Proceedings of KFIS Fall Conference, 2003. 5. 3

밀도 기반의 퍼지 C-Means 알고리즘을 이용한 클러스터 합병

Cluster Merging Using Density based Fuzzy C-Means algorithm

한진우, 전성해[0], 오경환
서강대학교 컴퓨터학과, 청주대학교 통계학과[0]

Jin-Woo,Han, Sung-Hea,Jun[0], Kyoung-Whan Oh
Dept.of Computer Science,Sogang University, Dept.of Statistics, Chongju University
E-mail : {hey_han@ailab,shjun@ailab,kwoh@ccs}.sogang.ac.kr

요 약

Fuzzy C-Means(FCM) 알고리즘은 초기 군집 중심의 개수와 위치에 따라 군집 결과의 성능차이가 많이 나타난다. 하지만 일반적인 경우에 군집 중심의 개수는 분석가의 주관에 의해 결정되고, 임의적으로 결정되기 때문에 원래 데이터의 구조와는 무관하게 수행되어 최적화된 군집화 수행을 실행하지 못하는 경우가 발생하게 된다. 따라서 본 논문에서는 원래의 데이터의 구조에 좀더 근접한 퍼지 군집화를 수행하기 위하여 격자를 바탕으로 한 데이터의 밀도를 이용한 FCM을 제안하고, 이러한 밀도 기반 FCM에 의해 결정된 군집의 합병 기법을 제안하였다. N-차원의 데이터 공간을 N-차원의 격자로 나누고, 초기 군집 중심의 개수와 위치는 각 격자의 밀도를 바탕으로 결정된다. 초기화 이후에 각 격자 내부에서 FCM을 이용하여 군집화를 수행하고, 계속해서 이웃 격자의 군집결과에 대하여 군집간의 유사도 측도를 이용하여 군집 합병을 수행함으로써 데이터의 자연적인 구조에 근접한 군집화를 수행하였다. 제안된 군집화 합병 기법의 향상된 성능은 UCI Machine Learning Repository 데이터를 이용하여 확인하였다.

1. 서론

목적 함수(objective function) 기반 퍼지 군집화(clustering) 알고리즘의 하나인 퍼지(fuzzy) c-means 알고리즘은 패턴인식, 이미지 처리 등의 다양한 분야에서 사용되었다. 퍼지 c-means 알고리즘은 데이터를 서로 겹쳐질 수 있는 집단(overlapping group)으로 나눈다. 퍼지 c-means 군집화 알고리즘은 초기 군집 중심의 개수와 위치에 의해 군집 결과의 성능에 영향을 받기 때문에 초기 군집수와 군집 위치 결정은 데이터 속성을 기반으로 결정되어야 한다. 따라서 본 논문에서는 적절한 초기 군집 중심의 개수와 위치를 데이터의 밀도를 이용하여 결정하고, 군집 결과를 군집 간의 유사도를 바탕으로 합병하는 방법을 제안하였다.

2장에서는 기존의 FCM 알고리즘과 초기 군집 개수와 위치 결정문제에 대해 알아보고, 3장에서는 이러한 문제를 해결하기 위한 밀도 기반 퍼지 c-means 알고리즘을 살펴볼 것이다. 제안된 알고리즘에 대한 평가를 위해 4장에서 기존의 기계학습 데이터인 Fisher의 Iris데이터를 이용하여 성능을 측정한 후, 마지막 5장에서 향후 연구과제와 결론을 보인다.

2. 퍼지 c-means(FCM) 알고리즘

퍼지 군집화에서는 군집 중심과 각 개체간의 유사도 정보를 가지는 분할 행렬 U를 사용하여 군집화를 수행한다. U의 각 원소인 μ_{ik}는 개체 i가 집단 k에 대한 소속 가능도인 멤버 함수값을

나타낸다.[2] 일반적으로 μ_{ik}는 다음의 조건식을 만족한다.

$$\mu_{ik} \in [0,1], \quad \sum_{i=1}^{C} \mu_{ik} = 1 \qquad \text{식(1)}$$

즉 한 개의 개체에 대하여 모든 군집에 대한 소속 가능도의 합은 1이 된다. FCM은 목적 함수 기반 퍼지 군집화 기법으로, 다음의 식(2)의 목적 함수를 최소화하여 군집화를 수행한다.

$$J(U, v_1, ..., v_K) = \sum_{i=1}^{n}\sum_{k=1}^{K} (\mu_{ik})^m d^2(x_i, v_k) \qquad \text{식(2)}$$

위 식에서 $v_{ik} = (v_{ka})(k=1, ..., K, \ a=1, ..., p)$는 집단 k의 중심값을 나타내고, $x_i = (x_{ip})(i=1, ...n, \ a=1, ..., p)$는 i번째 개체를 나타낸다. 그리고 $d^2(x_i, v_k)$는 x_i와 v_k간의 유클리디안 거리(Euclidean distance)를 나타낸다. m은 1에서 ∞까지의 값을 가지며 군집화의 퍼지화(fuzziness) 정도를 결정한다.

일반적인 FCM 알고리즘은 다음과 같다.

1. 군집 개수 K를 선택한다.
2. K개의 초기 중심을 구한다.
3. 각 개체와 군집 중심과의 거리를 구한다.
4. 개체와 군집 중심과의 멤버쉽 값을 구한다.

$$\mu_{ik}^{(t+1)}(x_k) = \sum_{j=1}^{K} \left(\frac{x_k - v_i^{(t)\,2}}{x_k - v_j^{(t)\,2}} \right)^{\frac{1}{m-1}} - 1$$

5. 군집의 중심을 보정한다.

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

6. 새로 생성된 군집 중심과 이전 군집 중심과의 거리가 임의의 값 이상이면 3으로 가서 반복하고 임의의 값 이하이거나 최대 반복회수에 도달하면, 종료한다.

FCM 알고리즘은 K-means 알고리즘과 마찬가지로 생성된 군집 중심에 따라 군집의 결과가 달라진다. 특히 초기 군집 중심을 어떻게 선택하는가에 따라 군집의 질(quality)가 크게 결정된다. 일반적인 FCM 알고리즘에서 군집의 수는 분석가에 의해 주관적으로 결정된다.

3 밀도 기반 FCM 알고리즘(DBFCM)

3.1 DBFCM

DBFCM은 격자의 밀도에 의해 결정된 초기 군집 중심으로 군집화를 수행한다. 격자를 기반으로 하여 실제 데이터가 모여 있는 지점에서 초기 군집 중심을 결정하므로, 데이터의 본래의 구조에 근접한 군집화를 수행할 수 있다[1][4].

밀도 기반 FCM 알고리즘은 적절한 초기 군집 중심의 개수와 위치를 결정하기 위해서 격자를 사용한다. 격자(grid)는 개체의 속성과 동일한 차원(dimension)을 갖는다. 즉 N차원의 속성에 대해서 N차원의 하이퍼 큐브(hyper cube)로 격자를 설정한다. 격자의 크기는 다음의 식에 의해 결정된다.

$$Edge_i = k\sigma_i \qquad \text{식(3)}$$

식(3)에서 $Edge_k$는 격자의 각 변의 길이이며, σ_k는 개체의 각 속성의 표준편차이다.

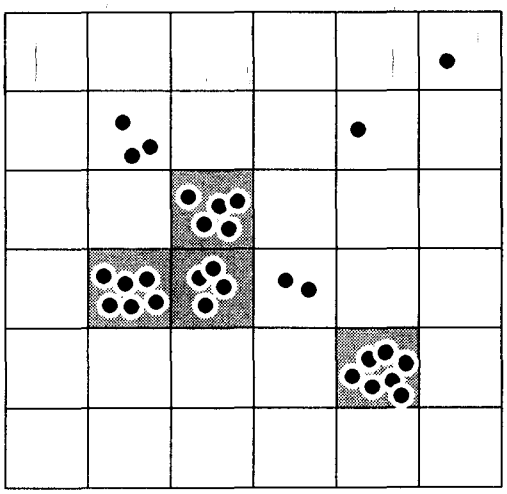

그림 1. 격자 공간내에서의 데이터 분포

각 개체는 그림 1.에서와 같이 한 개의 격자에 속하게 된다. 따라서 초기 군집 중심은 격자의 밀도에 의해서 결정된다. 격자의 밀도는 각 격자에 속해있는 개체의 수에 의해 결정되며, 초기 군집의 중심은 일정 수의 개체를 포함하고 있는

격자 내부의 임의의 위치로 결정된다. 격자의 밀도에 의해 결정된 군집 중심으로 FCM 알고리즘을 사용하여 군집화를 수행한다. 군집화 수행시 유사도 행렬 U는 다음의 식으로 보정된다.

$$v_i = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_k}{\sum_{k=1}^{n} \mu_{ik}}, \quad m > 1 \qquad \text{식(4)}$$

3.2 군집의 합병

퍼지 군집화 기법에서 군집의 수는 군집화의 성능을 좌우하는 중요한 요소이다. 기존의 군집화 기법에서는 군집의 수를 분석가의 판단에 의해 임의로 결정되거나 사전 지식을 바탕으로 결정되었다. 따라서 주어진 데이터에 대한 적절한 군집의 수를 자동적으로 결정하기 위해서 유사도를 기반으로한 군집의 합병 기법을 사용하였다. 군집간 유사도는 다음의 식에 의해 결정된다.[2]

$$S_{ij} = \frac{\sum_{k=1}^{n} min(\mu_{ik}, \mu_{jk})}{\min\left(\sum_{k=1}^{n} \mu_{ik}, \sum_{k=1}^{n} \mu_{jk}\right)} \qquad \text{식(5)}$$

위 식에서 S_{ij}는 군집 i와 j간의 유사도 이다. 각 군집들은 군집 간 유사도가 특정 임계치 이상일 때 합병된다. 임계치 $\alpha \in [0, 1]$는 데이터의 특성과 퍼지화 정도에 의존한다.

4 실험 및 결과

본 논문에서 구현한 전체 시스템은 격자를 설정하고 밀도를 구하는 모듈과 FCM 모듈, 그리고 유사도를 기반으로 군집 합병을 하는 모듈로 구성되어 있다.

4.1 실험 데이터

군집의 유효성을 측정하기 위해 기계 학습 데이터인 Fisher의 Iris데이터를 이용하였다.[6]

4.2 군집 유효성(cluster validity) 측도

본 논문에서는 군집의 유효성을 측정하기 위해서 분할 상관계수(partition coefficient)를 사용하였다.[5] 분할 상관계수는 다음의 식으로 구해진다.

$$F(\tilde{U}, c) = \sum_{k=1}^{n} \sum_{i=1}^{K} \frac{(\mu_{ik})^2}{n} \qquad \text{식(6)}$$

따라서 결정된 군집에 대한 분할 상관계수의 값은 우수한 군집에 대해 높은 값을 보이게 된다.

4.3 실험결과

4.3.1 최적 군집수 결정

군집수의 결정은 DBFCM에 의해 생성된 각 군집의 합벼은 특정 임계치 이상을 유지할 때까지 진행된다. 이때 이 임계치는 합병된 군집이 유효한 의미를 가지는 것을 의미하며, 이것은 분할 상관계수에 의해 결정된다. 하지만 이 값은 군집수가 증가함에 따라 높은 값을 가질 수 있으므로 이를 보완하기 위해 군집수 증가에 따라 다음식과 같은 불이익(penalty)을 부과한다.

$$penalty = \frac{1}{e^{-1/k}} \qquad \text{식(7)}$$

위의 식에 의해 군집수 k가 증가하면 전체 불이익은 증가하지만 그 값은 선형증가에 비해 낮은 폭으로 증가하게 된다. 따라서 임계치를 구하는 식은 다음과 같이 수정된다.

$$threshold = F(\tilde{U}, c) \times \frac{1}{e^{-1/k}} \qquad \text{식(8)}$$

그림 2.는 DBFCM에 의해 결정된 군집수에 따른 군집 유효성을 나타내고 있다.

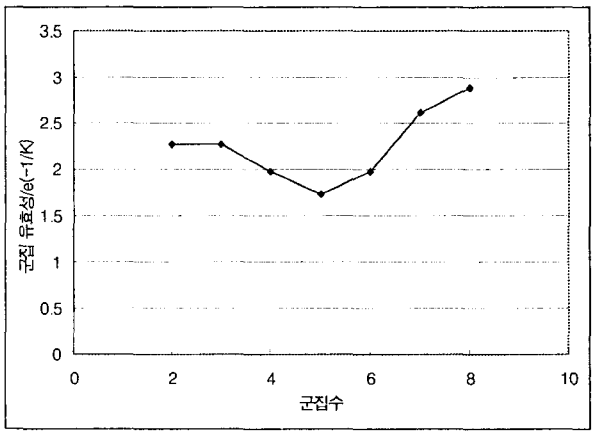

그림 2. 군집수에 대한 군집 유효성

4.3.1 군집 유효성 비교

기존의 FCM과 DBFCM을 비교하기 위해 군집 유효성을 10회 측정한 값이 표 1.에 나타나 있다. FCM의 초기 군집수는 3으로 결정하여 수행한 결과 초기 군집 위치 결정에 의해 군집 유효성 값의 변화폭이 넓음을 알 수 있다. 하지만 이에 반해 DBFCM은 거의 일정한 수준으로 FCM보다 향상된 값을 보이고 있다.

실험횟수	FCM	밀도기반 FCM
1	1.1857	1.6175
2	1.0652	1.4608
3	1.6192	1.5120
4	1.4988	1.5120
5	1.3862	1.5120
6	1.5101	1.4608
7	1.3670	1.4608
8	1.3567	1.5150
9	1.2080	1.6203
10	1.40569	1.5230

표 1. 군집 유효성 비교

5. 결론

본 논문에서는 퍼지 c-means 알고리즘에서 군집화의 성능에 여향을 미치는 초기 군집수의 적절한 개수 및 위치를 결정하기 위한 밀도기반의 퍼지 c-means 알고리즘과 군집 합병방법을 제안하였다. 제안된 알고리즘은 기존의 FCM에 비해 항상 최적화에 수렴됨을 확인할 수 있다. 이러한 연구 결과는 특히 데이터에 대한 사전지식이 부족한 경우 적절히 사용될 수 있을 것이다. 하지만 군집 유효성에 대한 보다 분명한 평가 기준의 마련과 이를 통한 최적화 군집 결정문제는 향후 계속 연구해야할 영역이다.

감사의 글

본 연구는 과학 기술부 주관 뇌신경 정보학 사업에 의해 지원되었음.

참고문헌

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

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

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

[4] J. Han and 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] http://www.ics.uci.edu/~mlearn