클러스터링 알고리즘을 이용한 항공기 이상 현상 분석

손일락*, 이상호*, 이근택*, 박노삼**
주식회사 애나, 연세대학교, 한국전자통신연구원
*ilrak.son@byanna.io, **sangholee@yonsei.ac.kr, *geuntaek.lee@byanna.io, ***siru23@etri.re.kr

Aircraft Anomaly Analysis Using Clustering Algorithm

*Il Rak Son, **Sang Ho Lee, *Geun Taek Lee, ***Noh Sam Park
*ANNA Inc., **Yonsei University, ***ETRI.

요 약

본 논문은 항공기 안전사고를 예방하기 위해 클러스터링 기법을 항공 데이터에 적용하여 항공기 이상 현상을 분석하는 연구로 착륙단계의 항공 데이터를 활용하여 이상 현상을 분류하였다. K-means, GMM, DBSCAN, OPTICS 클러스터링을 사용하여 이상 현상 분류 결과를 확인하고 성능을 비교하였다. 항공 데이터는 각 이상 현상이 정규 분포와 유사한 분포를 하고 있어 이상 현상을 분류하는데 GMM 클러스터링이 가장 높은 성능을 보였다. 본 논문에서 분류한 데이터를 기반으로 항공기 이상 현상을 분석하여 향후 실시간 항공 데이터를 기반으로 예측을 통해 안전사고를 예방하는데 활용될 것을 기대한다.

Ⅰ. 서 론

최근 머신러닝 기술이 발전됨에 따라 여러 산업 분야에 머신러닝 기법을 적용하여 산업 발전을 도모하고자 하는 수요가 증가하고 있다. 머신러닝을 활용하면 산업 방면에서 발생할 수 있는 이상 상황을 예측하여 리스크를 줄일 수 있다. 이러한 머신러닝 기술은 크게 지도학습, 비지도 학습, 시계열 분석, 클러스터링 등으로 다양하게 분류할 수 있다. 그중 클러스터링은 각 데이터의 유사성을 기반으로 데이터의 그룹을 분류하는 방법으로 데이터의 특징을 기반으로 데이터를 분석하는 데 많이 사용한다.

항공 분야는 다른 산업에 비해 사고로 인한 리스크가 매우 높은 분야이다. 항공 산업에서는 이러한 리스크를 줄이기 위해 항공안전관리체계를 구축하고 예측적 관리단계를 통해 데이터 기반으로 항공안전 관리 기술을 증진하고자 노력하고 있다.[1] 따라서 본 논문에서는 예측적 관리를 위해 항공 데이터를 기반으로 항공 이상 징후를 분류하는 클러스터링 기법을 적용하여 항공 분야의 산업 발전을 도모하고자 한다.

Ⅱ. 본론

본 논문에서는 머신러닝 알고리즘 중 하나인 클러스터링을 활용하여 항공기 이상 현상을 분류하고자 한다. 클러스터링 기법은 크게 중심 기반 알고리즘과 밀도 기반 알고리즘으로 구분되는데 중심 기반 알고리즘은 군집 중심점을 기준으로 데이터를 구분하는 방법으로 대표적으로 중심점으로부터 거리가 가까운 지점을 찾는 K-means[2]와 중심점으로부터 정규 분포를 찾는 GMM(Gaussian Mixtrue Model)[3]이 있다. 밀도 기반 알고리즘은 이웃한 데이터들이 같은 군집으로 분류되고 멀어진 데이터를 이상치나 다른 군집으로 분류하는 방법으로 대표적으로 DBSCAN[4]과 OPTICS[5]가 있다. 각각의 클러스터링 기법을 적용하여 항공기 이상 현상 분류 성능을 비교해 보고자 한다.

항공 데이터는 시간, 위치(위도, 경도, 고도), 속도, 각도 등 다양한 데이터들의 집합으로 이루어져 있다. 그중 이상 현상을 분류하기 위해 가장 중요한 데이터는 항공기 이상 현상을 분류하는 기준인 위치 데이터이다. 항공기 이상 현상은 절반 이상이 이륙 및 착륙단계에서 발생하며 그중 대부분이 착륙단계에서 발생한다. 착륙단계에서 발생하는 이상 현상은 대표적으로 착륙 실패로 인한 급격하게 고도를 상승하는 복행(Go-around)과 기상악화, 활주로 관제 등 외부 요인으로 인한 착륙 지연 등이 있다. 본 논문에서는 복행과 예측 불가 상황(UOC_D)으로 이상 현상을 구분하였다. 이러한 이상 현상은 위치 데이터를 기반으로 클러스터링을 통해 분류할 수 있다. 위치 데이터로 클러스터링을 진행하기 위해 데이터를 가공하여 클러스터링 데이터셋을 생성하였다. 가공 데이터는 착륙단계에서 최고 연속 고도 상승량, 고도 상승 구간에서의 평균 속도, 고도 상승 최고 속도로 이루어져 있다. 그림 1은 가공 데이터를 이상 상황 기준에 따라 분류한 3D chart 및 분류 결과에 따른 항적을 가시화한 그래프이다.

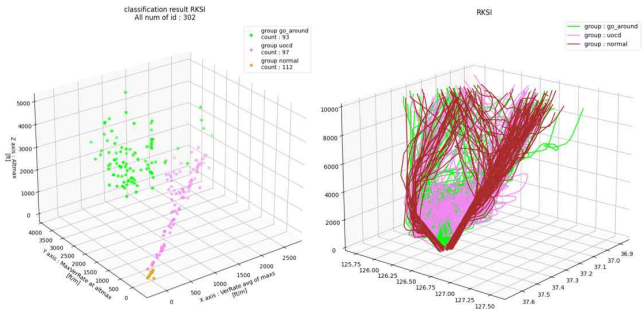

그림 1 가공 데이터셋 시각화 3D chart 및 항적 가시화 그래프

가공 데이터를 기반으로 K-mean, GMM, DBSCAN, OPTICS 클러스터링을 적용하기 위해 각 모델의 하이퍼 파라미터를 설정하였다. 중심 기반의 클러스터링 기법의 하이퍼 파라미터는 n_cluster로 k의 개수는 두 가지의 이상 상황과 정상 상황으로 전부 3개로 지정하여 결과를 확인하였고, 밀도 기반 클러스터링의 하이퍼 파라미터는 epsilon과 minPts로 각각 기본값인 1과 분류 개수인 3으로 지정하여 클러스터링을 적용하였다. 그림

2는 중심 기반 클러스터링인 K-means와 GMM 적용 결과이며, 그림 3은 밀도 기반 클러스터링인 DBSCAN 및 OPTICS 적용 결과이다.

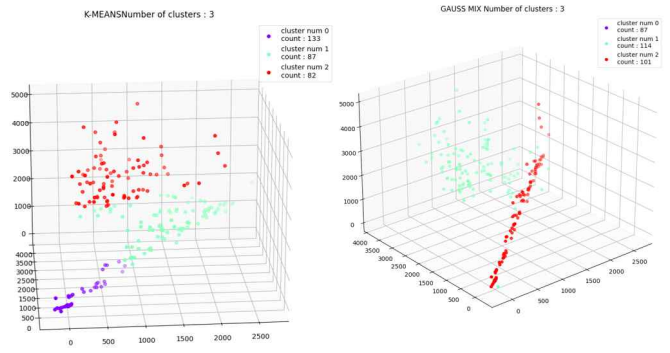

그림 2 K-means(왼) 및 GMM(오) 클러스터링 결과

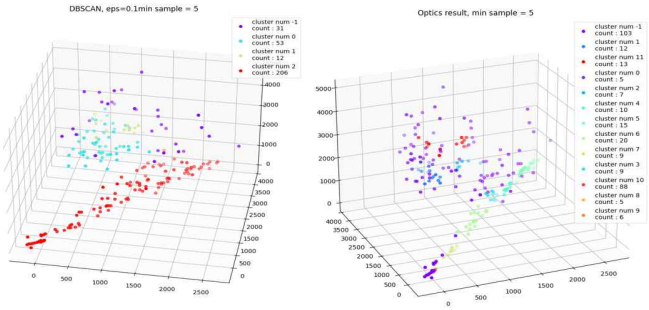

그림 3 DBSCAN(왼) 및 OPTICS(오) 클러스터링 결과

전체 데이터를 보면 예측 불가 상황 데이터가 넓게 분포하여 정상 데이터와 복행 데이터의 중심으로부터 가까운 데이터가 다수 존재하였다. K-means 클러스터링 결과를 살펴보면 정상 데이터와 인접한 이상 현상 데이터가 다수 존재하여 군집 결과가 잘 이루어지지 않는다. 이러한 결과가 발생하는 것은 데이터가 군집 중심으로 모이는 형태가 아닌 다른 형태를 띄고 있기 때문이다. GMM 클러스터링 결과를 살펴보면 정상 데이터의 일부가 예측 불가 상황으로 분류된 것을 제외하고 군집이 잘 이루어진 것을 확인할 수 있다. 이는 데이터가 각 현상에 대하여 정규 분포와 유사한 형태를 이루고 있는 것을 보여준다. DBSCAN 클러스터링 결과는 정상 데이터와 예측 불가 상황이 매우 근접하여 같은 군집을 이루고 있고 복행 데이터가 각각 다른 군집으로 나뉘며 군집을 이루지 못한 데이터도 확인된다. OPTICS도 DBSCAN과 유사하나 epsilon이 짧아 더 많은 군집이 생성되고 군집을 이루지 못한 데이터도 다수 존재하였다. 각 분류 기준에 맞는 데이터의 유사성에 따라 데이터의 거리를 근접시키고, 비 유사 데이터의 거리를 밀어낼 수 없다면 이러한 데이터에서는 밀도 기반의 클러스터링은 사용할 수 없다. 각 클러스터링 결과와 실제 이상 상황 분류의 성능을 평가하기 위해 confusion matrix를 통해 정확도, 정밀도, 재현율, f-score를 계산하고 각 성능을 비교해 보았다. 표 1은 각 기법에 따른 모델 성능 표이다.

표 1 이상 징후 분류 클러스터링 모델 성능

	ACC	PRECISION	RECALL	F_1 -score
GMM	0.889	0.747	0.771	0.759
K-menas	0.857	0.811	0.760	0.785
DBSCAN	0.778	0.472	0.453	0.463
OPTICS	0.703	0.649	0.240	0.350

각 클러스터링의 모델 성능을 살펴보면 GMM 클러스터링의 정확도 0.889, 재현율 0.771로 가장 좋은 성능을 보였고, K-means 클러스터링이 정밀도 0.811, F-score 0.785로 다음 높은 결과를 보인다. 이는 정상 데이터가 매우 밀집되어 있어 K-means 클러스터링에서 정상 데이터를 모두 같은 데이터 군집으로 지정되어 다른 군집에서 발생한 오차를 무시할 정도로 높은 성능이 나타난 것이다. 따라서 이상 현상 데이터가 많아지더라도 성능 결과가 좋아지지 못하는 한계점을 가진다. 하지만 GMM 클러스터링은 이상 현상 데이터가 증가함에 따라 전체 모델 성능이 증가 할 것으로 예상된다.

Ⅲ. 결론

본 논문에서는 클러스터링 알고리즘을 활용하여 항공기 이상 현상을 분류하고 각 클러스터링 알고리즘의 성능을 비교하였다. 항공기 이상 현상 데이터는 각 현상에 따라 정규 분포와 유사한 형태로 분포되어 GMM 클러스터링 결과가 가장 좋은 성능을 보였고, K-means가 다음으로 좋은 성능을 보였다. 하지만 K-means 및 밀도 기반 클러스터링은 모델 성능은 크게 낮지 않으나 실제 이상 현상을 분류하기에는 특성 데이터를 잘 분류하지 못해 실제적인 사용은 어려울 것으로 보인다. 이상 현상 데이터가 많아질수록 데이터의 정규 분포가 잘 이루어져 GMM 클러스터링의 성능이 높아질 것으로 보인다.

클러스터링을 통한 항공기 이상 현상 분류는 항공기 착륙 사후의 데이터를 통해 이상 현상을 분류하는 모델로 실시간 항공 데이터를 통해 이상 현상을 탐지하는 모델은 아니다. 실제적인 항공 안전사고 예방을 위해서는 미래에 발생할 이상 현상을 예측할 수 있는 모델이 필요하다. 이러한 문제를 해결하기 위해서 앞으로 본 논문을 통해 알 수 있는 항공기 이상 현상의 데이터를 기반으로 이상 현상 발생 전 데이터를 분석하여 향후 발생 가능성을 예측하는 모델의 연구가 진행되어야 할 것이다.

ACKNOWLEDGMENT

본 연구는 국토교통부 항공선진화사업의 일환으로, "빅데이터 기반 항공 안전관리 기술개발 및 플랫폼 구축(22BDAS-C158275-03)"사업을 통해 수행되었습니다.

참 고 문 헌

[1] Billings, C. E., Lauber. J. K., Funkhouser. H., Lyman. E. G., Huff. E. M. "NASA aviation safety reporting system," No. NASA-TM-X-3445. 1976.

[2] Bock. H. H. "Clustering Methods: A History of k-Means Algorithms", Selected Contributions in data analysis and classification, pp. 161-172, 2007

[3] Alpayd. E., "Soft vector quantization and the EM Algorithm", neural Network, Vol. 11, Issue. 3, pp. 467-477. 1998.

[4] Ester. M., Kriegel. H.P., Sander. J., and Xu. X., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in Proceeding of the ACM International Conference on Knowledge Discovery and Data Mining, pp. 226-231, Aug. 1996.

[5] Ankerst. M., Breunig. M. M., Kriegel. H. P., Sander. J. "OPTICS: Ordering points to identify the clustering structure." ACM Sigmod record, No.28, Vol.2, pp. 49-60, 1999