새로운 수렴특성을 이용한 클러스터 모델링

A Cluster modeling using New Convergence properties

김승석**, 백찬수*, 김성수***, 유정웅****
(Sung-Suk Kim, Chan-Soo Baek, Sung-Soo Kim, Joeng-Woong Ryu)

Abstract - In this paper, we propose a clustering that perform algorithm using new convergence properties. For detection and optimization of cluster, we use to similarity measure with cumulative probability and to inference the its parameters with MLE. A merits of using the cumulative probability in our method is very effectiveness that robust to noise or unnecessary data for inference the parameters. And we adopt similarity threshold to converge the number of cluster that is enable to past convergence and delete the other influence for this learning algorithm. In the simulation, we show effectiveness of our algorithm for convergence and optimization of cluster in given data set.

Key Words : Gaussian Mixture Model, Expectation-Maximization algorithm, Mountain Clustering, Convergence

1. 장 서론

인공지능 모델을 이용한 제어 및 패턴을 분류하고자 할 때 모델의 학습 성능에 의하여 전체 구조의 성능이 영향을 받는다. 학습된 최종 모델의 성능은 초기 파라미터 결정과 학습 알고리즘에 의하여 좌우된다[1-2]. 모델의 학습뿐만 아니라 초기 파라미터의 결정 역시 모델의 성능 개선에 영향을 미치며 이에 대하여 다양한 연구가 이루어져 왔다[2]. 기존의 격자 형태로 데이터 공간을 분할한 후 전체 데이터 영역을 이용하여 모델을 학습하는 알고리즘의 경우 전체 영역에 대하여 모두 규칙들을 부여할 수 있는 장점과는 달리 데이터의 입력 차원이 증가하거나 각 차원의 분할 영역이 증가하는 경우 학습 모델의 크기가 커지므로 모델의 학습이나 실제 구성에 문제가 발생할 수 있었다[2][6-7]. 반면 클러스터링을 이용하여 파라미터를 추정하는 경우 입력차원의 증가나 분할 영역의 증가에 크게 영향을 받지 않으면서도 적은 파라미터를 이용하여 데이터 공간을 표현할 수 있다[6-7].

클러스터링의 기본 개념은 유사도가 높은 데이터들을 같은 영역에 그렇지 않은 경우 다른 영역에 속하도록 하는 것이다. 또한 전체 데이터 공간을 모두 사용하는 것이 아니라 데이터가 존재하는 공간만을 이용하여 규칙을 추정함으로써 불필요한 공간에 규칙을 할당하는 것을 방지할 수 있다[2]. 클러스터링 기법으로는 크게 두 가지로 나눌 수 있다. 먼저 사전에 지정된 임계치을 이용하여 임계 조건을 만족하는 형태로 클러스터를 구성하여 클러스터의 수와 파라미터를 추정하는 방법과 사전에 지정된 클러스터의 수를 학습 알고리즘을 이용하여 최적화는 것이다. 전자의 경우 Mountain 알고리즘과 같이 미지의 데이터에 대하여 클러스터의 수를 추정할 수 있으나 임계치에 의하여 클러스터의 수가 쉽게 변할 수 있으며 최적화 과정이 따로 포함되어 있지 않다. 두 번째의 경우 Fuzzy C-Mean (FCM)이나 Gaussian Mixture Model (GMM)과 같이 주어진 클러스터의 수에 대하여 이들 파라미터를 최적화하지만 클러스터의 수를 알지 못하는 경우 모델 학습에 불필요한 파라미터를 추정할 수 가 있다[7].

본 논문에서는 Chen 알고리즘을 기반으로 클러스터 수의 수렴 및 파라미터 최적화를 실시하는 알고리즘을 제안한다[5]. 클러스터의 수렴 및 알고리즘 속도를 개선하기 위하여 Subtractive 알고리즘을 이용하였다. 또한 유사도 함수의 분산과 유사도 측정에 동시에 임계치를 적용하여 임계값의 변화에도 일정한 결과를 보이도록 하였다[4-5]. 이 경우, 클러스터 추정에 필요한 Gaussian 분산 등이 커지는 경우 클러스터 추정 범위를 제한하고 분산이 작아지는 경우 추정 범위를 넓히는 방식으로 서로 상반되는 파라미터를 동시에 임계치 적용 대상에 포함함으로써 임계치 변동에 대하여 강인한 특성을 보이도록 하였다. 또한 Maximum Likelihood Estimation (MLE) 의 Expectation-Maximization (EM) 알고리즘의 개념을 도입하여 클러스터 수의 추정과 동시에 이들 파라미터를 최적화하였다[8].

제안된 방법을 이용하여 주어진 데이터에 대하여 패턴에 대하여 시뮬레이션을 실시하여 클러스터의 수 추정 및 파라미터 최적화에 대한 결과 및 임계치 변동에 따른 강인한 클러스터 추정 능력을 보임으로써 제안된 알고리즘의 유용성을 보이자 한다.

2. 장 제안된 클러스터링 알고리즘

하나의 클러스터를 구성하는 조건으로 각 중심과 데이터 간의 유사도를 이용하여 결정하는 방법과 유사하게 본 논문

저자 소개
* 準 會 員 : 忠北大學 電氣工學科 碩士課程
** 正 會 員 : 忠北大學 電氣工學科 博士課程
*** 正 會 員 : 忠北大學 電氣工學科 助敎授 · 工博
**** 正 會 員 : 忠北大學 電氣工學科 敎授 · 工博

에서는 각 중심과 전체 데이터 간의 누적 유사도(누적 확률)를 이용하여 중심을 추정하였다. 각 데이터간의 유사도만을 이용하는 경우 알고리즘이 진행되는 동안 클러스터 파라미터의 변화(분산의 변화)에 의하여 특정한 영역이나 데이터에 의하여 클러스터의 수렴 형태에 영향을 발생할 수 있다[4-5]. 반면 각 중심에 대하여 전체 데이터의 누적 유사도를 이용할 경우 특정 영역이나 특정 데이터에 대한 영향이 적은 강인한 클러스터 추정을 할 수가 있다. 또한 클러스터의 수를 수렴시키기 위하여 임계값을 이용하여 누적 유사도가 적은 정보는 클러스터 추정에 영향을 주지 않도록 하였다.

각 데이터 x_i와 중심 μ_j와의 확률은 다음과 같이 계산한다.

$$s_{ij} = \exp\left(-\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)\right) \quad (1)$$

식( )에서 생성된 확률은 각각의 중심에 대하여 누적된 확률을 이용하여 사전에 지정된 임계치를 이용하여 클러스터의 수를 수렴시킨다.

$$U_{ij} = \begin{cases} 0 & \text{if } \dfrac{\sum_{i=1}^{m}(s_{ij})}{\max\left(\sum_{i=1}^{m} s_{ij}\right)} < \zeta \\ s_{ij} & \text{otherwise} \end{cases} \quad (2)$$

이 경우 누적확률이 임계치 이하로 되는 중심의 분할행렬 U_j는 다음 학습에서 제외가 된다[5]. 즉 데이터 분포에서 클러스터 추정에 불필요하거나 클러스터의 중심에서 벗어난 파라미터는 학습 규칙에 의하여 제거된다.

새롭게 생성된 분할행렬 U_{ij}를 이용하여 새로운 클러스터 파라미터를 다음과 같이 생성한다.

$$\mu_j = \frac{\sum_{i=1}^{m} U_{ij}x_i}{\sum_{i=1}^{m} U_{ij}} \quad (3)$$

클러스터의 중심이 결정되면 이의 분산은 MLE을 최적화하기 위하여 EM알고리즘의 개념을 이용하여 다음과 같이 추정한다[8].

$$p_{ij} = \frac{1}{(2\pi)^{d/2}\det(\Sigma_j)} \exp\left(-\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)\right) \quad (4)$$

사후 확률 PW은 p_{ij}에 사전확률을 곱하여 구하여 이를 이용하여 분할 행렬을 생성하면 다음과 같다.

$$U_{ij} = \frac{PW}{\sum_{j=1}^{d} PW} \quad (5)$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} U_{ij}(x_i-\mu_j)(x_i-\mu_j)^T}{\sum_{i=1}^{m} U_{ij}} \quad (6)$$

이를 이용하여 새로운 사전확률 w_j를 다음과 같이 구할 수 있다.

$$w_j = \frac{1}{m}\sum_{i=1}^{m} U_{ij} \quad (7)$$

초기 파라미터는 전체에 대하여 균일하게 설정하고 알고리즘이 진행되는 동안 임계치 ζ를 이용하여 자율적으로 수렴하도록 한다. 또한 알고리즘이 진행되는 동안 클러스터의 수가 변화하는 경우 클러스터 파라미터의 값을 초기화하여 이전의 특정 파라미터가 학습에 유리하게 작용하는 것을 방지하였다. 또한 클러스터의 수렴 속도를 개선하기 위하여 Subtractive 알고리즘 개념을 도입하여 특정 지역으로 군집하는 클러스터 파라미터를 묶는 방식을 통하여 알고리즘의 수렴 성능 및 속도를 개선하였다.

3. 장 시뮬레이션 및 결과

본 논문에서는 Jang[2]이 생성한 서로 다른 분포와 크기를 가지는 임의의 데이터를 이용하여 시뮬레이션을 실시하였다. 먼저 임의로 생성된 900개의 그림 1과 같은 데이터 집합을 연산의 편리성을 위하여 300개의 집합으로 축소한 후 알고리즘을 실시하였다.

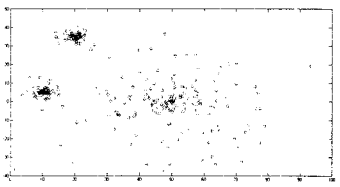

그림 1. 데이터 분포

각 데이터가 가지는 누적확률분포는 그림 2와 같다.

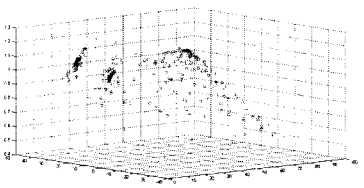

그림 2. 데이터의 누적분포

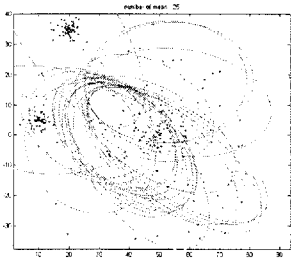

그림 3. 1회 학습후의 파라미터

그림 1에서 볼 수 있듯이 군집이 몰려 있는 공간과 그렇지

않은 공간에서의 누적확률 밀도는 차이를 보이며 임계치 이하의 누적확률을 가지는 클러스터 중심을 제거함으로써 알고리즘을 진행한다. 이를 이용하여 한번 학습하였을 때 클러스터 중심과 분산을 그림 3에 나타내었다.

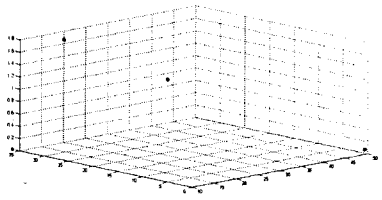

그림 4. 최종 누적확률

그림 4와 그림 5에서 최종 학습 알고리즘이 종료(수렴)한 후 각 클러스터의 중심 및 누적확률을 나타내었다. 또한 데이터 집합과 클러스터 중심 및 파라미터를 그림 에 나타내었다.

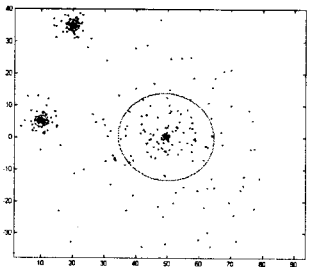

그림 5. 추정된 파라미터

이를 사전 임계값을 이용하는 Subtractive 클러스터링 알고리즘과 비교하였을 때 클러스터 추정 결과를 비교하면 다음과 같다.

표 1. 결과 비교

임계치	Subtractive	Proposed	비고
0.3	3	3	변동/동일
0.2	4	3	변동/동일
0.1	4	3	변동/동일
0.05	4	3	변동/동일

4. 장 결론

본 논문에서는 신경회로망이나 패턴 인식에 최적화된 초기 파라미터 결정에 필요한 알고리즘의 하나인 클러스터링의 자율적인 클러스터 추정에 대하여 제안하였다.

제안된 방법에서는 클러스터 파라미터와 데이터간의 유사도를 이용하여 자율적으로 클러스터링을 실시하며 알고리즘의 수렴 및 속도를 개선하였다. 데이터의 패턴이 명확하게 나누어지거나 각 패턴 간의 거리가 충분할 경우 대부분의 알고리즘이 원하는 성능을 보이지만 패턴 간의 특성 또는 분포가 다른 경우 원하는 성능을 보장하지 못하는 경우, 제안된 알고리즘에서 좀 더 강인한 특성을 보였다. 또한 클러스터 중심 추정시 각 데이터와 중심 간의 유사도를 이용함으로써 발생할 수 있는 노이즈 등과 같은 원치 않은 정보에 의한 성능저하를 누적확률을 이용함으로써 특정 데이터 또는 정보에 의한 영향에 더욱 강인한 특성을 보이도록 하였다.

향후 연구과제로는 제안된 알고리즘의 연산량을 줄이는 새로운 규칙 등을 개발하는 것과 이러한 결과를 신경회로망이나 패턴인식 모델의 초기값이나 또다른 모델로의 확장 등이 있다.

참 고 문 헌

[1] Simon Haykin, Neural Networks : A Comprehensive Foundation Second Edition, Prentice Hall, 1999.

[2] J. S. R. Jang, C. T. Sum, E. Mizutani, Neuro-Fuzzy and Soft Computing : A Computational Approach to Learning and Machine Intelligence, Prentice Hall, 1997.

[3] J. S. Jang, "Input Selection for ANFIS Learning", Proceedings of the Fifth IEEE International Conference on Fuzzy System, Vol. 2, pp. 1493-1499, 1996.

[4] R. R. Yager, D. P. Filev, "Generation of Fuzzy Rules By Mountain Clustering", Jounal of Intelligence and Fuzzy System, Vol. 12, pp. 209-230, 1994.

[5] Ching-Chang Wong, Chia-Chong Chen, Mu-Chun Su, "A novel algorithm for data clustering", Pattern Recognition, Vol. 34, Issue. 2, pp. 425-442, 2001.

[6] 김승석, 곽근창, 유정웅, 전명근, "계층적 클러스터링과 Gaussain Mixture Model을 이용한 뉴로-퍼지 모델링", 한국퍼지및지능시스템학회 논문지, Vol. 13, No. 5, pp. 512-519, 2003.

[7] 김승석, 김성수, 유정웅, "새로운 클러스터링 알고리즘을 적용한 향상된 뉴로-퍼지 모델링", 대한전기학회 논문지, Vol. 53D, No. 7, pp. 536-543, 2004.

[8] Guorong Xuan, Wei Zhang, Peiqi Chai, "EM algorithm of Gaussian Mixture Model and Hidden Markov Model", Image Processing Proceedings, International Conference on, Vol. 1, pp. 145-148. 2001.