A Machine Learning Based Temporary Base Station (BS) Placement Scheme in Booming Customers Circumstance

Qinglong Dai[1](✉), Li Zhu[2,3], Peng Wang[1], Guodong Li[1], and Jianjun Chen[1]

[1] China Academy of Electronics and Information Technology, Beijing 100041, China
qldaisd@126.com

[2] China Transport Telecommunications and Information Center, Beijing 100011, China
jolie.zhl@hotmail.com

[3] People's Public Security University of China, Beijing 100038, China

Abstract. Explosive increase of terminal users and the amount of data traffic give a great challenge for Internet service providers (ISPs). At the same time, this big data also brings an opportunity for ISPs. How to solve network planning problem in emergency or clogging situation, based on big data? In this paper, we try to realize effective and flexible temporary base station (BS) placement through machine learning in a booming customers situation, with ISPs' massive data. A machine learning based temporary BS placement scheme is presented. A K-means based model training algorithm is put forward, as a vital part of machine learning based temporary BS placement scheme. K-means algorithm is selected as a representative example of machine learning algorithm. The performances of BS position with random starting point, BS position iteration, average path length with different parameters, are conducted to prove the availability of our work.

Keywords: Network planning · Machine learning · Big data
K-means algorithm

1 Introduction

Nowadays, with the explosive increase of terminal users and the amount of data traffic, Internet service providers (ISPs) seems not equal to this trend. In China, until Dec 2017, the number of Internet users via cellphone is 753 million, accounting for 97.5% in Internet users. In 2017, the amount of data consumed by Chinese Internet users is 21.21 billion GB [2]. As for 2017, according to ITU-T's statistics, the number of cellphone in the world is 7.74 billion [1]. There will be a promis-

Supported by key special project of National Key Research and Development Program (Grant No. 2017YFC0803900)

X. Liu et al. (Eds.): ChinaCom 2018, LNICST 262, pp. 562–572, 2019.
https://doi.org/10.1007/978-3-030-06161-6_55

ing future for mobile internet usage, as global mobile data traffic is projected to increase nearly sevenfold between 2016 and 2021. In 2016, global mobile data traffic amounted to 84 EB [3]. With the development of Internet of things, industry 4.0 and Internet of vehicles, a mass of sensors are applied in varied scenarios [7,16]. The communications between these sensors and backup data centers may take up a large amount of bandwidth. These make ISPs' situation worse.

Everything has two sides. Increasing terminals and data traffic also bring a huge opportunity for ISPs, because of the emergence of big data. Big data encompasses unstructured, semi-structured and structured data and represents the information assets characterized by so-called 4 V, i.e., high volume, high variety, high velocity and low veracity [10]. Based on the numerous terminals and varied services, ISPs could acquire generous customer and service data. From these considerable data, some hidden knowledge or regular patterns about users, like shopping habit, preference, trip information and etc., can be obtained via a certain method. The common method that is used to extract knowledge or pattern from big data is machine learning.

Machine learning is widely known as a technique of artificial intelligence [4]. Machine learning is a field of computer science that gives computer systems the ability to learn (i.e. progressively improve performance on a specific task) with data, without being explicitly programmed [18]. With the increase of data scale, the complexity of machine learning algorithm becomes the important factor that should be considered. The application of machine learning based on big data would be more effective, compared with that based on traditional data. The specific algorithms or methods in machine learning usually come from many diverse existing fields, including statistics, probability theory, neural network and etc. Machine learning can be applied in many areas such as speech recognition, image processing and fraud detection [5].

Network planning is always a significant direction in communication area. Simply, network planning determines where to place the base stations (BSs) and how to connect them. Through BSs, users' terminals including cellphone, laptop, tablet PC and so on, can access to Internet service. BSs are normally hexagonal honeycomb distributed. However, in some special situations with booming customers in a brief time, such as concert or football match, extra surge terminal access and service data exceed the capability of ordinary deployed BSs. Temporary BSs are necessary. In traditional network planning, temporary BSs placement is on the basis of experience and estimation. With the help of machine learning, temporary BSs placement would be more effective and flexible.

The contributions of this paper are listed as follows:

1. A machine learning based temporary BS placement scheme is presented, in which ISPs' operation data about users is exploited as dataset. Proposed temporary BS placement scheme scheme is classified into two phases: training phase and verification phase. Each component in this scheme is also explained in detail.
2. A K-means based model training algorithm is put forward, as a vital part of machine learning based temporary BS placement scheme. K-means algorithm

is selected as a representative example of machine learning algorithm, because of its low complexity.

3. Numerical simulation is conducted. Typical performances including BS position with random starting point, BS position iteration, average path length with different parameters, are demonstrated, to prove the availability of proposed scheme.

The rest of this paper is organized as follows. Section 2 reviews related work on machine learning. Section 3 proposes a machine learning based temporary BS placement scheme. A K-means based model training algorithm is put forward in Sect. 4. Section 5 represents the numerical simulation. Finally, Sect. 6 concludes the paper.

2 Related Work

Around machine learning, a great number of works have been done.

Gregory Piatetsky-Shapiro, the co-founder of knowledge discovery and data mining conferences, deemed that data science or data mining, big data and machine learning were related. The relationship of them is shown in Fig.1. Any one of data science, machine learning and clustering algorithm had overlap regions with others. And clustering algorithm could be seen as a part of machine learning [19].

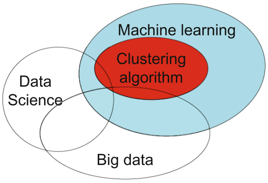


Fig. 1. The relationship of data science, big data, machine learning and clustering algorithm

Mohammadi et al. focused on the challenge of the big data generated by smart cities from a machine learning perspective. In their framework, machine learning was used to meet the cognitive services of smart city, based on the big data that generated by smart city sensors [15].

Most machine learning algorithms are categorized into supervised, unsupervised and reinforcement learning. Supervised learning deals with a labeled dataset, to build the relationship of input, output and other parameters. Unsupervised learning is provided with an un-labeled dataset, to classify data into different groups or clusters. In reinforcement learning, the agent learns by interacting with environment [5,9].

To detect mobile botnets and minimize the threat, a machine learning algorithm, exactly classification algorithm based approach was presented to identify anomalous behaviors in statistical features extracted from system calls by da Costa et al. [6]. The core part of their approach was a classifier. This classifier was responsible to classify normal or mobile botnet activity.

Li et al. extensively applied supervised machine learning in spam emails classification and evaluated different classifiers in three environments with 1000 users. Some supervised machine learning classifiers such as decision tree and support vector machine (SVM) were acceptable in real emails classification [11].

Liu et al. adopted deep learning to investigate the latent relationship between flow information and link usage, and then classified the used and unused links. Through this, the scale of network optimization problems was reduced [13]. Similarly, Rottondi et al. also introduced machine learning into the quality of transmission, as a novel way to achieve pre-deployment estimation [17].

As for video service accounting for large wireless traffic, Lin et al. applied supervised machine learning and SVM to forecast video starvation events, like the number of users existing in cell. They demonstrated the correlation of video starvation and recorded users' features for streaming with diverse characteristics [12].

Dargie et al. came up with a feasible way to locate wireless sensor network node in a 3D environment, using supervised neural networks with four input measurements, i.e., signal strength indicator, time of arrival, time difference of arrival and angle [8].

3 A Machine Learning Based BS Placement Scheme

In order to place temporary BSs more effectively, avoiding the uncertainty involved by experience and estimation, clustering algorithm is selected as a feasible method. Clustering algorithm is one of the most common machine learning applications. In clustering algorithm, the typical algorithm is K-means algorithm.

The K-means algorithm is used to recognize data with no labels into different classes. In here, no labels means that there is no output vector or non-judgmental property, like *yes* or *no*, *good* or *bad*. Each class is called as a cluster. The data in the same cluster have high similarity, meanwhile the data in different clusters have low similarity. Due to K-means algorithm's linear complexity, simple implementation and the data with no labels that collected by ISPs, it is selected by us.

We give up a machine learning based temporary BS placement scheme, as shown in Fig. 2. There are two phases in this scheme: training phase and verification phase. In training phase, based on a portion of the data from ISPs, all the terminals are divided into several different clusters. This division process is called model. To verify the rationality of this model, the other portion of the data from ISPs, is used to input the model. If the model output of verification data accords with already obtained result, the model is rational. Otherwise, the model is unqualified. And a feedback new model should be generated based on new data.

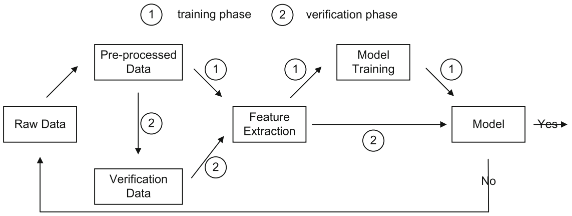


Fig. 2. A machine learning based temporary BS placement scheme

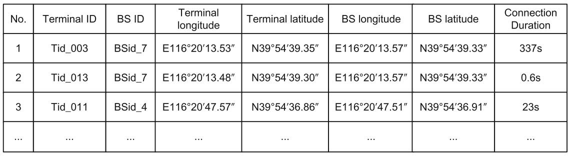
The components in our proposed scheme are interpreted as follows:

1. Raw data. The original data from ISPs' database. Raw data is generated by users requesting ISPs service. Under the large volume and chaotic data, the habit and preference of users can be gained. Because of the diversity of services and the existence of over the top (OTT) services from different companies, there are a lot of redundant, repeated, default and blank fields in raw data. Therefore, pre-process for raw data is necessary.
2. Pre-processed data. Through pre-process, raw data becomes pre-processed data. The redundant field is simplified. The repeated field is deleted. The default field is filled up with default value. The blank field is replenished by random value. Pre-processed data is split into two parts. One part of pre-processed data is used for model training, the others is used for the rationality verification of trained model.
3. Feature extraction. In machine learning, the scale of data is too large to be processed and it is suspected to be redundant (e.g., the same record in different sheets), then the data can be transformed into a reduced set of feature or a featured vector. Feature extraction facilitates the subsequent model training.
4. Model training. By import extracted data record successively, a general pattern is optimized in a iterated way, by the manner of confirming different parameters' coefficients.
5. Model. The model is the final result of model training, based on the aforementioned part of pre-processed data that is used in training phase. Note that the model is relaying on model training input data. In other words, the model is limited. As it were, all models are wrong, but some are useful.
6. Verification data. Verification data is used to verify the rationality of trained model. If the number of inapplicable data for trained model is inferior to a pre-set threshold value, the trained model is the final model; Otherwise, the trained model is unqualified. A brand new raw data for model retraining is necessary.

Note that the data pre-process is strongly related to machine learning purpose. A sample sheet in pre-processed data is shown in Fig. 3. For different machine learning purposes, some fields in sheet can be deleted. For instance, for a machine learning predicting users' connection duration, the information of terminal position is not needed.

Algorithm 1 K-means based Model Training Algorithm

```
Input:
    Dataset of previous terminal communications, $\mathbf{A}$, in which the $j$th entry is denoted as $a_j$;
    $K$ initial temporary BS location: $(lon_1,lat_1)$, $(lon_2,lat_2),\cdots,(lon_i,lat_i),\cdots,(lon_K,lat_K)$;
    Maximum distance of connection establishment, $D$;
Output:
for each entry in dataset $\mathbf{A}$, i.e., $Terminal_j$ do
    for each initial BS location, i.e. $BS_i$ do
        Compute the distance between $Terminal_j$ and $BS_i$, i.e., $D_{ij}$;
        if $D_{ij} \leq D$ then
            Replace $(lon_i,\ lat_i)$ with $((lon_i + lon_j)/2,\ (lat_i + lat_j)/2)$;
        else
            Continue;
        end if
    end for
end for
return $K$ final BS location, i.e., $(lon_1,lat_1)$, $(lon_2,lat_2),\ \cdots,(lon_i,lat_i),\cdots,(lon_K,lat_K)$;
```



No.	Terminal ID	BS ID	Terminal longitude	Terminal latitude	BS longitude	BS latitude	Connection Duration
1	Tid_003	BSid_7	E116°20′13.53″	N39°54′39.35″	E116°20′13.57″	N39°54′39.33″	337s
2	Tid_013	BSid_7	E116°20′13.48″	N39°54′39.30″	E116°20′13.57″	N39°54′39.33″	0.6s
3	Tid_011	BSid_4	E116°20′47.57″	N39°54′36.86″	E116°20′47.51″	N39°54′36.91″	23s
...	...	...	...	...	...	...	...

Fig. 3. A sample sheet in pre-processed data

4 K-means based Model Training Algorithm

In this section we select a typical clustering algorithm, i.e., K-means algorithm, to accomplish temporary BS placement, i.e., the model training in our proposed scheme. For a booming customers circumstance, a targeted model training algorithm is raised in Algorithm 1.

Model training algorithm is started, based on the dataset of previous terminal communications, initial temporary BSs locations and pre-set maximum distance of connection establishment. Each entry in the dataset is applied to training model. K initial temporary BSs locations are seen as the initial centroids for different clusters. Compute the distance between each terminal and each BS. Compare each time computation result with pre-set maximum distance of connection establishment. If computation result is smaller, replace corresponding BS location with the middle point location of this BS location and terminal location. Otherwise, continue the computation of next pair of BS and terminal. Finally, the ultimate BSs locations are training result.

In our proposed scheme, the location is recoded by geographic coordinate system. Therefore, for distance comparison conveniently, geographic coordinate has to be transformed into distance, according to Eq. 1. In Eq. 1, R is the radius of the earth, lat is short for latitude, lon is short for longitude, sin is sine function, cos is cosine function and $arcsin$ is arc-sin function.

$$D_{ij} = 2R * \arcsin\left(\sqrt{sin^2\left(\frac{lat_i - lat_j}{2}\right) + cos\left(lat_i\right) * cos\left(lat_j\right) * sin^2\left(\frac{lon_i - lon_j}{2}\right)}\right) \tag{1}$$

5 Numerical Results

The performances of proposed BS placement scheme are analyzed via simulation using *MATLAB*. The BSs placement in [14] is selected as a baseline method. With varied parameters, the performances of proposed BS placement scheme are depicted in detail.

In our simulation, unless explicitly stated, otherwise, the simulation parameters shown in Table 1 are used. The network area of terminals and BSs is a 100 km×100 km area. The terminal number in this simulation is 120. And these terminals are randomly distributed in network area. These are all demonstrated in Fig. 4. Based on these conditions and parameters, a series of evaluations are performed in different scenarios.

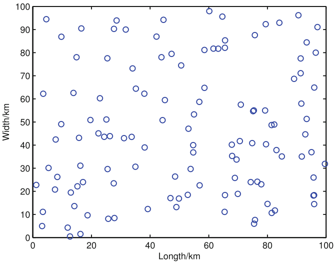
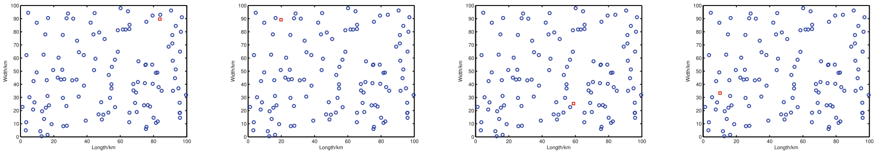


Fig. 4. User distribution in 100 km× 100 km area

Figure 5 is the temporary BS placement with different random starting points, namely, (100, 0), (100, 100), (0, 100) and (0, 0). The red rectangle in each sub-fig is the final BS placement. The maximum distance of connection establishment is 20 km. It is obviously that the temporary BS placement has a strongly correlation with starting point. Normally, from different starting points, different temporary BS placements are obtained. However, every result of different starting point is the best BS placement for the terminals within its coverage



(a) SP:(100,100)
(b) SP:(0,100)
(c) SP:(100,0)
(d) SP:(0,0)

Fig. 5. Base position with random starting point (SP)

with distance D. Therefore, to cover all the terminals, when D is fixed, increased starting points is a feasible solution.

In the process of BS placement, the BS position is varied after each iteration. The BS position iteration with starting point (0, 100) is shown in Fig. 6. Each rectangle in Fig. 6 is the result of a position re-calculation. Through N position changes, BS position moves from (0, 100) to final position. According to Algorithm 1, once there is one or more terminal emerging in BS's coverage, position re-calculation happens, until all the terminals in network area are considered. These lead to the BS movement.

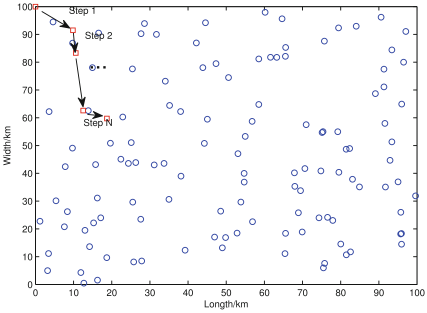


Fig. 6. BS position iteration

Average path length is negatively related to communication distance, D, as shown in Fig. 7. The number of added temporary BS is 1. With the increase of communication distance, average path length decreases, no matter proposed algorithm or baseline method. Greater communication distance means that more terminals can connected to a BS, becoming a cluster. Similarly, different clusters are linked together, like the way of terminals connecting to BS. The greater communication distance, the less cluster, and then the smaller average path length. Besides, no matter which one the starting point is, the average path lengths of baseline method with different starting points are almost the same. They are greater than the average path length of proposed algorithm all the time.

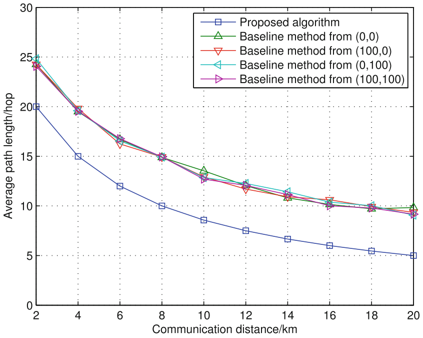


Fig. 7. World Map

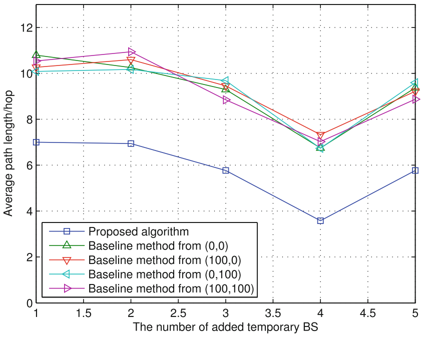


Fig. 8. APL vs. communication distance

In Fig. 8, the relationship of average path length and the number of added temporary BS is demonstrated. The distance of connection establishment is set as 10 km, using the average distance. Average path length of proposed algorithm is always smaller than those of baseline methods with different starting points. When the number of added temporary BS grows, average path length firstly decreases to a minimum value, and then increases. The connections among different clusters are considered in average path length computation. More added temporary BSs means more clusters. More clusters make an extra burden for average path length, if added temporary BSs number beyond a specific value. In this specific simulation for Fig. 8, the appropriate value of added temporary BS is 4.

6 Conclusion

With explosive increase of terminal users and the amount of data traffic, in order to place temporary BS effectively and flexibly in a booming customers situation. Based on ISPs collected big data, a machine learning based temporary BS placement scheme was presented by us. A K-means based model training algorithm was also put forward. K-means algorithm was selected as a representative example of machine learning algorithm. Through simulation, propose scheme was superior to compared work. In our future work, service-oriented machine learning application and the complexity analysis of different machine learning algorithms will be key emphases.

References

1. Key ict indicators for developed and developing countries and the world (totals and penetration rates). Technical report, ITU-T, Geneva, Switzerland (2017)
2. The 41st china statistical report on internet development. Technical report, China Internet Network Information Center, Beijing, China (2018)
3. Mobile internet usage worldwide - statistics and facts. Technical report, Statista, Hamburg, Germany (2018)
4. Alpaydin, E.: Introduction to Machine Learning, 2nd edn. The MIT Press, USA (2009)
5. Alsheikh, M.A., Lin, S., Niyato, D., Tan, H.P.: Machine learning in wireless sensor networks: algorithms, strategies, and applications. IEEE Commun. Surv. Tutor. **16**(4), 1996–2018 (2014). https://doi.org/10.1109/OMST.2014.2320099. Fourthquarter
6. da Costa, V.G.T., Barbon, S., Miani, R.S., Rodrigues, J.J.P.C., Zarpelao, B.B.: Detecting mobile botnets through machine learning and system calls analysis. In: 2017 IEEE International Conference on Communications (ICC). pp. 1–6 (2017). https://doi.org/10.1109/ICC.2017.7997390
7. Curry, E., Hasan, S., Kouroupetroglou, C., Fabritius, W., ul Hassan, U., Derguech, W.: Internet of things enhanced user experience for smart water and energy management. IEEE Internet Comput. **22**(1), 18–28 (2018). https://doi.org/10.1109/MIC.2018.011581514
8. Dargie, W., Poellabauer, C.: Localization. Wiley, New York (2010)
9. Klaine, P.V., Imran, M.A., Onireti, O., Souza, R.D.: A survey of machine learning techniques applied to self-organizing cellular networks. IEEE Commun. Surv. Tutor. **19**(4), 2392–2431 (2017). https://doi.org/10.1109/COMST.2017.2727878. Fourthquarter
10. LHeureux, A., Grolinger, K., Elyamany, H.F., Capretz, M.A.M.: Challenges and approaches. IEEE Access **5**, 7776–7797 (2017). https://doi.org/10.1109/ACCESS.2017.2696365
11. Li, W., Meng, W.: An empirical study on email classification using supervised machine learning in real environments. In: 2015 IEEE International Conference on Communications (ICC), pp. 7438–7443 (2015). https://doi.org/10.1109/ICC.2015.7249515
12. Lin, Y.T., Oliveira, E.M.R., Jemaa, S.B., Elayoubi, S.E.: Machine learning for predicting qoe of video streaming in mobile networks. In: 2017 IEEE International Conference on Communications (ICC), pp. 1–6 (2017). https://doi.org/10.1109/CC.2017.7996604
13. Liu, L., Cheng, Y., Cai, L., Zhou, S., Niu, Z.: Deep learning based optimization in wireless network. In: 2017 IEEE International Conference on Communications (ICC), pp. 1–6 (2017). https://doi.org/10.1109/ICC.2017.7996587
14. Liu, Y., Zhou, C., Cheng, Y.: Integrated bs/onu placement in hybrid epon-wimax access networks. In: GLOBECOM 2009–2009 IEEE Global Telecommunications Conference, pp. 1–6 (2009). https://doi.org/10.1109/GLOCOM.2009.5425770
15. Mohammadi, M., Al-Fuqaha, A.: Enabling cognitive smart cities using big data and machine learning: Approaches and challenges. IEEE Commun. Mag. **56**(2), 94–101 (2018). https://doi.org/10.1109/MCOM.2018.1700298
16. Nguyen, T.T.T., Armitage, G.: A survey of techniques for internet traffic classification using machine learning. IEEE Commun. Surv. Tutor. **10**(4), 56–76 (2008). https://doi.org/10.1109/SURV.2008.080406. Fourth

17. Rottondi, C., Barletta, L., Giusti, A., Tornatore, M.: Machine-learning method for quality of transmission prediction of unestablished lightpaths. IEEE/OSA J. Opt. Commun. Netw. **10**(2), A286–A297 (2018). https://doi.org/10.1364/JOCN.10.00A286
18. Samuel, A.L.: Some studies in machine learning using the game of checkers. Ibm J. Res. Dev. **3**(3), 210–229 (1959)
19. Taylor, D.: Battle of the data science venn diagrams (2016)