 그림 4(a)는 기존 K-means를 통해서 얻은 초기값을 삼각형으로 표시하고, 이것을 사용한 기존 EM의 추정된 새로운 중심값



(a) 기존 EM의 초기값과 추정된 중심값



(b) 기존 EM의 반복 과정

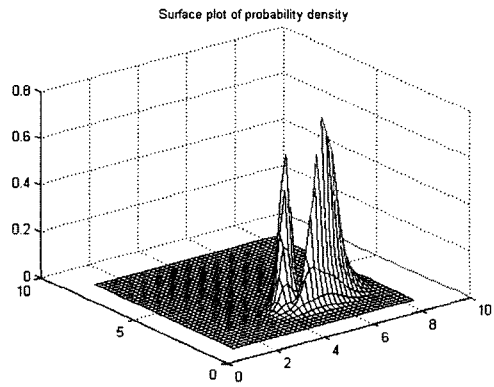


그림 4(c) 기존 EM의 3차원 확률 밀도 분포도

을 십자형 모양으로 나타내었다. 그림 4(b)는 기존 EM 알고리즘이 수행되는 과정을 보여주고 있다. 앞 2장에서 언급한 바와 같이 처음에 갱신된 파라미터 값을 가지고 다음 반복과정에서는 이전 값에서 구한 중심값에서 최종 수렴 조건을 만족할 때까지 EM 과정을 수행하게 된다. 결국 기존 EM 알고리즘은 실험에서 지정해준 최대 반복 회수 100회를 초과하여도 최소 점에 도달하지 못하게 된다.

다음의 그림 4(c)는 EM결과의 후행 확률값을 이용하여 얻은 확률 밀도를 3차원 그림으로 보여주고 있다. 실제 Iris 모델은 3개의 그룹으로 형성되어 있는 반면에, 기존 EM결과는 2개의 클러스터들 형성하고 있다. 비록 세 개의 중심값을 찾았다 하더라도 2개의 그룹으로 나누어짐을 알 수 있다.

다음에는 본 논문에서 제안하는 균등 영역분할방법을 사용한 새로운 K-means를 EM의 초기값으로 적용하고, 최대 가능성 후행 확률을 이용한 FEM알고리즘의 결과를 살펴본다. 그림 5(a)에서는 제안된 K-means에 의한 초기 중심값(삼각형 모양)과 최대-가능성 후행 확률이 적용된 FEM의 반복 과정을 나타내었다. 자세히 살펴보면, 전체 EM 과정이 종료되기 전까지 초기 중심값으로부터 거의 변화가 없는데, 그것은 균일 영역분할법에 의해 구해진 초기 값과, 이 초기값으로부터 계산된 후행 확률의 가능성을 최대로 하는 FEM의 파라미터 추정값의 결과이다. 기존 EM 과정에서는 최대 지정 반복 회수를 초과하여 알고리즘을 종료하였지만, 제안된 방법의 FEM은 단지 6번의 반복 과정을 거친 후 빠른 수렴하였을 뿐만 아니라, 향상된 파라미터 추정을 함으로서, 결과적으로 반복 과정의 단축과 정확한 파라미터 추정을 동시에 만족하고 있다. 제안된 EM결과의 확률 밀도 분포를 그림 5(b)에서 보여주고 있듯이 기존 EM의 결과와 다르게 세 개의 그룹으로 클러스터링 되었다는 것을 알 수 있다.

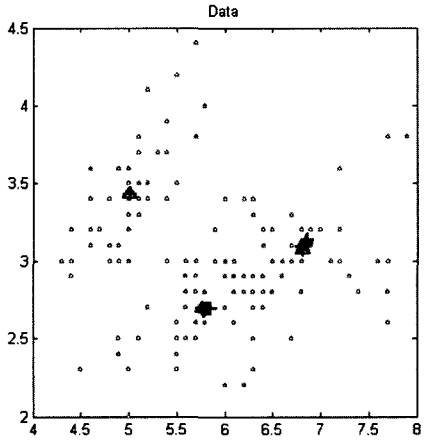
다음은 기존 EM과 제안된 FEM과의 결과를 표와 그림을 통해서, 비교 분석하였다. 우선, 2장의 EM 알고리즘에서 살펴본 바와 같이 매번 EM과정을 거치면서 계산되는 대수 에러 함수값 즉, 식 (11)의 관점에서 기존 EM과 제안된 FEM의 비교를 그림 6에 나타내었다.

그림 6에서 기존 EM의 전체 반복 횟수가 지정해준 최대 반복 횟수(100회)를 초과하고, 그 값의 차이가 크지 않기 때문에 제안된 EM과의 비교를 자세히 보이기 위해서 25회 이상은 나타내지 않았다. 제안된 FEM 결과 전체적인 오차 값이 기존의 EM보다 낮은 뿐만 아니라, 훨씬 적은 반복 횟수를 거치면서 수렴하는 결과를 보여주고 있다. 반면에, 기존 EM의 결과는 초기의 높은 에러 값을 보이면서 최종 반복 회수를 초과하도록 추정 과정을 지속하고 있음을 보여주고 있다.

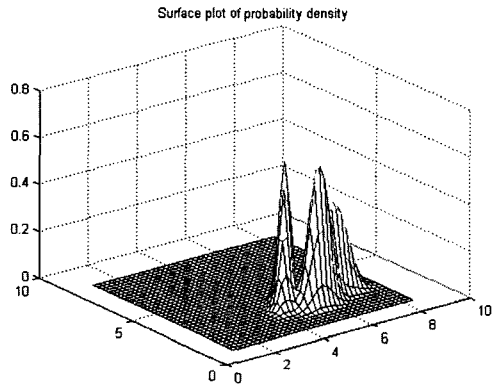
다음 표 1은 기존의 방법과 제안된 EM 결과의 최종 파라미터 값들을 분석하여 비교한 것이다. 기존의 K-

표 1 기존의 방법과 제안된 방법의 의한 시뮬레이션1의 결과 비교

	반복 횟수	초기 중심값	최종 평균중심값	분산값
기존의 EM	100회 초과	5.0130 3.3704 5.8855 2.7200 6.8805 3.0976	5.0158 3.4548 6.1041 2.8781 6.3491 2.8616	0.1195 0.1182 0.4815 0.1294 0.4313 0.0966
제안된 EM	6회	5.0060 3.4280 5.7736 2.6925 6.8128 3.0745	5.0060 3.4280 5.8492 2.6898 6.8561 3.1341	0.1218 0.1408 0.1696 0.0643 0.2186 0.0583



(a) 제안된 FEM 과정의 결과



(b) 제안된 FEM의 3차원 확률 밀도 분포도

그림 5

means와 제안된 K-means를 통한 초기 중심값을 비교하고, 전체 반복 횟수와 각 파라미터 값들 즉, EM 알고리즘 종료 후의 평균값과 분산값을 제시하였다. 제안된 FEM은 반복 횟수, 전체 에러 값뿐만 아니라 최종 파라미터인 평균값과 분산 값 또한 기존의 EM 알고리즘에 비하여 향상된 결과를 가져왔다.

앞에서 살펴본 바와 같이 기존 EM의 문제점은 크게 초기값 선정과 EM 과정에서의 후행 확률로 나누어 볼 수 있는데, 그 중에서 초기 값의 선정문제는 기존의 K-means 방법에서 랜덤하게 임의로 선정하기 때문에

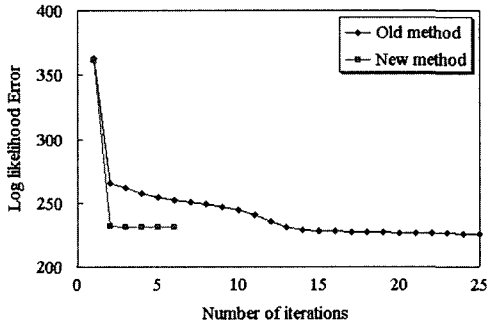


그림 6 기존 EM과 제안된 FEM의 전체 반복 횟수 대 에러 값의 비교

주어진 상황에 따라서 변하는 추정값을 얻게 되는 결과를 가져온다. 그러므로 기존의 K-means를 사용하여 초기 값을 선정한 기존의 EM 알고리즘 역시 그 K-means의 결과로 인해서 여러 가지 서로 다른 결과를 가져온다. 이는 2가지 경우에 대해서 기존의 K-means를 사용한 EM의 결과가 그림 7에서와 같이 서로 다름을 알 수 있다. 그림에서 원 안의 굵은 두 선분의 교차점은 각 클러스터의 평균 중심 값이고, 두 선분의 길이는 표준편차를 나타낸다. 따라서 평균 중심 값으로부터 각 클러스터에 속한 데이타들의 분산 정도를 표준편차를 이용하여 원을 그려줌으로써, 평균 중심 값으로부터의 분산정도를 쉽게 알아 볼 수 있다.

기존의 EM 결과와는 다른 제안된 FEM 알고리즘의 결과를 그림 8에서 보이고 있다. 앞에서 제시한 그림1에서 보면, 실제 Iris 모델은 클러스터들의 데이타들이 서로 뒤섞여 있기 때문에 단순히 2차원 정보에 의한 클러스터링에 어려움이 있다. 그럼에도 불구하고, 본 논문에서 제안한 FEM과정을 통한 파라미터 추정은 실제 데이타 모델과 매우 근접한 결과 값을 얻는 것을 알 수 있다.

본 논문에서는 데이타의 분포 특성을 이용한 균일한 영역분할 방법에 의한 타당성 있는 초기값을 선정하고, EM 과정에서 얻게 되는 후행 확률 값에 최대 가중치를 주어 빠른 수렴속도를 갖게 하였다. 이는 기존의 방법에 비하여 반복 횟수를 매우 감소 시켰으며 향상된 EM의 결과를 얻게 되었다.

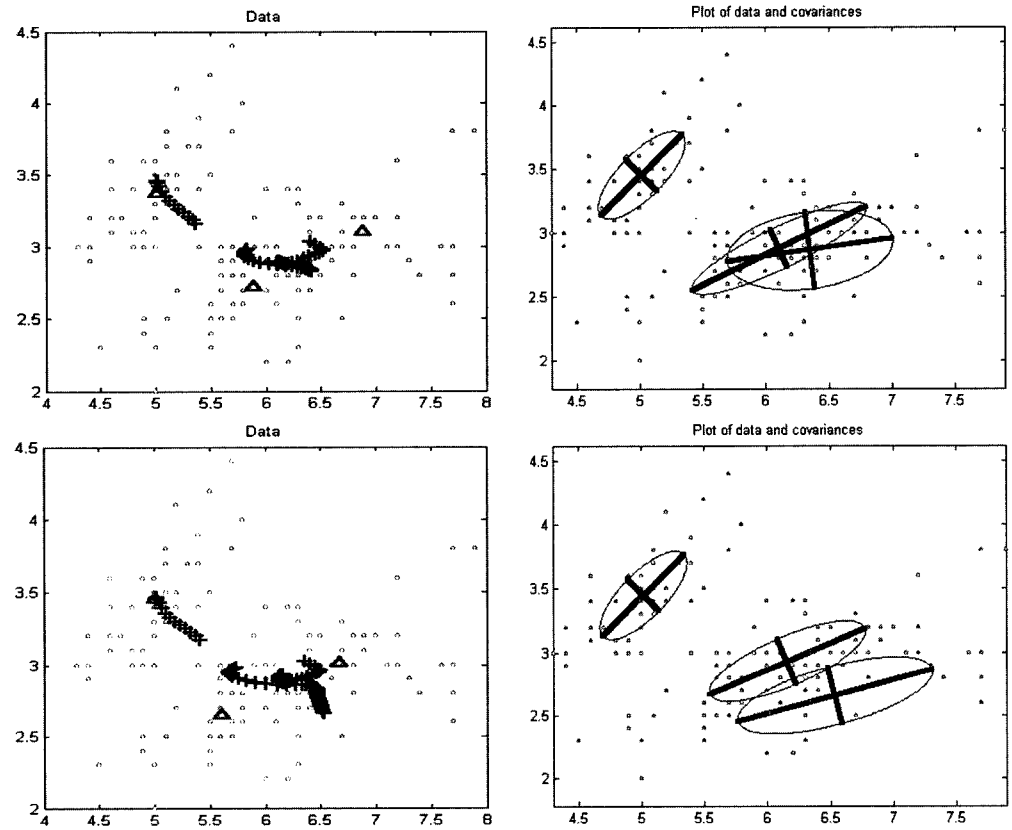


그림 7 기존 K-means를 초기값으로 한 반복과정과 각 클러스터의 평균 중심값과 분산값

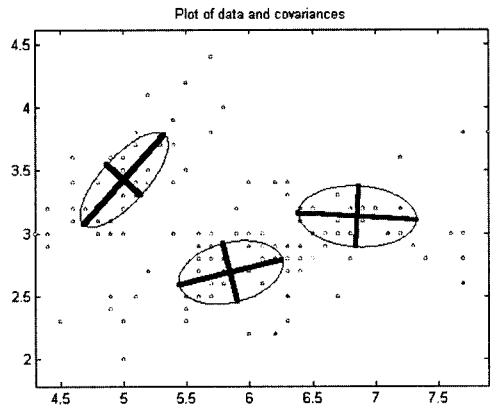


그림 8 제안된 FEM 알고리즘을 통한 평균값과 분산값의 표현

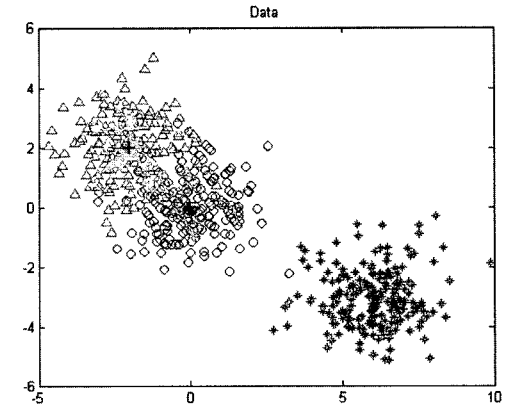


그림 9 시뮬레이션에 쓰인 실제 데이타 모델의 분포와 중심값

다음은 iris 데이타와 다른 특성을 지닌 데이타들에 대하여 제안된 알고리즘의 우수성을 점검해 보았다. 그림 9에서는 시뮬레이션에 사용된 실제 데이타 모델의 분포와 각 클러스터의 중심값을 나타내었다.

각각은 분산이 1인 가우시안 정규분포로 다음과 같이 C1:(-3, 2), C2:(0, 0), C3:(6,-3) 3개의 중심값을 가지는 데이타를 나타내고, 기존 EM의 반복 과정을 그림 10에 보여주었다.

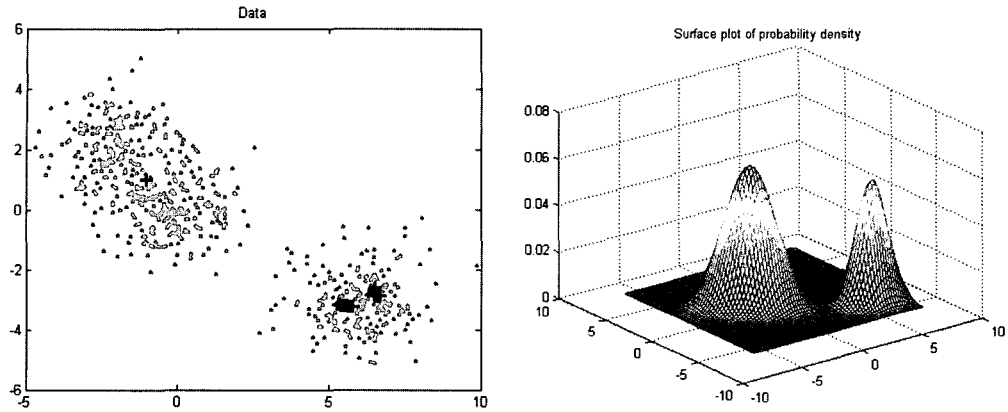


그림 10 (a) 기존 K-means를 초기값과 EM의 반복 과정 (b) 3차원 확률 밀도 분포도

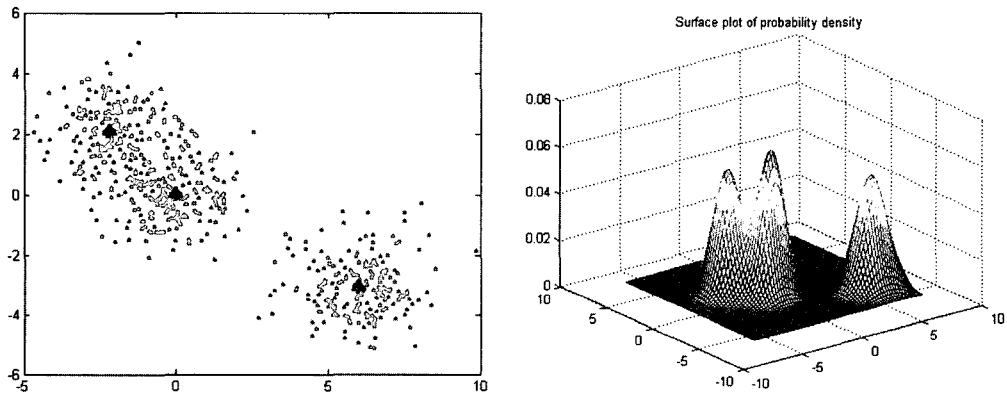


그림 11 (a) 제안된 FEM의 초기값과 반복 과정 (b) 3차원 확률 밀도 분포도

기존 EM에 의한 결과는 클러스터링이 제대로 이루어지지 않은 상태에서 최대 반복 지정 횟수 100번을 초과하더라도 국부 최소점에 도달하지 못하고 그림 10(b)의 3차원 확률 밀도 결과에서도 2개의 그룹으로 나누어짐을 알 수 있다. 반면에 제안된 FEM 알고리즘의 결과는 그림 11(a)을 통해 반복 과정을 살펴보면 초기값으로부터 최종 종료 조건을 만족시킬 때까지 불과 2번의 반복 과정을 거치고, 국부 최소점에 도달하게 된다. 아울러 제안된 FEM의 확률 밀도 분포를 그림 11(b)에서 3개의 데이타 분포를 제대로 찾는 결과를 보여준다.

EM과정을 거치면서 계산되는 에러값(Log likelihood Error)은 전체 EM 반복 계산량 만큼 얻게 되므로, 그림 12를 통해서 전체 반복 횟수에 따른 그 수렴 정도를 쉽게 파악할 수 있다. 제안된 EM이 기존의 방법보다 매우 적은 반복 횟수와 에러값을 갖는 반면 기존의 EM 결과는 최대 지정 반복 회수 100회를 초과하도록 국부 최소점에 도달하지 못한다. 또한 초기 에러값이 매우 큰 것을 볼 수 있는데 이것은 EM 과정에서 후행 확률값에

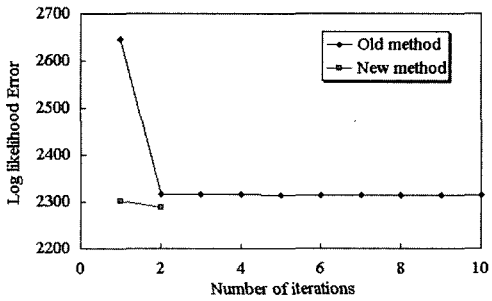


그림 12 기존 방법과 제안된 방법의 반복 횟수 대 에러값의 비교

최대 가능성의 과정을 거치지 않고 가중치의 역할을 제대로 발휘되지 않게 지정된 알고리즘의 결과라 볼 수 있다. 반면 제안된 FEM의 초기 에러값이 작은 것은 개선된 K-means에 의한 초기 중심값이 거의 최종 중심값과 비슷하기 때문이다. 또한 후행 확률에 의한 가중치를 적용한 결과 매우 짧은 반복 횟수로 수렴하게 된다.

표 2 기존의 EM과 제안된 FEM에 의한 결과 비교

	반복 횟수	초기 중심값	최종 평균중심값	분산값
기존의 EM	100회 초과	-0.1773 1.9747 4.6663 -2.5155 9.8782 -1.8896	-0.9990 0.9673 5.7976 -3.2499 6.5143 -2.6202	2.1359 1.9238 1.2242 0.6690 1.0612 1.1082
제안된 EM	2회	-2.1897 2.0942 -0.0081 0.0232 5.9984 -3.0531	-2.0820 2.0410 -0.0183 -0.0125 5.9983 -3.0531	0.9763 0.9343 1.0446 0.8490 1.3326 0.8314

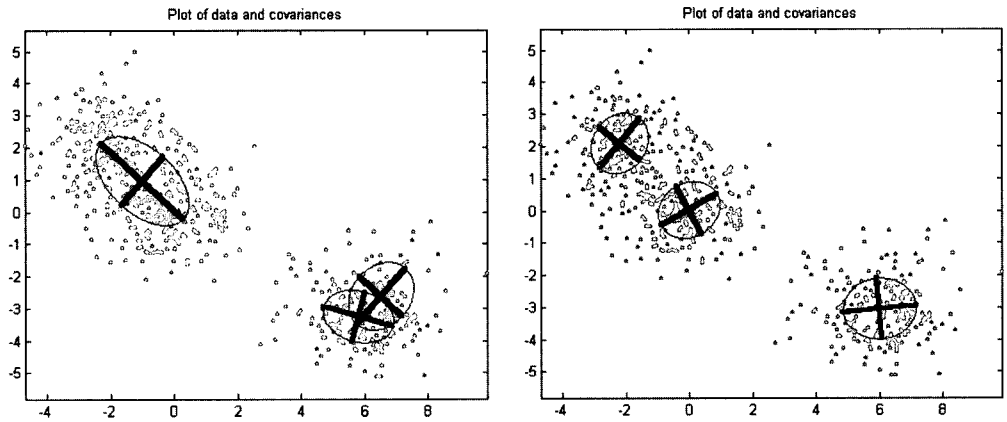


그림 13 (a) 기존 EM의 평균 중심값과 분산값, (b) 제안된 FEM의 평균 중심값과 분산값

다음의 표 2는 기존의 방법과 제안된 방법에 의한 EM 결과로 얻은 파라미터 값들을 분석하여 비교한 것이고, 그림 13를 통해서 그 차이를 보여주고 있다. 기존 EM 과정의 최종 파라미터 값들은 제대로 추정되지 않았고, 반면에 본 논문에서 제안한 FEM 알고리즘의 결과는 시뮬레이션에 이용한 실제 데이타 모델과 거의 일치하는 파라미터 값을 추정한 결과를 알 수 있다. 이러한 향상된 EM 결과를 얻을 수 있는 것은, 제대로 설정된 초기값과 그 초기 조건으로부터 구한 후행 확률(posterior)값에 최대 가능성의 의미를 부여한 결과이다.

일반적으로, 데이타 분포 특성에 따라서 K-means나 EM의 결과가 많이 다르다는 것은 잘 알려진 사실이다. 데이타들의 상관 정도가 매우 높은 경우에는 클러스터링이 그만큼 쉽기 때문에 K-means의 영향이 크게 발휘되지 않는다. 그러나, 클러스터링이 힘든, 즉 데이타들의 분산이 큰 경우는 상관관계가 작기 때문에 그만큼 K-means의 향상된 결과가 EM에서 적용된 영향이 크다.

다음은 비교적 상관관계가 높은 데이타 분포를 가지는 경우로서 각 클러스터의 분산정도가 서로 다른 여러 타입의 데이타들을 가지고 시뮬레이션에 사용해 보았다. 대체로 주어진 데이타들의 상관관계가 비교적 높기 때문에 K-means로 얻어진 초기 EM의 조건이 거의 동일하다고 간주한다면 기존의 EM과 본 논문이 제안한 최대 가능성 후행 확률의 영향을 쉽게 비교해 볼 수 있다. 그림 14, 15에서는 기존 EM과 제안된 FEM의 반복 과정과 최종 결과를 나타내고 있는데, 기존의 경우는 최대로 지정한 반복 횟수를 초과하여도 수렴하지 못하는 반면에 거의 비슷한 초기값으로 제안된 FEM의 결과는 짧은 수렴 속도를 갖는다. 이것은 초기 조건이 동일하다고 간주한다면, 제안된 최대 가능성의 후행 확률이 적용된 EM의 결과가 빠른 수렴 과정을 통해서 더 향상된 파라미터 값으로 추정되었다는 것을 알 수 있다. 비록 데이타들의 상관관계가 높아서 K-means에 의한 초기값 영향이 적다고 하여도 제안한 FEM에서는 후행확률에 의한 빠른 수렴 효과를 얻을 수 있다.

표 3은 여러 가지 분산 타입의 데이타들에 제안된 알고리즘을 적용하여 전체 수렴과정을 기존의 방법과 비교 분석한 결과이다. 시뮬레이션 과정과 결과를 생략하고 수렴과정의 반복 횟수만을 가지고 비교한 것은 상관관계가 높은 데이타의 경우는 충분한 반복 과정을 거친 뒤 국부 최소점에 도달되기 때문에 가장 큰 차이점인 반복 횟수만을 제시한 것이다.

지금까지 여러 가지 시뮬레이션을 통해 기존 EM과 제안된 FEM의 결과에 대해서 비교 분석하였다. 개선된 K-means에 의한 초기값과 최대 가능성 후행 확률을 적용한 FEM의 알고리즘은 기존의 EM과 비교할 때 모두 향상된 결과를 나타내는 우수한 특성을 보였다.

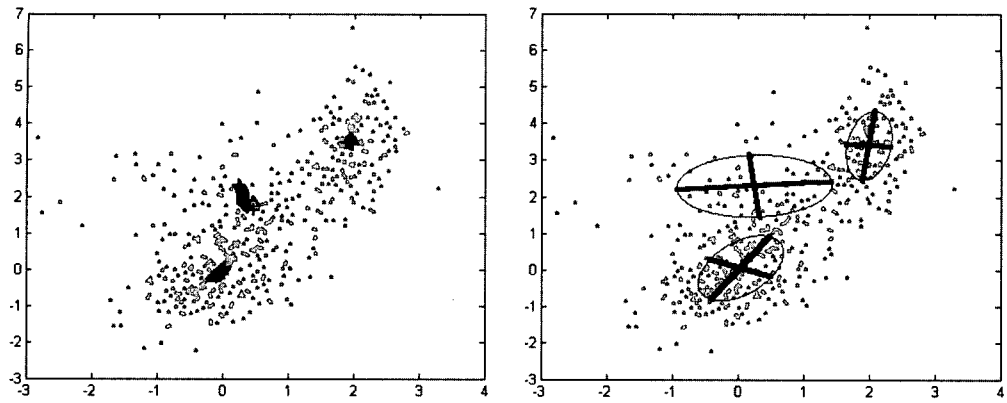


그림 14 (a) 기존 EM의 전체 반복 과정, (b) 최종 평균 중심값과 분산값

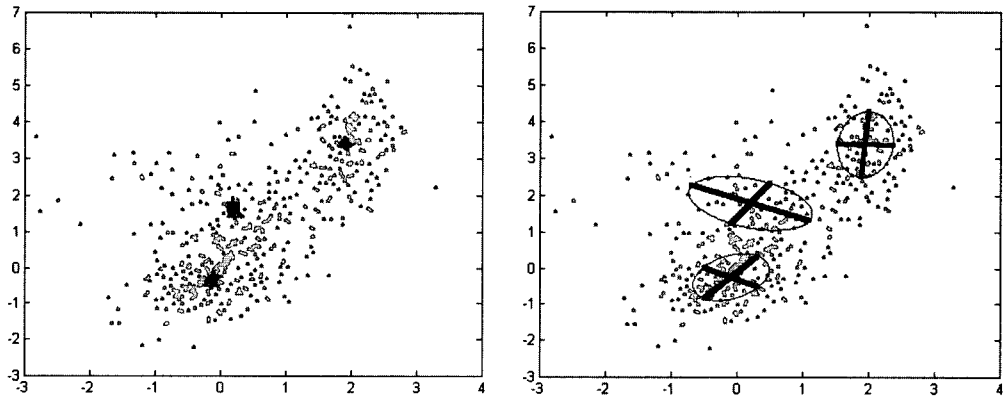


그림 15 (a) 제안된 FEM의 전체 반복 과정, (b) 최종 평균 중심값과 분산값

표 3 데이타 분산 모양에 따른 기존과 제안된 알고리즘의 결과 비교

데이타 모델의 분산 타입	총 반복 횟수	
	기존의 EM	제안된 EM
원모양 (Spherical)	48 회	5 회
수직 방향 (Diagonal)	72 회	9 회
대각선 방향 (Full)	100 회 초과	11 회
원과 대각선방향이 포함(ppca)	15 회	7 회

5. 결 론

본 논문에서는 여러 분야에서 널리 이용되고 있는 클러스터링 방법인 K-means와 파라미터 추정의 EM 알고리즘에서 빠르고 향상된 결과를 얻기 위한 고속 EM (FEM) 알고리즘을 제안하였다. 우선, 아무런 사전 정보 없이 데이타 모집단을 임의의 그룹으로 분류하는 K-means 알고리즘에서 초기값을 랜덤하게 선정하는 방법을 개선하였다. 랜덤한 임의의 초기값 보다는 주어진 데이타의 분포 특성을 이용함으로서, 짧은 반복 과정을 거치면서도 각 클러스터의 중심값과 데이타와의 분산이 가장 작은 최적의 클러스터링 결과를 얻을 수 있는 방법으로 균등 분할법을 초기값 설정에 적용하였다. 이와 같이 개선된 K-means는 클러스터링 뿐만 아니라, 실제로 많이 이용되는 EM에서도 향상된 결과를 가져온다. 이는 EM의 과정에서 초기값의 조건으로부터 구해지는 최종 파라미터의 추정값은 초기값에 민감한 영향을 받기 때문이다.

또한, FEM 알고리즘은 EM의 초기값 개선뿐만 아니라 국부 최소점에 도달하기까지 반복 과정의 횟수를 줄이고 향상된 파라미터 추정을 위한 최대 가능성 후행 확률(posterior)을 EM 과정에 적용시켰다. 이러한 제안된 FEM의 성능은 여러 타입의 데이타 모델에 대해서 빠른 수렴과정과 향상된 파라미터 추정 결과를 실험을 통해서 입증해 보였다.

참 고 문 헌

[1] C. N. Schizas and C. S. Pattichis, "Neural networks, genetic algorithms and the K-means algorithm: in search of data classification," COGANN-92. International Workshop, pp. 201-222, Jun 1992.

[2] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.

[3] 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.

[4] 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.

[5] Ian T. Nabney, NETLAB Algorithms for Pattern Recognition, Springer, 2001.

[6] R.O. Duda and P.E. Hart, Pattern Classification, Willey, 2001.

[7] Su Mu-Chun and Chou Chien-Hsing, "A modified version of the K-means algorithm with a distance based on cluster symmetry," IEEE Pattern Analysis and Machine Intelligence Trans, vol. 23, pp. 674-680, Jun 2001.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

김 성 수

1983년 2월 충북대학교 전기공학과(B.S) 1989년 2월 University of Arkansas-Fayetteville(M.S.). 1997년 12월 University of Central Florida (Ph.D.). 1998년 2월~1999년 3월 시스템공학연구소/전자통신연구원. 1999년 3월~2001년 8월 우석대학교 전기공학과 조교수. 2003년 5월~현재 충북대학교 전기공학과 조교수. 관심분야는 신호처리, 통신이론, 인공지능, 해석학

강 지 혜

2003년 2월 충북대학교 전기전자 컴퓨터공학부 졸업(공학사). 2003년 5월~현재 충북대학교 전기공학과 석사과정. 관심분야는 통계 신호처리, 패턴인식, 인공지능, 디지털 통신