RPMA Low-Power Wide-Area Network Planning Method Basing on Data Mining

Yao Shen, Xiaorong Zhu(✉), and Yue Wang

College of Telecommunication and Information Engineering,
Nanjing University of Posts and Telecommunications, Nanjing 210003, China
xrzhu@njupt.edu.cn

Abstract. A network planning method based on data mining was proposed for Random Phase Multiple Access (RPMA) low-power wide-area network (LPWAN) with large density of base stations and uneven traffic distribution. First, a signal quality prediction model was established by using the boosting regression trees algorithm, which was used to extract the coverage distribution spacial pattern of the network. Then, the weighted K-centroids clustering algorithm was utilized to obtain the optimal base station deployment for the current spacial pattern. Finally, according to the total objective function, the best base station topology was determined. Experimental results with the real data sets show that compared with the traditional network planning method, the proposed method can improve the coverage of low-power wide-area networks.

Keywords: Low power wide area network · Boosting regression trees
Weighted K-centroids · Base station deployment

1 Introduction

With the rapid development of the Internet of things, the number of interconnected devices will be expected to increase to 50 billion, and the traffic volume will increase by more than a thousand times [1]. Traditional short-range wireless technologies and cellular network technologies cannot meet the diversified IoT traffic requirements. Therefore, a new communication pattern, Low Power Wide Area Network (LPWAN) [2]. LPWAN [3] mainly includes NB-IoT, LORA, RPMA and other wireless communication technologies, which can support a large number of devices to access the network. Specially, RPMA can support 60x to 1300x more endpoints for a given network relative to Sigfox and LORA. Moreover, compared with their defect in capacity scalability, PRMA leverages a Time Division Duplex (TDD) approach to provide huge capacity. So, PRMA is a radically new technology with many performance benefits, which is worth studying.

However, for LPWAN such as RPMA, the large density of base station, 2–3 km coverage distance and uneven traffic distribution [4] make the deployment of base stations difficult. Therefore, LPWAN network planning has a great challenge. It should be properly deployed and optimized to improve the network service quality according to its own characteristics. For LPWAN network planning, base station deployment determines the overall performance of the network. However, the determination of the

J. Zheng et al. (Eds.): ADHOCNETS 2018, LNICST 258, pp. 115–125, 2019.
https://doi.org/10.1007/978-3-030-05888-3_11

base station site is a NP-hard problem. It is not scientific to use the traditional location model to analyze the various factors of the site problem, which may lead to the dimension catastrophe of the variables and constraints in the model. In addition, for network planning, when the coverage is considered, the traffic distribution also need to be concerned, where spatio-temporal characteristics need to be addressed and integrated, which makes the problem more complex and designing a reasonable network planning scheme more important.

Now, a lot of researches have been done in network planning. Wang et al. [5] developed an approximation algorithm to address the budgeted base station planning problem in Het Nets, where they aimed to maximize the traffic demand points number with a given budget. Ghazzai et al. [6] proposed an optimal LTE wireless planning method to determine the minimum number and the optimal location of base stations under the constraints of cell coverage and capacity. For network planning, besides the number and location of base stations, energy efficiency is also an important target. So Yang et al. [7] aimed to establish a mathematical model to minimize power consumption for LTE cell planning. Wang et al. [8] employed a cutting-edge territory division to deal with the cell planning problem in Het Nets with the use of load balancing, based on the goal to guarantee users QoS and seamless coverage. The method can reduce the total deployment cost and improve the system performance. The above schemes are mainly for cellular network planning. For LPWAN, most of them survey on its technology, and no reasonable planning scheme has yet been proposed. What's more, the proposed network planning methods were based on a large number of assumptions, which have limitations in application. An effective planning method to quickly plan and deploy a large number of base stations has not fundamentally proposed.

To solve the above problems, big data analytics are combined with network planning in this paper, based on the application of LPWAN in communication system. The paper transforms the base station location problem from the traditional model-driven to data-driven, with massive data as the main analysis line, which overcomes the shortcomings of the traditional network planning model and combines the clustering algorithm to explore a data-driven base station location method so as to improve the level of rationalization of the site selection.

2 System Model

RPMA was design to provide a secure, large coverage footprint with tremendous capacity and low-power connectivity in the global 2.4 GHz band. It is the ideal technology to build a public network to connect many billions of devices for both Brown Field applications, and the even more exciting Green Field applications.

According to the characteristics of RPMA network, a novel network planning method based on data mining is presented in this paper, as shown in Fig. 1. Firstly, considering the coverage objective of network planning, the measured data of RPMA network are collected. Based on the network planning knowledge database, preliminary cleaning and analysis of measured data are required by removing attributes with many repeated and default values to improve the quality of the data, which make them more

suitable for specific data mining methods. Through the analysis, we need determine the characteristics that affect the quality of signal coverage and save the analysis results in the knowledge base. Then, with the goal of minimizing the loss function, a signal quality prediction model is trained by inputting the above data into the boosting regression trees model for predicting the network coverage under the current base station deployment. According to the obtained network coverage, by extracting the coverage weight value, we employ the weighted K-centroids clustering algorithm with the location data of base station and test points to achieve the base station deployment that adapts to the current coverage. Finally, we set the total planning objective function to decide whether it is the best base station topology.

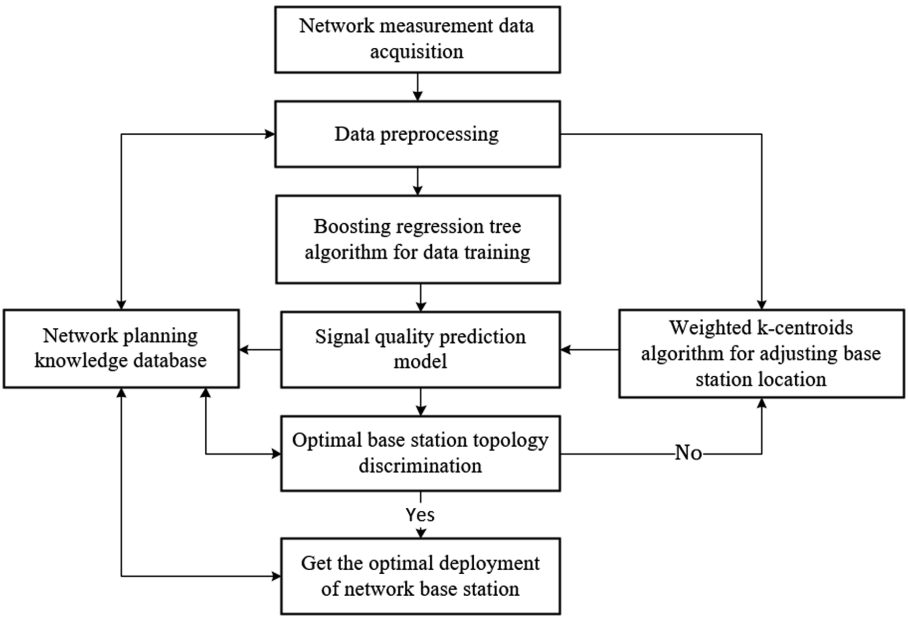

Fig. 1. Block diagram of network planning system based on data mining

3 Signal Quality Prediction Model

As shown in Fig. 1, we focus on the optimization of the coverage blind area and the weak coverage area, by analyzing the weak coverage problem in the wireless network, and adjust the location of base stations according to the network coverage, so that the adjusted base station topology can meet the required coverage effect. Generally, regional weak coverage is mainly caused by insufficient received signal strength, and the specific factors are involved in three aspects: (1) The factors affecting the coverage at base station, such as transmitting power, antenna azimuth, antenna height, antenna gain, etc.; (2) The factors of signal transmission path, such as path loss and shadow fading caused by obstruction; (3)The influence of interference on coverage, such as

co-channel interference in the overlapped areas of multiple adjacent base stations, and multipath interference caused by the surface reflections of buildings and mountains to the radio reflection.

All in all, the received signal quality at a certain location in the network is basically related to the three factors, which is the result of combining these factors. Therefore, we consider the mapping relationship between the signal quality and these factors in this paper, that is, to predict the quality of the signal, which is used to assist the final site location of base stations.

3.1 Data Feature Selection

First, we need to carry out preliminary cleaning and analysis of the data by removing the attributes with many repetitions and default values, such as UL Per, Network State and so on. For Deploy Region, latitude and longitude are considered to represent the location difference characteristics of base stations, so it can be eliminated. In addition, combined with the above coverage factors, the irrelevant attributes such as Last Connect Time and Last Connect Address are eliminated. Finally, base station location *B_loc* (including latitude and longitude), base station height *B_alt,* base station power *B_power*, antenna height *A_height*, and terminal location *P_loc* (including latitude and longitude) are selected as input features. The features are integrated into a record:

$$x_k = B\_loc_k, B\_alt_k, B\_power_k, A\_height_k, P\_loc_k \tag{1}$$

The set of these records serves as the training dataset for the signal quality prediction model. Since RPMA network employs power control, the uplink received signal strength is always near the reception sensitivity, so the terminal downlink received RSSI is used as an indicator for measuring signal quality, that is, the output variable. The process of establishing wireless network data model is to find the mapping function f between them by training existing data sets:

$$y_k = f(x_k) \tag{2}$$

The value y_k is the predicted signal quality value under given input characteristic variables x_k.

3.2 The Establishment of Signal Prediction Model

In this paper, we use the boosting regression trees algorithm [9] to construct the above function. Boosting Regression Trees (BRT) algorithm completes the learning task together by integrating multiple base learners–decision trees, which is one of ensemble Learning methods. Compared with the single regression algorithms, such as linear regression and logistic regression, BRT algorithm has better generalization performance by integrating multiple decision trees, thereby improving the prediction accuracy of the model. In addition, BRT algorithm can automatically fit the interaction of independent variables and is less prone to overfitting, so the generalization error is lower. The BRT model can be given by the addition model of M decision trees:

$$f_M(x) = \sum_{m=1}^{M} T(x; \gamma_m) \tag{3}$$

Each tree is given by:

$$T(x; \gamma) = \sum_{j=1}^{J} c_j I(x \in R_j) \tag{4}$$

where $\gamma = \{(R_1, c_1), (R_2, c_2), \ldots, (R_J, c_J)\}$ represent the divided areas $R_1, R_2, \ldots, R_J$ of each tree on the input variable set and the constants $c_1, c_2, \ldots, c_J$ on the corresponding area, J is the number of the leaf nodes of the decision tree.

BRT adopts the forward stepwise algorithm to learn each decision tree from the front to the back, that is, learning parameters of each tree by optimizing the following loss function:

$$\hat{\gamma} = \arg\min_{\gamma_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \gamma_m)) \tag{5}$$

In which the loss function is square error, which is the squared sum of the difference between the predicted value and the actual value of the sample:

$$L(y_i, f_{m-1}(x_i) + T(x_i; \gamma_m)) = (y_i - f_{m-1}(x_i) - T(x_i; \gamma_m))^2 \tag{6}$$

Where $e_m = y_i - f_{m-1}(x_i)$ means the residuals of the data fitted by the current model. Therefore, BRT algorithm is used to solve the regression problem, which only needs to fit the residual of each model. The specific algorithm process is shown as follows:

(1) Set $f_0(x) = 0$.
(2) For m = 1,2,…,M:
   a. Calculate the residuals of the current model:

$$e_m = y_i - f_{m-1}(x_i), i = 1, 2, \ldots, M$$

   b. Fit the residuals to learn a regression tree $T(x; \gamma_m)$ and update

$$f_m(x) = f_{m-1}(x) + T(x; \gamma_m)$$

(3) Get the BRT model of the problem:

$$f_M(x) = \sum_{m=1}^{M} T(x; \gamma_m)$$

4 Position Adjustment of Base Station

Typical K-means algorithm partitions the data set $X = \{x_1, x_2 \ldots, x_n\}$ with n points into K clusters according to the set distance similarity, and the cluster set is expressed by $C = \{c_1, c_2, \ldots, c_k\}$. Generally, the Euclidean distance is used as a similarity measure between two points, and data points are divided into the nearest clusters.

In typical K-means algorithm, each data point has same importance for locating the location of the cluster center. However, we treat the base station position selection as a weighted problem based on coverage distribution spatial patterns in this paper, which means that each point in the space no longer has an equivalent impact on cluster centers. A weight is introduced to measure influence degree of each point on the base station position, and a weighted K-centroids algorithm is proposed.

The input of this algorithm includes n terminal data point set $P = \{p_1, p_2, \ldots, p_n\}$ and initial base station positions $B = \{b_1, b_2, \ldots, b_k\}$. Planning network based on the existing base station site, the current sites and the number of base stations can be used as the initialization parameter of the algorithm, that is, the initialization centers and the number of clusters. In this algorithm, normalized distance is used to determine which cluster the data points belong to, which is called the membership function [10]:

$$f(b_j|p_i) = \frac{\|p_i - b_j\|^2}{\sum_{j=1}^{k} \|p_i - b_j\|^2} \tag{7}$$

After all data points are assigned, the location of base station is adjusted iteratively in the algorithm, which is mainly considered from distance influence and coverage weight. For distance influence, compared to the terminals close to the base station with smaller distance influence, the terminal signals far from the base station may be worse, which is due to obstructions from buildings and path loss of signal propagation, so forane terminals have a greater distance influence on base station. We employ the above membership function $f(b_j|p_i)$ to measure the distance influence. For coverage weight, the optimization of base station location aims to ensure that the received signal of the terminal within the coverage of the base station can be as good as possible, so we are concerned with the terminal with poor coverage and give it a greater impact weight on the location adjustment of the base station. According to the coverage spatial pattern obtained in the previous stage, a corresponding weight $w(p_i)$ is generated for each data point. With $f(b_j|p_i)$ and $w(p_i)$, the iterative formula for each base station location is given by:

$$b_j = \frac{\sum_{i=1}^{n} f(b_j|p_i) w(p_i) p_i}{\sum_{i=1}^{n} f(b_j|p_i) w(p_i)} \tag{8}$$

The weighted K-centroids algorithm process is shown as follows:

(1) Use the location and number of existing base stations as the initial cluster center locations and cluster number;
(2) By membership function $f(b_j|p_i)$, each data point p_i is assigned to the cluster where its nearest base station b_j is located;
(3) Adjust each base station location b_j with the membership function $f(b_j|p_i)$ and spatial pattern weight $w(p_i)$;
(4) Repeat steps (2) and (3) until b_j no longer changes.

The network topology obtained by the weighted K-centroids algorithm has been optimized for the current network coverage, but it is not necessarily the final optimal result. It still needs to carry on coverage prediction analysis and optimize base station positions again based on the analysis results. Until the following total objective function is met, an optimal network topology is finally obtained.

The total objective function of the entire planning process:

$$\min \sum_{i \in \{i|y_i \le \overline{y}\}} (y_i - \overline{y})^2 \tag{9}$$

Where y_i represents the coverage strength RSSI predicted by the BRT algorithm at some point, $\overline{y}$ is a theoretically good signal coverage threshold to meet the coverage standard, and $i \in \{i|y_i \le \overline{y}\}$ represent test points with signal quality values below the threshold in the area, which means that the signal coverage of the test point is poor. The least square error of the two is used as the objective function for the iteration termination of the entire planning process.

5 Simulation Analysis and Performance Evaluation

The experimental data is derived from the real measured data of 37 RPMA network and drive test data of 131454 test points after data cleaning, including base station basic information data, terminal test point data and corresponding geographic location data. The data is used to verify the feasibility of the proposed method for optimal base station deployment, and the experimental results are visible with python matplotlib tools.

5.1 Result and Analysis of Signal Prediction Model

Before applying the BRT algorithm, three parameters need to be determined to adjust its learning process. The first is the number of base learners. With the increase of its number, the BRT algorithm on training data may be improved. However, the number of base learners exceeds a certain value, which may cause over-fitting. The second is the size of the base learner, which represents the degree of interaction between multiple features captured by the BRT model, and the depth of the tree is used to control the size of the base learner. For the selection of the two parameters, GridSearchCV grid tracking method in sklearn is used in this paper, which can traverse multiple

combinations of parameter values that need to be optimized according to the given data set through cross validation until the optimum parameters are obtained. The number of base learners is 530, and the depth of trees is 11 in this paper. Finally, in order to prevent over-fitting on training data, the regularization factor (i.e., learning rate) is introduced to measure the impact of each base learner on the final result. This value is set to a smaller constant below 0.1, which is set to 0.1 in this paper.

Then, 85% of the dataset is selected as the training data set, and 15% is the test data set. The horizontal axis in Fig. 2 represents the number of iterations (i.e., the number of basic learners), the vertical axis represents loss error values, and two lines represent test errors and training errors of each iteration respectively. It is found that the training error and the test error are gradually decrease with the increase of the number of iterations, which indicates that the fitting effect on the data sets increases gradually with the increase of the number of iterations. The test shows that the test error is higher than the training error due to the difference between the test set and the training set, which makes the learning ability of the model on the unknown data set weaker than the original training data set and is a normal phenomenon. In addition, the trend of the two curves also indicates that the parameters obtained by GridSearchCV are appropriate.

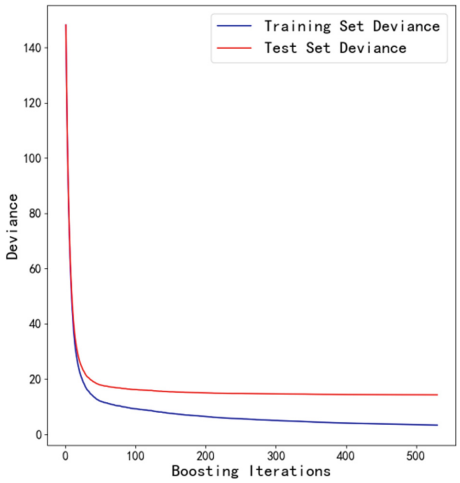

Fig. 2. Relationship between loss error and number of base learners

5.2 Determination of Optimal Base Station Deployment

The left of Fig. 3 shows the collected initial base station locations and test point distributions. The base station location is marked with a blue star point, and the dot marks the test point location. The color depth of the dot represents the RSSI value. The darker the red is, the lower the RSSI value is, and the worse the signal coverage is. It shows that there are still many weak coverage areas in the initial base station deployment. The RSSI unit in the Fig. 3 is dBm.

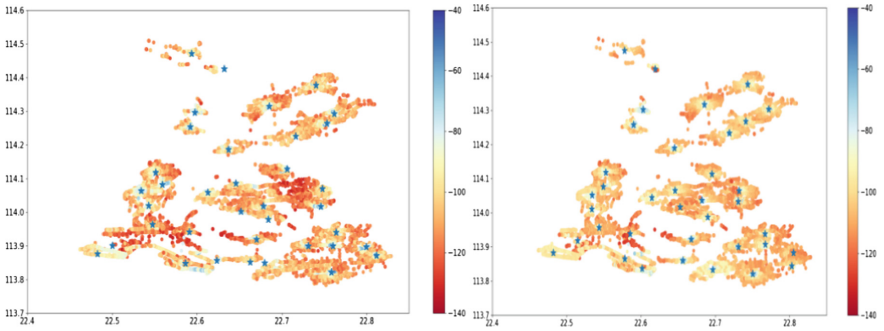

Fig. 3. The left is initial base station locations and change of RSSI value of test points, the right is final base station locations and change of RSSI value of test points (Color figure online)

In this paper, base stations are deployed based on the coverage distribution spatial pattern, and RSSI is used to measure the weight of the coverage strength. Combined with the weighted K-centroids algorithm, the base station location is determined, which is judged by the total objective function. Table 1 shows the total objective function value of each planning iteration, which means that the smaller the value, the better the coverage quality. We can see that the total objective function value gradually decreases with the increase of the number of iterations in the table. That is, the signal coverage gradually improves until the termination of the 10th iteration, and the minimum value of the total objective function is 535.41. The right of Fig. 3 shows the final base station locations and the change of RSSI value of test points, and the dark red region is less than that in the left of Fig. 3, which means that employing the proposed planning method helps improve signal coverage. The right is the corresponding clustering result, the location of base station is marked with a black star point and the location of test point is marked with a dot. Dots with the same color belong to the same base station cluster nearest to them.

Table 1. Total objective function values for each iteration

Iterations	Total objective function values
1	640.93
2	581.79
3	554.43
4	548.89
5	548.66
6	547.08
7	546.87
8	539.41
9	535.54
10	535.41

In order to verify the superiority of the proposed method, we compare the proposed method with the K-means based optimization method [11]. By calculating the total objective function value of each iteration, its iteration result tends to 584.22. Therefore, the method proposed in this paper can better improve the signal coverage rate compared with the K-means based optimization method.

6 Conclusion

In this paper, we have proposed a network planning method based on data mining. First, the overall network is preliminarily analyzed by using the measured data, and the features of coverage quality are selected. Then, the BRT algorithm and K-centroids algorithm are employed to extract the coverage distribution spatial pattern of the network, and the optimal RPMA network base station deployment is obtained. Finally, the feasibility of the proposed method is verified by using the measured data. Compared with conventional K-means based optimization method, this method can improve the coverage quality of LPWAN well, and has a certain reference value for the network planning.

In actual network planning, the base station deployment needs to consider many factors, and we only consider the coverage objective of the network planning in this paper. Therefore, in future work we will introduce capacity objective, and optimize the base station deployment combined with the two objectives, making the network planning more perfect.

Acknowledgements. This work was supported by National Science & Technology Key Project of China (2017ZX03001008), Natural Science Foundation of China (61871237), Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX17_0766) and Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (16KJA510005).

References

1. Patel, D., Won, M. Experimental study on low power wide area networks (LPWAN) for mobile internet of things. In: IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, NSW, pp. 1–5 (2017)
2. Hernandez, D.M, Peralta, G., Manero, L., et al.: Energy and coverage study of LPWAN schemes for industry 4.0. In: 2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM), Donostia-San Sebastian, pp. 1–6 (2017)
3. Xiong, X., Zheng, K., Xu, R., Xiang, W., Chatzimisios, P.: Low power wide area machine-to-machine networks: key techniques and prototype. Commun. Mag. IEEE **53**(9), 64–71 (2015)
4. Krupka, L., Vojtech, L., Neruda, M.: The issue of LPWAN technology coexistence in IoT environment. In: 2016 17th International Conference on Mechatronics - Mechatronika (ME), Prague, pp. 1–8 (2016)
5. Wang, S., Zhao, W., Wang, C.: Budgeted cell planning for cellular networks with small cells. IEEE Trans. Veh. Technol. **64**(10), 4797–4806 (2015)

6. Ghazzai, H., Yaacoub, E., Alouini, M.S., et al.: Optimized LTE cell planning with varying spatial and temporal user densities. IEEE Trans. Veh. Technol. **65**(3), 1575–1589 (2016)
7. Yang, Z.H., Chen, M., Wen, Y.P., et al.: Cell Planning based on minimized power consumption for lte networks. In: IEEE Wireless Communications and NETWORKING Conference. IEEE (2016)
8. Wang, S., Ran, C.: Rethinking cellular network planning and optimization. IEEE Wirel. Commun. **23**(2), 118–125 (2016)
9. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29** (5), 1189–1232 (2001)
10. Wen, R., Yan, W., Zhang, A.N.: Weighted clustering of spatial pattern for optimal logistics hub deployment. In: 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, pp. 3792–3797 (2016)
11. Kanungo, T., Mount, D.M., Netanyahu, N.S., et al.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 881–892 (2002)