# Telecom Big Data Based User Analysis and Application in Telecom Industry

Guanglu Shao, Weiwei Chen, Xinzhou Cheng, Lexi Xu(✉),
Tao Zhang, and Chen Cheng

China United Network Technology Corporation,
Beijing, People's Republic of China
xulx29@chinaunicom.cn

**Abstract.** Due to the fast progress of mobile internet and smart phone, telecom operators store massive telecom big data. In this paper, we utilize telecom big data for the user analysis and application. Initially, this paper studies the content of telecom big data. Then, 360-degree user portrait is drawn via three types of factors, including user's information, user's consumption behavior and service behavior. Based on the user portrait, this paper seeks 4G potential users via data mining technique, which includes the business understanding, data preparation, model building, model training, model evaluation and model deployment steps. Based on the 4G potential users, the marketing department can accelerate the transferring process from 2G/3G users to 4G users.

**Keywords:** Telecom big data · 4G · Telecom operator · Data mining
User portrait

## 1 Introduction

The telecom industry experiences fast progress in the past decade [1, 2]. The dramatically increased network capacity and transmission rate allow a large number of users to get access to networks [3, 4]. The telecom industries generate abundant information, which is called telecom big data [5]. It's generally known that telecom big data brings great opportunity for telecom operators [6, 7].

Recently, telecom operators utilize telecom big data to generate value and bring profits, the value chain of telecom big data includes four stages, as shown in Fig. 1.

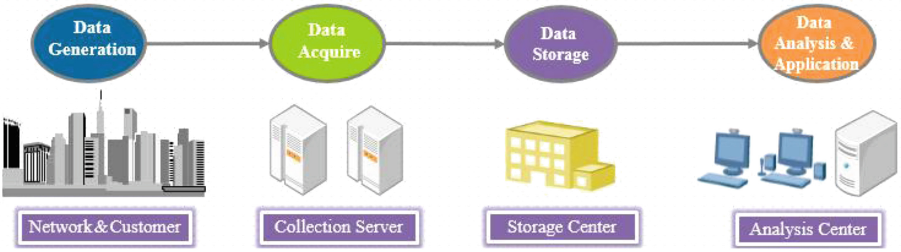


**Fig. 1.** Value chain of telecom big data

K. Long et al. (Eds.): 5GWN 2017, LNICST 211, pp. 99–109, 2018.
https://doi.org/10.1007/978-3-319-72823-0_11

The first stage is the data generation stage since customers generate abundant data during the telecom services [8]. As the channel to bear these services, networks also generate massive signaling data and service data [9, 10]. According to the data source, telecom big data includes BSS data and OSS data. BSS data includes customer personal information, voice billing details, data billing details, monthly billing etc. OSS data includes measurement report (MR), counter, performance indicator, call detailed record (CDR), engineering parameters etc. [11, 12].

The second stage is the data collection stage. The widely used method is telecom operators set acquisition equipment in the network interface, thus collecting the data stream, which pass the network interface. Typical network interfaces include A and Abis interface in the switching network, Gb and Gn interface in GPRS, lu-PS and lu-CS interface in 3G, S1 and X2 interface in LTE [13, 14].

After collecting the data stream, in the third data storage stage, telecom operators can resolve data stream and form new-structure data, which suit for storage. Then, telecom operators store these data. Since telecom big data has the characteristics of large volume, effective storage is vital and cloud mechanism is a promising technique to store telecom big data in the future [15].

The key of big data era is to seek the data value in the data analysis and application stage [16, 17]. On one hand, telecom big data can be utilized in the telecom industry, for example, telecom big data based network planning and optimisation, telecom big data assisted market strategy, value-added service, telecom big data analysis of network quality and performance [18, 19]. On the other hand, telecom big data can be also utilized for the relevant industries, for example, advertisement, smart finance, transport, epidemic control, human flow control, public sentiment monitoring etc. [20].

In order to study the telecom big data application in the telecom industry, this paper analyses the content of telecom big data and then draw the 360-degree user portrait, including the characteristics, the resident area, the mainstream services and other preferences of users. According to the user information and behavior, this paper uses data mining technique to seek 4G potential users for telecom operators.

This paper is organised as follows: Sect. 2 describes the user portrait. Section 3 introduces the data mining process and presents 4G potential users analysis in details. Conclusions are given in Sect. 4.

## 2 360 Degree User Portrait Analysis

Section 2 introduces the telecom big data and extracts user related information to draw the user portrait. According to the data source, there are twenty-three types of telecom big data, including seventeen types of OSS domain data and six types of BSS data [21]. The details of telecom big data are list in Table 1.

In this section, we extract user related information from above 23 types of telecom big data. Then, we get three types of factors to draw the 360-degree portrait for each user, as shown in Fig. 2.

The first type of factors is the user's information. It includes a series of static attributes, including the age, the gender, the member level, the network type, the user state etc. [15, 22].

**Table 1.** Telecom big data source

| Index | Category | Data source |
|---|---|---|
| 1 | OSS domain | Engineering parameters |
| 2 | | DT/CQT data |
| 3 | | IU-PS/Gn/S1 interface data |
| 4 | | CDR (call detailed record) data |
| 5 | | MR (measurement report) data |
| 6 | | Wireless traffic statistical data |
| 7 | | Wireless parameters |
| 8 | | Core network data |
| 9 | | Alarm data |
| 10 | | Wireless call trace data |
| 11 | | Equipment version and patch |
| 12 | | Equipment load data |
| 13 | | Wireless counter data |
| 14 | | Core network counter data |
| 15 | | Complaint data |
| 16 | | Voice detail data |
| 17 | | Resource configuration data |
| 18 | BSS domain | User detailed information |
| 19 | | Monthly bill data |
| 20 | | Voice service detail data |
| 21 | | Data service detail data |
| 22 | | Product information |
| 23 | | Terminal information |

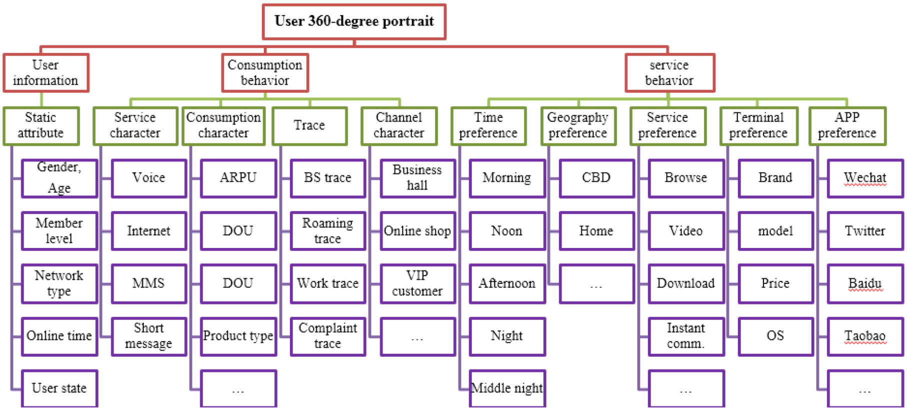


**Fig. 2.** 360-degree portrait for each user

The second type of factors is the user's consumption behavior. It includes four types of sub-factors, including the service character, the consumption character, the trace, and the channel character [17, 23].

The third type of factors is the service behavior. It includes five types of sub-factors, including the time preference, the geography preference, the service preference, the terminal preference, and the Application (APP) preference [23].

After drawing the 360-degree portrait of each user, we further analyze the user-group. This user-group analysis aims at studying the corresponding group of user (e.g., potential value user-group, Iphone terminal user-group, 4G migrated user-group, youth user-group etc.), according to a specific service requirement or a specific character [19, 24]. Typical user-group/clustering analysis algorithms include K-Means algorithm, Clarans algorithm, Focused-Clarans algorithm, K-Medoids algorithm etc. [25, 26].

On the basis of the 360-degree portrait and the user-group clustering analysis, we can jointly analyse the user information, the consumption behavior, the service behavior, as well as the network performance [27, 28]. In all, the 360-degree portrait and the user-group analysis help telecom operators identify serving users and be aware of the user-group precisely.

# 3 Data Mining Process for 4G Potential Users

## 3.1 Overall Process of 4G Data Mining

This section discusses the process of data mining for the 4G potential users targeting. The data mining process for working a machine learning (ML) problem is illustrated in Fig. 3. The whole process consists of six phases. This process provides a good coverage of the steps needed, starting with business understanding, data preparation, model building, model training, model evaluation, model deployment.

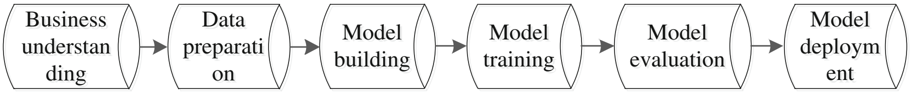


**Fig. 3.** Overall process of data mining for 4G potential users

## 3.2 Business Understanding

A data mining project starts with the understanding of the business problem [26, 29]. In this process, data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then formulated into a data mining problem.

As for the 4G potential users targeting problem, the purpose is to target the potential users from existing 2G or 3G users. It is because the potential 4G users will contribute a lot to the growth of the services traffic in the near future. In addition, LTE network deployment is undergoing rapidly in China, telecom operators are trying to persuade 2G or 3G users to camp on LTE network to provide more efficient data services and better user experience, as well as the guaranteed quality of service (QoS) [2, 21].

### 3.3 Data Preparation

It is generally known that machine learning algorithms learn from data. The quality of the training data is vital factor that impacts the model performance. Therefore, it is critical that ML algorithm has the appropriate data for the problem formulation.

For the data preparation, the first step is collecting the data related with the problem. And then we try to be aware of the details of the data. The next step is to identify data quality problems such as data missing problem, data transformation etc. With regarding the problem in this paper, the user basic information, user consumption information, user service behavior information are needed.

### 3.4 Model Building

Model building actually is defining the abstract problem in a mathematical way that can be understood by machine. For a certain problem, you can select and apply various mining functions because you can use different mining functions for the same type of data mining problem. For example, Logistic Regression, Decision Tree, Support Vector Machine can all be used to deal with the classification problem [23]. However, different kinds of algorithm has its own applicable scenario and require specific data types.

For this paper, apparently this is a problem of classification, which aims to distinguish the potential 4G users from all the telecom subscribers. The two different groups can be expressed in Eq. (1):

$$\text{4G}\_potential\_user\_(Q_i) = \begin{cases} 1, & Q_i\, is\, \text{4G}\_potential\_user \\ 0, & Q_i\, is\, not\, \text{4G}\_potential\_user \end{cases} \tag{1}$$

Then the problem of this paper can be showed in the following linear mathematical model:

$$y = \theta^T z = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + , \ldots, + \theta_n z_n \tag{2}$$

where $\theta$ is the weighting value, $z$ is the feature vector.

### 3.5 Model Training

The process of model training is to train the model built in previous part by using training data. After the training process, we can get a ML model that attempts to predict whether a new user will be a potential 4G user or not.

During the model training process, another import issue can determine the corresponding features. This process is called the feature engineering, which aims to transform original data into features that make ML algorithm work precisely.

The feature engineering can be divided into several parts. Firstly, select the relevant variables to the target based on the understanding of the data, as for the topic in this paper, the user smartphone network type, the user traffic consumption etc. will have great influence to the 4G user prediction. Secondly, feature construction and feature selection. Sometimes, the feature can be obtained directly from the raw data, and for other times, we have to do some transformation to generate some more powerful features. Then we can evaluate each feature's contribution to the accuracy of the model, the unimportant features will be eliminated. Thirdly, we will check weather this set of features will work with the model precisely. We will choose the set of features generated the better prediction result. Lastly, if the prediction result is not ideal, then go back to the beginning part to create more features until the model result is well. From the above steps, we can see that the process of feature engineering is an iterative process.

For model training part in this paper, because this is a classification problem, we will choose the classic LR algorithm [15] to predict the 4G potential users. LR algorithm is a binary classification problems predict a binary outcome (one of two possible classes). LR generates the coefficients of a formula to predict a logit transformation of the probability of presence of the target of interest.

$$\operatorname{logit}(p) = \beta_0 + \beta_1 z_1 + \ldots + \beta_k z_k \tag{3}$$

where $p$ is the probability of presence of the target of interest and total $k$ independent features $\mathrm{z} = (z_1, z_2, \ldots, z_k)$.

The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} \tag{4}$$

where p can be denoted in the Eq. (5):

$$P(Y = 1 \mid z) = \frac{1}{1 + e^{-g(z)}} \tag{5}$$

where $g(z) = \beta_0 + \beta_1 z_1 + \ldots + \beta_k z_k$, $\beta$ is the regression coefficient [26]. Similarity, we can define the probability of absence of the characteristic of interest:

$$P(Y = 0 \mid z) = 1 - P(Y = 1 \mid z) = 1 - \frac{e^{g(z)}}{1 + e^{g(z)}} = \frac{1}{1 + e^{g(z)}} \tag{6}$$

Based on Eqs. (5) and (6), the Eq. (4) can be expressed as:

$$\log(\frac{p}{1-p}) = \log(e^{g(z)}) = g(z) = \beta_0 + \beta_1 z_1 + \ldots + \beta_k z_k \tag{7}$$

In Eq. (7), the coefficient $\beta$ is what we want to learn from the training data. The cost function can be denoted as:

$$L(\beta) = \prod P(y_i = 1 \mid \mathrm{z}_i)^{y_i}(1 - P(y_i = 1 \mid z_i))^{1-y_i} \tag{8}$$

We can get the value of each $\beta$ coefficient via GD (Gradient Descent) algorithm.

### 3.6 Model Evaluation

Model evaluation metrics are used to assess goodness of fit between model and data, to compare different models, and to reveal the accuracy of the model predictions. Actually, there are many ways to do the model evaluation, such as the Confidence Interval, Confusion Matrix, Gain and Lift Chart and ROC curve etc. In terms of the classification problem in this paper, we introduce Confusion Matrix method to measure the model's performance.

A confusion matrix is a $N \times N$ matrices. $N$ is the number of classifications. A *2 × 2 confusion matrix* for two classes (Positive and Negative) is presented in Table 2.

**Table 2.** Parameters in confusion matrix

| | | Target | |
|---|---|---|---|
| | | *Positive* | *Negative* |
| Model | *Positive* | $T_{11}$ | $T_{12}$ |
| | *Negative* | $T_{21}$ | $T_{22}$ |

From Table 2, we can see that the performance of the classification models can be evaluated using the data in the matrix. $T_{11}$ is called correct positive prediction, these are cases in which we predicted yes (they are the potential 4G users), and they are actually potential 4G users. Similarly, $T_{12}$ is called incorrect positive prediction, $T_{21}$ is called correct negative prediction, $T_{22}$ is called incorrect negative prediction.

Various measures can be derived from a confusion matrix [26]. The first metric is $M_{Accuracy}$, calculated by Eq. (9). $M_{Accuracy}$ reflects the proportion of the total number of predictions that were correct.

$$M_{Accuracy} = \frac{T_{11} + T_{22}}{T_{11} + T_{21} + T_{12} + T_{22}} \tag{9}$$

The second metric is $M_{Precision}$, calculated by Eq. (10). $M_{Precision}$ reflects the proportion of actual positive cases which are correctly identified.

$$M_{\mathrm{Precision}} = \frac{T_{11}}{T_{11} + T_{12}} \tag{10}$$

The third metric is $M_{Recall}$, calculated by Eq. (11). $M_{Recall}$ reflects the proportion of positive cases that were correctly identified.

$$M_{\mathrm{Re}call} = \frac{T_{11}}{T_{11} + T_{21}} \tag{11}$$

Finally, we can get the ideal data model for the targeting for 4G potential users.

### 3.7 Model Deployment

After the model is validated, telecom operator can apply this model to 4G potential user classification. For each existing 2G or 3G user, this model will present the classification results. A data mining process continues after a solution is deployed. Data mining is an iterative process, since both the data and services are changing during the process.

### 3.8 Application Deployment and Results Analysis

This paper applies the proposed data mining algorithm based 4G potential users prediction in the city central area of China (named as A-city). Table 3 and Fig. 4 show the mining results. There are 1807674 2G mobile users in A-city. Employing our proposed data mining algorithm, 108000 of 2G users are predicted as 4G potential users. Then, market department of the telecom operator utilises this potential users list to take relevant marketing measures/methods to persuade these users transferring from 2G to 4G. Finally, 58423 2G users transfer to 4G successfully. Therefore, the *4G user conversion rate* reaches 54.10% (namely, 58423/108000) under 2G scenario.

Similarly, from Table 3 and Fig. 4, there are 1063955 3G users in A-city. The proposed data mining algorithm predicts that 335000 3G users are 4G potential users. After employing marketing measures/methods by the market department of the telecom operator, 216581 3G users transfer to 4G successfully. Hence, the *4G user conversion rate* reaches 64.65% (namely, 216581/335000) under 3G scenario.

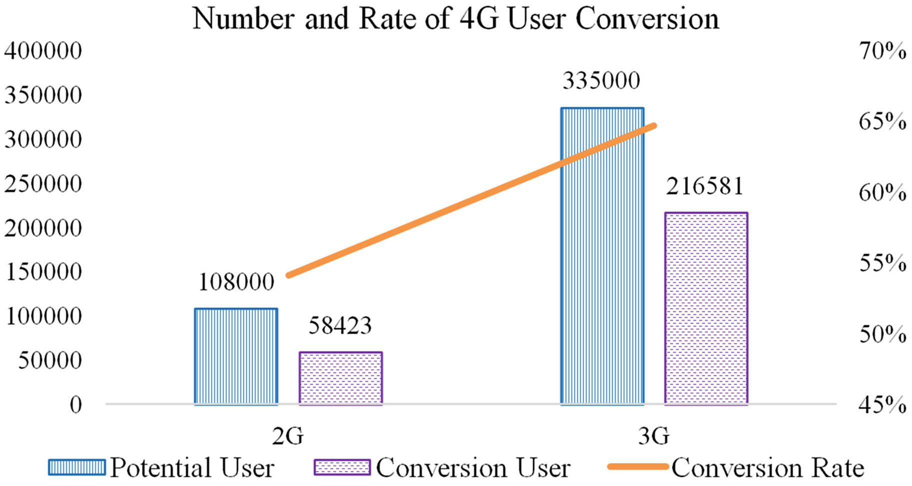
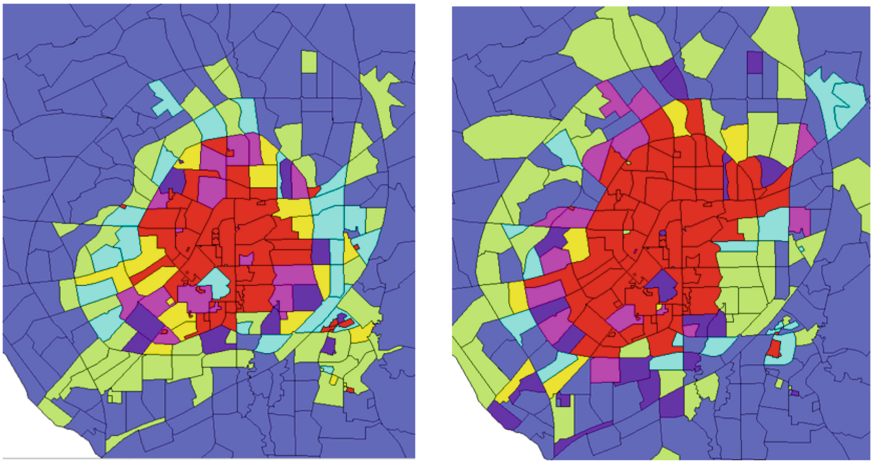


**Fig. 4.** Number and rate of 4G user conversion

**Table 3.** Application scenario and results

| Network type | *4G User conversion rate* | | | |
|---|---|---|---|---|
| | *Total user* | *4G Potential user prediction* | *4G Conversion user* | *Conversion rate* |
| 2G | 1807674 | 108000 | 58423 | 54.10% |
| 3G | 1063955 | 335000 | 216581 | 64.65% |

Figure 5 shows the 4G potential users distribution in the downtown of A-city. Red area reflects that this area has a large number of 4G potential users. From Fig. 5(a), most of 2G–to–4G potential users are gathered in the very city center. Since many 3G users surf on the mobile internet, these users are easily upgrade to 4G users. Compared to 2G–to–4G potential users, 3G–to–4G potential users have larger distribution, as shown in Fig. 5(b).



(a) 2G-to-4G potential users

(b) 3G-to-4G potential users

**Fig. 5.** Distribution of 4G potential users

## 4 Conclusion

This paper investigates on telecom big data based user analysis and application for telecom operators. Initially, this paper studies the content of telecom big data. Then, 360-degree user portrait is drawn via three types of factors, including user's information, user's consumption behavior and service behavior. Based on the user portrait, this paper employs data mining to seek 4G potential users. The data mining process includes the business understand, data preparation, model building, model train, model evaluation and model deployment steps. Overall, effectively targeting 4G potential users can benefit telecom operators.

## References

1. Cao, Y., Wang, N., Sun, Z., Cruickshank, H.: A reliable and efficient encounter-based routing framework for delay/disruption tolerant networks. IEEE Sens. J. **15**(7), 4004–4018 (2015)
2. Xu, L., Luan, Y., Cheng, X., Cao, X., Chao, K., Gao, J., Jia, Y., Wang, S.: WCDMA Data based LTE Site Selection Scheme in LTE Deployment. In: International Conference on Signal and Information Processing, Networking and Computers, pp. 249–260. CRC Press, Taylor & Francis Group, Beijing (2015)
3. Zhang, H., Dong, Y., Cheng, J., Hossain, M.J., Leung, V.C.M.: Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks. IEEE Wirel. Commun. **22**(3), 92–99 (2015)
4. Xu, L., Chen, Y., Chai, K.K., Schormans, J., Cuthbert, L.: Self-organising cluster-based cooperative load balancing in OFDMA cellular networks. Wiley Wireless Commun. Mob. Comput. **15**(7), 1171–1187 (2015)
5. Zhao, L., Al-Dubai, A.Y., Li, X., Chen, G.: A new efficient cross-layer relay node selection model for wireless community mesh networks. Comput. Electr. Eng. **61**, 361–372 (2017). https://doi.org/10.1016/j.compeleceng.2016.12.031
6. Cao, Y., Sun, Z., Wang, N., Riaz, M., Cruickshank, H., Liu, X.: Geographic-based Spray-and-Relay (GSaR): an efficient routing scheme for DTNs. IEEE Trans. Veh. Technol. **64**(4), 1548–1564 (2015)
7. Xu, L., Chen, Y., Gao, Y., Cuthbert, L.: A Self-Optimizing Load Balancing Scheme for Fixed Relay Cellular Networks. In: IET International Conference on Communication Technology and Application, pp. 306–311. IET Press, Beijing (2011)
8. Wang, W., Xu, L., Zhang, Y., Zhong, J.: A novel cell-level resource allocation scheme for OFDMA system. In: International Conference on Communications and Mobile Computing, pp. 287–292. IET Press, Kunming (2009)
9. Zhao, L., Li, Y., Meng, C., Gong, C., Tang, X.: A SVM based routing scheme in VANETs. In: IEEE International Symposium on Communications and Information Technologies, pp. 380–383. IEEE Press, Qingdao (2016)
10. Zhang, H., Dong, Y., Cheng, J., Hossain, M.J., Leung, V.C.M.: Fronthauling for 5G LTE-U ultra dense cloud small cell networks. IEEE Wirel. Commun. **23**(6), 48–53 (2016)
11. Deng, Y., Wang, L., Zaidi, S.A.R., Yuan, J., Elkashlan, M.: Artificial-noise aided secure transmission in large scale spectrum sharing networks. IEEE Trans. Commun. **64**(5), 2116–2129 (2016)
12. Xu, L., Cheng, X., Liu, Y., Chen, W., Luan, Y., Chao, K., Yuan, M., Xu, B.: Mobility load balancing aware radio resource allocation scheme for LTE-Advanced cellular networks. In: IEEE International Conference on Communication Technology, pp. 806–812, IEEE Press, Hangzhou (2015)
13. Xu, L., Chen, Y., Chai, K. K., Luan, Y., Liu, D.: Cooperative mobility load balancing in relay cellular networks. In: IEEE International Conference on Communication in China, pp. 141–146. IEEE Press, Xi'An (2013)
14. 3GPP TR 36.814: Further Advancements for E-UTRA Physical Layer Aspects (2010)
15. Zhang, T., Cheng, X., Yuan, M., Xu, L., Cheng, C., Chao, K.: Mining target users for mobile advertising based on telecom big data. In: IEEE International Symposium on Communications and Information Technologies, pp. 296–301. IEEE Press, Qingdao (2016)
16. Landset, S., Khoshgoftaar, T.M., Richter, A.N., Hasanin, T.: A survey of open source tools for machine learning with big data in the Hadoop Ecosystem. J. Big Data **2**(1), 24 (2015)

17. Cheng, C., Cheng, X., Yuan, M., Song, C., Xu, L., Ye, H., Zhang, T.: A novel cluster algorithm for telecom customer segmentation. In: IEEE International Symposium on Communications and Information Technologies, pp. 324–329. IEEE Press, Qingdao (2016)
18. Liu, Y., Xu, L., Chen, Y., Fan, Y., Xu, B., Nie, J.: A novel power control mechanism based on interference estimation in LTE cellular networks. In: IEEE International Symposium on Communications and Information Technologies, pp. 397–401. IEEE Press, Qingdao (2016)
19. Xing, H., Xu, L., Qu, R., Qu, Z.: A quantum inspired evolutionary algorithm for dynamic multicast routing with network coding. In: IEEE International Symposium on Communications and Information Technologies, pp. 186–190. IEEE Press, Qingdao (2016)
20. Cao, Y., Wang, N., Kamel, G., Kim, Y.J.: An electric vehicle charging management scheme based on publish/subscribe communication framework. IEEE Syst. J. **11**(3), 1822–1835 (2015). https://doi.org/10.1109/JSYST.2015.2449893
21. Xu, L., Luan, Y., Cheng, X., Fan, Y., Zhang, H., Wang, W., He, A.: Telecom big data based user offloading self-optimisation in heterogeneous relay cellular systems. Int. J. Distrib. Syst. Technol. **8**(2), 27–46 (2017)
22. Xu, L., Chen, Y., Chai, K.K., Liu, D., Yang, S., Schormans, J.: User relay assisted traffic shifting in LTE-Advanced systems. In: IEEE Vehicular Technology Conference, pp. 1–7. IEEE Press, Dresden (2013)
23. Cheng, X., Xu, L., Zhang, T., Jia, Y., Yuan, M., Chao, K.: A novel big data based telecom operation architecture. In: International Conference on Signal and Information Processing, Networking and Computers, pp. 385–396. CRC Press Taylor & Francis Group, Beijing (2015)
24. Xu, L., Luan, Y., Cheng, X., Xing, H., Liu, Y., Jiang, X., Chen, W., Chao, K.: Self-optimised joint traffic offloading in heterogeneous cellular networks. In: IEEE International Symposium on Communications and Information Technologies, pp. 263–267. IEEE Press, Qingdao (2016)
25. Cao, Y., Wang, N., Kamel, G.: A publish/subscribe communication framework for managing electric vehicle charging. In: IEEE International Conference on Connected Vehicles and Expo, pp. 318–324. IEEE Press, Vienna (2014)
26. Cheng, X., Yuan, M., Xu, L., Zhang, T., Jia, Y., Cheng, C., Chen, W.: Big data assisted customer analysis and advertising architecture for real estate. In: IEEE International Symposium on Communications and Information Technologies, pp. 312–317. IEEE Press, Qingdao (2016)
27. Cui, G., Lu, S., Wang, W., Zhang, Y., Wang, C., Li, X.: Uplink coordinated scheduling based on resource sorting. In: IEEE Vehicular Technology Conference, pp. 1–5. IEEE Press, Quebec (2012)
28. Cui, G., Lu S., Wang, W., Wang, C., Zhang, Y.: Decentralized antenna selection with no CSI sharing for multi-cell MU-MIMO systems. In: IEEE International Symposium on Personal and Indoor, Mobile Radio Communications, pp. 2319–2323. IEEE Press, Sydney (2012)
29. Xie, L., Pan, D.: On customer segmentation and retention of telecom broadband in Pearl River Delta. Chinese Control Conference, pp. 5564–5568 (2010)