Supervised Learning based Small Cell Planning

Seok-Yeong Bang, Sang-Yeon Lee, Byungsuk Kim*, Kyeongjun Shin*, Een-Kee Hong
KyungHee University, KT*

qkdtjr97@khu.ac.kr, sangyeon@khu.ac.kr, b.kim@kt.com*, shin.kh@kt.com, ekhong.khu.ac.kr

Abstract— As mobile communication system evolve from LTE to 5G, a wider range of services are being provided leading to a rapid increase of mobile data usage. The increase of traffic usage can be attributed to the commercialization of data-intensive content like video streaming, augmented/virtual reality (AR/VR), resulting in the installation of 5G base stations to accommodate the demand for these new services. Due to the increase in 5G base stations, there has been a growing interest in small cell planning. Among many methods for small cell planning, we focus on determining whether to target outdoor or indoor areas for base station deployment when additional installations are needed for efficient small cell planning. For this, we leveraged machine learning (ML) models to analyze base station data and categorized them into four main usage types, extracting patterns for each category. In this paper, we proposed a framework utilizing these extracted patterns to estimate indoor service ratio and provide criteria for indoor/outdoor small cell deployment. Furthermore, using the proposed framework in this paper, we estimated the indoor traffic ratio for office, shopping mall and Gangnam Station basestation data. Estimation results showed that the indoor traffic ratio for shopping mall data and office data was approximately 98% and 97%, respectively, approaching almost 100%. And the estimated indoor traffic ratio for the entire Gangnam Station area is approximately 68%, which closely aligns with the results reported in other studies ranging from 70-80%.

Keywords—mobile traffic, supervied learning, classification, small cell, cell planning

I. INTRODUCTION

As mobile communication system evolve from LTE to 5G, a wider range of services are being provided, leading to a rapid increase of mobile data usage. According to [1], global mobile data traffic reached 67 Exabytes per month in 2021 and is projected to reach 282 Exabytes per month by 2027. The increase of traffic usage can be attributed to the commercialization of data-intensive content like video streaming, augmented/virtual reality (AR/VR), resulting in the additional installation of 5G base stations to accommodate the demand for these new services. The authors in [1] suggested that the growth in traffic until 2027 will be influenced by the early adoption of extended reality (XR) services, including AR/VR, and mixed reality (MR).

In this context, it is crucial to study the cell planning that are suitable for the 5G era [2]. Traditionally, the network operators deployed base stations to support the demand for peak-hour traffic. However, during the periods when base stations operate at low load or even when no individual users are actively using them, such as during nights and weekends, the base stations consume power equivalent to that of peak traffic hour [3]. To solve this problem, we can leverage small cells to efficiently distribute and handle the large amount of data. To optimize the efficiency with a given budget, it is necessary to deployment of small cells in the areas where they are needed most. Furthermore, small cells can be easily turned on and off, allowing them to be deactivated during non-peak hours to save power consumption.

Research about small cell planning has evolved from addressing capacitiated facility location problems in 2G/3G cellular system to embracing the emergence of HetNets in the 4G cellular network. The 4G cellular network is quite different from the initial one where macro base stations were firstly deployed. This has steered researches toward small base station deployment suitable to the LTE mobile network. Meanwhile, many researches have focused on the small cell deployment [4]-[6]. Doru Calin et al. investigate insights on possible deployment architectures for femtocells along with an analysis framework for quantifying macro offloading benefits in realistic network deployment scenarios by means of advanced performance analysis techniques [4]. Shaowei Wang et al. proposed approximation algorithms to minimum cell planning problem in heterogeneous networks. The planning task involves selecting a subset of possible base station (BS) sites to minimize the overall deployment cost [5]. They also tackled the budgeted cell planning problem in HetNets [6]. Subsequently, [7, 8] proposed methods to address small cell planning problem taking into acount the interference caused by small cell deployment. Additionally, based on research that suggested 70-80% of mobile data is generated indoor [9], studeis related to small cell deployment indoor have also been conducted. In [10], Guo et al. considered a single femto cell in a 1-dimentional indoor scenario and derived the worst cell-edge throughput. And Weisi Guo et al. understand the significance of indoor-generated traffic and propose optimal locations for deploying indoor access points (APs) at building [11]. However, to the author's knowledge, there's no research to provide criteria for determining whether to target the outdoor area or indoor area when additional base station installations are needed.

Therefore, in this paper, to provide criteria that can be refered when installing additional base stations, we estimated the ratio of downlink traffic served to indoor

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

users and outdoor users by the deployed base stations. To estimate indoor traffic ratio, an analysis of the characteristics of traffic genereated indoor is necessary. According to the [12], indoor data traffic is analyzed to have unique characteristics. [12] propose Machine Learning(ML)-based method to analyze indoor data traffic.

This paper aims to propose a framework that provides criteria necessary for small cell planning by leveraging the capability of ML model to classify indoor data traffic. To achieve this, we propose extracting base station usage specific patterns using ML and using these patterns to estimate indoor/outdoor traffic ratio of base stations.

The remainder of this paper is organized as follows. In Section II, we explain the dataset and propose framework of estimation of the indoor/outdoor traffic ratio. ML model used for pattern extraction also introduced in this section. In Section III, Multiple classification results using various ML models and the optimal ML model-based estimation of the indoor/outdoor traffic ratio are presented. Conclusion is drawn in Section IV.

II. DATASET AND SYSTEM MODEL

In this section, we explain dataset used in the study and present an overview of the proposed framework for estimating the indoor/outdoor ratio. The dataset used for this research is obtained from a South Korea mobile carrier and consists of downlink traffic data collected hourly from each base station, from September 17, 2022, 00:00 to October 16, 2022, 23:00. The term "collected hourly" means to the summation of downlink traffic volume during 1 hour. This dataset is used for the pattern extraction, indoor/outdoor ratio estimation and validation of estimation result. The validation will be assessed by comparing the average estimated indoor/outdoor ratio value for the base station data from Gangnam Station with the ratio provided in [7]. And also by verifying whether the indoor ratio of the indoor base station data, which has been reliably labeled by the mobile carrier, approaches 1.

A. Dataset

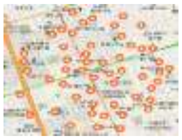

Fig 1. Gangnam

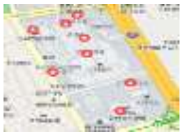

Fig 2. Coex

Fig 3. Office building

In this paper, to classify mobile data traffic generated indoors, downlink data of mobile traffic from base stations located within a 1km radius of Gangnam Station, COEX and Office building located in Gwanghwamun (as shwon in Figure 1 to 3) are used. The total number of base stations is 139, including 87 within a 1km radius of Gangnam Station, 36 at COEX, and 16 in Office building. The dataset is labeled based on the area purposes where the base stations are located. Labeling is conducted using both Kakao Map's area usage labeling or base station information provided by mobile carrier. For the Gangnam Station data, 25 were labeld as residential areas, 64 as commercial areas, 2 as other areas, and 5 as targeting roads. In 36 base station data at COEX, 30 are labeled as shopping mall areas, and 6 as office areas. And the office buidling data in Gwanghwamun, 16 out of the data are labeled as office areas. Among these, residential areas, shopping mall areas, and office areas are assumed to represent indoor data traffic, while data originating from roads are assumed to represent outdoor traffic.

B. Framework of Estimation Indoor/Outdoor Ratio

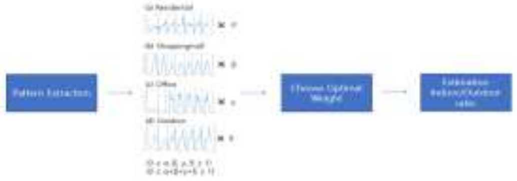

Fig 4. procedure of Estimation

The framework for estimating the indoor/outdoor ratio of base stations is illustrated in Figure 4 and follows the process as described. Firstly, ML model is used to train labeled base station data categorized into residential, shopping mall, and office usage. Data categorized into outage usage is reliably labeled by the mobile carrier. So, without procedure of classification for pattern extraction, it is assuemd that the average of the data represents the outage pattern. Patterns for each category are extracted by averaging the highly accurately classified data from the trained supervised learning model with each category. These extracted patterns, normarlized between 0 and 1 as time series data, are then multiplied by weights α, β, γ and δ for residential, shopping mall, office, and outdoor, respectively. The optimized weight set of each base station data traffic is decided when MAE (Mean Absolute Error) between base station traffic data and estimated data that made by combining the weighted patterns is minimized. In the optimized weights, the weigths corresponding to indoor data α, β and γ are used to estimate the ratio of indoor-generated data traffic relative to estimated data. The remaining weight δ is then used to estimate the ratio of outdoor generated data traffic.

C. Process of Pattern Extraction

The process of extracting patterns based on the purpose of the base station is as follows. Firstly, since the current data proportion is not balanced, a data balancing process is necessary. Among the labeled data, the category with the smallest dataset is data for office usage that consists of 22 labeled. Therefore, in order to balance the data proportions, 22 data for each category of residential, commercial, and shopping mall usage categories need to be selected. The selection process is as follows: Initially, a supervised learning model is trained using the train data, followed by classification of the test data. The ratio of target data is kept consistent between the train and test data. Classification is performed 100 times for each target category, and 22 data with the highest probabilities of being correctly classified based on the labeled information are selected for each category. This procedure helps create a balanced dataset for each base station usage categories. Following that, the newly generated dataset is used to process another round of classification. Similarly, classification is performed 100 times for each target category, and the data that has correctly classified possibility exceeding 0.8 are selected and averaged to extract patterns to each base station usage category.

D. Optimal Classification Model

In the previous study by [12], the characteristics of indoor data traffic were invetigated using a decision tree model. In addition to this, since we need to extract patterns specific to each base station usage category in this paper, we conducted research to determine the optimal model that can improve classification accuracy. 5 ML models commly used in time series classification - Decision Tree, RandomForest, AdaBoost, GradientBoost and XGBoost - were considered for the study. Using the balanced dataset created based on base station usage categories, each model performed 100 times classification for each target category. The model that yield higher probabilities correctly classified will be selected as the optimal model for pattern extraction.

III. PERFORMANCE EVALUATION

To find the optimal ML model for pattern classification, 5 different models were used to classify based on base station usage categories, and the classification results are shown in Table 1 and Table 2.

	residential	commercial	shopping mall	office
Decision Tree	0.389	0.681	0.666	0.714
Random Forest	0.24	0.798	0.728	0.881
AdaBoost	0.385	0.657	0.657	0.724
Gradient Boost	0.371	0.694	0.657	0.751
XGBoost	0.337	0.746	0.701	0.798

Table 1. Classification result of original dataset

	residential	Commercial	shopping mall	office
Decision Tree	0.798	0.738	0.58	0.852
Random Forest	0.738	0.702	0.803	0.905
AdaBoost	0.81	0.76	0.616	0.837
Gradient Boost	0.788	0.755	0.659	0.814
XGBoost	0.796	0.714	0.74	0.869

Table 2. Classification result of balanced dataset

Table 1 presents the classification results when the dataset for base stsation usage categories is not balanced, and the performance metric used for evaluation is the probabilities correctly classified after 100 times classification. Table 2 shows the classification results when the dataset is balanced, and the performance metric used for evaluation is the same as Table 1.

From the classification results, it can be observed that in most cases, classification accuracy improves after balancing the dataset and referring to Table 2, RandomForest and XGBoost model outperform the other 3 ML modles in terms of classification performance. However, it is becomes evident that the acccuracy of labeled information varies across different usage categories.

Specifically, the accuracy is relatively lower for the residential and commercial categories, as these were labeled using Kakao Map. In contrast, the accuracy is notably higher for the shopping mall and office categories, as they were labeled based on information provided by the mobile carrier. Taking this into consideration, we have selected the RandomForest model as the optimal choice for pattern extraction due to its superior classification accuracy in the datasets labeled for shopping mall and office categories.

The patterns extracted by RandomForest model are shown in Figure 5-9.

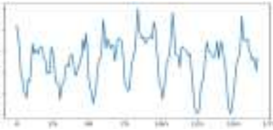

Fig 5. Residential

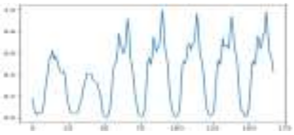

Fig 6. Commercial

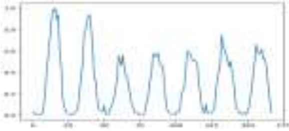

Fig 7. Shopping Mall

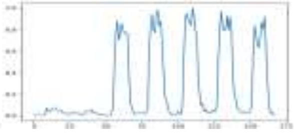

Fig 8. Office

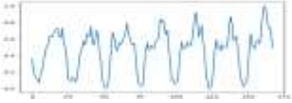

Fig 9. Outdoor

Upon analyzing the patterns, distinctive characteristics for each base station usage category were observed. For the residential pattern, a consistent pattern throughout the week with peaks during moring and evening hours was identified. The shopping mall pattern showed increased data generation during weekends, with the highest volume around 12 o'clock. The office pattern displayed minmal traffic on weekends and significant variations during weekday commuting hours. Additionally, the commercial pattern exhibited similarities to the outdoor pattern, showing consistent weekly patterns and daily traffic similarities. According to this, it is expected that the outdoor data ratio in commercial areas will be significantly high.

The results of estimating indoor ratios using the indoor/outdoor ratio estimation framework described in Section 2 are as follows. Firstly, we verify the estimated values for the base station data in office buildings, which were labeled based on information provided by the mobile carrier. The estimated indoor traffic ratio for the entire office dataset is 98%, signifying a remarkably high ratio. Notably, the weight γ associated with the Office pattern exhibited the highest magnitude across all datasets. The Coex base station data, which was also labeled based on information provided by the mobile carrier, exhibits an estimated indoor traffic ratio of 96% for the entire Coex dataset. The Coex dataset includes two types of data labeled as Office and Shopping Mall. It was observed that for all the data, except two, the weights closely matched the labeling information. The two excluded data points were labeled as Shopping Mall but showed high weights for the Office pattern. We plan to verify this with the mobile carrier in the future. Finally, when estimating indoor traffic for the base station data in Gangnam Station, the indoor traffic ratio is approximated at 68%, closely aligning with the suggested indoor traffic ratio of 70-80% from a previous study [9]. Earlier, we emphasized the similarity between commercial and outdoor patterns. As a result, we can observe that the base station data labeled as commercial areas exhibit a low indoor traffic ratio and a high outdoor traffic ratio. Additionally, there are even data points with indoor traffic ratios close to 0%.

There are some data points that require further verification of labeling information, but the estimation results show meaningful findings. The estimation results closely align with the previously suggested indoor traffic ratios, and the estimation values for base station data labeled as indoor traffic by the mobile carrier are close to 100%. This proves the effectiveness of the indoor/outdoor ratio estimation framework proposed in this paper. Therefore, we propose using the estimated indoor traffic ratio through this framework as a criteria for determining whether to target indoor or outdoor areas when additional base station installations are needed.

IV. CONCLUSION

In this paper, we propose a framework, utilizing machine learning models, to analyze base station usage patterns and estimate the indoor/outdoor service ratio of base station data using the extracted patterns. Through the process of finding the optimal machine learning model for pattern extraction, we have demonstrated the suitability of the RandomForest model for classifying base station data according to usage categories. Furthermore, by using the extracted patterns to estimate the indoor/outdoor ratio for each base station data traffic, we have confirmed that the average indoor ratio of the entire Gangnam Station data closely approximates the indoor ratio range suggested in [9]. Therefore, we propose using the estimated indoor traffic ratio through this framework as a criteria for determining whether to target indoor or outdoor areas when additional base station installations are needed.

ACKNOWLEDGMENT

"This research was conducted as part of the commisioned research project for the development of intelligent open small cell technology based on NR (New Radio) by the Electronics and Telecommunications Research Institute (ETRI) in South Korea" (ETRI-2018-0-0-01659)

REFERENCE

[1] Ericsson, "Ericsson Mobility Report," 2022.[Online]. Available:https://www.ericsson.com/en/reports-and-papers/mobilityreport/

[2] Q. Wu, L. Chen, X. Chen and W. Wang, "Cell planning for millimeter wave cellular networks," 2017 9th International Conference on Wireless Communications and Signal Processing (WCSP), Nanjing, China, 2017.

[3] Ericsson, Ericsson Mobility Report, June 2021.

[4] Doru Calin, Holger Claussen, Huseyin Uzunalioglu, "On femto deployment architectures and macro cell offloading benefits in joint macro-femto deployments," IEEE Communications Magazine, vol 48, Jan. 2010.

[5] Wentao Zhao, Shaowei Wang, "Cell planning for heterogeneous cellular networks," 2013 IEEE Wireless Communications and Networking Conference (WCNC), Apr. 2013.

[6] Shaowei Wang, Wentao Zhao, Chonggang Wang, "Budgeted Cell Planning for Cellular Networks With Small Cells," IEEE Transactions on Vehicular Technology, vol. 64, no. 10, Oct. 2015.

[7] Yosub Park, Jihaeong Heo, Hyunsoo Kim, Hano Wang, Sooyoung Choi, "Effective Small Cell Deployment with Interference and Traffic Consideration," 2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall), Sep, 2014.

[8] Yue Li, Kai Sun, Lin Cai, "Small-cell plnning based on uplink interference and traffic in two-ties cellular system," 2015 International Conference on Wireless Communications & Signal Processing (WCSP), Oct, 2015.

[9] Ericsson, "Planning in-building coverage for 5G: from rules of thumb to statistics and AI", 2021.

[10] W. Guo and S. Wang, "Interference-aware self-deploying femto-cell," *IEEE Wireless Commun. Letters*, vol. 1, no. 6, pp. 609–612, 2012.

[11] Weisi Guo, Siyi Wang, Xiaoli Chu, Jie Zhamg, Jiming Chen, Hui Song, "Automated small-cell deployment for heterogeneous cellular networks," IEEE Communications Magazie, vol. 51, May, 2013.

[12] 김영준, 유현민, 조영준, 이상연 홍인기, "Decision Tree 모델을 이용한 Indoor Traffic 분석," 2022 년도 한국통신학회 추계종합학술발표회.