한국빅데이터학회지
제6권 제2호, 2021, pp. 29-38 https://doi.org/10.36498/kbigdt.2021.6.2.29

공항 근처 ADS-B 항적 자료에서의 클러스터링 기법 비교*

Comparison of Clustering Techniques in Flight Approach Phase using ADS-B Track Data

박종찬 · 박헌진†

인하대학교 통계학과

요 약

항공안전관리에서 항공기 경로 이탈은 큰 사고로 이어질 수 있는 위험한 요인이다. 본 연구에서는 항공기 경로 이탈 문제를 예방하기 위해 클러스터링을 통해 항적을 분류하고, 클러스터 중심과의 거리를 계산하여 이상 점수를 산출하고자 한다. 1년 동안 수신된 ADS-B 항적 자료에서 공항을 기준으로 근방 100km 이내 항적을 추출하여 연구를 진행했다. 항적은 선형 보간법을 이용하여 벡터화하였다. 위도 · 경도 · 고도 3차원 좌표 자료를 사용하였다. PCA를 통해 전체 데이터 분산 90% 이상을 나타내는 축으로 차원을 축소하였고, k-평균 군집화, 계층적 군집화, PAM 기법을 적용하였다. 클러스터 개수는 실루엣 측도를 사용하여 선택하였고, 클러스터 중심과의 거리를 계산하여 이상 점수를 산출하였다. 본 연구에서는 각 클러스터 기법별로 클러스터 개수를 비교해보고, 실루엣 측도를 통해 클러스터링 결과를 평가하고자 한다.

■ 중심어 : 이상탐지, ADS-B 항적 자료, 클러스터링

Abstract

Deviation of route in aviation safety management is a dangerous factor that can lead to serious accidents. In this study, the anomaly score is calculated by classifying the tracks through clustering and calculating the distance from the cluster center. The study was conducted by extracting tracks within 100 km of the airport from the ADS-B track data received for one year. The wake was vectorized using linear interpolation. Latitude, longitude, and altitude 3D coordinates were used. Through PCA, the dimension was reduced to an axis representing more than 90% of the overall data distribution, and k-means clustering, hierarchical clustering, and PAM techniques were applied. The number of clusters was selected using the silhouette measure, and an abnormality score was calculated by calculating the distance from the cluster center. In this study, we compare the number of clusters for each cluster technique, and evaluate the clustering result through the silhouette measure.

■ Keyword : anomaly detection, ADS-B track data, clustering

2021년 11월 26일 접수; 2021년 12월 11일 수정본 접수; 2021년 12월 13일 게재 확정.

* 본 연구는 국토교통부의 '빅데이터 기반 항공안전관리 기술 개발 및 플랫폼 구축(21BDAS-B158275-02)' 연구의 지원에 의하여 이루어진 연구로서, 관계부처에 감사드립니다. This work is supported by the Korea Agency for Infrastructure Technology Advancement(KAIA) grand funded by the Ministry of Land, Infrastrucrue and Transport (Grant 21BDAS-B158275-02).

† 교신저자 (hjpark@inha.ac.kr)

Ⅰ. 서 론

항공 산업에서 항공안전관리는 중요한 이슈이다. 특히 항공기의 경로/고도 이탈은 큰 사고로 이어질 수 있으며 인명피해와 재산 피해 등 막대한 피해를 초래할 수 있다. 항공안전법 시행규칙 [별표3] 항공안전장애의 범위에 따르면, 비행고도 이탈 등 항공교통관제기관의 사전 허가받지 아니한 항행을 한 경우에는 항공안전 장애로 본다고 명시되어 있다. 이에 활주로에 착륙을 시도하는 항공기에 대하여 전파고도계를 통해 고도 문제에 대한 경고시스템 등이 존재하지만, 그 밖의 상황에서는 관제사의 판단과 명령에 의존해야 한다.

ASIAS(Aviation Safety Information Analysis and Sharing) 사고 분석 보고서에 따르면 항공사고 대부분은 공항 근처 비행 절차에서 발생하며[10][11], 인천공항의 경우 매년 이용객 수가 증가하는 점과 북한과 인접해 있어 북측 활주로(33R, 33L, 34R)로 이・착륙하는 항공기의 항적은 구부러진 형태를 띠는 점으로 보았을 때, 인천 공항 근처 항적 이상 탐지는 매우 중요하게 여겨진다. 따라서 본 연구에서는 인천공항 근처 항공기 경로 이탈에 대한 문제를 예방하기 위하여 군집화 기반 이상 탐지(anomaly detection)기법을 비교 분석하고자 한다.

이상 탐지는 자료에서 다른 패턴을 보이는 자료를 탐지하는 것으로 정상(normal)과 이상(anomaly)을 구별하는 문제를 의미한다. 이상 탐지는 다양한 산업 분야에서 각 분야에 맞게 정의되어 적용되고 있다. 대표적으로는 제조공정에서 관리도 기법(control chart)을 활용한 품질관리가 있으며, 금융 산업에서는 신용 사기 탐지, IT 산업에서는 네트워크 침입 탐지가 있다.

항공 산업에서도 이상 탐지에 관한 연구가 이어지고 있다. Adric. E.(2009)은 Tracon Radar Tracks에서의 각 항적을 보간(interpolation) 후, k-평균 군집화(k-means clustering) 방법을 통해 항적의 흐름을 식별해 냈다[3]. Lishuai et al.(2011)은 항공기의 위치, 속도 등의 다양한 정보를 차원 축소한 후 DBSCAN(Ester et al.,1996)을 사용하여 이상 탐지를 진행했다[7][4].

본 논문의 구성은 다음과 같다. 2장에서는 다변량 시계열 및 trajectory analysis에서 사용되는 이상 탐지 기법 관련 연구를 소개하고, 3장에서는 데이터셋 설명과 데이터 전처리, 군집화 기반 이상 탐지 과정을 소개한다. 4장은 3가지 군집화 기법으로 분류된 클러스터를 확인하고 기법 간 비교 결과를 보여준다. 5장은 분석 결과를 종합하고 결론으로 마무리 짓는다.

Ⅱ. 관련 연구

본 장에서는 관련 연구를 살펴보고자 한다. 가용할 수 있는 데이터셋인 ADS-B 항적 데이터의 특성인 다변량 시계열과 trajectory analysis 분야에서의 이상 탐지 방법을 소개하고, 각 방법에서 사용된 기계학습 및 알고리즘을 설명하고자 한다.

2.1 다변량 시계열 이상 탐지

시간의 경과에 따라 두 개 이상의 변수를 갖는 시계열 데이터를 다변량 시계열이라 한다. 다변량 시계열 데이터는 기상・항공・금융 등에서 다루어지며 예측・분류・군집화・이상 탐지 등 다양한 분석을 수행할 수 있다.

다변량 시계열에서의 이상 탐지는 크게 5가지 기반의 방법이 있다. 경계나 영역을 생성하여 정상과 비정상 데이터를 구분하는 영역 기반 방법과 데이터 개체 간 거리 및 유사성 함수를 정의하여 탐지하는 거리 기반 방법, 트리 모형 기반의 앙상블 모형을 활용한 앙상블 기반 방법, 통계 모형을 적합하여 확률을 계산하고 분포를

추정하는 통계학적(statistical) 방법, 데이터를 저차원 공간으로 매핑하는 재구성(reconstruction) 방법이 있다. 영역 기반 방법에는 원점과 데이터 간 최대 마진을 구하는 OC-SVM이 대표적이며, 거리 기반 방법에는 kNN과 군집화 기법이 대표적이다. 재구성 방법에는 PCA 기반 방법이나 오토인코더 방법이 주로 사용되며, 통계학적 방법에는 분포를 가정한 VAR 모형, GMM 모형이 사용된다. 앙상블 기반 방법은 Isolation forest 등의 트리 모형 기반 기법이 사용된다[1].

2.2 궤적 분석

ADS-B 항적 데이터는 궤적(trajectory) 데이터이다. trajectory는 궤적 또는 비행경로로 움직이는 물체가 시간의 함수로 표현되는 경로로, 보행자・택시・항공 교통・비디오 분야에서 주로 사용된다[9][12].

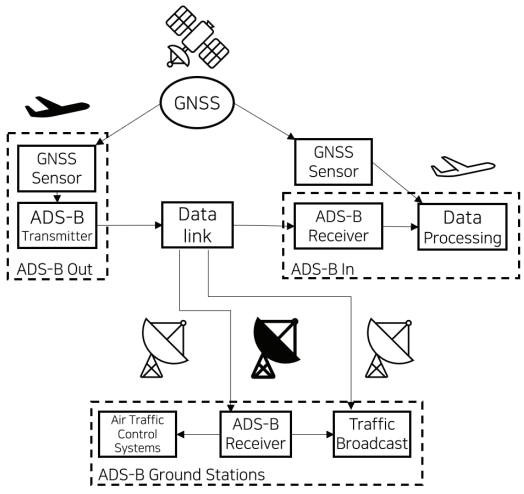
궤적 분석(trajectory analysis)에서 이상 탐지는 크게 두 단계로 이루어진다. 궤적의 주요 흐름을 분류하는 흐름 식별(Flow identification) 과정을 거쳐 이상 탐지가 진행되며, 이상 탐지에는 오토인코더, OC-SVM, 군집화, p-value를 활용한 기법 등이 사용된다.

본 연구에서는 ADS-B로부터 수신된 항적 데이터를 군집화 기법을 이용하여 흐름을 식별하고, 클러스터 중심과의 거리를 계산하여 이상 점수를 산출하고자 한다.

Ⅲ. 연구 방법

3.1 데이터셋

데이터는 2019년 1월부터 12월까지 수신된 ADS-B 데이터를 사용하였다. ADS-B(Automatic Dependent Surveillance-Broadcast)는 GPS(Global Positioning System) 위성 항법 시스템과 1,090MHz 전송 링크를 이용하여 항공기 감시 정보를 지상의 항공 교통 관제(ATC: Air Traffic Control) 및 다른 항공기에 자동으로 방송(broadcast)하는 항공기 감시 체계이다. 본 연구에서 사용된 ADS-B 자료는 2019년 1월부터 12월까지 수신된 자료이다. 한 번의 운항은 하나의 Flight ID로 표현되며, 운항 중 시간에 따라 여러 번 수신되므로 Track Data이다. 각 Track Point마다, 항공기가 수신된 시간인 Time(UTC), 수신될 때의 항공기 지리 좌표 정보(위도・경도・고도)인 Latitude(deg), Longitude(deg), Altitude(ft), 그리고 항공기 지면 속도 정보인 Groundspeed (kts) 등의 정보가 포함되어 있다. 이 중 인천공항으로 이・착륙하는 공항 중심 100km 이내 항적의 Latitude.deg, Longitude.deg, Altitude.ft를 추출하였다. 공항과의 거리는 ECEF(Earth-Centered Earth-Fixed) 좌표계를 사용하여 계산하였다. 총 287,972,941개 중 14,723,339개의 항적이 추출되었으며 각 항적은 수신된 구간이 일정하지 않으며 미수신구간이 존재하기도 한다.

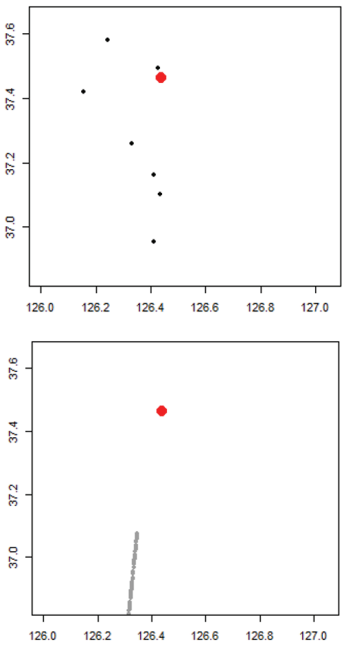


〈그림 1〉 ADS-B 시스템 아키텍처

3.2 데이터 전처리

3.2.1 불완전 항적 제거

ADS-B 항적 데이터는 수신 간격이 일정하지 않아 항적 간 데이터 포인트의 개수나 데이터 포인트 간 간격이 일정하지 않다. 특히, 데이터 개수가 상대적으로 적거나 간격이 긴 경우, 해당 운항편은 항적을 완전하게 표현하지 못하고 추후 분석 결과에 안 좋은 영향을 끼칠 수 있다. 따라서, 하나의 운항편에 대해서 연속된 데이터 포인트 간 거리가 20km 초과인 항적과 인천공항 반경 100km 이내에 항적의 포인트 개수가 20개 미만인 항적은 제거하였다.



〈그림 2〉 불완전 항적 예시
(x축 : 경도, y축 : 위도, 빨간색 점 : 인천공항)

3.2.2 데이터 벡터화

클러스터링을 진행하기 전, 항적 간 거리를 계산하기 위해 Track Data 형태의 자료에 대하여 벡터화를 진행하였다[3]. 벡터화 방법은 다음과 같다. 항적 자료의 위도, 경도, 고도 변수에 대하여 ECEF 좌표계 변환을 적용한 후, 반환된 (x,y,z) 좌표(km)를 사용하여 같은 Flight ID를 갖는 항적 자료마다 수신된 시간 순서에 따라 정렬한 후 선형 보간법을 적용하였다. Flight ID마다 항적의 길이를 계산하여, 선형 보간법 예측값(x,y,z)을 100 등분 한 지점에서 각각 추출하였다. 벡터의 첫 번째 값은 인천공항 좌표이다. 벡터화 결과 자료의 차원은 300, 이륙 항적 자료 개수는 201,072개, 착륙 항적 자료 개수는 199,564개이다.

3.3 이상 탐지

이상 탐지는 정상인지 이상인지 정보를 가진 자료 라벨(label)의 유무에 따라 크게 세 가지 방

〈표 1〉 항적 벡터화 데이터

	x 1	…	x 100	y 1	…	y 100	z 1	…	z 100
1402-1558399646-adhoc-0	-3010.234	…	-3044.464	4077.881	…	4055.089	3858.724	…	3839.409
2000-1557791603-adhoc-0:0	-3010.234	…	-3044.787	4077.881	…	4055.548	3858.724	…	3839.863
…	…	…	…	…	…	…	…	…	…
JJA2408-1566852000-schedule-0001:0	-3010.234	…	-3061.661	4077.881	…	4109.225	3858.724	…	3778.895
JJA4908-1566847200-schedule-0000:2	-3010.234	…	-3061.515	4077.881	…	4109.585	3858.724	…	3778.943

법으로 분류할 수 있다. 라벨이 있다면 지도학습을 통해 학습할 수 있다. 일반적인 산업 현장에서는 정상 자료에 비해 이상 자료가 적기 때문에 클래스 불균형(class-imbalance) 문제가 있다. 이에 정상 자료만을 사용하여 학습하는 준-지도학습(semi-supervised learning) 방법을 사용하기도 하며, Support Vector Machine(Vapnik., 1995)이 대표적인 알고리즘이다[2]. 라벨이 없으면 비지도 학습을 통해 진행된다.

본 연구는 정상인지 이상인지에 대한 라벨의 정보가 없는 자료로 비지도 학습을 기반으로 한 이상 탐지를 진행하고자 한다. 또한, 클러스터링 기반의 이상 탐지를 진행하여 정상 항적과 이상 항적을 구분하고자 한다.

3.3.1 클러스터링 기반 이상 탐지

클러스터링은 대표적인 비지도 학습 방법의 하나로 비슷한 개체들을 묶어 하나의 클러스터로 분류한다. 거리 기반 클러스터링 이상 탐지는 클러스터링 후 각 클러스터의 중심으로부터 각 개체 간의 거리를 계산하여 이상 점수(anomaly score)를 구축하고 이를 기반으로 정상과 이상을 구분한다. 이상 점수는 이상 탐지 모형을 통해 얻은 결괏값으로 연구자가 판단하여 직접 정상과 이상의 경계를 정하게 된다.

본 연구에서는 k-평균 클러스터링, 계층적 클러스터링(hierarchical clustering), PAM (Partitioning Around Medoids)(Kaufman et al., 1990)을 사용하고자 한다[6].

3.3.2 k-평균 클러스터링

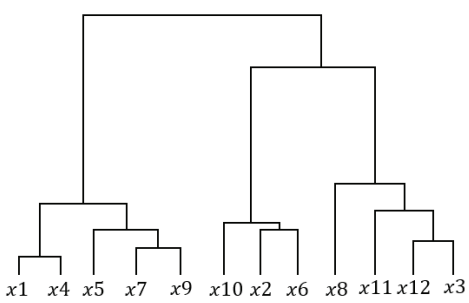
k-평균 알고리즘은 클러스터링 기법 중 가장 일반적으로 사용되는 알고리즘이다. 주어진 데이터를 k개의 클러스터로 분류하며, 알고리즘은 개체들과 해당 개체들이 속하는 클러스터의 중심으로부터 거리 차이의 분산을 최소화하는 방향으로 동작한다. 클러스터의 중심은 클러스터에 속한 개체들의 평균 또는 중심(centroid)이라 한다. k-평균의 목적 함수는 RSS(Residual Sum of Squares)이며, 이를 최소화하는 방향으로 동작한다. 잔차는 각 개체와 중심과의 거리로 계산한다.

〈표 2〉 k-평균 클러스터링 알고리즘

Input k : the number of clusters D : a data set containing n objects Output : A set of k clusters Method (1) Randomly extract k data objects from the data object set D, and set these data objects as the centroid of each cluster. (default setting) (2) For each data object in the set D, the distances from the k cluster center objects are respectively calculated, and the centroid of each data object is found with the highest similarity. And each data object is assigned to the found center point. (3) Recalculate the center point of the cluster. That is, the center point is recalculated based on the clusters reassigned in step 2. (4) Repeat steps 2 and 3 until the cluster belonging to each data object does not change.

3.3.3 계층적 클러스터링

계층적 클러스터링은 거리 기반의 계층적 트리 모형을 생성하여 차례대로 유사한 개체와 통합을 수행하는 알고리즘이다. 계층적 트리 모형은 덴드로그램이라 하며, <그림 3>과 같다.



〈그림 3〉 덴드로그램

알고리즘을 수행하기 위해선 각 개체 간 거리 행렬을 계산해야 한다. 덴드로그램을 생성하기 위해선 개체-개체, 개체-군집, 군집-군집 간의 거리를 정의해야 한다. 거리 연결 기준에는 대표적으로 두 클러스터 간 가장 가까운 거리를 사용하는 단일연결(single linkage), 두 클러스터 간 가장 먼 거리를 사용하는 완전연결(complete linkage), 클러스터 내 모든 데이터와 다른 클러스터 내 모든 데이터 사이의 평균 거리를 사용하는 평균연결(average linkage)가 있다. 항적 데이터는 데이터양이 많고 다양한 경로의 항적이 존재하기 때문에 단일연결(single linkage)를 사용하지 않았다. 또한, k-평균 클러스터링을 따로 적합하므로 평균연결(average linkage)를 고려하지 않고 완전연결(complete linkage)를 사용하였다.

3.3.4 PAM

k-중앙개체 클러스터링(k-medoids clustering)는 클러스터의 평균을 중심으로 하는 k-평균 클러스터링과 달리 클러스터 내 개체 중 하나를 대표 개체로 선택한다. 이상치가 존재하더라도 강건하여(robust) 이상치가 존재할 때 적합한 방법으로 사용된다. k-중앙개체 클러스터링은 다음과 같이 정의된다.

표본공간 $D=\{x_1,x_2,\cdots,x_n\}$가 주어졌을 때, 클러스터 $M=\{C_1,C_2,\cdots,C_k\}$는 다음 (1)을 최소화한다.

$$E=\sum_{i=1}^{k}\sum_{x\in C_j} dist(x,o_i) \quad (1)$$

x는 표본공간의 한 점, o_i는 C_j의 대표 개체이다. $dist(x,o_i)$는 두 점 사이의 거리이며 $L1$-norm으로 정의된다.

PAM은 k-중앙개체 클러스터링의 대표적인 알고리즘으로 대표 개체(medoid)를 선택하기 위해 표 3과 같이 작동한다(Han et al., 2011)[5].

〈표 3〉 PAM 알고리즘

Input k : the number of clusters D : a data set containing n objects Output : A set of k clusters Method (1) arbitrarily choose k objects in D as the initial representative objects of seeds; (2) repeat (3) assign each remaining object to the cluster with the nearest representative object; (4) randomly select a nonrepresentative object, o_{random}; (5) compute the total cost, S, of swapping representative object, o_j, with, o_{random}; (6) if $S<0$ then swap o_j with o_{random} to form the new set of k representative objects; (7) until no change;

3.3.5 클러스터 개수

각 클러스터링 기법별 하이퍼파라미터인 클러스터 개수는 실루엣(silhouette) 측도를 사용하여 선택했다. 실루엣 측도는 클러스터 내 개체가 잘 분류되었는지 평가하는 방법으로 내부적(internal) 평가 측도 중 하나이다. 각 개체의 실루엣 값은 같은 군집 내에 있는 다른 개체와의 거리를 평균한 값과 해당 개체가 속하지 않은 군집 중 가장 가까운 군집과의 평균 거리를 통해 계산된다. i번째 개체의 실루엣 값은 식 2를 통해 계산된다.

$$s(i)=\frac{(b(i)-a(i))}{\max(a(i),b(i))} \quad (2)$$

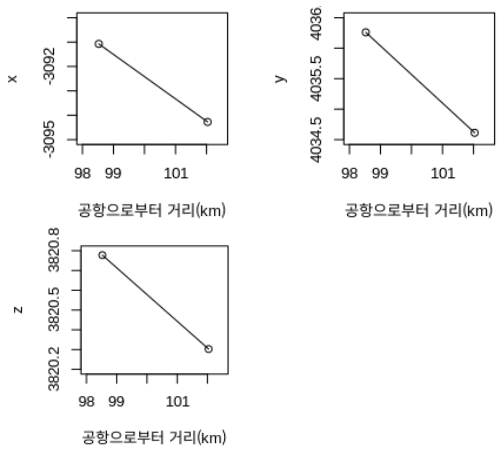
$a(i)$는 같은 군집 내에 있는 다른 개체와의 거리를 평균한 값, $b(i)$는 해당 개체가 속하지 않은 군집 중 가장 가까운 군집과의 평균 거리이다.

클러스터 개수를 늘려가면서, 모든 개체 실루엣값을 평균한 값인 평균 실루엣 너비(average silhouette width)를 산출하고, 가장 큰 값을 갖는

클러스터 개수를 선택한다. 인천공항에는 20개의 SID와 25개의 STAR이 존재한다. SID (Standard Instrument Departure, 표준계기출발절차)와 STAR(Standard Terminal Arrival Route, 표준계기도착절차)는 계기비행(IFR)을 통해 항공로에 나가고 들어오는 업무 절차의 종류이다. 절차들은 특정 지점들로 통과하는 방식으로 구성되어 있다. 계기비행 외에 항공기 조종 방법으로는 대표적으로 시계비행이 있다. 계기비행을 하는 항공기들은 계기비행절차를 따르기 때문에, 해당 항적들은 계기비행절차를 중심으로 분포되어 있다. 계기비행은 항공기 비행 방법 중 하나이므로, 공항 근처 클러스터는 계기비행절차 개수보다 많거나 같을 것이다. 이를 고려하여 최소 클러스터 개수는 계기비행절차 개수로 설정하였다. 또한, 클러스터 개수의 후보 범위는 최소 클러스터 개수를 포함하여 20개로 선정하였다. 이륙 항적은 클러스터 개수를 20부터 39까지, 착륙 항적은 25부터 44까지 증가하면서 평균 실루엣 너비를 계산하였다.

Ⅳ. 연구 결과

벡터의 시작점을 인천공항으로, 끝점을 인천공항으로부터 거리가 100km인 점으로 하여 벡터화를 진행하였다. 공항으로부터 거리가

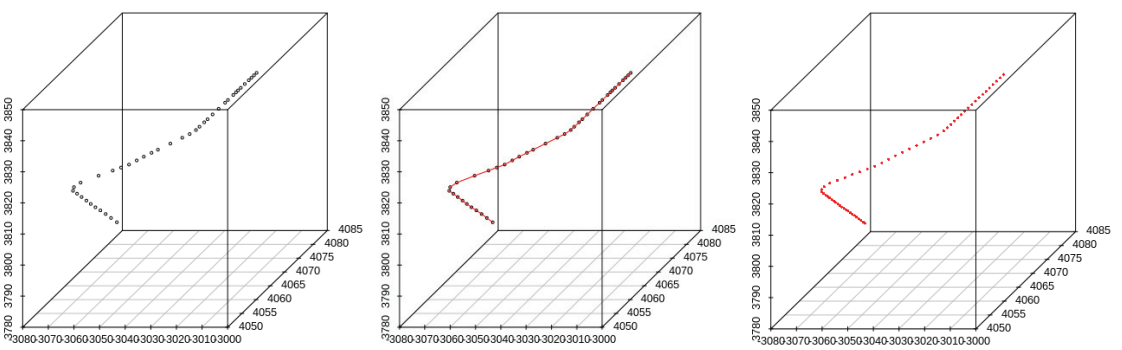


〈그림 4〉 선형보간법 적용 결과

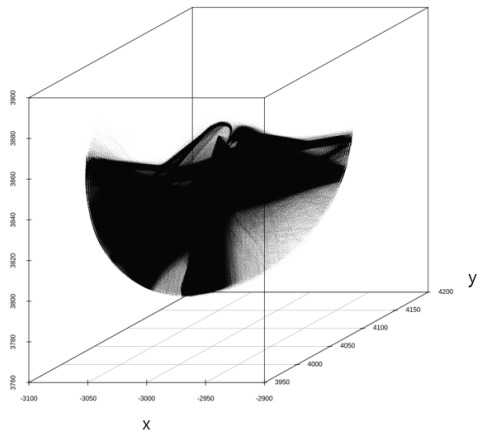
100km인 점은 선형 보간법을 사용하여 생성하였다. 100km 이내 항적 중 공항으로부터 가장 먼 항적 포인트와 100km 바깥의 항적 포인트 중 공항으로부터 가장 가까운 항적 포인트 두 점만을 사용하였다<그림 4>. 항적 포인트 하나는 세 개의 ECEF 좌표(x,y,z)를 갖기 때문에, 가로축을 공항으로부터 거리(km), 세로축을 각 좌표(x,y,z)로 하여 한 항적마다 3번의 선형 보간법을 적용하였다.

항적별 끝점 생성 후, 2-2에서 설명한 방식대로 운항편별 길이가 300인 벡터로 벡터화하였다<그림 5>.

벡터화한 항적 데이터는 300차원으로 클러스터링 알고리즘 적합 시 거리 계산 비용이 크다.

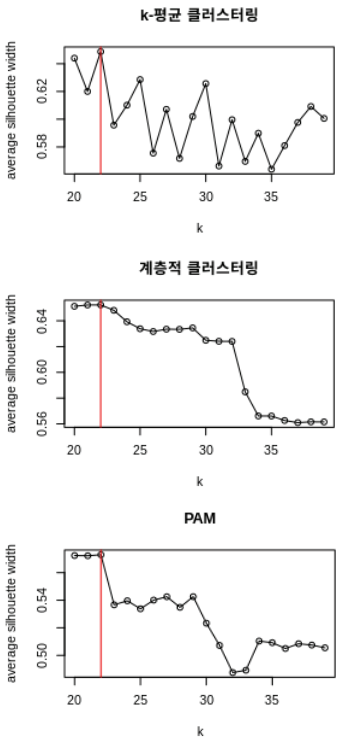


〈그림 5〉 항적 벡터화 과정(x축 : ECEF x 좌표, y축 : ECEF y 좌표, z축 : ECEF z 좌표)



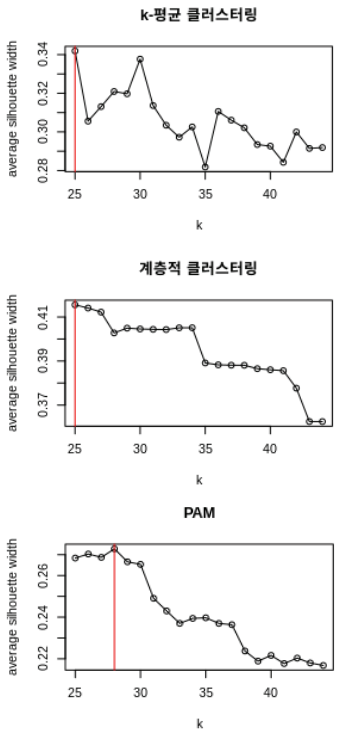
〈그림 6〉 항적 벡터화 결과

비용을 줄이기 위해, PCA를 적용하여 전체 데이터 90% 이상 분산을 갖는 축으로 차원을 축소하였다. 그 결과, 이륙 항적은 5개의 주성분으로, 착륙 항적은 10개의 주성분으로 차원을 축소하였다.



〈그림 7〉 클러스터 개수에 따른 평균 실루엣 너비 (이륙 항적)

평균 실루엣 너비를 기준으로 클러스터 개수를 선택한 결과, 이륙 항적의 경우, k-평균 클러스터링과 계층적 클러스터링, PAM 모두 22개의 클러스터 개수가 선택되었고, 착륙 항적의 경우 k-평균 클러스터링과 계층적 클러스터링은 25개, PAM은 28개의 클러스터 개수가 선택되었다.



〈그림 8〉 클러스터 개수에 따른 평균 실루엣 너비 (착륙 항적)

선택된 클러스터 개수를 이용하여 각 클러스터링 기법별 최종 적합을 진행하였다. 이상 점수는 각 항적과 해당 항적이 속한 클러스터의 중심으로부터 유클리드 거리를 계산하여 산출하였다. k-평균 클러스터링은 개체들의 평균을, PAM은 중심 개체(medoid)를 클러스터의 중심으로 선택하였고, 계층적 클러스터링의 경우 군집의 평균을 클러스터의 중심으로 선택하였다.

선택된 클러스터 개수에 따른 평균 실루엣 너비는 <표 4>와 같다. 착륙 항적 자료에 비해 이

〈표 4〉 최종 적합 결과

알고리즘	이륙 항적 평균 실루엣 너비	착륙 항적 평균 실루엣 너비
k-평균 클러스터링	0.6486959	0.3419156
계층적 클러스터링	0.6523367	0.4155161
PAM	0.5729625	0.2728285

륙 항적 자료의 평균 실루엣 너비가 높으며, 이륙 항적・착륙 항적 모두 계층적 클러스터링, k-평균 클러스터링, PAM의 순서로 평균 실루엣 너비 값을 갖는다.

산출한 이상 점수를 이용하여 이상 탐지를 진행하였다. 클러스터별 해당 클러스터에 속하는 개체들의 이상 점수를 이용하여 신뢰구간을 생성하고, 구간을 벗어나는 개체를 이상으로 판단하였다. 그 결과, 이상 항적의 개수는 <표 5>와 같다. 이륙 항적보다 착륙 항적에서 이상 항적이 더 탐지되었으며, 이륙 항적에서는 PAM, 착륙 항적에서는 k-평균 클러스터링이 가장 많은 이상 항적을 탐지하였다.

〈표 5〉 이상 항적 개수

알고리즘	이륙 이상 항적 개수	착륙 이상 항적 개수
k-평균 클러스터링	3190	3842
계층적 클러스터링	3350	3668
PAM	3471	3706

Ⅴ. 결론

본 논문은 ADS-B 수신기로부터 수집된 센서 데이터를 대상으로 클러스터링 기반의 이상 탐지 모델 설계를 제안하였다. 선형보간법을 이용하여 항적을 벡터화하고 클러스터링 기법을 적용하였다. 이상 점수는 클러스터 중심으로부터 거리를 계산하여 산출하였고, 클러스터링 알고리즘은 k-평균 클러스터링, 계층적 클러스터링, PAM을 활용하였다. 최종 적합 결과, 평균 실루엣 너비를 기준으로 계층적 클러스터링, k-평균 클러스터링, PAM 순서로 클러스터링 기법의 성능이 평가되었다.

본 연구에서 활용한 데이터는 1년간 수집이 완료된 데이터셋으로, 실제 항공 시스템에서 발생하는 데이터의 경우 실시간으로 이상 탐지를 진행하고 지시를 줄 수 있는 절차가 필요하다. 따라서, 클러스터 결과를 바탕으로 클러스터의 중심을 삼차 스플라인 등의 모형 적합을 통해 함수화를 진행하고, 연속적으로 표현하여 실시간으로 들어오고 나가는 항적과의 거리를 이상 점수로 산출할 수 있다. 또한, 회전하는 항적을 탐지하기 위해, 본 연구에서 사용한 유클리드(Euclidean) 거리 외에 삼각함수의 개념을 적용하거나 시간 및 속도를 고려하는 등 다양한 거리 계산 함수를 고려할 수 있다. 클러스터 내에 속한 데이터의 분포를 기반으로 계산하거나 항적을 회전(transformation)시켜 새로 변환된 공간에서 항적 간 거리 계산하는 방식을 사용할 수 있고 항적 간 면적을 적분하여 누적 거리 합을 구하는 방법도 수행될 수 있다.

정확하고 실용성이 높은 시스템을 구축하기 위해선, 검증하는 절차가 필요할 것이며, 수집된 데이터에 정상인지 아닌지 라벨을 부여하여 기계학습 범주 중 준-지도 학습(Semi-Supervised Learning)을 시행할 수 있다.

전조징후 탐지뿐만 아니라 항공 분야에서 발생하는 다양한 사고 원인을 찾기 위해선, 본 연구에서 사용하지 않은 기상 변수나 속도 및 시간 변수에 관한 연구 또한 지속해서 수행할 필요가 있다.

참 고 문 헌

[1] Basora, Luis, Xavier Olive, and Thomas Dubot. “Recent advances in anomaly detection methods applied to aviation.” Aerospace 6.11 (2019): 117.

[2] Cortes, Corinna, and Vladimir Vapnik. “Support-vector networks.” Machine learning 20.3 (1995): 273-297.

[3] Eckstein, Adric. “Automated flight track taxonomy for measuring benefits from performance based navigation.” 2009 Integrated Communications, Navigation and Surveillance Conference. IEEE, 2009.

[4] Ester, Martin, et al. “A density-based algorithm for discovering clusters in large spatial databases with noise.” kdd. Vol. 96. No. 34. 1996.

[5] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[6] Kaufman, Leonard, and Peter J. Rousseeuw. “Partitioning around medoids (program pam).” Finding groups in data: an introduction to cluster analysis 344 (1990): 68-125.

[7] Li, Lishuai, et al. “Anomaly detection in on-board-recorded flight data using cluster analysis.” 2011 IEEE/AIAA 30th Digital Avionics Systems Conference. IEEE, 2011.

[8] Olive, Xavier, and Luis Basora. “Identifying anomalies in past en-route trajectories with clustering and anomaly detection methods.” ATM Seminar 2019. 2019.

[9] Piciarelli, Claudio, Christian Micheloni, and Gian Luca Foresti. “Trajectory-based anomalous event detection.” IEEE Transactions on Circuits and Systems for video Technology 18.11 (2008): 1544-1554.

[10] *WEATHER-RELATED AVIATION ACCIENT STUDY,* Aviation Safety Information Analysis and Sharing, 2010.

[11] *Wrong Runway Departure,* Aviation Safety Information Analysis and Sharing, 2007.

[12] Zhang, Daqing, et al. “iBAT: detecting anomalous taxi trajectories from GPS traces.” Proceedings of the 13th international conference on Ubiquitous computing. 2011.

저 자 소 개

박 헌 진(Heon Jin Park)

·1990년 9월～1994년 8월 : SAS Institute Inc. Senior Research Statistician

·1994년～현재 : 인하대학교 통계학과 교수, 데이터사이언학과 학과장

·관심분야 : 데이터마이닝, 시계열, 통계계산

박 종 찬(Jong-Chan Park)

·2020년 2월 : 인하대학교 통계학과 (학사)

·2020년 3월～현재 : 인하대학교 통계학과 석사과정

·관심분야 : 데이터 마이닝, 빅데이터, 머신러닝, 이상 탐지