IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 23, 2016, accepted October 26, 2016, date of publication December 15, 2016, date of current version January 27, 2017.

*Digital Object Identifier 10.1109/ACCESS.2016.2635805*

# When Will You Have a New Mobile Phone? An Empirical Answer From Big Data

**QINGLI MA[1], SIHAI ZHANG[1], WUYANG ZHOU[1], SHUI YU[2], (Senior Member, IEEE), AND CHONGGANG WANG[3], (Fellow, IEEE)**

[1]Key Laboratory of Wireless-Optical Communications, Chinese Academy of Sciences University of Science and Technology of China, Hefei 230026, China
[2]School of Information Technology, Deakin University, Melbourne 3125, Australia
[3]InterDigital Communication, Pennsylvania 19428, USA

Corresponding author: S. Zhang (shzhang@ustc.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61461136002, in part by the Key Program of National Natural Science Foundation of China under Grant 61631018, in part by the National Programs for High Technology Research and Development under Grant 2014AA01A707, and in part by the Fundamental Research Funds for the Central Universities and Huawei Technology Innovative Research on Wireless Big Data.

**ABSTRACT** When and why people change their mobile phones are important issues in mobile communications industry, because it will impact greatly on the marketing strategy and revenue estimation for both mobile operators and manufactures. It is a promising way to take use of big data to analyze and predict the phone changing event. In this paper, based on mobile user big data, first through statistical analysis, we find that three important probability distributions, i.e., power-law, log-normal, and geometric distribution, play an important role in the user behaviors. Second, the relationships between eight selected attributes and phone changing are built, for example, young people have greater intention to change their phones if they are using the phones belonging to the low occupancy phones or feature phones. Third, we verified the performance of four prediction models on phone changing event under three scenarios. Information gain ratio was used to implement attribute selection and then sampling method, cost-sensitive together with standard classifiers were used to solve imbalanced phone changing event. Experiment results show our proposed enhanced backpropagation neural network in the undersampling scenario can attain better prediction performance.

**INDEX TERMS** Mobile big data, attribute selection, imbalance problem, phone changing prediction, machine learning.

## I. INTRODUCTION

During the past twenty years, we have witnessed the marvelous growth of mobile communications. Taking mainland China as an example, up to May 2015, China Mobile, China Unicom, and China Telecom possess 816 million, 290 million, and 190 million users, respectively. These enormous mobile user population support the prospect of mobile phone manufacture industry and stimulate the research on mobile user behavior. The living habitats of mobile users are changing due to the development of mobile phones and mobile payments [1], [2]. Thus, mobile user resources are the most important asset of mobile operators, and their user behavior are surely worthy to investigate.

In recent years, the development of information technology, such as data gathering, data storage, and data processing, have led to a deluge of data from diverse domains [3]. A large number of researches based on big data have emerged gradually, such as big data in health care [4], [5], big data in business [6], networking for big data [7], privacy protection for big data [8], [9], and so on. Mobile operators have also obtained numerous amount of data, including mobile user calling records, mobile user location information, wireless signalling data, status information of base stations and other devices and sensing data from mobile phones. Lots of research topics have been launched to investigate the power of big data analysis in mobile communication industry. For example, user services based on geographical location can help ease traffic pressure [10], even carry out criminal behavior detection [11]. By analyzing the user behavior [12], operators can make efficient strategy in marketing, sales, and other fields. Individual's stress assessment using smart phone interaction analysis [13] can help people to relax and reduce their stress level.

Mobile users also have dazzling chances to choose and replace their mobile phones for various reasons, such as phone damage or lost, new phone products, operators'


2169-3536 © 2016 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

marketing activities and so on. According to the up-to-date statistics of Chinese government, mobile phone coverage rose to 95.5% in mainland China by the end of 2015 [14], therefore mobile operators and phones manufacturers are facing a increasing bottleneck to develop new users. It is obvious that phone changing events are playing important role for both mobile phone manufacturers and mobile communication operators, although with different significance and applications. Thus, existing customers who change their phones should be paid more attention to. Leveraging the big data analysis to predict the phone changing events will bring significant effects, which is the major work of this paper.

Basically, there is one most important question in this topic: When will one mobile user change her/his mobile phone? Traditional survey methods have been adopted to investigate this question but contributed few satisfactory outcomes, and few research works using large scale user data set has been reported by now. Liu *et al.* [15] evaluated the user behavior of phone replacement from the perspective of users' economic conditions based on the traffic data, but actual prediction on phone changing was not performed. The case study in Turkey used 302 mobile phone consumers and clustered them into behaviorally different groups in terms of phone buying decision [16], but the conclusions are hard to be generalized due to the small number of samples.

In this paper, using the mobile user data with profound user behavior information, including user profiles, calling habits and social impact information provided by one Chinese telecommunication operator, we first present the statistics of phone using behaviors, and then discuss the impact of attributes on phone changing events using statistical analysis to recognize the key attributes, which may be closely related to phone changing events. Finally, we investigate the phone changing prediction in-depth, such as solving imbalance problem and comparing classifier algorithms. Our contributions are summarized as follows:

- We dispose the user data and witness three important distributions to describe user behavior. The phone brands conform to the power law distribution. The age and fee follow log-normal distribution. The number of phones used in past two years (NPTY), used months and call duration obey geometric distribution.
- Two interesting trends on phone changing are presented. (1) Young people are prone to changing their phones when they have feature phones or low occupancy phones, while old people not. (2) The more phones having been used, the higher the probability of changing the phone. The longer people use their phones, they have higher probability of changing their phones.
- We verified four prediction models on phone changing over three scenarios to predict phone changing events and get three important conclusions. (1) The undersampling scenario perform better than SMOTE and cost-sensitive. (2) Logistic with ridge regression algorithm is more stable than other three classifiers. (3) Our proposed E-BP classifier in undersampling scenario can achieve best prediction performance. In addition, information gain ratio algorithm was used to implement attribute selection and sampling method or ameliorating standard classifiers were used to solve imbalanced phone changing problem.

The remainder of this paper are organized as follows. Data formats and preprocessing is introduced in Section II. Basic statistics of these mobile users are presented in Section III. Attribute selection to analyze the phone changing events is introduced and discussed in Section IV. Evaluation measurements and approaches for solving imbalance are introduced in Section V. Algorithm design and experiment are implemented and discussed in Section VI. Finally, we conclude our work in section VII.

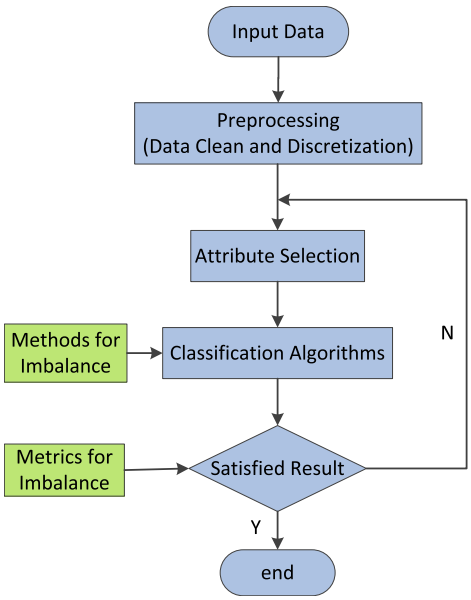



**FIGURE 1.** Processing procedure.

## II. DATA FORMATS AND PREPROCESSING

In this paper, the whole process is described in Fig. 1. The data set provided by one mobile operator contains 200,000 records collected in Dec. 2014, each record representing one user's phone using behavior. There are 16 attributes in each record which are exhibited in Table I. Some attribute abbreviations are explained as follows.

The 'Used Months' attribute denotes how many months the user has used her/his current mobile phone. The 'Size of Comm.' and 'Phones Changed in Comm.' are two attributes that consider the effect of social ties. The 'Size of Comm.' denotes the number of friends this user has, which is revealed by the phone calling history, and 'Phones Changed in Comm.' means the number of friends who changed their phones in current month.

The 'Label of Changing Phone' label denotes that, 1 means that user changed her/his phone in the next month, while 0 not. In this data set, there are 14,053 records with flag $= 1$, and 185,947 records with flag $= 0$.

**TABLE 1.** Attribute description of the data set.

| Attribute | Format | Comment |
|---|---|---|
| User ID | String | |
| User Age | Numeric | 16-70 Years |
| Phone Brand ID | Numeric | 846 Brands in total |
| Phone Brand Version | String | 6887 kinds in total |
| Smart Flag | Binary | 1 : Smart Phone<br>0 : Feature Phone |
| Used Months | Numeric | 1-37 |
| Data Traffic of Month before Last | Float | 0-21 GByte |
| Data Traffic of Last Month | Float | 0-38 GByte |
| Data Traffic of Current Month | Float | 0-45 GByte |
| Call Duration of Current Month | Numeric | 0-203 Hours |
| Fee of Current Month | Numeric | 0-2725 CNY |
| Date of Open Account | Date | 20100612-20140630 |
| NPTY | Numeric | 1-18 |
| Size of Comm | Numeric | 1-10 |
| Phones Changed in Comm | Numeric | 1-10 |
| Label of Changing Phone | Binary | 1: Changed,<br>0: Not Changed |

### A. DATA CLEANING

Several data cleaning operations have been performed.

- Phone Brand Integration. Two operations have been taken for brand related attributes. Firstly, there are different names for the same brand, for example, 'Xiaomi' and 'Xiaomi Tech' both denoting Xiaomi brand. Secondly, the phone brand version attribute is neglected, so that, the phone brands with different versions are considered as identical.
- Noisy Data Elimination. Firstly, some values are missing in some records due to unknown reasons, thus these records are cleaned. Secondly, the records with abnormal values, for example, 'Fee of Current Month' is 0, but call duration $>$ 1 hour and traffic of current month $>$ 1G bits, have been cleaned.
- Attribute Conversion. We obtain a new attribute, namely, Average Traffic, by averaging the traffic of last three months, for the following analysis.

After the data cleaning above, we have 199,910 records with 12 attributes including the class label for the analysis below.

### B. PHONE BRAND SEGREGATION

We separate all phone brands into four categories according to their market occupancy, as shown in Table II.

Apple and Samsung, the two famous international brands, are grouped into category $C1$, because they occupy the top 2 occupancy which both exceed 10 percentage. $C2$ consists of 6 brands whose occupancy are in the interval of (5.0%, 10.0%), which take the 3rd to 8th occupancy rate and contribute totally 41.8% occupancy in our data set. Other 10 brands with their occupancy in the interval of (0.5%, 5.0%), are grouped into category $C3$. Finally, all rest brands whose occupancy are below 0.5% are left into category $C4$, which contributes totally 12.5% occupancy. The occupancy rate reflect the users' confidence degree in the brand.

**TABLE 2.** Phone brand categories.

| Phone Brand | Occupancy | Category | Sum Occupancy |
|---|---|---|---|
| Apple | 16.0% | C1 | 30.7% |
| Samsung | 14.7% | | |
| Xiaomi | 9.1% | C2 | 41.8% |
| Lenovo | 8.0% | | |
| Vivo | 7.8% | | |
| OPPO | 6.6% | | |
| Huawei | 5.3% | | |
| Nokia | 5.0% | | |
| Coopad | 4.2% | C3 | 15.0% |
| ZTE | 2.2% | | |
| K-Touch | 1.6% | | |
| Gionee | 1.5% | | |
| HTC | 1.5% | | |
| Meizu | 1.4% | | |
| Hisense | 0.9% | | |
| Sony Ericsson | 0.6% | | |
| DOOV | 0.6% | | |
| Motorola | 0.5% | | |
| Others | 12.5% | C4 | 12.5% |

**TABLE 3.** Interval division of traffic amount.

| Category Number | Traffic Segmentation | Proportion |
|---|---|---|
| 1 | (0 )M | 19.8% |
| 2 | (1-15 )M | 6.1% |
| 3 | (16-44 )M | 6.3% |
| 4 | (45-74 )M | 6.2% |
| 5 | (75-99 )M | 6.1% |
| 6 | (100-129 )M | 6.1% |
| 7 | (130-164 )M | 6.1% |
| 8 | (165-206 )M | 6.2% |
| 9 | (207-259 )M | 6.2% |
| 10 | (260-332 )M | 6.1% |
| 11 | (333-433 )M | 6.2% |
| 12 | (434-572 )M | 6.2% |
| 13 | (573-846)M | 6.2% |
| 14 | (846- )M | 6.2% |

### C. DATA DISCRETIZATION

As the distribution of traffic is uneven, we separate the traffic attributes from real values into discrete intervals using equal frequency discretization (EFD) [17], as shown in Table III. Other attributes are processed using equal width discretization (EWD) unless specifically noted.

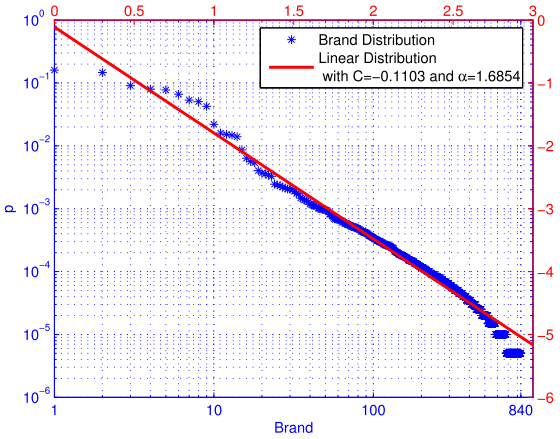



**FIGURE 2.** Phone brand distribution. each point of X axis denote a phone brand, which is listed from left to right according to their occupancy rates.

## III. BASIC STATISTICS

In this section, we present basic statistics on several attributes. Statistics shows that the number of phone brands is 840, but the top 18 brands almost occupy 90 percentage. We plot the brand occupancy distribution in Fig. 2 and the distribution can be fitted using power law very well, which has $c = 0.7757$ and $\alpha = 1.6854$. Here, the probability distribution of power law is $f(x) = cx^{-\alpha}$ and its logarithmic transformation is $\log f(x) = \log c - \alpha \log x$, which can be simplified as $Y = C - \alpha X$.

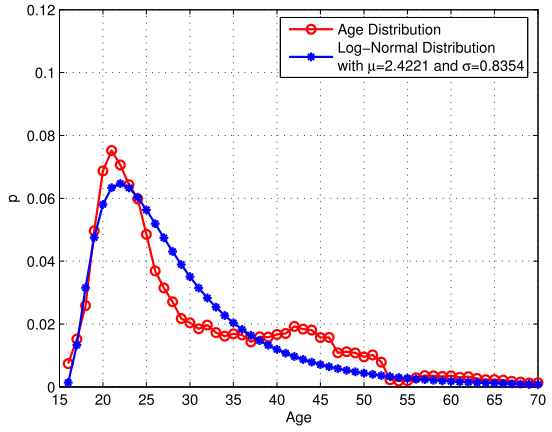



**FIGURE 3.** User age distribution.

The proportion of users in different ages (from 16 to 70) is presented in Fig. 3, which satisfies the discrete log-normal distribution with $\mu = 2.4221$ and $\sigma = 0.8354$. The probability mass function (PMF) of log-normal distribution is in the following.

$$P(x = k) = \frac{A(\mu, \sigma)}{k} \exp[-\frac{(\ln k - \mu)^2}{2\sigma}], \quad k = 1, 2, 3... \tag{1}$$

where $A(\mu, \sigma) = \{\sum_{k=1}^{\infty} \frac{1}{k} \exp[-\frac{(\ln k - \mu)^2}{2\sigma}]\}^{(-1)}$.

We find that the distribution of fee expense for current month of all mobile users also fits the log-normal distribution,

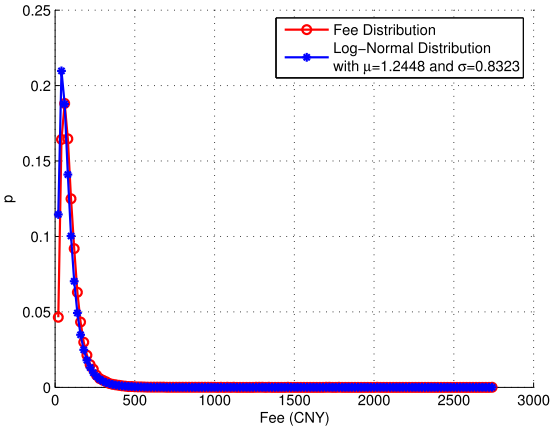



**FIGURE 4.** Fee distribution.

which has $\mu = 1.2448$ and $\sigma = 0.8323$ demonstrated in Fig. 4.

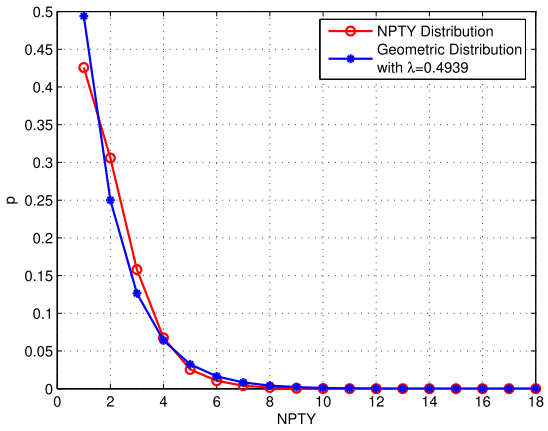



**FIGURE 5.** The distribution of NPTY.

Two important findings on NPTY information are presented in Fig. 5. Firstly, in average, nearly 42.6%, 30.6% and 15.8% of users keep using one, two and three mobile phones in the past two years, respectively, which can be referred to estimate the expected annual shipment of mobile phones if we know the population information of mobile users. Secondly, we find the NPTY fits the geometric distribution, which has the $\lambda = 0.4939$. The PMF of geometric distribution is: $P(x = k) = (1 - \lambda)^{k-1}\lambda, k = 1, 2, 3, ...$

From the data of mobile users with 'Label of Changing Phone' = 1, we can imply how long the mobile phones will be used, and the result is presented in Fig. 6. Several interesting findings can be obtained. Nearly 15.2%, 10.0%, 11.6% of mobile phones will be changed or discarded after only two, four and six months' usage, respectively. We also notice that just 1.4% of mobile phones can be used longer than 3 years and it fits the geometric distribution with the $\lambda = 0.0907$.

Similarly, call duration also fit the geometric distribution, which has $\lambda = 0.1658$. in the Fig. 7.

## IV. ATTRIBUTE SELECTION ON PHONE CHANGING

In this section, we investigate which attributes have important influence on phone changing event. Firstly, we conduct

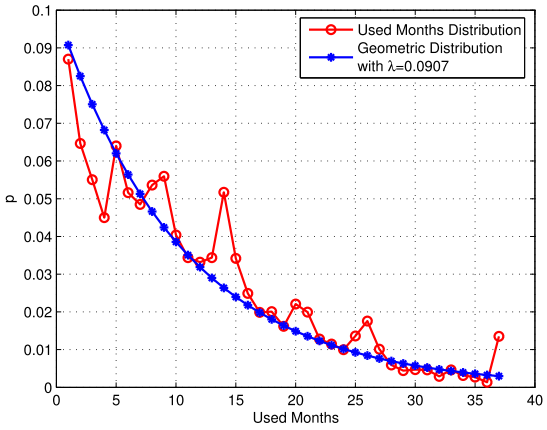



**FIGURE 6.** Used months distribution. each point of X axis means the used month interval.

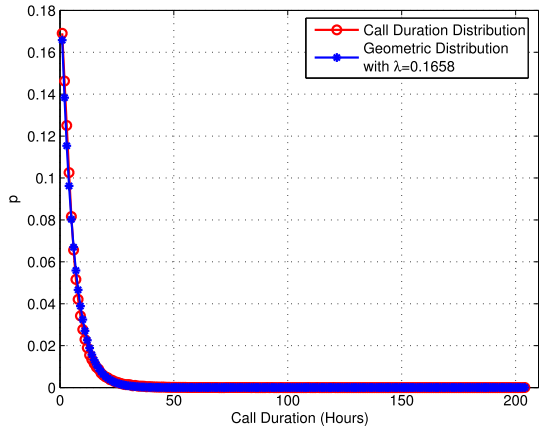



**FIGURE 7.** Call duration distribution.

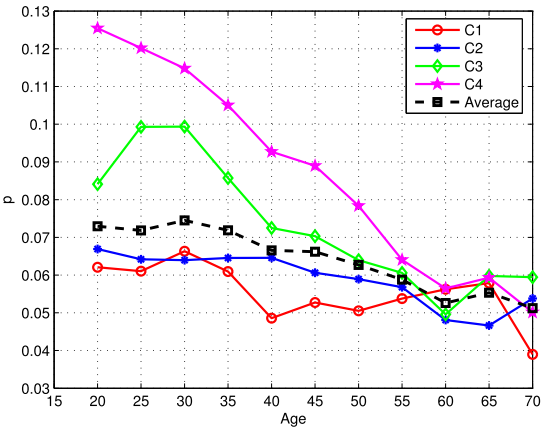



**FIGURE 8.** Phone change probability over age intervals. (interval=5)

attribute analysis based on user behavior. Secondly, machine learning algorithms are utilized to implement attribute selection which lays a foundation for prediction.

### *A. ATTRIBUTE ANALYSIS BASED ON USER BEHAVIOR*

#### 1) AGE AND BRAND INFLUENCE

Fig. 8 shows the phone changing rate over age intervals in average and different phone brands categories. It is clear

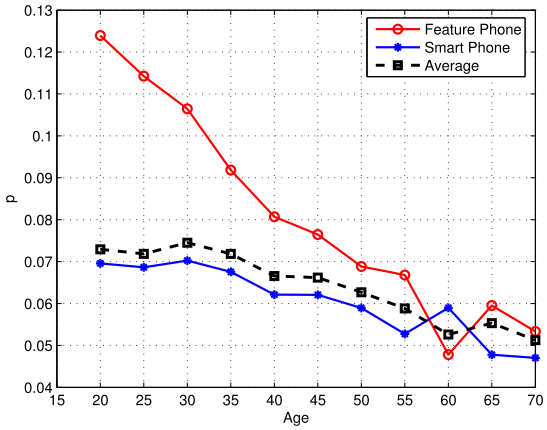



**FIGURE 9.** Change probability between smart phones and feature phones.

that phone brand has important impacts on phone changing decisions and the joint influence of brand categories and age intervals are more significant. Low occupancy phones, like C3 and C4, have higher changing rate in nearly all age intervals, which implies that, their life-cycle are much shorter than other kinds of mobile phones. On the contrary, the popular brands often correspond to better product quality, thus they are less frequently changed. Customers using these brands such as international brands, Apple and Samsung, will show high loyalty which is extremely valuable for phone manufacturers.

Smart phones are becoming more popular nowadays compared to feature phones and Fig. 9 presents the difference between these two kinds of phones on phone changing event. Considering the smart flag and the age jointly, young people are more prone to changing their phones if they are using feature phones, but old people wouldn't change too much, no matter whether they have smart phones or not. Meanwhile we can see that people who have smart phones greatly outnumber people that have feature phones because the average curve bias to the smart phone curve.

#### 2) DYNAMIC TRAFFIC INFLUENCE

We introduce the dynamic traffic influence on phone changing events, include traffic in Fig. 10, fee expense in Fig. 11 and call duration in Fig. 12 because these three factors are all closely related with dynamic behaviors of mobile users.

As the distribution of traffic is uneven, EFD method is adopted to separate the traffic attributes into 14 categories, presented in Table III. We analyze the probability of phone changing based on traffic categories in Fig. 10, and can see that, in average, mobile users who consume higher data traffic will have larger likelihood to change their phones. Especially, for the people who have low occupancy phones, like C4, the probability of phone changing increase more quickly with the traffic increasing. Fig. 11 shows that, the phone changing probability increase from 6.0% steadily to nearly 10.5% when the fee increases from 20 CNY per month to 300 CNY per month, which is very similar with Fig. 10 as

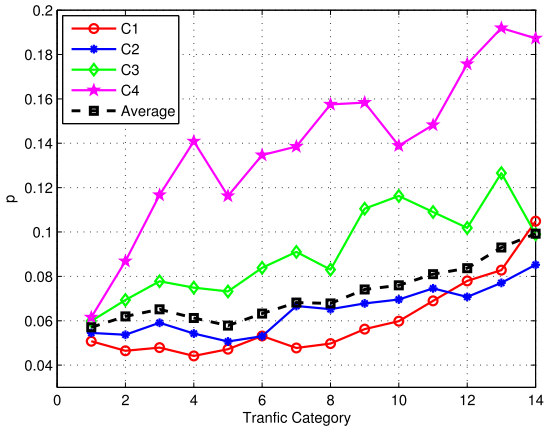



**FIGURE 10.** Phone change rate over traffic amount using the traffic interval division in Table III. The value of X axis denote corresponding categories.

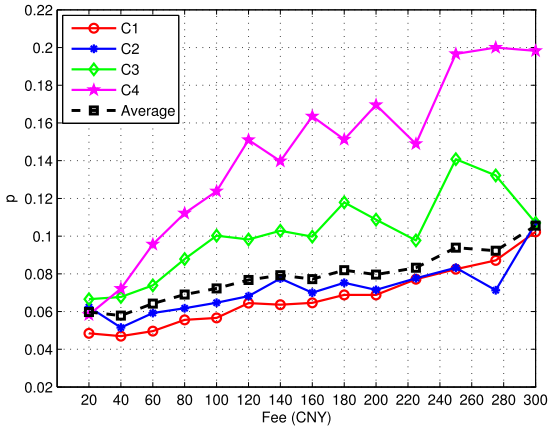



**FIGURE 11.** Phone change rate over fee intervals (interval=20 CNY).

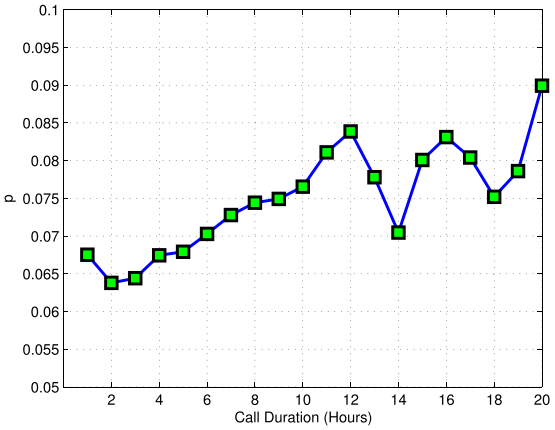



**FIGURE 12.** Phone change rate over duration time.

the traffic attribute have strong correlation with fee attribute. Fig. 12 shows that, the call duration time is only a little relevant to mobile phone changing events. From 1 to 12 hours, the probability of phone changing increase from 6.4% to 8.4%. After the 12 hours, the curve fluctuate due to insufficient data.

### 3) LONG TERM INFLUENCE FACTORS

Long term influence considers much longer duration than three months, and in this paper, we have the data attribute, say, NPTY, denoting the number of mobile phones one has used in the past two years, and we can calculate how many months the mobile phones have been used up to now.

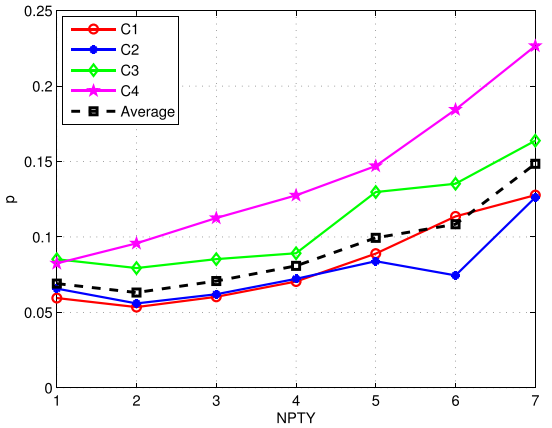



**FIGURE 13.** Phone change rate over NPTY.

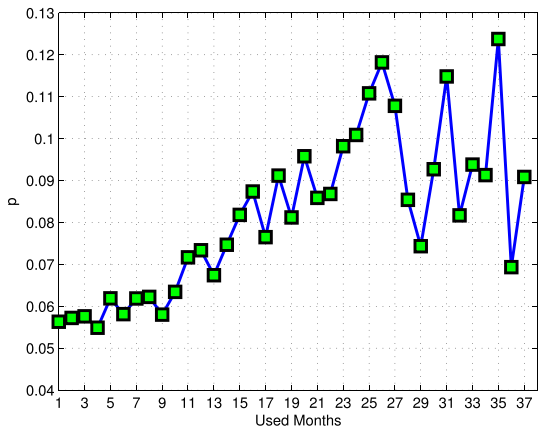



**FIGURE 14.** Phone change rate over used months.

Fig. 10, Fig. 11 and Fig. 12 only show mobile users' latest service consumption, but NPTY can reflect their consumption habits in longer period. We can notice this increasing trend of phone changing possibility when NPTY increases shown in Fig. 13, i.e, the more phones someone have used, the greater probability he will change the phone.

Fig. 14 demonstrates that, the longer people use their phones, the greater possibility they change their phones, which is a natural understanding not only for phone usage, but possible for all products. As the number of used months that exceed 25 get smaller and smaller, the phone changing rate fluctuate greatly. The probability that changing phone basically conform to linear growth trend with the used months increasing.

Note that the significance of Size of Comm and Phones Changed in Comm on phone changing events is not recognized and the reason might be incorrect data collections

before presenting to us. For example, the two attributes, the Size of Comm and Phones Changed in Comm, have been set to 10 as long as they are more than 10.

### B. ATTRIBUTE SELECTION USING MACHINE LEARNING ALGORITHM

The purpose of attribute selection is mainly three parts [18]: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. Information gain (IG) base on entropy is initially used in ID3 algorithm and also applied in attribute selection [19], but it is prone to choosing the biased attributes that possess more values. In this subsection, we use information gain ratio (IGR) [20] to choose effective attributes and validate whether the attributes are related with the class label.

$$IG_j = H(Class) - H(Class|Attribute_j) \tag{2}$$

$$IGR_j = \frac{IG_j}{H(Attribute_j)} \tag{3}$$

$H(Class)$ and $H(class|Attribute_j)$ give the entropy of class label before and after observing the $j$th attribute. Then, $IG_j$ indicates the information gain of $j$th attribute and $IGR_j$ denotes the information gain ratio using the information gain normalization which is divided by the entropy of $j$th attribute.

The results using information gain ratio is presented in Fig. 15. As the attributes are ranked, we can even construct a simple classifier using information gain ratio. Then, how many attributes should we select to implement prediction? One strategy is to define a confidence ratio metric in the following.

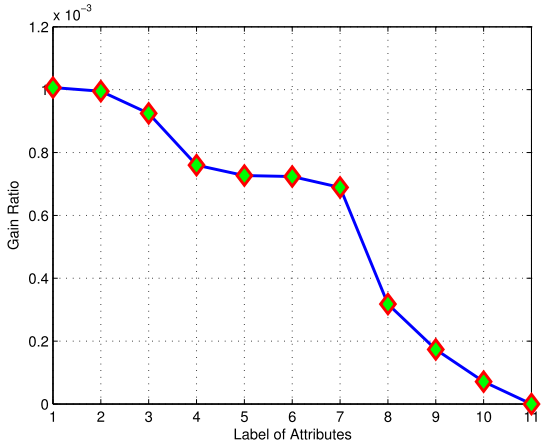



**FIGURE 15.** Gain ratio over different attributes. The values of X axis are the different attributes. They are, from left to right, phone brand(C1 C2 C3 C4), average traffic, used months, NPTY, fee, smart flag, call duration, age, size of comm, phones changed in comm and date of open account.

$$cr = \frac{\sum_{j=1}^{n} IGR_j}{\sum_{j=1}^{N} IGR_j}, \tag{4}$$

where $N$ is the number of all attributes.

Suppose the information gain ratio of attributes have been ranked in descending order just as Fig. 15. Generally we set confidence ratio to 95%, i.e., $cr \geq 95\,\%$. As a result, the value $n$ is 8. Hence, we select the top 8 attributes, i.e. Phone Brand (C1 C2 C3 C4), Average Traffic, Used Months, NPTY, Fee, Smart Flag, Call Duration, and Age, for the next experiment. So we discard Size of Comm, Phones Changed in Comm and Date of Open Account, for their little relevance with the class label.

As to the attributes selection, we consider the combinatorial cases by choosing 11 to 5 attributes in descending order after the attributes have been ranked and implement prediction analysis. We find that it can bring better prediction performance using 8 attributes, especially for E-BP algorithm in undersampling scenario, which would be introduced in the subsequent Section VI.

## V. MEASUREMENTS AND IMBALANCE

In practical scenarios, lots of class imbalance problems exist, such as customer churn prediction [21], [22], traffic accident prediction [23], and so on. As to phone changing, the imbalance problem should also be discussed. In this section, various measurements are introduced for evaluating the performance of classifiers and several methods to deal with data set imbalance are described.

**TABLE 4.** Confusion matrix.

| | Predict 1 | Predict 0 |
|---|---|---|
| Class 1 | True Positive (TP) | False Negative (FN) |
| Class 0 | False Positive (FP) | True Negative (TN) |

### A. PERFORMANCE MEASUREMENTS

The confusion matrix is presented in the Table IV and subsequently several measurements are introduced in the following.

#### 1) OVERALL ACCURACY (OA)

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

In learning imbalanced data, the overall classification accuracy is no longer an appropriate measurement of performance. A futile classifier that predicts every instance as the majority maybe achieve very high overall accuracy.

#### 2) RECALL

$$recall = \frac{TP}{TP + FN} \tag{6}$$

*Recall* is also named accuracy of minority class (MIA) or named true positive rate (TPR).

#### 3) PRECISION

$$precision = \frac{TP}{TP + FP} \tag{7}$$

*Precision* is sensitive to the distribution of class label, whereas *recall* not. It is difficult to decide which metric is more important, therefore it mainly depends on the application domains. For example, when predicting whether someone will get cancer or whether earthquake will occur, the *recall* is expected to be much higher, because for such severe disease, doctors prefer not obtaining the wrong prediction for actual existing cancer cases. But sometimes we focus more on *precision*, for example, the rule in dubio pro reo is supported by law. We hope that it is very precise to judge whether the suspect is guilty, even if we might miss some real criminals.

#### 4) $\beta - F_{MEASURE}$ :

$$\beta - F_{measure} = \frac{(1+\beta^2) \cdot recall \cdot precision}{\beta^2 \cdot recall + precision}, \tag{8}$$

where $\beta$ is the tuning factor.

If $\beta$ is set to 1, it give the same importance to the *recall* and *precision*. Obviously it is not suitable in the imbalanced condition. We follow the rule given in [24], with $\beta = \frac{C(+|-)}{C(-|+)}$ where $C(+|-)$ is the cost of misclassifying a negative instance as positive and where $C(-|+)$ is the cost of misclassifying a positive instance as negative. Apparently $\beta - Fmeasure$ is an integrated measurement of *recall* and *precision*, where $\beta$ denotes the importance level of *recall* and *precision*.

#### 5) THE AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC)

The receiver operating characteristic curve (ROC) provides a visual scene between $TP_{rate}$ and $FP_{rate}$, and AUC is the numeric indicator of ROC, which is a better measurement than accuracy [25] and provide effective method to evaluate the imbalanced data classification [26].

### B. APPROACHES FOR SOLVING IMBALANCE

The experimental data is imbalanced, i.e., people who don't change their phones greatly outnumber that of changing phones. Standard learning methods attempt to reach overall accuracy, but not take into consideration the data distribution, thus easily bias towards the majority class. As a result, the majority class instances are well classified whereas the minority instances are heavily misclassified. In order to solve the problem, several methods are proposed as follows.

- Undersampling. Random undersampling (RUS) randomly eliminates a part of the majority class instances to maintain basic balance with the minority. The majority of training set is pre-processed by random undersampling and the minority is retained. Then the new set is input into the classifier algorithms.
- SMOTE. Random oversampling (ROS) only randomly duplicate the minority class instances to balance the class distribution. Therefore, another method that synthetic minority oversampling technique (SMOTE) was presented and applied in many articles [27], [28]. In the SMOTE method, firstly the $k$ nearest neighbors of each instance are computed and randomly select one of them, then a new artificial minority instance is generated through interpolating between the existing instance and selected nearest neighbor. Similarly, the new set is trained directly using all kinds of classifier algorithms.
- Cost-Sensitive. Most classifiers assume that the costs of misclassification are the same. In practical application, this case is not true. In the cost-sensitive method, we can revise the cost of classifier while building the training model rather than change the distribution of data set. We define the metric $C(i|j)$, which is the cost of predicting an instance to be class $i$ while it belongs to class $j$ in fact. Especially, $C(i|i)$ is set to zero due to the accuracy prediction. As the imbalanced ratio is about 1:13, i.e., the number of people who don't change their phones is 13 times than that of people who change their phones, the cost ratio can be set to 13 by inverting prior class distributions [29], [30]. Notice that Cost Ratio=$C(0|1)/C(1|0)$.

## VI. IMPLEMENTATION AND PERFORMANCE COMPARISON

The data set is sampled randomly and divided into two subsets, training set and testing set, with the ratio of 6:4. Due to the imbalance of training set, it can't be directly used as input features and meantime, as the testing is also imbalanced, overall accuracy also can't be considered as effective measurement.

### A. FOUR CLASSIFIERS

Since there is no unique classifier that is likely to perform best for all problems [31], we choose four classifier algorithms, i.e. logistic regression (LR), back propagation (BP) neural network, support vector machine (SVM) and random forest (RF), to train and predict the phone data, so that we can compare the performances. In order to test and verify the performance under different data imbalance processing, every classifier algorithm is tested in three different scenarios, namely undersampling, SMOTE and cost-sensitive.

- Logistic Regression. Logistic regression applies sigmoid function to a multinomial of the data. As the conditional likelihood for logistic regression is concave, the optimum value is searched using the method of gradient ascent. To solve the problem of over fitting, the concept of structural risk minimization is proposed and regularization term is embedded into standard logistic regression. We choose L2 regularization, i.e., ridge estimator [32], which can improve the generalization performance and prediction accuracy.
- Random Forest. It is an ensemble classifier. In the RF, a decision tree, i.e. class and regression tree (CART), is used as a weak classifier. RF was recommended highly in this paper [33], but sometimes it still lead to over fitting due to the imbalance data. In our experiment, the number of trees is set to 10, the maximum depth of

the trees is unlimited, the number of randomly chosen attributes is set to $log_2$(number of attributes).

- Support Vector Machine. It is used widely in all kinds of applications [34], [35], especially binary classification, since it adopt the principle of structural risk minimization. In this paper, radial basis function is used as the SVM kernel to train and predict.
- Enhanced BP Neural Network (E-BP). Standard BP algorithm is a multi-layer feedforward network trained according to error back propagation. The steepest descent method is used to update weights during training process.In this paper, variable learning rate is utilized to improve and speed up the convergence. Let $\eta_i = \eta/(atan(i) + 1)$, where $\eta$ is learning rate, $i$ is the $i$th iteration. Initial learning rate $\eta$ is set to 0.3 and then subsequent learning rate $\eta_i$ will get smaller and smaller with the increasing of iterations.

**TABLE 5.** Algorithm comparison over Undersampling.

| Undersampling | | | | | |
|---|---|---|---|---|---|
| Algorithm | OA | Precision | Recall | $\beta-$ Fmeasure | AUC |
| LR | 61.2% | 9.7% | 54.5% | 53.1% | 61.1% |
| | 61.2% | 9.7% | 54.5% | 53.1% | 60.8% |
| RF | 64.0% | 9.2% | 46.4% | 45.3% | 58.2% |
| | 63.8% | 9.2% | 46.8% | 45.7% | 58.3% |
| SVM | 65.5% | 9.9% | 48.1% | 47.0% | 57.5% |
| | 65.8% | 9.8% | 47.3% | 46.2% | 57.3% |
| E-BP | 54.2% | 9.3% | 63.2% | 61.1% | 61.7% |
| | 53.6% | 9.3% | 64.0% | 61.9% | 61.7% |

**TABLE 6.** Algorithm comparison over SMOTE.

| SMOTE | | | | | |
|---|---|---|---|---|---|
| Algorithm | OA | Precision | Recall | $\beta-$ Fmeasure | AUC |
| LR | 59.4% | 8.3% | 47.5% | 46.2% | 55.2% |
| | 59.8% | 8.3% | 47.1% | 45.8% | 55.1% |
| RF | 92.6% | 16.9% | 1.3% | 1.3% | 55.7% |
| | 92.3% | 13.0% | 1.7% | 1.7% | 55.9% |
| SVM | 58.8% | 8.2% | 48.2% | 46.9% | 53.9% |
| | 61.7% | 8.5% | 45.4% | 44.3% | 54.2% |
| E-BP | 80.7% | 8.3% | 17.4% | 17.3% | 53.4% |
| | 89.0% | 9.2% | 6.5% | 6.5% | 52.9% |

### B. PERFORMANCE COMPARISON

Weka [36] is chosen for our experiments and the results of four classifiers are shown in Table V,VI, and VII, respectively. Considering the phone using behavior data in this paper, to maximize the interest, the commercial companies should develop useful strategy and try not to omit the persons who want to change their phones. Hence, the *recall* index is much more meaningful than *precision* to mobile phone manufacturers and mobile communication operators. In addition, $\beta - Fmeasure$ and $AUC$ are also paid attention to as assembling measurements. Therefore, we only consider the three measurements, namely *recall*, $\beta - Fmeasure$ and $AUC$.

**TABLE 7.** Algorithm comparison over cost-sensitive.

| Cost-Sensitive | | | | | |
|---|---|---|---|---|---|
| Algorithm | OA | Precision | Recall | $\beta-$ Fmeasure | AUC |
| LR | 62.6% | 9.9% | 53.3% | 52.0% | 61.3% |
| | 62.8% | 9.8% | 52.8% | 51.5% | 60.9% |
| RF | 91.7% | 16.5% | 4.4% | 4.4% | 57.2% |
| | 91.1% | 14.2% | 5.4% | 5.4% | 57.3% |
| SVM | 66.2% | 10.0% | 47.7% | 46.7% | 57.7% |
| | 66.9% | 10.0% | 46.3% | 45.3% | 57.4% |
| E-BP | 90.8% | 13.6% | 5.7% | 5.7% | 60.1% |
| | 85.1% | 12.2% | 18.2% | 18.1% | 60.9% |

In the three tables, every classifier has two rows, where first row indicates the results without attribute selection, second row denotes that with attribute selection, using information gain ratio algorithm to select 8 effective attributes for prediction. These two prediction accuracy don't have obvious difference, but the latter classifier will be simplified, with reduced computational complexity and required storage. In addition, we can draw important conclusions as follows.

First of all, we find it very difficult to predict whether people change their mobile phones with excellent performance. The maximum of *recall*, $\beta - F_{measure}$ and $AUC$ are 64.0%, 61.9% and 61.7%, respectfully. As to why all of classifiers in different scenarios can't achieve good performance, two possible reasons are as follows. One is that, some users may decide to change their phones due to unpredictable factors, such as phone damage, phone lost, sudden events and so on. In this case, the historical behaviors of these users can not be good indicators for coming phone changing. Another reason is that, even the behavioral attributes are good indicators for this prediction task, phone changing prediction, we believe, belongs to complex non-linear classification problem, because user behavior is not only in just this prediction scope but also in many other disciplines, such as psychology. In this paper, we do provide some insights for the further research of phone changing event.

Secondly, sampling methods play important roles in prediction results and in this data set, undersampling scenario performs the best due to its sufficient size. An important reason that we often ignore the undersampling method is that many researches only use limited scale of data sets [28], [37], [38]. For example, Huang [28] conducted experiments using 7 data sets, maximum scale of which is also no more than 1,000.

Thirdly, logistic regression method with a ridge estimator is more stable than other classifiers in different scenarios, namely undersampling, SMOTE, cost-sensitive. Inversely, the performances of other classifiers are heavily dependent on the three scenarios. For example, RF and E-BP perform more worsely especially in the SMOTE and cost-sensitive scenarios, although their overall accuracies exceed 90%.

Finally, Fernández-Delgado *et al.* [33] summarized lots of existing classifiers and drew a conclusion that random forest and SVM are the top two classifiers in classification application for most data sets, but SVM performs better than RF for binary classification. However, in our case, RF and SVM did not perform very well due to the data imbalance and data intrinsic characteristics, while E-BP classifier in undersampling scenario performs the best.

## VII. CONCLUSION

In this paper, we perform the statistical analysis on the mobile users data collected and provided by one mobile operator, recognize the key attributes that are significant for phone changing event and implement four existing classification models to predict whether people change their phones. Based on that, we do obtain some novel findings and draw some important conclusions in this paper.

We have made basic statistics on user behavior, and confirmed that, the brands people used satisfy the power law distribution, the age and fee obey the log-normal distribution, and geometric distribution are well fittable for the distribution of used months, the NPTY and call duration.

The joint influence of brand and age, smart flag and age are more significant. For example, young people are more prone to changing their phones when they have low occupancy phones or feature phones, which generally be consider as low-end phones, but old people seldom make the same decision. In addition, the more phones people used, the higher probability they change their phones again; the longer someone use the phone, the much stronger intention he will change the phone.

Taking use of information gain ratio algorithm to extract efficient attributes for prediction, 8 attributes are retained, and experiment prove that the prediction performance wouldn't change too much although we discard other three attributes. Nevertheless, the classifier models can be simplified with lower computational complexity. Data imbalance problem in this case can be solved by undersampling, SMOTE and cost-sensitive, which can improve the prediction accuracy. After comparison, our proposed E-BP classifier together with undersampling method can attain the best performance in phone changing event.

Through our work, we find that, phone changing is still very hard to predict. Therefore, new attribute extraction methods and prediction models should be considered. For example, principal component analysis (PCA) can be utilized to implement attribute transformation by using the correlation of attributes. Adaboost algorithm by assembling weak classifier [39], [40], should be brought into prediction scope in our future work.

## REFERENCES

[1] A. Brown, S. Dodini, A. Gonzalez, E. Merry, and L. Thomas, "Consumers and mobile financial services 2015," Board Governors Federal Reserve Syst., Washington, DC, USA, Tech. Rep., Mar. 2015.
[2] "Total retail 2015: Retailers and the age of disruption," PWC, London, U.K., Tech. Rep., 2015.
[3] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, Jul. 2014.
[4] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken, "The big data revolution in healthcare: Accelerating value and innovation," Mckinsey Company, New York, NY, USA, Tech. Rep, Jan. 2013.
[5] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.
[6] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, 2012.
[7] S. Yu, M. Liu, W. Dou, X. Liu, and S. Zhou, "Networking for big data: A survey," *IEEE Commun. Surveys Tut.*, to be published, doi: 10.1109/COMST.2016.2610963.
[8] M. Jensen, "Challenges of privacy protection in big data analytics," in *Proc. IEEE Int. Congr. Big Data, (Bigdata)*, Jun./Jul. 2013, pp. 235–238.
[9] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data," *IEEE Access*, vol. 4, pp. 2751–2763, 2016.
[10] Y. Zhao, "Mobile phone location determination and its impact on intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 1, pp. 55–64, Jan. 2000.
[11] M. A. Tayebi, M. Ester, U. Glässer, and P. L. Brantingham, "Crimetracer: Activity space based crime location prediction," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2014, pp. 472–480.
[12] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 1414–1425, 2012.
[13] M. Ciman and K. Wac, "Individuals' stress assessment using human-smartphone interaction analysis," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2016.2592504.
[14] N. China, "Statistical communiqué of the people's Republic of China on the 2015 national economic and social development," Nat. Bureau Statist. China, Beijing, China, Tech. Rep., Feb. 2016.
[15] J. Liu, Z. Lei, L. Chen, and Y. Zhou, "Understanding how users change their mobile phones by massive data analysis," in *Proc. 7th Int. Conf. Intell. Human-Mach. Syst. Cybern. (IHMSC)*, vol. 1. Aug. 2015, pp. 232–237.
[16] H. Kimiloglu, V. A. Nasir, and S. Nasir, "Discovering behavioral segments in the mobile phone market," *J. Consum. Marketing*, vol. 27, no. 5, pp. 401–413, 2010.
[17] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. New York, NY, USA: Elsevier, 2011.
[18] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[19] S. Zhao *et al.*, "Mining user attributes using large-scale APP lists of smartphones," *IEEE Syst. J.*, to be published, doi: 10.1109/JSYST.2015.2431323.
[20] J. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," *Appl. Soft Comput.*, vol. 13, no. 1, pp. 211–221, 2013.
[21] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4626–4636, 2009.
[22] Y. Xie, X. Li, E. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5445–5449, 2009.
[23] S. H. Park and Y. G. Ha, "Large imbalance data classification based on mapreduce for traffic accident prediction," in *Proc. 8th Int. Conf. Innov. Mobile Internet Services Ubiquitous Comput. (IMIS)*, Jul. 2014, pp. 45–49.
[24] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Mining Knowl. Discovery*, vol. 17, no. 2, pp. 225–252, 2008.
[25] C. X. Ling, J. Huang, and H. Zhang, "Auc: A statistically consistent and more discriminating measure than accuracy," in *Proc. IJCAI*, 2003, pp. 519–524.
[26] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.
[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[28] P. J. Huang, "Classification of imbalanced data using synthetic over-sampling techniques," M.S. thesis, Dept. Sci. Statist., Univ. California, Los Angeles, CA, USA, 2015.

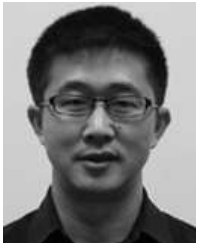
[29] N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2010, pp. 1–8.
[30] J. Bian, X.-G. Peng, Y. Wang, and H. Zhang, "An efficient cost-sensitive feature selection using chaos genetic algorithm for class imbalance problem," *Math. Problems Eng.*, vol. 2016, May 2016, Art. no. 8752181.
[31] S. Ali and K. A. Smith, "On learning algorithm selection for classification," *Appl. Soft Comput.*, vol. 6, no. 2, pp. 119–138, 2006.
[32] K. Månsson and G. Shukur, "On ridge parameters in logistic regression," *Commun. Statist.-Theory Methods*, vol. 40, no. 18, pp. 3366–3381, 2011.
[33] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems," *J. Mach. Learn. Res*, vol. 15, no. 1, pp. 3133–3181, 2014.
[34] P. Janik and T. Lobos, "Automated classification of power-quality disturbances using SVM and RBF networks," *IEEE Trans. Power Del.*, vol. 21, no. 3, pp. 1663–1669, Jul. 2006.
[35] Y. Lin *et al.*, "Large-scale image classification: Fast feature extraction and SVM training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1689–1696.
[36] *Weka 3: Data Mining Software in Java*, accessed on Oct. 2016. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html
[37] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Classification of imbalanced data by combining the complementary neural network and smote algorithm," in *Proc. Int. Conf. Neural Inf. Process.*, 2010, pp. 152–159.
[38] V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowl.-Based Syst.*, vol. 25, no. 1, pp. 13–21, 2012.
[39] J. Thongkam, G. Xu, and Y. Zhang, "AdaBoost algorithm with random forests for predicting breast cancer survivability," in *Proc. IEEE Inter. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 3062–3069.
[40] W. Hu, W. Hu, and S. Maybank, "Adaboost-based algorithm for network intrusion detection," *IEEE Trans. Syst., Man, B, Cybern.*, vol. 38, no. 2, pp. 577–583, Feb. 2008.

**QINGLI MA** received the B.S. and M.S. degrees in electronic engineering from PLA Information Engineering University, Zhengzhou, China, in 2005 and 2009, respectively. He is currently pursuing the Ph.D. degree in electronic engineering with the University of Science and Technology of China, Hefei, China. His research interests include speech coding, modulation recognition and big data.

**SIHAI ZHANG** received the Ph.D. degree from the Department of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China, in 2006. He has been with the PCNSS Laboratory, Department of Electronic Engineering and Information Science, USTC, since 2009. He is currently an Assistant Professor of Electronic Engineering with the Department of Electronic Engineering and Information Science, USTC. His research interests include wireless networks and intelligent algorithms. He has authored or co-authored over 60 technical papers, such as the IEEE TETC, the IEEE TVT, MONET, and WPC. He initiated the research field of wireless big data in 2014. He has participated in projects, including the National Science Foundation of China for Machine Type Communications, Key Program of the National Natural Science Foundation of China for Wireless Big Data. He has served over 15 international conferences as a member of organizing committee, TPC member or a reviewer, such as publication chair for SEAL 2006 and WCSP 2014. In 2016, he has co-chaired special sessions on Wireless Big Data in WCSP 2016, Machine Type Communications in WPMC 2016 and guest editor of special issue on Wireless Big Data for JCIN.

**WUYANG ZHOU** received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 1993, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 1996 and 2000, respectively. He is currently a Professor of Wireless Communication Networks with the Department of Electronic Engineering and Information Science, University of Science and Technology of China. He participated in the National 863 Research Project Beyond Third Generation of Mobile System in China (FUTURE Plan). He has been a Task Director in many projects, including innovative wireless campus experimental networks research on high frequency networking technologies, and research on transmission and networking technologies in satellite mobile communications. His research interests include co-operative communications, radio resource management, wireless networking, satellite mobile communications, and underwater acoustic communications.

**SHUI YU** (SM'12) is currently a Senior Lecturer (equivalent to Associate Professor in North America) with the School of Information Technology, Deakin University, Australia. He is a member of AAAS, the Vice Chair of Technical Subcommittee on Big Data Processing, Analytics, and the Networking of IEEE Communication Society, and a Big Data Standardization Committee. He has authored two monographs and edited one book, over 150 technical papers, including top journals and top conferences, such as the IEEE TPDS, the IEEE TC, the IEEE TIFS, the IEEE TMC, the IEEE TKDE, the IEEE TETC, and the IEEE INFOCOM. His research interest includes big data, networking theory, cybersecurity, and mathematical modeling. He initiated the research field of networking for big data in 2013. His h-index is 20. He actively serves his research communities in various roles. He is currently serving the editorial boards of IEEE Transactions on Parallel and Distributed systems, IEEE Communications Surveys and Tutorials, IEEE Access, and a number of other international journals. He has served over 50 international conferences as a member of organizing committee, such as publication chair for IEEE Globecom 2015 and IEEE infocom 2016, TPC Co-Chair of the IEEE BigDataService 2015, IEEE ATNAC 2014 and 2015.



**CHONGGANG WANG** (F'17) received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2002. He is currently a Member Technical Staff/Senior Manager of InterDigital Communications, Conshohocken, PA, USA. His current research interests include Internet of Things, mobile communication and computing, and big data management and analytics. He is as a Distinguished Lecturer of the IEEE Communication Society from 2015 to 2016. He is the founding Editor-in-Chief of the IEEE Internet of Things Journal from 2014 to 2016 and an associate Editor-in-Chief of the IEEE Transactions on Big Data. He has been on the editorial board for several journals including IEEE Access. He is also on the advisory board of IEEE-The Institute.