The Korean Communications
in Statistics Vol. 9, No. 3, 2002
pp. 715-723

A Density-based Clustering Method

Sung Mahn Ahn[1] and Sung Wook Baik[2]

Abstract

This paper is to show a clustering application of a density estimation method that utilizes the Gaussian mixture model. We define "closeness measure" as a clustering criterion to see how close given two Gaussian components are. Closeness measure is defined as the ratio of log likelihood between two Gaussian components. According to simulations using artificial data, the clustering algorithm turned out to be very powerful in that it can correctly determine clusters in complex situations, and very flexible in that it can produce different sizes of clusters based on different threshold values

Keywords : clustering method, closeness measure, Gaussian mixture model, maximum penalized likelihood, EM algorithm

1. 서론

최근 몇 년 전부터 정규혼합모형을 사용한 클러스터링에 대한 새로운 방법들이 등장하기 시작했고 이들 방법들은 여러 다양한 분야에 적용되고 있다. 그 대표적인 분야들은 주로 염색체(genome) (Baldi, 2000)와 유전 인자(gene) 정보 분석(McLachlan et. al., 2000; Medvedovic and Sivaganesan, 2002; Yeung et. al., 2001)을 연구하는 유전정보학(Bioinformatics) 분야와 컴퓨터 비전 분야이다. 컴퓨터 비전 분야에서는 얼굴 인식을 위한 얼굴영상 이미지분석 (Sadeghi et. al., 2001), 비디오의 연속적인 일련의 영상 이미지분석(Nitsuwat et. al., 2000), 의료영상 이미지분석 (Suckling et. al., 1999), 이미지 세그멘테이션(image segmenation) (Pauwels et. al., 2001) 등에서 응용되고 있다.

이 논문은 정규혼합모형을 사용한 클러스터링의 새로운 방법을 제시하며, 위에서 언급한 여러 분야에 응용될 수 있다.

2. 정규혼합모형과 최대벌점가능도 추정

1) Assistant Professor, School of Management Information Systems, Kookmin University, Seoul 136-702. E-mail: sahn@kookmin.ac.kr
2) Assistant Professor, Department of Digital Contents, Sejong University, Seoul 143-747. E-mail: E-mail : sbaik@sejong.ac.kr

정규혼합모형은 다음과 같다.

$$f(x;\ \pi, \mu, \Sigma) = \sum_{j=1}^{g} \pi_j N(x; \mu_j, \Sigma_j) \tag{1}$$

(1)에 있는 모형에서 모수를 추정하려면 반복적 절차(iterative procedure)를 이용해야만 가능한데, 일반적으로 많이 사용하는 기법은 EM 알고리즘이다 (Dempster et. al., 1977). EM알고리즘에 따른 로그 우도함수는 다음과 같다 (Titterington et. al., 1985).

$$\lambda_c(\Theta) = \sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \log[\pi_j N(x_i; \mu_j, \Sigma_j)] \tag{2}$$

(2)에서 z_{ij}는 0 혹은 1의 값을 가지는 변수로서 데이터 x_i가 j번째 성분에 의해 만들어 졌으면 1, 그렇지 않으면 0의 값을 가진다. 그러나 이 변수는 관측되어지지 못하므로 EM알고리즘에 의해 기대값으로 추정된다 (Dempster et. al., 1977).

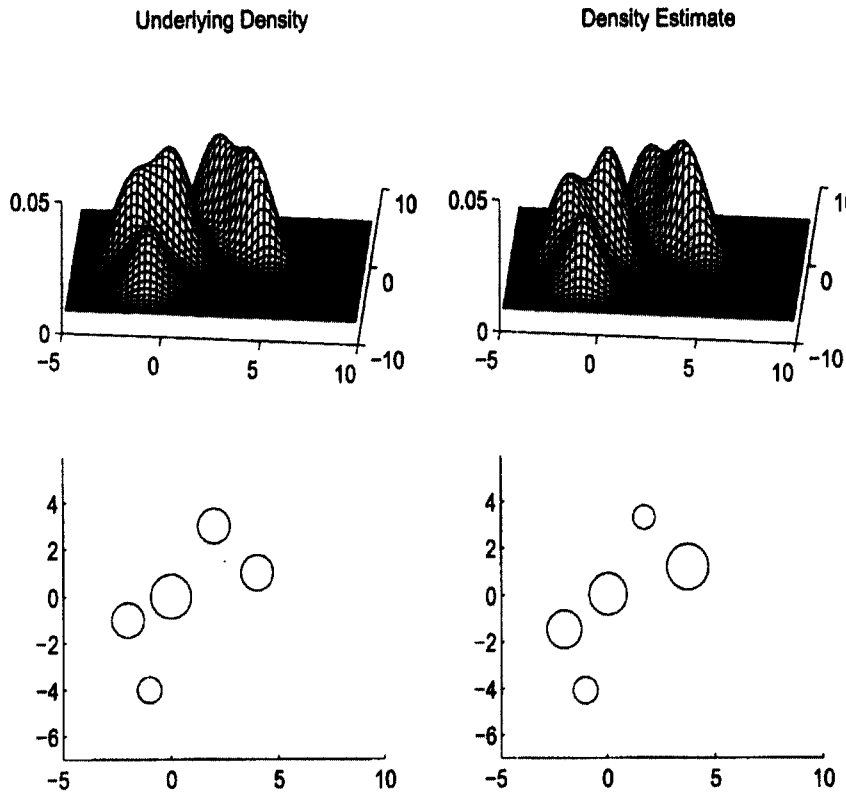
한편 정규혼합모형에서 까다로운 문제는 성분의 수((1)에서 g)를 추정해야 하는 것인데 이를 위해서는 앞의 로그우도함수 (2)에 벌점항목을 추가하여 벌점가능도 추정을 하는 것이 효율적이다. 이를 위하여 새로운 벌점가능도를 정의한 것이 다음과 같다 (Ahn, 2001).

$$\sum_{i=1}^{n} \sum_{j=1}^{g} z_{ij} \log[\pi_j N(x_i; \mu_j, \Sigma_j)] + \lambda n \sum_{j=1}^{g} (\alpha - 1) \log \pi_j \tag{3}$$

추가된 벌점항목을 간단히 설명하면 다음과 같다. EM알고리즘을 사용하면 (3)을 최대화하는 모수의 값을 찾아내는데, (3)에서 벌점항목은 π_j의 값이 0일 때 최대가 된다 $(0<\alpha<1,\ \lambda>0)$. 이때 $\pi_j=0$이 의미하는 것은 j번째 성분의 영향력이 0이라는 뜻이며, 결국은 (3)을 최대화함으로써 가능한 많은 수의 성분을 제거하게 된다. (모든 π_j가 0이 되지 않는 이유는 $\Sigma\pi_j=1$이라는 제약조건이 있기 때문이다) 더욱 자세한 설명은 Ahn (2001)에서 찾을 수 있다. 본 논문에서는 클러스터링을 위하여 (3)을 이용한 확률분포 추정을 사용할 것이다.

3. 클러스터링 알고리즘

이 절에서는 앞에서 설명한 확률분포 추정을 이용한 클러스터링 방법을 제안하고자 한다. 이를 위하여 본 논문에서는 2차원 예제를 사용하여 이해를 돕고자 한다. 물론 3차원 이상의 경우에도 적용이 가능하지만 그림으로 표현하기가 어렵기 때문에 2차원 예제로 한정하였다. 우선 첫번째 예제를 보자. <그림1>에서는 5개의 성분으로 이루어진 확률분포를 보여주고 있다.



<그림 1> 확률분포(좌)와 추정확률분포(우)

<그림1>에서 좌측에 있는 분포로부터 데이터가 만들어 졌으며, 그 데이터를 사용하여 앞의 확률분포 추정방법으로 추정된 것이 우측의 분포이다. <그림1>의 하단에 있는 그래프는 상단에 있는 확률분포의 성분구조를 설명해 주고 있다 (Solka et. al., 1995). 각각의 원은 하나의 성분을 설명하고 있는데, 각 원의 중심점의 x-y좌표는 해당하는 성분의 평균 벡터를 나타낸다. 원의 크기는 π의 상대적 크기를 나타낸다. 그러므로 그림에서의 확률분포는 5개의 성분으로 구성되어 있으며 각 성분은 정규분포이며 그 평균과 공분산행렬은 다음과 같다.

1. $N\left(\begin{bmatrix}-1\\-4\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$
2. $N\left(\begin{bmatrix}-2\\-1.5\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$
3. $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$
4. $N\left(\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}1 & -5\\-5 & 1\end{bmatrix}\right)$

5. $N\left(\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}1 & -5\\-5 & 1\end{bmatrix}\right)$

각 성분에 대한 혼합계수는 각각 0.15, 0.2, 0.25, 0.2, 그리고 0.2이다 위와 같은 분포를 구성하는 데이터가 몇 개의 군집을 이루는지에 대해서는 이견이 있을 수 있다. 즉 군집화 기준을 어떻게 설정하는가에 따라 2,3, 혹은 5개의 군집을 이룬다고 말할 수 있다. <그림1>의 경우에도 좌측에 있는 확률분포로부터 4000개의 무작위 데이터를 만든 뒤에 앞 절에 있는 방법을 이용하여 우측과 같은 추정확률분포를 만들어 내었다. 계속해서 추정확률분포를 이용한 군집화 기법을 제안한다.

3.1 군집화 기준

본 논문에서는 군집화 기준으로 두 성분에 대한 "근접척도"를 사용하는데, 그것은 다음과 같이 정의된다.

$$\text{근접척도} = \frac{\text{두 성분의 로그 우도값}}{\text{한 개의 성분을 이용한 모형의 로그 우도 값}}$$

근접척도는 대부분의 경우에 1보다 작은 값으로서 '두개의 성분으로 추정된 데이터를 하나의 성분으로 설명할 수 있는 정도'로 해석할 수 있다. 즉 근접척도의 값이 1에 가까우면 두개의 추정된 성분은 동일한 군집에 속한다고 볼 수 있다. 이 근접척도는 한 쌍의 성분에 대한 척도이기 때문에 모두 $\binom{g}{2} = \frac{g(g-1)}{2}$ 개의 값을 계산할 수 있다. 근접척도를 계산하는 과정을 순서대로 설명하면 다음과 같다.

1. 확률분포를 추정한다
2. 사후확률 (z_{ij})에 근거하여 각각의 데이터를 개별 성분에 할당한다. 예를 들면, 난수를 사용하여 z_{ij}에 따라 할당한다.
3. 두개의 성분과 그에 할당된 데이터를 선택하여 로그 우도값을 계산한다
4. 3에서 사용한 데이터를 정규분포로 가정하고 로그우도값을 계산한다
5. 근접척도를 계산한다

$$\frac{\text{3에서 계산한 값}}{\text{4에서 계산한 값}}$$

6. 다른 쌍의 성분을 선택하여 3번으로 간다

위와 같은 방법으로 근접척도를 모두 계산한 것이 <표1>에 있다. <표1>에는 총 10개의 근접척도가 있으며, 값이 큰 것부터 위에서 아래로 나열되어 있다. 그 중 근접척도가 가장 큰 값을 가지는 성분의 쌍을 보면, 해당하는 성분은 5와4 이며, 이 때의 근접척도는 0.9923이다. 즉 본 논문에서 정의한 근접척도에 따르면 5번성분과 4번 성분이 서로 가장 가까우며 따라서 하나의 군집에

포함될 가능성이 제일 높다고 하겠다. 그러므로 <표1>에 의거하여 우리는 필요한 만큼의 군집을 만들어 낼 수 있다. 2개의 군집을 원하면 근접척도의 분계점(threshold)을 0.9443과 0.9820 사이의 값으로 하면 되고, 3개의 군집을 원하면 분계점을 0.9820과 0.9901 사이의 값으로 하면 된다. 근접척도의 서로 다른 분계점에 따라 달라질 수 있는 군집을 요약한 것이 <표2>이다. 한편 <표1>에서 근접척도의 값을 보면, 0.9820과 0.9443의 차이가 상대적으로 큰 것을 알 수 있다. 그러므로 이를 참고로 하여 적절한 분계점의 위치를 결정할 수도 있을 것이다. <그림1>에서 보면 육안으로 3개 정도의 군집이 있다고 말할 수도 있는데, 근접척도의 분계점을 0.9820과 0.9443사이의 값으로 하면 육안으로 판단한 3개의 군집과 일치하게 된다. 다음에는 2개의 군집을 가지는 확률분포를 사용한 예를 보도록 하겠다.

<표 1> 로그우도값과 근접척도

성분의 쌍	두개의 성분을 사용한 모형의 로그우도값	단일 성분을 사용한 모형의 로그우도값	근접척도
5&4	-4874.01	-4911.59	.9923
3&2	-5980.46	-6040.45	.9901
2&1	-4882.56	-4972.10	.9820
3&4	-5022.03	-5318.20	.9443
3&1	-5310.07	-5670.90	.9364
3&5	-6807.94	-7349.81	.9263
2&4	-4837.45	-5654.24	.8555
5&2	-6609.61	-7855.44	.8414
5&1	-5542.75	-6991.22	.7928
4&1	-3850.42	-4908.12	.7845

<표 2> 여러 개의 분계점에 따른 군집들

분계점	군집의 수	군집
> 0.9923	5	(1)(2)(3)(4)(5)
> 0.9901	4	(4,5)(1)(2)(3)
> 0.9820	3	(4,5)(2,3)(1)
> 0.9443	2	(4,5)(1,2,3)
< 0.9443	1	(1,2,3,4,5)

3.2 두개의 군집을 가지는 경우

여기서 사용한 예는 좀 극단적인 예이다. 이 분포는 총 18개의 2차원 정규분포 성분으로 구성된다. 그 중 17개는 혼합비율이 1/20이며 나머지 한 개는 3/20이며, 공분산행렬은 항등행렬로 모두

동일하다. 이를 그림으로 표시한 것이 <그림2>의 왼쪽에 위치한 것이다. 이 분포로부터 4000개의 무작위 데이터를 만들어서 확률분포 추정을 한 결과가 <그림2>의 우측에 있는 것이다. 이 경우에는 총 6개의 성분이 추정되었으며 각 성분의 모수는 다음과 같다.

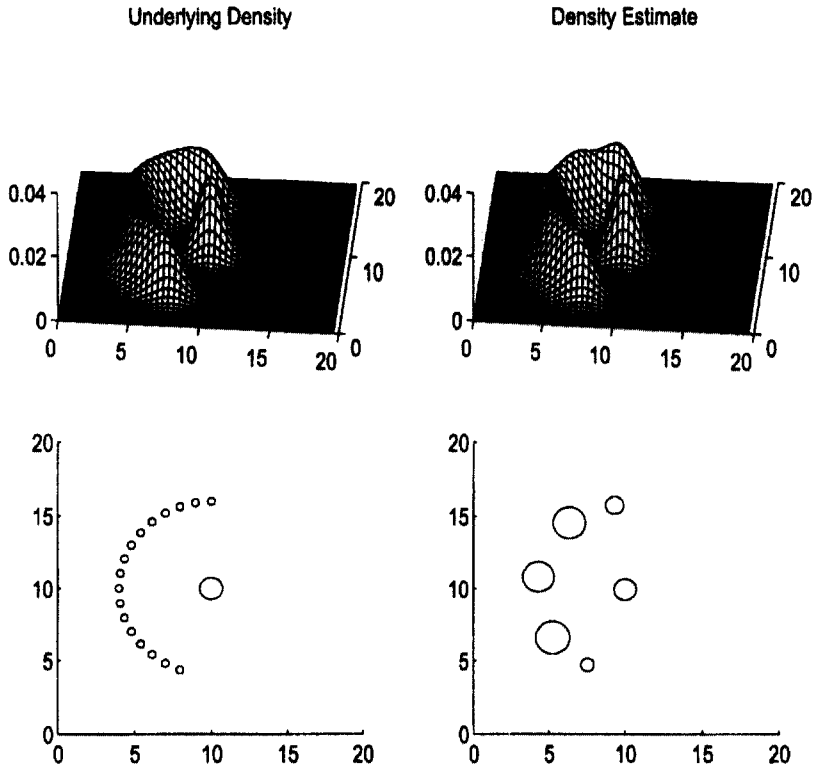
1. $\pi=0.15,\ \mu=\begin{bmatrix}10.02\\9.95\end{bmatrix},\ \Sigma=\begin{bmatrix}0.95 & -0.03\\-0.03 & 0.91\end{bmatrix}$
2. $\pi=0.09,\ \mu=\begin{bmatrix}7.54\\4.73\end{bmatrix},\ \Sigma=\begin{bmatrix}1.04 & -0.23\\-0.23 & 1.21\end{bmatrix}$
3. $\pi=0.23,\ \mu=\begin{bmatrix}5.26\\6.61\end{bmatrix},\ \Sigma=\begin{bmatrix}1.57 & -1.03\\-1.03 & 2.73\end{bmatrix}$
4. $\pi=0.21,\ \mu=\begin{bmatrix}4.32\\10.83\end{bmatrix},\ \Sigma=\begin{bmatrix}1.00 & 0.36\\0.36 & 1.00\end{bmatrix}$
5. $\pi=0.21,\ \mu=\begin{bmatrix}6.35\\14.56\end{bmatrix},\ \Sigma=\begin{bmatrix}2.16 & 1.18\\1.18 & 2.04\end{bmatrix}$
6. $\pi=0.12,\ \mu=\begin{bmatrix}9.32\\15.76\end{bmatrix},\ \Sigma=\begin{bmatrix}1.27 & 0.058\\0.058 & 0.93\end{bmatrix}$

위의 6개의 성분을 사용하여 근접척도를 구한 결과가 <표3>에 있다. <표3>에서 보면 근접척도의 간격이 0.9739와 0.9071이 매우 큼을 알 수 있다. 그러므로 분계점을 그 사이로 결정하게 되면, 우리는 2개의 군집을 만들어 낼 수 있다. 즉 (2,3,4,5,6)이 하나의 군집이고 (1)이 하나의 군집인데, 첫번째 군집은 그림2에서 반원모양의 형상을 하는 군집이며, 두 번째 군집은 오른쪽에 있는 작은 원 모양을 하는 군집임을 알 수 있다.

4. 요약

정규혼합모형에서 가장 어려운 문제는 성분의 수를 어떻게 추정하느냐 인데, 이를 위하여 지금까지 다양한 방법이 제시되어 왔으며 (Barron and Cover, 1991; Rissanen, 1987; Schwarz, 1978), 현재도 연구가 지속되는 형편이다. 정규혼합모형을 군집화에 응용한 논문인 Sadeghi et. al. (2001)에서는 예측적검증기법 (predictive validation method)이란 방법을 사용하였으며, Pauwels et. al. (2001)에서는 데이터와 비교하면서 성분의 수를 점진적으로 늘려가는 방법을 사용하였고, 또 Nitsuwat et. al. (2000)는 Agglomerate Gaussian Mixture Decomposition이란 방법을 사용하였다.

본 논문에서는 Ahn (2001)이 제시한 방법을 사용하여 개별 성분을 추정하였으며, 군집화를 위해서는 '근접척도'를 정의하여 군집을 결정하는 하나의 방법을 제시하였다. 그리고 제시한 방법의 성과를 알아보기 위해 2개의 예제를 사용하였다. 첫번째 예는 군집의 분리가 상대적으로 명확히 구분되지 않는 경우이며, 이 경우에는 분석가가 필요한 군집의 수에 따라서 분계점을 적절히 조정하여 사용할 수 있고, 두 번째 예는 군집이 상대적으로 잘 분리되는 경우인데, 이 때에는 근접척도들 간의 크기가 상대적으로 큰 곳을 분기점으로 결정하여 군집을 분리하면 되었다.



<그림 2> 확률분포(좌)와 추정확률분포(우)

<표 3> 로그우도값과 근접척도

성분	두개의 성분을 사용한 모형의 로그우도값	한 개의 성분을 사용한 모형의 로그우도값	근접척도
3&2	-4487.47	-4519.25	.9930
5&6	-4816.70	-4882.28	.9866
5&4	-6505.12	-6653.49	.9777
3&4	-6788.01	-6970.01	.9739
2&4	-4574.20	-5042.83	.9071
4&6	-5131.28	-5745.85	.8930
3&5	-7234.58	-8133.73	.8895
2&1	-3196.09	-3628.85	.8807
3&1	-5735.41	-6544.60	.8764
...	...	...	...

참고문헌

[1] Ahn, S.M. (2001). A Penalized Likelihood Method for Model Complexity Reduction in Gaussian Mixture Density, *The Korean Communications in Statistics*, vol. 8, no.1, 173-184.

[2] Baldi, P. (2000). On the Convergence of a Clustering Algorithm for Protein-coding Regions in Microbial Genomes, *Bioinformatics (Oxford, England)*, vol. 16, Issue 4, 367-371

[3] Barron, A. R. and Cover, T. M. (1991). Minimum Complexity Density Estimation, *IEEE trans. On Information Theory*, 37(4), 1034-1054

[4] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of Royal Statistical Society(B)*, 39, 1-38.

[5] McLachlan, G. J., Bean, R. W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data, *Bioinformatics (Oxford, England)*, vol. 18, Issue 3, 413-422

[6] Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles, *Bioinformatics (Oxford, England)*, Vol. 18, Issue 9, 1194-1206

[7] Nitsuwat, S., Jin, J.S., and Hudson, H.M. (2000). Motion-based video segmentation using fuzzy clustering and classical mixture model, *Proceedings of International Conference on Image Processing*, vol. 1, 300-303

[8] Pauwels, E.J., Frederix, G., and Caenen, G. (2001). Image segmentation based on statistically principled clustering, *Proceedings of 2001 International Conference on Image Processing*, vol. 2, 66-69

[9] Rissanen, J. (1987). Stochastic Complexity, *Journal of The Royal Statistical Society, Series B*, 49(3), 223-239 and 252-265

[10] Sadeghi, M., Kittler, J., and Messer, K. (2001). Spatial clustering of pixels in the mouth area of face images, *Proceedings of 11th International Conference on Image Analysis and Processing*, 36-41

[11] Schwarz, G. (1978). Estimating the Dimension of a Model, *The Annals of Statistics*, 6(2), 461-464

[12] Suckling, J., Sigmundsson, T., Greenwood, K., and Bullmore, E T (1999). A modified fuzzy clustering algorithm for operator independent brain tissue classification of dual echo MR images, *Magnetic Resonance Imaging*, vol. 17, Issue 7, 1065-1076

[13] Solka, J. L., Poston, W. L., and Wegman, E. J. (1995). A Visualization Technique for Studying the Iterative Estimation of Mixture Densities, *Journal of Computational and Graphical Statistics*, 4, 180-198.

[14] Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley.

[15] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data, *Bioinformatics (Oxford, England)*, vol. 17, Issue 10, 977-987

[2002년 3월 접수, 2002년 11월 채택]