다중 뷰 데이터에 대한 적응형 분광 군집화

손정우, 전준기, 김선중
한국전자통신연구원 방송통신미디어연구소
지능형융합미디어 연구부
e-mail:[jwson, jkjeon, kimsj]@etri.re.kr

Adaptive Spectral Clustering for Multiview Data

Jeong-Woo Son, Junekey Jeon, Sun-Joong Kim
Intelligent Convergence Technology Research Department
Broadcasting & Telecommunications Media Research Laboratory
Electronics and Telecommunications Research Institute

요 약

분광 군집화 기술은 Non-convex 군집에 대해 타 군집화 기술에 비해 강건하여 다양한 분야에서 활용되고 있다. 본 논문에서는 다중 뷰 데이터의 특성을 반영한 새로운 분광 군집화 기술을 제안한다. 제안한 방법은 협업학습의 접근 방법을 적용하되, 다수의 뷰가 서로 간에 가지는 독립성의 정도를 반영하여 유사도 그래프를 구축하고, 구축된 그래프를 기반으로 분광 군집화를 수행한다. 이를 통해 뷰들 간 서로 다른 정보 요구를 그래프에 반영함으로써 군집화 성능을 높인다. 세 개의 뷰를 가정한 가상의 데이터에서 제안한 방법은 기존 방법에 비해 최대 8.25%, 높은 성능을 보였다.

1. 서론

군집화(Clustering)는 대표적인 비지도 학습(Unsupervised learning) 기술로 문서 분류, 영상 처리 등 다양한 분야에서 활용되고 있다 [1]. 이 중, 분광 군집화는 데이터로부터 구축되는 군집의 형태가 non-convex하더라도 강건한 성능을 보일 수 있어 다양한 분야에서 활용되고 있다[2, 3]. 분광 군집화는 유사도를 기반으로 구축된 그래프의 라플라시안(Laplacian)을 구하고, 해당 라플라시안의 고유 벡터가 서브 그래프를 특정하는 특징을 이용하여 데이터를 분광 맵핑(spectral clustering)을 한다. 실제 군집화는 잘 알려진 군집화 알고리즘을 이용하여 맵핑된 공간에서 이루어진다. 분광 맵핑을 통해 특징 공간에서 non-convex한 군집을 이루는 데이터를 convex한 군집을 이루도록 함으로써 다양한 데이터 분포에 강건하게 동작한다.

본 논문에서는 다중 뷰를 가지는 데이터에 대한 새로운 분광 군집화 기술을 제안한다. 현실에서 데이터는 다수의 뷰를 통해 표현될 수 있다. 예컨대, 동영상은 영상 특징과 음성 특징 등으로 표현 가능하며, 웹 페이지는 텍스트와 링크 등으로 표현할 수 있다. 이러한 다중 뷰를 가지는 데이터를 다루는 대표적인 학습 기법으로 협업학습(Co-training)[4]을 들 수 있다. 협업 학습에서는 각각의 뷰에서 학습한 학습 모델을 통해 데이터에 대한 분류, 추정 등을 수행하고 해당 결과를 다른 뷰에 적용함으로써 뷰 간의 정보 교류가 일어나도록 유도한다. 기존의 다양한 연구에서 협업학습은 단일 뷰 혹은 뷰 간 단순 결합보다 높은 성능을 보여 왔다. 협업학습은 초기 반지도 학습(semi-supervised learning) 문제에만 적용되어 왔으나, 최근에는 비지도 학습까지 적용 범위를 넓혔다 [5]. 특히, Kumar와 Daumé는 분광 군집화에 협업 학습을 적용한 다중뷰 분광 군집화 기술을 제안한 바 있다 [6].

기존에 제안된 다중뷰 분광 군집화 기술의 경우 두 개의 뷰를 가정하고 개발되었기 때문에 셋 이상의 뷰를 가지는 데이터에서 뷰 간 서로 다른 의존성을 반영하지 못하는 단점이 있다. 예컨데 세 개의 뷰(x1, x2, x3)로 이루어진 데이터 x를 생각해보자. 실제 데이터에서 각 뷰들 간에 완전한 독립성을 보장하지 못한다. 즉, x1과 x2, x1과 x3는 온전히 독립이 될 수 없으며, 이들 간의 독립 정도도 차이가 날 수 밖에 없다. 이 경우, 뷰 간의 독립을 가정하는 협업 학습의 성능은 떨어질 수 있다.

본 논문에서는 이와 같이 다중의 뷰에서 나타날 수 있는 서로 다른 독립성을 군집화에 반영함으로써 새로운 다중뷰 분광 군집화 기술을 제안한다. 제안한 기술에서는 대상 뷰(x1)에 대해 최대한 독립을 이루는 새로운 뷰를 나머지 뷰들(x2, x3)을 조합하여 생성함으로써 뷰 간 서로 다른 독립성을 반영한다. 생성된 뷰의 고유 벡터를 이용하여 대상 뷰의 데이터를 분광 맵핑함으로써 대상이 되는 뷰는 자신이 가지지 못한 정보를 효율적으로 전달 받는다.

실험에서는 세 개의 뷰를 가지는 가상의 데이터를 이용하여 기존 기술과 제안한 기술 간의 군집화 성능을 측정하였다. 실험 결과 제안한 기술은 기존 기술에 비해 최대 8.25% 높은 성능을 보여 뷰 간 독립성에 대한 적응이 군집화 성능을 높이는데 효율적이었음을 보였다.

2. 다중 뷰 분광 군집화

본 장에서는 먼저 제안한 방법의 기반이 되는 다중 뷰

분광 군집화 기술을 소개한다. 다중 뷰 분광 군집화는 분광 군집화에 협업 학습의 방식을 도입한 것으로 각 뷰로부터 구축된 유사도 그래프 각각에 대해 라플라시안을 구하고, 고유벡터를 도출한 후, 서로 간의 고유벡트를 이용하여 유사도 그래프를 분광 맵핑한다. 이 과정을 수회 반복함으로써 각 그래프가 내포하는 유사도 값은 서로 간의 특성을 반영하게 된다. 표 1은 뷰가 두 개인 경우의 다중 뷰 분광 군집화의 알고리즘을 보여준다. 알고리즘에서 sym()은 입력된 행렬을 대칭행렬로 변환하는 함수이며 concat()은 입력된 벡터 간의 결합을 위한 함수이다.

표 1. 다중 뷰 분광 군집화 알고리즘

입력: 유사도 행렬 K_1, K_2
출력: k 군집에 대한 군집화 결과
초기화: 그래프 라플라시안 L_1, L_2,
고유벡터 U_1^0, U_2^0

for i = 0 to iter do
1: $S_1 = sym(U_2^{i-1} U_2^{i-1^T} K_1)$
2: $S_2 = sym(U_1^{i-1} U_1^{i-1^T} K_2)$
3: S_1, S_2를 이용하여 U_1^i, U_2^i를 계산
end for
4: U_1^i, U_2^i 정규화
5: $V = concat(U_1^i, U_2^i)$
6: V를 기반으로 k-means 군집화 수행
7: 군집화 결과 반환

세 개 이상의 뷰를 가지는 데이터의 경우 다중 뷰 분광 군집화에서는 각 뷰들이 동일하게 독립되어 있다 가정하고 분광 맵핑을 한다. 즉 알고리즘에서 1과 2에 해당하는 맵핑 과정이

$$S_v = sym((\sum_{i \neq v} U_i U_i^T) K_v) \quad (1)$$

로 바뀌어 적용된다.

3. 적용형 분광 군집화

기존의 다중 뷰 분광 군집화는 둘 이상의 뷰가 주어졌을 때, 이들의 중요도를 균일하게 보고 서로 간의 정보를 전달하였다. 하지만, 뷰가 셋 이상일 때, 실제 데이터에서는 이러한 가정이 맞지 않는 경우가 대부분이다. 실세계 데이터에서 각 뷰는 온전히 독립을 이루기는 힘들며, 이는 협업 학습에 대한 다수의 연구가 완전한 독립이 아닌 상황에서의 학습 능력 검증을 다루고 있는 이유이다. 본 논문에서는 셋 이상의 뷰가 주어졌을 때, 대상이 되는 뷰에 상대되는 뷰를 나머지 뷰의 선형 조합을 통해 생성하고자 한다. 생성되는 뷰는 대상 뷰에 대해 최대한 독립을 유지하도록 강제하여 기존 방법의 문제를 해결하고자 한다.

표 2. 적응형 분광 군집화 알고리즘

입력: 유사도 행렬 K_1, K_2, K_3
출력: k 군집에 대한 군집화 결과

1: 가중치 W_v 추정
2: K'_1, K'_2, K'_3 구축
3: K'_1, K'_2, K'_3를 이용하여 U_1^0, U_2^0, U_3^0 계산
for i = 0 to iter do
4: $S_j = sym(U_j^{i-1} U_j^{i-1^T} K_j)$
5: S_j를 이용하여 W_v 추정
6: W_v를 기반으로 S'_j 구축
7: U_1^i, U_2^i, U_3^i 계산
end for
8: U_1^i, U_2^i 정규화
9: $V = concat(U_1^i, U_2^i)$
10: V를 기반으로 k-means 군집화 수행
11: 군집화 결과 반환

세 개의 뷰에 대해 유사도 그래프 K_1, K_2, K_3가 주어졌다고 가정하자. 이에 대해 대응되는 세 개의 그래프 K'_1, K'_2, K'_3를 구축한다. 이 때,

$$K'_i = \sum_{v \neq i} w_v K_v,$$

이며, 각 뷰에 대한 가중치 $w_v \in R^n$는 생성되는 K'_i가 K_i와 최대한 독립이 되도록 추정한다. K_i와 K'_i 간 독립성은 상관 계수의 제곱을 계산함으로써 알 수 있다. 상관 계수는 -1에서 1 사이의 값을 가지는데, 완전히 독립일 경우 0을 가진다. 따라서 상관 계수의 제곱을 최소화하는 것을 통해 K_i와 K'_i 간 독립성을 보장할 수 있다.

간략화를 위해 K_i의 인스턴스 x_i에 대해 기술해보면, $x_i \in R^n$에 $X_{\neg i} \in R^{(|v|-1, n)}$가 다른 뷰로부터 구축될 수 있다. 이 때 $w_i \in R^{|v|-1}$는

$$w_i^* = \arg\min_{w_i \in W} (bA \cdot w_i^T)^2 + \frac{C}{2} |w_i|^2,$$

로 정의된다. 위의 수식에서 C는 L_2 regularization을 위한 사용자 파라미터이며,

$$b = x_i - E[x_i],$$
$$A = X_{\neg i} - E[X_{\neg i}],$$

로 정의된다. w_i를 기반으로 구축되는 K'_i는 유사도 행렬이어야 한다. 즉, 행렬의 모든 값이 0에서 1사이에 위치해야 한다. 이를 위해 아래 두 제약사항을 추가하여 최적화 문제를 정의하였다.

$$\forall j, 0 \leq w_{i,j} \leq 1.0 \ ,$$

$$\sum_{j}^{n} w_{i,j} = 1.0.$$

표2는 적응형 분광 군집화의 알고리즘을 보여준다. 표에서 알 수 있듯이, 매 반복마다 가중치 추정를 추정하고, 데이터를 맵핑하는 것을 알 수 있으며, 추정되는 가중치의 수는 $|v| \times (|v|-1) \times n$으로 $|v|$는 뷰의 수를 n은 인스턴스의 수를 의미한다.

4. 실험

4.1 데이터 셋

실험을 위한 데이터는 Kumar와 Daumé가 정의한 세계의 뷰로 이루어진 가공의 데이터를 샘플링하여 준비하였다. 데이터 셋은 두 개의 군집으로 이루어져 있는데, 각 군집의 데이터의 각 뷰 1, 2, 3은 아래 평균과 분산으로 이루어진 가우시안으로부터 샘플링되었다:

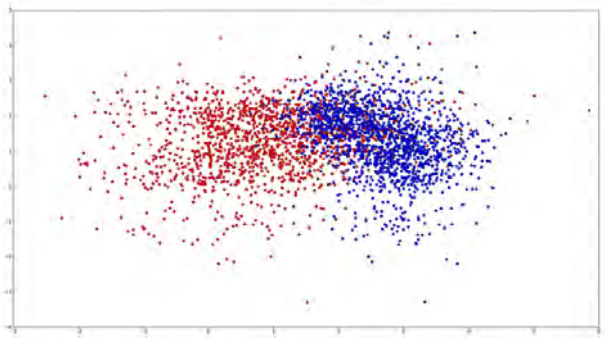
$$\mu_1^1 = (1,1), \mu_1^2 = (1,2), \mu_1^3 = (1,1),$$

$$\mu_2^1 = (3,4), \mu_2^2 = (2,2), \mu_2^3 = (3,3),$$

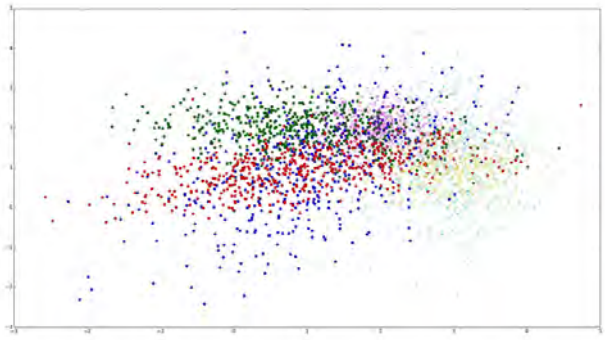
$$\Sigma_1^1 = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}, \Sigma_1^2 = \begin{pmatrix} 1.0 & -0.2 \\ -0.2 & 1.0 \end{pmatrix}, \Sigma_1^3 = \begin{pmatrix} 1.2 & 0.2 \\ 0.2 & 1.0 \end{pmatrix},$$

$$\Sigma_2^1 = \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.6 \end{pmatrix}, \Sigma_2^2 = \begin{pmatrix} 0.6 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}, \Sigma_2^3 = \begin{pmatrix} 1.0 & 0.4 \\ 0.4 & 0.7 \end{pmatrix}.$$

실험에서는 각 군집 당 500개씩 총 1000개의 데이터를 샘플링하여 정확도를 계산함으로써 성능을 측정하였다. 그림 1은 군집과 각 뷰에 따른 데이터 분포를 보여준다.



(a) 두 데이터 군집의 분포



(b) 뷰에 따른 데이터 분포

그림 2. 군집과 데이터 분포

4.1 성능

그림 2는 실험 결과를 보여 준다. 실험에서는 단일 뷰만을 이용한 군집화(single view), 커널 결합 기반의 군집화(kernel addition, kernel product), 다중 뷰 분광 군집(multiview Spectral clustering), 그리고 제안한 방법을 비교하였다. 그림에서 알 수 있듯이, 단일 뷰를 이용할 경우 성능은 90.9%로 가장 낮았다. 커널 결합 기반의 방법들과 다중 뷰 분광 군집화의 경우 유사한 성능을 보였다. 이는 Kumar와 Daumé의 결과와 동일하다. 반면 제안한 방법은 실험에 참가한 방법 중, 가장 높은 성능인 98.4%의 정확률을 보였다. 이는 다중 뷰들 간의 서로 다른 독립성을 고려한 결과라 할 수 있다.

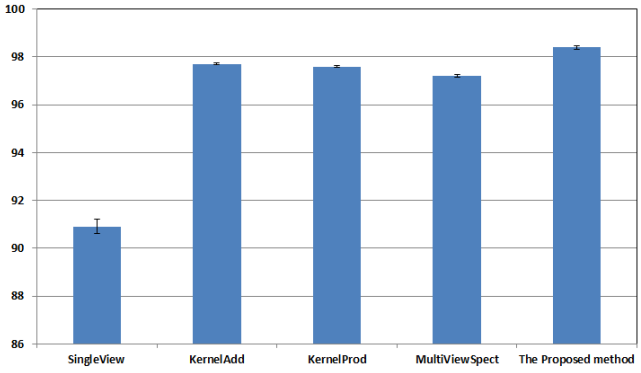


그림 1. 군집화 성능

2. 결론

본 논문에서는 둘 이상의 뷰를 가지는 다중 뷰 데이터에 대한 새로운 분광 군집화 기술을 제안하였다. 제안한 방법은 각 뷰에 대응하는 새로운 뷰를 나머지 뷰들의 선형 조합을 통해 생성함으로써 뷰들 간의 독립성을 보장하였다. 실험에서는 총 다섯가지 방법들을 비교하였으며, 그 결과 제안한 방법이 가장 높은 정확률을 보였다. 이를 통해 다수의 뷰가 서로 다른 독립성을 보일 때, 제안한 방법이 효율적임을 증명했다. 추후에는 실세계 데이터에 적용하여 성능을 검증하고자 한다.

사 사 (Acknowledge)

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.B0125-15-1002, 개방형 미디어 생태계 구축을 위한 시맨틱 클러스터 기반 시청상황 적응형 스마트방송 기술 개발)

참고문헌

[1] C. Bishop, "*Pattern Recognition and Machine Learning*," Springer, 2007.

[2] Y. Ng, M. Jordan, and Y. Weiss. "On spectral clustering: Analysis and an algorithm." *Advances in*

Neural Information Processing Systems, vol. 2 pp. 849-856, 2002.

[3] U. Luxburg, "A tutorial on spectral clustering." *Statistics and computing*, vol. 17, no. 4, pp. 395-416, 2007.

[4] A. Blum and T. Mitchell. "Combining labeled and unlabeled data with co-training." in *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92-100, 1998.

[5] A. Levin, P. Viola, and Y. Freund, "Unsupervised improvement of visual detectors using cotraining," In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pp. 626-633, 2003.

[6] A. Kumar and H. Daumé III, "A Co-training Approach for Multi-view Spectral Clustering," In *Proceedings of the 28th International Conference on Machine Learning*, pp. 393-400, 2011.