Fuzzy c-means 알고리즘에서의 가변학습 가중치의 효과

박 소 희 , 조 제 황
동신대학교 전기전자공학과

The Effect of Variable Learning Weights in Fuzzy c-means algorithm

So Hee Park , Che Hwang Cho
Dept. of Electrical & Electronic Eng., Dongshin Univ.
hee023@hanmail.net , chcho@dongshinu.ac.kr

요 약

기존의 K-means 알고리즘은 학습벡터가 단일군집에 할당되는 방법이 crisp 이므로 다른 군집에 할당될 확률을 무시하게 된다. 따라서 군집화 작업과 관련하여 반복적인 코드북 설계 과정에서 각 학습벡터를 다중 군집으로 할당하는 Fuzzy c-means를 사용한다. 또한 Fuzzy c-means 알고리즘의 학습과정에서 구해지는 각 클래스의 프로토타입에 가중치를 곱하여 다음 학습의 프로토타입으로 사용함으로써 Fuzzy c-means 알고리즘 적용결과 얻어지는 코드북의 성능을 기존 알고리즘과 비교하여 개선된 Fuzzy c-means 알고리즘을 찾기 위한 근거를 마련한다.

I. 서 론

코드북 설계는 벡터 양자화에 기반을 둔 영상압축에서 가장 중요한 과정으로 군집들의 원형들을 찾는 과정과 유사하다. 하나의 영상을 고정된 크기의 직각 블록으로 분해한 후 추출된 n 차원 벡터들은 n 차원 벡터 공간 안에서 몇 개의 군집(cluster)들을 형성하며 이러한 군집들의 원형이 코드북의 코드벡터들과 동일시된다. 이러한 코드북 설계는 학습벡터와 코드벡터간의 평균 왜곡양을 최소화 하는데 기반을 둔다.

코드북을 설계하는 알고리즘 중에서 가장 대표적인 방법은 K-means(혹은 c-means) 알고리즘으로[1], 이 알고리즘은 주어진 코드북에 대하여 최소거리 조건과 중심 조건을 이용하여 평균 거리 오차가 최소가 되는 코드북을 반복 조건에 따라 연속적으로 생성하는 것이다. 그러나 K-means 알고리즘은 국부적으로 최적화 되고, 그 성능이 초기 코드북에 크게 의존한다는 문제점을 가지므로 이를 보완하기 많은 방법들이 제시되었는데[2]-[4], 그 중에서 spliting 방법이 가장 대표적이다.

K-means 알고리즘과 거의 동일하지만 각 반복과정에서 새로운 코드벡터들을 구하는 방법만 다른 알고리즘을 Jancey가 제안했는데[5], Jancey의 방법은 현재코드벡터와 새로운 군집의 중심점과 일직선상에 있는 반대편의 점, 즉 거리의 가중치(δ)가 2.0인 점을 새로운 코드벡터로 사용하는 것으로 임의의 데이터가 수렴이 되지 않은 경우가 있을 수 있다. 이러한 문제를 보완한 것이 D.Lee가 제안한 Modified K-means 알고리즘으로 거리의 가중치(δ) 에 1.8을 사용하였다[6].

위에서 언급한 코드북 설계 기술들은 각 학습벡터가 어떤 평가기준에 따라 단일 군집으로 할당된다는 의미에서 crisp 결정에 의거하며[7], 각 학습벡터가 다른 군집에도 속할 수 있는 확률을 무시하는 경우가 발생될 수 있다. 이러한 문제를 해결하기 위해 각 군집을 퍼지 집합으로 생각하고, 소속 함수로써 각 학습벡터가 한 군집에 소속될 가능성을 측량하는 Fuzzy c-means 알고리즘을 사용한다[8].

본 논문에서는 코드북 설계 시 Fuzzy c-means 알고리즘을 사용하여 반복학습 시 코드벡터 간 거리의 가중치와 퍼지화 변수를 가변하는 방법을 제안한다. 제안한 방법에서 최적의 코드북을 생성해 내는 가중치와 퍼지화 변수를 실험에 의해 도출하고, 그 결과를 기존의 코드북 생성 알고리즘과 비교 분석한다.

II. K-means와 Fuzzy c-means 알고리즘

n 차원 영역에서 집합 Y를 크기 k의 코드북이라 하면 $Y=\{ y_1, y_2, \cdots, y_k \}$ 이고, M 크기의 학습 집합을 X라 하면 $X=\{ x_1, x_2, \cdots, x_M \}$ 이 되며, 코드북은 M개의 학습벡터를 k군집에 할당함으로써 설계된다. 이러한 코드북 설계의 질은 학습벡터를 군집에 할당하기 위해 사용된 방법과 최소화된 차이의 측정, 그리고 이러한 최소화를 이루기 위해 사용된 최적 방법 등에 의존하며, 코드북 설계의 질은 다음과 같은 평균 왜곡에 의해 측정된다.

$$D = \frac{1}{M} \sum_{i=1}^{M} d_{\min}(x_i) = \frac{1}{M} \sum_{i=1}^{M} \min_{y_i \in Y} d(x_i, y_j) \quad (1)$$

K-means 알고리즘은 최단거리조건에 근거하여 각 학습벡터를 어떤 한 군집에 할당하고, 이 알고리즘에 의해 $d(x_i, y_j) = d_{\min}(x_i) = \min_{y_j \in Y} d(x_i, y_j)$ 이면 학습벡터 x_i는 j번째 군집에 할당된다. 여기서, $d(x_i, y_j)$는 $d(x_i, y_j) = \| x_i - y_j \|^2$로 정의되는 학습벡터 x_i와 코드벡터 y_j와의 유클리드 거리의 제곱이다. 또한 최소거리 조건은 다음과 같이 정의되는 선택함수로 표현될 수 있다.

$$u_j(x_i) = \begin{cases} 1 & d(x_i, y_i) = d_{\min}(x_i) \text{ 일때} \\ 0 & \text{위조건이아닐때} \end{cases} \quad (2)$$

K-means 알고리즘은 최소거리조건에 근거하여 각 학습벡터를 단일 군집에 할당하는데, 이 경우에 crisp 결정에 근거하여 0 아니면 1을 할당한다. 그러나 Fuzzy c-means 알고리즘은 각 학습벡터에 0 과 1 사이의 소속값을 할당하며, 이 소속값은 특정 벡터가 어떤 군집에 어느 정도 소속되는 것으로 간주할 것인가를 가르킨다. 학습벡터의 퍼지 분할은 각 k 군집의 각 벡터의 소속 정도를 규정한다. Fuzzy c-means 알고리즘의 유도는 다음과 같은 목적 함수의 강제적인 최소화에 기초한다. 따라서 코드벡터는 다음과 같이 정의되는 왜곡측정을 최소화 함으로써 얻어진다.

$$J_m = \sum_{j=1}^{k} \sum_{i=1}^{M} u_j(x_i)^{\beta} \| x_i - y_j \|^2 \quad (3)$$

여기서 β는 퍼지화 정도를 나타내는 퍼지화 변수로써, $1 < \beta < \infty$ 범위이다. 주어진 일련의 코드북 벡터에 대하여 $u_j(x_i) \in [0,1] \quad \forall i, j, 0 < \sum_{i=1}^{M} u_j(x_i) < M$ 이고,

$$\sum_{j=1}^{k} u_j(x_i) = 1 \quad \forall i = 1, 2, \cdots, M \quad (4)$$

인 제한 하에서 $J_\beta = J_\beta(u_j, J=1,2,\cdots,k)$을 최소화할 때 다음과 같은 소속함수를 얻는다.

$$u_j(x_i) = \frac{1}{\sum_{l=1}^{k} \left(\frac{d(x_i, y_j)}{d(x_i, y_l)} \right)^{\frac{1}{\beta-1}}} \quad (5)$$

주어진 일련의 소속함수에 대한 코드벡터는 다음과 같이 $J_\beta = J_\beta(y_j, j=1,2,\cdots,k)$을 최소화 함으로써 평가될 수 있다.

$$y_j = \frac{\sum_{i=1}^{M} u_j(x_i)^{\beta} x_i}{\sum_{i=1}^{M} u_j(x_i)^{\beta}} \quad \forall j = 1, 2, \cdots, k \quad (6)$$

식 (6)에서 정의되는 코드벡터 y_j는 유클리드의 중심 혹은 j번째 군집에 할당된 모든 학습벡터의 중심이고, β가 1이고, 선택함수 $u_j(x_i)$가 식 (2)와 같을 경우 기존의 K-means 알고리즘에 의한 중심벡터이다.

III. 가중치와 β의 효과

가중치 가변의 Fuzzy c-means 알고리즘은 식 (6)에 의해서 구해지는 중심벡터를 식 (7)과 같이 가중치 δ에 의해 변경된 새로운 중심벡터를 사용한다.

$$y_j^{n+1} = y_j^n + \delta(C_j^{n+1} - y_j^n) \quad (7)$$

그림 1은 식 (7)을 설명한 것으로 여기서, y_j^n은 n번 반복시 j번째 코드벡터, y_j^{n+1}은 $n+1$번 반복시 j번째 코드벡터, C_j^{n+1}은 $n+1$번 반복시 j번째 코드벡터에 대응되는 중심벡터이다.

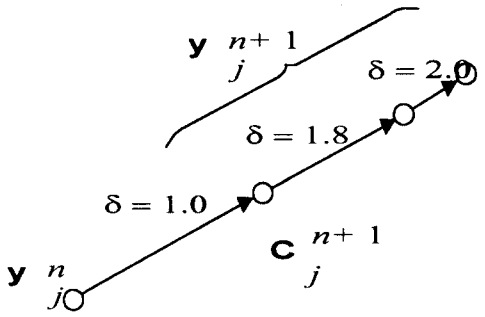

그림 1 코드벡터와 중심벡터를 결정하는 거리의 가중치(δ)

여기서, δ=1인 경우 기존의 K-means 알고리즘을 나타내고, δ=2 인 경우는 Jancey의 방법, δ=1.8 인 경우는 D. Lee의 방법이다.

Fuzzy c-means 알고리즘으로 생성된 군집의 퍼지화는 1보다 큰 변수 β로 제어되며, β가 1에 접근함에 따라 학습벡터의 공간의 분할이 거의 crisp 결정 과정이 된다. 이 변수를 증가시키는 것은 가장 퍼지한 상태로 만들어서 소속 정도를 낮추는 결과를 낳는다.

Ⅳ. 실험 및 결과

본 실험에서는 제안한 알고리즘과 기존 알고리즘을 비교하기 위해 256 그레이 레벨을 갖는 512×512 LENA 영상을 이용한다. 또한 동일한 규격의 20개 영상을 4×4 블록단위로 블록킹한 후 얻어지는 학습벡터를 사용하고, splitting 방법에 의해 얻은 256크기의 코드북을 사용한다. 원 영상과 복원된 영상을 비교 평가하기 위한 PSNR(Peak to Signal Noise Ratio)은 다음과 같다.

$$PSNR = 20 \log_{10}\left(\frac{255}{\sqrt{\frac{1}{512^2}\sum_{i=1}^{512}\sum_{i=1}^{512}(f_{ij}-g_{ij})^2}}\right) \quad (8)$$

여기서 f_{ij}는 원 영상의 화소 값이고, g_{ij}는 복원된 영상의 화소 값이다.

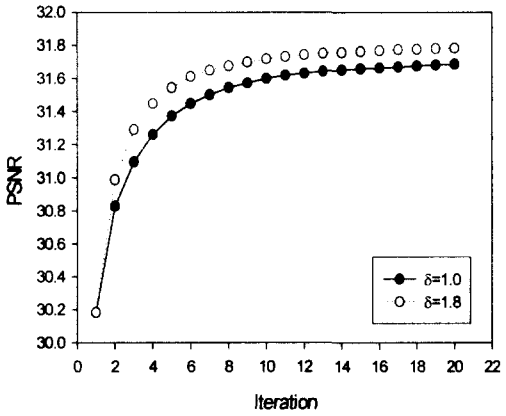

그림 2. K-means 알고리즘

그림 2는 δ가 1.0과 1.8인 경우 반복학습에 대한 PSNR을 각각 나타내며, 그림 3은 반복학습에 대한 코드벡터의 표준편차를 나타낸다. 그림 3에서 알 수 있는 바와 같이 δ가 1.8인 경우 1.0 보다 반복학습의 초기단계에서 표준편차가 크게 나타나지만 후반부로 갈수록 표준편차가 작아짐을 알 수 있다.

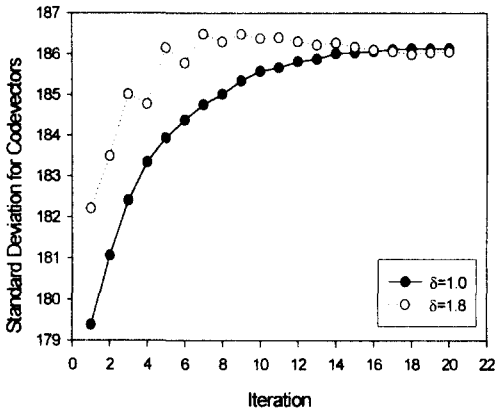

그림 3. K-means 알고리즘

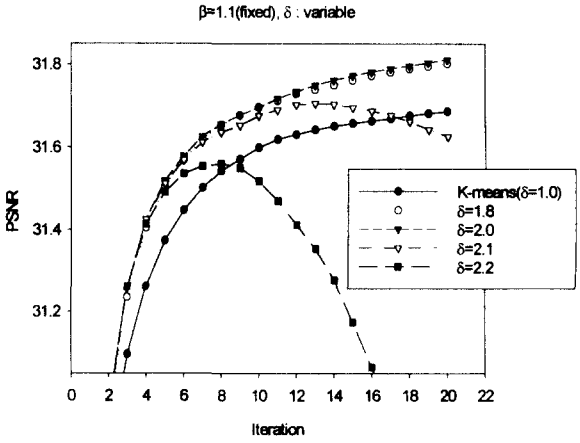

그림 4. 가중치의 변화에 따른 PSNR의 비교

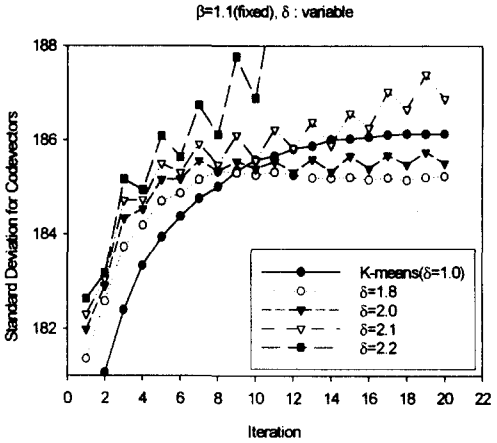

그림 5. 가중치 변화에 따른 코드벡터의 표준편차 비교

그림 4는 기존의 K-means 알고리즘과 Fuzzy c-means 알고리즘에서 δ를 변화시킬 때 반복학습에 대한 PSNR을 나타낸다. 위의 결과 δ가 2.0을 초과할 때 PSNR은 크게 감소하는데, 이것은 학습시 새로운 코드벡터의 발산결과이다.

그림 3에서와 같이 그림 5의 경우에서도 δ가 1.8, 2.0인 경우 반복학습의 초반부에서 코드벡터의 표준편차는 δ가 1인 기존의 K-means보다 크게 나타나지만, 후반부의 반복학습에서는 보다 적게 나나타남을 알 수 있다.

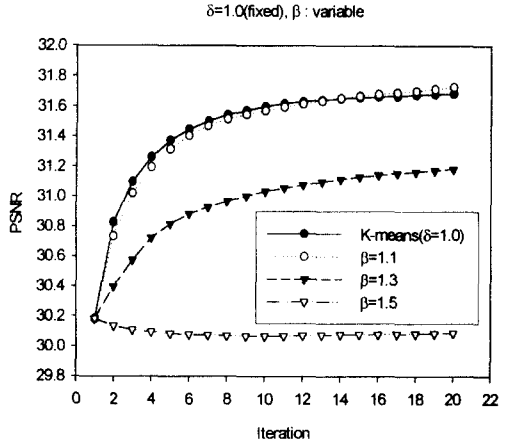

그림 6. β의 변화에 따른 PSNR 비교

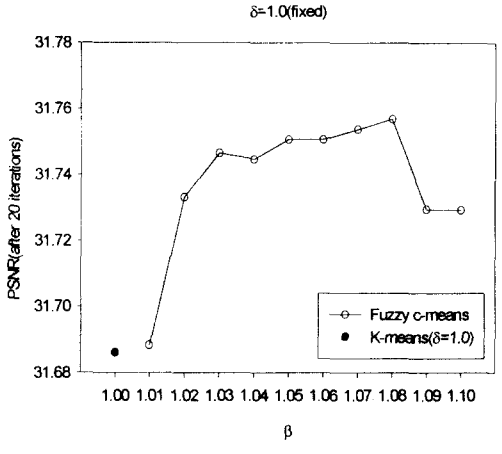

그림 7. Fuzzy c-means와 K-means의 PSNR 비교

그림 6은 δ가 1인 기존의 K-means 알고리즘과 Fuzzy c-means 알고리즘을 비교하기 위하여 δ=1로 하고, β를 각각 1.1, 1.3, 1.5로 한 후, 반복학습에 대한 PSNR을 구한 결과이다. 이 결과에서 알 수 있듯이 β가 1.1에서 Fuzzy c-means의 결과가 가장 우수함을 알 수 있다. 그림 7은 그림 6의 결과에 따라 β값의 변화에 따른 코드북의 성능을 비교하기 위해 β=1.01에서 β=1.10까지 변화해가며, PSNR을 나타낸 것이다.

V. 결 론

기존 K-means 알고리즘의 경우와 같이 학습반복시 중심벡터를 동일선상에서 가중치에 의해 변형된 벡터를 새로운 코드벡터로 결정할 때 보다 우수한 코드북이 설계된 것과 같이 Fuzzy c-means 알고리즘에서도 유사한 결과를 얻을 수 있었다. β가 1.1로 주어질 때 δ는 2.0이하에서 기존 K-means 알고리즘 보다 더 나은 결과를 얻을 수 있었다.

참고문헌

[1] Y.Linde, A.Buzo, and R.M. Gray, "An algorithm for vector quantizer design", IEEE Trans. Commun., vol. COM-28, pp. 84-95, 1980.

[2] W. H. Equitz, "A new vector quantization clustering algorithm", IEEE Trans. Acoust. Speech and signal Proc., pp. 1568-1575, 1989.

[3] I.Katsavounidis, C.C. Jay Kuo, and Z.Zhang, "A new initialization technique for generalized Lloyd Iteration", IEEE Signal Processing Letters, vol. 1, pp. 144-146, 1994.

[4] M.Rabbani and P.W. Jones, *Digital image compression techniques*, SPIE Press, 1991.

[5] M.R. Anderberg, *Cluster analysis for applications*, Academic, New York, 1973.

[6. D.Lee, S.Baek, and K.Sung, "Modified K-means algorithm for vector quantizer design", IEEE Signal Processing Letters, vol. 4, pp. 2-4, 1997.

[7] M. Friedman and A. Kandel, *Introduction to pattern recognition*, World Scientific, 1997.

[8] Nicolaos B. Karayiannis, and Pin-I Pai, "Fuzzy Vector Quantization Algorithm and Their Application in Image Compression", IEEE Trans Image Processing, vol. 4, pp. 1193-1201, 1995.