Proceedings of KFIS 2002 Spring Conference, 2002. 5. 25

선형 행렬 부등식을 이용한 타원형 클러스터링 알고리즘

Hyper-ellipsoidal clustering algorithm using Linear Matrix Inequality

이한성* · 박주영 · 박대희***

*고려대학교 전산학과 · **고려대학교 제어계측공학과

Hansung Lee* , Jooyoung Park and Daihee Park***

*Dept.of computer science, Korea University

**Dept.of control and instrumentation engineering, Korea University

E-mail : mohan@korea.ac.kr

ABSTRACT

본 논문에서는 타원형 클러스터링을 위한 거리측정 함수로써 변형된 가우시안 커널 함수를 사용하며, 주어진 클러스터링 문제를 각 타원형 클러스터의 체적을 최소화하는 문제로 해석하고 이를 선형행렬 부등식 기법 중 하나인 고유값 문제로 변환하여 최적화하는 새로운 알고리즘을 제안한다.

Keyword : hyper-ellipsoidal clustering, modified gaussian kernel function, LMI, EVP

1. 서론

문헌고찰에 의하면, 대부분의 분할 클러스터링(partitional clustering) 알고리즘은 거리 측정 함수로써 유클리디안 거리(Euclidean distance)를 사용하고 있다. 그러나 유클리디안 거리에 기반한 클러스터링은 거리 측도의 성격상 비슷한 크기의 구형 분포를 갖는 잘 분리되어 있는 데이터에 한해서만 효과적인 클러스터링이 수행된다는 단점이 있다[1][4]. 따라서, 이러한 문제를 해결하기 위한 한가지 대안으로, 구형이 아닌 타원형에 기반한 많은 클러스터링 알고리즘들이 제안되었다[1][3][4][5]. 타원형의 클러스터를 생성하기 위한 거리 측도로써, 마하라노비스 거리(mahalanobis distance)가 일반적으로 사용되고 있으나, 마하라노비스 거리를 클러스터링에 직접 사용하면 타원형의 클러스터링을 얻을 수 없음이 왕(Wang)등의 연구결과에 의해 이미 증명되었다[2][4]. 또한, 그들은 문제 해결의 대안으로 가우시안 커널 함수 등을 거리 측정 함수로 사용할 것을 제안하고 있으나, 구체적인 알고리즘을 제시하지는 않고 있다.

따라서, 본 논문에서는 타원형 클러스터링을 위한 거리측도 함수로써 "변형된 가우시안 커널 함수"(modified gaussian kernel function)를 사용하여 타원형 클러스터를 실체화 하고자 한다. 또한, 주어진 클러스터링 문제를 각 클러스터의 체적들을 최소화하는 문제로 해석하여 선형행렬 부등식(LMI : linear matrix inequality)기법 중 하나인 고유값 문제(EVP : eigen-value problem)로 변형하여 최적화하는 새로운 기법의 클러스터링 알고리즘을 제안하고자 한다.

2. 타원형 클러스터링 알고리즘

d-차원의 n 개의 입력 패턴, $\mathbf{x} = \{ x_i \in R^d \}_{i=1}^n$ 이 주어졌을 때, 클러스터링은 정의된 코스트 함수(cost function), $E_c(P)$을 최소화하는 방향으로 각 패턴들을 해당 분할행렬(partition matrix), $P=\{P_{ik} | P_{ik} \in \{0,1\};\ i=1,2,\ldots,n;\ k=1,2,\ldots,c\}$로 할당하는 문제로 정의된다.

$$P = \arg(P) \min E_c(P) \qquad (1)$$

여기서, c 는 클러스터의 개수이고, x_i가 k 클러스터에 속하면 $P_{ik}=1$, 그렇지 않을 경우 $P_{ik}=0$이 된다. 각 입력 패턴들은 반드시 하나의 클러스터에만 할당된다. 따라서, 다음 식 $\sum_{k=1}^{c} P_{ik}=1;\ i=1,2,\ldots,n$ 을 만족하여야만 한다.

입력 패턴 x_i와 클러스터 k의 중점 m_k 사이의 거리측정 함수 $D(x_i, m_k)$에 의해 정의되는 클러스터링 코스트 함수는 다음과 같다.

$$E_c(P)=\sum_{i=1}^{n}\sum_{k=1}^{c} P_{ik}D(x_i, m_k) \tag{2}$$

본 논문에서는 타원형 클러스터링을 위한 거리측정 함수로서 "변형된 가우시안 커널 함수"를 사용한다.

$$D(x_i, m_k)=(1-\lambda)\cdot(x_i-m_k)^T Q_k^{-1}(x_i-m_k) + \lambda\cdot\log(|Q_k|) \tag{3}$$

여기서, m_k와 Q_k는 각각 클러스터 k 의 중점 및 의사 공분산 행렬(pseudo covariance matrix)이다. λ는 가우시안 커널 함수의 첫 번째 항과 두 번째 항의 가중치를 조정하는 변수이다. 식 (3)의 첫 번째 항은 마하라노비스 거리를 나타내고 있으며, 두 번째 항은 의사 공분산 행렬 Q_k에 의해 표현되는 타원형 클러스터 k 의 체적이다. 따라서, 변형된 가우시안 커널 함수를 이용한 클러스터링 코스트 함수는 아래의 식으로 정의된다.

$$E_c(P)=\sum_{i=1}^{n}\sum_{k=1}^{c} P_{ik}[(1-\lambda)\cdot(x_i-m_k)^T Q_k^{-1}(x_i-m_k) + \lambda\cdot\log(|Q_k|)] \tag{4}$$

코스트 함수 $E_c(P)$의 최적화를 위한 필요 조건, $\frac{\partial E_c(P)}{\partial m_k^T}=0$ 으로부터 다음 식 $\sum_{i=1}^{n} P_{ik}Q_k^{-1}m_k=\sum_{i=1}^{n} P_{ik}Q_k^{-1}x_i$ 을 얻을 수 있으며, 변수 i 에 대하여, m_k와 Q_k^{-1}는 상수이므로, 클러스터 k의 중점 m_k을 식 (5)과 같이 구할 수 있다.

$$m_k=\frac{\sum_{i=1}^{n} P_{ik}x_i}{\sum_{i=1}^{n} P_{ik}} \tag{5}$$

정리 1: 변형된 가우시안 커널 함수를 이용한 클러스터링 코스트 함수는 다음과 같이 정리된다.

$$E_c(P)\cong\sum_{k=1}^{c}(n_k\log(|Q_k|)) \tag{6}$$

여기서, n_k 는 클러스터 k 에 속해있는 패턴들의 개수이다.

증명)

[2]의 *정리 1*로부터 $\sum_{i=1}^{n}(x_i-m_k)^T Q_k^{-1}(x_i-m_k)=d\cdot(n-1)$을 얻을 수 있다. 클러스터 k 에 속해있는 패턴들의 개수를 $n_k=\sum_{i=1}^{n} P_{ik}$; $\sum_{k=1}^{c} n_k=n$ 라고 정의하면, [2]의 *정리 2*로부터 다음의 식 $\sum_{i=1}^{n} P_{ik}D(x_i, m_k)=n_k\cdot(1-\lambda)\cdot d\cdot(n_k-1)+\lambda\cdot n_k\cdot\log(|Q_k|)$ 을 유도할 수 있다. 따라서,

$$E_c(P)=\sum_{k=1}^{c}[n_k\cdot(1-\lambda)\cdot d\cdot(n_k-1)+\lambda\cdot n_k\cdot\log(|Q_k|)]$$
$$=n\cdot(1-\lambda)\cdot d\cdot(n-c)+\sum_{k=1}^{c}[\lambda\cdot n_k\cdot\log(|Q_k|)]$$

위 식에서 n, d, c 그리고 λ는 상수임으로,

$$E_c(P)\cong\sum_{k=1}^{c}(n_k\log(|Q_k|))$$ ▣

Remark : 정리 1은 가우시안 커널 함수를 이용한 타원형 클러스터링 문제가 각 클러스터의 체적을 최소화하는 문제임을 보여준다. 각 클러스터의 체적들을 최소화한다는 것은 클러스터의 응집도를 최대화한다는 동치의 의미로 해석된다.

각 클러스터에 속한 패턴의 개수가 정해진 경우, 각 클러스터의 체적을 최소로 하는 Q_k는 공분산 행렬의 정의에 의해 아래의 조건에 만족하도록 구한다.

1) Q_k는 양 한정(positive definite) 행렬이며, 대칭(symmetric) 행렬이다.

2) 클러스터 k 에 속한 모든 입력 패턴은 Q_k 에 의해 표현되어지는 타원 안에 속해야 한다.

위의 문제정의는 아래의 식으로 표현될 수 있다.

$$\arg(Q_k)\min\log(|Q_k|) \tag{7}$$
subject to

$Q_k > 0$,

$(x_j - m_k)^T Q_k^{-1}(x_j - m_k) < 1$

; $k=1,2,\ldots,c$, $j=1,2,\ldots,n_k$

슈어 컴플리먼트(the schur complement) 에 의해 식 (7)은 Q_k를 변수로 하는 전형적인 선형 행렬 부등식 문제중 하나인 볼록 문제(CP : convex problem)로 변환된다.

$$\min \log(|Q_k|) \quad (8)$$

subject to

$Q_k > 0$,

$$\begin{bmatrix} 1 & (x_j - m_k)^T \\ (x_j - m_k) & Q_k \end{bmatrix} > 0$$

; $k=1,2,\ldots,c$, $j=1,2,\ldots,n_k$

Q_k가 대칭 행렬이고, 양 한정 행렬이라는 성질로부터 스펙트랄 분해(spectral decomposition) , $|Q_k| = |\cup^T||\cup||\wedge_k| = I_d|\wedge_k| = |\wedge_k| = \lambda_{1k} \cdot \lambda_{2k} \cdot \cdots \lambda_{dk}$,를 통해 식 (9)을 얻는다.

$$\begin{aligned} &\arg(Q_k)\min\log(|Q_k|) \\ &= \arg(Q_k)\min\log(\lambda_{1k}) + \log(\lambda_{2k}) + \cdots \quad (9) \\ &\quad + \log(\lambda_{dk});\ k=1, 2, \ldots, c \end{aligned}$$

의사 공분산 행렬 Q_k에 의해서 표현되는 타원체는 기본 축(principal axes)들의 방향이 Q_k의 고유벡터 u_i 에 의하여 결정되며, 타원체의 중점으로부터 각 축의 길이는 고유값의 루트, $\sqrt{\lambda_i}$ 값이 된다. 따라서, 식 (9)의 고유값을 작게 할수록 타원체의 체적 역시 작아지며, 체적을 최소화하는 문제는 각 고유값의 합을 최소화하는 문제로 근사(approximation) 표현이 가능하다. 따라서, 식 (8)의 볼록 문제는 다음의 고유값 문제로 근사 되어질 수 있다.

$$\min Trace(Q_k) \quad (10)$$

subject to

$Q_k > 0$,

$$\begin{bmatrix} 1 & (x_j - m_k)^T \\ (x_j - m_k) & Q_k \end{bmatrix} > 0$$

; $k=1,2,\ldots,c$, $j=1,2,\ldots,n_k$

본 논문에서 제안하는 클러스터링 알고리즘은 클러스터의 중점 식 (5)과 선형 행렬 부등식 (10)을 통한 반복알고리즘으로 다음과 같다.

STEP 1. 클러스터 개수 c를 지정하고, 임의의 c 개의 입력데이터를 각 클러스터의 중점으로 한다. $\sum_{k=1}^{c} P_{ik} = 1;\ i=1,2,\ldots,n$ 을 만족하도록 분할행렬을 초기화한다.

STEP 2. 해당 클래스의 새로운 중점 및 클러스터에 속한 데이터의 개수를 계산한다.

$$m_k = \frac{\sum_{i=1}^{n} P_{ik} x_i}{\sum_{i=1}^{n} P_{ik}}, \quad n_k = \sum_{i=1}^{n} P_{ik}$$

STEP 3. 의사 공분산 행렬 Q_k를 계산한다.

$$\min Trace(Q_k);\ k=1, 2, \ldots, c$$

subject to

$$Q_k > 0,$$

$$\begin{bmatrix} 1 & (x_j - m_k)^T \\ (x_j - m_k) & Q_k \end{bmatrix} > 0$$

STEP 4. 새로운 클러스터의 중점으로 분할행렬을 조정한다.

$P_{ik} = 1$ *if* $D(x_i, m_k) \le D(x_i, m_j)$

$P_{ik} = 0$ *otherwise*

STEP 5. 중점의 변화가 없으면 알고리즘을 종료하고 그렇지 않으면 STEP 2 - 4를 반복한다.

3. 실험 및 결과

본 논문에서 제안된 클러스터링 알고리즘의 타당성을 보이기 위하여 인공 데이터 및 Iris 데이터에 대하여 K-Means 알고리즘과 비교하였다. 본 논문에서 수행한 모든 실험 결과는 가중치 λ을 0.6으로 고정시켰을 때의 결과이다. 본 실험을 위한 알고리즘의 반복 수행 횟수는 100으로 제한하였으며, 반복 수행 횟수가 100을 넘을 경우는 마지막 100번째 결과를 선택하였다.

실험 1) 인공 데이터에 의한 실험

본 실험에서는 2개의 클러스터가 존재하는 14개의 데이터를 인공적으로 만들어 실험을 수행하였다. K-Means 알고리즘의 경우, 무작위 300회의 실험에서 단 두 회만 정상적으로 클러스터링 수행했으며, 제안된 알고리즘의 경우는 단 한번을 제외한 모든 경우에 성공적인 클러스터링을 수행하였다. 그림 1과 그림 2는 각 알고리즘에 의해 수행 된 결과를 보여주고 있다.

실험 2) Iris 데이터에 의한 실험

두 번째 실험은 클러스터링 알고리즘의 벤치마크 데이터로 알려진 Iris 데이터에 대한 실험을 수행하였다. 본 실험에서 K-Means 알고리즘과 제안된 알고리즘 모두 초기 중심 값에 무관하게 클러스터링을 수행하는 결과를 보였다. 그러나, 오판률에 있어서는 본 논문에서 제안된 알고리즘이 보다 좋은 성능을 보임을 알 수 있었다. 각 알고리즘의 클러스터링 결과는 표 1과 같다.

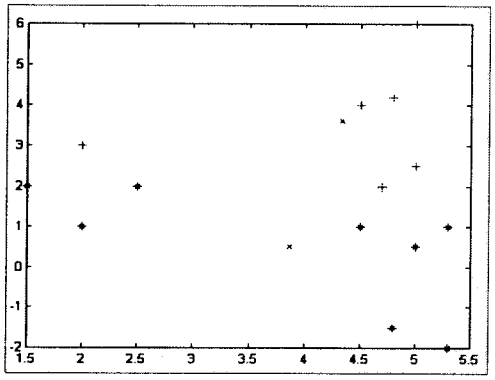

그림 1 K-Means 알고리즘에 의한 클러스터링

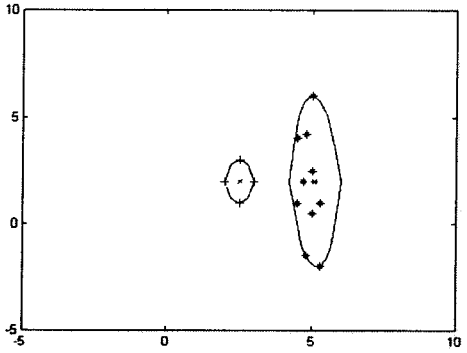

그림 2 제안된 알고리즘에 의한 클러스터링

표 1 Iris 데이터에 대한 분류 실험 결과

알고리즘 / 클래스	잘못 클러스터링 된 패턴의 개수	
	K-Means	제안된 알고리즘
Setosa	0	0
Versicolor	2	0
Virginica	14	11

4. 결론

본 논문에서, 우리는 클러스터링을 위하여 응집도가 고려된 변형된 가우시안 커널 함수를 이용하여 타원형 클러스터를 실체화했으며, 클러스터링 문제를 선형 행렬 부등식(LMI) 기법의 하나인 고유값 문제(EVP)로 변환하여 반복 알고리즘을 이용한 새로운 차원의 클러스터링 알고리즘을 제안하였다. 제안된 알고리즘은 다음과 같은 몇 가지 좋은 특성을 가지고 있다.

1) 제안된 알고리즘은 초기 조건에 비교적 덜 민감하다.
2) 고유값 문제로 얻어진 의사 공분산 행렬은 LMI 조건에 의해 non-singular 값을 갖는다.
3) 제안된 알고리즘은 좋은 지역 최소 값에 수렴하는 성질을 갖는다.
4) 제안된 알고리즘은 실제 클러스터의 중점으로 수렴한다.

그러나, 반복 알고리즘의 성격상 알고리즘 수행 속도가 다소 느리다는 단점을 구조적으로 피할 수는 없었다. 따라서, 향후 연구에서는 클러스터링의 속도를 개선하기 위한 연구 및 퍼지 클러스터링으로의 알고리즘 발전에 관한 연구를 수행할 계획이며, 본 알고리즘을 이용한 응용 분야에 관한 연구도 병행할 계획이다.

5. 참고문헌

[1] J. Mao and A. K. Jain, "A Self-Organizing Network for Hyperellipsoidal Clustering (HEC)", IEEE Trans. on Neural Network, Vol. 7, No. 1, pp. 16-29, 1996.

[2] Wang Song and Xia Shaowei, "Comments on [A Self-Organizing Network for Hyper-ellipsoidal Clustering(HEC)]", IEEE Trans. on Neural Network, Vol. 8, No. 6, pp. 1561-1562, 1997.

[3] H. Ichihashi, M. Ohue, and T. Miyoshi, "Fuzzy c-Means Clustering Algorithm with Pseudo Mahalanobis Distances", Proc. of the Third Asian Fuzzy Systems Symposium, pp. 148-152, 1998.

[4] W. Song, M. Feng, S. Wei, and X. Shaowei, "The Hyper-ellipsoidal Clustering Using Genetic Algorithm", IEEE Int. Conf. on Intelligent Processing Systems, pp. 592-596, 1997.

[5] R. Krishnapuram and J. Kim, "A clustering Algorithm Based on Minimum Volume", Proc. of the Fifth IEEE Int. Conf. on Fuzzy Systems, Vol. 2, pp.1387-1392, 1996.

[6] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, "Linear Matrix Inequalities in Systems and Control Theory", Philadelpia, PA: SIAM, pp. 68-71, 1994.