차감 HyperBox 알고리듬을 이용한 Unsupervised 클러스터 추정

Unsupervised Cluster Estimation using Subtractive HyperBox Algorithm

문 성환, 최 병걸, 강 훈
중앙대학교 공과대학 제어계측학과

ABSTRACT

Mountain Method 의 다른 형태인 Subtractive 클러스터링 알고리듬은 계산이 간단하고 기존의 클러스터링 방법들과는 달리 초기 클러스터 중심의 개수 선정이 필요 없기 때문에 클러스터를 추정하는데 효과적인 알고리듬이다. 또한 클러스터의 간격을 결정하는 파라미터의 값에 따라 클러스터의 개수를 다르게 할 수 있다. 그러나 이 파라미터에 의해 동일한 그룹(Class)내에서 여러 개의 클러스터 중심이 발생될 수도 있다. 본 논문에서는 Subtractive HyperBox 알고리듬을 사용하여 이 파라미터의 영향을 줄이고 발생한 클러스터 중심이 속한 그룹의 경계를 판정함으로서 같은 그룹 내에서 하나의 클러스터만 발생하도록 하고, 순차적으로 클러스터링 한 후 결과를 Subtractive 클러스터링 알고리듬과 비교하여 보았다.

I. 서론

Yager 와 Filev는 Mountain 함수의 밀도 측정에 기초한 클러스터 중심의 추정에 간단하면서도 효과적인 Mountain 클러스터링 방법[3,4]을 제시하였다. Mountain 클러스터링 방법의 한 형태인 Subtractive 클러스터링 방법 [1]은 계산량에 있어 Mountain 클러스터링 방법보다 현저히 줄일 수 있다. 그러나 클러스터의 개수가 초기 파라미터 값에 따라 매우 민감하게 달라지는 단점을 가지고 있다「」. 다시 말해 하나의 데이터 그룹(Class)내에서 여러 개의 클러스터가 발생할 수 있다는 것이다. Subtractive HyperBox Clustering 방법은 포텐셜 값의 차감으로 클러스터 중심을 구하고 이 클러스터 중심이 속해있는 그룹의 경계(Boundary)를 결정해 줌으로써 하나의 Class에서 하나의 클러스터 중심만이 발생할 수 있도록 하였다.

본 논문에서는 먼저 Subtract Clustering 방법을 살펴보고, 문제점과 원인을 파악한 후 Subtract HyperBox 알고리듬을 사용하여 해결방안을 모색하고 이후 두 가지 방법을 Simulation하여 얻은 결과를 통해 개선된 점을 살펴보았다.

II. 본론

1. Subtractive Clustering 방법

M-차원의 공간상에 어떤 HyperCube에 정규화(normalized)된 N개의 데이터 $\{X_1, X_2, X_3 \cdots X_N\}$가 주어졌을 때 각 데이터마다 포텐셜 P_i 를 구한다.

$$P_i = \sum_{j=1}^{N} \exp(-\alpha \|X_i - X_j\|^2) \qquad i=1,2,3\ldots N$$

파라미터 α는 $4/r_a^2$으로 주어지며 r_a는 양의 상수로 r_a밖의 데이터는 포텐셜 값에는 영향을 거의 주지 못하게 된다. 여기서 구한 N개의 포텐셜 값중 가장 높은 값을 P_1^* 라 놓고 이때의 데이터가 첫 번째 클러스터의 중심 X_1^* 가 된다. 첫 번째 클러스터 중심의 영향을 다른 포텐셜 P_i에서 제거한다. 즉, 첫 번째 클러스터 중심 근처에서 다음 클러스터의 중심에 발생하지 않도록 첫 번째

*본 연구는 1996년도 한국학술진흥재단 대학부설연구소과제 연구비에 의하여 연구되었음

클러스터의 영향을 제거하지 않으면 안된다. 첫 번째 클러스터 중심의 영향을 제거한 포텐셜 값 P_i'을 구한다.

$$P_i' = P_i - P_1^* \exp(-\beta \|X_i - X_1^*\|^2)$$

i=1, 2, 3...N

파라미터 β 는 $4/r_b$로 r_b는 r_a보다 큰 상수 값으로 클러스터의 중심 근처에 다음 클러스터 중심이 나타나지 않도록 한다. k 번째 클러스터 중심은 다음과 같이 구한다.

$$P_i' = P_i - P_{k-1}^* \exp(-\beta \|X_i - X_{k-1}^*\|^2)$$

i=1, 2, 3...N

이 과정을 충분한 수의 클러스터가 발생할 때까지 반복한다.

Subtractive 클러스터링 방법은 클러스터의 포텐셜을 결정하는 반경값(초기 파라미터 α)에 의해서 클러스터의 수가 결정되기 때문에 최적화된 클러스터링 방법이라고 볼 수는 없다. 그리고 Subtraction 과정에서 클러스터 중심의 영향을 제거한다 하더라도 그룹의 범위가 주어진 파라미터보다 크게되면 한 그룹 내에 또다른 클러스터 중심이 발생하게 된다. 즉 충분한 만큼의 클러스터 중심의 영향을 제거해 주지 못하게 되기 때문이다. 발생한 클러스터 중심의 영향을 제거하기 위해서는 우선 이 클러스터 중심이 속해 있는 그룹의 경계를 결정해야 하고, 그 그룹에 속하는 다른 데이터들의 포텐셜은 다른 그룹에 속해있는 것보다 충분이 낮아져야 한다. Subtractive HyperBox 알고리듬은 이러한 문제점에 대한 해결책을 제시한다.

3. Subtractive HyperBox Clustering Algorithm

3.1. 포텐셜(Potential function)의 결정

지수함수는 데이터의 기하학적 구조에 영향을 많이 받기 때문에 다음과 같은 포텐셜 함수의 사용으로 데이터들의 구조적 영향을 줄일 수 있다.

$$f(x) = \frac{1 - \tanh(\lambda(x - R))}{2}$$

파라미터 λ의 값을 크게 줌으로써 반경 R 내로 들어온 데이터들에 관해 높은 포텐셜 값을 줄 수 있고 차감 과정에서 클러스터 중심의 영향 또한 크게 제거할 수 있다.

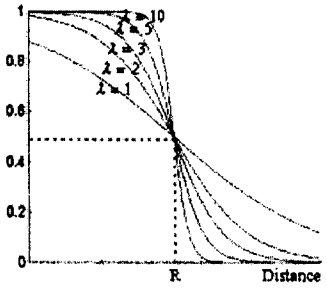


그림 1. 밀도 함수

어떤 HyperCube에 정규화(normalized)된 N개의 데이터 $\{X_1, X_2, X_3 \cdots X_N\}$가 주어졌을 때 각 데이터마다 포텐셜 P_i 를 구한다.

$$P_i = \sum_{j=1}^{N} \frac{1 - \tanh(\lambda(\|X_i - X_j\| - R))}{2}$$

포텐셜 P_i 값중 가장 큰 데이터가 첫 번째 클러스터 중심이 된다.

3.2 클러스터 중심이 속한 그룹의 경계 결정

클러스터 중심 X_k^* 가 결정되면 X_k^* 속한 그룹의 경계를 결정하게된다. 먼저, X_k^*를 기준으로 4분면을 만든 다음 (그림2). 각 사분면 마다 X_k^*으로부터 HyperBox $B(V, W)$를 형성하게 되는데 V는 HyperBox의 MIN Point 이고 W는 MAX Point를 말한다. [2]

$B(V, W)$의 초기값은 X_k^*로 놓는다. h 번째 데이터 X_h $(1 \le h \le N)$에 대해 먼저 X_h가 X_k^* 으로부터 어느 사분면에 속하는지 결정하게 된다. X_h 의 위치가 결정되면 자신이 속한 사분면의 HyperBox의 크기를 변경하게 된다.

$$\frac{\sum_{i=1}^{n}(MAX(W_{ji}, X_{hi}) - MIN(V_{ji}, X_{hi}))}{n} \le \theta$$

을 만족하게 되면 Min-Max [3] 방법을 통해 HyperBox의 크기가 변하게 된다.

$$V_{ji}^{new} = MIN(V_{ji}^{old}, X_{hi})$$

$$W_{ji}^{new} = MAX(W_{ji}^{old}, X_{hi})$$

여기서 i($1 \le i \le n$)는 데이터의 Dimension이고, j는 각 사분면을 말한다. 또한 파라미터 θ로 HyperBox의 최대 크기를 정해 줌으로써 직선과 같은 등간격의 데이터 또한 적당한 간격으로 클러스터링이 가능하게 된다. X_k^*로 부터 각 사분면의 HyperBox 와의 최대 거리를 구해 평균을 내면 이것이 현재 선택된 클러스터 중심이 속해있는 그룹의 경계 (R_k)가 된다(그림 3).

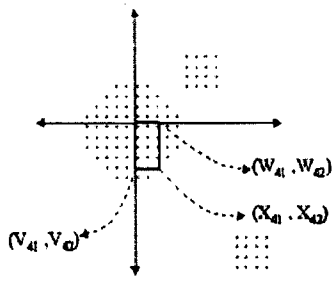


그림 2. HyperBox의 형성

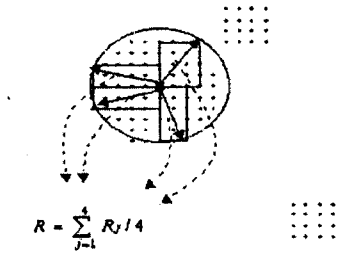


그림 3. 그룹의 경계 결정

3.3 클러스터링 수행 과정

얻어진 클러스터 중심 X_k^*, 와 R_k(그룹의 경계)로 Subtraction 과정을 거친다.

$$P_i' = P_i - P_k^* \left(\frac{1 - \tanh(\lambda(\|X_i - X_k^*\| - R_k))}{2} \right)$$

차감된 포텐셜 P_i'중 가장 큰 값이 클러스터 중심후보 X_{k+1}^*가되고

만약 $P_{k+1}^* > \bar{\varepsilon} P_1^*$ 이면 X_{k+1}^*를 클러스터의 중심으로 인정하고 차감과정을 계속하고

만약 $P_{k+1}^* < \underline{\varepsilon} P_1^*$ 클러스터링 수행과정은 끝나게 된다.

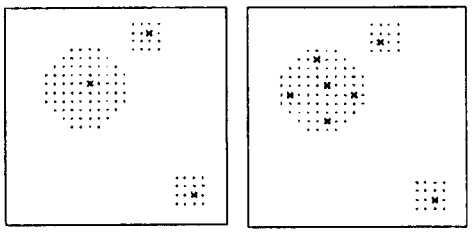
$\underline{\varepsilon} P_1^* \le P_{k+1}^* \le \bar{\varepsilon} P_1^*$ 인 경우 $\frac{d_c}{R_c} > 1.5$ 이고

$\frac{d_{\min}}{R_k} + \frac{P_1^*}{P_{k+1}^*} \ge 1$ 이면 X_{k+1}^* 클러스터 중심으로 결정하고 아니면 X_{k+1}^*의 포텐셜 $P_{k+1}^* = 0$으로 놓고 다음으로 높은 포텐셜 값을 갖는 데이터를 X_{k+1}^* 라놓고 다시 클러스터 중심으로서의 조건을 TEST한다. d_c는 현재 X_{k+1}^* 와 이미 구한 클러스터 중심값 X_c^* ($1 \le c \le k$)과의 거리이고 $d_{\min}$ 은 그중 가장 가까운 값이다. R_c는 각 그룹의 경계를 말하고, 이 조건으로 하나의 Class 내에서 여러개의 클러스터가 발생하는 것을 방지할 수 있다.

4. Subtractive Clustering 과 Subtractive HyperBox Algorithm의 비교 결과

그림 4 의 (a)와 (b)는 같은 데이터 셋으로 각각 Subtractive HyperBox 방법과 기존의 Subtractive Clustering을 한 것을 보여주고 있다. 데이터들은 정규화(그림 4에서는 -1에서 1사이) 되어있고 초기 파라미터의 값은 0.3, Subtractive HyperBox Algorithm 의 경우 HyperBox의 최대치를 0.3으로 놓았다. 그림 4.(b) 에서는 초기 파라미터의 값이 데

이터의 그룹보다 작기 때문에 한 그룹 내에서 여러 개의 클러스터 중심이 나타나는 것을 볼 수 있다. 그림 4.(a)의 경우 클러스터 중심의 영향과 동시에 중심이 속한 그룹 경계의 결정으로 그룹내의 다른 데이터의 포텐셜을 충분히 낮게 만들어 주었기 때문에 경계 내에서는 그룹으로 설정된 곳에서는 하나의 클러스터 중심만 나타난 것을 볼 수 있다.



(a) (b)

그림 4.

(a) Subtractive HyperBos Clustering
(b) 기존의 Subtractive Clustering

그림 5 와 그림 6은 균일한 분포를 가진 데이터 셋을 Subtract HyperBox Clustering 방법을 사용하여 나온 결과를 보여주고 있다. 이 경우 HyperBox의 최대 값을 지정해 줌으로써 비교적 일정한 간격의 클러스터 중심을 찾는 다는 것을 알 수 있다.

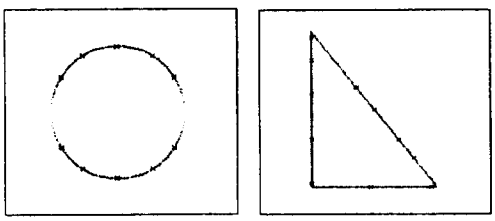


그림 5 그림 6

III. 결론

Subtactive HyperBox Algorithm은 데이터들의 포텐셜 측정과 클러스터 중심의 선택, 그리고 클러스터가 속한 그룹의 경계를 결정하여 충분한 만큼의 포텐셜을 제거해 주는 과정을 반복한다. 이러한 과정을 수행함으로써 기존의 Subtract clustering 방법보다 최적화된 클러스터 중심을 추정할 수 있다. 이렇게 추정한 클러스터 중심을 이용해 FCM(Fuzzy C-Means) 알고리듬[5] 같이 초기 파티션의 문제를 안고 있는 곳에 사용하여 초기 파티션의 문제를 해결할 수 있을 것으로 본다.

현재는 그룹의 경계는 HyperBox를 이용한 반경을 가지는 원의 형태로 나타난다. 앞으로 원의 경계에서 벗어나 좀더 정확한 그룹의 경계를 알아내는 것이 수행과제이다.

IV. 참고문헌

[1] S.L.Chiu. Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy systems*, 2(3),1994.

[2] P.K.Simpson.Fuzzy Min-Max Neural Networks-Part 1: Classification, *IEE Transactions. Neural Networks*, vol.3, Sept.1992.

[3] P.K.Simpson.Fuzzy Min-Max Neural Networks-Part 2: Clustering, *IEEE Transactions on Fuzzy Systems*. vol.1,No.1, February 1993.

[4] R.R.Yager and D.P. Filev. Approximate clustering via the mountain method. *IEEE Transactions on Systems, Man, and Cybertics*, 24:1297-1284, 1994.

[5] Ronald R.Yager and Dimitar P.Filev. *Essentials of fuzzy modeling and control*. John Wiley & Sons, Inc,1994.

[6] N.R.Pal and J.C.Bezdek. On Cluster Validity for the Fuzzy c-Means Model. *IEEE Transaction On fuzzy Systems*, Vol. 3,No.3. August 1995.