Research of Indicator System in Customer Churn Prediction for Telecom Industry

Qiu Yihui*
School of Economics and Management
Xiamen University of Technology
Xiamen, China
qiuyihui@xmut.edu.cn

Zhang Chiyu
School of Economics and Management
Xiamen University of Technology
Xiamen, China
zcy94@outlook.com

***Abstract*—Curse of dimensionality will occur if effective dimensionality reduction method is not applied in machine learning, especially in the telecom field. The existing researches on customer churn are still lack of a set of scientific, system theory and method and the single models methods for customer churn prediction also are unable to completely meet application needs. Therefore, it is important for theoretical and practical contribution to explore and study the customer churn prediction. Based on the substantive characteristics of customer churn in telecommunication, the indicator system of customer churn in telecommunication are studied in this paper. Firstly, we proposed a feature selection method based on pruning technique, which is called feature selection method based on orientation ordering pruning Method (OOPM). According to this algorithm attribute selection problem can be replaced by the pruning question of classifier combination and we structure an indicator system of customer churn. Secondly, in order to explore high-order statistical information in the properties, a feature extraction method based on Random Forest and Transduction (FE_RF&T) is proposed to extract multiple features from customer data. Experiments on the real data of telecom enterprise show that the feature selection method based on OOPM has more advantages than the feature selection method based on Random Forest, and compared to the PCA method the FE_RF&T method improves the performance of learning machine effectively.**

***keyword*—customer churn prediction, indicator system, feature selection, feature extraction**

I. INTRODUCTION

With the prosperity and development of the mobile communications industry market, the competition of mobile operators become intensified. According to the latest cost accounting structure of telecom, the cost of losing an existing customer is 5 times as much as the profit that a new customer can bring about[1]. The key to accomplish the development pattern, which for the sake of both scale and economies, is to control customer churn. The researches on customer churn prediction have come out with a lot of achievements on traditional statistics methods and machine learning methods. However, presenting original datasets is a great obstacle for using machine learning methods to predict telecom customer churn. Those datasets have many problems, such as sourcing being numerous, relationship among attributes being complicated, class imbalanced and so on, which will seriously impact the efficiency and effectiveness of prediction models. It is very important to perform feature selection and feature extraction as to get the improvement of model.

Feature selection, which includes filter model and wrapper model, is a technique for selecting appropriate subset of original features, which maintains the essential information of features. Feature extraction transforms the feature space of data usually by using a projection matrix, which will map the high-dimensional data to low-dimensional space of data. Fisher's ratio, F-score[2] and Chi-Square[3][4] are the filter models of feature selection. They select an optimal subset of feature by evaluating the statistics indexes between an input variable and the classes. But those methods cannot evaluate redundancy of each feature, the selected features may be full of redundancy, and it will reduce efficiency of classifications and increase risk of overfitting. Minimum redundancy and maximum relevance (mRMR) [2][5][6], widely used method in recent years, is the method which calculates mutual information to minimizes redundancy of selected feature and maximize relevance between features and classes. Mutual information feature selector (MIFS) and normalized mutual information feature selection (NMIFS)[7] are based on the calculating of mutual information. Those methods effectively deal with the redundancy problem of features, but the computing complexity soar with an increasing number of features and this algorithm will be inefficient. Kernel-Penalized Support Vector Machine (KP-SVM)[8] and Boosting-SVM[9] are the wrapper methods of feature selection. Wrapper methods is more efficient and determines simultaneously a classifier with high classification accuracy and an adequate feature subset. Yet, they depend on some special data type and not suit to telecom customer churn. The most widely used dimensionality reduction methods include principal component analysis (PCA)[10][11] method and linear discriminant analysis (LDA)[12][13] method. Both methods are induction learning methods which try to get a learning machine based in empirical risk minimization principle to make the best prediction to all the feature data. Adaptive slow feature discriminant analysis (ASFDA)[14] method and fuzzy local maximal marginal embedding (FLMME)[15] for feature extraction method seek projections

to minimize the difference between within-class variation and between-class variation based on the idea of maximize margin criterion (MMC). But those feature extraction methods also encounter the ill-posed problem and inefficiency in their real-word application, especially on telecom customer churn prediction. Because the drawbacks of present feature selection and extraction methods, we consider designing a more economic method of feature extraction. This paper proposed a feature extraction method based on Transduction[16][17][18] and Random Forest[19][20], and applied new method on the customer churn prediction model in telecom industry. Experimental results show that compared to PCA proposed method which makes full use of information of training data and test data improves the performance of prediction model.

II. RANDOM FOREST

Random Forest is an ensemble of decision trees that are individually created by means of drawing randomly the exactly number of samples from the original training data. These decision trees include 3 types of nodes that are root nodes, internal nodes and end nodes. As a real tree, decision tree has only one root node which is the set of whole data. Each internal node is a split question to split all the arrival samples by a certain feature. Every end node is a set of labeled data. A judge rule concludes from the root node to an end node. Decision tree method applies from-top-to-bottom greedy algorithm. Each internal node chooses the best split feature to divide the arrival samples into 2 parts or more. Follow this procedure the decision tree classifies all the training samples.

The key matter is to select a preferable split feature. There are many criterions for choosing split features such as information plus, Gini index and so on. Corresponding to different selection methods, there are many algorithms, for example, Iterative Dichotomizer 3 (ID3), C4.5 and Classification And Regression Tree (CART). In this paper we utilize C4.5 as base decision tree method whose split criterion is Gini index. Random Forest repeats the procedures above to construct a combination of multiple decision trees.

Given a collection of N samples with Q features, we intend generate M trees in forest. For each tree, N training samples are draw at random (with put back) from the original collection to create a decision tree. This process is so-called bagging. After M times bagging repeat, we can get M decision trees. During the growing of a single tree, we choose a better split feature from q (q<Q) features that are picked randomly at every internal node instead of from the whole Q features. By doing this, we can enhance the discrepancy of trees to prove the general error. Every tree is growing to the limitation without pruning. When the forest is done, each tree will make its own conclusion for a new-coming sample. We can get the final label through average votes of all trees. Compared to other classifiers, Random Forest is unlikely to over-fit the training samples if the number of trees is large enough. It is proved that the upper bound for general error of Random Forest is given by $\overline{\rho}(1-s^2)/s^2$, where $\overline{\rho}$ is the mean value of correlation among trees and s is the strength of the individual decision trees in the forest.

In this paper, we use Random Forest as the elementary churn prediction model. Based on this model, we go further into the discussion of dimensional reduction.

III. FEATURE SELECTION METHOD BASED ON ORIENTATION ORDERING PRUNING METHOD

A. Feature selection method based on OOPM

G. Martinez-Munoz and A. Suarez[21-23] proposed an feature selection method based on forward recursive technique, Orientation Ordering Pruning Method (OOPM). In each iteration, OOPM selects a basic classifier from the original group of classifiers into the optimal subset, so as to make the target function to approach the optimal solution. Since the OOPM is a forward recursive method, we apply it to the feature selection of customer churn prediction, and transform the feature selection into the pruning of ensemble classifier. The basic classifier OOPM required is a simple classifier which hardly appears overfitting. In the feature selection of telecom customer churn, therefore, we structure the most simply Signal to Noise Ratio (SNR) classifier for each attribution, then rank the classifiers of whole attributions by OOPM method, and select the optimal subset of attribution.

There are some continuous attributions, e.g. monthly expenditure and call duration, and some discrete attributions, e.g. brands and customer city. And there are many distinctions for those two types data to structure classifiers.

Firstly, the method which structures SNR classifier for continuous data is introduced. For each continuous attribution, s_i represents the SNR of the ith attribution, the definition of it as following Eq.1:

$$s_i = \frac{\mu_+^{(i)} - \mu_-^{(i)}}{\sigma_+^{(i)} + \sigma_-^{(i)}} \tag{1}$$

Where $\mu_+^{(i)}$ can be seen as the mean of ith attribution of all positive samples, $\mu_-^{(i)}$ as the mean of ith attribution of all negative samples, $\sigma_+^{(i)}$ as the variance of ith attribution of all positive samples, $\sigma_-^{(i)}$ as the variance of ith attribution of all negative samples. For each sample, $x(i)$ represents the value of ith attribution of x th sample. The decision function of classifier which based on each attribution as following Eq.2:

$$f_i(x) = sign(s_i(x(i) - b_i)) \tag{2}$$

Where b_i can be seen as the mean of ith attribution of all samples and defined as following Eq.3:

$$b_i = (\mu_+^{(i)} + \mu_-^{(i)})/2 \tag{3}$$

Then, according to the steps above, classifier for all continuous attributions are structured.

We will introduce some discriminating principles of single attribution classifiers of continuous attributions. Given

following characteristics of *i*th attribution: the mean of all positive samples is greater than it of all negative samples, namely $s_i > 0$. We can draw the following conclusion: when the *i*th attribution value of a sample is greater than the mean of all samples, namely $x(i) > b_i$, we are more convinced that it is a positive sample, vice versa. So every awaiting sample is compared its *i*th attribution value with the mean of *i*th attribution in the first place. If its value is greater than mean, it will be discriminated as positive class, on the contrary, it will be discriminated as negative class.

For the discrete attributions, we use the simple counting method, which is attribution value oriented method, to structure single attribution classifiers. Suppose the *i*th discrete attribution has m different kinds of attribution value, $i_k, k \in \{1, 2, \cdots, m\}$, then for each attribution value i_k, counting the number of positive sample as i_k^+ and the number of negative sample as i_k^-, $b_{i_k} = i_k^+ - i_k^-$, so the decision function of classifier of discrete attributions is as follows Eq.4:

$$f_i(x) = sign(b_{x(i)}) \tag{4}$$

Where $x(i)$ represents the value of *i*th attribution of *x*th sample. According to the steps above, classifiers of all discrete attributions are structured.

Then, we introduce some discriminating principles of single attribution classifiers of discrete attributions. Suppose some attribution value of *i*th attribution has some following characteristics of: the number of all positive samples which are the attribution value is greater than it of all negative samples, namely $b_{i_k} = i_k^+ - i_k^- \geq 0$, there are more opportunities for positive samples to have this attribution value, and we can draw a following conclusion: when the *i*th attribution value of a sample is this special value, we are more confident to believe that it is a positive sample, vice versa. So an attribution value of an awaiting sample is the value of the majority positive samples, it will be discriminated as a positive class, on contrary, it will be discriminated as a negative class.

If some single attribution classifier shows a good classification performance, it explains that this attribution importance of decision, namely its distribution can describe the distributions of the positive and the negative in some degree and it is the attribution which will be selected by us.

Back to the prediction of customer churn, for a given dataset:

$$Train = \{(x_i, y_i) \mid x_i \in R^n, y_i \in \{1, -1\}, i = 1, 2, \cdots, l\}$$

Where l is the amount of samples, n is the amount of attributions.

Using the methods above to structure single attribution classifiers of all attributions, including continuous attributions and discrete attributions, and making up a group of classifiers, then selecting a subgroup which has a better classification performance via OOPM, so the corresponding attributions of the classifiers in this subgroup is the optimal attributions which is selected by feature selection of OOPM. And the detail of algorithm is as follows:

For each attribution, using above method structures single attribution classifiers, then get n single attribution classifiers: $H = \{h_t(x) \in \{1, -1\} : t = 1, \cdots, n\}$ and make up a group of classifiers.

To discriminate all training samples by using every single attribution classifier h_t, and calculating its l-dimensional signature vector c_t, which the element is $c_{ti} = 2I(h_t(x_i) = y_i) - 1, i = 1, 2, \cdots, l$. When this classifier correctly classifies the *i*th sample x_i, $c_{ti} = +1$, on the contrary $c_{ti} = -1$.

Calculating the average signature vector c_{ens},

$$c_{ens} = \frac{1}{T} \sum_{t}^{T} c_t \tag{5}$$

Calculating

$$\lambda = -o \cdot c_{ens} / \left| c_{ens} \right|^2 \tag{6}$$

where o is the vector which gets along with the diagonal of first quadrant, get Eq.7:

$$c_{ref} = o + \lambda c_{ens} \tag{7}$$

Calculating the angle between c_t, which is the signature vector of every single attribution classifier, and c_{ref}, and ranking the results with an ascending order.

Finally, selecting the attributions which are the corresponding ones of top classifiers as the final subset of attributions, or selecting the attributions which are the corresponding ones of classifiers whose angle between c_{ref} and itself is less than $\frac{\pi}{2}$ as the final attribution subset.

B. Experiment of feature selection

The data for the following experiment is the real telecom industry dataset which has pre-processed and cleaned. We will introduce relevant pre-processed works in later section. The dataset has 12688 samples which including 6639 positive class samples, 6049 class samples and 44 attributions. In the experiment, the feature selection method based on OOPM, which is proposed in this paper, compares with the representing method of wrapper, which is based on Random Forest.

In the feature selection method of Random Forest, the features are ranked by importance, then using forward recursive feature selection method: adding an attribution, which is selected from the set of whole attributions, into a new subset to structure a classifier and evaluating the performance of this subset.

In the feature selection method of OOPM, the features are ranked by the angle between signature vector of single attribution classifier and reference vector, then successively

adding attributions into the subset of attributions to structure classifier and evaluating the performance of subset according to the forward recursive method.

All experiments have been realized in MATLAB and Random Forest is realized by Fortran language, the amount of predictors sampled for splitting at each node be set as $mtry = \sqrt{p}$ where p is the number of whole attributions, the amount of trees grown be set as 2000.

The experiment uses 10-fold cross validation method, namely successively using each 9-fold datasets as the training dataset, the rest 1-fold dataset as the testing dataset. On the training dataset, respectively using the two feature selection method above to select features, then applying the new subset of attributions to structure classifier, and evaluating the classifier by using testing dataset. Using the forward recursive selection method, recording the process of attribution increasing and the prediction accuracy of each attribution subset, and repeating it 100 times.

In order to objectively evaluate the performances of those two method, Random Forest feature selection method and OOPM feature selection method, we use k-Nearest Neighbor algorithm (KNN) as a basic classifier, and do evaluation by the mean of predication accuracy. Since KNN is sensitive for the distance measurement, the appropriate feature selection can get a better performance and get more obviously visual sense. We adopt the city block distance as the distance measurement, and K is 7. At the same time, we use the following indexes to evaluate the performance of feature selection:

(1) The accuracy means of 100 times 10-fold cross validation of each different attribution subset

(2) For the different size subset of attribution, calculating the ratio of repeated attribution in each 10-fold cross, and get the mean of 100 times, which be called as index of repetition rate. This index is bigger the influence, which of the feature selection by training dataset, is smaller and more stable.

At first, we compare the performance of those two feature selection methods in each ranking of attributions. According to the attribution ranking which is ranked by Random Forest method, then adding an attribution by one, using KNN to predict customer churn, and getting the performance of prediction models on hit rate, coverage rate, average accuracy and lift coefficient.

According to the ranking of attribution which is ranked by OOPM method, adding an attribution by one, using KNN to predict customer churn, and getting the performance of prediction models on hit rate, coverage rate, average accuracy and lift coefficient.

We can discover that the generalization performance of predication model is improved by feature selection method, and that the highest accuracy doesn't appear on the model which contains whole attributions, while appear on the model which contains less attributions. It indicates that there are full of redundancy and irrelevant information in the attributions and emphasize the necessary of feature selection. With the increasing number of attribution, the hit rate, coverage rate, average accuracy and lift coefficient of prediction model present following tendency: firstly, the indexes gradually increase, reach the highest values when the number of attribution reach around 20, then gradually decrease. The main reason of this phenomenon is as follow: the performance is poor due to the less information when few attributions are selected; the performance is the best when the selected attributions contain sufficient information for prediction; the performance decreases due to more redundancy information when the size of attribution subset greater than the best threshold.

We are going to compare the second evaluation index of the feature selection method, repetition rate. We study the repetition status of each attribution subset in the processing of forward recursion, and get repetition rate. As can be seen from the graph, when the size of the selected attributions is quite small, the repetition rate is also little in each 10-fold dataset which demonstrates that the two feature selection methods are reliable to the training dataset in some extent. When the size of attribution subset is middle, the feature selection method based on OOPM has more advantages than the Random Forest, which indicates that the OOPM method is less dependent to the training datasets, therefore it is more stable and reliable.

We have made a further selection for the attributions of the customer churn prediction model. The attribution subset will be selected which is the highest accuracy in 10-fold cross validation and the high repetition rate. We select the subset of attributions via the following steps: ranking attributions by their repetition rates, selecting the top attributions by the professional knowledge. Finally, there are 22 attributions selected, the ultimate attribution list is displayed in table 1.

TABLE I.
FINAL ATTRIBUTION LIST

ID	attribution description
1	Brands type
2	User credit
3	Halting duration
4	Variation of daily expenditure
5	Duration of daily calls
6	Effective Duration of package
7	Times of halting
8	Degree of member points
9	Degree of last moth expenditure
10	The mean of expenditures in last three months
11	Degree of arrearage
12	The number of massages
13	Degree of GPRS flow
14	Variation of expenditure
15	Expenditure degree of roaming calls
16	Expenditure degree of voice calls
17	Degree of classing times
18	Duration of local calls
19	The number of calling recipients
20	Service type of call center
21	Sensitivity degree of expenditure
22	Roaming type

IV. Feature Extraction method based on Random Forest and Transduction

A. Proximity Matrix

Random Forest is a combination of multiple decision trees, therefore we can apply the frequency of each 2 samples appear on the same end node of single tree, which means the chances of 2 samples belong to the same category, as the measurement of proximity of these 2 samples. Following this illumination, we are going to introduce the Proximity Matrix of Random Forest which is called *Prox* matrix.

Given a training set with N samples, we originally create a N*N zero *Prox* matrix whose single row represents an individual sample, then use every tree to run through all the samples, each sample will arrive at a certain end node of this tree; take 2 samples n and k as an example, if they end up on the same end node, we add 1 to the element of row n and column k; repeat this procedure till all the trees done, divide every element of matrix with the number of trees to normalize the matrix, and then we get the final *Prox* matrix which is a symmetry matrix with diagonal element 1. The elements of matrix can be defined as the proximity between 2 samples. All job above will be done soon after the Random Forest done. It is safe to tell from the matrix that the more samples a category has, the more 1 element will be on the corresponding rows. Furthermore, there are reasons for us to believe that the samples with more zero elements on their rows have much less proximity with other samples.

B. Dimensional Scale Transfer based on Proximity Matrix

Prox matrix can be treated as the mapping from original attribute space to proximity space, which uses proximity between each 2 samples to describe the relation of data. During this procedure, Random Forest is applied as a mapping tool. With its help we seek for a further transformer in the proximity space.

We apply below transform on *Prox* matrix which leads to the normalized *Cv* matrix as follows:

$$cv(n.k) = \frac{1}{2}(prox(n,k) - prox(n,-) - prox(-,k) + prox(-,-)) \quad (8)$$

Where $prox(-,k)$ is the mean of the first coordinate, $prox(n,-)$ is the mean of the second coordinate, $prox(-,-)$ is the mean of both coordinates. Just like the *Prox* matrix, *Cv* matrix is also symmetry and bounded matrix. Let $\lambda(j)$ to be the eigenvalue of *Cv* matrix, which is ranged by its magnitude in descending order, has eigenvector $v_j(n)$, then we can represent vector $x(n)$ as follows:

$$x(n) = (\sqrt{\lambda(1)}v_1(n), \sqrt{\lambda(2)}v_2(n), \cdots) \quad (9)$$

Where $\sqrt{\lambda(j)}v_j(n)$ can be seen as the jth scaling coordinate.

Multidimensional scaling is to approximately represent vector $x(n)$ by the preceding scaling coordinate, this procedure can be performed as the following steps: first range all the eigenvalues of *Cv* matrix in descending order, and then utilize the preceding eigenvalues and their eigenvectors to represent vector $x(n)$. Usually the first and the second scaling coordinates can provide sufficient useful information of data, and in a more general situation 3 scaling coordinates or 4 scaling coordinates can describe data pretty well.

Random Forest can be treated as a favorable mapping tool to mapping samples from original attribute space to proximity space. But it is noticeable that the linear transform varies with the dimensions of proximity matrix, i.e. changing when the number of samples changes. Because of this particular trait we cannot get the scaling coordinates of test data after the Random Forest was done. How to make the full use of training set and test set to get the extracted features of test data by pursuing the internal relation between training samples and test samples. This is called Transduction that is going to be introduced following.

C. Transduction

Transduction, which is different from induction and deduction, is also called transductive learning. For induction learning, the learning machine induces a discriminant function based on available labeled samples to get a small expectation discriminant error of potential distribution estimation as much as possible, i.e. to get a classifier. Furthermore, deduction uses the classifier to label the unknown sample, while transduction learning is an approach to identify certain unknown samples directly from known samples.

Random Forest can handle high-dimensional data and torrent with the noise and its decision trees provides a new way to explore proximity of data. Therefore, this paper developed a new feature extraction which applies proximity matrix of Random Forest to express the proximity between training data and test data to perform multidimensional scaling transform.

D. Feature extraction based on Random Forest and Transduction, FE_RF&T

This paper applies Random Forest as a kernel mapping from original feature space to another feature space, and uses transduction to express the relation between training data and test data, and then gets the multidimensional scaling transform coordinates which combined distribution information of both training data and test data, is called feature extraction method based on Random Forest and Transduction. Steps of FE_RF&T are as follows:

(1) Use training samples to build Random Forest, and

save it.
(2) Run both training samples and test data through Random Forest, all data will end up on certain end nodes of trees.
(3) Get proximity matrix of all data including training samples and test data.
(4) Perform multidimensional scaling transform on the matrix to get extracted features.

After Random Forest is done, we do not need to build new RF model to get the proximity matrix when new unlabeled samples arrive. Based on the above procedures, feature extraction method not only absorbs all the information of training samples but also makes full use of distribution information of test data to assess the potential distribution of the whole data space. The next experiments show the good performance of proposed feature extraction method.

V. DATA DESCRIPTION

A. Data characteristic

In 2002, China Mobile Communications Corporation started to establish their business analysis system in order to cope with the increasing market competition, and its' Business & Operation Support System (BOSS) has generated a great number of precious information resources. To provide timely, precise and scientific decision supports for decision makers at all levels, China Mobile integrated the information of relevant support systems, structured business analysis platform and intelligently processed information. The information, which contains in the data warehouse and data mart, will show for users of business analysis system by report form, Online Analytics Processing (OLAP), ad-hoc query and data mining. Based on their own special management and business needs, departments establish data marts which fit for their own applications on the data warehouse, of which the customer churn analysis is an important part. In the telecom business analysis system, the main algorithms of customer churn analysis are decision tree, neural network and logistic regression analysis. In this section, we will introduce some definitions of customer churn and churn abnormal statues.

Customer churn (or off-net customer) is the client who once used the mobile communication service of a telecom company and no longer used the service of the telecom company since a point of time.

The customer churn generally includes two categories: voluntary loss and reluctant loss. The reluctant loss occurs when service is forced to halt by telecom operations because of the arrearage of customers or the customers failing to fulfill their obligations. There are two reasons for the voluntary loss. The first one is caused by the subjective reasons, such as relocation, death or payment adjustment, which disables customer to afford the communication fee. The other reason is that the customer voluntarily abandons serves due to more preferential policies of opponents or dissatisfaction of presenting service.

When the customers are losing, they appear different abnormal characteristics. According to particular situations, China Mobile has done some deeply analyses and categorizes customer statues into the following types:

(1) normal customer;
(2) dismantling customer: include arrearage and initiative account cancellation;
(3) disappearing customer: no call and massage in last two months;
(4) silence customer: calls and massages less than 5 in last month;
(5) low-cost customer: the mean of calls and massages less than 10.

Except normal customer, the others are all suspected to relate with customer churn in some degree. General speaking, dismantling customer is losing absolutely, and it is an indication of possible customer churn that the calls or telephone fees recently plunged. So as said above, disappearing customer, silence customer and low-cost customers are called losing abnormal statue by telecom industry.

Combining the specific condition of customers with the general definition of telecom industry, the company provides a definition of churn customer:

(1) The customer initiatively closes account or experiences an arrearage more than 30 days
(2) The customer didn't consume, no call or massage, in the past two months.

The data we use in this paper derives from the business analysis system of China Mobile. Based on the definitions above, we extract the available customers to mark. The churn customer and normal customers are respectively categorized into negative class and positive class.

B. Data sampling

Since the actual proportions of customer churn is around 3%, if we randomly sample all customer data without replacement, the available churn customer samples could be only 3% of all data. It may lead to imbalance for sample distribution, and this dataset would cause overfitting of the predication models. Consider an extreme situation, the model may directly distinguish all samples into positive class, while the accuracy can still exceed 95%. Hence, we should not randomly sample the whole data without replacement in our experiments. This paper uses the following methods to extract data: firstly, all customer data would be classified as different classes by the above definitions of churn customer, the customer data categorize as positive sample dataset (normal customer) and negative sample dataset (churn customer); secondly, we respectively and equally sample data in positive sample dataset and negative sample dataset. It is an efficient method to avoid the imbalance distribution of data and the overfitting problem, so to acquire more useful information of churn customer.

16920 samples are selected by the steps above, including normal customer, churn customer, testing telephone number and group customer. Among the enormous customer data, the free customer (or group customer) cannot represent normal customer behaviors and is easier to become a noise for category and prediction. Therefore, we select all customer data

above except group customer data. To ensure all research data contain complete information, we select the customers who have been using their accounts more than 90 days.

C. Outlier detection

There are many anomalous samples, which are produced by various system errors and artificial deviation, in the numerous telecom industry datasets. Since those anomalous samples impact the accuracy of churn customer prediction in some degree, it is a priority requirement that we need by structuring perdition models to preprocess data and to delete anomalous samples. The Writer has proposed detection method of anomalous sample, which based on Random Forest method. The simulation experiments showed that compared with the other two anomalous detection methods, which based on distance measure, this method improves the accuracy of the model better, has a more robust result as well, and remarkably decreases the computing time facing large-scale datasets. In this paper we use this method to detect anomalous samples of prediction data.

According to the experiment result, we can discover that the outlier measurement of most sample are less than 4, so 4 is regarded as the threshold value of outlier measurement and any samples more than 4 are deleted. After deleting such kind of samples, the dataset remains 12688 samples with 6639 positive samples and 6049 negative samples.

VI. Application of Feature Extraction Method Based on Random Forest and Transduction in the Customer Churn Prediction

The telecom data for building Random Forest model is selected randomly from business analysis system. Description of data is showed in table 2.

TABLE II
DISCRIPTION OF DATA

	Whole set	Positive samples	Negative samples
All	12688	6639	6049
Train set	7613	3983	3630
Test set	5075	2656	2419

First, we use training samples to build an elementary churn prediction model without feature extraction. Table 3 shows the accuracy of different churn prediction models using the

TABLE III
ACCURACY OF DIFFERENT MODELS

	model	accuracy
Train set	C4.5	73.2%
	NN	67.05%
	Logistic	65.1%
	Random Forest	88.6%
Test set	C4.5	69.0%
	NN	66.2%
	Logistic	64.9%
	Random Forest	80.22%

Now we perform feature extraction on all data to get the extracted features of training samples and test data, and then use the training samples that perform dimensional reduction to rebuild a Random Forest model. The accuracy comparison of models before and after the projection is showed as table 4.

TABLE IV
ACCURACY OF MODELS BEFORE AND AFTER PROJECTION

	model	Before	After
Train set	C4.5	73.2%	75.9%
	NN	67.05%	70.3%
	Logistic	65.1%	67.8%
	Random Forest	88.6%	91.2%
Test set	C4.5	69.0%	70.2%
	NN	66.2%	66.4%
	Logistic	64.9%	65.7%
	Random Forest	80.22%	83.5%

We can see from table 4 that the proposed method improved the performance of all models and is proved effective.

VII. Conclusion

This paper proposed a feature selection method based on OOPM. According to this algorithm, we proposed an indicator system for customer churn prediction of telecom industry. Then in order to explore high-order statistical information in the properties, a new feature extraction method based on Random Forest and Transduction is proposed, this method used

the proximity matrix of Random Forest to transfer the information of test set into training process and perform multidimensional scale to get the extracted features. Experiments on the real data of telecom enterprise show that the feature selection method based on OOPM has more advantages than the feature selection method based on Random Forest, and compared to the PCA method the FE_RF&T method improves the performance of learning machine effectively. It is safe to make a conclusion from the above interpretation that feature selection method based on OOPM and feature extraction method based on Random Forest and Transduction are two effective methods of feature selection and extraction and the indicator system for customer churn prediction is also effective.

VIII. REFERENCES

[1] Webb A R, Copsey K D. Statistical Pattern Recognition, Third Edition[M]//Statistical Pattern Recognition. Wiley & Sons, 2010:183–190.

[2] Idris A, Khan A, Lee Y S. Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification[J]. Applied Intelligence, 2013. 39(3):659-672.

[3] Y. Huang, B.Q. Huang, M.T. Kechadi: A New Filter Feature Selection Approach for Customer Churn Prediction in Telecommunications[C]. Proceedings of the 2010 IEEE IEEM, 2010: 338-342

[4] Mesleh A M A. Chi square feature extraction based SVMs Arabic Language Text Categorization system[J]. Journal of Computer Science, 2007, 3(6):430-435.

[5] Hanchuan P, Fuhui L, Chris D. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(8):1226-38.

[6] Adnan Amin, Faisal Rahim, Imtiaz Ali, etc.: A Comparison of Two Oversampling Techniques (SMOTE vs MTDF) for Handling Class Imbalance Problem: A Case Study of Customer Churn Prediction[J]. New Contributions in Information Systems and Technologies, 2015, vol 1: 215-225

[7] Estévez P A, Michel T, Perez C A, et al. Normalized mutual information feature selection[J]. IEEE Transactions on Neural Networks, 2009, 20(2):189-201.

[8] Maldonado S, Weber R. A wrapper method for feature selection using Support Vector Machines[J]. Information Sciences, 2009, 179(13):2208-2217.

[9] Wang C W, You W H. Boosting-SVM: effective learning with reduced data dimension. Appl Intell[J]. Applied Intelligence, 2013, 39(3):465-474.

[10] Morris J, Coombes K, Koomen J K, et al. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum[J]. Bioinformatics, 2005, 21(9):1764-1775.

[11] Kuncheva L I, Faithfull W J. PCA feature extraction for change detection in multidimensional unlabeled data[C]// Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012:1140-1143.

[12] Burges C J C. GEOMETRIC METHODS FOR FEATURE EXTRACTION AND DIMENSIONAL REDUCTION[M]// Data Mining and Knowledge Discovery Handbook. Springer US, 2005:59-91.

[13] Ghassabeh Y A, Rudzicz F, Moghaddam H A. Fast incremental LDA feature extraction[J]. Pattern Recognition, 2015, 48(6):1999-2012.

[14] Gu X, Liu C, Wang S, et al. Feature extraction using adaptive slow feature discriminant analysis[J]. Neurocomputing, 2015, 154:139-148.

[15] Zhao C, Lai Z, Liu C, et al. Fuzzy local maximal marginal embedding for feature extraction[J]. Soft Computing, 2012, 16(1):77-87.

[16] Chen Y S, Wang G P, Dong S H. A Progressive Transductive Inference Algorithm Based on Support Vector Machine[J]. Journal of Software, 2003, 14(3):451-460.

[17] Qiu D H, Chen C B, Jin X J. Confidence Support Vector Machine Based on Algorithmic Theory of Randomness and its Application on Signature Verification[J]. Mini-micro Systems, 2004, 25(12):2131-2134.

[18] Yang L I, Fang B X, Guo L, et al. Network anomaly detection based on TCM-KNN and genetic algorithm[C]// ACM Symposium on Information, Computer and Communications Security, ASIACCS 2007, Singapore, March. 2007:13-19.

[19] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5--32.

[20] Breiman L I, Friedman J H, Olshen R A, et al. Classification and Regression Trees (CART)[J]. Lecture Notes in Computer Science, 1984, 81(1):17–23.

[21] Martinez-Munoz G, Suarez A. Pruning in Ordered Bagging Ensembles[C]. Proceeding of Conference of Machine Learning. 2006: 8-25.

[22] Martinez-Munoz G, Suarez A. Aggregation Ordering in Bagging[C]. Proceeding of Conference of Artificial Intelligence and Applications. 2004: 6-18.

[23] G.Martinez-Munoz, Suarez A. Using Boosting to Prune Baaging Ensembles [J]. Pattern Recongition Letters, 2007, 28(1):10-23.