Original Article

https://doi.org/10.12985/ksaa.2021.29.2.084
ISSN 1225-9705(print) ISSN 2466-1791(online)

항공사 기단의 상태변화 시각화에 관한 연구

이용화*, 이주환*, 이금진**

A Study on the Visualization of an Airline's Fleet State Variation

Yonghwa Lee*, Juhwan Lee*, Keumjin Lee**

ABSTRACT

Airline schedule is the most basic data for flight operations and has significant importance to an airline's management. It is crucial to know the airline's current schedule status in order to effectively manage the company and to be prepared for abnormal situations. In this study, machine learning techniques were applied to actual schedule data to examine the possibility of whether the airline's fleet state could be artificially learned without prior information. Given that the schedule is in categorical form, One Hot Encoding was applied and t-SNE was used to reduce the dimension of the data and visualize them to gain insights into the airline's overall fleet status. Interesting results were discovered from the experiments where the initial findings are expected to contribute to the fields of airline schedule health monitoring, anomaly detection, and disruption management.

Key Words : Airline Schedule(항공사 스케줄), Airline Fleet(항공사 기단), Machine Learning(기계학습), Data Visualization(데이터시각화), Health Monitoring(건강진단)

Ⅰ. 서 론

항공사의 스케줄은 고객에게 제공하는 가장 기본적인 항공 정보이며, 마케팅의 결과이자 항공사가 수익을 실현할 수 있는 기초라고 할 수 있다. 스케줄은 회사 이윤 추구뿐만 아니라, 충성 고객 확보, 회사 이미지 제고 등의 측면에서 매우 중요하며, 항공당국으로부터 배정받은 운수권, 공항 슬롯(slot), 운항허가 등에 구속된 항공사 고유의 자산으로 분류될 수 있다.

Received: 03. May. 2021, Revised: 10. May. 2021, Accepted: 25. May. 2021
* 한국항공대학교 항공교통물류학과
연락저자 E-mail : yongfa97@gmail.com
연락저자 주소 : 경기도 고양시 덕양구 항공대학로 76
** 한국항공대학교 항공교통물류학부 교수

항공사 스케줄에 대한 기존 연구의 대부분은 초기 스케줄에 대한 계획과 비정상적인 항공기의 운항스케줄을 원래의 스케줄로 복원하는 데 필요한 최적화 기법에 초점이 맞춰져 있다. 항공기뿐만 아니라, 승무원의 스케줄을 최적화하는 연구도 진행되어 왔으며(Gurkan, 2016; Barnhart, 2004), 항공기 결함, 기상 등으로 스케줄에 변동이 생겼을 때 이를 정상으로 되돌리는 복원 기법 등에 관한 연구도 활발히 진행되어 왔다(Evler, 2021; Clarke, 1998; Mathaisel, 1996).

항공기와 노선이 증가할수록 글자와 숫자의 조합으로 표현된 스케줄을 사람이 인지하는 데 한계가 있으며, 이를 해결하기 위해 시각 보조물인 간트차트(Gantt chart)를 대부분의 항공사들이 활용하고 있다(Wilson, 2003). 간트차트(Fig. 1)는 실시간으로 특정 항공기가 시간대별로 어떠한 노선에 배정되어 있는지 보여줄 뿐만 아니라, 각종 운항 정보나 제한사항들을 임의로 삽입

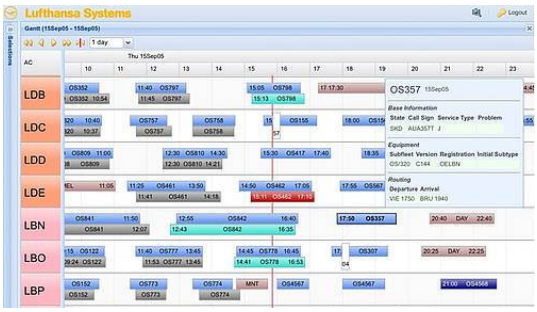

Fig. 1. Gantt chart(Lufthansa systems, 2015)

하여 항공기별 스케줄을 직관적으로 관리할 수 있는 데 그 장점이 있다(Jo, 2014).

최근 빅데이터의 중요성이 높아지면서 데이터 기반 의사 결정기법의 활용 사례가 증가하고 있으며(Shihab, 2019; Provost, 2013), 항공산업에서도 데이터를 기반으로 자료를 시각화하여 분석하는 것이 회사 경영의 중요한 축으로 자리매김하고 있다(DeGiovanni, 2017; Davenport, 2013).

항공사 스케줄은 수요 예측을 통해 노선, 시간, 기종, 정비사항 등이 확정된 영업스케줄(planned schedule)과 실제 운항 전 기상, 공항 시설, 항로제한 사항이 반영되어 조정된 운항스케줄(actual schedule)로 구분할 수 있다. 결항, 회항 등의 문제가 심각할 경우, 운항스케줄과 영업스케줄 간에는 현저한 시간 차이가 발생할 수 있으며, 운항스케줄의 정상화가 지연되면 대외적인 이미지 실추로도 이어질 수 있다. 본 연구에서는 기계학습[1] 기법을 사용해 항공사 스케줄이 정상적으로 운영되고 있는지를 시각적으로 표출하는 방법을 제안하였다. 제안된 방법은 항공기가 시간에 따라 계획된 노선을 운항하는지에 대한 실시간 데이터를 활용하여 항공사 기단(fleet)의 상태변화를 시각적으로 확인하고자 한다. 특정 시간에서의 기단의 상태가 다른 시간대와 비교하여 어떠한지를 확인할 수 있으며, 이를 통해 비정상적인 스케줄 패턴을 찾아낼 수 있다. 이는, 항공사 스케줄의 건강 상태를 측정할 수 있는 도구로써, 운항통제센터, 스케줄 기획자, 회사 경영진 등에게 활용될 수 있을 것으로 기대된다.

II. 연구 방법론

2.1 t-SNE

본 연구에서는 t-분포 확률적 임베딩(t-stochastic neighbor embedding; t-SNE)을 통해 항공사 기단의 상태변화를 시각적으로 표현하였다. 가장 전통적인 차원 축소(dimension reduction) 방법은 주성분 분석(principal component analysis; PCA)으로 고차원 데이터를 기존 변수들의 선형결합으로 변환한 후, 분산을 가장 잘 설명하는 저차원의 축으로 데이터를 표현하는 방식이다(Jolliffe, 2002). PCA는 원리가 수학적으로 이해하기 쉽고, 대부분 분야에 자료가 적용될 수 있을 정도로 활용성이 뛰어나지만 데이터의 시각화에 있어서 여러 군집을 명확히 구분하는 데 한계가 있다(Lu, 2008; Platzer, 2013). PCA 외 데이터의 시각화와 관련된 대표적 기법은 Sammon Mapping(Sammon, 1969), Isomap(Tenenbaum, 2000), t-SNE(Maaten, 2008) 등이 있다. 이 중, t-SNE는 데이터의 세부 구조를 탐색하여 군집 결과를 시각화하는 데 뛰어난 성능을 보여준다. t-SNE는 생물학 유전자의 시각화에 자주 활용되고 있으나(Kobak, 2019; Platzer, 2013), 다른 기계학습 사례(Barrat, 2019; Hong, 2015)와 달리 항공교통분야에 적용된 사례는 없으며, 본 연구에서는 항공교통, 특히 항공사 데이터가 t-SNE를 통해 효과적으로 학습이 가능한지를 살펴보았다.

t-SNE의 기본 원리는 SNE에 관한 기존 연구(Hinton, 2002)를 통해 확인할 수 있으며, 개략적인 내용은 다음과 같다. SNE는 고차원 공간에서의 데이터 점간 유사도를 조건부확률로 변환한다. 점 x_i를 중심으로 하는 가우시안(Gaussian) 확률밀도에 비례하여 이웃으로 x_j를 선택한다면 x_i가 x_j를 이웃으로 선택할 조건부확률 $p_{j|i}$는 식 (1)과 같다.

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\Sigma_{k \neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)},\ p_{i|i} = 0 \quad (1)$$

인근 점들 간 $p_{j|i}$는 상대적으로 값이 크고, 멀리 이격된 점들은 값이 매우 작아질 수 있다. σ는 점 x_i를 중심으로 하는 가우시안에 대한 표준편차이다. 고차원 공간에 대응하는 저차원 공간의 점 y_i와 y_j간의 관계는 식 (2)와 같은 조건부확률 $q_{j|i}$로 나타낼 수 있다. 여기서 적용되는 표준편차는 $q_{j|i}$가 점 y_i의 함수로 표현될 수 있도록 고차원 공간에서의 표준편차와 차별을 두었으며, $1/\sqrt{2}$로 고정하였다. 다른 고정값을 사용하더라도 저차원 공간에서 점의 분포 비율만 달라지는 특성이 있다.

1) 컴퓨터가 어떤 작업을 위해 특정 경험(데이터)으로부터 학습하여 성능에 대한 측정을 향상시키는 학문(Mitchell, 1997).

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\Sigma_{k \neq i} \exp(-\|y_i - y_k\|^2)}, \; q_{i|i} = 0 \quad (2)$$

SNE는 $p_{j|i}$와 $q_{j|i}$의 차이가 최소화될 수 있는 최적의 저차원 데이터를 찾기 위해 KL발산(Kullback-Leibler divergence)의 합인 비용함수(cost function)를 경사 하강법(gradient descent method)을 이용해 최소화한다. 비용함수는 식 (3)과 같으며, P_i, Q_i는 점 x_i, y_i와 이웃 점들간 조건부확률분포를 각각 나타낸다. 경사도(gradient)는 비용함수를 y_i에 대하여 미분한 형태로 식 (4)와 같다.

$$C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad (3)$$

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j) \quad (4)$$

t-SNE는 대칭형 SNE를 사용함으로써 비용함수를 최적화하기에 용이한 형태로 개선한 것으로, 데이터를 고차원에서 저차원으로 변환할 때 자주 겪는 밀집현상(crowding problem)을 가우시안이 아닌 t분포(Student-t distribution)로 대체함으로써 해결하였다. t-SNE는 PCA가 전역(global) 구조를 읽을 수 있는 것과 달리 국부(local) 구조를 심층적으로 이해하는 데 유용하고, 군집 특성과 같은 중요한 전역 구조를 표출하기도 한다. 아울러, 하이퍼 파라미터인 perplexity[2], learning rate[3] 등에 따라 결과가 다양하게 나와 반복 계산하는 어려움이 있으나, 다른 기법보다 점들의 군집 구분이 시각화하기에 상대적으로 명확하고 SNE보다 속도가 빠르다는 장점이 있다(Wattenberg, 2016; Maaten, 2008).

2.2 데이터 모델링

2.2.1 기단 상태변수 정의

시간 t_i에서 n개의 항공기로 구성된 전체 기단의 운항상태에 관한 상태변수 S는 식 (5)와 같다. s_{ij}는 항공기 k_j의 시간 t_i에서의 상태값으로 운항 구간 또는 주기공항에 대응한다. 즉, 항공기가 특정 노선을 비행 중이었다면 운항하고 있는 구간에 부여된 '출발공항-도착공항' 형태의 코드가 할당되고, 항공기가 착륙하여 지상에 대기 중이었다면 해당 공항에 부여된 코드가 할당된다.

$$S_{(t_i)} = \begin{bmatrix} s_{i1} \\ s_{i2} \\ \cdot \\ \cdot \\ s_{in} \end{bmatrix} \quad (5)$$

항공사는 통상적으로 편명을 기준으로 구간을 구분하고 있으나, 동일 편명으로 서로 다른 구간(이원구간, 회항 후 재운항, 부정기편 등)을 운항할 수 있으므로 편명이 아닌 출발, 도착공항의 조합으로 구간을 구분하였다.

Table 1은 기단의 크기가 4인 경우, 시간 $t_1 \sim t_4$에 따른 상태변수를 표로 나타난 예시이다. 상태변수 $S(t_1)$에서 항공기 k_1, k_4는 ICN-NRT(인천-나리타)와 GMP-CJU(김포-제주)를 각각 운항 중인 상태이며, 항공기 k_2, k_3는 ANC(앵커리지)와 SIN(싱가포르)에서 각각 주기 중인 상태이다.

상태변수의 시간 단위는 5분 단위로 추출하였으며, 시간 간격이 커질수록 데이터 처리 속도 또는 기계학습 계산 속도가 빨라지는 장점이 있으나, 기단의 상태변화 감지 능력이 떨어지는 것을 확인하였다(Appendix 1).

2.2.2 데이터 부호화

상태변수의 각 상태값은 범주형(categorical) 값으로 기계가 읽을 수 있도록 부호화(encoding) 과정을

Table 1. Example of state variable

항공기	$S(t_1)$	$S(t_2)$	$S(t_3)$	$S(t_4)$
k_1	ICN-NRT	ICN-NRT	ICN-NRT	ICN-NRT
k_2	ANC	ANC-JFK	ANC-JFK	ANC-JFK
k_3	SIN	SIN	SIN	SIN
k_4	GMP-CJU	GMP-CJU	GMP-CJU	CJU

2) 2^H(H: 섀넌 엔트로피)로 정의되며, 정보 값에 대한 측정치이다. 이는 특정 점에서의 유효한 인근점 개수를 측정하며 고차원 데이터의 조건부확률의 표준편차인 σ을 조정한다.

3) 국부 최적해에 수렴하기 위해 현재의 경사도에 특정 수를 곱하여 선형 증가하도록 조정하는 변수 즉, 단일 학습에서 변수가 변화되는 정도를 나타낸다.

거쳐야 한다. 연속된 정수(integer)로 부호화할 경우, 임의로 부여된 숫자값 간의 크기 차이로 t-SNE를 이용한 거리 계산이 왜곡될 수 있다. 따라서 본 연구에서는 범주형 변수를 부호화하는 대표적인 방법인 One Hot Encoding(이하, OHE)을 적용하였다.

OHE는 상태변수 전체 크기만큼의 벡터로 상태값이 표현되며, 각각의 상태값을 표현할 수 있도록 벡터 내 서로 다른 위치에 1을, 나머지 위치에 0을 할당하는 방식이다. OHE 벡터는 상태값 간 서로 중복하지 않고 등거리(equidistant) 및 직교(orthogonal)하는 성격을 가지므로 서로를 물리적으로 동등하게 대변할 수 있다(Cerda, 2019; Cohen, 2003). Table 2는 Table 1의 6개 상태값을 OHE한 예시이다. 실제 항공사 데이터를 OHE한 데이터 차원 결과는 3.1에서 확인할 수 있다.

III. 분석 결과

3.1 연구 대상

본 연구에서는 2016년 1월 23일 제주공항의 폭설로 인해 스케줄에 큰 변화가 생겼던 기간을 대상으로 제안된 방법을 확인하였다. 2016년 1월 23일 제주도에 12cm의 기록적인 폭설이 내리면서 같은 날 08:50 UTC (17:50 KST)부터 1월 25일 05:48 UTC까지 약 45시간 동안 공항 활주로가 폐쇄되어 국내선 전체항공사 운항편 528대가 결항 되고, 제주공항에 체류객 9만여명이 발생하였다(김아미, 2016). 1월 25일 공항 운항을 재개하고, 1월 26일 14:56 UTC까지 추가 공급석을 제공하면서 2일간 체류객 수송을 마치며 스케줄이 정상화되었다.

Table 2. Example of OHE

개수	s_{ij}	One hot encoding
1	ANC-JFK	[**1** 0 0 0 0 0]
2	GMP-CJU	[0 **1** 0 0 0 0]
3	ICN-NRT	[0 0 **1** 0 0 0]
4	ANC	[0 0 0 **1** 0 0]
5	CJU	[0 0 0 0 **1** 0]
6	SIN	[0 0 0 0 0 **1**]

스케줄 데이터는 폭설이 내린 기간을 포함한 2016년 1월 1일에서 2월 29일까지 A항공사의 실제 운항 실적을 사용하였다. 해당 기간동안 비정상 운항편을 포함한 A항공사의 운항 편수는 Table 3과 같으며, 운항 실적 데이터 중 운항 날짜, 항공기 기번, 실제 운항 시간(block to block), 운항구간(출발 공항과 도착 공항의 조합), 주기공항 자료를 사용하였다. A항공사는 제주공항에 국제선을 운항하지 않고 있으며, 제주 폭설에 따른 기단의 국내선 상태변화를 보기 위해 국제선 운항 실적은 포함하지 않았다. Fig. 2는 폭설이 내린 1월 23일 전후 1월 14일부터 1월 31일까지의 국내선 결항 현황이다.

대상 스케줄 데이터를 분석하기 위해 상태변수를 OHE한 결과는 다음과 같다. 연구 대상의 분석 기간인 2016년 1월~2월 (60일) 스케줄 데이터 내에서 기단은 총 85대의 항공기로 구성되며, 상태 공간의 크기는 422개(운항 구간 315개, 주기공항 107개)이다. 따라서 OHE를 적용할 경우, 기단의 상태변수 S는 35,870 (85×422)의 크기를 가진다. 5분 단위로 60일간의 상태변수를 추출한 결과, 샘플 데이터의 개수는 17,280 (12×24×60)가 되며, 따라서 최종 데이터 차원 크기는 17,280×35,870이 된다(Table 4).

항공기별 상태값의 카디널리티[4]는 해당 분석 기간

Table 3. Operation status from January to February 2016

편수	국제선	국내선	합계
운항	13,164	5,848	19,012
결항	51	314	365
회항	13	10	23
리턴	26	11	37
합계	13,254	6,183	19,437

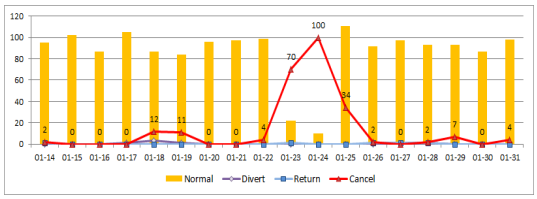

Fig. 2. Domestic operation status from 14 to 31 January 2016

4) Cardinality(관계수): 하나의 벡터에 중복되는 변수가 적을수록 카디널리티가 높음.

Table 4. OHE data dimension and cardinality

구분	내용
분석 기간	2016.1.1~2016.2.29 (60일)
추출 단위	5분
샘플 데이터 개수	17,280
기단 크기	85
상태공간 크기 (운항구간/주기공항)	422 (315/107)
상태변수 S 크기	35,870
데이터 차원(샘플 개수 × 상태변수 크기)	17,280×35,870
카디널러티	Avg13 / Max32 / Min2

내 평균 13개로 높지 않았음을 확인하였다. 통상 서로 중복되지 않는 고유한 데이터가 100개 이상인 경우 높은 카디널러티에 해당되며(Moeyersoms, 2015), 기계학습 결과에 통계적 혹은 계산상의 문제가 발생할 수 있다(Cerda, 2019).

3.2 t-SNE 하이퍼 파라미터 설정

t-SNE를 적용할 때는 일반적인 기계학습 기법과 같이 모델의 성능에 영향을 미칠 수 있는 하이퍼 파라미터를 다양하게 실험해야 한다(Claesen, 2015; Maaten, 2008). 가장 중요한 하이퍼 파라미터인 perplexity는 통상 5와 50 사이가 전형적인 값이지만, 데이터의 특성에 따라 더 큰 값에서 좋은 결과가 나올 수 있다(Kobak, 2019; Cao, 2017). OHE를 적용할 경우, 데이터가 드문(sparse) 특성을 가지게 됨을 감안하여, 본 연구에서는 perplexity 값을 10에서부터 최대 290까지 20단위씩 증가시켜 가면서 결과를 확인하였다. 또, 다른 하이퍼 파라미터인 learning rate는 매우 큰 경우(10,000 이상) 점들간 구분이 어려운 경우가 종종 발생하므로 우리는 500, 1,000, 1,500을 각각 적용하였다. 그 밖에, 반복 실행 횟수(iteration)는 10,000으로 고정하였다.

데이터 크기 및 차원이 커질수록 각 점간 확률분포를 계산하는 시간 및 메모리 할당량이 급격히 증가하여 인근 점들을 그룹으로 묶어 계산하는 Barnes-Hut 근사법을 적용하였다(Maaten, 2013). 또한, t-SNE 실행 시마다 초기 점들이 무작위로 설정되는 것을 막기 위해 인위적으로 초기값을 고정하여 동일 초기값에 대한 각 하이퍼 파라미터별 변화를 일관되게 확인하였다.

데이터 차원이 커질수록 점간 유사도 측정이 어려워지게 되며(Aggarwal, 2001), 계산 속도 또한 느려져(Maaten, 2008) 기계학습 전 PCA와 같은 차원 축소를 시행하지만, 본 연구의 OHE 데이터의 경우 비록 고차원이라도 t-SNE를 통한 직접적인 군집 및 시각화에 무리가 없다고 판단하여 차원 축소를 하지 않았다. 실제로 PCA를 통해 차원 축소(주성분 50개 이상 기준)를 시행하더라도 결과에 유의미한 변화가 없음을 별도로 확인하였다(Appendix 2). 분석을 위해 MATLAB (ver. R2020b)을 이용하였다.

3.3 t-SNE 분석 결과

2016년 1월~2월 국내선 데이터에서 5분 단위로 추출된 기단의 상태변수를 OHE 후 부호화된 자료에 t-SNE를 적용하였다. t-SNE는 각 상태변수에 대하여 2차원 값을 반환하므로, Table 4의 17,280개의 샘플 데이터를 2차원 공간에 그릴 수 있다. Perplexity 30에서 170, learning rate 1,000의 계산 결과는 Fig. 3과 같으며, 실험 결과 learning rate보다 perplexity 값에 결과가 민감하게 반응함을 확인하였다. Perplexity 110 이상부터 육안으로 군집이 형성되는 것을 확인할 수 있고, 200을 넘어가는 경우 비약적인 변화는 볼 수 없어 결과에서는 생략하였다. 그림에 나타난 바와 같이, 기단의 상태가 주(week) 단위로 군집이 되는 것을 알 수 있는데, 이는 항공사 스케줄이 주 단위로 반복되는 특징을 가지기 때문이다.

Fig. 4는 Fig. 3의 perplexity 150, learning rate 1,000의 결과에 대해 제주 폭설 기간에 해당하는 영역을 확대한 것이다(Fig. 3 적색 사각형). 그림의 상단은 일자별(1월 23일~1월 31일), 하단은 UTC 기준 시간대별(0시~23시) 레이블을 적용하였다. 확대된 분포를 보면 일자별, 시간대별 점들이 선 형태로 연결되어 있는 것을 볼 수 있는데, 일반적으로 직교(orthogonal) 성분 자료를 t-SNE에 적용하면 데이터의 군집이 선 형태로 나타나게 된다(Wattenberg, 2016). 거의 모든 날짜에 대해 14:00~21:00 UTC에 해당하는 데이터는 작은 타원형 군집을 이루는 것이 확인된다. 이는 해당 시간대에는 김포공항을 포함한 국내선 일부 공항에 심야 운항 제한(이하, curfew)이 적용되기 때문이다. 즉, 국내선의 기단 상태의 경우 시간에 따라 일정한 폭으로 표류하는 형태를 보이다가, 운항이 전무한 curfew 시간에는 표류를 멈추고 타원 군집을 형성 후 다시 표류하는 과정을 반복함을 알 수 있다.

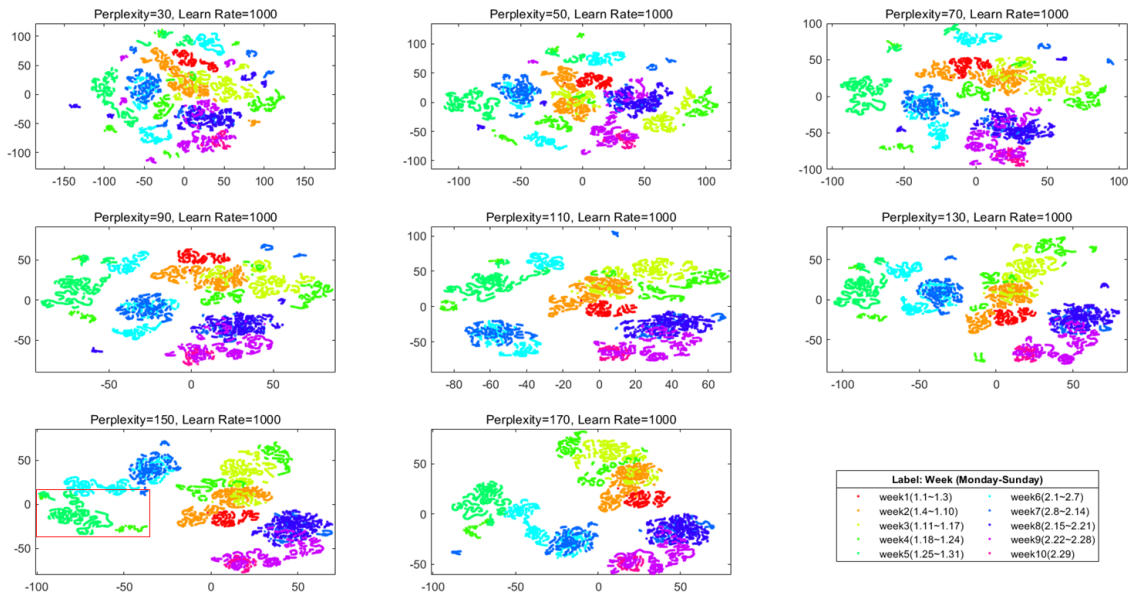

Fig. 3. Visualization of fleet state from January to February 2016 using t-SNE(labels are in weeks)

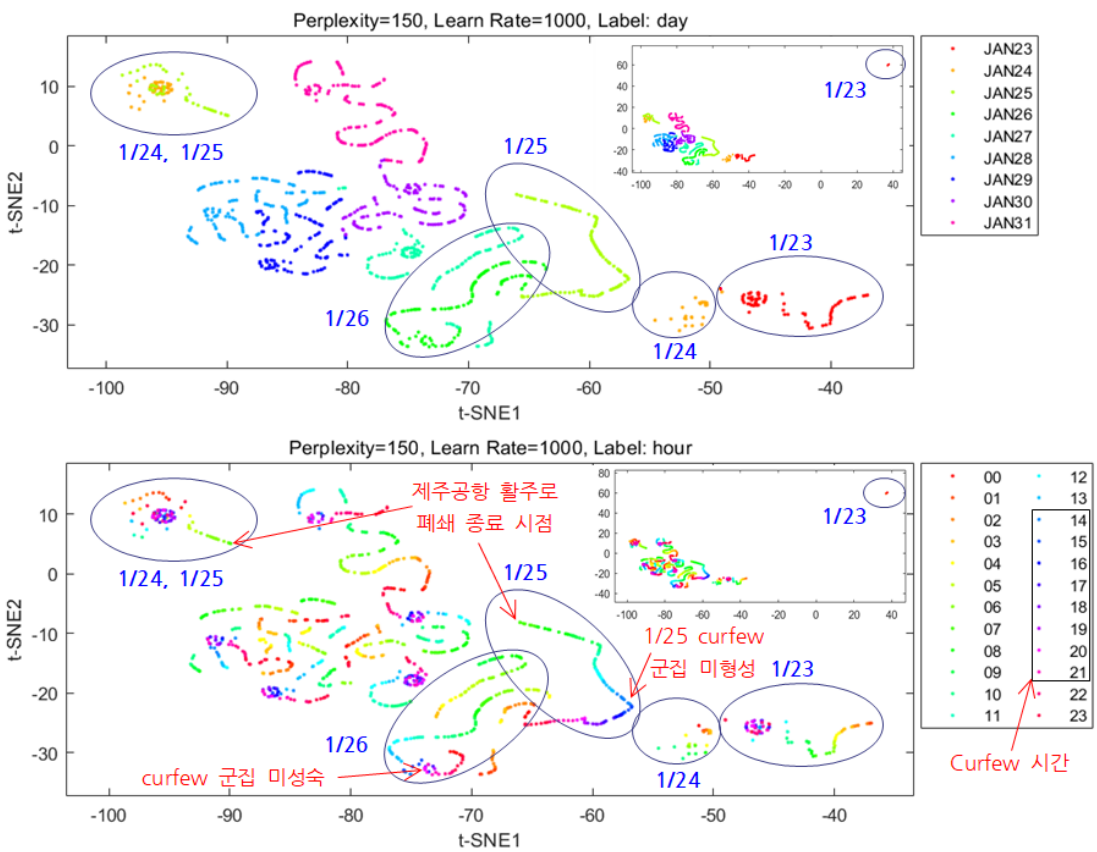

Fig. 4. Enlarged view of Fig. 3 for schedule disruption period at Jeju Airport
(Labels: Days for upper figure, hours for lower figure)

제주 폭설로 결항이 다수 발생한 2016년 1월 23일~1월 25일에는 기단의 상태변수가 다른 날들과는 구별되게 행동함을 알 수 있다. 우선 제주공항 활주로가 폐쇄되기 전 폭설이 내리기 시작하여 결항이 실제 시작된 1월 23일 01:00 UTC부터 1월 24일 사이의 기단의 상태변수 값이 해당 주(1월 4째주)의 군집에서 매우 멀리 떨어져 있음을 알 수 있는데, 이는 비정상적인 상황이 발생했음을 말해 주고 있다. 또한, 제주공항 활주로가 처음 폐쇄된 1월 23일 09:00 UTC 전후 점들은 일부 흩어져 있으나 육안으로 변화를 인지하기 힘들 정도로 상태변수의 변화가 적으며, 특히 비행이 완전히 중단된 1월 24일은 선 형태가 아닌 산개한 구조로 기단의 상태가 표출되어 있어 기단의 상태가 비정상적임을 쉽게 알 수 있다. 또한, 제주공항 활주로 폐쇄가 종료된 1월 25일 06:00부터 점들의 정상적인 표류 흐름이 재개된 것은 눈여겨볼 만한 특징이다.

1월 23일과 1월 24일은 항공기 운항시간대인 01:00~12:00 UTC에도 결항편이 많이 발생하여 해당 시간대의 일부 점들이 curfew에 편입되면서 다른 날보다 타원 군집이 다소 커진 것을 볼 수 있다. 또한, 1월 25일은 제주공항 운항 재개 이후 체류객 수송을 위해 김포공항 curfew가 최초로 해제(강동효, 2016)되면서 19편이 심야 시간에 운항하였는데, 이로 인해 1월 25일은 curfew 시간대에 타원형 군집을 형성하지 못하고 일반 운항시간대와 같이 표류하는 흐름을 보이고 있다. 1월 26일은 다소 적은 3편이 심야 시간에 운항하여 타원 군집이 다른 날 대비 성숙하지 못한 것도 확인이 된다.

t-SNE 적용 시 제주공항 활주로 폐쇄 등과 같은 정보를 반영하지 않았음에도 불구하고, 시각화된 데이터를 통해 해당 기간 동안 기단의 스케줄이 비정상적으로 작동하였음이 드러나게 되었다는 사실에 주목할 필요가 있다. 즉, 항공기 스케줄이 일반적으로 이행되는 경우 t-SNE 점들이 거의 일정한 간격으로 표류하듯 흘러간다는 것을 알 수 있었으며, 반대로 비운항 시간대에는 t-SNE 점들의 움직임이 멈춰 해당 시간대 스케줄이 없다는 점을 알 수 있었다. 또한 일반적인 스케줄을 벗어난 비정상적인 상황에서는 점들의 흐름이 기존과 다르게 산개되거나 갑작스럽게 다른 공간으로 튀는 현상이 발생한다는 점을 확인할 수 있었다. 이는 제안된 기법을 이용하여, 스케줄의 정상, 비정상 상황을 감지할 수 있을 뿐만 아니라, 정상 스케줄 중에서도 운항, 비운항 사항을 구분할 수 있음을 의미하고, 이와 같은 정보는 항공사의 기단 운영상태를 모니터링하는 데 활용될 수 있다.

Ⅳ. 결 론

기업의 데이터를 기계 학습하여 고객에게 서비스의 질과 업무 효율성을 높이려는 움직임이 최근 대부분의 산업 분야에서 각광을 받고 있다. 하지만 항공산업에서의 항공사 자료에 대해 기계학습을 적용한 사례는 제한적이다. 본 연구에서는 항공사 기단의 상태변화를 t-SNE 기법을 활용해 모니터링할 수 있는 방법이 제안되었다. 이를 위해 기단의 상태변수를 정의하고, 항공사 스케줄 데이터를 추출 및 부호화한 후 t-SNE를 적용하여 시각화하였다. 본 연구에서 제안한 방법을 통해, 기단의 상태를 일자별, 시간대별로 구분할 수 있음을 확인하였고, 특히 비정상 운항에 대한 상태변화를 충분히 감지할 수 있음을 확인하였다.

본 연구는 특정 항공사의 운항 네트워크 및 항공기 규모 기준의 결과만을 보여주고 있지만, 유사한 규모의 항공사에 동일 방법론의 적용이 가능할 것으로 판단된다. 또한, 과거 실적뿐만 아니라, 항공사의 실시간 스케줄을 이용해 기단의 상태가 특정 경계 이상으로 변하는 시점을 확인함으로써 특이 상태를 탐지하는 데 활용할 수 있을 것으로 기대된다. 또한 항공사의 주요 미래 스케줄(신규 노선, 시즌 스케줄 변화, 운휴 후 재운항, 항공기 도입, 항공기 운항정지 후 재운항 등)을 활용하여 기단의 건강 상태를 사전에 진단할 수 있을 것으로 기대된다. 이러한 후속 연구를 통해 항공사의 안전관리시스템(safety management system)에서 추구하는 예방형, 예측형 안전활동에 기여할 수 있다.

향후 연구로는 항공사 데이터에 적합한 보다 다양한 기계학습 기법을 적용해 볼 필요가 있다. 범주형 자료에 대한 부호화 방법을 달리하고(예, target encoding 등), 시각화 툴을 다양하게 적용(예, Isomap 등)하여 결과를 개선할 수 있다. 또한, 기단의 상태변화를 확인하기 위해 스케줄의 미세한 증감에서부터 기타 항공대란급 사건에 대해 제안된 방법을 폭넓게 적용해 볼 필요가 있으며, 화물기와 여객기, 정기편과 부정기편, 대형기와 소형기 등으로 대상을 다변화하여 기단의 상태를 확인하는 작업이 필요하다.

본 연구가 건강진단, 비정상 상태 탐지의 툴로 활용되기 위해서는 현재, 미래 스케줄을 접목하는 것도 중요하지만, 기단 상태가 특이 상태로 전환되는 경계(threshold)에 관한 후속 연구가 필수적이다. 마지막으로, 항공사의 규모는 서로 다를 수 있으므로 각 규모에 적합한 다양한 기계학습 기법을 실험할 필요가 있다.

후 기

본 연구는 국토교통과학기술진흥원의 “데이터기반 항공교통관리 기술개발” 과제의 일환으로 수행되었으며 지원에 감사드립니다.

References

1. Gürkan, H., Gürel, S., and Aktürk, M. S., “An integrated approach for airline scheduling, aircraft fleeting and routing with cruise speed control”, Transportation Research Part C 68, 2016, pp.38-57.
2. Barnhart, C., and Cohn, A., “Airline schedule planning: Accomplishments and opportunities”, Manufacturing & Service Operations Management, 6(1) Winter, 2004, pp.3-22.
3. Evler, J., Asadi, E., Preis, H., and Fricke, H., “Airline ground operations: Optimal schedule recovery with uncertain arrival times”, Journal of Air Transport Management 92, 2021, DOI: 10.1016/j.jairtraman.2021.102021
4. Clarke, M. D. D., “Irregular airline operations: A review of the state-of-the-practice in airline operations control centers”, Journal of Air Transport Management 4, 1998, pp.67-76.
5. Mathaisel, D. F. X., “Decision support for airline system operations control and irregular operations”, Computers & Operations Research, 23(11), 1996, pp.1083-1098.
6. Wilson, J. M., “Gantt charts: A centenary appreciation”, European Journal of Operational Research, 149, 2003, pp.430-437.
7. Jo, J., Huh, J., Park, J., Kim, B., and Seo, J., “LiveGantt: Interactively visualizing a large manufacturing schedule”, IEEE Transactions on Visualization and Computer Graphics, 20(12), 2014.
8. Shihab, S. A. M., Logemann, C., Thomas, D. G., and Wei, P., “Autonomous airline revenue management: A deep reinforcement learning approach to seat inventory control and overbooking”, arXiv:1902.06824 [cs.AI], 2009.
9. Provost, F., and Fawcett, T, “Data science and its relationship to big data and data-driven decision making”, Mary Ann Liebert, Inc., 1(1), Feb. 13, 2013.
10. DeGiovanni, J. J., “Seeing the data: United airlines implements new methods of analyzing safety data and improving performance”, Flight Safety Foundation, 2017, https://flightsafety.org/asw-article/seeing-the-data
11. Davenport, T. H., “At the big data crossroads: Turning towards a smarter travel experience”, Amadeus IT Group, 2013, https://amadeus.com/documents/en/blog/pdf/2013/07/amadeus-big-data-report.pdf
12. Lufthansa Systems, “Manage Your Airline Operations by Exception”, Lufthansa Systems GmbH & Co. KG, 2015, https://www.lhsystems.com/static/dde9d5c2f582d72ba75c3cf938346263/pb_netline_ops_0.pdf
13. Mitchell, T. M., “Machine Learning”, McGraw-Hill Science, Engineering, Math, New York, NY, USA, 1997, pp.2.
14. Jolliffe, I. T., “Principal Component Analysis, Second Edition”, Springer Verlag, New York, NY, 2002, pp.10-28.
15. Lu, H., Plataniotis, K. N., and Venetsanopoulos, A. N., “MPCA: Multilinear principal component analysis of tensor objects”, IEEE Transactions on Neural Networks, 19(1), 2008.
16. Platzer, A., “Visualization of SNPs with t-SNE”, PLoS ONE 8(2), 2013, e56883, DOI: 10.1371/journal.pone.0056883
17. Sammon Jr, J. W., “A nonlinear mapping for data structure analysis”, IEEE Transactions on Computers, C-18(5), 1969.
18. Tenenbaum, J. B., Silva, V. D., and Langford, J. C., “A global geometric framework for nonlinear dimensionality reduction”, Science, 290, 2000, pp.2319-2323.
19. Maaten, L. V. D., and Hinton, G., “Visualizing data using t-SNE”, Journal of Machine Learning Research, 9, 2008, pp.2579-2605.
20. Kobak, D., and Berens, P., “The art of using

t-SNE for Single-cell Transcriptomics", Nature Communications 10(5416), 2019, DOI: https://doi.org/10.1038/s41467-019-13056-x
21. Barratt, S. T., Kochenderfery, M. J., and Boyd, S. P., "Learning probabilistic trajectory models of aircraft in terminal airspace from position data", IEEE Transactions on Intelligent Transportation Systems, 2019, DOI: https://doi.org/10.1109/TITS.2018.2877572
22. Hong, S., and Lee, K., "Trajectory prediction for vectored area navigation arrivals", Journal of Aerospace Informations Systems, 12(7), 2015.
23. Hinton, G. E., and Roweis, S. T., "Stochastic Neighbor Embedding", Advances in Neural Information Processing Systems, The MIT Press, Vol. 15, Cambridge, MA, USA, 2002, pp.833-840.
24. Wattenberg, M., Viégas, F., and Johnson, I., "How to use *t*-SNE effectively", Distill, 2016, DOI: http://doi.org/10.23915/distill.00002
25. Kim, A. M., "Jeju Airport Resumes Operations at 14:48. Evacuation Will Take Three Days", Herald Economy, 2016, URL: http://news.heraldcorp.com/view.php?ud=20160125001029
26. Cerda, P., and Varoquaux, G., "Encoding High-Cardinality String Categorical Variables", ffhal02171256v1, 2019.
27. Cohen, J., Cohen, P., West, S. G., and Aiken, L. S., "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences Third Edition", Lawrence Erlbaum Associates, Inc., Publishers, Mahwah, NJ, USA, 2003, pp.303-320.
28. Moeyersoms, J., and Martens, D., "Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector", Decision Support Systems, 72, 2015, pp.72-81.
29. Claesen, M., and De Moor, B., "Hyperparameter Search in Machine Learning", 2015, arXiv:1502.02127
30. Cao, Y., and Wang, L., "Automatic selection of *t*-SNE Perplexity", 2017, arXiv:1708.03229
31. Maaten, L. V. D., "Barnes-Hut-SNE", 2013, arXiv:1301.3342v2
32. Aggarwal, C. C., Hinneburg, A., and Keim, D. A., "On The Surprising Behavior of Distance Metrics in High Dimensional Space", Van den Bussche J., Vianu V. (Eds.) Database Theory, ICDT 2001, Berlin, Heidelberg, 2001, pp.420-434.
33. Kang, D. H., "The Strongest Cold Wave in 15 Years, Gimpo, Gimhae Airport's Curfew Suspension, Historical Overnight Operations", Seoul Economy, 2016, URL: https://www.sedaily.com/NewsVIew/1KRCIE4WAQ

Appendix

본문에서는 상태변수의 시간 간격을 5분으로 설정하였다. 시간 간격이 커질수록 기단의 상태변화 감지 능력이 떨어지는 것을 확인하기 위해, 시간을 5분, 30분, 60분 간격으로 추출한 경우에 대한 *t*-SNE 결과 비교는 Fig. A.1과 같다.

t-SNE의 데이터 분류 성능을 향상시키기 위해 PCA를 통한 차원 축소를 적용할 수 있다. 하지만 본 연구

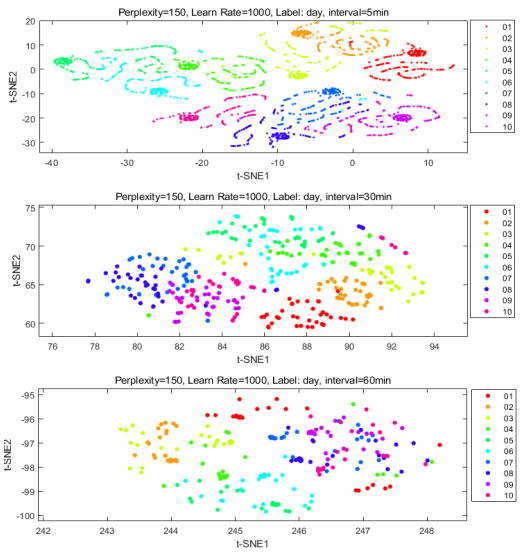

Fig. A.1. Comparison of *t*-SNE results for different time intervals (Jan 1 to 10, 2016)

에서 사용한 데이터의 경우 PCA 차원 축소로 인한 효과가 크지 않았는데, Fig. A.2는 PCA 적용 전과 70, 50개의 주성분으로 차원 축소 후 t-SNE를 적용하여 비교한 결과이다.

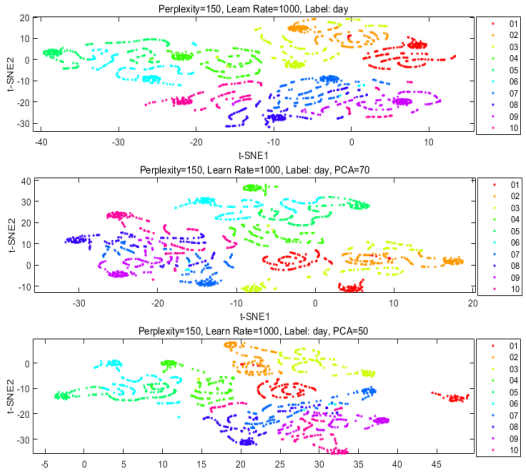

Fig. A.2. Comparison of t-SNE results before and after applying PCA (Jan 1 to 10, 2016)