얼굴 데이터의 실시간 클러스터링을 위한 주요 비지도 학습 알고리즘 비교 연구

*최희조 **장일식 ***박구만
*서울과학기술대학교 일반대학원 미디어IT공학과
**서울과학기술대학교 나노IT디자인융합대학원 정보통신미디어공학전공
***서울과학기술대학교 전자미디어IT공학과
*heejo0624@gmail.com

Comparisons of Ten Unsupervised Learning Models in Real time Clustering of Face Images

*Choi, Hee-jo **Chang, il-sik ***Park, Goo-man
*Dept. of Media IT Engineering
**Dept. of Information Technology and Media Engineering
***Dept. of Electronics and IT Media Engineering
Seoul National University of Science and Technology

요약

본 연구에서는 고차원 데이터에 대한 차원축소 및 군집 분석과 같은 비지도 학습 알고리즘에 대해 알아보기 위해서 얼굴 이미지 데이터 셋을 사용한다. 얼굴 데이터 셋에 대하여 주요 비지도 학습 알고리즘을 이용하여 실시간으로 클러스터링하고, 그 성능을 비교한다. 비디오에서 추출된 영상 속의 7명의 인물에 대하여 Scikit-learning 라이브러리에서 제공하는 클러스터링 알고리즘과 더불어 주요 차원축소 알고리즘(Dimension Reduction Algorithm)을 사용하여 총 10개의 알고리즘에 대하여 분석한다. 또한, 클러스터링 성능 검사를 통해 알고리즘의 성능을 비교해보고, 이를 통하여 앞으로의 연구 방향에 대해 고찰한다.

1. 서론

정보통신기술의 고속 성장에 따라 고도화된 IT기술과 센서 환경에 의해 데이터를 분류하거나 값을 예측하는 등의 데이터 처리와 분석에 대한 중요도가 증가하고 있다. 데이터의 형태와 쓰임새가 복잡해짐에 따라 비지도 학습을 통한 데이터의 처리에 대한 또한 요구가 늘어나고 있다.

비지도 학습과 지도 학습의 차이는 라벨(label) 존재의 여부이다. 비지도 학습의 종류는 크게 3가지 군집 분석(Clustering), 차원축소, 유사도 비교로 나눠 볼 수 있다. 클러스터링이란, 개체들이 명확하게 분류되지 않은 상황에서 주어졌을 때, 각 데이터의 특성을 고려하여 유사성을 측정하고 유사도가 높은 집단을 분류하고, 개체들을 몇 개의 부분 그룹인 클러스터로 나누는 과정을 의미한다. 클러스터링은 계층 군집과 비계층 군집으로 나눌 수 있다. 그림1의 내용과 같이 계층 클러스터링 분석과 비계층 클러스터링 분석의 차이는 순차적으로 그룹을 할당하는지의 여부에 따라 나뉜다. 계층적인 방법은 가까운 대상끼리 순차적으로 군집을 묶어간다면, 비계층적인 방법은 랜덤하게 군집을 나눈다.

이때, 고차원의 데이터는 차원의 희소성(Sparsity)을 증가시켜 효율적인 표현에 어려운 부분이 있다. 주요 클러스터링 및 차원 축소 알고리즘을 비교하여 얼굴 데이터의 클러스터링 성능을 비교하고 분석하고자 한다. 이 연구를 통하여 주요 비지도학습 알고리즘 비교를 통하여 얼굴 데이터의 실시간 클러스터링의 성능을 알아보고 앞으로의 얼굴 이미지 분석 연구 및 비지도 학습에 대한 연구 방향을 고찰한다.

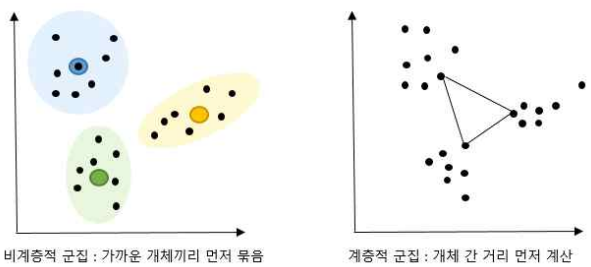

그림 1 비계층적 군집 분석과 계층적 군집분석의 차이점 비교

2. 계층적 군집 분석에 의한 10개 모델 비교

1) 실험 방법

고차원 데이터의 차원축소를 통한 클러스터링의 성능을 알아보기 위해서 얼굴 이미지 데이터셋을 사용한다. 이때 얼굴 클러스터링의 흐름도는 아래의 그림4와 같다. 동영상에서 프레임을 추출하여 전처리 과정을 거친 후, 프레임 내의 얼굴 위치를 찾아서 인코딩한다. 이때 180개의 얼굴을 인식하며, 인코딩된 얼굴 정보에 대하여 클러스터링 알고리즘을 통하여 얼굴에 대한 클러스터를 실시간으로 분석하며, 클래스 별로 폴더에 저장된다. 그 결과는 그림2와 같다. dataset은 동영상 내의 7명의 등장인물에 대해서 총 10가지의 알고리즘을 사용하여 클러스터링 한다.

클러스터링 성능 검사 방법에는 Adjusted Rand Index, Mutual Information based scores, Homogeneity completeness and V-measure, Fowlkes-Mallow scores, Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index,Contingency Matrix 등의 약 8종류가 있다. Adjusted Rand Index를 포함한 앞의 4가지 방법은 성능 검사 시, 수동으로 어노테이션을 할당해야한다는 한계가 있다. 반면, Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index 방법은 Ground Truth를 수동으로 할당하지 않아도 성능을 측정할 수 있도록 제공하고 있다. 라벨을 따로 부여하지 않아도 되는 편의성을 제공하여 실험의 클러스터링 성능 검사 방법으로 채택하였다. 여러 알고리즘 중 성능 검사가 가능한 알고리즘을 추려서 정리하였다. 얼굴 데이터의 실시간 클러스터링을 위한 주요 비지도학습 알고리즘 비교 연구에 대한 Silouttte Coeffient score, Davies-Bouldin Index 성능 검사의 내용은 표1과 같다.

2) 실험 결과

얼굴 데이터에 적합한 알고리즘을 알아보기 위해 10개의 알고리즘으로 실험해본 결과는 그림 2와 같다. 계층 클러스터링에 해당하는 알고리즘들은 자료의 크기가 크면 군집 분석이 어려운 경우가 있기 때문에 비계층 클러스터링에 비해서 성능 평가가 대체로 저조한 것을 관찰하였다. 비계층적 군집분석에는 가장 대표적인 예로 알려진 Vanilla K-means 클러스터링이 가장 우수한 성능을 보이는 것으로 관찰되었다. Vanilla K-means 클러스터링은 미리 클러스터의 수와 중심을 정하고, 각각의 요소가 어느 클러스터에 속하는지를 계산하여 분석하며, 이 작업을 계속 반복하여 최적의 클러스터 중심을 찾는 알고리즘이다. Mini-batch k-means는 Vanilla k-means 클러스터링 분석 방법에 Mini-batch를 이용하여 수렴 시간을 줄이고, 연산량이 많은 부분을 보완하기 위해 고안된 방법이다. PCA + k-means clustering 분석 방법은 고차원의 데이터를 PCA로 차원을 축소하고, k-means로 클러스터링하는 방법을 사용한다. Silhoutte Coefficient와 calinski harabasz 점수를 보았을 때, K-means 군집분석에 대한 성능 개선여부는

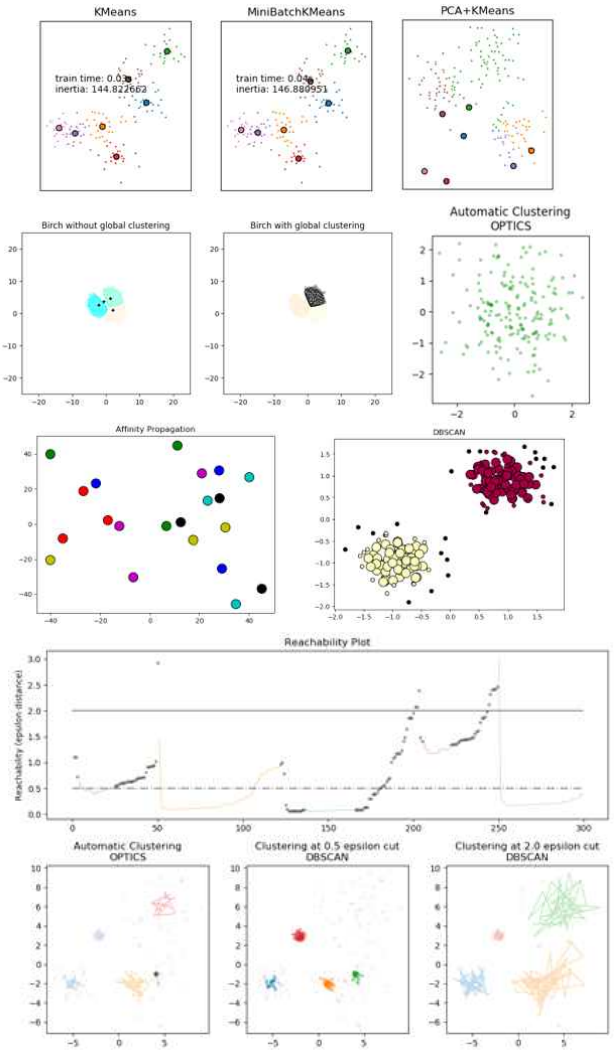

그림 2 클러스터링 알고리즘에 대한 얼굴 데이터의 실시간 클러스터링 결과

	비계층적 클러스터링				계층적 클러스터링		
클러스터링 알고리즘	K-Means	Mini batch K-means	PCA+ K_Means	Spectral Clustering	Affinity Propag-ation	DBSCAN	OPTICS
Silhoutte Coefficient score	0.11310104560	0.10402096982	0.09937628945	0.10679315690	0.08557775095	0.26921544488	-0.0657091303
calinski harabasz score	12.4419033426	12.6988070818	11.3053534914	12.2123313302	7.11148434244	12.6907548426	4.12643088605

표 1 얼굴 이미지 데이터에 대한 비계층적, 계층적 클러스터링 성능 평가

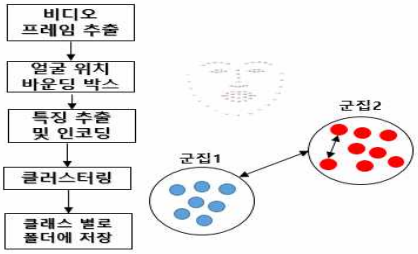

그림 3 얼굴 이미지 데이터 클러스터링의 과정

미미하다. Spherical k-means clustering k-means 클러스터링을 기반으로 기존의 유클리드 거리가 아닌 cosine 거리를 사용하는 알고리즘이다. 총 10개의 클러스터를 분리했지만, 클러스터 내의 정확성이 매우 높은 점을 확인할 수 있었다. 더불어, 그래프 기반 군집화 방법인 Spectral Clustering 또한 8개의 클러스터를 만들었고, Silhouette Coefficient score와 calinski harabasz score에서 우수한 성능을 보였다. 다음으로, 계층적 군집분석 방법인 Affinity propagation의 경우는 24개의 클러스터로, Birch(Balanced Iterative Reducing and Clustering Hierarchies)의 경우도 오직 1개의 클러스터만 찾아서 두 알고리즘 모두 비교적 낮은 성능을 보였다. 반면, OPTICS(Ordering Points to Identify Clustering Structure)의 경우 5개의 클러스터를 찾았는데, 그 클러스터 내에 비교적으로 정확한 성능을 냈고, Mean Shift는 클러스터의 개수는 맞게 나눴지만 클러스터 내의 정확도는 낮은 것으로 확인 되었다. 마지막으로, DBSCAN(Density-Based Spatial Clustering of Application with Noise)는 성능평가 점수가 우수해 이미지 데이터에 대해 좋은 성능 낼 것으로 예상했으나, 크게 좋은 성능을 보이지는 못했다. DBSCAN가 큰 데이터에 적용하기 힘들다는 점을 보완하기위해 t-SNE(t-Stochastic Neighbor Embedding)를 통해 차원을 먼저 축소한 후 DBSCAN 알고리즘을 사용하여 T-SNE + DBSCAN 결과를 보았다. 고차원 데이터를 축소하는 알고리즘을 사용하여 축소된 데이터의 군집한 결과 180개의 개체가 하나의 클러스터로만 분석되어 예상만큼 크게 성능이 나아지지 않는 것을 확인하였다.

3. 결론

표 1은 얼굴 이미지 클러스터링을 위한 차원축소 알고리즘의 성능을 나타낸 것이다, 그래프와 같이 비계층적 클러스터링에서는 K-means와 Spectral Clustering 모두 전반적으로 높은 성능 평가는 내는 것을 확인할 수 있었고, 계층적 클러스터링은 데이터의 차원을 축소하여 적용하였음에도 비교적 부진한 성능을 냈다.

이 실험 결과를 기반으로 더욱 높은 성능을 내는 군집분석 알고리즘을 연구할 것이며, 비지도 학습에 또다른 형태에 해당하는 GAN(Generative Adversarial Network) 모델에도 얼굴 및 객체 생성 및 검출 그리고 인식의 성능을 높이기 위한 연구를 진행할 예정이다.

Acknowledgement

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2019-0-01106, Sub-ppb급 가스성분 감지를 위한 후각지능 기술 개발)

참고문헌

[1] Florian Schroff, Dmitry Kalenichenko, James Philbin, FaceNet : A Unified Embedding for Face Recognition and Clustering, IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2015, June, 2015.

[2] 허경용, 김광백, 우영운 (2008), "스펙트럼 군집화에서 블록 대각 형태의 유사도 행렬 구성", 한국멀티미디어학회논문지, 11(9), 1302-1309.

[3] Hartigan, J. A.; Wong, M. A. (1979). “Algorithm AS 136: A K-Means Clustering Algorithm”. 《Journal of the Royal Statistical Society, Series C》 28 (1): 100–108. JSTOR 2346830

[4] Hamerly, G. and Elkan, C. (2002). 〈Alternatives to the k-means algorithm that find better clusterings〉 (PDF). 《Proceedings of the eleventh international conference on Information and knowledge management (CIKM)》.

[5]Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[6]“k-means++: The advantages of careful seeding” Arthur, David, and Sergei Vassilvitskii, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics (2007)

[7]“Mean shift: A robust approach toward feature space analysis.” D. Comaniciu and P. Meer, IEEE Transactions on Pattern Analysis and Machine Intelligence (2002)

[8]“A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise” Ester, M., H. P. Kriegel, J. Sander, and X. Xu, In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231. 1996

[9]“DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). In ACM Transactions on Database Systems (TODS), 42(3), 19

[10]“OPTICS: ordering points to identify the clustering structure.” Ankerst, Mihael, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. In ACM Sigmod Record, vol. 28, no. 2, pp. 49-60. ACM, 1999.

[11]Tian Zhang, Raghu Ramakrishnan, Maron Livny BIRCH: An efficient data clustering method for large databases.