JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 11: 1528~1533, Nov. 2020

기계학습 기반의 실내 측위 성능 향상을 위한 학습 데이터 전처리 기법

김대진[1] · 황치곤[2] · 윤창표[3*]

Learning data preprocessing technique for improving indoor positioning performance based on machine learning

Dae-Jin Kim[1] · Chi-Gon Hwang[2] · Chang-Pyo Yoon[3*]

[1]Invited Professor, Institute for Image & Cultural Contents, Dongguk University, Seoul, 04626 Korea
[2]Assistant Professor, Dept. of Computer Engineering, IIT, Kwangwoon University, Seoul, 01897 Korea
[3*]Assistant Professor, Dept. Of Computer & Mobile Convergence, GyeongGi University of Science and Technology, Siheung-si, 15073 Korea

요 약

최근 Wi-Fi 전파 지문을 이용한 실내 위치 인식 기술이 다양한 산업 분야 및 공공 서비스에서 적용되어 운영되고 있다. 기계학습 기술의 관심과 함께 단말 주변의 무선 신호 데이터를 사용한 기계학습 기반의 위치 인식 기술이 빠르게 발전하고 있다. 이때 기계학습에 필요한 무선 신호 데이터의 수집 과정에서 왜곡되거나 학습에 적합하지 않은 데이터가 포함되어 위치 인식의 정확도가 낮아지는 결과가 발생한다. 또한 특정 위치에서 수집된 데이터를 기반의 위치 인식을 수행하는 경우 학습에 포함되지 않은 주변 위치에서의 위치 인식에 문제가 발생한다. 본 논문에서는 수집된 학습 데이터의 전처리 과정을 통해 향상된 위치 인식 결과를 얻기 위한 학습 데이터 전처리 기법을 제안한다.

ABSTRACT

Recently, indoor location recognition technology using Wi-Fi fingerprints has been applied and operated in various industrial fields and public services. Along with the interest in machine learning technology, location recognition technology based on machine learning using wireless signal data around a terminal is rapidly developing. At this time, in the process of collecting radio signal data required for machine learning, the accuracy of location recognition is lowered due to distorted or unsuitable data for learning. In addition, when location recognition is performed based on data collected at a specific location, a problem occurs in location recognition at surrounding locations that are not included in the learning. In this paper, we propose a learning data preprocessing technique to obtain an improved position recognition result through the preprocessing of the collected learning data.

키워드 : 기계학습, 실내 측위, 전파 지문, 랜덤 포레스트
Keywords : Machine Learning, Indoor Positioning, Wi-Fi Fingerprint, Random Forest

Received 13 October 2020, **Revised** 20 October 2020, **Accepted** 24 October 2020

*** Corresponding Author** Chang-Pyo Yoon(E-mail:cpyoon@gtec.ac.kr, Tel:+82-31-496-6410)
Assistant Professor, Dept. Of Computer & Mobile Convergence, GyeongGi University of Science and Technology, Siheung-si, 15073 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.11.1528 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

실내 위치 인식과 관련된 연구는 스마트 기기의 보급과 함께 급격히 증가하고 있다. 기존의 연구는 통신사의 기지국 또는 주변 Wi-Fi와 같은 장치의 신호 값을 이용하여 단말의 위치를 특정하는 방법이 사용되고 있다. 그러나 신호를 이용한 삼각 측량기법과 GPS를 이용한 위치 인식은 위치의 오차범위가 환경에 따라 다르게 나타나는 문제와 실내와 같은 환경 그리고 주변 장애물에 의해 신호가 왜곡된 경우에서는 정확도가 급격히 떨어진다. 특히 실내 환경에서 인접한 두 장소의 사용자 위치가 서로 뒤바뀌는 현상은 너무나 자주 발생하는 문제이다[1]. 따라서 기존의 실내 위치 인식 기법을 개선하는 기술의 연구가 필요하다.

최근 기계학습에 관한 관심이 늘어남에 따라 이와 관련된 연구가 다양한 적용 분야에서 진행되고 있으며 실내 위치 인식 기술에 적용하고자 하는 연구가 활발하게 이루어지고 있다. 기계학습은 학습 데이터를 통해 학습하여 데이터를 판단하는 기술이다[2]. 이는 지도학습과 비지도, 강화 학습을 통해 문제의 해답을 찾는 과정으로 설명할 수 있다. 이때 학습에 사용하는 데이터의 정답을 알고 있는지에 따라 크게 두 가지의 학습 기술로 구분한다. 그중 문제의 정답을 알고 있는 데이터 집합을 이용하여 모델의 parameter 값을 학습하는 지도학습은 새로운 데이터가 입력될 때 정답을 추론하는 문제에 많이 적용된다. 즉 주어진 입력 벡터가 어떤 종류의 값인지 표시하는 분류에서 주어진 문제를 해결할 수 있는 기계학습 알고리즘들인 결정 트리, Support Vector Machine (SVM)[3][4], K-Nearest Neighbors(KNN)[5] 알고리즘을 통해 학습하고 입력값을 판단한다. 이와 같은 분류 문제를 해결할 수 있는 알고리즘을 통해 실내 위치 인식 문제에 기계학습을 적용하여 해결할 수 있다[6].

기계학습을 통해 특정 위치를 인식하고 정답을 결정하는 과정은 간단하고 단순한 과정으로 진행된다. 분류 문제로 구분되는 기계학습 기반의 실내 위치 인식 기술의 연구는 위치 라벨이 결정된 학습 데이터의 수집 과정과 기계학습 알고리즘을 통한 학습 과정 그리고 훈련 과정을 거치며 단말이 수신한 신호 데이터들을 기준으로 위치를 결정 또는 분류하는 과정을 수행한다. 그러나 학습에 사용되는 데이터의 수집 과정에서 단시간에 출현하는 신호 데이터와 같이 학습에 적합하지 않은 데이터가 포함되는 문제가 발생한다. 또한 수집된 학습 데이터에 위치 결정에 적합하지 않은 데이터를 다수 포함하게 된다. 따라서 수집된 학습 데이터의 전처리 과정을 통해 기계학습에 적합한 데이터를 필터링하는 과정이 필요하다.

본 논문의 구성은 다음과 같다. 2장에서 수집된 학습 데이터의 문제점과 데이터 전처리 과정의 필요성, 그리고 실내 위치 인식에 사용되는 기계학습 알고리즘들을 기술한다. 3장에서는 제안 기법의 데이터 전처리 기술 그리고 적용 알고리즘을 기술한다. 4장에서는 제안 기술을 통한 실험과 결과를 확인한다. 5장에서는 제안 기술의 결론과 보완해야 할 문제점을 기술한다.

Ⅱ. 학습 데이터

본 논문은 실내 위치 기반 서비스에 필요한 학습 과정의 문제점을 크게 두 가지로 바라본다. 첫 번째는 학습 데이터의 수집 과정에 포함되는 학습에 적합하지 않은 데이터의 수집 문제이며 두 번째 특정 위치들 사이의 모호한 경계로 발생 가능한 위치 결정의 장애 발생 문제이다.

2.1. 데이터 수집

실내 위치 인식이 요구되는 장소를 기준으로 주변의 무선 신호 수집은 해당 장소를 기준으로 장소 내부에서 임의의 위치에 수신되는 무선 신호 값들을 저장한다. 이 과정은 수집 위치에서 주변 무선 신호를 수집하여 생성되는 신호 테이블이다. 이러한 신호 테이블에 포함하는 값들은 측정 시간, 측정 위치, 수신된 무선 MAC 주소들과 해당 장치별 무선 신호의 세기 값들을 포함한다. 아래 표 1에 학습 데이터의 개요를 나타냈다.

Table. 1 Overview of traing data

Time Stemp	Location Label	MAC address	MAC...
1.55799E+12	326	-73	...
1.55782E+12	320	-	...
1.55783E+12	320	-82	...
...	...	...	...

학습 데이터의 수집 과정은 다음과 같다. 학습에 필요한 데이터는 측정 시간 값, 측정 위치 라벨, 수신된 주변 무선 신혼들의 세기 값으로 구성한다. 특정 위치의 분류

를 위해 측정 위치들을 라벨링하고 해당 위치에서 수신되는 Wi-Fi beacon 신호 세기 값을 수집한다[7]. 즉 측정 위칫값을 기준으로 수신된 무수히 많은 주변 무선 신호들로 구성한다.

신호 수집 위치는 일정 거리 간격 또는 무작위로 위치 결정하여 수집하며 장소 정보들과 장소별 신호 값들로 구성한다.

2.2. 학습 데이터의 문제점

신호 테이블에 포함된 특정 위치의 신호 값 중에는 학습에 적절하지 않은 왜곡되거나 일시적으로 출현한 신호와 다른 위치와 중복되지만, 효율이 떨어지는 미약한 신호 값이 포함된다[8]. 또한 앞서 언급한 두 번째 문제로 위치를 특정할 수 없으며 측위자가 장소로 구분되는 블록의 경계에 위치하였을 때 어떠한 장소에 속하는 위치인지를 결정하는데 문제가 발생한다. 즉 수집된 데이터는 장소로 구분하여 라벨링 되고 특정된 위치와 그렇지 않은 위치에 대한 경계가 모호하게 되어 라벨링들 사이에 존재하는 중간 위치의 경우 해당 위치가 어느 라벨에 속하는지 결정하는 과정에서 판정이 잘못되는 문제가 발생한다. 이처럼 라벨링 되어 특정된 위치와 그렇지 않은 주변 위치에 중복으로 나타나지만, 그 신호가 미약하거나 지속적이지 않은 데이터를 포함한다면 위치의 판단은 매우 어렵게 된다. 따라서 발생 가능한 문제를 해결하기 위해 학습 데이터의 전처리 과정이 요구된다.

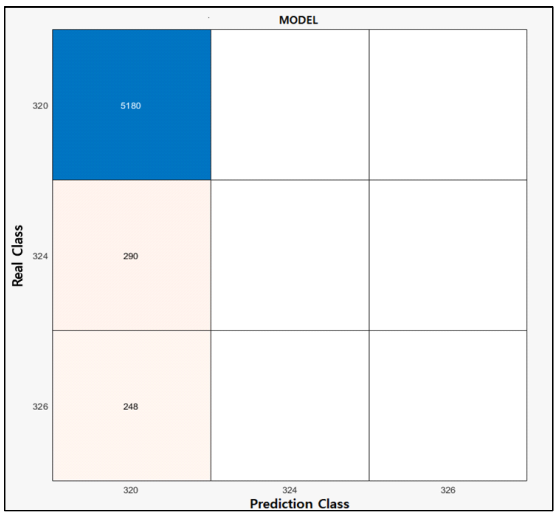

Fig. 1 Classification table

그림 1에 문제가 발생하여 위치 결정에 문제를 갖는 정오 분류표를 나타내었다. 이때 서포트 벡터 머신 중에서 기본적인 선형 SVM으로 실험을 진행한 결과이며 전처리 과정을 거치기 전의 결과이다. 이는 위치 인식에 데이터 집합이 적었을 때 정답률이 0%에 이르는 결과가 도출됨을 확인할 수 있다. 이때 실내 위치 인식을 위해 일반적으로 사용되는 분류 기법인 SVM과 KNN 알고리즘을 이용한 결과이다.

Ⅲ. 제안 기술

기계학습에 필요한 학습 데이터의 전처리 과정은 향상된 실내 위치 인식 결과를 얻기 위해 데이터를 정제하고 분류하는 데이터 관리 방법이 필요하다. 이처럼 관리된 데이터를 기계학습에 적용할 때 적절한 학습 알고리즘의 선택은 전체 시스템의 성능에 큰 영향을 준다. 따라서 적용 가능한 알고리즘의 선택은 매우 중요하다[9].

3.1. 데이터 전처리

제안하는 데이터의 전처리 기법은 측정된 신호 테이블(PT)로부터 유효한 데이터를 추출하기 위해 수식 1과 같은 방법을 사용한다.

$$\Pi_{p,\alpha_i}\left(\sigma_{p=\theta_j \wedge \alpha_i \geq \theta_B}(PT)\right) \quad (1)$$

수식 1의 각 요소와 조건에 이용되는 변수는 다음과 같다. p는 wifi 신호를 측정한 장소이고, θ_j는 장소(p)의 레이블이다. α_i는 측위에 따라 검출된 신호 값이며, θ_B는 wifi 신호의 수신율을 판단하는 기준값(-85dBm)이다. 그리고 PT는 전체 신호 테이블이다.

수식 1에서 의해 추출된 선택된 값은 해당 장소의 내부를 나타내는 inbound 신호이고, 그 외의 신호는 측정된 어느 장소에도 포함되지 않는 외부 위치를 나타내는 outbound 신호로 구분된다.

전체 신호 테이블(PT)에서 검출된 장소(θ_j)가 찾고자 하는 장소(p)이고 검출된 wifi 신호 값(α_i)이 wifi 검출을 위한 기준값인 -85dBm 이상의 값을 갖는 장소(p)와 그때의 wifi 신호 값 (α_i)을 검출하기 위한 공식이다. 이 과정은 각 장소의 유효한 신호들을 추출하기 위함이다. 따라서 지정 장소에서 측위를 할 때, inbound 신호와

outbound 신호를 구분하여 장소를 인식한다. 지정 장소에서는 inbound 신호 데이터를 통해 위치의 장소를 확인하지만, 지정 장소가 아닌 곳일 경우에는 outbound 신호 데이터를 통해 확인해야 한다. 다만 이로 인해 inbound를 위한 측위와 outbound를 위한 측위를 별도로 해야 하는 번거로움이 발생한다.

이와 같은 문제를 해결하기 위해 측위가 진행 될 때 inbound 신호와 outbound 신호를 분리하여 p와 θ_j를 이용하여 지정 장소 내부인지 외부인지를 확인한다. 이에 따라 inbound 신호를 이용한 측위와 outbound 신호를 이용한 측위를 선택한다.

3.2. 학습 알고리즘

앙상블(ensemble method)는 여러 전문가로부터 얻은 복수의 의견을 적절한 방법으로 결합하여 좀 더 적절한 의사결정을 하는 접근법이다. 기계학습에서 앙상블은 여러 개의 기계학습 모델을 결합하여 하나의 강한 모델을 만드는 기법이다. 이때 부족한 데이터의 양이나 품질을 극복하기 위해 재샘플링 기법을 이용한다[10].

랜덤 포레스트는 결정 트리의 앙상블 모델이며, 재샘플링 기법으로 배깅을 이용한다[11]. 일반적으로 트리를 제작할 때 결정 트리는 순도(Homogeneity)를 증가시키고 불순도(Impurity)와 불확실성(Uncertainty)을 감소시키는 방향으로 제작하여야 한다[12].

이를 위해 우리는 엔트로피를 기준으로 가지치기를 수행하여 트리를 구성한다. 엔트로피란 불순도를 정량화한 수치로 최소가 되는 방향으로 분류를 하는 것이 최적의 분류라고 할 수 있고, 이를 통해 과대적합(overfitting) 문제를 해결할 수 있다. 엔트로피가 높은 집단은 그 집단의 특징을 찾기 힘들고, 과대적합은 의사결정 트리에서 높은 정확성을 제공할 수 있지만, 새로운 데이터에 대해서는 정확도가 감소한다. 따라서 이를 위해 본 논문의 제안 기술이 적용된 실내 위치 인식 과정의 정확도 향상을 위해 랜덤 포레스트 기법을 사용한다.

Ⅳ. 실 험

본 논문은 실내 위치 인식 시스템의 향상된 위치 인식률을 위해 신호 데이터 테이블의 전처리 과정과 그로 인해 구분되는 데이터 값의 분류를 제안하였다. 이에 이 장에서 제안 기술의 성능을 실험하고 제안 기술을 평가한다.

실험 환경은 AMD Ryzen 9 3900X CPU와 32GB 메모리로 구성된 PC 환경에서 Windows 10 운영체제를 기반으로 이루어졌다. 그리고 실험에 사용된 SW는 MATLAB의 분류 학습기를 사용하였으며 실험에 사용된 데이터 세트는 세 구역으로 구분된 장소에서 수집된 272개의 MAC 주소의 신호 데이터를 사용하여 실험하고 평가하였다.

실내 위치 인식은 측위 된 시간과 위치가 중요한 데이터가 되고, 측위의 변수에 따라 데이터의 양이나 정확도에 차이가 발생할 수 있다. 이에 따라 본 논문에서 적용할 기계학습기법으로 앙상블 기법의 하나로 랜덤 포레스트 기법을 적용하였다. 이에 따른 방식으로 테스트 결과는 표 2와 같고, 테스트를 위한 설정은 앙상블 기법의 랜덤 포레스트, AdaBoost, 결정 트리로 하고 최대 분할 수는 20을 기본으로 하였다.

Table. 2 Initial setup for learning

	Number of learners	Learning rate	accuracy	Total misclassification cost
case1	60	0.1	99.80%	10
case2	50	0.1	99.80%	10
case3	40	0.1	95.60%	250
case4	30	0.1	97.80%	70
case5	20	0.1	96.40%	68
case6	10	0.1	95.60%	250

각 case는 학습기 수, 학습률을 차등하여 변화를 주었으며 이에 따른 정확도와 총 오 분류 비용은 표 2와 같이 나왔으며, 학습기 수가 50을 초과하면 정확도와 총 오 분류 비용의 차이가 거의 없는 것으로 확인되었다. 이에 inbound와 outbound를 분리하여 측정함으로써 측위 결과에 대한 위치 판단에 대한 정확성을 향상시킬 수 있다.

그림 2는 각 학습기 수에 따른 학습률과 정확도 그리고 오 분류 비용을 그래프로 나타낸다. 50 이상의 학습기 수에서 정확도는 99.8%의 높은 정확도를 나타낸다. 즉 제안 기술을 적용하지 않고 트리 또는 SVM 알고리즘을 적용한 실험 결과와는 큰 성능 차이를 보인다.

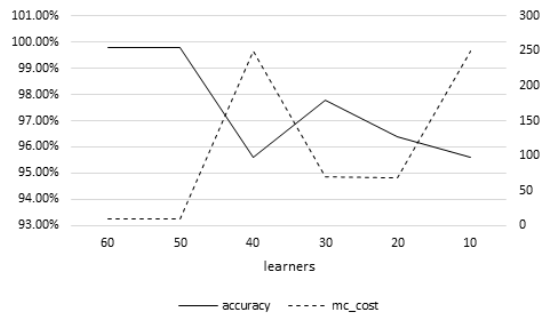

Fig. 2 Graph of learning result by initial value

그림 3은 제안 기술을 적용한 실험에 대한 정오 분류표이다. 제안 기술은 적은 수의 장소 데이터를 이용하더라도 위치를 인식하는 결과를 얻을 수 있음이 확인됐다.

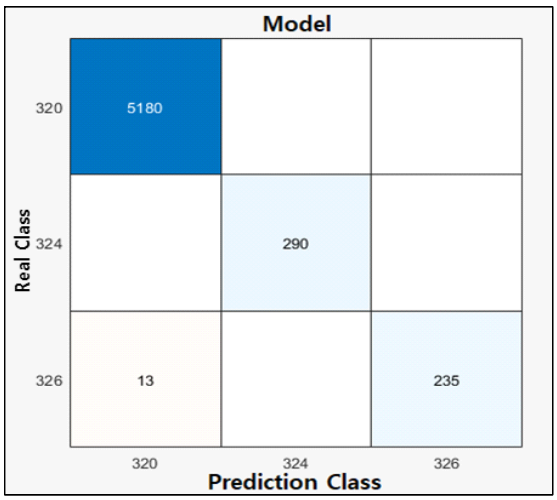

Fig. 3 Graph of learning result by initial value

Ⅴ. 결 론

본 논문은 실내 위치 인식률의 향상을 위하여 학습 데이터의 전처리를 통한 데이터 분리와 분리된 데이터를 기반으로 기계학습을 수행하는 전처리 기법을 채택하였다. 이러한 방식은 위치 정보가 저장된 데이터에서 위치의 정확도 개선과 장소 외부의 판단을 하나의 데이터로 처리하기 위함이다. 이에 측위자의 위치를 더욱 정확하게 결정하기 위해 기계학습 기법 중 랜덤 포레스트를 이용한 실내 위치 인식 기술을 제안하였다. 이는 스마트 기기에서 측정된 무선 신호 데이터를 이용하여 실내 사용자의 위치를 판단하기 위해 연구되었으며, 무선 신호를 사용하는 측위 시스템의 많은 분야에서 적용할 수 있다.

REFERENCES

[1] C. P. Yoon and C. G Hwang, “Efficient indoor positioning systems for indoor location-based service provider,” *KIICE*, vol.19, pp. 1368-1373, Jun. 2015.

[2] R. Sheikhpour, M. A. Sarram, S. Gharaghani and M. A. Z. Chahooki, “A Survey on semi-supervised feature selection methods,” *Pattern Recognition*, vol. 64, pp. 141-158, Apr. 2017.

[3] A. M. Abd and S. M. Abd, “Modelling the strength of lightweight foamed concrete using support vector machine (SVM),” *Case studies in construction materials* 6, pp. 8-15, 2017.

[4] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, “Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection.” *IEEE Access* 6, pp. 33789-33795, 2018.

[5] N. Papernot and M. Patrick “Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning.” *arXiv preprint arXiv*:1803.04765, pp. 1-18, 2018.

[6] J. Behmann, K. Hendriksen, U. Müller, W. Büscher, and L. Plümer, “Support Vector machine and duration-aware conditional random field for identification of spatio-temporal activity patterns by combined indoor positioning and heart rate sensors,” *Geoinformatica*, vol. 20, no. 4, pp. 693-714, 2016.

[7] D. Kim, S. H. Park, and H.K. Jung, “Fingerprint-Based Indoor Logistics Location Tracking System,” *KIICE*, vol.24, no.7, pp. 898-903, 2020.

[8] C. P. Yoon, I. K. Lee, and C. G. Hwang, “The iBeacon Signal Optimization Methods for Improving the Reliability of Indoor Positioning Systems,” *Journal of Engineering and Applied Sciences*, vol. 12, no. 10, pp. 2692-2696, 2017.

[9] E. S. Lohan, J. Torres-Sospedra, H. Leppäkoski, P. Richter, Z. Peng, and J. Huerta, “Wi-Fi crowdsourced fingerprinting dataset for indoor positioning,” *Data*, 2017.

[10] G.James, D.Witten, T.Hastie, and R.Tibshirani, “An introduction to statistical learning: with applications in R,” *Spinger*, 2013.

[11] A.Criminisi, J.Shotton, and E.Konukoglu, “Decision forests: A unified framework for classification, regression, density

estimation, manifold learning and semi-supervised learning," *Foundations and Trends in Computer Graphics and Vision* 7.2–3, pp. 81-227, 2012.

[12] L.Rokach and O.Maimon, "Top-down induction of decision trees classifiers-a survey," *IEEE Transactions on Systems, Man, and Cybernetics*, Part C(Applications and Reviews), vol. 35, no. 4, pp. 476-487, 2005.

김대진(Dae-Jin Kim)

1998년 대진대학교 전자공학과 (공학사)
2000년 대진대학교 전자공학과 (공학석사)
2010년 대진대학교 전기전자통신공학과 (공학박사)
2017년~현 재: 동국대학교 영상문화콘텐츠연구원 교수
※관심분야 : 코덱, 멀티미디어 플랫폼, 콘텐츠 DNA, 워터마크, 딥러닝 등

황치곤(Chi-Gon Hwang)

2012년 광운대학교 컴퓨터과학과 (공학박사)
2006년~2015년:(주)인찬 연구원
2016년~2018년: 경민대학교 인터넷정보과 교수
2019년~현 재: 광운대학교 정보과학교육원 컴퓨터공학과 교수
※관심분야 : 모바일 클라우드, 온톨로지, 기계학습, NLP

윤창표(Chang-Pyo Yoon)

1998년 광운대학교 전자계산학과 (이학사)
2001년 광운대학교 컴퓨터과학과 (공학석사)
2012년 광운대학교 컴퓨터과학과 (공학박사)
2012년~현 재: 경기과학기술대학교 컴퓨터모바일융합과 교수
※관심분야 : 기계학습, 모바일 시스템, 네트워크 보안, 무선 네트워크