JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 7: 1078~1090, Jul. 2022

모바일 기기에서 이상치 데이터 처리 정책에 따른 배터리 잔여 시간 예측 기법의 평가

탁성우*

Performance Evaluation of Battery Remaining Time Estimation Methods According to Outlier Data Processing Policies in Mobile Devices

Sungwoo Tak*

*Professor, School of Computer Science and Engineering, Pusan National University, Busan, 46241 Korea

요 약

모바일 기기 배터리의 잔여 시간 예측은 배터리 잔량별 사용 시간 데이터의 분포 특성에 영향을 받는다. 특히 이상치 데이터가 존재하는 경우, 통계적 회귀 기법의 예측 성능을 왜곡시킬 수 있다. 이에 본 논문에서는 통계적 회귀 기법의 예측 성능 향상을 위해 이상치 데이터를 탐지 및 처리하는 프레임워크를 제안하였다. 제안한 프레임워크는 먼저 배터리 잔여 시간 예측에 영향을 주는 이상치 데이터를 탐지한다. 탐지된 이상치 데이터는 평활 과정을 통해 새로운 값으로 치환된 후, 이상치 데이터와 치환된 데이터 간의 차이를 개별 데이터에 분배한다. 마지막으로 개별 데이터를 재강화하여 예측 성능을 향상시키고자 한다. 제안한 프레임워크의 성능 분석을 수행한 결과, 배터리 잔여 시간의 예측 성능이 향상됨을 확인하였다.

ABSTRACT

The distribution patterns of battery usage time data per battery level are able to affect the performance of estimating battery remaining time in mobile devices. Outliers may mainly affect the estimation performance of statistical regression methods. In this paper, we propose a software framework that detects and processes outliers to improve the estimation performance of statistical regression methods. The proposed framework first detects outliers that degrade the estimation performance. The proposed framework replaces outliers with smoothed data. The difference between an outlier and its replaced data will be properly distributed into individual data. Finally, individual data are reinforced to improve the estimation performance. The numerical results obtained by experimenting the proposed framework confirmed that it yielded good performance of estimating battery remaining time.

키워드 : 모바일 기기, 이상치 데이터, 배터리 잔여 시간, 회귀, 예측 성능

Keywords : Mobile Device, Outlier Data, Battery Remaining Time, Regression, Estimation Performance

Received 12 May 2022, **Revised** 2 June 2022, **Accepted** 8 June 2022

*** Corresponding Author** Sungwoo Tak(E-mail:swtak@pusan.ac.kr, Tel:+82-51-510-2387)
Professor, School of Computer Science and Engineering, Pusan National University, Busan, 46241 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.7.1078 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

모바일 기기 배터리의 잔여 시간 예측에 통계적 회귀 기법 사용을 고려할 수 있다. 측정 데이터의 변화량 분포가 일정한 범위 내에 있는 경우, 통계적 회귀 분석 기법의 예측 성능은 우수하다. 그러나 데이터 분포의 일정한 범위에서 벗어나 있는 이상치 데이터가 존재하는 경우, 통계적 회귀 기법의 예측 정확도를 왜곡시킬 수 있다[1]. 이상치 데이터 탐지 기법으로는 시스템 방정식 기반, 이상 신호 및 거리 기반, 그리고 통계적 분포 기반 이상치 탐지 기법이 있다. 이와 관련된 기존 연구를 살펴보면 다음과 같다.

시스템 방정식 기반 이상치 탐지에서는 측정 데이터와 시스템 방정식을 통해 추정된 데이터간의 차이가 임계치를 초과하면, 해당 측정 데이터는 이상치 데이터로 간주된다[2]. 참고 문헌 [2, 3]에서는 예측값 추정에 시스템 방정식 기반 칼만 필터를 사용하였다. 그러나 스마트폰과 같은 모바일 기기에서는 배터리 잔량별 배터리 사용 시간 패턴에 대한 규칙성이 없기에 시스템 방정식을 세우는 것은 어렵다. 이상 신호 기반 이상치 탐지에서는 측정 데이터가 기준 변화율에서 벗어나 있는 경우, 이상치 데이터로 간주한다. 참고 문헌 [4]에서는 측정 데이터가 평균 변화율에서 벗어나 있는 경우, 이상치 데이터로 간주한다. 거리 기반 이상치 탐지에서는 측정 데이터와 기존 데이터 간의 거리가 임계치를 초과하면, 해당 측정 데이터를 이상치 데이터로 간주한다[5,6]. 통계 분포 기반 이상치 탐지에서는 측정 데이터가 통계적 편차 범위에서 벗어나 있는 경우, 이상치 데이터로 간주된다. 참고 문헌 [7, 8]에서는 데이터 분포의 중간 값과 표준 편차를 사용하여 편차 범위를 설정하였다.

지금까지 살펴 본 기존 연구에서는 예측에 영향을 줄 수 있는 이상치 데이터를 측정 데이터에서 제거하였다. 그러나 이상치 데이터도 측정 데이터이기에, 본 논문에서는 이상치 데이터를 제거하는 대신 가공 처리하여 예측 성능을 향상시키는 프레임워크를 제안하였다. 본 논문의 구성은 다음과 같다. II장에서는 제안한 프레임워크에 적용 가능한 이상치 탐지 기법을 기술하였다. III장에서는 이상치 데이터의 치환 및 분배, 그리고 재강화 정책을 기술하였다. IV장에서는 제안한 프레임워크의 성능을 분석하였다. 마지막으로, V장에서는 결론을 기술하였다.

Ⅱ. 이상치 데이터 탐지 기법

제안한 프레임워크의 설계에 사용되는 공통 변수들은 다음과 같다.

- K: 초기 배터리 잔량 ($1 < K \leq 100$)
- BT_k, BT_i: 배터리 잔량 k와 i에서 측정한 배터리 사용 시간 ($1 \leq k, i < K$)
- $\widehat{BT_k}$, $\widehat{BT_i}$: 배터리 잔량 k와 i에서 통계적 회귀 기법이 예측한 배터리 사용 시간
- X: 배터리 잔량 $K(BL_K)$부터 $k(BL_k)$까지의 데이터 집합 ($X = \{BL_K, \cdots, BL_k\}$)
- Y: 배터리 잔량 K부터 k까지 측정한 배터리 사용 시간 데이터의 집합 ($Y = \{BT_K, \cdots, BT_k\}$)
- $\hat{Y}$: 집합 Y에 대하여 통계적 회귀 기법이 생성한 예측 데이터의 집합 ($\hat{Y} = \{\widehat{BT_K}, \widehat{BT_{K-1}}, ..., \widehat{BT_k}\}$)
- RE_i: 배터리 잔량 i에서 BT_i와 $\widehat{BT_i}$의 차이 값 ($RE_i = BT_i - \widehat{BT_i}$)
- SSE: 배터리 잔량 K부터 k까지 계산한 RE_i의 제곱합 ($SSE = \sum_{i=k}^{K} (RE_i)^2$)
- $OutlierBT_i$: 이상치 데이터로 판단된 BT_i
- $Outlier$: 배터리 잔량 K부터 k까지의 이상치 데이터 집합 ($Outlier = \{OutlierBT_K, \cdots, OutlierBT_k\}$)
- $ResidualBT_i$: 원 데이터 (BT_i)와 이상치 데이터 ($OutlierBT_i$) 간의 차이 ($ResidualBT_i = BT_i - OutlierBT_i$)
- $Residual$: $ResidualBT_i$ 데이터의 집합 ($Residual = \{ResidualBT_K, \cdots, ResidualBT_k\}$)

2.1. 이상 신호 및 통계 분포 기반 이상치 탐지 기법

이상 신호 기반 이상치 탐지 기법에서는 이상치 탐지에 많이 사용되는 *DFBETAS* (Difference in Betas)와 *DFFITS* (Difference in Fits)를 사용하였다[4]. 표 1은 *DFBETAS* 기반 *ODDB* (Outlier Detection with DFBETAS) 기법의 동작 과정을 보여준다.

Table. 1 Outlier detection with *ODDB*

1. procedure *ODDB*(Y, $\hat{Y}$, X)

1.1 $\hat{H} \leftarrow Y(Y^TY)^{-1}Y^T$, $where\ \hat{h}_{ij} \in \hat{H} \wedge K \leq i, j \leq k$,

1.2 $C \leftarrow (Y^TY)^{-1}$, $where\ c_{ij} \in C \wedge K \leq i, j \leq k$

1.3 $SSE \rightarrow \sum_{i=k}^{K} (RE_i)^2$, $where\ RE_i \leftarrow BT_i - \widehat{BT_i}$

1.4 $i \leftarrow K$, $N \leftarrow len(X)$

1.5 $while\ (i \geq k)$ {

1.6 $RE_i \leftarrow BT_i - \widehat{BT_i}$

1.7 $MSE_i \leftarrow \dfrac{(SSE - RE_i^2)/(1-\hat{h}_{ii})}{N-p-1}$

1.8 $Dfbetas \leftarrow \left((Y^TY)^{-1} \cdot \dfrac{RE_i}{1-\widehat{h_{ii}}}\right) / \sqrt{MSE_i \cdot c_{ii}}$

1.9 $if\ (Dfbetas \geq 2/\sqrt{N})\ SaveOutlier(BT_i)$

1.10 $i \leftarrow i-1$

1.11 }

2. function *SaveOutlier* (BT_i)

2.1 $OutlierBT_i \leftarrow BT_i$

2.2 $Outlier \leftarrow Outlier \cup \{OutlierBT_i\}$

ODDB 기법은 Y가 $\hat{Y}$에 미치는 영향값이 행렬 요소로 사용되는 영향도 행렬 ($\hat{H}$)를 사용한다 (동작 과정 1.1). $\hat{h}_{ii}$는 $\hat{H}$의 대각 원소를 나타내며, BT_i가 $\widehat{BT_i}$에 미치는 영향력은 0과 1사이 값으로 나타낸다. $\hat{h}_{ii}$가 1에 가까워지고 나머지 원소 값들이 0에 가까워지면, BT_i와 $\widehat{BT_i}$는 거의 일치한다[9]. 행렬 $(Y^TY)^{-1}$ 계산 결과를 행렬 C에 저장한다. c_{ij}는 C의 원소를 나타낸다 (동작 과정 1.2), RE_i를 사용하여 SSE를 계산한다 (동작 과정 1.3). 변수 i는 초기 배터리 잔량 K로 설정되고, N은 X 원소의 개수로 설정된다 (동작 과정 1.4). 배터리 잔량 k가 배터리 잔량 i보다 크거나 같은 경우, 이상치 탐지 절차를 수행한다 (동작 과정 1.5부터 1.11까지). BT_i와 $\widehat{BT_i}$간의 차이 (RE_i)를 계산한다 (동작 과정 1.6). BT_i를 제외한 나머지 측정 데이터의 평균 잔차 제곱합 (MSE_i)을 계산한다 (동작 과정 1.7). MSE_i에 대한 자유도는 N-p-1이다. p는 회귀 모형에서 사용하는 설명 변수의 개수를 나타낸다. 회귀 분석 기법에서 설명 변수인 배터리 잔량별에 대하여 반응 변수인 배터리 사용시간을 추정하기에, p는 1로 설정된다. 동작 과정 1.8에서는 이상치 기준 값으로 사용되는 *Dfbetas*를 계산한다. 이상치 데이터 기준은 *Dfbetas*가 $2/\sqrt{N}$보다 크거나 같은 경우이다 (동작 과정 1.9)[4]. 이상치 데이터로 탐지된 BT_i는 *SaveOutlier*() 함수를 통해 *Outlier*에 저장된다 (동작 과정 1.9와 단계 2). 표 1에서 사용한 변수 Y, $\hat{Y}$, X, SSE, RE_i, BT_i, $\hat{h}_{ii}$, p, 그리고 N은 다른 기법의 동작 과정에서도 동일하게 사용되었다.

Table. 2 Outlier detection with *ODDF*

1. procedure *ODDF*(Y, $\hat{Y}$, X)

1.1 $MSE \leftarrow SSE/(N-p)$, $i \leftarrow K$, $N \leftarrow len(X)$

1.2 $while\ (i \geq k)$ {

1.3 $Dffits \leftarrow RE_i \cdot \sqrt{\dfrac{\widehat{h_{ii}}}{1-\widehat{h_{ii}}}}\ \sqrt{\dfrac{N-p-1}{SSE \cdot 1-\widehat{h_{ii}} - RE_i^2}}$

1.4 $if\ (Dffits \geq 2/\sqrt{\dfrac{p}{N}})\ SaveOutlier(BT_i)$

1.5 $i \leftarrow i-1$

1.6 }

Table. 3 Outlier detection with *ODPI*

1. procedure *ODPI*(Y, $\hat{Y}$, X)

1.1 $i \leftarrow K$, $N \leftarrow len(X)$

1.2 $while\ (k < i)$ {

1.3 $LB,\ UB =\ IntervalOfResponse(Y, \hat{Y}, X)$

1.4 $if\ (BT_i < LB \vee BT_i > UB)\ SaveOutlier(BT_i)$

1.5 $i \leftarrow i-1$

1.6 }

2. *function* *IntervalOfResponse*(Y, $\hat{Y}$, X)

2.1 $SEE \leftarrow \sqrt{Var(\widehat{BT_i} - BT_i)}$

2.2 $LB \leftarrow \widehat{BT_i} - t_{\alpha/2} \cdot SEE$, $UB \leftarrow \widehat{BT_i} + t_{\alpha/2} \cdot SEE$

2.3 *return LB and UB*

표 2는 *DFFITS* 기반 *ODDF* (Outlier Detection with DFFITS) 기법의 동작 과정을 보여준다. MSE는 SSE를 자유도 N-p로 나눈 평균 잔차 제곱 합을 나타낸다 (동작

과정 1.1). *p*는 표 1과 동일하게 1로 설정하였다. 동작 과정 1.3에서는 이상치 기준 값으로 사용되는 *Dffits*를 계산한다. *Dffits* 값이 $2/\sqrt{p/N}$보다 크면 BT_i를 *Outlier*에 저장한다 (동작 과정 1.4) [4]. 통계 분포 기반 이상치 탐지 기법은 스튜던트 *t*-분포를 사용하여, 배터리 잔량별 배터리 사용시간에 대한 예측 구간을 추정한다. 표 3은 예측 구간 기반 *ODPI* (Outlier Detection with Prediction Intervals) 기법의 동작 과정을 보여준다. *ODPI* 기법은 *InternalOfResponse* () 함수 (단계 2)를 이용하여 예측 구간의 하한 (LB: Lower Bound)과 상한 (UB: Upper Bound)에서 벗어난 데이터를 이상치로 판별한다 (동작 과정 1.4). 단계 2에서는 BT_i와 $\widehat{BT_i}$간의 오차 분산 (SEE: Standard Error of Estimation)을 계산한다 (동작 과정 2.1). 스튜던트 *t*-분포에서 $t_{\alpha/2}$ 분포 값, SEE, 그리고 $\widehat{BT_i}$를 사용하여 예측 구간의 *LB*와 *UB*를 유도한다 (표 3의 동작 과정 2.2). IV장의 성능 분석에서 α를 0.1로 설정하여 90%의 예측 신뢰구간을 사용하였다.

2.2. 거리 기반 이상치 탐지 기법

표 4는 개별 측정 데이터와 평균 간의 거리 차이를 이용한 *ODVM* (Outlier Detection with distance between observed Values and Mean) 기법의 동작 과정을 보여준다.

Table. 4 Outlier detection with *ODVM*

1. procedure *ODVM*(Y, $\hat{Y}$, X)

1.1 $i \leftarrow K$, $N \leftarrow len(X)$, $\mu \leftarrow \frac{1}{N}\sum_{i=k}^{K} BT_i$

1.2 $\delta \leftarrow \sqrt{\frac{1}{N}\sum_{i=k}^{K}(BT_i - \mu)^2}$

1.3 $while\,(i \geq k)$

1.4 $if\,(|BT_i - \mu| \geq \omega \cdot \sigma)\, SaveOutlier(BT_i)$

1.5 $i \leftarrow i - 1$

1.6 }

먼저, Y의 평균 (μ)과 표준편차 (δ)를 계산한다 (동작 과정 1.1과 1.2). 배터리 잔량 i가 배터리 잔량 k 이상인 경우, 이상치 데이터 탐지를 수행한다 (동작 과정 1.3부터 1.6까지). BT_i와 μ에 대한 거리 차이가 δ에 가중치 (ω)를 곱한 값보다 큰 경우, 이상치 데이터로 간주하여 BT_i를 *Outlier*에 저장한다 (동작 과정 1.4). 이상치 데이터에 의해 평균과 표준편차가 부풀려져 있으면, 이상치 데이터를 탐지할 수 없는 경우가 *ODVM*에서 발생할 수 있다. IV장의 성능 분석에서 ω를 2로 설정하였다.

Table. 5 Outlier detection with *ODCD*

1. procedure *ODCD*(Y, $\hat{Y}$, X)

1.1 $i \leftarrow K$, $N \leftarrow len(X)$

1.2 $while\,(i \geq k)$

1.3 $CookDistance \leftarrow \frac{RE_i^2}{p \cdot MSE} \cdot \left[\frac{\widehat{h_{ii}}}{(1-\widehat{h_{ii}})^2}\right]$

1.4 $CumulProb \leftarrow F_{CDF}(CookDistance, p, N-p-1)$

1.5 $if\,(CumulProb \geq \theta_{threshold})\, SaveOutlier(BT_i)$

1.6 $i \leftarrow i - 1$

1.7 }

Table. 6 Outlier detection with *ODWT*

1. procedure *ODWT*(Y, $\hat{Y}$, X)

1.1 $Y_{sort} \leftarrow AscendingSort(Y)$, $where$
$Y_{sort} = \{y_1, y_2, \ldots, y_k\}, y_1 \leq y_2 \ldots \leq y_k, \forall y_i \in Y$
$i \leftarrow K$, $N \leftarrow len(X)$

1.2 $LN \leftarrow \lfloor LP \cdot N \rfloor$, $UN \leftarrow \lfloor UP \cdot N \rfloor$, $where$
$0 \leq (LP, UP) \leq 1 \wedge 0 \leq LP + UP \leq 1$

1.3 $i \leftarrow 1$

1.4 $while\,(i \leq N)$ {

1.5 $if\,(i \leq LN \vee i \geq UN)\, SaveOutlier(y_i)$

1.6 $i \leftarrow i - 1$

1.7 }

표 5는 *Cook's Distance* (동작 과정 1.3의 $CookDistance$)를 사용한 *ODCD* (Outlier Detection with Cook's Distance) 기법의 동작 과정을 보여준다. *p*는 표 1과 동일하게 1로 설정하였다. 동작 과정 1.3에서는 배터리 잔량 i (BL_i)에 대한 $CookDistance$를 계산한다. 분모의 자유도가 *N-p*-1이고 분자의 자유도가 *p*인 *F*분포에서 $CookDistance$의 누적 확률 ($CumulProb$)을 계산한다 (동작 과정 1.4). $CumulProb$가 $\theta_{threshold}$보다 큰 경우, 이상치 데이터로 간주하여 BT_i를 *Outlier*에 저장한다 (동작 과정 1.5). IV장의 성능 분석에서

$\theta_{threshold}$를 0.2로 설정하였다. 표 6은 측정 데이터 허용 길이 구간의 절사 비율을 사용한 *ODWT* (Outlier Detection with Winsorized Trim) 기법의 동작 과정을 보여준다. 이상치 판별 기준은 측정한 배터리 잔량별 사용 시간 데이터 (y_i)의 분포에서 가장 먼 거리에 있는 양쪽 가장 자리 구간의 크기로 설정된다. 먼저, *Y*를 오름차순으로 정렬하여 저장소 Y_{sort} 에 저장한다 (동작 과정 1.1). *LP*와 *NP*는 양 구간의 가장자리 비율을 나타낸다. *LP*와 *NP*는 0과 1사이의 실수 값이며, 합이 0과 1사이에 있어야 한다 (동작 과정 1.2). *LN*과 *UN*은 측정 데이터의 왼쪽 및 오른쪽 가장자리에 위치한 원소의 개수를 나타낸다. 첨자 *i*가 *LN* 및 *UN*보다 작거나 큰 경우, 이상치 데이터로 간주하여 y_i를 *Outlier*에 저장한다 (동작 과정 1.5).

Table. 7 Outlier detection with *ODOM*

1. procedure *ODOM*(Y, $\hat{Y}$, X):

1.1 $M_{edian} \leftarrow Median(Y)$

$MAD \leftarrow Median(|BT_K - M_{edian}|, |BT_{K-1} - M_{edian}|, \cdots, |BT_k - M_{edian}|)$

1.2 $i \leftarrow K$, $N \leftarrow len(X)$, $\hat{\sigma} \leftarrow \frac{MAD}{P_{value}}$

1.3 $while\,(k < i)$ {

1.4 $if\ \left(|BT_i - M_{edian}| > K_{Huber} \cdot \hat{\sigma}\right) SaveOutlier(BT_i)$

1.5 $i \leftarrow i - 1$

1.6 }

표 7은 개별 측정 데이터와 전체 데이터의 중위수 간의 거리 차이를 이용한 일단계 *M*-추정량 기반 *ODOM* (Outlier Detection with One step M-estimator) 기법의 동작 과정을 보여준다 [10]. 표 4에서 기술한 *ODVM* 기법의 단점을 제거하기 위해 평균 대신에 절대 편차 중앙값 (MAD: Median Absolute Deviation)을 사용한다 (동작 과정 1.1). Y의 중위수는 M_{edian}에 저장된다. IV장의 성능 분석에서 P_{value} 값을 0.6745로 설정하였다 (동작 과정 1.2). 정규 분포에서 약 68%의 값들이 평균에서 양쪽으로 표준 편차 (σ)가 1인 범위 내에 존재한다. 따라서 이 값은 정규분포의 평균에서 1σ 범위 내의 신뢰구간을 나타낸다. 그리고 MAD/P_{value}는 모집단의 $\hat{\sigma}$를 추정한다. K_{Huber} 상수는 휴버 (Huber) 추정량 상수이며, IV장의 성능 분석에서 1.28로 설정하였다[10]. BT_i와 M_{edian} 간의 차이가 K_{Huber}에 $\hat{\sigma}$를 곱한 값보다 큰 경우, 이상치 데이터로 간주하여 BT_i를 *Outlier*에 저장한다 (동작 과정 1.4).

Ⅲ. 이상치 데이터 치환 및 분배 기법

데이터 분포의 일정한 범위에서 벗어나 있는 이상치 데이터는 회귀 기법의 예측 성능을 왜곡시킬 수 있다. 이에 탐지된 이상치 데이터를 제거한 후, 새로운 데이터로 치환한다. 그리고 치환 후 남은 이상치 데이터를 개별 측정 데이터에 분배하여 회귀 기법의 예측 성능을 향상시키고자 한다.

3.1. 데이터 이상치 치환 기법

Table. 8 Outlier Replacement with *ORME*

1. procedure *ORME*(Y, *Outlier*, X):

1.1 $i \leftarrow K$, $N \leftarrow len(X)$

1.2 $while\,(k < i)$ {

1.3 $if\ Outlier BT_i\ exists$

1.4 $BT_i \leftarrow \frac{1}{N}\sum_{j=i}^{K} BT_j$;
$SaveResidual(BT_i, Outlier BT_i)$

1.5 $i \leftarrow i - 1$

1.6 }

2. function *SaveResidual* (BT_i, $Outlier BT_i$):

2.1 $Residual BT_i \leftarrow Outlier BT_i - BT_i$

2.2 $Residual \leftarrow Residual \cup \{Residual BT_i\}$

표 8은 이상치 데이터를 평균 추정량으로 치환하는 *ORME* (Outlier Replacement with Mean Estimation) 기법의 동작 과정을 보여준다. 먼저 이상치 데이터 BT_i (동작 과정 1.3의 $Outlier BT_i$)를 측정 데이터의 평균으로 치환한다 (동작 과정 1.4). 치환 후 남은 배터리 사용량 ($Residual BT_i$)을 이상치 잔여 데이터 집합 ($Residual$)에 저장한다 (동작 과정 2.1과 2.2).

Table. 9 Outlier Replacement with *OROM*

1. procedure *OROM*(*Y*, *Outlier*, *X*):

1.1 $i \leftarrow K, N \leftarrow len(X), M_{edian} \leftarrow Median(Y)$

1.2 $MAD \leftarrow Median(|BT_J - M_{edian}|, |BT_{J-1} - M_{edian}|, \cdots, |BT_k - M_{edian}|)$

1.3 $\hat{\sigma} \leftarrow \frac{MAD}{P_{value}}$, $NonOutlier \leftarrow Y - Outlier$

1.4 $U \leftarrow sizeof\left(\begin{Bmatrix} OutlierBT_i | OutlierBT_i \in Outlier \\ \wedge\ OutlierBT_i > M_{edian} \end{Bmatrix}\right)$

1.5 $L \leftarrow sizeof\left(\begin{Bmatrix} OutlierBT_i | OutlierBT_i \in Outlier \\ \wedge\ OutlierBT_i < M_{edian} \end{Bmatrix}\right)$

1.6 $while\,(k < i)$ {

1.7 $if\ OutlierBT_i\ exists$

$$BT_i \leftarrow \frac{K_{Huber} \bullet \hat{\sigma} \bullet |U-L| + \sum NonOutlierBT_i}{N - |U-L|}$$

$SaveResidual(BT_i,\ OutlierBT_i)$

1.8 $i \leftarrow i - 1$

1.9 }

표 9는 이상치 데이터를 일단계 *M*-추정량으로 치환하는 *OROM* (Outlier Replacement with One-step M-estimation) 기법의 동작 과정을 보여준다. *Y*에서 중위수(M_{edian})를 저장한다 (동작 과정 1.1). BT_i와 M_{edian}간 절대 편차값 중에서 중위수를 *MAD*에 저장한다. 표 7과 동일하게 P_{value}를 0.68로 설정하였다 (동작 과정 1.3). *NonOutlier*는 *Y*에서 *Outlier*의 원소를 제외한 측정 데이터 ($NonOutlierBT_i$)의 집합이다 (동작 과정 1.3). *U*와 *L*은 M_{edian}보다 크거나 작은 이상치 데이터의 개수이다 (동작 과정 1.4와 1.5), $OutlierBT_i$는 일단계 *M*-추정량으로 치환된다 (동작 과정 1.7). $\sum NonOutlierBT_i$ / (*N*-*L*-*U*)는 *NonOutlier* 집합 원소 ($NonOutlierBT_i$)들의 평균이다. $K_{huber} \bullet \hat{\sigma}$는 $| BT_i - M_{edian} | / \hat{\sigma} \le K_{huber}$ 관계식을 통해 BT_i와 중앙값 간의 절대편차 ($| BT_i - M_{edian} |$)로 대응된다. 상위 이상치 데이터 개수 (*U*)와 하위 이상치 데이터 개수 (*L*)가 균등하게 배분되어 있지 않은 경우, $K_{huber} \bullet \hat{\sigma}$에 차이 개수 |*U*-*L*|만큼 곱한 양을 (*N*-|*U*-*L*|)개의 측정 데이터 에게 균등 배분한다. 앞서 계산한 *NonOutlier* 집합 원소들의 평균에 균등 배분한 값을 더하여 새로운 BT_i로 치환한다.

Table. 10 Outlier Replacement with *ORCA*

1. *procedure* *ORCA*(*Y*, $\hat{Y}$, *X*):

1.1 $i \leftarrow K, N \leftarrow len(X)$

1.2 $while\,(k < i)$ {

1.3 $if\ OutlierBT_i\ exists$

1.4 $LB,\ UB = IntervalOfResponse(Y, \hat{Y}, X)$

1.5 $if\ (BT_{i+1} \le BT_i)\ BT_i \leftarrow Random(\widehat{BT_i}, UB)$

1.6 $else\ BT_i \leftarrow Random(LB, \widehat{BT_i})$

1.7 $SaveResidual(BT_i,\ OutlierBT_i)$

1.8 $i \leftarrow i - 1$

1.9 }

Table. 11 Outlier Replacement with *ORTM*

1. procedure *ORTM*(*Y*, $\hat{Y}$, *X*):

1.1 $Execute\ the\ same\ as\ the\ same\ steps\ as\ steps\ 1.1\ through\ of\ Table\ 6$

1.2 $\mu \leftarrow \frac{1}{N-(LN-UN+1)} \sum_{LN < i < UN}^{K} y_i$

1.3 $while\,(k < i)$ {

1.4 $if\ OutlierBT_i\ exists$

$BT_i \leftarrow \mu;\ SaveResidual(BT_i,\ OutlierBT_i)$

1.5 $i \leftarrow i - 1$

1.6 }

표 10은 평균값 신뢰 구간 기반 *ORCA* (Outlier Replacement with Confidence Interval of Average response) 기법의 동작 과정을 보여준다. *IntervalOf Response*() 함수 (표 3의 단계 2)가 생성한 평균 예측 값 신뢰 구간의 최소값 (*LB: Lower Bound*)과 최대값 (*UB: Upper Bound*)을 사용하여 $OutlierBT_i$는 새로운 BT_i로 치환된다 (동작 과정 1,4부터 동작 과정 1,6까지). BT_i가 이전 배터리 사용 시간보다 크거나 같은 경우, $\widehat{BT_i}$와 *UB* 값 범위 내 임의 값으로 치환된다. 이와 반대인 경우, $\widehat{BT_i}$와 *LB* 값 범위 내 임의 값으로 치환된다. 표 11은 절사 평균 기반 *ORTM* (Outlier Replacement with Trimmed Mean)의 동작 과정을 보여준다. 먼저 표 6의 동작 과정 1.1 및 1.2와 동일한 절차가 실행된다 해

당 동작 과정에서 *Y*를 오름차순으로 정렬한 후, 상위 데이터의 *UP* 비율만큼 절사하고 하위 데이터의 *LP* 비율만큼 절사한다. *UP* 및 *LP* 비율만큼 절사된 $OutlierBT_i$ 개수는 *LN*과 *UN*에 저장된다. 이상치 데이터가 절사된 남은 데이터의 평균 (μ)을 계산한다 (동작 과정 1.2). μ는 새로운 BT_i로 치환된다 (동작 과정 1,4).

Table. 12 Outlier Replacement with *ORWM*

1. procedure *ORWM*(Y, $\hat{Y}$, X):

Step 1.2 in Table 11 (*ORTM*) *will be replaced with the followings*:

1.1 $\widetilde{Y}$←*Winsorize*(Y, X, *LN*, *UN*); $\mu \leftarrow \frac{1}{N}\sum_{i=k}^{K}\widetilde{y_i}$, $\widetilde{y_i} \in \widetilde{Y}$

Steps 1.3 through 1.6 in Table are executed

2. function *Winsorize*(Y, X, *LN*, *UN*):

2.1 $\widetilde{Y} \leftarrow Y_{sort}$, $N \leftarrow sizeof(X)$

2.2 $\widetilde{Y}[1:LN] \leftarrow \widetilde{Y}[LN+1]$

2.3 $\widetilde{Y}[N-UN+1:N] \leftarrow \widetilde{Y}[N-UN]$

2.4 *return* $\widetilde{Y}$

표 12는 윈저화 평균 기반 *ORWM* (Outlier Replacement with Winsorized Mean)의 동작 과정을 보여준다. *ORWM* 기법은 *Winsorize*() 함수를 통해 가공 처리된 측정 데이터 집합 $\widetilde{Y}$의 윈저화 평균 (μ)을 사용한다 (동작 과정 1.1). *Winsorize*() 함수는 먼저 *Y*의 데이터를 오름차순으로 정렬하여 (동작 과정 2.1의 Y_{sort}) $\widetilde{Y}$에 저장한다. *LP* 비율에 해당되는 첨자 1부터 첨자 *LN*까지의 하위 구간 데이터 ($\widetilde{Y}[1:LN]$)는 $\hat{Y}[LN+1]$으로 치환된다 (동작 과정 2.2). 이와 유사하게 *UP* 비율에 해당되는 상위 구간 데이터 ($\widetilde{Y}[N-UN+1:N]$)는 $\widetilde{Y}[N-UN]$로 치환된다 (동작 과정 2.3). 이와 같이 $\widetilde{Y}$의 μ는 새로운 BT_i로 설정된다. 표 10의 *ORCA* 기법에서 *Y* 대신 $\widetilde{Y}$를 사용하는 경우, *ORWT* (Outlier Replacement with Winsorized T Confidence Interval of Average Response) 기법으로 명명하였다.

3.2. 이상치 잔여 데이터 분배 정책

표 8의 *SaveResidual*() 함수는 BT_i가 $OutlierBT_i$로 치환된 후 남은 배터리 사용 시간을 $Residual$에 저장한다. 데이터 크기가 *n*일 때, 평균과 분산의 붕괴점 (breakdown point) 비율은 1/*n*이다. 따라서 평균과 분산 기반 통계 정보를 이용하는 회귀 기법의 붕괴점 비율도 1/*n*이다. 이에 이상치 데이터 1개만이 있는 경우에도 회귀 기법의 예측값과 측정값의 오차가 매우 클 수 있다. 이러한 오차를 줄이기 위해 $Residual$에 저장된 값을 개별 데이터에 분배하여 회귀 기법의 예측 성능을 향상시키고자 한다. 먼저 수식 (1)부터 (2)까지는 이상치 잔여 데이터를 균등 배분하는 *REDP* 정책 (*outlier Residual data with Equal Distribution Policy*)을 유도한다.

$$T \leftarrow \sum_{i=k}^{K} ResidualBT_i \tag{1}$$

$$BT_i \leftarrow BT_i + \frac{T}{K-k+1}, \forall_i = K, K-1, \cdots, k \tag{2}$$

REDP 정책에서는 $ResidualBT_i$의 총합 (수식 (1)의 T)을 초기 BT_K부터 현재 측정 데이터 BT_k에 균등 배분한다, (수식 (2)). 이와 같이 $ResidualBT_i$를 균등 배분하여 개별 데이터 간의 편차를 줄이고, 평균과 분산의 붕괴점이 되는 이상치 데이터의 크기를 감소시키고자 한다.

DT is initialized into 0

Select index i subject to

$$Gap_i = Min\ \{GAP | Gap_K, \cdots, Gap_k\}, \quad where\ Gap_i = ResidualBT_i - \mu \tag{3}$$

$$t \leftarrow max(t, 1), \quad where\ DT + N \bullet t \le T, \forall_t = 1, 2, \cdots, T \tag{4}$$

$$BT_i \leftarrow BT_i + t,\ DT \leftarrow DT + t,\ T \leftarrow T - t,\ Gap_i \leftarrow Gap_i + t \tag{5}$$

수식 (3)부터 (5)까지는 최소 편차를 가진 데이터에게 이상치 잔여 값을 우선 분배하는 *RMDP* 정책 (outlier Residual data distributions with Minimum Deviation first Policy)의 동작 과정을 보여준다. 분배된 값은 *DT*에 저장된다. *DT*의 초기값은 0으로 설정된다. 수식 (3)에서는 $Residual$의 평균 (μ)과 $ResidualBT_i$ 간의 편차 (Gap_i)은 *GAP*에 저장된다. 그리고 편차가 가장 작은 데이터의 인덱스 *i*를 선택한다. *T*가 측정 데이터 개수 (*N*)보다 큰 경우 *N* 배수만큼 남은 이상치 잔여 데이터를 BT_i에 *t*만큼 분배할 수 있다 (수식 (4)). 이상치 잔여값

(t)는 편차 Gap_i가 가장 작은 BT_i에게 우선적으로 할당한다. 이를 통해 편차가 작은 데이터의 값을 증가시켜 편차가 큰 데이터에 근접하게 하여 개별 데이터의 편차를 균등하게 유지하고자 한다. 수식 (4)에서는 모든 데이터의 편차가 동일한 경우에 모든 데이터에 t만큼 균등하게 분배하기 위해 N 배수를 고려하였다. 수식 (5)에서는 분배된 t를 DT에 저장한다. 그리고 T는 t만큼 감소한다. Gap_i는 분배된 t만큼 증가한다. 수식 (4)에서 T가 N보다 작은 경우, N개의 모든 데이터에게 균등 분배할 수 없다. 이에 t를 1로 설정하여 편차가 작은 데이터 순서로 1씩 분배한다.

REDP 및 *RMDP*정책을 수행한 후에도 이상치 데이터가 탐지되는 경우, 전체 데이터를 정제된 값으로 재강화하는 *DRDS* 정책 (Deep Reinforcement with Data Smoothing policy)을 수행한다. *DRDS* 정책은 통계 분포에서 꼬리에 존재하는 데이터를 평균 근처로 오도록 한다. 그리고 이상치 데이터가 발생할 확률을 감소시켜 회귀 기법의 예측 성능을 향상시키고자 한다. *DRDS* 정책은 이미 알려져 있는 확률 분포를 사용하는 기존 확률 분포 모형과 측정 데이터를 사용하여 모델링한 경험적 추정 확률 분포 모형을 혼합한다. 수식 (6)부터 수식 (13)까지는 $\dot{y}_k$를 유도한다.$\dot{y}_k$는 기존 확률 분포 모형을 사용하여 배터리 잔량 k에서 사용한 배터리 사용시간을 추정한 값이다 (수식 (6)). $\dot{y}_k$ 유도에 Z_{k+1}과 A_{k+1}가 사용된다. Z_{k+1}는 배터리 잔량 k+1에서 사용한 배터리 사용 시간을 정제한 값이다. A_{k+1}은 배터리 잔량 $k+1$번까지 사용한 배터리 사용 시간의 평균 (μ_{k+1}) 및 평균의 허용 오차 (a_{k+1})를 더한 값이다 (수식 (13)).

$$\dot{y}_k = Z_{k+1} + A_{k+1} \tag{6}$$

subject to

$$SSE = \sum_{i=k+1}^{K} (BT_i - \mu_{j+1})^2, \quad where\ \mu_{k+1} = \frac{1}{K-k}\sum_{i=k+1}^{K} BT_i \tag{7}$$

$$s^2 = \sum_{i=k+1}^{K} \left(BT_i - \widehat{BT_i}\right)^2/(K-k-2) \tag{8}$$

$$Bound = t_{\alpha/2} \cdot s\sqrt{1/N + (BT_k - \mu_{k+1})^2/SSE,} \tag{9}$$

$$LB \le \mu_{k+1} \le UB, where \quad LB = \widehat{BT}_{k+1} - Bound,\ UB = \widehat{BT}_{k+1} + Bound \tag{10}$$

$$if\ \mu_{k+1} \le LB \vee \mu_{k+1} \ge UB \quad then, \mu_{k+1} = (LB + UB)/2 \tag{11}$$

$$\omega_{k+1} = random\left(f_{N\mu_{k+1},\sigma^2_{k+1}}(BT_K), \cdots, f_{N\mu_{k+1},\sigma^2_{k+1}}(BT_{k+1})\right) \tag{12}$$

$$A_{k+1} = f_{N\mu_{k+1},\sigma^2_{k+1}}(\omega_{k+1})^{-1} = \mu_{k+1} + a_{k+1} \quad where\ A_{k+1} = Min(Max(A_{k+1}, Min_{BT}), Max_{BT}) \tag{13}$$

수식 (7)부터 수식 (10)까지는 μ_{k+1}의 신뢰 구간을 유도한다. BT_i 와 μ_{k+1}간의 제곱 오차 합 (SSE: Sum of Square Errors)을 계산한다 (수식 (7)). BT_i와 $\widehat{BT_i}$간의 평균 제곱 오차 (s^2)를 유도한다 (수식 (8)). 수식 (9)에서 유도한 μ_{k+1}의 신뢰 구간 경계 (*Bound*)를 사용하여 μ_{k+1}의 신뢰 구간은 (10)과 같이 유도된다. μ_{k+1}의 신뢰 구간 경계 (*Bound*)는 수식 (9)와 같다. 수식 (9)를 사용하여 μ_{k+1}의 신뢰 구간을 유도한다 (수식 (10)). 만약 μ_{k+1}가 하위 경계 *LB*와 상위 경계 *UB* 사이를 벗어난 경우 *LB*와 *UB*의 평균으로 설정된다 (수식 (11)). 수식 (12)에서는 μ_{k+1}와 σ^2_{k+1}를 모수로 하는 정규분포로부터, 배터리 사용 시간에 대한 확률 밀도 값을 임의로 추출하여 ω_{k+1}로 설정한다. 정규 분포의 역함수를 사용하여 ω_{k+1}에 대응하는 배터리 사용 시간 (A_{k+1})을 생성한다. 수식 (13)에서 생성된 A_{k+1}은 μ_{k+1}과 a_{k+1}로 분리할 수 있다. σ^2_{k+1}가 작은 경우, 측정값의 대부분이 μ_{k+1}쪽으로 집중화된다. 따라서 A_{k+1}은 μ_{k+1}에 근접될 것이다. σ^2_{k+1}가 큰 경우, 확률 밀도 값이 여러 값으로 퍼져 있어 μ_{k+1}로부터 다소 떨어져 있는 값으로 설정될 수 있다. A_{k+1}은 이전 최소 배터리 사용 시간 (Min_{BT})과 최대 측정 배터리 사용 시간 (Max_{BT}) 범위 내로 설정된다.

$$\ddot{y}_k = BT_{k+1} + B_{k+1} \tag{14}$$

subject to

$$f_Y^{k+1}(BT_k - \mu_{k+1}) \tag{15}$$

$$\rightarrow \frac{1}{K-k}\sum_{i=k+1}^{K} \Phi_h((BT_k - \mu_{k+1}) - (BT_i - \mu_{k+1})) \tag{16}$$

$$\rightarrow \frac{1}{(K-k)h}\sum_{i=k+1}^{K}\Phi((BT_k - BT_i)/h) \quad (17)$$

$$b_{k+1} \leftarrow Random(BT_K-\mu_{k+1},\cdots, BT_{k+1}-\mu_{k+1},\ weight = f_Y^{k+1}(BT_K-\mu_{k+1}),\ \cdots,\ f_Y^{k+1}(BT_{k+1}-\mu_{k+1})) \quad (18)$$

$$B_{k+1} = \mu_{k+1} + b_{k+1} \quad (19)$$

$$where\ Min_{BT} \le B_{k+1} \le Max_{BT}$$

$$Z_{k+1} = p\dot{y}_{k+1} + (1-p)\ddot{y}_{k+1} = \ddot{y}_{k+1} + p(\dot{y}_{k+1} - \ddot{y}_{k+1})$$
$$, where\ p = a_{k+1}/(a_{k+1}+b_{k+1}) \quad (20)$$

수식 (14)부터 수식 (19)까지는 $ddoty_k$를 유도한다. $ddoty_k$는 경험적 추정 확률 분포 모형을 사용하여 배터리 잔량 *k*에서 사용한 배터리 사용시간을 추정한 값이다 (수식 (14)). $ddoty_k$ 유도에 BT_{k+1}과 B_{k+1}가 사용된다. B_{k+1}은 μ_{k+1} 및 경험적 추정 확률 분포를 통해 유도한 평균 허용 오차 b_{k+1}를 더한 값이다 (수식 (19)). $ddoty_k$ 값 추정에 커널 밀도 추정 함수 ($f_Y^{k+1}()$)를 사용한다 (수식 (15)). 커널 밀도 함수는 정규분포로 설정하였다, μ_{k+1}과 BT_i차이 기반 커널 함수 ($\Phi_h()$)를 생성한 후 (수식 (16)), 데이터 개수 (*K-k*)로 나눈 $f_Y^{k+1}()$를 생성하였다 (수식 (17)). 수식 (17)에서 *h*가 커질수록 분포 그래프의 뾰족한 형태가 완만한 모양을 가진다. 따라서 *h*에 대한 적절한 크기가 필요하다. IV장의 성능 분석에서는 1부터 100까지의 *h* 값 중에서 예측 오차율이 가장 작게 나온 3으로 설정하였다. 확률 밀도 값 ($f_Y^{j-1}(BT_i-\mu_{k+1})$)을 가중치로 사용하여 μ_{k+1}과 BT_i간의 차이 (b_{k+1})를 임의로 선택한다 (수식 (18)). b_{k+1}에 μ_{k+1}를 더해 B_{k+1}를 생성한다 (수식 (19)). B_{k+1}과 BT_{k+1}을 더하여 $\ddot{y}_k$를 유도한다 (수식 (14)). 수식 (13)와 (19)에서 유도한 a_{k+1}와 b_{k+1}를 $\dot{y}_k$와 $ddoty_k$에 대한 가중치로 곱해져 정제된 데이터 Z_{k+1}를 유도한다 (수식 (20)).

Ⅳ. 성능 분석

II절의 이상치 탐지 기법과 III절의 이상치 데이터 치환 및 분배, 그리고 재강화 정책을 기존 회귀 기법에 적용하여 성능 분석을 수행하였다. 성능 평가에 적용한 일반적인 회귀 기법은 자기 회귀 (AR: Auto Regression), 선형 회귀 (LR: Linear Regression), 그리고 단순 평균 회귀 (SAR: Simple Average Regression) 기법이다[11]. 그리고 이상치 데이터의 영향력을 고려한 휴버 회귀 (HR: Huber Regression), 분위수 회귀 (QR: Quantile Regression), 랜색 회귀 (RANR: Random Sample Consensus Regression), 리지 회귀 (RR: Ridge Regression), 시겔 회귀 (SRMR: Siegel Repeated Median), 테일 센 회귀 (TSR: TheilSen Regression), 그리고 윈저화 상관 회귀 (WCR: Winsorized Correlation) 기법이다[12]. 참고 문헌 [13]에서 분석한 안드로이드 기반 스마트 폰 파워 프로파일 (Power Profile) 정보를 기반으로 하여 모바일 기기의 배터리 잔량별 사용 시간 정보를 생성하는 시뮬레이터를 구현하였다. 배터리 잔량별 사용 시간의 편차별 개별 배터리 사용량 정보를 각각 20개씩 생성한 후, 실험한 결과의 평균값을 사용하였다.

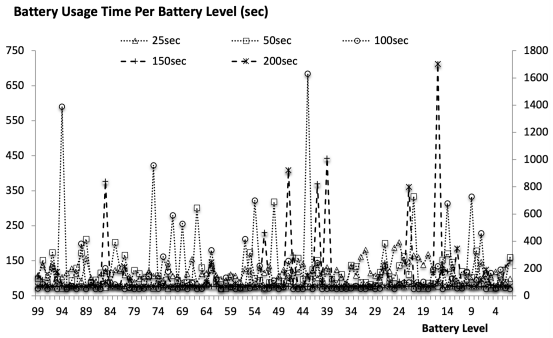

Fig. 1 Battery usage time per battery level

그림 1은 모바일 기기의 이용률 80%에서 배터리 잔량별 사용 시간의 편차가 25, 50, 100, 150, 그리고 200초인 경우, 배터리 초기 잔량 100부터 1까지 배터리 사용시간을 보여준다. 비교적 높은 편차인 100초 이상부터 이상치 데이터가 많이 분포되어 있음을 보여준다.

Table. 13 Number of outliers detected by outlier detection methods

Outlier Detection Method	Utilization				
	10%	20%	40%	60%	80%
ODDB	7	7	12	9	3
ODDF	7	7	12	9	3

Outlier Detection Method	Utilization				
	10%	20%	40%	60%	80%
ODPI	0	6	1	9	2
ODVM	6	4	6	4	3
ODCD	7	9	12	9	3
ODWT	40	40	40	40	40
ODOM	22	16	27	43	45

Table. 14 Mean absolute error ratio by regression methods

Method	Utilization (Number of Outliers)					
	10% (18)	20% (10)	40% (27)	60% (43)	80% (45)	Mean
AR	9.3	25.9	16.6	20.5	15.6	17.58
HR	14.2	12.5	17.4	10.4	26.8	16.26
LR	4.7	6.8	8.9	10.4	23.0	10.76
QR	4.7	8.0	8.8	12.0	23.9	11.48
RANR	4.7	6.8	9.0	10.4	23.6	10.9
RR	6.0	7.7	7.8	24.0	24.0	13.9
SAR	4.0	5.9	7.8	20.2	20.3	11.64
SRMR	5.0	7.5	10.7	27.9	27.9	15.8
TSR	4.7	7.0	10.2	26.0	26.1	14.8
WCR	14.7	6.9	18.3	32.3	32.3	20.9

표 13은 배터리 잔량별 사용 시간의 편차가 200초에서 탐지된 이상치 데이터의 개수를 보여준다. *ODWT*는 배터리 잔량 100부터 1까지의 구간에서 양쪽 가장 자리에 20% 비율로 절사하기에 탐지된 이상치 데이터의 개수는 40개가 된다. 평균 잔차 제곱합을 사용하는 ODDB, *ODDF* 그리고 *ODCD*는 서로 유사한 결과를 생성하였다. *ODPI*는 스튜던트 *t*-분포의 90% 신뢰 구간을 사용하였다. *ODVM*에서는 개별 데이터와 평균에 대한 거리 차이가 표준편차의 2배보다 큰 경우, 이상치 데이터로 간주하였다, 비교적 넓은 구간이기에 이상치 데이터의 탐지 개수가 작게 나왔다.ODOM이 비교적 많은 이상치 데이터를 탐지하였다. *ODOM*인 경우, 배터리 잔량별 사용 시간의 편차가 커질수록 측정 데이터와 데이터 중위수 간의 차이도 증가하여 탐지된 이상치 데이터의 개수도 증가하였다. 표 14는 회귀 기법에서 예측한 배터리 잔여 시간의 평균 절대 오차율을 보여준다. *LR*, *QR*, *RANR*, 그리고 *SAR* 등이 전반적으로 낮은 예측 오차율을 보여 주었다.

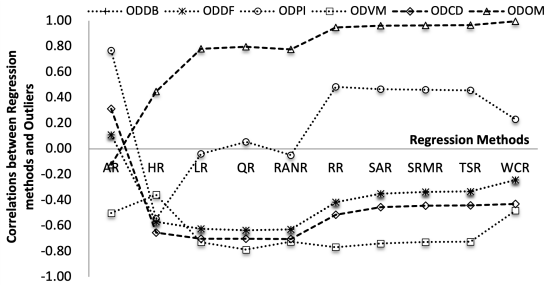

Fig. 2 Correlations between regression methods and outliers' numbers

그림 2는 표 13의 이상치 데이터 개수와 평균 절대 오차율간의 상관계수를 보여준다. 상관계수가 0에 가까울수록 이상치 데이터 개수와 평균 절대 오차율간의 상관관계가 약하며, 상관계수의 절대 값이 1에 가까울수록 두 항목간의 상관관계가 강함을 나타낸다. 전반적으로 배터리 이상치 개수의 변화량과 평균 절대 오차율 간의 상관관계가 있음을 보여준다. 특히 *ODOM*인 경우, 매우 높은 상관관계를 보여주었다. 편차에 따른 이상치 데이터의 개수가 항상 일정한 *ODWT*인 경우, 상관관계를 계산할 수 없기에 그림 2에 포함되어 있지 않았다.

표 15와 표 16은 3.1절에서 기술한 이상치 치환 기법들과 3.2절에서 기술한 *REDP* 및 *RMDP* 기반 이상치 잔여 데이터 분배 정책을 적용하여 나온 평균 절대 오차율을 보여준다. 사용한 회귀 기법은 이상치 데이터의 영향력을 고려하지 않는 회귀 기법이면서, 표 14에서 평균적으로 우수한 결과를 보여준 *LR* 기법이다. 실험 환경은 표 14에서 평균 절대 오차율이 가장 높게 나온 모바일 기기의 이용률 80%이다. 배터리 잔량별 사용 시간 편차는 25, 50, 100, 150, 그리고 200초이다. 표 15에서 보는 바와 같이, *ODDB*, *ODWT*, 그리고 *ODOM*를 통해 탐지된 이상치 데이터를 여러 이상치 치환 기법 및 *REDP*에 적용하여 나온 절대 오차율의 평균은 4.19%, 2.86%, 2.31%이다. 특히 *ODOM*인 경우, 표 14에서 기술한 *LR* 회귀 기법의 평균 10.76% 대비 2.1배 성능 향상을 보여주었다. *ODOM* 기법에 *OROM* 및 *ORWT* 기법을 *LR* 회귀 기법에 각각 적용한 경우, 표 14에서 기술한 *LR* 회귀 기법의 평균 10.76% 대비 약 10.95배 및 10.2배의 성능 향상을 보여 주었다. 표 13에서 보는 바와 같이, *ODOM*과 *ODWT*가 비교적 많은 이상치 데이터를 탐지한다. 이에 다양한 이상치 치환 기법을 통해 많은 이상치 데이터

를 정제할 수 있어 *ODOM*과 *ODWT*가 *ODDB*보다 낮은 예측 오차율을 보여주었다.

Table. 15 Mean absolute error ratio generated by REDP

Deviation of Battery Usage Time (ODDB)						
	25s	50s	100s	150s	200s	Mean
ORME	1.37	2.90	2.26	7.75	4.69	3.79
OROM	5.78	2.36	2.49	10.99	6.85	5.69
ORTM	1.35	2.90	2.38	8.88	5.20	4.14
ORWM	1.37	2.90	2.26	7.75	4.69	3.79
ORCA	1.51	2.86	2.76	7.56	3.49	3.64
ORWT	1.34	2.79	2.29	8.82	5.16	4.08
Mean	2.12	2.79	2.41	8.63	5.01	4.19
ODWT						
ORME	1.03	2.23	7.08	3.57	0.51	2.89
OROM	6.16	2.76	4.50	5.35	3.76	4.51
ORTM	0.94	2.73	5.17	0.15	2.61	2.32
ORWM	1.03	2.23	7.08	3.57	0.51	2.89
ORCA	1.69	1.70	5.47	3.46	4.28	3.32
ORWT	0.58	1.63	2.09	0.09	1.89	1.26
Mean	1.91	2.21	5.23	2.70	2.26	2.86
ODOM						
ORME	1.29	1.87	4.61	5.29	2.51	3.11
OROM	1.19	1.31	0.75	0.02	1.25	0.90
ORTM	1.19	1.66	1.62	0.09	2.60	1.43
ORWM	1.29	1.87	4.61	5.29	2.51	3.11
ORCA	1.59	1.56	3.57	2.81	12.12	4.33
ORWT	1.19	1.31	1.00	0.13	1.17	0.96
Mean	1.29	1.60	2.70	2.27	3.69	2.31

표 16은 *RMDP* 정책을 적용한 결과를 보여준다. *ODDB, ODWT*, 그리고 *ODOM*를 통해 탐지된 이상치 데이터를 *RMDP* 정책에 적용한 결과, *REDP* 정책 대비 28%, 4%, 0.3% 성능 향상을 보여 주었다. 배터리 잔량별 사용 시간의 편차가 200초인 경우, *ODOM* 기반 이상치 데이터 탐지 기법과 *OROM* 기반 이상치 치환 기법이 적용된 경우, 예측 오차율이 가장 작게 나왔다.

Table. 16 Mean absolute error ratio generated by RMDP

Deviation of Battery Usage Time (ODDB)						
	25s	50s	100s	150s	200s	Mean
ORME	1.75	3.13	1.42	7.59	2.47	3.27
OROM	3.17	2.83	2.41	6.76	2.26	3.49
ORTM	1.69	3.13	1.36	7.07	2.19	3.09
ORWM	1.75	3.13	1.42	7.59	2.47	3.27
ORCA	1.93	3.09	1.53	7.32	3.28	3.43
ORWT	1.67	3.04	1.34	7.08	2.21	3.07
Mean	1.99	3.06	1.58	7.24	2.48	3.27
ODWT						
ORME	1.06	2.18	6.06	3.89	2.23	3.08
OROM	5.61	2.97	6.75	4.85	1.51	4.34
ORTM	0.97	2.64	3.68	0.27	0.81	1.67
ORWM	1.06	2.18	6.06	3.89	2.23	3.08
ORCA	1.72	1.64	4.20	3.77	6.27	3.52
ORWT	0.60	1.54	0.93	0.41	0.59	0.81
Mean	1.84	2.19	4.61	2.85	2.27	2.75
ODOM						
ORME	1.47	1.93	4.12	5.56	3.56	3.33
OROM	1.30	1.35	0.44	0.40	0.40	0.78
ORTM	1.33	1.70	0.77	0.43	1.31	1.11
ORWM	1.47	1.93	4.12	5.56	3.56	3.33
ORCA	1.84	1.60	2.94	3.02	12.98	4.47
ORWT	1.29	1.35	0.45	0.48	0.44	0.80
Mean	1.45	1.64	2.14	2.57	3.71	2.30

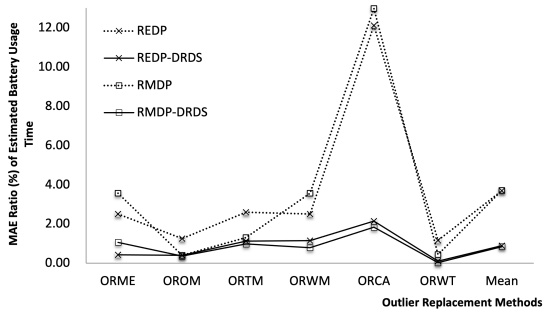

Fig. 3 Mean absolute error ratio generated by DRDS

그림 3은 예측 오차가 높게 나온 편차 200초, *ODOM*, 그리고 *ORCA*가 적용된 실험 환경에서, *DRDS* 정책을 통해 향상된 평균 절대 오차율의 결과를 보여준다. *REDP-DRDS* 및 *RMDP-DRDS*는 *REDP* 및 *RMDP*의 결과를 *DRDS* 정책으로 각각 재강화한 것이다. *REDP-DRDS* 정책 결과에 대한 평균은 0.89이며, *RMDP-DRDS* 정책 결과에 대한 평균은 0.84이다. 평균 대비 각각 3.14 배 및 3.4배의 성능 향상을 보여 주었다.

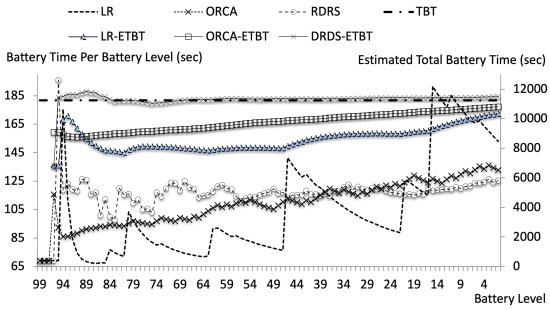

Fig. 4 Estimated battery time per battery level and total battery time by DRDS under ODOM, ORCA, and RMDP

그림 3에서 가장 높은 예측 오차율을 보여준 *ORCA-RMDP*를 재강화한 *DRDS* 정책의 결과는 그림 4와 같다. 전체 배터리 사용 시간 (TBT: Total Battery Time)은 11240초이다. 적용한 회귀 기법은 *LR* 기법이다. 그림 4에서 가로축은 배터리 잔량, 왼쪽 세로축은 배터리 잔량별 예측 사용 시간, 그리고 오른쪽 세로축은 배터리 잔량마다 예측한 전체 사용 시간을 나타낸다. *LR-ETBT*, *ORCA-ETBT*, 그리고 *DRDS- ETBT*는 각각 *LR*, *ORCA*, 그리고 *DRDS*가 적용되어 배터리 잔량별 예측 사용 시간 및 전체 사용 시간을 나타낸다. 그림 4에서 사용한 원 데이터의 변화량을 보여준 그림 1에서 높은 첨도를 가지는 배터리 잔량은 95, 81, 63, 47, 43, 22, 16, 12, 그리고 7이다. 그리고 해당 배터리 잔량에서 사용 시간은 175, 244, 922, 146, 798, 1701, 346, 118, 그리고 192초이다. 이들 데이터는 이상치 데이터로 탐지될 경향이 높다. 그림 4에서 보는 바와 같이 *LR* 기법만을 사용한 경우, 높은 첨도를 가지는 배터리 잔량 근처에서 예측한 배터리 잔량별 사용 시간 그래프에서도 높은 첨도를 보여준다. 이에 *LR* 기법은 높은 첨도를 가지는 이상치 데이터의 영향을 그대로 받기에, 비교적 높은 23%의 예측 오차율을 보여주었다. 표 16에서 *OROM-ORCA*의 예측 오차율은 12.98%이었다. *DRDS-ETBT*를 통해 *OROM-ORCA*에 의해 생성된 데이터를 재강화한 경우, 1.83%의 예측 오차율을 보여 주었다. 이는 *DRDS* 재강화 정책을 통해 정제된 데이터에 대해서는 *LR*과 같은 기존 회귀 기법만으로도 우수한 예측이 가능함을 보여준다.

Ⅴ. 결 론

제안한 이상치 데이터 처리 소프트웨어 프레임워크를 요약하면 다음과 같다. 첫째, 배터리 잔량별 사용 시간 데이터 중에서 이상치 데이터를 탐지하였다. 둘째, 이상치 데이터 개수가 증가함에 따라 통계적 회귀 기법의 예측 정확도가 낮아지는 현상이 발생하였다. 셋째, 탐지된 이상치 데이터의 치환 및 분배, 그리고 재강화 정책을 통해 가공 처리된 데이터를 사용하는 경우, 단순 회귀 기법을 사용하더라도 예측 오차율이 향상될 수 있음을 확인하였다. 이는 이상치 데이터가 포함된 측정 데이터를 정제된 데이터로 가공 처리되어진다면, 기존 회귀 기법을 적용하여도 우수한 배터리 잔여 시간 예측이 가능함을 보여준다.

ACKNOWLEDGEMENT

This work was supported by a 2-Year Research Grant of Pusan National University.

References

[1] R. Wilcox, *Introduction to robust estimation and hypothesis testing*, 5th ed. San Diego, CA: Academic Press, 2021.

[2] H. Wang, H. Li, J. Fang, and H. Wang, “Robust Gaussian Kalman Filter with Outlier Detection,” *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1236-1240, Jun. 2018.

[3] H. M. Khalid, Q. Ahmed, and J. C. -H. Peng, “Health monitoring of Li-Ion battery systems: A median expectation diagnosis approach (MEDA),” *IEEE Transactions on Transportation Electrification*, vol. 1, no. 1, pp. 94-105, Jun. 2015.

[4] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 6th ed. Hoboken, NJ: Wiley & Sons, 2021.

[5] T. Kim, A. Adhikaree, R. Pandey, D. Kang, M. Kim, C. Oh, and J. Back, “Outlier mining-based fault diagnosis for multicell lithium-Ion batteries using a low-priced microcontroller,” in *Proceeding of IEEE Applied Power Electronics Conference and Exposition*, San Antonio: TX, USA, pp. 3365-3369, 2011.

[6] J. Jiang, X. Cong, S. Li, C. Zhang, W. Zhang, and Y. Jiang,

"A Hybrid Signal-based Fault Diagnosis Method for Lithium-ion Batteries in Electric Vehicles," *IEEE Access*, vol. 9, pp. 19175-19186, Jan. 2021.

[7] Q. Quyang, J. hen, and J. Zheng, "State-of-Charge Observer Design for Batteries with Online Model Parameter Identification: A Robust Approach," *IEEE Transactions on Power Electronics*, vol. 35, no. 6, pp. 5820-5831, Jun. 2020.

[8] B. Kovacevie, M. M. Milosavljevie, and M. D. Veinovic, "Robust recursive AR speech analysis," *Signal Processing*, vol. 44, no. 2, pp. 125-138, Jun. 1995.

[9] D. C. Hoaglin and R. E. Welsch, "The hat matrix in regression and ANOVA," *The American Statistician*, vol. 32, no. 1, pp. 17-22, Jan. 1977.

[10] R. R. Wilcox, *Fundamentals of Modern Statistical Methods*, 2nd ed. Los Angeles, CA: Springer, 2010.

[11] W. Mendenhall and T. T. Sinich, *A second course in statistics: regression analysis*, 8th ed. Hoboken, NJ: Pearson, 2019.

[12] M. Kuhn and K. Johnson, *Applied predictive modeling*, 2nd ed. New York, NY: Springer, 2018.

[13] A. Saksonov, "Method to derive energy profiles for android platform," M. S. thesis, University of Oldenburg, Oldenburg, Lower Saxony, 2014.

탁성우(Sungwoo Tak)

2003년 2월 미국미주리주립대학교 Computer Science 박사
2004년~현재 부산대학교 정보컴퓨터공학부 교수
※관심분야 : 유무선 네트워크, 임베디드 시스템, 위치인식, 최적화, 예측 기법, 빅데이터, 딥러닝