퍼지 및 지능시스템학회 논문지 2004, Vol. 14, No. 5, pp. 517-524

개선된 밀도 기반의 퍼지 C-Means 알고리즘을 이용한 클러스터 합병

Cluster Merging Using Enhanced Density based Fuzzy C-Means Clustering Algorithm

*한진우, **전성해, *오경환

*Jin-Woo Han, **Sung-Hae Jun, *Kyung-Whan Oh

*서강대학교 컴퓨터학과

**청주대학교 통계학과

*Department of Computer Science Sogang University

**Department of Statistics Cheongju University

요 약

1960년대 퍼지 이론이 소개된 이후 데이터 마이닝을 포함한 기계 학습 분야의 군집화 작업에서 퍼지 이론이 폭넓게 사용되었다. 퍼지 C-평균 알고리즘은 가장 많이 사용되는 퍼지 군집화 알고리즘이다. 이 알고리즘은 하나의 데이터 개체가 서로 다른 소속 정도를 가지고 각 군집에 할당될 수 있도록 한다. 퍼지 C-평균 알고리즘도 K-평균 알고리즘과 같은 일반적인 군집화 알고리즘과 마찬가지로 초기 군집수와 군집 중심의 위치에 의해 최종 군집 결과의 성능 차이가 나타난다. 군집화를 위한 이러한 초기 설정은 주관적이며 이 때문에 적절치 못한 결과를 얻게 될 수도 있다. 본 논문에서는 이 문제를 해결할 수 있는 방법으로 주어진 학습 데이터의 속성을 기반으로 한 초기 군집수와 군집 중심을 결정하는 개선된 밀도 기반의 퍼지 C-평균 알고리즘을 제안하였다. 제안 방법은 격자를 사용하여 초기 군집 중심의 위치와 군집수를 결정하였다. 기존에 많이 이용되었던 객관적인 기계 학습 데이터를 이용하여 제안 알고리즘의 성능비교를 수행하였다.

Abstract

The fuzzy set theory has been wide used in clustering of machine learning with data mining since fuzzy theory has been introduced in 1960s. In particular, fuzzy C-means algorithm is a popular fuzzy clustering algorithm up to date. An element is assigned to any cluster with each membership value using fuzzy C-means algorithm. This algorithm is affected from the location of initial cluster center and the proper cluster size like a general clustering algorithm as K-means algorithm. This setting up for initial clustering is subjective. So, we get improper results according to circumstances. In this paper, we propose a cluster merging using enhanced density based fuzzy C-means clustering algorithm for solving this problem. Our algorithm determines initial cluster size and center using the properties of training data. Proposed algorithm uses grid for deciding initial cluster center and size. For experiments, objective machine learning data are used for performance comparison between our algorithm and others.

Key words : FCM, DBFCM, En-DBFCM.

1. 서 론

Zadeh에 의해 퍼지 집합이 소개된 이후 현재까지 퍼지에 대한 많은 연구가 이루어지고 있다. 기계학습 분야의 군집화 전략에도 퍼지 이론이 널리 사용되고 있다. 특히 Dunn에 의해 제안된 퍼지 C-means(FCM) 알고리즘은 Bezdek에 의해 확장되면서 가장 많이 사용되고 있는 군집화 알고리즘 중의 하나가 되었다[3]. FCM은 하나의 데이터 개체가 서로 다른 소속 정도를 가지고 각 군집에 할당되는 것이 가능하도록 하였다. 이것은 많은 군집화 과정에서 전통적인 crisp 군집화에 비해 더 융통성 있는 것으로 알려져 있다. FCM도 K-means 알고리즘과 같은 전통적인 군집화 알고리즘과 마찬가지로 초기 군집 중심의 개수와 위치에 의해 군집화 결과에 대한 성능에 많은 영향을 받는다. 하지만 군집화를 위한 초기 군집수와 군집 중심의 위치 결정은 주관적이며 이를 효과적으로 결정하는 것은 매우 어려운 작업이다. 초기의 군집수와 군집 위치의 결정은 최종적인 군집결과에 많은 영향을 미치기 때문에 잘못된 초기 결정에 의해 군집 결과의 성능이 좋지 않게 나올 수 있다. 이러한 문제를 해결할 수 있는 방법으로 초기의 군집수와 군집 중심의 위치 결정을 주관적으로 하지 않

접수일자 : 2004년 1월 14일
완료일자 : 2004년 3월 28일

고 주어진 학습 데이터의 개개의 개체들의 분포 속성에 기반하는 전략이 있다. 즉, 학습 데이터의 산점도(scatter plot)을 이용하여 시각적으로 또는 거리 등과 같은 측도를 이용하여 데이터의 분포 특성을 이용하여 군집수와 군집 중심의 위치와 같은 초기 결정을 객관적으로 결정할 수 있게 된다. 이러한 전략에 기반하여 본 논문에서는 개선된 밀도 기반의 FCM 알고리즘을 제안하였다. 제안 방법은 격자(grid)를 사용하여 초기 군집 중심의 위치와 개수를 결정하였다. 즉, 격자 내부의 밀도를 측정함으로서 적절한 초기 군집에 대한 정보를 알 수 있었고, 또한 적절한 군집 합병 기법을 통하여 최적의 군집 결과를 얻을 수 있었다. 2절에서는 일반적인 FCM에 대하여 알아보았고 3절에서는 밀도 기반의 FCM과 제안하는 개선된 밀도 기반의 FCM 알고리즘에 대한 제안 설명과 이 방법을 통하여 개선된 점과 새로운 군집 유효성 측도에 대해 살펴보았다. 4절에서는 기존의 기계 학습 알고리즘의 성능 평가에서 많이 사용되었던 UCI Machine Learning Repository의 객관적인 데이터를 이용하여 제안한 알고리즘의 향상된 성능을 다른 알고리즘과 비교하였다[14]. 마지막으로 결론과 향후 연구 과제에 대하여 5절에서 기술하였다.

2. 관련 연구

2.1 퍼지 C-means 군집화

일반적으로 군집화 알고리즘은 데이터 집합을 분할하여 군집을 형성할 때 한 개의 개체는 반드시 하나의 군집에만 할당하게 된다. 하지만 경우에 따라서는 각 데이터 개체를 단지 한 개의 군집에만 할당하는 것이 적절하지 않을 경우도 있다. 반면에 퍼지 군집화 알고리즘은 데이터 개체들을 서로 겹쳐질 수 있는 집단(overlapping group)으로 나눈다. 따라서 퍼지 군집화 알고리즘이 원래 데이터 집합이 가지고 있는 구조를 파악하는 데 있어서 일반적인 군집화 알고리즘에 비해 더 효과적인 결과를 제공할 수도 있다[5]. 퍼지 군집화 알고리즘에서 데이터 집합, $X=\{x_1,\ldots,x_n\}\subseteq R^p$는 p-차원(dimension) 벡터 공간 R^p의 부분집합이다. 데이터 집합의 각 개체, $x_k=(x_{k_1},\ldots,x_{k_p})\in R^p$는 속성 벡터(feature vector)로 표현된다. x_{k_j}는 x_k의 j번째 속성이다. 군집의 위치는 $v=(v_1,\ldots,v_c)\in R^{cp}$와 같은 군집 중심(cluster center)의 벡터로 나타낸다. v_i는 일반적으로는 데이터 집합 X의 원소와는 관계가 없다. 군집 중심 결정을 위한 초기 분할의 성능 향상을 위하여 주로 사용되는 기준은 분산 기준(variance criterion)이다. 즉, 군집 중심과 데이터 개체와의 부동성을 유클리드 거리로 측정한다. 거리d_{ik}는 다음과 같다.

$$d_{ik}=\sqrt{\sum_{j=1}^{p}(x_{kj}-v_{ij})^2} \tag{1}$$

분산 기준은 다음 식을 최소화하는 것으로 결정된다.

$$\min z(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m\|x_k-v_i\|^p \tag{2}$$

식 (2)에서 v_i는 다음 식과 같이 표현된다.

$$v_i=\frac{1}{\sum_{k=1}^{n}\mu_{ik}}\sum_{k=1}^{n}(\mu_{ik})^m x_k\ ,\ m>1 \tag{3}$$

즉, v_i는 소속 정도에 대해 m의 가중치를 갖는 x_k의 평균이다. 즉 높은 소속성을 가지는 x_k는 v_i에 더 많은 영향을 끼치게 되고, 그 반대의 경우는 더 적은 영향을 주게 된다. 이러한 경향은 퍼지화기(fuzzifier)인 지수 가중치(exponential weight) m에 의해서 강화될 수 있다. 또한 이것은 군집 중심의 계산과 목적 함수 값을 계산할 때 데이터의 잡음(noise) 영향을 감소시키는 역할을 한다. m은 군집 중심으로부터 먼 개체의 영향은 감소시키고, 가까운 개체의 영향은 강화시킨다. 위 기준을 위하여 노름(norm)을 일반화하면 다음과 같다.

$$\|x_k-v_i\|_G^p=(x_k-v_i)^T G(x_k-v_i) \tag{4}$$

따라서 m>1일 때 퍼지 분할 문제는 다음을 만족한다.

$$\min z_m(\tilde{U},v)=\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m\|x_k-v_i\|_G^p \tag{5}$$

위 식에서 $\tilde{U}$는 다음과 같다.

$$\tilde{U}\in M_{fc},\ \ v\in R^{cp} \tag{6}$$

위의 목적 함수에서 $\sum_{i=1}^{G}\mu_{ik}=1$의 조건을 만족시키면서 v_i와 μ_{ik}에 대한 편미분을 통하여 최적 해를 구한다[5].

2.2 군집 유효성

군집 유효성 문제는 군집화 알고리즘에 의해 최종 생성된 분할의 성능(quality)과 직접적인 관계가 있다. 이 문제는 정확한 군집의 수 C를 찾는 문제로 축소될 수 있기 때문에 군집 유효성 문제는 시작 분할(starting partition)의 수를 결정하는 문제와 관계가 있다. 최적의 최종 분할을 얻는 것은 시작 분할 $\tilde{U}^{(0)}$의 초기화에 의존하기 때문에 서로 다른 시작 분할로부터 얻어진 최종 분할의 안정성(stability)은 정확한 군집의 수를 찾는데 사용된다.

2.2.1 분할 상관계수

$\tilde{U}\in M_{fc}$를 n개의 데이터 개체들의 퍼지 c-분할이라 하면 군집 수는 c이고 소속 행렬의 열의 수는 n이다. 이 때 $\tilde{U}$의 분할 상관계수(partition coefficient)는 다음과 같다.

$$F(\tilde{U},c)=PC=\sum_{k=1}^{n}\sum_{i=1}^{c}\frac{(\mu_{ik})^2}{n} \tag{7}$$

분할 상관계수의 값은 $[\frac{1}{c},1]$의 범위를 갖는다. 퍼지 분할에 대한 모든 소속 값이 동일하다면($\mu_{ik}=\frac{1}{c}$) 분할 상관계수의 값도 1/c에 가까워진다. 분할 상관계수의 값이 1/c로 근접할수록 군집은 더 퍼지화 된다. 만약 이 값이 1/c가 되면 주어진 데이터 집합 내부에 군집을 형성하려는 경향이 없거

나, 군집화 알고리즘의 수행이 실패했다는 것을 의미한다.

2.2.2 분할 엔트로피 (Partition Entropy)

$\tilde{U} \in M_{fc}$를 n 데이터 개체들의 퍼지 c-분할이라 할 때, $\tilde{U}$의 분할 엔트로피는 다음과 같다.

$$H(\tilde{U}, c) = PE = -\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}log_e(\mu_{ik}) \qquad (8)$$

분할 엔트로피의 값은 $[0, log_e c]$의 범위를 갖는다. 분할 엔트로피의 값이 0에 가까울수록 군집은 전통적인 군집에 가까워진다. 분할 상관 계수와 마찬가지로 분할 엔트로피의 값이 $log_e c$라는 것은 주어진 데이터에 군집화 할 수 있는 어떠한 구조도 없거나, 군집화가 실패했다는 것을 의미한다.

3. 개선된 밀도 기반의 퍼지 C-Means 군집화 알고리즘

제안하는 개선된 밀도 기반의 퍼지 C-Means(Enhanced density based FCM: En-DBFCM) 알고리즘에서는 초기 군집수와 군집 중심의 위치에 따라 군집 결과의 성능 보장이 어려운 FCM 알고리즘의 문제점을 해결하기 위하여 격자를 사용하였다. DBFCM은 격자 밀도를 사용함으로서 초기 군집수와 위치를 결정하였다[8]. 격자 밀도는 해당 격자가 포함하고 있는 개체 수에 의해 구해진다. 따라서 밀도가 높은 격자는 격자 내부에 많은 개체들이 존재하고, 이 위치에서 군집 중심이 결정될 확률이 높게 된다. 밀도가 높은 격자 내부에서 군집 중심을 결정하므로, 데이터가 가진 본래의 구조에 근접한 군집화를 수행할 수 있다[1, 4].

3.1 밀도 기반의 퍼지 C-Means 군집화

3.1.1 군집화 알고리즘의 수행 과정

DBFCM 알고리즘에서도 핵심 군집화 모듈은 FCM과 같다. 하지만 군집 중심수와 위치 결정은 격자 생성 모듈을 별도로 사용한다. 적절한 군집수 결정을 위하여 군집 합병 모듈도 포함하고 있다. DBFCM의 수행과정은 다음과 같다.

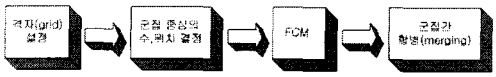

그림 1. DBFCM 알고리즘의 수행과정
Fig. 1. Processing of DBFCM algorithm

3.1.2 격자의 설정

DBFCM은 적절한 군집 중심수와 위치를 결정하기 위하여 격자를 사용한다. 격자는 데이터 개체가 갖는 차원과 동일한 하이퍼큐브이다. 즉, N차원의 속성에 대하여 N차원 하이퍼큐브로 격자를 설정한다. 격자 설정을 위한 격자의 각 변의 크기는 다음 식에 의해 결정된다.

$$Edge_i = k\sigma_i \qquad (9)$$

위 식에서 $Edge_i$는 격자의 각 변의 길이이며, σ_i는 데이터 개체가 가지고 있는 각 속성의 표준편차(standard deviation)이다. k는 상수이며 이 값이 클수록 속성의 편차 정도가 격자 크기에 많은 영향을 미치게 된다. 일반적으로 1이 많이 사용되며 본 논문에서도 1로 하였다. 많은 경우에서 이 값은 경험적(heuristic)으로 결정된다.

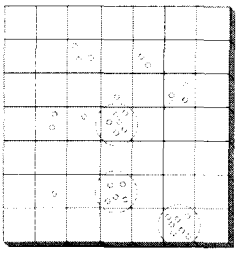

그림 2. DBFCM에 의해 형성된 격자
Fig. 2. Formed grid by DBFCM

각 데이터 개체는 그림 2에서와 같이 배타적으로 단 하나의 격자에만 속하게 된다. 따라서 격자 내부의 개체 수가 증가하면, 격자의 밀도가 증가하게 된다. 즉, 격자의 밀도는 격자 내부의 개체 수에 의해 계산되며, 시작 군집 중심은 일정한 수 이상의 개체를 포함하고 있는 격자 내부에 위치하게 된다. 따라서 시작 군집 중심의 개수와 위치는 밀도가 높은 격자의 수와 위치에 따라 결정된다. 만약 격자 경계상에 존재하는 개체가 있으면 3.2.3절의 재 군집화 과정을 통해 기존의 한 개의 격자로 편입시킨다. 격자는 데이터 공간 전체에 걸쳐 형성되는 것이 아니라, 개체가 위치한 공간에 대해서만 설정되며 격자의 한 변의 길이는 객체의 각 속성에 따라 결정된다. 따라서 격자를 계산하는 데 걸리는 시간은 $O(p \times n)$이다. 여기에서 p는 속성의 수이고, n은 데이터 개체의 수이다.

3.1.3 군집화 과정

DBFCM의 군집화는 기존의 FCM과 동일한 방법으로 수행된다. 군집화 수행시 유사도 행렬과 군집 중심은 다음의 식으로 보정된다.

$$v_i = \frac{1}{\sum_{k=1}^{n}(\mu_{ik})^m}\sum_{k=1}^{n}(\mu_{ik})^m x_k, \ i = 1, ..., c \qquad (10)$$

$$\mu_{ik} = \frac{\left(\frac{1}{\|x_k - v_i\|^2}\right)^{\frac{1}{(m-1)}}}{\sum_{j=1}^{c}\left(\frac{1}{\|x_k - v_j\|^2}\right)^{\frac{1}{(m-1)}}} \qquad (11)$$

3.1.4 군집 합병

원 데이터의 구조 파악은 퍼지 군집화에서 중요하다. 따라서 본 논문에서는 군집 간 유사도를 바탕으로 하는 합병(merging) 기법을 사용하였다. 먼저 최대 군집수를 측정한 후 군집 간 유사도를 평가한다. 만약 군집 간 유사도가 특정한 임계치, $\alpha \in [0,1]$를 넘을 경우 가장 유사한 두 군집 간의 합병을 시도한다. 일반적으로 0.01와 0.03 사이의 작은 값을 사용하지만 이 값이 지나치게 작게 되면 군집 결과가 무의미해 질 수 있다. 이 방법으로 군집화를 수행할 경우 지도(supervised) 퍼지 군집화 방식과는 다르게 군집화 수행 과정에서 특별히 최적화 과정이 필요하지 않게 된다. 하지만 군집 간 합병을 위해 적절한 유사도 임계치의 설정이 필요하

다. 두 퍼지 군집 간의 퍼지 유사도 측도(similarity measure)는 다음과 같이 정의된다[10, 11].

$$S_{ij}=\frac{\sum_{k=1}^{n}\min(\mu_{ik},\mu_{jk})}{\min\left(\sum_{k=1}^{n}\mu_{ik},\ \sum_{k=1}^{n}\mu_{jk}\right)}. \quad (12)$$

3.1.5 군집 유효성 측도

앞에서 설명한 분할 상관 계수와 분할 엔트로피는 단조성으로 인하여 적절한 최종 분할을 얻지 못할 수 있다. 또한 이 측도들은 멤버쉽 값만으로 군집 유효성을 계산하므로 각 개체에 의한 영향은 무시된다. 이러한 단점을 해결하기 위한 여러 가지 측도들이 제시되었다. 첫 번째로 Xie-Beni 인덱스(XB)는 군집 내부의 응집성(compactness)과 군집 간의 분리성(separation)에 대한 유효성 측도이다[9][12]. 군집 i내부에 위치한 개체 x_j에 대한 퍼지 편차(deviation) $\hat{d}_{ij}$는 다음의 식으로 정의된다.

$$\hat{d}_{ik}=\mu_{ik}\|x_k-v_i\| \quad (13)$$

군집 i에 대한 퍼지 편차의 제곱의 합은 데이터 객체에 대한 퍼지 분산(variation)이라 하고, σ_i로 나타낸다. 모든 군집의 분산 합 σ는 데이터 집합의 전체 분산이라 하고 군집 i의 응집성은 $\pi=(\sigma_i/n_i)$로 나타낸다. 여기서 n_i는 군집 i에 속한 데이터 개체의 수이고, π는 군집의 평균 분산이다. 퍼지 분할의 분리성은 다음과 같이 군집 중심 간의 최소 거리로 정의된다.

$$d_{\min}=min\|v_i-v_j\| \quad (14)$$

따라서 XB는 다음과 같이 정의된다.

$$XB=\frac{\pi}{n\cdot d_{\min}}=\frac{\sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ik}^{m}|x_k-v_i|^2}{n\cdot \min_{i,j}|v_i-v_j|^2} \quad (15)$$

XB의 값이 작을수록 뭉쳐 있고, 서로 잘 분리된 군집이다. 그러나 XB는 군집의 수가 매우 큰 값, 즉 n에 가까운 값을 가질 때 단조 감소하는 경향이 있다. 이러한 경향을 줄이기 위해서 단조성이 나타나는 군집의 수 $c_{\max}$를 결정하고 XB의 최소값을 $[2, c_{\max}]$의 범위에서 찾아야 한다. 또한 XB는 퍼지화기 m에 의존하므로, m이 증가함에 따라서 XB의 값도 증가하게 된다. 두 번째로 Fukuyama-sugano(FS)는 XB와 같은 범주에 속하는 측도로서 다음과 같이 정의된다[9, 13].

$$FS_m=\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}\left(|x_k-v_i|^2_G-|v_i-v|^{2c}\right) \quad (16)$$

위 식에서 v는 평균 벡터이고, G는 양정치(positive-definite) 행렬이다. FS_m이 작은 값을 가질수록 응집되고 서로 잘 분리된 군집 결과를 얻을 수 있다. 첫 번째 항은 군집의 응집성이고 두 번째는 군집 중심들의 거리 측도이다. 세 번째로 분할 분리도(PS)는 정규화된 분할 상관계수와 각 군집에 대한 지수적 분리도 측도를 사용한다. PS는 다음과 같이 정의된다[7].

$$PS_i=\sum_{k=1}^{n}\mu_{ik}^2/\mu_M-\exp\left(-\min\left\{|v_i-v_j|^2/\beta_T\right\}\right) \quad (19)$$

위 식에서 μ_M, β_T, 그리고 PS는 각각 다음 식과 같다.

$$\mu_M=\max\left\{\sum_{k=1}^{n}\mu_{ik}^2\right\},\ \beta_T=\frac{\sum_{i=1}^{c}|v_i-v|^2}{c},\ PS=\sum_{i=1}^{c}PS_i \quad (20)$$

분할 분리도는 정규화된 분할 상관계수와 지수적 분리도를 이용하여 구한다. PS_i가 클수록 응집되고 잘 분리된 군집을 얻을 수 있다.

3.2 개선된 밀도 기반의 퍼지 C-Means 알고리즘

본 논문에서 제안하는 En-DBFCM 알고리즘은 각 속성의 산포 척도(scattered degree)를 각 속성에 대한 가중치로 사용하였다. 즉 입력 벡터의 각 속성 중에서 학습 데이터로부터 계산된 분산(variance)인 흩어짐의 정도가 클수록 이 속성이 분석 결과에 많은 영향을 미치기 때문에 이들 속성들에 대하여 상대적으로 큰 가중치를 주었다. 다음 식은 가중치 w와 산포 척도 v와의 관계를 나타내고 있다.

$$w_i=cv_i\ (i=1,\cdots p) \quad (21)$$

위 식은 i번째 속성의 가중치를 나타낸 것이고 p는 입력 벡터의 차원이다. c의 값을 조절하여 산포 정도를 가중치에 어느 정도로 할지 결정할 수 있다. 본 논문에서는 c의 값을 1로 하였다. 이는 산포 척도값을 그대로 가중치로 사용한 것이다. 일반적으로 산포 척도 값이 크면 이 값을 작게 하고 산포 척도가 작으면 이 값을 작게 한다. En-DBFCM에서는 가중치가 높은 속성에 대해서만 군집화를 위한 격자를 설정한다. 제안 알고리즘은 속성에 대한 가중치를 사용하여 이전의 DBFCM에 비해서 향상된 군집 결과를 얻을 수 있었다. 또한 DBFCM에 비해 격자 설정에 필요한 계산 시간도 단축시켰다.

3.2.1 산포 척도

본 논문의 En-DBFCM에서 사용되는 산포 척도는 다음가 같다[6].

$$V_i=\frac{(F_{\max}-F_{\min})_i}{\sigma_i},\ i=1,\ldots,p \quad (22)$$

위 식에서 $F_{\max}$는 i번째 속성에서의 최대값, $F_{\min}$은 최소값을 나타낸다. σ_i는 i번째 속성 값의 표준 편차(standard deviation)이고, p는 속성의 크기이다. 제안 알고리즘에서는 데이터 개체가 갖는 각 속성의 산포 척도를 데이터 개체와 군집 중심 간의 거리를 계산할 때 가중치로 사용하였다. 즉, 어떤 속성이 높은 산포 척도값을 갖게 되면, 이 속성은 높은 분할 우선순위를 얻게 된다. 따라서 이 속성은 다른 속성에 비해서 보다 큰 가중치를 갖게 된다.

3.2.2 격자의 설정

격자 설정은 산포 척도에 의한 가중치 정보를 이용하여 선택된 속성들을 사용하였다. 최종적인 격자의 구조는 주어

진 전체 데이터의 속성 크기보다 낮은 차원의 하이퍼 큐브(hyper cube)가 된다. 산포 측도를 정규화한 후에 특정 임계치(threshold) 이상의 값을 갖는 속성들만으로 격자를 만들게 된다. 임계치의 결정은 주어진 데이터 분포에 따라서 경험적(heuristic)으로 결정된다. En-DBFCM에서 격자 설정에 소용되는 시간은 $O(\frac{k}{p} \cdot n)$이다. 여기서 k는 선택된 속성 크기이고, p는 전체 속성 크기이다.

3.2.3 재 군집화

최종 군집이 형성된 후에도 여러 군집들 사이에 중복해서 속해있는 개체들에 대하여 가장 유사한 한 개의 군집에 할당하기 위하여 En-DBFCM에서는 재 군집화(re-clustering) 방법을 사용하였다. 최종 군집화 이후 각 데이터 개체는 모든 군집들에 대하여 유사도를 나타내는 소속 함수값을 갖는다. 이 값을 이용하여 데이터 개체의 각 군집에 대한 소속 함수 값들 중에서 가장 큰 값과 두 번째 큰 값의 차이가 임계치보다 작을 경우, 이러한 특성을 갖는 데이터 개체들을 모두 제거한 후, 남은 데이터 개체만으로 또 한번 군집화를 수행한다. 재 군집화 과정이 끝나면 형성된 군집에 대해 제거했던 개체들의 소속 함수값을 다시 계산한다. 이렇게 계산된 소속 함수값으로 이들 개체들에 대한 군집 할당을 마지막으로 한다.

4. 실험의 설계 및 결과

4.1 실험 설계

4.1.2 군집화 알고리즘의 구현

본 논문의 제안 시스템은 격자를 설정하고 밀도를 구하는 모듈과 FCM 모듈, 그리고 유사도 기반의 군집 합병 모듈로 구성되어 있다. 각 모듈은 java를 이용하여 직접 구현하였다.

4.1.2 데이터 집합

제안된 En-DBFCM의 성능 평가를 위하여 UCI Machine Learning Repository로부터 3개의 학습 데이터를 사용하였다. 아래의 표는 이들의 일반적인 속성을 나타내고 있다.

표 1. 학습 데이터의 일반적인 속성
Table 1. General properties of training data

	IRIS	WDBC	BREAST
속성수	4	30	9
전체 패턴수	150	569	699
목적 속성수	3	2	2

4.1.3 데이터 전처리

실험을 위하여 사전에 의미 없는 속성을 제거하고 정규화(normalization)를 수행하였다. 즉, 한 속성이 다른 속성들의 연관성에 의한 상대적인 변이성(variability)에 심하게 의존하기 때문에 각 속성의 영향력을 공정하게 반영하기 위해서 정규화와 의미 없는 중복 속성 제거를 수행하였다.

4.1.4 군집화 평가 기준

군집 유효성과 오분류율을 사용하여 제안 알고리즘의 성능 향상을 측정하였다. 각 알고리즘의 수행 시간에 대한 비교를 통해서 제안 알고리즘의 시간 단축 효과도 확인하였다. 군집 유효성 측도는 FS와 XB의 두 가지 인덱스를 사용하였다.

4.2 실험 결과

4.2.1 군집 유효성 비교

IRIS, WDBC, BREAST 데이터의 최종 군집 결과에 대한 FS와 XB 측도를 사용하여 각 알고리즘의 성능을 평가하였다. XB 값은 민감하여 값의 변화폭이 매우 크게 나타나서 그래프에서의 XB 값은 log 함수를 통한 변환 값으로 나타내었다. 아래 그림은 각 데이터의 군집 합병에 따른 군집 수에 대한 군집 유효성 측도의 변화에 대한 그래프이다.

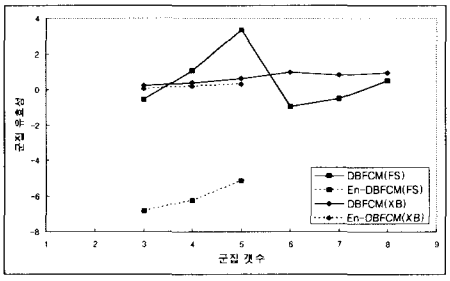

그림 3. IRIS 데이터의 군집 유효성
Fig. 3. Clustering validity of IRIS data

IRIS 데이터의 경우 군집 합병에 의해 3개의 최종 군집을 얻을 수 있으며, 위의 그래프에 의해서 FS, XB의 두 값 모두 En-DBFCM이 DBFCM보다 작게 나타났다.

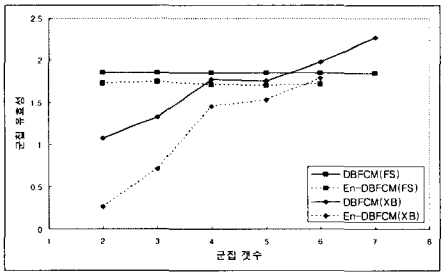

그림 4. WDBC 데이터의 군집 유효성
Fig. 4. Clustering validity of WDBC data

그림 4에서는 FS값이 log(XB)값에 비해 너무 커서 XB값의 변화 추이를 제대로 확인할 수가 없기 때문에 군집 유효성 값은 모두 log 함수를 사용하여 나타내었다. WDBC 데이터는 군집 합병 기법에 의해 최종 2개의 군집을 얻을 수 있었다.

4.2.1.1 BREAST 데이터

BREAST 데이터도 WDBC 데이터와 마찬가지로 두 군집 유효성 값을 log 함수를 이용하여 나타내었다. 그림 5를 보면 BREAST 데이터도 위의 두 데이터와 마찬가지로 En-DBFCM이 좀 더 좋은 군집 결과를 나타내고 있음을 알 수 있다.

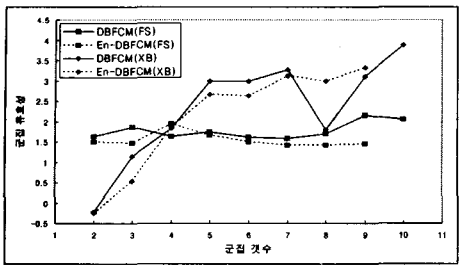

그림 5. BREAST 데이터의 군집 유효성
Fig. 5. Clustering validity of BREAST data

다음은 최종 군집과 재-군집화를 수행한 후 각각에 대한 군집 유효성 값을 정리한 것이다.

표 2. 최종 군집에 대한 군집 유효성
Table 2. Cluster validity of final clustering

		FCM	DBFCM	En-FCM
IRIS	FS	-0.593	-0.560	-6.823
	XB	1.722	1.679	1.063
WDBC	FS	72.363	71.338	53.884
	XB	17.542	11.869	1.847
BREAST	FS	42.887	42.616	31.562
	XB	0.604	0.602	0.548

표 3. 재 군집화 후의 군집 유효성
Table 3. Cluster validity after re-clustering

		FCM	DBFCM	En-FCM
IRIS	FS	-0.640	-0.698	-6.833
	XB	1.734	1.682	1.063
WDBC	FS	71.458	71.206	51.328
	XB	18.094	12.799	1.643
BREAST	FS	41.862	41.766	31.320
	XB	0.598	0.598	0.547

위의 2개의 표들로부터 실험에 사용된 모든 데이터 집합에서 En-DBFCM의 성능이 다른 알고리즘들에 비해 향상된 것을 확인할 수 있었다. WDBC와 BREAST 데이터의 경우 재 군집화를 수행할 경우 약간의 성능 향상이 있다는 것을 두 표를 통해 알 수 있었다. IRIS 데이터는 각 군집이 다른 두 데이터에 비해 잘 분리되어 있어 군집 사이에 애매하게 걸쳐있는 데이터가 상대적으로 적기 때문에 재 군집화를 통한 성능 향상이 거의 없다고 할 수 있다.

4.2.2 수행 시간 비교

En-DBFCM 알고리즘의 수행시간 단축 효과를 측정하기 위해서 각각의 알고리즘의 수행 시간을 비교하였다. 다음 표는 각 알고리즘의 수행시간을 비교한 것이다.

표 4. 각 알고리즘의 수행시간
Table 4. Computing time of each algorithm

	FCM	DBFCM	En-DBFCM
IRIS	0.249	0.469	0.313
WDBC	0.375	0.937	2.376
BREAST	0.312	8.344	1.797

DBFCM과 En-DBFCM 알고리즘은 FCM 알고리즘에 비해 많은 수행시간을 필요로 하였다. 이것은 군집 중심의 결정을 위한 격자 설정 시간과 군집화 수행 중에 군집간 합병에 많은 시간을 필요로 하기 때문이다. 하지만 En-DBFCM은 DBFCM에 비해 상대적으로 짧은 수행 시간을 필요로 하였다. 이 결과는 En-DBFCM이 격자 설정에 필요한 시간을 DBFCM 알고리즘보다 최대 1/2까지 줄일 수 있기 때문이다. 그러나 WDBC 데이터와 같이 산포가 큰 속성에 데이터가 많이 의존하는 경우 격자 생성에 필요한 시간은 줄어들지만 더 많은 군집 중심이 생성되므로 군집 합병에 대한 시간이 늘어나게 된다. 이러한 경우 En-DBFCM은 DBFCM보다 알고리즘의 수행에 있어서 많은 시간을 필요로 하게 된다. 각 데이터 개체들을 적절한 군집에 할당하기 위한 재군집화 과정의 수행시간을 다음의 표로 나타내었다.

표 5. 재 군집화 수행 시간
Table 5. Computing time of re-clustering

	FCM	DBFCM	En-DBFCM
IRIS	0.047	0.031	0.031
WDBC	0.047	0.016	0.016
BREAST	0.078	0.062	0.046

재 군집화에 소요된 시간은 전제 군집화 알고리즘이 수행된 시간에 비해 매우 작기 때문에 재 군집화 과정이 전체 알고리즘 수행시간에 미치는 영향은 작음을 알 수 있다. 또한 군집화 알고리즘에 의해 생성된 최종 군집이 잘 형성되어 있을 경우 재 군집화에 필요한 시간이 줄어든다는 것도 표 5를 통해 알 수 있었다.

4.2.3 분류 정확도 비교

각 알고리즘에 의해 형성된 최종 군집의 분류 정확도를 비교하기 위해서 최종 군집에 대한 Confusion matrix를 작성하였다. 각 표에서 괄호 안의 숫자는 재 군집화 수행 후 측정된 결과이다. 이것은 재 군집화에 의해 형성된 군집이 이전 군집에 대해 성능 향상이 있었는지를 평가하기 위한 것이다. 재 군집화 수행 후의 결과가 이전과 같은 것은 표기하지 않았다.

표 6. FCM에 의한 IRIS 데이터의 Confusion matrix
Table 6. IRIS Confusion matrix by FCM

	FCM에 의한 분류 결과			
actual	class 1	class 2	class 3	total
class 1	43	7		50
class 2	4	46		50
class 3			50	50
total	47	53	50	150

표 7. DBFCM에 의한 IRIS 데이터의 Confusion matrix
Table 7. Confusion matrix by DBFCM

	DBFCM에 의한 분류 결과			
actual	class 1	class 2	class 3	total
class 1	41(43)	9(7)		50
class 2	3(4)	47(46)		50
class 3			50	50
total	44(47)	56(53)	50	150

표 8. En-DBFCM에 의한 IRIS 데이터의 Confusion matrix
Table 8. IRIS Confusion matrix by En-DBFCM

	En-DBFCM에 의한 분류 결과			
actual	class 1	class 2	class 3	total
class 1	48	2		50
class 2	4	46		50
class 3			50	50
total	52	48	50	150

다음은 오분류율에 대한 정리이다.

표 9. IRIS 데이터의 오분류율
Table 9. Misclassification rate of IRIS data

	FCM	DBFCM	En-DBFCM
오분류 비율	7.33%	8% (7.33%)	**4%**

표 10. FCM에 의한 WDBC 데이터의 Confusion matrix
Table 10. WDBC Confusion matrix by FCM

	FCM에 의한 분류 결과		
actual	class 1	class 2	total
class 1	197(198)	15(14)	212
class 2	30(31)	327(326)	357
total	227(229)	342(340)	569

표 11. DBFCM에 의한 WDBC 데이터의 Confusion matrix
Table 11. WDBC Confusion matrix by DBFCM

	DBFCM에 의한 분류 결과		
actual	class 1	class 2	total
class 1	197(197)	15(15)	212
class 2	29(30)	328(327)	357
total	226(227)	343(342)	569

표 12. En-DBFCM에 의한 WDBC 데이터의 Confusion matrix
Table 12. WDBC Confusion matrix by En-DBFCM

	En-DBFCM에 의한 분류 결과		
actual	class 1	class 2	total
class 1	197(197)	15(15)	212
class 2	27(25)	330(332)	357
total	224(222)	345(347)	569

다음은 오분류 비율(Misclassification Rate)에 대한 표이다.

표 13. WDBC 데이터의 오분류율
Table 13. Misclassification rate of WDBC data

	FCM	DBFCM	En-DBFCM
오분류 비율	7.91%	7.73% (7.91%)	**7.38% (7.02%)**

표 14. FCM에 의한 BREAST 데이터의 Confusion matrix
Table 14. BREAST Confusion matrix by FCM

	FCM에 의한 분류 결과		
actual	class 1	class 2	total
class 1	227	14	241
class 2	12	446	458
total	239	460	699

표 15. DBFCM에 의한 BREAST 데이터의 Confusion matrix
Table 15. BREAST Confusion matrix by DBFCM

	DBFCM에 의한 분류 결과		
actual	class 1	class 2	total
class 1	227(227)	14(14)	241
class 2	12(11)	446(447)	458
total	239(238)	460(461)	699

표 16. En-DBFCM에 의한 BREAST 데이터의 Confusion matrix
Table 16. BREAST Confusion matrix by En-DBFCM

	En-DBFCM에 의한 분류 결과		
actual	class 1	class 2	total
class 1	223	18	241
class 2	11	447	458
total	234	465	699

다음은 오분류 비율(Misclassification Rate)에 대한 표이다.

표 17. BREAST데이터의 오분류율
Table 17. Misclassification rate of BREASTdata

	FCM	DBFCM	En-DBFCM
오분류 비율	3.72%	**3.72% (3.57%)**	4.14%

IRIS와 WDBC 데이터에 대해서는 En-DBFCM의 오분류율이 가장 작음을 확인할 수 있었다. 그러나 BREAST 데이터에 대해서는 En-DBFCM의 오분류 비율이 가장 높음을 알 수 있다. 데이터 집합을 분류하는 데 있어서 DBFCM이 일부의 데이터 집합에 대해서는 일부 좋은 성능을 나타내는 것도 알수 있었다. 그러나 전반적으로 En-DBFCM의 성능이 향상되었음을 확인할 수 있었다.

5. 결론 및 향후 연구과제

En-DBFCM이 DBFCM이나 FCM에 비해서 성능 향상이 되었음을 본 논문의 실험을 통하여 확인하였다. 군집 유효성 값을 비교한 결과 실험에 사용된 대부분의 데이터 집합에 대하여 En-DBFCM이 좋은 결과를 보여 주었으며, 오분류율도 작게 나타났다. 제안 알고리즘은 DBFCM 알고리즘의 격자 생성시간을 감소시키기 위해서 흩어진 정도를 사용하여 격자 생성에 필요한 시간을 최대 1/2까지 줄였다. 하지만 데이터 집합의 특성에 따라 초기 군집 중심이 많이 생성되면 군집 합병에 필요한 시간이 기존의 알고리즘에 비해 증가할 수도 있었다. 실험에 사용된 데이터의 개체 수가 크지 않았기 때문에 제안 알고리즘의 격자 생성 시간이 전체 알고리즘의 수행 시간에 대하여 큰 비중을 차지하지는 않았지만 수백만 혹은 수천만의 데이터 개체들을 포함하는 실제 데이터의 경우에는 격자 생성 시간이 문제가 될 수 있을 것이다. 개체나 속성 수가 많은 데이터 집합을 처리하기 위한 효율적인 격자 생성 방법에 대한 연구가 필요할 것이다. 이는 향후 연구과제 남겼다. 제안 알고리즘에서는 각 군집에 애매하게 걸쳐있는 데이터들을 좀 더 적절한 군집에 할당하기 위하여 재 군

집화를 수행하였다. 하지만 재 군집화 과정은 제거되었던 데이터들을 각 군집에 할당해야 하는데, 단순히 제거되었던 데이터 개체의 소속 함수 값만으로 군집 할당이 이루어짐으로써 적절치 못한 결과가 나타날 수도 있었다. 따라서 이에 대한 효과적인 해결 방안에 대한 연구도 아울러 필요할 것이다.

참 고 문 헌

[1] A. Hinneburg, D. A. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise", KDD'98, New York, 1998.

[2] U. Kaymak, M. Setnes, "Fuzzy Clustering With Volume Prototypes and Adaptive Cluster Margin", IEEE Transactions on Fuzzy Systems, Vol. 10, No. 6, 2002.

[3] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, 1987.

[4] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

[5] H. J. Zimmermann "Fuzzy Set Theory and Its Applications", Kluwer Academic Publishers. 2001.

[6] M. C. Hung, D. L. Yang, "An Efficient Fuzzy C-Means Clustering Algorithm", IEEE International Conference on Data Mining, pp. 225-232, 2001.

[7] M. S. Yang, K. L. Wu, "A New Validity Index For Fuzzy Clustering", IEEE International Conference on Fuzzy Systems, vol. 1, pp. 89-92, 2001.

[8] 한진우, 전성해, 오경환, "밀도 기반의 퍼지 C-Means 알고리즘을 이용한 클러스터 합병", 한국정보과학회 춘계학술대회 발표논문집, 2003.

[9] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "Clustering Validity Checking Method: Part II", ACM SIGMOD Record archive Vol. 31, Issue 3, 2002.

[10] D. Dubois, H. Prade, "A Unifying View of Comparison Indices in a Fuzzy Set-Theoretic Framework", Fuzzy Sets and Possibility Theory: Recent Developments, 1982.

[11] B. Kosko, "Neural Networks and Fuzzy Systems", Prentice-Hall, 1992.

[12] X. L. Xie, G. Beni, "A Validity Measure for Fuzzy Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.13, No.4, pp. 841-847, 1991.

[13] Y. Fukuyama, M. Sugeno, "A New Method of Choosing the Number of Clustering for the Fuzzy C-Means Method", Fuzzy Systems Symposium. 1989.

[14] http://www.ics.uci.edu/~mlearn

저 자 소 개

한진우(Min-Jae Park)

2000년 : 서강대 컴퓨터학과 학사
2003년 : 서강대학교 컴퓨터학과 석사
2003년~현재 : LG이노텍 소프트웨어그룹 연구원

관심분야 : 퍼지 클러스터링, 멀티에이전트, 신호분석
Phone : 02-703-7626
Fax : 02-704-8278
E-mail : heyhan@ailab.sogang.ac.kr

전성해(Sung-Hae Jun)

1993년 : 인하대 통계학과 (학사)
1996년 : 인하대 통계학과 (이학석사)
2001년 : 인하대 통계학과 (이학박사)
2003년 : 서강대학교 컴퓨터학과 (공학박사 수료)
2003년~현재 : 청주대학교 통계학과 전임강사

관심분야 : 데이터마이닝, 기계학습, 데이터공학
Phone : 043-229-8205
Fax : 043-229-8432
E-mail : shjun@cju.ac.kr

오경환(Kyung-Whan Oh)

1978년 : 서강대학교 수학과 (학사)
1985년 : Florida State University, Computer Science(공학석사)
1988년 : Florida State University, Computer Science(공학박사)
1989년~현재 : 서강대학교 컴퓨터학과 교수

관심분야 : 퍼지로직, 인공지능, 다중에이전트
Phone : 02-703-7626
Fax : 02-704-8278
E-mail : kwoh@ccs..sogang.ac.kr