이상치 검출 방식

1. 결측치 채워넣기 Missing Value Imputation

1) 중심 경향 값 넣기(평균, 중앙값, 최빈값 등) – 분산이 줄어들고, 소수의 평균이 전체를 대표하는 경우가 생김. 극단값에 의해 평균이 영향 받음.

Mid-minimum spacing: 양측 5%제거하고 평균 (예) 피겨스케이팅 점수 계산)

2) 랜덤 추출(분포 기반) – 랜덤에 의해 자주 나타나는 값이 채워짐

3) Regression Imputation(회귀 삽입) – 변수 내의 값들의 평균이 아닌 각 관측치의 특성을 고려하여 삽입, 기초 Imputation 먼저하고 회귀식에 의해 타겟 소실 데이터 채움

4) EM algorithm – 기초 Imputation – 회귀분석 - y데이터 변형 – 회귀분석 - y데이터변형

변화량이 작을 때까지 반복

5) Multiple Imputation – 다양한 모델 여러 번 반복

2. 이상치 Outlier 검출(이상치: 속성의 값이 일반적인 값보다 편차가 큰 값)

1) Variance: 정규분포에서 97.5% 이상 또는 2.5%의 이하에 포함되는 값을 이상치로 판별

2) Likelihood: 베이즈 정리에 의해 데이터 셋이 가지는 두가지 샘플(정상/이상)에 대한 발생 확률(Likelihood)로 이상치 판별

3) Nearest-neighbor: 모든 데이터 쌍의 거리를 계산하여 이상치 검출

4) Density: 샘플의 LOF(Local Outlier factor)를 계산하여 값이 가장 큰 데이터를 이상치로 추정, 밀도있는 데이터 셋으로부터 먼 거리

5) Clustering: 데이터를 여러 클러스터로 구분한 후 작은 크기의 클러스터나 클러스터 사이의 거리를 계산하여 먼 경우 해당 클러스터를 이상치로 판별

3. 이상치 대체(Outlier Replacement)

1) 하한값과 상한값을 결정한 후 하한값보다 적으면 하한값으로 대체, 상한값보다 크면 상한값으로 대체

2) 평균의 표준편차: 하한값 = 평균 – n*표준편차, 상한값 = 평균 + n*표준편차 (일반적으로 3시

그마, 99.7% 이상 혹은 이하를 이상치로 제거하거나 대체)(n=3 or 2.75)

3) 평균 절대 편차, 평균 절대 오차(Mean Absolute Deviation, Mean Absolute Error) : 중위수로부터 n편차 큰 값을 대체

모든 절대 오차의 평균

-> PCA에서 모델 검증 시 사용

MAE=(예측값-실제값)의 절댓값 평균

4) 극 백분위수: 상위 p번째 백분위수보다 큰 값을 대체

4. 잡음 (Waveform의 한 부분이지만, 입력 신호가 아닌 것, 실제 입력되지 않았지만 입력되었다고 잘못 판단된 값)

1) MA Filter(Moving Average): window가 이동하면서 주위 값들에 비해 높거나 낮을 경우 평균으로 대체

2) Median Filter: 일정 범위의 중간 값을 해당 지점의 값으로 지정(잡음이 클 경우 MAF보다 좋은 성능)

3) Curve Fitting and Splines : 최적의 파형을 기준으로 해당 파형에 유사한 신호 검출

4) Digital Filter: 고정된 시간 간격 단위의 필터(LPF, HPF, BPF)

5) Pivoting: 데이터 셋을 설정한 축 Pivot을 기준으로 카운팅하여 새로운 통계값 생성

5. 데이터 변환

1) Smoothing: 잡음 제거, 데이터 추세에 벗어나는 데이터 변환, 너무 타이트하면 overfitting, 너무 루즈하면 예측 정확도 감소 가능성

2) Aggregarion: 데이터 요약

3) Generalization: 특정 구간에 분포하는 값으로 스케일 변화

4) Normalization: min-max normalization, z-score normalization

$$Y=(X-\mu)/\sigma$$

5) feature construction: 새로운 속성이나 특징을 만드는 방법(PCA, FFT)

PCA: 데이터를 대표하는 주성분을 찾아 변수의 차원(개수)을 줄이는 목적

변수에 의한 데이터의 overlap을 감소시키는 목적

PCA 수행 전 반드시 normalization을 통해 데이터를 일반화해야 함

FFT: 시간 영역의 신호를 주파수 영역으로 변환하는 Fourier 알고리즘을 통해 눈에 보이지 않는 신호의 특징 추출

출처: https://pubdata.tistory.com/52

이상치 탐색

1. 이상치 탐색의 개념

- 통계학 측면에서 이상치는 관측치들이 주로 모여있는 곳에서 멀리 떨어져 있는 관측치로 정의됨

- 이상치는 비합리적인 이상치와 합리적인 이상치로 구분

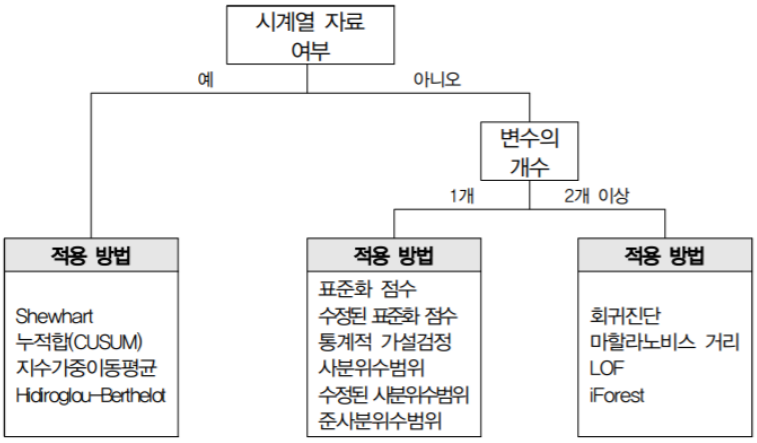
- 비합리적인 이상치: 입력 오류 등 자료의 오염으로 인해 발생한 이상치를 의미

- 합리적 이상치: 정확하게 측정되었으나 다른 자료들과 전혀 다른 경향이나 특성을 보이는 이상치

2. 이상치 탐색 방법의 분류

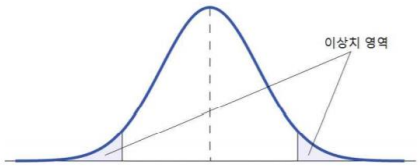
접근 방법	이상치 탐색 방법 분류
자료의 크기	소표본, 대표본
자료의 차원	일차원, 이차원, 다차원
변수의 개수	일변량, 이변량, 다변량
목표 변수의 유무	지도 방법, 비지도 방법
통계적 방법	모수적 방법, 비모수적 방법, 준모수적 방법

– 이상치 탐색 시 활용하는 변수의 개수와 시계열 자료 여부에 따라 이상치 탐색 방법을 분류함



[그림 3] 자료의 구조에 따른 이상치 탐색 방법의 분류

- **단변량 자료에서 이상치 탐색 방법은 이상치 영역을 정의하여 이상치를 탐색하는 방법임**
- **단변량 자료의 이상치 탐색 방법은 오염된 관측치를 탐색하는 방법이 아닌 정의된 이상치 영역의 포함 여부에 대한 판단 개념임**



[그림 4] 단변량 자료의 이상치 탐색 원리

- **다변량 자료에서 이상치 탐색 방법은 연관성이 존재하는 2개 이상의 변수 정보를 활용하여 관측치 사이의 거리, 밀도 등을 기반으로 이상치를 탐색하는 방법임**
- **시계열 자료에서 이상치 탐색 방법은 단변량 자료의 이상치 탐색 방법과 유사한 개념 이며, 본 연구에서는 감시(surveillance)를 위한 기법을 중심으로 검토함**

3. 이상치 탐색 방법

3-1. 단변량 자료에서 이상치 탐색

1) 표준화 점수(Z-score)를 활용한 이상치 탐색

○ 표준화 점수는 평균이 m이고, 표준편차가 o인 정규분포를 따르는 관측치들이 자료의 중심(평균)에서 얼마나 떨어져 있는지를 나타냄

- 표준화 점수 산출을 위해서는 관측치들이 정규분포를 따른다는 가정을 만족해야 함 - 정규분포를 만족하지 않는 경우, 로그변환, Box-Cox 변환2)을 적용하여 정규분포를 하도록 관측치를 변환하는 방법이 있음

○ n개의 각 관측치에 대한 표준화 점수는 다음과 같이 정의함

$$Z_i = \frac{x_i - \mu}{\sigma}, \quad i = 1,2,3,\cdots,n$$

○ 일반적으로 표준화 점수의 절대값이 3보다 큰 경우에 이상치로 정의하며, 연구마다 이상치 정의를 위한 기준은 다양하게 제시함

- 미국 국립표준기술연구소(National Institute of Standards and Technology)에서는 표준화 점수의 절대값이 3.5를 초과하는 경우 이상치로 정의함

- Aggarwal(2013)은 표준화 점수의 절대값이 3을 초과하는 경우 이상치로 정의함

- 어느 정도의 threshold 값을 매기냐에 따라 검출 가능한 이상치의 범위가 정해진다.

- 데이터 포인트의 68%는 +-1 표준 편차 사이에 있다.

- 데이터 포인트의 95%가 +-2 표준 편차 사이에 있다.

- 데이터 포인트의 99.7%가 +-3 표준 편차 사이에 있다.

2) 수정된 표준화 점수(Modified Z-score)를 활용한 이상치 탐색 (AE에서 활용)

○ 표준화 점수는 평균과 표준편차에 의존하므로, 산출 과정에 이상치의 영향을 받는 문제점이 있음

○ 수정된 표준화 점수는 표준화 점수의 문제점을 보완하기 위해 중앙값(x_bar)과 중앙값 절대편차(median absolute deviation, MAD)를 이용하여 산출함

- 중앙값은 관측치를 오름차순으로 정렬하였을 때, 중앙에 위치한 관측치를 의미하며 관측치의 수가 짝수인 경우에는 중앙에 위치한 두 값의 평균으로 산출됨

- 중앙값의 절대편차는 관측치와 중앙값 차이의 절대값에 대한 중앙값으로 아래 수식과 같이 정의됨

$$MAD = median(|x_i - \tilde{x}|),\ \tilde{x}\text{는 중앙값}$$

○ n개의 각 관측치에 대한 수정된 표준화 점수는 다음과 같이 정의됨

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD},\ i = 1,2,3,\cdots,n$$

○ 수정된 표준화 점수를 활용한 이상치 탐색 방법은 관측치의 수가 적은 경우에 적합한 방법으로 알려져 있음

○ Iglewicz와 Hoaglin(1993)은 수정된 표준화 점수의 절대값이 3.5보다 큰 경우에 이상치로 판단하는 것을 제안함

time	loss_mae	Threshold	Anomaly
2021-05-01 00:00:00	0.095340	0.27	False
2021-05-01 00:10:00	0.076679	0.27	False
2021-05-01 00:20:00	0.041020	0.27	False
2021-05-01 00:30:00	0.078962	0.27	False
2021-05-01 00:40:00	0.042152	0.27	False
...	...	...	...
2021-05-08 23:00:00	0.086101	0.27	False
2021-05-08 23:10:00	0.119305	0.27	False
2021-05-08 23:20:00	0.063611	0.27	False
2021-05-08 23:30:00	0.043346	0.27	False
2021-05-08 23:50:00	0.082896	0.27	False

925 rows × 3 columns

3) 통계적 가설검정을 활용한 이상치 탐색

○ 통계적 가설검정 방법은 최소값 혹은 최대값의 이상치 여부에 대한 검정임

○ 이상치로 판단된 관측치를 제외해 나가면서 이상치가 존재하지 않을 때까지 반복적으로 검정을 수행하여 이상치를 정의함

- **딕슨의 Q 검정(**딕슨의 Q 검정은 오름차순으로 정렬된 데이터에서 범위에 대한 관측치 간의 차이 (gap)에 대한 비율을 활용하여 이상치 여부를 검정하는 방법임)

- **그럽스 T-검정(**그럽스 T-검정은 정규분포를 만족하는 단변량 자료에서 이상치를 검정하는 방법임)

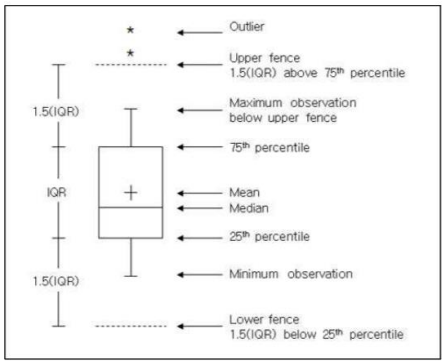
- **Generalized ESD(Extreme Studentized Deviate) Test(**그럽스 T-검정을 일반화한 방법으로 여러 개의 이상치에 대한 검정이 가능함)

- 카이제곱 검정(Chi-Square Test)(카이제곱 검정은 데이터가 정규분포를 만족하나, 자료의 수가 적은 경우에 이상치를 검정하는 방법임)

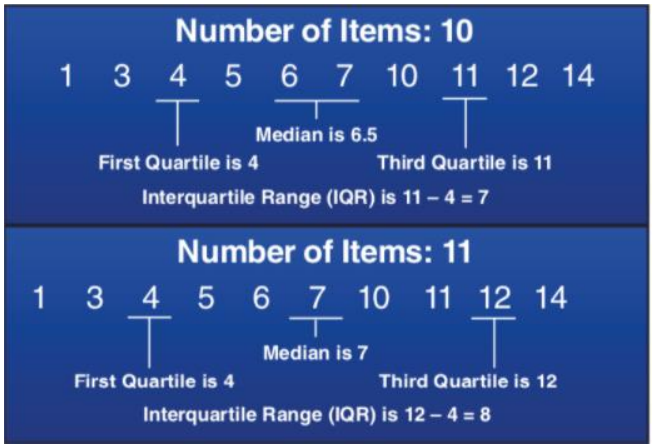
4) 사분위수범위를 활용한 이상치 탐색(Turkey Fence)

○ 상자그림(boxplot)은 최소값, 최대값, 제 1사분위수(Q1), 제 2사분위수(Q2), 제 3 사분위수(Q3)를 활용하여 데이터를 시각적으로 요약한 그래프임

- 상자그림에서 표현되는 최소값과 최대값은 이상치를 제외한 데이터의 최대값과 최소값을 의미하며, 이상치는 사분위수범위를 활용하여 정의함



[그림 6] 상자그림 그리는 방법



○ 사분위수범위는 제 1사분위수(ˆ)와 제 3사분위수(ˆ)의 차이로 정의되며, 사분위수 범위를 활용한 이상치 정의 수식은 아래와 같음

$$(Q_1 - c \times IQR,\ Q_3 + c \times IQR),\ IQR = Q_3 - Q_1,\ c\text{는 상수}$$

○ 일반적으로 상수c는 1.5나 3을 적용하며, 사분위수범위의 1.5배를 초과하는 관측치는 약한 이상치, 3배를 초과하는 관측치는 강한 이상치로 정의함

3-2. 다변량 자료에서 이상치 탐색

1) 회귀진단(Regression Diagnostics)에서 이상치 탐색

○ 회귀진단은 추정된 회귀식에 대한 전반적인 검토를 의미하며, 회귀식 추정에 영향을 미치는 극단치를 탐색하는 것을 포함함

○ 회귀진단을 통한 이상치 탐색 방법에는 레버리지, 표준화 잔차, 스튜던트 잔차, 스튜 던트 제외 잔차, 쿡의 거리, DFFITS, DFBETAS 등이 있음

2) 마할라노비스 거리(Mahalanobis Distance)를 활용한 이상치 탐색(PCA에서 활용)

○ 마할라노비스 거리는 데이터의 분포를 고려한 거리 측도로, 관측치가 평균으로부터 벗어난 정도를 측정하는 통계량임

○ 이상치 탐색을 위해 고려되는 모든 변수 간에 선형관계를 만족하고, 각 변수들이 정규 분포를 따르는 경우에 적용할 수 있는 전통적인 접근법임

○ 마할라노비스 거리를 산출하는 수식은 아래와 같으며, x_bar는 평균, V는 공분산 행렬임

$$MD_i = \sqrt{(x_i - \bar{x})^T V^{-1} (x_i - \bar{x})}, \quad i = 1,2,3,\cdots,n$$

○ 마할라노비스 거리의 이상치 기준은 k개의 변수에 대해, 자유도가 k인 카이제곱 분포의 임계값을 초과하는 경우에 이상치로 정의함

$$MD_i > \sqrt{\chi^2(k, 1-\alpha)}, \ \alpha\text{는 유의수준}$$

```
anomaly = pd.DataFrame()
anomaly['Mob dist']= dist_test
anomaly['Thresh'] = threshold
# If Mob dist above threshold: Flag as anomaly
anomaly['Anomaly'] = anomaly['Mob dist'] > anomaly['Thresh']
anomaly.index = X_test_PCA.index
anomaly.head()
```

	Mob dist	Thresh	Anomaly
2004-02-13 23:52:39	1.032676	5.082727	False
2004-02-14 00:02:39	1.148163	5.082727	False
2004-02-14 00:12:39	1.509998	5.082727	False
2004-02-14 00:22:39	1.849725	5.082727	False
2004-02-14 00:32:39	0.701075	5.082727	False

3) LOF(Local Outlier Factor)

○ LOF는 관측치 주변의 밀도와 근접한 관측치 주변의 밀도의 상대적인 비교를 통해 이상치를 탐색하는 기법임

○ 각 관측치에서 ▸¶번째 근접이웃까지의 거리를 산출하여 해당 거리 안에 포함되는 관 측치의 개수를 나눈 역수 값의 개념으로 산출되며, 다음과 같이 산출됨

- (k-distance) 특정 관측치에서 k번째 근접이웃까지의 거리를 k-distance라고 하며, A 관측치에서 k-distance 안에 포함되는 관측치의 수를 Nk(A)로 표기함

- (reachability distance) 특정 관측치(A)와 이웃한 관측치(B)의 거리와 k-distnace 중 큰 값을 의미하며, 아래와 같이 수식으로 표현할 수 있음

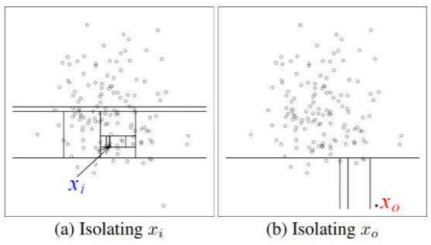
$$reachability\ distance_k = \max\{k-distance(B), d(A,B)\}$$

4) iForest(Isolation Forest)

○ iForest 기법은 관측치 사이의 거리 또는 밀도에 의존하지 않고, 데이터마이닝 기법인 의사결정나무(Decision tree)를 이용하여 이상치를 탐지하는 방법임

○ 의사결정나무 기법으로 분류모형을 생성하여 모든 관측치를 고립시켜나가면서 분할 횟수로 이상치를 탐색함

- 데이터의 평균적인 관측치와 멀리 떨어진 관측치일수록 적은 횟수의 공간 분할을 통해 고립시킬 수 있음



[그림 7] 공간 분할을 통한 이상치 탐색 원리

3-3. 시계열 자료에서 이상치 탐색

○ 시계열 자료에서 이상치 탐색은 대부분 모형 적합을 통해 관측치 사이의 연관성을 제거한 잔차를 산출한 후, 잔차에 대해 방법을 적용함

○ 감시(surveillance) 목적의 통계적 공정관리(Statistical Process Control, SPC) 기법을 시계열 자료의 이상치 탐색에 활용할 수 있음 - 대표적인 방법으로 슈하르츠(Shewhart) 관리도, 누적합(Cumulative Sum, CUSUM) 관리도, 지수가중이동평균(Exponentialy Weighted Moving Average, EWMA) 관리도가 있음

○ 감시를 위한 기법 또한 관측치 사이의 독립성을 만족해야 적용이 가능하나, 일반적 으로 독립성 가정에 대한 부분은 무시되는 경우가 많음

출처: 이상치 탐색을 위한 통계적 방법과 활용 방안, 건강보험심사평가원

이상치의 임계값(Treshold) 설정 방식 (자유 또는 선택 가능)

- Variance : 정규분포에서 97.5% 이상 또는 2.5% 이하에 포함되는 값을 이상치로 판별

- 표준편차 방식 : 하한값 = 평균 - n * 표준편차, 상한값 = 평균 + n * 표준편차 (n은 3 or 2.75)

- Reconstruction Error : recon_error_train['Anomaly'] = recon_error_train['Loss_mae'] > recon_error_train['Threshold'] # THRESHOLD = 0.3

- 또는, 전통적으로 Raw data의 Waveform에서 early sign (노랑), waring sign (주황), critical sign (빨강 데이터 포인트) 으로도 추가 표현 등 가능

- Early anomaly sign : RMS X 1.5, Waring anomaly sign : RMS X 3, Critical anomaly sign : RMS X 4.5

- Anomaly -> Fault -> Failure 3단계 중, Anomaly Detection이 핵심이며, Fault Detection은 옵션이며, failure는 목표 타스크는 아닙니다.

출처: https://www.sktaifellowship.com/d9788784-5df2-4782-bb12-c2ec7f651e39

결론

0. 모델 적용

- 차원 축소(PCA/AE)

1. Threshold 기준 설정

- Variance : 정규분포에서 97.5% 이상 또는 2.5% 이하에 포함되는 값을 이상치로 판별

- 표준편차 방식 : 하한값 = 평균 - n * 표준편차, 상한값 = 평균 + n * 표준편차 (n은 3 or 2.75)

- Reconstruction Error : recon_error_train['Anomaly'] = recon_error_train['Loss_mae'] > recon_error_train['Threshold'] # THRESHOLD = 0.3

2. 이상치 검출(다변량, 단변량 알고리즘에 따라 적합하게 적용)

- MAE

time	loss_mae	Threshold	Anomaly
2021-05-01 00:00:00	0.095340	0.27	False
2021-05-01 00:10:00	0.076679	0.27	False
2021-05-01 00:20:00	0.041020	0.27	False
2021-05-01 00:30:00	0.078962	0.27	False
2021-05-01 00:40:00	0.042152	0.27	False
...	...	...	...
2021-05-08 23:00:00	0.086101	0.27	False
2021-05-08 23:10:00	0.119305	0.27	False
2021-05-08 23:20:00	0.063611	0.27	False
2021-05-08 23:30:00	0.043346	0.27	False
2021-05-08 23:50:00	0.082896	0.27	False

925 rows × 3 columns

- Mahalanobis Ditstance

```
anomaly = pd.DataFrame()
anomaly['Mob dist']= dist_test
anomaly['Thresh'] = threshold
# If Mob dist above threshold: Flag as anomaly
anomaly['Anomaly'] = anomaly['Mob dist'] > anomaly['Thresh']
anomaly.index = X_test_PCA.index
anomaly.head()
```

	Mob dist	Thresh	Anomaly
2004-02-13 23:52:39	1.032676	5.082727	False
2004-02-14 00:02:39	1.148163	5.082727	False
2004-02-14 00:12:39	1.509998	5.082727	False
2004-02-14 00:22:39	1.849725	5.082727	False
2004-02-14 00:32:39	0.701075	5.082727	False

3. Threshold와 비교

- Threshold 이상이면 Anomaly detect

4. 이상치로 판별