# Combinatorial Data Augmentation 을 활용한 Soft Information 기반 측위 알고리즘

김상혁, 고승우
인하대학교

inhaase20@inha.edu, swko@inha.ac.kr

# Soft Information-Based Localization via Combinatorial Data Augmentation

Sang-Hyeok Kim, Seung-Woo Ko
Inha Univ.

요 약

6G 의 정밀한 위치 측위 요구조건을 만족하기 위해 likelihood 함수를 측정 데이터 기반으로 추정해내는 soft information (SI) 기반 측위 알고리즘이 많은 관심을 받고 있다. 본 논문은 실시간 측위 환경 정보 반영한 likelihood 추정을 위해 combinational data augmentation (CDA)-기반 SI 위치 측위 알고리즘을 제안한다. 제안하는 기법은 기존 거리 기반 데이터를 활용한 SI 측위 기법에 비해 적은 양의 학습 데이터로도 정밀 측위가 가능함을 시뮬레이션을 통해 검증하였다.

## Ⅰ. 서 론

6G 시스템은 센티미터 레벨의 위치 측위 성능을 요구함에 따라 기존의 추정 파라미터 사이의 지리적 특징을 활용한 signal value estimation (SVE) 기법을 뛰어넘는 새로운 방법론에 대해 많은 관심을 받고 있다. 그 중 하나의 방법론은 soft information (SI) 기반 측위로, 학습 데이터를 기반으로 likelihood 함수를 추정하여 maximum likelihood (ML) 기법에 적용하는 방법이다 [1]. 효율적인 likelihood 추정을 위해서는 주변의 환경 정보를 반영하고 있는 데이터 확보가 필요하다. 이를 위해서 본 논문에서는 데이터 사이의 조합을 통해 여러 개의 사전 위치 정보를 추출하는 combinatorial data augmentation (CDA)를 통해 likelihood 함수를 추정하는 기술을 소개한다. 제안하는 CDA-기반 SI 측위 알고리즘은 기존의 거리 기반 SI 기법 대비 base station (BS) 가 6 개이고 각 셀 당 data point 의 수가 500 개인 상황에서 RMSE 성능이 46 % 향상됨을 시뮬레이션을 통해 보인다.

## Ⅱ. 시스템 모델

본 논문에는 $N$개의 BS 가 특정 위치에 배치되어 있고, UE 위치는 랜덤하게 주어진 환경을 고려한다. 측위에 사용되는 positional feature 는 $i$번째 BS 와 UE 사이의 거리 $d_i$ 만을 이용한다. UE 의 위치 $\boldsymbol{p} \in \mathbb{R}^{1\times 2}$ 에 따라 결정되는 $d_i$ 들의 set 을 $\{d_i(\boldsymbol{p})\}_{i\in\mathcal{N}}$ 와 같이 정의한다. 이때 $i \in \mathcal{N} = \{1, 2, \cdots, N\}$는 BS 의 인덱스이다. $\{\hat{d}_i\}_{i\in\mathcal{N}}$은 BS 와 UE 사이의 측정된 거리들의 set 을 의미한다. $\hat{d}_i$ 다음과 같이 나타낼 수 있다.

$$\hat{d}_i = d_i(\boldsymbol{p}) + n_i + \mathbb{I}_{NLoS}(d_i) \times b_i \quad \cdots(1)$$

$n_i$는 노이즈이고, $\mathbb{I}_{NLoS}$는 지시함수 그리고 $b_i$는 bias 이다.

3GPP 기반 측위 기법들은 BS 와 UE 사이의 거리 $\{\hat{d}_i\}_{i\in\mathcal{N}}$ 와 같은 측정치 기반의 SVE 에 의존한다. 일반적으로 SVE 기반 측위 기법은 삼각측량 기법 즉, 최소제곱법에 의해 UE 의 위치 추정치 $\hat{\boldsymbol{p}}$이 얻어진다. 수식은 다음과 같다.

$$\hat{\boldsymbol{p}} = \underset{\tilde{\boldsymbol{p}}}{argmin}(\hat{\boldsymbol{d}} - \boldsymbol{d}(\tilde{\boldsymbol{p}}))^T \left(\hat{\boldsymbol{d}} - \boldsymbol{d}(\tilde{\boldsymbol{p}})\right)$$

이때 $\hat{\boldsymbol{d}} \in \mathbb{R}^{N\times 1}$ 과 $\boldsymbol{d}(\boldsymbol{p}) \in \mathbb{R}^{N\times 1}$ 은 각각 $\{\hat{d}_i\}_{i\in\mathcal{N}}$ 과 $\{d_i(\boldsymbol{p})\}_{i\in\mathcal{N}}$을 벡터화 한 변수이다.

SVE 기반의 측위 기법은 거리와 관한 정보만 사용하여 NLoS 에 의해 bias 가 많은 상황에서 성능이 현저하게 떨어진다. 이 한계를 극복하기 위해 SI 기반 측위 기법은 반드시 연구되어야 한다.

## Ⅲ. CDA 기법을 적용한 SI 기반 측위 기법

### A. 거리기반 SI 측위

SI 기반 측위 기법은 두단계로 나뉜다. 먼저 오프라인 단계에서 $\hat{\boldsymbol{d}}$에 해당하는 dataset 여러 개를 생성한 후 그에 맞는 likelihood 함수를 찾는다. 그리고 온라인 단계에서 $\hat{\boldsymbol{d}}$ 을 측정한 후 그에 따른 likelihood 값을 계산하여 위치를 추정한다. 그러므로 SI 기반의 측위 기법은 NLoS 에 의한 bias 도 고려하여 측위를 할 수 있다.

먼저 $\hat{\boldsymbol{d}}$을 그대로 dataset 으로 하여 측위를 한 경우를 살펴본다. 이때 Likelihood 함수는 다음과 같이 나타낼 수 있다.

$$\mathcal{L}_{\hat{d}}(d(\boldsymbol{p})) \propto f(\hat{d}; d(\boldsymbol{p}))$$

이때 $f$는 Gaussian Mixture Model (GMM)로 표현한 어떤 유저의 위치 $\boldsymbol{p}$에 대한 $\hat{d}$ 의 확률분포함수이다. 즉, Likelihood 는 $f$의 조건부 확률과 비례한다. SI 기반의 측위 기법은 측정치들인 SVE 이 들어왔을 때 SVE 에 대한 $f$의 함수 값, 즉 likelihood 가 최대가 되게 하는 $\boldsymbol{p}$를 추정위치 $\hat{\boldsymbol{p}}$ 로 선택하는 과정으로 진행된다. 이를 Maximum Likelihood Estimation (MLE)라고 한다. 수식으로 나타내면 다음과 같다.

$$\hat{\boldsymbol{p}} = \underset{\tilde{\boldsymbol{p}}}{\text{argmax}}\, f(\{\hat{d}_i\}_{i\in\mathcal{N}_{BS}}; \tilde{\boldsymbol{p}}) = \underset{\tilde{\boldsymbol{p}}}{\text{argmax}} \prod_{i\in\mathcal{N}_{BS}} \mathcal{L}_{\hat{d}_i}(d_i(\tilde{\boldsymbol{p}}))$$

위에서 설명한 GMM 기반의 함수 $f$는 $\hat{\boldsymbol{d}} \in \mathbb{R}^{N\times 1}$와 같은 형태의 데이터셋을 이용하여 학습된다. 이 데이터셋의 장

점은 $N_{BS}$차원이므로 차원이 낮아서 트레이닝 오버헤드가 낮고 그에 따라 오버피팅이 잘 생기지 않는다는 점이다. 하지만 그만큼 $N_{BS}$ 의 수가 적은 환경에서는 데이터의 차원이 낮아서 NLoS 의 편향이 반영되기 힘들 수 있다는 단점이 있다. 이러한 단점들을 어느정도 극복하고자 본 논문에서는 데이터셋 $\hat{\boldsymbol{d}} \in \mathbb{R}^{N_{BS}\times 1}$ 에 CDA 기법을 적용한다.

### B. CDA 기법을 활용한 SI 측위 기법

CDA 기법을 적용하기 위해 M 개의 $\hat{d}$를 인풋으로 받고 삼각 측량을 통해 2D 위치를 아웃풋으로 하는 함수 $g_M$: $\mathbb{R}^M \to \mathbb{R}^2$ 를 정의한다. 구체적으로 subset 크기가 $M$인 BS 의 $l$번째 부분집합 $\mathcal{M}_l$이 주어지면 $g_M$의 output $\boldsymbol{z}$는 다음과 같다.

$$\boldsymbol{z_l} = g_M(\{\hat{d}_m\}_{m\in\mathcal{M}_l}) = [x_l, y_l] \in \mathbb{R}^{1\times 2}$$

이때 부분집합 $\mathcal{M}_l$은 $l$번째 부분집합을 의미한다. 모든 $\mathcal{M}$의 set 을 $\{\mathcal{M}_l\}_{l\in\ell}$와 같은 형태로 나타낼 수 있고, 이때 $\ell = \{1,2,\cdots,L\}$이고 $L = \binom{N_{BS}}{M}$ 이다. 각각의 BS 의 조합은 $g_M$을 통해 위치 $\boldsymbol{z_l}$와 1 대 1 로 매핑 되며, 모든 BS 의 부분집합 $\mathcal{M}$에 대해 $\hat{\boldsymbol{z}}$을 나타내면 다음과 같다.

$$\hat{\boldsymbol{Z}} = \{\boldsymbol{z}_l\}_{l\in\ell} = g_M(\{\hat{d}_m\}_{m\in\mathcal{M}_l})_{l\in\ell}$$

즉 CDA 기법을 적용하여 $\hat{\boldsymbol{d}} \in \mathbb{R}^{N\times 1} \to g_M \to \hat{\boldsymbol{Z}} \in \mathbb{R}^{L\times 2}$ 의 과정으로 data 를 변환하였다. 이때 본 논문에서는 $\hat{\boldsymbol{Z}}$를 x 좌표에 관한 dataset $\hat{\boldsymbol{Z}}_{\boldsymbol{x}} \in \mathbb{R}^{L\times 1}$과 y 좌표에 해당하는 데이터 $\hat{\boldsymbol{Z}}_{\boldsymbol{y}} \in \mathbb{R}^{L\times 1}$로 분할하여 각각의 GMM $h_x$와 $h_y$를 생성했다. CDA 기법을 적용한 dataset 의 likelihood 즉, CDA 기반 SI 는 아래의 식과 같이 분할한 두 데이터에 의한 Likelihood 함수의 곱 형태로 나타난다.

$$\mathcal{L}_{\hat{\boldsymbol{Z}}} = \mathcal{L}_{\hat{\boldsymbol{Z}}_x}(\mathcal{Z}_x(\boldsymbol{p}))\mathcal{L}_{\hat{\boldsymbol{Z}}_y}(\mathcal{Z}_y(\boldsymbol{p}) \propto h_x(\hat{\mathcal{Z}}_x; \mathcal{Z}_x(\boldsymbol{p}))h_y(\hat{\mathcal{Z}}_y; \mathcal{Z}_y(\boldsymbol{p}))$$

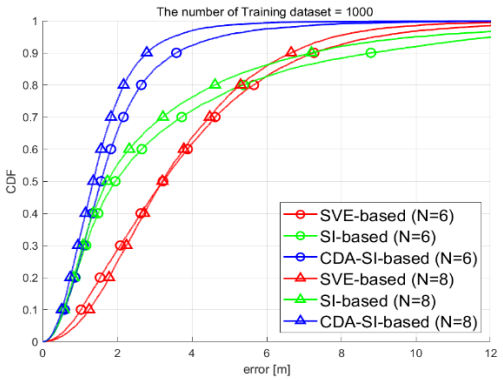
이를 통해 MLE 알고리즘을 사용하여 $\hat{\boldsymbol{p}}$을 구하는 수식은 다음과 같다.

$$\hat{\boldsymbol{p}} = \underset{\tilde{\boldsymbol{p}}}{\text{argmax}}\ \mathcal{L}_{\hat{\boldsymbol{Z}}_x}(\mathcal{Z}_x(\boldsymbol{p})) \times \mathcal{L}_{\hat{\boldsymbol{Z}}_y}(\mathcal{Z}_y(\boldsymbol{p})$$
$$= \underset{\tilde{\boldsymbol{p}}}{\text{argmax}}\ h_x(\hat{\mathcal{Z}}_x; \mathcal{Z}_x(\boldsymbol{p})) \times h_y(\hat{\mathcal{Z}}_y; \mathcal{Z}_y(\boldsymbol{p}))$$

CDA 를 활용한 SI 기반의 측위 기법은 기존 $N$차원이었던 dataset $\hat{\boldsymbol{d}}$을 $L$차원의 크기로 확장함으로써 적은 수의 BS 가 있는 환경에서도 NLoS 환경의 특징을 GMM 모델에 반영할 수 있음을 의미한다. 또한 dataset $\hat{\boldsymbol{d}}$ 와 달리 CDA 를 이용한 dataset $\hat{\boldsymbol{Z}}$는 $g_M$함수를 거치면서 BS 들의 위치에 관한 정보가 dataset 에 반영되면서 훨씬 많은 양의 정보를 포함될 수 있게 되었다.

## Ⅳ. 시뮬레이션

100m×100m 크기의 환경에 BS 는 반지름이 35m 인 원 위에 6 그리고 8 개가 배치된 상황을 각각 고려한다. 전파 환경은 3GPP 에서 제안하는 Urban Micro (UMi) 시나리오를 고려한다. UMi 는 가혹한 무선 전파환경을 가지고 있는 시나리오로 높은 Non-Line of Site (NLoS) 확률을 가지는 특징이 있다. BS 와 UE 사이의 거리 $d$에 따른 Line of Site (LOS)가 될 확률을 나타내는 함수 $P_{LoS}(d)$ 는 3GPP UMi 시나리오에서 인용하였다. 수식(1)에서 $n$은 zero mean Gaussian noise 이고, NLoS 상황에서의 $b$ 는 평균이 5 인 지수분포를 따른다고 가정하였다. 100m×100m 그리드를 2m 간격으로 나누어 각 cell 마다 GMM 을 생성하였다. 각 cell 당 dataset 의 수는 500 개, 1000 개로 하였다. 또한 대조군은 SVE 기반 측위, SI 기반 측위 그리고 CDA 기법이 적용된 SI 기반 측위를 사용하였다.




[그림 1] dataset=1000 일때의 추정 위치에 RMSE에 대한 CDF

| BS # (N) | Data points | 측위 방식 | RMSE | Med. | Std Dev. | 95th Perc |
|---|---|---|---|---|---|---|
| 6 | 500 | SVE | 3.82 | 3.24 | 2.67 | 8.89 |
| | | SI | 3.63 | 1.98 | 4.29 | 12.4 |
| | | CDA_SI | 1.97 | 1.59 | 1.71 | 4.82 |
| | 1000 | SVE | 3.81 | 3.22 | 2.73 | 8.91 |
| | | SI | 3.56 | 1.94 | 4.22 | 12.09 |
| | | CDA_SI | 1.93 | 1.56 | 1.68 | 4.72 |
| 8 | 500 | SVE | 3.65 | 3.24 | 2.23 | 7.79 |
| | | SI | 3.23 | 1.83 | 3.71 | 10.51 |
| | | CDA_SI | 1.95 | 1.34 | 5.81 | 3.84 |
| | 1000 | SVE | 3.66 | 3.20 | 2.24 | 7.81 |
| | | SI | 3.06 | 1.73 | 3.55 | 9.98 |
| | | CDA_SI | 1.67 | 1.35 | 3.51 | 3.49 |

[표 1] CDA 기법을 활용한 SI 측위법과 기존 기법의 성능

그림[1]과 [표 1]을 통해 어떠한 경우에서도 CDA 를 활용한 측위 방식이 성능이 더 좋다는 사실을 알 수 있다. CDA 과정에서 dataset 이 가지고 있는 정보와 차원을 증가시켰기 때문에 정확도가 높아지는 것은 자명하다. 하지만 차원이 높아짐에 따라 오버피팅 현상이 조금 발생하여 백분위 99th 이상의 RMSE 값이 커지는 현상이 있었다.

## Ⅴ. 결론

본 논문에서는 SI 기반 측위 기법에 CDA 기법을 적용할 것을 제안했다. 기존의 SI 기반 측위에 사용되던 dataset 에 CDA 기법을 적용하였더니 의미 있는 성능 향상을 보였다. 향후 연구에서는 CDA 를 통해 차원을 증강시킨 후 모든 차원의 데이터를 쓰는 것이 아닌 필터 또는 샘플링과 같은 기법을 통해 정보가 중첩되어 있는 데이터들을 없앤 후 GMM 학습에 사용해볼 것이다. 이렇게 되면 성능은 올라가고 트레이닝 오버헤드는 상당히 줄어들 것으로 예상된다. 또한 데이터의 차원을 조절할 수 있게 되므로 CDA 기반 측위의 단점인 오버피팅 현상도 줄일 수 있을 것이다.


## ACKNOWLEDGMENT

이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2022R1F1A1072911).


## 참 고 문 헌


[1]. S. M. Yu, K. Han, J. Park, S.-L. Kim, and S.-W. Ko, “Combinatorial data augmentation : A key enabler to bridge geometry and data-driven WiFi positioning,” arXiv preprint arXiv:2105.07475, 2021.

[2] F. Morselli, S. M. Razavi, M. Z. Win, and A. Conti, “Soft information based localization for 5G networks and beyond,” IEEE Trans. Wireless Commun., vol. 22, no. 12, pp. 9923-9938, Dec. 2023.