데이터 클러스터링을 위한 가우시안 혼합 모델을 이용한 퍼지 정보량 측정

Gaussian Mixture Model for Data Clustering using Fuzzy Entropy Measures

임채주, 최병인, 이정훈
한양대학교 전자공학과

Chae-Ju Lim, Byung-In Choi, Frank Chung-Hoon Rhee
Dept. of Electronic Engineering in Hanyang University
E-mail : {cjlim, bichoi, frhee}@fuzzy.hanyang.ac.kr

요 약

본 논문에서는 기존의 정보량(Entropy) 기반 클러스터링 기법을 향상시키기 위한 방법으로서 퍼지 정보량을 이용하였다. 가우시안 혼합 모델을 이용하면, 프로토타입의 목적 함수를 이용하는 클러스터링 기법보다 향상된 결과를 얻을 수 있고, Parameter의 조정이 요구되지 않는다. 그러나, 가우시안 혼합 모델의 사용은 주어진 패턴 집합을 클러스터링하는데 계산량의 증가를 초래하게 된다. 본 논문에서는 가우시안 혼합 모델의 정형화에 요구되는 계산량을 감소시키는 방법을 제시한다. 또한 퍼지정보량(Fuzzy Entropy)을 적용하여 기존의 정보량 기반의 클러스터링 결과와 비교 분석하였다.

1. 서론

현재 영상처리, 음성인식등 다양한 분야에서 데이터를 분석하고자 클러스터링 기법이 많이 적용되고 있다. 널리 쓰이고 있는 방식은 프로토타입 기반의 클러스터링 방법이다. 이 기법은 특정한 형태의 데이터 분포에 대해서는 잘 구분하는 것에 비해 복잡한 형태에 대해서는 정확도가 감소할 수 있다.

Gausian Mixture Model(GMM)을 이용한 정보량 기반의 클러스터링은 특정한 프로토타입이나 형태와 상관없이 정보량의 안정화에 따라 클러스터링을 수행한다. 그러나 가우시안 혼합모델을 정형화(fitting) 시키는 작업에 많은 연산량이 요구되므로 인식률과 연산량의 두 조건을 충족시키지 못하므로 실제 응용화에는 어려움이 따른다.

이러한 단점을 극복하기 위해, 본 논문에서는 가우시안 혼합 모델을 빠르게 정형화하기 위한 방법으로 패턴인식 클러스터링 기법인 Kmeans Algorithm 을 이용한다. 그리고 기존의 정보량 기반의 클러스터링 방법 MCP[1](Maximum Certainty Partitioning)에서 퍼지 정보량 방법을 적용하였다.

앞으로 2절에서는 GMM의 정형화의 구체적인 방법에 대해서 언급하고, 3절에서는 기존의 Shannon 정보량과, 퍼지정보량에 대해서 다룰 것이다. 4절에서는 본 논문에서 제시하는 알고리즘에 대해 서술하고, 마지막으로 5절에서는 몇가지 실험 데이터의 결과를 비교 분석할 것이다.

접수일자 : 2004년 1월 1일
완료일자 : 2012년 12월 31일
감사의 글 : 본 연구는 한국과학기술원 영상정보특화연구센터를 통한 국방과학연구소의 연구비 지원으로 수행되었습니다.

2. 가우시안 혼합 모델링

본 논문에서 제안하는 방법은 크게 두 단계로 나눌 수 있다.

첫번째 단계는 주어진 데이터를 N개의 GMM (가우시한 혼합 모델)로 정형화 시키는 단계이고, 두 번째는 정형화된 GMM을 가지고 퍼지 정보량을 이용해서 모든 데이터가 가지는 퍼지 정보량의 합산이 최소가 되는 방향으로 클러스터링을 수행하는 단계이다..

가우시안 혼합모델의 정형화를 위해 먼저, k개의 혼합모델에서 각 패턴이 각 혼합모델에 소속될 조건부 확률 $P^*(j|X)$값을 구해야 한다. 기존의 GMM 클러스터링[5] 에서는 각 혼합 모델에서의 조건부 확률 $P^*(j|X)$이 iterative 방법을 이용해서 구현한다. 즉, 각iteration 마다 Covariance Matrix를 계산, 반복적으로 다음의 가우시안 함수로부터 $P^*(j|X)$를 구한다.

$$P^*(j|X) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \times exp[-\frac{1}{2}(X-M_j)^T\Sigma_j^{-1}(X-M_j)] \quad (1)$$

여기서, d는 벡터X 의 차수, M_j는 혼합 모델의 센터이다.

각 iteration 마다 특정 패턴에서 하나의 모델에 소속될 조건부 확률이 계속 변화되다가 혼합모델의 센터 M_j 이 수렴 단계에 접어들면 각 조건부 확률이 변함없이 일정한 값을 가지게 된다. 그러나 가우시안 혼합모델의 정형화에는 많은 연산량이 요구된다.

본 논문에서는 가우시안 혼합모델의 빠른 정형화를 위해 패턴인식에서 많이 사용되는 K-Means Algorithm[3]을 이용했다.

K-Means Algorithm은 Euclidian Distance를 이용해 혼합모델의 수만큼 클러스터링을 수행한다. 수행후 얻어지는 센터와 Covariance Matrix로부터 (1)식을 이용해서 혼합모델들의 정형화를 얻을 수 있다. 따라서 Covariance Matrix를 단 한번 연산하기 때문에 기존의 방법보다는 정형화에 소요되는 연산량은 대폭 감소하게 된다.

3. 퍼지 정보량

가우시안 혼합모델이 정형화 되면, 다음 단계로 퍼지 정보량을 이용해서 클러스터링을 수행하게 되는데, 먼저 퍼지 정보량에 대해서 살펴보면,

$$H(X) = -\sum_i P(i|X)lnP(i|X) + (1-P(i|X))ln(1-P(i|X)) \quad (2)$$

(2)식의 퍼지 정보량은 de Luca 와 Termini [2] 가 제안했는데, 기존의 Shannon의 정보량을 이용했을때, [0,1] 사이의 값을 가지는 퍼지 집합에 대해서 편향성을 가진다. 그러나 퍼지 정보량은 0.5를 기준으로 대칭적인 구조를 가지기 때문에 퍼지 집합의 특성을 잘 반영할 수 있다.

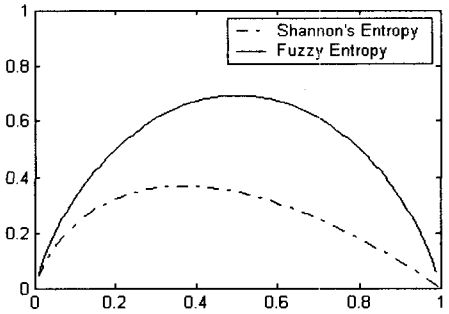


그림1. 정보량 특성의 비교

그림1에서 Shannon의 정보량은 0.4 부근에서 정보량의 최대값을 가지는 반면 퍼지 정보량은 0.5에서 최대값을 가진다. 즉 [0,1] 값을 가지는 퍼지 집합 대해서 어떠한 클러스터에 소속되지 않은 값 0.5에서 정보량의 최대치가 형성된다. (2)식에서 P(i|X)가 0 또는 1쪽으로 공평하게 minimize를 시키기 위해서는 퍼지 정보량을 이용하는게 더 효과적이다. 퍼지 정보량의 식은 Gradient의 변화가 더 급격하므로, 정보량 최소화에 수렴되는 시간은 빨라질 것이다.

4. 퍼지 정보량 최소화

클러스터링을 위해 가우시안 혼합모델과 각 클러스터들의 상관관계는 다음과 같다.

특정 데이타의 한 패턴이 모든 가우시안 혼합모델에 소속될 확률의 합이 1이다. 또한 모든 혼합 모델이 하나의 클러스터에 소속될 조건부 확률의 합도 1이다. 이는 다음과 같이 수식할 수 있다.

$$1 = \sum_{j=1}^{M} P^*(j|X) \quad (3)$$

$$P(i|X) = \sum_j W_{ij} P^*(j|X) \quad (4)$$

$$1 = \sum_i P(i|X) = \sum_i \sum_j W_{ij} P^*(j|X) \quad (5)$$

여기서 j는 모델의 index이고, i는 클러스터의 index이다. 행열 W_{ij}는 j번째 혼합 모델이 i번째 클러스터에 소속도를 나타내는 Weight 성분이다. 따라서 열 방향으로의 합은 1이 된다.

FEC에서 두번째 단계는 MCP[1]에서 제시하는 것처럼 (2)식의 정보량을 최소화시키는 방향으로 가중치 행렬 W_{ij}를 보정시켜야 한다.

W_{ij}은 다음과 같이 파라미터 θ_{ij}로 정의 할 수 있다.

$$W_{ij} = \frac{exp(\theta_{ij})}{\sum_{i'} exp(\theta_{i'j})} \quad (6)$$

여기서 θ_{ij}의 초기값은 1/C (C는 클러스터의 개수)을 가지고 다음과 같이 보정시킨다.

$$\theta_{ij(new)} = \theta_{ij} + \alpha \frac{\partial H}{\partial \theta_{ij}} \quad (7)$$

(4)식에서 $\frac{\partial H}{\partial \theta_{ij}}$는 Chain rule에 따라 다음과 같이 전개할 수 있다.

$$\frac{\partial H}{\partial \theta_{ij}} = \sum_{i'} \frac{\partial H}{\partial W_{ij}} \frac{\partial W_{ij}}{\partial \theta_{i'j}} \quad (8)$$

$$\frac{\partial W_{ij}}{\partial \theta_{i'j}} = W_{ij}\delta_{i'j} - W_{i'j}W_{ij} \quad (9)$$

($\delta_{i'i} = 1$ if $i' = i$ and 0 otherwise)

(2),(4)식 으로부터 다음을 구할 수 있다.

$$\frac{\partial H}{\partial W_{i'j}} = \frac{1}{N} \sum_{n} \frac{\partial H(X_n)}{\partial P(i'|X_n)} \frac{\partial P(i'|X_n)}{\partial W_{i'j}} \quad (10)$$

여기서 $P(i'|X_n)$는 (4)식에서 구할 수 있다. 위 수식들을 이용해서 W_{ij} 를 Gradient Descendent 방법으로 구할 수 있다.

본 논문에서 제시하는 FEC(Fuzzy Entropy Clustering)은 다음과 같은 절차와 같다.

Step 1. GMM Fitting

(Using K-means Algorithm)

Intialize Random k GMM center M_k

Do : Compare Distance each M_k

And Update nearist Center M_k

Until : ($|M_k^{(n-1)} - M_k^{(n)}| < \eta$)

Compute Covariance Matrix Σ_k

Using Σ_k, M_k compute $P^*(j|X)$ from(1)

Step 2. Fuzzy Entropy Minimization

(Using Gradient Descendent Method)

Intialize $\theta_{ij} = 1/C$, α(Learning rate)

Do : Compute $\frac{\partial H}{\partial \theta_{ij}}$

Update $\theta_{ij(new)}$(7) and W_{ij}(6)

Compute $P(i|X)$(4), $H(X)$(2)

Until : ($|H(X)_{new} - H(X)_{old}| < \eta$)

5. 실험결과

본 절에서는 여러 예제들을 통해 제시된 방법의 유용성을 보이도록 한다. 각 예제들은 기존의 FCM(Fuzzy C Means Algorithm)과 본 논문에서 제시하는 FEC(Fuzzy Entropy Clustering)와 비교하고 , 연산 수행의 측면에서 기존의 MCP와 FEC에 결과를 비교, 분석하도록 한다.

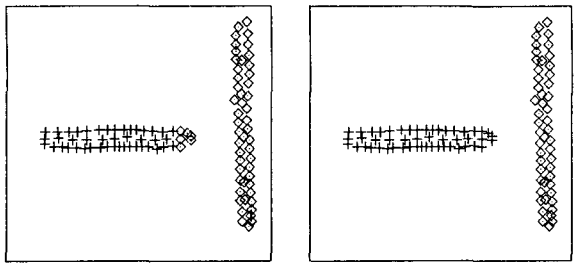
실험에 사용된 데이터는 Line Data(100개의 데이타 2개의 Feature) 와 Ring Data(120개의 데이터 2개의 Feature) 그리고 Iris Data(150개의 데이터, 2개 Feature)를 사용했다.

혼합모델의 개수는 데이터의 복잡도를 고려해 Lines Data는 10개의 혼합모델을 사용하였고, Ring Data는 12개의 혼합모델을 사용하였고, Iris Data는 20개의 혼합모델을 이용했다.

그림 2, 3, 4는 각각의 실험 데이터에 대해서 FCM(Fuzzy C Means) 알고리즘과 FEC의 결과를 비교한 그림이다.

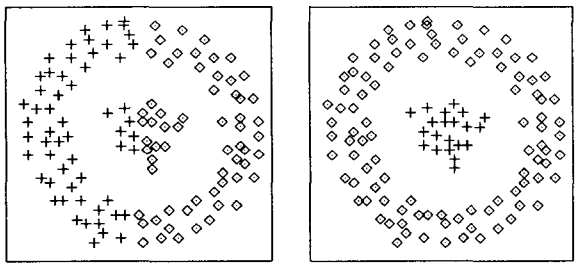
그림 2, 3에서처럼 특정한 모양에 대해서는 FCM은 인식률이 떨어지는데 비해 FEC는 모양에 상관없이 좋은 결과를 나타냈다. 그림 4. Iris 데이터에 대해서도 FCM은 94.5퍼센트의 결과를 FEC는 96퍼센트의 결과를 보였다. 결국 인식률의 측면에서는 FCM 보다는FEC가 높은 인식률을 나타냈다.

그리고, 그림 5는 Iteration에 따른 정보량의 변화를 나타낸다. 그림에서 볼 수 있듯이 FEC의 수렴속도가 MCP보다 훨씬 빠르다는 것을 알 수 있다. 이는 2절에서 살펴본 것처럼 퍼지 정보량 식(2) 의 특성을 반영한 것이라 할 수 있다.



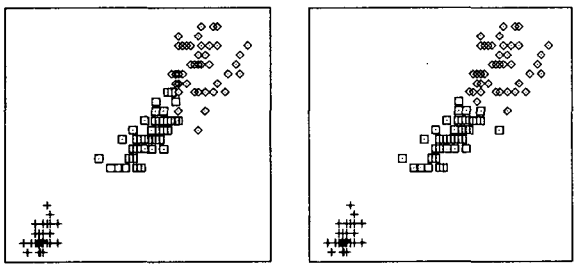
(a) (b)

그림 2. Line Data (a) FCM (b)FEC



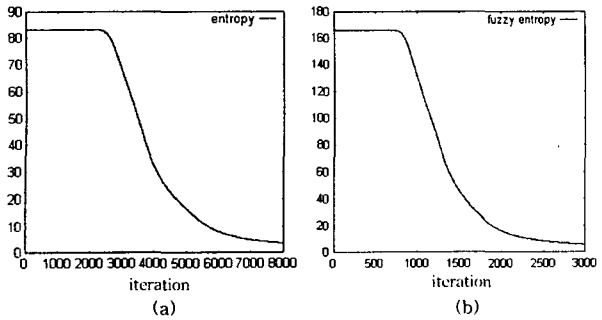
(a) (b)

그림 3. Ring Data (a) FCM (b)FEC



(a) (b)

그림 4. Iris Data (a) FCM (b)FEC



(a) (b)

그림 5. Ring Data의 Iteration별 정보량 변화
(a) MCP using Shannon's Entropy
(b) FEC using Fuzzy Entropy

5. 결론 및 향후과제

본 논문에서는 기존의 프로토타입 기반의 클러스터링 보다 높은 인식률을 유지하면서 MCP(Maximum Certainty Partitioning)보다 높은 연산수행능력 꽤하고자 퍼지 정보량(Fuzzy Entropy)을 도입했다. 실험데이터로터 클러스터링 수행한 결과, 기존의 방법보다 빠른 연산으로 클러스터링을 수행함을 보였다.

또, 가우시안 혼합모델(GMM)의 빠른 정형화를 위해서 K-means Algorithm을 사용한 FEC(Fuzzy Entropy Clustering)을 제시했다.

향후 이 FEC에서의 개선점은 가우시안 혼합모델을 정형화 하는 방법을 탈피해서 퍼지 맴버쉽을 이용한 혼합모델을 만들고, 이 혼합모델의 개수를 줄임과 동시에 성능향상을 꽤하는 방향으로 나아갈 것이다.

6. 참고문헌

[1] S. Roberts, R. Everson, I. Rezek, "Maxi -mum certainty data partitioning" ,*Pattern Recognition.*, vol. 33, pp. 833-839, 2000

[2] A de Luca and S. Termini, "A definition of a non probablilistic entropy in the setting of fuzzy set theory." *Inf. Contr.*, vol. 20, pp. 301-312, 1972

[3] J. T. Tou, R. C. Gonzalez, *Pattern Recognition Principles.* Addison-Wesley,. pp. 94-97

[4] S. Roberts, D. Husmeier, I. Rezek, W. Penny, Bayesian approaches to Gaussian mixture modelling, *IEEE Trans. Pattern Anal.* Mach. Intell. 20(11)(1998) 1133-1142

[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd. Academic Press. pp. 508-533