Journal of Korea Multimedia Society Vol. 11, No. 9, September 2008(pp. 1302-1309)

스펙트럼 군집화에서 블록 대각 형태의 유사도 행렬 구성

허경용[†], 김광백[††], 우영운[†††]

요 약

K-means나 퍼지 군집화와 같은 전통적인 군집화 기법들이 원형(prototype)을 기반으로 하고 볼록한 형태의 집단들에 적합한 반면, 스펙트럼 군집화(spectral clustering)는 국부적인 유사성을 기반으로 전역적인 집단을 찾아내는 기법으로 오목한 형태의 집단들에도 적용할 수 있어 커널을 기반으로 하는 SVM과 더불어 각광을 받고 있다. 하지만 SVM이 그러하듯이 스펙트럼 군집화에서도 커널의 폭은 성능에 지대한 영향을 끼치는 요인으로, 이를 결정하기 위한 다양한 방법이 시도되었지만 여전히 휴리스틱에 의존하는 실정이다. 이 논문에서는 유사도 행렬이 보다 명백한 블록 대각 형태를 가지도록 하기 위해 국부적인 커널의 폭을 거리 히스토그램을 바탕으로 적응적으로 결정하는 방법을 제시한다. 제안한 방법은 스펙트럼 군집화에 사용되는 유사도 행렬(affinity matrix)이 블록 형태의 대각 행렬을 이룰 때 이상적인 결과를 낸다는 사실에 기반하고 있으며, 이를 위해서 전통적인 유클리디안 거리와 무작위 행보 거리(random walk distance)를 함께 사용한다. 제안한 방법은 기존의 방법들에서 사용하는 유사도 행렬에 비해 명확한 블록 대각 행렬을 나타내고 있음을 실험 결과를 통해 확인할 수 있다.

Magnifying Block Diagonal Structure for Spectral Clustering

Gyeongyong Heo[†], Kwang-Baek Kim[††], Young Woon Woo[†††]

ABSTRACT

Traditional clustering methods, like *k*-means or fuzzy clustering, are prototype-based methods which are applicable only to convex clusters. On the other hand, spectral clustering tries to find clusters only using local similarity information. Its ability to handle concave clusters has gained the popularity recent years together with support vector machine (SVM) which is a kernel-based classification method. However, as is in SVM, the kernel width plays an important role and has a great impact on the result. Several methods are proposed to decide it automatically, it is still determined based on heuristics. In this paper, we proposed an adaptive method deciding the kernel width based on distance histogram. The proposed method is motivated by the fact that the affinity matrix should be formed into a block diagonal matrix to generate the best result. We use the tradition Euclidean distance together with the random walk distance, which make it possible to form a more apparent block diagonal affinity matrix. Experimental results show that the proposed method generates more clear block structured affinity matrix than the existing one does.

Key words: Spectral Clustering(스펙트럼 군집화), Kernel Width(커널 폭), Affinity Matrix(유사도 행렬), Block Diagonal Matrix(블록 대각 행렬)

※ 교신저자(Corresponding Author) : 우영운, 주소 : 부산시 부산진구 가야동 산 24(614-714), 전화 : 051)890-1712, FAX : 051)890-2706, E-mail : ywwoo@deu.ac.kr
접수일 : 2007년 12월 24일, 완료일 : 2008년 8월 4일
[†] 준회원, Dept. of Computer and Information Sci. and Eng., University of Florida, 박사과정
(E-mail : hgycap@hotmail.com)
[††] 종신회원, 신라대학교 컴퓨터정보공학부 교수
(E-mail : gbkim@silla.ac.kr)
[†††] 정회원, 동의대학교 멀티미디어공학과 교수

1. 서 론

군집화는 주어진 N 개의 데이터 포인트들을 K 개의 집단에 할당하는 비지도 학습의 일종으로, 이 때 동일한 집단 내 포인트들의 유사도는 크고, 서로 다른 집단에 속하는 포인트들의 유사도는 작게 만드는 것이 그 기본 목표이다. 군집화에서 집단의 개수 K를 정하는 문제는 또 다른 문제로, 이 논문에서는 알려진 것으로 가정한다. 최근 스펙트럼 군집화 기법은 선형 분리가 불가능한 데이터를 선형 분리가 가능한 데이터로 변환하여 처리할 수 있는 특성으로 인해 SVM과 같은 다른 커널 기반의 방법들과 더불어 기계 학습의 영역에서 각광을 받고 있다. 스펙트럼 군집화는 주어진 데이터 포인트들 사이의 유사도(affinity)를 나타내는 유사도 행렬 A를 구성하고, 이 행렬의 고유값과 고유벡터를 이용하여 원본 데이터를 군집화 하는 기법이다.

유사도 행렬 A는 일반적으로 그래프의 인접 행렬로 해석될 수 있으며 따라서 스펙트럼 군집화는 "유사도 행렬을 구하는 과정"과 "그래프 분할의 과정"으로 나누어 생각해 볼 수 있다. 후자의 경우 그래프 분할과 군집화의 연관성은 널리 연구되어 이론적인 정립이 이루어졌지만[1], 전자의 경우 상대적으로 널리 연구되고 있지 않다. 하지만 행렬 섭동 이론(matrix perturbation theory)에 따르면[2], 스펙트럼 군집화는 유사도 행렬이 블록 대각 행렬을 이룰 때 최적의 결과를 보여주는 것으로 알려져 있으므로 이 논문에서는 명백한 블록 대각 형태를 띠는 유사도 행렬을 구하는 방법에 주목한다.

주어진 데이터 포인트들 사이의 유사도는 일반적으로 가우시안 커널을 이용하여 식 (1)과 같이 계산된다.

$$A(x_i,x_j)=\exp\left(-\frac{d^2(x_i,x_j)}{\sigma^2}\right) \quad (1)$$

이 때 $d(\cdot,\cdot)$는 두 데이터 포인트 사이의 거리로 일반적으로 유클리디안 거리로 주어지며, σ는 커널의 폭을 나타낸다. 이 중 σ는 군집화 결과에 지대한 영향을 미치며, 서로 다른 값의 σ를 실험한 후 목적 함수를 최적화하는 값을 결정하는 방법[2]이 제시되었지만, 실제 데이터의 경우 전역적으로 하나의 σ를 결정하는 일이 불가능한 경우도 있다.

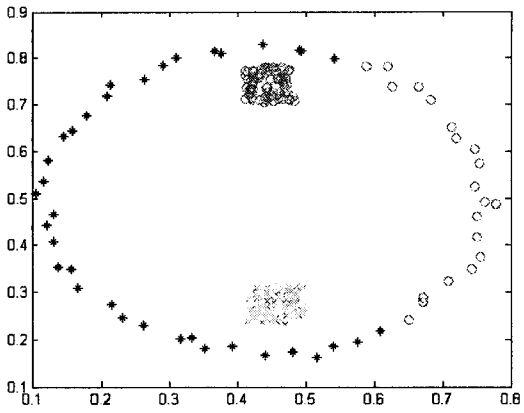

그림 1. 밀도가 다른 집단들의 군집화 결과

그림 1은 서로 다른 밀도를 가지는 3개의 집단으로 구성된 데이터를 군집화한 결과로, k-means 뿐만이 아니라 전역적인 σ 값으로도 정확한 결과를 얻을 수 없는 경우를 나타낸다. 따라서 이 논문에서는 각 데이터 포인트 x_i에 대해 국부적인 σ_i 값을 결정하는 방법을 제시한다. 국부적으로 σ_i를 결정하는 방법은 대칭인 유사도 행렬을 바탕으로 한 방법[3]과 비대칭인 행렬을 바탕으로 한 방법[4]으로 나눌 수 있으며, 이 논문에서는 대칭인 유사도 행렬을 사용한다. 또한 이상적인 경우 스펙트럼 군집화의 결과는 유사도 행렬이 블록 대각 행렬일 때 최적의 결과를 나타내므로, 유사도 행렬이 가능한 이러한 형태를 띠도록 하기 위해 무작위 행보 거리[5-7]를 함께 사용하여 유사도 행렬의 블록 구조가 명확하게 나타나도록 하였다.

이 논문의 구성은 다음과 같다. 먼저 2장에서는 스펙트럼 군집화에 대해 개략적으로 살펴보고, 3장에서는 무작위 행보 거리에 대해 살펴본다. 이 논문에서 제안한 방법은 4장에서 설명하며, 5장에서는 실험을 통해 제안한 방법이 기존의 방법에 비해 블록 대각 구조가 명확히 나타남을 보인다. 결론 및 향후 연구 방향에 대해서는 6장에서 언급한다.

2. 스펙트럼 군집화(Spectral Clustering)

스펙트럼 군집화는 데이터 군집화를 그래프 분할 문제로 바꾼 것으로, 특정 목적 함수를 최적화하는 그래프 분할을 찾는 문제로 귀결된다. 이 때 널리 사용되는 목적 함수 중의 하나는 정규화된 분할(normalized cut)[8]로, 식 (2)와 같이 정의된다.

$$Ncut(C_1,...,C_k)=\sum_{i=1}^{k}\frac{cut(C_i,\overline{C_i})}{vol(C_i)} \quad (2)$$

식 (2)는 주어진 데이터를 k 개의 집단으로 나누는 경우로, 각 데이터 포인트들 사이의 유사도(affinity)는 그래프에서 각 노드들 사이를 연결하는 에지의 강도로 설정되며, $cut(\cdot,\cdot)$은 식 (3)과 같이 그래프에서 두 노드 집합을 연결하는 에지의 강도 합으로 정의되고, $vol(\cdot)$은 식 (4)와 같이 한 노드 집합 내에서 각 노드들을 연결하는 에지의 강도 합으로 정의된다.

$$cut(A,B)=\sum_{i\in A,j\in B} w_{ij} \quad (3)$$

$$vol(A)=\sum_{i\in A,j\in A} w_{ij} \quad (4)$$

식 (3), (4)에서 w_{ij}는 노드 i와 노드 j를 연결하는 에지의 강도로, 두 데이터 포인트 사이의 유사도를 나타낸다. 기본적으로 군집화는 어떤 데이터 포인트가 특정 집단에 속하는 경우와 속하지 않는 경우의 두 가지로 판별되며, 따라서 군집화 결과를 나타내는 지시 벡터(indicator vector)는 2가지의 값만을 가질 수 있다. 하지만 이 경우 식 (2)를 최소화시키는 값을 구하는 것은 NP-Hard 문제이므로, 지시 벡터가 연속적인 값을 가질 수 있도록 허용함으로써 보다 쉽게 근사해를 구할 수 있다. Rayleigh 정리에 의해 n개의 데이터 포인트들을 k 집단에 할당하는 $n \times k$ 지시 행렬(indicator matrix)은 정규화된 라플라시안(Laplacian) 행렬의 고유벡터들로 구성할 수 있다. 보다 자세한 사항은 [1,2,8]을 참고하면 된다. 그림 2는 본 논문에서 사용한 스펙트럼 군집화 알고리즘을 나타낸 것이다[8].

입력 : 유사도 행렬 A, 집단 개수 k

1. 라플라시안 행렬(L)을 구한다.

$$D_{ij}=\begin{cases}\sum_{k=1}^{n}A_{ik} & \text{if } i=j\\ 0 & otherwise\end{cases}$$

$L = D - A$

2. 일반화된 eigenproblem $Lv=\lambda Dv$를 풀어 큰 값을 갖는 k 개의 고유값에 해당하는 고유벡터, $v_1,v_2,...,v_k$를 구한다.
3. $v_1,v_2,...,v_k$로 $n \times k$ 행렬 R을 구성한다.
4. 행렬 R의 각 행을 새로운 데이터 포인트로 하여 k-means 알고리즘을 통해 k 개 집단을 찾아낸다. R 행렬의 i 번째 행은 i 번째 데이터 포인트에 해당한다.

그림 2. 스펙트럼 군집화 알고리즘

그림 2의 알고리즘은 계산된 유사도 행렬을 입력으로 하고 있다. 따라서 이 논문은 그림 2의 알고리즘에 입력할 유사도 행렬이 블록 대각 형태를 띠도록 만드는 것에 중점을 둔다.

3. 무작위 행보 거리

앞 장에서 언급한 바와 같이 이 논문은 유사도 행렬이 블록 대각 형태를 가지도록 하는 것을 목표로 한다. 그림 3은 유클리디안 거리와 식 (1)의 가우시안 커널을 사용하여 유사도 행렬을 구한 예를 보이고 있다. 이 때 커널의 폭은 각 데이터 포인트에서 ($2D$ + 1)번째로 가까운 데이터 포인트까지의 거리로 모든 데이터 포인트에 대해 다르게 설정되었으며, D는 데이터의 차원을 나타낸다.

그림 3에서 알 수 있듯이, 유사도 행렬은 각 데이터 포인트에 인접한 몇 개의 데이터 포인트들만이 영이 아닌 유사도 값을 가지게 된다. 또한 이 결과는 커널의 폭 설정에 매우 민감하여 그 결과가 달라질 수 있다. 이러한 민감성은 유클리디안 거리가 절대적인 위치에 영향을 받기 때문이다. 그림 3-A에서 포인트 1과 2는 포인트 3에서 서로 다른 유클리디안 거리에 있지만 직관적으로는 비슷한 정도의 유사도를 가져야 한다. 따라서 식 (1)의 가우시안 커널로는 이를 만족시킬 수 없다. 본 논문에서는 그림 3-B의 유사도

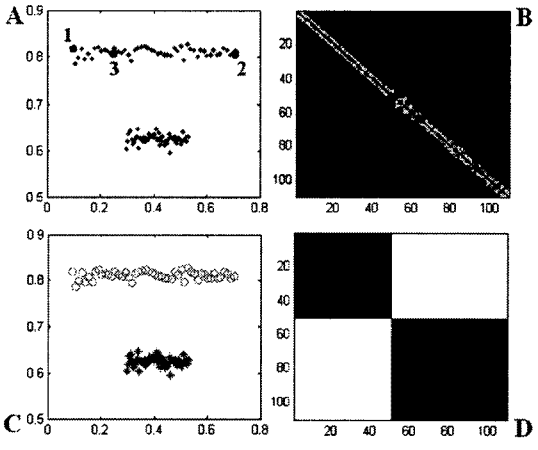

그림 3. A : 샘플 데이터, B : 유사도 행렬, C : 군집화 결과, D : 무작위 행보 거리

행렬을 바탕으로 무작위 행보 거리를 구하고, 이 거리로부터 최종 유사도 행렬을 구한다.

무작위 행보 거리는 Klein과 Randic[7]에 의해 처음 정립되었으며, 저항 거리(resistance distance)라고도 불린다. 이는 마코프(Markov) 체인의 특성에 기반한 2가지 거리 측도, 평균 최초 도달 시간, $m(j|i)$와 평균 왕복 시간(commute time), $n(i, j)$를 정의한다[5]. 전자는 노드 i에서 출발하여 노드 j에 처음으로 도달하기까지의 평균 시간을 나타내며, 후자는 노드 j에 도달한 이후 다시 노드 i로 되돌아가기까지의 평균 시간을 나타낸다. 즉 $n(i, j) = m(j|i) + m(i|j)$를 만족한다. 이 때 전자는 비대칭 이지만, $m(j|i) \neq m(i|j)$, 후자는 대칭인 점이 다르다. 전통적인 최단 경로 거리가 잡음에 민감하고 경로의 수를 고려하지 않는 반면, 이들 측도는 두 노드를 연결하는 경로의 수가 많아지거나 경로 중 하나의 시간이 작은 경우 줄어드는 특성이 있다. 따라서 그림 3-A에서 포인트 1과 2가 유클리디안 거리로는 포인트 3에서 차이가 큰 값을 가지지만, 무작위 행보 거리로는 유사한 값을 가지게 된다. 그림 3-D는 평균 왕복 시간을 나타낸 행렬로 블록 구조가 명확히 드러나고 있음을 볼 수 있다. 이들 측도들은 그래프 라플라시안의 의사 역행렬로부터 식 (5), (6)을 통해 구할 수 있다[5].

$$m(j|i) = \sum_{k=1}^{n} (l_{ik}^{+} - l_{ij}^{+} - l_{jk}^{+} + l_{jj}^{+}) D_{jj} \tag{5}$$

$$n(i,j) = V_G (l_{ii}^{+} + l_{jj}^{+} - 2l_{ij}^{+}) \tag{6}$$

이 때 l_{mn}^{+}은 그래프 라플라시안의 의사 역행렬 L^{+}의 m 행 n 열 원소를 나타내며, V_G는 그래프의 부피(volume)로 식 (7)로 정의된다.

$$V_G = \sum_{k=1}^{n} D_{kk} \tag{7}$$

식 (5)와 (6)의 차이 중 하나는 대칭과 비대칭의 차이이다. 일반적으로 거리 행렬은 대칭 행렬이어야 하므로 식 (5)를 사용하는 경우에는 최종 거리 행렬을 대칭으로 만들어줄 필요가 있다. 이 논문에서는 식 (6)의 평균 왕복 시간을 거리로 사용하였다.

그림 4는 샘플 데이터에 대해 유클리디안 거리를 구하고 이를 유사도로 변환한 후(이 과정은 다음 장에서 자세히 다룬다.), 유사도 행렬로부터 무작위 행보 거리를 구한 예를 보여주고 있다. 일반적으로 무작위 행보 거리 행렬은 유클리디안 거리 행렬에 비해 블록 대각 행렬 구조가 명확함을 알 수 있다. 하지만 한 가지 문제점은 대각선이 아닌 영역에서도 일부 영이 아닌 값을 가지는 것이다. 이는 무작위 행보 거리로부터 최종 유사도 행렬을 구하는 과정에서 최적의 커널 폭을 선택함으로써 완화할 수 있다.

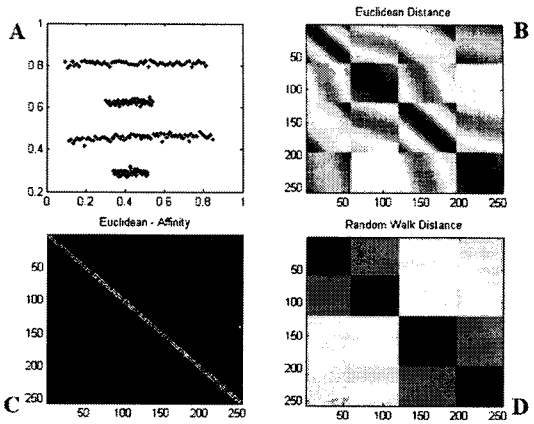

그림 4. A : 샘플 데이터, B : 유클리디안 거리 행렬, C : 유사도 행렬, D : 무작위 행보 거리 행렬

4. 적응적 커널 폭 결정

데이터 포인트 집합이 주어지는 경우 유사도 행렬은 3단계를 거쳐 계산된다. 먼저 식 (8)을 통해 유클리디안 유사도 행렬(A_1)을 구하고 (그림 4-C), 식 (6)과 A_1을 바탕으로 무작위 행보 거리 행렬(d_{RW})을 구한다 (그림 4-D). 끝으로 각 데이터 포인트에 대해 커널의 폭을 적응적으로 결정함으로써 최종 유사도 행렬(A_2)을 구한다.

이 과정에서 2번의 커널 폭을 결정할 필요가 있다. 하지만 커널의 폭을 선택하는 목적은 서로 다르다. 첫 번째 단계에서는 국부적으로 이웃한 점들을 찾아내어 무작위 행보 거리 계산을 위한 전이 행렬(transition matrix)을 만드는 것이 목표이므로, 앞 장에서 언급한 바와 같이 데이터의 차원(D)을 고려하여 ($2D$ + 1)번째로 가까운 데이터 포인트까지의 거리로써 커널 폭을 결정하는 것으로 충분하다[9]. 하지만 이 값으로 식 (1)의 가우시안 커널을 적용하는 경우 구해지는 유사도 행렬은 비대칭을 이루게 된다. 따라서 식 (8)을 통해 대칭 행렬을 이루도록 하였다[3].

$$A_1(x_i,x_j)=\exp\left(-\frac{d^2(x_i,x_j)}{\sigma_i\sigma_j}\right) \quad (8)$$

이 때 $d(\cdot,\cdot)$는 두 점 사이의 유클리디안 거리를 나타내며 σ_i는 데이터 포인트 x_i에서 $(2D+1)$번째로 가까운 데이터 포인트까지의 거리를 나타낸다.

세 번째 단계에서 커널의 폭은 각 데이터 포인트들의 이웃(neighbor)들을 파악하여 대각선이 아닌 영역에서의 유사도 행렬 값을 영으로 만들기 위해 필요하며, 동일한 집단에 속하는 모든 점들을 포함할 수 있도록 하는 블록의 크기를 추정하는 것으로 볼 수 있다. 이를 위해 본 논문에서는 무작위 행보 거리 히스토그램(distance histogram)과 가우시안 혼합 모델을 사용하였다. 거리 히스토그램은 하나의 데이터 포인트에서 자신을 제외한 모든 점들까지의 거리를 히스토그램으로 나타낸 것으로, 그림 5에서 그 예를 보이고 있다.

그림 5의 데이터 x에서 유클리디안 거리 히스토그램을 구성하는 경우에는 그림 5와 같은 형태가 나타나지 않고 하나의 피크만이 나타난다. 이는 x와 동일한 클러스터에 속하는 데이터 포인트들까지의 거리뿐만이 아니라 다른 클러스터에 속하는 데이터 포인트들까지의 거리가 동일한 분포를 따르기 때문이다. 하지만 무작위 행보 거리 히스토그램을 구하는 경우에는, 이상적인 경우 클러스터의 개수에 상관없이 2개의 피크를 나타낸다.

즉, 동일한 클러스터에 속하는 데이터 포인트들까지의 거리 분포가 하나의 가우시안 요소를 형성하고, 다른 클러스터에 속하는 데이터 포인트들까지의 거리가 다른 하나를 형성한다. 이는 그림 4의 B(유클리디안 거리)와 D(무작위 행보 거리)를 통해 명확하게 알 수 있다. 따라서 무작위 행보 거리 히스토그램은 2개의 요소를 가지는 가우시안 혼합 모델로 나타낼 수 있으며, 세 번째 단계에서의 커널 폭은 EM 알고리즘[10]을 통해 구한 평균 및 분산값의 함수로 나타낼 수 있다.

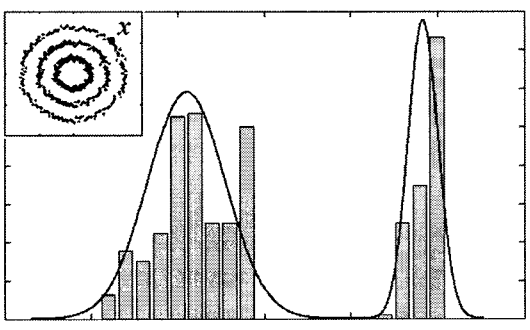

그림 5. 데이터 포인트 x의 거리 히스토그램과 가우시안 혼합 모델

$$\sigma_i' = f(\mu_1,\mu_2,\Sigma_1,\Sigma_2)=\frac{\mu_1+\mu_2}{2} \quad (9)$$

식 (9)는 본 논문에서 커널 폭을 결정하기 위해 사용한 식으로, μ와 Σ는 EM 알고리즘을 통해 구한 각 가우시안 요소의 평균과 분산을 나타낸다. 식 (9)에서와 같이 두 피크 위치의 평균값으로 각 데이터 포인트의 커널 폭을 설정함으로써 모든 실험 데이터에서 만족할만한 결과를 얻었으며, 다양한 실험 데이터 및 실제 데이터를 통해 보다 일반적인 커널 폭 결정 함수 f를 결정하기 위한 연구가 진행 중에 있다. 이처럼 각 데이터의 커널 폭이 결정되면, 식 (10)을 통해 최종 유사도 행렬 A_2를 계산한다.

$$A_2(x_i,x_j)=\exp\left(-\frac{d_{RW}^2(x_i,x_j)}{\sigma_i'\sigma_j'}\right) \quad (10)$$

식 (10)에서 $d_{RW}(\cdot,\cdot)$는 두 점 사이의 무작위 행보 거리를 나타내며 σ'은 식 (9)를 통해 결정된 커널의 폭을 나타낸다.

5. 실험 결과

본 논문에서 제안한 지역적 커널 폭 결정은 이미 여러 기존의 방법이 존재하지만[3,4], 이 논문에서는 [4]와 마찬가지로 유사도 행렬이 명확한 블록 구조를 가지도록 하는 것을 목표로 한다. 테스트한 여러 방법들 중 [4]의 방법이 모든 실험 데이터에서 오류 없는 군집화 결과를 보여주었으며, 그 목표 중 하나 역시 블록 구조에 두고 있으므로, 이 논문에서는 [4]의 방법과 비교하였다.

블록 구조를 평가하는 방법은 여러 가지가 있을 수 있지만, 기본적인 조건은 첫째, 블록 내의 유사도 값들은 균일한 값을 가져야 하며, 둘째, 서로 다른 블록 내의 값들은 그 차이가 커야 한다는 것이다. 따라서 균일성을 비교하기 위해 식 (11)을 계산하고 비교하였다. 식 (11)은 동일한 집단에 속하는 포인트들 사이의 유사도 분포 범위와 서로 다른 집단에 속하는 포인트들 사이의 유사도 분포 범위가 겹치는 정도를

계산한 것으로, μ_{within}과 σ_{within}은 동일한 집단에 속하는 포인트들 사이의 유사도 평균과 분산을 나타내고, $\mu_{between}$과 $\sigma_{between}$은 서로 다른 집단에 속하는 포인트들 사이의 유사도 평균과 분산을 나타낸다. 식 (11)의 값이 음의 값을 가지는 것은 두 범위가 서로 겹치는 것을 나타낸다. 이 때 분포 범위는 유사도 히스토그램이 정규분포를 따른다는 가정 하에 95.4%의 분포를 포함하는 $(\mu \pm 2\sigma)$ 범위를 사용하였다.

$$Z=(\mu_{within}-2\sigma_{within})-(\mu_{between}+2\sigma_{between}) \quad (11)$$

표 1은 Fisher의 방법[4]과 본 논문에서 제안한 방법의 Z 값을 비교한 것으로, 사용한 실험 데이터는 그림 6과 같다.

표 1에서 알 수 있듯이 Fisher의 방법은 모든 데이터 집합에서 음의 값을 보였다. 즉, 집단 내 유사도와 집단 간 유사도가 일부 겹치고 있는 것으로, 블록 구조가 명확하지 않음을 나타낸다. 제안한 방법의 경우 데이터 집합 1이나 3과 같이 집단들이 명확히 분리되어 있는 경우에는 큰 값을 가지며, 데이터 집합 5와 같이 서로 다른 밀도의 집단들에 대해서도 큰 값을 보여주었다. 하지만 데이터 집합 2와 같이 잡음이 첨가되는 경우에는 분리도가 떨어지고 음의 값을 가지게 된다. 이는 Fisher의 방법에서 사용한 거리가 식 (5)의 평균 최초 도달 시간(average first passage time)인데 비해 이 논문에서는 식 (6)의 평균 왕복 시간(average commute time)을 사용한데서 일부 원인을 찾을 수 있다. 식 (5)를 사용하는 경우 최종 유사도 행렬은 비대칭을 이루게 되므로 최종 유사도를 $A_{ij}=\min(A'_{ij},A'_{ji})$과 같이 대칭을 이루도록 변환하였다. 이는 잡음의 영향을 줄이는데 효과적이다. 실제로 식 (5)를 사용하고 제안한 방법으로 σ를 구한 경우 -0.125의 분리도 값을 얻을 수 있었다. 하지만 식 (5)는 식 (6)에 비해 보다 많은 연산을 필요로 하고, 대칭 행렬을 사용하는 것이 일반적이므로 본 논문에서는 식 (6)을 사용하였다.

또 한 가지 다른 점으로는, 그림 6-2에 대한 유사도 행렬(A_2)에서 동일한 집단 내에 속하는 점들의 유사도 값이 보다 넓은 범위에 분포하는 반면, 서로 다

표 1. 유사도 행렬의 분리 정도

Data Set	Fisher et al.	Heo et al.
1	-0.152	0.939
2	-0.199	-0.311
3	-0.179	0.752
4	-0.213	-0.027
5	-0.063	0.291

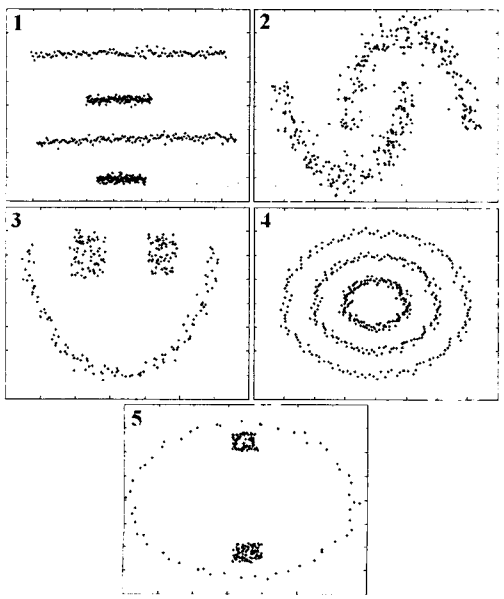

그림 6. 샘플 데이터 집합

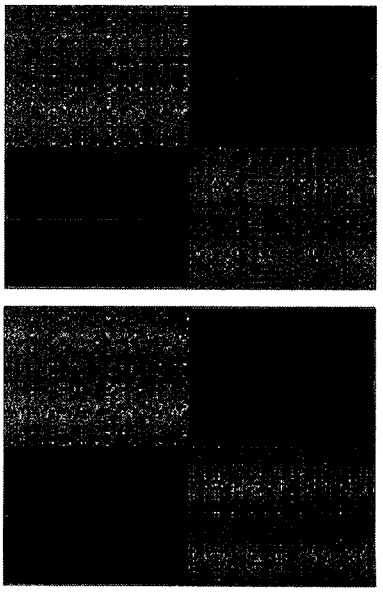

그림 7. Fisher 등의 방법(위)과 제안한 방법(아래)의 유사도 행렬

른 집단에 속하는 점들 사이의 유사도는 대부분 0으로 좁은 범위에 피크를 이루며 분포했다는 점이다. 하지만 Fisher의 방법은 전자가 상대적으로 좁은 범위에 분포한 반면, 후자는 상대적으로 넓은 범위에 고르게 분포하였다. 즉 Z 값에 영향을 크게 미치는 σ_{within} 값이 제안한 방법에서 크게 나타났다.

이러한 사실은 그림 7에서 명확히 알 수 있으며, 이는 두 방법의 근본적인 차이점이기도 하다. 즉, 제안한 방법이 서로 다른 집단에 속하는 점들 사이의 유사도를 효과적으로 억제하는 반면 동일한 집단에 속하는 점들 사이의 유사도를 강화시키지 못하는 경향이 있으며, 특히 잡음이 첨가된 경우 블록 구조가 명확해지지 않는다. 이처럼 잡음이 첨가된 경우 블록 구조를 강화시키는 방안에 대해서는 현재 연구 중에 있으며, 이는 잡음을 효과적으로 억제하는 방법이 포함된다.

6. 결 론

본 논문에서는 기존 스펙트럼 군집화 방법에서 중요한 역할을 하는 커널의 폭(σ)을 정확하게 결정하여 유사도 행렬이 블록 대각 행렬을 가지도록 하는 방법을 제안하였다. 제안한 방법은 세 단계로 이루어져 있으며, 첫 번째 단계에서 근접한 이웃들만을 고려한 커널의 폭을 결정하고, 두 번째 단계에서는 유클리디안 거리를 바탕으로 무작위 행보 거리를 구하며, 끝으로 거리 히스토그램을 바탕으로 커널 폭을 계산한다.

제안한 방법은 실험 데이터에서 만족할만한 군집화 결과를 보여 주었으며, 기존의 방법과 비교하였을 때, 보다 좋은 분리도를 보여주었다. 즉 유사도 행렬이 보다 명확한 블록 대각 형태를 가짐을 알 수 있다. 하지만 잡음이 증가함에 따라 분리 정도는 나빠지며 이를 위해 현재 잡음을 효과적으로 제거하는 방법과 잡음 정도를 이용하여 각 데이터 포인트에 가중치를 주는 방법 등에 관해 연구 중에 있다. 또한 Fisher의 방법에서 사용한 비대칭 거리의 적용 및 보다 안정적인 커널 폭 결정 함수 f를 결정하는 방법에 관한 연구도 진행 중이다.

참 고 문 헌

[1] U. von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, Vol.17, No.4, pp. 395-416, 2007.

[2] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, Vol.14, pp. 849-856, 2002.

[3] L. Zelnik-Manor and P. Perona, "Self-Tuning Spectral Clustering," *Advances in Neural Information Processing Systems*, Vol.17, pp. 1601-608, 2004.

[4] Igor Fischer and Jan Poland, "Amplifying the Block Matrix Structure for Spectral Clustering," *Proceedings of the 14th Annual Machine Learning Conference of Belgium and the Netherlands*, pp. 21-28, 2005.

[5] Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens, "Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation," *IEEE Transactions on Knowledge and Data Engineering*, Vol.19, No.3, pp. 355-369, 2007.

[6] Bernd Fischer, Volker Roth, and Joachim M. Buhmann, "Clustering with the Connectivity Kernel," *Advances in Neural Information Processing Systems*, Vol.16, pp. 89-96, 2003.

[7] D.J. Klein and M. Randic, "Resistance Distance," *Journal of Mathematical Chemistry*, Vol.12, pp. 81-85, 1993.

[8] Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, No.8, pp. 888-905, 2000.

[9] Igor Fischer and Jan Poland, "New Methods for Spectral Clustering," *Technical Report No. IDSIA-12-04*, Dalle Molle Institute for Artificial Intelligence, 2004.

[10] Arthur Dempster, Nan Laird and Donald Rubin, "Maximum likelihood from imcomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, Vol.39, No.1, pp. 1-38, 1977.

허 경 용

1994년 2월 연세대학교 전자공학과 공학사
1996년 8월 연세대학교 전자공학과 공학석사
2004년 9월~현재 Dept. of Comp. and Info. Sci. and Eng., University of Florida
관심분야 : 영상처리, Machine Learning, Bayesian Network

김 광 백

1999년 부산대학교 전자계산학과 이학박사
1997년~현재 신라대학교 컴퓨터정보공학부 교수
2005년~현재 한국멀티미디어학회 학술이사 및 논문지 편집위원
2005년~현재 한국해양정보통신학회 논문지 편집위원
관심분야 : Neural Networks, Image Processing, Fuzzy Logic, Medical Imaging and Biomedical System, Support Vector Machines

우 영 운

1991년 8월 연세대학교 전자공학과 공학석사
1997년 8월 연세대학교 전자공학과 공학박사
1997년 9월~현재 동의대학교 멀티미디어공학과 부교수
관심분야 : 패턴인식, 지식표현, 퍼지이론, 의료정보