

데이터 변수의 유효성 검증과 이상치 정보제공에 대한 연구

박현진¹, 백명선^{2,3}

¹ 청주대학교, ² 한국전자통신연구원, ³ 과학기술연합대학원대학교

thddnjs45@cj.ac.kr, sabman@etri.re.kr

A Study on the Validation of Data Variables and the Provision of Outliers

Park Hyun Jin¹, Myung-Sun Baek^{2,3}

¹Cheongju Univ., ²ETRI, ³UST

요 약

본 논문에서는 연합 디지털 트윈 동작을 위한 다종의 데이터 유효성 검증 방법을 제시한다. 다종의 데이터 유효성을 검증하기 위해, 시계열 데이터와 분류 유형 데이터가 고려되었다. 또한 유효성 검증 방법론 제시를 위해 해당 데이터에 대한 변수의 유효성을 검증하고 이상치 탐지 및 이상원인을 파악할 수 있는 방법을 연구하였다. 본 연구를 통해 시계열 데이터의 변수들과 분류유형 데이터와의 상관관계가 도출되었으며, 도출된 상관관계 결과를 반영하여 데이터 전처리 및 관리도 연산을 통한 데이터 이상치의 정보제공이 가능하다.

I. 서 론

디지털 트윈이란 가상공간에 현실 세계에 있는 실물과 똑 같은 가상물체를 만들어 현실에서는 여러 제약조건으로 해볼 수 없는 다양한 모의시험을 통해[1], 현실세계의 문제들을 해결하거나 가치를 창출하기 위해 개발된 기술이다. 최근 다양한 분야에서 설계 검토, 작업 환경 시뮬레이션 등 상기 디지털 트윈을 적용하는 사례들이 점차 증가하고 있다.

그러나 도시, 국가 단위의 거대 규모 디지털 트윈을 구축하는 경우, 디지털 트윈에 포함된 독립변수 및 종속 변수의 수 등이 크게 증가하기 때문에 구현의 오류 발생률과 관리 복잡도의 문제가 발생될 수 있다. 따라서 여러 단일 디지털 트윈을 연계하여 거대한 연합 디지털 트윈을 구현하는 것이 대체적인 방법이 될 수 있다[2]. 연합 디지털 트윈을 구현하는 과정에서 다양한 개별 디지털 트윈들의 데이터가 유효하지 않거나 오류를 포함한 데이터인 경우, 연합 디지털 트윈의 동작 및 출력에도 오류가 발생할 수 있다.

따라서 상기 연합 디지털 트윈의 오류 발생률을 감소시키기 위해 단일 디지털 트윈 데이터 및 동작의 이상징후 발생 여부와 이상원인을 파악하는 기술이 요구될 수 있다.

본 논문에서는 다종의 데이터 유형 중 분류 유형 데이터의 변수의 유효성을 검증하고, 분류 유형 데이터와 시계열 데이터가 연계된 출력 값의 이상치 정보 및 이상원인을 파악하는 방안을 연구하였다.

II. 데이터 및 분석 모델

본 논문에서는 2016년 1월 1일부터 2017년 12월 31일까지 시계열데이터(서울시 기상관측데이터)와 분류 유형 데이터(서울시 공공자전거 이용현황)의 회귀분석을 통해 변수의 유효성을 검증한다. 시계열데이터의 변수로는 평균기온, 최저기온, 최고기온, 일 강수량, 최대순간풍속, 평균풍속 등이 있다.

시계열 데이터의 변수와 분류 유형 데이터 사이의 상

관관계 및 각 변수들의 유의성을 파악하기 위해 회귀분석을 진행하였다. 회귀분석 수행 시 분석기준으로 계절(봄:3월~5월, 여름:6월~8월, 가을:9월~11월, 겨울:12월~1월)을 선택하여, 각 계절 별로 데이터를 분할하여 분석하였다. 표 1은 9월부터 11월까지의 가을에 해당하는 기간 동안의 회귀분석 결과 중 회귀모형을 보여준다.

표 1. 서울시 가을(9~11월) 기후와 자전거 대여내역 데이터 회귀모형

모형	비표준화 계수 B	표준화 계수		t	유의확률
		표준화 오류	베타		
1 (상수)	123.571	144.555		0.855	0.393
평균기온C	113.272	52.630	0.871	2.152	0.032
최저기온C	-60.250	28.008	-0.470	-2.151	0.032
최고기온C	4.061	29.159	0.032	0.139	0.889
일강수량mm	-32.865	6.107	-0.150	-5.382	<0.001
최대순간풍속ms	6.683	20.126	0.016	0.332	0.740
평균풍속ms	122.435	57.596	0.100	2.126	0.034

a. season_1 = 3.00 가을

b. 종속변수: 대여내역

회귀모형을 기반으로 다음과 같은 가설을 세울 수 있다.

- 귀무가설(H0): 각 독립변수와 종속변수는 독립이다.
- 대립가설(H1): 각 독립변수와 종속변수는 독립이 아니다.

각 변수의 유의확률 p 중 유의수준 0.05 보다 작은 값($p < 0.05$)을 갖는 변수들의 경우에는 귀무가설이 기각되어, 종속변수인 자전거 대여내역에 유의한 영향을 미치는 변수가 된다. 또한 변수의 영향력을 나타내는 베타계수(β)의 부호로 종속변수와 음/양의 상관관계 유형을 알 수 있다. 표 1에서와 같이 가을(9월~11월)의 경우 평균기온, 최저기온, 일 강수량, 평균풍속 변수들이 자전거 대여내역과 연관이 있음을 알 수 있고, 유의한 변수들의 베타계수를 이용하여 대여내역에 대한 아래의 회귀식을 작성할 수 있다.

$$y = 113.272 \cdot \bar{T} - 60.250 \cdot T_L - 32.865 \cdot R + 122.435 \cdot \bar{W} \quad (1)$$

식 (1)에서 y 는 자전거 대여내역이고 $\bar{T}$ 와 T_L 은 평균기온과 최저기온을 각각 나타낸다. 또한 R 과 $\bar{W}$ 는 일 강수량과 평균풍속을 의미한다. 상기 회귀식을 사용하면 주어진 기상데이터 기반의 대략적인 자전거 대여내역을 계산할 수 있으며, 계산된 값과 큰 차이가 나는 자전거 대여내역의 경우 이상치로 판단할 수 있다.

III. 데이터 이상치 탐지 방법

표 2. 공공자전거 대여횟수 데이터 예시

지역구분 측정시점	1 번째 구 (광진구)	2 번째 구 (마포구)	3 번째 구 (서대문구)	4 번째 구 (성동구)	5 번째 구 (영등포구)	6 번째 구 (중구)
1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	X_{n1}	X_{n2}	X_{n3}	X_{n4}	X_{n5}	X_{n6}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	X_{N1}	X_{N2}	X_{N3}	X_{N4}	X_{N5}	X_{N6}

본 절에서는 통계적 공정관리 도구 중 관리도[3]를 이용한 데이터의 이상치 탐지, 정보 제공의 방법을 제시한다. 관리도 기반의 이상치 탐지를 위해 2016년 가을 서울시 6 개구의 일일 공공자전거의 대여횟수를 고려하였다. 고려된 공공자전거 대여횟수 데이터에 대한 예시는 표 2와 같이 나타낼 수 있다. 본 연구에서는 표 2에서와 같은 6 개구에 대한 일일 평균 자전거 대여횟수의 평균값을 활용하여 이상치 탐지를 수행하였다. 관리도 기반의 이상치 탐지의 경우 데이터의 평균값인 $\bar{X}$ 와 데이터의 변화범위인 R 값을 이용한 $\bar{X}$ -bar-R chart를 이용하게 되며, $\bar{X}$ -bar-R chart에 이용되는 변수값은 다음과 같이 구할 수 있다.

$$\bar{X}_n = \frac{1}{6}(X_{n1} + X_{n2} + X_{n3} + X_{n4} + X_{n5} + X_{n6}) \quad (1)$$

$$\bar{\bar{X}} = \frac{1}{N} \sum_{n=1}^N \bar{X}_n \quad (2)$$

$$R_n = \max(X_{n1}, X_{n2}, X_{n3}, X_{n4}, X_{n5}, X_{n6}) - \min(X_{n1}, X_{n2}, X_{n3}, X_{n4}, X_{n5}, X_{n6}) \quad (3)$$

$$\bar{R} = \frac{1}{N} \sum_{n=1}^N R_n \quad (4)$$

$$UCL = \bar{\bar{X}} + A_2 \cdot \bar{R} \quad (5)$$

$$LCL = \bar{\bar{X}} + A_2 \cdot \bar{R} \quad (6)$$

상기 식에서 $\bar{X}_n$ 는 6 개구의 일일 평균대여횟수이고 $\bar{\bar{X}}$ 는 관찰구간인 N 일 동안의 6 개구의 전체 대여횟수의 평균값이다. 또한 UCL 과 LCL 은 각각 관리상한(Upper Control Limit)과 관리하한(Lower Control Limit)을 나타내며, A_2 는 통계적 공정관리 도구에서 사용되는 계수로 본 논문에서는 가장일반적으로 사용되는 0.483을 적용하였다 [3].

그림 1은 상기 변수를 적용한 자전거 대여횟수에 대한 $\bar{X}$ -bar-R chart를 보여준다. 그림에서 각 점들은 6 개구의 일일 평균대여횟수($\bar{X}_n$)를, 파란색 선은 모든 점들의 평균($\bar{\bar{X}}$)을, 빨간색 선들은 관리한계로 UCL 과 LCL 을 각각 나타낸다. 그림 1에서 관리한계를 벗어난 데이터를 이상치로 판단할 수 있다.

그림 2는 관리도 기반의 이상치 탐지/정보 제공 기술을 활용한 데이터 모니터링 및 이상치 검출 시스템 활용예시를 보여준다. 그림에서와 같이 데이터 모니터링 중 이상치가 검출되면 해당 이상치 데이터를 시스템 화면에 표출하고, 그와 관련된 연계 데이터를 오른쪽에 표시하여

이상치 검출데이터 및 이상치 발생 사유를 쉽게 파악할 수 있도록 하는 것이 가능하다.

$\bar{\bar{X}}$ UCL LCL
588.6848 1110.7343 66.6353

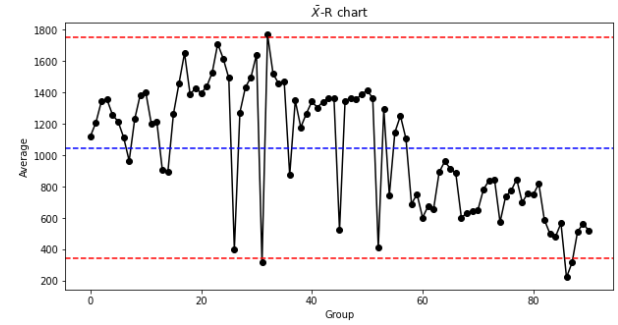


그림 1. $\bar{X}$ -bar-R 차트(2016년 가을, 9월~11월)

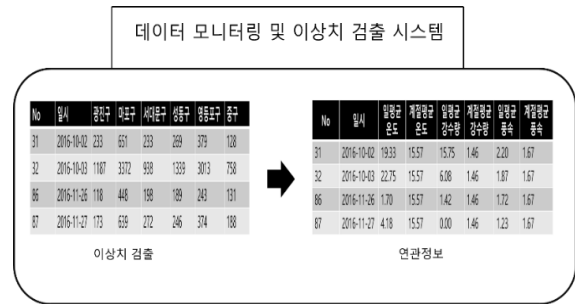


그림 2. 데이터 모니터링 및 이상치 검출 시스템 활용예시

상기 코드를 이용하여 얻은 이상치의 정보를 바탕으로 이상원인을 파악할 수 있다. 제안된 방법을 단일 디지털 트윈에 적용시켜 이상징후가 발생될 경우 이상원인을 파악하고 문제점을 해결하여 연합 디지털 트윈의 오류 발생률을 감소시킬 수 있다.

III. 결론

거대규모 디지털 트윈환경을 구축하는 경우 다수개의 단일 디지털 트윈이 연합된 형태의 연합 디지털 트윈의 구현이 필요하다[2]. 이러한 연합디지털 트윈의 원활한 동작을 위해서는 단일 디지털 트윈 데이터의 유효성 및 출력 값의 이상치, 이상원인을 탐지하는 기술이 필요하다. 본 논문에서는 연합디지털 트윈 동작 오류를 감소시키기 위해 다종의 데이터검증 방안을 시계열 데이터와 분류 유형 데이터를 예로 들어 연구하였다. 제안된 방법을 통하여 데이터 및 개별 트윈의 공정상태를 검증하게 되면 원활한 연합 디지털 트윈 동작이 가능하다.

ACKNOWLEDGMENT

이 논문은 2022년도 정부 (과학기술정보통신부)의 재원으로 정보통신기획평가원의 100% 지원을 받아 수행된 연구임 (No. 2022-0-00545, (2세부) 지능형 디지털 트윈 연합 객체 구성 및 데이터 프로세싱 기술 개발)

참고 문헌

- [1] 정보통신기획평가원, "디지털 트윈 기술 K-로드맵"
- [2] 백명선 외, "연합 디지털 트윈을 위한 데이터 유효성 검증", 한국통신학회 하계학술대회, 2022. 06.
- [3] Ju-Hyun Lee, "Statistical_Quality_Control/SQC_chart.py" (https://github.com/lee-ju/Statistical_Quality_Control/tree/main/Statistical_Quality_Control).